Machine Learning Techniques applied in risk assessment related to food safety

IZSTO[*]
G. Ru
M.I. Crescio
F. Ingravalle
C. Maurella

UBESP[†]
D. Gregori
C. Lanera
D. Azzolina
G. Lorenzoni
N. Soriani
S. Zec

DSCB[‡]
P. Berchialla
S. Mercadante

ZETA[§]
F. Zobec
M. Ghidina
S. Baldas
B. Bonifacio
A. Kinkopf
D. Kozina
L. Nicolandi
L. Rosat

Approved 29 May, 2017
ver: 1.5.0

Abstract

In 2014 European Food Safety Authority (EFSA) commissioned this evaluation of the potential use of Machine Learning Techniques (MLTs) to provide insights for the elaboration of a guidance document and to facilitate the harmonisation in EFSA's assessments. Four objectives were provided: 1. To produce an inventory of MLTs that could be of use in the EFSA risk assessment activities; 2. To carry out a classification of EFSA opinions to identify the questions most commonly asked; 3. To assess the performance of ML techniques compared to non-MLTs and to propose a decision tree to help in the choice of the most appropriate methodology; 4. To develop, if possible, machine learning algorithms tailored to answer EFSA specific questions. The extensive literature search on 22 online databases led to an inventory of more than 2.6 million MLTs references: 213,070 abstracts were classified as relevant for EFSA and labelled by applying a Support Vector Machine and a Name co-Occurrences analysis. The application of Latent Dirichlet Allocation and Correlated Topic Modeling to the text of 3,744 EFSA scientific documents allowed the description of 28 main topics characterising the overall activity of assessment carried out by EFSA. Moreover the most common statistical techniques applied in EFSA to address the topics have been identified by text mining and by a questionnaire survey that involved 49 EFSA staff. Six different examples were used to show and compare the different performances of MLTs and non-MLTs techniques: this activity served to develop a decision tree that on the basis of a set of predefined criteria provides a guideline for the selection of fit for purpose MLT. Finally to better address some specific issues, data from the European Union Summary Reports on Zoonoses and on Antimicrobial Resistance were used to develop case studies where existing MLTs were expressly modified.

key words: Machine Learning, MLTs, Extensive Literature Search, Decision Tree, Risk Assessment, Food Safety, Case studies

Question number: EFSA-Q-2014-00467
Correspondence: amu@efsa.europa.eu

[*] Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle D'Aosta

[†] Unità di Biostatistica, Epidemiologia e Sanità Pubblica del Dipartimento di Scienze Cardiologiche, Toraciche e Vascolari dell'Università degli Studi di Padova

[‡] Dipartimento di Scienze Cliniche e Biologiche dell'Università degli Studi di Torino

[§] Zeta Research s.r.l., Trieste

Disclaimer: The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

Suggested citation: Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle D'Aosta; Unità di Biostatistica, Epidemiologia e Sanità Pubblica del Dipartimento di Scienze Cardiologiche, Toraciche e Vascolari dell'Università degli Studi di Padova; Dipartimento di Scienze Cliniche e Biologiche dell'Università degli Studi di Torino; Zeta Research s.r.l., Trieste, 2017. EFSA supporting publication 2017:EN-1254. 311 pp.

doi:10.2903/sp.efsa.2017.EN-1254

Summary

MLT deals with the study, the design and the development of algorithms that give computers the ability to learn from experience without being explicitly programmed. In the last two decades, Machine Learning (ML) has become increasingly important with the aim of automatically learning from data, particularly in situations where large collections of data or large, multi-dimensional and heterogeneous datasets are available and/or there is not a recommended mathematical approach. These techniques are particularly suitable in the field of Risk Assessment, in which classification problems are commonly encountered: e.g. recently clustering methods have been explored to extract knowledge and identify patterns in the field of antimicrobial resistance.

In 2014 EFSA commissioned this scientifically-based evaluation of the potential use of MLT to provide insights for the elaboration of a guidance document and to facilitate the harmonisation in EFSA's assessments. The evaluation had four main objectives:

- to produce an inventory and a proposal of classification of available MLTs that could be of use in the EFSA risk assessment activities;
- to carry out a classification of EFSA risk assessment published opinions to identify categories of questions most commonly asked to EFSA within its remit;
- to assess the performance of ML techniques compared to non-ML techniques when applied to those risk questions and to propose a decision tree to help in the choice of the most appropriate methodology;
- to develop, if possible, new machine learning algorithms tailored to answer some specific issues.

Regarding the first objective, an Extensive Literature Search was carried out on 22 Data Bases and Search Engines for literature relevant to MLT. Selected resources were: Arχiv, Association for Computing Machinery, CiteseerX, Cochrane Library, Cumulative Index to Nursing and Allied Health Literature, Current Index of Statistics, Directory of Open Access Journal, EconLit, IEEE Xplore Digital Library, Ingenta Connect, JSTOR, MathSciNet, Medical Literature Analysis and Retrieval System Online, PsycINFO, PubMed, Research Papers in Economics, ScienceDirect, Scopus, Web of Science Arts&Humanities Citation Index, Web of Science Core Collection, Web of Science Citation Index and Web of Science Social Science Citation Index.

More than 2.6 million references were retrieved and the inventory has been stored in a My-SQL database, which is available through an ad hoc web interface (WEBi) connecting at dedicated web site (mlt-webi.zetafield.eu). Focusing on papers with abstracts in English and with at least 700 characters, overall, about 1.65 million were classified as relevant to MLT field using a SVM classifier. To facilitate the navigation into the WEBi, a further Name co-Occurrences (NO) analysis was carried out with the aim of labelling abstracts, according to some methodological and application-related MLT aspects. As a result, 213,070 relevant abstracts were labelled.

To address the second objective, 3,744 EFSA scientific documents (mainly Scientific Opinions) were retrieved. The parallel application of two topic modeling techniques (i.e. Latent Dirichlet Allocation and Correlated Topic Modeling) allowed the identification of 28 main themes and relative issues characterising the overall activity of assessment carried out by EFSA. Then the most common statistical techniques applied in EFSA to address the risk questions associated to those themes have been identified by both text mining of the mentioned EFSA documents using an ad hoc developed vocabulary of statistical techniques and carrying out a questionnaire survey that involved 49 EFSA staff.

Focusing on the risk questions emerged from the classification of EFSA published opinions, performance of MLT compared to non-MLT has been investigated in terms of both reliability and robustness of the outcomes. Based on predefined criteria, the assessment included also pros and cons. A range of the main MLT approaches were described in details highlighting properties and limitations and the evaluation was carried out investigating six exploratory case studies.

The exploratory case studies allowed to derive a taxonomy for MLT useful for the development of a decision tree/recipe book task oriented, i.e. from the problem to the approach, which provides a guideline in the choice of the most appropriate methodology. Whereas any Non-MLT approach relies on data modelling, MLT rely on the predictive accuracy of models.

The decision tree includes the main parameters that have to be taken into account by a MLT user to choose the proper data mining technique in a real application; such parameters are: (i) the main goal of the problem to

be solved (supervised/unsupervised problem), (ii) the structure of the data (inputs and outputs characteristics, linearity, scalability, sample size, sparsity, dimensionality), (iii) the difference between MLT that rely on the predictive accuracy of models, and Non-MLT approach that relies on data modelling (stability, robustness).

Finally to better address EFSA specific questions, data from the European Union Summary Reports on Zoonoses and on Antimicrobial Resistance were used to develop case studies to illustrate the potential for the use of MLT on addressing biological hazards. The first case study was developed to provide an automated procedure, which could be potentially embedded in data quality assurance processes. The second case study aimed at providing a fit for purpose technique when the detection of epidemiological latent patterns is of interest. Three further case studies are focusing on food borne outbreaks and antimicrobial resistance. A case study was based on food borne outbreaks data and aimed at illustrating the potential use of MLT for exploring patterns related to food borne outbreaks severity and developing a predictive model for the risk of hospitalization. A case study focused on antimicrobial resistance data was aimed at illustrating the use of MLT for monitoring purposes and for understanding the relationship between prevalence of zoonoses and antimicrobial resistance. A final case study on antimicrobial resistance data was developed to monitor and describe similarities in zoonotic agents.

Finally, to adapt MLT to some specific issues, like small sample size, the need of generalization of the results, modifications to standard MLT were proposed and illustrated into two ad hoc case studies.

Table of contents

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

List of Figures

List of Tables

Introduction

This contract was awarded by EFSA to:

- Contractor: a consortium made up of four institutions i.e. Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle d'Aosta (IZSTO), Italy, Project Leader; Unità di Biostatistica, Epidemiologia e Sanità Pubblica (UBESP), del Dipartimento di Scienze Cardiologiche, Toraciche e Vascolari (DSCTV), University of Padua, Italy; Dipartimento di Scienze Cliniche e Biologiche (DSCB), Università di Torino, Italy; Zeta Research s.r.l. (ZETA), Trieste, Italy. The units participating in the consortium have a history of long-standing cooperation in the areas of the present call, with the proven capability of team-working in complex and long-lasting projects related in various ways to the matter of the current call. In particular with MLT, RA, Risk analysis education and dissemination, R software programming, extensive literature search (ELS) and Meta-analysis.
- contract title: Machine Learning techniques applied in risk assessment related to food safety;
- contract number: OC/EFSA/AMU/2014/02.

Background and objectives as provided by EFSA for this procurement procedure.

EFSA's role is to assess and communicate on all risks associated with the food chain.. Much of the work of this organization is to respond to specific scientific advice, undertaking also works developed on its own initiative when needed. The goal is to give always a coherent, accurate and timely answer on the scope of food safety problems.

The scientific advice made by EFSA concern the areas of food and feed safety, nutrition, animal health and welfare, plant protection and plant health. Some panels have been created with the specific fields of action to manage this amount of workload. Independent scientific advice are carried out by the following EFSA's Panels:

- Panel on Animal Health and Animal Welfare (AHAW);
- Panel on Biological Hazards (BIOHAZ);
- Panel on CONTAM;
- Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids (CEF);
- Panel on PLH;
- Panel on Additives and Products or Substances used in Animal Feed (FEEDAP);
- Panel on Dietetic Products, Nutrition and Allergies (NDA);
- Panel on ANS;
- Panel on Genetically Modified Organisms (GMO);
- Panel on Plant Protection Products and their Residues (PPR).

When performing RA some of the most frequent questions EFSA has to address are essentially fitting a limited set of typologies:

1. Questions that deal with the identification of factors that can modify a given feature;
2. Questions that deal with classification issue: EFSA is frequently asked to classify a given unit of interest into positive or negative according to its risk of being a case. Some formal questions could have the following formulation: "is probiotic X effective?" (diagnosis) or it is also quite common the need of answering to questions like: "will the threshold be crossed by pesticide P in 6 month from now?". In each case, a dichotomous yes/no decision has to be made;

3. Questions that deal with Risk Prediction: in this case the outcome is not positive or negative anymore. EFSA is usually asked to address questions like "What is the probability that compound C is toxic?".

Machine Learning deals with the study, the design and the development of algorithms that give computers the capability to learn without being explicitly programmed. Therefore these techniques are particularly suitable in the field of risk assessment to extract knowledge and identify patterns in situations where large collections of data or large, multi-dimensional and heterogeneous datasets are available and/or there is not a recommended mathematical approach.

Various research groups / developers in the field of ML have attributed a different level of attention to these two latter types of questions (classification and risk prediction) and have proposed different ways of tackling the related problems. As an example, classification is dealt mainly using non-parametric approaches by the ML community, but also parametric methods have been developed. On the other hand, estimation of probabilities is generally approached by statisticians using parametric methods, such as the logistic regression model.

Probability estimation at individual (unit of interest) level has a long-standing tradition in biostatistics as it provides more detailed information than a simple yes/no answer. For this reason, applications can be found in all areas related to food safety. Since in the biostatistical community the term risk prediction is used with reference to therapies (and thus to investigate on treatment response probabilities or side effects probabilities), the more general term of probability estimation will be used. It is important to emphasize that neither classification nor probability estimation automatically follow from association results.

Predictions models are widely applicable to the type of questions that EFSA has to address and many units make use of them. Nevertheless, classification problems are also commonly encountered when dealing with RA and, recently, clustering methods have been explored in the area of antimicrobial resistance. Of course, these methods are not exclusively used in this field: they could be potentially used, e.g., also when dealing with animal and plant health issues, biohazards, etc.

The aim of the procurement procedure is to explore other techniques that could be of use in EFSA when dealing with such problems. In particular EFSA commissioned this scientifically-based evaluation of the potential use of MLT to provide insights for the elaboration of a guidance document and to facilitate the harmonisation in EFSA's assessments.

The objectives of the contract as provided by EFSA for this procurement procedure were as follows:

Objective 1. To produce an Inventory of available MLT that could be of use in risk assessment.

The inventory must be produced by performing an Extensive Literature Search (ELS) followed by a screening for relevance process. In this phase there is no need to restrict the field to topics related to food safety: the focus must be on the methodology (e.g. underpinning statistical approach) itself, regardless of the specific field where the methodology has been applied. Morover a proposal for criteria for clustering the so-identified MLT must be proposed in a way that is useful for EFSA. Within the report, this objective is addressed by Part I.

Objective 2. To carry out a classification of EFSA risk assessment published opinions to identify the category of questions most commonly asked to EFSA within its remit.

Part II of this report deals with this objective.

Objective 3. For each identified clusters of risk questions, to assess the performance of ML techniques compared to non-ML techniques, for each cluster of risk questions, in terms of reliability and robustness of the outcomes. In particular the report must to include: i) the outcome of the matching exercise; ii) a full description and a summary table of the pros and cons assessment (comparing ―classical‖ approaches with MLT techniques) for each cluster of risk questions; iii) a decision tree/recipe book ―from the problem to the approach‖ - to help in the choice of the most appropriate methodology.

Objective 4. For clusters of risk questions where no appropriate MLT methods are already available, to explore the possibility (for a maximum of 2 clusters) of developing new machine learning algorithms tailored to answer those specific questions.

The last two objectives are addressed by the rest of the report. Six different examples of possible risk assessment exercises were used to show and compare the different performances of MLT and non-MLT techniques. Then a range of the main MLT approaches were described in details highlighting properties and limitations. This activity served to develop a decision tree that on the basis of a set of predefined criteria (such as supervision, inputs and outputs characteristics, linearity, scalability, sample size, stability, sparsity, dimensionality, robustness) provides a guideline for the selection of fit for purpose MLT. Finally to better address EFSA specific questions, data from the European Union Summary Reports on Zoonoses and on Antimicrobial Resistance were used to develop case studies where existing MLT were expressly modified in a way that may be useful for EFSA.

1 Extensive literature search on Machine Learning Technique

1.1 Introduction

MLTs are particularly suitable in the field of risk assessment to extract knowledge and identify patterns in situations where large collections of data or large, multi-dimensional and heterogeneous datasets are available and/or there is not a recommended mathematical approach.
Here we presents a detailed description of the activity performed for the identification/application of an appropriate ELS strategy, screening of ELS results and finally identification of criteria for classifying the identified MLT and carry out the task. To be performed properly, ELS must be based on a large set of potential candidates. In this context, ELS on MLT must be based on searching both for methods strictly adhering to the MLT world and for techniques belonging to different backgrounds that can be effectively used for the purposes of risk analysis and risk classification.
As a general preliminary consideration, having taken into account our interpretation of EFSA needs, in performing ELS priority has been assigned :

- to sensitivity (i.e. probability that a relevant paper is correctly classified as relevant) with respect to specificity (i.e. probability that a non-relevant paper is correctly classified as non-relevant) (Hirschman et al., 2002), on the basis that, in principle, the price for EFSA of losing important information would be higher than the price due to a slightly inefficiency in retrieving some unnecessary results.
- to the perspective of a practitioner needing concrete results with respect to that of a pure mathematician or theoretical bio-informatics expert.

1.1.1 Aim

The present chapter aims at addressing the issue of finding an appropriate and extensive class of documents each of which is classified in a suitable way for identification and classification of MLT that may be potentially useful for EFSA.

1.2 Methods

After the agreement between the Consortium and EFSA, DBs and Search Enginess (SEs) (from now on referred to as resources) were defined (see sec. 1.2.1) as well as appropriate SSs, for the retrieval process of the BCs useful for the objective 1 (see sec. 1.2.2).

Available resources with their characteristics and their limitations were analyzed too (see sec. 1.2.3). The retrieval of all the identified BCs was conducted (see sec. 1.2.4) in order to import them in a single DB and proceed with the identification of duplicates. Record identified as no-duplicated BCs were then exported with the purpose of conducting an analysis of pertinence of the results obtained with:

- an Active Learning (AL) algorithm;
- a particular type of MLT.

In the meanwhile, a DB validation was carried out (see sec. 1.2.5).

After that, using an algorithm based on Name co–Occurences (NO), a text mining technique, a first classification of the BCs was done according to topics specific to the field of ML (see sec. 1.2.6).

1.2.1 Information sources

In the first part of this section the selection criteria of the resources (adopted by the Consortium) that led to the identification of the final set of 22 resources used for the research are described. A brief description of the

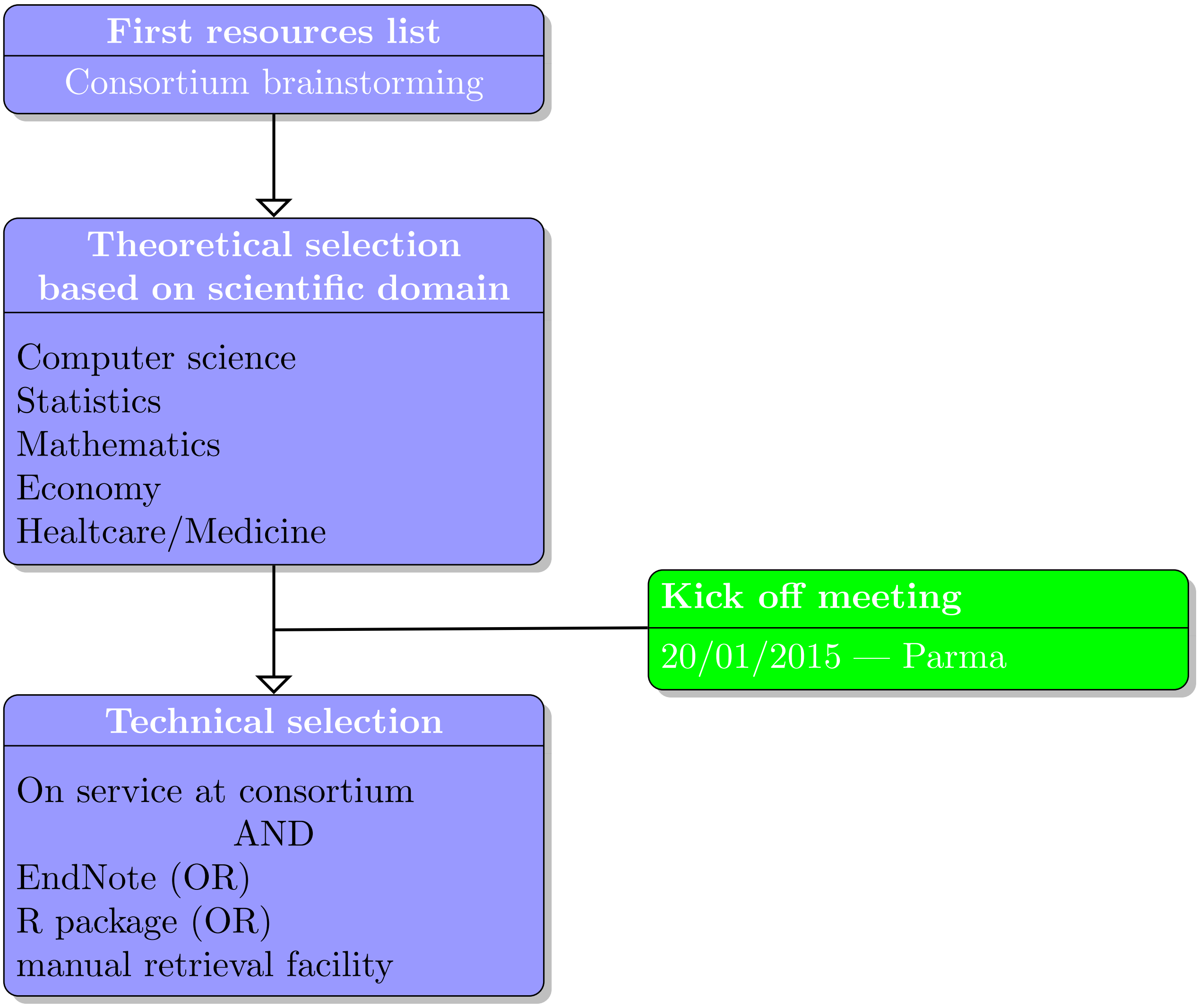


Figure 1: Flowchart of steps followed for the resources selection.

resources is provided in the second part of the section. An overview of tasks related to the approach adopted is provided in Figure 1.

Selected resources are: Arχiv, Association for Computing Machinery (ACM), CiteseerX, Cochrane Library, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Current Index of Statistics (CIS), Directory of Open Access Journal (DOAJ), EconLit, IEEE Xplore Digital Library, Ingenta Connect, JSTOR, MathSciNet, Medical Literature Analysis and Retrieval System Online (MEDLINE), Psycinfo, PubMed, Research Papers in Economics (REPEC), ScienceDirect, Scopus, Web of Science Arts & Humanities Citation Index (WOS-AHCI), Web of Science Core Collection (WOS-CORE), Web of Science Core Collection (WOS-CORE), Web of Science Social Science Citation Index (WOS-SSCI).

Criteria for selection/inclusion of datasets

In December 2014, the first step (aimed at selecting resources) was the creation of a list of academic DBs and SEs, deriving from the experience gained by the consortium members in the field of meta-analysis.

In the call in question, EFSA required the use of the software EndNote, for the bibliographic management, and R, for the part of programming. Therefore, the next step was to identify, within the identified group of resources, DBs and SEs concerning the following areas: computer science, statistics, economy, biomedicine, mathematics or multidisciplinary.

After this initial selection, the list was submitted to EFSA during the kick-off meeting of the project (January 20th, 2015). On that occasion, the need for a second screening was expressed (as there was the possibility to retrieve an additional considerable amount of BCs).

Considering, therefore, the properties of the different DBs and in order to achieve the maximum benefit from each one of them, the resources where further screened for possessing *at least one* of the following parameters, in addition to the basic and obvious one of being accessible by the Consortium:

- Possibility to connect by EndNote;
- Availability of one *ad hoc* R package (e.g. `aRxiv`) or the existence of a bijective correspondence from each BC and its url page showing Meta-Data;
- Availability of downloadable results after querying the resource's web interface.

The final list of 22 resources has been approved by EFSA during the web meeting held on January 26th, 2015. A flowchart representing the steps followed to decide how to perform the search on each resources is provided in Figure 2. Results can be found in Table 1.

After a first analysis on the 22 selected resources (web searching for "machine learning OR artificial intelligence" and considering the order of magnitude af the results), it was expected about one and a half million references. Therefore, a decision has been initially taken to start the retrieval process by using the aforementioned software. For this purpose 9 resources for which EndNote admitted a connection to the recovery of BCs were identified. Subsequently, among the remaining resources, it has been explored the possibility to write R scripts to retrieve the related BCs (this include the cases for which it was possible to use existing R packages (ArXiv, etc...) and the cases for which it is necessary to write *ad hoc* R-code), identifying 7 additional resources. After that, being still present 6 resources for which it was not possible to find an EndNote connection nor to approach the problem with an R script, a person belonging to the consortium (see sec. 1.2.3) has been dedicated for performing the manual retrieval of the BCs.

Table 1: Resources division by retrieval type.

EndNote	**R**[a]	**Manually retrieval**
cinahl	acm[b]	Cochrane
EconLit	Arχiv[b]	IEEE Xplore
medline	cis	MathSciNet
PsycInfo	CiteseerX	Scopus
PubMed	doaj[b]	Science Direct
WoSahci	Ingenta	JSTOR
WoScore	RePEc	
WoSsci		
WoSssci		

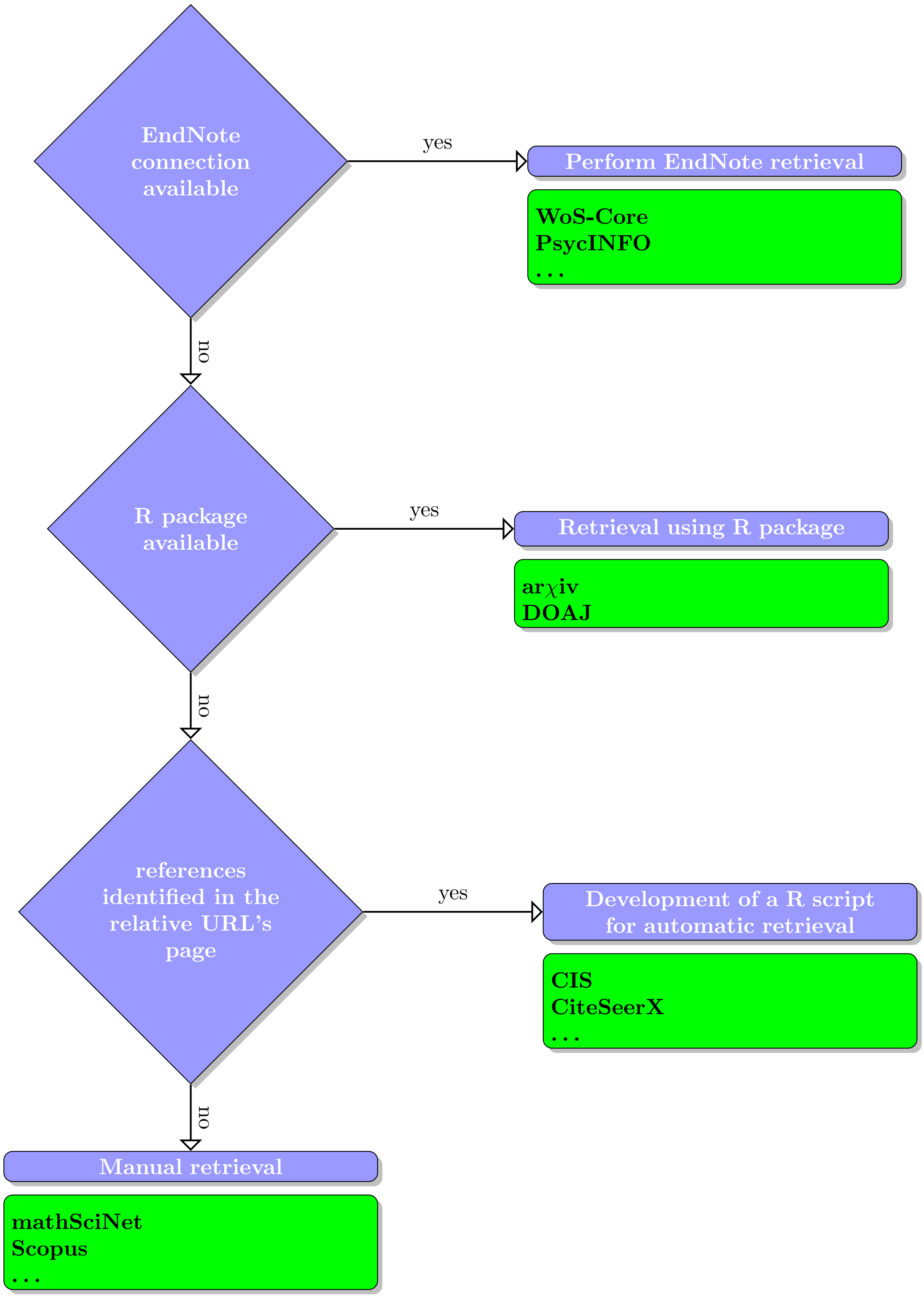
[a] Development R script for automatic retrieve.

[b] Using R package directly connected to the resource.

Description of datasets

Arχiv: in August 1991 a central repository mailbox was created and stored at the Los Alamos National Laboratory which could be accessed from any computer. It is a highly-automated electronic archive and distribution server for research articles. Covered areas include physics, mathematics, computer science, nonlinear sciences, quantitative biology and statistics. Arχiv is maintained and operated by the Cornell University Library. It is an openly accessible, moderated repository of scientific papers in the fields of mathematics, physics, astronomy, computer science, quantitative biology, statistics, and quantitative finance, which can be accessed online. By the end of 2014 hit a million article milestone. Website: `www.arXiv.org`

Association for Computing Machinery (ACM): originally it was established as the Eastern Association for Computing Machinery at a meeting at Columbia University in New York on 1947. It is a not-for-profit professional membership group and actually is the world's largest educational and scientific society, combining computing educators, researchers and professionals. The Association for Computing Machinery


EndNote connection available
yes
Perform EndNote retrieval
WoS-Core
PsycINFO
...
no
R package available
yes
Retrieval using R package
arχiv
DOAJ
no
references identified in the relative URL's page
yes
Development of a R script for automatic retrieval
CIS
CiteSeerX
...
no
Manual retrieval
mathSciNet
Scopus
...

· · · : see table 1.

Figure 2: Procedure for the classification of the resources.

(ACM) Digital Library contains a comprehensive archive, starting from '50s, of the organization's journals, magazines, newsletters and conference proceedings. All metadata in the Digital Library is open to the world, including abstracts, linked references and citing works, citation and usage statistics, as well as all functionality and services. Website: www.acm.org

CiteSeerX: it was developed in 1997 at the NEC Research Institute, Princeton, USA. It became public in 1998 and the service transitioned to the Pennsylvania State University's College of Information Sciences and Technology in 2003. It is an evolving scientific literature digital library and search engine that has focused primarily on the literature in computer and information science. It is often considered to be the first automated citation indexing system and was considered a predecessor of academic search tools. CiteSeer freely provided Open Archives Initiative metadata of all indexed documents and links indexed documents when possible to other sources of metadata such as Digital Bibliography & Library Project (DBLP) and the ACM Portal. Website: citeseerx.ist.psu.edu/

Cochrane Library: The Cochrane Database of Systematic Reviews was published in 1988. Actually it is a collection of six databases (Cochrane Reviews, Database of Abstracts of Reviews of Effects (DARE), Cochrane Central Register of Controlled Trials (CENTRAL), Methodology Register, Health Technology Assessment (HTA), National Health Service - Economic Evaluation Database (NHS EED)) in medicine and other healthcare specialties provided by the Cochrane Collaboration and other organizations. Cochrane researchers perform searches of medical DB including Medical Literature Analysis and Retrieval System Online (MEDLINE), PubMed and Excerpta Medica dataBASE (EMBASE). It is located at Cardiff University. Website: www.cochranelibrary.com

Cumulative Index to Nursing and Allied Health Literature (CINAHL): it is an index of journal articles about nursing, allied health, biomedicine and healthcare. The index was first published as Cumulative Index to Nursing Literature (CINL) in 1961. The title changed to Cumulative Index to Nursing and Allied Health Literature (CINAHL) in 1977 and first went online in 1984. This research database provides full text for over 700 nursing and allied health journals indexed in the CINAHL database, and includes a higher number of records, additional journals, records dating back to 1937 and expanded content. Website: www.ebscohost.com/nursing/products/cinahl-databases/the-cinahl-database

Current Index of Statistics (CIS): the Current Index to Statistics is an online database published by the Institute of Mathematical Statistics and the American Statistical Association that contains bibliographic data of articles in statistics, probability, and related fields. The on-line Current Index of Statistics Extended Database indexes the entire contents of over 160 ”core journals”, in most cases from 1975 (or first issue if later) to the current end year, and pre-1975 coverage for some, and about 11 000 books in statistics published since 1975. Website: www.statindex.org

Directory of Open Access Journal (DOAJ): the Open Society Institute funded various open access related projects after the Budapest Open Access Initiative; the Directory was one of those projects. After the first Nordic Conference on Scholarly Communication in 2002, Lund University became the organization to set up and maintain the Directory of Open Access Journal (DOAJ). The aim of the DOAJ is to increase the visibility and ease of use of open access scientific and scholarly journals, thereby promoting their increased usage and impact. The database contains more then a million and a half articles and records for more then 10 000 journals. Website: www.doaj.org

EconLit: it is an academic literature abstracting database service published by the American Economic Association. EconLit has added indexed records for journal articles from 1886 to 1968. EconLit is available at libraries and on university Web sites throughout the world, licensed from information service providers, who provide search engines, links to libraries' full-text subscriptions, and other enhancements to assist users in document retrieval. It uses the Journal of Economic Literature (JEL) classification codes for classifying papers by subject. Website: www.aeaweb.org/econlit/

IEEE Xplore Digital Library: it is a scholarly research database that indexes, abstracts, and provides full-text for articles and papers on computer science, electrical engineering and electronics. The database mainly covers material from the Institute of Electrical and Electronics Engineers (IEEE) and the Institution of Engineering and Technology (IET). IEEE Xplore provides Web access to more than 3 million full-text documents from some of the most highly cited publications. The content in IEEE Xplore comprises over 160 journals, over 1 200 conference proceedings and more than 3 800 technical standards. Approximately 25 000 new documents are added to IEEE Xplore each month. Website: www.ieee.org/ieeexplore

Ingenta Connect: founded in May 1998, today Ingenta Connect hosts content from over 250 publishers, with an aggregated database of over 13 500 publications and over 4.5 million articles, including both journals and eBooks. The company is headquartered at Publishing Technology Plc in Oxford, UK. Publishing Technology is the largest supplier of technology and related services for the publishing industry. Working with eight of the ten largest publishers in the world. Website: www.ingentaconnect.com

JSTOR: short for Journal Storage, is a digital library founded in 1995. Originally containing digitized back issues of academic journals, it now also includes books and primary sources, and current issues of journals. It provides full text searches of almost 2 000 journals, in more than 50 disciplines. Most access is by subscription, but some older public domain content is freely available to anyone. The service does not offer full-text, although academics may request that from JSTOR, subject to a non-disclosure agreement. Website: www.jstor.org

MathSciNet: published by the American Mathematical Society (AMS), is an electronic publication offering access to a maintained and searchable database of reviews, abstracts and bibliographic information for much of the mathematical sciences literature. Over 100 000 new items are added each year, most of them classified according to the Mathematics Subject Classification. MathSciNet contains information on about 2 million articles from 1 900 mathematical journals. Bibliographic data from retrodigitized articles dates back to the early 1800s. Website: www.ams.org/mathscinet

Medical Literature Analysis and Retrieval System Online (MEDLINE): it was launched by the National Library of Medicine (NLM) in 1964. It is a bibliographic database of life sciences and biomedical information that includes information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE contains over 21 million references to journal articles in life sciences with a concentration on biomedicine and it is freely available on the Internet and searchable via PubMed. Website: http://www.nlm.nih.gov/pubs/factsheets/medline.html

Psycinfo: it is a database of abstracts of literature in the field of psychology. It is produced by the American Psychological Association (APA) and distributed on the association's APA Psycnet. Contained more than 3.7 million records, some dating back to 1887, and includes abstracts from Psychological Abstracts back to 1927, Psychological Bulletin from 1921-1926, and all APA journals and the American Journal of Psychology (AJP) back to their first issues. Currently, there are 2 561 journals covered in the Psycinfo database. Website: http://www.apa.org/pubs/databases/psycinfo/

PubMed: first released in January 1996, ushered in the era of private, free, home- and office-based MEDLINE searching. The PubMed system was offered free to the public in June 1997. It comprises more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books. Has over 24.6 million records going back to 1966, selectively to the year 1865, and very selectively to 1809; about 500,000 new records are added each year. As of the same date, 13.1 million of PubMed's records are listed with their abstracts, and 14.2 million articles have links to full-text. Website: http://www.ncbi.nlm.nih.gov/pubmed

Research Papers in Economics (REPEC): it is a decentralized database of working papers, preprints, journal articles, and software components. The project started in 1997 with the aims to enhance the dissemination of research in economics. Sponsored by the Research Division of the Federal Reserve Bank of St. Louis and using its IDEAS database, Research Papers in Economics (REPEC) provides links to over 1 200 000 full text articles. Most contributions are freely downloadable, but copyright remains with the author or copyright holder. Website: http://repec.org/

ScienceDirect: it is a leading full-text scientific database offering journal articles and book chapters from nearly 2 500 journals, 13.3 million articles and more than 30 000 books. The journals are grouped into four main sections: Physical Sciences and Engineering, Life Sciences, Health Sciences, and Social Sciences and Humanities. For most articles abstracts are freely available; access to the full text generally requires a subscription or pay-per-view purchase. Website: www.sciencedirect.com

Scopus: it is the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings. Delivering a overview of the research output in the fields of science, technology, medicine, social sciences, and arts and humanities. This database provides nearly 55 million records, 21 915 titles and 5 000 publishers, records dating back to 1966. Scopus can be integrated with Open Researcher and Contributor ID (ORCID). Website: www.scopus.com

Web of Science Arts & Humanities Citation Index (WOS-AHCI): it was originally developed by the Institute for Scientific Information. It is a citation index, with abstracting and indexing for more than 1 700 arts and humanities journals, and coverage of disciplines that includes social and natural science journals. Part of this database is derived from Current Contents records. This database, and the following 3 databases, can be accessed online through Web of Science.

Website: http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-sea arts-humanities-citation-index.html

Web of Science Core Collection (WOS-CORE): multidisciplinary content covers over 12 000 of the highest impact journals worldwide, including Open Access journals and over 150 000 conference proceedings. Is possible to find current and retrospective coverage in the sciences, social sciences, arts, and humanities, across more than 250 disciplines. It provide nearly 2.6 million records and backfiles dating back to 1898.

Website: http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-sea web-of-science-core-collection.html

Web of Science Science Citation Index (WOS-SCI): it is a citation index originally produced by the Institute for Scientific Information (ISI) and it was officially launched in 1964. The larger version (Science Citation Index Expanded) covers more than 6 500 notable and significant journals, across 150 disciplines, from 1900 to the present. The index is made available online through different platforms, such as the Web of Science and SciSearch. Exist also markets several subsets of this database such as the Neuroscience Citation Index and the Chemistry Citation Index.

Website:http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-sea science-citation-index-expanded.html

Web of Science Social Science Citation Index (WOS-SSCI): it is an interdisciplinary citation index. Like the 3 latter databases, it was product of Thomson Reuters' Healthcare & Science division and it was developed by the ISI from the Science Citation Index. This citation database covers some 2 474 of the world's leading journals of social sciences across more than 50 disciplines. It is made available online through the Web of Science service for a fee.

Website:http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-sea social-sciences-citation-index.html

1.2.2 Eligibility criteria for the records

The SSs have been created using only the "or" logical operator between keywords. This was intended to simplifying the logistics in performing the search, allowing to adopt a unique SS across multiple resources, each one with his own characteristics. Indeed, using only or logic-connection allows (when needed) to search one keyword a time, without changing the global behavior of the search. Moreover, the results of the search on the second SS can be merged with the results of the first SS obtaining the same result as if the search had been conducted on a single SS containing all keywords. The final DB consists of not duplicate BCs, coming from each resources. The process and unification criteria are described in section 1.2.5.

During December 2014, the Consortium has drafted a proposal regarding the first SS starting from that made by Unità di Biostatistica, Epidemiologia e Sanità Pubblica (UBESP), realized thanks to the contribution and experience of its members. After that, a comparison with the other members of the Consortium has been carried out, adding, removing or modifying the terms, until the proposal was approved by each member of the Consortium.

This SS was presented during the kick off meeting of January 20th, 2015 and discussed and approved by "EFSA representatives" during the web meeting of January 26th, 2015, specifically dedicated to the definition of the SS. On this occasion EFSA required to conduct a research on an additional SS of cross validation, composed of keywords available in "Cran Task View: Machine Learning and Statistical Learning".

An overview of actions performed in relation to the definition of the SS is provided in Figure 3. Below are listed the keywords of the first SS agreed with EFSA, separated by a comma. It is worth noticing that the search was carried out by connecting the keywords with the function "or" and these keywords represent the inclusion criteria of all the records.

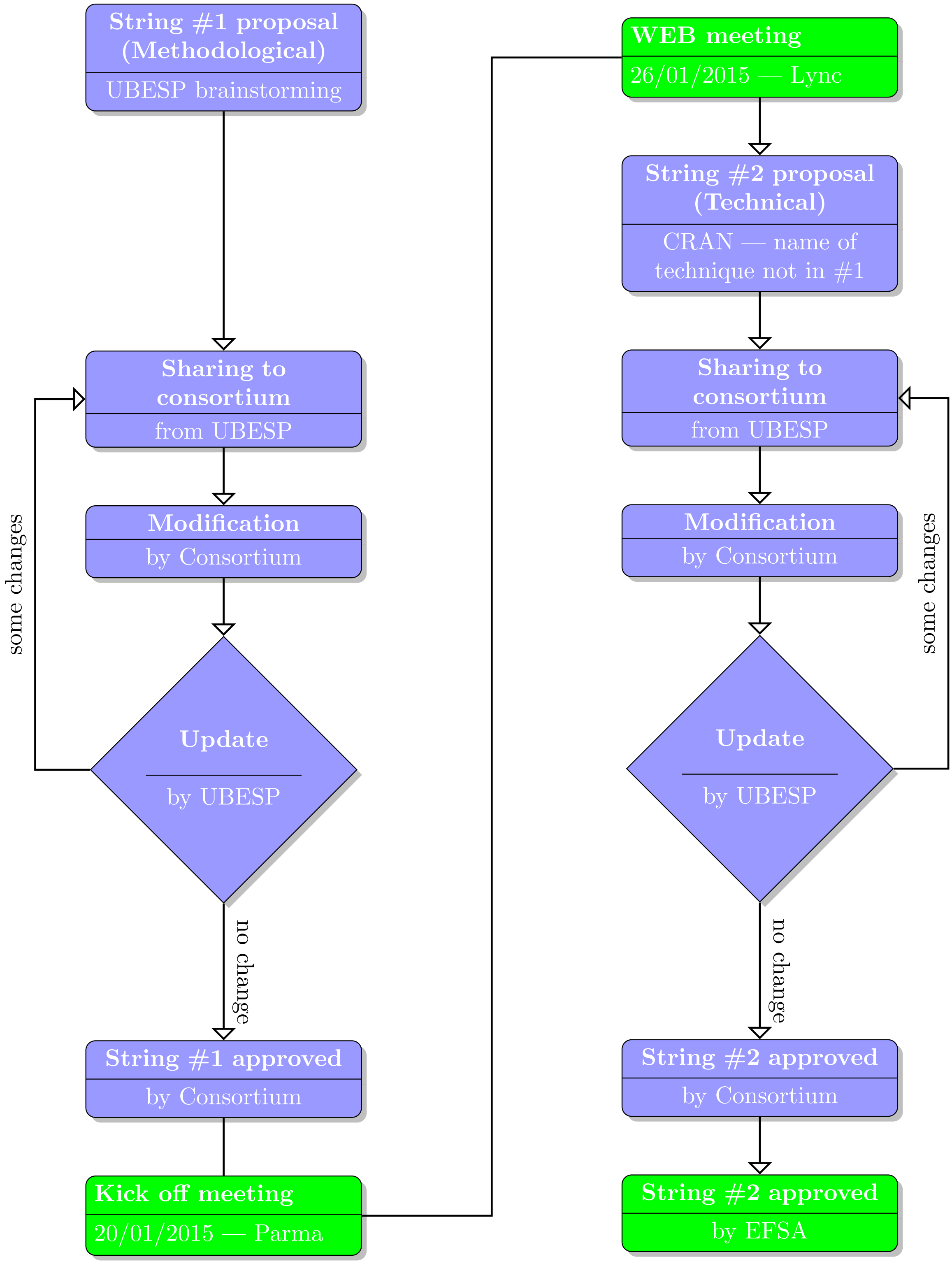


Figure 3: Procedure of activities to define the search strings.

Table 2: Main topics resources

	Computer science	Economy	Healtcare/Medicine	Math	Statistics
Arχiv	x			x	x
acm	x				x
CiteseerX	x			x	x
Cochrane			x		
cinahl			x		
cis					x
doaj	x	x	x	x	x
EconLit		x			
ieee Xplore	x				
Ingenta Connect	x	x	x	x	x
JSTOR	x	x	x	x	x
MathSciNet				x	
medline			x		
PsycINFO			x		
PubMed			x		
RePEc		x			
ScienceDirect	x	x	x	x	x
Scopus	x	x	x	x	x
WoS-ahci			x		x
WoScore	x				x
WoSsci	x				x
WoSssci			x		x

x: topic mainly treated.

String #1 (main): artificial intelligence, bayes, belief network(s), classification algorithm(s), classifier(s), data mining, kernel estimation, machine learning, neural network(s), instance based method(s), k-nearest neighbor, learning vector quantization, self organizing map(s), decision tree(s), classification tree(s), iterative dichotomizer(s), iterative dichotomiser(s), chi-squared automatic interaction(s), random forest(s), gradient boosting machine(s), kernel method(s), support vector machine(s), clustering method(s), k-means, association rule learning, a-priori algorithm(s), eclat algorithm(s), naive bayes, bayesian network(s), bayesian belief network(s), hidden markov model(s), perceptron(s), back propagation, hopfield network(s), ensemble method(s), deep learning, restricted boltzmann machine(s), convolutional network(s).

With regard to the second SS, keywords from the text of reference were identified. Then, all those already present in the first SS were eliminated. Below are reported the keywords of the second SS, also sought using the function ”or”:

String #2 (cross validation): recursive partitioning, rule-based model(s), logic regression(s), logic forest(s), regularized method(s), shrinkage method(s), lasso, elastic-net regularization path(s), ridge penalized regression model(s), shrinkage path(s), semiparametric additive hazards model(s), high-throughput ridge regression(s), heteroskedastic effects model(s), kernel learning, bayesian additive regression tree(s), BART, genetic algorithm(s), memetic algorithm(s), fuzzy rule-based system(s), rough set theory, fuzzy rough set theory.

With the definition of the above-mentioned SS started the search of the BCs. The use of the first string, along with the search results, helped in providing a concrete estimation of the overall workload related to computational time and algorithm development.

The amount of obtained results has required a re-planning of work, organizing the search on a cluster of multiple computers, with the participation of more people and the preparation of a dedicated mySQL DB (see sec. 1.2.3). Furthermore, the amount of results for the first SS has highlighted the impossibility to repeat the process for the second string and stay on what foreseen in the project workflow.

For this reason, the resources that needed a manual retrieval and those considered less relevant were removed from our list. The criterion of relevance for the second SS was the ”technicality” of the resource. All resources handled manually plus Web of Science Social Science Citation Index (WOS-SSCI), Web of Science Arts & Humanities Citation Index (WOS-AHCI) and CINAHL were excluded because the resources are not suitable to perform a search based on the names of the algorithms used in the field of ML.

The process of selection of the second SS has therefore narrowed the search to the resources reported in Table 3.

Table 3: Resources division for retrieval of second SS.

EndNote	R[a]	Manually retrieval
	acm[b]	
EconLit	Arχiv[b]	
medline	cis	
PsycInfo	CiteseerX	
PubMed	doaj[b]	
	Ingenta	
WoScore	RePEc	
WoSsci		

[a] Development R script for automatic retrieve.

[b] Using R package directly connected to the resource.

For what concerns the exclusion criteria of the research, steps were taken to instruct a SVM, with the aim of obtaining a score of relevance (see sec. 1.2.5).

Finally, Consortium experts on ML have selected nine books, relevant to the context, and widely used in graduate, post-graduate courses and research settings in MLT, in order to be able to proceed with the validation of the DB. From a selection of BCs drawn from these books it was therefore created the validation set (sec. sec. 1.2.5).

1.2.3 Technology Specifications

Software specifications

EndNote For the bibliographic management, EFSA clearly requested the use of EndNote software. EndNote is produced by Thomson Reuters and when it was introduced in the 1989, its only competitor on the Macintosh platform was a program called "REF52." For years it was an application only for Macintosh, but in 1995, it was expanded to Windows machines as well.

EndNote is one of the family programs which are used to manage bibliography references and quotations of documents. It is a database manager where the data (secondary, Meta-Data) defines the documents (e.g. books, articles, contribution to conferences, websites, ...).

The family of this software is called personal reference, or citation, or Bibliography management system. Therefore EndNote is part of a family more restricted than of the Database Management System (DBMS), because in particular it processes data and bibliography functions. The last version available for Macintosh and Windows is EndNote X7.3 , released April 1st, 2015. The last version utilized by the consortium is X7.3 from when it was released, before it was used the previous version X7.2

The main characteristics of EndNote x7.3 utilized in the project are:

- Possibility to execute automatic searches on some resource that the software had the connections (see tab. 1).
- Ability to subdivide into groups the BCs inserted in a library. Thanks to this option is it possible to divide the results from consecutive researches (see sec. 1.2.4).
- Ability to import files in Research Information Systems (RIS) e BibTeX (BIB) format. Thanks to this ability it has been able to exploit the manual downloading of BCs from resources untreatable in another way, but being able to import them into EndNote and to treat them as if they were downloaded from this software (see sec. 1.2.4).

- Ability to export the data in a personalized way by building *ad hoc* data that EndNote called export style. In this way, it has been able to build an export format that avoided some problems (see sec. 1.2.5).

In addition to this, the criteria considered in EndNote to identify duplicates were used to set up a My Structured Query Language (MYSQL) procedure (see sec. 1.2.5) used to identify the duplicate BCs. These criteria have been exploited to search for the duplicate BCs evenly for every single resource. (see sec. 1.2.5).

A number of limitations or "bugs" of EndNote are described below:

- In EndNote it is possible to insert a maximum of 10 search phrasal keywords for each search. Considered that keywords in the SSs used are all connected by or operator, they are grouped by 10 as described in section 1.2.4.
- It was found in some engines the presence of a bug in the connection file when the $ character is in the text. This problem was identified thanks to the control procedure "check" described in sec. 1.2.5. The fields identify as problematic were corrected manually, verifying on the interested resource.
- The import of the file (RIS or BIB) is limited up to 100 MiB. For this reason the files to import were combined up to the limit, before importing them individually in EndNote as described in the sec. 1.2.4.
- In the case where the first character of the field was a punctuation mark (e.g. `.23` in the number of volume), the software deleted the last character of the End of Field (EOF). The two adjacent fields to the damaged EOF become part of only one field. A check of fields allowed to identify the problem and the errors were manually corrected.
- A drawback of EndNote is its unability to manage millions of citations in a single library, in fact, the research on databases of that order of magnitude, it takes minutes, while the import and the export it takes hours (considering the machines utilized as described in the sec. 1.2.3). For this reason, we decided to prepare a DB MYSQL.

R The use of R programming language was specifically requested by EFSA.. R is a free software since it is distributed under the GNU is Not Unix (GNU) General Public License (GPL), and it is available for different operating system.

The language object–oriented derived directly from the S pack distributed with a non-open source and developed by John Chambers and others at the BellLaboratories. The R language was closely modeled on the S Language for Statistical Computing conceived by John Chambers, Rick Becker, Trevor Hastie, Allan Wilks and others at Bell Labs in the mid '70s, and made publically available in the early '80s.

The popularity of R is due to the wide availability of packages distributed with the GPL and organized in a special website called Comprehensive R Archive Network (CRAN), in analogy to Comprehensive TEXArchive Network (CTAN) and Comprehensive Perl Archive Network (CPAN). Those packages allows a broad extension of the ability of the program. The last available version is: R 3.2.0 dated April 16th, 2015.

The version utilized in the project was 3.1.2. then updated to 3.1.3. As R interfaces two Graphical User Interfaces (GUIs) have been used:

- R's GUI;
- RStudio (Multiplatform software, version 0.98).

For the developed scripts, we used the following packages, which we report within a brief description as it appears in the original documentation.

R package description

aRxiv: it is an interface to the API for arXiv, a repository of electronic preprints for computer science, mathematics, physics, quantitative biology, quantitative finance, and statistics. It has a MIT license. — ver: 0.5.10

Data.table: it's used for fast aggregation of large data (e.g. 100GB in RAM), fast ordered joins, fast add/modify/delete of columns by group using no copies at all, list columns and a fast file reader (fread). Offers a natural and flexible syntax, for faster development. It's licensed by GPL-2 and GPL-3. — ver: 1.9.4

Httr: useful tools for working with HTTP organized by HTTP verbs (GET(), POST(), etc). Configuration functions make it easy to control additional request components (authenticate(), add_headers() and so on). It's distributed under the MIT license. — ver: 0.6.1

OAIHarvester: Harvest metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) version 2.0. This package is licensed by GPL-2. — ver: 0.1-7

RCurl: is an R-interface to the libcurl library that provides HTTP facilities. This allows us to download files from Web servers, post forms, use HTTPS (the secure HTTP), use persistent connections, upload files, use binary content, handle redirects, password authentication, etc. This is distributed under the BSD license. — ver: 1.95-4.6

RSelenium: The RSelenium package provides a set of R bindings for the Selenium 2.0 WebDriver using the JsonWireProtocol. Selenium automates web browsers (commonly referred to as browsers). Using RSelenium you can automate browsers locally or remotely. It's licensed by AGPL-3. — ver: 1.3.5

RTextTools: is a machine learning package for automatic text classification. The package includes nine algorithms for ensemble classification (svm, slda, boosting, bagging, random forests, glmnet, decision trees, neural networks, maximum entropy), comprehensive analytics, and thorough documentation. It has a GPL-3 license. — ver: 1.4.2

Stringr: to use set of wrappers around the fantastic 'stringi' package. All function and argument names (and positions) are consistent, all functions deal with "NA"'s and zero length vectors in the same way, and the output from one function is easy to feed into the input of another. Also this package is licensed by GPL-2. — ver: 1.0.0

Textcat: Text categorization based on n-gram profile db for 26 languages based on the European Corpus Initiative Multilingual Corpus I. This is distributed under the GPL-2 license. — ver: 1.0-2

Tm: is a framework for text mining applications within R. Create content transformers, i.e., functions which modify the content of an R object. It's licensed by GPL-3. — ver: 0.6

XML: this package provides many approaches for both reading and creating XML (and HTML) documents (including DTDs), both local and accessible via HTTP or FTP. It also offers access to an XPath "interpreter". Also this package is licensed by BSD. — ver: 3.98-1.1

MySQL MYSQL is a Relational Database Management System (RDBMS) composed of a client and a server. Available for both systems Unix, Unix-like and Windows; the major reference platforms are Linux and Oracle Solaris. MYSQL is Free Software, released with a dual license, including the GNU GPL. The source code for MYSQL was originally owned by the company MYSQL AB, but it was released with the GNU GPL license as well as a commercial license. Up to the version 4.1 also a good part of the client code was released with the GNU Lesser GPL. From version 4.1 and so on, the whole client code is distributed with the license GNU GPL. The last available version is: MYSQL 5.6.19 released on February 2nd, 2015.

Computing and human resources

With regard to BCs retrieval, entering and analysis of data in a single MYSQL DB were quite challenging from the computational point of view and the amount of hardware resources needed. A strong cooperation between University of Study of Padua (UNIPD) and Zeta Research s.r.l. (ZETA) helped in lowering the impact of time needed to perform the requested tasks.

Retrieval For BCs retrieval, UBESP exploited the following hardware resources:

- Two virtual machines was running on a Dell PowerEdge PowerEdge 2950 physical machine with Debian 6 64 bit Intel(R) Xeon(R) CPU E5420 @ 2.50GHz:
  - virtualized machine running Debian Linux 6 64 bit We assigned to the virtual machine 4Gb of RAM DDR2 and 1 virtual processor Intel(R) Xeon(R) CPU E5420 @ 2.50GHz;
  - virtualized machine running Window 7 Professional 64 bit We assigned to the virtual machine 8Gb of RAM DDR2 and 1 virtual processor Intel(R) Xeon(R) CPU E5420 @ 2.50GHz.

- Eleven desktop remotely controlled:
  - OS : Windows 8.1 Pro 64bit;
  - CPU : Intel Core i5-650 Processor 3.2 GHz, 4M total cache 2 cores/4 threads Integrated Intel® HD Graphics Intel® CoreTM processor with vProTM technology Intel® Stable Image Platform Program (SIPP);
  - RAM : 4GB PC3-10600 MEMORY (2X2GB).
- Five desktop locally controlled:
  - OS : Windows 8.1 Pro 64bit;
  - CPU : Intel(R) Core(TM) i5-4570 CPU @ 3.2 GHz;
  - RAM : 8GB.
- A laptop locally controlled:
  - OS :Windows 8.1 Home 64bit
  - CPU : AMD A4-5000 with Radeon(TM) HDGraphics 1.5GHz
  - RAM : 6GB
- A laptop locally controlled:
  - OS : Windows 7 Pro 32bit
  - CPU : Intel(R) Pentium(R) Dual CPU T2370 1.73GHz,
  - RAM : 2GB

To complete the retrieval of BCs from those resources that have been manually retrieved, the task has been assigned to a person belonging to the consortium, with a BS degree in nursing, expertise in evidence revision and properly trained.

MySQL-Validation-Cleaning-SVM-NO Concerning the creation of the DB MYSQL and the analysis of the BCs contained in it, ZETA and Dipartimento di Scienze Cliniche e Biologiche (DSCB) have exploited the hardware resource made available by ZETA:

- A virtualized machine running CentOS (Community Enterprise Operating System) 7 operating system. A CentOS operation system is a free operating system released under Linux distribution. We assigned to the virtual machine 24Gb of RAM (DDR3) and 8 virtual processors (8 Intel(R) Xeon(R) CPU E5520 @ 2.27GHz). The Machine was running on a Dell PowerEdge R410 physical machine.

For the manual training of SVM, the task has been assigned to a person belonging to the consortium, with a MS degree in pharmacy, mastered in clinical trial and pharmacovigilance, and properly trained.

Time span Concerning to EndNote the processes started 2nd February and ended 14th May, 2015; the time for maximum retrieval (see the above limitations) of the BCs for one resource was variable from one to three days, and it was processed in parallel in 6 computers provided and controlled locally.

Concerning R, the processes started 28th January and ended 14th May, 2015; the time for the overall retrieval of the BCs for one resource was variable from some hours in one computer for a week in the set of 20 computers provided by UBESP, local or remote.

Regarding the manual retrieve of BCs from resources handled manually, the task lasted from 1st February to 6th April, 2015, used an Intel(R) laptop;

Regarding to the processes of MYSQL started 9th April and ended 14th May, 2015, in particular the import process with the verification of duplicates for individual resources and the process of global import in one DB with control of the duplicates had an average duration of 57 hours each; while the import of the classification and relevance classification lasted almost one hour each, while the inclusion of these data in the DB lasted about a day and an half. ZETA provided the servers for these processes.

Regarding the SVM and NO: the task is started on 16th and ended 19th May, 2015, in particular SVM took about 7 hours for processing and 2 more hours for AL, while the NO algorithm took about 4 hours to complete the process of classification.

1.2.4 Search

EndNote

After performing a check in order to understand whether an EndNote connection was available to perform the search, and considering the amount of expected BCs (see sec. 1.2.1), the characteristics of the program (see sec. 1.2.3) and of the SSs (i.e.: sole presence of the connector or), it was decided to work in parallel on multiple computers (see sec. 1.2.3) also for retrieving BCs from the resources connected through EndNote (see sec. 1.2.1). For these reasons, the SSs were then divided in the following way:[1]

sub-string 1.1: artificial intelligence, bayes, belief network, belief networks, classification algorithm, classification algorithms, classifier, classifiers, data mining, kernel estimation;

sub-string 1.2: machine learning, neural network, neural networks, instance based method, instance based methods, k-nearest neighbor, learning vector quantization, self organizing map, self organizing maps, decision tree;

sub-string 1.3: decision tree, decision trees, classification tree, classification trees, iterative dichotomizer, iterative dichotomizers, iterative dichotomiser, iterative dichotomisers, chi-squared automatic interaction, chi-squared automatic interactions, random forest;

sub-string 1.4: random forest, random forests, gradient boosting machine, gradient boosting machines, kernel method, kernel methods, support vector machine, support vector machines, clustering method, clustering methods, k-means;

sub-string 1.5: association rule learning, a-priori algorithm, a-priori algorithms, eclat algorithm, eclat algorithms, naive bayes, bayesian network, bayesian networks, bayesian belief network, bayesian belief networks;

sub-string 1.6: hidden markov model, hidden markov models, perceptron, perceptrons, back propagation, hopfield network, hopfield networks, ensemble method, ensemble methods, deep learning;

sub-string 1.7: restricted boltzmann machine, restricted boltzmann machines, convolutional network, convolutional networks;

sub-string 2.1 recursive partitioning, rule-based model, rule-based models, logic regression, logic regressions, logic forest, logic forests, regularized method, regularized methods, shrinkage method;

sub-string 2.2 shrinkage methods, lasso, elastic-net regularization path, elastic-net regularization paths, ridge penalized regression model, ridge penalized regression models, shrinkage path, shrinkage paths, semiparametric additive hazards model, semiparametric additive hazards models;

sub-string 2.3 high-throughput ridge regression, high-throughput ridge regressions, heteroskedastic effects model, heteroskedastic effects models, kernel learning, bayesian additive regression tree, bayesian additive regression trees, BART, genetic algorithm, genetic algorithms;

sub-string 2.4 memetic algorithm, memetic algorithms, fuzzy rule-based system, fuzzy rule-based systems, rough set theory, fuzzy rough set theory.

The research was carried out on `Title`, `Keywords` and `Abstract`. Since search options change depending on the selected resource (see fig. 5), for each search was selected the option that included less search fields but, at the same time, at least three of our interest.

With regard to the insertion of the keywords, an "exact match" search was run. This type of search is not possible for all EndNote connections, the following message <No matching references found> is displayed for keywords in quotation marks (e.g. "machine learning"). Consequently, to retrieve significant BCs (see sec. 1.2.5) it was decided to proceed, for these resources, with a search of the unlisted keys (e.g.: only machine learning).

Table 4 shows a summary of the EndNote search settings used for each resource. Figure 5 shows a configuration examples for three engines. It should be noted that, among the options available to connect to CINAHL, the option "all fields" (present with PubMed) is absent, the same applies to the field "Title/Keywords/Abstract" available in Web of Science Core Collection (WOS-CORE).

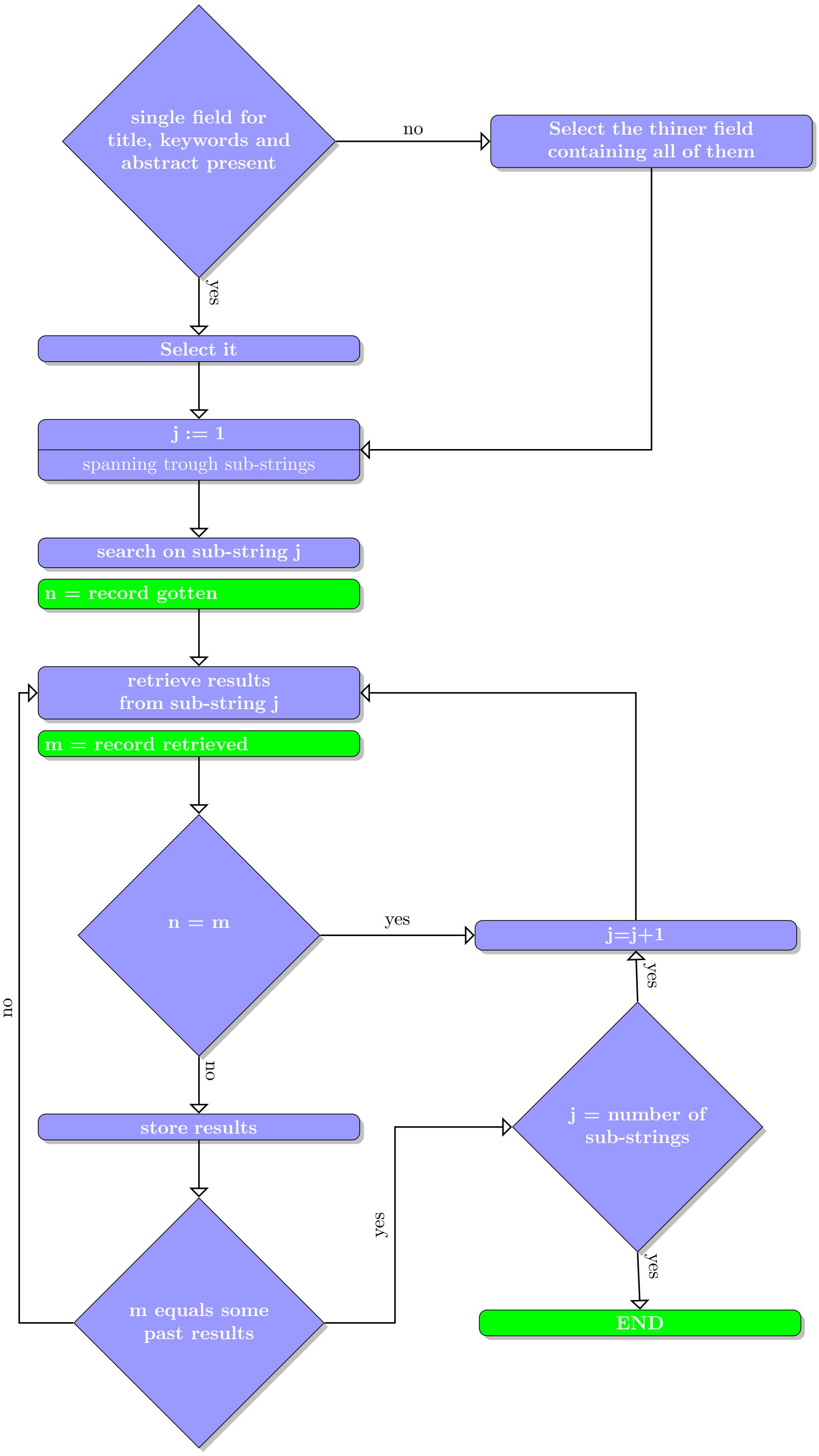
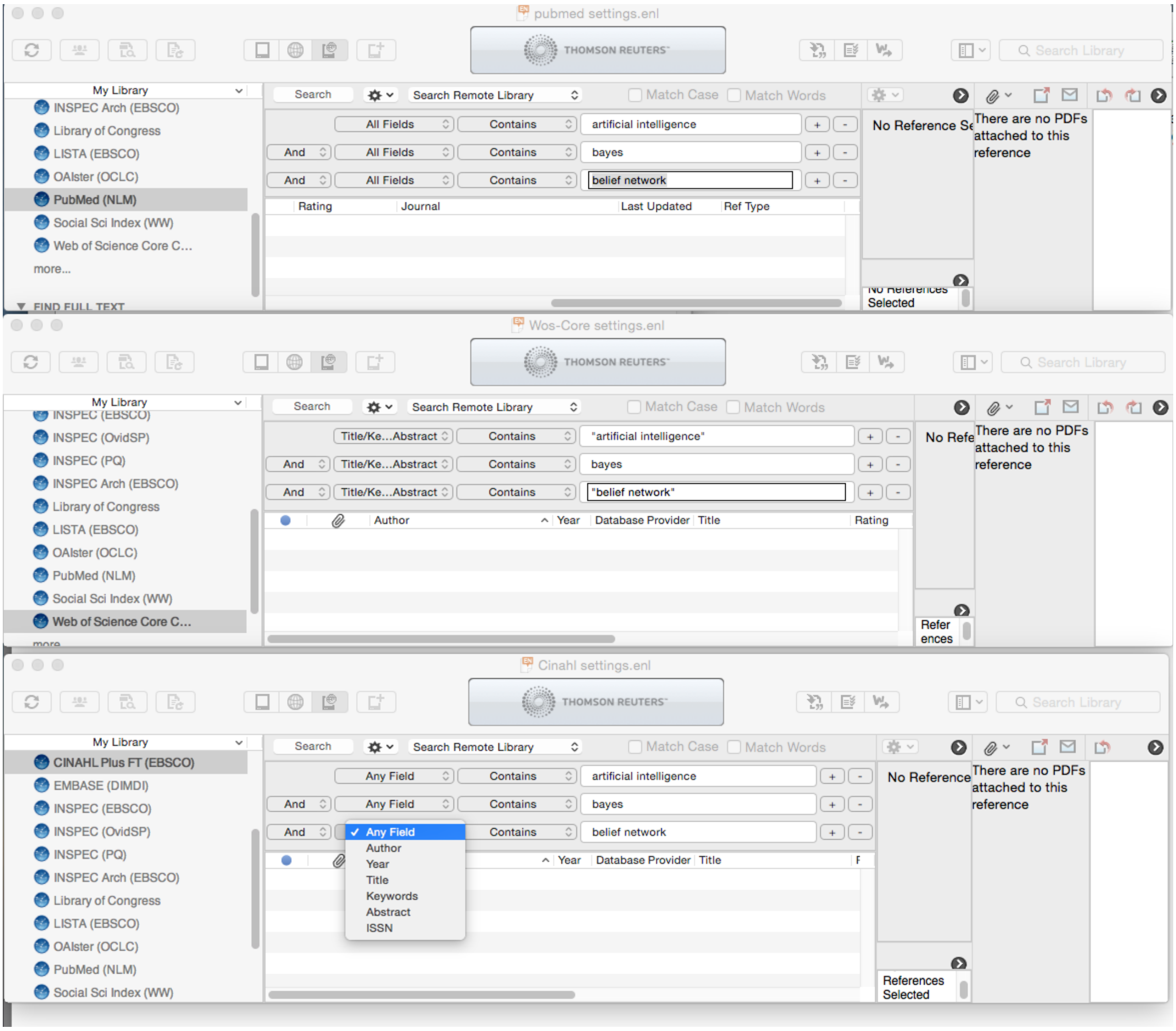

single field for title, keywords and abstract present
no
Select the thiner field containing all of them
yes
Select it
j := 1
spanning trough sub-strings
search on sub-string j
n = record gotten
retrieve results from sub-string j
m = record retrieved
n = m
yes
j=j+1
yes
no
no
store results
j = number of sub-strings
yes
yes
m equals some past results
END

Figure 4: Process of BCs retrieval using EndNote.



pubmed settings.enl

Search Remote Library — PubMed (NLM)

All Fields | Contains | artificial intelligence
And | All Fields | Contains | bayes
And | All Fields | Contains | belief network

Wos-Core settings.enl

Search Remote Library — Web of Science Core C...

Title/Ke...Abstract | Contains | "artificial intelligence"
And | Title/Ke...Abstract | Contains | bayes
And | Title/Ke...Abstract | Contains | "belief network"

Cinahl settings.enl

Search Remote Library — CINAHL Plus FT (EBSCO)

Any Field | Contains | artificial intelligence
And | Any Field | Contains | bayes
And | Any Field | Contains | belief network

Figure 5: Snapshot of three different EndNote configurations for selected resources.

The BCs obtained from EndNote (see `Find` column in Table 10) are often more than the number of BCs actually retrieved (see `Get` column in Table 10). To try to retrieve the largest number of BCs from each resources, a search and download procedure was iterated until all BCs were retrieved or the amount were stabilized.

This has generated many duplicates in the retrieved BCs. Moreover, since the SSs were divided into multiple sub-strings, each research generated more duplicates. Table 10 shows a summary of the results obtained with EndNote for each SS. Exploiting the potential of EndNote (see sec. 1.2.3) an export style format *ad hoc*, for citations, was created (see sec. 1.2.5). Retrieved BCs were then exported and post-processed, as described in sec. 1.2.5, where a check for duplicates was carried out.

R

When a connection to EndNote was not available for a resource, an R script was used to retrieve the relevant BCs. In accordance to the procedure reported in Figure 2, we firstly explored the availibility of a suitable R package, finding `aRxiv` (suitable for arχiv) and `OAIHarvester` (suitable for DOAJ). For resources without a suitable R package, an R script was developed to inteface the resource web page.

[1] sub-string n.m := m-th subset of the n-th strings.

Table 4: EndNote search settings used for each resource.

Resource	Field of search[a]	Quotation admitted (yes/no)
CINAHL	Any field	n
EconLit	Any field	n
MEDLINE	Title/Keyword/Abstract	y
PsycInfo	Any field	n
PubMed	All fields	n
WOS-AHCI	Title/Keyword/Abstract	y
WOS-CORE	Title/Keyword/Abstract	y
WOS-SCI	Title/Keyword/Abstract	y
WOS-SSCI	Title/Keyword/Abstract	y

y := e.g. it is possible to search "machine learning" (quoted)
n := every tried quoted search returns <No matching references found>.

[a] Exactly as it appears into EndNote interface. E.g. `Any field` (CINAHL, EconLit e PsycInfo) and `All field` (PubMed) means the same but with different interface (see fig. 5).

The only condition was the availability of a bi-directional correspondence between the web-page of the search and its actual web address. In other words, the web page of the resource, after having performed the search, was needed to provide in the http header the access of the specific page of the retrieved references (usually separated by &).

In Table 5 we report the strategy adopted to retrieve the BCs of the resources managed by R.

Table 5: Type of R procedure adopted for each resource.

Resource	Package found (yes/no)	Main Package (name)	Connection (URL)
Acm	n	-	`http://dl.acm.org/`
Arχiv	y	aRxiv	—provided by package—
Cis	n	-	`http://www.statindex.org/`
CiteSeerX	n	-	`http://citeseer.ist.psu.edu/`
Doaj	y	oaiharvester	`http://doaj.org/oai.article`
Ingenta	n	-	`http://www.ingentaconnect.com/`
repec	n	-	`https://ideas.repec.org/`

-: no dedicated package used

Manually retrieval

Finally, for those resources for which neither an EndNote connection was available nor a solution through R script was found, the manual retrieval was implemented. A person was dedicated for performing the task and properly trained (see sec. 1.2.3).

All but MathSciNet admit a RIS downloadable file. For all these files, the following three post-processing steps were performed:

1. Because of the amount and sizes of the RIS files resulting from each resource (see tab. 13), BCs were grouped into files each of which with a size of less than 100 MiB (see sec. 1.2.3). To make it possible, the `copy` function of Windows' *Command NT Interface* was used. Specifically, file were gathered together in sub-folders, each one with a maximal size admitted of 100 MiB (see sec. 1.2.3), and then in each folder the following command was ran:

```
copy *.ris <resource>_<n>.ris
```

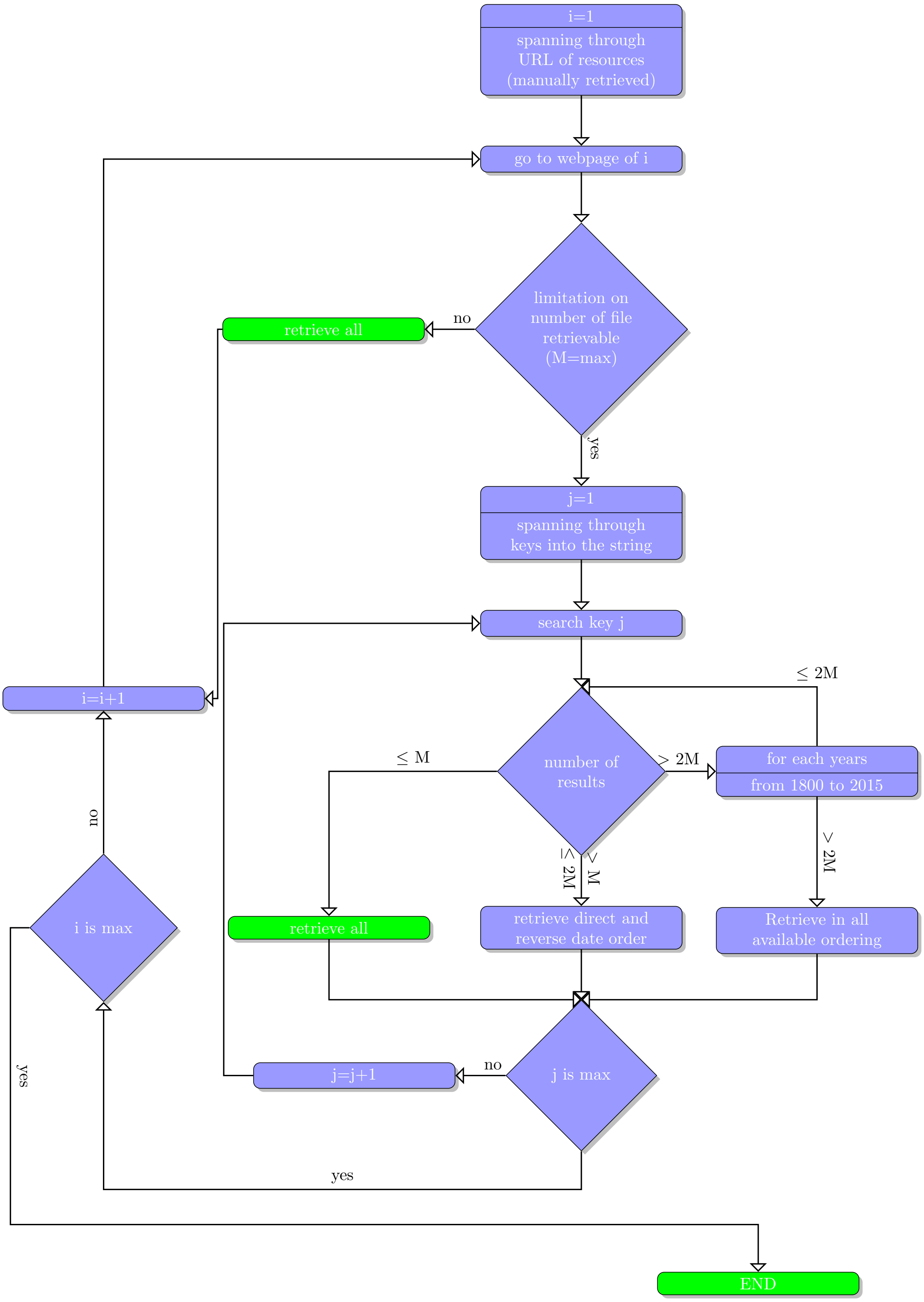

i=1
spanning through URL of resources (manually retrieved)
go to webpage of i
limitation on number of file retrievable (M=max)
no
retrieve all
yes
j=1
spanning through keys into the string
search key j
i=i+1
≤ 2M
number of results
≤ M
> 2M
for each years
from 1800 to 2015
≤ 2M
> M
> 2M
no
i is max
retrieve all
retrieve direct and reverse date order
Retrieve in all available ordering
yes
j=j+1
no
j is max
yes
END

Figure 6: Procedure to manually retrieve BCs.

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

with n varying in the number of sub-folder created.

2. Manual import all these files into a single EndNote library, one for each resource.
3. Export all references with the export style provided (see sec. 1.2.5).

As MathSciNet has a web interface which provide a BIB export fomat, we use the following method. Among the information included in BIB by MathSciNet, the abstract of the various BCs was not inserted because it was in the page dedicated to the specific BCs. Therefore, each abstract was recovered and inserted in the corresponding BIB file, imported in a specific EndNote library and exported as described for the other resources.

A description of the procedure adopted for the six resources considered for manual retrieval is provided in the following sub-paragraphs.

An overview of acts related to the general manually retrieval procedure for the selected resources is provided in Figure 6.

In Table 6 the URLs of the six resources are shown. Each URL allows to access to the relative interface of string search.

Cochrane: In the advanced search we set the option to admit word variation, in order to be able not to explicitily insert plurals for the keywords. After that, the `Title, Abstract, Keywords` field was set and the search was performed in a single line with each keyword in parenthesis and separated by an or. Once the search is performed, Cochrane allows to download a different RIS file each type of references which include all references in that type: i.e. Cochrane Review, Other Reviews, Trials, Methods Studies, Technology Assessments, Economic Evaluations, Cochrane Groups. We have downloaded all except for the last, which is empty.

IeeeXplore: in the advanced search, all eleven lines of searching have been activated, inside of each one the or operator was selected and at a latter stage the `Metadata Only` type of search was checked. Afterwards we had inserted a keyword for each line and performed the search. For every group of keywords each page resulting from the search has been downloaded by selecting firstly `select all on page` and then by `Download Citations` with the options `Citations & Abstract` and `EndNote, Procite, RefMan` as an output format.

MathSciNet: In all downloadable format provided by MathSciNet, excluded the html ones, the abstract isn't included. For that reason it was decided to download all the html pages of our research. The `preferences` were set to provide 100 results per page and to activate numbering (to follow the work more easily). Then, the `Anywhere` and the or operator for all the four field provided by the search have been set. After that, four keywords at a time were inserted and the search was performed. MathSciNet does not provide an automatic "all on page" selection but only "Retrieved First 50" or "Retrieved Marked" so, after the first 50 citations, it was necessary to select each reference in the page until the end, and then retrieve the corresponding html export page.

Scopus: Scopus allows to retrieve only 2 000 citation for each search in a couple of steps. To maximize the number of citations to retrieve, it was decided to perform the following algorithm for downloading the references: searching a single keyword at a time, if the result was less than 2 000 the option "select all" followed by RIS`export` was set. Otherwise, if the result was more than 2 000, it has been grouped limiting the date (starting from 1800) at the minimum range to reduce the results under that limit. If a single keyword in a single year had more than 2 000 citations the result was split in groups by the wider sub-group of "Document Type" in which the result was less than 2 000. If a single keyword in a single year had more than 2 000 citations of the same `Document type` but less than 4 000 then both ascending and descending date ordered BCs were downloaded. If the result had more than 4 000, all the seven types of search sorting were downloading: date, cited by, relevance, date (oldest), First Author (A-Z), Firs Author (Z-A), Source Title (A-Z).

Science Direct: Science Direct limits the download in a single action to 1 000 BCs. In this case, the search was performed by using a single keyword, for a wider range of date starting from 1800. In case a single keyword in a single year had more than 1 000 BCs the results were sub-set by the three `Content type`. Those still presenting more than 1 000 BCs were processed using both type of ordering admitted (date and relevance).

JSTOR: JSTOR also limits the download to 1 000 BCs. However, in a single action the BCs in each page must be selected and a ris file was retrieved for each one. Therefore, the search were split by single keyword, for a single year (if BCs were still more than 1 000), and then in each `Journals`, `Books` or `Pamphlets` category (if BCs were still more than 1 000) and finally in both direct or inverse ordering by date and if the BCs were still more than 1 000 by relevance too.

Table 6: Url for resources manually retrieved.

Resource	URL
Cochrane	`http://onlinelibrary.wiley.com/cochranelibrary/search`
IEEE Xplore	`http://ieeexplore.ieee.org/search/advsearch.jsp`
MathSciNet	`http://ams.math.uni-bielefeld.de/mathscinet`
Scopus	`http://www.scopus.com/search/form.url`
Science Direct	`http://www.sciencedirect.com/science/search`
JSTOR	`http://www.jstor.org/action/showAdvancedSearch`

1.2.5 Study selection

Post-processing

All BCs coming from different resources were merged in an unique DB MYSQL (see sec. 1.2.3). A comma-separated values (CSV) format was prepared after developing an ad hoc separation format, to deal with the presence of commonly used CSV separation formats (e.g. punctuation, peak's series) in fields like abstracts.

Considering the bibliographic structure of EndNote as principal reference, in the developed export style all the fields managed by EndNote were included. Furthermore, an initial field was introduced to provide an identification code to each BC, and a final field constantly filled with a character `1`, meant for checking exportation errors.

This format was implemented both as EndNote export style (see sec. 1.2.3) and in every BCs retrieval procedure conducted using R[2].

Specifically, the developed format use the following sequence for the EOF: `\t&F$4\t`, in which `\t` represents a tabulation and the following for the End of Record (EOR):`\t&nd`. In addition to the two fields just mentioned, the remaining fields are filled with all the information managed by EndNote, in the same order in which are indicated in the software.[3] The specific pattern of the export style format provided is:

Listing 1.1: Export format pattern.

```
Bibliography Number\t&F$4\tReference Type
\t&F$4\tAuthor\t&F$4\tYear\t&F$4\tTitle
\t&F$4\tSecondary Author\t&F$4\t
Secondary Title\t&F$4\tPlace Published
\t&F$4\tPublisher\t&F$4\tVolume\t&F$4\t
Number of Volumes\t&F$4\tNumber\t&F$4\t
Pages\t&F$4\tSection\t&F$4\tTertiary Author
\t&F$4\tTertiary Title\t&F$4\tEdition
\t&F$4\tDate\t&F$4\tType of Work\t&F$4\t
Subsidiary Author\t&F$4\tShort Title\t&F$4\t
Alternate Title\t&F$4\tISBN/ISSN\t&F$4\tDOI
\t&F$4\tOriginal Publication\t&F$4\t
Reprint Edition\t&F$4\tReviewed Item\t&F$4\t
Custom 1\t&F$4\tCustom 2\t&F$4\tCustom 3
\t&F$4\tCustom 4\t&F$4\tCustom 5\t&F$4\t
Custom 6\t&F$4\tCustom 7\t&F$4\tCustom 8
```

[2] All BCs manually retrieved are exported in export style after they have been imported in EndNote (see sec. 1.2.4).

[3] In the `Reference Type` preferences panel.

```
\t&F$4\tAccession Number\t&F$4\tCall Number
\t&F$4\tLabel\t&F$4\tKeywords\t&F$4\tAbstract
\t&F$4\tNotes\t&F$4\tResearch Notes\t&F$4\t
URL\t&F$4\tFile Attachments\t&F$4\t
Author Address\t&F$4\tFigure\t&F$4\tCaption
\t&F$4\tAccess Date\t&F$4\tTranslated Author
\t&F$4\tTranslated Title\t&F$4\t
Name of Database\t&F$4\tDatabase Provider
\t&F$4\tLanguage\t&F$4\t1\t&nd
```

The main DB was developed incorporating all files created following the procedure described below. An overview of actions related to the importing procedure of BCs into the mySQL DB is provided in Figure 7.

In particular, once the data were imported separately for each single resource, the field check was verified to guarantee the correct import of each record for all fields. In case of erroneous records, their analysis was performed in order to understand the problem, which could be attributable either to human error during the code writing or to an intrinsic limit of the used software (see sec. 1.2.3). The original files (EndNote, RData, export of script and txt) were controlled and corrected as long as the record was imported correctly.

When all BCs of a resource were imported, the DOI field was cleaned and duplicates were controlled. Therefore, after resources import, main DB was developed, collecting all the BC and removing the duplicated.

To monitor the original resource, for each DB record a column was added for each resource, while for every item a marker was added to point out the source. The first BC found during import procedure was inserted in the DB, that is the one corresponding to the first mark signed in the corresponding record. Therefore import order become not negligible and it was decided after resources organization, both for number of not empty abstracts and DOI.

When this process was made for both SSs, they were inserted in a DB without duplicated items, storing and underlying also in this case the original string.

MySQL procedure The process of automation of bibliography import, through MYSQL resources, can be divided in two main phases In the first, every resource's BC is imported in a specific table (henceforth Resource Table (RT)). At the same time, the identification of duplicated BC (henceforth Internal Duplicate (INDUP)) imported in the same table is performed. In the second phase, every BC deriving from different RT is mixed in a unique table (henceforth Overall Table (OT)) after INDUP, verifying possible duplicates already present in the OT and deriving from another resource previously imported. Both in BC itself and in OT the resource of the BC is specified.

Fase 1 For each resource a folder was created, in which one or more file in .txt format containing extracted BC are present. Every file is imported in the equivalent RT and an univocal ID is assigned to each record. For each imported file, its name, the number of present records (identified counting the number of EOR \t&nd) and the records included in the specific RT MYSQL, are integrated in a log file. This information gathers all the records inserted in RT, so, that if more than one file is imported for a single resource, the value reflects the number of imported records, and not the last imported file.

To identify import errors, a control field at the end of each record (a field populated '1') was added; when the import of every BC of the single resource is concluded, the records presenting no compliance in control field are identified. Encountered errors are corrected at the source and the resource is imported again. Subsequently data are than cleaned: the spaces before and after each field are removed, the values "", "NA", "NULL" are gathered in NULL. The field DOI, containing the DOI of BC and later used to verify INDUP, is equalized to the various resources, removing from original string, the text that precede 10 (taking into account that every DOI have this prefix).

The possible INDUP at the resource are then identified, using two identification criteria: i) use of DOI of BC (only if it is available), ii) use of complete criteria to control EndNote duplicates (used only for records without filled DOI). In case of one or more records have the same Reference_Type, Author, Year, Title, Secondary_Title, Place_Published, Publisher, Volume, Number, Pages, Section, Tertiary_Title, Subsidiary_Author, Short_Title e Label they are considered INDUP . Among INDUP records of the

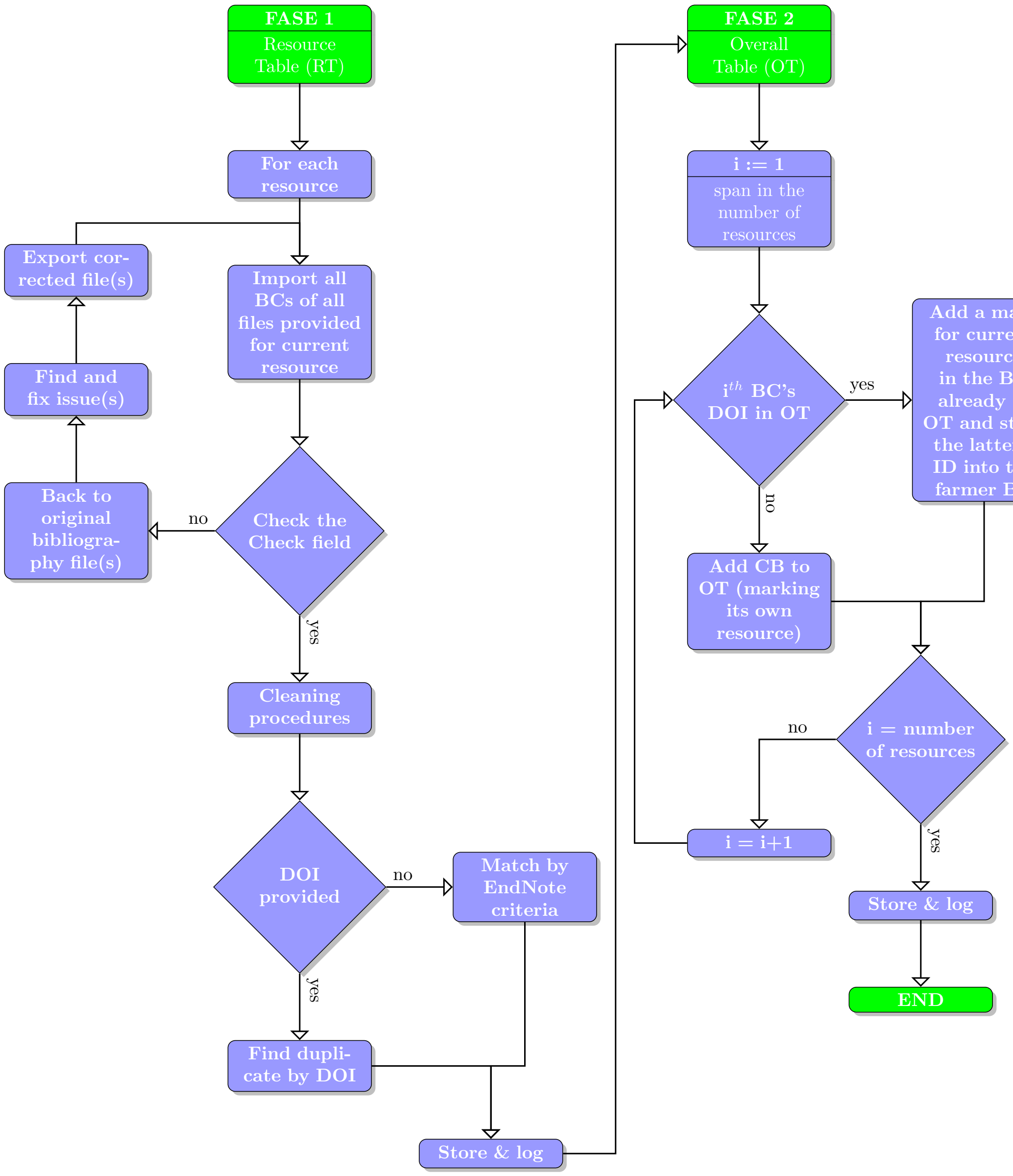

FASE 1
Resource Table (RT)
For each resource
Import all BCs of all files provided for current resource
Check the Check field
no
Back to original bibliography file(s)
Find and fix issue(s)
Export corrected file(s)
yes
Cleaning procedures
DOI provided
no
Match by EndNote criteria
yes
Find duplicate by DOI
Store & log
FASE 2
Overall Table (OT)
i := 1
span in the number of resources
i^{th} BC's DOI in OT
yes
Add a mark for current resource in the BC already in OT and store the latter's ID into the farmer BC
no
Add CB to OT (marking its own resource)
i = number of resources
no
i = i+1
yes
Store & log
END

Figure 7: Flowchart of procedures act to import BCs into the main DB.

same BC, the one who has the lower ID is considered the primary while the others the duplicates. These latter are underlined inserting primary record ID in the field `Duplicate`. A file CSV with the equivalent records is created. In the end, some information related to imported resource are saved underlining, separately, the number of cases in which `Abstract`, `Title`, `Keywords` and `DOI` are missing or otherwise the minimum, maximum and medium number of characters contained in `Abstract`, `Title` and `Keywords` (excluded missing fields).

Fase 2 The OT containing all BC items and the resources from which they derives, is developed in this phase. MYSQL RT records are compared with OT records through `DOI`, to find BC already imported, deriving from other resources. In this case, the ID of the BC included in the OT is inserted in the RT. No other criteria is used for the search of BCs already existing. At the end of this process, the item not already included are imported in the OT and a file CSV containing equivalent BCs with original RT indication is created. Finally, a file log of the resources imported in the OT and other imported records is produced.

The processes of duplicates' import and check are realized through store parameterized procedure of

MYSQL. For the procedure's management and the parameterization a R-script was created, which allows to control their execution through a Open Database Connectivity (ODBC) connection.

Scoring references by relevance

Many words used in the SSs like for example *classification* or *bayesian* may appear in the abstracts in a context different from that of the ML field.

Given the huge number of BCs retrieved, a first analysis was carried out to discriminate between abstracts relevant to the field of ML and potential abstracts selected by the bibliographic search but not relevant to ML.

With regards to relevant abstracts, , all those in which the identified keywords used in the SSs were semantically related to ML were considered. For combined words like *neural network* or *Support Vector Machine*, it was self–evident, but for other words like *classifier* or *bayesian* the relevance of abstracts in which they appeared was not automatically guaranteed. In order to perform this analysis of relevance, SVM were chosen with the aim to categorize abstracts as relevant/non relevant. The theoretical considerations that led to the choice of using SVMs as text classifiers are shortly reported in the following ((Joachims, 1998), (Y. Yang and Pedersen, 1997))

- **Few irrelevant features:** a typical assumption when working with high dimensions input space is to consider the most of the feature as irrelevant and feature selection is carried out to determine these irrelevant features. In text categorization, there are only very few irrelevant features. In fact, the usual assumption is that good classifiers should combine many features (dense concept problem) and that aggressive feature selection may result in a loss of information.
- **High dimensional input space:** the input space is made of all the words contained in the abstracts. SVM have the potential to handle these large feature spaces since they use overfitting protection, which does not necessarily depend on the number of features.
- Working on many documents with potential different vocabulary leads to work on sparse matrices. Both theoretical and empirical evidence show that SVMs are well suited for problems with dense concepts and sparse instances.
- **Most text categorization problems are linearly separable.** Empirical evidences show that often in text categories are linearly separable. And SVMs outperformed other text classifiers in finding such linear (or polynomial or radial basis) separators.

Pre-processing In order to perform text categorization using SVMs is necessary to transform documents, which are strings of characters, into a representation suitable for the learning algorithm and the classification task. According to Information Retrieval research, the *bag-of-words* assumption, in which the order of words in the document is ignored because it is considered of minor importance, was made. This assumption led to an attribute–value representation of text: each distinct word corresponded to a feature, with the number of times the word occurred in the document as its value. To avoid unnecessarily large feature vectors, words were considered as features only if they were not stop-words (like and, or, etc.).

The overall BCs returned from SSs #1 and #2 were thus checked for:

- missing abstracts;
- number of characters;
- English language.

BCs with:

- missing abstracts;
- abstracts too short (abstracts with less than 700 characters);
- non-English abstracts

were discarded.

The choice to discard abstracts with less than 700 characters was dictated by the need to have a sufficient amount of text on which classify. The number of 700 characters guaranteed the abstracts of at least 100 hundred words were considered.

Finally the text of the remaining abstracts was further processed in order to discard mathematical formulas, HTML code, TEX/LATEX code, Unicode sequences as detailed below:

- only the characters belonging to the Lower ASCII Character Set (character codes 0-127) were retained, since no plain English words use the Extended (Higher) ASCII Character Set (character codes 128-255).
- characters

  `\r`

  and

  `\n`

  that are used as "carriage return" or "line feed" printing commands were removed.
- TEX/LATEX commands, which are single-words preceded by a

  `\`

  (backslash character) were removed.
- HTML tags, in one of the forms

  `<tag_name>`

  or

  `</tag_name>`

  or

  `<tag_name/>`

  were removed.
- HTML representations of ASCII codes, in the form

  `&\#digits;`

  were removed.
- Unicode representations of the characters belonging to the Unicode Standard, in the form

  `<U+digitsletters>`

  were removed.
- The following symbols:

  `\"^+−*/=%${}[]()<>.,;:?!`

  were removed.
- Lastly, all digits were also removed.

Active Learning procedure (AL) Since SVMs are supervised algorithms, a training dataset of labelled relevant abstracts was built for their implementation. To reduce the manual annotation efforts without sacrify the classification accuracy, a sample selection strategy, or an AL procedure was developed.

The AL is an approach for developing supervised learning algorithm while reducing the labeling cost. The AL procedure iteratively selects a sample of data to be labeled based on some selection strategy, which suggests to pick the data that most deserves to be labeled ((B. Yang et al., 2009)).

The selection strategy can be defined in order to:

a) iteratively label the unlabeled data on which the current hypothesis is most uncertain *(uncertainty sampling)*;
b) label data to minimize the expected error on the unlabeled data *(expected error reduction)*;
c) label data that have largest disagreement among several classifiers *(Query-by-Committee)*.

The pool-based AL approach, which is the most popular paradigm of AL, was adopted. The pool-based AL approach works in the following way: it assumes that a set of partially labeled documents, which is typically small in size, is given. At the beginning, a classifier is trained using the initial labeled set. Based on this classifier, according to the learning strategy chosen, a sample from the unlabeled documents is drawn out and asks for its true labels. Then, the newly labeled documents are incorporated into the initial set of labeled documents and the classifier is trained on this new set. The training and the labeling process runs iteratively after a certain number of iterations or when the classifier achieves a sufficient accuracy.

The AL procedure is described in the following algorithm:

a) The SVM classifier was trained on the initial dataset of manually labelled abstracts;

b) the cross-validated accuracy was computed;

c) 1000 abstracts were randomly selected from the DB of all the BCs but those ones in the training set;

d) for each of the 1 000 abstracts, SVM classifier was run to predict the label relevant/not–relevant;

e) for each of the 1 000 labelled abstracts $\hat{a}_i, \quad i)1\ldots1000$, the expected loss reduction was computed using the following formula

$$\frac{1-\hat{a}_i \times \text{SVM}(a_i)}{2}$$

where a_i denote the i–th abstract, for $i = 1, \ldots, 1000$ and $SVM(a_i)$ is the size of the version space of the classifier associated with target class i and learnt from the labelled data;

f) all the scores computed in the step above were sorted in decreasing order and abstracts with a score greater than the 97.5 percentile were retained and pulled together with the abstracts in the initial training set;

g) the SVM classifier was trained on the updated training set.

Then steps from b) to g) were repeated until the cross-validated accuracy reached the threshold of 0.98. In every iteration, once the selected data were incorporated, the AL retrained a new classifier on the expanded labelled set.

In Figure 8 the iterative procedure of AL is depicted.

The criteria used to manually label the initial set of abstracts were the following:

- the key words used in the SS were identified into the abstracts;
- the entire sentence in which the keywords appeared was assessed if it was semantically related to ML. In case of doubt previous and subsequent sentences were also evaluated in order to decide if the abstract should be considered as relevant.

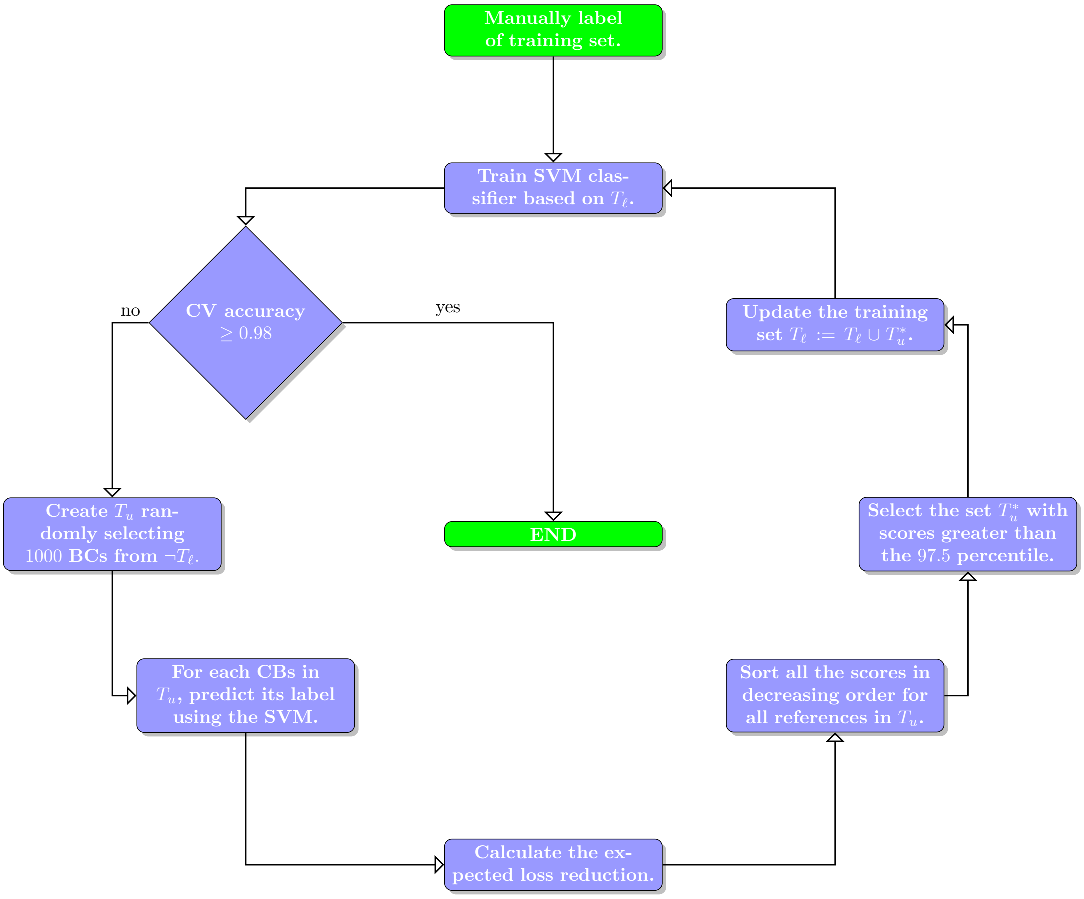
The manual annotation was done by a person properly trained (see sec. 1.2.3).

Score probability of relevance After the trained classifier reached the sufficient accuracy, it was used to classify the remaining unlabelled abstracts, which were returned with a score probability to be relevant to the ML field.

Validation of DB

A test set of selected BCs extracted from a list of works reported in the bibliography of nine relevant ML books (provided by the Consortium experts) was created in order to check the accuracy of the search. An overview on the adopted procedure is provided in Figure 9.

From the nine books, 230 pages were acquired, in a single pdf file bearing their BCs (CitForVal .pdf). Consortium experts selected relevant BCs and DSCB produced, through the application of the optical character recognition (OCR), a CSV file showing the text content. The file had 3 412 lines (from now on First Data Entry (FDE)). A member of the staff selected and trained by ZETA produced a CSV suitable for research within the DB, using the following specifications and procedures.


Manually label of training set.
Train SVM classifier based on T_ℓ.
CV accuracy ≥ 0.98
no
yes
END
Create T_u randomly selecting 1000 BCs from $\neg T_\ell$.
For each CBs in T_u, predict its label using the SVM.
Calculate the expected loss reduction.
Sort all the scores in decreasing order for all references in T_u.
Select the set T_u^* with scores greater than the 97.5 percentile.
Update the training set $T_\ell := T_\ell \cup T_u^*$.

T_ℓ := labelled training set;
CV := Cross Validated;
T_u := unlabelled dataset;
$\neg T_\ell$:= dataset of all BCs excepts those uses in T_ℓ.

Figure 8: Flowchart for the pertinent SVM classifier labelling of BCs.

Specifications: in order to perform a correct verification of BCs inclusion in the DB, the following fields were checked

- First author;
- Year;
- Title (until first full stop).

Procedure: for a correct creation of the CSV file and verification of the validity of the entered data, an internal procedure [`Zeta SOI ZR 19 Standard Operating Procedure for Data Management`], usually used for clinical trials, has been adapted.

Single data entry by first operator: the FDE was divided by an IT operator ZETA in four record intervals of 853 lines each, each BC interval has been processed by a different operator. Fields described in the specifications have been compiled in a `.xlsx` file (now called Second Data Entry (SDE)). The IT operator has, in the `.xlsx` file, eliminated blank lines in the file FDE, divided BCs that were on the same line and corrected erroneous beginning of lines. With these four files together, the SDE had 3 231 lines, i.e. 3 231 different BCs.

A visual check was performed by a second operator: Resumed the FDE, this was divided into nine parts, one for each book and the same was did with the SDE. Five different operators carried out the check, to which no one has been attributed a portion corresponding to the one made in the previous phase. The check was between what is recorded on the FDE and what was entered on the

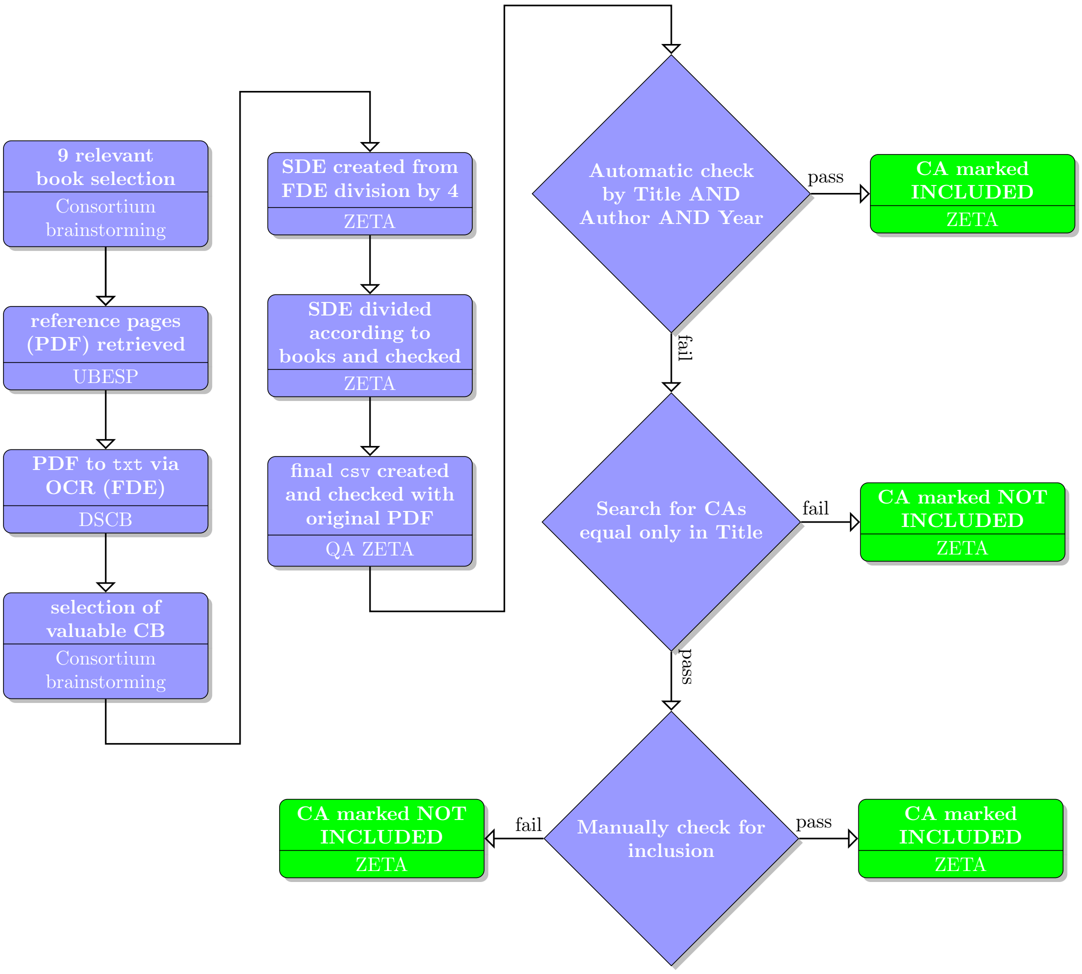

9 relevant book selection
Consortium brainstorming
reference pages (PDF) retrieved
UBESP
PDF to txt via OCR (FDE)
DSCB
selection of valuable CB
Consortium brainstorming
SDE created from FDE division by 4
ZETA
SDE divided according to books and checked
ZETA
final csv created and checked with original PDF
QA ZETA
Automatic check by Title AND Author AND Year
pass
CA marked INCLUDED
ZETA
fail
Search for CAs equal only in Title
fail
CA marked NOT INCLUDED
ZETA
pass
Manually check for inclusion
fail
CA marked NOT INCLUDED
ZETA
pass
CA marked INCLUDED
ZETA

Figure 9: Flowchart representing activities to validate DB.

SDE for each BC. Matching the list of BCs aligned the two files. In case of doubt, the operator has carried out a verification source file directly on `CitForVal.pdf`.

Validation via Monitoring by Quality Assurance (QA): Quality Assurance (QA) performed a random visual check.

In table 7 you can find the subdivisions made for FDE and SDE.

Table 7: Document partitions for FDE and SDE

Operator (code)	**FDE** (lines range)	**SDE** (Book ID)[a]
A	-	B1 (1-845) + B8 (2 627-2 929)
B	1 707-2 559	B5 (1 351-1 395) +B3 (676-1 051)
C	1-853	B4 (1 052-1 350) + B9 (2 930-3 250[b])
D	2 560-3 412[b]	B6 (1 396-2 522)
E	854-1 706	B2 (486-675) + B7 (2 523-2 626)

-: not present
B1–B9: Book #1 – Book #9

[a] Ordered as in CitForVal.pdf (SDE lines range provided between brackets)
[b] Last line on FDE is not equal to SDE one because the white lines cut and the multiple lines split.

At this point the CSV format was used to query the DB for a first automatic control of the inclusion of BCs. The parameters of control were:

- Title in CSV included in the `Title` field of the DB;
- Fields same `Year`;
- Field `Author` CSV included in one of DB.

When all the three controls were successful the BC was marked ”included”, othewise it was given to the control only by title and author.A file with all the possible matches was created and passed to the manual control of the entire record to monitor the actual membership. Identified, possibly a corresponding BC, the ID of this BC was reported as a correspondent and BC has been marked as included. In other cases, the BCs has been marked as not included. A person was dedicated for performing the task (see sec. 1.2.3). The results of the audit can be found in Table 8. Out of 3231 BCs identified in the validation books, 1093 were excluded according to the inclusion criteria listed above. Out of the remaining (2138), 986 have been identified in the first control (Title + Author + Year), 386 in the second (Title + Author) and 274 in the third (Title only), up to a total of 1646 BCs, that represents a 79% of the sample considered.

Table 8: Validation of DB

References	Considered	Included	Not Included
3231	2138	1646	492

1.2.6 Records classification criteria

Classification of the MLT

In order to classify all the MLT in a useful way for EFSA, a data sheet which will be compiled for each technique had been created. Moreover, as explained in the sub-section 1.3.2, all the *relevant* references will be provided (by the svm) with a score for each of the following entry. In this way, the article can be searched for all the following topics, ordered by specific rank of relevance.

On the other hand, this classification will be the first step leading to the implementation of the main decision tree on road-vm.

Some general and methodological aspects and concepts which might be useful for an effective implementation of MLT to the analysis of the topics addressed by EFSA in its current scientific activity have been identified. In this activity, the chosen perspective is that of a statistician willing to extend her/his analytic horizons toward the use of MLT. It has been adopted in this sense, a terminology and a way of conceptualizing topics and issues closer to the statistical terminology than other (e.g.: computational or mathematical formulations of the problems). This decision has been taken in view of the practical workflow in a data-analysis context, like the one foreseen by EFSA, where the statistician is leading the process. This lead to identify few macro-categories as a guidance for a more detailed classification:

- **Type of MLT**. Traditionally, classification, regression and clustering, although strictly inter-connected techniques, are treated as independent machineries, with their own tools and a unique analytical perspective. In general, the classification problem is focused on providing the best possible assignment of a set of observation to a set of labels representing the classes, on a basis of observed individual features (Richard M Cormack, 1971). Small if no-attention is paid to evaluating how the single (or a set of) features impact the belonging of an individual to a given class. In regression the main focus is usually the identification of the effects of a features on the response variable (Dasgupta et al., 2011), associated to a strong emphasis on prediction (Copas, 1983). Clustering is a pillar of the classical multivariate analysis and of current data-mining approaches, where the focus is to identify subgroups or patients more similar to each other than to others. Here, the emphasis is given to the homogeneity of the identified cluster, with no prior knowledge on how ”labeling” the clusters nor to impact of features to the cluster belonging (Gnanadesikan, 2011). If we adopt this schema, a simple, oversimplifying yet perhaps useful table could be derived (see tab. 9).
- **User Interaction**. Feature selection techniques do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of

Table 9: Gross taxonomic identification of mlt. Combinations not listed here are mixed methods and hybrid approaches.

	A priori Known	Matter of interest	Field of interest
Classes	Y	Y	Classification
Feature's effects	Y	Y	Regression
Classes	N	Y	Clustering

Y := yes.
N := no.

the variables, hence, offering the advantage of interpretability by a domain expert (Saeys, Inza, and Larrañaga, 2007a). Feature selection can be applied to both supervised and unsupervised learning, depending from the nature of the problem, as stated in the previous item. Supervised learning (classification), is the case where the class labels are known beforehand. Unsupervised learning (clustering) instead, is a more complex issue and occurs when labels are not a priori known (Varshavsky et al., 2006). Hybrid techniques are those identified in a recent review (Alpaydin, 2014) and commonly indicated under the names of *reinforcement* (Cuayáhuitl et al., 2013), *deep learning* (Bengio, 2009) and *active learning* (Bordes et al., 2005). The proposed classification is hierarchically organized first selecting among ”pure” supervised, unsupervised and hybrid approaches, and then, for the latter, including a list of potential approaches, like the ones just listed above, open for further refinements if needed.

- **Type of input variables**. Most MLT are limited to the treatment of one or more type of statistical variables, providing biased or inefficient -if not useless at all- results of the analysis if applied wrongly to a set of variables not suitable for the specific method. Heuristics behind the choice has been proposed and investigated (Nisbett et al., 1983). The taxonomy presented here is a scholarly evaluation of the most common types as occurring in practical research.
- **Methodological aspects**. There are several aspects related to practical implementation of MLT. Such aspects are typically shared across most of the statistical work: they range from minimal/optimal number of observations or subjects for the technique being validly applied, the minimum/maximum/optimal number of features admitted by the technique or the magnitude of the ratio $\frac{\sharp\text{ of subjects}}{\sharp\text{ of variables}}$ (Harrell Jr et al., 1985). Such aspects cover also the distributional assumptions needed by the methods, both in classical and Bayesian analysis or, if not needed, by non-parametric or hybrid approaches like empirical Bayes or density estimators (Norman L Johnson and Kotz, 1970)(Norman Lloyd Johnson, Kotz, and Balakrishnan, 1995)(Norman Lloyd Johnson, Kotz, and Balakrishnan, 1997). Other well known statistical issues will also enter in this field, like missing values, goodness of fit and robustness to model/approach assumptions. Both issues have been treated in the context of MLT, and excellent reviews have been published on the topic (W. Z. Liu et al., 1997)(Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas, 2006). Goodness of fit is limited to model-based MLT, and it has been reviewed fairly recently both in classical approaches (Lavesson and Davidsson, 2007) and in clustering (Anil K Jain, Murty, and Flynn, 1999). In wider terms, goodness of fit can also be viewed in the perspective of feature selection: in this case, the goal is to gain efficiency in reducing overfitting (Guyon and Elisseeff, 2003). Efficiency involves also the decision on the form of the functional relationships between features and response(s), well discussed in a seminal reference of Blum (Blum and Langley, 1997) and recently updated (Sotiris B Kotsiantis, I. Zaharakis, and P. Pintelas, 2007). Robustness is a transversal concept in MLT, ranging from the development of MLT fitting the robust methods chapter of statistical analysis (Trafalis and Gilbert, 2006) or discussing robustness to the model/approach assumptions or characteristics, like in recent review, all focused on specific MLT (De La Torre and Black, 2003) (Shami and Verhelst, 2007a)(Arel, Rose, and Karnowski, 2010).
- **Output of MLT**. From the applied point of view, there are two main aspects that can help in discriminating the MLT to apply, i.e.:
  a) the possibility of obtaining more or less direct estimation of an effect of features on the response, possibly via a parametric estimation compatible with a biological or epidemiological interpretation;
  b) the possibility to gain from the MLT model a prediction (or forecasting in longitudinal studies terminology) on the potential occurrence of the response of interest.

Not all MLT allow for either one or both of them, as pointed out recently in comparative studies (Kruppa, Ziegler, and König, 2012)(Y. Liu, 2004).

- **Computational aspects of MLT.** An important aspect to consider is the availability of appropriate software packages for performing the MLT analysis or the evaluation of the proper algorithm for gaining computational efficiency. The Comprehensive R Archive Network provide an excellent list of available R Libraries at cran.r-project.org, maintained by Torsten Hothorn, while algorithms available and most efficient are specific for any MLT, and specific reviews are available (Snoek, Larochelle, and R. P. Adams, 2012)(Parsons, Haque, and H. Liu, 2004).

The chosen approach foresee a mixture of horizontal and vertical selection of relevant aspects within the extracted MLT literature. Basically, first a choice is expected to be made within the main topic, proceeding further deeper in the selection tree for getting more focused results.

- Type of mlt
  - classification,
  - parametrization/regression,
  - clustering,
  - hybrid approaches;
- Interaction
  - supervised,
  - unsupervised,
  - hybrid
    - reinforcement,
    - deep learning,
    - active learning;
- Input
  - categorical
  - numerical,
  - mixed,
  - ordered,
  - nominal,
  - dichotomic,
  - discrete,
  - continuous;
- Methodological aspects
  - Output;
  - Number of subjects;
  - Number of variables;
  - Magnitude of the relation subj/var;
  - Distributional assumptions
    - Not needed (non-parametric)
    - Fully specified
      - categorical
      - monovariate,
      - multivariate,
      - Prior,
      - Normal,
      - Gamma,
      - Dirichlet,
      - Others,
    - Hybrid approaches
  - Robustness,
  - Missing value,
  - Goodness of fit
  - Efficency/complexity
    - space
    - time,
    - constant,
    - sublinear,
    - linear,
    - superlinear,
    - quadratic,
    - polynomial,
    - exponential;
- Output of MLT
  - Parameter interpretation
    - direct
    - indirect;
  - Prediction and Forecasting;
- R-package
  - cran
  - GitHub,
  - others,
  - none;
- Algorithms
  - one
  - more.

There are two operational considerations worth to be made with regard to the classification criteria, one about complexity and the other about flexibility. For what concern complexity, the management of a large set of classification criteria in retrieving records is most likely impossible to be managed without a proper interface. Therefore, we updated the webi interface to incorporate a search based on the classification criteria proposed. The procedure is described (with examples) in Appendix C. For what concerns flexibility, we do expect that, as the work will go further, in particular involving the evaluation of the specific EFSA needs, as foreseen for the subsequent deliverable, there will be the need for refinements in the classification procedure. Indeed, the proposed classification, is intrinsically dependent on the purposes on which it has been built (Clancey, 1993) and we expect to incorporate or amend proposed classification once new concepts or needs will emerge. Therefore, among the potential class of classification mechanisms we favored those having more flexibility. In this case, all classification algorithms which are based on a heavy training session should be excluded because hardly fitting the requirement of a seamless updating. Among those showing a high degree of flexibility, the "name occurrence" has the advantage of being flexible and accurate, representing, by definition, a wider set incorporating all those stemming from other classification algorithms (Aggarwal and Zhai, 2012).

Name Co–occurrence NO was chosen as the methodology to classify abstracts by topic annotation. The methods based on NO quantify the relatedness of two domain variables by the relative frequency of co-occurrence of their names and possibly synonyms in documents from a corpus. In (Burgun and Bodenreider, 2001) the evaluation of the co–occurrence of MeSH terms in Medline abstracts against the manually curated UMLS Semantic Net similarly showed the effectiveness of this approach.

To quantify the relatedness of two domain variables, a vector representations of their textual descriptions, which are called kernels, is needed. Each component of the vector represents the weight of a single word in the document, which can be derived from simple counting statistics or eventually from more elaborate weighting techniques. Thus, the relatedness of two variables can be based on either direct similarity or indirectly by the corelevance of their kernels.

Direct similarity means that domain variables (concepts) are related if their descriptions are similar while corelevance (indirect similarity) means that variables are related if the same documents are similar to their descriptions. Usually, the similarity measure of documents is defined by common concept presence.

In order to perform the NO analysis, the same pre-processing procedure used for cleaning the text of the abstracts described in 1.2.5 was carried out.

The same attribute–value representation of the text (i. e., each distinct word corresponded to a feature, with the number of times the word occurred in the document as its value) adopted for the text categorization performed with SVMs was considered. Even in this case to avoid unnecessarily large feature vectors, words were considered as features only if they were no stop-words.

Sparse terms, i.e. terms occurring only in very few documents, were removed. Normally, this reduces dramatically the size of the matrix representation of the features into documents without losing significant relations. The final result of this pre–processing step was a document–term matrix with documents as rows and terms as columns; the matrix elements are term frequencies.

To derive a vector representation for each MLT topic identified as worthy to be annotated, associations between terms were searched for. Finding associations is a form of count-based evaluation method, and it was done by computing correlations between terms, i.e. the correlations between all terms in the document–term matrix were computed and then those higher than the correlation threshold were filter out. The minimal correlation for identifying association was set to 0.3, in order to be not too restrictive in detecting relatedness among variables (Feinerer, Hornik, and Meyer, 2008).

1.3 Results

1.3.1 Search outcome

Individual outcome

In Table 10 the total number of abstracts retrieved using Endnote by search string #1 and search string #2 is provided, stratified according to the bibliographic database resources. The column **Get** provides the number of abstracts that EndNote was able to retrieve among all those found, which is thus less or equal to the respective number reported in column **Find**.

WoS–Core and Wos–SCI are the resources from which the major number of references were retrieved. It turned out that both of them contained also the majority of replicated abstracts (about half of the total abstract found), according to what can be seen in the (**Unique** column). This is due to the order by which the databases were investigated. Indeed, the bibliographic resources with higher quality of the "abstract" field (basically according to the degree of its completion) received a higher priority (e.g.: Medline, which is part of PubMed, since it was searched before PubMed, references found in the latter database almost surely contained those found in MedLine).

Table 10: Results for resources gotten using EndNote.

(a) *String #1*

Resource	Find	Get	Unique[a]
Cinahl	15 708	15 696	13 349
EconLit	10 478	10 478	8909
MedLine	82 395	82 395	66 519
PsycInfo	25 723	2661	2435
PubMed	308 878	333 756	188 914
wos–ahci	1597	1580	1487
wos–core	492 548	328 760	244 149
wos–sci	418 003	280 060	234 567
wos–ssci	22 119	20 375	17 782

(b) *String #2 (selection of resources)*

Resource	Find	Get	Unique[a]
EconLit	3296	3296	2858
MedLine	44 626	44 626	35 083
PsycInfo	1553	1430	1359
PubMed	65 007	65 007	38 522
wos–core	188 574	187 984	145 262
wos–sci	117 965	117 965	91 073

[a] Not replicated records

The same information is provided also for abstracts retrieved using R as a retrieval-interface. In Table 11 the number of files and the number of references contained in the files is reported according to search string.

Regarding the results obtained from the BCs, it's useful to report ACM because it contains the largest number of results, due to the fact that this resource is concerned specifically the area of computer machinery and CiteSeerX, which contains the most number of retrieved files. This latter fact is caused by the research methodology adopted: admitting the recovery of up to 500 BCs per call, the automated search has been divided for individual keywords and per year, thus generating a large number of empty files.

In Table 12 the number of references retrieved manually are reported (only for string #1 since search string #2 was not carried out manually).

Regarding the manually retrieved files, it can be observed very high results, as the amount of the unique voices. Despite the many types of sorting used for the retrieval, was still reduced the overlap between the research.

A overview of the quantity of files retrieved and their size (split by string) can be found in Table 13.

To summarize all the results, in the Figure 10 we have reported the number of unique, obtained and founded (the latter only for EndNote) citations for each investigated resources.

Table 11: Results for resources gotten using R.

(a) *String #1*

Resource	Files	Records	Unique
Acm	21 854	449 981	449 157
Arχiv	92	45 775	45 757
Cis	747	20 509	13 635
CiteSeerX	648 000	15 515	10 887
Doaj	127	34 904	34 904
Ingenta	1499	68 042	66 451
repec	3595	36 136	36 054

(b) *String #2*

Resource	Files	Records	Unique
acm	21 854	462 519	449 981
Arχiv	6	2803	2803
cis	113	2615	1570
CiteSeerX	270 000	14 002	9289
doaj	127	9736	9720
Ingenta	271	12 085	10 325
repec	339	4480	4400

Table 12: Results for resources manually retrieved.

(a) *String #1 (only)*

Resource	Records	Unique
Cochrane	2205	2189
IEEE Xplore	433 458	315 768
JSTOR	562 363	298 973
MathSciNet	34 150	34 128
Scopus	515 816	431 252
Science Direct	222 764	167 365

MySQL outcome

The procedure described in the previous subsections was adopted for both string. The union of string #1 and string #2 was performed excluding duplicates only by doi. The summary of these procedures are reported in table 14 and shown on figure 11.

In Figure 11, unique and duplicated records and their relative size is shown. After merging the BCs from various resources and from the two SSs, there are still present duplicates because the control was performed only by DOI.[4] Without altering the contents of the DB, the following analysis of BCs has been conducted to estimate the number of duplicates. In the first place, all BCs were grouped by the field *Title*; next, within each group, the number of BCs with different *Abstract* field and the number BCs with different *Author* field were counted. Subsequently, if at least one the two latter numbers was inferior to the number of BCs in the corresponding group, that group was regarded as containing at least one duplicate. Finally, the number of groups that include duplicates were counted, amounting to a total of 352797 groups of BCs (approximately 13.29 % with respect to the total number of BCs and 17,33% with respect to the groups). On the one hand, one can observe that this number could be overestimated to the extent that the same team of authors might have written two different articles with the same title or that two different articles have the same title and the same abstract. On the other hand, it is underestimated in the sense that some fields of different BCs in the DB

[4] A search through WEBi (see App. **??**) empirically confirms this observation

Table 13: Quantity and relative size of files retrieved.

(a) *String #1 (all resources)*

Resource	Files	Size (MiB)
acm	21 854	417
arχiv	92	59
Cinahl	1	16
cis	747	16
CiteSeerX	648 000	1480
Cochrane	6	4
doaj	127	3650
EconLit	1	12
IeeeXplore	4158	810
Ingenta	1499	135
JSTOR	5755	765
MathSciNet	806	673
MedLine	1	134
PsycInfo	1	7
PubMed	1	422
repec	3595	24
Science Direct	367	342
Scopus	495	4140
wos–ahci	1	2
wos–core	1	449
wos–sci	1	426
wos–ssci	1	32

(b) *String #2 (selection resources)*

Resource	Files	Size (MiB)
acm	21 854	417
arχiv	6	4
cis	113	2
CiteSeerX	270 000	623
doaj	127	3650
EconLit	1	4
Ingenta	271	22
MedLine	1	73
Psycinfo	1	3
PubMed	1	88
repec	339	2
wos–core	1	263
wos–sci	1	173

Table 14: Imported references into mysql DB.

String	Considered	Duplicate by EndNote	Duplicate by DOI	Imported
#1	3 621 863	950 424	323 971	2 347 468
#2	360 349	8636	15 696	336 017
$\#1 \cup \#2$[a]	2 683 485[b]	NA	28 120	2 655 365

∪: the union set operator.
NA := Not Applicable.

[a] The merging of the two strings.
[b] #1 imported + #2 imported.

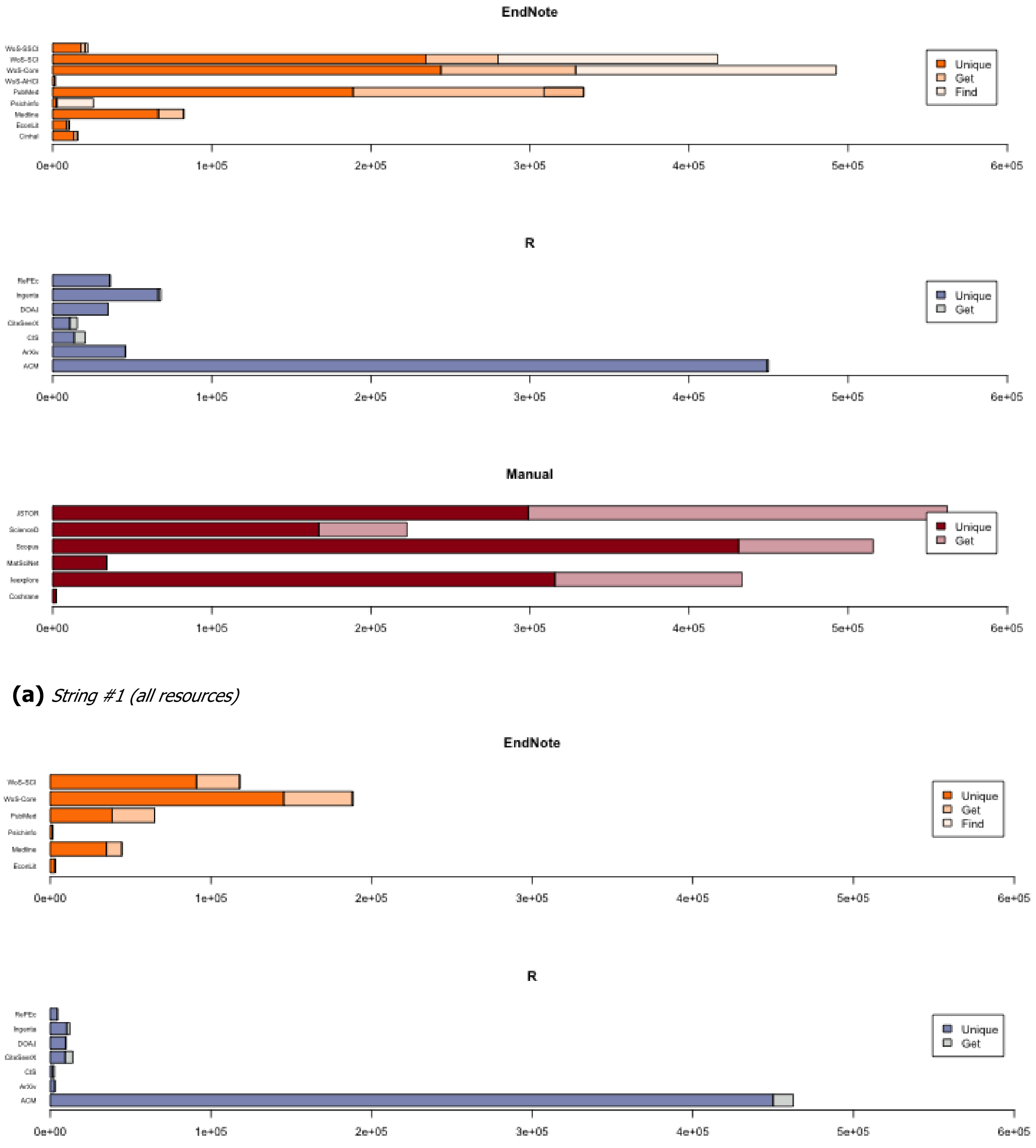
are written in a different manner but are, *de facto*, equal.

In Table 15 it reported the number of BCs that are coming from one or more resources, in the Figure 12 there is a representation of it.

Therefore the major number of the BCs were retrieved from just one resource, it be negligible to consider the quantity of those retrieved from more resources.

1.3.2 Study selection outcome

In the table 16, the overall number of abstracts retrieved and the number of abstracts retained after the checks in step 2 are reported according to the resource.



(a) *String #1 (all resources)*

(b) *String #2 (selection of resources)*

Figure 10: Citations *unique-gotten-founded* for each resource.

SVM relevance classification

Overall, 866 abstracts were manually classified as relevant/non-relevant and constituted the starting training set. From the overall number of no-missing abstracts, a sample of 1,000 abstract was drawn in order to keep the representativity of the 21 resources (CIS was excluded because abstracts are missing). Finally from the sample, abstracts with less than 700 characters were filtered out. Filtering out abstracts was not a problem in term of size of the training set since working on a small amount of training data is very typical in AL (B. Yang

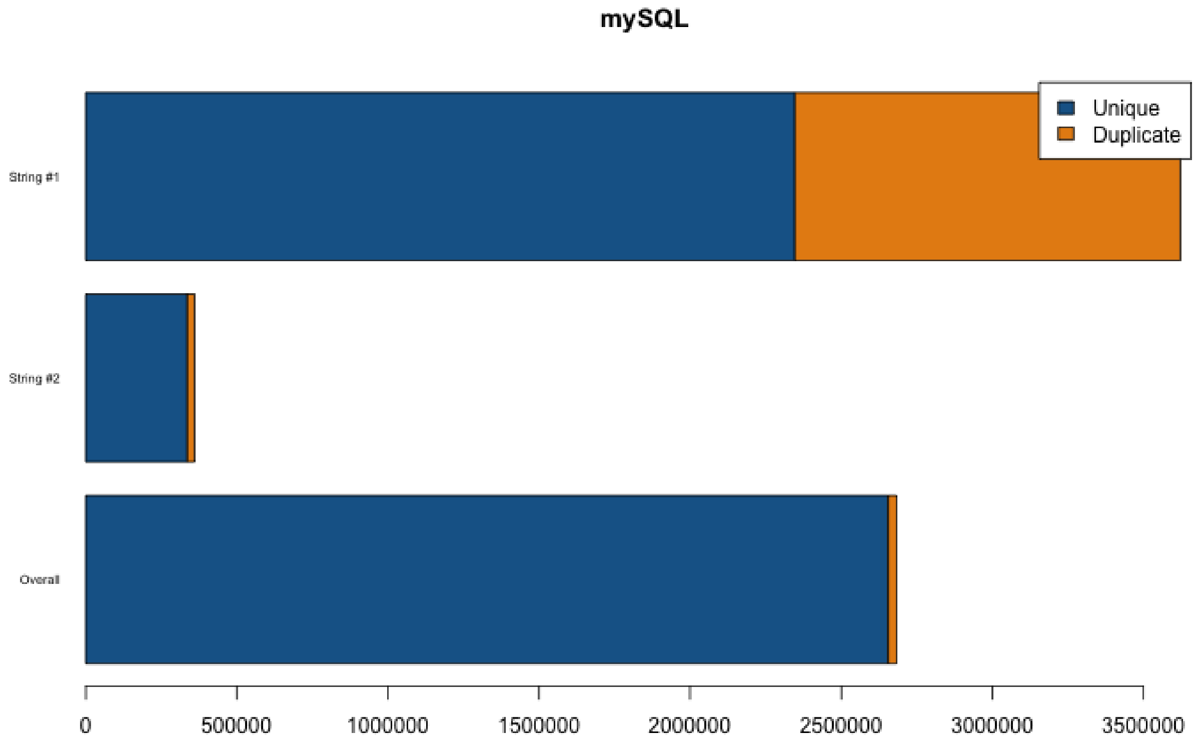


Figure 11: Unique and Duplicate records imported in mySQL and their relative size

Table 15: Number of BCs into multiple resources.

Number of resources considered	Amount of BCs
1	2 405 659
2	172 217
3	60 037
4	16 571
5	873
6	8

et al., 2009). Among the 866 abstracts, 1.5% of them were manually annotated as non-relevant.

The kernel of the SVM (radial basis) and the hyper-parameters gamma (equal to 0.0001061684) and costs (equal to 100) were tuned over a grid of specified values in order to get the best combination which minimized the misclassification rate. The initial misclassification rate was equal to 95.27%.

The size of the final training set was of 2 267 abstracts. The accuracy was computed using a 10fold-crossvalidation procedure. The final accuracy was equal to 98.26%.

A total number of 1 670 088 abstracts in english language with more than 700 characters were considered for relevance-non relevance classification.

In the Table 17, the number of abstracts classified as relevant was reported according to the bibliographic database.

In table 17, references that are present in more than one resources are counted more than once.

Overall, 1 649 076 out of 1 670 088 were classified as pertinent.

To assess the accuracy of the SVM classification, on a randomly selected subset of 20 no-relevant abstracts, the SVM turned out to be 100% accurate. On a randomly selected subset of 100 relevant abstracts, the SVM turned out to classify correctly 99 abstracts. Based on this data, reported in table 18, the recall, or the sensitivity in identifying relevant paper, is 99%. Specificity is 100%. Precision (or Positive Predictive Value,i.e.

Table 16: Overall number of abstracts retrieved and the number of abstracts retained after the checks in step 2

Resource	Total N of abstracts	no–missing abstracts	$>$ 700 chars length abstracts	english abstracts
IEEE Xplore	315 768	308 829	208 199	208 186
JSTOR	298 973	213 212	165 637	158 625
Science Direct	167 276	151 100	121 991	121 743
ACM	446 477	446 124	230 595	230 583
Arχiv	47 501	47 501	31 973	31 967
CiteSeerX	20 176	20 176	12 574	12 538
DOAJ	44 624	41 824	35 995	34 367
PsycInfo	2352	2349	2261	2261
WOS-SSCI	108 790	105 845	92 785	92 778
PubMed	185 542	177 927	164 075	164 014
MEDLINE	101 602	98 363	91 464	91 464
WOS-SCI	29 988	28 788	24 459	24 457
Scopus	257 768	249 805	182 776	182 758
REPEC	27 273	27 271	19 789	19 750
MathSciNet	22 358	14 711	11 655	6960
WOS-AHCI	1487	1111	933	933
WOS-CORE	271 737	263 582	209 642	209 627
Cochrane	1540	600	567	567
CIS	15 205	0	0	0
CINAHL	13 345	6656	5995	5992
Econlit	11 767	9470	7048	7036
Ingenta Connect	67 881	67 880	64 394	63 482

Table 17: Number of abstracts classified as relevant by SVM according to bibliographic resource

Resource	N
ACM	230 050
Arχiv	32 596
CINAHL	5738
CIS	0
CiteSeerX	12 466
Cochrane	1160
DOAJ	33 930
Econlit	6973
IEEE Xplore	207 179
Ingenta Connect	68 477
JSTOR	155 004
MathSciNet	14 417
MEDLINE	89 890
PsycInfo	3448
PubMed	191 028
REPEC	19 487
Science Direct	120 077
Scopus	320 798
WOS-AHCI	922
WOS-CORE	291 948
WOS-SCI	108 630
WOS-SSCI	14 002

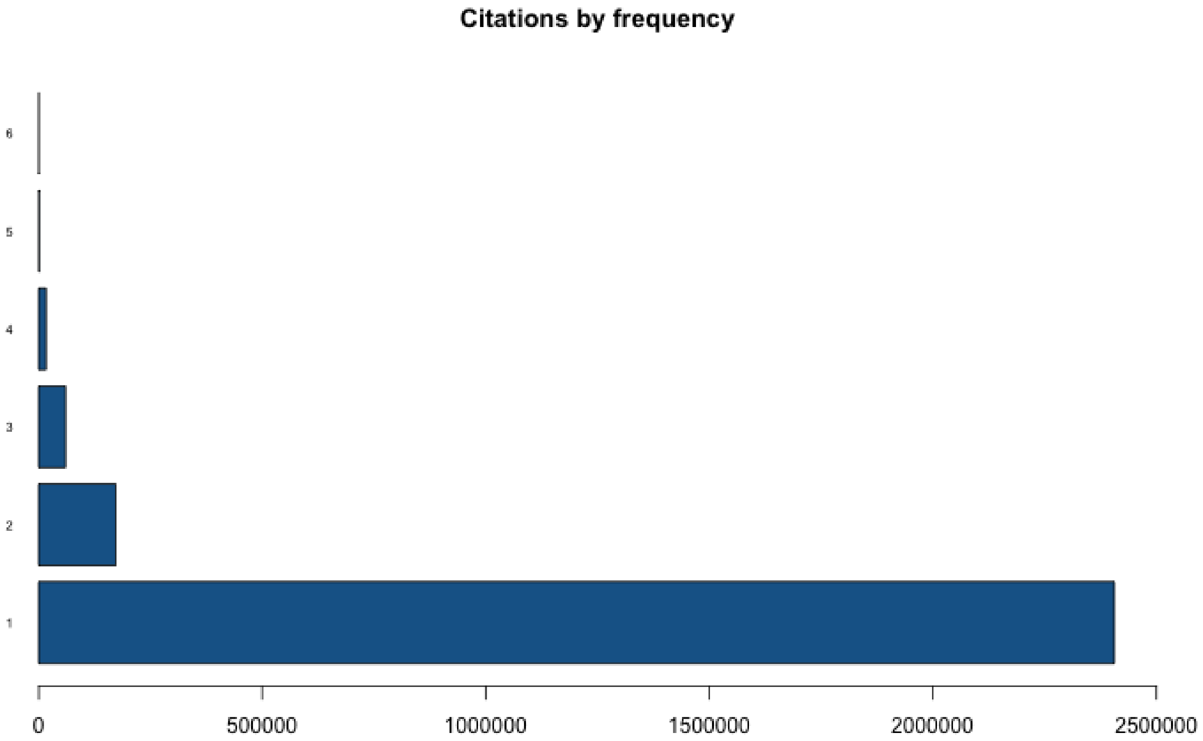


Figure 12: BCs belonging in multiple resources.

Table 18: The specificity and sensitivity of the relevance score used to identify relevant/non-relevant

Actual/Predicted	Relevant	Non-relevant	Total
Relevant	99	1	100
Non-relevant	0	20	20
Total	99	21	120

the probability a randomly selected abstract scored as relevant is really relevant) is equal to 100%. Negative Predictive Value (probability to get a real non-relevant, given a predicted non-relevant paper by the score) is equal to 95%.

SVM topic classification and scoring

The following topics were chosen among those listed in sec. 1.2.6.

- algorithms
- classification
- clustering
- computation
- decision
- discovery knowledge
- efficiency
- expert
- food
- forecasting
- hybrid
- missing values
- optimization
- regression

- risk assessment
- robustness
- sample size

The vector representation of the topic-concept is reported below:

algorithms=c(initial, identification, prediction, attribute, complicated, decomposition, interpolation, multiobjective, posterior);

classification=c(offline, symbol, bootstrapping, resampling, auc, crossvalidation, roc, hematuria, nocturia, pathologists, administrative, asymptomatic, bisphosphonate, bisphosphonates, nuclear, compliance, discriminated, inputlayer, pneumonia, stepwise, accurately, inflammatory, diseases, assessed, pathways, contribution, resuscitation, severity, trauma, fuzzylogic, emergency, administration, quartile, develop, biomarkers, surgical, breast–cancer–specific, nodal, transarterial, pharmacokinetic, chemoembolization, finder, assess, enrichment, crises, endothelin, visits, maxent, c–statistic, multi–wavelet, curve, infinity, mesothelin, ovarian, up–regulation, modelled, malignant, derivation, chest, addition, signs, arthritis, cord, neoplasia, traumatic, confounders, diagnostic, glycan, glycomics, spectrometry, empirical, boost, readmission, reason, visit, polarimetric, wishart, high–sensitivity, masses, characterization, fuzzification, significance, standardization, incident, incidents, f–measure, recall, inhibition, intermodality, cohort, bipartite, objective, kernel-rank, evaluate, hypovolemia, hypovolemic, multisensor, malware, classifiers, dissolved, maximization, benign, lesion, specificity, mutation–based, hyperplanes, resubstitution, classifier, localisation, intercept, varying, iso–performance, lift, breast, clinical, improve, invasive, trial, cancer, women, models);

clustering =c(antimicrobial, atom, atomistic, atoms, bilingual, bimetallic, biotyper, bonds, bruker, clustering, clusters, corpora, creation, degranulation, dislocation, dispersion, divisive, duplication, elliptic, errors, explore, galaxies, genetic, hexagonal, infrasonic, k-modes, loadings, marginality, merge, neutron irradiated, outbreaks, parahox, paralogon, pathogens, purity, radius, relational, scan, semisupervised, structures, surges, surveillance, tables, tick, validity, vertebrate, ward);

computation=c(analog, anticulture, api, articles, artificial, attributable, balanced, binned, binning, bioinformatics, biological, biology, challenges, characterization, checking, cholesky, coefficients, computation, computational, computationally, computing, covariables, data–intensive, disciplines, discretization, drift, ecological, ecology, elliptic, exomecopy, exonic, focus, generalized, genotypic, gradient–based, gradients, hertz, identity, implementations, indicator, inhomogeneous, instance, intelligent, intensive, interaction, intruder, involving, journal, kinship, legged, literary, literature, matrix, methodologies, modeling, nanoscience, nanotechnology, , node, out–of–core, paradigms, preconditioned, predict, prediction, propensity, protein, residuals, residues, ridge-regression, soft, special, sprint, stickers, stranded, strands, streamline, submissions, symmetrized, systemic, time, took, turing, ultrascan, unbalanced, update);

decision=c(access, acute, alfa, arc, attribute, attributes, authorities, balloon, barley, behavioural, random, bucking, campaigns, categories, chose, collective, colonies, combinatorially, complicated, considering, core, corn, diagrams, diesel, digital, doctors, ecosystem, elephants, end-of-life, enroute, episodes, farm, farmer, farmers, finance, forecast, forests, functioning, fund, homogeneity, irrigate, irrigation, jobs, knowledge-based, last, lenders, makers, making, military, motivation, need, nursing, occupational, operations, pasture, prediction, propose, provision, psychotic, purchase, realm, reasons, refocus, relocation, rpd-agent, season, seekers, sensory, slower, societies, staffing, streamflow, supply–demand, tactical, trapped, value, value-focussed, wind, within–cluster, work, workload);

knowledge–discovery=c(abstract, abstracted, abstracts, academy, adequate, ad-hoc, affecting, agonists, argue, attribute-oriented, become, behaviors, benchmark, big–picture, biophysical, book, boundary-spanning, brain-writing, brand, branders, brands, cards, category, cause-consequence, cerebellar, chance, clear, clones, colleagues, communications, confrontation, conservation-based, consumers, contextual, control-flow, core, creative, creativity, cross-functional, customers, day-to-day, descriptive, development, directory, disciplines, discrimination, discuss, discussed, distancing, driven, eco-tilling, efforts, employees, enhance, enterprise, enterprises, entitled, example, experimentation, expert, explanation, exploratory, explores, extension, extensions, external, failures, faults, firm, firms, formal, fragment-based, fundamental, governance, harm, hemiparetic, highvalue, idea, ideas, identifications, immunoprecipitation, implementation, implementations, informal, innovating, innovation, innovations, innovative, insights, institutions, internal, internet-drafts, interviews, intrusive, issues, justification, kind, kinds, know-how, learning-before-doing, learning-by-doing, malicious, manage, management, managers, managing, marketplace, meaningful, medicinal, members, mindset, modern, money, monte-carlo, motif, move, neighbor, neurofuzzy, nuclear, nucleotide, offer, ontology, opportunities, organization, organizational, organizations, pathologies, pool, position-specific, practices, pre–plant, prestige-oriented, primitives, principles, product, products, progress, protocol, representing, requesters, retrieve, rewards, roles, routines, science, scientific, screens, sequence-based, service, signal-to-noise, silico, simplification, single-copy, skills, stepdown, stepup, still, strategic, subgroup, succeed, success, successful, suggest, suggests, support, sustained, technology, transcript, understand, understandings, understood, updated, validated, variety, vision, vitamin, walnut, well-known, workers, written);

efficient=c(abstract, abstracted, abstracts, academy, adequate, affecting, another, assets, behaviors, behaviour, biogenesis, biomimetic, biosorption, bogac, bottom, bottoms, boundary-spanning, brain-writing, brand, branders, broadcast, calculus, cards, centric, coding-decoding, communications, confrontation, consumers, core, cross-functional, currencies, currency, day-to-day, delivering, destruction, diffuser, disciplines, display, dissertation, distancing, efficiency-based, efficiently, efforts, electricalmechanical, emergent, employees, encouraging, enhance, entities, envelopment, epsilon-constraint, exergy, exploration, extension, extensions, failures, faster, financed, firm, firms, formal, frontiers, fundamental, gastruct, genome-pop, haplotype-disease, higherlevel, highvalue, hypergraphs, idea, ideas, innovating, innovation, innovations, innovative, interacting, interviews, investment, know-how, k-partitioning, latter, launches, launching, leaders, learning-before-doing, learning-by-doing, limited, lnorm, macroeconomic, macrolevel, managing, marketplace, mature, meaningful, mentioned, metapopulation, microlevel, micropump, mindset, minimizers, offer, operational, opportunities, optimised, organizational, organizations, out-of-core, place, polyplexes, portfolio, portion, power-train, practices, prestige-oriented, principles, privately, progress, proposes, referenced, relatively, reliance, rely, reproducing, requests, rewards, risk allocation, roles, savings, semifree, share, simplification, simulating, skipping, slice, slicing, stepdown, stepup, still, streamline, subgraph, subgraphs, subpopulation, subsequences, subunits, supervisory, trenches, tube, twosided, understand, understood, unit–cells, view, vision, voxels, waveguide, weeder);

expert=c(agreement, aids-defining, allocation, alpha-agglutinin, ambient, amino, amputation, analysts, asian, assess, assessment, attribute-generalization, benchmark, cart, characteristics, chase, cleaning, cloud, compositionally, computerized, connectionist, connector, consensus, contents, crashinvolved, critic, critics, crossdomain, delphi, delta-function, desorption, diagnoses, diesel, dietetics, disappearance, discovered, discrepancy, discussion, disease-resistant, disease-susceptible, disjoint, dna-binding, domain, driving, dyspepsia, elderly, encodings, ends, entropic, exercises, experience, expert, expertise, experts, explanation, factin, families, feedback, finite-difference, formulation, fourier, fractional, framebased, frames, free, funcional, functioning, homogeneity, homogeneous, homologous, image-based, interdomain, intervention, intuitive, isoplotter, jobs, kinase, knowledge, labeled, learn, linguistic, machinelearning, magazine, majority, markup, mathematical, membranebinding, monolayers, multidomain, multitask, nonconvex, nursing, nutrition, occupational, operate, opinion, opinions, options, organ, overall, perceived, persistence, persons, players, polybayes, powerlaw, practice, pragmatic, preclinical, presents, printing, prion, prions, probabilities, professionals, proteins, proteomes, question, radiologists, rankings, rate, reflective, reformulation, regimens, relevance, research, returns, rnabinding, rulebased, satisfaction, scored, sectors, significativa, spurious, statement, statistics, structured, superfamily, ssyntax, taxonomic, technicians, topics, tortuosity, traffic, transporter, tuberculosisdiagnosis, usability, validate, vocabulary, vote, voting, wellbeing, withincluster, words, zoonoses);

food/nutrition=c(addiction, affect, alimentares, allele, allergens, antioxidative, assurance, bioenergy, biomagnification, bmi, body, changing, choice, clinoptilolite, colonies, competition, cook, cooking, cultural, days, described, diabetes, diagnosis, diet, dietetics, eater, eating, ecosystem, endogenous, eutrophication, expert, exploitation, exploration, exploring, fed, final, fodder, following, foods, foragers, foraging, gender, generaciones, gridbased, guidelines, health, homogenate, ibd, increase, increased, insulin, intake, intakes, international, leptin, lifestyle, long, , mesh-free, metabolic, microbial, microbiological, neighbourhood, neonatal, nutritional, obese, obesity, obestatin, obtaining, overeating, palatable, participants, patch, patches, patterns, perishable, personal, photoperiod, postpartum, predation, predator, predators, prediabetes, preparacin, processed, producing, professionals, recommendations, related, reported, restricted, richness, schemas, scrounging, searching, shocks, short, siberian, signaling, situations, sleep, soybean, standards, status, subject, subjects, taboos, tactics, taste, treatments, trophic, unrestricted, utilization);

forecasting=c(achieve, adjusters, arima, autoregressive, basins, random, blooms, brownian motion, catchments, closing, cointegration, compartments, competitions, conflict, consensus, consumer, created, decisions, define, disadvantages, discourse, downturn, exchange, expenditures, expressions, faced, false, fama, financial, fluctuation, fluctuations, forecasts, generalization, geometric, handset, harvesting, hydrologic, impact, independent, inflow, injuries, injury, intervals, intraday, knearest, literature, markets, multiscale, neighbours, occupant, occupants, occurrence, operation, performs, period, precipitation, prices, probabilistic, profits, project-profit, rates, realtime, regression-corrected, release, respective, river, season, series, short-term, skill, snowpack, socioeconomic, spring, stock, svm, tourism, trading, transparent, universe, vapor, wind);

hybrid=c(alfa, analysis, attribute, building, comprising, distinctiveness, expansive, final, heterotic, hybrids, indiscernibility, inner, layer, lights, manual, metallic, microhybrid, none-mbedded, nonheterotic, optilux, parallel, retrievability, robotic, robots, sectioned, seeding, singleobjective, spurious, stainlesssteel, stent, stepwise, ultralume, wire);

missing values=c(accomplished, attribute, bushings, calling, chromophase, closest, correctly, corrupted, covariability, dairy, dealing, empty, estimations, extra, imputation, imputations, impute, imputed, imputing, indiscernibility, inpainting, knn-cat-impute, microarray, mim, missingness, mixedattribute, nearest, phasing, quartet, regressionbased, reliable, responsible, standard, superresolution, tagging, ttest, underestimated, values);

optimization=c(areas, controller, discuss, fuzzy, genetic, membership, noisy, problems, robot, settings, solving, stochastic, targeting, watershed);

regression=c(blup, framework, marginal, noise, ridge, robust, selection, shall, significance, squared);

risk assessment=c(appropriate, aquatic, assessing, autism, autistic, aversion, base-case, capability, capital, cause, caused, choice, cities, citizens, city, coastal, communities, community, comprehensive, computerized, concept, conflict, conflict-risk, considering, deficiencies, development, developmental, disabled, efforts, evidence, executive, experts, external, extrascientific, fibromyalgia, generalization, gradation, hazards, high-risk, histological, homogeneity, identity, inadequate, includes, inner, inquiry, instrument, intellectually, inundation, leaching, lifestyle, living, low-risk, machinery, macroporosity, making, managing, master, methodological, municipality, n-butane, neighborhood, no-spam, obesity, observation, opls, opportunities, oral, others, outer, parks, participation, percent, pgic, pilot, place, planning, plasmaporosity, pollutants, position, practitioner, prevention, probabilistic, providing, purposes, rankings, rating, regeneration, regional, renewability, review, risk-informed, safety, satisfy, sience, simultaneously, situation, social, soil-landscape, stakeholders, stand, stenosis, strategic, strategies, strengths, stronger, subcompartment, subterrain, sustainability, systemic, tailored, teachers, thesis, threats, timedependent, toddlers, translated, uncertain, uncertainty, urban, urbanism, utility, validated, weaknesses, web-based, within-cluster, worksite);

robustness=c(algebraic, antagonistic, biology, boiler, cancellation, canonical, changeability, code, codes, codon, consumer, crosstalk, damping, dedicated, diameters, distribution, disturbance, disturbances, elastic, evolvability, extrinsic, fidelity, fluctuations, formulations, frameworks, functionals, gate, imperceptibility, interval, intrinsic, likely, low-level, marker-locus, multiproduct, mutational, nodal, noise, noises, organization, parameter, penalties, perform, phenotype, portfolios, pre-mirnas, quadratic-optimal, reason, regression, robuststable, scalefree, shared, small-scale, smoothers, sparsity-aware, stabilizability, stabilization, standard, synthetic, theories, transcriptiontranslation, fuzzy-model-based, uncertainties, unmodeled, variable-length, varied, viroid, visualisation, watermarking);

sample size=c(bartletts, class, clinically, clumped, cochrans, complexity, differentially, equal, equivalence, hypertension, jackknife, learning, lifetimes, many-to-many, multi-rule-based, national, nhanes, number, occurrence, one-to-one, open, pathologic, prevalence, radiographic, reasoner, selfreported, semiexposed, size, sizes, smaller).

All the words with a correlation coefficient greater than 0.3 with term *algorithm* were used to identify the abstracts to label.

In the Table 19 and 20 the number of abstracts classified as relevant and labeled with one of the topics above is reported.

Overall a total of 217 915 abstracts were labelled. This is reasonable in view of the fact that most of the retrieved title and abstract provide information only on the particular type of MLT involved and not on the specific kind of statistical or methodological aspects addressed (if any) in the paper. In all such cases, the information contained in the abstracts does not consent to categorize abstracts.

Based on the labelled abstract, the trend over the years of the topics selected is shown in Figure 13. From the figure is evident the exponential growth of topics like *discovery knowledge*, *clustering* and *food* in the 20's years. On the other hand, trend of topics like *algorithms* or *risk assessment* appears to be constant over time.

The cross classification of topics is thus shown in Table 21.

1.3.3 Synthesis of results

Overall, the following features have been provided by the procedure adopted:

- 3 982 212 references retrieved and analyzed;
- MYSQL DB of 2 655 365 references imported after duplicate exclusion;

Table 19: Number of abstracts labelled according to search string.

Keyword	String #1	String #1 ∪ string #2
algorithms	1430	1441
classification	24 568	25 015
clustering	12 321	12 582
computation	21 709	22 270
decision	41 364	41 571
discovery knowledge	9185	9729
efficient	22 204	22 400
expert	35 105	35 903
food	16 840	17 310
forecasting	11 879	12 282
hybrid	5566	5647
missing values	10 177	10 421
optimization	4177	4191
regression	5183	5196
risk assessment	10 002	10 484
robustness	13 539	13 950
sample size	17 997	18 149

Table 20: Abstracts classified by topic

	Labeled	Not labeled	Total
non considered	NA	985 277	985 277
no relevant	4845	16 167	21 012
relevant	213 070	1 436 006	1 649 076
total	217 915	2 437 450	2 655 365

NA := Not Applicable

Table 21: Cross-classification of topics

	alg.	class.	clust.	comput.	decision	disc. knowl.	efficient	expert	food	forec.	hybrid	miss. val.	opt.	regr.	RA	robust.	sample size
algorithms	1441	16	24	76	61	4	60	94	39	35	10	18	376	5	21	21	41
classification	16	25 015	536	1016	1488	213	1113	1077	755	363	120	373	110	175	269	507	892
clustering	24	536	12 582	359	1278	146	526	1095	282	244	270	252	78	109	493	248	214
computation	76	1016	359	22 270	1339	427	859	1099	692	368	247	264	197	707	455	572	877
decision	61	1488	1278	1339	41 571	2044	2420	1980	1512	792	329	651	300	357	1119	588	1733
discovery knowledge	4	213	146	427	2044	9729	319	432	339	124	90	131	56	76	467	170	147
efficient	60	1113	526	859	2420	319	22 400	1959	1652	493	280	277	100	215	318	609	598
expert	94	1077	1095	1099	1980	432	1959	35 903	1019	777	719	386	200	385	2641	543	710
food	39	755	282	692	1512	339	1652	1019	17 310	278	99	172	89	32	403	386	955
forecsating	35	363	244	368	792	124	493	777	278	12 282	135	211	76	62	223	208	271
hybrid	10	120	270	247	329	90	280	719	99	135	5647	62	41	77	350	201	404
missing values	18	373	252	264	651	131	277	386	172	211	62	10 421	43	78	124	192	132
optimization	376	110	78	197	300	56	100	200	89	76	41	43	4191	17	21	171	182
regression	5	175	109	707	357	76	215	385	32	62	77	78	17	5196	153	431	208
risk assessment	21	269	493	455	1119	467	318	2641	403	223	350	124	21	153	10 484	130	289
robustness	21	507	248	572	588	170	609	543	386	208	201	192	171	431	130	13 950	558
sample size	41	892	214	877	1733	147	598	710	955	271	404	132	182	208	289	558	18 149

- SVM classification of relevant and no relevant score on a subset of 1670088 abstracts: the subset was based on English abstracts with more than 700 characters;
- accuracy of SVM in detecting non relevant abstracts was maximized;
- 1 649 076 abstracts were classifies as relevant;
- abstracts classification according to general and methodological aspects and concepts were applied, based on 17 different categories labeling 217915 abstracts;
- an overall 213 070 relevant abstracts were labelled by NO analysis.

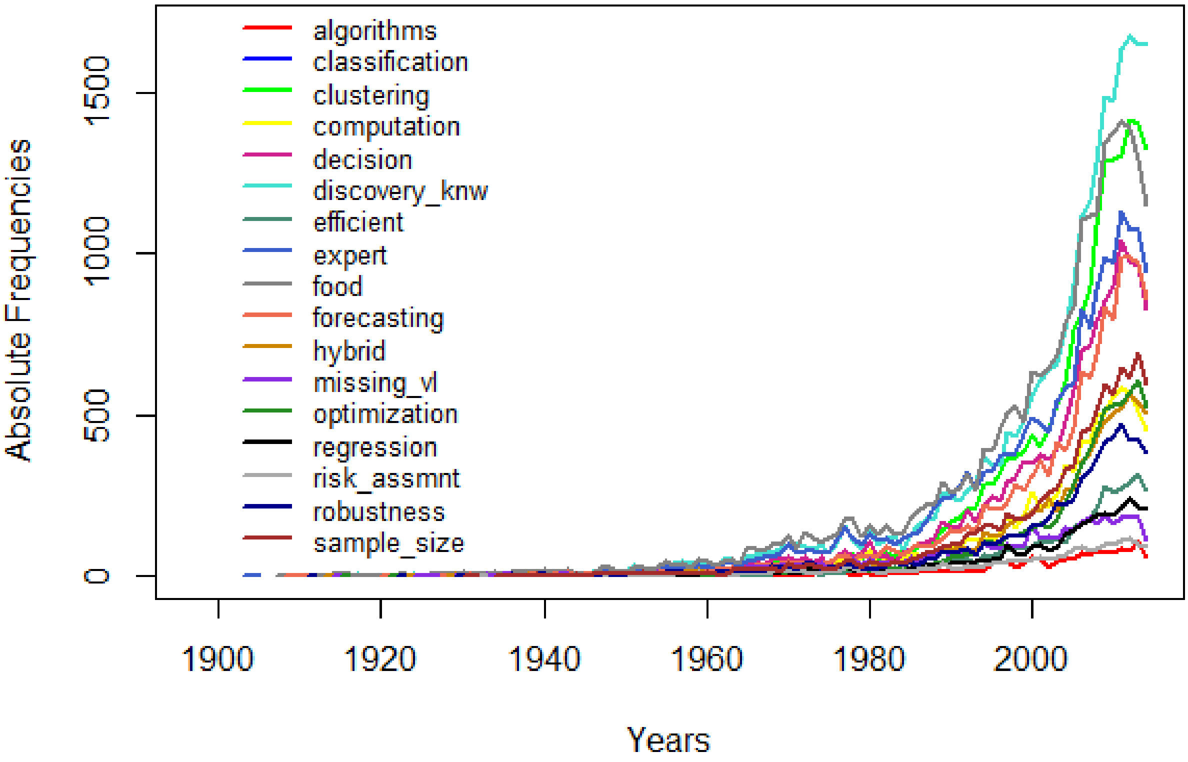


Figure 13: Trend of publication in MLT fields, according to classification keys.

The overall procedure is summarized on the flowchart 14.

1.4 Discussion

1.4.1 Limitations

Retrieval

In most cases search has been conducted without any restriction. The only exceptions have been applied to resources manually retrieved and resources originating from R-Scripts that need explicit *date* settings: in these cases the researches were restricted to records retrieved from year 1800 to year 2015.

A limit in the retrieval is that not all resources can be automatically approached, and even if an automatic approach is allowed (e.g EndNote) the retrieval can be not complete. A further limitation arise when considering the encoding adopted by each resources, that slows the data merging procedures.

Besides these theoretical limits there are some technical limitations, linked to the different strategies of retrieval:

- EndNote main limitation is its incapability of handling libraries of millions of records, becoming not only very computationally slow, but also losing features like the import-export of libraries or the reference updates.
- R is an engine tha works exploiting exclusively the RAM memory: very complex processes can be run very fastly, but the time of processing increases exponentially accordingly to the size of the databases.
- Manual retrieval has its main limitation in the time and costs.

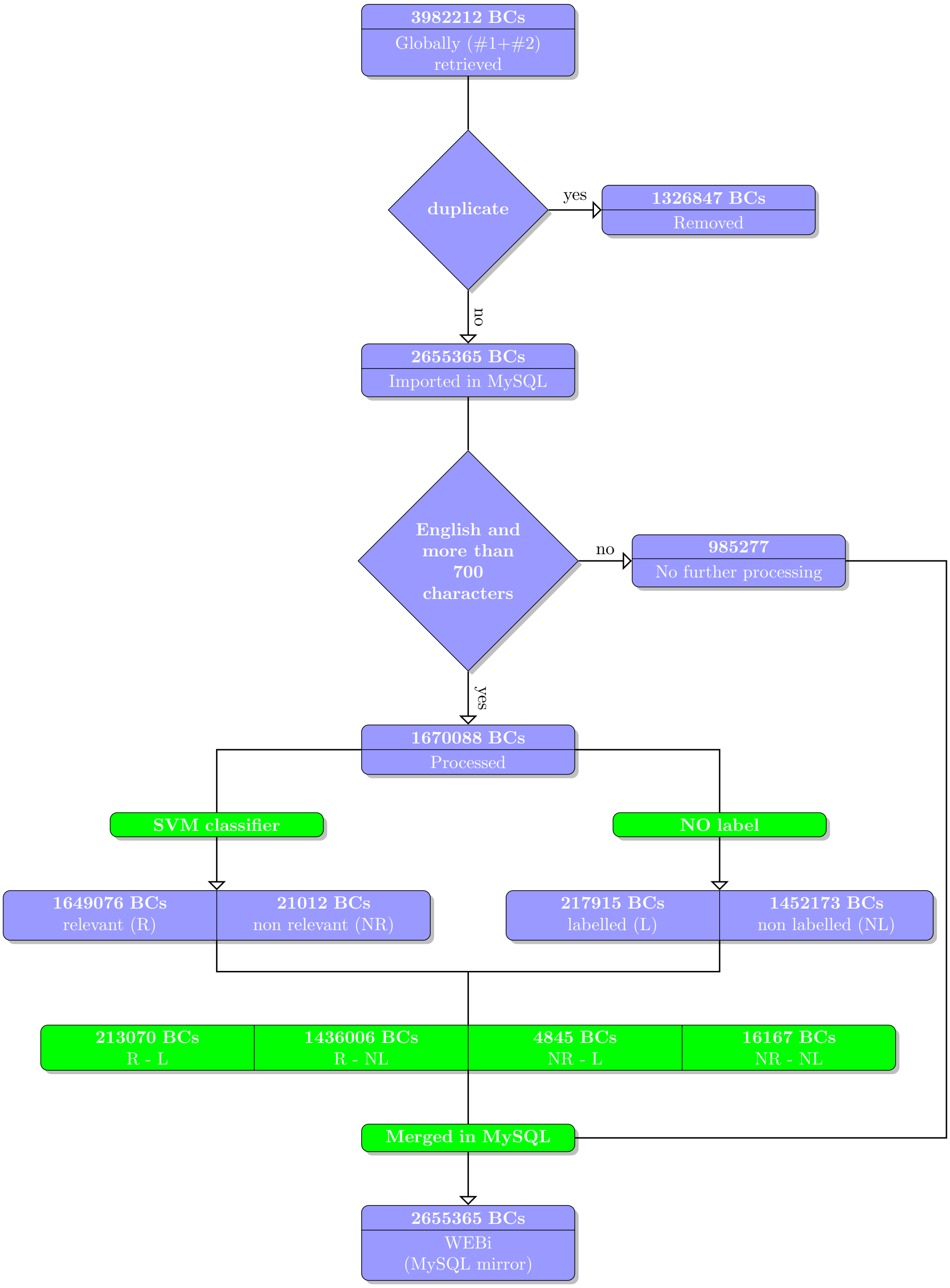


Figure 14: Flow of selection and processing of BCs.

SVM

One first limitation in the AL procedure is related to the procedure followed for manual annotation. In fact, manual annotation was done by one reviewer only. A full procedure should have been performed by two reviewers independently and in case of disagreement, consensus should have been sought by discussion and in case of further disagreement a third reviewer should have been consulted.

A more extensive direct validation of SVM classifier is needed by selecting a larger sample of abstracts and assessing the correctness of the classification.

NO

Documents annotation was done by using natural language–processing, in particular using only information about co–occurrence of terms in abstracts. NO, which is widely recognized because it allows for easy implementation and efficient processing of huge amounts of texts, was based on measures of similarity or relatedness between the terms used to define topics and the terms distribution in the abstracts.

The main limitations in the actual analysis can be summarize in:

- the lacking of a validation of documents annotation, by checking both a sample of abstracts annotated and a sample of abstracts not annotated;
- an improvement of the co–occurrence analysis could be achieved by using thesaurus to describe the concepts. In the analysis carried out, the concepts, i.e. the words used to describe the topics, were found out looking for terms associated to the label terms and occurring into the abstracts. Using thesaurus would allow also synonyms to be mapped in the same concept, which would reduce the noise caused by natural language variation.

1.4.2 Lessons Learnt

EndNote use

EndNote is a tool at the same time very powerful but very slow for our purposes. The ability to automatically connect to the resources linked, to create groups of references, as well as the ability to create customized output formats for export, were very usefull features. On the other hand it is very slow in downloading of references, not precise with respect to the overlap between the reference "found" and those actually "downloaded" and with coding issues that have led to export mistakes to be reviewed *at hoc*, as well as the *practical* impossibility of managing libraries with data exceeding 100 MiB with no one of the computers at our disposal (see. 1.2.3). In future similar work, EndNote can be used as an excellent tool for the preliminary investigation, as well as to manage the standard references of the work, but for the download itself of the BCs subject of the work we will try to make more use of algorithms written in a language programming, as R, which allow more flexibility and structure.

Relevance of relevance

Given the estimated small number of non-relevant abstracts (1.5% estimated on the initial training set of 866 randomly selected abstracts) the usefulness of a relevant score computed using a machine approach - which is not error free - it achieves an accuracy of 98.26% is debatable. However, since it provide a probability of relevance (thus is not a dicothomic indicator only) it can be regarded as a tool for help in managing all the abstracts retrieved. Indeed, since the relevance score is on a continuous scale, it can allow for setting greater thresholds than 0.5 when deciding to explore only a limited subset of the entire database. For example, if a more tailored search is needed, one could decide to explore only the subset of the abstracts with a relevance score greater than 0.7.

1.4.3 Conclusions

A huge number of references related to MLT has been retrieved from available bibliographic resources and they have been imported into a sql database, which is available on a online platform.

SVM classifiers were used to facilitate navigation into the bibliographic database by adding a relevance score, in order to help filtering out abstracts no–relevant to the field.

Furthermore NO analysis was carried out to label abstracts according to a set of topics that are common to MLT and RA.

The main limitations of the actual work rely on the need of further validation tasks to assess the performance of the SVM classifier and the *co–occurrence* annotation analysis, which could be improved by adopting thesaurus to broaden the dictionary (the terms) used to define topics.

ELS bibliography

[Alp14] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.

[ARK10] Itamar Arel, Derek C Rose, and Thomas P Karnowski. "Deep machine learning-a new frontier in artificial intelligence research [research frontier]". In: *Computational Intelligence Magazine, IEEE* 5.4 (2010), pp. 13–18.

[AZ12] Charu C Aggarwal and ChengXiang Zhai. "A survey of text classification algorithms". In: *Mining text data*. Springer, 2012, pp. 163–222.

[Ben09] Yoshua Bengio. "Learning deep architectures for AI". In: *Foundations and trends® in Machine Learning* 2.1 (2009), pp. 1–127.

[Ber+12] Yoav Bergner et al. "Model-Based Collaborative Filtering Analysis of Student Response Data: Machine-Learning Item Response Theory." In: *International Educational Data Mining Society* (2012).

[BL97] Avrim L Blum and Pat Langley. "Selection of relevant features and examples in machine learning". In: *Artificial intelligence* 97.1 (1997), pp. 245–271.

[Cla93] William J Clancey. "Notes on "Epistemology of a rule-based expert system"". In: *Artificial intelligence* 59.1 (1993), pp. 197–204.

[Cop83] John B Copas. "Regression, prediction and shrinkage". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1983), pp. 311–354.

[Cor71] Richard M Cormack. "A review of classification". In: *Journal of the Royal Statistical Society. Series A (General)* (1971), pp. 321–367.

[DB03] Fernando De La Torre and Michael J Black. "A framework for robust subspace learning". In: *International Journal of Computer Vision* 54.1-3 (2003), pp. 117–142.

[GE03] Isabelle Guyon and André Elisseeff. "An introduction to variable and feature selection". In: *The Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[Gna11] Ram Gnanadesikan. *Methods for statistical data analysis of multivariate observations*. Vol. 321. John Wiley & Sons, 2011.

[Har+85] Frank E Harrell Jr et al. "Regression models for prognostic prediction: advantages, problems, and suggested solutions." In: *Cancer treatment reports* 69.10 (1985), pp. 1071–1077.

[Hir+02] Lynette Hirschman et al. "Accomplishments and challenges in literature data mining for biology". In: *Bioinformatics* 18.12 (2002), pp. 1553–1561.

[JB04] Aleks Jakulin and Ivan Bratko. "Testing the significance of attribute interactions". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM. 2004, p. 52.

[JK70] Norman L Johnson and Samuel Kotz. *Distributions in Statistics: Continuous Univariate Distributions: Vol.: 2*. Houghton Mifflin, 1970.

[JKB95] Norman Lloyd Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distributions. Vol. 1*. Vol. 2. Wiley New York, 1995.

[JKB97] Norman Lloyd Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Discrete multivariate distributions*. Vol. 165. Wiley New York, 1997.

[JMF99] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. "Data clustering: a review". In: *ACM computing surveys (CSUR)* 31.3 (1999), pp. 264–323.

[Joa98] Thorsten Joachims. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. 1998.

[KZK12] Jochen Kruppa, Andreas Ziegler, and Inke R König. "Risk estimation and risk prediction using machine-learning methods". In: *Human genetics* 131.10 (2012), pp. 1639–1654.

[KZP06] Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas. "Machine learning: a review of classification and combining techniques". In: *Artificial Intelligence Review* 26.3 (2006), pp. 159–190.

[KZP07b] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. *Supervised machine learning: A review of classification techniques*. 2007.

[LD07] Niklas Lavesson and Paul Davidsson. "Evaluating learning algorithms and classifiers". In: *International Journal of Intelligent Information and Database Systems* 1.1 (2007), pp. 37–52.

[Liu+97] Wei Zhong Liu et al. "Techniques for dealing with missing values in classification". In: *Advances in Intelligent Data Analysis Reasoning about Data*. Springer, 1997, pp. 527–536.

[Liu04] Ying Liu. "A comparative study on feature selection methods for drug discovery". In: *Journal of chemical information and computer sciences* 44.5 (2004), pp. 1823–1828.

[Nis+83] Richard E Nisbett et al. "The use of statistical heuristics in everyday inductive reasoning." In: *Psychological Review* 90.4 (1983), p. 339.

[PHL04] Lance Parsons, Ehtesham Haque, and Huan Liu. "Subspace clustering for high dimensional data: a review". In: *ACM SIGKDD Explorations Newsletter* 6.1 (2004), pp. 90–105.

[SIL07a] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics". In: *Bioinformatics* 23.19 (2007), pp. 2507–2517. doi: 10.1093/bioinformatics/btm344. eprint: http://bioinformatics.oxfordjournals.org/content/23/19/2507.full.pdf+html. url: http://bioinformatics.oxfordjournals.org/content/23/19/2507.abstract.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. "Practical Bayesian optimization of machine learning algorithms". In: *Advances in Neural Information Processing Systems*. 2012, pp. 2951–2959.

[SV07a] Mohammad Shami and Werner Verhelst. "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech". In: *Speech Communication* 49.3 (2007), pp. 201–212.

[TG06] Theodore B Trafalis and Robin C Gilbert. "Robust classification and regression using support vector machines". In: *European Journal of Operational Research* 173.3 (2006), pp. 893–909.

[Var+06] Roy Varshavsky et al. "Novel Unsupervised Feature Filtering of Biological Data". In: *Bioinformatics* 22.14 (2006), e507–e513. doi: 10.1093/bioinformatics/btl214. eprint: http://bioinformatics.oxfordjournals.org/content/22/14/e507.full.pdf+html. url: http://bioinformatics.oxfordjournals.org/content/22/14/e507.abstract.

[YP97] Yiming Yang and Jan O. Pedersen. "A Comparative Study on Feature Selection in Text Categorization". In: Morgan Kaufmann Publishers, 1997, pp. 412–420.

2 A classification of EFSA RA published opinions to identify the category of questions most asked to EFSA in its specific remit

2.1 Assessment of the relationships between cluster of words (Topics) from Topic Modeling (TM) and specific RQs addressed by EFSA in its specific remits

2.1.1 Introduction

Objective 2 of this project was to carry out a classification of EFSA RA published opinions to identify the category of questions most commonly asked to EFSA within its remit. In particular it was requested to screen all EFSA Scientific Opinions in order to classify them into clusters of RQs, e.g. risk factor analysis, classification, prediction, etc. Moreover an automated procedure was indicated as the preferred approach to perform this task.

Recently, TM has emerged as unsupervised technique to address the clustering task in text mining (D. M. Blei, A. Y. Ng, and M. I. Jordan, 2003) and it has been shown to outperform other clustering techniques (2013).

The clustering techniques that are more commonly used to perform document clustering are: agglomerative hierarchical and K-means. Agglomerative hierarchical algorithm is often portrayed as better than K-means, even if it is less computationally efficient. Hierarchical techniques produce a nested sequence of partitions, with a single, all inclusive cluster at the top and singleton clusters of individual points at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level (or splitting a cluster from the next higher level). The result of a hierarchical clustering algorithm can be graphically displayed as a tree, called a dendrogram, which for document clustering, provides a taxonomy. Among hierarchical algorithms, agglomerative techniques are more common and with respect to k-means techniques, they do not require the specification of the desired number of clusters (topic).

Recently TM has been successfully applied for discovering the main themes that pervade a large collection of documents (2010).

The core of TM methodology consists in capturing the mutual connection between the documents assuming that each document is comprised of several topics, with some proportions. This is achieved assigning probabilities to each word in the document and assigning a probability distribution over a fixed vocabulary (topic). The intuition behind TM is that documents exhibits multiple topics. For example, a scientific opinion of the Panel on Contaminants in the Food Chain, which provides independent scientific advice on contaminants in the food chain and undesirable substances such as natural toxicants, mycotoxins and residues of unauthorized substances, will contain a vocabulary of terms relevant to contaminants, a vocabulary of terms relevant to the food chain and another one to risk assessment field. If we would highlight the words of this different vocabulary with different colors we would see that the document blends three different topic in different proportions. Some documents might share topics and two documents are semantically closer to each other if they share topics with similar proportions.

TM can be achieved by using LDA algorithm, which automatically learns itself to assign probabilities to every word in the document, thus allowing to classify text into topics (D. Blei and Lafferty, 2006). LDA considers a probabilistic model at document level and works iteratively. Using a Gibbs sampling approach, for each document, LDA assigns each word to one of the topics,for which the number of topics has been chosen. For each word in a document and for each topic it computes the proportion of words that are currently assigned to the topic and the proportion of assignments to the topic over all documents that come from the word and eventually reassign the word to a new topic according to the probability that the topic generated that word. After repeating the previous step a large number of times, Gibbs sampling converges into the posterior distribution of the model parameters or word–topic assignments. So these assignments can be used to estimate:

1. the topic mixtures of each document, by counting the proportion of words assigned to each topic within that document;
2. the words associated to each topic by counting the proportion of words assigned to each topic overall.

One limitation of the LDA is the assumption that topics are unrelated. The independence assumption between topics comes from using a Dirichlet prior over topics, in which the correlation is not taken into account.

To address this limitation, (D. Blei and Lafferty, 2006) suggested to carry out CTM as an effective extension of the LDA by replacing Dirichlet prior with Logistic Normal Distribution, which captures better inter-component correlations (JONATHAN Huang and Malisiewicz, 2009).

2.1.2 Aim

To identify the main RQs addressed by EFSA in its specific remits on the basis of the output of a TM activity.

2.1.3 Methods

A Classification of EFSA Risk Analysis Domains has been carried out adopting a ML unsupervised approach.

Scientific documents as pdf files have been provided by EFSA. Starting from a folder of 4451 documents in pdfformat, firstly the ”main” body text of each Opinion was separated from the Annexes. Using the ”pdftotext” unix command (v3.04) the pdfs were converted into `txt` files. Manually inspecting `txt`-generated files revealed some corrupted `txt` files that were excluded from any further analysis: the final dataset was made by 3744 documents (84.1%) and based on a dictionary of 15475 words. Numeric filenames were then converted into a fixed 4-digits format. The TM was then carried out on these files.

TM was carried out through the parallel application of two different approaches available within R i.e. the LDA and the CTM.

Two different set of 20 topics were obtained respectively using LDA and CTM. Each topic was build up of a set (column) of 100 words.

To obtain a meaningful interpretation of the identified topics two different activities have been performed. In both cases the a priori knowledge of the experts of the Consortium on the scientific activities of EFSA through its Units or Panel teams was always used.

- As a first strategy to address the need of identifying the RQs, for each topic, the whole list of words was inspected to consider their individual meaning, the combination within the list and accounting for the ranking of each listed word. The following categories of items were considered: agents (e.g. microbiological organisms, chemicals, allergens, pests); effects (e.g. efficacy, toxicity, nutritional requirements, safety, claim, spread, exposure); populations of concern (e.g. animals, plants, infants, environments etc.); mentioned statistical technique (e.g. ANOVA).

  The words inspection was used to identify one or more main themes per topic and to compare the contents between topics either within the output (20 topics) of each algorithm or between the two series (LDA, CTM) of 20 topics.

- A second strategy of analysis of the contents of the topics was based on the retrieval of sets of EFSA documents linked to a specific list of 100-word list (topic). For each topic a string of words was created by selecting the top 10 ranked words, assumed to be highly associated with the topic. Plural terms were replaced following the ranking within each list.

  Starting with the 20 topics obtained through LDA, the 10-word strings were used to retrieve 3-4 EFSA documents per topic directly from the EFSAwebsite.thisFor the purpose we used the EFSA’s search engine that allows to execute full text searches of all EFSA’s outputs regardless of format, and when needed to refine the search according to type (scientific opinion in the EFSA Journal, news story, topic, etc), date and scientific panel. The search engine also contextualises search queries to suggest similar or related terms and corrects mistyped queries. A wide time span of publication was preferred in the selection. Each retrieved document was manually screened and classified by two independent evaluators provided by Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle d’Aosta (IZSTO). The following items were considered: type of EFSA publication, Unit/Panel/Committee producing it, main addressed subject and outputs, nature and number of Term of References (TORs), type of assessment (narrative review vs. data analysis), statistical techniques actually employed to address specific TOR. A spreadsheet was created with the findings of this screening activity (Annex I). Moreover a qualitative analysis of the contents of the TOR and the thorough reading of the documents were carried out.

After the completion of the mentioned activities, a first attempt was carried out to characterize and interpret the general content of each of the 40 topics. Combining the findings from the word inspection and the manual screening of the documents, the contents of the topics were compared to identify any overlapping between the

LDA and CTM output. Pairs of topics from the two sets (LDA and CTM) linked by a similar interpretation were searched or distinct subsets of independent topics were defined. Each pair or each distinct topic was assigned with a title based on both word inspection and document screening: this title was considered the best available proxy of the topic's "RQ".When pair of topics were identified, the degree of agreement was assessed. For his purpose, firstly, per each pair of topics, the number of shared terms (out of 100) was calculated; then each common term was assigned with a pair of values corresponding to the ranks associated by respectively LDA and CTM. Based on the pair of rank values assigned to the shared terms a Spearman's rho was calculated per each pair of topics. As a comparison, the same two-step procedure was carried out matching one independent topic from LDA (not previously paired with any CTM topic) with a couple of independent CTM topics. Since the preliminary retrieval of EFSA documents had been done focusing on the 20 LDA topics, a second round of retrieval was carried out focusing on the CTM topics that were not matching any LDA one. Again each document was manually screened and classified. The output of the final screening was used to further refining and finalizing the "RQ" identification by topic. Finally, the relationship between topics automatically identified through the CTM algorithm was checked and interpreted on the basis of their logical plausibility.

2.1.4 Results

Based on the first strategy described in the Materials & Methods (M&M) section and taking into account the main remits of EFSA, agents, effects and populations of concern were helpful in interpreting the combination of the 100 words within each topic allowing a preliminary and provisional identification of general themes (e.g. GMO, AHAW, health claims, etc.). In many cases, apparently a similar theme was shared by more than one topic identified by LDA and/or CTM whereas in other cases the relationship was one (theme) to one (topic). Annexes G and H include the list of words per topic. The application of the second strategy resulted in the retrieval of 91 EFSA documents(control dataset) obtained on the basis of the 10-word top ranking strings (the list of the manually screened documents is reported in the *Second Strategy Bibliography* at the end of the section. In few cases the same document was associated to more than one topic. After having identified the EFSA Panel authoring the retrieved documents, the most represented were NDA (n=20), AHAW (n=10), BIOHAZ (n=9) and FEEDAP (n=9). A group of documents (17) were produced directly by EFSA without any direct or specific link with Panels teams. Out of the 91 documents, only a small proportion was different from Scientific Opinions (i.e. 5 Reasoned Opinions, 3 Scientific Reports, 3 Conclusions on pesticide peer review, 1 Statement).

The number of TORs per Opinion were mostly less than 3 (respectively one in 42 Scientific Opinions and 2 in other 21). When considering how the TORs were addressed, narrative assessments/reviews without any application of statistical techniques were the most common approach (67/91 i.e. 74%) . When statistical or mathematical techniques were mentioned, in 12 cases the analyses were based merely on the application of basic descriptive or inferential techniques; in as many cases, RQs were addressed through a few models ("deterministic", "stochastic", "spatial", "linear regression", "poisson or negative binomial regression", "benchmark dose", "margin of exposure approach", "predictive microbiology growth models").

On the basis of both the analysis of the contents of the TORs and the thorough reading of the documents, the provisional identification of general themes associated to each topic was revised with a final attribution/identification of a title (as mentioned above, the best available proxy for a RQ) per topic. The combination of the series of topics from the two different TM techniques (LDA and CTM) resulted in 28 individually distinct topics that in 12 cases were shared by the twoapproach i.e. the clustering was based on almost the same words. The agreement between paired topics was confirmed by the calculation of the Spearman's rho that ranged between 0.36 and 0.999 (mean=0.77, SD=0.21); the mean number of the shared terms was 71.5 (SD= 20.2). As a comparison the independent LDA Topic 6 (nutritional claims, nutritional safety) shares respectively just 9/100 terms with the CTM Topic 15 (nutritional, health claims) and 14/100 terms with the CTM Topic 19 (nutritional claims / nutritionally requirements): based on this low share no Spearman's rho is worth to be calculated..

The list of the topics by algorithm showing when they were shared by both the subsets is in Figure 15 and in Table 22. The word clouds useful to describe each topic and based on the post probabilities associated to each of the top ranked 30 words are shown in Figure 16, 17 and 18.

A deeper explanation of the reason why we intended such topic as Risk Questions:

LDA *Clusters*

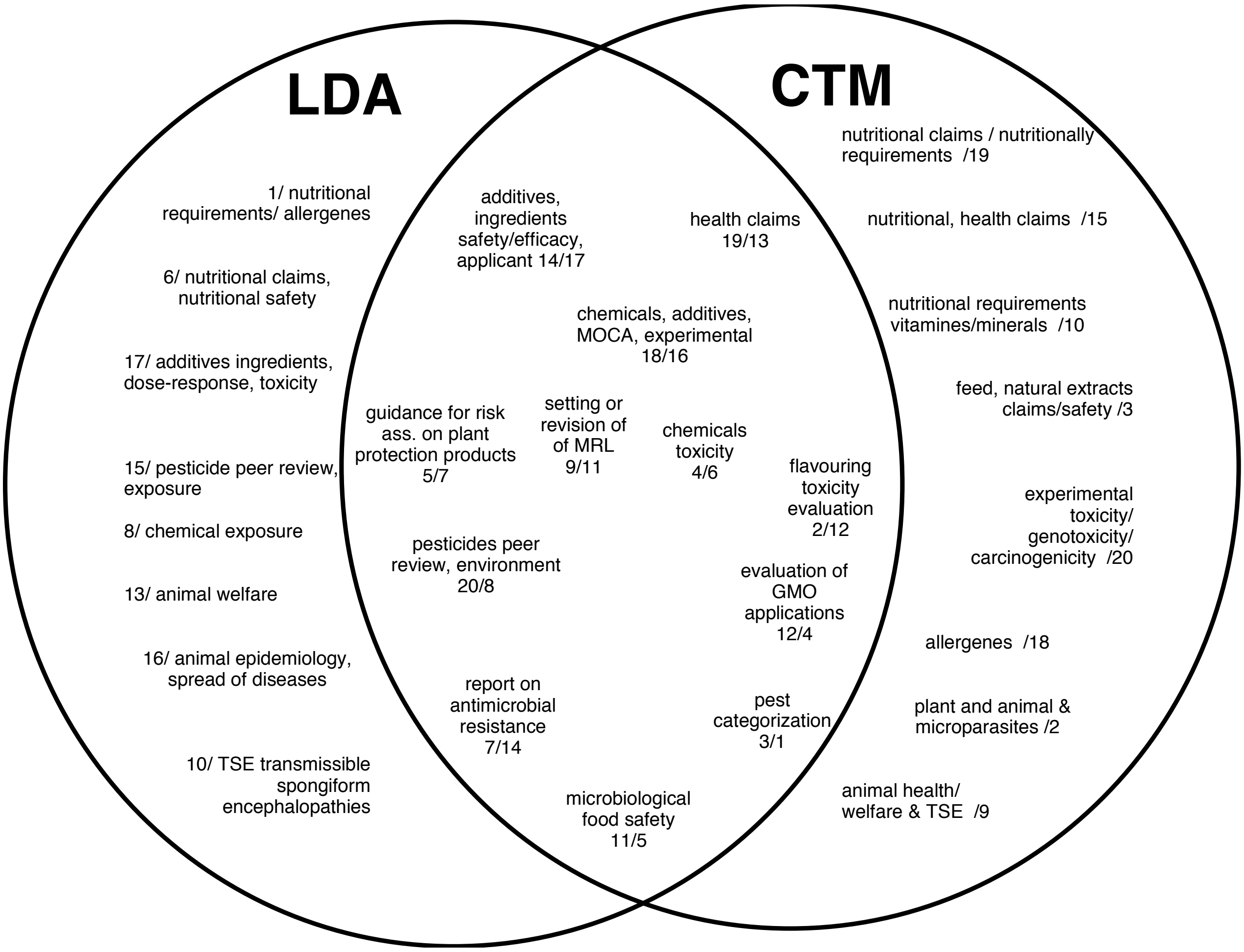

LDA
CTM
1/ nutritional requirements/ allergenes
6/ nutritional claims, nutritional safety
17/ additives ingredients, dose-response, toxicity
15/ pesticide peer review, exposure
8/ chemical exposure
13/ animal welfare
16/ animal epidemiology, spread of diseases
10/ TSE transmissible spongiform encephalopathies
additives, ingredients safety/efficacy, applicant 14/17
health claims 19/13
chemicals, additives, MOCA, experimental 18/16
guidance for risk ass. on plant protection products 5/7
setting or revision of of MRL 9/11
chemicals toxicity 4/6
flavouring toxicity evaluation 2/12
pesticides peer review, environment 20/8
evaluation of GMO applications 12/4
report on antimicrobial resistance 7/14
pest categorization 3/1
microbiological food safety 11/5
nutritional claims / nutritionally requirements /19
nutritional, health claims /15
nutritional requirements vitamines/minerals /10
feed, natural extracts claims/safety /3
experimental toxicity/ genotoxicity/ carcinogenicity /20
allergenes /18
plant and animal & microparasites /2
animal health/ welfare & TSE /9

Figure 15: Risk questions identified through LDA and CTM topic models. Each topic was assigned a title as explained in the body text. Numbers refer to the ordering of the topics as presented in the outputs of the two models (#/ from LDA, /# from CTM). The shared areas shows the topics whose list of words was mostly the same in the specific topics defined either one or the other model. The order of the topics within the picture is based on a subjective attempt to keep closer topics with similar meanings.

Topic 1 (nutritional requirements/allergens): Its main topic is the evaluation of dietary requirements and intakes of infants and young children.

Topic 2 (CTM's 12) (flavoring toxicity evaluation): It focuses on the evaluation of toxicity of flavoring substances previously assessed by agencies other than EFSA.

Topic 3 (CTM's 1) (pest categorization): It focuses on the assessment of the risk posed by pests and their potential to affect crops.

Topic 4 (CTM's 6) (chemical toxicity): It is to evaluate the toxicity of organic and inorganic substances and their thresholds for safety.

Topic 5 (CTM's 7) (guidance for RA on plant protection products): To develop a RA scheme for plant protection products.

Topic 6 (nutritional claims, nutritional safety): Aiming at scientific substantiating a health claim related to particular concerns for fats.

Topic 7 (CTM's 14) (report on antimicrobial resistance): Evaluation of the resistance of the main zoonotic bacterial isolates based on data submitted by Member States.

Topic 8 (chemical exposure): Assessment of dietary exposure to some chemicals such as Arsenic or Lead.

Table 22: Risk questions identified through LDA and CTM topic models. Each topic was assigned a title as explained in the body text. Numbers refer to the ordering of the topics as presented in the outputs of the two models (# from LDA, # from CTM). The matching of parts of the topics between the two lists obtained from respectively LDA and CTM is shown; per each pair the number of shared terms and the Spearman's rho are also reported.

LDA topic	TITLE	CTM topic	Number of shared terms	Spearman's rho
1	nutritional requirements/ allergenes			
2	flavouring toxicity evaluation	12	91	0.9731
3	pest categorization	1	67	0.7329
4	chemicals toxicity	6	48	0.5018
5	guidance for RA on plant protection products	7	84	0.9084
6	nutritional claims, nutritional safety			
7	report on antimicrobial resistance	14	60	0.6605
8	chemical exposure			
9	setting or revision of of Maximum Residue Limit (MRL)	11	100	0.9988
10	transmissible spongiform encephalopathies (TSE) transmissible spongiform encephalopathies			
11	microbiological food safety	5	43	0.3611
12	evaluation of GMO applications	4	97	0.9876
13	animal welfare			
14	additives, ingredients safety/efficacy, applicant	17	81	0.8936
15	pesticide peer review, exposure			
16	animal epidemiology, spread of diseases			
17	additives ingredients, dose-response, toxicity			
18	chemicals, additives, MOCA, experimental	16	43	0.5344
19	health claims	13	77	0.9046
20	pesticides peer review, environment	8	67	0.7543
	plant and animal & microparasites	2		
	feed, natural extracts claims/safety	3		
	animal health/welfare & TSE	9		
	nutritional requirements vitamines/minerals	10		
	nutritional, health claims	15		
	allergenes	18		
	nutritional claims / nutritionally requirements	19		
	experimental toxicity/genotoxicity/carcinogenity	20		

Topic 9 (CTM's 11) (setting or revision of MRL): Aiming at providing a reasoned opinion combining the MRL reviews of some active substances.

Topic 10 (TSE): It focuses on updating TSE epidemiological situation and reviewing risks given by Specified Risk Materials (SRM).

Topic 11 (CTM's 5) (microbiological food safety): It focuses on identifying public health risks related to the food chain and the maintenance of particular parameters such as temperature or correct sampling.

Topic 12 (CTM's 4) (evaluation of GMO applications): It focuses on specific uses of GMOs for food and feed, their import and management

Topic 13 (animal welfare): It focuses on animal welfare in farm, during transport and at the slaughterhouse.

Topic 14 (CTM's 17) (additives, ingredients safety/efficacy, applicant): Evaluation of the safety and efficacy of additives added to the diet of livestock.

Topic 15 (pesticide peer review, exposure): It is asked to take conclusion on the peer review of the pesticide risk assessment of particular active substances affecting animals.

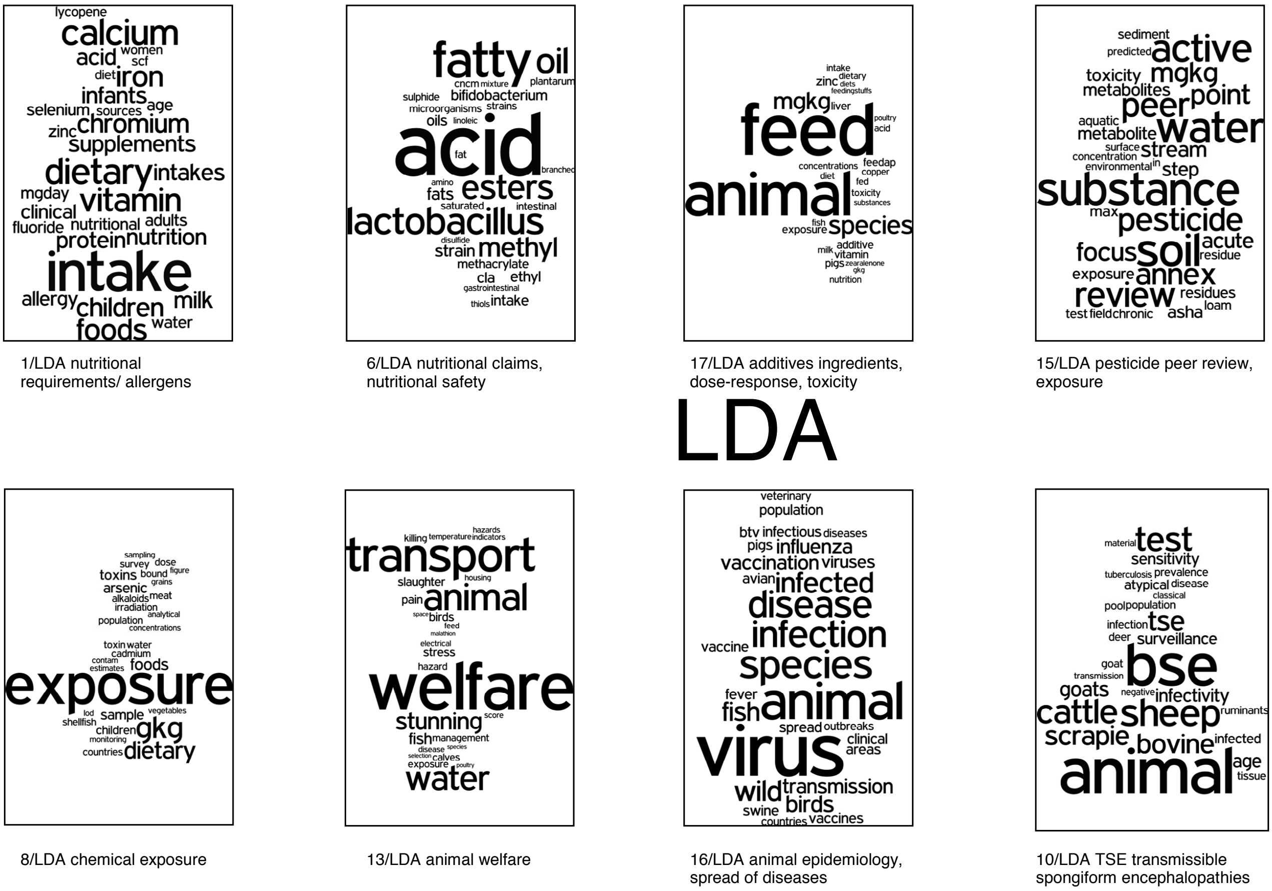
Topic 16 (animal epidemiology, spread of diseases): It focuses on the causes of spread and transmissibility of specific infectious diseases and on the efficacy of vaccines.

Topic 17 (additives ingredients, dose-response, toxicity): Evaluation of the safety and of the correct dose of certain additives added to the animal diet.

Topic 18 (CTM's 16) (chemicals, additives, MOCA, experimental health claims): It aims at the assessment of the exposure to chemicals which could contaminate food.

Topic 19 (CTM's 13) (health claims): It addresses the scientific substantiation of health claims in relation to specific substances.

Topic 20 (CTM's 8) (pesticides peer review, environment): It takes conclusions on the peer review of the assessment of the risks posed by certain pesticides on the environment.



1/LDA nutritional requirements/ allergens

6/LDA nutritional claims, nutritional safety

17/LDA additives ingredients, dose-response, toxicity

15/LDA pesticide peer review, exposure

8/LDA chemical exposure

13/LDA animal welfare

16/LDA animal epidemiology, spread of diseases

10/LDA TSE transmissible spongiform encephalopathies

Figure 16: Word clouds based on the 30 words with the highest posterior probabilities showing the RQs identified through LDA only. Numbers refer to the ordering of the topics as presented in the output of LDA.

CTM *Clusters* As twelve of the CTM topics were almost indistinguishable from those obtained through LDA, the following list include only the eight topics with features that were dissimilar with those identified by LDA.

Topic 2 (plant and animal & microparasites): It evaluates the risks related to the introduction of viruses and parasites in animal and plant populations.

Topic 3 (feed, natural extracts claims/safety): It focuses on the harmful effect related to natural active ingredients in herbs or plants for human consumption.

Topic 9 (animal health, welfare & TSE): It addresses the risks posed by particular breeding management practices and/or by those concerning some diseases.

Topic 10 (nutritional requirements vitamins/minerals): Aiming at advising about some nutritional requirements.

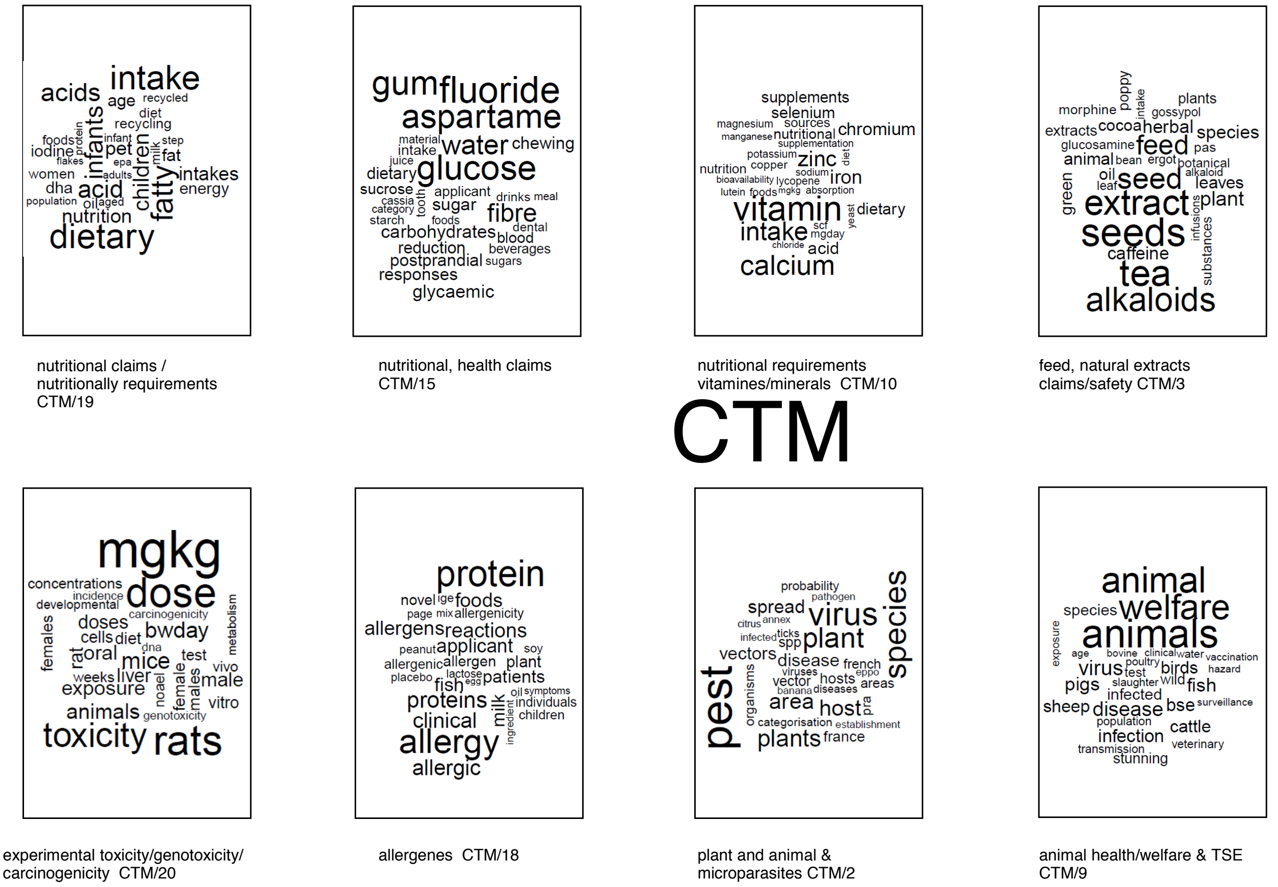
Topic 15 (nutritional, health claims): Aiming at scientifically substantiating some health claims on specific nutritional ingredients.

Topic 18 (allergens): Its main focus is to establish the maximum acceptable level of certain allergens or the likelihood of adverse effects.

Topic 19 (nutritional claims/nutritional requirements): Aiming at advising about some nutritional requirements and the admitted level in a healthy diet.

Topic 20 (experimental toxicity/genotoxicity/carcinogenicity): Aiming at evaluating the toxicity of some substances through the evaluation of animal experiments.

The parallel application of the two topic models provided an added value in the capability of characterizing the RQs within the remits of EFSA: LDA led to distinct topics with regards the issues linked to animal epidemiology/spread of disease, transmissible spongiform encephalopathies (TSEs) and animal welfare, whereas all this



nutritional claims / nutritionally requirements CTM/19

nutritional, health claims CTM/15

nutritional requirements vitamines/minerals CTM/10

feed, natural extracts claims/safety CTM/3

experimental toxicity/genotoxicity/ carcinogenicity CTM/20

allergenes CTM/18

plant and animal & microparasites CTM/2

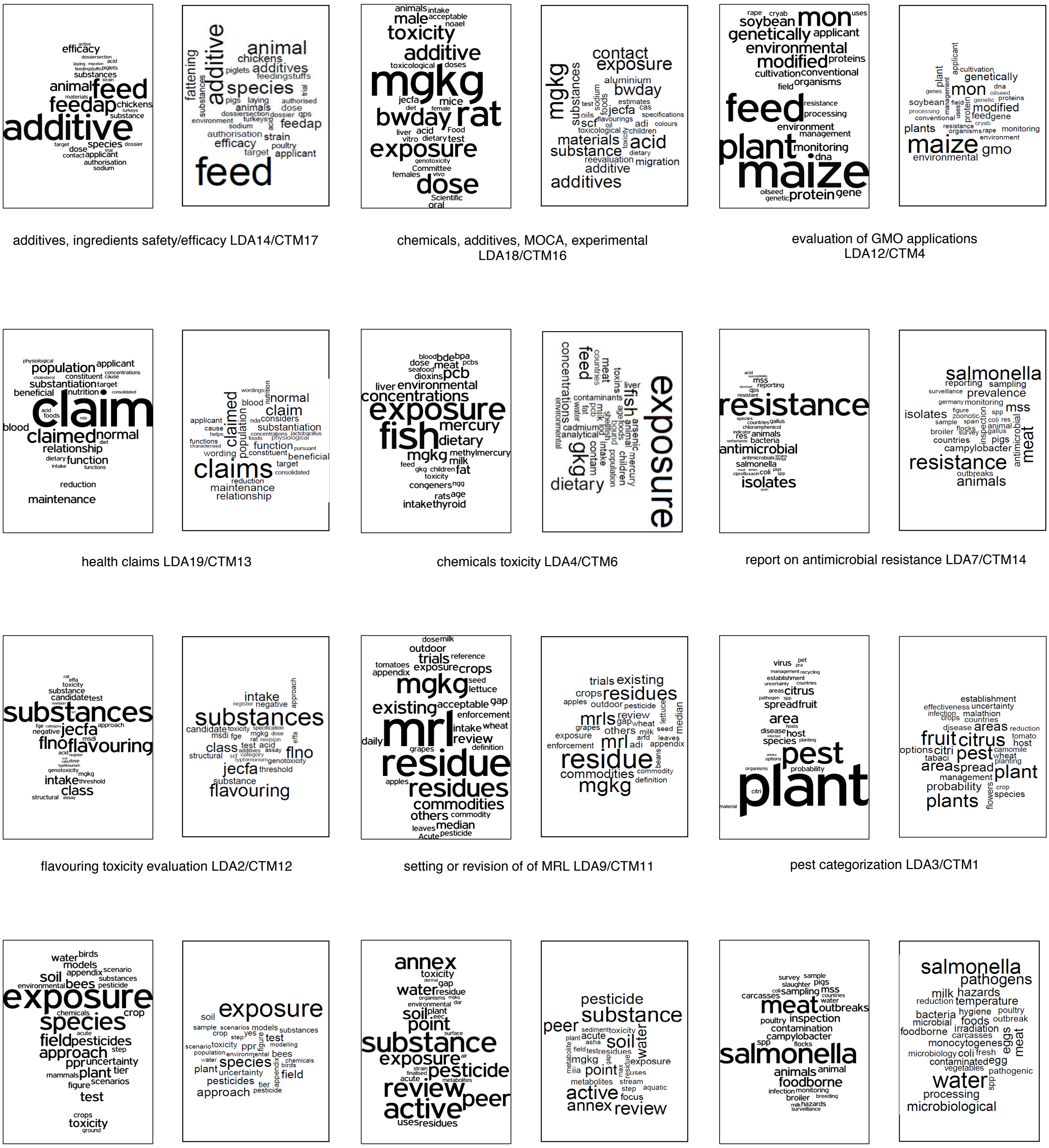
animal health/welfare & TSE CTM/9

Figure 17: Word clouds based on the 30 words with the highest posterior probabilities showing the RQs identified through CTM only. Numbers refer to the ordering of the topics as presented in the output of CTM.

issues were combined in one topic by CTM; moreover a specific LDA topic was associated to environmental chemical exposure. On the other hand, based on the CTM modeling, a specific topic was associated, for instance, with: issues associated with the exposure to allergens, pest/microparasites affecting either plants or animals, experimental toxicity/genotoxicity/carcinogenicity, safety assessment or claims regarding natural extracts in feeds. Both topic models led to the identification of a number of main RQs easily recognizable within the classical remits of EFSA (e.g. evaluation of GMO applications; health claims; pesticides peer reviews, microbiological RA). A relevant result was the identification of topics related to specific EFSA's periodical activities like the production of the European Union Summary Reports on antimicrobial resistance or the issuing of Reasoned Opinions on setting/revision of MRLs. Moreover the CTM allowed the ranking of the topics based on the expected topic proportions within the documents dataset (Figure 19). The top five topics were: setting or revision of MRL (CTM_topic 11); additives, ingredients safety/efficacy, applications (17); health claims (13); pesticides peer review, environment (8); flavouring toxicity evaluation (12).The negligible impact of the elimination of the Panels acronyms was evident in particular when inspecting this graph as only one topic shifted of one place its ranking in the list.

CTM also allowed the graphical representation of the relationships between topics. In Figure 20 the relationships among the 20 CTM topics are shown along with the labels of the RQs: the resulting general pattern is interpretable and plausible if put on the context of the remits of EFSA.

Given the distribution over the clusters and given the distribution over the words in each documents, LDA or CTM provided a posterior distribution of each document's membership in each cluster. When having manually revised the 91 EFSA documents a good consistency of the automated classification with respect to the RA domains was evident. The impact of the stemming procedure was evident in both outcomes of LDA and CTM resulting in some differences in the set of the obtained topics. However that does not necessary mean that these alternative outcomes are more valid or more meaningful. By comparing the list of words column by column, in the case of the LDA outputs only 12 out of the 20 original topics didn't change substantially their



additives, ingredients safety/efficacy LDA14/CTM17

chemicals, additives, MOCA, experimental LDA18/CTM16

evaluation of GMO applications LDA12/CTM4

health claims LDA19/CTM13

chemicals toxicity LDA4/CTM6

report on antimicrobial resistance LDA7/CTM14

flavouring toxicity evaluation LDA2/CTM12

setting or revision of of MRL LDA9/CTM11

pest categorization LDA3/CTM1

guidance for RA on plant protection products LDA5/CTM7

pesticides peer review, environment LDA20/CTM8

microbiological food safety LDA11/CTM5

Figure 18: Pairs (see in the body text for details on matching) of word clouds based on the 30 words with the highest posterior probabilities showing the RQs identified through LDA and CTM. Numbers refer to the ordering of the topics as presented in the outputs of the two models (#/ from LDA, /# from CTM).

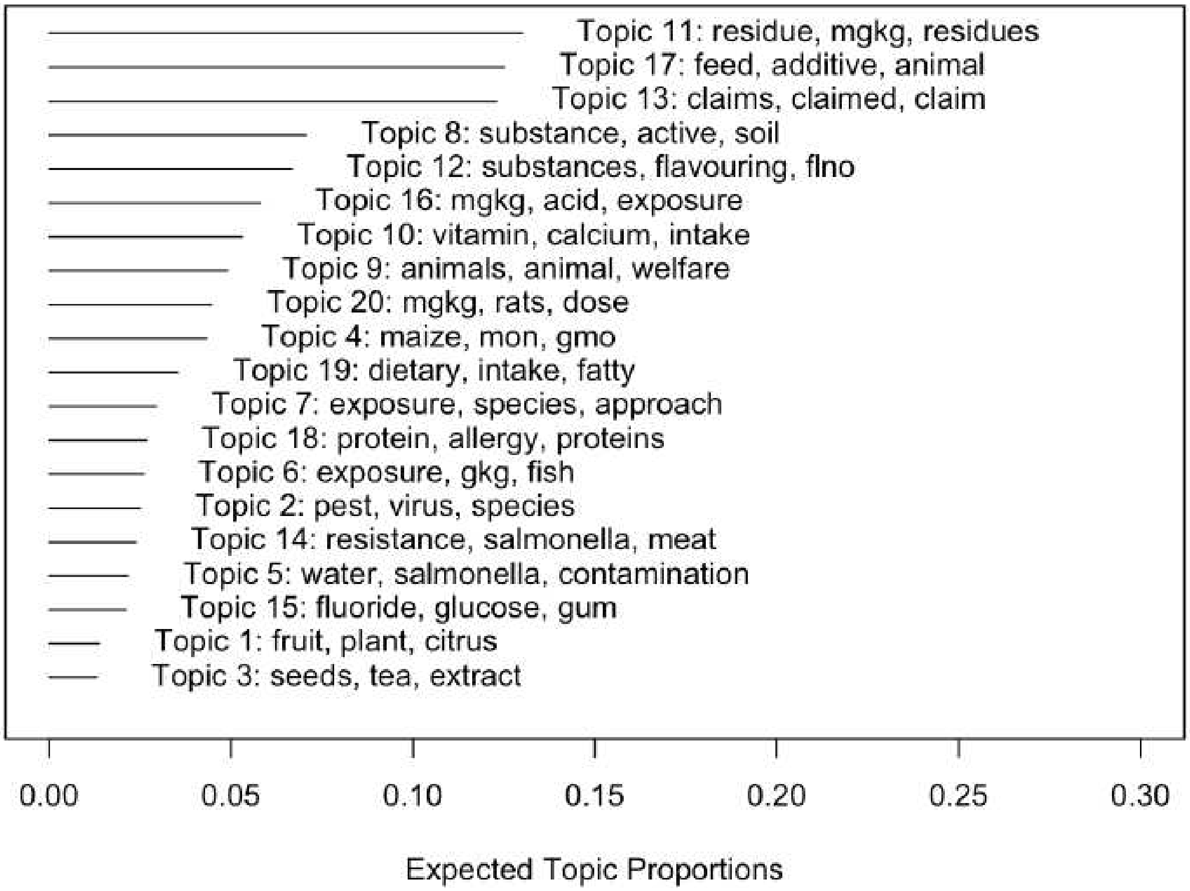


Figure 19: Ranking of the CTM topics based on the expected topic proportions within the documents dataset. Numbers refer to the CTM ordering of the topics. For each topic the three words associated with the highest post probability are listed.

ranking of words and meaning. The interpretation of new topics obtained after stemming was not always as easy as was with the original ones; in particular:

- two topics, including among the top ranked terms e.g. residues, MRLs were mentioned, were similar without suggesting different specific issues: in the original output the Topic 9 (setting or revision of MRL) was meaningful by itself.
- two topics were related with experimental exposure to chemicals hazards and food contact materials but also in this case a clear interpretation of the difference was not straightforward;
- the pesticides peer reviews, originally split in two topics (15 and 20), in the new output were merged in one and the same holds for one additional topic focussing on dietary and environmental chemical exposure (combining terms reported in the original Topic 4 (chemical toxicity) and Topic 8 (chemical exposure);
- a new topic was obtained mixing two different issues i.e. animal welfare (the original Topic 13) and allergens (that was an issue linked to nutritional items within the original Topic 1);
- finally it was not possible to identify any topic referring to important issues as TSE (the original Topic 10 (TSE)) or nutritional claims/safety (the original Topic 6 (nutritional claims/safety)) .

When considering the CTM output after stemming, 15 out of the 20 the original topic columns didn't change their word ranking and meaning. As a consequence of the stemming the main changes were:

- the original CTM Topic 9 (animal health, welfare & TSE) was splitted in three topics i.e. (1) animal health (2) animal welfare and (3) TSE; therefore the same distinction of animals related themes as identified

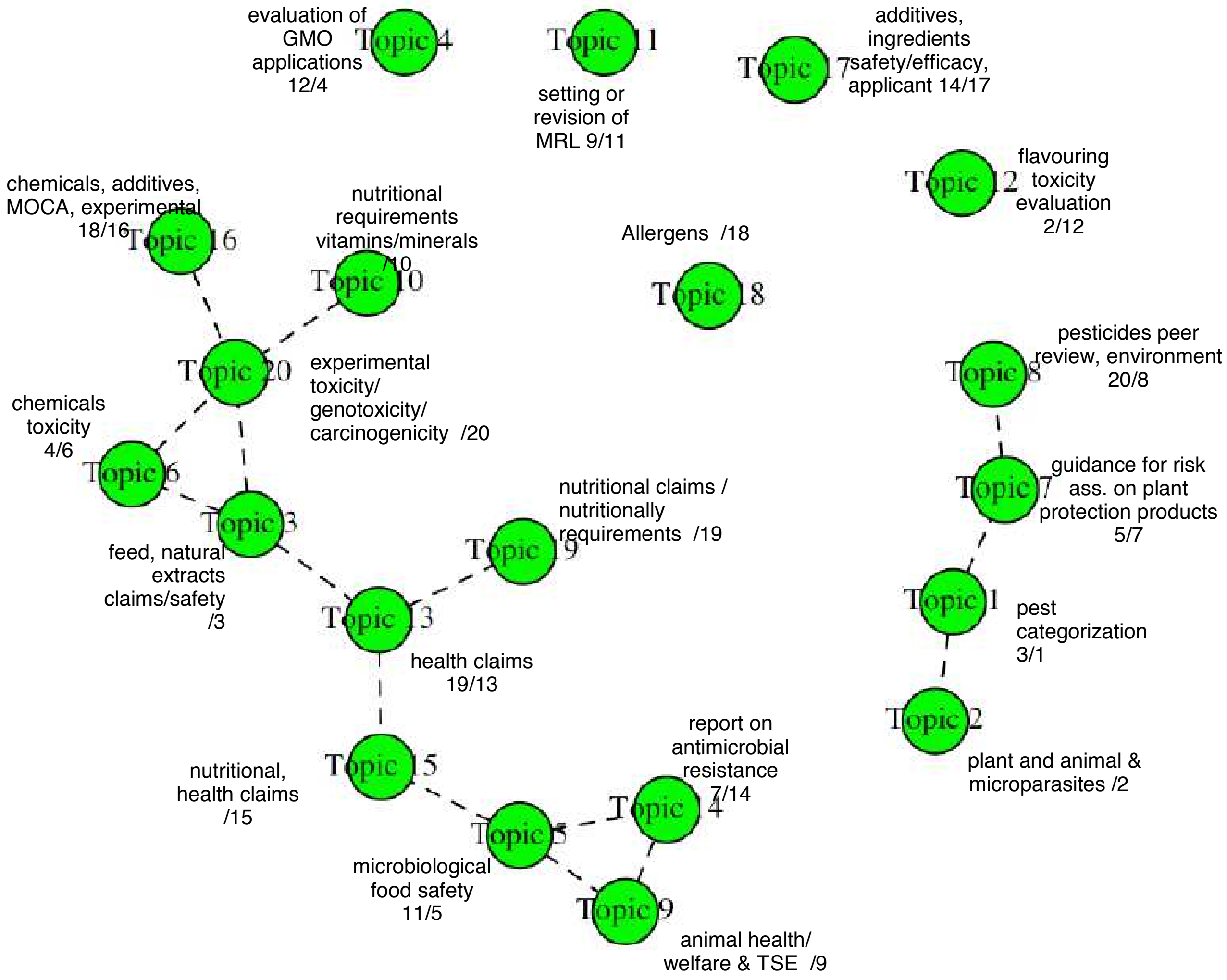

evaluation of GMO applications 12/4
Topic 4
Topic 11
setting or revision of MRL 9/11
Topic 17
additives, ingredients safety/efficacy, applicant 14/17
Topic 12
flavouring toxicity evaluation 2/12
chemicals, additives, MOCA, experimental 18/16
Topic 16
nutritional requirements vitamins/minerals /10
Topic 10
Allergens /18
Topic 18
Topic 20
experimental toxicity/ genotoxicity/ carcinogenicity /20
pesticides peer review, environment 20/8
Topic 8
chemicals toxicity 4/6
Topic 6
Topic 3
feed, natural extracts claims/safety /3
nutritional claims / nutritionally requirements /19
Topic 19
guidance for risk ass. on plant protection products 5/7
Topic 7
Topic 13
health claims 19/13
Topic 1
pest categorization 3/1
Topic 2
plant and animal & microparasites /2
nutritional, health claims /15
Topic 15
report on antimicrobial resistance 7/14
Topic 14
Topic 5
microbiological food safety 11/5
Topic 9
animal health/ welfare & TSE /9

Figure 20: Graphical representation of the relationships among the individual topics identified as best available proxy of RQs. The topic labels were assigned to each topic following the procedure described in the body text. The representation is restricted to the 20 topics from CTM. LDA does not provide such an output.

thanks to the original (no stemming) LDA; as an additional and coherent consequence the original CTM Topic 2 (plant and animal & microparasites) changed in a topic focussed on plant populations;

- two different topics, originally associated to claims (Topic 13 and 19), were merged in one;
- a new topic emerged and focussed on the reassessment of the potential adverse effects of additives in food and beverages;
- a topic focussing on claims regarding the use of natural extracts in feed (the original Topic 3) was no longer detected as a specific topic.

Summarizing, a preliminary stemming may lead to an increase in the discriminating ability of the topic modeling but also to some potential for a reduced level of interpretability of the outcomes (apparently more in case of LDA than CTM): the availability of either alternative set of outputs may help in a better interpretation of the data of interest.

2.2 Detection of the commonest statistical techniques applied within the EFSA assessment activities and their matching to the RQs

2.2.1 Aim

To detect and describe the commonest statistical techniques applied within the assessment activities carried out by EFSA and to match them with the main RQs, the following two strategies were simultaneously carried out:

1. Supervised retrieval of the statistical techniques from the dataset of the EFSA Opinions;
2. On line survey among all the officers involved in the scientific activities of the EFSA Units or Panels.

2.2.2 Supervised retrieval of the statistical techniques from the datasets of the EFSA Opinions

Materials

All EFSA documents included in the Opinions dataset were considered. Each of those documents needed to be classified according to the presence and application of a statistical technique. The supervised retrieval was also used as a classification tool. To assess the correct ability to classify each document we used the subset of 91 EFSA documents already mentioned when TM was described.

Methods

Preliminary, a "statistical vocabulary" was created as a list of keywords associated to various statistical techniques. For the purpose, a list of relevant books on statistics, epidemiology and risk assessment (provided by the Consortium experts) were consulted paying attention to both the tables of contents and the body texts. The same statistical vocabulary was used as the basis for building the list of statistical techniques inserted in a multiple choice menu within the questionnaire used for the on line survey. In particular the following books were screened:

- Douglas G Altman. *Practical statistics for medical research*. CRC press, 1990. isbn: 0412276305 (Douglas G Altman, 1990)
- Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014. isbn: 1439819181 (Banerjee, Carlin, and Gelfand, 2014)
- Lyle D Broemeling. *Bayesian methods in epidemiology*. CRC Press, 2013. isbn: 1466564970 (Broemeling, 2013)
- Louis Anthony Cox Jr. *Risk analysis foundations, models, and methods*. Vol. 45. Springer Science & Business Media, 2012. isbn: 1461508479 (Cox Jr, 2012)
- Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013. isbn: 1118625617 (Fleiss, Levin, and Paik, 2013)
- Johan Giesecke. *Modern infectious disease epidemiology*. Edward Arnold (Publisher) Ltd., 1994. isbn: 0340592370 (Giesecke, 1994)
- Government Document. 2015 (2015)
- Theodore R Holford. *Multivariate methods in epidemiology*. Oxford University Press, 2002. isbn: 0195124405 (Holford, 2002)
- Ettore Marubini and Maria Grazia Valsecchi. *Analysing survival data from clinical trials and observational studies*. Vol. 15. John Wiley & Sons, 2004. isbn: 0470093412 (Marubini and Valsecchi, 2004)
- John Neter, William Wasserman, and Michael H Kutner. *Applied linear regression models*. Vol. 1127. Irwin Homewood, IL, 1989 (Neter, Wasserman, and Kutner, 1989)
- Daryl S Paulson. *Handbook of regression and modeling: Applications for the clinical and pharmaceutical industries*. CRC Press, 2006. isbn: 1420017381 (Paulson, 2006)
- David Vose. *Risk analysis: a quantitative guide*. John Wiley & Sons, 2008. isbn: 0470512849 (Vose, 2008)

- Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*. Vol. 368. John Wiley & Sons, 2004. isbn: 0471662674 (Waller and Gotway, 2004)
- Robert F Woolson and William R Clarke. *Statistical methods for the analysis of biomedical data*. Vol. 371. John Wiley & Sons, 2011. isbn: 111803130X (Woolson and Clarke, 2011)

A list of about 60 different techniques was constructed. To maximise the ability to find them when performing a text-based search, the statistical vocabulary underwent to a fine tuning process of seven rounds of supervised retrieval.

A result of the supervised retrieval applied to the overall Opinion dataset was considered "positive" if a single occurrence of at least one of the terms listed within the statistical vocabulary was found in the document; otherwise the result was considered "negative". Accordingly, each EFSA document was classified and labelled as "positive" or "negative". A positive result was unrelated with the number of statistical techniques detected within the document. The detection of the individual technique was based on the retrieval of at least one term out of a small set of keywords that were identified as likely associated to each technique.

To assess the ability of assigning a correct classification through the automated retrieval, a comparison between the results of the supervised retrieval and those obtained through the direct inspection of a subset of EFSA document was performed. As described previously the manual inspection and the thorough reading of 91 EFSA documents (control dataset) allowed the identification of any practical application of a statistical technique as a tool to address every TOR: the result of such an inspection was considered as a "gold standard" result. As mentioned previously in this particular subset of documents, 25 documents were classified as (true) positives and 66 as (true) negatives. Due to a problem of mismatching between this subset and the overall one, only 76 documents were used for comparison (54 true positive and 22 true negative): part of the documents of the control dataset were recently published and therefore not included in the available overall Opinion dataset.

Based on the comparison outcome, sensitivity and specificity of the supervised retrieval were calculated. In this context sensitivity was defined as the ability of finding any "directly applied statistical technique" when it was actually mentioned and used whereas specificity was defined as the ability of labelling as "negative" any document where statistical techniques were neither mentioned nor used to address and answer a TOR.

The outcomes of the retrieval were organised looking at the distribution of the occurrence of each statistical techniques respectively in the overall dataset and within each subset of EFSA documents associated to a specific "RQ". The latter, as it was previously described, was identified on the basis of the topic modelling. In particular LDA made it possible to match each EFSA document with one or more topics (with a decreasing probability of association). In this way each document was used to link a RQ to one or more statistical technique.

With regards to the association between each EFSA document and one or more statistical technique actually mentioned and directly applied, the main risk of misclassification was deemed to incur in a low level of specificity. The main reason to consider this risk as the most likely one was the high probability to detect any citations of statistical methods (or keyword) without necessary meaning that they were also applied. Therefore a strategy to increase the specificity (i.e. reduce the false positive rate) was considered as a priority and was developed.

The supervised retrieval was therefore fine tuned over seven subsequent rounds. At each one, sensitivity and specificity were calculated using the mentioned subset of EFSA document. Moreover to improve the performance of the supervised retrieval a subjective assessment of the obtained outcomes was carried out. In particular, topic by topic, the analysis focussed on the ranking of the frequency of each statistical technique screened: for instance when a statistical technique was reoccurring as a top one in most or many of the topics (e.g. factor analysis, discriminant analysis) this was considered a potential for false positive results; a similar potential for a false positive result was considered when a really specialised technique (e.g. Moran's I) or field-specific approach (Non Observed Adverse Effect Level (NOAEL) assessment) were detected across the topics.

After each retrieval round, the list of techniques and keywords (including the synonyms) were reviewed. Potential sources of misclassification were investigated by means of targeted retrieval of individual EFSA documents based on particular keywords included in the statistical vocabulary and suspected to be associated with low specificity. In this way each of them was searched and their use and meaning in context were checked. After the checking, a decision was taken about keeping or excluding the keyword from the vocabulary. After seven rounds, this iterative procedure was stopped and a final evaluation of the global sensitivity and specificity of the supervised retrieval was calculated. Finally the frequencies of occurrence of each statistical techniques were calculated both at the level of the whole dataset and at the level of the "RQ" (i.e. topics). This resulted in the final matching of each topic with a number of statistical techniques ranked on the base of their frequency of detection among the EFSA documents associated with that specific topic.

Results

Based on the outcome of the final round of supervised retrieval, 829 EFSA documents out 3,774 (22.14%) were identified as mentioning one or more than one statistical techniques as listed in the refined version of statistical vocabulary. In particular Table 23 shows that when detected, the statistical techniques in two third of the cases were represented by a unique type (565, i.e. 68.15%) whereas only in less than 1% of the cases more than seven different techniques were detected.

Table 23: Frequency (absolute number of occurrence, % and cumulative frequency) of statistical techniques detected in the overall dataset of EFSA documents (N= 3744). 829 documents included at least one technique. One to eleven techniques may have been identified in a individual document.

Number of different statistical techniques detected in an individual EFSA document	Number of EFSA documents	Frequency (%)	Cumulative Frequency (%)
1	565	68.2	68.2
2	156	18.8	87.0
3	44	5.3	92.3
4	32	3.9	96.1
5	15	1.8	98.0
6	6	0.7	98.7
7	5	0.6	99.3
8	4	0.5	99.8
9	1	0.1	99.9
11	1	0.1	100.0

This general outcome was the result of the final seventh round of supervised retrieval. The comparison of the classification of the subset of EFSA documents that undewent to manual inspection and thorough reading resulted in respectively a sensitivity equal to 81.82% and a specificity equal to 74.07%.

When detected each technique was counted once per document leading to an overall occurrence of 1,335: Table 24 shows the set of statistical techniques accounting for the 99% of the occurrence in the overall dataset of 3,744 documents. Nine techniques (i.e. meta.analysis, Analysis of Variance (ANOVA), benchmark dose methods, linear.regression, ROC modeling, simulation methods, logistic regression and generalized linear models, dose response models) accounted for about 80% of all the mentioned statistical methods.

In Figure 21 the result of the final matching of each RQ (topic) based on LDA modeling with a specific set of statistical techniques is shown after ranking them on the basis of the absolute frequency of occurrence (tables were truncated after the tenth top ranked technique).

2.2.3 On-line survey among all the officers involved in the scientific activities of the EFSA Units or Panels

Questionnaire

Based on the outcome of the TM a standardized questionnaire (Annex J) has been designed to collect relevant information from EFSA officers directly involved in the scientific activities.

The questionnaire (Annex J) has been devised in order to be able to identify the relationship linking the main EFSA RA domains, specific risk questions and statistical techniques. Therefore it was organized with:

1. an opening section allowing firstly the identification of the participant and her/his professional background, then the link of the subsequent answers to an EFSA RA domain by asking the Unit/Panel team of the participant and the years of involvement in it;

2. a second section lists 25 items identifying either assessment activities (e.g. exposure assessment, outbreak analysis) or outcomes of assessment (e.g. efficacy, morbidity/mortality): the selection of these 25 items is based on the topic lists of words, obtained through LDA and CTM and that identify the main risk questions to be addressed by EFSA. The items represent a variety of concepts within the EFSA remit

Table 24: Frequency (absolute number of occurrence, % and cumulative frequency) of each statistical technique within the overall dataset of EFSA documents (N= 3744). Table was truncated: the listed techniques accounted for the 99% cumulative frequency.

Statistical techniques	Number of occurrences	Frequency (%)	Cumulative Frequency (%)
meta analysis	313	23.45	23.45
anova	143	10.71	34.16
benchmark dose methods	126	9.44	43.6
linear regression	117	8.76	52.36
receiver operating characteristic	89	6.67	59.03
simulation	86	6.44	65.47
logistic regression	66	4.94	70.41
generalized linear models	60	4.49	74.91
dose response models	55	4.12	79.03
non parametric test	40	3	82.02
anderson hauck	34	2.55	84.57
complementary log log regression	28	2.1	86.67
survival analysis	23	1.72	88.39
ancova	18	1.35	89.74
manova	14	1.05	90.79
poisson regression	14	1.05	91.84
network analysis	13	0.97	92.81
chi square test	12	0.9	93.71
principal component analysis	11	0.82	94.53
multivariate regression	10	0.75	95.28
hierarchical models	8	0.6	95.88
non linear regression	7	0.52	96.4
discriminant analysis	6	0.45	96.85
generalized estimating equations	6	0.45	97.3
random effect models	5	0.37	97.68
probit regression	4	0.3	97.98
generalized linear mixed models	3	0.22	98.2
inverse distance	3	0.22	98.43
mancova	3	0.22	98.65
negative binomial regression	3	0.22	98.88

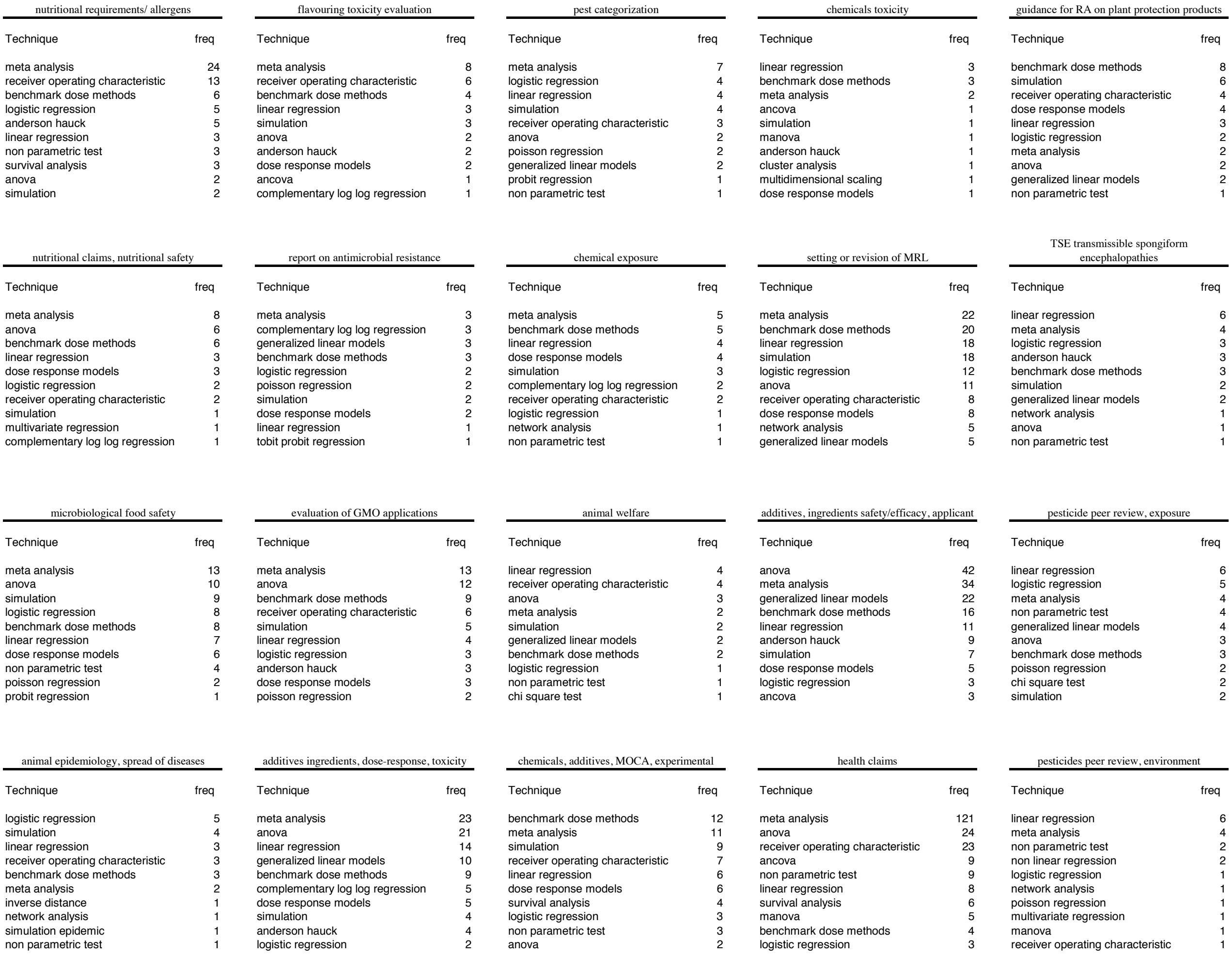
and are easily linkable with the application of statistical techniques; the participant, referring to her/his own personal experience, is asked to validate the list identifying the questions most frequently addressed to/by EFSA;

3. within the same section of the questionnaire, associated to each item there is a multiple choice menu listing all the statistical techniques (about 60) that were identified in the previously mentioned "statistical vocabulary": the menu allows the participant to identify and match any statistical analysis applied by a Unit/Panel team with the individual items (i.e. assessment activities or outcomes);

4. next section allows the participant to add up to five additional items if the previous item list is considered not complete; again for each of the eventually added item the participant is requested to identify the statistical techniques applied to address the item.

Finally, each participant is requested to provide contact details that are collected to allow, after the analysis of the survey, the potential recruitment of a subset of EFSA staff from different Units/Panels, for a face to face interview to further clarify specific issues that may arise from the survey results.

An ad hoc "Note on the processing of personal data in the context of survey" to be provided electronically to each participant before filling the on line survey was prepared and agreed with EFSA.



nutritional requirements/ allergens

Technique	freq
meta analysis	24
receiver operating characteristic	13
benchmark dose methods	6
logistic regression	5
anderson hauck	5
linear regression	3
non parametric test	3
survival analysis	3
anova	2
simulation	2

flavouring toxicity evaluation

Technique	freq
meta analysis	8
receiver operating characteristic	6
benchmark dose methods	4
linear regression	3
simulation	3
anova	2
anderson hauck	2
dose response models	2
ancova	1
complementary log log regression	1

pest categorization

Technique	freq
meta analysis	7
logistic regression	4
linear regression	4
simulation	4
receiver operating characteristic	3
anova	2
poisson regression	2
generalized linear models	2
probit regression	1
non parametric test	1

chemicals toxicity

Technique	freq
linear regression	3
benchmark dose methods	3
meta analysis	2
ancova	1
simulation	1
manova	1
anderson hauck	1
cluster analysis	1
multidimensional scaling	1
dose response models	1

guidance for RA on plant protection products

Technique	freq
benchmark dose methods	8
simulation	6
receiver operating characteristic	4
dose response models	4
linear regression	3
logistic regression	2
meta analysis	2
anova	2
generalized linear models	2
non parametric test	1

nutritional claims, nutritional safety

Technique	freq
meta analysis	8
anova	6
benchmark dose methods	6
linear regression	3
dose response models	3
logistic regression	2
receiver operating characteristic	2
simulation	1
multivariate regression	1
complementary log log regression	1

report on antimicrobial resistance

Technique	freq
meta analysis	3
complementary log log regression	3
generalized linear models	3
benchmark dose methods	3
logistic regression	2
poisson regression	2
simulation	2
dose response models	2
linear regression	1
tobit probit regression	1

chemical exposure

Technique	freq
meta analysis	5
benchmark dose methods	5
linear regression	4
dose response models	4
simulation	3
complementary log log regression	2
receiver operating characteristic	2
logistic regression	1
network analysis	1
non parametric test	1

setting or revision of MRL

Technique	freq
meta analysis	22
benchmark dose methods	20
linear regression	18
simulation	18
logistic regression	12
anova	11
receiver operating characteristic	8
dose response models	8
network analysis	5
generalized linear models	5

TSE transmissible spongiform encephalopathies

Technique	freq
linear regression	6
meta analysis	4
logistic regression	3
anderson hauck	3
benchmark dose methods	3
simulation	2
generalized linear models	2
network analysis	1
anova	1
non parametric test	1

microbiological food safety

Technique	freq
meta analysis	13
anova	10
simulation	9
logistic regression	8
benchmark dose methods	8
linear regression	7
dose response models	6
non parametric test	4
poisson regression	2
probit regression	1

evaluation of GMO applications

Technique	freq
meta analysis	13
anova	12
benchmark dose methods	9
receiver operating characteristic	6
simulation	5
linear regression	4
logistic regression	3
anderson hauck	3
dose response models	3
poisson regression	2

animal welfare

Technique	freq
linear regression	4
receiver operating characteristic	4
anova	3
meta analysis	2
simulation	2
generalized linear models	2
benchmark dose methods	2
logistic regression	1
non parametric test	1
chi square test	1

additives, ingredients safety/efficacy, applicant

Technique	freq
anova	42
meta analysis	34
generalized linear models	22
benchmark dose methods	16
linear regression	11
anderson hauck	9
simulation	7
dose response models	5
logistic regression	3
ancova	3

pesticide peer review, exposure

Technique	freq
linear regression	6
logistic regression	5
meta analysis	4
non parametric test	4
generalized linear models	4
anova	3
benchmark dose methods	3
poisson regression	2
chi square test	2
simulation	2

animal epidemiology, spread of diseases

Technique	freq
logistic regression	5
simulation	4
linear regression	3
receiver operating characteristic	3
benchmark dose methods	3
meta analysis	2
inverse distance	1
network analysis	1
simulation epidemic	1
non parametric test	1

additives ingredients, dose-response, toxicity

Technique	freq
meta analysis	23
anova	21
linear regression	14
generalized linear models	10
benchmark dose methods	9
complementary log log regression	5
dose response models	5
simulation	4
anderson hauck	4
logistic regression	2

chemicals, additives, MOCA, experimental

Technique	freq
benchmark dose methods	12
meta analysis	11
simulation	9
receiver operating characteristic	7
linear regression	6
dose response models	6
survival analysis	4
logistic regression	3
non parametric test	3
anova	2

health claims

Technique	freq
meta analysis	121
anova	24
receiver operating characteristic	23
ancova	9
non parametric test	9
linear regression	8
survival analysis	6
manova	5
benchmark dose methods	4
logistic regression	3

pesticides peer review, environment

Technique	freq
linear regression	6
meta analysis	4
non parametric test	2
non linear regression	2
logistic regression	1
network analysis	1
poisson regression	1
multivariate regression	1
manova	1
receiver operating characteristic	1

Figure 21: Result of the final matching of each RQ (topic) based on LDA modeling with the associated statistical techniques. Per each topic (N=20) the distribution of the frequency of statistical techniques is shown. Table was truncated after the tenth top ranked technique

Survey data analysis

A .csv file was obtained from the on line system used to collect the answers to the survey; the data manipulation and analysis was carried out using Stata 14. The original list of both scientific items and statistical techniques were further reduced by creating new subsets. In particular a set of 10 new summary topics (Table 25) were obtained by the following combinations of the original 25 items describing activities or outcomes consistent with the assessment carried out in EFSA.

In this case the criterion used was essentially based on the similarity or closeness of the items as considered by the consortiums experts. A similar procedure was used to create a subset of 12 general statistical approaches based on the combinations of the original 60 listed options included in the questionnaire menu. The new list of statistical approaches is:

- Hypothesis testing
- Variance/Covariance analysis
- Regression models
- Dose/response models
- Classification
- Bayesian analysis

Table 25: New summary topics obtained by the combinations of the original 25 items describing activities or outcomes consistent with the assessment carried out in EFSA.

Ten summary topics	Twentyfive original items
Diagnostics evaluation	Performances of analytical/diagnostic.
Dietary reference values	Other
Dose response	Benchmark dose/NOAEL
	Dose-response assessment
	Morbidity
	Mortality
Efficacy/risk benefit	Efficacy/effectiveness
	Risk benefit
Exposure assessment	Exposure assessment
Hazard ident/charact/ranking	Hazard identification
	Hazard characterization
	Risk ranking/classification
	Toxicity classification of chemicals
Pest/environm RA	Pest risk assessment
	Environmental risk assessment
Risk charact/uncertainty /EKE	Risk characterization
	Uncertainty
	Expert knowledge elicitation
Risk factors/prediction	Outbreak data analysis
	Spatial modeling for risk factors
	Risk prediction
Surveillance	Surveillance-monitoring
	Freedom from disease
	Spatial analysis
	Disease mapping

- Simulation
- Spatial analysis
- Survival analysis
- Time series
- Meta-analysis
- Other

Lists, frequency tables and graphs were used to summarise the answers.

Survey data analysis

Out of 160 invited participants, 65 (40.6%) accessed the on line questionnaire and 49 (30.6%) were able to complete it.

All the EFSA teams were represented (Table 26), even if a unique respondent was available in the case of ANS, CONTAM and PLH. The background of participants was mostly from the fields of veterinary, biology, food sciences and chemistry or toxicology (Table 27). More than two third of the respondents have gained a good experience in EFSA's specific topics as they have been working in their current Unit for more than three years (Table 28).

In general all the items included were considered by the respondents and only an additional one was added in the category "Other" in one questionnaire ("Dietary reference values"). Therefore the general impression is that the list of items devised from the topic modeling output was actually helpful to capture the main topic addressed in the EFSA remit.

Table 26: Distribution of the 49 respondents by EFSA team.

Team	Frequency	%
AHAW	2	4.08
AMU	6	12.2
ANS	1	2.04
BIOHAZ	3	6.12
CEF	2	4.08
CONTAM	1	2.04
DATA	6	12.2
FEEDAP	5	10.2
GMO	3	6.12
NDA	2	4.08
PLH	1	2.04
PPR	11	22.5
SCER	6	12.2

Table 27: Scientific background of the EFSA's respondents by EFSA team.

Background	AHAW	AMU	ANS	BIOHAZ	CEF	CONTAM	DATA	FEEDAP	GMO	NDA	PLH	PPR	SCER	Total
Agricultural Sciences	0	0	0	0	0	0	0	0	0	0	1	0	2	3
Biology	0	0	0	0	1	0	0	0	2	0	0	1	2	6
Statistics	0	4	0	0	0	0	1	0	0	0	0	0	0	5
Environ. Engineering/Sciences	0	0	0	0	0	0	0	0	0	0	0	3	0	3
Epidemiology	0	1	0	1	0	0	2	0	0	0	0	0	0	4
Food Sciences	0	0	0	0	0	1	1	0	0	1	0	3	1	7
Human medicine	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Informatics/Mathematics	0	0	0	0	0	0	1	0	1	0	0	0	0	2
Pharm/toxic/chem	0	0	1	0	1	0	0	0	0	0	0	4	1	7
Veterinary	2	1	0	2	0	0	1	5	0	0	0	0	0	11
Total	2	6	1	3	2	1	6	5	3	2	1	11	6	49

Table 28: Distribution of the respondents by team and the number of years of work in their current team.

	Years of activity in the team				
Team	1	2	3	>3	Total
AHAW	0	0	0	2	2
AMU	1	2	0	3	6
ANS	0	0	0	1	1
BIOHAZ	1	0	0	2	3
CEF	1	0	1	0	2
CONTAM	0	0	0	1	1
DATA	1	1	0	4	6
FEEDAP	0	0	1	4	5
GMO	1	1	0	1	3
NDA	0	0	0	2	2
PLH	0	0	0	1	1
PPR	3	0	1	7	11
SCER	0	0	1	5	6

This is also confirmed by the results obtained when looking at the Unit level. Based on the number of statistical techniques that were actually applied to address each item, they may be ranked, team by team, suggesting different scientific priorities. Focussing on the first five ranked items per team, for instance both "Prevalence" and "Surveillance and monitoring" represent a relevant issue (i.e. ranked among the first five ones) for BIOHAZ and AHAW; on the other hand ANS, CEF,CONTAM,FEEDAP, GMO,NDA and PPR share a main interest in both "Hazard identification" and "Hazard characterisation" whereas only staff from FEEDAP andNDA ranked "Efficacy/effectiveness" among the first five item. "Dose –Response" is not included among the top-five issues only for AHAW, BIOHAZ and GMO. Besides the teams directly connected with a thematic

Panel, Assessment and Methodological Support Unit (AMU), Evidence Management Unit (DATA) and Scientific Committee and Emerging Risks (SCER) share main involvement in "Dose-Response assessment" and "Hazard characterization" (AMU & SCER), and in dealing with "Uncertainty" (DATA &SCER).

An overall view is provided by the Figure 22 showing the relationship between each specific Unit linked to Panels and the shortened list of items previously described. The diagram was built after taking into account the different number of the respondents by team: therefore the relationship with each item is based on the relative frequency within the team instead of referring to absolute numbers. In Figure 23 a similar diagram show the relationships when focusing on AMU, DATA and SCER.

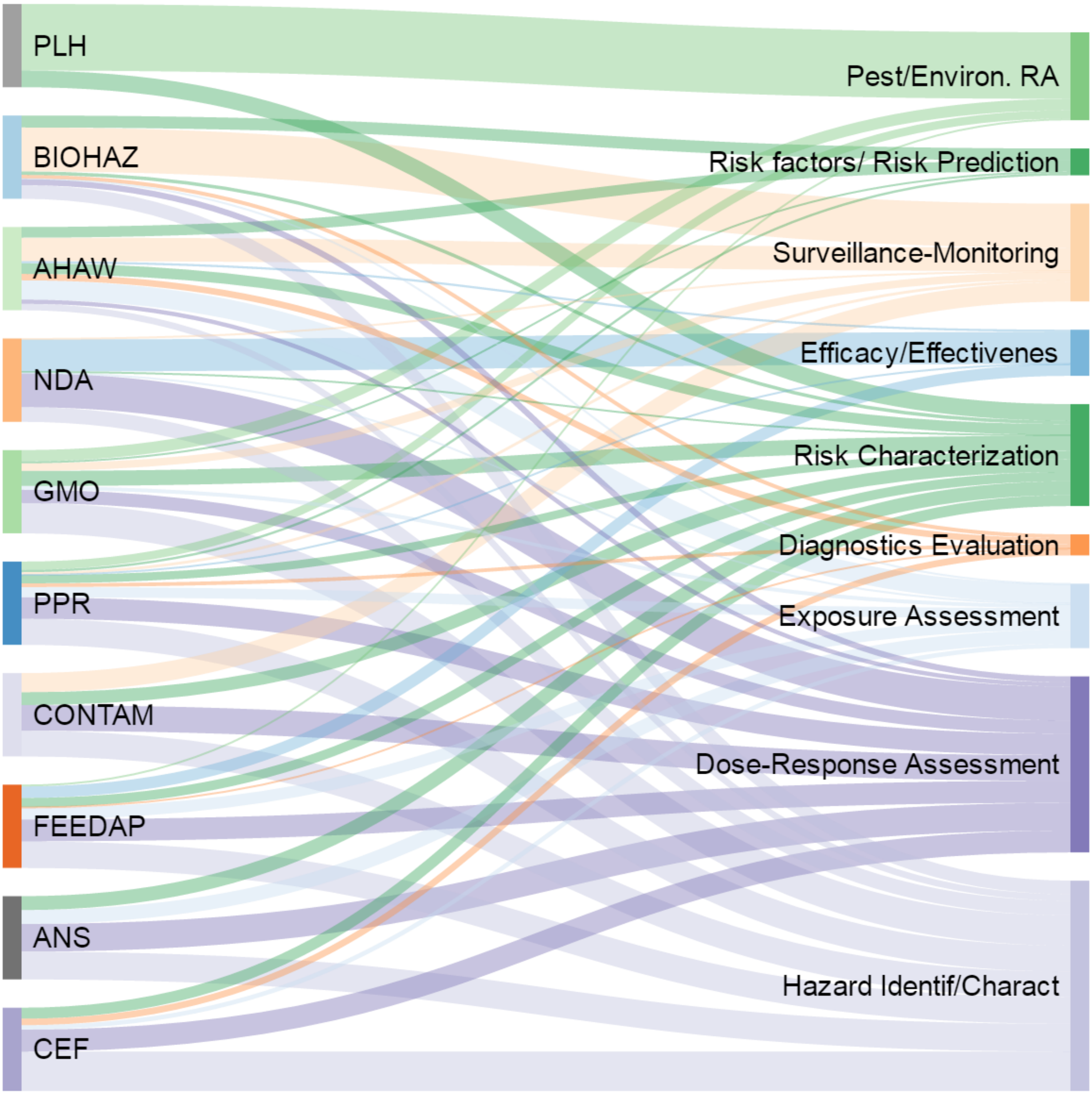


Figure 22: Sankey diagram showing the relative importance for each Panel team of the main 10 summary topics based on the combination of the items included by the questionnaire.

All the 25 items were matched with at least one or more statistical techniques which were mentioned 1,251 times: however the answer "No need of statistical technique" was given 121 times, indicating that a qualitative approach may be the preferred one. Just one EFSA staff was not able to indicate at least one statistical technique used to address a specific item.

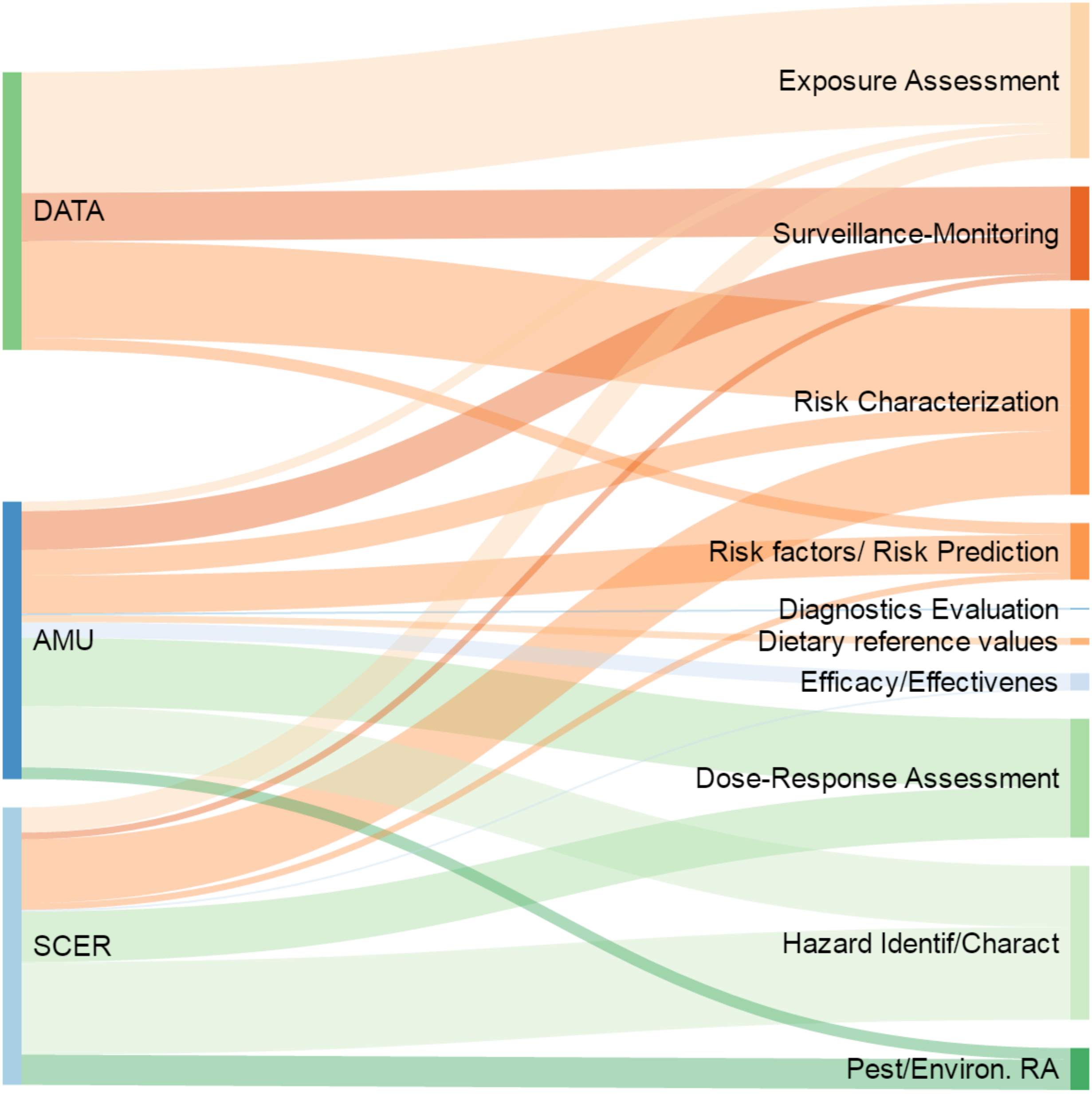


Figure 23: Sankey diagram showing the relative importance for the AMU, DATA and SCER teams of the main 10 summary topics based on the combination of the items included by the questionnaire

When looking at the need of applying any statistical technique to address each of the 25 specific items investigated by the survey, the main effort is focussed on the classical steps of any risk assessment i.e. hazard characterization, dose-response assessment, hazard identification and exposure assessment (Table 29). Surveillance and Risk characterization/Uncertainty emerge as important issues if looking at the subset of the 10 summary topics (Table 30).

The analysis of the application of each statistical technique was carried out either by team or by items: given the huge number of possible cross-tabulations, here the results are mostly summarised focussing on the main ten summary topics and the 12 general statistical approaches. However, regardless the item of application, the statistical techniques that were most mentioned (top-ten) are reported in Table 31: they accounted for more than 50% of all the mentioned techniques and include the most classical approaches (e.g. ANOVA, tests of hypothesis, dose-response models, linear models). If looking at the general approaches, a similar proportion is accounted for by regression models, dose-response models and hypothesis testing (Table 32).

Table 29: Distribution of the respondents by team and the number of years of work in their current team.

Item	Frequency	%	Cumulative %
Hazard characterization	154	12.31	12.31
Dose-response assessment	117	9.35	21.66
Hazard identification	105	8.39	30.06
Exposure assessment	92	7.35	37.41
Mortality	76	6.08	43.49
Efficacy/effectiveness	72	5.76	49.24
Surveillance-monitoring	71	5.68	54.92
Risk characterization	62	4.96	59.87
Benchmark dose/NOAEL	54	4.32	64.19
Prevalence	54	4.32	68.51
Uncertainty	51	4.08	72.58
Environmental risk assessment	42	3.36	75.94
Spatial modeling for risk factors	36	2.88	78.82
Morbidity	33	2.64	81.45
Toxicity classification of chemical/c..	33	2.64	84.09
Risk prediction	31	2.48	86.57
Performances of analytical/diagnostic..	27	2.16	88.73
Outbreak data analysis	24	1.92	92.65
Spatial analysis	24	1.92	94.56
Freedom from disease	22	1.76	96.32
Pest risk assessment	21	1.68	98
Expert knowledge elicitation	9	0.72	98.72
Other	8	0.64	99.36
Disease mapping	4	0.32	99.68
Risk benefit	4	0.32	100
Total	1251	100	

Table 30: The ten summary topics sorted by the number of statistical techniques that were associated to each of them.

10 summary topics	Frequency	%	Cumulative%
Hazard ident/charact/ranking	317	25.34	25.34
Dose response	280	22.38	47.72
Surveillance	175	13.99	61.71
Risk charact/uncertainty /EKE	122	9.75	71.46
Exposure assessment	92	7.35	78.82
Risk factors/prediction	91	7.27	86.09
Efficacy/risk benefit	76	6.08	92.17
Pest/environm RA	63	5.04	97.2
Diagnostics evaluation	27	2.16	99.36
Dietary reference values	8	0.64	100
Total	1251	100	

Coming back to the main aim i.e. to detect the commonest statistical techniques applied within the assessment activities carried out by EFSA and to match them with the main risk questions, the survey allows to match each item to a subset of statistical techniques that were identified by the EFSA staff as applied tools. To summarise the overall picture again a Sankey diagram was built based on the ten summary topics and the twelve general statistical approaches. Even in this case the relationship with each statistical approach is based on the relative frequency within each summary topic instead of referring to absolute numbers (Figure 24). For instance "Surveillance" needs in particular the application of both regression models and statistical spatial techniques whereas when dealing with risk characterization and its inherent uncertainty, a qualitative approach, that does not rely on statistical techniques, plays a role equally as important as regression models.

Table 31: Statistical techniques that were most mentioned (top-ten), regardless the topic of application.

Statistical tecnique	Frequency	%	Cumulative %
ANOVA	87	6.95	6.95
Benchmark Dose Methods	71	5.68	19.26
NOAEL (no observed adverse effect lev..	70	5.6	24.86
Dose-response models	66	5.28	30.14
Nonparametric tests of hypotheses	59	4.72	34.85
Simulation (bootstrap/Monte Carlo, etc)	58	4.64	39.49
Generalized linear models	55	4.4	43.88
Linear regression	54	4.32	48.2
Meta-analysis	49	3.92	52.12

Table 32: General statistical approaches.

Statistical approach	Frequency	%	Cumulative %
Regression models	418	33.41	33.41
Dose/response	207	16.55	49.96
Hypothesis testing	151	12.07	62.03
Variance/Covariance analysis	105	8.39	70.42
Spatial analysis	83	6.63	77.06
Classification	70	5.6	82.65
Simulation	68	5.44	88.09
Meta-analysis	49	3.92	92.01
Bayesian analysis	38	3.04	95.04
Time series	23	1.84	96.88
Other	21	1.68	98.56
Survival analysis	18	1.44	100
Total	1251	100	

The team-specific application of the statistical techniques is shown in Figure 25. In the case of ANS, the unique available respondent was not able to identify any statistical technique used to address one of the items (each of the steps of a classical risk assessment) that she had indicated. A complete picture, even if a bit more complex, is shown in Figure 26 where the relationships between Units linked to Panels, summary topics and general statistical approaches are presented altogether.

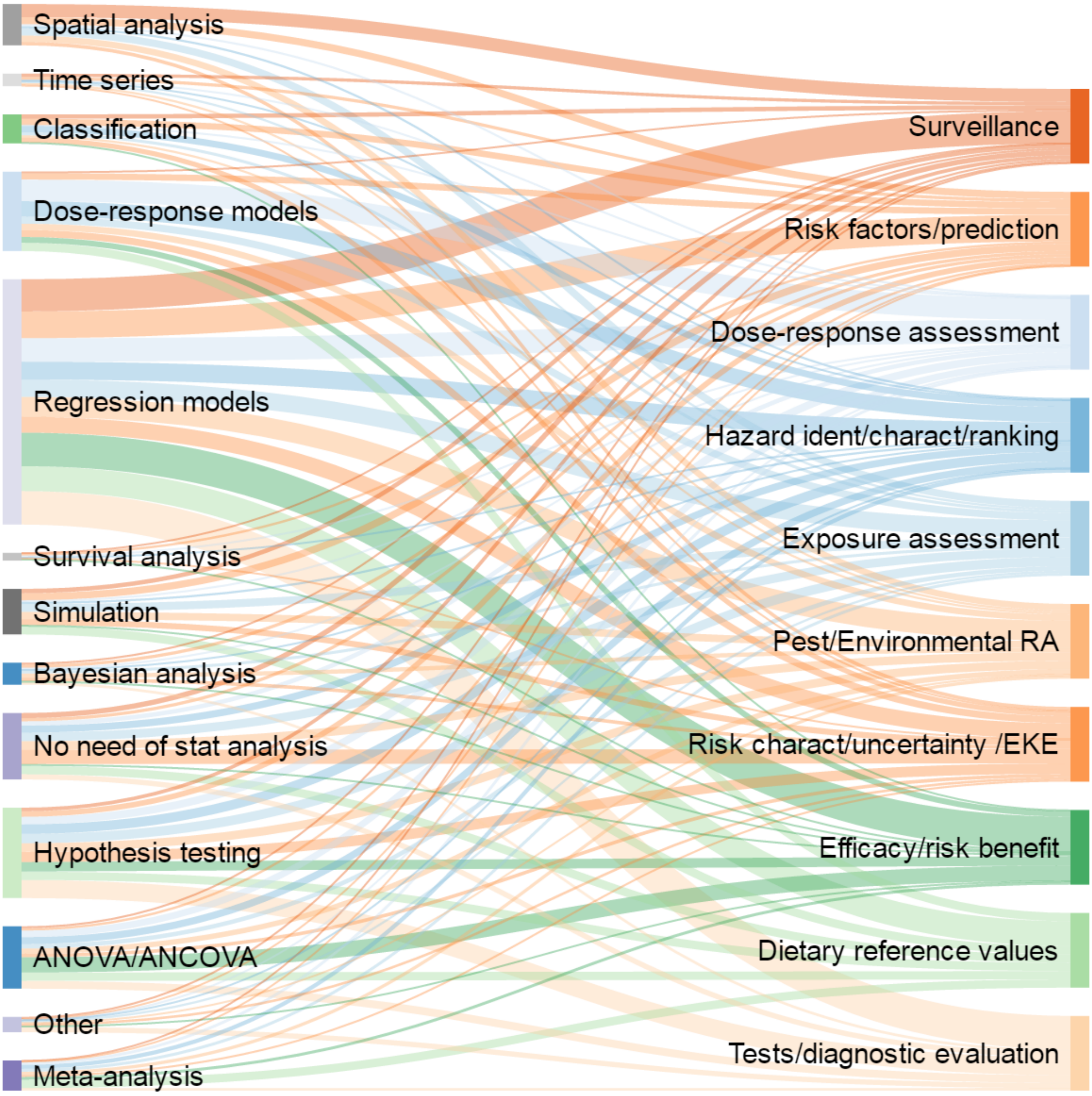


Figure 24: Sankey diagram showing the relative importance of the 12 general statistical approaches to address the main 10 summary topics in the EFSA remits. For instance "Surveillance" needs in particular the application of regression models and statistical spatial techniques.

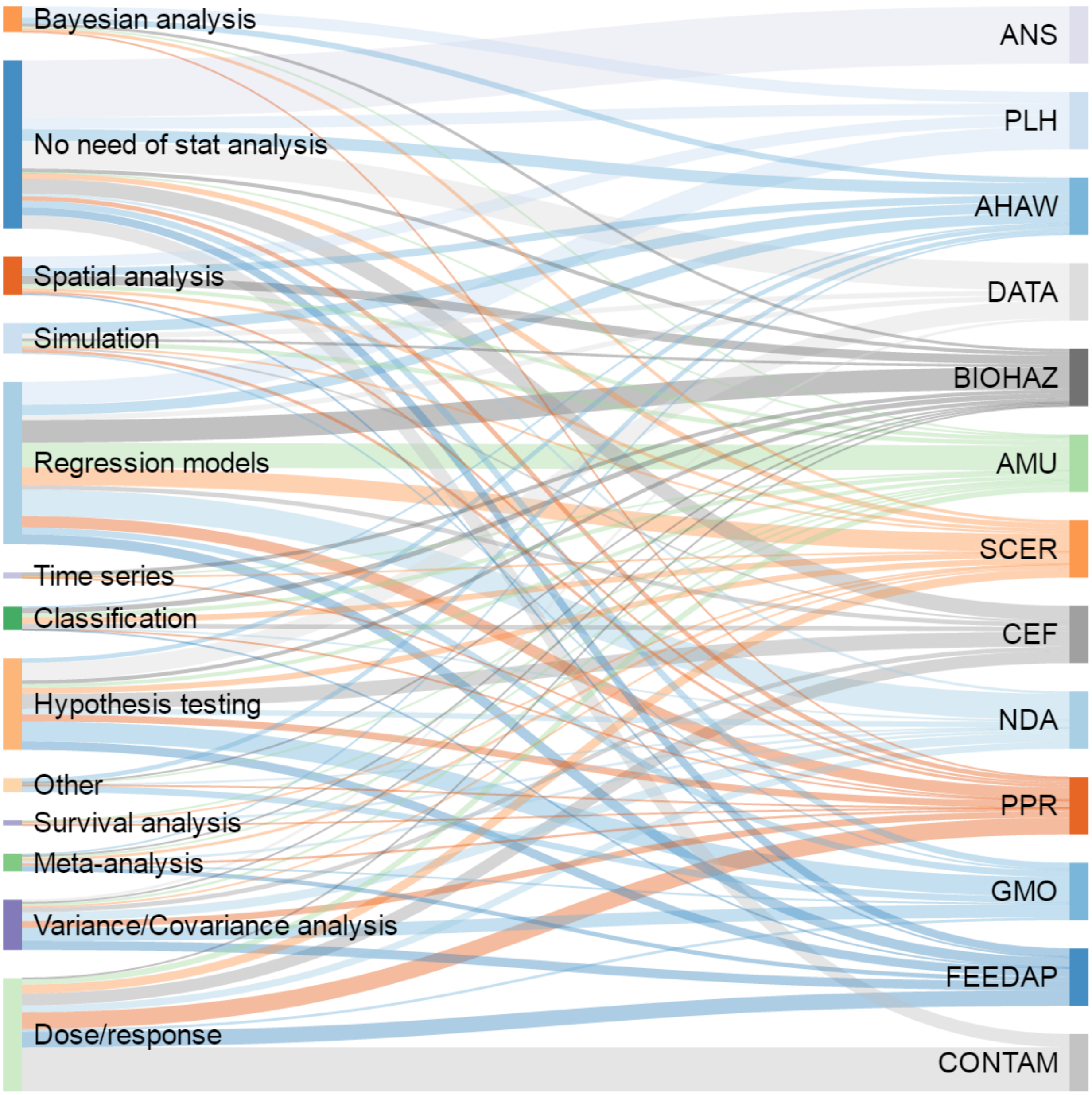


Figure 25: Sankey diagram showing the team-specific use of the 12 general statistical approaches (in the case of ANS, CONTAM and PLH he answers from one unique respondent were available for analysis)

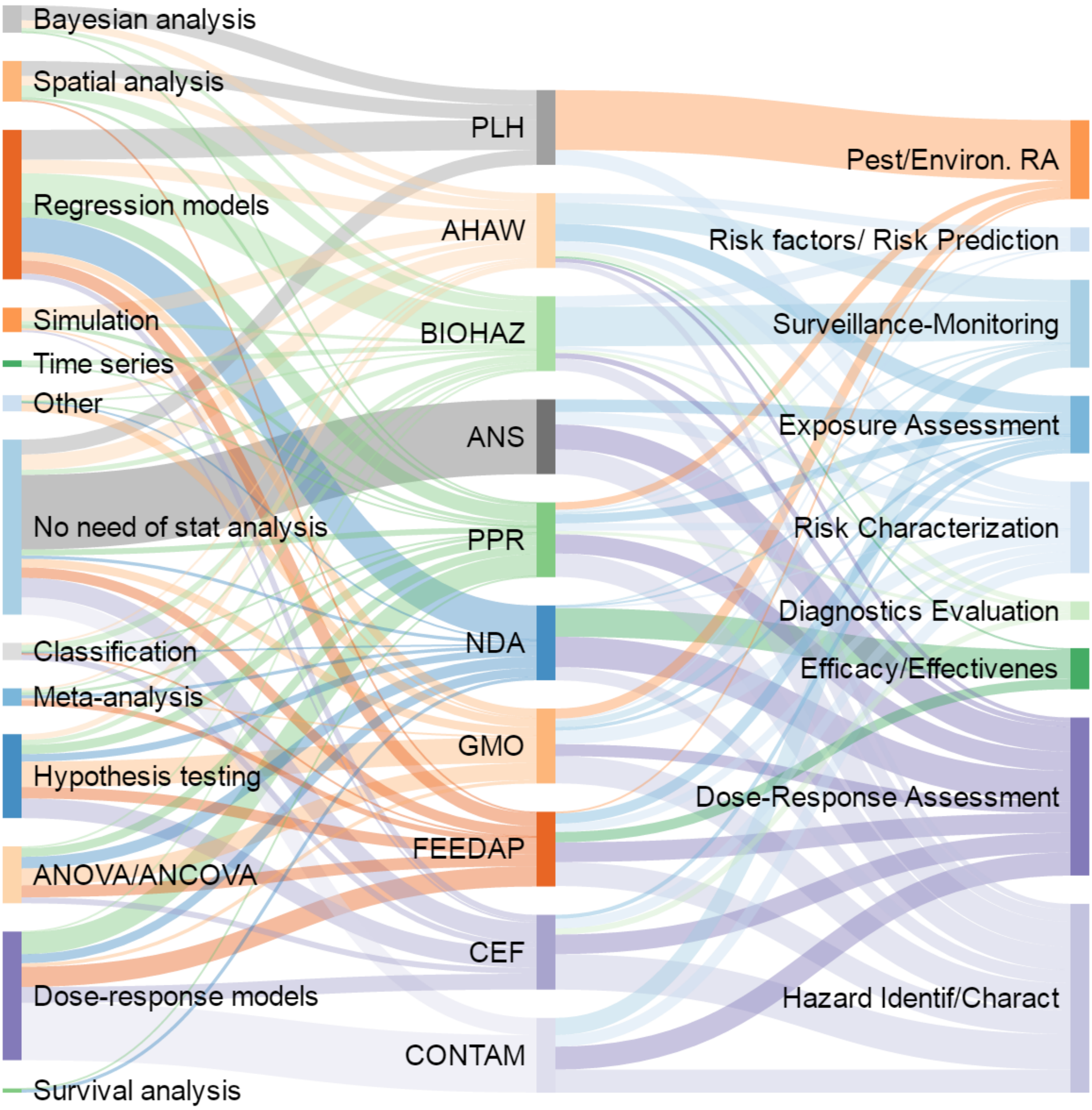


Figure 26: Sankey diagram showing at the same time the team-specific use of the 12 general statistical approaches and the topics within the scope of each of the involved Panels. Per each Panel the total contribute of either the topics or the statistical approach sums up to 100%

Opinions classification bibliography

[BK10] Michael W Berry and Jacob Kogan. “Text Mining”. In: *Applications and Theory. West Sussex, PO19 8SQ, UK: John Wiley & Sons* (2010).

[BL06] David Blei and John Lafferty. “Correlated topic models”. In: *Advances in neural information processing systems* 18 (2006), p. 147.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[HM09] JONATHAN Huang and TOMASZ Malisiewicz. *Fitting a hierarchical logistic normal distribution*. 2009.

[XX13] Pengtao Xie and Eric P Xing. “Integrating document clustering and topic modeling”. In: *arXiv preprint arXiv:1309.6874* (2013).

[15] Government Document. 2015.

[Alt90] Douglas G Altman. *Practical statistics for medical research*. CRC press, 1990. isbn: 0412276305.

[BCG14] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Crc Press, 2014. isbn: 1439819181.

[Bro13] Lyle D Broemeling. *Bayesian methods in epidemiology*. CRC Press, 2013. isbn: 1466564970.

[Cox12] Louis Anthony Cox Jr. *Risk analysis foundations, models, and methods*. Vol. 45. Springer Science & Business Media, 2012. isbn: 1461508479.

[FLP13] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013. isbn: 1118625617.

[Gie94] Johan Giesecke. *Modern infectious disease epidemiology*. Edward Arnold (Publisher) Ltd., 1994. isbn: 0340592370.

[Hol02] Theodore R Holford. *Multivariate methods in epidemiology*. Oxford University Press, 2002. isbn: 0195124405.

[MV04] Ettore Marubini and Maria Grazia Valsecchi. *Analysing survival data from clinical trials and observational studies*. Vol. 15. John Wiley & Sons, 2004. isbn: 0470093412.

[NWK89] John Neter, William Wasserman, and Michael H Kutner. *Applied linear regression models*. Vol. 1127. Irwin Homewood, IL, 1989.

[Pau06] Daryl S Paulson. *Handbook of regression and modeling: Applications for the clinical and pharmaceutical industries*. CRC Press, 2006. isbn: 1420017381.

[Vos08] David Vose. *Risk analysis: a quantitative guide*. John Wiley & Sons, 2008. isbn: 0470512849.

[WC11] Robert F Woolson and William R Clarke. *Statistical methods for the analysis of biomedical data*. Vol. 371. John Wiley & Sons, 2011. isbn: 111803130X.

[WG04] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*. Vol. 368. John Wiley & Sons, 2004. isbn: 0471662674.

Second Strategy bibliography

[AfDC14] EFSA (European Food Safety Authority), ECDC (European Centre for Disease Prevention, and Control). “The European Union Summary Report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2012.” In: *EFSA Journal* 12.3, 3590 (2014), p. 336. doi: 10.2903/j.efsa.2014.3590.

[AfDC15] EFSA (European Food Safety Authority), ECDC (European Centre for Disease Prevention, and Control). “EU Summary Report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2013”. In: *EFSA Journal* 13.2, 4036 (2015), p. 178. doi: 10.2903/j.efsa.2015.4036.

[AfDC16] EFSA (European Food Safety Authority), ECDC (European Centre for Disease Prevention, and Control). “The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2014.” In: *EFSA Journal* 14.2, 4380 (2016), p. 207. doi: 10.2903/j.efsa.2016.4380.

[Aut06] European Food Safety Authority. "Conclusion on the peer review of the pesticide risk assessment of the active substance oxyfluorfen." In: *EFSA Journal* 8.11, 1906 (2006), p. 78. doi: 10.2903/j.efsa.2010.1906.

[Aut10a] EFSA (European Food Safety Authority). "Analysis of the baseline survey on the prevalence of Campylobacter in broiler batches and of Campylobacter and Salmonella on broiler carcasses in the EU, 2008 - Part A: Campylobacter and Salmonella prevalence estimates." In: *EFSA Journal* 8.3, 1503 (2010), p. 100. doi: 10.2903/j.efsa.2010.1503.

[Aut10b] European Food Safety Authority. "Conclusion on the peer review of the pesticide risk assessment of the active substance cyproconazole." In: *EFSA Journal* 8.11, 1897 (2010), p. 73. doi: 10.2903/j.efsa.2010.1897.

[Aut10c] European Food Safety Authority. "Conclusion on the peer review of the pesticide risk assessment of the active substance pyridaben." In: *EFSA Journal* 8.6, 1632 (2010), p. 71. doi: 10.2903/j.efsa.2010.1632.

[Aut11a] European Food Safety Authority. "Modification of the existing MRLs for amidosulfuron in bovine fat, kidney, liver and milk." In: *EFSA Journal* 9.7, 2325 (2011), p. 34. doi: 10.2903/j.efsa.2011.2325.

[Aut11b] European Food Safety Authority. "Setting of temporary MRLs for nicotine in tea, herbal infusions, spices, rose hips and fresh herbs". In: *EFSA Journal* 9.3, 2098 (2011), p. 50. doi: 10.2903/j.efsa.2011.2098.

[Aut12] European Food Safety Authority. "Conclusion on the peer review of the pesticide risk assessment of the active substance kieselgur (diatomaceous earth)." In: *EFSA Journal* 10.7, 2797 (2012), p. 35. doi: 10.2903/j.efsa.2012.2797.

[Aut13a] European Food Safety Authority. "Conclusion on the peer review of the pesticide risk assessment of the active substance fenazaquin." In: *EFSA Journal* 11.4, 3166 (2013), p. 80. doi: 10.2903/j.efsa.2013.3166.

[Aut13b] European Food Safety Authority. "Reasoned opinion on the modification of the existing MRLs for indoxacarb in various salad plants and in spinach-like plants". In: *EFSA Journal* 11.5, 3247 (2013), p. 31. doi: 10.2903/j.efsa.2013.3247.

[Aut13c] European Food Safety Authority. "Statement of EFSA on host plants, entry and spread pathways and risk reduction options for Xylella fastidiosa Wells et al." In: *EFSA Journal* 11.11, 3468 (2013), p. 50. doi: 10.2903/j.efsa.2013.3468.

[Aut14a] EFSA (European Food Safety Authority). "Conclusion on the peer review of the pesticide risk assessment for aquatic organisms for the active substance imidacloprid." In: *EFSA Journal* 12.10, 3835 (2014), p. 49. doi: 10.2903/j.efsa.2014.3835.

[Aut14b] EFSA (European Food Safety Authority). "Explanatory statement for the applicability of the Guidance of the EFSA Scientific Committee on conducting repeated-dose 90-day oral toxicity study in rodents on whole food/feed for GMO risk assessment." In: *EFSA Journal* 12.10, 3871 (2014), p. 25. doi: 10.2903/j.efsa.2014.3871.

[Aut14c] European Food Safety Authority. "Dietary exposure to inorganic arsenic in the European population." In: *EFSA Journal* 12.3, 3597 (2014), p. 68. doi: 10.2903/j.efsa.2014.3597.

[Aut15a] EFSA (European Food Safety Authority). "Reasoned opinion on combined review of the existing maximum residue levels (MRLs) for the active substances metalaxyl and metalaxyl-M." In: *EFSA Journal* 13.4, 4076 (2015), p. 56. doi: 10.2903/j.efsa.2015.4076.

[Aut15b] European Food Safety Authority. "Outcome of a public consultation on the Draft Scientific Opinion of the EFSA Panel on Animal Health and Welfare (AHAW) on the welfare risks related to the farming of sheep for wool, meat and milk production". In: *EFSA Supporting publication*, EN-738 (2015), p. 25. doi: 10.2903/j.efsa.2009.738.

[Aut16] EFSA (European Food Safety Authority). "Reasoned opinion on the modification of the existing maximum residues levels (MRLs) for dimethomorph in various crops." In: *EFSA Journal* 14.1, 4381 (2016), p. 19. doi: 10.2903/j.efsa.2016.4381.

[Com11] EFSA Scientific Committee. "EFSA guidance on conducting repeated-dose 90-day oral toxicity study in rodents on whole food/feed". In: *EFSA Journal* 9.12, 2438 (2011), p. 21. doi: 10.2903/j.efsa.2011.2438.

[Com14] EFSA SC (EFSA Scientific Committee). "Scientific Opinion on a Qualified Presumption of Safety (QPS) approach for the safety assessment of botanicals and botanical preparations". In: *EFSA Journal* 12.3, 3593 (2014), p. 38. doi: 10.2903/j.efsa.2014.359.

[Com15] EFSA Scientific Committee. "Statement on the benefits of fish/seafood consumption compared to the risks of methylmercury in fish/seafood." In: *EFSA Journal* 13.1, 3982 (2015), p. 36. doi: 10.2903/j.efsa.2015.3982.

[EFS05a] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Opinion of the NDA Panel related to a notification from ONIVINS on milk products, egg products and fish products used as fining agents in wines pursuant to Article 6 paragraph 11 of Directive 2000/13/EC". In: *EFSA Journal*, 184 (2005), p. 5. doi: 10.2903/j.efsa.2005.184.

[EFS05b] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Opinion of the Scientific Panel on Dietetic products, nutrition and allergies [NDA]related to nutrition claims concerning omega-3 fatty acids, monounsaturated fat, polyunsaturated fat and unsaturated fat". In: *EFSA Journal*, 253 (2005), p. 29. doi: 10.2903/j.efsa.2005.253.

[EFS07] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Opinion of the Scientific Panel on Dietetic Products, Nutrition and Allergies related to a notification from DWV and VINIFLHOR on milk (casein) products used as fining agents in wine pursuant to Article 6 paragraph 11 of Directive 2000/13/EC". In: *EFSA Journal*, 534 (2007), p. 7. doi: 10.2903/j.efsa.2007.534.

[EFS09] EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA). "Scientific Opinion on the substantiation of health claims related to alpha-linolenic acid and maintenance of normal blood cholesterol concentrations (ID 493) and maintenance of normal blood pressure (ID 625) pursuant to Article 13(1) of Regulation (EC) No 1924/2006 on request from the European Commission." In: *EFSA Journal* 7.9, 1252 (2009), p. 17. doi: 10.2903/j.efsa.2009.1252.

[EFS10a] EFSA Panel on Dietetic Products, Nutrition, and Allergies (NDA). "Scientific Opinion on Dietary Reference Values for carbohydrates and dietary fibre." In: *EFSA Journal* 8.3, 1462 (2010), p. 77. doi: 10.2903/j.efsa.2010.1462.

[EFS10b] EFSA Panel on Dietetic Products, Nutrition, and Allergies (NDA). "Scientific Opinion on Dietary Reference Values for fats, including saturated fatty acids, polyunsaturated fatty acids, monounsaturated fatty acids, trans fatty acids, and cholesterol." In: *EFSA Journal* 8.3, 1461 (2010), p. 107. doi: 10.2903/j.efsa.2010.1461.

[EFS10c] EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA). "Scientific Opinion on the substantiation of health claims related to Camellia sinensis (L.) Kuntze (tea), including catechins in green tea and tannins in black tea, and protection of DNA, proteins and lipids from oxidative damage (ID 1103, 1276, 1311, 1708, 2664), reduction of acid production in dental plaque (ID 1105, 1111), maintenance of bone (ID 1109), decreasing potentially pathogenic intestinal microorganisms (ID 1116), maintenance of vision (ID 1280), maintenance of normal blood pressure (ID 1546) and maintenance of normal blood cholesterol concentrations (ID 1113, 1114) pursuant to Article 13(1) of Regulation (EC) No 1924/2006." In: *EFSA Journal* 8.2, 1463 (2010), p. 29. doi: 10.2903/j.efsa.2010.1463.

[EFS11a] EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA). "Scientific Opinion on the substantiation of a health claim related to "low fat and low trans spreadable fat rich in unsaturated and omega-3 fatty acids" and reduction of LDL-cholesterol concentrations pursuant to Article 14 of Regulation (EC) No 1924/2006". In: *EFSA Journal* 9.5, 2168 (2011), p. 13. doi: 10.2903/j.efsa.2011.2168.

[EFS11b] EFSA Panel on Dietetic Products, Nutrition and Allergies (NDA). "Scientific Opinion on the substantiation of health claims related to sodium bicarbonate and maintenance of normal blood pressure (ID 1404) pursuant to Article 13(1) of Regulation (EC) No 1924/2006." In: *EFSA Journal* 9.6, 2262 (2011), p. 12. doi: 10.2903/j.efsa.2011.2262.

[EFS11d] EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids (CEF). "Scientific Opinion on Flavouring Group Evaluation 50, Revision 1 (FGE.50Rev1): Consideration of pyrazine derivatives evaluated by JECFA (57th meeting) structurally related to pyrazine derivatives evaluated by EFSA in FGE.17Rev2 (20109". In: *EFSA Journal* 9.5, 1921 (2011), p. 41. doi: 10.2903/j.efsa.2011.1921.

[EFS13a] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Scientific Opinion on nutrient requirements and dietary intakes of infants and young children in the European Union." In: *EFSA Journal* 11.10, 3408 (2013), p. 103. doi: 10.2903/j.efsa.2013.3408.

[EFS13b] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Scientific Opinion on the substantiation of a health claim related to a combination of Tuscan black cabbage, "tri-coloured" Swiss chard, "bicoloured" spinach and "blu savoy" cabbage and maintenance of normal blood LDL-cholesterol concentration pursuant to Article 13(5) of Regulation (EC) No 1924/2006." In: *EFSA Journal* 11.10, 3415 (2013), p. 7. doi: 10.2903/j.efsa.2013.3415.

[EFS14a] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Scientific Opinion on Dietary Reference Values for chromium". In: *EFSA Journal* 12.10, 3845 (2014), p. 25. doi: 10.2903/j.efsa.2014.3845.

[EFS14b] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Scientific Opinion on health benefits of seafood (fish and shellfish) consumption in relation to health risks associated with exposure to methylmercury." In: *EFSA Journal* 12.7, 3761 (2014), p. 80. doi: 10.2903/j.efsa.2014.3761.

[EFS14c] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Scientific Opinion on the essential composition of infant and follow-on formulae." In: *EFSA Journal* 12.7, 3760 (2014), p. 106. doi: 10.2903/j.efsa.2014.3760.

[EFS14d] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Scientific Opinion on the evaluation of allergenic foods and food ingredients for labelling purposes". In: *EFSA Journal* 12.11, 3894 (2014), p. 286. doi: 10.2903/j.efsa.2014.3894.

[EFS14e] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Scientific Opinion on the substantiation of a health claim related to high-fibre sourdough rye bread and reduction of post-prandial glycaemic responses pursuant to Article 13(5) of Regulation (EC) No 1924/2006." In: *EFSA Journal* 12.10, 3837 (2014), p. 11. doi: 10.2903/j.efsa.2014.3837.

[EFS15a] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Scientific Opinion on the essential composition of total diet replacements for weight control." In: *EFSA Journal* 13.1, 3957 (2015), p. 25. doi: 10.2903/j.efsa.2015.3957.

[EFS15b] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies). "Scientific Opinion on the substantiation of a health claim related to FRUIT UP® and a reduction of post-prandial blood glucose responses pursuant to Article 13(5) of Regulation (EC) No 1924/2006." In: *EFSA Journal* 13.5, 4098 (2015), p. 12. doi: 10.2903/j.efsa.2015.4098.

[EFS16] EFSA CEF Panel (EFSA Panel on Food Contact Materials, Enzymes, Flavourings and Processing Aids). "Scientific Opinion on Flavouring Group Evaluation 75, Revision 1 (FGE.75Rev1): Consideration of tetrahydrofuran derivatives evaluated by JECFA (63rd meeting) structurally related to tetrahydrofuran derivatives evaluated by EFSA in FGE.33 (2008)". In: *EFSA Journal* 14.1, 4335 (2016), p. 26. doi: 10.2903/j.efsa.2016.4335.

[EoGC08] EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies), EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms), and EFSA Scientific Committee (Scientific Committee). "Safety of 'Ice Structuring Protein (ISP) - Scientific Opinion of the Panel on Dietetic Products, Nutrition and Allergies and of the Panel on Genetically Modified Organisms." In: *EFSA Journal*, 768 (2008), p. 18. doi: 10.2903/j.efsa.2008.768.

[IP05] EFSA FIP (Food Ingredients and Packaging). "Opinion of the Scientific Panel on food additives, flavourings, processing aids and materials in contact with food (AFC) related to Bis(2-ethylhexyl)phthalate (DEHP) for use in food contact materials." In: *EFSA Journal*, 243 (2005), p. 20. doi: 10.2903/j.efsa.2005.243.

[IP07] EFSA FIP (Food Ingredients and Packaging). "Opinion of the Scientific Panel on food additives, flavourings, processing aids and materials in contact with food (AFC) related to an application on the use of polyethylene glycol (PEG) as a film coating agent for use in food supplement products." In: *EFSA Journal*, 414 (2007), p. 22. doi: 10.2903/j.efsa.2007.414.

[IP11] EFSA FIP (Food Ingredients and Packaging). "Scientific Opinion on flavouring group evaluation 96 (FGE.96): consideration of 88 flavouring substances considered by EFSA for which EU production volumes / anticipated production volumes have been submitted on request by DG SANCO. Addendum to FGE.51". In: *EFSA Journal* 9.12, 1924 (2011), p. 60. doi: 10.2903/j.efsa.2011.1924.

[oAA11] EFSA Panel on Animal Health and Welfare (AHAW). "Scientific Opinion concerning the welfare of animals during transport." In: *EFSA Journal* 9.1, 1966 (2011), p. 125. doi: 10.2903/j.efsa.2011.1966.

[oAA12] EFSA Panel on Animal Health and Welfare (AHAW). "Scientific Opinion on infectious salmon anaemia." In: *EFSA Journal* 10.11, 2971 (2012), p. 22. doi: 10.2903/j.efsa.2012.2971..

[oAoS09] EFSA Panel on Additives and Products or Substances used in Animal Feed (FEEDAP). "Scientific Opinion on the use of cobalt compounds as additives in animal nutrition." In: *EFSA Journal* 7.12, 1383 (2009), p. 45. doi: 10.2903/j.efsa.2009.1383.

[oAoS12a] EFSA Panel on Additives and Products or Substances used in Animal Feed (FEEDAP). "Scientific Opinion on the safety and efficacy of AviPlus® as feed additive for chickens and minor avian species for fattening and reared for laying and minor porcine species (weaned)." In: *EFSA Journal* 10.5, 2670 (2012), p. 11. doi: 10.2903/j.efsa.2012.2670.

[oAoS12b] EFSA Panel on Additives and Products or Substances used in Animal Feed (FEEDAP). "Scientific Opinion on the safety and efficacy of Toyocerin® (Bacillus cereus) as a feed additive for sows, piglets, pigs for fattening, cattle for fattening, calves for rearing, chickens for fattening and rabbits for fattening." In: *EFSA Journal* 10.10, 2924 (2012), p. 34. doi: 10.2903/j.efsa.2012.2924.

[oAoS12c] EFSA Panel on Additives and Products or Substances used in Animal Feed (FEEDAP). "Scientific Opinion on the safety and efficacy of vitamin D3 (cholecalciferol) as a feed additive for chickens for fattening, turkeys, other poultry, pigs, piglets (suckling), calves for rearing, calves for fattening, bovines, ovines, equines, fish and other animal species or categories, based on a dossier submitted by DSM." In: *EFSA Journal* 10.12, 2968 (2012), p. 26. doi: 10.2903/j.efsa.2012.2968.

[oAoS13a] EFSA Panel on Additives and Products or Substances used in Animal Feed (FEEDAP). "Scientific Opinion on the safety and efficacy of Bonvital (Enterococcus faecium) for chickens reared for laying and minor avian species." In: *EFSA Journal* 11.4, 3167 (2013), p. 10. doi: 10.2903/j.efsa.2013.3167.

[oAoS13b] EFSA Panel on Additives and Products or Substances used in Animal Feed (FEEDAP). "Scientific Opinion on the safety and efficacy of vitamin C (ascorbic acid and sodium calcium ascorbyl phosphate) as a feed additive for all animal species based on a dossier submitted by VITAC EEIG." In: *EFSA Journal* 11.2, 3103 (2013), p. 25. doi: 10.2903/j.efsa.2013.3103.

[oAoS15] EFSA FEEDAP Panel (EFSA Panel on Additives and Products or Substances used in Animal Feed). "Scientific Opinion on the safety and efficacy of Liderfeed® (eugenol) for chickens for fattening." In: *EFSA Journal* 13.11, 4273 (2015), p. 16. doi: 10.2903/j.efsa.2015.4273.

[oAoS16a] EFSA FEEDAP Panel (EFSA Panel on Additives and Products or Substances used in Animal Feed). "Scientific opinion on the safety and efficacy of guanidinoacetic acid for chickens for fattening, breeder hens and roosters, and pigs." In: *EFSA Journal* 14.2, 4394 (2016), p. 39. doi: 10.2903/j.efsa.2016.4394.

[oAoS16b] EFSA FEEDAP Panel (EFSA Panel on Additives and Products or Substances used in Animal Feed). "Scientific opinion on the safety and efficacy of selenium compounds (E8) as feed additives for all animal species: sodium selenite, based on a dossier submitted by Retorte GmbH Selenium Chemicals and Metals." In: *EFSA Journal* 14.2, 4398 (2016), p. 26. doi: 10.2903/j.efsa.2016.4398.

[oAW04a] EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare). "Opinion of the Scientific Panel on Animal Health and Welfare (AHAW) on a request from the Commission related to the welfare of animals during transport". In: *EFSA Journal* 4, 44 (2004), p. 219. doi: 10.2903/j.efsa.2004.44.

[oAW04b] EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare). "Opinion of the Scientific Panel on Animal Health and Welfare (AHAW) on a request from the Commission related to welfare aspects of the main systems of stunning and killing applied the main commercial species of animals". In: *EFSA Journal*, 45 (2004), p. 270. doi: 10.2903/j.efsa.2004.45.

[oAW06a] EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare). "Opinion of the Scientific Panel on Animal Health and Welfare (AHAW) on a request from the Commission related with animal health and welfare risks associated with the import of wild birds other than poultry into the European Union". In: *EFSA Journal*, 410 (2006), p. 218. doi: 10.2903/j.efsa.2006.410.

[oAW06b] EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare). "Opinion of the Scientific Panel on Animal Health and Welfare (AHAW) on a request from the Commission related with the welfare aspects of the main systems of stunning and killing applied to commercially farmed deer, goats, rabbits, ostriches, ducks, geese." In: *EFSA Journal*, 326 (2006), p. 89. doi: 10.2903/j.efsa.2006.326.

[oAW07a] EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare). "Possible vector species and live stages of susceptible species not transmitting disease as regards certain fish diseases - Scientific Opinion of the Panel on Animal Health and Welfare". In: *EFSA Journal*, 584 (2007), p. 163. doi: 10.2903/j.efsa.2007.584.

[oAW08a] EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare). "Animal health and welfare aspects of avian influenza and the risk of its introduction into the EU poultry holdings - Scientific opinion of the Panel on Animal Health and Welfare." In: *EFSA Journal*, 715 (2008), p. 162. doi: 10.2903/j.efsa.2008.715.

[oAW15] EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare). "Scientific opinion on Echinococcus multilocularis infection in animals". In: *EFSA Journal* 13.12, 4373 (2015), p. 129. doi: 10.2903/j.efsa.2015.4373.

[oBio+11] EFSA Panels on Biological Hazards (BIOHAZ) et al. "Scientific Opinion on the public health hazards to be covered by inspection of meat (swine)." In: *EFSA Journal* 9.10, 2351 (2011), p. 198. doi: 10.2903/j.efsa.2011.2351.

[oBio+12] EFSA Panels on Biological Hazards (BIOHAZ) et al. "Scientific Opinion on the public health hazards to be covered by inspection of meat (poultry)." In: *EFSA Journal* 10.6, 2741 (2012), p. 179. doi: 10.2903/j.efsa.2012.2741.

[oBio04] EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards). "Opinion of the Scientific Panel on biological hazards (BIOHAZ) on the interpretation of results of EU surveillance of transmissible spongiform encephalopathies (TSEs) in ovine and caprine animals, culling strategies for TSEs in small ruminants and the TSE". In: *EFSA Journal*, 12 (2004), p. 6. doi: 10.2903/j.efsa.2004.12.

[oBio07] EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards). "Opinion of the Scientific Panel on biological hazards (BIOHAZ) on the quantitative risk assessment on the residual BSE risk in sheep meat and meat products." In: *EFSA Journal*, 442 (2007), p. 44. doi: 10.2903/j.efsa.2007.442.

[oBio10] EFSA Panel on Biological Hazards (BIOHAZ). "Scientific Opinion on BSE/TSE infectivity in small ruminant tissues." In: *EFSA Journal* 8.12, 1875 (2010), p. 92. doi: 10.2903/j.efsa.2010.1875.

[oBio11] EFSA Panel on Biological Hazards (BIOHAZ). "Scientific Opinion on assessment of epidemiological data in relation to the health risks resulting from the presence of parasites in wild caught fish from fishing grounds in the Baltic Sea." In: *EFSA Journal* 9.7, 2320 (2011), p. 40. doi: 10.2903/j.efsa.2011.2320.

[oBio14a] EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards). "Scientific Opinion on the public health risks related to the maintenance of the cold chain during storage and transport of meat. Part 1 (meat of domestic ungulates)." In: *EFSA Journal* 12.3, 3601 (2014), p. 81. doi: 10.2903/j.efsa.2014.3601.

[oBio14b] EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards). "Scientific Opinion on the scrapie situation in the EU after 10 years of monitoring and control in sheep and goats." In: *EFSA Journal* 12.7, 3781 (2014), p. 155. doi: 10.2903/j.efsa.2014.3781.

[oCon06] EFSA Panel on Contaminants in the Food Chain. "Scientific Opinion on marine biotoxins in shellfish – Emerging toxins: Ciguatoxin group." In: *EFSA Journal* 8.6, 1627 (2006), p. 38. doi: 10.2903/j.efsa.2010.1627.

[oCon10] EFSA Panel on Contaminants in the Food Chain (CONTAM). "Scientific Opinion on Lead in Food". In: *EFSA Journal* 8.4, 1570 (2010), p. 171. doi: 10.2903/j.efsa.2010.1570.

[oCon12] EFSA Panel on Contaminants in the Food Chain (CONTAM). "Scientific Opinion on the risk for public health related to the presence of mercury and methylmercury in food." In: *EFSA Journal* 10.12, 2985 (2012), p. 241. doi: 10.2903/j.efsa.2012.2985.

[oCon14] EFSA CONTAM Panel (EFSA Panel on Contaminants in the Food Chain). "Scientific Opinion on Chloramphenicol in food and feed." In: *EFSA Journal* 12.11, 3907 (2014), p. 145. doi: 10.2903/j.efsa.2014.3907.

[oFatF10] EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS). "Scientific Opinion on the safety of sucrose esters of fatty acids prepared from vinyl esters of fatty acids and on the extension of use of sucrose esters of fatty acids in flavourings on request from the European Commission." In: *EFSA Journal* 8.3, 1512 (2010), p. 36. doi: 10.2903/j.efsa.2010.1512.

[oFatF16] EFSA ANS Panel (EFSA Panel on Food Additives and Nutrient Sources added to Food). "Statement on the refined exposure assessment of tertiary-butyl hydroquinone (E 319)." In: *EFSA Journal* 14.1, 4363 (2016), p. 26. doi: 10.2903/j.efsa.2016.4363.

[oFla11] EFSA Scientific Opinion on Flavouring Group. "Evaluation 78, Revision 1 (FGE.78Rev1): Consideration of aliphatic and alicyclic and aromatic hydrocarbons evaluated by JECFA (63rd meeting) structurally related to aliphatic and aromatic hydrocarbons evaluated by EFSA in FGE.25Rev2." In: *EFSA Journal* 9.6, 2178 (2011), p. 69. doi: 10.2903/j.efsa.2011.2178.

[oFS13] EFSA ANS Panel (EFSA Panel on Food Additives and Nutrient Sources). "Scientific Opinion on safety evaluation of Ephedra species in food". In: *EFSA Journal* 11.11, 3467 (2013), p. 79. doi: 10.2903/j.efsa.2013.3467.

[oGen07a] EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms). "Opinion of the Scientific Panel on Genetically Modified Organisms on an application (Reference EFSA-GMO-NL-2005-12) for the placing on the market of insect-resistant genetically modified maize 59122, for food and feed uses, import and processing under Regulation (EC) No 1829/2003, from Pioneer Hi-Bred International, Inc. and Mycogen Seeds, c/o Dow Agrosciences LLC." In: *EFSA Journal*, 470 (2007), p. 25. doi: 10.2903/j.efsa.2007.470.

[oGen07b] EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms). "Opinion of the Scientific Panel on Genetically Modified Organisms on an application (Reference EFSA-GMO-NL-2005-18) for the placing on the market of the glufosinate tolerant soybean A2704-12, for food and feed uses, import and processing under Regulation (EC) No 1829/2003 from Bayer CropScience". In: *EFSA Journal*, 524 (2007), p. 22. doi: 10.2903/j.efsa.2007.524.

[oGen15] EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms). "Scientific Opinion on an application (EFSA-GMO-BE-2011-98) for the placing on the market of herbicide-tolerant genetically modified soybean FG72 for food and feed uses, import and processing under Regulation (EC) No 1829/2003 from Bayer CropScience". In: *EFSA Journal* 13.7, 4167 (2015), p. 29. doi: 10.2903/j.efsa.2015.4167.

[oPla14a] EFSA Panel on Plant Health (PLH). "Scientific Opinion on the pest categorisation of Citrus tristeza virus". In: *EFSA Journal* 12.12, 3923 (2014), p. 32. doi: 10.2903/j.efsa.2014.3923.

[oPla14b] EFSA Panel on Plant Health (PLH). "Scientific Opinion on the pest categorisation of Tomato yellow leaf curl virus and related viruses causing tomato yellow leaf curl disease in Europe". In: *EFSA Journal* 12.10, 3850 (2014), p. 27. doi: 10.2903/j.efsa.2014.3850.

[oPla14c] EFSA PLH Panel (EFSA Panel on Plant Health). "Scientific Opinion on pest categorisation of Xanthomonas campestris pv. pruni (Smith) Dye." In: *EFSA Journal* 12.10, 3857 (2014), p. 25. doi: 10.2903/j.efsa.2014.3857.

[oPla15] EFSA Panel on Plant Health (PLH). "Scientific Opinion on pest categorisation of Circulifer haematoceps and C. tenellus". In: *EFSA Journal* 13.1, 3988 (2015), p. 32. doi: 10.2903/j.efsa.2015.3988.

[oPtR09] EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues). "The usefulness of total concentrations and pore water concentrations of pesticides in soil as metrics for the assessment of ecotoxicological effects[1] - Scientific Opinion of the Panel on Plant Protection Products and their Residues (PPR)." In: *EFSA Journal*, 922 (2009), p. 90. doi: 10.2903/j.efsa.2009.922.

[oPtR12] EFSA Panel on Plant Protection Products and their Residues (PPR). "Scientific Opinion on the science behind the development of a risk assessment of Plant Protection Products on bees (Apis mellifera, Bombus spp. and solitary bees)." In: *EFSA Journal* 10.5, 2668 (2012), p. 275. doi: 10.2903/j.efsa.2012.2668.

[oPtR14] EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues). "Scientific Opinion addressing the state of the science on risk assessment of plant protection products for non-target terrestrial plants". In: *EFSA Journal* 12.7, 3800 (2014), p. 157. doi: 10.2903/j.efsa.2014.3800.

[oPtR15] EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues). "Scientific Opinion addressing the state of the science on risk assessment of plant protection products for non-target arthropods." In: *EFSA Journal* 13.2, 3996 (2015), p. 2012. doi: 10.2903/j.efsa.2015.3996.

3 Classical Statistical Techniques vs MLTs

In the following, some css are considered in order to compare the techniques most commonly used by the EFSA with some of the main MLTs.

Considered css are:

1. RA of processed meat consumption
2. RA of safety of the "F" feed additive
3. RA of Countries trend
4. Daphnia Magna (DAMA)
5. The food pyramid and portions
6. Assessment of analytical error for model stability in multi-level design in clinical research

3.1 RA of processed meat consumption

Red meat refers to unprocessed mammalian muscle meat—for example, beef, veal, pork, lamb, mutton, horse, or goat meat—including minced or frozen meat; it is usually consumed cooked. Regulation (EC) No. 853/2004 defines meat products "processed products resulting from the processing of meat or from the further processing of such processed products, so that the cut surface shows that the product no longer has the characteristics of fresh meat". In this document *processed meat* refers to meat that has been transformed through salting, curing, fermentation, smoking, or other processes to enhance flavor or improve preservation. Most processed meats contain pork or beef, but might also contain other red meats, poultry, offal (e.g. liver), or meat byproducts such as blood. Meat processing, such as curing and smoking, can result in formation of carcinogenic chemicals, including N-nitroso-compounds (NOC) and polycyclic aromatic hydrocarbons (PAH). Cooking improves the digestibility and palatability of meat, but can also produce known or suspected carcinogens, including heterocyclic aromatic amines (HAA) and PAH. High-temperature cooking by pan-frying, grilling, or barbecuing generally produces the highest amounts of these chemicals. This document focuses on the risk posed to human health by the consumption of red meat and/or processed meat on a daily basis.

3.1.1 Terms of References

- To assess the public health risk arising from the consumption of red meat and/or processed meat.

In particular, the document should consider any new developments regarding the carcinogenicity related to the consumption of red meat.

3.1.2 Methods

The wide known European Prospective investigation into Cancer and Nutrition (EPIC) study has been used as a reference to build up a simulated dataset. EPIC was designed to investigate the relationships between diet, nutritional status, lifestyle and environmental factors, and the incidence of cancer and other chronic diseases in 10 EU countries.

Based on the selected variables, one can predict health status after a 5-year period from baseline (T_0). In particular, since the main outcome of interest is the incidence of colorectal cancer at T_{0+5}, which is a binary outcome, this represents mainly a **classification problem**.

3.1.3 Definition of objectives

Current biological databases are populated by vast amounts of experimental data. ML has been widely applied to life sciences data (Cios, Kurgan, and Reformat, 2007).

At present, with various learning algorithms available in the literature, researchers are facing difficulties in choosing the best method that should be applied to their data. We performed an empirical study on 14 classification learning algorithm and provide some performance comparisons method in order to choose the best algorithm suitable for the case study.

In this case study a simulation of EPIC has been performed selecting a limited set of information, similar to real data, but with correlation levels arbitrarily chosen aiming at enhancing MLT exercise.

3.1.4 Outcome

y5 : Cancer in 5 years following the baseline.

The variable is binary and generated following the model:

$$\begin{aligned} y_5 &= logit^{-1} \quad (0.35 * p\_meat + 0.002 * nit + 0.015 * age + 0.01 * sex + 2.1 * weight \\ & \qquad + 0.05 * height + 5.2smk + 2.3 * paw + u + e) \end{aligned} \tag{1}$$

rumour: $u \sim N(0, 5)$
bug: $e \sim Bernoulli(0.01) * N(1.5)$

3.1.5 Tumour markers

mark_ca19.9 : CA 19_9 biomarker for gastrointestinal tumours. It is an alternative outcome expressed as a non negative bounded continuous variable. With regard to cancer, it is considered normal if < 37 U/ml, whereas values over 120 U/ml are generally considered to be caused by tumour.

In this case, we simulated it as a function of y5. It is a mixture of χ^2 and normal distribution in order to obtain positive bimodal overall distribution with two peaks: on <37 and >120 u/ml.

Listing 3.1: definition of marker ca19-9

```
mark_ca19.9 <- y5 * apply(
                matrix(
                    4*rchisq(n = n, df = 7) + 110,
                    rnorm(n = n, mean = 130, sd = sqrt(500)),
                    ncol = n
                ),
                2,
                mean
               ) +
          (1-y5) * 5 * rchisq(n = n, df = 4.5)
```

3.1.6 Hazard variables

p_meat : Processed meat consumed in g/day.

We simulated a triangular distribution, where the minimum value is set to 0, the maximum value is set to 160, and the mean value is set to 40. Values are chosen to resemble the EPIC study.

nit : assumed Nitrose-compounds (cancerogenic) quantity in μg (according to literature, processed meat has on average 1 microgram per 100 g).

It is constructed as

$$nit = 0.01 * p\_meat + N(0, 2.1).$$

3.1.7 Baseline characteristics

sex : Random generated in order to have 52% of women.

age : We generated a population of adults from a Uniform$(30, 70)$.

ht : Height in cm;
Female population: N$(165, 11)$,
Male population: N$(195, 11)$.

wt : Weight in kg; we assume that it is partly correlated with height;
Female population: N$(70, 20) + 0.01 * height$,
Male population: N$(80, 20) + 0.01 * height$.

Hemoglobin : Hemoglobin in g/dl;
Female population: N$(13, 2.1)$,
Male population: N$(14, 2.1)$.

Hematocrit : Hematocrit (%);
Female population: Uniform$(36, 47)$,
Male population: Uniform$(40, 52)$.

mcv : Mean corpuscular volume (fl);
Triangle distribution with min = 0, max = 100 and most likely value = 80.

rdw : Red blood cell distribution width (%);
Uniform$(11.5, 14.5)$

paw : Hours of physical activity a week;
This is a positive skewed distribution $5*$Beta$(n, 1, 10)$

smk : Smoker (Yes/No). Random generated in order to have 30% of smokers.

3.1.8 Data summary

A descriptive table of the database, respect to cancer, not cancer and overall sample, is reported in Table 33. Continuous variables are reported by I/II/III quartiles, while discrete variables are reported using absolute and relative frequencies. For qualitative variables, absolute and relative frequencies with respect to the considered outcome variable are reported.

The data present a greater prevalence of cancer disease between the male component of the simulated sample. The age distribution of different groups is similar. The data are generated considering an overall overweight population.

The data are simulated considering smoking habit and Nitrose-compounds quantities assumed as cancer risk factors; in fact, in cancer subgroups there is a greater percentage of smoking subjects with a distribution of Nitrose-compounds quantity with an interquartile range defined on higher values for cancer population.

3.1.9 Graphical description

Pairs plot in Figure 27 refers to continuous variables. Histograms with kernel density estimation curve were reported on the main diagonal, Pearson correlation coefficients above the diagonal, and spline regression curves (for each combination of variables) with correlation ellipses below the diagonal. The shape of the ellipse is determined by the Pearson correlation coefficient, r. Strong correlation means a long 'a' (major semiaxis) and a short 'b' (minor semiaxis). Also, the orientation of the ellipse also depends on r. The pairs plot show that the data are not much correlated, excluding a 'natural' relation between height and weight. Weak relations are visible for the height and weight with hematocrit concentration, and also between the meat intake and nitroso with hematic concentration. It is evident that data are not affected from multicollinearity problems.

With respect to cancer disease factor, Figure 28 reports a boxplot for each continuous variable. The distributions look, more or less, balanced between cancer and not cancer populations, but there is a slightly

Table 33: Descriptive Statistics by cancer

	N	Not Cancer[a]	Cancer	Combined
		$N = 5664$	$N = 4336$	$N = 10000$
sex : M	10000	41% (2349)	57% (2451)	48% (4800)
age	10000	39 50 60	40 50 60	40 50 60
height	10000	164 168 174	166 171 176	165 170 175
weight	10000	70 75 80	73 79 83	71 76 82
hemoglobin	10000	12 13 14	13 14 15	12 13 15
hematocrit	10000	40 43 46	41 44 47	41 44 46
mvc	9985	87 90 93	87 90 93	87 90 93
rdw	10000	12 13 14	12 13 14	12 13 14
smk	9980	1% (80)	67% (2920)	30% (3000)
paw	9982	0.12 0.27 0.51	0.20 0.47 0.84	0.14 0.33 0.64
p_meat	10000	40 61 89	41 65 94	40 62 91
nit	10000	1.1 1.6 2.4	1.1 1.7 2.4	1.1 1.7 2.4
mark_ca19.9	10000	11 19 30	119 132 146	17 41 128

[a] a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.N is the number of non–missing values.Numbers after percents are frequencies.

shifted distribution on higher values in inter–quartile range for weight, meat intake, hours of physical activity, and hematocrit concentration.

May be interesting to show the relation between the BMI, instead of the single weight components, and the outcome variables, as showed in Figure 29, containing a scatter plot for tumour marker variable (*mark_ca19.9*) and a boxplot for cancer disease variable. The relation between tumour marker and BMI seems to be non-linear, while for the cancer factor variable it seems that the BMI distribution is shifted towards greater values for cancer sub-population. In both cases the distribution is, more or less, symmetric. The third plot in Figure 29 shows the histogram plot for the BMI distribution: the variable is symmetric and its distribution may be considered normal.

3.1.10 Structure of the analysis

The MLTs considered in the analysis are:

naïve : most common class

ctree : Conditional Inference Trees (CTREE)

rpart : Recursive Partitioning and Regression Trees (RPART)

kknn : k-Nearest Neighbour (KNN)

mlp : Multi–Layer Perceptron (MLP) with one hidden layer

mlpe : Multi–Layer Perceptron Ensemble (MLPE)

svm : Support Vector Machine (SVM)

randomForest : Random Forest (RF)

bagging

boosting

lda : Linear Discriminant Analysis (LDA)

lr : Logistic Regression (LR)

nb : Naïve Bayes (NB)

qda : Quadratic Discriminant Analysis (QDA)

Algorithms were implemented using R for statistical computing and Rminer package.

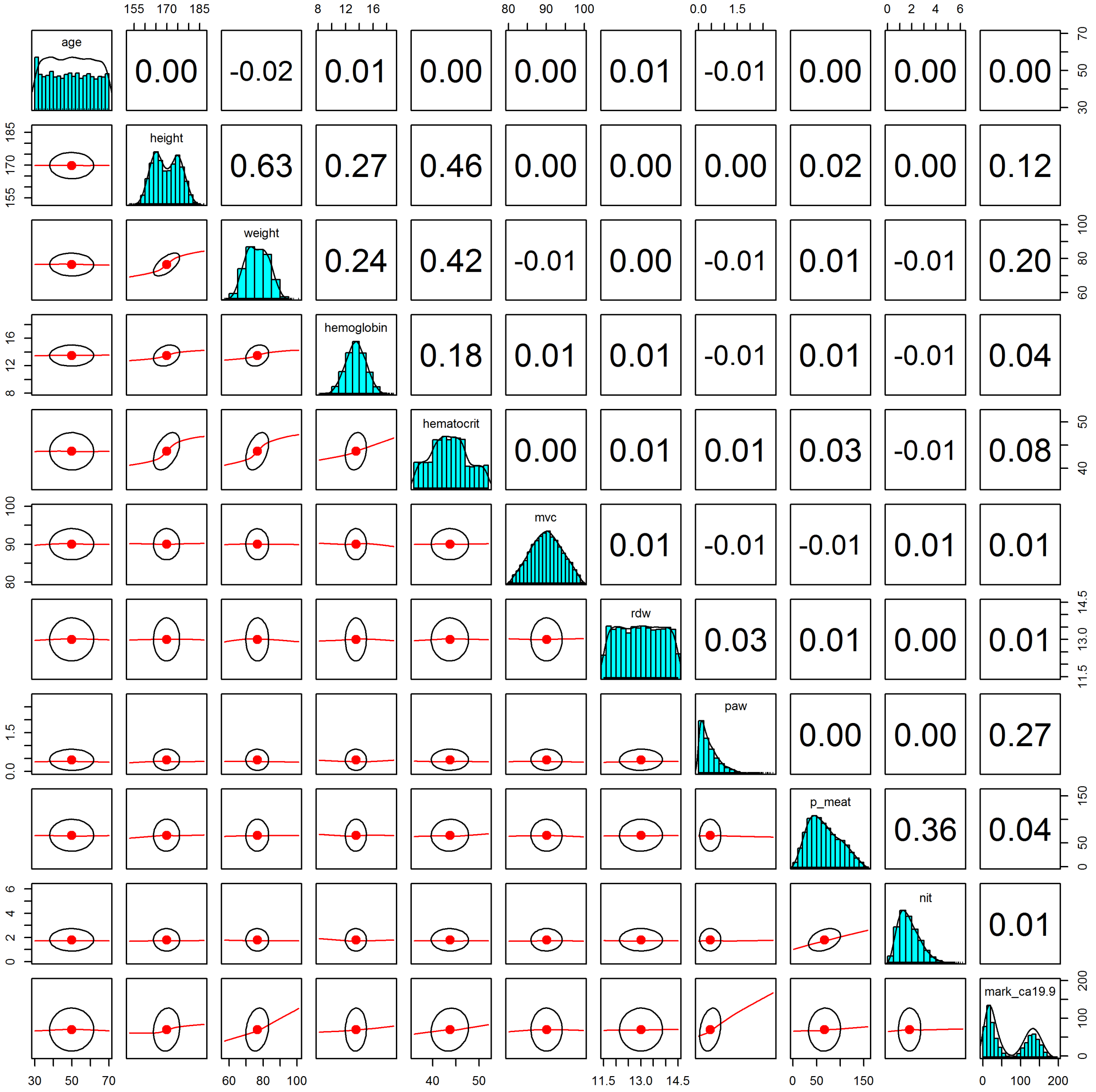


Figure 27: Pairs plot for continuous database variables

3.1.11 Model setting

Four models are computed considering different explanatory covariate settings:

All : a full model including all explanatory covariates

Relevant : only the relevant features $(p\_meat, age, sex, nit, weight, height, smk, paw)$ used to simulate the outcome variable are included

Not Relevant : only not relevant features are included

Only Meat : only the p_meat variable is considered

These models were applied to each MLT to evaluate the ML behaviours to build a good classification rule with different levels of available information.

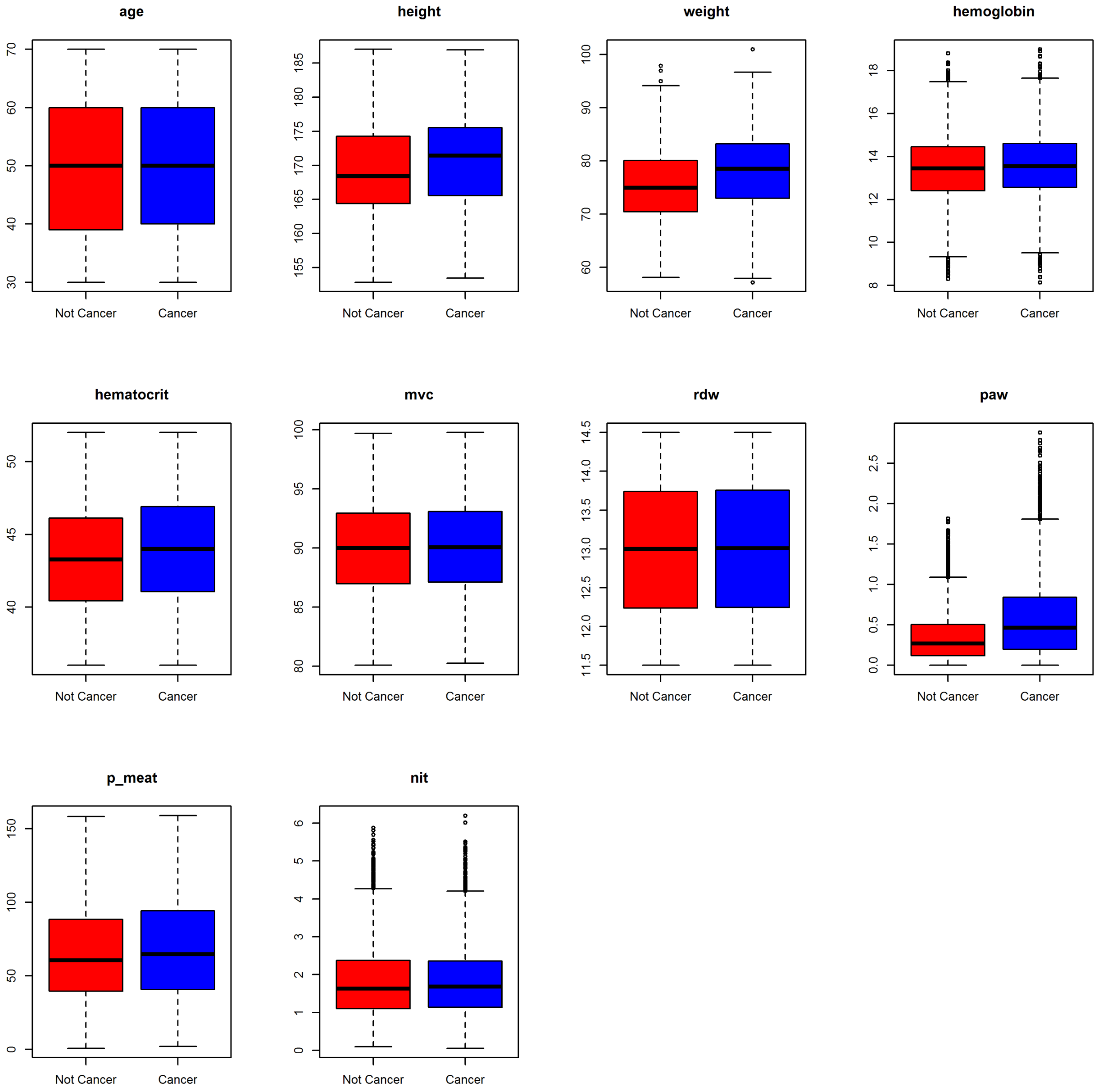


Figure 28: Boxplots for continuous database variables according to cancer/not cancer condition

3.1.12 Performance measures

Several performance measures were adopted in order to compare the different MLTs:

- **Classification accuracy** is the number of correct predictions, divided by the total number of predictions, expressed as a percentage.

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}} \tag{2}$$

- **Precision** (also called sensitivity in the context of diagnostic tests) is defined as the proportion of the true positives against all the positive results (both true positives and false positives):

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positives}} \tag{3}$$

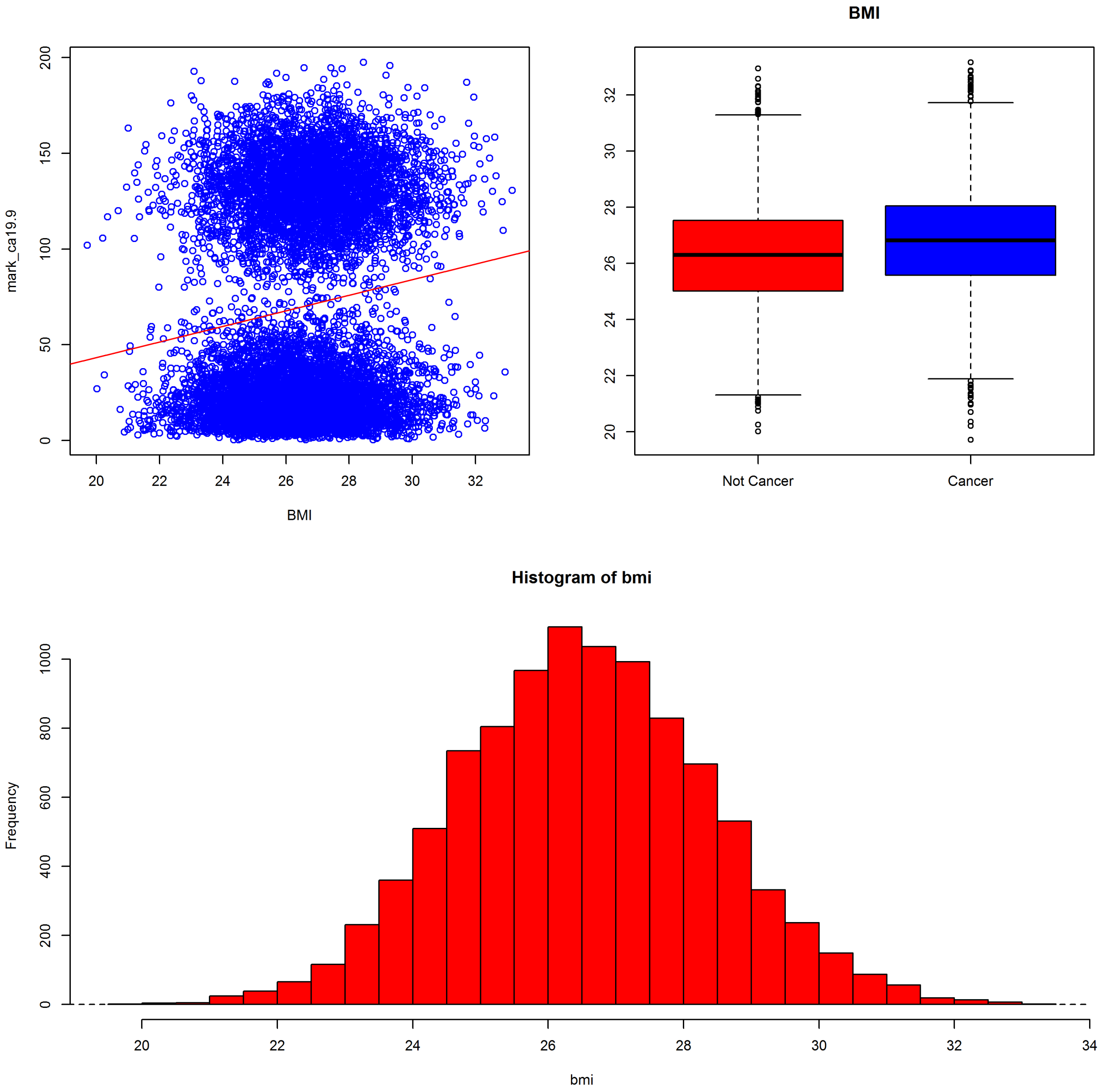


Figure 29: BMI respect tumour marker and cancer/no–cancer condition

- **AUC** is an abbreviation for area under the ROC curve. In ROC curves, the sensitivities are plotted against 1 minus specificity shifting the threshold value. It is used to determine how good are the values predicted by the classifier. The closer AUC for a model comes to 1, the better it is. So models with higher AUC are preferred over those with lower AUC.

- **F1 score**, also known as F Score or F Measure, conveys the balance between the precision and the recall. *Recall* is the number of true positives divided by the number of true positives and the number of false negatives. In other terms it is the number of positive predictions divided by the number of positive class values in the test data.

$$\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

- **Lift score** is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model. The greater is the area between the lift curve and the baseline, the better is the model.

3.1.13 Techniques evaluation

A 10 fold cross-validation procedure has been used for each combination of model and MLT. The procedure is based on a search method to tune the parameters.In this case a grid search has been performed testing all multiple combinations of search parameters. The procedure is based on a finite set of reasonable values for each combination of possible parameters; grid search then trains the technique for each combination of parameters and evaluates their performance on a validation set (or by internal cross-validation on the training set, in this case the selected techniques are trained multiple times per combination). Finally, the grid search algorithm outputs the settings that achieved the highest score in the validation procedure.

A table has been generated for each MLT (Table 34-47), comparing the model according to their precision, AUC, F1 score and Lift score. To help reading the results, among each block of models, the best value of each performance measure has been emphasized in bold.

For example, considering the bagging algorithm, on the accuracy, the maximum value of the performance including all covariates is 88.91; the same value is obtained, tuning the parameters in each run, using only relevant feature as represented in bold face. For relevant feature model a slightly better performance is obtained: 88.93% highlighted in bold.

The better performance for each model and each measure is reported in Table 48, while in Table 49 the ranking for each MLT is reported.

As can be seen in table Table 48, across all the performance measure, the worst result is obtained using most common classification techniques (**naïve**), which serves as the most basic benchmark in comparison with MLT. With regard to accuracy, the best performance is achieved by MLPE (89.61% both on overall and relevant variables models), and MLP, followed by logistic regression, while excluding **naïve** the worst performances are for KNN (87.57% on relevant variables model) and LDA. Considering the Predictive Positive Value (Precision 1) **boosting** results to achieve the best ranking(88.46% on overall model), followed by LR and MLPE, whereas SVM and LDA are ranked worst (84.87% and 83.99% respectively). The best performance for Negative Predictive Value (Precision 2) is achieved by SVM (96.83%), LDA and QDA, whereas **boosting** and KNN (90.79% and 90.26%)are ranked worst. For the AUC, MLPE achieves the best result (95.5% both on overall and relevant variables models) and LR , while the worst values are achieved by **bagging** and **rpart** (88.93% and 88.99% respectively). Similar results are given for F1, with MLPE achieving the best value (91.16% on overall and relevant variables models) followed by MLP and LR. The smallest values belong to SVM and KNN (90.38% and 89.43% respectively). Similarly, the best Lift score is achieved by MLPE, MLP and LR (0.758 on overall and relevant models) and the worst values is given for **bagging** and RPART (0.734 and 0.726 respectively).

It is important to consider that the differences between the best and the worst values is minimal across the MLTs (ranging from 2% to 7%). As can be seen in Table 34-47, for all MLTs, the performances obtained including in the model all the variables is similar (or even worst in some cases) to those obtained considering only the relevant features. Differently, it can be observed a large difference (ranging from 20% to 30%) in the performance values of models including the relevant features and models that include not relevant or only meat features. In most cases the performance of the model that containing only the meat intake feature is very similar to that one of the model that include not relevant features. In such case, the performance is comparable the **naïve** benchmark model ”naive”.

3.1.14 Feature selection

In a general context, a researcher is not informed on the relative importance of the feature influencing the outcome of interest. However it is important, for improving the capability of the model to generalize to external data, to make feature selection and include in the model only the features, which appear to be relevant according with some specific criteria.

Some MLT, like for example decision trees or random forests, have built in mechanisms for reporting the variable importance. For those MLT algoriths, which do not embed feature selection procedure, the variable importance can be estimated through a ROC analysis by assessing the impact of each variable on the accuracy of the model.

Table 34: Performance measures obtained using bagging algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	88.84	86.86	92.04	0.911	90.57	0.732
All	88.88	86.95	91.98	0.91	90.6	0.732
All	88.79	86.84	91.92	0.913	90.53	0.733
All	88.86	86.85	92.11	**0.915**	90.59	0.734
All	88.71	86.66	92.04	0.914	90.47	**0.734**
All	88.89	86.95	92.01	0.911	90.61	0.732
All	88.88	86.94	92.01	0.91	90.6	0.731
All	88.89	**86.99**	91.94	0.911	90.6	0.732
All	88.86	86.95	91.92	0.912	90.58	0.732
All	**88.91**	86.92	**92.12**	0.91	**90.63**	0.731
Relevant	88.9	86.98	91.99	0.91	90.61	0.732
Relevant	88.86	86.91	92	**0.914**	90.58	**0.734**
Relevant	88.8	86.87	91.9	0.912	90.53	0.733
Relevant	88.9	**86.99**	91.97	**0.914**	90.61	0.733
Relevant	88.84	86.87	92.02	0.911	90.57	0.732
Relevant	**88.91**	86.89	92.17	0.913	90.63	0.733
Relevant	88.9	86.94	92.06	0.91	90.62	0.731
Relevant	**88.93**	86.95	92.13	0.913	**90.65**	0.733
Relevant	88.82	86.91	91.89	0.91	90.55	0.732
Relevant	88.85	86.77	**92.22**	0.909	90.59	0.731
Not Relevant	57.68	58.92	52.6	0.534	69.09	0.517
Not Relevant	**57.69**	**58.93**	**52.63**	**0.538**	**69.1**	**0.52**
Not Relevant	**57.69**	**58.93**	52.62	0.536	69.09	0.518
Not Relevant	57.62	58.9	52.46	0.537	69.04	**0.52**
Not Relevant	57.66	58.91	52.55	0.536	69.08	0.519
Not Relevant	57.63	58.9	52.5	0.536	69.06	0.519
Not Relevant	57.66	58.91	52.54	0.537	69.07	**0.52**
Not Relevant	57.67	58.92	52.57	0.536	69.07	0.519
Not Relevant	57.67	58.92	52.57	0.536	69.08	0.519
Not Relevant	**57.69**	**58.93**	52.62	0.536	69.09	0.519
Only Meat	56.69	57.09	50.34	0.525	71.27	0.514
Only Meat	56.63	57.06	49.83	0.523	71.21	0.513
Only Meat	**56.82**	**57.14**	**51.52**	0.522	**71.41**	0.512
Only Meat	56.58	57.01	49.37	0.522	71.25	0.512
Only Meat	56.69	57.06	50.37	**0.526**	71.34	**0.515**
Only Meat	56.62	57.05	49.74	0.521	71.23	0.511
Only Meat	56.61	57.03	49.65	0.522	71.25	0.512
Only Meat	56.56	57.08	49.33	0.524	71.03	0.513
Only Meat	56.58	57.03	49.41	0.525	71.19	0.513
Only Meat	56.62	57.04	49.74	0.521	71.24	0.512

Feature selection is an important step for data mining which is often characterized by data sets with far too many variables for model building. There are two main approaches to select the features (variables):

1. **minimal-optimal** feature selection which identifies a small (ideally minimal) set of variables that gives the best possible classification result (for a class of classification models);

2. **all-relevant** feature selection which identifies all variables that are in some circumstances relevant for the classification.

The **all-relevant** Boruta algorithm was introduced by Miron B. Kursa and Witold R. Rudnicki. It is based on the more general idea that by adding randomness to a system and then collecting results from random samples of the bigger system, it is possible to reduce the misleading impact of randomness in the original sample.

For the implementation, the Boruta package relies on a random forest classification algorithm. This provides an intrinsic measure of the importance of each feature, known as the $Z - score$. While this score is not directly a statistical measure of the significance of the feature, the Boruta algorithm can compare it to random permutations of the variables to test if it is higher than the scores from random variables.

Performing the Boruta algorithm on our data, it seems that, as it could be seen in Figure 30, the relevant variables are those considered in our *relevant* model.

The better performance is evident for the neural network, specifically for MLPE. Artificial Neural Network (ANN) ensembles are techniques used to improve the generalization of a single MLP, by combining set of ANNs, which leads to achieve results better than those obtained by any single algorithm. Furthermore, the software

Table 35: Performance measures obtained using boosting algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	88.93	88.19	90.03	0.951	90.48	0.755
All	88.53	87.71	89.75	0.951	90.16	0.755
All	88.8	88.28	89.56	0.95	90.35	0.755
All	88.72	88.19	89.5	**0.951**	90.28	**0.755**
All	88.71	88.02	89.73	0.951	90.29	0.755
All	**89.06**	88.4	**90.04**	0.949	**90.58**	0.755
All	89	**88.45**	89.81	0.95	90.52	0.755
All	88.76	88.02	89.86	0.949	90.34	0.755
All	88.83	88.31	89.59	0.95	90.37	0.755
All	88.65	87.93	89.72	0.949	90.25	0.754
Relevant	89.05	88.05	90.56	0.952	90.62	0.756
Relevant	89.15	88.21	90.56	0.952	90.7	0.756
Relevant	89.04	88.32	90.11	0.951	90.58	0.755
Relevant	**89.19**	**88.34**	90.45	**0.952**	**90.72**	**0.756**
Relevant	89.03	88.09	90.43	0.951	90.59	0.756
Relevant	88.94	87.96	90.41	0.951	90.52	0.755
Relevant	89.14	88.05	**90.79**	0.952	90.71	0.756
Relevant	89.16	88.15	90.69	0.951	90.71	0.756
Relevant	89.02	88.22	90.21	0.952	90.57	0.756
Relevant	89.04	88.05	90.54	0.951	90.61	0.756
Not Relevant	57.36	58.36	52.21	0.553	**69.65**	0.53
Not Relevant	57.4	58.41	52.29	0.554	69.62	0.531
Not Relevant	57.27	58.3	51.94	0.551	69.6	0.529
Not Relevant	57.25	58.29	51.87	**0.557**	69.58	**0.532**
Not Relevant	57.3	58.35	51.98	0.553	69.54	0.53
Not Relevant	57.37	58.38	52.21	0.553	69.62	0.53
Not Relevant	**57.42**	**58.42**	**52.35**	0.552	69.63	0.53
Not Relevant	57.22	58.32	51.72	0.555	69.45	0.531
Not Relevant	56.99	58.13	51.06	0.555	69.41	0.531
Not Relevant	57.15	58.2	51.59	0.547	69.58	0.527
Only Meat	56.59	56.99	49.43	0.526	71.32	0.515
Only Meat	56.66	**57.06**	50.09	0.524	71.28	0.514
Only Meat	**56.77**	57.03	**51.34**	0.525	**71.59**	0.514
Only Meat	56.71	57	50.65	0.523	71.53	0.513
Only Meat	56.64	57.03	49.91	0.525	71.33	0.514
Only Meat	56.73	57.05	50.78	0.526	71.44	0.515
Only Meat	56.66	57.02	50.1	**0.528**	71.38	**0.516**
Only Meat	56.71	57.02	50.63	0.526	71.5	0.515
Only Meat	56.46	56.98	48.41	0.527	71.09	**0.516**
Only Meat	56.54	56.96	48.94	0.524	71.3	0.514

implementation of ANN makes quite easy to overcome overfitting situation by allowing some tuning parameters, in addition to feature selection procedures.

Generally, ensemble of MLTs (like randomForest, bagging, boosting) perform better than single MLT algorithms (like decision trees). The ensemble methods guarantees a better performance but a greater computational effort. A very good performance is evident for LR. In this case study, this fact may be explained because the outcome variable is generated from a logistic function. In the real case a logistic function may be useful for the interpretation, but doesn't handle multicollinearity problem. An advantage of LR is that the output can be interpreted as a probability, leading to a not difficult interpretation of the results, for example, for ranking instead of classification.

A disadvantage of LR is that it can be trained for problem that are linearly separable only, while other techniques like neural network are much more flexible.

A good performance is evident for the Ensemble Decision Tree methods, especially RF algorithm, considering the overall fitting measures. Tree Ensembles main advantages are that they do not expect linear features or even features that interact linearly, as well as how they handle very high dimensional spaces and large number of training examples.

Table 36: Performance measures obtained using CTREE algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	89.08	86.19	**94.05**	0.948	90.89	0.754
All	**89.13**	86.49	93.58	0.948	**90.9**	0.754
All	88.94	86.43	93.15	0.948	90.72	0.754
All	88.94	86.56	92.89	0.947	90.71	0.753
All	88.84	86.49	92.74	0.949	90.62	**0.754**
All	88.92	86.53	92.89	0.948	90.69	0.754
All	89.08	86.54	93.34	0.948	90.84	0.754
All	88.94	86.71	92.6	0.948	90.69	0.754
All	88.91	86.4	93.12	0.947	90.7	0.753
All	**89.13**	**86.71**	93.14	**0.949**	90.87	0.754
Relevant	89.07	86.19	**94.02**	0.949	90.88	0.754
Relevant	89.1	86.44	93.6	0.948	90.88	0.754
Relevant	88.92	86.4	93.14	0.949	90.71	0.754
Relevant	88.98	86.54	93.04	0.947	90.75	0.753
Relevant	88.85	86.58	92.58	**0.949**	90.62	**0.754**
Relevant	88.9	86.49	92.91	0.949	90.68	0.754
Relevant	89.03	86.64	92.98	0.948	90.78	0.754
Relevant	88.99	**86.65**	92.86	0.948	90.74	0.754
Relevant	88.93	86.35	93.26	0.947	90.72	0.753
Relevant	**89.16**	86.61	93.42	0.949	**90.91**	0.754
Not Relevant	57.69	58.94	52.62	0.543	69.09	0.524
Not Relevant	57.68	58.93	52.59	0.54	69.08	0.523
Not Relevant	**57.71**	**58.95**	**52.67**	0.546	**69.1**	0.526
Not Relevant	57.62	58.89	52.44	**0.546**	69.03	**0.526**
Not Relevant	57.67	58.92	52.56	0.544	69.07	0.525
Not Relevant	57.69	58.93	52.62	0.535	69.09	0.52
Not Relevant	57.67	58.92	52.56	0.543	69.07	0.524
Not Relevant	57.69	58.94	52.61	0.541	69.08	0.523
Not Relevant	57.69	58.94	52.62	0.543	69.09	0.524
Not Relevant	57.67	58.92	52.56	0.536	69.07	0.52
Only Meat	56.65	56.65	0	0.515	**72.33**	0.508
Only Meat	**56.68**	**56.71**	**52.24**	0.513	72.23	0.507
Only Meat	56.65	56.65	0	**0.517**	**72.33**	**0.51**
Only Meat	56.65	56.65	0	0.517	**72.33**	0.509
Only Meat	56.65	56.65	0	0.517	**72.33**	0.509
Only Meat	56.65	56.65	0	0.514	**72.33**	0.508
Only Meat	56.65	56.65	0	0.515	**72.33**	0.508
Only Meat	56.65	56.65	0	0.516	**72.33**	0.509
Only Meat	56.65	56.65	0	0.516	**72.33**	0.509
Only Meat	56.6	56.68	47.13	0.515	72.14	0.509

Finally, in all the MLT considered, the Precision1 (predictive positive value PPV) is greater respect to Precision2 (predictive negative value PNV) indicating a better performance on cancer outcome prediction.

Table 37: Performance measures obtained using KNN algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	87.34	85.67	90.03	0.925	89.3	0.74
All	87.41	85.77	90.05	**0.927**	89.36	**0.741**
All	**87.49**	85.78	**90.26**	0.925	**89.43**	0.74
All	87.22	85.7	89.65	0.924	89.18	0.74
All	87.1	85.59	89.51	0.923	89.08	0.739
All	87.34	85.66	90.06	0.924	89.3	0.74
All	87.23	85.59	89.88	0.924	89.21	0.74
All	87.29	85.8	89.67	0.924	89.23	0.74
All	87.22	**85.82**	89.44	0.922	89.16	0.739
All	87.07	85.68	89.28	0.921	89.04	0.738
Relevant	87.23	86.5	88.33	0.933	89.06	0.745
Relevant	87.46	86.66	88.67	0.933	89.27	0.746
Relevant	87.27	86.5	88.44	0.932	89.11	0.745
Relevant	87.46	86.59	88.79	**0.934**	89.28	**0.746**
Relevant	87.36	86.52	88.64	0.932	89.19	0.745
Relevant	**87.57**	86.64	**89**	0.932	**89.38**	0.745
Relevant	87.37	86.56	88.61	0.931	89.19	0.744
Relevant	**87.49**	**86.73**	88.64	0.933	89.29	0.745
Relevant	87.5	86.65	88.8	0.933	89.31	0.745
Relevant	87.52	86.63	88.88	0.933	89.33	0.745
Not Relevant	52.75	57.69	44.98	0.518	59.89	0.51
Not Relevant	**53.08**	**58.03**	**45.47**	0.52	59.98	**0.512**
Not Relevant	52.22	57.31	44.37	0.508	59.26	0.505
Not Relevant	52.83	57.84	45.16	0.519	59.73	0.511
Not Relevant	52.92	57.85	45.22	**0.521**	59.98	0.511
Not Relevant	52.61	57.61	44.84	0.517	59.65	0.51
Not Relevant	52.96	57.8	45.2	0.515	**60.22**	0.508
Not Relevant	51.99	57.08	44.03	0.513	59.17	0.507
Not Relevant	53.03	57.95	45.37	0.517	60.05	0.51
Not Relevant	52.59	57.63	44.86	0.517	59.53	0.51
Only Meat	51.57	56.84	43.63	0.506	58.53	0.502
Only Meat	**52.85**	**57.62**	**44.96**	**0.516**	**60.36**	**0.509**
Only Meat	52.07	57.07	44.02	0.513	59.48	0.507
Only Meat	52.4	57.42	44.55	**0.516**	59.53	textbf0.509
Only Meat	51.95	56.93	43.8	0.512	59.55	0.507
Only Meat	52.26	57.31	44.38	0.512	59.4	0.507
Only Meat	52.11	57.17	44.16	0.51	59.32	0.506
Only Meat	52.62	57.4	44.61	0.514	60.28	0.508
Only Meat	52.2	57.35	44.41	0.51	59.08	0.505
Only Meat	52.54	57.43	44.61	0.51	59.95	0.506

Table 38: Performance measures obtained using LDA algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	88.21	83.93	96.56	0.954	90.39	0.757
All	**88.24**	**83.95**	**96.59**	0.954	**90.42**	**0.757**
All	88.17	83.88	96.56	**0.954**	90.36	0.757
All	88.21	83.93	96.56	0.954	90.39	0.757
All	88.22	83.95	96.53	0.954	90.4	0.757
All	88.2	83.9	96.59	0.954	90.39	0.757
All	88.21	83.92	96.59	0.954	90.39	**0.757**
All	88.21	83.93	96.56	0.954	90.39	0.757
All	**88.24**	**83.95**	**96.59**	0.954	**90.42**	0.757
All	88.19	83.9	96.56	0.954	90.38	0.757
Relevant	88.22	83.94	96.56	0.954	90.4	0.757
Relevant	**88.27**	**83.99**	96.59	0.954	**90.44**	0.757
Relevant	88.22	83.93	96.59	0.955	90.4	**0.757**
Relevant	88.17	83.88	96.56	0.954	90.36	0.757
Relevant	88.22	83.95	96.53	0.955	90.4	0.757
Relevant	88.25	83.97	96.59	0.954	90.42	0.757
Relevant	88.25	83.97	96.59	0.955	90.42	0.757
Relevant	88.23	83.94	96.59	0.954	90.41	0.757
Relevant	88.21	83.92	96.59	**0.955**	90.39	0.757
Relevant	88.21	83.91	**96.61**	0.954	90.4	0.757
Not Relevant	57.14	57.56	52.8	0.546	71.01	0.526
Not Relevant	57.2	57.59	53.17	0.548	71.07	0.527
Not Relevant	57.07	57.52	52.42	0.548	70.98	0.527
Not Relevant	57.23	57.6	53.36	0.548	71.1	0.527
Not Relevant	**57.35**	**57.68**	**53.99**	0.548	**71.15**	0.527
Not Relevant	57.12	57.53	52.77	**0.548**	71.05	**0.527**
Not Relevant	57.2	57.58	53.21	0.547	71.09	0.527
Not Relevant	57.08	57.51	52.52	**0.548**	71.01	0.527
Not Relevant	57.12	57.53	52.75	0.547	71.04	0.527
Not Relevant	57.19	57.58	53.14	**0.548**	71.08	0.527
Only Meat	56.67	**56.67**	56.25	**0.526**	72.31	0.515
Only Meat	56.65	**56.67**	50	0.525	72.28	0.514
Only Meat	56.62	56.64	36.36	**0.526**	72.29	0.515
Only Meat	56.66	56.66	53.85	**0.526**	72.31	0.515
Only Meat	56.6	56.63	30.77	**0.526**	72.27	**0.515**
Only Meat	56.65	56.66	50	**0.526**	72.3	0.515
Only Meat	56.66	**56.67**	52.94	**0.526**	72.3	0.515
Only Meat	**56.68**	**56.67**	**60**	**0.526**	**72.32**	0.515
Only Meat	56.58	56.63	34.78	**0.526**	72.24	0.515
Only Meat	56.62	56.65	40	**0.526**	72.28	0.515

Table 39: Performance measures obtained using LR algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	89.5	88.21	**91.47**	**0.955**	91.03	**0.758**
All	**89.52**	**88.27**	**91.43**	0.955	**91.05**	0.758
All	89.48	88.23	91.38	0.955	91.01	0.758
All	89.44	88.19	91.35	0.955	90.98	0.758
All	89.47	88.21	91.4	0.955	91	0.758
All	89.5	88.21	**91.47**	0.955	91.03	0.758
All	89.51	88.25	91.43	0.955	91.04	0.758
All	89.47	88.18	91.44	0.955	91.01	0.758
All	89.49	88.24	91.41	0.955	91.02	0.758
All	89.45	88.2	91.36	0.955	90.99	0.758
Relevant	89.48	88.22	91.41	**0.955**	91.01	**0.758**
Relevant	89.5	88.25	91.41	0.955	91.03	0.758
Relevant	89.45	88.22	91.33	0.955	90.98	0.758
Relevant	89.48	88.23	91.38	0.955	91.01	0.758
Relevant	89.5	88.25	91.41	0.955	91.03	0.758
Relevant	89.48	88.23	91.38	0.955	91.01	0.758
Relevant	89.47	88.19	91.42	0.955	91.01	0.758
Relevant	**89.52**	**88.27**	**91.43**	0.955	**91.05**	0.758
Relevant	89.48	88.22	91.41	0.955	91.01	0.758
Relevant	89.45	88.2	91.36	0.955	90.99	0.758
Not Relevant	57.14	57.55	52.83	0.546	71.03	0.526
Not Relevant	57.16	57.56	52.96	0.548	71.06	0.527
Not Relevant	57.02	57.48	52.15	0.548	70.96	0.527
Not Relevant	57.22	57.59	53.33	0.548	71.1	0.527
Not Relevant	**57.3**	**57.65**	**53.73**	0.548	**71.13**	0.527
Not Relevant	57.07	57.5	52.49	**0.548**	71.03	**0.527**
Not Relevant	57.18	57.56	53.11	0.547	71.08	0.527
Not Relevant	57.1	57.52	52.64	0.548	71.03	0.527
Not Relevant	57.1	57.52	52.66	0.547	71.04	0.527
Not Relevant	57.11	57.52	52.71	0.548	71.04	0.527
Only Meat	56.66	56.66	53.33	0.526	72.31	0.515
Only Meat	56.64	56.66	48	0.525	72.28	0.514
Only Meat	56.64	56.65	44.44	0.526	72.31	0.515
Only Meat	56.65	56.66	50	0.526	72.31	0.515
Only Meat	56.6	56.63	30.77	**0.526**	72.27	**0.515**
Only Meat	56.65	56.66	50	0.526	72.31	0.515
Only Meat	56.66	56.66	53.33	0.526	72.31	0.515
Only Meat	**56.69**	**56.68**	**66.67**	0.526	**72.33**	0.515
Only Meat	56.58	56.63	34.78	0.526	72.24	0.515
Only Meat	56.61	56.64	35.71	0.526	72.28	0.515

Table 40: Performance measures obtained using MLP algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	89.37	87.8	91.83	0.955	90.96	0.758
All	89.47	88.04	91.68	0.955	91.02	0.758
All	89.48	**88.11**	91.59	**0.955**	91.03	**0.758**
All	89.38	88.08	91.38	0.955	90.93	0.758
All	89.41	88.04	91.52	0.955	90.97	0.758
All	**89.5**	88.05	91.75	0.954	91.05	0.757
All	**89.5**	87.92	**91.96**	0.955	**91.07**	0.758
All	89.39	87.96	91.6	0.955	90.96	0.758
All	89.41	87.94	91.69	0.955	90.98	0.758
All	89.41	87.96	91.67	0.954	90.98	0.757
Relevant	89.35	87.78	91.8	0.955	90.94	0.758
Relevant	89.49	88.1	91.64	0.954	91.04	0.757
Relevant	89.36	87.92	91.59	0.954	90.93	0.757
Relevant	89.47	88.14	91.51	0.955	91.01	0.758
Relevant	89.51	87.98	91.9	0.955	91.07	0.758
Relevant	89.38	87.87	91.72	0.955	90.96	0.758
Relevant	89.52	88.05	91.79	**0.955**	91.07	**0.758**
Relevant	89.56	**88.16**	91.69	0.955	91.09	0.758
Relevant	**89.56**	88	**92**	0.955	**91.12**	0.758
Relevant	89.47	87.95	91.83	0.955	91.03	0.758
Not Relevant	56.63	57.53	49.91	0.537	70.06	0.521
Not Relevant	**57.16**	**57.68**	**52.56**	**0.547**	70.78	**0.527**
Not Relevant	56.75	57.42	50.52	0.545	70.56	0.525
Not Relevant	56.95	57.57	51.48	0.542	70.61	0.524
Not Relevant	57.07	57.57	52.26	0.544	**70.86**	0.525
Not Relevant	56.85	57.44	51.09	0.544	70.73	0.525
Not Relevant	56.81	57.63	50.68	0.54	70.19	0.523
Not Relevant	56.96	57.63	51.44	0.547	70.48	0.526
Not Relevant	56.9	57.54	51.24	0.543	70.58	0.524
Not Relevant	56.66	57.35	50.05	0.54	70.55	0.523
Only Meat	56.79	57.11	51.28	0.521	71.41	0.512
Only Meat	56.45	56.82	47.5	0.52	**71.47**	0.511
Only Meat	56.84	57.19	51.53	0.523	71.31	0.513
Only Meat	56.63	56.98	49.79	0.523	71.43	0.513
Only Meat	56.81	**57.21**	51.22	0.523	71.21	0.513
Only Meat	56.82	57.17	51.41	**0.526**	71.33	**0.515**
Only Meat	56.71	57.04	50.58	0.523	71.42	0.513
Only Meat	56.48	56.86	48.04	0.524	71.42	0.514
Only Meat	**56.86**	57.17	**51.81**	0.521	71.39	0.512
Only Meat	56.63	57.08	49.84	0.522	71.18	0.512

Table 41: Performance measures obtained using MLPE algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	**89.61**	87.96	**92.18**	0.955	**91.16**	0.758
All	89.42	87.97	91.67	0.955	90.98	0.757
All	89.42	88.05	91.54	0.954	90.98	0.757
All	89.33	87.95	91.46	0.955	90.9	0.758
All	89.47	87.91	91.91	0.955	91.04	0.758
All	89.34	87.94	91.5	0.955	90.91	0.758
All	**89.61**	**88.19**	91.81	**0.955**	91.14	**0.758**
All	89.46	**88.07**	91.61	0.955	91.01	0.758
All	89.43	87.93	91.75	0.955	91	0.758
All	89.41	87.96	91.67	0.955	90.98	0.758
Relevant	**89.61**	87.93	**92.22**	0.955	**91.16**	0.758
Relevant	89.47	87.93	91.87	0.955	91.04	0.758
Relevant	89.35	87.96	91.5	0.955	90.92	0.758
Relevant	89.53	87.89	92.1	0.955	91.1	0.758
Relevant	89.42	87.8	91.97	0.955	91.01	0.758
Relevant	89.42	87.78	91.99	0.955	91.01	0.758
Relevant	89.42	87.96	91.69	0.955	90.99	0.758
Relevant	89.39	87.9	91.7	0.955	90.97	0.758
Relevant	89.59	88.02	92.05	0.955	91.14	0.758
Relevant	89.46	**88.07**	91.61	**0.955**	91.01	**0.758**
Not Relevant	57.07	**57.73**	51.88	0.54	70.48	0.523
Not Relevant	56.92	57.66	51.19	0.545	70.33	0.525
Not Relevant	57.12	57.69	52.27	**0.548**	70.68	**0.527**
Not Relevant	57.13	57.7	52.3	0.547	70.67	0.527
Not Relevant	**57.18**	57.6	**52.96**	0.547	**71**	0.526
Not Relevant	57.11	57.67	52.26	0.547	70.71	0.527
Not Relevant	57.09	57.64	52.2	0.545	70.73	0.526
Not Relevant	56.86	57.49	51.08	0.545	70.63	0.525
Not Relevant	56.63	57.36	49.9	0.546	70.47	0.526
Not Relevant	57.02	57.55	51.98	0.546	70.81	0.526
Only Meat	56.77	56.96	51.68	0.52	71.75	0.511
Only Meat	56.75	56.87	52.02	0.523	71.94	0.513
Only Meat	56.94	**57.06**	53.85	**0.525**	71.83	**0.514**
Only Meat	56.79	57.02	51.64	0.519	71.64	0.511
Only Meat	56.74	57.02	50.98	0.524	71.54	0.514
Only Meat	56.58	56.88	49.07	0.523	71.59	0.513
Only Meat	56.73	56.91	51.28	0.524	71.81	0.514
Only Meat	**56.95**	56.94	**57.43**	0.523	**72.16**	0.513
Only Meat	56.72	56.9	51.11	0.525	71.8	0.514
Only Meat	56.78	56.91	52.28	0.522	71.9	0.512

Table 42: Performance measures obtained using most common classes of algorithms (naïve)

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
All	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Not Relevant	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.497**

Table 43: Performance measures obtained using NB

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	88.62	87.48	90.36	0.942	90.28	0.75
All	88.63	87.51	90.34	0.942	90.28	0.75
All	88.53	87.42	90.21	0.942	90.2	0.75
All	88.55	87.42	90.28	0.942	90.22	0.75
All	88.56	87.42	90.3	0.942	90.23	0.75
All	88.61	87.44	90.39	0.942	90.27	0.75
All	**88.68**	**87.52**	**90.45**	**0.942**	**90.33**	**0.75**
All	88.56	87.46	90.24	0.942	90.22	0.75
All	88.57	87.45	90.28	0.942	90.23	0.75
All	88.58	87.45	90.31	0.942	90.24	0.75
Relevant	**88.68**	87.07	91.2	0.945	90.39	0.752
Relevant	88.66	87.05	91.2	0.945	90.37	0.752
Relevant	88.62	87.03	91.13	0.945	90.34	0.752
Relevant	88.66	87.09	91.12	0.945	90.37	0.752
Relevant	88.66	87.06	91.18	0.945	90.37	0.752
Relevant	**88.7**	87.09	**91.23**	0.945	**90.41**	0.752
Relevant	88.67	87.05	91.22	**0.945**	90.38	**0.752**
Relevant	88.66	87.08	91.14	0.944	90.37	0.752
Relevant	88.6	87.01	91.1	0.945	90.32	0.752
Relevant	88.67	**87.1**	91.14	0.945	90.38	0.752
Not Relevant	57.03	57.84	51.49	0.546	70.13	0.526
Not Relevant	**57.27**	57.96	**52.48**	**0.548**	**70.34**	**0.527**
Not Relevant	57.13	57.88	51.92	0.547	70.24	0.527
Not Relevant	57.16	57.92	52.01	0.548	70.23	0.527
Not Relevant	57.26	57.97	52.41	0.548	70.31	0.527
Not Relevant	57.17	57.92	52.05	0.547	70.23	0.527
Not Relevant	57.19	57.94	52.13	0.546	70.25	0.526
Not Relevant	57.16	57.91	52.01	0.547	70.23	0.527
Not Relevant	57.15	57.91	51.98	0.546	70.23	0.526
Not Relevant	57.25	**57.97**	52.35	0.547	70.28	0.527
Only Meat	56.75	57.03	51.06	0.523	71.53	0.513
Only Meat	56.74	57.02	50.98	0.523	71.55	0.513
Only Meat	**56.89**	**57.11**	**52.52**	**0.525**	**71.61**	**0.514**
Only Meat	56.67	56.99	50.22	0.525	71.49	0.514
Only Meat	56.72	57.02	50.73	0.524	71.49	0.514
Only Meat	56.78	57.05	51.37	0.525	71.55	0.514
Only Meat	56.75	57.04	51.05	0.524	71.52	0.514
Only Meat	56.86	57.09	52.23	0.524	71.6	0.514
Only Meat	56.78	57.06	51.33	0.524	71.52	0.513
Only Meat	56.75	57.03	51.07	0.523	71.54	0.513

Table 44: Performance measures obtained using QDA algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	88.62	85.34	94.46	0.951	90.57	0.756
All	88.65	85.34	**94.57**	0.951	90.6	0.756
All	88.63	85.34	94.49	0.952	90.58	0.756
All	88.62	85.34	94.46	0.952	90.57	0.756
All	88.58	85.26	94.51	0.952	90.54	0.756
All	**88.69**	**85.41**	94.52	0.951	**90.63**	0.756
All	88.59	85.29	94.48	**0.952**	90.55	**0.756**
All	88.61	85.35	94.41	0.952	90.56	0.756
All	88.64	85.34	94.52	0.951	90.59	0.755
All	88.6	85.31	94.46	0.952	90.56	0.756
Relevant	88.65	85.34	**94.57**	0.952	90.6	0.756
Relevant	88.67	85.33	**94.64**	0.951	90.62	0.756
Relevant	88.67	85.35	94.59	0.952	90.62	0.756
Relevant	88.65	85.31	94.62	0.952	90.6	0.756
Relevant	88.62	85.3	94.56	0.952	90.58	0.756
Relevant	88.62	85.33	94.49	0.951	90.57	0.756
Relevant	88.64	85.31	94.59	**0.953**	90.59	**0.756**
Relevant	88.6	85.31	94.46	0.952	90.56	0.756
Relevant	88.66	85.34	94.59	0.951	90.61	0.756
Relevant	**88.68**	**85.35**	94.62	0.952	**90.63**	0.756
Not Relevant	56.92	57.6	51.27	0.543	70.49	0.524
Not Relevant	56.86	57.55	51	**0.546**	70.47	**0.526**
Not Relevant	56.93	57.6	51.33	0.544	70.51	0.525
Not Relevant	56.82	57.54	50.8	0.546	70.42	0.526
Not Relevant	56.99	57.62	51.63	0.546	70.57	0.526
Not Relevant	56.96	57.62	51.46	0.544	70.52	0.525
Not Relevant	**56.89**	57.58	51.13	0.542	70.46	0.524
Not Relevant	56.98	57.63	51.55	0.544	70.53	0.525
Not Relevant	56.97	57.62	51.52	0.542	70.54	0.524
Not Relevant	**57.06**	**57.68**	**51.93**	0.545	**70.58**	0.525
Only Meat	56.75	57.03	51.06	0.523	71.53	0.513
Only Meat	56.74	57.02	50.98	0.523	71.55	0.513
Only Meat	**56.89**	**57.11**	**52.52**	**0.525**	**71.61**	**0.514**
Only Meat	56.67	56.99	50.22	0.525	71.49	0.514
Only Meat	56.72	57.02	50.73	0.524	71.49	0.514
Only Meat	56.78	57.05	51.37	0.525	71.55	0.514
Only Meat	56.75	57.04	51.05	0.524	71.52	0.514
Only Meat	56.86	57.09	52.23	0.524	71.6	0.514
Only Meat	56.78	57.06	51.33	0.524	71.52	0.513
Only Meat	56.75	57.03	51.07	0.523	71.54	0.513

Table 45: Performance measures obtained using RF algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	89.35	87.62	**92.08**	0.949	90.96	0.754
All	**89.36**	**87.7**	91.98	**0.95**	**90.96**	**0.755**
All	89.12	87.51	91.64	0.949	90.75	0.755
All	89.21	87.55	91.81	0.949	90.84	0.755
All	89.16	87.41	91.94	0.948	90.81	0.754
All	89.08	87.49	91.57	0.949	90.72	0.754
All	89.18	87.5	91.83	0.949	90.81	0.754
All	89.2	87.51	91.86	0.949	90.83	0.754
All	89.06	87.26	91.91	0.948	90.73	0.754
All	89.14	87.51	91.69	0.949	90.77	0.754
Relevant	**89.29**	87.69	91.79	**0.95**	**90.89**	**0.755**
Relevant	89.14	87.4	91.89	0.95	90.79	0.755
Relevant	89.12	87.25	92.1	0.949	90.79	0.754
Relevant	89.01	87.1	92.08	0.947	90.7	0.753
Relevant	89.13	87.4	91.86	0.949	90.78	0.755
Relevant	89.18	**87.76**	91.38	0.949	90.78	0.755
Relevant	89.28	87.64	91.85	0.95	90.89	0.755
Relevant	89.24	87.39	92.19	0.949	90.89	0.755
Relevant	89.14	87.06	**92.5**	0.949	90.83	0.754
Relevant	89.13	87.28	92.08	0.95	90.8	0.755
Not Relevant	54	57.7	45.7	0.529	63.43	0.516
Not Relevant	54.31	57.98	46.26	0.529	63.54	0.517
Not Relevant	54.48	58.01	46.44	0.529	63.91	0.517
Not Relevant	54.36	58.02	46.35	0.533	63.57	0.519
Not Relevant	54.8	**58.31**	**47.03**	0.533	63.99	0.518
Not Relevant	54.48	58.15	46.58	0.528	63.57	0.516
Not Relevant	54.12	57.78	45.88	0.528	63.55	0.516
Not Relevant	**54.81**	58.17	46.92	**0.533**	**64.37**	**0.519**
Not Relevant	54.28	57.93	46.19	0.53	63.58	0.517
Not Relevant	54.07	57.77	45.83	0.526	63.45	0.515
Only Meat	51.39	57.01	43.83	0.508	57.38	0.504
Only Meat	52.08	57.71	44.73	**0.518**	57.69	**0.51**
Only Meat	51.42	57.13	43.97	0.515	57.12	0.509
Only Meat	51.78	57.37	44.32	0.514	57.65	0.508
Only Meat	51.68	57.36	44.27	0.512	57.36	0.507
Only Meat	51.6	57.28	44.18	0.511	57.29	0.506
Only Meat	51.61	57.27	44.16	0.511	57.37	0.506
Only Meat	**52.13**	**57.76**	**44.8**	0.516	**57.7**	0.509
Only Meat	51.71	57.27	44.2	0.511	57.69	0.506
Only Meat	51.66	57.29	44.2	0.511	57.47	0.506

Table 46: Performance measures obtained using RPART algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	88.73	86.93	91.6	0.896	90.46	0.724
All	**88.99**	87.3	91.66	0.896	**90.66**	0.724
All	88.84	87.19	91.43	**0.899**	90.53	**0.726**
All	88.88	87.25	91.44	0.897	90.56	0.725
All	88.77	86.98	91.61	0.897	90.49	0.725
All	88.86	87.08	91.7	0.898	90.56	0.726
All	88.79	86.99	91.66	0.899	90.51	0.726
All	88.85	87.17	91.5	0.898	90.54	0.726
All	88.69	86.81	**91.7**	0.898	90.43	0.726
All	**88.99**	**87.35**	91.57	0.897	90.65	0.725
Relevant	88.73	86.93	91.6	0.896	90.46	0.724
Relevant	**88.99**	87.3	91.66	0.896	**90.66**	0.724
Relevant	88.84	87.19	91.43	**0.899**	90.53	**0.726**
Relevant	88.88	87.25	91.44	0.897	90.56	0.725
Relevant	88.77	86.98	91.61	0.897	90.49	0.725
Relevant	88.86	87.08	91.7	0.898	90.56	0.726
Relevant	88.79	86.99	91.66	0.899	90.51	0.726
Relevant	88.85	87.17	91.5	0.898	90.54	0.726
Relevant	88.69	86.81	**91.7**	0.898	90.43	0.726
Relevant	**88.99**	**87.35**	91.57	0.897	90.65	0.725
Not Relevant	57.7	58.94	52.65	**0.532**	69.09	**0.518**
Not Relevant	57.69	58.93	52.62	0.53	69.09	0.517
Not Relevant	**57.72**	**58.95**	**52.69**	0.53	**69.11**	0.517
Not Relevant	57.62	58.9	52.46	0.532	69.04	0.518
Not Relevant	57.51	58.79	52.2	0.53	69.03	0.517
Not Relevant	57.68	58.92	52.6	0.53	69.08	0.517
Not Relevant	57.68	58.93	52.59	0.531	69.08	0.517
Not Relevant	57.7	58.94	52.64	0.528	69.09	0.516
Not Relevant	57.7	58.94	52.65	0.53	69.09	0.517
Not Relevant	57.68	58.93	52.59	0.53	69.08	0.517
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	0.496
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	0.498
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	0.497
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	0.497
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	0.498
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	**0.498**
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	0.498
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	0.498
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	0.498
Only Meat	**56.65**	**56.65**	**0**	**0.5**	**72.33**	0.496

Table 47: Performance measures obtained using SVM algorithm

Model	Accuracy	Precision1	Precision2	AUC	F1	Lift
All	87.75	83.94	94.97	0.942	89.97	0.75
All	87.56	83.39	95.75	0.946	89.88	0.753
All	87.2	82.73	96.28	0.947	89.65	0.753
All	87.51	83.27	95.9	0.944	89.85	0.751
All	87.9	84.11	95.02	0.944	90.07	0.751
All	87.2	82.66	96.48	0.947	89.66	0.753
All	**88.34**	**84.87**	94.65	0.945	**90.38**	0.752
All	88.16	84.8	94.22	0.946	90.21	0.752
All	87.39	82.88	**96.56**	**0.947**	89.8	**0.753**
All	87.27	83	95.76	0.945	89.67	0.752
Relevant	88.22	84.43	95.3	0.946	90.33	0.753
Relevant	87.67	83.66	95.38	0.945	89.93	0.752
Relevant	87.08	82.57	96.3	0.947	89.56	0.753
Relevant	86.94	82.21	**96.83**	**0.947**	89.5	**0.754**
Relevant	**88.32**	**84.78**	94.8	0.945	**90.37**	0.752
Relevant	87.84	83.77	95.67	0.945	90.07	0.752
Relevant	87.86	84.17	94.73	0.945	90.03	0.752
Relevant	87.22	82.6	96.74	0.946	89.69	0.753
Relevant	87.08	82.59	96.24	0.947	89.56	0.753
Relevant	87.7	83.61	95.66	0.946	89.97	0.752
Not Relevant	56.79	56.85	53.65	**0.544**	**72.07**	**0.525**
Not Relevant	56.56	56.8	48.44	0.54	71.75	0.523
Not Relevant	56.78	57.01	51.55	0.532	71.65	0.518
Not Relevant	56.76	56.89	52.05	0.534	71.91	0.519
Not Relevant	56.76	56.99	51.37	0.54	71.67	0.523
Not Relevant	56.85	56.91	**54.46**	0.537	72.05	0.521
Not Relevant	56.75	56.84	52.4	0.535	72.02	0.52
Not Relevant	56.59	56.87	49.18	0.538	71.63	0.521
Not Relevant	56.58	56.73	47.93	0.526	71.98	0.515
Not Relevant	**56.88**	**57.04**	52.99	0.538	71.78	0.522
Only Meat	56.67	56.78	50.59	0.51	72.03	0.505
Only Meat	56.64	56.73	49.61	0.507	72.09	0.504
Only Meat	56.63	56.69	48.72	0.518	72.17	0.51
Only Meat	56.68	**56.82**	50.67	0.517	71.94	0.51
Only Meat	56.54	56.65	43.53	0.512	72.1	0.507
Only Meat	56.69	56.71	53.7	0.512	72.26	0.507
Only Meat	56.66	56.8	50.23	**0.519**	71.95	**0.511**
Only Meat	**56.72**	56.8	52.05	0.517	72.06	0.51
Only Meat	56.62	56.7	48.31	0.51	72.15	0.506
Only Meat	56.71	56.72	**55.56**	0.511	**72.27**	0.506

Table 48: Summary of the best algorithms according to their performance measures

MLT	Accuracy	Precision1	Precision2	AUC	F1	Lift
bagging	88.93	86.99	92.22	0.915	90.65	0.734
boosting	89.19	**88.45**	90.79	0.952	90.72	0.756
ctree	89.16	86.71	94.05	0.949	90.91	0.754
knn	87.57	86.73	90.26	0.934	89.43	0.746
lda	88.27	83.99	96.61	0.955	90.44	0.757
lr	89.52	88.27	91.47	0.955	91.05	0.758
mlp	89.56	88.16	92	0.955	91.12	0.758
mlpe	**89.61**	88.19	92.22	**0.955**	**91.16**	**0.758**
naive	56.65	56.65	0	0.5	72.33	0.497
naiveBayes	88.7	87.52	91.23	0.945	90.41	0.752
qda	88.69	85.41	94.64	0.953	90.63	0.756
randomForest	89.36	87.76	92.5	0.95	90.96	0.755
rpart	88.99	87.35	91.7	0.899	90.66	0.726
svm	88.34	84.87	**96.83**	0.947	90.38	0.754

Table 49: Ranking of algorithms according to their performance measures

MLT	Accuracy	Precision1	Precision2	AUC	F1	Lift
bagging	8	8	6	12	8	12
boosting	5	1	12	6	6	6
ctree	6	10	4	8	5	8
knn	13	9	13	11	13	11
lda	12	13	2	4	10	4
lr	3	2	10	2	3	2
mlp	2	4	8	3	2	3
mlpe	1	3	7	1	1	1
naive	14	14	14	14	14	14
naiveBayes	9	6	11	10	11	10
qda	10	11	3	5	9	5
randomForest	4	5	5	7	4	7
rpart	7	7	9	13	7	13
svm	11	12	1	9	12	9

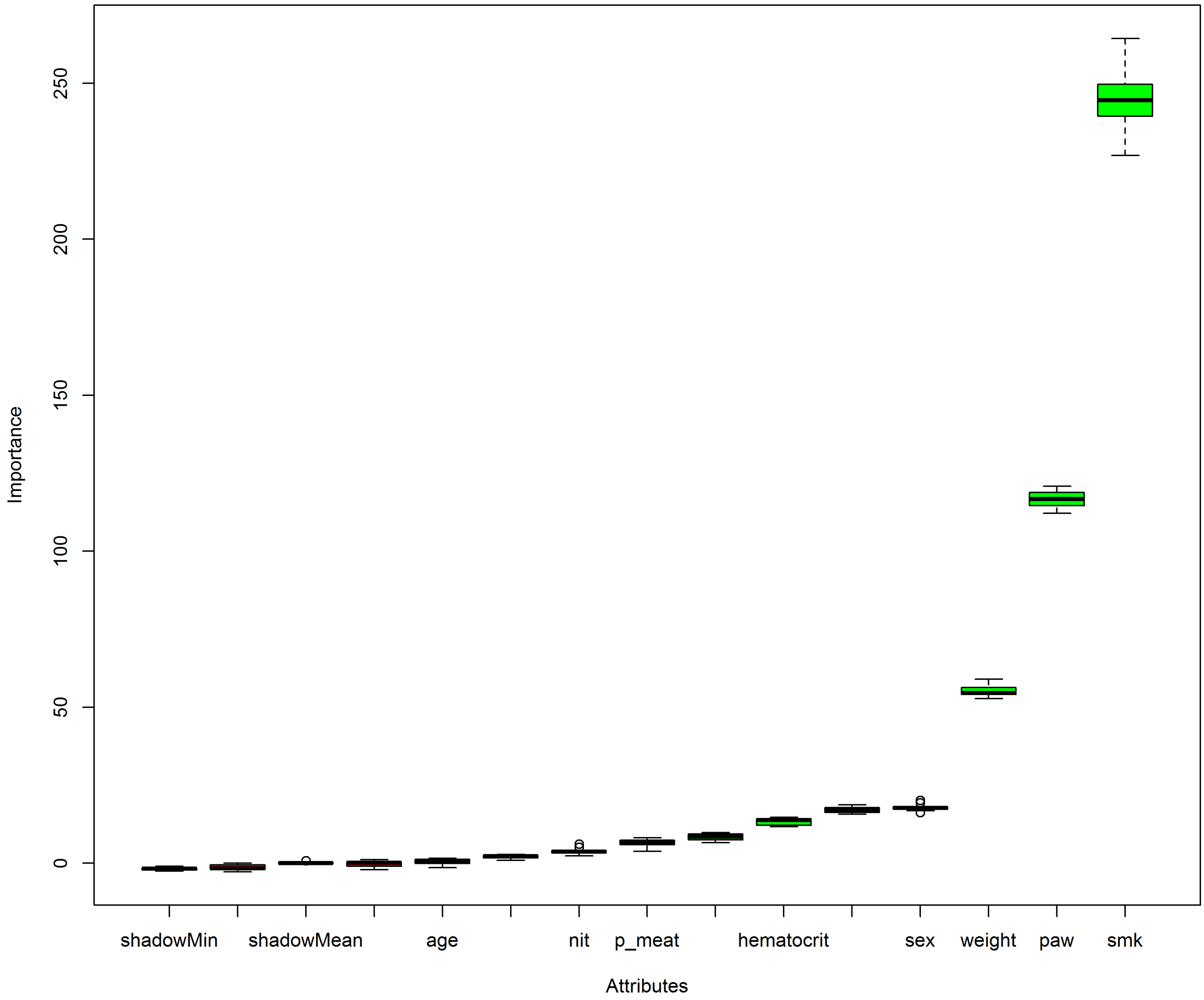


Figure 30: Importance of database variables computed by Boruta algorithm

3.2 RA of safety of the “F” feed additive

Regulation (EC) No. 1831/20031 establishes the rules governing the Community authorization of additives for use in animal nutrition. In particular, Article 4(1) of that Regulation lays down that any person seeking authorization for a feed additive or for a new use of a feed additive shall submit an application in accordance with Article 7. According to Article 7(1) of Regulation (EC) No 1831/2003, the Commission forwarded the application to EFSA as an application under Article 4(1). EFSA received directly from the applicant the technical dossier in support of this application. According to Article 8 of that Regulation, EFSA, after verifying the particulars and documents submitted by the applicant, shall undertake an assessment in order to determine whether the feed additive complies with the conditions laid down in Article 5. The feed additive “F” appears to be quite effective, but it shows a certain degree of toxicity. The private company provided EFSA with a set of data, as the result of a few experiments where the impact of different doses on different parameters are recorded.

3.2.1 Terms of References

- To estimate an appropriate BMDL based on the toxicological studies provided.

In particular, the document should consider Akaike’s information criterion (AIC) for model choice.

3.2.2 Definition of objectives

The estimation of the dose-response curve is a continuous problem which may have many types of Benchmark Responses (BMRs). Ideally, the BMR must reflect an effect size that is negligible or non-adverse. We choose a BMR corresponding to a 5% increase in response compared to the background, the recommendation for continuous data collected on animal studies (Woutersen et al., 2001).

3.2.3 Methods

The Benchmark Dose (BMD) approach required is the accepted tool for the analysis of dose-response data for RA aimed at overcoming the limitations of NOAEL in deriving an appropriate BMDL. The BMDL is used as reference point to derive a health-based guidance value from animal data (Crump, 1995).

In the present explorative case studies we apply some MLTs for carrying out BMD in order to analyze the three datasets provided. All the ML algorithms are trained on all the three sets of feeds separately and next used to estimate the dose-response curve on a pseudo-continuous range, i.e. 10^5 points of equally separated dose spanning from 0 to 150, the maximum dose actually considered. Next the BMD, BMDL, the ratio from both and AIC, used to select the best model, are computed.

To summarize we consider the following ML algorithms: naïve (the simple mean of the output, used as a benchmark), conditional inference tree, decision tree, k-nearest neighbor, multilayer perceptron with one hidden layer, multilayer perceptron ensemble, support vector machine, randomForest, multiple regression using artificial neural networks with zero hidden layer and with linear output function, multivariate additive regression splines, principal component regression, partial last squared regression, canonical powered partial last squares and relevance vector machine. All of them are detailed described in the relative sections of the section 4.1.

Data summary and BMD approach

Data have been organized in three distinct databases, with the general aim of identifying BMD. A description of the database is reported in Table 50.

The classical methodology adopted is based on the maximum likelihood estimator (MLE) approach. Applying this methodology, the analysis is carried out according to the common practice employed, which chooses the BMDL estimates from an adequately fitting model with the lowest Akaike Information Criteria/Bayesian Information Criteria or in alternative, when the range of the BMDL models is broad, it considers the lowest BMDL.

For continuous response, the recommended models to use are the Exponential (3 or 4 parameters) or the Hill family models (3 or 4 parameters). Once the models are fitted, the procedure of choice of the model and the corresponding confidence intervals for BMD has been followed as suggested in (Slob, 2002) and (European Food Safety Authority, 2017).

Table 50: Descriptive Statistics by dose and feed

	Study 1			Study 2			Study 3		
Dose	N	Mean	SD	N	Mean	SD	N	Mean	SD
0.00	10	29.13	8.08	15	13.42	4.84	20	2,68	0,08
1.38	10	28.94	10.57	15	14.93	6.73	20	2,71	0,12
34.70	10	40.15	6.60	15	17.43	17.43	20	2,76	0,09
104.70	10	76.57	8.69	15	24.28	24.28	20	2,85	0,07

To summarize, the following models have been suggested and are reported below along with their functional form:

Reference models:

- Full model: set of observed means at each dose
- Null model: $y = a$

Exponential family models:

- 3 parameters: $y = a\,\text{exp}(bx^d)$
- 4 parameters: $a[c - (c-1)\,\text{exp}(-bx^d)]$

Hill family models:

- 3 parameters: $y = a[1 - x^d/(b^d + x^d)]$
- 4 parameters: $y = a[1 + (c-1)x^d/(b^d + x^d)]$

For the three feeds, BMD estimates are provided in Table 51, Table 52 and Table 53 .

Table 51: BMD analysis of Feed 1 data

model	N. parameters	Log-likelihood	AIC	BMD	BMDL	BMDU	converged	Accepted AIC
Null model	1	-28.346	58.692				yes	
Full model	3	-3.087	12.174				yes	
Exponential								
3 parameter	3	-3.207	12.414	3.944	0.634	13.419	yes	Yes
4 parameter	4	-3.14	14.28	22.626	1.533	29.768	yes	yes
Hill family								
3 parameters	3	-3.303	12.606	2.188			IC not converged	Yes (but no IC)
4 parameters	4	-3.14	14.28	22.140	1.535	29.595	yes	yes

Table 52: BMD analysis of Feed 2 data

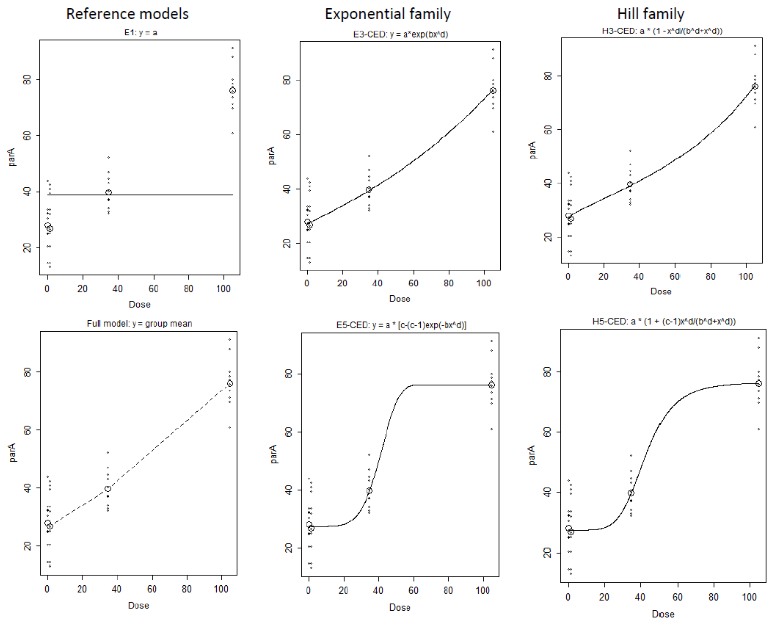
model	N. parameters	Log-likelihood	AIC	BMD	BMDL	BMDU	converged	Accepted AIC
Null model	1	-34.009	70.018					
Full model	3	-21.047	48.094					
Exponential								
3 parameter	3	-21.148	48.296	2.77	0	25.343	yes	yes
4 parameter	4		50.376	4.249	0.026	29.91	yes	no
Hill family								
3 parameters	3	-21.106	48.212	1.635			yes	yes
4 parameters	4	-21.188	50.376	4.252	0.027	29.93	yes	no

According to the BMD guidance illustrated in (European Food Safety Authority, 2017), which recommends to use all the models that are compatible with the data and choose the smallest BMDL and the largest BMDU among them, the confidence interval for Feed 1 is $(0.634, 29.768)$, for Feed 2 is $(0, 29.93)$ and for Feed 3 is $(24.338, 146, 1)$. The graphical representation of the models is shown in Figure 31, Figure 32 and Figure 33.

Table 53: BMD analysis of Feed 3 data

model	N. parameters	Log-likelihood	AIC	BMD	BMDL	BMDU	converged	Accepted AIC
Null model		143.22	-284.44					
Full model		159.867	-313.734					
Exponential								
3 parameter		158.75	-311.5	76.824	24.665	144.17	yes	no
4 parameter		158.073	-316.146	131.244	24.338	146.1	yes	yes
Hill family								
3 parameters		158.77	-311.54	76.410	24.804	142.99	yes	no
4 parameters		158.073	-308.146	141.51	24.402	145.85	yes	no

Figure 31: Dose–response relationship for all models (Feed 1 data)



MLT for BMD approach

The following MLT have been implemented to replicate classical analysis under the ML scenario:

- cppls
- ctree
- knn
- mars
- mlp
- mlpe
- mr
- naive
- pcr
- plsr
- randomForest

Figure 32: Dos–response relationship for all models (Feed 2 data)

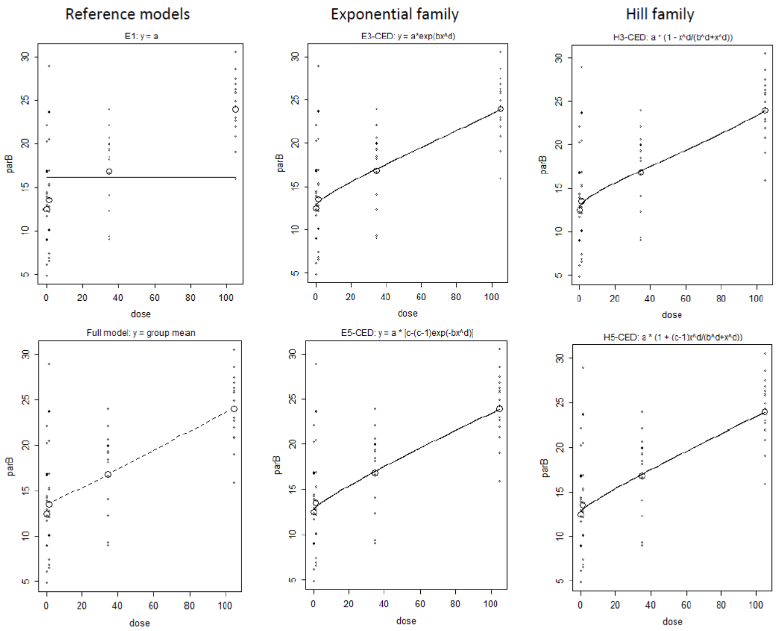
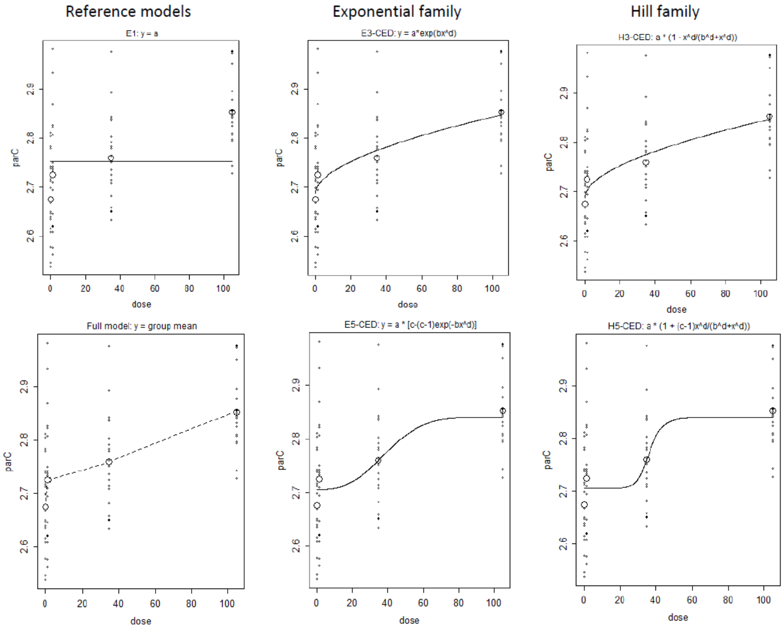


Figure 33: Dose–response relationship for all models (Feed 3 data)



- rpart
- svm

Purposely, all techniques have been left at the default parameterizations settings, to avoid interference from the point of a more or less sophisticated fine-tuning.

ML models are computational efficient in fitting the data. Results are reported in Table 54. Beside NB, not being able to produce sensitive results, in terms of goodness-of-fit, all ML models performed quite well, however with a strong tendency to overfitting the data. This ends up in BMD estimates which tend to be higher than those derived from the classical models. Noticeably, the functional form of the data fitting differs remarkably across the MLT adopted in the analysis Figure 34, 35 and 36. Despite the simplicity of fitting the data without imposing any distributional form, BMD analysis relies on the biological assumption of a monotonic dose-response relationship. As can be seen in Figure 34, 35 and 36, MLT based on splitting procedures tend to fit a step function, whereas MLT like mars, SVM and pcr are able to fit a monotonic dose-response curve. Constraining MLT to fit a monotonic dose-response relationship is not straightforward. MLTs based on splitting rules like recursive binary trees or random forest are not recommended. As it is usually done also in the classical analysis for the choice of the final model, the visual inspection of the fitted curve along with the performance measure can help in choosing the best model.

Table 54: ML analysis of the three feeds

MLT	Feed 1			Feed 2			Feed 3		
	R^2	BMD	BMDL	R^2	BMD	BMDL	R^2	BMD	BMDL
cppls	1.00	3.07	3.04	1.00	3.07	3.04	1.00	3.07	3.04
ctree	0.67	3.11	2.44	0.67	3.11	2.44	0.67	3.11	2.44
knn	0.79	12.38	11.51	0.79	12.38	11.51	0.79	12.38	11.51
mars	0.99	6.04	5.42	0.99	6.04	5.42	0.99	6.04	5.42
mlp	0.95	13.48	12.19	0.95	13.48	12.19	0.95	13.48	12.19
mlpe	0.93	14.42	13.15	0.93	14.42	13.15	0.93	14.42	13.15
mr	1.00	3.04	3.01	1.00	3.04	3.01	1.00	3.04	3.01
naive									
pcr	1.00	3.06	3.03	1.00	3.06	3.03	1.00	3.06	3.03
plsr	1.00	3.03	3.00	1.00	3.03	3.00	1.00	3.03	3.00
randomForest	0.79	11.73	10.99	0.79	11.73	10.99	0.79	11.73	10.99
rpart	0.78	16.75	16.18	0.78	16.75	16.18	0.78	16.75	16.18
svm	0.99	11.17	10.61	0.99	11.17	10.61	0.99	11.17	10.61

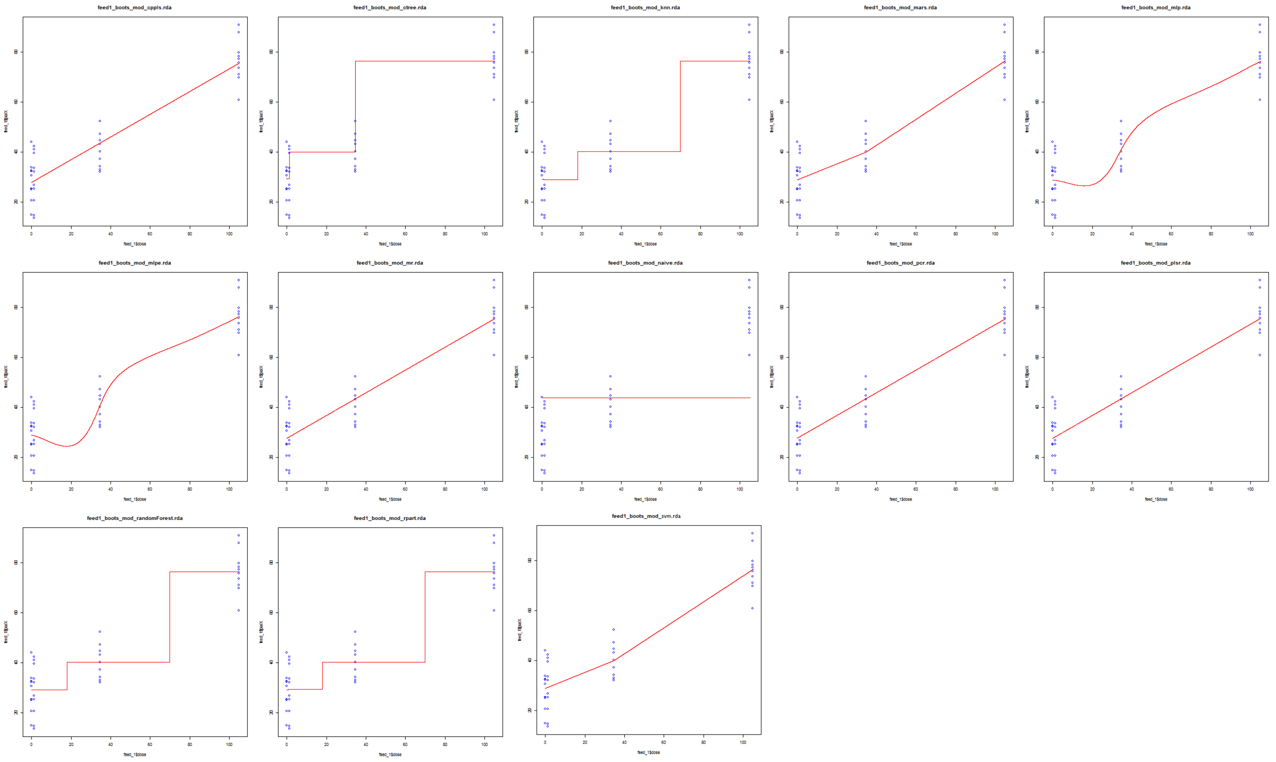


Figure 34: ML fitting of dose-response curves. Feed 1.

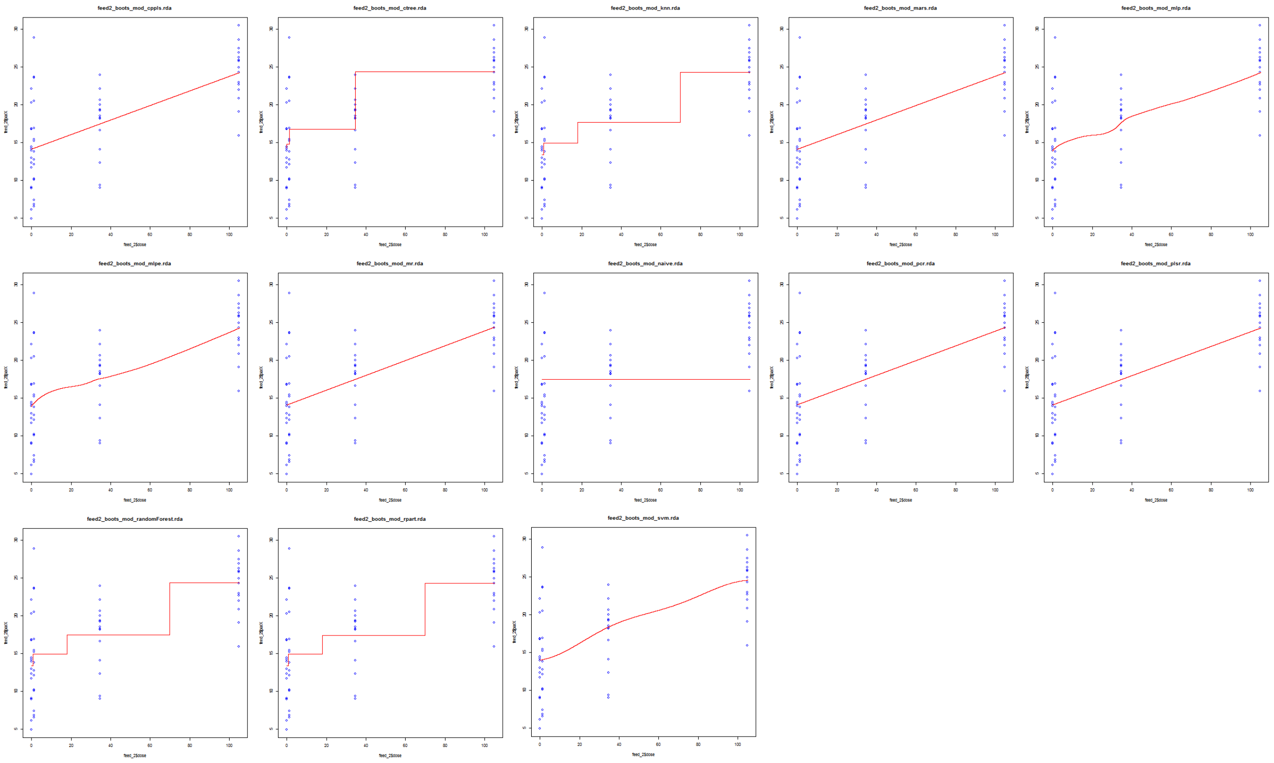


Figure 35: ML fitting of dose-response curves. Feed 2.

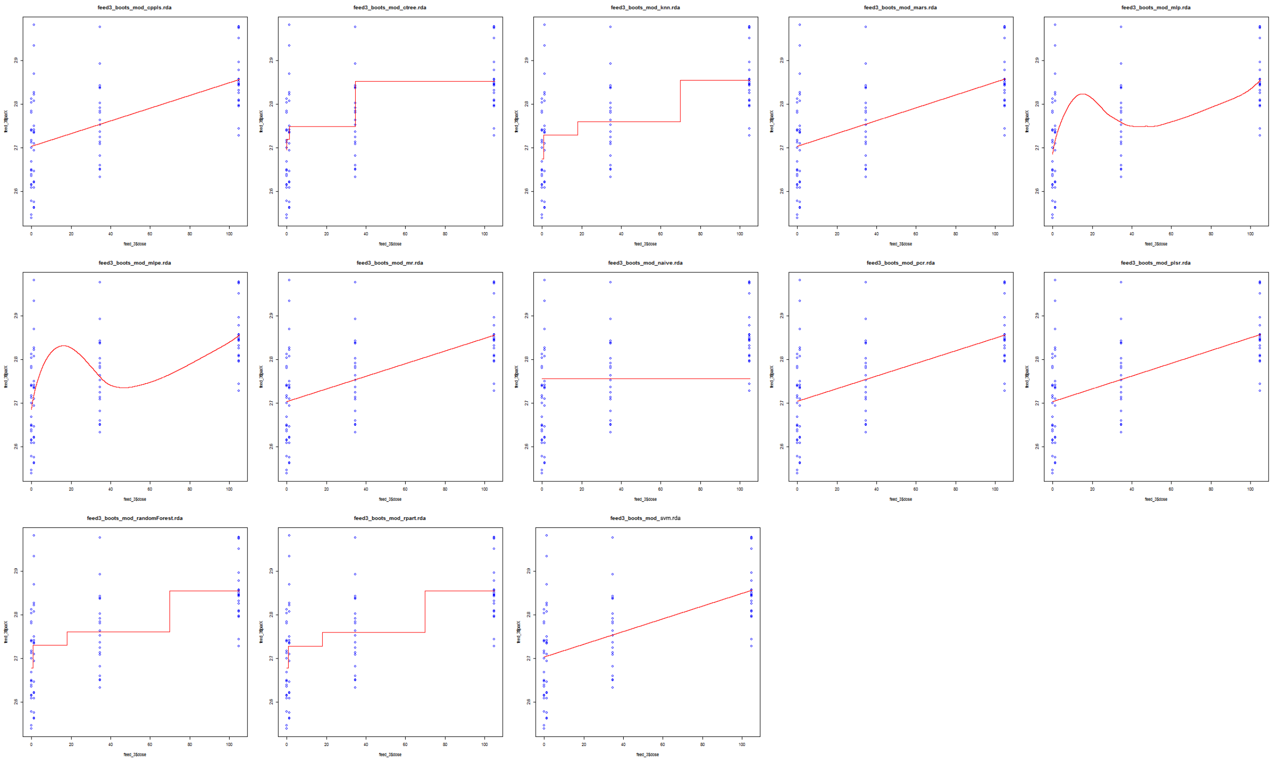


Figure 36: ML fitting of dose-response curves. Feed 3.

3.3 Veterinary infections, assessment and forecasting

The situation of scrapie has been actively surveyed in the Member States since the implementation of a compulsory programme for the monitoring of TSEs in sheep and goats in 2002, on the basis of a random sampling of healthy slaughtered animals on one hand and fallen stock on the other. Targeted population and sample sizes have evolved over time. Mandatory eradication measures have simultaneously been enforced by Regulation (EC) No 999/2001, in holdings where TSE cases were confirmed, combining culling, movement restrictions and reinforced surveillance measures. Recognising that some polymorphisms of the PRNP gene are associated with differences in the phenotypic expression of prion diseases in sheep (incubation period, physiopathology and clinical signs), several Member States have been implementing since the 1990s breeding programmes aimed at increasing the level of alleles associated with resistance (ARR) and decreasing the frequency of alleles associated with susceptibility (VRQ) in their sheep population. As of 2004, the EU made compulsory the introduction by the Member States of a breeding programme to be applied to the flocks of high genetic merit, until it became facultative again in 2007. This global strategy for monitoring and controlling TSEs in sheep and goats has now been in place for approximately 10 years. Among other goals, one of its underlying objectives is the eradication of Classical scrapie in EU population of sheep and goats. Today, the situation of Classical scrapie appears to be heterogeneous among the Member States, with no clear trend perceived by the Commission with regards to the evolution of its prevalence rate at the scale of the European Union. In order to assess the progress accomplished and evaluate the measures in place, the Commission needs a better understanding of the dynamics of the epidemiologic situation of Classical scrapie (CS) and Atypical scrapie (AS).

3.3.1 Terms of References

EFSA is requested to provide a scientific opinion on the following question:

- On the basis of the results of the TSE monitoring programme laid down in the TSE Regulation, what is the trend since 2002 of the situation of Classical scrapie and Atypical scrapie in sheep and in goats respectively, in the EU as a whole and in the 27 Member States individually?

3.3.2 Results

In the classical approach, data analysis was conducted separately by species (sheep vs. goats) and disease (CS vs. AS). In each individual subset, descriptive frequency tables were produced showing the breakdown of animals tested, and number of cases by country, year, surveillance stream (SHC and NSHC) and rapid test. The precision and validity of the crude prevalence rates obtained through the analysis of active surveillance data may have been affected by the targeted and sample-based design of both the SHC and NSHC surveys. Country-specific temporal trends are in general heterogeneous, precluding any meaningful interpretation of the overall temporal trend at the EU27-level. Therefore the analysis and interpretation of the temporal trends has been conducted only at MS level. The potential for a confounding effect of stream in the case of CS in both sheep and goats became evident after comparing the stream-specific prevalence and the different distribution of the number of tests carried out in each stream by country or by year. Non-significant differences in the prevalence of AS by stream were observed therefore in this case no need of adjustment on stream was considered. Negative binomial models were used to fit ”count of cases detected” and ”year” to estimate the country-specific and stream-adjusted annual prevalence ratios (PRs). Significance levels of the slope of the linear function for individual MS and years were used to determine statistically significant temporal trends (Figure 37).

Data are organized in ”long” format as an R representation of longitudinal data.

```
           country year tested positives route      type species    prevalence
1          AUSTRIA 2002   2017         0   SHC classical   sheep 0.000000e+00
2          AUSTRIA 2002   2017         0   SHC  atypical   sheep 0.000000e+00
...
35         AUSTRIA 2011     20         0   SHC classical   sheep 0.000000e+00
36         AUSTRIA 2011     20         0   SHC  atypical   sheep 0.000000e+00
37         AUSTRIA 2011   4943         0  NSHC classical   sheep 0.000000e+00
38         AUSTRIA 2011   4943         4  NSHC  atypical   sheep 8.092252e−04
39         AUSTRIA 2012     34         0   SHC classical   sheep 0.000000e+00
...
56         BELGIUM 2005     10         0   SHC  atypical   sheep 0.000000e+00
57         BELGIUM 2005   1451         1  NSHC classical   sheep 6.891799e−04
58         BELGIUM 2005   1451         1  NSHC  atypical   sheep 6.891799e−04
59         BELGIUM 2006   7292         0   SHC classical   sheep 0.000000e+00
```

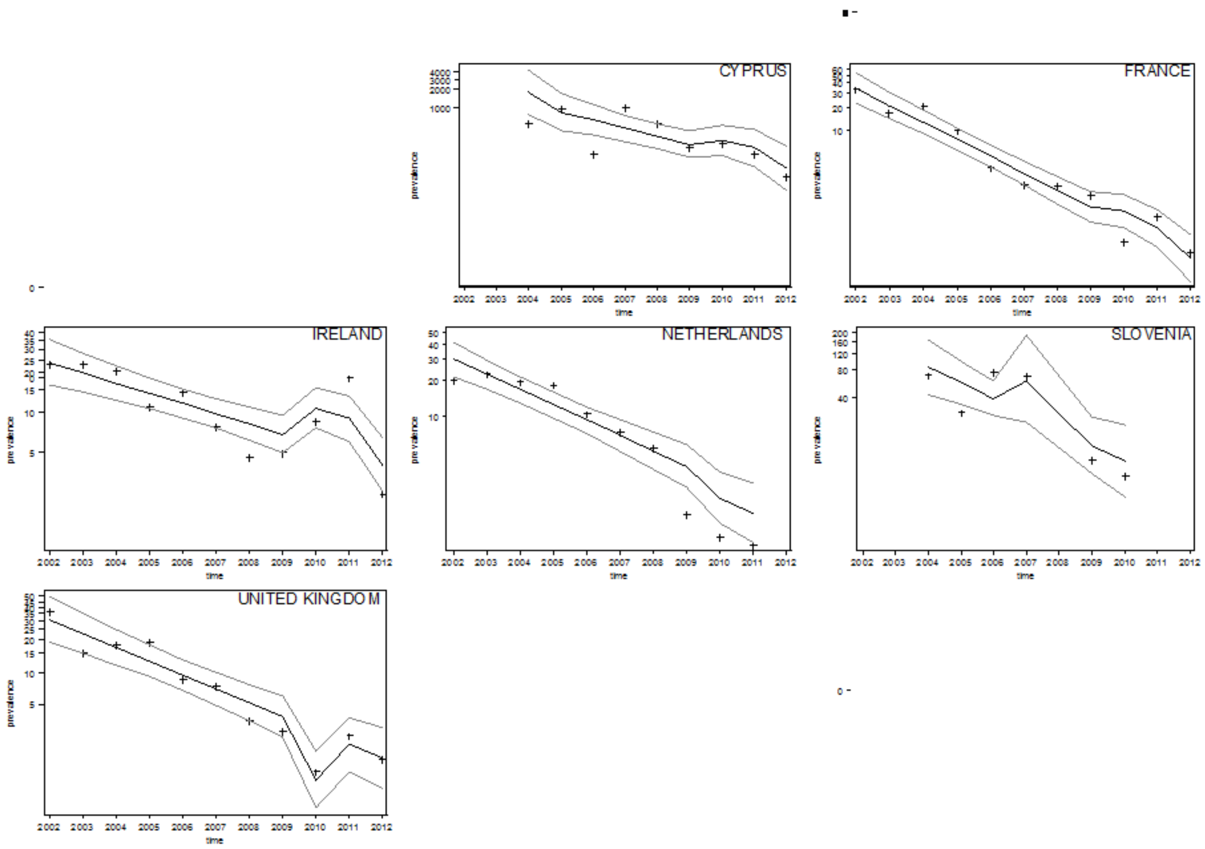


Figure 37: Temporal trend of CS in sheep in countries where a statistically significant decreasing trend was confirmed. Crosses (+) indicate the annual stream-adjusted prevalence (cases per 10,000 rapid tests) whereas the lines show respectively the linear trend (black line) with its 95% confidence limits (grey lines). The adjustment on stream was obtained by fitting a negative binomial model (internal reference).

```
60          BELGIUM  2006    7292          1    SHC   atypical    sheep  1.371366e−04
...
```

The idea behind the use of MLT is to conceptualize a dynamic usage of such information, where, like a dynamic Markov process, each year the information up to then is used to forecast (predict) future incidence rates.

This task has been approached by using a set of MLT

- cppls
- ctree
- knn
- mars
- mlp
- mlpe
- mr
- naive
- pcr

- plsr
- randomForest
- rpart
- svm

The idea is thus to use ML to predict future behaviors in terms of discrepancies between the actual and the forecasted trends. Such exercise is applied to 2006-2011 data, used to forecast what would have happened in 2012. Predicted differences can be estimated (assuming a perfect fit) as an indication of discontinuity in trends. This is performed using RF (Table 55). Ireland, France and Slovakia are those at higher risk.

Table 55: RF estimates

	N	Mean	Lower	Upper
AUSTRIA	1	-0.77		
BELGIUM	0			
BULGARIA	0			
CYPRUS	3	-0.048	-0.64	1.07
CZECH REPUBLIC	0			
DENMARK	1	-0.45		
ESTONIA	1	-1.00		
FINLAND	0			
FRANCE	4	7.17	0.40	20.17
GERMANY	4	1.29	-0.39	2.97
GREECE	6	0.12	-0.33	0.70
HUNGARY	2	0.45	-0.23	1.14
IRELAND	2	0.93	0.58	1.27
ITALY	7	0.06	-0.30	0.46
LATVIA	0			
LITHUANIA	0			
LUXEMBOURG	0			
MALTA	0			
NETHERLANDS	3	4.69	-0.53	14.48
POLAND	2	0.66	-0.25	1.56
PORTUGAL	4	1.11	0.24	1.98
ROMANIA	2	0.47	-0.79	1.73
SLOVAKIA	4	0.31	0.19	0.41
SLOVENIA	1	-0.81		
SPAIN	8	0.64	0.05	1.39
SWEDEN	1	0.47		
UNITED KINGDOM	6	1.75	-0.38	4.50
Overall	62	1.17	0.41	2.35

On the other hand, using Principal Component Regression (PCR) gives a completely different result, which is undoubtedly an overestimation of the discrepancies among countries. (Table 55).

The results obtained using statistical techniques in Figure 37 showed an inherent difficulty of capturing the temporal trend as it can see from the point data outside the estimated confidence interval. On the other hand, when using MLT, variability in prediction is very high across the algorithms. This result can be partially explained by the time-dependent nature of the data, which many MLT are not able to handle in a suitable way.

As recommended in chapter 4, a complete analysis based on MLT should include the implemention of more than one MLT to check their robustness and consequently choose the algorithm that is most apt to the problem.

Table 56: RF estimates

N	Mean	Lower	Upper	
AUSTRIA	1	-9.66e+14		
BELGIUM	0			
BULGARIA	0			
CYPRUS	3	-2.67e+13	-5.23e+13	-1.03e+13
CZECH REPUBLIC	0			
DENMARK	1	-9.40e+14		
ESTONIA	1	-7.81e+12		
FINLAND	0			
FRANCE	4	-5.40e+15	-8.44e+15	-2.37e+15
GERMANY	4	-3.46e+15	-6.44e+15	-1.25e+15
GREECE	6	-1.70e+15	-3.43e+15	-3.71e+14
HUNGARY	2	-1.37e+15	-2.11e+15	-6.24e+14
IRELAND	2	-4.21e+15	-8.00e+15	-4.27e+14
ITALY	7	-1.98e+15	-3.08e+15	-9.07e+14
LATVIA	0			
LITHUANIA	0			
LUXEMBOURG	0			
MALTA	0			
NETHERLANDS	3	-4.96e+15	-9.13e+15	-1.83e+15
POLAND	2	-4.51e+15	-7.97e+15	-1.04e+15
PORTUGAL	4	-4.14e+15	-7.42e+15	-8.67e+14
ROMANIA	2	-8.23e+14	-1.60e+15	-4.84e+13
SLOVAKIA	4	-4.90e+14	-6.75e+14	-3.89e+14
SLOVENIA	1	-4.02e+14		
SPAIN	8	-2.70e+15	-3.98e+15	-1.67e+15
SWEDEN	1	-1.81e+15		
UNITED KINGDOM	6	-1.20e+15	-2.08e+15	-3.33e+14
Overall	62	-2.38e+15	-3.09e+15	-1.79e+15

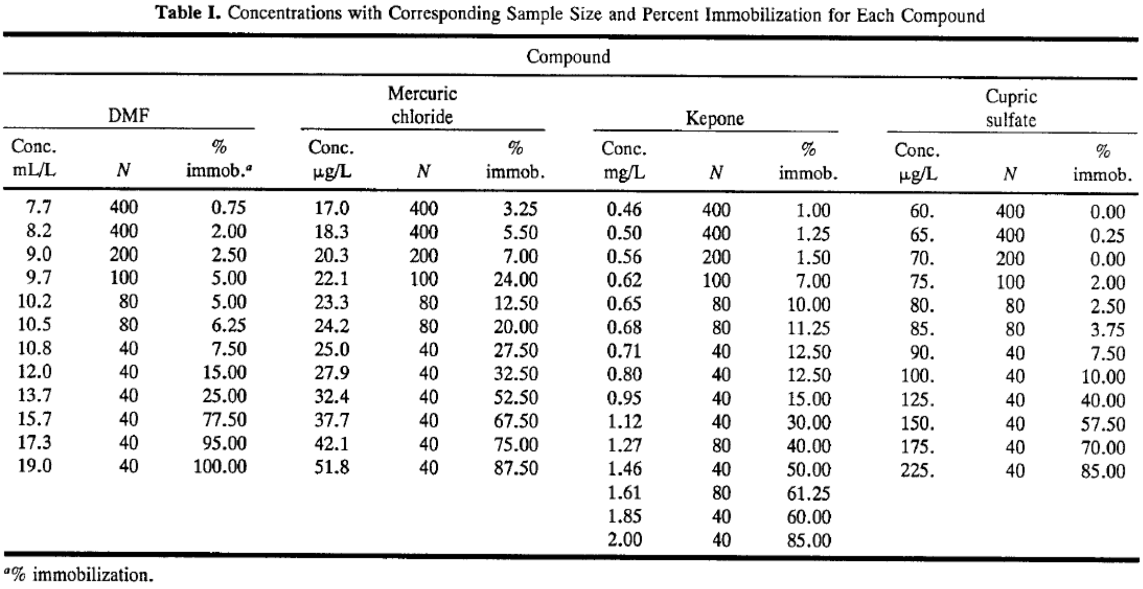


Table I. Concentrations with Corresponding Sample Size and Percent Immobilization for Each Compound

Compound											
DMF			Mercuric chloride			Kepone			Cupric sulfate		
Conc. mL/L	N	% immob.[a]	Conc. µg/L	N	% immob.	Conc. mg/L	N	% immob.	Conc. µg/L	N	% immob.
7.7	400	0.75	17.0	400	3.25	0.46	400	1.00	60.	400	0.00
8.2	400	2.00	18.3	400	5.50	0.50	400	1.25	65.	400	0.25
9.0	200	2.50	20.3	200	7.00	0.56	200	1.50	70.	200	0.00
9.7	100	5.00	22.1	100	24.00	0.62	100	7.00	75.	100	2.00
10.2	80	5.00	23.3	80	12.50	0.65	80	10.00	80.	80	2.50
10.5	80	6.25	24.2	80	20.00	0.68	80	11.25	85.	80	3.75
10.8	40	7.50	25.0	40	27.50	0.71	40	12.50	90.	40	7.50
12.0	40	15.00	27.9	40	32.50	0.80	40	12.50	100.	40	10.00
13.7	40	25.00	32.4	40	52.50	0.95	40	15.00	125.	40	40.00
15.7	40	77.50	37.7	40	67.50	1.12	40	30.00	150.	40	57.50
17.3	40	95.00	42.1	40	75.00	1.27	80	40.00	175.	40	70.00
19.0	40	100.00	51.8	40	87.50	1.46	40	50.00	225.	40	85.00
						1.61	80	61.25			
						1.85	40	60.00			
						2.00	40	85.00			

[a] % immobilization.

Figure 38: Experimental setting of toxicity study using DAMA (Sebaugh, 1991)

3.4 DAMA study illustration

DAMA is a commonly used test animal in aquatic toxicology. DAMA is specified to be used in the Organisation for Economic Co-operation and Development (OECD) Guidelines for the Testing of Chemicals. The reference for this case study is the test No. 202: a 48 hour acute toxicity study, where young DAMAs are exposed to different concentrations of toxically substances.

Current setting of exposure standards are largely based on the two key concept of NOAEL and Lowest Observed Adverse Effect Level (LOAEL) for some exposure effects at which no outcomes are observed. In conjunction, traditional analysis assumes a probit model for estimating the dose-response. However, traditional models for binary data, such as probit model, fail to predict in an accurate manner the response in the tails. This represents an ideal setting where MLT offer flexible tools and yield valuable insights about the shape of the dose-response. We aim to outline rigidities in the traditionally used approach and suggest different reasons for which MLT may be a valid alternative to the more conservative framework.

3.4.1 Methods

Sebaugh (Sebaugh et al., 1991) performed four set of experiments in order to compare the various model in the logistic family using different toxicants: dimethylformamide, mercuric ion, cupric ion, kepone, each at 12 levels of concentration. The recorded response was immobilization of *Daphnia magna* to a specific dose-level of the given toxicant. The aim is to identify concentration levels that correspond to the lowest *"effective concentrations"*: EC01, EC05 and EC10, that is the levels of concentration that correspond to 1, 5 and 10% of immobilized DAMA.

3.4.2 Methods comparison

Several alternative to those models have been compared in literature (Sebaugh et al., 1991). Some recent generalization in the logistic family have been proposed by Stukel (Stukel, 1983), Prentice (Prentice, 1976), Aranda-Ordaz (Aranda Ordaz, 1981).

One of the widely used ones is the generalization of Stukel (Stukel, 1983; Stukel, n.d.), which consists in adopting a parameterized version of the non-canonical link, of the form log$(\mu/(1-\mu)) = \mathcal{H}_\alpha(\eta)$ where $\mathcal{H}_\alpha(\eta)$

Generalized Logit	Statistic	Compound			
			Mercuric		Cupric
		DMF	chloride	Kepone	sulfate
4-parameter	χ^2	3.29	11.13	11.49	2.81
	df	8	8	11	8
	p	0.92	0.19	0.40	0.95
$\alpha_1 = -\alpha_2$	χ^2	3.29	11.27	15.30	2.84
	df	9	9	12	9
	p	0.95	0.26	0.23	0.97
α_1 model	χ^2	7.19	12.75	18.69	6.04
$(\alpha_2 = 0)$	*df*	9	9	12	9
	p	0.62	0.17	0.10	0.74
α_2 model	χ^2	4.21	11.11	14.62	2.92
$(\alpha_1 = 0)$	*df*	9	9	12	9
	p	0.90	0.27	0.26	0.97
$\alpha_1 = \alpha_2$	χ^2	8.42	11.53	13.79	4.53
	df	9	9	12	9
	p	0.49	0.24	0.31	0.87
$\alpha_1 = \alpha_2 = .165$	χ^2	29.69	12.09	14.43	8.03
Probit	*df*	10	10	13	10
	p	0.001**	0.28	0.35	0.63
$\alpha_1 = \alpha_2 = 0$	χ^2	18.95	14.43	18.69	14.44
Logit, logistic	*df*	10	10	13	10
	p	0.04*	0.15	0.13	0.15

Table 57: χ^2 Goodness-of-fit statistics from testing the null hypothesis of adequate model fit obtained from testing each of the models using all concentrations for the four compounds. Legend: $^*, p < .05;^{**}, p < .01$, Source: Sebaugh (1991)

takes the form for $\eta \geq 0$

$$\mathcal{H}_\alpha(\eta) = \begin{cases} \alpha_1^{-1}(\exp(\alpha_1|\eta|) - 1) & \text{if } \alpha_1 > 0 \\ \eta & \text{if } \alpha_1 = 0 \\ \alpha_1^{-1}(\log(1 - \alpha_1|\eta|)) & \text{if } \alpha_1 < 0 \end{cases} \tag{5}$$

and for $\eta \leq 0$ the same but with α_2 in the place of α_1. This generalization of logistic family models allows for some flexibility in the tails, introducing asymmetry, modeled by the choice of α_1 and α_2. The full model is assumed to provide the best fit, although the drawback might be that the model is overspecified. Symmetric versions of this models are the ones that assume $\alpha_1 = \alpha_2$, with unspecified α_1, $\alpha_1 = \alpha_2 = 0.165$ (probit model) and $\alpha_1 = \alpha_2 = 0$ (logit model). The asymmetric counterparts of the models are the ones with $\alpha_1 = -\alpha_2$ and the $\alpha_1 = 0$.

Statistics χ^2 were used as goodness-of-fit statistics from testing the null hypothesis of adequate model fit obtained from testing each of the models using all concentrations for the four compounds. Table 57 shows the results, and the best models are shown in Figure 39.

The best fit for each toxicant can be found in figure below . As a final result of the cited study, probit model seems to be the best model for kepone and mercuric chloride but it tended to overestimate the concentrations corresponding to 5, 10 and 50% for DMF and cupric sulfate. The limit for estimating risks below the 5% threshold cannot be overcome using a probit model. In this context although the probit models has advantages in terms of interpretability, a more complex new model that accounts for asymmetry, can be of great utility, and this study in a nutshell highlights the intuition for such need.

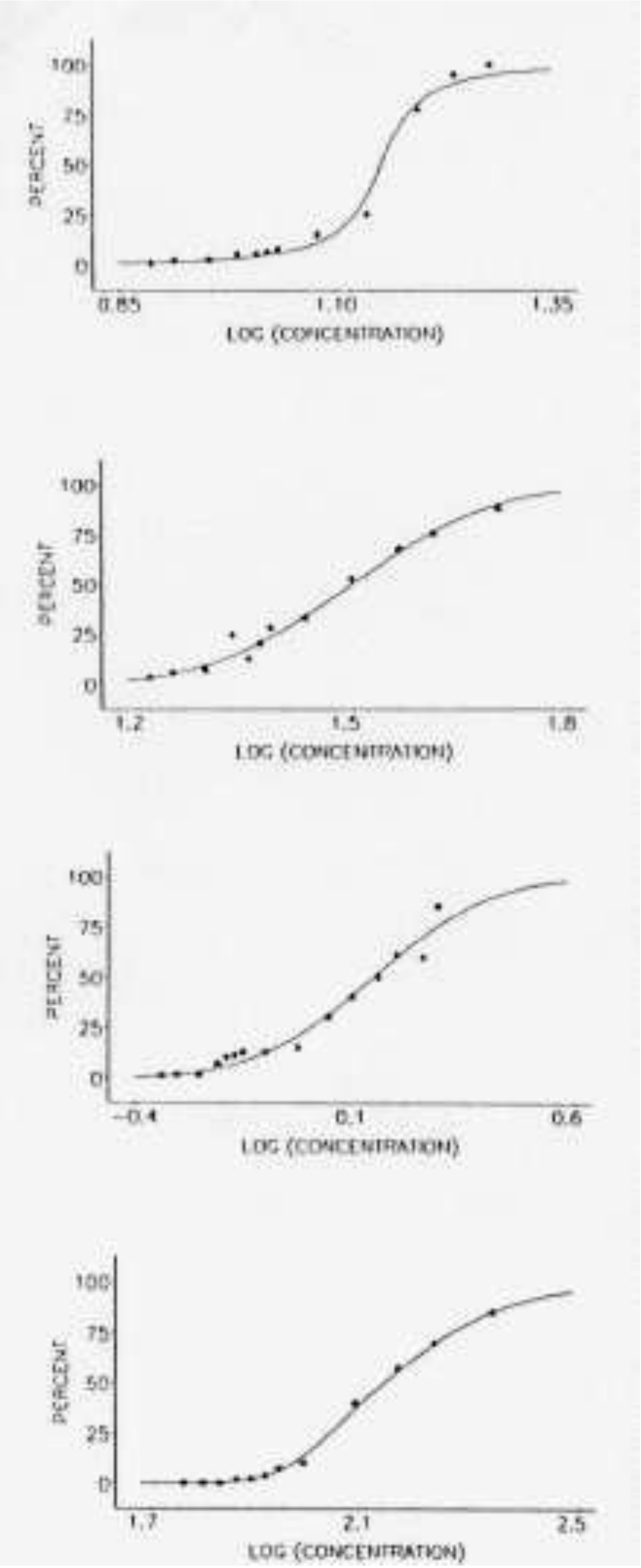


Figure 39: Best fitting model and observed immobilization for: (a) DMF, $\alpha_1 = -\alpha_2$ generalized linear model; (b) mercuric chloride, $\alpha_1 = \alpha_2 = .165$, probit model; (c) kepone, $\alpha_1 = \alpha_2 = .165$, probit model; (d) cupric sulphate, α_2 generalized logistic model.

Note that in the above example, the entire study relies on strong parametric assumptions. Although some flexibility has been introduced, it remains a model with a rather rigid functional form. Also, in terms of model validation and goodness-of-fit evaluation, it remains quite basic. The result is that the final conclusions about the toxicity levels might not be predicted with accuracy. MLT have many advantages including no need of parametric assumptions, as well as high power and flexibility. Generalized logistic function would be considered as one of the special cases considered and compared with other models.

In general, ML covers a wide range of classification and prediction algorithms and, unlike approaches that assume a fixed statistical model, for example, a Generalized Linear Model (GLM), ML aims to extract the relationship between the endpoint and covariates through a learning algorithm .

In this case, ML algorithms that should be considered are:

- naive: most common class
- CTREE
- decision tree
- KNN
- MLP ensemble
- SVM
- RF
- LDA
- logistic regression

Table 58: Ranking table of techniques by performance measures

MLT	Accuracy	Precision1	Precision2	AUC	F1	Lift
naive	4	3	2	4	2	3
CTREE	5	5	6	5	6	6
rpart	9	9	8	7	8	7
KNN	8	7	7	8	9	8
MLP ensemble	1	2	1	1	2	1
SVM	7	8	9	9	7	9
RF	6	6	5	6	5	5
LDA	3	4	4	3	4	4
logistic regression	2	1	3	1	3	2

3.4.3 Conclusion

The purpose of a predictive toxicology model is to predict the toxicity of untested compounds on the basis of existing experimental data of other compounds (training data). Putting it in a formal way machine learning involves seeking a function for predicting new (unseen) cases. A learning algorithm identifies this function by searching in a set of suitable functions (the hypothesis space) in order to identify a function that minimizes the empirical error (i.e. the difference between predictions and real values).

The purpose of this cs is not to identify the best possible techniques in order to predict the ”true” dose-response. It is to highlight and illustrate two main advantages of ML computational tools over traditional analysis and estimation. The first is the fact that we can easily accomodate more flexible non parametric models, that do not require too rigid functional form assumptions, but which allow the data to ”speak freely for themselves”. Sometimes adding constraints is essential, given the nature of the data. Constraints can be added during the optimization procedures of the algorithms, for example considering random forest, constraints can be added during the optimization of the purity index of the individual trees. However, even in this case, adding constraints beyond tune paramters is not straighforward.

Another advantages is that it goes beyond comparison of the same family of models but it extends them to different modeling families.

Finally, automated algorithms can easily illustrate and rank the techniques according not only standard traditional goodness of fit, but according to different measures of accuracy and precision, described in detail in the Case Study 1 (section 3.1).

3.5 The food pyramid and portions

3.5.1 Introduction

The Food Pyramid, developed by the US Department of Agriculture (USDA), is a tool to lead the subject to make healthy food choices. The food pyramid suggests portions size according to nutrient intake controlling the amount of calories, fat, saturated fat, cholesterol, sugar or sodium in the diet.

MyPiramid dataset has been considered; the data are provided by the US Department of Agriculture (USDA) to give information on the total calories; calories from solid fats, added sugars, and alcohol (extras) according with food group and subgroup amounts of over 1,000 commonly eaten foods with corresponding commonly used portion amounts.

Groups in the MyPiramid dataset have been constructed according to the approximate nutritional properties grouping foods into seven broad categories as described below:

```
group
"1"    "Grains"
"1"    "Whole\_Grains"
"2"    "Vegetables"
"2"    "Orange\_Vegetables"
"2"    "Darkgreen_Vegetables"
"2"    "Starchy\_vegetables"
"2"    "Other\_Vegetables"
"3"    "Fruits"
"4"    "Milk"
"4"    "Meats"
"5"    "Soy"
"5"    "Drybeans\_Peas"
"5"    "Oils"
"5"    "Solid_Fats"
"6"    "Added\_Sugars"
"6"    "Alcohol"
"7"    "Calories"
"7"    "Saturated\_Fats"
```

An application of Least Absolute Shrinkage and Selection Operator (LASSO) to the supervised learning of the contribution that nutrients provide to the effective portion size is shown. The portion size, representing one equivalent, is determined using the FNDDS metric provided by the USDA's technical files for analyzing food and nutrient intakes; it has have developed over several decades of food surveys.

Table 59: Food pyramid Composition Data

	(0,1]	(1,4]	(4,25]	P-value
	$N = 1855$	$N = 98$	$N = 61$	
Grains	0.00 0.00 0.72	0.00 0.00 0.64	0.00 0.00 0.00	< 0.001
Whole_Grains	0 0 0	0 0 0	0 0 0	0.18
Vegetables	0.00 0.00 0.04	0.00 0.00 0.00	0.00 0.00 0.30	< 0.001
Orange_Vegetables	0 0 0	0 0 0	0 0 0	0.034
Drkgreen_Vegetables	0 0 0	0 0 0	0 0 0	0.2
Starchy_vegetables	0 0 0	0 0 0	0 0 0	< 0.001
Other_Vegetables	0 0 0	0 0 0	0 0 0	0.002
Fruits	0 0 0	0 0 0	0 0 0	0.31
Milk	0 0 0	0 0 0	0 0 0	0.42
Meats	0.00 0.00 0.61	0.00 2.62 4.00	0.00 0.00 0.00	< 0.001
Soy	0 0 0	0 0 0	0 0 0	0.57
Drybeans_Peas	0 0 0	0 0 0	0 0 0	0.2

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
Test used:
Kruskal-Wallis test

Table 59: *(continued)*

	(0,1] $N = 1855$	(1,4] $N = 98$	(4,25] $N = 61$	P-value
Oils	0.00 0.00 0.06	0.00 0.00 0.00	0.00 0.00 0.00	< 0.001
Solid_Fats	0.0 5.1 41.7	0.0 23.1 51.5	0.0 0.0 10.6	< 0.001
Calories	53 113 208	128 194 248	34 96 132	< 0.001
Added_Sugars	0.0 0.0 12.1	0.0 0.0 1.5	0.0 0.0 0.0	< 0.001
Alcohol	0 0 0	0 0 0	0 0 0	< 0.001
Saturated_Fats	0.08 0.85 2.84	0.62 1.96 3.79	0.03 0.29 1.35	< 0.001

$a\ b\ c$ represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
Test used:
Kruskal-Wallis test

A descriptive statistic table is reported in Table 3.5.1 showing that there is a significant difference in portion size group for Grains component, vegetables meat and the overall fat and sugar component of dietary intake.

3.5.2 Methods

A Lasso regression method has been performed. LASSO allows for carrying out both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the regression model. Moreover, a grouped penalized regression is considered to analyze the data in order to allow for predefined blocks of covariates to be selected to enter the model.

The method is based on the minimization of the sum of squared errors; it imposes a constraint on the sum of the absolute values of the coefficients by defining an upper bound.

Figure 40 shows the inconsistency of all nutrients w.r.t. portion size definition. A ”flat” line is indicating a stable coefficient.

Same result is obtained by the ”lasso” fitting in ”lars” library (Figure 41). Some method are compared in Figure 42, showing the Mallow's C_p and degrees of freedom for lasso, lar stepwise and stagewise penalization. As pointed out by several authors, beside some slight benefits from the computational perspective, all methods agree (Figure 42) in selecting models (model components for which the smallest Mallow's C_p is observed).

Grouped regularized regression The package `grpreg` fits several options for the penalty functions (Breheny and Huang, 2015), like lasso, MCP (like the group lasso, but with an MCP penalty on the norm of each group), SCAD (like the group lasso, but with a SCAD penalty on the norm of each group), cMCP (a hierarchical penalty which places an outer MCP penalty on a sum of inner MCP penalties for each group) (Breheny and Huang, 2009), GEL (Group exponential lasso) and gBridge (a penalty which places a bridge penalty on the L1-norm of each group) (Huang et al., 2009). There are two general classes of methods involving grouped penalties. Bi-level means performs the selections of variables at the group level and for singular covariates, selecting important groups and important members of those groups. Reversely, the group selection selects important groups, and not members within the group.

3.5.3 Methods comparison

Figure 43 shows the λ plotted agains the estimated $\hat{\beta}$, for ungroupedwithout ridge penalty. Both MCP and SCAD are provided consistent results, but with different variables selection according to grouping In a two-penalties setting, parameter α controls the proportional weight of the regularization parameters of these two penalties. The group penalties' regularization parameter is $\lambda \times alpha$, while the regularization parameter of the ridge penalty is $\lambda \times (1 - \alpha)$.

Bi-level penalties applied to the MyPiramid dataset produced the selection in Figure 44.

An interesting feature of the `grpreg` library is to allow for *seemingly unrelated regressions* and *multitask learning*. They can be carried out by passing a matrix to y. In this case, X will be used in separate regressions for each column of y, with the coefficients grouped across the responses. In other words, each column of x will form a group with m members, where m is the number of columns of 'y'. For multiple Gaussian responses, it is recommended to standardize the columns of y prior to fitting, in order to apply the penalization equally across columns.

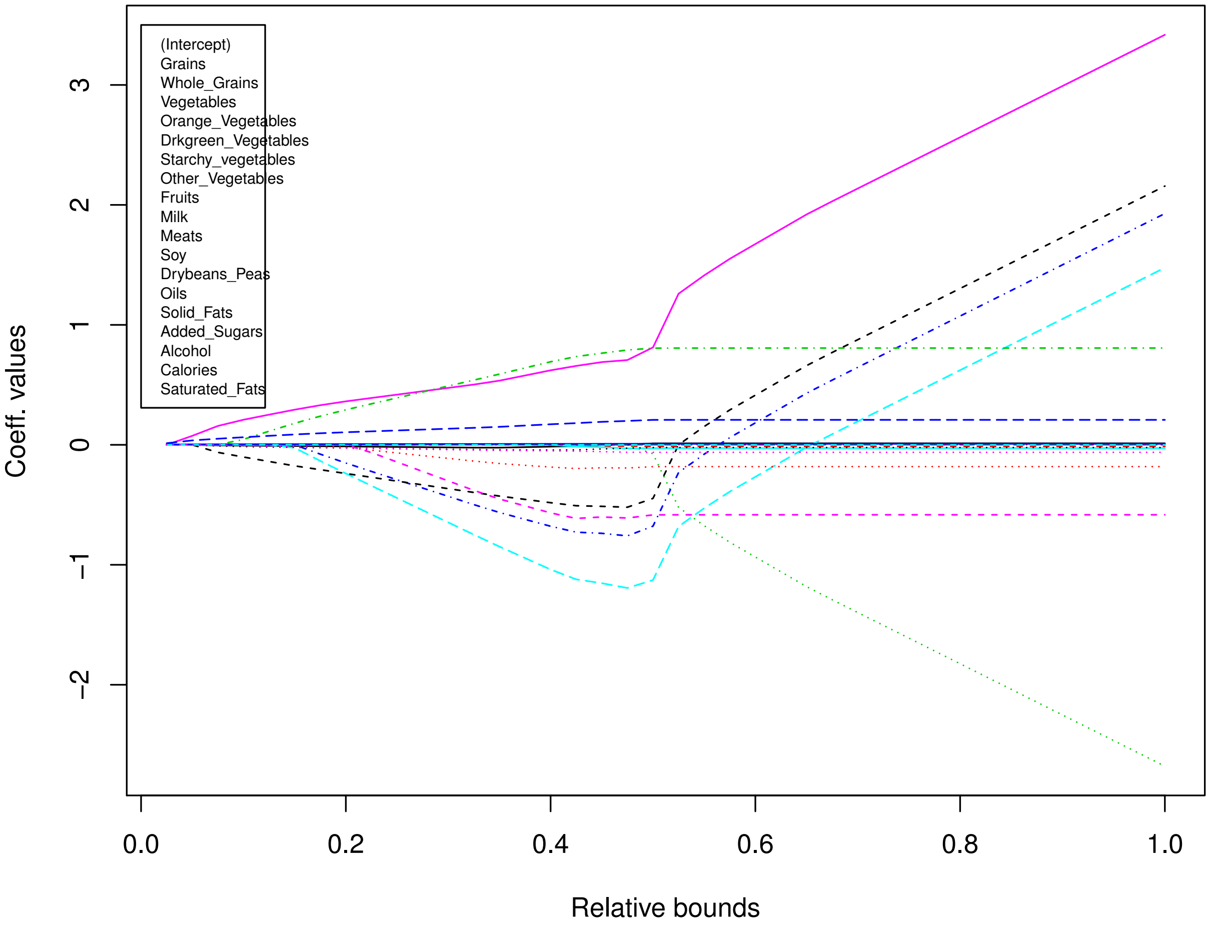


Figure 40: L1-constrained lasso coefficients for varying boundaries

In examining the association in menu composition between vegetables and diary and meat products, the application of the seemingly unrelated regression clearly shows the relevance of diary (milk in particular) on the vegetable composition of the diet, including, not surprisingly, cereals and grains.

3.5.4 Conclusion

Observing the results of the analysis it is evident that the algorithms providing penalized regression on grouped covariates are stable. In several cases there is a general coherence between the results provided by single penalization methods and the grouped penalization method separately in bi-level form and grouped form. The standard algorithm penalizing linear model assumes that the model matrices in each group are not correlated. This penalization in group lasso takes into account the correlation within each group. Considering the Pyramid Dataset is useful to define a group penalized regression taking into account that some nutrients may be represented in broad categories, a single lasso regression algorithm doesn't consider the correlation structure which define the relation in each nutrients group in the nutritional Pyramid.

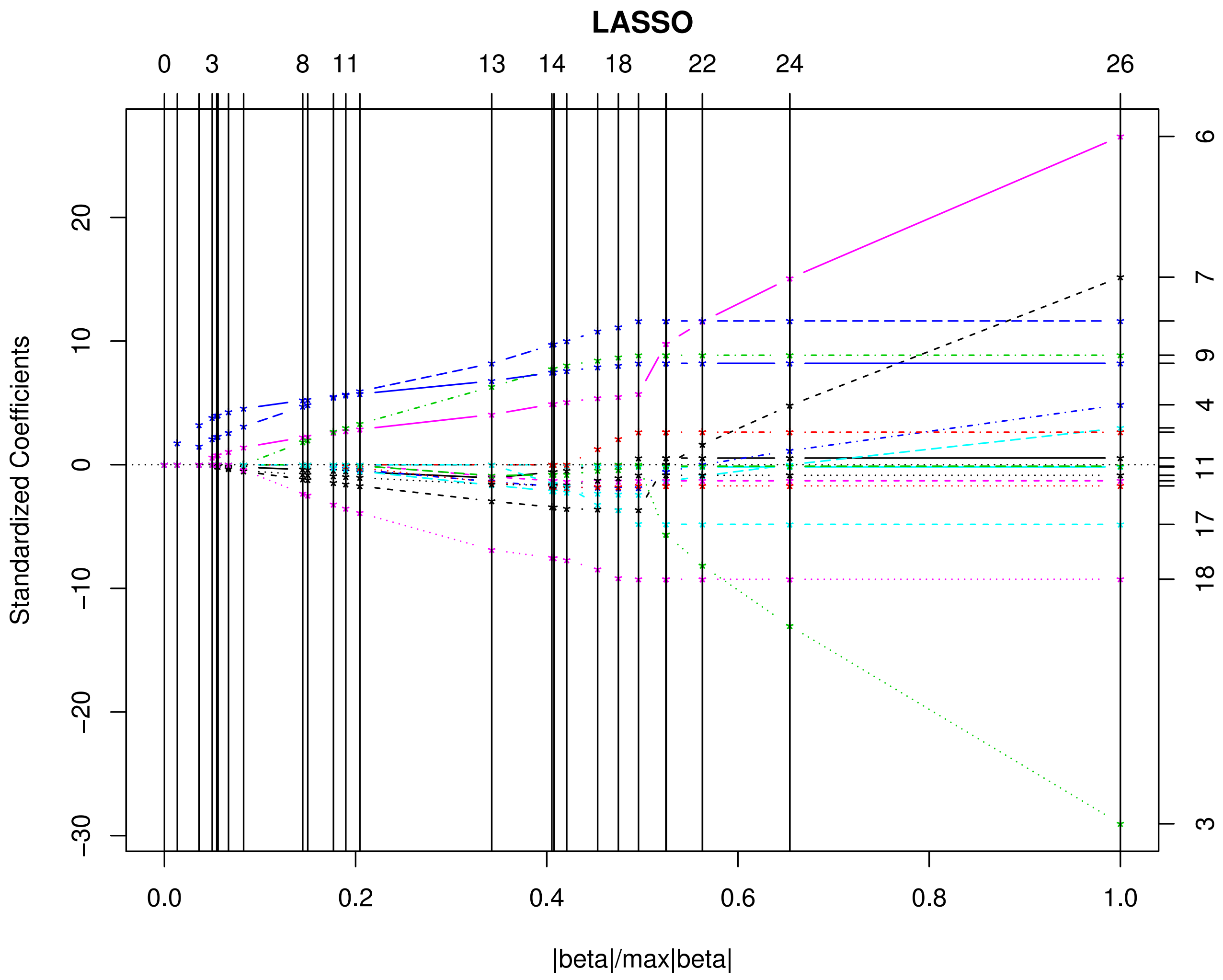


Figure 41: Coefficients plotted against the L1 norm of the coefficient vector, as a fraction of the maximal L1 norm

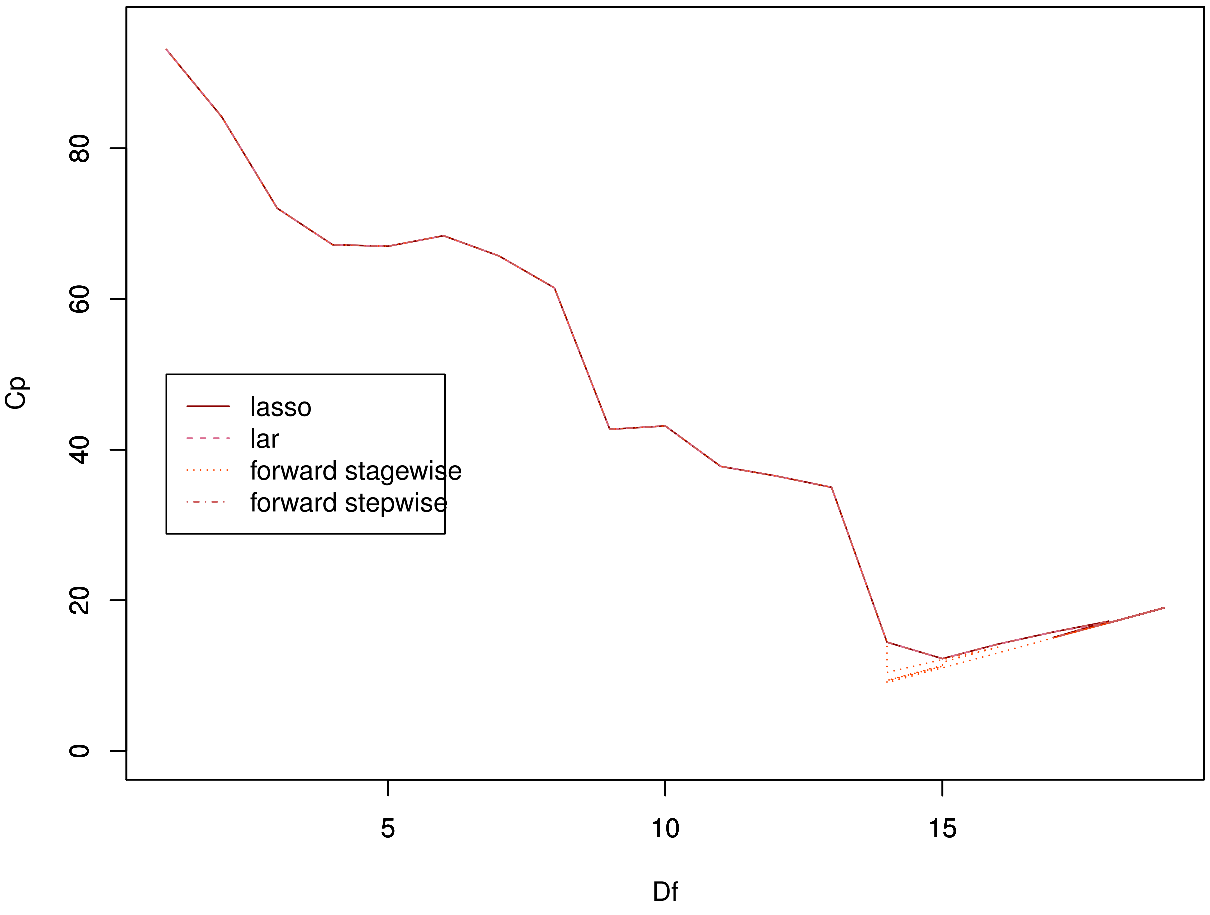


Figure 42: Mallow's C_p plotted against Df. Interpretation of Df for stagewise and stepwise must be taken cautiously.

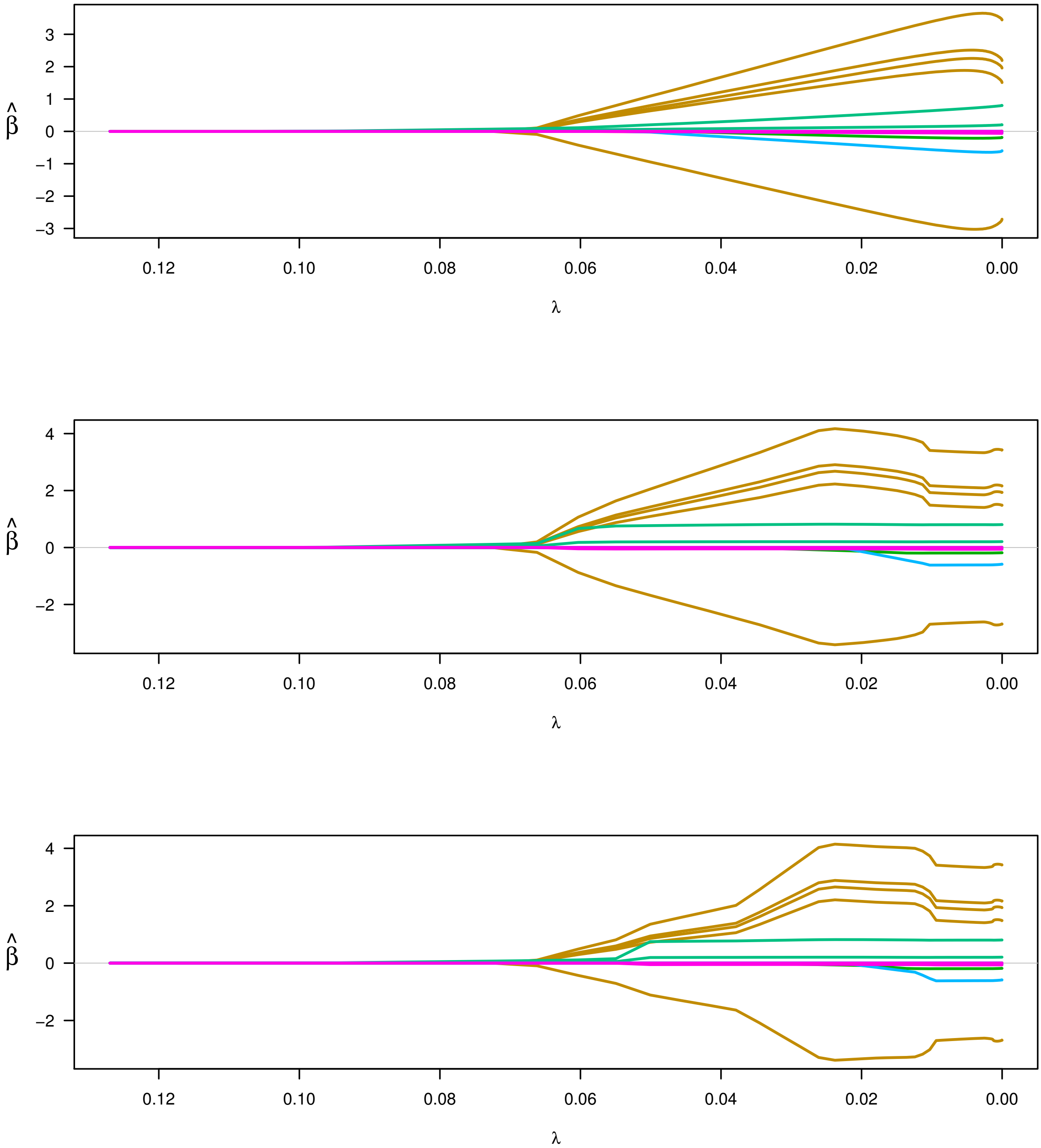
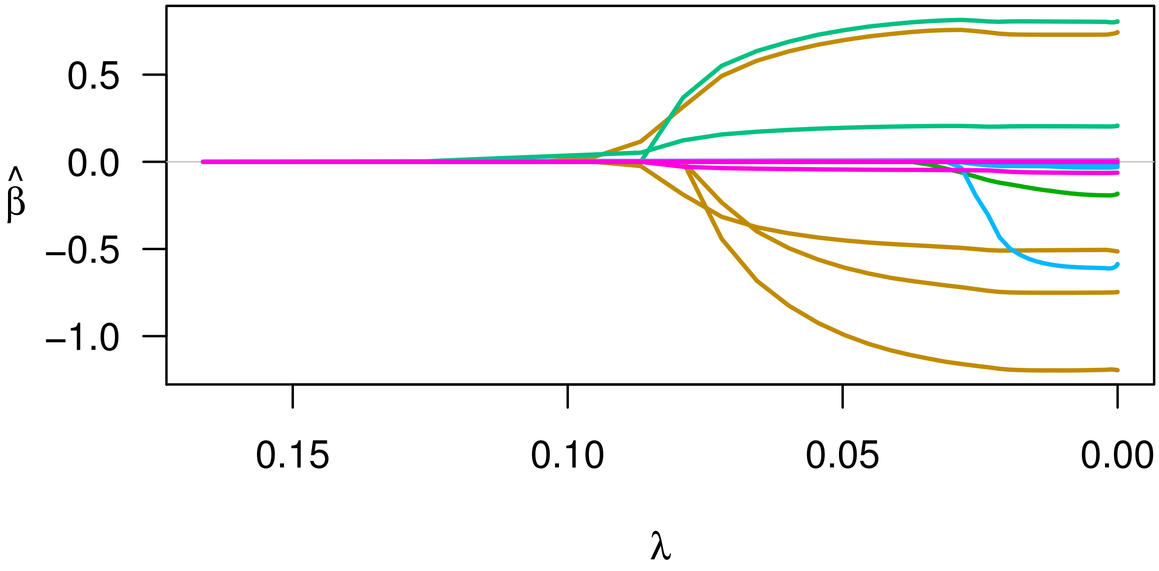


Figure 43: Grouped regularized techniques, from upper panel down, lasso, MCO and SCAD.

	Nutrient	GEL	MCP
Grains	Grains	0.000	0.000
WholeGrains	WholeGrains	0.000	0.000
Vegetables	Vegetables	0.000	0.000
OrangeVegetables	OrangeVegetables	−0.233	−0.045
DrkgreenVegetables	DrkgreenVegetables	−0.442	−0.167
Starchyvegetables	Starchyvegetables	0.492	0.427
OtherVegetables	OtherVegetables	−0.316	−0.249
Fruits	Fruits	0.000	0.000
Milk	Milk	0.551	0.470
Meats	Meats	0.157	0.144
Soy	Soy	0.000	0.000
DrybeansPeas	DrybeansPeas	0.000	0.000
Oils	Oils	0.000	−0.008
SolidFats	SolidFats	0.000	0.000
AddedSugars	AddedSugars	0.000	0.000
Alcohol	Alcohol	0.006	0.007
Calories	Calories	0.000	0.000
SaturatedFats	SaturatedFats	−0.036	−0.037

Table 60: Bi-level regularized regression with GEL and composite MCP criteria. Data are estimated coefficients. Analysis based on food pyramid Composition Data



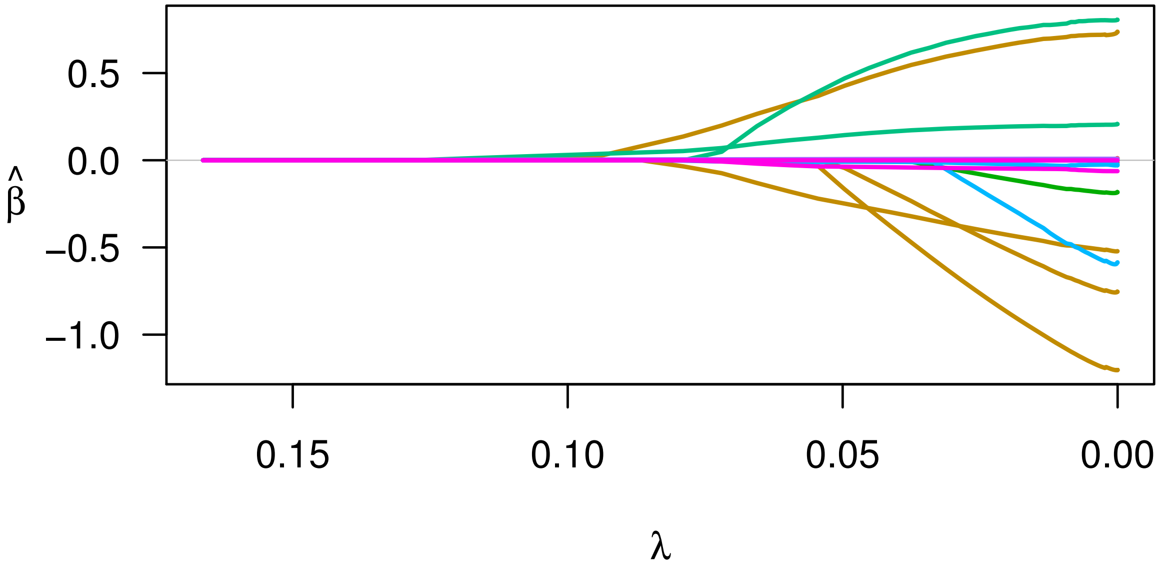


Figure 44: Bi-level regularized techniques, from upper panel down, GEL and composite MCP.

3.6 Assessment of analytical error for model stability in multilevel design in clinical research

3.6.1 Introduction

Nephron function can be assessed by measurements which are influenced by both interanimal (between animals) and intra-animal (within animals) fluctuations.

The accuracy of estimates depends on the number of animals, n, and the number of replications, k, in each animal. Usually, each parameter is estimated by calculating the mean value for each animal and then, by taking the overall mean of n animal values. Quintuplicate determinations of the parameters measured in studies of glomerular dynamics revealed that the intra-animal variance is larger than the corresponding interanimal variance (Glantz and Slinker, 1990). Indeed, substrains of rats used in previous studies exhibited a degree of internephron functional heterogeneity in a number of different parameters of glomerular dynamics that was measured (TUCKER, 1977). This was significantly greater than the interanimal variability of mean values.

Former statistical analysis revealed that the precision of estimates of both the measured and the derived parameters in glomerular dynamic studies is appreciably affected by ignoring the intra-animal effect (Breslow, 1990). The importance of the intra-animal variance in studies on glomerular dynamics is maximal when only one or two samples of each measured parameter are obtained in every rat (k=1 or 2) and least when k is large. However, triplicate sampling provides combined standard errors (SE) that are not disturbingly larger than those obtained with k=5 and offers the best cost-benefit ratio in studies of glomerular dynamics.

Due to technical restrictions, it has been customary to measure each parameter in a different nephron within a given kidney, the mean parameter value being derived from a limited number of outer cortical nephrons. It is assumed that the nephrons sampled will provide reasonably representative values for the superficial nephron population as a whole. The precision of such estimates in each of a series of animals, however, depends on the degree of internal variation of the different parameters measured and on the number of nephrons included in the mean (Aitkin, 1987).

In small series of rats, the means should show less variability than individual values sampled from the same animals since, under the latter circumstance, extreme values should be measured purely by chance. Usually, the precision of estimate of the mean value for any given parameter is increasing as the number of measurements increases (Bartoo and Puri, 1967). Although a large sample size is required to improve the precision of estimates of mean values in the face of marked internephron heterogeneity, distinct limitations on the number of repetitions are imposed by limits in the laboratory and in the experiment itself (JACKSON and BLEST, 1982) and thus it is important to define the minimum effective number of measurements for each parameter (Scott L. Zeger and Kung Yee Liang, 1992)

A reasonable compromise between precision and practicality thus must be struck in performing glomerular dynamic studies. One of the principle concerns in renal micropuncture studies of glomerular dynamics is the determination of single nephron glomerular filtration rates (SNGFR), which is then used in deriving estimates of glomerular plasma and blood flows, mean and end capillary filtration pressures, individual pre- and postglomerular vascular resistances and glomerular capillary hydraulic conductivity. Customarily, SNGFR is measured in different nephrons, individual values being averaged with the implicit assumption that the mean values for the nephrons sampled are reasonably representative of the superficial nephron population. Studies of glomerular dynamics are thus expected to be strongly influenced by the degree of variability in SNGFR within a given kidney. SNGFR values of individual nephrons in a given rat kidney exhibit significant variation; thus, a degree of true heterogeneity of SNGFR among different nephrons in a given kidney is not unexpected (Romano, Yang, and O'Malley, 1996).

The controversies over the results obtained by different laboratories are important and concern basic aspects of renal physiology. One of the reasons underlying different results could be represented by the large experimental error inherent to the method. Consequently, the sample size could be insufficient in many experiments and give false statistical differences or obscure the detection of true differences in other circumstances. The error of micropuncture measurements has been studied in different ways. The analytical techniques have been investigated in terms of reproducibility of the results and recoveries of known amounts of chemical.

The accuracy of measurements of chemical insulin, of radioactive glomerular markers, of Na and K concentrations has been studied and published (Romano, Sesma, et al., 1995) (Bartoli et al., 1996). The error due to inter-nephron variability can be eliminated by the recollection technique, which yields reproducible data when the measurements are performed from the same site of the same nephrons. However there is disagreement on the reproducibility and the precision of measurements obtained from different sites of the same nephrons, since many authors report a significant difference in SNGFR measured at the early distal (ED) with respect to the late proximal (LP) tubule, possibly endowed with an important physiologic significance linked to the effect mediated by the Macula Densa (Bartoli et al., 1996). Others did not find such difference. The third method to assess the error of micropuncture techniques deals with intra-animal variability and was studied with the computation of coefficients of variation (Romano, Sesma, et al., 1995).

3.6.2 Objectives

The aim of the case study is to consider a supervised learning approach without explicitly account for correlation and compare the results, in terms of model stability, with those obtained using statistical tecniques that allow for modeling correlation across observations. The reason of developing such case study is that in predictive modeling there is not need to explicitly account for the correlation when training the model, even if when using a holdout set for validation or computation of out-of-sample error it should be ensured that each single observation appears only in one set, either training or validation but not both. The reason is that if the aim is to test hypotheses about coefficients by checking statistical significance, the correlation across observations must be modelled because otherwise the standard errors will be too small. However, coefficients are unbiased even with correlated observations, and thus such model can be used for prediction. In this example, discussion about this issue is shown considering linear regression as machine learning approach (predictive models without accounting for correlation) and comparing results with statistical approaches that correct for non-independence between observations. To compute the error by accounting for all variability sources, the inter-site and intra-site variability, the inter-nephron and intra-nephron variability and the intra-animal and inter-animals variability are considered.

3.6.3 Data

Data were collected from 75 rats studied with micropuncture and clearance techniques.

To be used collectively for the analysis, the following minimal requirements are met in each animal:

- at least two measurements of whole kidney GFR (GFR in uL/min);
- at least one measurement of SNGFR from the early distal tubule (ED-SNGFR in nl/min), paired to a measurement from the last proximal (LP-SNGFR) sampling site of the same nephron. These pairs will represent a measurement of intersite variability;
- at least one paired measurement of SNGFR from the same puncturing site (collection-recollection pair), either from the ED, LP or early proximal (EP) segment of the same nephron. These pairs will represent a measurement of intrasite variability.
- at least two measurements of SNGFR from different nephrons. These values will represent measurements of internephron variability.

3.6.4 Methods

Longitudinal data consist of a number of relatively short non-stationary time series in which the trends $\mu(t)$ are of direct interest. The situation arises in experiment which involve comparisons among trends associated with different treatments. It also applies to growth studies. This involves a change in the practical emphasis of the investigation. In this case we seek only to accommodate serial dependence within the individual time series in order to make valid inferences about the corresponding $\mu_i(t)$.

There are several steps in building a model for longitudinal data; they involve some degree of exploratory analysis, the setup of a reasonable model for the expectation and typically for the variance function of the response variable. The choice of the linear predictor, including fixed and, if it is the case, random effects, is usually crucial in this contest (Smyth, 1989).

In our analysis we used three models

- **naive model**, where independence is assumed between observations, with the main purpose of showing the bias in doing inference discarding the knowledge about the longitudinal nature of the data;
- **marginal model**, where we are interested in the effect of predictors in terms of population average, treating the inter-observation correlation as a nuisance parameter;

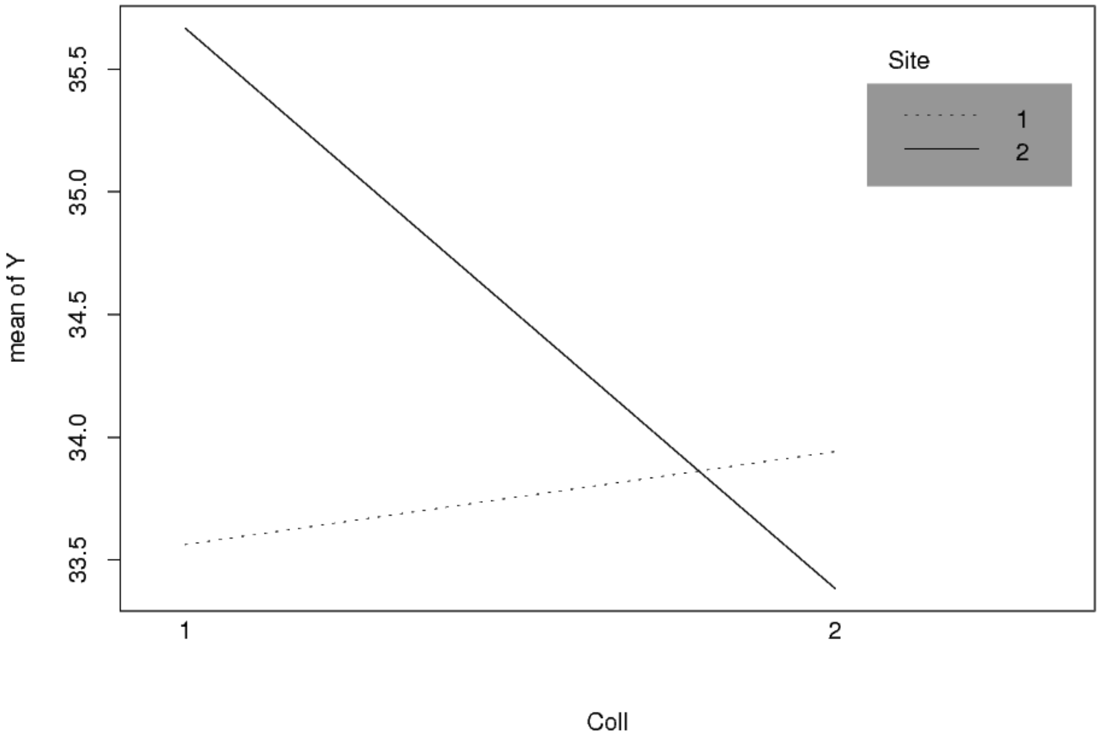


Figure 45: Interaction Plots for Site and Collection

- **conditional model**, where the subject specific effects of the predictors are of interest, as well as the population averaged ones

From the point of view of the linear predictor, a quick exploratory analysis reveals that the relationship between `Site`, `Collection` and `SNGFR` should be modeled including an interaction term between the two independent variables (see Figure 45).

We will discuss now the specification of the models under the three different settings outlined above. We will indicate the `SNGFR` level with Y_{ij}, where $i = 1, \dots, 75$, $j = 1, 2$ and $k = 1, 2$.

Naive (independence) model We assume a linear model for the response,

$$E(Y_{ijk}) = \mu_{jk} \;\; \text{Var}(Y_{ijk}) = \sigma^2\mathbf{I}$$

where $\mathbf{I}$ is, in principle, a 300×300 identity matrix, which implies the observations are independent. The expectation μ_{jk} is linked identically to the linear predictor η_{jk}

$$\eta_{jk} = \beta_0 + \beta_1\text{Site} + \beta_2\text{Coll} + \beta_3\text{Site} \times \text{Coll}$$

The study of the deviance associated in a forward fashion with each term is shown in Table 61. All terms including, as expected, the interaction, are significant, although the Likelihood Ratio Test (LRT) is known to be not conservative if the data are dependent. In fact, the Wald test on the coefficients indicates no significant effect of any of the predictors but the intercept (see Table 62).

Indeed, the residual vary form a minimum of -30.576 to a maximum of 68.927. The variance σ^2 is estimated equal to 351.721.

Marginal model In the marginal approach, the main idea is to use the information about dependency in the data to adjust the estimates of the variability of the coefficients. There's no interest, however, in modeling the

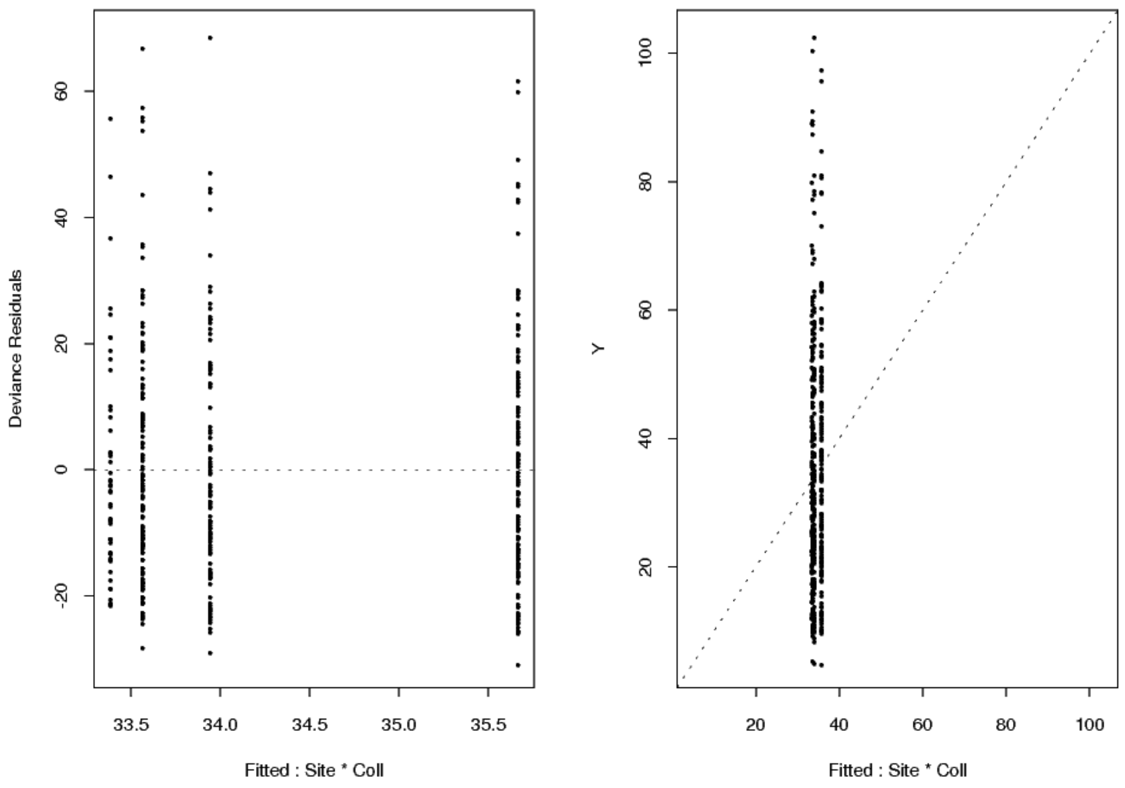


Figure 46: Diagnostics for Independence Model (GLM)

	Df	Deviance	Resid. Df	Resid. Dev
Intercept			442	154809.1
Site	1	204.1300	441	154605.0
Coll	1	37.7571	440	154567.2
Site:Coll	1	161.6537	439	154405.6

Table 61: Analysis of Deviance Table. Naive model assuming independence. Terms added sequentially (first to last).

Coefficients	Value	Std. Error	t value
(Intercept)	34.138	0.981	34.784
Site	0.386	0.981	0.393
Coll	-0.475	0.981	-0.484
Site:Coll	-0.665	0.981	-0.677

Table 62: Naive model (Independence).

correlation at a subject level, which is treated essentially as a nuisance parameter (S. L. Zeger, K. Y. Liang, and Self, 1985; S. L. Zeger and K. Y. Liang, 1986).

The model specification is the same as in the naive setting, with the only difference regarding the variance function, which is now block diagonal, with each block of dimension 4×4 parameterized as a function of the correlation, according to several specifications.Three structures for the working correlation matrix were used

- *independence*, which corresponds to the identity matrix as in the naive approach;
- *unstructured*, which uses 6 parameters, one for each term in the 4×4 block;

- *exchangeable*, which uses only one parameter.

The fit of the models is shown in Table 63, where there's no evidence of `Site` or `Collection` effect, even if the correct robust Standard Errors are used instead of the naive ones. The fit is not particularly satisfactory, with the residuals ranging from -30.57 to 68.92.

Notice that, because of the identical specification of the expectation the marginal and the naive model give exactly the same fitted values and residuals: the only part where the model differs is the variance function specification. The σ^2 is estimated, under the three parameterizations, as respectively 340.173, 339.814, 340.447.

Coefficients	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
Independence Working Correlation					
(Intercept)	34.138	0.965	35.370	1.640	20.806
Site	0.386	0.965	0.400	0.843	0.457
Coll	-0.475	0.965	-0.492	0.907	-0.524
Site:Coll	-0.665	0.965	-0.689	0.893	-0.744
Unstructured Working Correlation					
(Intercept)	33.733	1.176	28.671	1.519	22.204
Site	0.171	0.930	0.183	0.797	0.214
Coll	-0.836	0.909	-0.920	0.852	-0.981
Site:Coll	-0.659	0.909	-0.725	0.812	-0.811
Exchangeable Working Correlation					
(Intercept)	34.152	1.426	23.944	1.477	23.116
Site	0.565	0.871	0.648	0.749	0.754
Coll	-0.386	0.839	-0.460	0.897	-0.437
Site:Coll	-0.485	0.871	-0.557	0.770	-0.630

Table 63: GEE models for entire dataset

Conditional model The conditional model, in contrast with the marginal one, assumes that the observations are independent after the response has been conditioned to the random effect parameters. This implies that the expectation term is modified to incorporate the random parameters, and can be written, in general, as

$$E(y_{ijk}|b_i) = \mu_{ijk}$$

with

$$\eta_{ijk} = \beta_0 + b_0 + (\beta_1 + b_1)\text{Site} + (\beta_2 + b_2)\text{Coll} + (\beta_3 + b_3)\text{Site} \times \text{Coll}$$

where the β terms indicates the fixed effects and the b terms the random effects, which are considered *iid* normal distributed. As shown in Table 64, the random effect slopes are not significant using mixed chi-squared distributions for random effects likelihood ratio tests (Pinheiro and Bates, 2000), leading to a simplified random intercept model:

$$\eta_{ijk} = \beta_0 + b_0 + \beta_1\text{Site} + \beta_2\text{Coll} + \beta_3\text{Site} \times \text{Coll}$$

	Df	Log-likelihood
b_0	1	-1890.65
$b_0 + b_1$	1	-1889.25
$b_0 + b_2$	1	-1890.29
$b_0 + b_1 + b_2 + b_3$	1	-1888.98

Table 64: Log-likelihood for the various random effects.

No fixed effect term is significant, as in the previous two models, as shown in Table 65; the random intercept has a high variability, both in range from rat to rat and in standard deviation, equal to 9.938 (see Figure 47). The cluster residual variance is quite high (246.305), but the residuals are much smaller than in the naive and marginal model, ranging from -2.065 to 4.308 with a median of -0.151, indicating that a subject specific effect is necessary to account for the variability unexplained by the fixed effects.

.

Table 65: Mixed Effect Model

	Value Approx.	Std.Error	z ratio(C)
(Intercept)	34.148	1.472	23.183
Site	0.573	0.861	0.665
Coll	-0.381	0.828	-0.460
Site:Coll	-0.477	0.861	-0.554

3.6.5 Model Stability for selected subsamples

In this kind of study, there are two main approaches in the design of the study, depending on the costs and availability of resources. The choice is usually in between

- a relatively high number of rats, with one or two measurements for each rat;
- a high number of observations for a relatively small number of rats.

The two designs are quite different, and in this sense, they are expected to provide different insight into the mechanism studied. In particular, we investigated the performance of the models discussed before, under three alternative designs, keeping however the total number of observations constant, equal to 22.

One nephron per rat In this case, we selected the 22 rats for which only one nephron has been studied. The fit for the naive model is shown in Table 66; the measurements done in the two `Collection` occasions result to be significantly different. The collection and the interaction term are significantly different in the marginal model (see Table 67), where the correlation parameter has been estimated equal to 0.339. Variance term σ^2 is estimated under the naive model as 277.12, and in the marginal model under assumption of diagonal working correlation matrix as 264.019; finally, under assumption of exchangeability, equal to 264.041. The unstructured model did non converge, due to an overparametrization of the model.

Coefficients	Value	Std. Error	t value
(Intercept)	28.256	3.326	8.495
Site	-1.717	3.326	-0.516
Coll	-6.527	3.326	-1.962
Site:Coll	-5.141	3.326	-1.545

Table 66: Naive model. Subsample of 22 rats, one nephron per rat

The conditional model shown in Table 68 reveals a significant difference among measurements, with a standard deviation of the random effects unchanged, equal to 9.487 and a cluster residual variance of 171.424.

Two nephrons per rat In this case, we fitted the models on the basis of two nephrons per rat, for a total number of 22 observations and 11 rats. All models did not show any particular difference with the main model, the naive (Table 69), the marginal (Table 70) and the conditional (Table 71). The estimated variance σ^2 is equal to 495.256 for the naive model and 467.49, 474.337 and 468.538 for the marginal model, respectively with independent, unstructured and exchangeable working correlation matrix. The estimated inter-cluster correlation is equal to 0.301. For the conditional model, the cluster residual variance is 325.526, with a standard deviation of the random intercept equal to 11.89. The fit is quite good, with residuals ranging from -1.725 to 2.921.

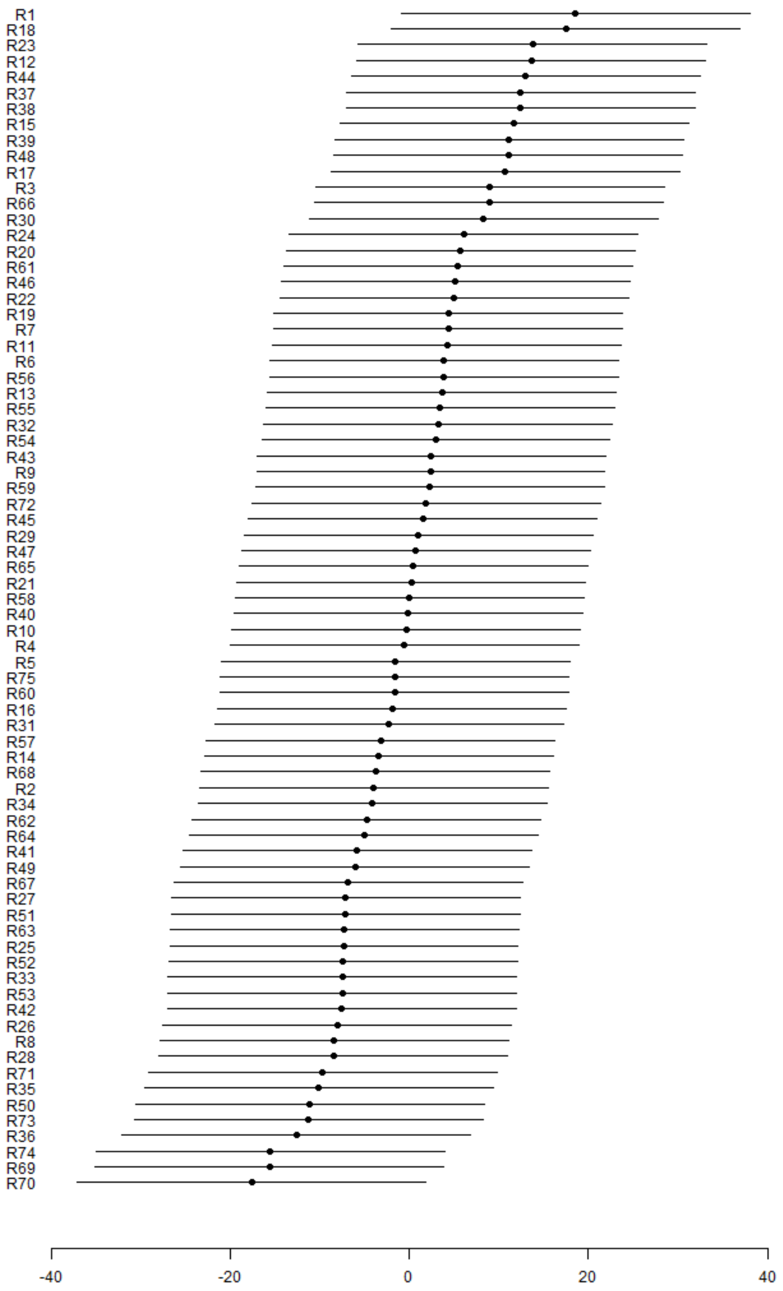


Figure 47: Caterpillar plot of random effects

Coefficients	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
Independence model					
(Intercept)	28.256	3.245	8.705	1.968	14.354
Site	-1.717	3.245	-0.529	1.196	-1.435
Coll	-6.527	3.245	-2.010	1.743	-3.743
Site:Coll	-5.141	3.245	-1.583	0.954	-5.384
Exchangeable model					
Coefficients	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	28.404	3.470	8.185	2.338	12.146
Site	-1.453	2.864	-0.507	1.455	-0.998
Coll	-6.378	2.823	-2.259	1.754	-3.635
Site:Coll	-4.877	2.864	-1.702	1.315	-3.708

Table 67: Marginal model. Subsample of 22 rats, one nephron per rat

Fixed Effects	Estimates Value	Approx. Std.Error	z ratio(C)
(Intercept)	28.407	3.454	8.223
Site	-1.449	2.841	-0.510
Coll	-6.376	2.800	-2.277
Site:Coll	-4.873	2.841	-1.715
Random Effects (Conditional Modes) for case 2			
rat	(Intercept)		
2	-3.525		
4	3.240		
5	-0.820		
8	-8.480		
13	5.056		
20	7.341		
21	1.223		
24	7.790		
25	-7.220		
26	-8.029		
27	-7.554		
28	-8.535		
29	2.056		
30	10.235		
32	4.558		
33	-7.420		
34	-3.699		
35	-10.411		
38	14.845		
39	13.471		
40	0.792		
42	-4.914		

Table 68: Conditional model. Subsample of 22 rats, one nephron per rat

More than five nephrons per rat The last dataset is based on only 4 rats, for which either 5 or 6 nephrons have been considered. All models do not show any particular difference with the main model, either the naive (Table 72), the marginal (Table 73) or the conditional (Table 74). The variance σ^2 is estimated equal to 356.634 for the naive model, and 345.828 and 345.836 for the marginal model respectively with independence and exchangeable working correlation matrix. The estimated correlation is 0.437, 50% higher than in the previous case. The conditional model indicates, as usual, a good fit, the residuals ranging from -1.422 to 1.959, and a cluster residual variance equal to 63.080. The standard deviation of the random coefficient is 50% higher,

Coefficients	Value	Std. Error	t value
(Intercept)	40.595	3.490	11.631
Site	0.648	3.490	0.185
Coll	0.342	3.490	0.098
Site:Coll	-2.686	3.490	-0.769

Table 69: Naive model. Subsample of 11 rats, two nephrons per rat

Coefficients	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	40.595	3.390	11.971	4.620	8.785
Site	0.648	3.390	0.191	3.392	0.191
Coll	0.342	3.390	0.101	2.699	0.126
Site:Coll	-2.686	3.390	-0.792	3.390	-0.792
Unstructured model					
(Intercept)	38.642	4.343	8.895	4.350	8.883
Site	-0.339	2.361	-0.143	3.135	-0.108
Coll	-0.904	2.696	-0.335	2.269	-0.398
Site:Coll	-5.639	2.551	-2.210	3.765	-1.497
Exchangeable model					
(Intercept)	40.171	4.604	8.724	4.749	8.458
Site	-0.282	2.935	-0.096	3.311	-0.085
Coll	-0.081	2.888	-0.028	2.466	-0.032
Site:Coll	-3.617	2.935	-1.232	2.994	-1.208

Table 70: Marginal model. Subsample of 11 rats, two nephrons per rat.

Fixed Effects	Estimates Value	Approx. Std.Error	z ratio(C)
(Intercept)	40.170	4.600	8.732
Site	-0.284	2.928	-0.097
Coll	-0.082	2.881	-0.028
Site:Coll	-3.618	2.928	-1.235
Random Effects (Conditional Modes)			
Rat	(Intercept)		
1	14.669		
3	4.947		
6	-0.316		
9	-2.803		
14	-8.441		
18	12.634		
23	8.873		
31	-7.605		
36	-18.189		
37	7.482		
41	-11.250		

Table 71: Conditional model. Subsample of 11 rats, two nephrons per rat.

equal to 15.022.

3.6.6 Conclusion

Several useful indications result from the study. There is a strong indication that the subject specific models give a better fit in situation like this, where the intra animal variability is high, than the population average models, even if adjusted for the cluster effect (Chesher, 1984). As a consequence, the model is more stable if

Coefficients	Value	Std. Error	t value
(Intercept)	33.690	2.465	13.664
Site	-1.005	2.465	-0.407
Coll	3.667	2.465	1.487
Site:Coll	0.460	2.465	0.186

Table 72: Naive model. 4 rats, 5 or 6 nephrons per rat.

Model for independence					
Coefficients	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	33.690	2.427	13.876	7.212	4.671
Site	-1.005	2.427	-0.413	3.244	-0.309
Coll	3.667	2.427	1.510	3.084	1.188
Site:Coll	0.460	2.427	0.189	3.059	0.150
Model for exchangeable Correlation					
(Intercept)	34.117	6.418	5.315	7.272	4.691
Site	-0.863	1.918	-0.450	1.771	-0.487
Coll	3.667	1.820	2.014	3.106	1.180
Site:Coll	0.602	1.918	0.313	1.375	0.437

Table 73: Marginal model. 4 rats, 5 or 6 nephrons per rat.

Fixed Effects Estimates	Value	Approx.Std.Error	z ratio(C)
(Intercept)	30.449	7.806	3.900
Site	1.796	2.380	0.754
Coll	1.581	2.173	0.727
Site:Coll	1.253	2.455	0.510
Random Effects (Conditional Modes)			
Rat	(Intercept)		
12	19.565		
19	8.293		
63	-11.732		
69	-16.127		

Table 74: Conditional model. 4 rats, 5 or 6 nephrons per rat.

at least 2 observations (nephrons) are taken on each animal. The conclusions using a one-nephron-design can yield in fact to results that could be misleading.

There are however some cautions in using the methods: the model, used here for comparison, based on only one measurement in each rat, is saturated. Moreover, the risk of overparameterizing the model is always present, limiting the degrees of freedom available for the analysis.

Examples bibliography

[Ait87] M. Aitkin. “Modeling Variance Heterogeneity in Normal Regression Using Glim”. In: *Applied Statistics-Journal of the Royal Statistical Society Series C* 36.3 (1987), pp. 332–339. url: %3CGo%20to%20ISI%3E://WOS:A1987K880300009.

[Ara81] F. J. Aranda Ordaz. “On two families of transformations to Additivity for Binary Response data”. In: *Biometrika* 68 (1981), pp. 357–363.

[Bar+96] M. Bartoli et al. “Modeling Litz-wire winding losses in high-frequency power inductors”. In: *PESC Record. 27th Annual IEEE Power Electronics Specialists Conference*. Vol. 2. June 1996, 1690–1696 vol.2. doi: 10.1109/PESC.1996.548808.

[BH09] Patrick Breheny and Jian Huang. “Penalized methods for bi-level variable selection”. In: *Statistics and its interface* 2.3 (2009), p. 369. url: http://www.ncbi.nlm.nih.gov/pubmed/20640242.

[BH15] Patrick Breheny and Jian Huang. “Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors”. In: *Statistics and computing* 25.2 (2015), pp. 173–187. issn: 0960-3174.

[BP67] J. B. Bartoo and P. S. Puri. “On optimal asymptotic tests of composite statistical hypotheses”. In: *Annals of Mathematical Statistics* 38 (1967), pp. 1845–52.

[Bre90] N. Breslow. “Tests of Hypothesis in overdispersed regression and other quasi-likelihood models”. In: *Journal of the American Statistical Association* 85 (1990), pp. 1–26.

[Che84] A. Chesher. “Testing for neglected heterogeneity”. In: *Econometrica* 52 (1984), pp. 1–26.

[CKR07] K. J. Cios, L. A. Kurgan, and M. Reformat. “Machine learning in the life sciences”. In: *IEEE Engineering in Medicine and Biology Magazine* 26.2 (Mar. 2007), pp. 14–16. issn: 0739-5175. doi: 10.1109/MEMB.2007.335579.

[Cru95] Kenny S. Crump. “Calculation of Benchmark Doses from Continuous Data”. In: *Risk Analysis* 15.1 (1995), pp. 79–89. issn: 1539-6924. doi: 10.1111/j.1539-6924.1995.tb00095.x. url: http://dx.doi.org/10.1111/j.1539-6924.1995.tb00095.x.

[Eur17] European Food Safety Authority. “Update: use of the benchmark dose approach in risk assessment”. In: *EFSA Journal* 15.1 (2017). e04658, e04658–n/a. issn: 1831-4732. doi: 10.2903/j.efsa.2017.4658. url: http://dx.doi.org/10.2903/j.efsa.2017.4658.

[GS90] S. A. Glantz and B. K. Slinker. *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill, 1990.

[Hua+09] Jian Huang et al. “A group bridge approach for variable selection”. In: *Biometrika* 96.2 (2009), pp. 339–355. issn: 0006-3444.

[JB82] R. R. JACKSON and A. D. BLEST. “Short Communication: The Distances at Which a Primitive Jumping Spider, Portia Fimbriata, Makes Visual Discriminations”. In: *Journal of Experimental Biology* 97.1 (1982), pp. 441–445. issn: 0022-0949. eprint: http://jeb.biologists.org/content/97/1/441.2.full.pdf. url: http://jeb.biologists.org/content/97/1/441.2.

[PB00] José C Pinheiro and Douglas M Bates. *Mixed-effects models in S and S-PLUS*. New York, NY [u.a.]: Springer, 2000. isbn: 0387989579 9780387989570 9781441903174 1441903178. url: http://www.worldcat.org/search?qt=worldcat_org_all&q=1441903178.

[Pre76] R. L. Prentice. “A generalization of the probit and logit methods for dose response curves”. In: *Biometrics* 32.4 (1976), pp. 761–8. url: http://www.ncbi.nlm.nih.gov/pubmed/1009225.

[Rom+95] Carmelo Romano, Michael A. Sesma, et al. “Distribution of metabotropic glutamate receptor mGluR5 immunoreactivity in rat brain”. In: *The Journal of Comparative Neurology* 355.3 (1995), pp. 455–469. issn: 1096-9861. doi: 10.1002/cne.903550310. url: http://dx.doi.org/10.1002/cne.903550310.

[RYO96] Carmelo Romano, Wan-Lin Yang, and Karen L. O’Malley. “Metabotropic Glutamate Receptor 5 Is a Disulfide-linked Dimer”. In: *Journal of Biological Chemistry* 271.45 (1996), pp. 28612–28616. doi: 10.1074/jbc.271.45.28612. eprint: http://www.jbc.org/content/271/45/28612.full.pdf+html. url: http://www.jbc.org/content/271/45/28612.abstract.

[Seb+91] J. L. Sebaugh et al. "A study of the shape of dose-response curves for acute lethality at low response: a megadaphnia study". In: *Risk Analysis* 11.4 (1991), pp. 633–640.

[Slo02] Wout Slob. "Dose-Response Modeling of Continuous Endpoints". In: *Toxicological Sciences* 66.2 (2002), p. 298. doi: 10.1093/toxsci/66.2.298. url: +%20http://dx.doi.org/10.1093/toxsci/66.2.298.

[Smy89] G. K. Smyth. "Generalized Linear Models with varying dispersion". In: *Journal of the Royal Statistical Society, series B* 51 (1989), pp. 1–26.

[Stu] Thèrése A. Stukel. "Implementation of an Algorithm for fitting a Class of Generalized Logistic Models". In: ed. by R. Gilchrist. Springer Verlag, pp. 160–167.

[Stu83] Thèrése A. Stukel. "Generalized Logistic Models". In: (1983).

[TUC77] LOIS E. TUCKER. "Regulation of Ions in the Haemolymph of the Cockroach Periplaneta Americana During Dehydration and Rehydration". In: *Journal of Experimental Biology* 71.1 (1977), pp. 95–110. issn: 0022-0949. eprint: http://jeb.biologists.org/content/71/1/95.full.pdf. url: http://jeb.biologists.org/content/71/1/95.

[Wou+01] S Woutersen et al. "Hydrogen-bond lifetime measured by time-resolved 2D-IR spectroscopy: N-methylacetamide in methanol". In: *Chemical Physics* 266.2–3 (2001), pp. 137–147. issn: 0301-0104. doi: http://dx.doi.org/10.1016/S0301-0104(01)00224-5. url: http://www.sciencedirect.com/science/article/pii/S0301010401002245.

[ZL86] S. L. Zeger and K. Y. Liang. "Longitudinal data analysis for discrete and continuous outcomes". In: *Biometrics* 42.1 (1986), pp. 121–30. url: http://www.ncbi.nlm.nih.gov/pubmed/3719049.

[ZL92] Scott L. Zeger and Kung Yee Liang. "An overview of method for the analysis of longitudinal data". In: 11 (1992), pp. 1825–39.

[ZLS85] S. L. Zeger, K. Y. Liang, and S. G. Self. "The Analysis of Binary Longitudinal Data with Time-Independent Covariates". In: *Biometrika* 72.1 (1985), pp. 31–38. url: %3CGo%20to%20ISI%3E://WOS:A1985AEB8600004%20http://biomet.oxfordjournals.org/content/72/1/31.

4 A decision tree/recipe book to help the choice of the most appropriate methodology — From the problem to the approach

4.1 Description of ML algorithms

In order to chose the right technique for a given problem from the highly heterogeneous MLT repertoire, a MLT taxonomy is developed as a support for a task-oriented decision tree.

The main concepts, bridging the MLT and the classical statistical literature, are those of variable characteristics, supervision, scalability, data size (both as sample size and dimensionality), and robustness. Techniques are commonly chosen in agreement with their capability of matching one or more of those aspects.

- **Supervised, semi-supervised, clustering and unsupervised techniques**. In a classification problem, we have a set of elements divided into classes. Given an element (or instance) of the set, a class is assigned according to some of the element's features and a set of classification rules. In many real-life situations, this set of rules is not known, and the only information available is a set of labelled examples (i.e. a set of instances associated with a class). Supervised classification paradigms are algorithms that induce the classification rules from the data. In two-group supervised classification, there is a feature vector $x \in \mathcal{R}$ whose components are called predictor variables and a label or class variable $\mathcal{C} \in \{0, 1\}$. Hence, the task is to induce classifiers from training data, which consists usually of a set of N independent observations drawn from the joint probability distribution $p(x, c)$. Unsupervised learning is a type of ML algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.

  Semi-supervised classification is a special form of the general classification (Zhu, 2005). Traditional classifiers use only labeled data (feature and label pairs) to train. Labeled instances however are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced human annotators. Meanwhile unlabelled data may be relatively easy to collect, but there has been few ways to use them. Semi-supervised learning addresses this problem by using large amount of unlabelled data, together with the labeled data, to build better classifiers. The problem in such acontext is that they can under-perform because of bad matching of problem structure with model assumptions. This can lead to degradation in classifier performance. For example, quite a few semi-supervised learning methods assume that the decision boundary should avoid regions with high $p(x)$. For example, quite a few semi-supervised learning methods assume that the decision boundary should avoid regions with high p(x); these methods include:

  - transductive SVMs,
  - information regularization,
  - Gaussian processes with null category noise model,
  - graph-based methods if the graph weights is determined by pairwise distance.

  Nonetheless if the data is generated from two heavily overlapping Gaussian, the decision boundary would go right through the densest region, and these methods would perform badly.

- **Features and labels characteristics**. A traditional taxonomy of MLT is based on the characteristics of both inputs and outputs (Saeys, Inza, and Larrañaga, 2007), identifying vector-like data (the conventional data-matrix) and the general object concept, where any object can be used either as feature or output. A first distinction to be drawn is between methods able to deal with only one output at time and the others being capable to address multiple outcomes simultaneously. Noticeably, inputs and outputs intersection is providing a good way for matching MLT techniques (Kotsiantis, Zaharakis, and Pintelas, 2007). This distinction in output type has led to a naming convention for the prediction tasks: regression when the aim is to predict quantitative outputs, and classification when the aim is to predict qualitative outputs. Both can be viewed as a task in function approximation.

Table 75

inputs\outputs	real numbers	binary	categorical	ordered, sequencing	general objects
real number	•	•	•	•	
binary	•	•	•	•	
categorical	•	•	•	•	
sequence	•	•	•	•	
general objects	•	•	•	•	

- **Scalability and number of instances**. Scalability and data efficiency is a growing issue in the MLT world, as the issue of MLT computing by facing an increasing number of instances (J. Lin and Kolcz, n.d.) (possibly recursively (Syed et al., 1999)) is becoming relevant. Stochastic and randomized algorithms are the major solution in this field, although not necessarily as an exclusive approach.

- **Sample size**. The sample size needed to solve a MLT problem depends on the method used to find the parameters of the classification rule, the number of features, the asymptotic probability of misclassification (error rate), and the desired learning accuracy. Raudys and Jain (Raudys and Jain, 1991) provide extensive discussion on the relationship between the sample size N and classification accuracy for a two class, Gaussian distribution data set. They show that the increase in classification error of the parametric classifiers is proportional to $1/N$ and depends on the dimensionality of the feature space p; for the linear classifiers, the relationship is linear and for quadratic classifiers the relationship is quadratic (only for large p). For non-parametric classifiers such as Parzen windows or k nearest-neighbor, the increase in classification error is proportional to $1/N$ or $1/\sqrt{(N)}$. These estimates are for Gaussian distributions with equal covariances. Additional factors will influence these estimates for data with unequal covariances and different number of samples per class. For other data distributions, and more than two class problems, it is recommended to estimate the metric $\hat{\Delta}_N = \hat{P}_c - \frac{\hat{P}_R}{2}$, where $\hat{\Delta}_N$ is the increase in classification error, $\hat{P}_c$ is the leave-one-out estimate of classification error, and $\hat{P}_R$ is the re-substitution estimate of the classification error.

  As a rule of thumb to understand whether an actual sample size is appropriate or not, is to evaluate if the difference $\hat{\Delta}_N$ is small in comparison with the empirical estimate of asymptotic probability of misclassification (error rate). If this is the case, then the sample size is considered sufficient.

  Mukherjee et al. (Mukherjee et al., 2003) address a similar question on what is the relationship between sample size and classification performance. Approaches within the statistics and pattern recognition communities have used power calculations (Adcock, 1997; Guyon et al., 1998), but assume data normality and independence of variables—assumptions that may not necessarily hold. They compute bounds or estimates of a quantity's deviation from its expected value as a function of the number of samples. Unfortunately, these methods are not suitable for predicting the future performance of a classifier as the sample size is increased. Learning curves estimate the empirical error rate as a function of the training set for a given classifier and data set. These learning curves are well characterized by inverse-power laws: $e(N) = aN - \alpha + \beta e(N) = \alpha N - \alpha + \beta$, where $e(N)$ is the expected error rate, N is the number of samples, α is the learning rate, and β is the Bayes error which is the minimum error rate achievable. These parameters take on different values depending on the type of classifier and data set being used. As the data set increases in size asymptotically, the error rate approaches β. This equation holds well for a number of classifiers. Using this power-law scaling model as a basis, one can use the empirical error rates of a classifier over a range of training set sizes drawn from a data set to fit an inverse-power law model. The fitted inverse-power law model can be used to extrapolate the error rate to larger data sets. Tests have been developed (Mukherjee et al., 2003) to detect when this model fails (especially with very small sample sizes), such that this part of the curve is ignored when fitting and extrapolating.

- **Robustness**. The performance of the algorithm robustness is defined as the case where estimates does not deteriorate too much when training and testing with slightly different data are considered in the analysis (either by adding noise or by taking other dataset), hence, algorithm is prone to overfitting. This robustness property is also known as algorithmic stability.

- **Sparsity, stability and high-dimensionality**. Stability and Sparsity have both emerged as important properties of machine learning algorithms (H. Xu, Caramanis, and Mannor, 2012). In a broad sense, stability means that an algorithm is well-posed, so that given two very similar data sets, the algorithm's output varies little. More specifically, an algorithm is stable if its output is nearly identical on two data sets differing on only one sample (this is known as the leave-one-out error). Stability itself is a desirable property for learning algorithms. For example, in feature-selection, one might seek algorithms that select nearly the same feature set when run on very similar data sets. Following the landmark work in (Bousquet and Elisseeff, 2002), stability is also pivotal for proving generalization performance of an algorithm.

  Sparsity is another useful property of machine learning algorithms. An algorithm yields a sparse result when only a small number of coefficients are nonzero, among all those it has estimated. In term of statistical properties, sparsity is associated with fast evaluation and fast optimisation models (i.e. parameters optimization in SVM), stability itself and the capability to address regularization (i.e. LASSO classification and regression)(Hastie, Robert Tibshirani, and Wainwright, 2015).

  Recently, sparse machine learning algorithms have emerged as a powerful tool to get models of high-dimensional data with high degree of interpretability at low computational cost. Generally, problems with high-dimensional data arise from the fact that a fixed number of data points become increasingly sparse as dimensionality increases. Principal component analysis is a classic technique to overcome this problem. Sparse principal component analysis is a variant of Principal Component Analysis (PCA) that allows to find sparse directions with high variance (Zou, Hastie, and Robert Tibshirani, 2006a) and it is one of the main algorithm in sparse machine learning along with regularization algorithms, sparse graphical models and the analogous sparse Bayesian Learning methods (Hastie, Robert Tibshirani, and Wainwright, 2015), (Tipping, 2001).

The large error rate of a classifier can be usually attributed to the inherent difficulty of the classification problem. However, in finite sample situations, the following factors can also decrease the performance of the classifier: (*i*) small number of samples; (*ii*) large number of features; (*iii*) complexity of the classification rule (e.g. quadratic versus linear discriminant function); (*iv*) presence of outliers and (*v*) inappropriate width for a classifier involving non-parametric kernel density estimation.

4.1.1 Main MLT approaches

Before going into the details of the decision tree built as a tool useful to assist the choice of the right MLT for a given problem, the main groups of techniques that will be involved in the decision tree itself are described. In the following sections, *training set* is the name given to the data on which the MLTs are fine tuned, whereas the performances of each MLT are measured against a *test set*. By comparison, the *external error rate* is then estimated.

Regression-based algorithms

Regression models are such a wide area in MLT that it is impossible to cover all of them. The sketch provided here is based on the common notation provided by the GLM.

Let $\boldsymbol{y}$ be a vector of n observations assumed to be a realization of a random variable Y, whose components are independently distributed with mean μ and constant variance σ^2.

The distribution of Y belongs to a simple exponential family. Further let $\boldsymbol{x}$ be a p component vector of covariate that produces a linear predictor

$$\eta = \sum_{j=1}^{p} \beta_j x_j \tag{6}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters of dimension p.

The GLM relates the mean μ to the linear predictor η by a smooth invertible function given by

$$\eta = g(\mu) \qquad \mu = h(\eta).$$

In such a case, the function g is called the *link* function. A specific GLM is characterized by the exponential family and the link function. As in classical linear models, the β's represent the regression coefficients.

The *link function* relates the linear predictor η to μ, the mean of given y. A link function is selected according to the distribution of the response variable. For the binomial distribution, we have $0 < \mu < 1$, and therefore, the link function should map the interval [0,1] onto the whole real line.

The principal link functions in such cases are

$$logit \tag{7}$$
$$probit \tag{8}$$
$$complementarylog - log \tag{9}$$
$$log - log \tag{10}$$

A common link function for count data is the log-link. Reciprocal link is used for continuous data exhibiting constant coefficient of variation.

In all of applied statistics, linear least squares regression (identity link) ranks among the most versatile and often used methods of data analysis. Its limitations are twofold:

- the mean response is a linear function of the regression parameters
- the error variance is the same for all observations

Regression models are usually implemented under different names, according to the specific problem and the cultural habit of the researcher.

Linear regression Normality, linearity, and homoscedasticity are generally required, but these requirements could be relaxed still remaining within a linear regression model. In any case for classical linear regression the assumption of independence is also made.

Multiple linear regression Predicts the value of a quantitative variable for a new instance as a linear combination of several numerical variables. Requires normality, linearity, homoscedasticity and independence

ANOVA Predicts the value of a quantitative variable for a new instance as a linear combination of one or two qualitative variables. Requires conditional normality, linearity, homoscedasticity and independence.

GLM Predicts the value of a qualitative or quantitative variable for a new instance as a linear combination of several numerical and qualitative variables.

Statistical clustering

Clustering methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality.

Clustering consists in partitioning a set of elements into subsets according to the differences between them. In other words, it is the process of grouping similar elements together. The main difference from the supervised classification is that, in clustering, we have no information about how many classes there are (Larrañaga et al., 2006). Cluster analysis, also called data segmentation, has a variety of goals. All relate to grouping or segmenting a collection of objects into subsets or 'clusters', such that those within each cluster are more closely related to one another than objects assigned to different clusters. Sometimes the goal is to arrange the clusters into a natural hierarchy. This involves successively grouping the clusters themselves so that, at each level of the hierarchy, clusters within the same group are more similar to each other than those in different groups. Central to all of the goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects being clustered. A clustering method attempts to group the objects based on the definition of similarity supplied to it. This can only come from subject matter considerations.

The most popular clustering algorithms are (i) k-Means, (ii) k-Medians, (iii) Expected Minimization (EM) and (iv) Hierarchical Clustering.

K-means K-means is one of the simplest unsupervised learning algorithms. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assuming k clusters) fixed apriori.

The starting point is to define k centers, one for each cluster as much as possible far away from each other. The second step is to take each point belonging to a given data set and associate it to the nearest center until no more any point is left.

Then, k new centroids are calculated as barycenter of the clusters resulting from the previous step and a new binding has to be done between the same data set points and the nearest new center. Following this procedure, the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} D(x_i - v_j)^2$$

where $D(\cdot)$ is the Euclidean distance, c_i is the number of data points in the ith cluster and c is the number of cluster centers.

The algorithm is fast, robust and relatively efficient. However the main disadvantages it that it requires a-priori specification of the number of cluster centers.

Furthermore, it uses Exclusive Assignment, i.e. if there are two highly overlapping data then k-means will not be able to resolve the problem by indicating that there are two clusters.

k-medians It is a variation of the k-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median. This has the effect of minimizing error over all clusters with respect to the 1-norm distance metric (for example, Manhattan distance) as opposed to the Euclidean distance.

Hierarchical Clustering Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called Hierarchical Agglomerative Clustering (HAC) . Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached.

Dimensionality Reduction Algorithms

Dimensionality reduction seek and exploit the inherent structure in the data, usually in an unsupervised manner, to summarize or describe data using less information. Many of these methods can be adapted for use in classification and regression. They include PCA, PCR, Partial Least Squares Regression (PLS), Sammon Mapping, Multidimensional Scaling (MDS), Projection Pursuit, LDA, Mixture Discriminant Analysis (MDA), QDA, Flexible Discriminant Analysis (FDA).

PCA is a commonly used data reduction technique (Abdi and L. J. Williams, 2010). This method seeks to find linear combinations of the predictors, known as Principal Components (PCs), which capture the most possible part of the variance of the response variable. The first PC is defined as the linear combination of the predictors that captures variability of all possible linear combinations. Then, subsequent PCs are derived such that these linear combinations capture most of the remaining variability while also being uncorrelated with all previous PCs:

$$PC_j = (a_{j1} \times Predictor_1) + (a_{j2} \times Predictor_2) + \ldots + (a_{jP} \times Predictor_P).$$

The coefficients $a_{j1}, a_{j2} \ldots, a_{jP}$ are the component weights and help us understand which predictors are most important to each PC.

The primary advantage of PCA is that it creates components that are uncorrelated. Some predictive models reqiuire predictors to be uncorrelated (or at least low correlation) in order to find solutions and to improve the model's numerical stability. PCA preprocessing creates new predictors with desirable characteristics for these kinds of models.

Since PCA seeks linear combinations of predictors that maximize variability, it will naturally first be drawn to summarizing predictors that have more variation. If the original predictors are on measurement scales that differ in orders of magnitude, then the first few components will focus on summarizing the higher magnitude predictors while latter components will summarize lower variance predictors. This means that the PC weights will be larger for the higher variability predictors on the first few components. In addition, it means that PCA will be focusing its efforts on identifying the data structure based on measurement scales rather than based on the important relationships within the data for the current problem. Centering and scaling variables allow to overcome this issue, making PCA independent from the measurement scale.

The second caveat of PCA is that it does not consider the modeling objective or response variable when summarizing variability. Since PCA is blind to the response, it si *an unsupervised technique*.

For data sets with many predictor variables, a heuristic approach for determining the number of components to retain is to create a scree plot, which contains the ordered component number (on the x-axis) and the amount of summarized variability (on the y-axis). The first few PCs will summarize a majority of the variability, and the plot show a steep descent; variation decreases for the remaining components. Generally, the component number prior to the decreasing of variation is the maximal component that is retained.

PCR Pre-processing predictors via PCA prior to performing regression is known as PCR (Massy, 1965). PCR is a technique for analyzing multiple regression data that suffer from multicollinearity.

Collinearity is the situation where a pair of predictor variables have a substantial correlation with each other. When this kind of situation occurs between multiple predictors at once, then it is called *multicollinearity*.

To carry out a PCR, a set of linear combinations of predictors it is chosen and regressed the outcome.

The particular combinations used are the sequence of principal components of the inputs, and are uncorrelated and ordered by decreasing variance.

PC of some input data points. The largest PC is the direction that maximizes the variance of the projected data, and the smallest PC minimizes that variance.

PCR has been widely applied in the context of problems with inherently highly correlated predictors or problems with more predictors than observations. While this two-step regression approach (dimension reduction, then regression) has been successfully used to develop predictive models under these conditions, it can easily be misled. Specifically, dimension reduction via PCA does not necessarily produce new predictors that explain the response.

Indeed, in selecting independent components, response is not taken into consideration by PCA. In fact, principal components are variables that explain variation in the predictors space.

Discriminant analysis The typical discriminant analysis problem deals with a population consisting of two groups, π_1 and π_2. By observing a $k \times 1$ vector **x**, the idea is to assign the individual whose measurements are given by **x** to π_1 or π_2.

PLS originated with Herman Wold's Nonlinear Iterative Partial Least Squares (NIPALS) algorithm (Wold 1966, 1982) which linearized models that were nonlinear in the parameters subsequently adapted Wold et al. (1983) for the regression setting with correlated predictors.

As PCA, PLS finds linear combinations of the predictors. These linear combinations are commonly called components or latent variables. While the PCA linear combinations are chosen to maximally summarize predictor space variability, the PLS linear combinations of predictors are chosen to maximally summarize covariance with the response.

In other words, PLS finds components that maximally summarize the variation of the predictors while simultaneously requiring these components to have maximum correlation with the response. PLS therefore strikes a compromise between the objectives of predictor space dimension reduction and a predictive relationship with the response. In other words, PLS can be viewed as a *supervised* dimension reduction procedure. Projection Pursuit regression further enhances this approach providing a joint estimation of both dimensionality reduction and effect of the dimensions derived from the analysis.

Instance-based learning

Instance-based learning uses historical data to classify a new instance of a problem in a predefined set of classes. Instance-based Algorithms model a decision problem with instances or examples of training data that are deemed important or required by the model. Such methods typically build up a database of example data and compare new data to the database using a similarity measure in order to find the best match and make a prediction. For this reason, instance-based methods are also called winner-take-all methods and memory-based learning. Focus is put on representation of the stored instances and similarity measures used between instances. The most popular instance-based algorithms are KNN, Learning Vector Quantization (LVQ), Self-Organizing Map (SOM), Locally Weighted Learning (LWL)

KNN algorithm is a non-parametric method useful for classification and regression problems. In both cases, the method is based on the assumption that the data points are in a metric space. The data can be scalars or possibly even multidimensional vectors. Since the points are related to each other in therm of distance, a metric has to be chosen. To this purpose, the most commonly used one is the Euclidean distance. Let x_i be an observation sample with p covariates $(x_{i1}, x_{i2}, \dots, x_{ip})$, and n as total sample size $(i = 1, 2, \dots, n)$ with p the total number of features $(j = 1, 2, \dots, p)$. The Euclidean distance between observation x_i and x_l $(l = 1, 2, \dots, n)$ is defined as:

$$d(\mathbf{x}_i, \mathbf{x}_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + (x_{ip} - x_{lp})^2} \tag{11}$$

The algorithm decides which of the points from the training set are similar enough to be considered when choosing the class to predict a new observation the k closest data points to the new observation, and to take the most common class among these or the average weighted by inverse distance in a regression problem. The procedure is ”lazy learning” algorithm not requiring explicit training or model. The choice of nearest neighbors k parameters is very important; k value is like a smoothing parameter. In fact small values of k are responsible for greater variances in predictions. Alternatively, setting k to a large value increases bias. The k parameters should be chose minimizing the probability of misclassification and have to be small enough so that the k nearest points are not very distant to the considered point. As is the case of smoothing parameters, k admits an optimal value with respect to the trade off between bias and variance. This parameter may be estimated using cross validation. The general idea is to divide the sample in g folds. Given a values of k, the KNN model is used to make predictions on the $g-$th subgroup of sample and evaluate the prediction error, that for a regression problem is mean square error (MSE) and for classification is the accuracy. The process is repeated for each fold of subsample, finally the error are averaged. The cross validation is repeated for each k, then is selected the k minimizing error.

Since the predictions are based on the assumption that similar elements in the space are those who are the closest to each other, it makes sense to introduce one weight related to the point k for each of its nearest neighbour:

$$W_{ik} = \frac{exp(-d(\mathbf{x}_i, \mathbf{x}_l))}{\sum_k(-d(\mathbf{x}_i, \mathbf{x}_l))} \tag{12}$$

Considering these weights the estimated values are:

$$y_i = \sum_k W_{ik} y_{ik} \tag{13}$$

For a classification problems, the maximum of the equation is considered respect in k nearest classes.

LVQ The Learning Vector Quantization algorithm is a supervised neural network that uses a competitive (winner-take-all) learning strategy (Kohonen, n.d.). It is related to other supervised neural networks such as the Perceptron and the Back-propagation algorithm and to competitive learning neural networks such as the SOM algorithm (Kohonen, Barna, and Chrisley, n.d.).

LVQ is designed for those classification problems who have some existing data sets that can be used to supervise the learning process.

LVQ is non-parametric, thus it does not rely on assumptions about that structure of the function that it is approximating. Real-values in input vectors should be normalized such that $x \in (0, 1)$.

Euclidean distance:

$$\sum_{i=1}^{n}(x_i - c_i)^2$$

(where n is the number of attributes,x_i the input vectors and c_i the given instance based vectors) is commonly used to measure the distance between real-valued vectors, although other distance measures may be used (such as dot product), and data specific distance measures may be required for non-scalar attributes. There should be sufficient training iterations to expose all the training data to the model multiple times. The learning rate is typically linearly decayed over the training period from an initial value to close to zero. The more complex the class distribution, the more codebook vectors that will be required, some problems may need thousands. Multiple passes of the LVQ training algorithm are suggested for more robust usage, where the first pass has a large learning rate to prepare the codebook vectors and the second pass has a low learning rate and runs for a long time (perhaps 10-times more iterations).

LWL Locally Weighted Learning is a class of function approximation techniques, where a prediction is done by using an approximated local model around the current point of interest.

LWL is a non-parametric method and it is the classic approach to solve the function approximation problem locally (Atkeson, Moore, and Schaal, 1997).

The basic idea behind LWL is that instead of building a global model for the whole function space, for each point of interest a local model is created based on neighboring data of the query point, i.e the set of information to use to make a specific prediction. For this purpose each data point becomes a weighting factor which expresses the influence of the data point on the prediction. In general, the shorter the distance from a data point to the current query point, the higher the weight it receives. LWL is also called lazy learning, because the processing of the training data is shifted until a query point needs to be answered. This approach makes LWL a very accurate function approximation method where it is easy to add new training points.

Given a standard regression model $y = f(x) + \varepsilon$ where $f(x)$ is a continuous function, in order to approximate the $f(x)$ the general LWL solution methods try to find the coefficient β_q that minimizes the equation for the current point x_q

$$J = \frac{1}{2}\sum_{i=1}^{n} w_i(x_q)(y_i - x_i\beta_q)^2$$

where, (x_i, y_i) for $i = 1, 2, \dots, n$ represent the training data, where data point x_i has a corresponding output value y_i; and x_q is the current data point for which a prediction $\hat{y}_q$ is made. An important difference to global least square methods is that β_q depends on the current point x_q.

LWL computes the weights w_i into two separate steps:

- using a *distance function* to measure the relevance of training points for the current prediction. Typically, the Euclidean distance $\sqrt{(x-q)D(x-q)}$ with a distance metric D is used.
- computing a weight w_i, in terms of a kernel function (for example $K(d) = \exp(-d^2)$, for each distance value. The smoothness of the used kernel influences the smoothness of the output function.

LWL is called Memory-Based Learning, because all training data is kept in memory to calculate the prediction. The single steps of LWL are outlined in algorithm has a complexity of $O(n^2)$, where n is the size of the training dataset.

SOM Self-Organizing Map is a model very similar to neural networks (Kohonen and Somervuo, 1998) (Kohonen, 2012) and it can be used as an unsupervised, exploratory technique or in a supervised fashion for prediction (Melssen, Wehrens, and Buydens, 2006).

Self-organizing neural networks are used to cluster input patterns into groups of similar patterns. They're called *maps* because a topological structure among their cluster units and map weights to input data. Each weight is representative of a certain input. Input patterns are shown to all neurons simultaneously.

The structure of a SOM involves m cluster units, arranged in either a one- or two-dimensional array, with vectors of n input signals.

Like the brain, which organizes similar or related functions in distinct and interconnected anatomical locations, Kohonen networks group similar clusters in close proximity and dissimilar clusters at greater distances. Therefore, unlike other pattern recognition algorithms, the relative position of the clusters identified in a Kohonen network have additional value in that clusters that are relatively close share more similarities than those positioned at greater distance on the map.

Unsupervised learning is a means of modifying the weights of a neural network without specifying the desired output for any input patterns. The advantage is that it allows the network to find its own solution, making it more efficient with pattern association. The disadvantage is that other programs or users have to figure out how to interpret the output.

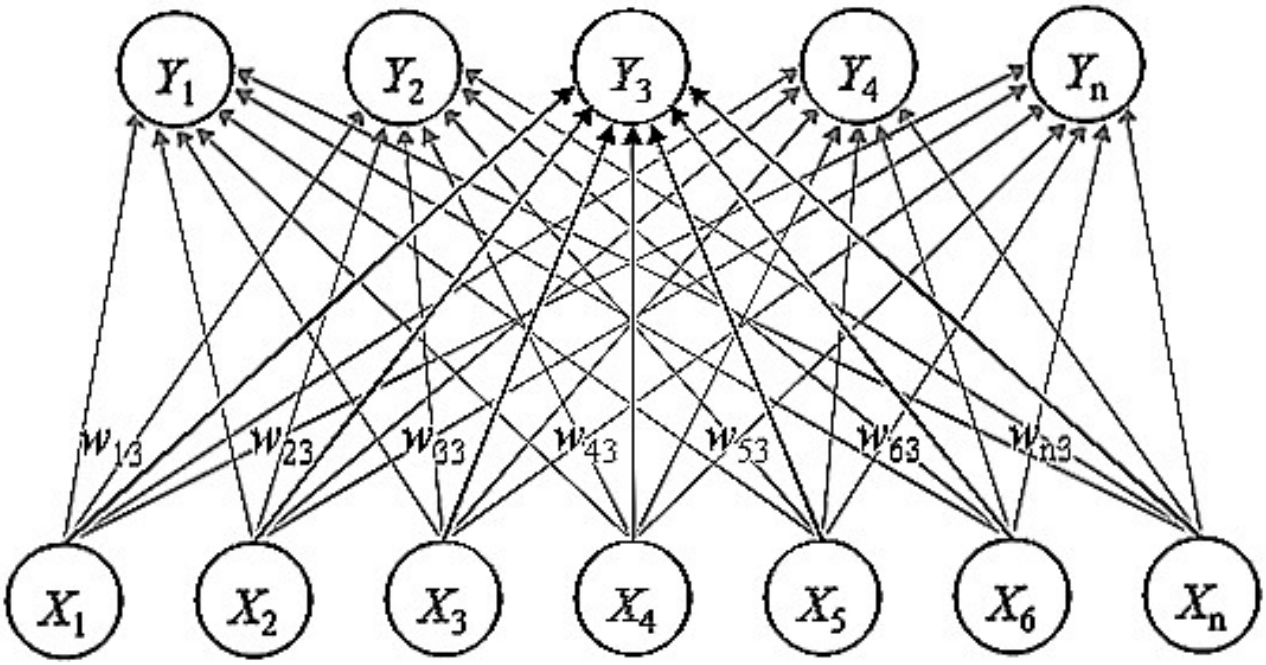


Figure 48: Self-organizing network with 5 cluster units Y_i and 7 input units X_i

The weight vectors define each cluster. Input patterns are compared to each cluster, and associated with the cluster it best matches. The comparison is usually based on the square of the minimum Euclidean distance. When a best match is found, the associated cluster gets its weights and its neighbouring units updated.

Weight vectors are arranged into lines or various grid structures. Some neighborhoods closer to the ends or edges will have smaller weights (Fausett, 1994).

Rule-based classifiers

Rule-based classifiers provide a set of classification rules that can be used later to evaluate a new case and classify it in a predefined set of classes.

Association rules are among the most popular representations for local patterns in data mining, by extracting rules that best explain observed relationships between variables in data.

The framework of association rules was originally developed for large sparse transaction data sets. The concept can be directly generalized to non-binary variables taking a finite number of values.

Algorithms for finding association rules find all rules satisfying the frequency and accuracy thresholds. If the frequency threshold is low, there might be many frequent sets and hence also many rules. One of the research challenges in using association rules for data mining is to develop methods for selecting potentially interesting rules from among the mass of discovered rules.

The rule frequency tells how often a rule is applicable. In many cases, rules with low frequency are not interesting, and this assumption is indeed built into the definition of the association rule-finding problem. The accuracy of an association rule is not necessarily a very good indication of its interestingness.

The statistical significance of an association rule $A \Rightarrow B$ can be evaluated using standard statistical significance testing techniques to determine whether the estimated probability $p(B = 1|A = 1)$ differs from the estimated background probability of $B = 1$, and whether this difference would be likely to occur by chance. This is equivalent to testing whether $p(B = 1|A = 1)$ differs from $p(B = 1|A = 0)$.

Given an association rule $\vartheta \Rightarrow \varphi$, its *accuracy* $c(\vartheta \Rightarrow \varphi)$ (also sometimes referred to as the *confidence*) is the fraction of rows that satisfy φ among those rows that satisfy ϑ In terms of conditional probability notation, the empirical accuracy of an association rule can be viewed as a maximum likelihood (frequency-based) estimate of the conditional probability that φ is true, given that ϑ is true.

The most popular association rule learning algorithms are Apriori (Agrawal and Srikant, n.d.; Borgelt, n.d.) and ECLAT (Zaki et al., n.d.; Borgelt, n.d.).

Apriori algorithm It works by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. It uses a bottom up approach: i.e. frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. Then it stops when no further successful extensions are found.

ECLAT algorithm Opposite to Apriori algorithm, ECLAT algorithm generates frequent items only once.

Decision tree methods

Tree methods aims at constructing a model of decisions using actual values of attributes in the data. Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning. The most popular decision tree algorithms are Classification and Regression Tree (CART), Iterative Dichotomizer 3 (ID3), C4.5 and C5.0, Chi-squared Automatic Interaction Detection (CHAID), Decision Stump, M5 and Conditional Decision Trees.

CART Decision tree learning is a method commonly used in for classification and regression objective. The purpose is to create a predictive model for a response variable based on several explanatory variables. Classification and Regression Tree are commonly used to predict a response or a class Y considering some covariates $X_1, X_2, ..., X_n$. If Y is a continuous outcome it is defined as a regression tree, while, if the response is categorical, it is called a classification tree.

Considering a classification problem, in this case each element of the outcome variable is a class. In a classification tree each internal node is labeled with an explanatory variable or feature. The tree's leaf is indicated considering a class or a probability distribution over the class of the response variable.

A tree comes from a recursive partition based on an attribute. This process is repeated on each derived subset. In order to stop the recursion, a stopping rule may be defined: the procedure halts when the subset corresponding to a node shares the same values of the explanatory covariates or when an additional split adds no values to predictions. This process is known as a top-down greedy algorithm.

The greedy algorithm chooses, at each step, a variable optimizing the homogeneity in the subset. There are different measures to define the optimal subset split based on the minimum heterogeneity:

- Gini impurity index was proposed for decision trees by (Leo Breiman et al., 1984). The Gini Index, as originally defined, measures the probability of misclassification of a set of instances. It is minimum when equal to zero: this happens when all cases, in the terminal node, are classified in only one of m category (Minimum Heterogeneity).

$$\sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i=1}^{m}(f_i - {f_i}^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} {f_i}^2 = 1 - \sum_{i=1}^{m} {f_i}^2 = \sum_{i \neq k} f_i f_k \tag{14}$$

- The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (in other word the most homogeneous branches).

$$I_E(f) = -\sum_{i=1}^{m} f_i \log_2 f_i \tag{15}$$

- Variance reduction is used for regression tree, specifically for continuous outcome, is defined as the total reduction of the variance of a considered variable consequently to splitting node.

Decision trees algorithm may be useful because it easy to interpret its results and lead to manage big data entries. The problem of the method is in overfitting: in fact greater complex trees do not generalize well the predictions to the test data. Some pruning procedures have been introduced to overcome overfitting problem or some alternative version of the algorithm, like CTREE, that does not require pruning.

C4.5 and C5.0 Another approach for classification trees is the C4.5 model (J Ross Quinlan, 2014). The main difference with CART is how the splitting criteria is carried out. For two class data, based on (Shannon, 1949), an information statistics, which represents the information content of the data prior to the splitting, is defined as

$$-\frac{n_{1+}}{n} \times log_2 \frac{n_{1+}}{n} - \frac{n_{2+}}{n} \times log_2 \frac{n_{2+}}{n}$$

where $\frac{n_{1+}}{n}$ is the probability of class 1.

The information after the split would be the sum of the information values from each of the resulting partitions, and the total information after the split is a weighted average of these values where the weights are related to the number of samples in the leaves of the split.

For continuous predictors, a tree could be constructed by searching for the predictor and single split that maximizes the information gain.

C5.0 is a more advanced version of Quinlan's C4.5 classification model (J Ross Quinlan, 2014), introducing additional features, such as boosting and unequal costs for different types of errors. C5.0 combines non occurring conditions for splits with several categories and it also conducts a final global pruning procedure that attempts to remove the sub-trees with a cost-complexity approach.

These kind of improvements typically allow to generate smaller trees.

M5 One limitation of the basic regression trees is that each terminal node uses the average of the training set outcomes in that node for prediction. As a consequence, these models may not do a good job when facing samples whose true outcomes are extremely high or low.

One approach to dealing with this issue is to use a different estimator in the terminal nodes. M5 algorithms is the model tree approach described by (John R Quinlan, n.d.), which is similar to regression trees except:

- the splitting criterion is different;
- the terminal nodes predict the outcome using a linear model (as opposed to the simple average);
- when a sample is predicted, it is often a combination of the predictions from different models along the same path through the tree.

Like simple regression trees, the initial split is found using an exhaustive search over the predictors and training set samples, but the expected reduction in the node's error rate is used. If $S_1, \ldots, S_P$ are the P subsets of the data after splitting, the split criterion would be

$$SD(S_1, \ldots, S_P) - \sum_{i=1}^{P} \frac{n_i}{n} \times SD(S_i)$$

where SD is the standard deviation and n_i is the number of sample partitions i. This metric determines if the total variation in the splits, weighted by sample size, is lower than in the pre split data.

The split that is associated with the largest reduction in error is chosen and a linear model is created within the partitions using the split variable in the model and the process is repeated until there are no further improvements. Once the tree is fully grown, there is a linear model for every node in the tree.

Once the complete set of linear models have been created, each of them undergoes a simplification procedure to potentially drop some of the terms. For a given model, an adjusted error rate is computed. First, the absolute differences between the observed and predicted data are calculated and then multiplied by a term that penalizes models with large numbers of parameters

$$\frac{n^* + p}{n^* - p} \sum_{i=1}^{n*} |y_i - \hat{y}_i|$$

where n^* is the number of training set data points that were used to build the model and p is the number of parameters.

Finally, M5 incorporates usually some smoothing to decrease the potential for over-fitting. The technique is based on the *recursive shrinking* methodology of Hastie and Pregibon (Hastie and Pregibon, 1990). When predicting, the new sample goes down the appropriate path of the tree, and moving from the bottom up, the linear models along that path are combined. These two prediction are put together by

$$\hat{y}_{(p)} = \frac{n_{(k)}\hat{y}_{(k)} + c\hat{y}_{(p)}}{n_{(k)} + c}$$

where $\hat{y}_{(k)}$ is the prediction from the child node, $n_{(k)}$ is the number of training set data points in the child node, $\hat{y}_{(p)}$ is the prediction from the parent node, and c is a constant with a default value of 15.

This kind of smoothing can have a significant positive effect on the model.

Once the tree is fully grown, it is pruned back by finding inadequate subtrees and removing them. Starting at the terminal nodes, the adjusted error rate with and without the sub-tree is computed. If the sub-tree does not decrease the adjusted error rate, it is pruned from the model. This process is continued until no more sub-trees can be removed.

Conditional Decision Trees Recursive partitioned trees and Conditional Decision Trees, recursively perform univariate splits of the dependent variable based on values assumed by a set of covariates. A remarkable summary of the main features for the decision tree, to be examined deeper in the following paragraphs is in Table 77. While pruning procedures are able to overcome the overfitting problem, the variable selection bias affects the interpretability. Conditional Inference Trees, may indeed be useful against selection bias problems to select variables that have many possible splits or many missing values. Both the response variable and the dependent variables may be measured at any arbitrary scale. The conditional distribution of the response variable given the covariates depends on a function of the explanatory covariates expressed as $D(Y|X) = D(Y|X_1, \ldots, X_m) = D(Y|f(X_1, \ldots, X_m))$ Considering a sample of N observations an algorithm can be formulated weighting the cases with $w = (w_1, \ldots, w_n)$. The node elements may be represented considering the non-zero weights component if the observation is in the node and zero otherwise. The algorithm start testing for case weights hypothesis of independence between covariates and the outcome variable. The procedure is stopped if the hypothesis is rejected for a specific alpha value. In other cases, the covariate more associated with the response variable is selected. From the values $B \subset X_j$, is chosen to split X_j, into two disjoint sets. The original case w_{left} and w_{right} are splitted determining two subgroups with $w_{left,i} = w_i I(X_{ji} \in B)$ and $w_{right,i} = w_i I(X_{ji} \notin B)$ where $I(\cdot)$ denotes the indicator function, defining membership of an element in a specific subset. The procedure is repeated recursively changing weight. The procedures employees a significance test to select variables, instead of selecting the variable that maximizes an information measure (like Gini coefficient); the multiple significance tests, computed at each start of the algorithm, are permutation tests.

Ensemble methods

Ensemble methods are models composed by multiple weaker models that are independently trained and whose predictions are combined in some way to make the overall prediction. Much effort is put into what types of weak learners to combine and the ways in which to combine them. This is a very powerful class of techniques and as such is very popular. They include Boosting, Bootstrapped Aggregation (Bagging), AdaBoost, Stacked Generalization (blending), Gradient Boosting Machines (GBM), Gradient Boosting Regression Tree (GBRT) and RF

RF (L. Breiman, 2001) are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.

RF algorithm is an ensemble method based on the computation of many decision trees at training time, obtaining a prediction class that is the mode of the classes for each tree (classification) or the mean result of the prediction (regression). This technique seems to correct the problem of overfitting on training set. Each tree, singularly, overfit their training sets, because they have low bias, but very high variance. RF averaging multiple trees built on different parts of the same training set induce a loss in variance but increase the bias component and, in the same time, improve the performance of the predictive model. The algorithm is based on **bagging**, that, basically, draws a replacement random sample for B times on the training estimating trees on each sample; final predictions are made on average of the predictions from all the singular trees for regression or on greater vote in case of classification trees.

The RF algorithm is based on the same procedure of the bagging algorithm but, in addition, it uses a random subset of the covariates (feature bagging) for each split of growing trees. The feature bagging may be useful because if some predictors are related with the response variable, these features may be selected in the greater part of generated trees, so they would be correlated. For a classification problem, some authors, suggests to use $\sqrt{p}$ covariates to consider in each node split, while, for a regression problem, they recommend to consider $p/3$ covariates and a minimum split node size of 5.

The RF algorithm consider the relative importance of the predictors by using the variable importance measures; first it is estimated a RF on the data and, for each observation is computed the out-of-bag error and is averaged considering each tree in the forest. The importance of the m variable is measured permuting the variable on training set and the out-of-bag error on the data set obtained with permutation. Averaging the differences of out-of-bag error pre and post permutation it is possible to obtain an importance measure normalized respect to its standard deviation. The variables are ranked according to the variable importance computed.

The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to AdaBoost, but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

Boosting Trees A boosting procedure, specifically a Schapire's AdaBoost.M1 Algorithm, does not draw a succession of independent bootstrap samples. Instead, it weights each observation weight for each instance: a higher weight indicates more influences in the classification procedure. At each step of the procedure the weights are adjusted reflecting the relative performance, increasing the weights of misclassified observation. The obtained final classifiers using votes aggregates each classifiers and the votes are related to classificator's accuracy. The data modifications at each so-called boosting iteration consist of applying weights $w_1, w_2, ..., w_N$ to each of the training samples, initially all weights are $w_i = 1/N$, so that the first step simply trains a weak learner on the original data. For each successive iteration, the sample weights are individually modified and the learning algorithm is reapplied to the reweighted data. At a given step, those training examples that were incorrectly predicted by the boosted model induced at the previous step have their weights increased, whereas the weights are decreased for those that were predicted correctly. As iterations proceed, examples that are difficult to predict receive ever-increasing influence. Each subsequent weak learner is thereby forced to concentrate on the examples that are missed by the previous in the sequence.

Specifically, if w_i^t denotes the weight of observation i at trial t, and at the first step $w_i^1 = 1/N$. At each step the classifier C^t for $t = 1, ..., T$ is computed considering the weights w^t. The classification error e^t is calculated as a sum of the weights of misclassified observation for each classifier at t step. If the classification error is greater than 0.5 the procedure is stopped and the classifiers becomes $t-1$. If the classification error is 0, the procedure is stopped and the classifiers are $T = t$. Otherwise the procedure generates the weight for next classifiers w^{t+1} multiplying the weights observation not misclassified for a normalization factor $\beta^t = e^t/(1-e^t)$, leading the weights sum to one. The final classifier is obtained summing the votes of single classifier that is equal to $log(1/\beta^t)$.

Bayesian Regression Trees (BART) BART is a statistical sum of trees model (Chipman, George, and McCulloch, 2010). It can be considered a Bayesian version of ML tree ensemble methods where the individual trees are the base learners.

For datasets where the number of variables p is large (e.g. $p > 5000$) the algorithm can become prohibitively computationally expensive.

SVMs

SVM is a supervised learning method useful for classification and regression analysis. SVM provides discriminant functions to distinguish between two predefined classes that can be non-linearly separable. The method is based on the construction of one or more hyperplanes, in high or infinite dimension, leading to a separation in the data according to the output variable. For this purpose, the better hyperplane has the largest functional margin (distance to the nearest data belonging to a class): the greater the margin, the lower the classification error.

Formally an hyperplane is given by

$$\beta_0 + \beta^T x \tag{16}$$

The optimal hyperplane can be represented in an infinite number of ways scaling of β and β_0. The chosen representation is

$$|\beta_0 + \beta^T x| = 1 \tag{17}$$

where x is the training observation closest to the hyperplane they are the *support vectors*. This representation is the *canonical hyperplane*.

Geometrically the distance d between a point and the hyperplane is given by (β, β_0):

$$\text{distance} = \frac{|\beta_0 + \beta^T x|}{||\beta||} \tag{18}$$

In a canonical Hyperplane the distance may be expressed as:

$$\frac{|\beta_0 + \beta^T x|}{||\beta||} \tag{19}$$

The margin is defined M, is twice the distance to the closest observation:

$$M = \frac{2}{||\beta||} \tag{20}$$

Now the optimal hyperplane, which is the one that maximizes M, is like to minimize a function $G(\beta)$ under some conditions:

$$\min_{\beta,\beta_0} G(\beta) = \frac{1}{2}||\beta||^2 \text{ subject to } y_i(\beta^T x_i + \beta_0) \geq 1 \; \forall i \tag{21}$$

where y_i is a class of the training observation.

The Lagrange optimization can be obtained using Lagrange multipliers: in this way the weight β and the bias β_0 of the minimum margin hyperplane are obtained.

Neural Networks

ANNs are models inspired by the structure and/or function of biological neural networks. A huge variety of pattern-matching algorithms falls into this category: they actually form an enormous subfield, and are commonly used for regression and classification problems of any kind.Artificial neural networks models allow to predict the value of one or more variables for a new instance on the basis of non-linear combination of the values of several input variables and intermediary layers. Among ANN, Deep Learning methods are a modern update to ANN that exploit abundant cheap computation. They are concerned with building much larger and more complex neural networks, often in the framework of semi-supervised learning problems where large datasets contain very little labelled data. The most popular deep learning algorithms are Deep Boltzmann Machine (DBM), Deep Belief Networks (DBN), Convolutional Neural Network (CNN) and Stacked Auto-Encoders

The most popular ANN algorithms are (i) Perceptron, (ii) back–propagation, (iii) Hopefield Network and RBFN.

Perceptron MLP with one hidden layer may be considered as a logistic regression where the covariates are preprocessed using a non-linear transformation Φ. Projecting the input data in a space (intermediate hidden layer) where it becomes linearly separable. In much more cases one hidden layer is sufficient to approximate the input relation.

The MLP is a function projecting the points in covariates space of D dimension to a subspace of dimension L as size output vector $f(x)$ $f : R^D \rightarrow R^L$:

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x))) \tag{22}$$

Where the matrices $W^{(1)}, W^{(2)}$ are the weights component and $b^{(1)}, b^{(2)}$ the bias component with activation functions G and s

The hidden layer is $h(x) = \Phi(x) = s(b^{(1)} + W^{(1)}x)$, to connect the input to the hidden layer is possible to use $W^{(1)} \in R^{D \times D_h}$ weight matrix connecting the input vector to the hidden layer. The activation function s may be $tanh$, with $tanh(a) = (e^a - e^{-a})/(e^a + e^{-a})$, or for binary classification, a the logistic function $sigmoid(a) = 1/(1 + e^{-a})$. In this case outputs larger or equal to 0.5 are assigned to the first class otherwise to second class.

In the case of more than two classes, the `softmax` function is used, which is written as:

$$\text{softmax}(x)_i = \frac{\text{exp}(x_i)}{\sum_{l=1}^{k} \text{exp}(z_l)} \tag{23}$$

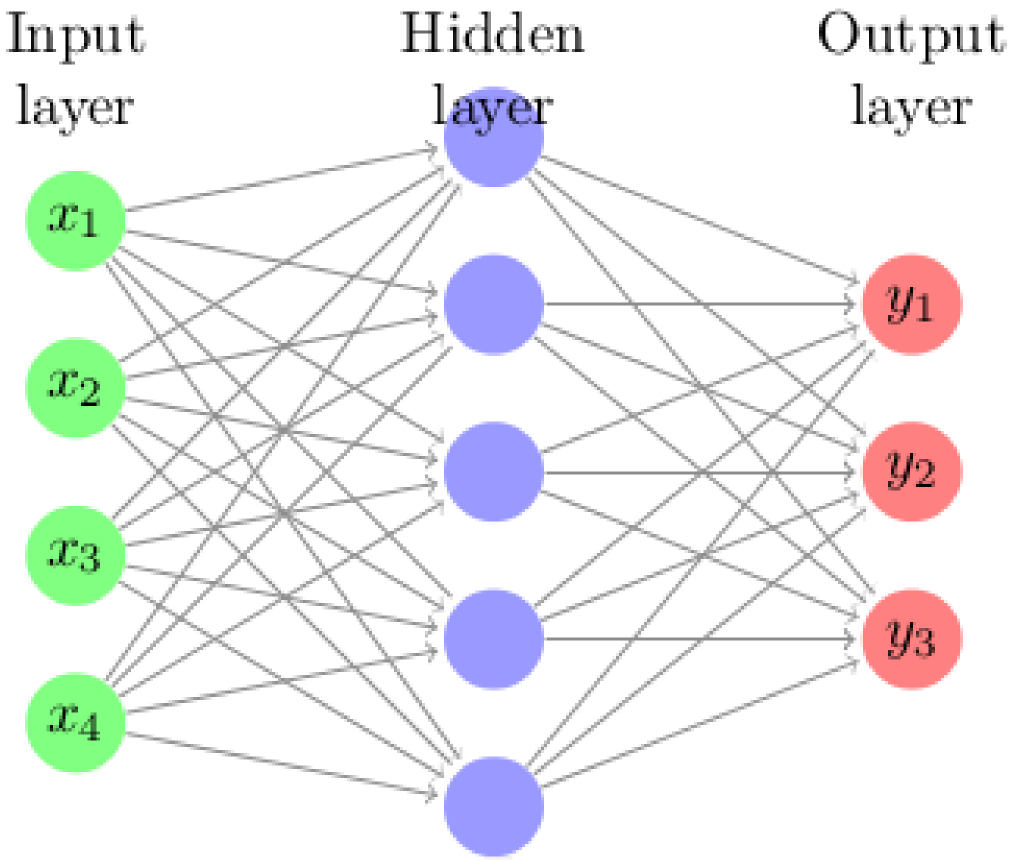


Figure 49: MLP with one hidden layer

The information in the network is stored in the weights, the back–propagation method is used to learn the method:

- First is chosen the architecture for the network, which will contain input, hidden and output units, all of which will contain sigmoid functions.
- The weights between all the nodes (generally small numbers between -0.5 and 0.5) are randomly assigned.
- Each training is used to redefine the weights component.
- Different initial random weight lead to converge to different local minimum doesn't finding an unique point minimizing error to stop the algorithm. Some authors, as Mitchell suggests to learn different networks averaging the results. Alternative are proposed in order to overcome the local minima problem.

A combination of multiple classifiers may increase the overall predictive accuracy. For this reason, computing an Ensemble MLP may be more useful than considering a single classifier or regression method. In general, an ensemble is built in two steps: first, multiple individual classifiers are trained; then, they are combined using average for regression or votes for classification. Some algorithm implements the procedure sequentially, for example AdaBoost. Other ensemble algorithm parallelizes the procedure like Bagging algorithm.

Back–Propagation (Rumelhart, Hinton, and R. J. Williams, 1986) is a highly efficient methodology that works with derivatives to find the optimal parameters.

The back–propagation algorithm looks for the minimum of the error function in weight space using the gradient descent. The combination of weights which minimizes the error function is considered to be a solution of the learning problem. Since this method requires computation of the gradient of the error function at each iteration step, the assumption of continuity and differentiability of the error function is required.

While perceptrons use step functions as activation functions, back–propagation networks use the sigmoid:

$$f(x) = \frac{1}{1 + e^{-cx}}$$

The constant c can be selected arbitrarily and its reciprocal is called the temperature parameter in stochastic ANN. Higher values of c bring the shape of the sigmoid closer to that of the step function and in the limit $c \to \infty$ the sigmoid converges to a step function at the origin.

Hopefield Network consists of a set of interconnected neurons, i.e. where each neuron is connected to every others but not to itself and the connection strengths or weights are symmetric in that the weight from node i to node j is the same as that from node j to node i.

The connections are weighted and depends on the sign of the weight they can be intercepting or activating; e.g. when a neuron become active, then also all neurons which are connected to it with a positive weight become active. There is a threshold value for every neuron which the sum of the input values must reach to produce activity.

At the beginning of the calculation of the network output, the neuron's activation corresponds to the pattern to recognize. Then the network is iterated, which means that the state of the neurons is recalculated until the network is stable, i.e. the network state doesn't change any more. This is possible in a finite amount of time and iterations for Hopfield networks. This can also be seen as the minimization of the energy in the net, so that the final state is a minimum of an energy function called *attractor*.

*Radial Basis Function Network (*RBFN*)* is embedded in a two layer neural network, where each hidden unit employs a radial activated function. The output units apply a weighted sum of hidden unit outputs. The input into an RBFN is nonlinear while the output is linear. RBFN have a static Gaussian function as the nonlinearity for the hidden layer processing elements. The Gaussian function responds only to a small region of the input space where the Gaussian is centered. The key to a successful implementation of these networks is to find suitable centers for the Gaussian functions.

In this case, the k-means clustering algorithm is used to derive the Gaussian centers and the widths from the input data. These centers are encoded within the weights of the unsupervised layer using competitive learning. During the unsupervised learning, the widths of the Gaussians are computed based on the centers of their neighbors. The output of this layer is derived from the input data weighted by a Gaussian mixture.

Once the unsupervised layer has completed its training, the supervised segment then sets the centers of Gaussian functions (based on the weights of the unsupervised layer) and determines the width (standard deviation) of each Gaussian. RBFN has been successfully applied in astronomy, for example, solar flare prediction (Qahwaji and Colak, 2007), stellar spectra classification (L. Zhang and Bai, 2005), separation of stars and galaxies (Joe Qin, 2003).

Bayesian methods

Methods that are explicitly applying Bayes' Theorem for problems such as classification and regression. The most popular Bayesian algorithms are Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Averaged One-Dependence Estimators (AODE), Bayesian Belief Network (BBN) and Bayesian Network (BN). The latter provides graphical interpretation of causal relationships between variables together with conditional probabilities

Glsnb NB is the simplest Bayesian classifier (Domingos and Pazzani, 1997). It is built upon the assumption of conditional independence of the predictive variables given the class. Although this assumption is violated in numerous occasions in real domains, the paradigm performs well in many situations (Berchialla, Foltran, and Gregori, 2013). The most probable a posteriori assignment of the class variable is calculated as

$$arg \max_c p(c|x_1, \ldots, x_n) = arg \max_c p(c) \Pi_{i=1}^{n} p(x_i|c)$$

A slightly improvement of NB is the Tree Augmented Naïve Bayes (TAN) (N. Friedman, Geiger, and Goldszmidt, 1997) classifier, which is able to take into account relationships between the predictive variables by extending a NB structure with a tree structure among the predictive variables. This tree structure can be obtained adapting the algorithm proposed by (Chow and Liu, 1968) and calculating the conditional mutual information for each pair of predictive variables, given the class. The TAN classification model is limited by the number of parents of the predictive variables. A predictive variable can have a maximum of two parents: the class and another predictive variable. The k dependence Bayesian (KDB) classifier (Sahami, n.d.) avoids this restriction by allowing a predictive variable to have up to k parents aside from the class.

AODE is considered an improvement on the NB and an interesting alternative to other semi-naive approaches. It provides a good trade-off between efficiency and performance. To maintain efficiency, the AODE is restricted to the exclusive use of one estimator for the dependency relationship. Specifically, the AODE can be considered as an ensemble of SuperParent One-Dependence Estimators (SPODEs) because every attribute depends on the class and another shared attribute, which is designated as the super-parent

BN is a graphical representation of the joint probability distributions over a set of random variables. It consists of a series of nodes, representing variables connected by arrows, or directed arcs, forming a graph that has no cycles. The relationships in the networks are usually described as in human genealogies. So, for example, a parent-child relationship (X, Y) is present when there is an arrow from node X (parent) to node Y (child). The arcs specify the probabilistic relationships that hold between nodes. In general, there may be many arcs going into and out of each node, creating a complex network. The most important restriction is that the arcs must not create cycles within the network (Nielsen and Jensen, 2009). On the contrary, the absence of any direct arc between two variables X and Y points out their marginal independence, i.e. conditional probability of Y given X is equal to the probability of Y and vice-versa; however the two variables become dependent if they have a common child. Each node of the network is associated with a set of probability tables. For nodes without ingoing arcs, the probability distribution is a prior distribution which requires supplying a set of initial values. For variables with *parents*, each entry in the tables contains a conditional probability for that variable being in a specific state, given a specific configuration of the states of its parents. Both the structure and the numerical parameters of a BN can be learned entirely from data (Cooper and Herskovits, 1992)

BART is a statistical sum of Bayesian CART models (Chipman, George, and McCulloch, 2010). It can be viewed as the Bayesian version of ML tree ensemble methods where the individual trees are the base learners. It consists of two part: a sum-of-trees model and a regularization prior on the parameters of that model:

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2) \tag{24}$$

where to each binary regression tree T_j, which is made up of a set of interior node decision rules and a set of terminal nodes, is associated a terminal node parameter M_j. T_j is a binary tree in the sense that the decision rules are binary splits over the predictor space, i.e. they are of the form $\{x \in A\}$ vs $\{x \notin A\}$ for categorical variable Xs and of the form $\{x \leq c\}$ vs $\{x > c\}$ for continuous variables. The conditional mean $E(Y|x)$ equals the sum of all the terminal node assigned to x by the tree $g(x; T_j, M_j)$. This implies that the sum-of-trees model can incorporate both main effects and interaction effects; since Equation 24 may be based on trees of varying sizes, the interaction effects may be of varying orders. With a large number of trees, a sum-of-trees model gains increased representation flexibility with good predictive capabilities. However, for datasets where the number of variables is large, typically, when greater than 5,000, the algorithm can get computationally very expensive. To complete the specification of the BART model, a *regularization prior* is imposed over all the parameters of the sum of trees model $(T_1, M_1), \ldots, (T_m, M_m)$ and σ. To facilitate the prior specification, it is recommended (Chipman, George, and McCulloch, 2010) to reduce the e prior formulation problem to the specification of just a few interpretable hyper-parameters which govern the prior probabilities on T_j, M_j and σ.

Regularization Algorithms

An extension made to another method (typically regression methods) that penalizes models based on their complexity, favoring simpler models that are also better at generalizing. They are generally slight modification made to other methods. The most popular Regularization Algorithms (REGULA) are (i) Ridge Regression, (ii) LASSO, (iii) Elastic Net and (iv) Least-Angle Regression (LARS).

Ridge Regression When the model over–fits the data, or when there are issues with collinearity, the linear regression parameter estimates may become inflated. Controlling or regularizing these parameter estimates reduces the sum-of-squared errors (SSE)

$$SSE = \sum_{i=1}^{n}(y_i + \hat{y}_i)^2$$

Ridge regression ((Hoerl and Kennard, 1970)) tries to do this by adding a penalty to the SSE if the estimates become large.

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i + \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2 \tag{25}$$

where L_2 means that a second—order penalty (i.e., the square) is being used on the parameter estimates.

The effect of this penalty is that the parameter estimates are only allowed to become large if there is a proportional reduction in SSE. In effect, this method shrinks the estimates towards 0 as the λ penalty becomes large. These techniques are sometimes called *shrinkage methods*.

While ridge regression shrinks the parameter estimates towards 0, the model does not set the values to absolute 0 for any value of the penalty. Even though some parameter estimates become negligibly small, this model does not carry out any kind of *feature selection*.

LASSO Initially proposed by Tibshirani (Robert Tibshirani, 1996), it has widely developed over the years.

Give a set of input measurements $x_1, x_2, \cdots, x_p$ and an outcome measurement y, the LASSO fits a linear model

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

The criterion it uses is to minimize the $\sum(y - \hat{y})^2$ subject to $\sum|b_j| \leq s$.

The first sum is taken over observations (cases) in the dataset. The bound ”s” is a tuning parameter. When ”s” is large enough, the constraint has no effect and the solution is just the usual multiple linear least squares regression of y on $x_1, x_2, \cdots x_p$. This is equivalent as setting the parameter λ to zero in Equation 25.

However for smaller values of s $(s \geq 0)$ the solutions are shrunken versions of the least squares estimates. Often, some of the coefficients b_j are zero. Choosing ”s” is like choosing the number of predictors to use in a regression model, and cross-validation is a good tool for estimating the best value for ”s”.

The computation of the LASSO solutions is a quadratic programming problem, and can be tackled by standard numerical analysis algorithms. Tibshirani (Robert Tibshirani, 1996) proposes standardizing each regressor so that it has (sample) mean zero and (sample) variance one and standardizing the dependent variable to have mean zero. This standardization amounts to incorporating an intercept term that is orthogonal to all other regressors, not part of the penalty and estimated by the mean of the dependent variable. Essentially, this procedure shrinks the ordinary least squares towards zero, typically setting some of them to be equal to zero. Thus, it seems to behave as a compromise between subset selection and ridge regression and may therefore be a useful tool for variable selection. An improvement to the original algorithm was proposed by Osborne (Osborne, Presnell, and Turlach, 2000), which can still be applied if there are more regressors than observations (Osborne, Presnell, and Turlach, 1998).

Elastic Net is a generalization of the LASSO model (Zou and Hastie, 2005). It combines the two type of penalties L_1 and L_2

$$SSE_{Enet} = \sum_{i=1}^{n}(y_i + \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{P} \beta_j^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j|$$

The advantage of this model is that it enables effective regularization via the ridge-type penalty with the feature selection quality of the LASSO penalty. Both the penalties require tuning to achieve optimal performance.

Adaptive Lasso was introduced by Zou (Zou, 2006) for linear regression and by Zhang and Lu (H. H. Zhang and Lu, 2007) for proportional hazards regression.

For linear regression, the approach is to use the LARS after re-weighting the X matrix. As tuning parameter using the ”known variance” version of Bayesian Information Criteria (BIC) with full model MSE for estimating the error variance.

An important approximate adaptive LASSO approach for many types of regression modeling was proposed by Wang and Leng (H. Wang and Leng, 2008).

Small sample performances of LASSO and in general of Least Absolute Deviations are focused on the accuracy of model selection. Hurvich and Tsai (Hurvich and C. L. Tsai, 1990) developed a small sample criterion (*L1cAIC*) for the selection of least absolute deviations regression models. In contrast to AIC (Akaike, 1973), L1cAIC provides an exactly unbiased estimator for the expected Kullback—Leibler information, assuming that the errors have a double exponential distribution and the model is not under-fitted.

4.1.2 MLT applied properties

In Table 76 (Bhaskar, Hoyle, and Singh, 2006) and 77 (2007) a core information is provided on the main limits and characteristics of each MLT.

Table 76: Properties of machine learning classification and prediction algorithms.

Classification tools	Parameters	Linear (L)/non-linear	Effect of small sample(NL) /feature ratio	Computational complexity	Data assumptions	Noise and outlier effect	Transparency	Incremental learning
Multi Layer Perceptron ANN	High	NL	Medium	High	None	Low	Poor	Poor
RBFN ANN	High	NL	Medium	Medium	None	Low	Good	Poor
SOM	Medium	NL		Medium	None	Low	Poor	Poor
Probabilistic ANN	High	NL	Medium	Medium	None	Low	Good	Poor
SVM	Low	L/NL	Low	Medium	Variable	Low	Good	Medium
LDA	Low	L	Low	Low	Gaussian<comma> equal variance	Medium	Good	Medium
QDA	Low	NL	Low	Low	Gaussian unequal variance	Medium	Good	Medium
KNN	Low	NL	Low	High	None	Low	Good	Good
Gaussian mixture model	Medium	NL	High	High	Variable	High	Good	Poor
Naive bayes	Low	NL	High	Low	None	Low	Good	Poor
Decision trees	Low	NL	Medium	Medium	None	Low	Good	Poor
Neuro-fuzzy systems	Low	NL	Medium	High	None	Low	Good	Poor
Clustering tools								
SOM	Low	–	Medium	High	None	Low	Poor	Medium
k-means	Low	–	High	Medium	Spherical clusters	High	Poor	Good
Fuzzy c-means	Low	–	High	Medium	Spherical clusters	High	Poor	Good
Hierarchical clustering	Medium	–	High	Low	None	Low	Good	Good
Dimensionality reduction								
PCA	None	L	Low	Low	Gaussian densities	High	Good	Medium
LDA	Low	L	Low	Low	Gaussian densities	High	Good	Poor
Sammon’s mapping	Low	NL	Low	High	None	Medium	Poor	Poor
Multi-dimensional scaling	Low	NL	Low	Low	None	High	Good	Medium
Independent components analysis	Low	NL	Medium	Medium	Variable	High	Good	Medium

Table 77: Comparing learning algorithms (●●●● represent the best and ● the worst performance)

	Decision Trees	ANN	Naive Bayes	KNN	SVM	Rule-learners
Accuracy in general	●●	●●●	●	●●	●●●●	●●
Speed of learning with respect to number of attributes and the number of instances	●●●	●	●●●●	●●●●	●	●●
Speed of classification	●●●●	●●●●	●●●●	●	●●●●	●●●●
Tolerance to missing values	●●●	●	●●●●	●	●●	●●
Tolerance to irrelevant attributes	●●●	●	●●	●●	●●●●	●●
Tolerance to redundant attributes	●●	●●	●	●●	●●●	●●
Tolerance to highly interdependent attributes	●●	●●●	●	●	●●●	●●
Dealing with discrete, binary or continuous attributes	●●●●	●●●(not discrete)	●●●(not continuous)	●●●(not directly discrete)	●●(not discrete)	●●●(not directly continuous)
Tolerance to noise	●●	●●	●●●	●	●●	●
Dealing with danger of overfitting	●●	●	●●●	●●●	●●	●●
Attempts for incremental learning	●●	●●●	●●●●	●●●●	●●	●
Explanation ability/transparency of knowledge/classifications	●●●●	●	●●●●	●●	●	●●●●
Model parameter handling	●●●	●	●●●●	●●●	●	●●●

4.1.3 The building blocks of the decision tree

The process of taking a decision on the most suitable MLT is hardly unique. Most likely, it ensembles a complex interaction among a deep knowledge of the problem from a substantive point of view, the data structure and the final gain expected to be from the analysis. In this context, the decision was taken to privilege generality over precision. The main parameters taken into account by humans to choose the proper data mining technique in a real application are:

- The main goal of the problem to be solved
- The structure of the available data

Labels and singularity

The main purpose is to define the principal characteristics of the problem and those of the data. At this stage, two major information are necessary, one related to the availability of explicit information on labels, the second on the singularity of the outcome.

1. Labels
    a) Known
        i. Output
            A. Single
            B. Multiple
    b) Unknown
    c) Partially known

As derived from the analysis of opinions and topics, virtually no evidence of a use of multiple outcomes simultaneously in practical EFSA work is available. A noticeable exception is the opinion on GMO, raising the call for more research in this context (on Genetically Modified Organisms, 2010). Therefore, the importance of such an item is conceptual more than practical.

Type of outcomes

Outcome types in statistics and probability provide most information on the analyses and the approaches to be followed. In the MLT context, the approach is similar

1. General objects
2. Vector-based objects
    a) Continuous
    b) Binary
    c) Categorical
    d) Correlated
    e) Sequencing/time series

General objects are of limited interest for the purposes of EFSA. Data evaluated in opinions are quite structured and referred to setups where a classical analysis is foreseen. Regarding the vector-based objects, the most used in EFSA opinions are continuous and binaries, referring to the conceptual framework of the regression under the GLM umbrella. Time series data are quite often used, for the purpose of detecting outbreaks or to derive an association with some explanatory phenomena in terms of shared trends (on Plant Health, 2009b). Extending the concept of sequencing beyond the time domain toward correlation in data, several opinions deal with correlated data (both longitudinally and spatially, addressed by random effects (EFSA Panel on Dietetic Products, Nutrition and Allergies, 2005b) or generalized estimating equations models (on Contaminants in the Food Chain, 2009b).

Type of inputs

Input and output types are basically the same, with the further notice that general, non-vector objects, are even less used in EFSA.

1. General objects
2. Vector-based objects
   a) Continuous
   b) Binary
   c) Categorical
   d) Correlated
   e) Sequencing/time series

As a general comment, most models used by EFSA are accepting any kind of vector-type inputs. Limitations are represented by the most commonly used models for NOAEL or BMDL, requiring a continuous input, and time-series models for trends, requiring usually continuous, time trends data. Collinearity in the input/feature matrix is also a concern for some techniques, although not very common in EFSA activities (Authority, 2010).

Sparsity and missing data

Sparse data matrices (i.e.: a matrix in which most elements are zero) is an important part of data modeling (DiMaggio et al., 2010). The concept of sparsity is useful in ML and network theory, when it is interpreted as a low density of significant data or connections. Sparse data is a common problem in microbial resistance (on Biological Hazards, 2008), contaminants (on Contaminants in the Food Chain, 2008; on Contaminants in the Food Chain, 2005a; on Contaminants in the Food Chain, 2009c) or welfare risk (E. A. P. (P. on Animal Health and Welfare), 2006; Authority, 2010; E. P. on Animal Health and Welfare, 2009; E. P. on Animal Health and Welfare, 2004). Missing data is a closely related situation from the modeling point of view, and it is very common in EFSA work shared by several topics, having received attention also for the methodological aspects (E. P. on Animal Health and Welfare, 2007a; E. P. on Animal Health and Welfare, 2007d; E. P. on Animal Health and Welfare, 2007c; E. P. on Animal Health and Welfare, 2007b; Authority, 2010; Ingredients and Packaging), 2010b; on Genetically Modified Organisms, 2010; (Assessment and methodological support), 2009; on Plant Health, 2009a; on Biological Hazards, 2009; on Plant Protection Products and their Residues, 2008a; Ingredients and Packaging), 2008). MLT interact with the missing data problem both as a viable medium for imputing missing data and from the other side, as any other classical technique, being able or not to deal with missing data in the matrix.

Linear separability

Although not always and directly appreciable, linear separability, i.e.: the possibility to discriminate among two ore more sets of data points via a linear function (lien or hyperplane), is a property of data, and thus under the major control of the investigator. Basically, a data point is viewed as a p-dimensional vector, and the aim is to know whether such points can be separated by a linear classifier with a $(p-1)$-dimensional hyperplane. The investigator can obtain a certain degree of information on the linear separability of her/his data by scatter plots, both bi-and multidimensional, inspection. Noticeably, but more complex, is always possibly to use linear methods for non-linear problems by virtually adding additional dimensions to make a non-linear problem linearly separable (this is the so-called kernel trick, which basically consists in simply computing the inner products between the images of all pairs of data in the feature space).

1. Linear
2. Non-linear

Most basic linear classifiers have been used by EFSA, relying on more additional assumptions on the covariance matrix (E. P. on Animal Health and Welfare, 2008; Ingredients and Packaging), 2010a).

Data characteristics

Classical statistical problems in data, like (heavy) skewness, outliers, collinearity in the feature matrix affect also MLT in terms of robustness of their finding. Skewness is in particular very common in characterizing probability distributions of hazards (on Plant Protection Products and their Residues, 2005; on Biological Hazards, 2005; on Plant Protection Products and their Residues, 2008b). Commonly associated with skewness are outliers, a transversal problem in EFSA opinions (on Plant Protection Products and their Residues, 2006; on Additives and or Substances used in Animal Feed, 2005; on Contaminants in the Food Chain, 2005b; EFSA Panel on Dietetic Products, Nutrition and Allergies, 2005a; on Contaminants in the Food Chain, 2009a; Ingredients and Packaging), 2009).

1. Skewness
2. Outliers

Management and logistic issues

Three main aspects are under the control of investigator, having major impact on how the MLT process would managed and what is practically expected from it. They are scalability, computational complexity and sample size/dimensionality. Scalable ML occurs when statistics, systems, ML and Data Mining (DM) techniques are combined into flexible, often nonparametric, and scalable techniques for analyzing large amounts of data at internet scale. Computational complexity is basically the time necessary to perform the MLT analysis. Beside the obvious considerations about computer infrastructure is a function of sample size and dimensionality, for the latter being intended the number of features considered in the analysis.

In terms of sample size, the proposed distinction goes between very small and very high sample size. The first are usually related to experimental settings, of the order of 10-50 observations, the latter exceeding thousands of data. Everything in between is considered as medium and its computational impact depends heavily on the number of features considered. In terms of dimensionality, the major issue is to understand if the number of features p exceeds or not the number of observations n.

Computational complexity is usually a function of both. As an example, RF have a computing requirement factor which is increasing by $O(np\log(n))$. In practice, sparseness of solutions is closely related to the dimensionality. Sparse machine learning refers to a collection of methods to learning that seek a trade-off between some goodness-of-fit measure and sparsity of the result, the latter property allowing better interpretability. In classification task for instance, the aim is to provide not only a high-performance classifier, but one that only involves a few features, allowing researchers to focus their research efforts on those. There is an extensive literature on the topic of sparse machine learning, with terms such as compressed sensing (Donoho, 2006; Candès and Plan, 2009), $L1$-norm penalties and convex optimization (Tropp, 2006), often associated with the topic.

Expected results

Closely related to the management aspects is the issue of what kind of outcome the investigator is looking for. Indeed, in particular in classification, but not restricted to, there might be the need of having a good performance of the ML model, possibly for making prediction, or of having a sort of interpretable effects out of the model. The latter case is often less restrictive than expected, provided the model is parametric and somehow associated with differentiable probability densities or moments. Indeed, in all cases where a ”nice” and simple solution, like e.g.: in the logistic regression, where coefficients are the logarithm of the Odds Ratios, does not exists, alternatives for interpreting parameters are available by computing the partial derivatives of the first moment on the covariate(s). Table 78 shows such reasoning for the parametric extension of the logit model.

Close to prediction efficiency and effect estimation is the concept of feature selection.

4.2 A decision tree

Attempting to summarize the discussion in the chapters above, a decision tree has been developed (Table 79). The decision tree is built around the basic concepts of ML, also in view of the practical usage which is foreseeable on the basis of EFSA opinions in the previous chapters.

The decision tree starts from the information on the availability of labels:

Table 78: Rate of change $\partial\pi(x)/\partial x$ for different choices of the link function, for non-zero values of λ.

Link	$\partial\pi(x)/\partial x$
Aranda Ordaz Symmetric	$\beta\frac{[\pi(x)^\lambda+(1-\pi(x))^\lambda]^2}{4\pi(x)^{\lambda-1}(1-\pi(x))^{\lambda-1}}$
Aranda Ordaz Asymmetric	$\beta\lambda^{-1}\frac{(1-\pi(x))^{-\lambda}-1}{(1-\pi(x))^{-\lambda-1}}$
Stukel	$\begin{cases}\beta\frac{e^{-1/\lambda(e^{\lambda\lvert\eta\rvert}-1)-\lvert\eta\rvert}}{\left[1+e^{-1/\lambda(e^{\lambda\lvert\eta\rvert}-1)}\right]^2} & \lambda>0 \quad \eta<0\\ \beta\frac{(1-\lambda\lvert\eta\rvert)^{1/\lambda-1}}{\left[1+(1-\lambda\lvert\eta\rvert)^{1/\lambda}\right]^2} & \lambda<0 \quad \eta<0\end{cases}$
Czado	$\beta\frac{(-\eta+1)^{\lambda-1}e^{-\frac{(-\eta+1)^\lambda-1}{\lambda}}}{[1+e^{-\frac{(-\eta+1)^\lambda-1}{\lambda}}]^2}$
Pregibon	$\frac{\beta}{\pi(x)^{\alpha-\delta-1}+[1-\pi(x)]^{\alpha+\delta-1}}$
Gosset	$\beta\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\eta^2\nu^{-1}\right)^{-((\nu+1)/2)}$

- Supervised
- Unsupervised
- Both

Only two families of techniques allow approaching the problem both as a supervised and a unsupervised technique, indicated by "both" in the tree.

The second step is to reason about the output, both in terms of cardinality and quality: first, it will be indicated if the outcome of interest is single (like one response) or multiple (more responses simultaneously, and second the type

- continuous
- categorical/binary
- time series and longitudinal

Then, some information is requested about the characteristics of the problem in terms of linearity. If not sure about that, choosing non-linearity is highly advisable. Scalability of the ML is usually well known in terms of the structure of the data, as well as its foreseeable use and enhancement. For example, most of the problems addressed in EFSA opinions are scalable in principle, but there's the impression that data might not be available in that sense. The ratio between n and p is very important and it is usually very well known when starting the analysis. Since the beginning of the analysis It is also know the purpose of the study, i.e. whether the aim is to estimate an effect or to make a prediction.

As noticed, more than one MLT is available to address the same problem. To properly tackle the issue of robustness, it is highly advisable to run all of them and look for stability and agreement in estimates.

As a general remark, most of the MLT are not directly applicable to the food safety and nutrition field. They require a deep rephrasing of the problem, to make it compatible with the setup of MLT. Solutions are mostly ad-hoc, and they will be illustrated in the following chapters.

Table 79: A decision tree/recipe book — from the problem to the approach — to help in the choice of the most appropriate methodology

Label	**Input** (type)	**Output** (#)	**Output** (type)	**Linear**	**Scalable** (instance)	**Sample size**	**Relation** $(n-p)$	**Miss**	**Complexity**	**Effect**	**Predict**	**Robust**	**MLT**
Both	-	Multiple	Any	Non-linear	Multiple	Medium to high	$n > p$	PreProc.	High	No	No	High [5]	ANN (Artificial Neural Networks)
Both	-	Single	Any	Non-linear	Multiple	Small	$n > p$	Yes	High [6]	No	Yes	Yes [7]	KNN (K-nearest neighbor)
Supervised	-	Multiple	Any	Non-linear	Single	Small	-	PreProc.	High	No	Yes	Yes	EA (Evolutionary Algorithm)
Supervised	-	Multiple	Continuous	Non-linear	Single	High	$n > p$	No	High	No	No	No [8]	GMDH (Group Method of Data Handling)
Supervised	multiple	Multiple	Any	Non-linear	Multiple	Medium	$n \leq p$	No	Low	no	No	Yes	IBL (Instance-based learning)
Supervised	multiple	Multiple	Any	Non-linear	Multiple	High	$n > p$	PreProc.	Medium	no	No	Yes	SVM (Support-Vector Machine)
Supervised	-	Multiple	Categorical	Non-linear	Yes	Medium to high	$n \leq p$	Yes	Medium	Yes	Yes	Yes	Bayesian Networks
Supervised	-	Multiple	Categorical	Non-linear	Yes	Medium	$n > p$	No	High	Yes	No	No	Hidden Markov Model
Supervised	-	Single	Categorical, Binary	Linear	Multiple	Small	$n > p$	PreProc.	High	No	No	-	AODE (Average one-dependence estimator)
Supervised	-	Single	Continuous, Time series	Linear	Single	Small [9]	$n > p$	No	Low	Yes	No	Yes	Kriging (Gaussian process regression)
Supervised	Continuous	Single	Continuous	Linear	Single	Small	$n > p$	No	Low	No	No	No	ANOVA (Analysis Of Variance)
Supervised	Continuous	Single	Binary	Linear	Single	Small	$n > p$	No	Low	No	No	No	Fischer's Linear Discriminant
Supervised	Any	Single	Binary	Linear	Single	Small	$n > p$	No	Low	Yes	Yes	No	Logistic Regression
Supervised	Any	Single	Categorical	Linear	Single	Small	$n > p$	No	Low	Yes	Yes	No	Multinomial Logistic Regression
Supervised	any	Single	Any	Linear	Multiple	Small	$n > p$	Yes	Medium	yes	No	yes	Naive Bayes Classifier
Supervised	any	Single	Binary	Non-linear	Single	High	$n > p$	No	Medium	Yes	No	Yes [10]	LMT (Logistic Model Tree)
Supervised	any	Single	Categorical	Non-linear	Multiple	Small	$n \leq p$	PreProc.	High	No	No	Yes	Random Forest
Supervised	-	Single	Binary	Non-linear	Yes	Small	$n > p$	No	Low	No	No	No	Quadratic Classifier
Supervised	-	Single	Categorical	Non-linear	Multiple	Small	$n \leq p$	Yes	High [11]	No	Yes	No [12]	Decision Tree [13]

Table 79: continue on the following page...

[5] Highly variable with ANN implementation Basheer and Hajmeer, 2000.

[6] Improved versions of traditional approaches are strongly improving performances Wu, Ianakiev, and Govindaraju, 2002.

[7] K-NN methods are robust even to semi-labeled or wrongly-labeled datasets Chi and Bruzzone, 2006.

[8] The major issue is the tendency of the algorithm to stuck in local minima points Tang et al., 1996.

[9] Sampling error and model instability increases with sample size and Smaller spatial autocorrelation Curran and H. Williamson, 1986.

[10] Very high, also in presence of nuisance features Niels Landwehr, Hall, and Frank, 2003.

[11] Highly scalable and distributed computing versions are available Scott et al., n.d.

[12] Specific versions exists to improve robustness in specific cases, like in presence of high error rates in data John, n.d., in speech analysis Shami and Verhelst, 2007 and in health care Yao et al., n.d.

[13] Including ID3, C4.5, CART.

Table 79: ...continue from the previous page

Label	**Input** (type)	**Output** (#)	**Output** (type)	**Linear**	**Scalable** (instance)	**Sample size**	**Relation** ($n - p$)	**Miss**	**Complexity**	**Effect**	**Predict**	**Robust**	MLT
Supervised	-	Single	Categorical	Non-linear	Yes	Small	$n \leq p$	Yes	Medium	Yes	Yes	No	BART
Supervised	-	Single	Time series	Non-linear	Yes	High	$n > p$	No	High [14]	No	No	Yes [15]	Vector Quantization
Supervised	-	Single	Any	Non-linear	Single	Small	$n > p$	No	High	Yes	Yes	No	Moment methods and EM algorithms
Supervised	-	Single	Time series	Non-linear	Yes	Small	$n > p$	PreProc.	-	No	No	-	CWM (Cluster Weighted modeling)
Supervised	-	Multiple	Categorical	Non-linear	Single	Medium	$n > p$	No	High [16]	No	No	-	MRF (Markov Random Fields) [17]
Supervised	General objects	Multiple	General objects	Non-linear	Yes	Small	-	PreProc.	-	No	No	Yes [18]	ILP (Inductive Logic Programming)
Unsupervised	-	Multiple	Continuous	Linear	Yes [19]	-	-	PreProc.	-	No	No	Medium [20]	Hierarchical clustering
Unsupervised	-	Multiple	Continuous, Time series	Linear	No	Small [21]	$n > k$	PreProc.	Medium [22] to high [23]	No	No	High [24].	OPTICS, DBSCAN [25]
Unsupervised	-	Multiple	Continuous [26]	Linear [27]	Multiple	Small to medium [28]	-	PreProc.	High [29]	No	No	-	K-means
Unsupervised	-	Multiple	Continuous	Non-linear	Yes	Small	$n > p$	Yes	High [30]	No	No	-	Self-Organizing Maps (SOM)

Table 79: continue on the following page...

[14] There's an exponential growth of encoding complexity which can be partially addressed by some algorithms Cheng et al., n.d.

[15] In case of highly noised data, averaging strongly improves accuracy Paliwal and Atal, 1993.

[16] Depends heavily from the search algorithm used Sutton and McCallum, 2006.

[17] Including CRF, Conditional Random Fields.

[18] Accuracy has been shown to be quite high even in presence of language biases Chesani et al., 2009.

[19] High scalability is one of the main characteristics of the methods Unrau et al., 1995.

[20] Depends on applications. Very good for gene expression data Herrero, Valencia, and Dopazo, 2001; H. Y. Chang et al., 2005, lower in not spherical databases.

[21] Large spatial data are considered using algorithm modifications A. Zhou et al., 2000.

[22] Computational complexity increases linearly with sample size in legacy implementation of the technique.

[23] DBSCAN in its hierarchical implementation outperforms CLARA and CLARANS Ng and Han, 2002 by an order of magnitude of 100 W. Wang, Yang, and Muntz, n.d.

[24] In general accuracy and robustness depend from the field of application. In industrial applications is very high, not the same in biological settings. Improved versions of DBSCAN in hybrid with ANN and Bayesian Networks are raising accuracy up to 99% Liang et al., 2015

[25] Ordering points to identify the clustering structure (OPTICS) is an algorithm for finding density-based clusters in spatial dataAnkerst et al., n.d. Its basic idea is similar to DBSCAN Ester, H.-p. Kriegel, et al., 1996 but it addresses the problem of detecting meaningful clusters in data of varying density.

[26] Extensions are proposed for dealing with categorical data Z. Huang, 1998; San, Huynh, and Nakamori, 2004.

[27] In legacy implementation Bishop, 2006.

[28] Up to large in some approaches Bradley, Fayyad, and Reina, n.d.

[29] Scalable approaches may reduce computational requirements by several order of magnitude Bradley, Fayyad, and Reina, n.d.

[30] SOM apply competitive learning as opposed to error-correction learning (such as backpropagation with gradient descent), using a neighborhood function to preserve the topological properties of the input space. This makes computational complexity higher than common ANN Barbalho et al., n.d.

Table 79: ...continue from the previous page

Label	**Input** (type)	**Output** (#)	**Output** (type)	**Linear**	**Scalable** (instance)	**Sample size**	**Relation** $(n-p)$	**Miss**	**Complexity**	**Effect**	**Predict**	**Robust**	MLT
Unsupervised	-	Multiple	Continuous	Non-linear	Yes [31]	Small	-	PreProc.	High [32]	No	No	No [33]	Adaptive Resonance Theory (ART) G. A. Carpenter, 2001
Unsupervised	-	Multiple	Any	Non-linear	Yes	small to medium [34]	$n > k$	Yes	Medium	No	No	Medium [35]	Biclustering [36]
Unsupervised	-	Multiple	Continuous	Non-linear	Yes	Small	$n > k$	PreProc.	-	No	No	Low	Group models
Unsupervised	-	Multiple	Continuous	Non-linear	Yes	Small [37]	$n > k$ [38]	PreProc.	High	No	No	Low [39]	HCS clustering (semi-cliques)
Unsupervised	-	Multiple	Continuous	Non-linear	Yes	small to medium [40]	$n > k$	Yes [41]	Medium	No	No	Low[42]	Fuzzy (soft) clustering [43]
Unsupervised	-	Single	Continuous	Linear	Single	Small to medium	$n > p$	No	Medium	No	No	No	Principal component analysis
Unsupervised	-	Single	Continuous	Linear	Single	medium to high [44]"	$n > p$	No	High [45]	Yes	Yes	No [46]	Independent component analysis

[31] In adapted versions Mulder and D. C. Wunsch, n.d.

[32] Fast adaptive-search algorithms reduce complexity Gail A Carpenter and Grossberg, 1988.

[33] In gene-expression analysis, the fuzzy version of ART Yom-Tov and Inbar, 2002 has been shown to be highly robust to noise in data.

[34] Extensions have been proposed to cover large datasetsDolnicar et al., 2012.

[35] Good capacity to deal with noise. In gene expression studies, biclustering is however poor in case of overlapping clusters Eren et al., 2013.

[36] Including block clustering Govaert and Nadif, 2008, co-clustering or two-mode clustering Govaert and Nadif, 2013; Van Mechelen, Bock, and De Boeck, 2004 and other data mining technique which allows simultaneous clustering of the rows and columns of a matrix.

[37] New algorithms like CLICK Sharan and Shamir, n.d. in gene-expression analysis extended the capability of HCS and in general graph-based clustering to very large (100000+) datasets.

[38] In partition-based approaches, it allows n<k.

[39] HTC methods are highly sensitive to noise and outliers Shu and Schaeffer, n.d. Once an object is assigned to a cluster, it will not be considered again, which means that HC algorithms are not capable of correcting possible previous misclassificationR. Xu and D. Wunsch, 2005.

[40] Extensions have been proposed to cover large datasetsZ. Huang, 1998.

[41] Missing data is provided by extensions like missing data partitioning and clustering Timm, Döring, and Kruse, 2004

[42] Pal and Bezdek, 1995

[43] Including C-Means and Soft K-Means

[44] If performed in high dimensions with an insufficient sample size, this may lead to generation of artifactual source signals due to over-learning (or overfitting). The existence of strong time-correlations in the data increases the probability of the occurrence of artifacts. These results are essentially independent of the particular algorithm used for ICA Hyvarinen, Sarela, and Vigário, n.d.

[45] In its Maximum Likelihood version it faces problems of local maxima Hyvärinen and Erkki Oja, 2000.

[46] It is heavily relying on the independence assumption of all (non-Gaussian) subcomponent signals.

4.3 R packages for ML

In Table 80 we have reported all the ML packages provided by R with the relative functions and the principal features. Some notes and the bibliography used by the authors of the packages are reported.

Table 80: R packages & functions with principal features

Family/ package/**function**	**Interact**	**Class**	**Regr**	**Clust**	**Filter**	**Dim**	**Input**	**Citations**
Artificial Neural Networks								
nnet	— Note:							Venables and B. D. Ripley, 1994
multinom	S	no	yes	no	no	Any	real	
nnet	S	yes	yes	no	no	Any	real	B. Ripley, n.d.
RSNNS	— Note: —							Bergmeir and Benítez Sánchez, n.d.; Andreas Zell et al., 1998
art1	UnS	no	no	yes	no		binary	Gail A Carpenter and Grossberg, 1987; Grossberg, 1976; Herrmann, 1992; A Zell, 1994
art2	UnS	no	no	yes	no		real	
artmap	S	no	no	yes	no		binary	
assoz	UnS	no	no	yes	no		binary	Palm, 1980; Rojas, 1996
dlvq	S	yes	?	no	no	1	?	Kohonen and Somervuo, 1998
jordan	S	yes	yes	no	no		?	Jordan, 1986
elman	S	yes	yes	no	no		?	Elman, 1990
rbf	S	yes	yes	no	no		real	Poggio and Girosi, 1989; Vogt, 1992
rbfDDA	S	yes	?	no	no		real	Berthold and Diamond, 1995
som	UnS	yes	no	no	no	2	real	Kohonen and Somervuo, 1998
FCNN4R	— Note: Networks can be exported to C functions in order to integrate them into virtually any software solution. —							
mlp_teach_bp	S	?	yes	no	no		?	Bryson, 1975
mlp_teach_rprop	S	yes	yes	no	no		real	Riedmiller and Braun, n.d.; Riedmiller, 1994
mlp_teach_grprop	S	yes	yes	no	no		real	
mlp_teach_sa	S	yes	yes	no	no		binary	
mlp_teach_sgd		yes	yes	no	no		real	
Recuryesve Partitioning								
rpart	— Note: —							
rpart	S	yes	yes	no	no		real	Leo Breiman et al., 1984

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/**function**	**Interact**	**Class**	**Regr**	**Clust**	**Filter**	**Dim**	**Input**	**Citations**
tree	— Note: —							
tree	S	yes	yes	no	no		real	Leo Breiman et al., 1984; B. Ripley, n.d.
RWeka	— Note: NOTE: need java —							I. H. Witten and Frank, 2005
Apriori	UnS	?	?	?				Agrawal and Srikant, n.d.
Tertius	UnS	?	?	?				Flach and Lachiche, 2001
LinearRegression	S	no	yes	no				
Logistic	S	no	yes	no				
SMO	S	yes	no	no				J. C. Platt, 1999
IBk	S	yes	no	no				Aha, Kibler, and Albert, 1991
LBR	S	yes	no	no				Zheng and Webb, 2000
AdaBoostM1	S	yes	no	no				Freund and Schapire, n.d.
Bagging	S	yes	no	no				Leo Breiman, 1996a
logitBoost	S	yes	no	no				J. Friedman, Hastie, and Robert Tibshirani, 2000
MultiBoostAB	S	yes	no	no				Webb, 2000
Stacking	S	yes	no	no				Wolpert, 1992
CostSenyestiveClasyesfier	S	yes	no	no				
JRip	S	yes	?	no				Cohen, n.d.
M5Rules	S	yes	no	no				Holmes, Hall, and Prank, 1999
OneR	S	yes	no	no				Holte, 1993
PART	S	yes	no	no				Frank and I. H. Witten, n.d.
J48	S	yes	no	no				J Ross Quinlan, n.d.; 1993
LMT	S	yes	?	no				N Landwehr, 2003; Niels Landwehr, Hall, and Frank, 2005
M5P	S	yes	no	no				I. Witten and Y. Wang, n.d.; J Ross Quinlan, n.d.; 1993
DeciyesonStump	S	yes	?	no				
Cobweb	UnS	no	no	yes				D. H. Fisher, 1987; Gennari, Langley, and D. Fisher, 1989
FarthestFirst	UnS	no	no	yes				Hochbaum and Shmoys, 1985
yesmpleKMeans	UnS	no	no	yes				
XMeans	UnS	no	no	yes				Pelleg and Moore, n.d.
DBScan	UnS	no	no	yes				Ester, H.-P. Kriegel, et al., n.d.
Normalize	UnS	no	no	no	yes		numeric	Irani, 1993
Discretize	S	no	no	no	yes		numeric	
LovinsStememr	UnS	no	no	no	no		chr vctr	Lovins, 1968
IterateLovinsStemmer	UnS	no	no	no	no		chr vctr	

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/function	Interact	Class	Regr	Clust	Filter	Dim	Input	Citations
Cubist	— Note: —							
cubist	S	no	yes	no	no		numeric	John R Quinlan, n.d.; J Ross Quinlan, n.d.; 1993; John R Quinlan, n.d.; Y. W. Wang, n.d.
C50	— Note: —							
C5.0	S	yes	no	no	no			J Ross Quinlan, n.d.; 1993
party	— Note: —							
cforest	S	yes	yes	no	no	Any	numeric	L. Breiman, 2001; Hothorn, Lausen, et al., 2004; Hothorn, Bühlmann, et al., 2006; Hothorn, Hornik, and Zeileis, 2006a; Strobl, J. Malley, and Tutz, 2009; Strobl, Boulesteix, et al., 2007
ctree	S	yes	yes	no	no	Any	numeric	Strasser and Weber, 1999; Hothorn, Hornik, Van De Wiel, et al., 2012
mob	S	yes	yes	no	no	Any	numeric	Zeileis, Hothorn, and Hornik, 2008
vcrpart	— Note: —							
fvcm	S	?	yes	no	no	Any	numeric	Leo Breiman, 1996a; L. Breiman, 2001; J. Friedman, Hastie, and Robert Tibshirani, 2001
fvcolmm	S	?	yes	no	no	Any	numeric	
fvcglm	S	?	yes	no	no	Any	numeric	
olmm	S	?	yes	no	no	<4	numeric	
tvcglm	S	?	yes	no	no	Any	numeric	Leo Breiman et al., 1984; J. C. Wang and Hastie, 2014; Bürgin and Ritschard, 2015
tvcm	S	?	yes	no	no	Any	numeric	Zeileis, Hothorn, and Hornik, 2008; J. C. Wang and Hastie, 2014; Hothorn and Zeileis, 2014; Bürgin and Ritschard, 2015
tvccolmm	S	?	yes	no	no	Any	numeric	Zeileis and Hornik, 2007; Sela and Simonoff, 2012; Hajjem, Bellavance, and Larocque, 2011
LogicReg	— Note: —							

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/function	Interact	Class	Regr	Clust	Filter	Dim	Input	Citations
logreg	S	yes	yes	no	no	Any		I. Ruczinski, Kooperberg, and M. LeBlanc, 2003a; I. Ruczinski, Kooperberg, and M. LeBlanc, 2003b; Kooperberg and I. Ruczinski, 2005; C. K. I. Ruczinski, M. L. LeBlanc, and L. Hsu, 2001
REEMtree	— Note: —							
REEMtree	S	?	yes	no	no	Any	numeric	Sela and Simonoff, 2012
RPMM	— Note: —							
blcTree	UnS	no	no	yes	no	Any	numeric	Houseman et al., 2008
glcTree	UnS	no	no	yes	no	Any	numeric	
partykit	— Note: —							
cforest	S	?	yes	no	no	Any	numeric	L. Breiman, 2001; Hothorn, Lausen, et al., 2004; Hothorn, Bühlmann, et al., 2006; Hothorn, Hornik, and Zeileis, 2006a; Hothorn and Zeileis, 2014; Meinshausen, 2006; Strobl, Boulesteix, et al., 2007
ctree	S	?	yes	no	no	Any	numeric	Hothorn, Hornik, Van De Wiel, et al., 2012; Strasser and Weber, 1999
glmtree	S	?	yes	no	no	Any	numeric	Zeileis, Hothorn, and Hornik, 2008
lmtree	S	?	yes	no	no	Any	numeric	
mob	S	?	yes	no	no	Any	numeric	
evtree	— Note: —							
evtree	S	yes	yes	no	no	Any	numeric	Grubinger, Zeileis, and Pfeiffer, 2011
oblique.tree	— Note: —							
glmpath	S	?	yes	no	no	Any	numeric	Park and Hastie, 2007
oblique.tree	S	yes	yes	no	no	Any	numeric	Truong, 2009; B. Ripley, n.d.
Random Forest								
randomForest	— Note: —							
randomForest	Any	yes	yes	?	no	Any	numeric	L. Breiman, 2001; Leo Breiman, 2002

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/function	Interact	Class	Regr	Clust	Filter	Dim	Input	Citations
ipred	— Note: —							
bagging	S	yes	yes	no	no	Any	numeric	Leo Breiman, 1996a; Leo Breiman, 1996b; Leo Breiman, 1998; Büchlmann and Yu, 2002; Hothorn and Lausen, 2005; Hothorn, Lausen, et al., 2004
inbagg	S	yes	yes	no	no	Any	numeric	Hand, H. G. Li, and Adams, 2001; Peters et al., 2003
inclass	S	yes	no	no	no	Any	numeric	
slda	S	yes	no	no	yes	Any	numeric	Kai-Tai and Yao-Ting, 1990; Kropf, 2000; Läuter, 1992; Lauter, Glimm, and Kropf, 1998
party	— Note: —							
cforest	S	yes	yes	no	no	Any	numeric	L. Breiman, 2001; Hothorn, Lausen, et al., 2004; Hothorn, Bühlmann, et al., 2006; Hothorn, Hornik, and Zeileis, 2006a; Strobl, Boulesteix, et al., 2007; Strobl, Boulesteix, et al., 2007
ctree	S	yes	yes	no	no	Any	numeric	Strasser and Weber, 1999; Hothorn, Hornik, Van De Wiel, et al., 2012
mob	S	yes	yes	no	no	Any	numeric	Zeileis, Hothorn, and Hornik, 2008
randomForestSRC	— Note: parallel processing via parallel::mclapply —							L. Breiman, 2001; Ishwaran, 2007; Ishwaran, Kogalur, Gorodeski, et al., 2010; Ishwaran, Gerds, et al., 2014; Ishwaran, 2015
rfsrc	S	yes	yes	yes	no	Any	numeric	Leo Breiman et al., 1984; Cutler and Zhao, 2001; Gray, 1988; Harrell et al., 1982; Hothorn and Lausen, 2003b; Ishwaran, 2007; Y. Lin and Jeon, 2006; M. LeBlanc and Crowley, 1993; Loh and Shih, 1997; Mogensen, Ishwaran, and Gerds, 2012; Segal, 1988
rfsrcSyn	S	yes	yes	no	no	Any	numeric	
var.select	S	no	no	no	yes	Any	numeric	Ishwaran, Kogalur, Chen, et al., 2011

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/function	Interact	Class	Regr	Clust	Filter	Dim	Input	Citations
quantregForest	— Note: —							
quantregForest	S	no	yes	no	no	Any	numeric	Meinshausen, 2006
LogicForest	— Note: —							
LBoost	S	yes	?	no	yes	Any	binary	Wolf, Hill, Slate, et al., 2012
logforest	S	yes	?	no	yes	Any	binary	Wolf, Hill, and Slate, 2010
LogicReg	— Note: —							
logreg	S	yes	yes	no	no	Any		I. Ruczinski, Kooperberg, and M. LeBlanc, 2003a; I. Ruczinski, Kooperberg, and M. LeBlanc, 2003b; C. K. I. Ruczinski, M. L. LeBlanc, and L. Hsu, 2001; Kooperberg and I. Ruczinski, 2005
varSelRF	— Note: —							
randomVarImpsRF	UnS	no	no	no	yes	Any	numeric	L. Breiman, 2001; Diaz-Uriarte and de Andrés, 2005; Svetnik et al., 2004
varSelImpSpecRF	UnS	no	no	no	yes	Any	numeric	Jerome H Friedman and Meulman, 2004
varSelRF	UnS	no	no	no	yes	Any	numeric	
Boruta	— Note: —							
Boruta	Any	yes	yes	no	yes	Any	vctr, fctr, num	Kursa and Rudnicki, 2010
TentativeoughFix	Any	no	no	no	yes	Any	vctr, fctr, num	
ranger	— Note: —							
ranger	S	yes	yes	no	yes	Any	vctr, fctr, num	Wright and Ziegler, 2015; L. Breiman, 2001; Ishwaran, Kogalur, Blackstone, et al., 2008; J. D. Malley et al., 2012
Rborist	— Note: —							
Rborist	S	yes	yes	no	no	Any	vctr, fctr, num	
Regularized and Shrinkage Methods								
lasso2	— Note: —							
gl1ce	S	no	yes	no	no	Any	numeric	Lokhorst, 1999
l1ce	S	no	yes	no	no	Any	numeric	Osborne, Presnell, and Turlach, 2000; Robert Tibshirani, 1996

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/function	Interact	Class	Regr	Clust	Filter	Dim	Input	Citations
lars	— Note: —							
lars	S	no	yes	no	no	Any	numeric	Efron et al., 2004
grplasso	— Note: *early test release* —							
grplasso	S	no	yes	no	no	Any	numeric	Meier, Van De Geer, and Bühlmann, 2008
grpreg	— Note: —							Yuan and Y. Lin, 2006; Breheny and J. Huang, 2009; J. Huang, Breheny, and Ma, 2012; Breheny, 2015
gBridge	S	?	yes	no	no	Any	numeric	Breheny and J. Huang, 2009
grpreg	S	?	yes	no	no	Any	numeric	
glmpath	— Note: —							
coxpath	S	?	yes	no	no	Any	list	Park and Hastie, 2007
glmpath	S	?	yes	no	no	Any	numeric	
elasticnet	— Note: —							
enet	S	?	yes	no	no	Any	numeric	Zou and Hastie, 2005
spca	UnS	?	no	yes	yes	Any	numeric	Zou, Hastie, and Robert Tibshirani, 2006b
glmnet	— Note: —							
glmnet	S	yes	yes	no	no	Any	numeric	J. Friedman, Hastie, and Rob Tibshirani, 2010 Robert Tibshirani et al., 2012
penalized	— Note: —							
penalized	S	?	yes	no	no	Any	numeric	Goeman, 2010
RXshrink	— Note: —							
RXlarlso	S	no	yes	no	no	Any	numeric	Leo Breiman, 1995; Efron et al., 2004; 2004
RXridge	S	no	yes	no	no	Any	numeric	Goldstein and Smith, 1974; Burr and Fry, 2005
RXtrisk	S	no	yes	no	no	Any	numeric	
RXtyesmu	S	no	yes	no	no	Any	numeric	
RXuclars	S	no	yes	no	no	Any	numeric	
ahaz	— Note: —							
ahaz	S	no	yes	no	no	Any	numeric	D. Lin and Ying, 1994
ahaziyess	S	no	yes	no	no	Any	numeric	Gorst-Rasmussen and T. H. Scheike, 2011

Table 80: continue on the following page

Table 80: continue from the last page

Family/package/function	Interact	Class	Regr	Clust	Filter	Dim	Input	Citations
ahazpen	S	no	yes	no	no	Any	numeric	Gorst-Rasmussen and T. Scheike, 2011; Leng and Ma, 2007; Martinussen and T. H. Scheike, 2009
relaxo	— Note: —							
relaxo	S	?	yes	no	no	Any	numeric	Meinshausen, 2007
penalizedLDA	— Note: —							
PenalizedLDA	S	yes	?	no	no	Any	numeric	D. M. Witten and Robert Tibshirani, 2011
earth	— Note: —							
earth	S	no	yes	no	no	Any	numeric	Faraway, 2005; J. H. Friedman, 1991; Jerome H Friedman and Silverman, 1989; Leathwick et al., 2005; Miller, 2002
penalizedSVM	— Note: —							
lpsvm	S	yes	no	no	no	Any	numeric	Fung and Mangasarian, 2004
scadsvc	S	yes	no	no	no	Any	numeric	H. H. Zhang, Ahn, et al., 2006
svm.fs	S	yes	no	no	no	Any	numeric	Becker et al., 2009
hda	— Note: —							
hda	UnS	yes	no	no	yes	Any	numeric	Burget, n.d.; Kumar and Andreou, 1998; Szepannek, Harczos, et al., n.d.
rda	— Note: —							
rda	S	yes	no	no	no	Any	numeric	Guo, Hastie, and Robert Tibshirani, 2005
sda	— Note: —							
sda	S	yes	no	no	yes	Any	numeric	Ahdesmäki and Strimmer, 2010
LiblineaR	— Note: A wrapper around the LIBLINEAR C/C++ library formachine learning (available at http://www.cyese.ntu.edu.tw/ cjlin/liblinear). —							
heuristicC	UnS	no	no	no	no	Any	numeric	
LiblineaR	S	yes	yes	no	?	Any	numeric	Fan et al., 2008
ncvreg	— Note: —							
ncvreg	S	?	yes	no	no	Any	numeric	Breheny and J. Huang, 2011
ncvsurv	S	no	yes	no	no	Any	numeric	Breheny and J. Huang, 2011,
bigRR	— Note: —							

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/function	Interact	Class	Regr	Clust	Filter	Dim	Input	Citations
bigRR	S	no	yes	no	no	Any	numeric	Shen et al., 2013; Ronnegard, Shen, and Alam, 2010
hugeRR	S	no	yes	no	no	Any	numeric	
bmrm	— Note: —							
bmr	S	?	yes	no	no	Any	numeric	Teo et al., n.d.
Boosting								
ada	— Note: —							
ada	S	yes	no	no	no	2	numeric	J. Friedman, Hastie, and Robert Tibshirani, 2000; Jerome H Friedman, 2001; Jerome H Friedman, 2002; Culp, Johnson, and Michailidis, 2016
gbm	— Note: —							
gbm	S	?	yes	no	no	Any	any	Freund and Schapire, 1997; Ridgeway, 1999; J. Friedman, Hastie, and Robert Tibshirani, 2000; Jerome H Friedman, 2001; Jerome H Friedman, 2002; Kriegler, 2007; Burges, 2010
bst	— Note: —							
bst	S	yes	yes	no	no	Any	numeric	Z. Wang, 2011; Bühlmann and Hothorn, 2010
mbst	S	yes	no	no	no	Any	numeric	Z. Wang, 2012
mhingebst	S	yes	no	no	no	Any	numeric	
mhingeova	S	yes	no	no	no	Any	numeric	
rbst	S	yes	no	no	no	Any	binary	
rbstpath	S	yes	no	no	no	Any	binary	
rmbst	S	yes	no	no	no	Any	numeric	
GAMBoost	— Note: —							
GAMBoost	S	yes	yes	no	no	Any	numeric	Harald Binder and Schumacher, 2009; Harald Binder and Schumacher, 2008; Hurvich, Simonoff, and C.-L. Tsai, 1998; Eilers and Marx, 1996; Tutz and Harald Binder, 2007; Tutz and Harald Binder, 2006
GLMBoost	S	yes	yes	no	no	Any	numeric	

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/function	Interact	Class	Regr	Clust	Filter	Dim	Input	Citations
mboost	— Note: —							
blackboost	S	no	yes	no	no	Any	numeric	Bühlmann and Hothorn, 2007; Hothorn, Hornik, and Zeileis, 2006b; Freund and Schapire, n.d.; Jerome H Friedman, 2001; Ridgeway, 1999
gamboost	S	?	yes	no	no	Any	numeric	
glmboost	S	no	yes	no	no	Any	numeric	
mboost	S	no	yes	no	no	Any	numeric	
CoxBoost	— Note: —							
CoxBoost	S	no	yes	no	no	Any	numeric	H Binder et al., 2013; Harald Binder, Allignol, et al., 2009; Harald Binder and Schumacher, 2009; Harald Binder and Schumacher, 2008; Tutz and Harald Binder, 2007; Fine and Gray, 1999
GMMBoost	— Note: —							
bGamm	S	yes	yes	no	no	Any	numeric	Groll and Tutz, 2012
bGLMM	S	yes	yes	no	no	Any	numeric	Tutz and Groll, 2010
OrdinalBoost	S	yes	yes	no	no	Any		Tutz and Groll, 2013
gamboostLSS	— Note: —							
mboostLSS	S	?	yes	no	no	Any	numeric	Hofner, Mayr, and Schmid, 2014; Mayr et al., 2012; Schmid et al., 2010; Rigby and Stasinopoulos, 2005; Bühlmann and Hothorn, 2007
glmboostLSS	S	?	yes	no	no	Any	numeric	
gamboostLSS	S	?	yes	no	no	Any	numeric	
blackboostLSS	S	?	yes	no	no	Any	numeric	
Support Vector Machine								
e1071	— Note: —							
bclust	UnS	no	no	yes	no	Any	numeric	Leisch, 1999
cmeans	UnS	no	no	yes	no	Any	numeric	Bezdek, 2013; Chung and T. Lee, 1994; Pal, Bezdek, and Hathaway, 1996
cshell	UnS	no	no	yes	no	Any	numeric	Dave, 1990
ica	UnS	no	no	yes	yes	Any	numeric	E Oja, 1991; Karhunen and Joutsensalo, 1995
lca	UnS	no	no	yes	no	Any	numeric	

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/function	Interact	Class	Regr	Clust	Filter	Dim	Input	Citations
naiveBayes	S	yes	no	no	no	Any	numeric	
svm	S	yes	yes	no	no	Any	numeric	C.-C. Chang and C.-J. Lin, 2011
kernlab	— Note: —							
gausspr	S	yes	yes	no	no	Any	numeric	C. K. Williams and Barber, 1998
kha	UnS	no	no	yes	yes	Any	numeric	Kim, Franz, and Schölkopf, 2003
kkmeans	UnS	no	no	yes	no	Any	numeric	Dhillon, Guan, and Kulis, 2004
kpca	UnS	no	no	yes	yes	Any	numeric	Schölkopf, A. Smola, and Müller, 1998
kqr	S	no	yes	no	no	Any	numeric	Takeuchi et al., 2006
ksvm	S	yes	yes	no	no	Any	numeric	C.-C. Chang and C.-J. Lin, 2011; J. Platt, 1999; H.-T. Lin, C.-J. Lin, and Weng, 2007; C.-W. Hsu and C.-J. Lin, 2002; Crammer and Singer, 2002; Weston and Watkins, 1998
lssvm	S	yes	yes	no	no	Any	numeric	Suykens and Vandewalle, 1999
onlearn	S	yes	yes	no	no	Any	binary	Kivinen, A. J. Smola, and R. C. Williamson, 2004
ranking	Semi-S	yes	yes	yes	no	Any	vctr, fctr, num	D. Zhou et al., 2004
rvm	S	yes	yes	no	no	Any	numeric	Tipping, 2001
specc	UnS	no	no	yes	no	Any	numeric	
klaR	— Note: —							
corclust	UnS	no	no	yes	no	Any	numeric	
greedy.wilks	UnS	no	no	no	yes	Any	numeric	
kmodes	UnS	no	no	yes	no	Any	categorical	Z. Huang, n.d.; MacQueen, n.d.
loclda	UnS	yes	no	no	no	Any	numeric	Tutz and Harald Binder, 2005
NaiveBayes	S	yes	no	no	no	Any	vctr, fctr, num	
nm	UnS	no	no	yes	yes	Any	numeric	
pvs	UnS	yes	no	no	yes	Any	numeric	Szepannek and Weihs, 2006
rda	S	yes	yes	no	no	Any	numeric	
sknn	UnS	no	no	yes	no	Any	numeric	
svmlight	S	yes	no	no	no	Any	numeric	
Bayeyesan Methods								
BayesTree	— Note: —							

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/**function**	**Interact**	**Class**	**Regr**	**Clust**	**Filter**	**Dim**	**Input**	**Citations**
bart	S	yes	yes	no	no	Any	vctr, fctr, num	Chipman, George, and McCulloch, 2010; Chipman, George, and McCulloch, 2007; J. H. Friedman, 1991
tgp	— Note: —							
blm	S	yes	yes	no	no	Any	numeric	Gramacy, 2007; Gramacy and Taddy, 2014; Gramacy and Taddy, 2014; Gramacy and H. K. Lee, 2008; Gramacy and Lian, 2012; Gramacy and H. K. Lee, 2009; Chipman, George, and McCulloch, 1998; Chipman, George, and McCulloch, 2002; Schonlau, Welch, and Jones, 1998
btlm	S	yes	yes	no	no	Any	numeric	
bcart	S	yes	yes	no	no	Any	numeric	
bgp	S	yes	yes	no	no	Any	numeric	
bgpllm	S	yes	yes	no	no	Any	numeric	
btgp	S	yes	yes	no	no	Any	numeric	
btgpllm	S	yes	yes	no	no	Any	numeric	
bnclasyesfy	— Note: —							
Model selection and Validation								
e1071	— Note: —							
bclust	UnS	no	no	yes	no	Any	numeric	Leisch, 1999
cmeans	UnS	no	no	yes	no	Any	numeric	Bezdek, 2013; Chung and T. Lee, 1994; Pal, Bezdek, and Hathaway, 1996
cshell	UnS	no	no	yes	no	Any	numeric	Dave, 1990
ica	UnS	no	no	yes	yes	Any	numeric	E Oja, 1991; Karhunen and Joutsensalo, 1995
lca	UnS	no	no	yes	no	Any	numeric	
naiveBayes	S	yes	no	no	no	Any	numeric	
svm	S	yes	yes	no	no	Any	numeric	C.-C. Chang and C.-J. Lin, 2011
ipred	— Note: —							

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/**function**	**Interact**	**Class**	**Regr**	**Clust**	**Filter**	**Dim**	**Input**	**Citations**
bagging	S	yes	yes	no	no	Any	numeric	Leo Breiman, 1996b; Leo Breiman, 1996a; Leo Breiman, 1998; Büchlmann and Yu, 2002; Hothorn and Lausen, 2003a; Hothorn and Lausen, 2005; Hothorn, Lausen, et al., 2004
inbagg	S	yes	yes	no	no	Any	numeric	Hand, H. G. Li, and Adams, 2001; Peters et al., 2003
inclass	S	yes	no	no	no	Any	numeric	
slda	S	yes	no	no	yes	Any	numeric	Kai-Tai and Yao-Ting, 1990; Kropf, 2000; Läuter, 1992; Lauter, Glimm, and Kropf, 1998
svmpath	— Note: —							
svmpath	S	no	no	no	no	Any	vctr, fctr, num	
ROCR	— Note: —							
hdi	— Note: —							
hdi	Semi-S	no	no	no	yes	Any	numeric	Meinshausen, Meier, and Bühlmann, 2012; Meinshausen and Bühlmann, 2010
stability	S	no	yes	no	yes	Any	numeric	Bühlmann, Kalisch, and Meier, 2014
stabs	— Note: —							
lars.lasso	S	no	no	no	yes	Any	numeric	
lars.stepwise	S	no	no	no	yes	Any	numeric	
glmnet.lasso	S	no	no	no	yes	Any	numeric	
glmnet.lasso_maxCoeff	S	no	no	no	yes	Any	numeric	
stabsel	S	no	no	no	yes	Any	numeric	
Meta MLT								
caret	— Note: —							
mlr	— Note: —							
SuperLearner	— Note: —							
SuperLearner	S	yes	yes	no	no	Any	numeric	
h2o	— Note: —							
GUI								
rattle	— Note: GUI —							

Table 80: continue on the following page

Table 80: continue from the last page

Family/ package/**function**	**Interact**	**Class**	**Regr**	**Clust**	**Filter**	**Dim**	**Input**	**Citations**
Repoyestory packages								
CORElearn	— Note: —							
coremodel	S	yes	yes	no	no	Any	vctr, fctr, num	Robnik-Šikonja and Kononenko, 2003; L. Breiman, 2001; Robnik-Šikonja, 2004; Robnik Šikonja, n.d.; Sikonja and Kononenko, 1995
rminer	— Note: —							
fit	S	yes	yes	no	yes	Any	vctr, fctr, num	Cortez, 2015; Cortez et al., 2009; C.-M. Huang et al., 2007
mining	S	yes	yes	no	yes		vctr, fctr, num	

Decision Tree Bibliography

[04] Generic. 2004.

[07] Generic. 2007.

[93] Generic. 1993.

[Adc97] CJ Adcock. “Sample size determination: a review”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 46.2 (1997), pp. 261–283. issn: 1467-9884.

[Aka73] Htrotugu Akaike. “Maximum likelihood identification of Gaussian autoregressive moving average models”. In: *Biometrika* 60.2 (1973), pp. 255–265. issn: 0006-3444.

[AKA91] David W Aha, Dennis Kibler, and Marc K Albert. “Instance-based learning algorithms”. In: *Machine learning* 6.1 (1991), pp. 37–66. issn: 0885-6125.

[Ams09] EFSA AMU (Assessment and methodological support). “Guidance of the Scientific Committee on Use of the benchmark dose approach in risk assessment”. In: *EFSA Journal*, 1150 (2009), p. 72. doi: 10.2903/j.efsa.2009.1150.

[AMS97] Christopher G Atkeson, Andrew W Moore, and Stefan Schaal. “Locally weighted learning for control”. In: *Lazy learning*. Springer, 1997, pp. 75–113. isbn: 904814860X.

[Ank+] Mihael Ankerst et al. “OPTICS: ordering points to identify the clustering structure”. In: *ACM Sigmod Record*. Vol. 28. ACM, pp. 49–60. isbn: 1581130848.

[AS] Rakesh Agrawal and Ramakrishnan Srikant. “Fast algorithms for mining association rules”. In: *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215, pp. 487–499.

[AS10] Miika Ahdesmäki and Korbinian Strimmer. “Feature selection in omics prediction problems using cat scores and false nondiscovery rate control”. In: *The Annals of Applied Statistics* 4.1 (2010), pp. 503–519. issn: 1932-6157.

[Aut10] European Food Safety Authority. “Model-based comparative assessment of the Australian and European hygiene monitoring programmes for meat production”. In: *EFSA Journal* 8.6, 1450 (2010), p. 52. doi: 10.2903/j.efsa.2010.1450.

[AW10] Hervé Abdi and Lynne J. Williams. “Principal component analysis”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4 (2010), pp. 433–459. issn: 1939-0068. doi: 10.1002/wics.101. url: http://dx.doi.org/10.1002/wics.101.

[Bar+] Jose M Barbalho et al. “Hierarchical SOM applied to image compression”. In: *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*. Vol. 1. IEEE, pp. 442–447. isbn: 0780370449.

[BB] Cristoph Norbert Bergmeir and José Manuel Benítez Sánchez. “Neural networks in R using the Stuttgart neural network simulator: RSNNS”. In: American Statistical Association.

[BD95] Michael R Berthold and Jay Diamond. “Boosting the performance of rbf networks with dynamic decay adjustment”. In: (1995).

[BE02] Olivier Bousquet and André Elisseeff. “Stability and generalization”. In: *The Journal of Machine Learning Research* 2 (2002), pp. 499–526. issn: 1532-4435.

[Bec+09] Natalia Becker et al. “penalizedSVM: a R-package for feature selection SVM classification”. In: *Bioinformatics* 25.13 (2009), pp. 1711–1712. issn: 1367-4803.

[Bez13] James C Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013. isbn: 147570450X.

[BF05] Tom L Burr and Herbert A Fry. “Biased regression: The case for cautious application”. In: *Technometrics* 47.3 (2005), pp. 284–296. issn: 0040-1706.

[BFG13] Paola Berchialla, Francesca Foltran, and Dario Gregori. “Naive Bayes classifiers with feature selection to predict hospitalization and complications due to objects swallowing and ingestion among European children”. In: *Safety science* 51.1 (2013), pp. 1–5. issn: 0925-7535.

[BFR] Paul S Bradley, Usama M Fayyad, and Cory Reina. “Scaling Clustering Algorithms to Large Databases”. In: *KDD*, pp. 9–15.

[BH00] IA Basheer and M Hajmeer. "Artificial neural networks: fundamentals, computing, design, and application". In: *Journal of microbiological methods* 43.1 (2000), pp. 3–31. issn: 0167-7012.

[BH07] Peter Bühlmann and Torsten Hothorn. "Boosting algorithms: Regularization, prediction and model fitting". In: *Statistical Science* (2007), pp. 477–505. issn: 0883-4237.

[BH09] Patrick Breheny and Jian Huang. "Penalized methods for bi-level variable selection". In: *Statistics and its interface* 2.3 (2009), p. 369. url: http://www.ncbi.nlm.nih.gov/pubmed/20640242.

[BH10] Peter Bühlmann and Torsten Hothorn. "Twin boosting: improved feature selection and prediction". In: *Statistics and Computing* 20.2 (2010), pp. 119–138. issn: 0960-3174.

[BH11] Patrick Breheny and Jian Huang. "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection". In: *The annals of applied statistics* 5.1 (2011), p. 232.

[BHS06] H. Bhaskar, D. C. Hoyle, and S. Singh. "Machine learning in bioinformatics: a brief survey and recommendations for practitioners". In: *Comput Biol Med* 36.10 (2006), pp. 1104–25. issn: 0010-4825 (Print), 0010-4825 (Linking). doi: 10.1016/j.compbiomed.2005.09.002. url: http://www.ncbi.nlm.nih.gov/pubmed/16226240.

[Bin+09] Harald Binder, Arthur Allignol, et al. "Boosting for high-dimensional time-to-event data with competing risks". In: *Bioinformatics* 25.7 (2009), pp. 890–896. issn: 1367-4803.

[Bin+13] H Binder et al. "Tailoring sparse multivariable regression techniques for prognostic single-nucleotide polymorphism signatures". In: *Statistics in medicine* 32.10 (2013), pp. 1778–1791. issn: 1097-0258.

[Bis06] Christopher M Bishop. "Pattern Recognition". In: *Machine Learning* (2006).

[BKM14] Peter Bühlmann, Markus Kalisch, and Lukas Meier. "High-dimensional statistics with a view toward applications in biology". In: *Annual Review of Statistics and Its Application* 1 (2014), pp. 255–278. issn: 2326-8298.

[Bor] Christian Borgelt. "Efficient implementations of apriori and eclat". In: *FIMI'03: Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations*.

[BR15] Reto Bürgin and Gilbert Ritschard. "Tree-based varying coefficient regression for longitudinal ordinal responses". In: *Computational Statistics & Data Analysis* 86 (2015), pp. 65–80. issn: 0167-9473.

[Bre+84] Leo Breiman et al. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, 1984.

[Bre01] L. Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

[Bre02] Leo Breiman. "Manual on setting up, using, and understanding random forests v3. 1". In: *Statistics Department University of California Berkeley, CA, USA* (2002).

[Bre15] Patrick Breheny. "The group exponential lasso for bi-level variable selection". In: *Biometrics* 71.3 (2015), pp. 731–740. issn: 1541-0420.

[Bre95] Leo Breiman. "Better subset regression using the nonnegative garrote". In: *Technometrics* 37 (1995), pp. 373–384.

[Bre96a] Leo Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pp. 123–140. issn: 0885-6125.

[Bre96b] Leo Breiman. *Out-of-bag estimation*. Report. Citeseer, 1996.

[Bre98] Leo Breiman. "Arcing classifier (with discussion and a rejoinder by the author)". In: *The annals of statistics* 26.3 (1998), pp. 801–849. issn: 0090-5364.

[Bry75] Arthur Earl Bryson. *Applied optimal control: optimization, estimation and control*. CRC Press, 1975. isbn: 0891162283.

[BS08] Harald Binder and Martin Schumacher. "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models". In: *BMC bioinformatics* 9.1 (2008), p. 14. issn: 1471-2105.

[BS09] Harald Binder and Martin Schumacher. "Incorporating pathway information into boosting estimation of high-dimensional risk prediction models". In: *BMC bioinformatics* 10.1 (2009), p. 18. issn: 1471-2105.

[Bur] Lukas Burget. "Combination of speech features using smoothed heteroscedastic linear discriminant analysis". In: *INTERSPEECH*.

[Bur10] Christopher JC Burges. "From ranknet to lambdarank to lambdamart: An overview". In: *Learning* 11 (2010), pp. 23–581.

[BY02] Peter Büchlmann and Bin Yu. "Analyzing bagging". In: *Annals of Statistics* (2002), pp. 927–961. issn: 0090-5364.

[Car01] G. A. Carpenter. "Neural network models of learning and memory: Leading questions and an emerging framework". In: *Trends in Cognitive Sciences* 5.3 (2001). PMID: 11239811, pp. 114–118. issn: 1364-6613.

[CB06] Mingmin Chi and Lorenzo Bruzzone. "An ensemble-driven k-NN approach to ill-posed classification problems". In: *Pattern recognition letters* 27.4 (2006), pp. 301–307. issn: 0167-8655.

[CG87] Gail A Carpenter and Stephen Grossberg. "A massively parallel architecture for a self-organizing neural pattern recognition machine". In: *Computer vision, graphics, and image processing* 37.1 (1987), pp. 54–115. issn: 0734-189X.

[CG88] Gail A Carpenter and Stephen Grossberg. "The ART of adaptive pattern recognition by a self-organizing neural network". In: *Computer* 21.3 (1988), pp. 77–88. issn: 0018-9162.

[CGM02] Hugh A Chipman, Edward I George, and Robert E McCulloch. "Bayesian treed models". In: *Machine Learning* 48.1-3 (2002), pp. 299–320. issn: 0885-6125.

[CGM07] Hugh A Chipman, Edward I George, and Robert E McCulloch. "Bayesian ensemble learning". In: *Advances in neural information processing systems* 19 (2007), p. 265. issn: 1049-5258.

[CGM10] Hugh A Chipman, Edward I George, and Robert E McCulloch. "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* (2010), pp. 266–298. issn: 1932-6157.

[CGM98] Hugh A Chipman, Edward I George, and Robert E McCulloch. "Bayesian CART model search". In: *Journal of the American Statistical Association* 93.443 (1998), pp. 935–948. issn: 0162-1459.

[CH92] Gregory F Cooper and Edward Herskovits. "A Bayesian method for the induction of probabilistic networks from data". In: *Machine learning* 9.4 (1992), pp. 309–347. issn: 0885-6125.

[Cha+05] Howard Y Chang et al. "Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.10 (2005), pp. 3738–3743. issn: 0027-8424.

[Che+] De-Yuan Cheng et al. "Fast search algorithms for vector quantization and pattern matching". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.* Vol. 9. IEEE, pp. 372–375.

[Che+09] Federico Chesani et al. "Exploiting inductive logic programming techniques for declarative process mining". In: *Transactions on Petri Nets and Other Models of Concurrency II*. Springer, 2009, pp. 278–295. isbn: 3642008984.

[CJM16] Mark Culp, Kjell Johnson, and George Michailidis. *ada: The R Package Ada for Stochastic Boosting*. R package version 2.0-5. 2016. url: https://CRAN.R-project.org/package=ada.

[CL11] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: a library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011), p. 27. issn: 2157-6904.

[CL68] CK Chow and CN Liu. "Approximating discrete probability distributions with dependence trees". In: *Information Theory, IEEE Transactions on* 14.3 (1968), pp. 462–467. issn: 0018-9448.

[CL94] Fu Lai Chung and Tong Lee. "Fuzzy competitive learning". In: *Neural Networks* 7.3 (1994), pp. 539–551. issn: 0893-6080.

[Coh] William W Cohen. "Fast effective rule induction". In: *Proceedings of the twelfth international conference on machine learning*, pp. 115–123.

[Cor+09] Paulo Cortez et al. "Modeling wine preferences by data mining from physicochemical properties". In: *Decision Support Systems* 47.4 (2009), pp. 547–553. issn: 0167-9236.

[Cor15] Paulo Cortez. "A tutorial on using the rminer R package for data mining tasks". In: (2015).

[CP09] Emmanuel J Candès and Yaniv Plan. "Near-ideal model selection by l1 minimization". In: *The Annals of Statistics* 37.5A (2009), pp. 2145–2177. issn: 0090-5364.

[CS02] Koby Crammer and Yoram Singer. "On the learnability and design of output codes for multiclass problems". In: *Machine learning* 47.2-3 (2002), pp. 201–233. issn: 0885-6125.

[CW86] PJ Curran and HD Williamson. "Sample size for ground and remotely sensed data". In: *Remote sensing of environment* 20.1 (1986), pp. 31–41. issn: 0034-4257.

[CZ01] Adele Cutler and Guohua Zhao. "Pert-perfect random tree ensembles". In: *Computing Science and Statistics* 33 (2001), pp. 490–497.

[Dav90] Rajesh N Dave. "Fuzzy shell-clustering and applications to circle detection in digital images". In: *International Journal Of General System* 16.4 (1990), pp. 343–355. issn: 0308-1079.

[DdA05] Ramon Diaz-Uriarte and Sara Alvarez de Andrés. "Variable selection from random forests: application to gene expression data". In: *arXiv preprint q-bio/0503025* (2005).

[DGK04] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. *A unified view of kernel k-means, spectral clustering and graph cuts*. Citeseer, 2004.

[DiM+10] Peter A DiMaggio et al. "A Novel Framework for Predicting in vivo Toxicities from in vitro Data using Optimal Methods for Dense and Sparse Matrix Re-ordering and Logistic Regression". In: *Toxicological Sciences* (2010), kfq233. issn: 1096-6080.

[Dol+12] Sara Dolnicar et al. "Biclustering overcoming data dimensionality problems in market segmentation". In: *Journal of Travel Research* 51.1 (2012), pp. 41–49. issn: 0047-2875.

[Don06] David L Donoho. "Compressed sensing". In: *Information Theory, IEEE Transactions on* 52.4 (2006), pp. 1289–1306. issn: 0018-9448.

[DP97] Pedro Domingos and Michael Pazzani. "On the optimality of the simple Bayesian classifier under zero-one loss". In: *Machine learning* 29.2-3 (1997), pp. 103–130. issn: 0885-6125.

[Efr+04] Bradley Efron et al. "Least angle regression". In: *The Annals of statistics* 32.2 (2004), pp. 407–499. issn: 0090-5364.

[EFS05a] EFSA Panel on Dietetic Products, Nutrition and Allergies. "Opinion of the Scientific Panel on Dietetic products, nutrition and allergies [NDA] related to an application on the use of a-tocopherol-containing oil suspension of lycopene from Blakeslea trispora as a novel food ingredient." In: *EFSA Journal*, 212 (2005), p. 29. doi: 10.2903/j.efsa.2005.212.

[EFS05b] EFSA Panel on Dietetic Products, Nutrition and Allergies. "Opinion of the Scientific Panel on Dietetic products, nutrition and allergies [NDA] related to the safety and suitability for particular nutritional use by infants of formula based on whey protein partial hydrolysates with a protein content of at least 1.9 g protein/100 kcal". In: *EFSA Journal*, 280 (2005), p. 16. doi: 10.2903/j.efsa.2005.280.

[Elm90] Jeffrey L Elman. "Finding structure in time". In: *Cognitive science* 14.2 (1990), pp. 179–211. issn: 0364-0213.

[EM96] Paul HC Eilers and Brian D Marx. "Flexible smoothing with B-splines and penalties". In: *Statistical science* (1996), pp. 89–102. issn: 0883-4237.

[Ere+13] Kemal Eren et al. "A comparative analysis of biclustering algorithms for gene expression data". In: *Briefings in bioinformatics* 14.3 (2013), pp. 279–292. issn: 1467-5463.

[Est+] Martin Ester, Hans-Peter Kriegel, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: *Kdd*. Vol. 96, pp. 226–231.

[Est+96] Martin Ester, Hans-peter Kriegel, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise". In: AAAI Press, 1996, pp. 226–231.

[Fan+08] Rong-En Fan et al. "LIBLINEAR: A library for large linear classification". In: *The Journal of Machine Learning Research* 9 (2008), pp. 1871–1874. issn: 1532-4435.

[Far05] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press, 2005. isbn: 0203492285.

[Fau94] Laurene Fausett. "Fundamentals of neural networks: architectures, algorithms, and applications". In: (1994).

[FG99] Jason P Fine and Robert J Gray. "A proportional hazards model for the subdistribution of a competing risk". In: *Journal of the American statistical association* 94.446 (1999), pp. 496–509. issn: 0162-1459.

[FGG97] Nir Friedman, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers". In: *Machine learning* 29.2-3 (1997), pp. 131–163. issn: 0885-6125.

[FHT00] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)". In: *The annals of statistics* 28.2 (2000), pp. 337–407. issn: 0090-5364.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin, 2001.

[FHT10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1 (2010), p. 1.

[Fis87] Douglas H Fisher. "Knowledge acquisition via incremental conceptual clustering". In: *Machine learning* 2.2 (1987), pp. 139–172. issn: 0885-6125.

[FL01] Peter A Flach and Nicolas Lachiche. "Confirmation-guided discovery of first-order rules with Tertius". In: *Machine Learning* 42.1-2 (2001), pp. 61–95. issn: 0885-6125.

[FM04a] Jerome H Friedman and Jacqueline J Meulman. "Clustering objects on subsets of attributes (with discussion)". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66.4 (2004), pp. 815–849. issn: 1467-9868.

[FM04b] Glenn M Fung and Olvi L Mangasarian. "A feature selection Newton method for support vector machine classification". In: *Computational optimization and applications* 28.2 (2004), pp. 185–202. issn: 0926-6003.

[Fri01] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232. issn: 0090-5364.

[Fri02] Jerome H Friedman. "Stochastic gradient boosting". In: *Computational Statistics & Data Analysis* 38.4 (2002), pp. 367–378. issn: 0167-9473.

[Fri91] J. H. Friedman. "Multivariate adaptive regression splines (with discussion)". In: *Annals of Statistics* 19 (1991), pp. 1–141.

[FS] Yoav Freund and Robert E Schapire. "Experiments with a new boosting algorithm". In: *ICML*. Vol. 96, pp. 148–156.

[FS89] Jerome H Friedman and Bernard W Silverman. "Flexible parsimonious smoothing and additive modeling". In: *Technometrics* 31.1 (1989), pp. 3–21. issn: 0040-1706.

[FS97] Yoav Freund and Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139. issn: 0022-0000.

[FW] Eibe Frank and Ian H Witten. "Generating accurate rule sets without global optimization". In: *ICML*. Vol. 98, pp. 144–151.

[GHT05] Yaqian Guo, Trevor Hastie, and Robert Tibshirani. "Regularized discriminant analysis and its application in microarrays". In: *Biostatistics* 1.1 (2005), pp. 1–18.

[GL08] Robert B Gramacy and Herbert KH Lee. "Gaussian processes and limiting linear models". In: *Computational Statistics & Data Analysis* 53.1 (2008), pp. 123–136. issn: 0167-9473.

[GL09] Robert B Gramacy and Herbert KH Lee. "Adaptive design and analysis of supercomputer experiments". In: *Technometrics* 51.2 (2009), pp. 130–145. issn: 0040-1706.

[GL12] Robert B Gramacy and Heng Lian. "Gaussian process single-index models as emulators for computer experiments". In: *Technometrics* 54.1 (2012), pp. 30–41. issn: 0040-1706.

[GLF89] John H Gennari, Pat Langley, and Doug Fisher. "Models of incremental concept formation". In: *Artificial intelligence* 40.1-3 (1989), pp. 11–61. issn: 0004-3702.

[GN08] Gérard Govaert and Mohamed Nadif. "Block clustering with Bernoulli mixture models: Comparison of different approaches". In: *Computational Statistics & Data Analysis* 52.6 (2008), pp. 3233–3245. issn: 0167-9473.

[GN13] Gérard Govaert and Mohamed Nadif. *Co-clustering*. John Wiley & Sons, 2013. isbn: 1118649508.

[Goe10] Jelle J Goeman. "L1 penalized estimation in the Cox proportional hazards model". In: *Biometrical journal* 52.1 (2010), pp. 70–84. issn: 1521-4036.

[Gra07] Robert B Gramacy. "tgp: an R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models". In: *Journal of Statistical Software* 19.9 (2007), p. 6.

[Gra88] Robert J Gray. "A class of K-sample tests for comparing the cumulative incidence of a competing risk". In: *The Annals of statistics* (1988), pp. 1141–1154. issn: 0090-5364.

[Gro76] Stephen Grossberg. "Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors". In: *Biological cybernetics* 23.3 (1976), pp. 121–134. issn: 0340-1200.

[GS11a] Anders Gorst-Rasmussen and Thomas Scheike. *Coordinate descent methods for the penalized semiparametric additive hazards model*. Report. Department of Mathematical Sciences, Aalborg University, 2011.

[GS11b] Anders Gorst-Rasmussen and Thomas H Scheike. "Independent screening for single-index hazard rate models with ultra-high dimensional features". In: *arXiv preprint arXiv:1105.3361* (2011).

[GS74] M Goldstein and Adrian FM Smith. "Ridge-type estimators for regression analysis". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1974), pp. 284–291. issn: 0035-9246.

[GT12] Andreas Groll and Gerhard Tutz. "Regularization for generalized additive mixed models by likelihood-based boosting". In: *Methods of Information in Medicine* 51.2 (2012), p. 168. issn: 0026-1270.

[GT14] Robert B Gramacy and Matthew Taddy. "Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an R package for treed Gaussian process models". In: *Journal of Statistical Software* 33.i06 (2014).

[Guy+98] Isabelle Guyon et al. "What size test set gives good error rate estimates?" In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 20.1 (1998), pp. 52–64. issn: 0162-8828.

[GZP11] Thomas Grubinger, Achim Zeileis, and Karl-Peter Pfeiffer. *evtree: Evolutionary learning of globally optimal classification and regression trees in R*. Report. Working Papers in Economics and Statistics, 2011.

[Har+82] Jr. Harrell F. E. et al. "Evaluating the yield of medical tests". In: *JAMA* 247.18 (1982), pp. 2543–6. issn: 0098-7484 (Print) 0098-7484 (Linking). url: http://www.ncbi.nlm.nih.gov/pubmed/7069920%20http://jama.jamanetwork.com/article.aspx?articleid=372568.

[HBL11] Ahlem Hajjem, François Bellavance, and Denis Larocque. "Mixed effects regression trees for clustered data". In: *Statistics & probability letters* 81.4 (2011), pp. 451–459. issn: 0167-7152.

[HBM12] Jian Huang, Patrick Breheny, and Shuangge Ma. "A selective review of group selection in high-dimensional models". In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 27.4 (2012).

[Her92] Kai-Uwe Herrmann. "ART-Adaptive Resonance Theory-Architekturen, Implementierung und Anwendung". Thesis. 1992.

[HHP99] Geoffrey Holmes, Mark Hall, and Eibe Prank. *Generating rule sets from model trees*. Springer, 1999. isbn: 3540668225.

[HHZ06a] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. "Unbiased recursive partitioning: A conditional inference framework". In: *Journal of Computational and Graphical statistics* 15.3 (2006), pp. 651–674. issn: 1061-8600.

[HHZ06b] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. "Unbiased recursive partitioning: A conditional inference framework". In: *Journal of Computational and Graphical statistics* 15.3 (2006), pp. 651–674. issn: 1061-8600.

[HK70] Arthur E. Hoerl and Robert W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[HL02] Chih-Wei Hsu and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines". In: *Neural Networks, IEEE Transactions on* 13.2 (2002), pp. 415–425. issn: 1045-9227.

[HL03a] Torsten Hothorn and Berthold Lausen. "Double-bagging: Combining classifiers by bootstrap aggregation". In: *Pattern Recognition* 36.6 (2003), pp. 1303–1309. issn: 0031-3203.

[HL03b] Torsten Hothorn and Berthold Lausen. "On the exact distribution of maximally selected rank statistics". In: *Computational Statistics & Data Analysis* 43.2 (2003), pp. 121–137. issn: 0167-9473.

[HL05] Torsten Hothorn and Berthold Lausen. "Bundling classifiers by bagging trees". In: *Computational Statistics & Data Analysis* 49.4 (2005), pp. 1068–1078. issn: 0167-9473.

[HLA01] David J Hand, Hua Gui Li, and Niall M Adams. "Supervised classification with structured class definitions". In: *Computational Statistics & Data Analysis* 36.2 (2001), pp. 209–225. issn: 0167-9473.

[HMS14] Benjamin Hofner, Andreas Mayr, and Matthias Schmid. "gamboostLSS: An R package for model building and variable selection in the GAMLSS framework". In: *arXiv preprint arXiv:1407.1774* (2014).

[HO00] Aapo Hyvärinen and Erkki Oja. "Independent component analysis: algorithms and applications". In: *Neural networks* 13.4 (2000), pp. 411–430. issn: 0893-6080.

[Hol93] Robert C Holte. "Very simple classification rules perform well on most commonly used datasets". In: *Machine learning* 11.1 (1993), pp. 63–90. issn: 0885-6125.

[Hot+04] Torsten Hothorn, Berthold Lausen, et al. "Bagging survival trees". In: *Statistics in medicine* 23.1 (2004), pp. 77–91. issn: 0277-6715.

[Hot+06] Torsten Hothorn, Peter Bühlmann, et al. "Survival ensembles". In: *Biostatistics* 7.3 (2006), pp. 355–373. issn: 1465-4644.

[Hot+12] Torsten Hothorn, Kurt Hornik, Mark A Van De Wiel, et al. "A Lego system for conditional inference". In: *The American Statistician* (2012).

[Hou+08] E Andres Houseman et al. "Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions". In: *BMC bioinformatics* 9.1 (2008), p. 1. issn: 1471-2105.

[HP90] Trevor Hastie and Daryl Pregibon. *Shrinking trees*. AT & T Bell Laboratories, 1990.

[HS85] Dorit S Hochbaum and David B Shmoys. "A best possible heuristic for the k-center problem". In: *Mathematics of operations research* 10.2 (1985), pp. 180–184. issn: 0364-765X.

[HST98] Clifford M Hurvich, Jeffrey S Simonoff, and Chih-Ling Tsai. "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.2 (1998), pp. 271–293. issn: 1467-9868.

[HSV] A Hyvarinen, J Sarela, and Ricardo Vigário. "Spikes and bumps: Artefacts generated by independent component analysis with insufficient sample size". In: *First International Workshop on Independent Component Analysis and Signal Separation*.

[HT90] Clifford M Hurvich and Chih Ling Tsai. "Model selection for least absolute deviations regression in small samples". In: *Statistics & probability letters* 9.3 (1990), pp. 259–265. issn: 0167-7152.

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

[Hua] Zhexue Huang. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining". In: *DMKD*.

[Hua+07] Chien-Ming Huang et al. "Model selection for support vector machines via uniform design". In: *Computational Statistics & Data Analysis* 52.1 (2007), pp. 335–346. issn: 0167-9473.

[Hua98] Zhexue Huang. "Extensions to the k-means algorithm for clustering large data sets with categorical values". In: *Data mining and knowledge discovery* 2.3 (1998), pp. 283–304. issn: 1384-5810.

[HVD01] Javier Herrero, Alfonso Valencia, and Joaquın Dopazo. "A hierarchical unsupervised growing neural network for clustering gene expression patterns". In: *Bioinformatics* 17.2 (2001), pp. 126–136. doi: 10.1093/bioinformatics/17.2.126. url: http://bioinformatics.oxfordjournals.org/content/17/2/126.abstract.

[HZ14] Torsten Hothorn and Achim Zeileis. *partykit: A modular toolkit for recursive partytioning in R*. Report. Working Papers in Economics and Statistics, 2014.

[IP08] EFSA FIP (Food Ingredients and Packaging). "Assessment of the results of the study by McCann et al. (2007) on the effect of some colours and sodium benzoate on childrens behaviour". In: *EFSA Journal*, 660 (2008), p. 139. doi: 10.2903/j.efsa.2008.660.

[IP09] EFSA FIP (Food Ingredients and Packaging). "The use of taurine and D-glucurono-gamma-lactone as constituents of the so-called energy drinks". In: *EFSA Journal*, 935 (2009), p. 31. doi: 10.2903/j.efsa.2009.935.

[IP10a] EFSA FIP (Food Ingredients and Packaging). "Flavouring Group Evaluation 32 (FGE.32): Flavonoids (Flavanones and dihydrochalcones) from chemical groups 25 and 30". In: *EFSA Journal* 8.9, 1065 (2010), p. 61. doi: 10.2903/j.efsa.2010.1065.

[IP10b] EFSA FIP (Food Ingredients and Packaging). "Scientific Opinion on Flavouring Group Evaluation 90 (FGE.90): Consideration of Aliphatic, acyclic and alicyclic terpenoid tertiary alcohols and structurally related substances evaluated by JECFA (68th meeting) structurally related to aliphatic, alicyclic". In: *EFSA Journal* 8.2, 1336 (2010), p. 32. doi: 10.2903/j.efsa.2010.1336.

[Ira93] Keki B Irani. "Multi-interval discretization of continuous-valued attributes for classification learning". In: (1993).

[Ish+08] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, et al. "Random survival forests". In: *The annals of applied statistics* (2008), pp. 841–860. issn: 1932-6157.

[Ish+10] Hemant Ishwaran, Udaya B Kogalur, Eiran Z Gorodeski, et al. "High-dimensional variable selection for survival data". In: *Journal of the American Statistical Association* 105.489 (2010), pp. 205–217. issn: 0162-1459.

[Ish+11] Hemant Ishwaran, Udaya B Kogalur, Xi Chen, et al. "Random survival forests for high-dimensional data". In: *Statistical analysis and data mining* 4.1 (2011), pp. 115–132. issn: 1932-1872.

[Ish+14] Hemant Ishwaran, Thomas A Gerds, et al. "Random survival forests for competing risks". In: *Biostatistics* 15.4 (2014), pp. 757–773. issn: 1465-4644.

[Ish07] Hemant Ishwaran. "Variable importance in binary regression trees and forests". In: *Electronic Journal of Statistics* 1 (2007), pp. 519–537. issn: 1935-7524.

[Ish15] Hemant Ishwaran. "The effect of splitting on random forests". In: *Machine Learning* 99.1 (2015), pp. 75–118. issn: 0885-6125.

[Joe03] S Joe Qin. "Statistical process monitoring: basics and beyond". In: *Journal of chemometrics* 17.8-9 (2003), pp. 480–502. issn: 1099-128X.

[Joh] George H John. "Robust Decision Trees: Removing Outliers from Databases". In: *KDD*, pp. 174–179.

[Jor86] MI Jordan. *Serial order: a parallel distributed processing approach. Technical report, June 1985-March 1986*. Report. California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science, 1986.

[KA98] Nagendra Kumar and Andreas G Andreou. "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition". In: *Speech communication* 26.4 (1998), pp. 283–297. issn: 0167-6393.

[KBC] Teuvo Kohonen, György Barna, and Ronald Chrisley. "Statistical pattern recognition with neural networks: Benchmarking studies". In: *Neural Networks, 1988., IEEE International Conference on*. IEEE, pp. 61–68.

[KFS03] Kwang In Kim, Matthias O Franz, and Bernhard Schölkopf. "Kernel Hebbian algorithm for iterative kernel principal component analysis". In: (2003).

[KJ95] Juha Karhunen and Jyrki Joutsensalo. "Generalizations of principal component analysis, optimization problems, and neural networks". In: *Neural Networks* 8.4 (1995), pp. 549–562. issn: 0893-6080.

[Koh] Teuvo Kohonen. "Improved versions of learning vector quantization". In: *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*. IEEE, pp. 545–550.

[Koh12] Teuvo Kohonen. *Self-organization and associative memory*. Vol. 8. Springer Science & Business Media, 2012. isbn: 3642881637.

[KR05] Charles Kooperberg and Ingo Ruczinski. "Identifying interacting SNPs using Monte Carlo logic regression". In: *Genetic epidemiology* 28.2 (2005), pp. 157–170. issn: 1098-2272.

[KR10] Miron B. Kursa and Witold R. Rudnicki. "Feature Selection with the Boruta Package". In: *Journal of Statistical Software* 36.11 (2010), pp. 1–13. url: http://www.jstatsoft.org/v36/i11/.

[Kri07] Brian Kriegler. *Cost-sensitive stochastic gradient boosting within a quantitative regression framework*. University of California at Los Angeles, 2007. isbn: 0549130845.

[Kro00] Siegfried Kropf. *Hochdimensionale multivariate Verfahren in der medizinischen Statistik*. Shaker, 2000. isbn: 3826582055.

[KS98] Teuvo Kohonen and Panu Somervuo. "Self-organizing maps of symbol strings". In: *Neurocomputing* 21.1 (1998), pp. 19–30. issn: 0925-2312.

[KSW04] Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. "Online learning with kernels". In: *Signal Processing, IEEE Transactions on* 52.8 (2004), pp. 2165–2176. issn: 1053-587X.

[KY90] Fang Kai-Tai and Zhang Yao-Ting. *Generalized multivariate analysis*. Science Press, 1990. isbn: 3540176519.

[KZP07] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. "Machine learning: a review of classification and combining techniques". In: *Artificial Intelligence Review* 26.3 (2007), pp. 159–190. issn: 1573-7462. doi: 10.1007/s10462-007-9052-3. url: http://dx.doi.org/10.1007/s10462-007-9052-3.

[Lan03] N Landwehr. "Logistic Model Trees, Master's Thesis". In: *Institute for Computer Science, University of Freiburg, Germany* (2003).

[Lar+06] Pedro Larrañaga et al. "Machine learning in bioinformatics". In: *Briefings in bioinformatics* 7.1 (2006), pp. 86–112. issn: 1467-5463.

[Läu92] J Läuter. "Stabile multivariate verfahren". In: *Diskriminanzanalyse, Regressionsanalyse, Faktor* (1992).

[LC93] Michael LeBlanc and John Crowley. "Survival trees by goodness of split". In: *Journal of the American Statistical Association* 88.422 (1993), pp. 457–467. issn: 0162-1459.

[Lea+05] JR Leathwick et al. "Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish". In: *Freshwater Biology* 50.12 (2005), pp. 2034–2052. issn: 1365-2427.

[Lei99] Friedrich Leisch. "Bagged clustering". In: (1999).

[LGK98] Jurgen Lauter, Ekkehard Glimm, and Siegfried Kropf. "Multivariate tests based on left-spherically distributed linear scores". In: *Annals of Statistics* (1998), pp. 1972–1988. issn: 0090-5364.

[LHF03] Niels Landwehr, Mark Hall, and Eibe Frank. "Logistic Model Trees". In: *Machine Learning: ECML 2003: 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings*. Ed. by Nada Lavrač et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 241–252. isbn: 978-3-540-39857-8. doi: 10.1007/978-3-540-39857-8_23. url: http://dx.doi.org/10.1007/978-3-540-39857-8_23.

[LHF05] Niels Landwehr, Mark Hall, and Eibe Frank. "Logistic model trees". In: *Machine Learning* 59.1-2 (2005), pp. 161–205. issn: 0885-6125.

[Lia+15] Zhu Liang et al. "Fault detection and diagnosis of belt weigher using improved DBSCAN and Bayesian Regularized Neural Network". In: *Mechanics* 21.1 (2015), pp. 70–77. issn: 2029-6983.

[LJ06] Yi Lin and Yongho Jeon. "Random forests and adaptive nearest neighbors". In: *Journal of the American Statistical Association* 101.474 (2006), pp. 578–590. issn: 0162-1459.

[LK] Jimmy Lin and Alek Kolcz. "Large-scale machine learning at twitter". In: *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, pp. 793–804. isbn: 1450312470.

[LLW07] Hsuan-Tien Lin, Chih-Jen Lin, and Ruby C Weng. "A note on Platt's probabilistic outputs for support vector machines". In: *Machine learning* 68.3 (2007), pp. 267–276. issn: 0885-6125.

[LM07] Chenlei Leng and Shuangge Ma. "Path consistent model selection in additive risk model via Lasso". In: *Statistics in medicine* 26.20 (2007), pp. 3753–3770. issn: 1097-0258.

[Lok99] Justin Lokhorst. "The lasso and generalised linear models". In: *Honors Project, The University of Adelaide, Australia* (1999).

[Lov68] Julie B Lovins. *Development of a stemming algorithm*. MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.

[LS97] Wei-Yin Loh and Yu-Shan Shih. "Split selection methods for classification trees". In: *Statistica sinica* (1997), pp. 815–840. issn: 1017-0405.

[LY94] DY Lin and Zhiliang Ying. "Semiparametric analysis of the additive risk model". In: *Biometrika* 81.1 (1994), pp. 61–71. issn: 0006-3444.

[Mac] James MacQueen. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. Oakland, CA, USA., pp. 281–297.

[Mal+12] James D Malley et al. "Probability machines: consistent probability estimation using nonparametric learning machines". In: *Methods of Information in Medicine* 51.1 (2012), p. 74.

[Mas65] W.F. Massy. "Principal components regression in explanatory statistical research". In: *Journal of American Statistical Association* 60.4 (1965), pp. 234–256.

[May+12] Andreas Mayr et al. "Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61.3 (2012), pp. 403–427. issn: 1467-9876.

[MB10] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010), pp. 417–473. issn: 1467-9868.

[Mei06] Nicolai Meinshausen. "Quantile regression forests". In: *The Journal of Machine Learning Research* 7 (2006), pp. 983–999. issn: 1532-4435.

[Mei07] Nicolai Meinshausen. "Relaxed lasso". In: *Computational Statistics & Data Analysis* 52.1 (2007), pp. 374–393. issn: 0167-9473.

[MIG12] Ulla B Mogensen, Hemant Ishwaran, and Thomas A Gerds. "Evaluating random forests for survival analysis using prediction error curves". In: *Journal of statistical software* 50.11 (2012), p. 1.

[Mil02] Alan Miller. *Subset selection in regression*. CRC Press, 2002. isbn: 1420035932.

[MMB12] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. "P-values for high-dimensional regression". In: *Journal of the American Statistical Association* (2012).

[MS09] Torben Martinussen and Thomas H Scheike. "Covariate selection for the semiparametric additive risk model". In: *Scandinavian Journal of Statistics* 36.4 (2009), pp. 602–619. issn: 1467-9469.

[Muk+03] S. Mukherjee et al. "Estimating dataset size requirements for classifying DNA microarray data". In: *Journal of Computational Biology* 10.2 (2003), pp. 119–142. doi: 10.1089/106652703321825928. url: http://www.scopus.com/inward/record.url?eid=2-s2.0-0038237368&partnerID=40&md5=46d6efff0b62706958b0183e20ec49fe.

[MVB08] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. "The group lasso for logistic regression". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1 (2008), pp. 53–71. issn: 1467-9868.

[MW] Samuel A Mulder and Donald C Wunsch. "Using adaptive resonance theory and local optimization to divide and conquer large scale traveling salesman problems". In: *Neural Networks, 2003. Proceedings of the International Joint Conference on*. Vol. 2. IEEE, pp. 1408–1411. isbn: 0780378989.

[MWB06] Willem Melssen, Ron Wehrens, and Lutgarde Buydens. "Supervised Kohonen networks for classification problems". In: *Chemometrics and Intelligent Laboratory Systems* 83.2 (2006), pp. 99–113. issn: 0169-7439. url: http://www.sciencedirect.com/science/article/pii/S016974390600027X.

[NH02] Raymond T Ng and Jiawei Han. "CLARANS: A method for clustering objects for spatial data mining". In: *Knowledge and Data Engineering, IEEE Transactions on* 14.5 (2002), pp. 1003–1016. issn: 1041-4347.

[NJ09] Thomas Dyhre Nielsen and Finn Verner Jensen. *Bayesian networks and decision graphs*. Springer Science & Business Media, 2009. isbn: 0387682821.

[oAoS05] EFSA Panel on Additives and Products or Substances used in Animal Feed. "Opinion of the Scientific Panel on additives and products or substances used in animal feed (FEEDAP) on the safety of the enzyme preparation Bio-Feed Combi for use as feed additive for chickens for fattening and piglets". In: *EFSA Journal*, 261 (2005), p. 6. doi: 10.2903/j.efsa.2005.261.

[oAW04] EFSA Panel on Animal Health and Welfare. "Opinion of the Scientific Panel on Animal Health and Welfare (AHAW) on a request from the Commission related to the risk of transmission of Mycobacterium avium subsp. paratuberculosis via bovine semen". In: *EFSA Journal*, 110 (2004), p. 59. doi: 10.2903/j.efsa.2004.110.

[oAW06] EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare). "Opinion of the Scientific Panel on Animal Health and Welfare (AHAW) on a request from the Commission related with animal health and welfare risks associated with the import of wild birds other than poultry into the European Union". In: *EFSA Journal*, 410 (2006), p. 218. doi: 10.2903/j.efsa.2006.410.

[oAW07a] EFSA Panel on Animal Health and Welfare. "Animal Welfare aspects of the killing and skinning of seals". In: *EFSA Journal*, 610 (2007), p. 165. doi: 10.2903/j.efsa.2007.610.

[oAW07b] EFSA Panel on Animal Health and Welfare. "Opinion of the Scientific Panel on Animal Health and Welfare (AHAW) on a request from the Commission concerning Brucellosis Diagnostic Methods for Bovines, Sheep, and Goats". In: *EFSA Journal*, 432 (2007), p. 249. doi: 10.2903/j.efsa.2007.432.

[oAW07c] EFSA Panel on Animal Health and Welfare. "Opinion of the Scientific Panel on Animal Health and Welfare (AHAW) on the Framework for EFSA AHAW Risk assessments". In: *EFSA Journal*, 550 (2007), p. 46. doi: 10.2903/j.efsa.2007.550.

[oAW07d] EFSA Panel on Animal Health and Welfare. "The risks associated with tail biting in pigs and possible means to reduce the need for tail docking considering the different housing and husbandry systems". In: *EFSA Journal*, 611 (2007), p. 109. doi: 10.2903/j.efsa.2007.611.

[oAW08] EFSA Panel on Animal Health and Welfare. "Tuberculosis testing in deer". In: *EFSA Journal*, 645 (2008), p. 200. doi: 10.2903/j.efsa.2008.645.

[oAW09] EFSA Panel on Animal Health and Welfare. "Guidance on Good Practice in Conducting Scientific Assessments in Animal Health using Modelling". In: *EFSA Journal* 7.12, 1419 (2009), p. 38. doi: 10.2903/j.efsa.2009.1419.

[oBio05] EFSA Panel on Biological Hazards. "Opinion of the Scientific Panel on biological hazards (BIOHAZ) on the Quantitative risk assessment of the animal BSE risk posed by meat and bone meal with respect to the residual BSE risk". In: *EFSA Journal*, 257 (2005), p. 30. doi: 10.2903/j.efsa.2005.257.

[oBio08] EFSA Panel on Biological Hazards. "Microbiological risk assessment in feedingstuffs for food-producing animals". In: *EFSA Journal*, 720 (2008), p. 84. doi: 10.2903/j.efsa.2008.720.

[oBio09] EFSA Panel on Biological Hazards. "Quantitative estimation of the impact of setting a new target for the reduction of Salmonella in breeding hens of Gallus gallus". In: *EFSA Journal*, 1036 (2009), p. 114. doi: 10.2903/j.efsa.2009.1036.

[oCon05a] EFSA Panel on Contaminants in the Food Chain. "Opinion of the Scientific Panel on contaminants in the food chain [CONTAM] related to the presence of non dioxin-like polychlorinated biphenyls (PCB) in feed and food". In: *EFSA Journal*, 284 (2005), p. 262. doi: 10.2903/j.efsa.2005.284.

[oCon05b] EFSA Panel on Contaminants in the Food Chain. "Opinion of the Scientific Panel on contaminants in the food chain [CONTAM] related to the safety assessment of wild and farmed fish". In: *EFSA Journal*, 236 (2005), p. 118. doi: 10.2903/j.efsa.2005.236.

[oCon08] EFSA Panel on Contaminants in the Food Chain. "Perfluorooctane sulfonate (PFOS), perfluorooctanoic acid (PFOA) and their salts". In: *EFSA Journal*, 653 (2008), p. 131. doi: 10.2903/j.efsa.2008.653.

[oCon09a] EFSA Panel on Contaminants in the Food Chain. "Cadmium in food". In: *EFSA Journal*, 980 (2009), p. 139. doi: 10.2903/j.efsa.2009.980.

[oCon09b] EFSA Panel on Contaminants in the Food Chain. "Scientific Opinion on Arsenic in Food". In: *EFSA Journal* 7.10, 1351 (2009), p. 199. doi: 10.2903/j.efsa.2009.1351.

[oCon09c] EFSA Panel on Contaminants in the Food Chain. "Uranium in foodstuffs, in particular mineral water". In: *EFSA Journal*, 1018 (2009), p. 59. doi: 10.2903/j.efsa.2009.1018.

[oGen10] EFSA Panel on Genetically Modified Organisms. "Statistical considerations for the safety evaluation of GMOs". In: *EFSA Journal* 8.1, 1250 (2010), p. 59. doi: 10.2903/j.efsa.2010.1250.

[Oja91] E Oja. "Learning in nonlinear constrained Hebbian networks". In: *Artificial neural networks* (1991), pp. 385–390.

[oPla09a] EFSA Panel on Plant Health. "Mortality verification of pinewood nematode from high temperature treatment of shavings". In: *EFSA Journal*, 1055 (2009), p. 19. doi: 10.2903/j.efsa.2009.1055.

[oPla09b] EFSA Panel on Plant Health. "Pest risk assessment and additional evidence provided by South Africa on Guignardia citricarpa Kiely, citrus black spot fungus CBS". In: *EFSA Journal*, 925 (2009), p. 153. doi: 10.2903/j.efsa.2009.925.

[OPT00] Michael R Osborne, Brett Presnell, and Berwin A Turlach. "On the lasso and its dual". In: *Journal of Computational and Graphical statistics* 9.2 (2000), pp. 319–337. issn: 1061-8600.

[OPT98] Michael R Osborne, Brett Presnell, and Berwin A Turlach. "Knot selection for regression splines via the lasso". In: *Computing Science and Statistics* (1998), pp. 44–49.

[oPtR05] EFSA Panel on Plant Protection Products and their Residues. "Opinion of the PPR Panel related to the appropriate variability factor(s) to be used for acute dietary exposure assessment of pesticide residues in fruit and vegetables." In: *EFSA Journal*, 177 (2005), p. 61. doi: 10.2903/j.efsa.2005.177.

[oPtR06] EFSA Panel on Plant Protection Products and their Residues. "Opinion of the Scientific Panel on Plant protection products and their residues (PPR) a request from the Commission on the Guidance Document on Estimating Persistence and Degradation Kinetics from Environmental Fate Studies on Pesticides in EU". In: *EFSA Journal*, 300 (2006), p. 13. doi: 10.2903/j.efsa.2006.300.

[oPtR08a] EFSA Panel on Plant Protection Products and their Residues. "Opinion on a request from EFSA related to the default Q10 value used to describe the temperature effect on transformation rates of pesticides in soil". In: *EFSA Journal*, 622 (2008), p. 154. doi: 10.2903/j.efsa.2008.622.

[oPtR08b] EFSA Panel on Plant Protection Products and their Residues. "Risk Assessment for Birds and Mammals - Revision of Guidance Document under Council Directive 91/414/EEC (SANCO/4145/2000 final of 25 September 2002)". In: *EFSA Journal*, 734 (2008), p. 790. doi: 10.2903/j.efsa.2008.734.

[PA93] Kuldip K Paliwal and Bishnu S Atal. "Efficient vector quantization of LPC parameters at 24 bits/frame". In: *Speech and Audio Processing, IEEE Transactions on* 1.1 (1993), pp. 3–14. issn: 1063-6676.

[Pal80] Günther Palm. "On associative memory". In: *Biological cybernetics* 36.1 (1980), pp. 19–31. issn: 0340-1200.

[PB95] Nikhil R Pal and James C Bezdek. "On cluster validity for the fuzzy c-means model". In: *Fuzzy Systems, IEEE Transactions on* 3.3 (1995), pp. 370–379. issn: 1063-6706.

[PBH96] Nikhil R Pal, James C Bezdek, and Richard J Hathaway. "Sequential competitive learning and the fuzzy c-means clustering algorithms". In: *Neural Networks* 9.5 (1996), pp. 787–796. issn: 0893-6080.

[Pet+03] A Peters et al. "Diagnosis of glaucoma by indirect classifiers". In: *Methods of information in medicine* 42.1 (2003), pp. 99–103. issn: 0026-1270.

[PG89] Tomaso Poggio and Federico Girosi. *A theory of networks for approximation and learning*. Report. DTIC Document, 1989.

[PH07] Mee Young Park and Trevor Hastie. "L1-regularization path algorithm for generalized linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.4 (2007), pp. 659–677. issn: 1467-9868.

[Pla99a] John Platt. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74.

[Pla99b] John C Platt. "12 fast training of support vector machines using sequential minimal optimization". In: *Advances in kernel methods* (1999), pp. 185–208.

[PM] Dan Pelleg and Andrew W Moore. "X-means: Extending K-means with Efficient Estimation of the Number of Clusters". In: *ICML*. Vol. 1.

[QC07] R Qahwaji and Tufan Colak. "Automatic short-term solar flare prediction using machine learning and sunspot associations". In: *Solar Physics* 241.1 (2007), pp. 195–211. issn: 0038-0938.

[Quia] J Ross Quinlan. "Combining instance-based and model-based learning". In: *Proceedings of the Tenth International Conference on Machine Learning*, pp. 236–243.

[Quib] John R Quinlan. "Learning with continuous classes". In: *5th Australian joint conference on artificial intelligence*. Vol. 92. Singapore, pp. 343–348.

[Qui14] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014. isbn: 0080500587.

[RB] Martin Riedmiller and Heinrich Braun. "RPROP-A fast adaptive learning algorithm". In: *Proc. of ISCIS VII), Universitat*. Citeseer.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1". In: ed. by David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group. Cambridge, MA, USA: MIT Press, 1986. Chap. Learning Internal Representations by Error Propagation, pp. 318–362. isbn: 0-262-68053-X. url: http://dl.acm.org/citation.cfm?id=104279.104293.

[Rid99] Greg Ridgeway. "The state of boosting". In: *Computing Science and Statistics* (1999), pp. 172–181.

[Rie94] Martin Riedmiller. *Rprop-Description and Implementation Details: Technical Report*. Inst. f. Logik, Komplexität u. Deduktionssysteme, 1994.

[Rip] BD Ripley. "Pattern recognition and neural networks. 1996". In: *Cambridge Uni. Press, Cambridge* ().

[RJ91] S. J. Raudys and A. K. Jain. "Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.3 (1991), pp. 252–264. doi: 10.1109/34.75512. url: http://www.scopus.com/inward/record.url?eid=2-s2.0-0026120032&partnerID=40&md5=c196878d37136f200d7a856084f63d6a.

[RK03] Marko Robnik-Šikonja and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF". In: *Machine learning* 53.1-2 (2003), pp. 23–69. issn: 0885-6125.

[RKL03a] Ingo Ruczinski, Charles Kooperberg, and Michael LeBlanc. "Logic regression". In: *Journal of Computational and Graphical Statistics* 12.3 (2003), pp. 475–511. issn: 1061-8600.

[RKL03b] Ingo Ruczinski, Charles Kooperberg, and Michael LeBlanc. "Logic regression—methods and software". In: *Nonlinear Estimation and Classification*. Springer, 2003, pp. 333–343. isbn: 0387954716.

[RLH01] Charles Kooperberg Ingo Ruczinski, Michael L LeBlanc, and Li Hsu. "Sequence analysis using logic regression". In: *Genetic epidemiology* 21.1 (2001), S626–S631.

[Rob] Marko Robnik Šikonja. "CORE-a system that predicts continuous variables". In: *Proceedings of Electrotechnical and Computer Science Conference (ERK'97), Portoroz, Slovenia. Ljubljana: Slovene Section of IEEE*, B145–8.

[Rob04] Marko Robnik-Šikonja. "Improving random forests". In: *Machine Learning: ECML 2004*. Springer, 2004, pp. 359–370. isbn: 3540231056.

[Roj96] Raúl Rojas. "Neural Networks-A Systematic Introduction Springer-Verlag". In: *New York* (1996).

[RS05] Robert A Rigby and D Mikis Stasinopoulos. "Generalized additive models for location, scale and shape". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54.3 (2005), pp. 507–554. issn: 1467-9876.

[RSA10] Lars Ronnegard, Xia Shen, and Moudud Alam. "hglm: A Package for Fitting Hierarchical Generalized Linear Models". In: *The R Journal* 2.2 (2010), pp. 20–28. url: http://journal.r-project.org/archive/2010-2/RJournal%5C_2010-2%5C_Roennegaard~et~al.pdf.

[Sah] Mehran Sahami. "Learning Limited Dependence Bayesian Classifiers". In: *KDD*. Vol. 96, pp. 335–338.

[Sch+10] Matthias Schmid et al. "Estimation and regularization techniques for regression models with multidimensional prediction functions". In: *Statistics and Computing* 20.2 (2010), pp. 139–150. issn: 0960-3174.

[Sco+] Steven L Scott et al. "Bayes and big data: The consensus Monte Carlo algorithm". In: *EFaBBayes 250 conference*. Vol. 16.

[Seg88] M. R. Segal. "Regression Trees for Censored-Data". In: *Biometrics* 44.1 (1988), pp. 35–47. issn: 0006-341X. doi: Doi10.2307/2531894. url: %3CGo%20to%20ISI%3E://WOS:A1988M501700004.

[Sha49] Claude E Shannon. "Communication theory of secrecy systems*". In: *Bell system technical journal* 28.4 (1949), pp. 656–715. issn: 1538-7305.

[She+13] Xia Shen et al. "A novel generalized ridge regression method for quantitative genetics". In: *Genetics* 193.4 (2013), pp. 1255–1268. issn: 0016-6731.

[SHN04] Ohn Mar San, Van-Nam Huynh, and Yoshiteru Nakamori. "An alternative extension of the k-means algorithm for clustering categorical data". In: *International Journal of Applied Mathematics and Computer Science* 14.2 (2004), pp. 241–248. issn: 1641-876X.

[SIL07] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics". In: *bioinformatics* 23.19 (2007), pp. 2507–2517. issn: 1367-4803.

[SK95] Marko Robnik Sikonja and Igor Kononenko. "Discretization of continuous attributes using ReliefF". In: (1995).

[SM06] Charles Sutton and Andrew McCallum. "An introduction to conditional random fields for relational learning". In: *Introduction to statistical relational learning* (2006), pp. 93–128.

[SMT09] Carolin Strobl, James Malley, and Gerhard Tutz. "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests". In: *Psychological methods* 14.4 (2009), p. 323. issn: 1939-1463.

[SSa] Roded Sharan and Ron Shamir. "CLICK: a clustering algorithm with applications to gene expression analysis". In: *Proc Int Conf Intell Syst Mol Biol*. Vol. 8, p. 16.

[SSb] Lingyan Shu and Jonathan Schaeffer. "HCS: Adding Hierarchies to Classifier Systems". In: *ICGA*, pp. 339–345.

[SS12] Rebecca J Sela and Jeffrey S Simonoff. "RE-EM trees: a data mining approach for longitudinal and clustered data". In: *Machine learning* 86.2 (2012), pp. 169–207. issn: 0885-6125.

[SSM98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. "Nonlinear component analysis as a kernel eigenvalue problem". In: *Neural computation* 10.5 (1998), pp. 1299–1319.

[Str+07] Carolin Strobl, Anne-Laure Boulesteix, et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1 (2007), p. 1. issn: 1471-2105.

[SV07] Mohammad Shami and Werner Verhelst. "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech". In: *Speech Communication* 49.3 (2007), pp. 201–212. issn: 0167-6393.

[SV99] Johan AK Suykens and Joos Vandewalle. "Least squares support vector machine classifiers". In: *Neural processing letters* 9.3 (1999), pp. 293–300. issn: 1370-4621.

[Sve+04] Vladimir Svetnik et al. "Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules". In: *Multiple Classifier Systems*. Springer, 2004, pp. 334–343. isbn: 3540221441.

[SW06] Gero Szepannek and Claus Weihs. *Variable selection for discrimination of more than two classes where data are sparse*. Springer, 2006. isbn: 3540313133.

[SW99] Helmut Strasser and Christian Weber. "On the asymptotic theory of permutation statistics". In: (1999).

[SWJ98] Matthias Schonlau, William J Welch, and Donald R Jones. "Global versus local search in constrained optimization of computer models". In: *Lecture Notes-Monograph Series* (1998), pp. 11–25. issn: 0749-2170.

[Sye+99] Nadeem Ahmed Syed et al. "Incremental learning with support vector machines". In: (1999).

[Sze+] Gero Szepannek, Tamas Harczos, et al. "Extending features for automatic speech recognition by means of auditory modelling". In: *Proceedings of European Signal Processing Conference (EUSIPCO)*, pp. 1235–1239.

[Tak+06] Ichiro Takeuchi et al. "Nonparametric quantile estimation". In: *The Journal of Machine Learning Research* 7 (2006), pp. 1231–1264. issn: 1532-4435.

[Tan+96] Kit-Sang Tang et al. "Genetic algorithms and their applications". In: *Signal Processing Magazine, IEEE* 13.6 (1996), pp. 22–37. issn: 1053-5888.

[TB05] Gerhard Tutz and Harald Binder. "Localized classification". In: *Statistics and Computing* 15.3 (2005), pp. 155–166. issn: 0960-3174.

[TB06] Gerhard Tutz and Harald Binder. "Generalized Additive Modeling with Implicit Variable Selection by Likelihood-Based Boosting". In: *Biometrics* 62.4 (2006), pp. 961–971. issn: 1541-0420.

[TB07] Gerhard Tutz and Harald Binder. "Boosting ridge regression". In: *Computational Statistics & Data Analysis* 51.12 (2007), pp. 6044–6059. issn: 0167-9473.

[TDK04] Heiko Timm, Christian Döring, and Rudolf Kruse. "Different approaches to fuzzy clustering of incomplete datasets". In: *International Journal of Approximate Reasoning* 35.3 (2004), pp. 239–249. issn: 0888-613X.

[Teo+] Choon Hui Teo et al. "A scalable modular convex solver for regularized risk minimization". In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 727–736. isbn: 1595936092.

[TG10] Gerhard Tutz and Andreas Groll. "Generalized linear mixed models based on boosting". In: *Statistical Modelling and Regression Structures*. Springer, 2010, pp. 197–215. isbn: 3790824127.

[TG13] Gerhard Tutz and Andreas Groll. "Likelihood-based boosting in binary and ordinal random effects models". In: *Journal of Computational and Graphical Statistics* 22.2 (2013), pp. 356–378. issn: 1061-8600.

[Tib+12] Robert Tibshirani et al. "Strong rules for discarding predictors in lasso-type problems". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.2 (2012), pp. 245–266. issn: 1467-9868.

[Tib96] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288. issn: 0035-9246.

[Tip01] Michael E Tipping. "Sparse Bayesian learning and the relevance vector machine". In: *The journal of machine learning research* 1 (2001), pp. 211–244. issn: 1532-4435.

[Tro06] Joel A Tropp. "Just relax: Convex programming methods for identifying sparse signals in noise". In: *Information Theory, IEEE Transactions on* 52.3 (2006), pp. 1030–1051. issn: 0018-9448.

[Tru09] Alfred Kar Yin Truong. "Fast growing and interpretable oblique trees via logistic regression models". Thesis. 2009.

[Unr+95] Ronald C Unrau et al. "Hierarchical clustering: A structure for scalable multiprocessor operating system design". In: *The Journal of Supercomputing* 9.1-2 (1995), pp. 105–134. issn: 0920-8542.

[VBD04] Iven Van Mechelen, Hans-Hermann Bock, and Paul De Boeck. "Two-mode clustering methods: astructuredoverview". In: *Statistical methods in medical research* 13.5 (2004), pp. 363–394. issn: 0962-2802.

[Vog92] Michael Vogt. "Implementierung und Anwendung von'Generalized Radial Basis Functions' in einem Simulator neuronaler Netze". Thesis. 1992.

[VR94] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S-Plus*. Springer-Verlag, 1994.

[Wan] Y Witten Wang. "IH: Inducing Model Trees for Predicting Continuous Classes". In: *Proceedings of European Conference on Machine Learning. University of Economics. Prague*.

[Wan11] Zhu Wang. "HingeBoost: ROC-based boost for classification and variable selection". In: *The International Journal of Biostatistics* 7.1 (2011), pp. 1–30. issn: 1557-4679.

[Wan12] Zhu Wang. "Multi-class HingeBoost". In: *Methods of information in medicine* 51.2 (2012), pp. 162–167. issn: 0026-1270.

[WB98] Christopher KI Williams and David Barber. "Bayesian classification with Gaussian processes". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 20.12 (1998), pp. 1342–1351. issn: 0162-8828.

[Web00] Geoffrey I Webb. "Multiboosting: A technique for combining boosting and wagging". In: *Machine learning* 40.2 (2000), pp. 159–196. issn: 0885-6125.

[WF05] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005. isbn: 008047702X.

[WH14] Jianqiang C Wang and Trevor Hastie. "Boosted varying-coefficient regression models for product demand prediction". In: *Journal of Computational and Graphical Statistics* 23.2 (2014), pp. 361–382. issn: 1061-8600.

[WHS10] Bethany J Wolf, Elizabeth G Hill, and Elizabeth H Slate. "Logic forest: an ensemble classifier for discovering logical combinations of binary markers". In: *Bioinformatics* 26.17 (2010), pp. 2183–2189. issn: 1367-4803.

[WIG02] Yingquan Wu, Krassimir Ianakiev, and Venu Govindaraju. "Improved k-nearest neighbor classification". In: *Pattern recognition* 35.10 (2002), pp. 2311–2318. issn: 0031-3203.

[WL08] Hansheng Wang and Chenlei Leng. "A note on adaptive group lasso". In: *Computational Statistics & Data Analysis* 52.12 (2008), pp. 5277–5286. issn: 0167-9473.

[Wol+12] Bethany J Wolf, Elizabeth G Hill, Elizabeth H Slate, et al. "LBoost: A Boosting Algorithm with Application for Epistasis Discovery". In: *PloS one* 7.11 (2012), e47281. issn: 1932-6203.

[Wol92] David H Wolpert. "Stacked generalization". In: *Neural networks* 5.2 (1992), pp. 241–259. issn: 0893-6080.

[WT11] Daniela M Witten and Robert Tibshirani. "Penalized classification using Fisher's linear discriminant". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.5 (2011), pp. 753–772. issn: 1467-9868.

[WW] IH Witten and Y Wang. "Induction of model trees for predicting continuous classes". In: *Proc. Poster Papers Europ. Conf. Machine Learning*.

[WW98] Jason Weston and Chris Watkins. *Multi-class support vector machines*. Report. Citeseer, 1998.

[WYM] Wei Wang, Jiong Yang, and Richard Muntz. "STING: A statistical information grid approach to spatial data mining". In: *VLDB*. Vol. 97, pp. 186–195.

[WZ15] Marvin N. Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *arXiv preprint arXiv:1508.04409* (2015).

[XCM12] Huan Xu, Constantine Caramanis, and Shie Mannor. "Sparse algorithms are not stable: A no-free-lunch theorem". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34.1 (2012), pp. 187–193. issn: 0162-8828.

[XW05] Rui Xu and Donald Wunsch. "Survey of clustering algorithms". In: *Neural Networks, IEEE Transactions on* 16.3 (2005), pp. 645–678. issn: 1045-9227.

[Yao+] Zheng Yao et al. "R-C4. 5 Decision tree model and its applications to health care dataset". In: *Services Systems and Services Management, 2005. Proceedings of ICSSSM'05. 2005 International Conference on*. Vol. 2. IEEE, pp. 1099–1103. isbn: 0780389719.

[YI02] E. Yom-Tov and G. F. Inbar. "Feature selection for the classification of movements from single movement-related potentials". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 10.3 (Sept. 2002), pp. 170–177. issn: 1534-4320. doi: 10.1109/TNSRE.2002.802875.

[YL06] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67. issn: 1467-9868.

[Zak+] Mohammed Javeed Zaki et al. "New Algorithms for Fast Discovery of Association Rules". In: *KDD*. Vol. 97, pp. 283–286.

[ZB05] L Zhang and Yun Fei Bai. "Genetic algorithm-trained radial basis function neural networks for modelling photovoltaic panels". In: *Engineering applications of artificial intelligence* 18.7 (2005), pp. 833–844. issn: 0952-1976.

[Zel+98] Andreas Zell et al. "SNNS: Stuttgart Neural Network Simulator-Manual Extensions of Version 4.0". In: (1998).

[Zel94] A Zell. "Simulation neuronaler Netze, vol. 1". In: *Aufl. Bonn: Addison-Wesley Verlag* (1994).

[ZH05] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320. issn: 1467-9868.

[ZH07] Achim Zeileis and Kurt Hornik. "Generalized M-fluctuation tests for parameter instability". In: *Statistica Neerlandica* 61.4 (2007), pp. 488–508. issn: 1467-9574.

[Zha+06] Hao Helen Zhang, Jeongyoun Ahn, et al. "Gene selection using support vector machines with non-convex penalty". In: *Bioinformatics* 22.1 (2006), pp. 88–95. issn: 1367-4803.

[ZHH08] Achim Zeileis, Torsten Hothorn, and Kurt Hornik. "Model-based recursive partitioning". In: *Journal of Computational and Graphical Statistics* 17.2 (2008), pp. 492–514. issn: 1061-8600.

[Zho+00] Aoying Zhou et al. "Approaches for scaling DBSCAN algorithm to large spatial databases". In: *Journal of computer science and technology* 15.6 (2000), pp. 509–526. issn: 1000-9000.

[Zho+04] Dengyong Zhou et al. "Ranking on data manifolds". In: *Advances in neural information processing systems* 16 (2004), pp. 169–176.

[ZHT06a] Hui Zou, Trevor Hastie, and Robert Tibshirani. "Sparse Principal Component Analysis". In: *Journal of Computational and Graphical Statistics* 15.2 (2006), pp. 265–286. doi: 10.1198/106186006X113430.

[ZHT06b] Hui Zou, Trevor Hastie, and Robert Tibshirani. "Sparse principal component analysis". In: *Journal of computational and graphical statistics* 15.2 (2006), pp. 265–286. issn: 1061-8600.

[Zhu05] Xiaojin Zhu. "Semi-supervised learning literature survey". In: (2005).

[ZL07] Hao Helen Zhang and Wenbin Lu. "Adaptive Lasso for Cox's proportional hazards model". In: *Biometrika* 94.3 (2007), pp. 691–703. issn: 0006-3444.

[Zou06] Hui Zou. "The adaptive lasso and its oracle properties". In: *Journal of the American statistical association* 101.476 (2006), pp. 1418–1429. issn: 0162-1459.

[ZW00] Zijian Zheng and Geoffrey I Webb. "Lazy learning of Bayesian rules". In: *Machine Learning* 41.1 (2000), pp. 53–84. issn: 0885-6125.

5 Case studies based on the data from the 2011-2014 annual European Union Summary Reports on Zoonoses.

5.1 Introduction

Salmonella is a bacterium that can cause an illness called salmonellosis in humans. In the European Union over 90,000 salmonellosis cases are reported every year. EFSA has estimated that the overall economic burden of human salmonellosis could be as high as EUR 3 billion a year. To combat human salmonellosis it is important to reduce Salmonella in animals and derived products so that food is safer for consumers. The occurrence of Salmonella in humans, animals and food is monitored and analysed in EU Summary Reports prepared by EFSA and the European Centre for Disease Prevention and Control (ECDC) each year to provide up-to-date information on the current situation in Europe.

Two case studies are presented to illustrate the potential use of MLTs on biological hazards, which are the object of the European Union Summary Reports on Zoonoses. In particular, two different risk questions were explored as being of potential interest for EFSA:

1. data quality assurance: currently, in EFSA the national submitted data require a large amount of quality checks and this procedure could be carried out in a more efficient way. Such a huge amount of data restrict the possibility of control just to a comparison on the submitted data with those referring to the previous year. For the involved EFSA staff, the availability of a MLT base tool, able to provide an automated process data, could be of some interest;
2. detection of latent pattern of epidemiological concern: data from multiple Member States relating to multiple matrices and zoonotic agents may hide transnational epidemiological patterns of relevant interest at the EU level; the identification of such patterns is another useful potential application of MLT techniques to the zoonotic agents data considered by the EU Reports.

5.2 Methods

5.2.1 Data

The focus was on data regarding the following annual report:

- 2014: http://www.efsa.europa.eu/en/efsajournal/pub/4329
- 2013: http://www.efsa.europa.eu/en/efsajournal/pub/3991
- 2012: http://www.efsa.europa.eu/en/efsajournal/pub/3547
- 2011: http://www.efsa.europa.eu/en/efsajournal/pub/3129

For these case studies, the datasets used for producing the European Union Summary Reports on Zoonoses over the period 2011-2014 were retrieved from the EFSA website and reviewed. For illustration purposes, the case studies are focused on the data regarding *Salmonella*. A further focus was on food, feedingstuff and animal, excluding foodborne outbreaks data.

Salmonella "L3_disease status" were selected for each year and a single file was obtained. Over the 2011-2014 period,31 EU Countries submitted data on *Salmonella* prevalence in food, feedingstuff and animals, summing up to $n = 101064$ records.[47] Twelve variables were candidated for data analysis:

1. **repYear**: report year;
2. **SpeciesType**: the matrix type in which *Salmonella* is tested:

[47] The dataset obtained by appending the internet available data contained $n = 103732$ observations: negative or missing records for the variable totUnitsTested were not considered.

- Food;
- Animal;
- Feed;

3. **matrix_L1**: the matrix of interest at level 1 (Species level);
4. **zoonotic_agent**: the *Salmonella* specific serotype;
5. **repCountry**: Country (EU) which submitted data;
6. **totUnitsTested**: total units tested;
7. **totUnitsPositive**: total units positive;
8. **sampUnit**: unit of measure of sample;
9. **sampWeight_num**: absolute value of the sample weight;
10. **sampWeight_unit**: unit of measure of the sample weight;
11. **record-specific_prevalence**: totUnitsPositive/totUnitsTested;
12. **crude_prevalence**: overall zoonotic agent prevalence over the year.

With regards to animal, feedingstuff, food there were respectively 100, 30 and 84 distinct and mutually exclusives matrices, with the only exception for fish that belongs both to animal and food. Each record is the unique combination of repYear, repCountry, zoonotic_agent, matrix_L1 and SpeciesType.

Sample unit and weight

After data inspection, sampUnit and sampWeight_num and sampWeight_unit were excluded from the analysis as they were highly heterogeneous: the units of measure are qualitative whereas the weights are a mix of weights, volumes and areas.

Prevalence

The record-specific prevalence at serovar level for each countries were considered in order to have a model able to predict the expected prevalence at country by year level. The crude zoonotic agent prevalence was computed by merging all country-specific dataset on the variable *Salmonella* serovar. This data manipulation that is useful since *Salmonella* serovar is the primary information for clustering the Countries when the aim consists in detecting epidemiological latent patterns that might be of concern.

A description of the data can be found in the Table 81.

Table 81: summary of numerical variables

	min	q1	median	mean	q3	max	sd
totUnitsTested	1.00	4.00	19.00	454.46	108.00	212245.00	4082.20
totUnitsPositive	0.00	0.00	0.00	25.18	1.00	69608.00	777.55
record-specific_prevalence	0.00	0.00	0.00	0.04	0.01	1.00	0.15
crude_prevalence	0.00	0.01	0.01	0.03	0.03	1.00	0.06

5.2.2 Exploratory EFSA Case Study 1: Data control

The first objective could be accomplished training a machine learning algorithm to perform the two following tasks:

1. To detect errors in current data submission (current year) by comparing the pattern of presence/absence of national submitted testing data by matrix over the available previous years;

2. To detect errors in current data submission (current year) by predicting the expected prevalence of a *Salmonella*'s serovar current year based on MLT training based on the dataset, cumulated over the available previous years;

These tasks represent an automated preliminary evaluation of the data provided by the Countries. Both of them can be achieved by using random forrest MLTs (see section 4.1.1). Following the guidelines found in Trigal et al. (2013), the following two RFs were trained.

Presence/absence of Country reports on matrices

A random forest RF was constructed by considering the submission of samples in 2014 by each Country and each matrix. To train the RF, 500 bootstrap samples were generated and $\sqrt{p}$ was the number of variables randomly sampled as candidates at each split, where p is the total number of explanatory variables (in this case study $p = 12$). The bootstrap procedure guarantees that on average, trees are trained on about 2/3rd of the data. For each tree, the leftover observations were used to compute the missclassification (out-of-bag) error rate. Averaging the error from all trees determined the overall out-of-bag error rate. The out-of-bag errors were used for the internal validation of the overall procedure.

The training set was based only on the `repCountry`, `matrix_L1` variables plus the `totUnitsTested` value for each year.

The area under the ROC curve was used to evaluate the models performance. Probability provided by the RF were converted into a binary output variable, based on a probability threshold obtained by minimizing the distance between the top left corner of the ROC diagram and the threshold value.

Expected record-specific prevalence

To train a RF for predicting record-specific prevalence in each Country that provided at least one record submission for a specific matrix in the period 2010-2013, the configuration was the same except for:

1. Data format was kept "long" (i.e. the years are the content of a single variable, repYear, and the totUnitsTested is a single variable too, exactly as provided by EFSA).

2. The variables used to train this RF were: repYear, matrix_L1, repCountry, record-specific_prevalence and serotype.

5.2.3 Exploratory EFSA Case Study 2: latent epidemiological pattern discovery

The identification of latent epidemiological *Salmonella* patterns, through the detection of clusters of countries over time, is a second potential useful application of MLT to zoonotic agent data.

To address this question, unsupervised MLTs can be applied to find similarities comparing *Salmonella* serovars stratified by matrix, Country and year. A comparison over the whole period (all years pool together) is also possible.

5.3 Results

5.3.1 Case Study 1: Data control

The ROC curve of the prediction of the absence/presence of a record submission is shown in Figure 50. On the curve, the values of the estimated prevalence, corresponding to the different thresholds at which 1-specificity and sensitivity of the random forest are computed. As benchmark, a logistic regression model on the same data was considered (Figure 50).

In Table 82, the performance achieved by the first RF is shown. Not unexpectedly, logistic regression and random forest had the same predictive capability. This is because random forests are poorly exploited in their potential given the small number of covariates in the dataset. On the other side, logistic regression in presence of highly stratified predictors (i.e., with large number of levels for any covariate), tends to overfit data, with the consequence of low apparent error rates. Investigation on the model stability would inform better on the true model performances.

In Figure 51 the national probabilities of record submission (presence) by group of matrices are shown along with a threshold line. The validity of each prediction is defined by the shape of the symbols (i.e. a triangle

Table 82: Value and standard deviation (where applicable) of the total predicted percent, sensitivity, specificity and area under the ROC (AUC) matrix for the prediction of presence/absence of a record submission in 2014 for a Country in each matrix.

	value-rf	sd-rf	value-glm	sd-glm
PosPred	0.81	0.01	0.81	0.01
Sens	0.82	0.01	0.82	0.01
Spec	0.81	0.01	0.81	0.01
AUC	0.88		0.87	

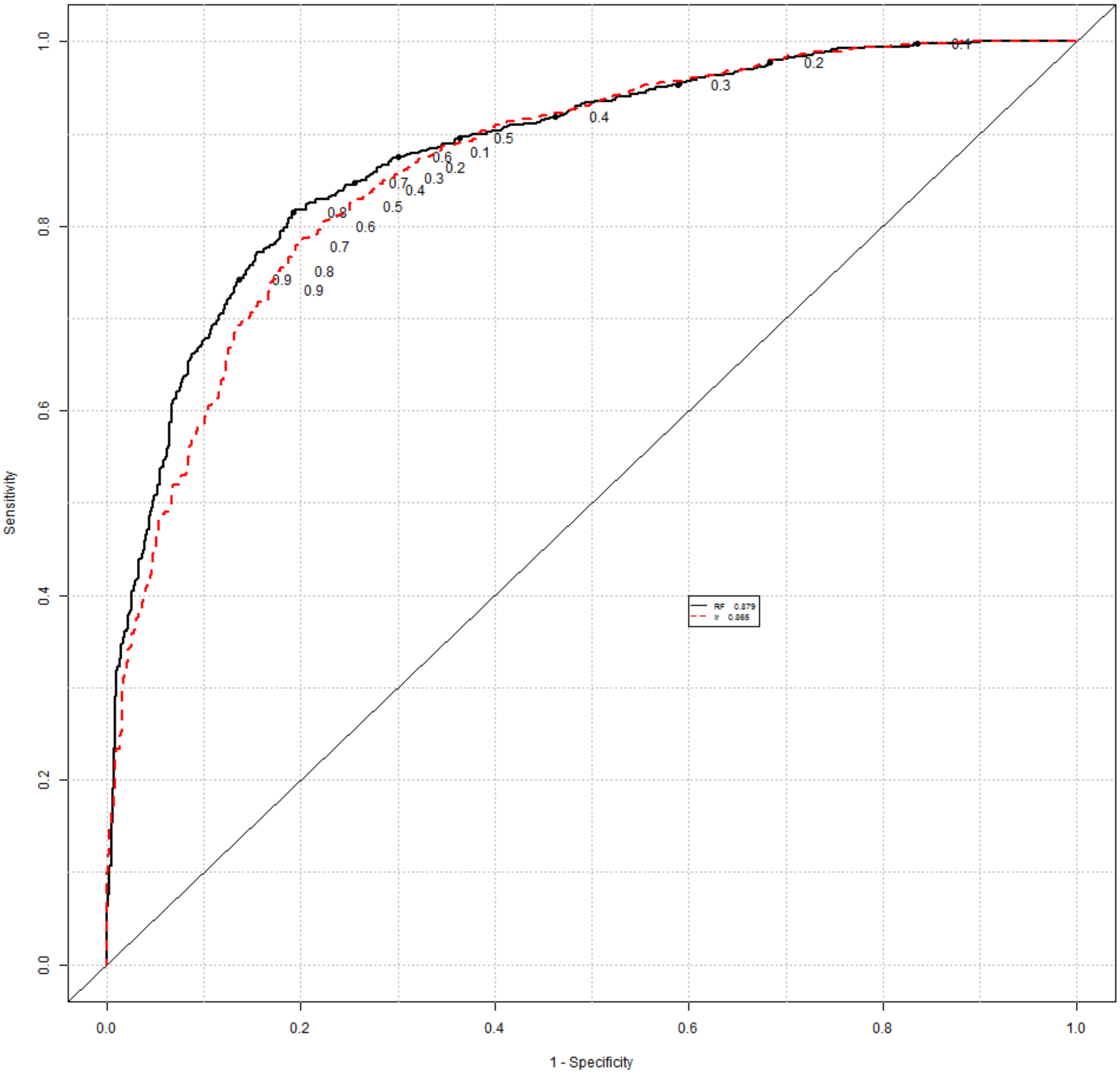


Figure 50: ROC for report presence/absence prediction in 2014 for a Country in each matrix (Random Forest and Logistic Regression)

is a country's behavior consistent with the expected value whereas a circle is representing a not expected behavior).

Looking at a national chart, as one of those in figure 51, attention should be focused on the circles when they are above the threshold line (top of the chart), which means that a record submission expected is missing. A circle below the threshold indicates an improvement in the record submission, as it means that a record submission not expected is indeed present. Predictions are available only for those countries that have submitted records at least in one year.

5.3.2 Case study 2: Latent pattern discovered

Aiming at analyzing latent patterns, the distribution of zoonotic agents and their prevalences by year and Country was considered. Main findings are summarized in a heatmap (Figure 52). Yearly behavior in produced reports can be read horizontally along with the yearly prevalence rates. Such information is structured by agent and by its prevalence. A change in color toward green indicates lower prevalence and toward red the opposite.

To further explore these data, a cluster identification exercise has been performed. Dissimilarity among records has been investigated. For illustration purpose only results for the year 2011 are reported.

Gower's coefficients (Gower, 1971) have been used to build the dissimilarity matrix. Using this approach, variables are standardized where the "distance" between two units is the sum of all the variable-specific distances. PAM (Partitioning Around Medoids) algorithm has been used to process dissimilarity matrix. Compared to the k-means approach, PAM also accepts a dissimilarity matrix; it is more robust because it minimizes a sum of dissimilarities instead of a sum of squared euclidean distances and it provides a novel graphical display, the silhouette plot: for each observation i, a bar is drawn, representing its silhouette width s(i) grouped per cluster, starting with cluster 1 at the top. Observations with a large s(i) (almost 1) are very well clustered, a small s(i) (around 0) means that the observation lies between two clusters, and observations with a negative s(i) are probably placed in the wrong cluster.

Preliminary analyses for several values of k (the number of clusters), i.e. from 2 to 30 (see Figure 53) were performed. The silhouette width gives a measure of the space between clusters: if the cluster coesion is good (i.e. the average distance between each points and all other points in the same cluster) and the cluster separation is large (i.e. the average distance between each point in a cluster and all points in the nearest one)then the silhouette width will be large. The value of $k = 16$ corresponding to the one with the largest average silhouette width, equal to 0.25, was chosen.

For the choice of $k = 16$ the resulting silhouette and the relative clusplot (see Kaufman and Rousseeuw, 1990) of a cluster partition (a two-dimensional representation of the observations, in which the clusters are indicated by ellipses) are shown in Figure 54.

5.4 Discussion

Two potentially useful applications of MLT to data from any of the zoonotic microorganism considered in the EU reporting on zoonoses are presented. A set of MLT based techniques may serve to automate some steps of the process of quality assurance. Experts' knowledge is of paramount importance in this exploratory case studies to assess the value of MLT in this specific field. The epidemiological added value may be particularly relevant for refining the approaches that EFSA adopts when addressing these kind of food safety related issues.

5.5 MLT modification to fit specific issues

5.5.1 Case Study 1: addressing generalization in small samples

Generally MLTs are successfully applied to scenarios with large datasets, i.e. hundreds of thousand or millions records. When small datasets arise, they are trickier to deal with. In the case study of Data Control, due to aggregation of the data on countries, year, matrix of interest and species type, the resulting dataset consists of 2,013 records, which can be considered a small dataset for a typical MLT application. More generally speaking, this issue can arise when dealing with rare phenomena, working with time series or when data are aggregated (or in any situations in which sampling is expensive or the population itself is limited).

In these situations, over-fitting is a hard issue to avoid, not only on the training data but also on the validation set, and outliers get more dangerous since they can have an influential role on the model fitting.

One straightforward method to limit the overfitting is reducing the complexity of the tree cutting down its depth. In the simplest way this can be achieved by pruning the trees of the random forest. Instead of just

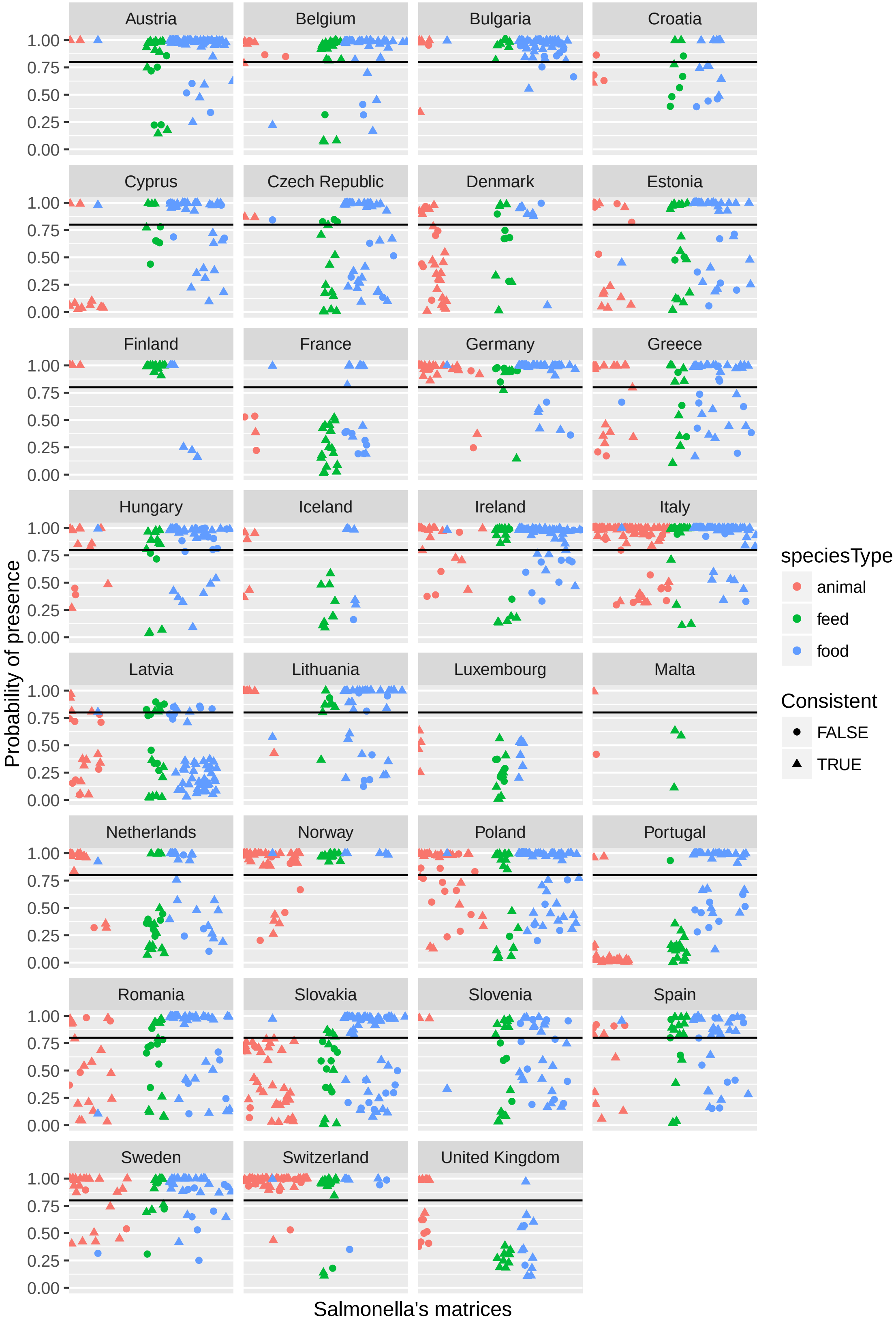


Figure 51: Estimated probability presence for matrices each Country in 2014. Triangles indicates record submission consistent with the expected. If they are above the solid black line they indicates that an expected record has been submitted; if they are below they indicates that a not expected record has not been submitted. Circles indicates record submission not consistent with the expected. If they are above the solid black line they indicates that an expected record has not been submitted; if they are below they indicates that a not expected record has been submitted. Colors indicates the matrix type in which Salmonella is tested.

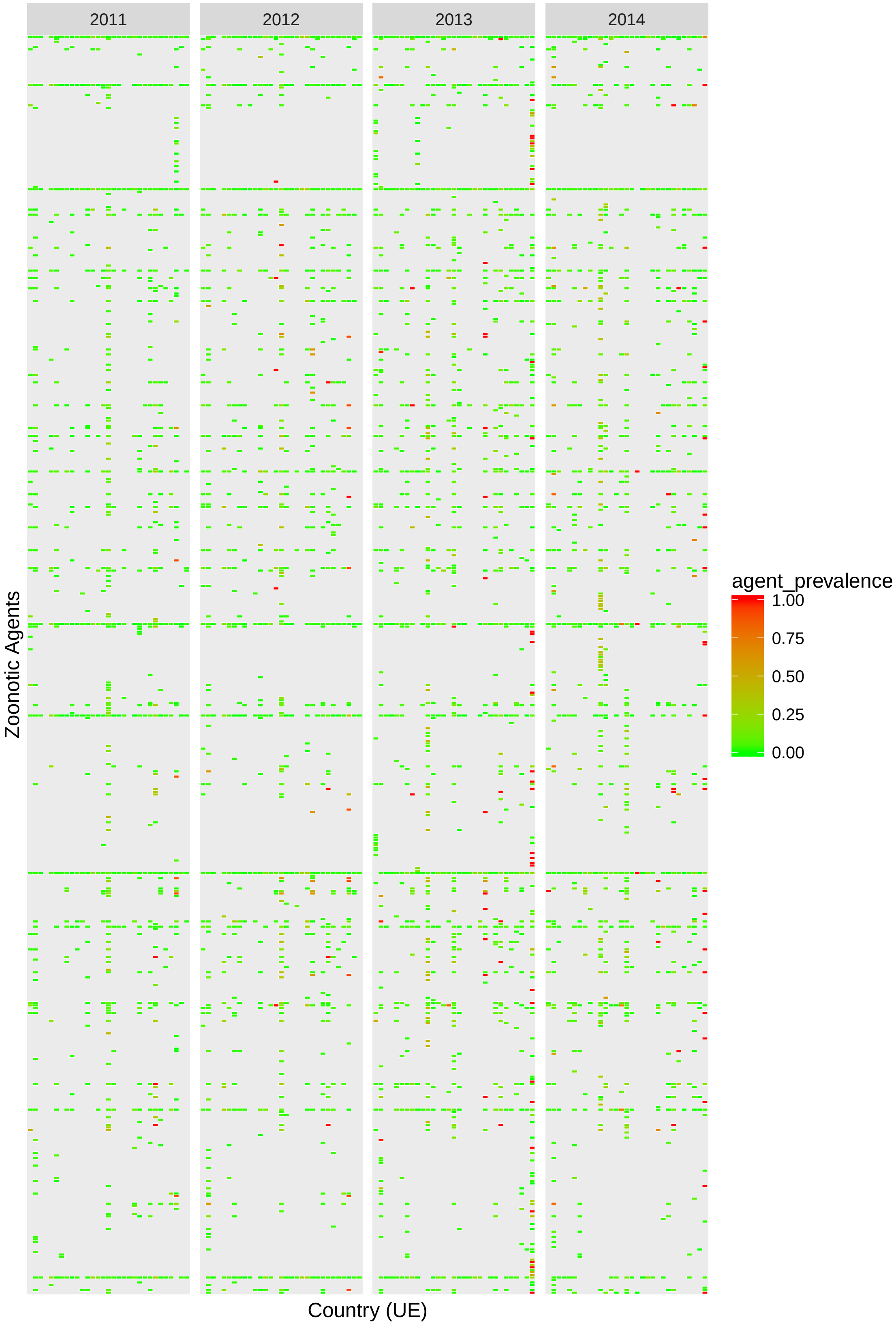


Figure 52: Heatmap of zoonotic agents per Countries (UE) prevalence over the Years

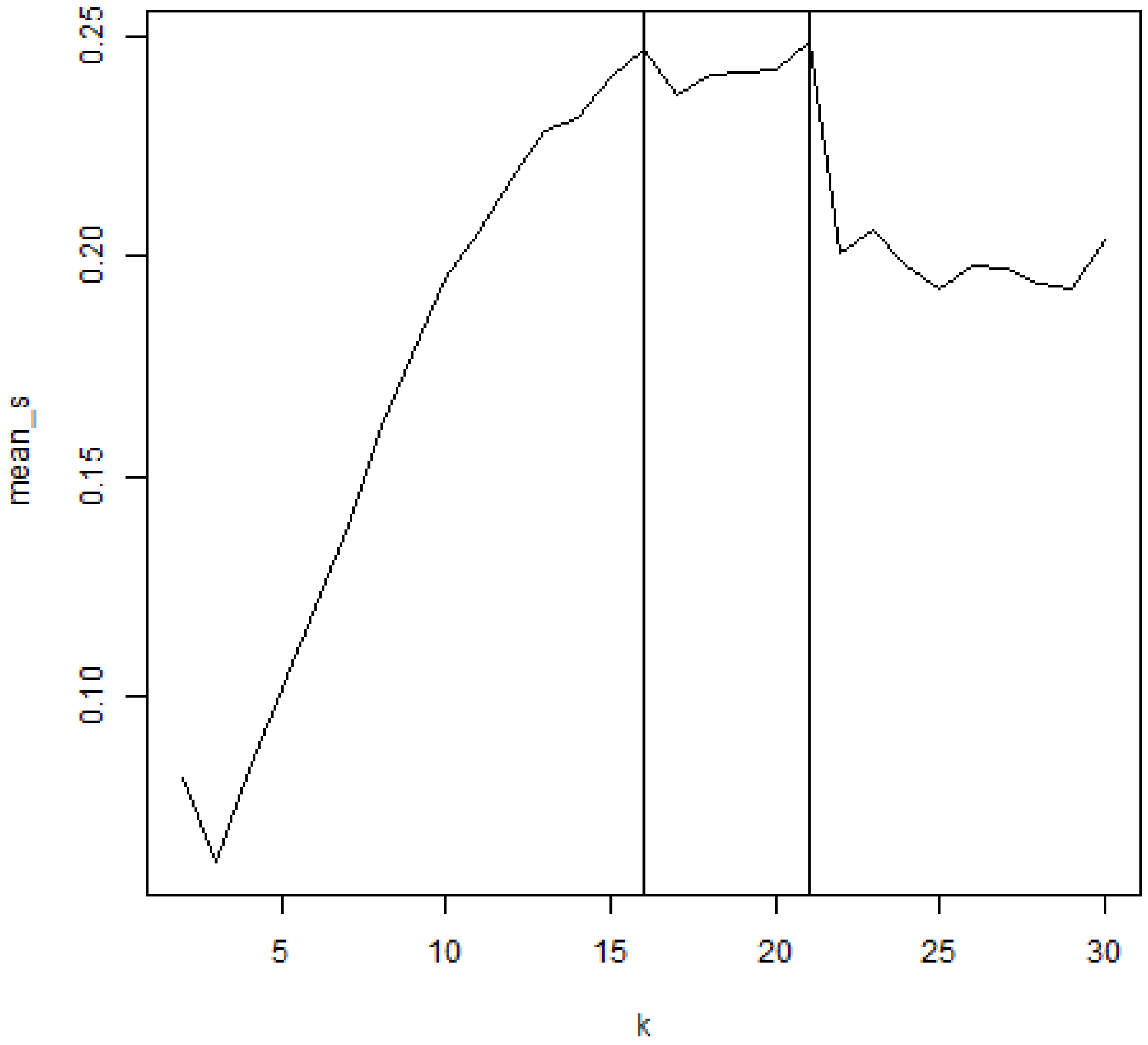


Figure 53: Plots of mean silhouettes value with highlighted its two maximum value. There were two maximum value, i.e. k=16 and k=21. The value k=16 has been chosen because in the case k=21 there was a cluster with many observations allocated to the wrong cluster, negative silhouette width

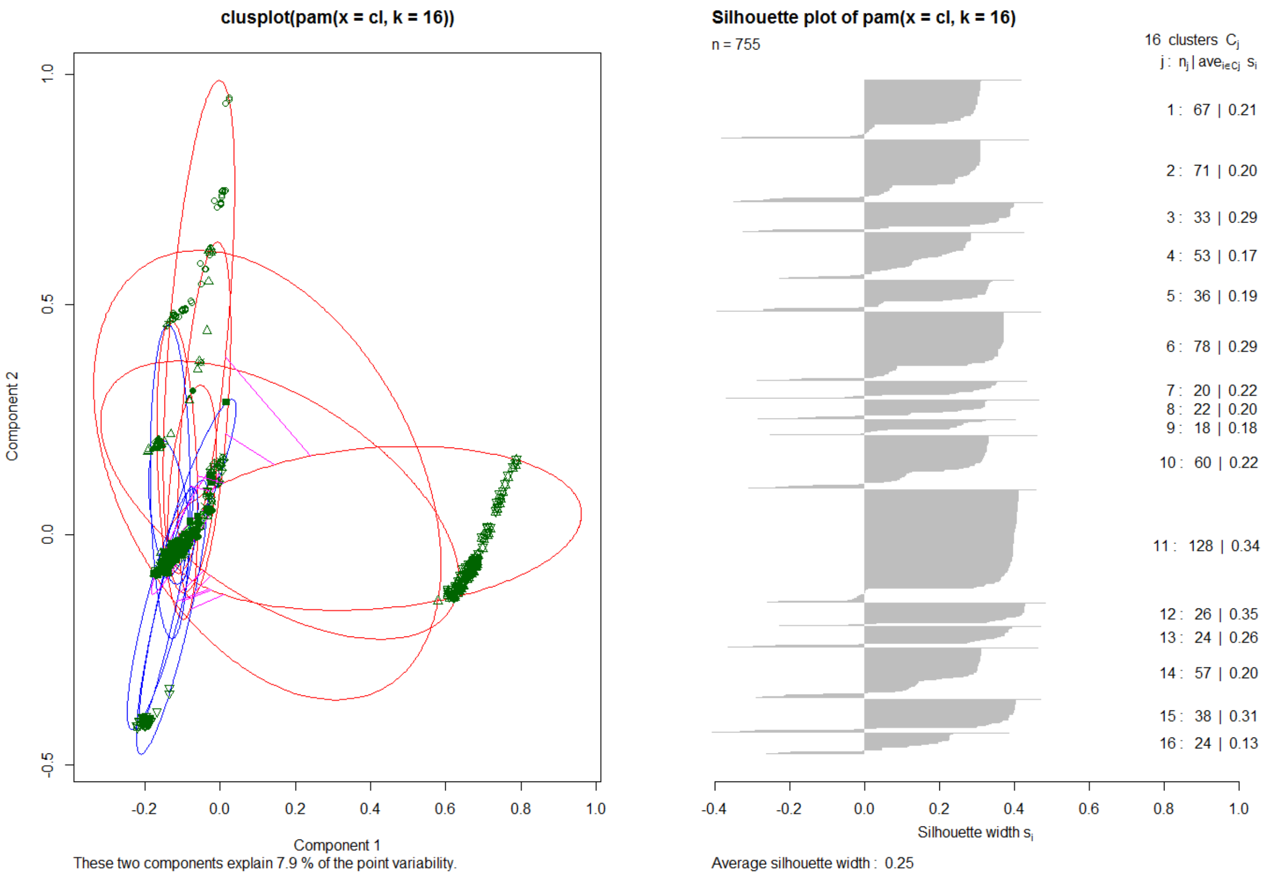


Figure 54: Heatmap of zoonotic agents per Countries (UE) prevalence over the Years

limiting the depth of the tree, i.e. the number of the splits, this can be done reaching a trade-off between accuracy and complexity of the model using regularization, i.e. adding a compexity penalty to the loss function that must be minimized: $L + \lambda|w|_1$, where $|\cdot|$ is the L1-norm regularization.

A potential fruitful way is to build a random forest using the training dataset and applying a regularization so that some of the rules that define the random forest can be deleted.

Since each observation goes down from the root node to only one and one leaf node of each of the trees that form the random forest, for each leaf node the set of rules, which encodes the leaf node assignment, can be captured.

Defining X the matrix of the rules, this means that at each observation y in the sample corresponds a set of rules so that:

$$y = a + \boldsymbol{X} \cdot w^T$$

where the coefficient w indicates the importance of the rule. Thus the limitation of the number of rules can be reformulated as a linear regression problem of learning the weight parameters. This learning problem can be addressed using the standard 1-norm regularization.

Implementation is quite straightforward in R using the package *inTrees* for extracting the rules from a random forest built using the *randomForest* package. The code is reported below:

```
library(randomForest)
library(inTrees)

X <− Data[,−7]
Y <− Data[,”repYear_2014”]
rf <− randomForest(X ,Y ,ntree=100)

rules <− extractRules(RF2List(rf), X)
```

```
unique_rules <- unique(rules)
ruleMetric <- getRuleMetric(unique_rules, X, target)
```

The command `getRuleMetric` in the package *inTrees* allows for extracting the rules built by the random forest, which are reported in Table 83.

Table 83: rules built by the random forest

	len	freq	err	condition	pred
1	2	0.097	0.229	X[,1] %in% c('feed') & X[,2] %in% c('Czech Republic','Denmark','France','Iceland','Latvia','Luxembourg','Malta','Netherlands','Portugal','Slovakia','Slovenia','Spain','United Kingdom')	0.354
2	1	0.276	0.237	X[,1] %in% c('animal')	0.614
3	1	0.497	0.202	X[,1] %in% c('food')	0.718
4	1	0.497	0.202	X[,1] %in% c('food')	0.718
5	1	0.228	0.245	X[,1] %in% c('feed')	0.572
6	1	0.228	0.245	X[,1] %in% c('feed')	0.572
7	2	0.002	0	X[,1] %in% c('animal') & X[,2] %in% c('Luxembourg')	0
8	2	0.017	0.195	X[,1] %in% c('animal') & X[,2] %in% c('Latvia','United Kingdom')	0.735
9	4	0.036	0.148	X[,1] %in% c('animal','feed') & X[,2] %in% c('Czech Republic','Denmark','France','Iceland','Malta','Netherlands','Portugal','Slovakia','Slovenia','Spain') & X[,4]>0.5 & X[,5]<=0.5	0.180
10	3	0.065	0.164	X[,1] %in% c('animal','feed') & X[,2] %in% c('Czech Republic','Denmark','France','Iceland','Malta','Netherlands','Portugal','Slovakia','Slovenia','Spain') & X[,5]<=0.5	0.208
11	1	0.503	0.241	X[,1] %in% c('animal','feed')	0.595
12	1	0.503	0.241	X[,1] %in% c('animal','feed')	0.595
13	1	0.497	0.202	X[,1] %in% c('food')	0.718
14	1	0.338	0.214	X[,2] %in% c('Belgium','Croatia','Cyprus','Estonia','Germany','Greece','Hungary','Lithuania','Poland','Romania')	0.690
15	1	0.034	0.155	X[,2] %in% c('Belgium')	0.809
16	1	0.063	0.230	X[,2] %in% c('Cyprus','Greece')	0.643
17	1	0.102	0.223	X[,2] %in% c('Croatia','Estonia','Poland')	0.663
18	1	0.276	0.237	X[,1] %in% c('animal')	0.614
19	1	0.724	0.220	X[,1] %in% c('feed','food')	0.672
20	1	0.038	0.150	X[,2] %in% c('Germany')	0.816
21	1	0.034	0.217	X[,2] %in% c('Hungary')	0.681
22	1	0.314	0.147	X[,2] %in% c('Austria','Bulgaria','Finland','Ireland','Italy','Norway','Sweden','Switzerland')	0.822
23	1	0.314	0.147	X[,2] %in% c('Austria','Bulgaria','Finland','Ireland','Italy','Norway','Sweden','Switzerland')	0.822
24	1	0.075	0.142	X[,2] %in% c('Austria','Switzerland')	0.828
25	1	0.239	0.148	X[,2] %in% c('Bulgaria','Finland','Ireland','Italy','Norway','Sweden')	0.820
26	3	0.163	0.078	X[,2] %in% c('Austria','Bulgaria','Finland','Ireland','Italy','Norway','Sweden','Switzerland') & X[,3]>0.5 & X[,4]>0.5	0.915
27	1	0.306	0.247	X[,5]<=0.5	0.447
28	1	0.376	0.248	X[,4]<=0.5	0.538
29	1	0.503	0.241	X[,1] %in% c('animal','feed')	0.595
30	2	0.03	0.193	X[,1] %in% c('animal','feed') & X[,2] %in% c('Slovakia')	0.262
31	2	0.493	0.126	X[,4]>0.5 & X[,5]>0.5	0.852
32	1	0.497	0.202	X[,1] %in% c('food')	0.718
33	2	0.493	0.126	X[,4]>0.5 & X[,5]>0.5	0.852
34	2	0.44	0.127	X[,3]>0.5 & X[,5]>0.5	0.851
35	3	0.362	0.140	X[,1] %in% c('animal','food') & X[,3]>0.5 & X[,4]>0.5	0.831
36	1	0.413	0.237	X[,3]<=0.5	0.614
37	1	0.308	0.250	X[,2] %in% c('Cyprus','Denmark','France','Latvia','Lithuania','Luxembourg','Malta','Portugal','Slovakia','Spain','United Kingdom')	0.490
38	1	0.692	0.197	X[,2] %in% c('Austria','Belgium','Bulgaria','Croatia','Czech Republic','Estonia','Finland','Germany','Greece','Hungary','Iceland','Ireland','Italy','Netherlands','Norway','Poland','Romania','Slovenia','Sweden','Switzerland')	0.730
39	2	0.44	0.127	X[,3]>0.5 & X[,5]>0.5	0.851
40	1	0.376	0.248	X[,4]<=0.5	0.538
41	1	0.772	0.217	X[,1] %in% c('animal','food')	0.681
42	1	0.228	0.245	X[,1] %in% c('feed')	0.572
43	2	0.107	0.232	X[,2] %in% c('Cyprus','Latvia','Netherlands','Portugal','Romania','Spain','United Kingdom') & X[,4]<=0.5	0.366
44	1	0.694	0.188	X[,5]>0.5	0.748
45	1	0.624	0.198	X[,4]>0.5	0.727
46	1	0.228	0.245	X[,1] %in% c('feed')	0.572
47	4	0.032	0.109	X[,1] %in% c('feed') & X[,2] %in% c('Austria','Belgium','Bulgaria','Cyprus','Denmark','Finland','Hungary','Latvia','Lithuania','Luxembourg','Norway','Romania','Slovakia','Slovenia','Switzerland') & X[,3]<=0.5 & X[,4]<=0.5	0.875
48	3	0.043	0.136	X[,1] %in% c('feed') & X[,4]>0.5 & X[,5]<=0.5	0.163
49	2	0.024	0.220	X[,1] %in% c('feed') & X[,2] %in% c('Belgium','Denmark','Luxembourg')	0.673
50	1	0.48	0.227	X[,2] %in% c('Austria','Belgium','Bulgaria','Cyprus','Denmark','Finland','Hungary','Latvia','Lithuania','Luxembourg','Norway','Romania','Slovakia','Slovenia','Switzerland')	0.651

The set of rules like the following ones (reported for the first 10 observations for illustration purpose)

```
    ”X[,1]␣%in%␣c(’feed’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Czech␣Republic’,’Denmark’,’France’,’Iceland’,
␣␣␣␣␣’Latvia’,’Luxembourg’,’Malta’,’Netherlands’,
␣␣␣␣␣’Portugal’,’Slovakia’,’Slovenia’,’Spain’,
␣␣␣␣␣’United␣Kingdom’
␣␣␣␣␣)␣&
␣␣␣␣␣X[,4]␣<=␣0.5”

    ”X[,1]␣%in%␣c(’animal’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Czech␣Republic’,’Denmark’,’France’,’Iceland’,
␣␣␣␣␣’Latvia’,’Luxembourg’,’Malta’,’Netherlands’,
␣␣␣␣␣’Portugal’,’Slovakia’,’Slovenia’,’Spain’,
␣␣␣␣␣’United␣Kingdom’
␣␣␣␣␣)␣&
␣␣␣␣␣X[,4]␣<=␣0.5”

    ”X[,1]␣%in%␣c(’food’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Czech␣Republic’,’Denmark’,’France’,’Iceland’,
␣␣␣␣␣’Latvia’,’Luxembourg’,’Malta’,’Netherlands’,’Portugal’,
␣␣␣␣␣’Slovakia’,’Slovenia’,’Spain’,’United␣Kingdom’
␣␣␣␣␣)␣&
␣␣␣␣␣X[,4]␣<=␣0.5”

    ”X[,1]␣%in%␣c(’food’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Latvia’,’Luxembourg’,’United␣Kingdom’)␣&
␣␣␣␣␣X[,4]␣>␣0.5”

    ”X[,1]␣%in%␣c(’feed’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Luxembourg’)␣&
␣␣␣␣␣X[,4]␣>␣0.5”

    ”X[,1]␣%in%␣c(’feed’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Latvia’,’United␣Kingdom’)␣&
␣␣␣␣␣X[,4]␣>␣0.5”

    ”X[,1]␣%in%␣c(’animal’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Luxembourg’)␣&
␣␣␣␣␣X[,4]␣>␣0.5”

    ”X[,1]␣%in%␣c(’animal’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Latvia’,’United␣Kingdom’)␣&
␣␣␣␣␣X[,4]␣>␣0.5”

    ”X[,1]␣%in%␣c(’animal’,’feed’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Czech␣Republic’,’Denmark’,’France’,
␣␣␣␣␣’Iceland’,’Malta’,’Netherlands’,’Portugal’,
␣␣␣␣␣’Slovakia’,’Slovenia’,’Spain’
␣␣␣␣␣)␣&
␣␣␣␣␣X[,3]␣<=␣0.5␣&
␣␣␣␣␣X[,4]␣>␣␣␣0.5␣&
␣␣␣␣␣X[,5]␣<=␣0.5”

   ”X[,1]␣%in%␣c(’animal’,’feed’)␣&
␣␣␣␣␣X[,2]␣%in%␣c(’Czech␣Republic’,’Denmark’,’France’,
␣␣␣␣␣’Iceland’,’Malta’,’Netherlands’,’Portugal’,
␣␣␣␣␣␣'Slovakia','Slovenia','Spain'
␣␣␣␣␣␣)␣&
␣␣␣␣␣␣X[,3]␣>␣␣␣0.5␣&
␣␣␣␣␣␣X[,4]␣>␣␣␣0.5␣&
␣␣␣␣␣␣X[,5]␣<=␣0.5"
```

evaluated at each observation $(X[i,1], X[i,2], \ldots, X[i,5])$, for $i = 1, \ldots, n$ are transformed in a FALSE/TRUE or 0/1 matrix. Then the standard glmet function in the glmnet package can be used to run the LASSO procedure and cross-validation can be applied to tune tuning the λ parameter.

Finally, in the Table 84 the regularized rules are reported along with the prediction error computed on the test set.

Table 84: final regularized rules

	len	freq	err	condition	pred
1	3	0.0909090909090909	0	repYear_2011<=0.5 & repYear_2012<=0.5 & repYear_2013<=0.5	1
2	2	0.0357675111773472	0	speciesType %in% c('feed') & repCountry %in% c('Czech Republic','France','Iceland','Malta','Portugal','United Kingdom')	0
3	3	0.0183805265772479	0	speciesType %in% c('animal') & repCountry %in% c('Iceland','Luxembourg','Portugal','Slovakia') & repYear_2012<=0.5	0
4	4	0.014903129657228	0	speciesType %in% c('animal') & repCountry %in% c('Bulgaria','Denmark','Estonia','Germany','Iceland','Norway','Portugal','Romania','Spain','Sweden','Switzerland') & repYear_2012>0.5 & repYear_2013<=0.5	0
5	3	0.0134128166915052	0	speciesType %in% c('food') & repCountry %in% c('Latvia','Luxembourg','Malta') & repYear_2011<=0.5	0
6	4	0.0134128166915052	0	speciesType %in% c('feed') & repCountry %in% c('Belgium','Cyprus','Denmark','Estonia','Latvia','Luxembourg','Malta','Portugal','Romania','Slovakia','Spain') & repYear_2011>0.5 & repYear_2013<=0.5	0
7	4	0.0124192747143567	0	speciesType %in% c('feed') & repCountry %in% c('Belgium','Bulgaria','Hungary','Ireland','Lithuania','Netherlands','Switzerland') & repYear_2011<=0.5 & repYear_2012<=0.5	1
8	2	0.0114257327372081	0	repCountry %in% c('Finland') & repYear_2012>0.5	1
9	1	0.789369100844511	0.084	Else	1

5.5.2 Conclusions

In proposing this slight modification of the tree pruning procedure, a regularization process has been introduced in order to penalize those rules that have less impact in minimizing the loss function/squared prediction error. The choice of the loss function to minimize can potentially influence the performance of the procedures. In this specific example, L1 regularization has been preferred to L2 regularization. Generally, L1 regularization induces fewer non-zero parameters, effectively performing implicit feature selection, which could be desirable for explainability of performance in production. On the other hand, L2 regularization induces closer to zero parameters and it is thought of having strong zero-centered a-priori for the parameters. At our knowledge there is not a standard implementation of such procedure and thus it requires a more exaustive assessment possibly on already studied dataset, in order to have benchmark on which to compare results. The motivation of such proposal relies in the requirement of dealing with small sample size, thus with an inherent difficulty in generalizing the results to larger population. To avoid overfitting and have a better chance to capture only the more relevant pattern in the data, strategies enabling to cut down the rules produced by the random forest are of benefit.

Case studies bibliography

[Gow71] J. C. Gower. “A general coefficient of similarity and some of its properties”. In: *Biometrics* 27 (1971), pp. 857–874.

[KR90] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Ed. by Wiley. 1990.

6 Case studies based on food borne outbreaks and antimicrobial resistance.

6.1 Introduction

Antimicrobial resistance (AMR) is a major threat to global health and in the European Union (EU) it is estimated that each year approximately 25,000 people die from infections resistant to therapy. Both in the medical and in the veterinary field the problem is associated to the massive and not proper use of antimicrobial drugs. In cattle, the phenomenon is exacerbated by the practice of administering mass therapy and by the administration of antimicrobials orally via the water or powder supply: the spread of resistance to humans is due to direct contact with infected animals, the consumption of food contaminated with resistant microorganisms and finally through the spread into the environment of manure containing antibiotics residues or resistant bacteria or their resistance genes. The spread and persistence in the environment favour the further propagation through the horizontal transfer of resistance genes between bacterial species, not necessarily pathogenic. Periodically EFSA and ECDC publish a report providing the results of the analysis of data submitted by Member States to provide up-to-date information on the AMR situation in Europe, currently the 2014 data are available.

6.2 Data

The focus was on data regarding the following annual report:

- 2014: https://www.efsa.europa.eu/en/efsajournal/pub/4380
- 2013: http://www.efsa.europa.eu/it/efsajournal/pub/4036
- 2012: http://www.efsa.europa.eu/it/efsajournal/pub/3590
- 2011: http://www.efsa.europa.eu/en/efsajournal/pub/3196
- 2010: http://www.efsa.europa.eu/it/efsajournal/pub/2598

In this section it is provided a short description of the EFSA data that will be used in the Case Studies: Food-Borne Outbreaks (FBO) and Antimicrobial Resistance datasets.

FBO data

FBO data are collected to monitor the number of food-borne outbreaks, the number of human cases, along with the number of hospitalized individuals and deaths due to the intake of infected food. Data are available for years 2011-2013. Missing values, coded in the datasets as "-1", were coded in *NA* values R standard format. The outcome of analysis was the ratio between number of hospitalized individuals and number of human cases with food-borne outbreaks; the attention was focused on the number of hospitalized individuals in order to consider the proportion of the most serious food-borne outbreak cases.

The **disease** dataset, which contains information about the type of outbreak, of the same year was used to add information about zoonosis.

In order to set the granularity of the data, **Disease** and **FBO** dataset were merged by the country (*repCountry* variable), zoonosis (*zoonosis* variable), food type (*matrix* variable) and type of disease status units (*unitsDS* variable), which was then aggregated in order to get the sum of all this units.

The dataset was finally reshaped setting the country as the indentifier variable and the zoonosis type, the food type and the type of disease status units as the measured variables. Variables containing comments were removed.

On the merged dataset, covariates were identified: in this set of variables were included the type of zoonosis, the number of outbreaks and all the variables containing information about the number of the type of food infected and the number of disease type status unit.

The dataset was thinned out, removing variables that did not contain any values, reducing the number of explanatory variable from 34 to 9. Finally, additional missing values were removed from the outcome, covariate were identified and training and testing data were prepared, where: the training set was created randomly sampling 90% of dataset records, while the testing set was created taking the remaining 10% of dataset records.

Antimicrobial Resistance data

For case studies focused on Antimicrobial Resistance, two datasets were considered: **AMU_AMR_VALIDATED_ALL_YEARS2.csv**, which contains information about Antimicrobial Resistance of zoonosis to a certain type of substance (years (2009-2014) and **PREVALENCE_ALLYEARS.csv**, which contains information about the prevalence of zoonosis, given by the number of units of food tested in every laboratory.

First, duplicated variables were removed. Then, in the **prevalence** dataset, the granularity of the data was set by grouping the dataset by country (*repCountry* variable), zoonosis type (*zoonosis_L1* variable), food type (*matrix_L1* variable) and year (*repYear* variable). Then, two variables containing the sum of tested positive units and the sum of tested units were created.

The aggregated **prevalence** dataset was merged with **AMR** dataset by year, country, type of zoonosis and type of food analyzed. In the merged dataset, a variable, called *resistance*, indicating the resistance of the zoonosis to the substance, was created in the following way: if the Microbial Concentration (MIC) of the substance in the food was greater than a certain cutoff value, the variable *resistance* would get the value 1, indicating that the zoonosis was resistant to that substance, otherwise the variable would get the value 0, indicating that the zoonosis was not resistant to that substance. Finally, missing values, coded in the datasets as -1, were recoded as *NA* values.

6.3 Monitoring antimicrobial resistance

The aim of the analysis is to understand relationships between prevalence of zoonosis and antimicrobial resistance. In the first instance, data will be analyzed with respect a crude classification as "resistant" or "not resistant". Distribution over years of resistance is provided in table 85.

Table 85: Distribution over years of resistance.

	non resistant	resistant	%
2010	126195	28421	18.38
2011	122961	24776	16.77
2012	180610	47173	20.07
2013	205588	49541	19.48
2014	193133	48316	20.01

Factors associated with resistance can be explored using several techniques, ranging from classical regression tools up to classification trees and random forests. In the current analysis, a random forest classifier has been used, in particular for its flexibility, which allow to deal with different type of response variable. Overall, identification of zoonotic agent and substance combinations can be addressed by exploring interactions as derived from a random forest model. Considering zoonotic agent, substance and matrix detailed at L1 level only, the data are composed by a 1048575 x 154 matrix for all years. In particular, considering Germany in the year 2014, the dataset contains 36,860 records and 152 explanatory variables. Although this dataset is not high–dimensional, computational performances can be an issue. Trees tend to favor splits on continuous variables and factors with large numbers of levels (Loh and Shih, 1997).

A randomize version of a splitting rule can be adopted to mitigate this issue, and to considerably improve computational speed. A maximum of pre-specified split points are chosen randomly for each of the pre-specified number of variables within a node. The splitting rule is applied to the random split points and the node is split on that variable and random split point yielding the best value (as measured by the splitting rule). Pure random splitting can be also used. For each node, a variable is randomly selected and the node is split using a random split point (Cutler and Zhao, 2001; Lin and Jeon, 2006). Large sample consistency results provides a rationale for this approach. Indeed, under random splitting, if the number of splits k_n used to grow the trees satisfies the rule $\frac{k_n}{n} \to 0$ and $k_n \to \infty$, then the partitioning classifiers approximates the true classification rule.

The resulting random forest is a classification tree that has a satisfactory performance, with an overall error rate of about 19.89%, mostly due to the 6880 cases of resistance misclassified as non-cases (85.11%

conditional error rate) and the 453 non-resistance cases wrongly classified as case (1.57% conditional error rate). Target error rate is reached after less than 50 grown trees for non-resistance cases, whereas is not reached for resistance cases (Figure 55) out of the 1000 generated, with an average number of nodes examined equal to 92.59 and 13 variables tried at each split. Classification has been based on Gini index splitting rule. Other alternatives, like unweighted and heavy weighted Gini index splitting rules have been tested without showing improvement in classification performances. This is in line with literature where the best performances of Gini have been highlighted (Ishwaran, 2015).

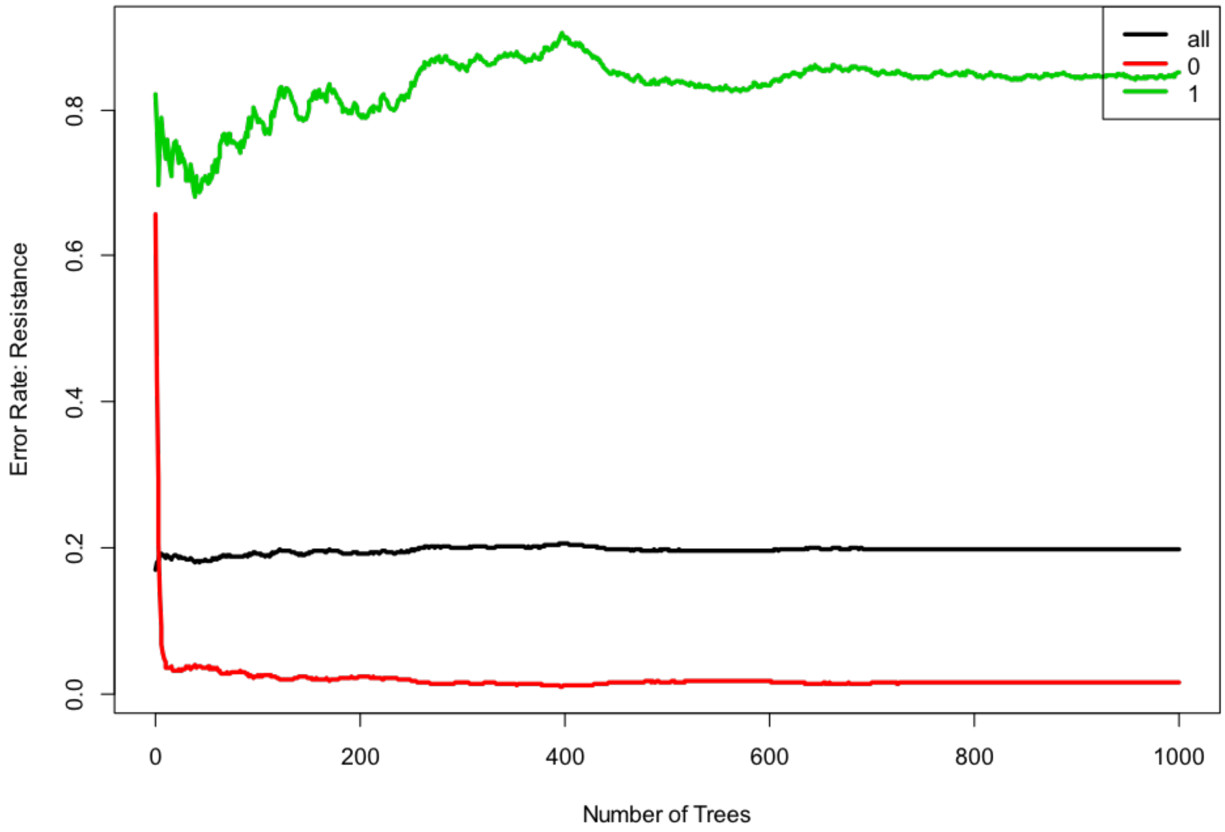


Figure 55: Classification error rate for resistance vs. grown number of trees (Germany, 2014).

Clearly, variables characterizing resistance are more specific than sensitive, and this is a common situation in the classification problem. We will see that this issue is almost impossible to overcome but by inserting more information in the response variable. We will accomplish this by studying the quantity of microbial resistance. Nevertheless, also in this classification problem, one essential aspect is to understand which variable is considered by the model as relevant. This is caught by the concept of variable importance, which has been calculated both by random daughter assignment and by using Breiman-Cutler joint estimator.

Importance of variable is a measure that can be estimated using several algorithms (Ishwaran, 2007), but usually the choice is limited to random daughter assignment or by random permutation of the variable(s).

Joint importance of variables was overall 6.07%, 1.16% for classifying non-resistance and 6.04% for classifying cases. Variable importance is illustrated in Table 86. For the purpose of the current analysis, the most relevant information is given by the values in the column named *resistance*, which reports the importance of the variables, including subsets of zoonotic agents, substance and environmental situations, with respect resistance classification. Variables omitted had importance value equal to 0.

Importance can be inspected also by plotting the importance measure against all variables considered in the analysis. Clusters of relevance are tentatively and visually identified in the upper parts, with Salmonella, Rodents and Quinolones as major determinants of resistance (Figure 56).

Table 86: Variable importance for classifying resistances.

	non resistant	resistance
Campylobacter	0.0154769600	0.0003036863
Enterococcus..non.pathogenic	0.0000000000	0.0000000000
Escherichia.coli..non.pathogenic	0.0000000000	0.0082615037
Escherichia.coli..pathogenic	0.0024189950	0.0052370114
Salmonella	0.0000000000	0.0000000000
Staphylococcus	0.0000000000	0.0053959885
All.animals	0.0004317139	0.0075087828
All.feedingstuffs	0.0000000000	0.0000000000
Compound.feedingstuffs.for.cattle	0.0020258167	-0.0013681540
Compound.feedingstuffs.for.fish	0.0082675656	0.0000000000
Mice	0.0060920193	0.0006011954
Milk.from.other.animal.species.or.unspecified	0.0080802820	0.0000000000
Other.products.of.animal.origin	0.0000000000	0.0016789263
Otter	0.0004427996	0.0012382484
Parrots	0.0000000000	0.0000000000
Partridges	0.0000000000	0.0046694047
Pet.food	0.0011811579	0.0036277833
Pheasants	0.0000000000	0.0000000000
Rats	0.0014517564	-0.0010024326
Ready.to.eat.salads	0.0060025977	0.0000000000
Reptiles	0.0018328451	0.0001362246
Rodents	0.0025901781	0.0000000000
Solipeds..domestic	0.0052557410	0.0028406311
Spices.and.herbs	-0.0012976249	0.0000000000
Ionophores	0.0000000000	0.0019997792
Lincosamides	-0.0013817765	0.0082726373
Macrolides	0.0000000000	0.0000000000
Trimethoprim	0.0050907800	0.0017764804
Trimethoprim...Sulfonamides	0.0022143741	0.0000000000

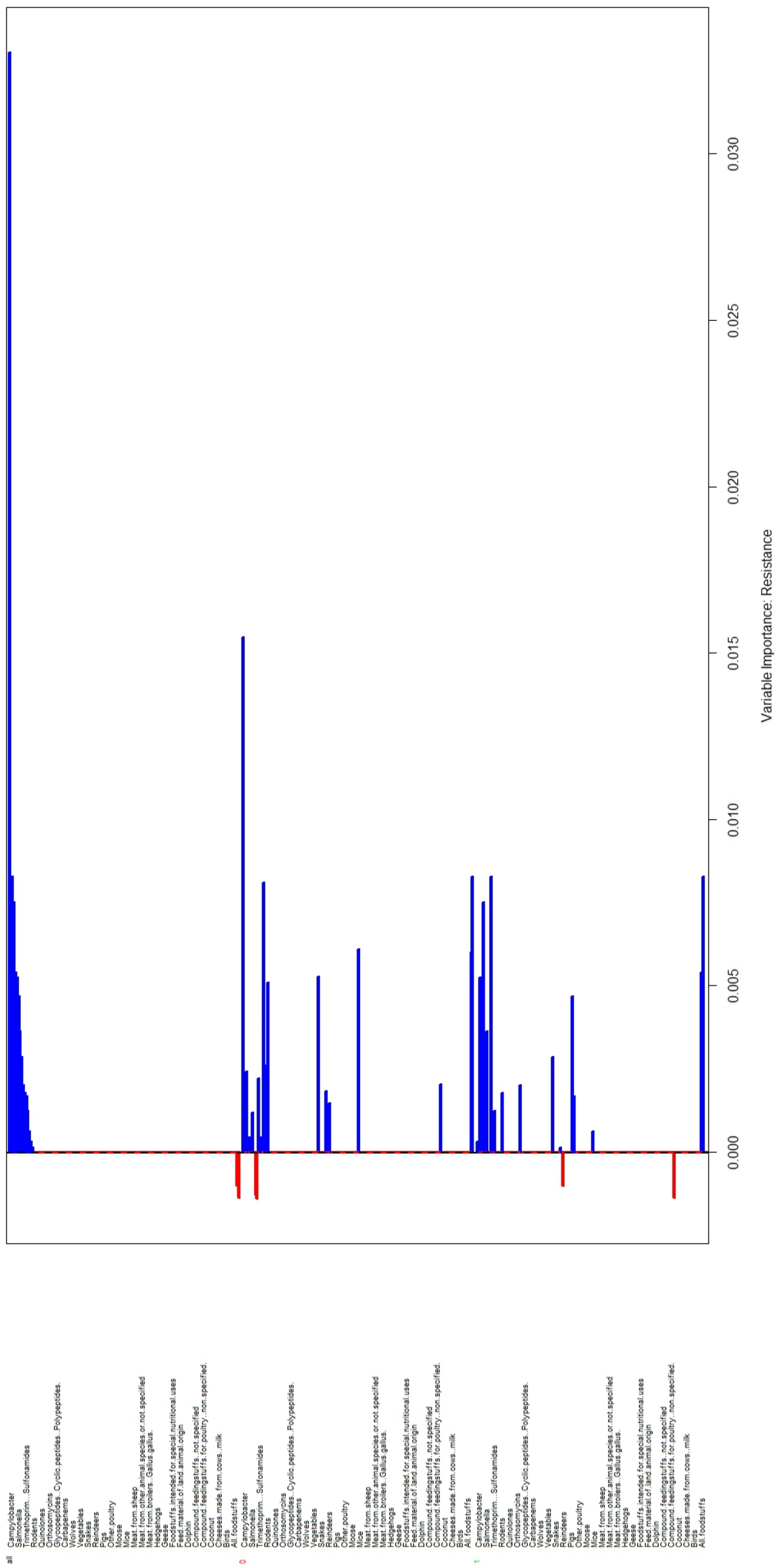


Figure 56: Variable importance plot.

Noticeably, such heuristic selection of relevant variables should be associated with more formal assessments. Indeed, selection of zoonotic agents and substances associated with onset of resistance is a common tool within the random forest building process. Variable selection can be based on the minimal depth variable selection criterion, which uses all data and all variables simultaneously.

Alternatives are consisting in implementing features selection strategies, which are generally used for problems where the number of variables is substantially larger than the sample size (e.g., p/n is greater than 10). Using training data from a stratified K-fold subsampling (stratification based on the outcomes), a forest is fit using a pre-specified number of randomly selected variables (variables are chosen with probability proportional to weights determined using an initial forest fit). The variables are then ordered by increasing minimal depth and added sequentially (starting from an initial model determined using minimal depth selection) until joint importance measure no longer increases (meaning that other variables are of no importance in the final model).

A forest is refit to the final model and applied to the test data to estimate prediction error. The process is repeated n times. Final selected variables are the top P ranked variables, where P is the average model size (rounded up to the nearest integer) and variables are ranked by frequency of occurrence.

Control over the selection process can be further enhanced by setting the level of conservativeness of the threshold rule used in minimal depth selection, i.e.: (i) high, using the most conservative threshold, (ii) medium, using the default less conservative tree-averaged threshold and (iii) low, using the more liberal one standard error rule for the minimal depth of the maximal subtree.

The maximal subtree for a variable x is the largest subtree whose root node splits on x. Thus, all parent nodes of x's maximal subtree have nodes that split on variables other than x. The largest maximal subtree possible is the root node. In general, however, there can be more than one maximal subtree for a variable. A maximal subtree may also not exist if there are no splits on the variable. The minimal depth of a maximal subtree (the first order depth) measures predictiveness of a variable x. It equals the shortest distance (the depth) from the root node to the parent node of the maximal subtree (zero is the smallest value possible). The smaller the minimal depth, the more impact x has on prediction. The mean of the minimal depth distribution is used as the threshold value for deciding whether a variable's minimal depth value is small enough for the variable to be classified as strong.

The low level is presented with reference to the classification problem in Table 87. No major differences arise when increasing conservativeness, except the inclusion of Cephalosporins...lactamase.inhibitores (depth 12.383442) in the set of most relevant variables also .

Selection process provides a clear indication on the impact of zoonotic agents, substance used and environmental condition where microbiotic resistance is more likely. From the perspective of the practitioner, a common question arises toward the identification of the combination of such factors in increasing the risk of resistance. In the random forest approach, this task is accomplished by an exhaustive search of pairwise interactions for all pairs of variables considered in the analysis (either all of them or just the most relevant as emerging from the variable selection process).

A first approach is by using a maximal subtree analysis (Ishwaran, Kogalur, Gorodeski, et al., 2010; Ishwaran, Kogalur, Chen, et al., 2011), which return a symmetric matrix, whose principal diagonal contains the normalized minimal depth of variable [i] relative to the root node (normalized with respect to the size of the tree) and the off-principal diagonal elements contain the normalized minimal depth of a variable [j] wrt the maximal subtree for variable [i] (normalized wrt the size of [i]'s maximal subtree). Smaller values in the diagonal indicate predictive variables. Small number values in the off-diagonal provides insights on a potential interaction between variables.

Table 87: Variable selection by minimal depth algorithm and low conservativeness.

	depth
Tetracyclines	4.472767
Turkeys	4.491285
Gallus.gallus..fowl.	4.583878
Campylobacter	5.044662
Fluoroquinolones	5.069717
Salmonella	5.165577
Aminoglycosides	5.189542
Penicillins	5.229847
Meat.from.turkey	5.351852
Meat.from.broilers..Gallus.gallus.	5.360566
Escherichia.coli..non.pathogenic	5.367102
Cephalosporins	5.676471
Sulfonamides	5.733115
Pigs	5.791939
Carbapenems	5.802832
Quinolones	5.899782
Glycylcyclines	6.117647
Other.poultry	6.587146
Cattle..bovine.animals.	6.611111
Macrolides	6.834423
Meat.from.pig	6.964052
Polymyxins	7.092593
Meat.from.other.animal.species.or.not.specified	7.591503
Trimethoprim	8.039216
Amphenicols	8.630719
Meat.from.bovine.animals	8.735294

Table 88: Maximal subtree analysis of the interactions among 10 randomly chosen variables (for illustration purposes only).

	Campylobacter	Salmonella	Escherichia.coli..non.pathogenic	Enterococcus..non.pathogenic	Escherichia.coli..pathogenic	Staphylococcus	All.animals	All.feedingstuffs	All.foodstuffs	Alpacas
Campylobacter	0.3374742	0.6117846	0.6269071	1	1	1	1	1	1	1
Salmonella	0.8631577	0.3460774	0.8442647	1	1	1	1	1	1	1
Escherichia.coli..non.pathogenic	0.8181295	0.8258853	0.3563723	1	1	1	1	1	1	1
Enterococcus..non.pathogenic	1	1	1	1	1	1	1	1	1	1
Escherichia.coli..pathogenic	1	1	1	1	1	1	1	1	1	1
Staphylococcus	1	1	1	1	1	1	1	1	1	1
All.animals	1	1	1	1	1	1	1	1	1	1
All.feedingstuffs	1	1	1	1	1	1	1	1	1	1
All.foodstuffs	1	1	1	1	1	1	1	1	1	1
Alpacas	1	1	1	1	1	1	1	1	1	1

Campylobacter, Salmonella and non.pathogenic *Escherichia coli* interact and lead to an increase of the probability of developing resistances. A second approach is based on the joint-importance approach. Two variables are paired and their paired importance measure is calculated (referred to as 'Paired' importance). Thus the importance for each separate variable is also calculated. The sum of these two values is referred to as 'Additive' importance. A large positive or negative difference between 'Paired' and 'Additive' indicates an association worth pursuing if the univariate importance for each of the paired-variables is reasonably large. In order to avoid computational burden, analysis is shown for 10 selected variables. For example, environmental variables(i.e. feedingstuffs, foodstuffs, alpacas) are associated with *Escherichia coli* but not with *Enterococcus* in developing resistance. Then such method can be used for identifying clusters of variables (in the example, environmental variables), which can help in indentifying factors associated to the risk of developing resistance.

Table 89: Joint-importance based analysis of the interactions among 10 randomly chosen variables (for illustration purposes only).

	Var 1	Var 2	Paired	Additive	Difference
Campylobacter:Enterococcus..non.pathogenic	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:Escherichia.coli..non.pathogenic	0.016067403	0.015913286	0.023349995	0.031980689	-0.008630694
Campylobacter:Escherichia.coli..pathogenic	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:Salmonella	0.016067403	0.015913286	0.021599815	0.031980689	-0.010380874
Campylobacter:Staphylococcus	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:All.animals	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:All.feedingstuffs	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:All.foodstuffs	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:Alpacas	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Enterococcus..non.pathogenic:Escherichia.coli..non.pathogenic	0.000000000	0.000000000	0.007648699	0.000000000	0.007648699
Enterococcus..non.pathogenic:Escherichia.coli..pathogenic	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:Salmonella	0.000000000	0.000000000	0.005823052	0.000000000	0.005823052
Enterococcus..non.pathogenic:Staphylococcus	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:All.animals	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:All.feedingstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..non.pathogenic:Escherichia.coli..pathogenic	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:Salmonella	0.007752569	0.007219941	0.010372372	0.014972509	-0.004600137
Escherichia.coli..non.pathogenic:Staphylococcus	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:All.animals	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:All.feedingstuffs	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:All.foodstuffs	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:Alpacas	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..pathogenic:Salmonella	0.000000000	0.000000000	0.005823052	0.000000000	0.005823052
Escherichia.coli..pathogenic:Staphylococcus	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..pathogenic:All.animals	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..pathogenic:All.feedingstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..pathogenic:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..pathogenic:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Salmonella:Staphylococcus	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Salmonella:All.animals	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Salmonella:All.feedingstuffs	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Salmonella:All.foodstuffs	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Salmonella:Alpacas	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Staphylococcus:All.animals	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Staphylococcus:All.feedingstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Staphylococcus:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Staphylococcus:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.animals:All.feedingstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.animals:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.animals:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.feedingstuffs:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.feedingstuffs:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.foodstuffs:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000

As discussed above, additional information can be incorporated in the model by considering levels of microbial resistance. A random regression forest achieved an explained variance of about 47.7%, with 197.44 terminal nodes, 5 minimum depth. Variable importance is shown in Figure 57, where the importance measure represents the proportion of variance explained.

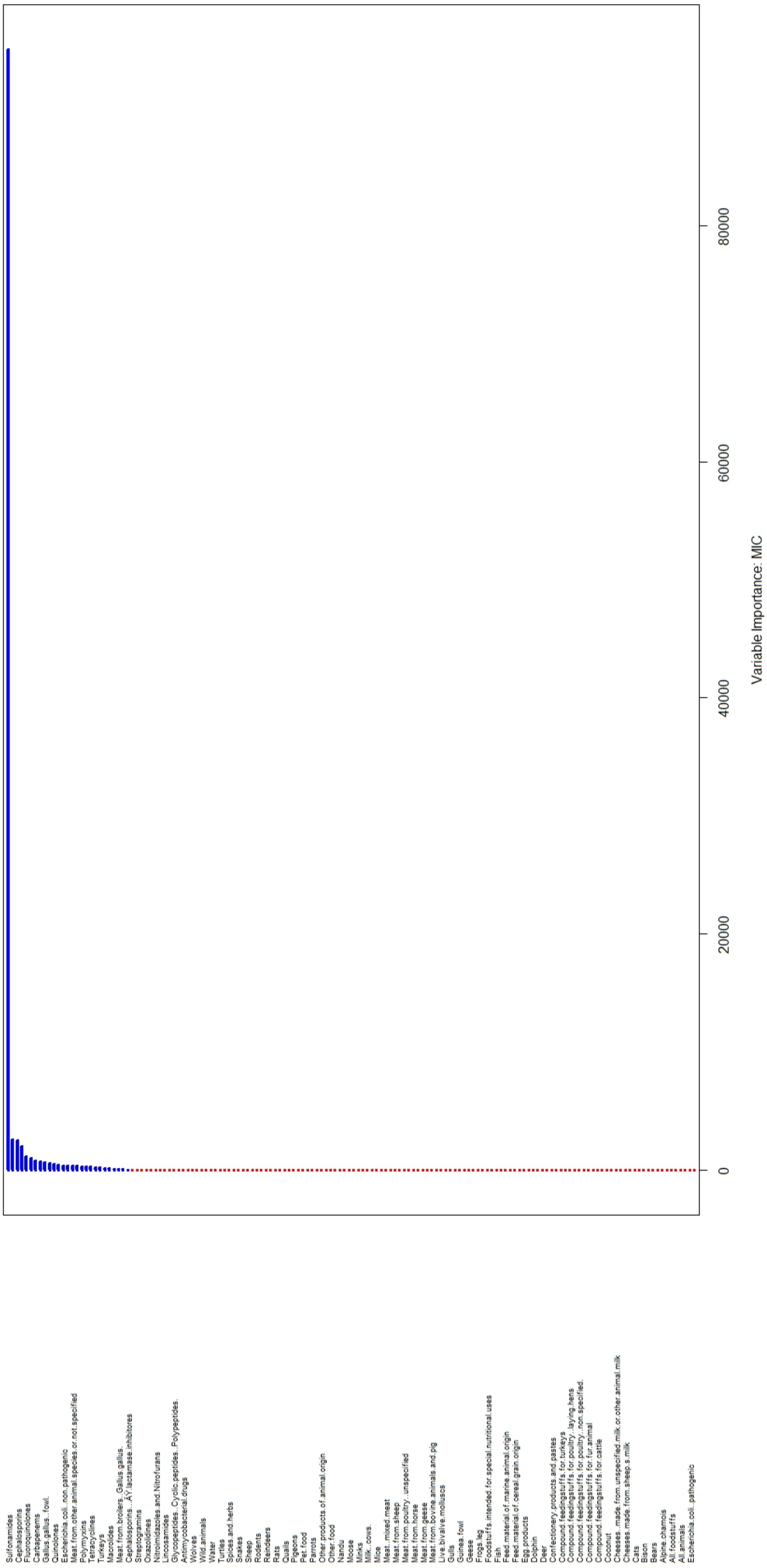


Figure 57: Variable importance for regression on levels of microbial resistance.

The selection process is similar to the one of the classification random forest. Sulfonamides, quinolones, pigs and non pathogenic *Escherichia coli* and *Salmonella* ends up being the most important variables. All together, 25 variables are selected.

Table 90: Variables selected (minimal depth) using a high conservative approach for level of microbial resistance.

Top variables	depth
Sulfonamides	1.117
Pigs	2.206
Quinolones	2.411
Gallus.gallus..fowl.	2.552
Escherichia.coli..non.pathogenic	2.660
Salmonella	2.899
Meat.from.broilers..Gallus.gallus.	2.926
Turkeys	3.013
Tetracyclines	3.137
Campylobacter	3.739
Penicillins	3.953
Meat.from.turkey	4.016
Meat.from.pig	4.399
Amphenicols	5.191
Cattle..bovine.animals.	5.278
Other.poultry	5.725
Macrolides	5.964
Aminoglycosides	6.417
Meat.from.other.animal.species.or.not.specified	6.888
Trimethoprim	7.147
Meat.from.bovine.animals	7.716
Cephalosporins	7.815
Carbapenems	8.270
Fluoroquinolones	9.299

The risk analysis of the levels of risk of resistance observed in different environments, for different zoonotic agents, for the use of antibiotics is based on the analysis of the interactions among such variables, resulting in very similar conclusions to the classification problem previously shown (Table 91).

6.4 MLT modification to fit specific issues

6.4.1 Addressing the issue of feature selection

In the previous case study, the variable importance scores are crucial for identifying those factors associated to an increase of the risk of developing resistances. Variable importace scores are computed as the average of an impurity index, usually the Gini index, over all the trees that form the random forest. Another way to evaluate a split is the entropy. The Gini information gain of splitting a node by a variable is the difference between the entropy at that node and the weighted average of entropies at its child nodes, i.e. the difference between the entropy in children nodes and the parent node. Regularization strategies can be embedded in the training of random forests in order to perform feature selection making advantage of the ranking given by the variable importance score. As for the modification proposed in the Case Study 1, the idea is to reformulate the problem in the terms of a regularization problem. A very simple strategy is to impose a penalty to the Gini information gain. This can be done by building a random forest and get the importance score for each variable. Then normalizing the importance score $\frac{VarImp(X_i)}{max_{i=1,...nVar}VarImp(X_i)}$ and using it to penalized the Gini information gain of using the variable X_i to split the node:

$$Penalized_{Gain}(X_i) = \frac{VarImp(X_i)}{max_{i=1,...nVar}VarImp(X_i) \times Gain(X_i)}$$

In this way, features with smaller importance scores are penalized more and overall the random forest is forced to use a smaller number of features.

Table 91: Joint-importance based analysis of the interactions among 10 randomly chosen variables (for illustration purposes only) for the regression problem of levels of resistance.

	Var 1	Var 2	Paired	Additive	Difference
Campylobacter:Enterococcus..non.pathogenic	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:Escherichia.coli..non.pathogenic	0.016067403	0.015913286	0.023349995	0.031980689	-0.008630694
Campylobacter:Escherichia.coli..pathogenic	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:Salmonella	0.016067403	0.015913286	0.021599815	0.031980689	-0.010380874
Campylobacter:Staphylococcus	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:All.animals	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:All.feedingstuffs	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:All.foodstuffs	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Campylobacter:Alpacas	0.016067403	0.015913286	0.016067403	0.031980689	-0.015913286
Enterococcus..non.pathogenic:Escherichia.coli..non.pathogenic	0.000000000	0.000000000	0.007648699	0.000000000	0.007648699
Enterococcus..non.pathogenic:Escherichia.coli..pathogenic	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:Salmonella	0.000000000	0.000000000	0.005823052	0.000000000	0.005823052
Enterococcus..non.pathogenic:Staphylococcus	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:All.animals	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:All.feedingstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Enterococcus..non.pathogenic:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..non.pathogenic:Escherichia.coli..pathogenic	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:Salmonella	0.007752569	0.007219941	0.010372372	0.014972509	-0.004600137
Escherichia.coli..non.pathogenic:Staphylococcus	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:All.animals	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:All.feedingstuffs	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:All.foodstuffs	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..non.pathogenic:Alpacas	0.007752569	0.007219941	0.007752569	0.014972509	-0.007219941
Escherichia.coli..pathogenic:Salmonella	0.000000000	0.000000000	0.005823052	0.000000000	0.005823052
Escherichia.coli..pathogenic:Staphylococcus	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..pathogenic:All.animals	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..pathogenic:All.feedingstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..pathogenic:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Escherichia.coli..pathogenic:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Salmonella:Staphylococcus	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Salmonella:All.animals	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Salmonella:All.feedingstuffs	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Salmonella:All.foodstuffs	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Salmonella:Alpacas	0.005795101	0.003475631	0.005795101	0.009270732	-0.003475631
Staphylococcus:All.animals	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Staphylococcus:All.feedingstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Staphylococcus:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Staphylococcus:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.animals:All.feedingstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.animals:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.animals:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.feedingstuffs:All.foodstuffs	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.feedingstuffs:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
All.foodstuffs:Alpacas	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000

Implementation in R is quite straightforward using the package randomForest and slightly modifying the *randomForest* function to penalize the Gini information gain with the importance score normalized.

For illustration purposes only, the procedure will be shown only on the subset of data from UK.

```
rf <- randomForest(Data[,-1], as.factor(Data[,'Resistance'])
imp <- rf$importance[,"MeanDecreaseGini"]
impVarNorm <- imp/max(imp)
```

where `Data[,-1]` represents the X matrix and `Data[,'Resistance']` the Y variable.

From an operative point of view, the procedure consists in redefining the `impout` matrix inside the function randomForest which calls a C++ routine. The simplest way is to define the `impout` parameter. The `impout` parameter is a matrix with number of rows equal to the number of covariates and nclass + 2 columns, where nclass is the number of categories of the Y variable. Defining the new `impout` parameter as `impout` in our specific example VIMP is a matrix with 4 columns (2 columns, one for each level of the outcome variable plus two other columns). Since the last column of the `impout` matrix contain the Gini Gain score, VIMP is a matrix with all columns of 1's but the last one with the variable importance values normalized.

The standard random forest assigns to 17 features out of the 150 present in the dataset a variable importance greater than 0 and it produces an out-of-bag error rate equal to 11.49%. In the table 92, the confusion matrix is reported.

Table 92: Random forest confusion matrix

	0	1	class.error
0	6361.00	0.00	0.00
1	908.00	0.00	1.00

The importance variable values are reported in the table 93. Only the variable with values greater than 0 are reported

Table 93: importance variable values

	imp_p[imp_p > 0]
Campylobacter	15.76
Salmonella	13.87
Gallus.gallus..fowl.	1.85
Turkeys	2.02
Aminoglycosides	9.98
Amphenicols	1.56
Carbapenems	1.60
Cephalosporins	4.96
Fluoroquinolones	3.35
Glycylcyclines	0.37
Macrolides	5.80
Penicillins	0.21
Polymyxins	1.05
Quinolones	4.16
Sulfonamides	9.90
Tetracyclines	43.00
Trimethoprim	0.15

.

Modifying the `randomForest` function as explained, in order to allow for using a penalization of the Gini information gain, 14 features, instead of 17, have variable importance greater than 0. Moreover, the out-of-bag error is slightly better, and it is equal to 11.49%. The confusion matrix is shown in Table 94.

Things change little when considering the classes are imbalanced (1 case every 7 controls). Setting the random forest in order to take into account this problem, means setting the parameter `sampSize` of the function

Table 94: Random Forest confusion matrix

	0	1	class.error
0	6239.00	122.00	0.02
1	713.00	195.00	0.79

randomForest downsampling the number of controls (for example *sampSize = c(1500, 908)*. In Table 95 the confusion matrix is shown.

Table 95: Random Forest confusion matrix with adjustment for imbalanced classes

	0	1	class.error
0	6252.00	109.00	0.017
1	693.00	215.00	0.76

The selected variables along with their importance score are reported in the Table 96.

Table 96: selected variables along with their importance score

	imp_p[imp_p > 0]
Salmonella	92.83
Turkeys	49.95
Aminoglycosides	3.29
Amphenicols	2.52
Carbapenems	1.47
Cephalosporins	0.78
Fluoroquinolones	44.94
Glycylcyclines	2.73
Macrolides	1.65
Polymyxins	2.29
Quinolones	44.50
Sulfonamides	49.72
Tetracyclines	157.20
Trimethoprim	6.32

In the following figure 58 the variable importances for the two approaches are plotted togheter.

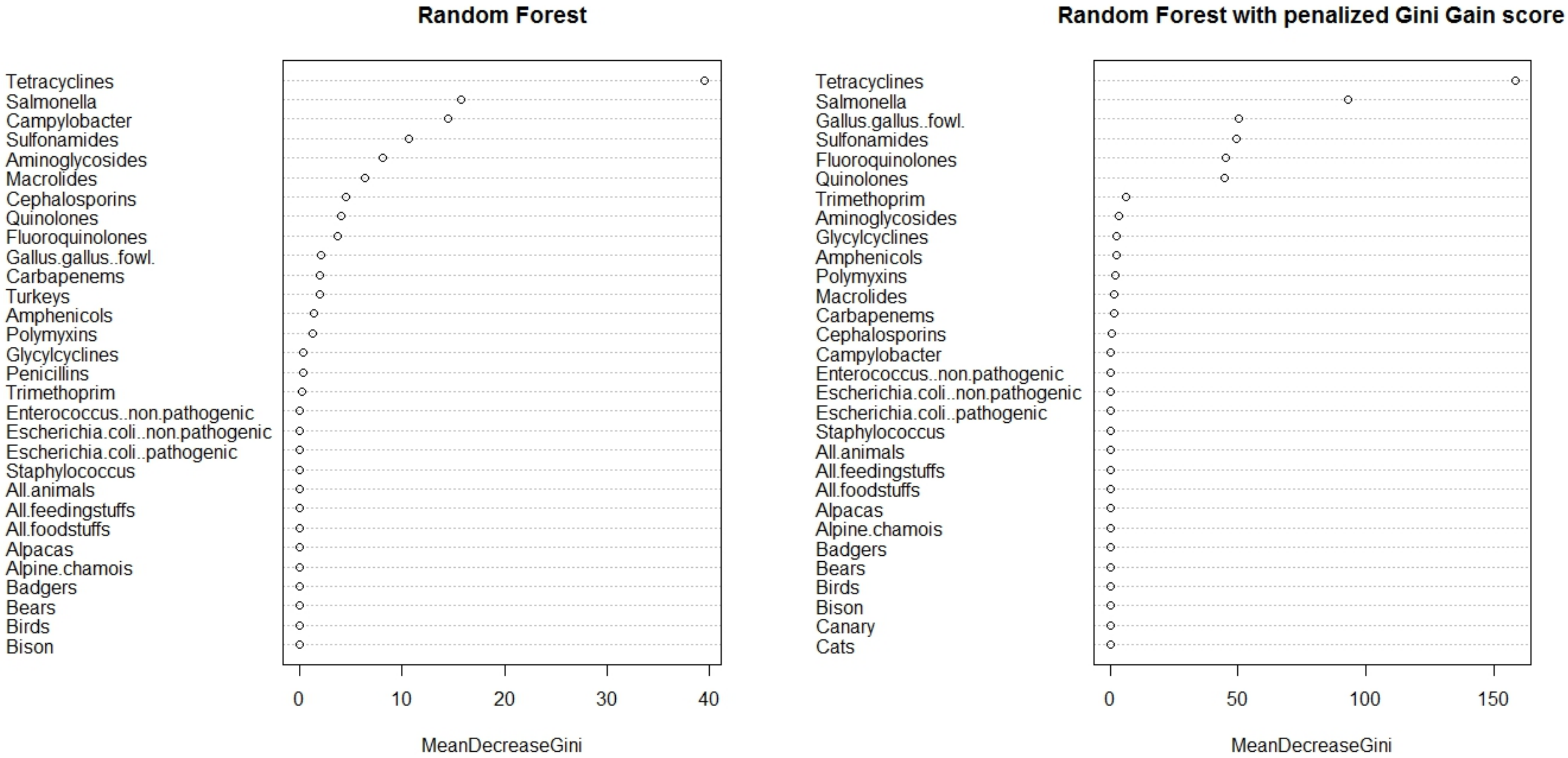


Figure 58: Comparison of variable importance.

6.4.2 Conclusions

In this section a slight modification of the impurity measure used at each split of the node for rule assignment has been proposed. Random forests provide variable importance scores of features, which facilitate understanding the contribution of each independent variable. However, when there is a huge number of variables, it could be of benefit to have a stronger feature selection procedure. In our particular example, 17 out of 150 variables resulted in an importance score greater than 0. However implementing a more stringent procedure, a random forest with 14 out of 150 variable ended up with a better out-of-bag estimation error. In this procedure, a penalized version of Gini Gain Information based on normalized variable importance score is presented and it is motivated by the fact that smaller importance scores are more penalized and thus force the random forest to be more selective using variables for the rule assignment. This procedure is depicted for reducing also the computation burden of building random forests when p (the number of features) is very large. However, as in the case of the previous modification it requires a more extensive assessment, possibly on already studied dataset, in order to have a benchmark on which to compare results.

6.5 Food-borne outbreak

A total of 5,648 food-borne outbreaks were reported in the European Union affecting 69,553 human cases with 7,125 hospitalisations and 93 deaths. Most of the reported outbreaks were caused by: Salmonella, bacterial toxins, Campylobacter and viruses.

The outbreak with most human cases was caused by Shiga toxin-producing Escherichia coli/verotoxigenic Escherichia coli and associated with sprouted seeds. The most important food sources of the outbreaks are eggs and egg products, mixed food and fish and fish products. In 2011, 11 waterborne outbreaks caused by Campylobacter, calicivirus, Cryptosporidium hominis and verotoxigenic Escherichia coli were reported.

The objective of this case study is to train a Superlearner to predict risk of hospitalization in food borne outbreak.

In order to develop the case study, data on food borne outbreaks occurred in 2011 have been considered along with data about animal disease and the SuperLearner trained have been tested on data about food borne outbreak occurred in 2012. The aim of the analysis consists in using a SuperLearner as a predictive model of human hospitalization and exploring some potential pattern related to the food-borne outbreak severity.

6.6 Ensemble learning

Ensemble methods in Machine Learning use more than one learning algorithm to obtain better predictive performance than could be otherwise obtained from any of the single base learning algorithms.

The main reason in developing an ensemble algorithm is that if the set of base learners does not contain the true prediction function, the ensemble can give a good approximation of that function.

There are two mainly algorithms for developing ensemble models: 1. Super Learner algorithm 2. Subsemble algorithm.

The Super Learner algorithm is a loss-based supervised learning approach that finds the optimal combination of a collection of prediction algorithms. It performs asymptotically as well as best possible weighted combination of the base learners.

The Subsemble algorithm is a generalization of the Super Learner and is aimed at combining subset-specific algorithm fits by dividing the dataset randomly into J subsets and fitting a Super Learner in each of them. When $J = 1$, the Subsemble is equivalent to the Super Learner algorithm

6.6.1 Methods

In order to develop a Super Learner algorithm, it is necessary to:

1. define a base learner library of learners $\Psi_1, \dots, \Psi_L$;
2. specify a meta-learning method, Φ;
3. get a partition of the training observation into V-folds to carry out the cross-validation to evaluate its performance.

The following outline summarize how the Super Learner algorithm works:

1. it generates a matrix Z of size $n \times L$ of cross-validated prediction:

  - during the cross-validation, it obtains fits $\hat{\Psi}^l_{-v}$ defined as fitting Ψ^l on the observations that are not in fold v;
  - predictions are generated for the observations in the v^{th} fold;

2. it finds the optimal combination of subset-specific fits according to the specified meta-learner algorithm $\hat{\Phi}$ with a new matrix Z
3. it fits L models, one for each base learning algorithms, on the original training set X and save the L individual model fit object along with $\hat{\Phi}$.

The ensemble model obtained in step 3 can be used to make predictions on new data.

6.6.2 The Subsemble algorithm

In order to develop a Subsemble algorithm, it is necessary to:

1. define a base learner library of learners $\Psi_1, \ldots, \Psi_L$;
2. specify a meta-learning method, Φ;
3. divide the dataset into J subset at random;
4. for each of the J subset get a partition of the training observation into V-folds.

The following outline summarize how the Subsemple algorithm works:

1. it generates a matrix Z of size $n \times (J \times L)$ of cross-validated prediction:
    - during the cross-validation, it obtains fits $\hat{\Psi}^l_{j,-v}$ defined as fitting Ψ^l on the observations that are in the j-esim subset but not in fold v over the subsets;
    - predictions are generated for the observations in the v^{th} fold;
2. it finds the optimal combination of subset-specific fits according to the specified meta-learner algorithm $\hat{\Phi}$ with a new matrix Z
3. it fits $J \times L$ models on the J subsets, one for each base learning algorithms, on the original training set X and save them along with $\hat{\Phi}$.

The ensemble model obtained in step 3 can then be used to make predictions on new data.

6.6.3 Ensemble modeling packages in R

The are three packages in R to carry out ensemble modeling:

1. SuperLearner package, developed by Eric Polley, University of California - Berkeley
2. Subsemple package, developed by Erin LeDell, University of California - Berkeley
3. h2oEnsemble, which is an interface for H20 Java machine learning library (whose algorithms have distributed implementations that work over clusters), developed by H20.ai team

All three packages are available for download from the CRAN. However they have some enhaced features in the version available for download from Git

6.7 Super learner applied to Italian FBO data

The objective of this case study is to train a super learner to predict risk of hospitalization in food borne outbreak.

In order to develop the case study, data on food borne outbreaks occurred in 2011 will be considered along with data about animal disease and the Super Learner trained will be tested on data about food borne outbreaks occurred in 2012

Descriptive data about FBOs in Italy in 2011 are reported in the table 97

In order to build the SuperLearner, data on disease animals, were grouped by Country, zoonosis, matrix and unitDS and then were merged by Country. The analyses were then run considering all countries

Table 97: Description of food borne outbreak, year 2011, country: Italy

n	13	
fboStrengthStrong = N (%)	13	(100.0)
fboAgentGroup (%)		
Bacillus - B. cereus	1	(7.7)
Campylobacter - Campylobacter	1	(7.7)
Clostridium - Cl. perfringens	1	(7.7)
Listeria - Listeria monocytogenes	1	(7.7)
Other Bacterial agents - Shigella	1	(7.7)
Salmonella - Other serovars	1	(7.7)
Salmonella - S. Enteritidis	1	(7.7)
Salmonella - S. Typhimurium	1	(7.7)
Staphylococcal enterotoxins - Staphylococcal enterotoxins	1	(7.7)
Unknown agent - Unknown agent	1	(7.7)
Viruses - Hepatitis viruses	1	(7.7)
Viruses - Norovirus	1	(7.7)
Viruses - Other Viruses	1	(7.7)
numOutbreaks (mean (sd))	69.85	(190.20)
numHumanCases (mean (sd))	299.00	(723.15)

6.8 SuperLearner

Implementation of the SuperLearner requires the specification of all the algorithms to enter in the esemble model. All the algorithms used and combined in the ensembled SuperLearner are reported below

```
SL.complete.library <- c("SL.bartMachine","SL.bayesglm", "SL.cforest",
                         "SL.gbm","SL.glm","SL.glm.interaction","SL.glmnet",
                         "SL.ipredbagg","SL.leekasso","SL.loess","SL.mean",
                         "SL.nnet","SL.nnls","SL.polymars","SL.randomForest",
                         "SL.rpart",SL.rpartPrune","SL.step","SL.step.forward",
                         "SL.step.interaction", "SL.stepAIC", "SL.svm")
```

Since the computational time to run an algorithm and crossvalidate the SuperLearner in order to achieve an internal validation is about 3.87 hours, for the presentation of the case study, the ensemble SuperLearner will be limited to the following algorithms

```
SL.library <- c("SL.glm.interaction","SL.gbm","SL.ipredbagg",
                "SL.randomForest", "SL.rpartPrune", "SL.cforest",
                "SL.svm","SL.loess")
```

Following a short description of Random Forest (randomforest), Gradient Boosting Machine (gbm) and Support Vector Machine is provided, since they resulted to be the algorithms more preminent in the final model.

6.8.1 Random forest

According to the definition in (Breiman, 2001), Random Forests are formally defined as a combination of tree structured classifiers $h(x, \vartheta_k), k = 1, 2, \ldots, K$, where ϑ_k is a random vector that meets the independent and identically distributed assumption.

The procedure to build a random forest as reported in

- Draw a bootstrap sample from the dataset
- Train a decision tree
  - Until the tree is maximum size
    - Choose next leaf node
    - Select m attributes at random from the p available

    - * Find the best attribute/split based on some impurity measure

- Measure out-of-bag error
  - Evaluate against the samples that were not selected in the bootstrap
  - Provides measures of strength, correlation between trees and variable importance

6.8.2 Support Vector Machine

Support Vector Regression (SVR) is applied for forecasting in regression framework by introducing an alternative loss function. The loss function is modified to include a distance measure. It employs a rich class of non-linear modeling functions via the use of kernels. For the current research, svmPoly kernel was used to obtain the support vectors. This kernel takes in three parameters namely degree, scale and cost. A grid search was performed to choose these parameters automatically. Root mean square error was the metric used to select the efficient parameters for each and every model

6.8.3 Gradient Boosting Machine

Gradient Boosting Machine is a tree based model involving a recursive addition to the initial learning from the residuals. It fits a tree based model on the residuals using the specified list of variables at hand and explains the variance in the residuals. Total number of trees specified for model building was 500 with interaction depth as 5 and learning weight of iteration was 0.1. Interaction depth more than 2 imply a model with potential level of interaction higher than 2-ways. A small value of learning weight (the standard parameter in 1) allows for reducing the effect of new tree, avoiding overfitting, since new trees tend to fit the training data.

6.8.4 The SuperLearner

Table 98: The resulting SuperLearner

	Risk	Coef
SL.glm.interaction_All	5.66074647	0.0000000
SL.gbm_All	0.07443113	0.3051609
SL.ipredbagg_All	0.07613827	0.0000000
SL.randomForest_All	0.07379368	0.3953483
SL.rpartPrune_All	0.07700123	0.1091586
SL.cforest_All	0.08351686	0.0000000
SL.svm_All	0.08150051	0.1903322
SL.loess_All	0.10305000	0.0000000

Coefficients are computed using as meta-learner algorithm the non-linear least square. Then, a V-fold crossvalidation was implemented in order to validate internally the SuperLearner.

Table 99: The resulting SuperLearner

Algorithm	Ave	se	Min	Max
Super Learner	0.073533	0.0042910	0.068121	0.086661
Discrete SL	0.074266	0.0041975	0.068793	0.087630
SL.glm.interaction_All	15.727183	12.1476026	0.375031	74.662581
SL.gbm_All	0.074767	0.0040816	0.068636	0.088478
SL.ipredbagg_All	0.077039	0.0042533	0.069170	0.090698
SL.randomForest_All	0.074266	0.0041975	0.068793	0.087630
SL.rpartPrune_All	0.078931	0.0044068	0.070510	0.093366
SL.cforest_All	0.084082	0.0043691	0.073763	0.101016
SL.svm_All	0.082896	0.0052631	0.075373	0.093158
SL.loess_All	0.102664	0.0047551	0.095172	0.115994

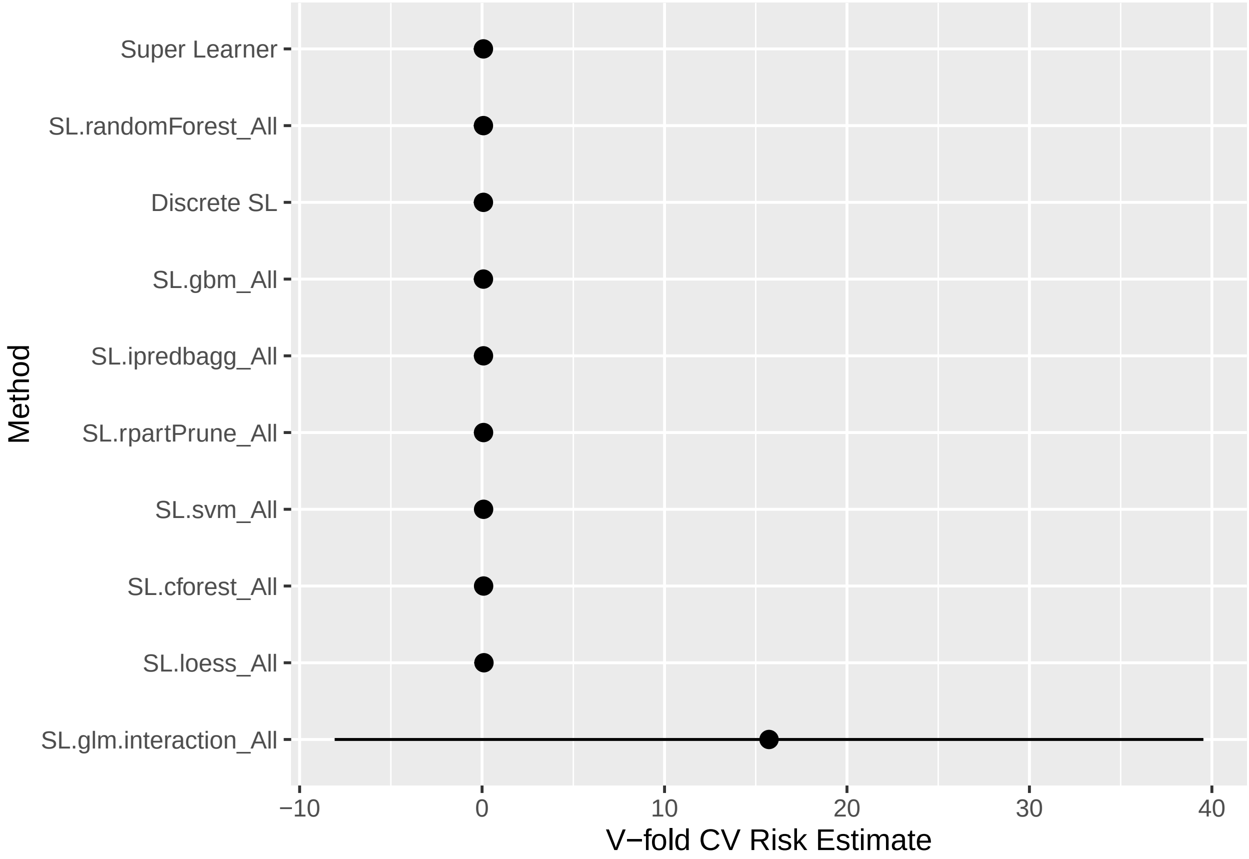


Figure 59: Error rate of the base learners and the Super Learner

6.8.5 Using the SuperLearner as predictive model

The SuperLearner can be used as predictive model to predict the expected risk of hospitalization for the year 2012. The prediction for each of the algorithm used to build the Super Learner is reported in table 100)

Table 100: The resulting SuperLearner

SL.glm.interaction_All	SL.gbm_All	SL.ipredbagg_All	SL.randomForest_All	SL.rpartPrune_All	SL.cforest_All	SL.svm_All	SL.loess_All
1.04E+12	0.2323803	0.08817702	0.2171971	0.0991308	0.2658688	0.2180109	0.1785453
1.04E+12	0.5123602	0.45829264	0.7609412	0.7954545	0.2782993	0.2180109	0.1785453
1.04E+12	0.2364158	0.08817702	0.2302078	0.0991308	0.2703081	0.2180109	0.1785453
1.04E+12	0.2673008	0.08817702	0.2209358	0.0991308	0.2660564	0.2180109	0.1785453
1.04E+12	0.499446	0.45915351	0.7828503	0.7954545	0.2779359	0.2180109	0.1785453
1.04E+12	0.5482822	0.46892672	0.4936924	0.7954545	0.2720939	0.2180109	0.1785453

Finally, it provides the overall prediction given by the weighted combination of the single algorithms' prediction. Weights correspond to the coefficients estimated during the cross-validation.

Prediction obtained from the Super Learner can be used to explore relationships with predictors. In the figure 60 is depicted the relationship between the estimated risk of hospitalization and the variable Cattle_bovine_animals_Number_of_infected_herds_Mycobacterium.

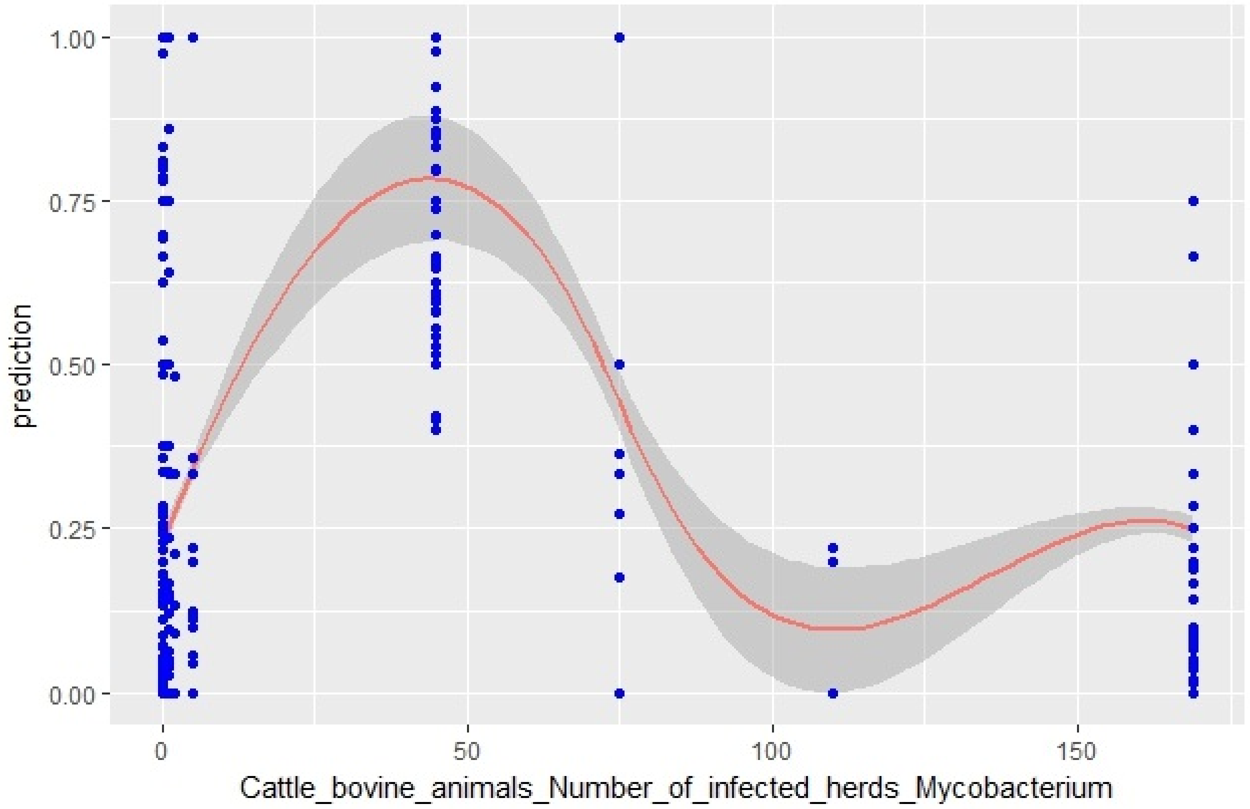


Figure 60: Pattern in prediction.

6.9 Monitoring similarities in zoonotic agent

The aim of the case study is to find similar predicting pattern in AMR between different zoonotic agents. Many MLT approaches are used in this field when genetics data are available ((Lupolova et al., 2016; Niehaus et al., 2014; Zhu et al., 2011)). However, little has been done using data mining perspective on observational data (Giannopoulou et al., 2007).

Based on the data collected a MLT can be trained to predict the lower resistant concentration each pair of zonotic-antimicrobial every year. We start to explore only a selection of data which ara complete for every zonotic-year subset and present with more than a single unique value in all of them. This way each algorithm trained on a zonotic-year pair can be used to predict the lower resistant concentration for every other subset considered.

The final selection considers:

- every years from 2010 to 2014:
- the zoonoses_L1: Campylobacter, Enterococcus (non-pathogenic)
- the variables: lowest, repCountry, matrix_L1, matrix_L2, matrix_L3, sampStage, sampType_L1, sampType_L2, sampler, sampArea, anMethCode, substance_L1, substance_L2.

So, finally we trained 15 (5 years times 3 agents) MLT where each of them is used to predict the results of the other 14 ones.

The path used to chose the MLT follow the decision tree (Table 79) in the following way:

- Label = supervised (train + test)
- Input = any (categorical + continuous cutoff)
- Output = single/any (concentration level/ continuous-categorical)
- Linear = non linear
- Scalable = multiple (one train vs many tests)
- Sample size = small/medium (<10k rows)
- Relation = n > p
- Miss = no (only complete rows with a minimum concentration level)

Leading to the SVM algorithm.

The choice of the R-implementation (i.e. package and function) to be used was based on the package table of the decision tree (Table 80)

- Interact = supervised
- Class = yes (discrete concentration)
- Regression = no
- Cluster = no
- Filter = indifferent
- Dim = any
- Type of input = numeric (factor, i.e. not character)

Table 101: 5-Fold Crossvalidated accuracy each years for SVMs trained

agent	Acc 2010 (%)	Acc 2011 (%)	Acc 2012 (%)	Acc 2013 (%)	Acc 2014 (%)
Campylobacter	97.656	94.015	94.776	95.92	92.892
Enterococcus, non-pathogenic	97.874	84.34	90.506	87.259	90.767
Escherichia coli, non-pathogenic	95.657	93.426	93.152	92.898	98.305

Table 102: Confusion matrix for 2010.Campylobacter.pred-Enterococcus, non-pathogenic predictions.

	Reference																	
Prediction	0.008	0.015	0.016	0.03	0.06	0.12	0.25	0.5	1	128	16	2	256	32	4	512	64	8
0.008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.06	0	0	0	0	0	0	564	630	0	0	0	0	0	0	0	0	0	0
0.12	0	0	0	0	0	415	415	1128	564	457	797	0	0	0	564	0	0	0
0.25	0	0	0	0	0	0	0	0	1247	0	0	0	0	0	0	0	0	0
0.5	0	0	0	0	0	0	564	1623	0	0	564	0	0	0	0	0	0	0
1	0	0	833	0	0	0	810	911	1509	1291	0	2435	0	373	417	0	876	914
128	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	1029	0	0	564	0	0	0
256	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
512	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
64	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Leading to choose the `e1071::svm()` implementation.

The 5-fold crossvalidated accuracies ranges from 84% to 98% on the trained machines (Table 101)

The High 5-crosvalidate accuracy show how a glssvm are capable to predict the lower resistant concentration for a zoonotic agent in given year based on the variable used.

The result of applying each SVM trained to predict the lower resistant concentration of a different zoonotic agent for the same year leads to 30 prediction. In the Table 102 an example of the resulting confusion matrix is reported.

The overall accuracies for all the 30 predictions are reported in Table 103

It can be seen that none of the algorithms achieves a satisfactory level of accuracy to suggest overall similarities between agents in the same year.It could be of interest to see the accuracy of the prediction of each level of resistant concentration. An example is reported in Table 104

6.10 Conclusions

With the aim of illustrating the potential for application of MLT in Risk Assessment, five case studies have been proposed based on data from the European Union Summary Reports on Zoonoses and on Antimicrobial Resistance. Random forests, clustering methods and ensemble models have been illustrated and specific strategies like cross-validation to address well-known issues like over-fitting have been shown. The choice of fit for purpose MLTs in these case studies can be made by using the decision tree developed in 4. Finally, to adapt MLT to some specific issues, like small sample size and the need of generalization of the results, modifications to standard MLT have proposed and illustrated into two ad hoc case studies. Such results are provided to give rise to discussion on the epidemiological added value of using MLT and eventually to refine the approaches adopted when addressing food safety related issues.

Table 103: Overall accuracy for the predicted lower resistant concentration

year.train.predict	overall accuracy (%)
2010.Campylobacter.pred-Enterococcus, non-pathogenic	0.21289662
2010.Campylobacter.pred-Escherichia coli, non-pathogenic	0.19212748
2010.Enterococcus, non-pathogenic.pred-Campylobacter	0.19931479
2010.Enterococcus, non-pathogenic.pred-Escherichia coli, non-pathogenic	0.19135300
2010.Escherichia coli, non-pathogenic.pred-Campylobacter	0.25728873
2010.Escherichia coli, non-pathogenic.pred-Enterococcus, non-pathogenic	0.16544152
2011.Campylobacter.pred-Enterococcus, non-pathogenic	0.16174552
2011.Campylobacter.pred-Escherichia coli, non-pathogenic	0.20286975
2011.Enterococcus, non-pathogenic.pred-Campylobacter	0.13790479
2011.Enterococcus, non-pathogenic.pred-Escherichia coli, non-pathogenic	0.09616057
2011.Escherichia coli, non-pathogenic.pred-Campylobacter	0.26785220
2011.Escherichia coli, non-pathogenic.pred-Enterococcus, non-pathogenic	0.16868090
2012.Campylobacter.pred-Enterococcus, non-pathogenic	0.13910823
2012.Campylobacter.pred-Escherichia coli, non-pathogenic	0.18656572
2012.Enterococcus, non-pathogenic.pred-Campylobacter	0.16541353
2012.Enterococcus, non-pathogenic.pred-Escherichia coli, non-pathogenic	0.10325594
2012.Escherichia coli, non-pathogenic.pred-Campylobacter	0.15248648
2012.Escherichia coli, non-pathogenic.pred-Enterococcus, non-pathogenic	0.09162455
2013.Campylobacter.pred-Enterococcus, non-pathogenic	0.18092403
2013.Campylobacter.pred-Escherichia coli, non-pathogenic	0.18450387
2013.Enterococcus, non-pathogenic.pred-Campylobacter	0.17929103
2013.Enterococcus, non-pathogenic.pred-Escherichia coli, non-pathogenic	0.16150678
2013.Escherichia coli, non-pathogenic.pred-Campylobacter	0.15930942
2013.Escherichia coli, non-pathogenic.pred-Enterococcus, non-pathogenic	0.13600392
2014.Campylobacter.pred-Enterococcus, non-pathogenic	0.23234670
2014.Campylobacter.pred-Escherichia coli, non-pathogenic	0.14334825
2014.Enterococcus, non-pathogenic.pred-Campylobacter	0.27383583
2014.Enterococcus, non-pathogenic.pred-Escherichia coli, non-pathogenic	0.06650423
2014.Escherichia coli, non-pathogenic.pred-Campylobacter	0.01123216
2014.Escherichia coli, non-pathogenic.pred-Enterococcus, non-pathogenic	0.23153854

Table 104: Accuracy stratified by lower level of concentration for 2010.Campylobacter.pred-Enterococcus, non-pathogenic prediction. (NA indicates no report at the level)

	Accuracy (%)
0.008	NA
0.015	NA
0.016	0.5000000
0.03	NA
0.06	NA
0.12	0.9068979
0.25	0.4674259
0.5	0.6562858
1	0.4835042
128	0.5000000
16	0.5000000
2	0.6328871
256	NA
32	0.5000000
4	0.5000000
512	NA
64	0.5000000
8	0.5000000

Case Studies in RA Bibliography

[Bre01] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. issn: 1573-0565. doi: 10.1023/A:1010933404324. url: http://dx.doi.org/10.1023/A:1010933404324.

[CZ01] Adele Cutler and Guohua Zhao. "Pert-perfect random tree ensembles". In: *Computing Science and Statistics* 33 (2001), pp. 490–497.

[Gia+07] Eugenia G Giannopoulou et al. "A Large Scale Data Mining Approach to Antibiotic Resistance Surveillance". In: *Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*. IEEE. 2007, pp. 439–444.

[Ish+10] Hemant Ishwaran, Udaya B Kogalur, Eiran Z Gorodeski, et al. "High-dimensional variable selection for survival data". In: *Journal of the American Statistical Association* 105.489 (2010), pp. 205–217. issn: 0162-1459.

[Ish+11] Hemant Ishwaran, Udaya B Kogalur, Xi Chen, et al. "Random survival forests for highdimensional data". In: *Statistical analysis and data mining* 4.1 (2011), pp. 115–132. issn: 1932-1872.

[Ish07] Hemant Ishwaran. "Variable importance in binary regression trees and forests". In: *Electronic Journal of Statistics* 1 (2007), pp. 519–537. issn: 1935-7524.

[Ish15] Hemant Ishwaran. "The effect of splitting on random forests". In: *Machine Learning* 99.1 (2015), pp. 75–118. issn: 0885-6125.

[LJ06] Yi Lin and Yongho Jeon. "Random forests and adaptive nearest neighbors". In: *Journal of the American Statistical Association* 101.474 (2006), pp. 578–590. issn: 0162-1459.

[LS97] Wei-Yin Loh and Yu-Shan Shih. "Split selection methods for classification trees". In: *Statistica sinica* (1997), pp. 815–840. issn: 1017-0405.

[Lup+16] Nadejda Lupolova et al. "Support vector machine applied to predict the zoonotic potential of E. coli O157 cattle isolates". In: *Proceedings of the National Academy of Sciences* 113.40 (2016), pp. 11312–11317.

[Nie+14] Katherine E Niehaus et al. "Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis". In: *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2014, pp. 618–621.

[Zhu+11] Xiaojin Zhu et al. "Machine learning for zoonotic emerging disease detection". In: *ICML 2011 Workshop on Machine Learning for Global Challenges, Bellevue, WA, USA*. Citeseer. 2011.

Glossary

ACM Association for Computing Machinery. 18, 20, 47, 271

AHAW Animal Health and Animal Welfare. 14, 65, 79

AIC Akaike's information criterion. 119, 172

AJP American Journal of Psychology. 21

AL Active Learning. 16, 28, 39, 40, 50, 60, 271

AMS American Mathematical Society. 21

AMU Assessment and Methodological Support Unit. 80

ANN Artificial Neural Network. 103, 104, 167, 169, 172

ANOVA Analysis of Variance. 75, 157

ANS Food Additives and Nutrient Sources added to Food. 9, 14, 78, 79, 83, 85

AODE Averaged One-Dependence Estimators. 169, 170

APA American Psychological Association. 21

AUC Area Under the Curve. 12, 214

BART Bayesian Regression Trees. 166, 170

BBN Bayesian Belief Network. 169

BC Bibliographic Citation. 9, 11, 16, 18, 22, 24–32, 34–38, 40–43, 47–49, 51, 53, 59, 60, 271, 272

BIB BibTeX. 25, 26, 34

BIC Bayesian Information Criteria. 172

BIOHAZ Biological Hazards. 14, 65, 79

BMD Benchmark Dose. 119, 120, 122, 123

BMDL The statistical lower bound of BMD. 119, 120, 123, 174

BMI Body Mass Index. 9, 98, 101

BMR Benchmark Response. 119

BN Bayesian Network. 169, 170

CART Classification and Regression Tree. 163, 164, 170

CAWI Computer Assisted Web Interviewing. 302

CEF Food Contact Materials, Enzymes, Flavourings and Processing Aids. 14, 79

CENTRAL Cochrane Central Register of Controlled Trials. 20

CHAID Chi-squared Automatic Interaction Detection. 163

CINAHL Cumulative Index to Nursing and Allied Health Literature. 20, 25, 29, 32

CINL Cumulative Index to Nursing Literature. 20

CIS Current Index of Statistics. 271

CNN Convolutional Neural Network. 167

CONTAM Contaminants in the Food Chain. 9, 14, 78, 79, 85

CPAN Comprehensive Perl Archive Network. 26

CRAN Comprehensive R Archive Network. 26

cs Case Study. 10, 95, 132

CSV comma-separated values. 35, 37, 40, 41, 43, 271

CTAN Comprehensive TEXArchive Network. 26

CTM Correlated Topic Models. 9, 64–72, 75

CTREE Conditional Inference Trees. 11, 98, 105, 132, 163

DAMA Daphnia Magna. 6, 10, 95, 129

DARE Database of Abstracts of Reviews of Effects. 20

DATA Evidence Management Unit. 80

DB Data Base. 9, 16–18, 20, 22, 25–28, 35, 36, 40–43, 48, 56, 271, 273

DBLP Digital Bibliography & Library Project. 20

DBM Deep Boltzmann Machine. 167

DBMS Database Management System. 25

DBN Deep Belief Networks. 167

DM Data Mining. 175

DMS Document Management System. 271

DOAJ Directory of Open Access Journal. 20, 31, 271

DSCB Dipartimento di Scienze Cliniche e Biologiche. 28, 40

EFSA European Food Safety Authority. 1, 3, 4, 6, 11, 14–18, 22, 25, 26, 43, 46, 63–66, 68, 69, 72–80, 82, 95, 119, 173–176, 211–213, 215, 271, 302, 303

ELS Extensive Literature Search. 15, 16

EM Expected Minimization. 157

EMBASE Excerpta Medica dataBASE. 20

EOF End of Field. 26, 35

EOR End of Record. 35, 36

EPIC European Prospective investigation into Cancer and Nutrition. 95, 96

export style Struttura di esportazione dati di EndNote. 26, 31, 34, 35, 271

FDA Flexible Discriminant Analysis. 158

FDE First Data Entry. 40–42

FEEDAP Additives and Products or Substances used in Animal Feed. 14, 65, 79

GBM Gradient Boosting Machines. 165

GBRT Gradient Boosting Regression Tree. 165

GLM Generalized Linear Model. 132, 156, 157, 173

GMO Genetically Modified Organisms. 14, 65, 67, 69, 79, 173

GNU GNU is Not Unix. 26, 27

GPL General Public License. 26, 27

GUI Graphical User Interface. 26

HAC Hierarchical Agglomerative Clustering. 158

HTA Health Technology Assessment. 20

ID3 Iterative Dichotomizer 3. 163

IEEE Institute of Electrical and Electronics Engineers. 20

IET Institution of Engineering and Technology. 20

INDUP Internal Duplicate. 36, 271

ISI Institute for Scientific Information. 22

IZSTO Istituto Zooprofilattico Sperimentale del Piemonte, Liguria e Valle d'Aosta. 64

JEL Journal of Economic Literature. 20

KDB k dependence Bayesian. 170

KNN k-Nearest Neighbour. 11, 98, 102, 106, 132, 160, 172

LARS Least-Angle Regression. 170, 172

LASSO Least Absolute Shrinkage and Selection Operator. 134, 135, 156, 170–172

LDA Latent Dirichlet Allocation. 9, 11, 63–70, 72, 74, 75, 77, 98, 102, 107, 132, 158, 172

LOAEL Lowest Observed Adverse Effect Level. 129

LR Logistic Regression. 12, 98, 102, 104, 108

LVQ Learning Vector Quantization. 160, 161

LWL Locally Weighted Learning. 160, 161

M&M Materials & Methods. 65

MDA Mixture Discriminant Analysis. 158

MDS Multidimensional Scaling. 158

MEDLINE Medical Literature Analysis and Retrieval System Online. 20, 21

ML Machine Learning. 3, 6, 10, 12, 16, 25, 38, 40, 64, 96, 99, 119, 121–124, 127, 132, 154, 166, 170, 174–176, 180

MLE maximum likelihood estimator. 119

MLP Multi Layer Perceptron. 10, 12, 98, 102, 103, 109, 132, 167, 168

MLPE Multi Layer Perceptron Ensemble. 12, 98, 102, 103, 110

MLT Machine Learning Technique. 1, 3, 4, 6, 9, 15, 16, 25, 43–46, 56, 58, 60, 95, 96, 98–100, 102, 104, 105, 119, 121, 123, 126, 127, 129, 132, 154–156, 172–179, 211–252

MRL Maximum Residue Limit. 67, 69, 71

MSE mean square error. 160, 172

MYSQL My Structured Query Language. 26–28, 35–38, 56, 271

NB Naïve Bayes. 12, 98, 112, 122, 169, 170

NDA Dietetic Products, Nutrition and Allergies. 14, 79

NHS EED National Health Service - Economic Evaluation Database. 20

NIPALS Nonlinear Iterative Partial Least Squares. 159

NLM National Library of Medicine. 21

NO Name co–Occurences. 16, 28, 46, 57, 60

NOAEL Non Observed Adverse Effect Level. 74, 119, 129, 174

OCR optical character recognition. 40

ODBC Open Database Connectivity. 38

OECD Organisation for Economic Co-operation and Development. 129

ORCID Open Researcher and Contributor ID. 21

OT Overall Table. 36, 37, 271

PC Principal Component. 158, 159

PCA Principal Component Analysis. 156, 158, 159, 172

PCR Principal Component Regression. 127, 158, 159

PLH Plant Health. 9, 14, 78, 85

PLS Partial Least Squares Regression. 158, 159

PPR Plant Protection Products and their Residues. 14, 79

PRISMA Preferred Reporting Items for Systematic reviews and Meta-Analyses. 260

QA Quality Assurance. 42

QDA Quadratic Discriminant Analysis. 12, 98, 102, 113, 158, 172

QUOROM QUality Of Reporting Of Meta-analyses. 260

RA Risk Assessment. 5, 6, 14, 15, 60, 63, 66, 67, 69, 75, 95, 119

RASA Risk Assessment and Scientific Assistance. 302

RBFN Radial Basis Function Network. 169, 172

RCT Randomized Controlled Trial. 260

RDBMS Relational Database Management System. 27

REGULA Regularization Algorithms. 170

REPEC Research Papers in Economics. 21, 271

REPRO Scientific Evaluation of Regulated Products. 302

Resource General term for reffering to one of the 22 database/search engines considered. 9, 11, 16–18, 22, 24–26, 28, 29, 31, 32, 34–37, 47–49, 51, 53, 58, 60

RF Random Forest. 12, 98, 104, 114, 127, 128, 132, 165, 175, 213

RIS Research Information Systems. 25, 26, 32, 34

ROC Receiver Operator Curve. 10, 12, 213, 214

RPART Recursive Partitioning and Regression Trees. 12, 98, 102, 115

RQ Risk Question. 5, 9, 63–65, 68–70, 72–75, 77

RT Resource Table. 36, 37, 271

SCER Scientific Committee and Emerging Risks. 80

SDE Second Data Entry. 41, 42

SE Search Engines. 16, 17

SOM Self-Organizing Map. 160, 161, 172

SPODE SuperParent One-Dependence Estimator. 170

SRM Specified Risk Materials. 67

SS Search String. 11, 16, 22, 24–26, 29, 31, 36, 38, 40, 48

SSE sum-of-squared errors. 171

SVM Support Vector Machine. 11, 12, 25, 28, 38–40, 46, 51, 52, 57, 60, 61, 98, 102, 116, 123, 132, 154, 156, 166, 172, 249, 250, 271

TAN Tree Augmented Naïve Bayes. 169

TM Topic Modeling. 63–65, 73, 75

TOR Term of Reference. 64, 65, 74

TSE transmissible spongiform encephalopathies. 67, 68, 71

UBESP Unità di Biostatistica, Epidemiologia e Sanità Pubblica. 22, 27, 28

UNIPD University of Study of Padua. 27

Unique Not replicated records. 47

WOS-AHCI Web of Science Arts & Humanities Citation Index. 25

WOS-CORE Web of Science Core Collection. 29

WOS-SSCI Web of Science Social Science Citation Index. 25

ZETA Zeta Research s.r.l.. 27, 28, 40, 41, 302

Part I

Appendix

A References for validation

Below are listed the references used for the citations, to validate the accuracy of the database, as described in the sub-section 1.2.5. A copy of the original pages of the books with all citations `CitForVal.pdf` are reported in the attached file.

Validation Bibliography

[Dav08] Taniar David. *Data Mining and Knowledge Discovery Technologies*. Hershey, PA, USA: IGI Global, 2008. doi: 10.4018/978-1-59904-960-1. url: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-59904-960-1.

[Fla12] Peter Flach. "Machine Learning: The Art and Science of Algorithms that Make Sense of Data". In: 2012. Chap. 409. isbn: 978-1107422223.

[LDL14] Xin Liu, Anwitaman Datta, and Ee-Peng Lim. *Computational Trust Models and Machine Learning*. Chapman and HALL/CRC, 2014. isbn: 978-1-4822-2666-9.

[Lin09] Tsau Young Lin. *Foundations and Novel Approaches in Data Mining*. Springer-Verlag, 2009, p. 388. isbn: 9783642066504.

[McC07] Colleen McCue. *Data Mining and Predictive Analysis*. Ed. by Colleen McCue. Burlington: Butterworth-Heinemann, 2007, p. 332. isbn: 978-0-7506-7796-7. doi: http://dx.doi.org/10.1016/B978-075067796-7/50024-6. url: http://www.sciencedirect.com/science/article/pii/B9780750677967500246.

[Mur12] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012, p. 1096. isbn: 9780262018029.

[Tor10] Luis Torgo. *Data Mining with R: Learning with Case Studies*. Chapman & Hall/CRC, 2010, p. 305. isbn: 9781439810187.

[Tri10] Evangelos Triantaphyllou. *Data Mining and Knowledge Discovery via Logic-Based Methods*. Springer US, 2010, p. 350. isbn: 978-1-4614-2613-4. doi: 10.1007/978-1-4419-1630-3.

[WF05b] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., 2005. isbn: 0120884070.

B 27 PRISMA checklist

Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) is an evidence-based minimum set of items aimed at helping authors to report a wide array of systematic reviews and meta-analyses that assess the benefits and harms of a health care intervention. PRISMA focuses on ways in which authors can ensure a transparent and complete reporting of this type of research. In 1996, an international group of 30 clinical epidemiologists, clinicians, statisticians, editors, and researchers convened The QUality Of Reporting Of Meta-analyses (QUOROM) conference to address standards for improving the quality of reporting of meta-analyses of clinical Randomized Controlled Trials (RCTs). The PRISMA checklist includes 27 items pertaining to the content of a systematic review and meta-analysis, which include the title, abstract, methods, results, discussion and funding.

See the table 105 of PRISMA Checklist.

Table 105: 27 PRISMA checklist

Topic	#	Checklist Item	Page
		Title	
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
		Abstract	
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	1
		Introduction	
Rationale	3	Describe the rationale for the review in the context of what is already know.	16
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	**??**
		Methods	
Protocol and registration	5	Indicate if a review protocol exists if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	NA
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g.,years considered, language, publication status) used as criteria for eligibility, giving rationale.	22
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched	18-22
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated	29
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis)	35
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators	29-35
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made	NA

Page := (range of) report's page(s) where related topic is treated
NA := not applicable

Table 105: continue on the following page

Table 105: continue from the last page

Topic	#	Checklist Item	Page
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be use in any data synthesis	NA
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means)	NA
Synthesis of results	14	Describe the methods of handling data and comining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	43
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	NA
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indications which were pre-specified	NA
		Results	
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram	46
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	NA
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome-level assessment (see Item 12).	NA
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group and (b) effect estimates and confidence intervals, ideally with a forest plot.	NA
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency	56
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	NA
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16])	NA
		Discussion	
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., health care providers, users, policy makers).	NA
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	58
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	60
		Funding	
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	16

Page := (range of) report’s page(s) where related topic is treated
NA: not applicable

Table 105: end from the last page

C WEBi

C.1 Description

With the aim of establishing an active and ongoing communication with efsa that allows also to view and query the realized mysql database, we implemented a web interface (webi). To access the webi simply connect to the url mlt-webi.zetafield.eu and enter the assigned credentials.

C.2 WEBi manual (ver: 13/07/2015)

C.2.1 Introduction

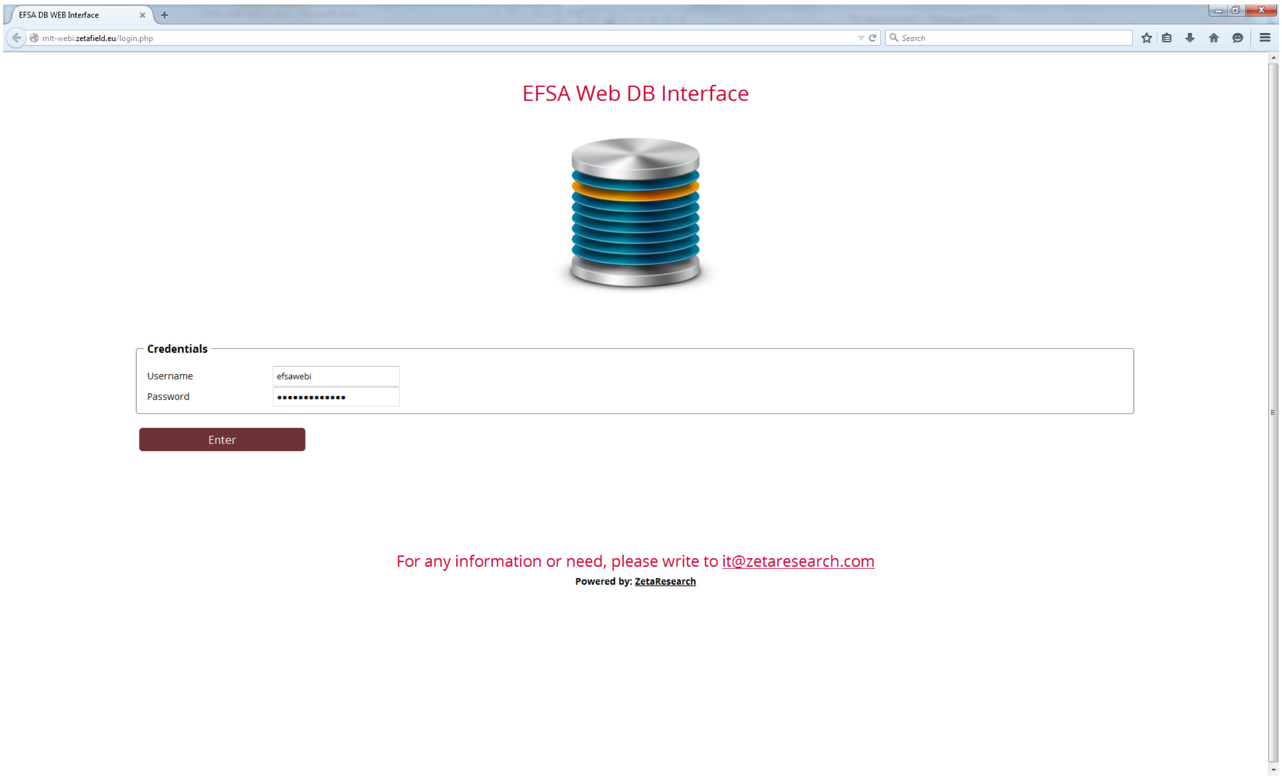
With the aim of establishing an active and ongoing communication with EFSA that allows also to view and query the realized MySQL database, we implemented an interactive interface.

To access the interface simply connect to the url http://mlt-webi.zetafield.eu.

C.2.2 Logging in with your account

We recommend users to access the page using Google Chrome or Mozilla Firefox as a web browser, with those credentials:

- Username: efsawebi
- Password: 0700EFSAwebDB



EFSA Web DB Interface

Credentials

Username efsawebi

Password •••••••••••••

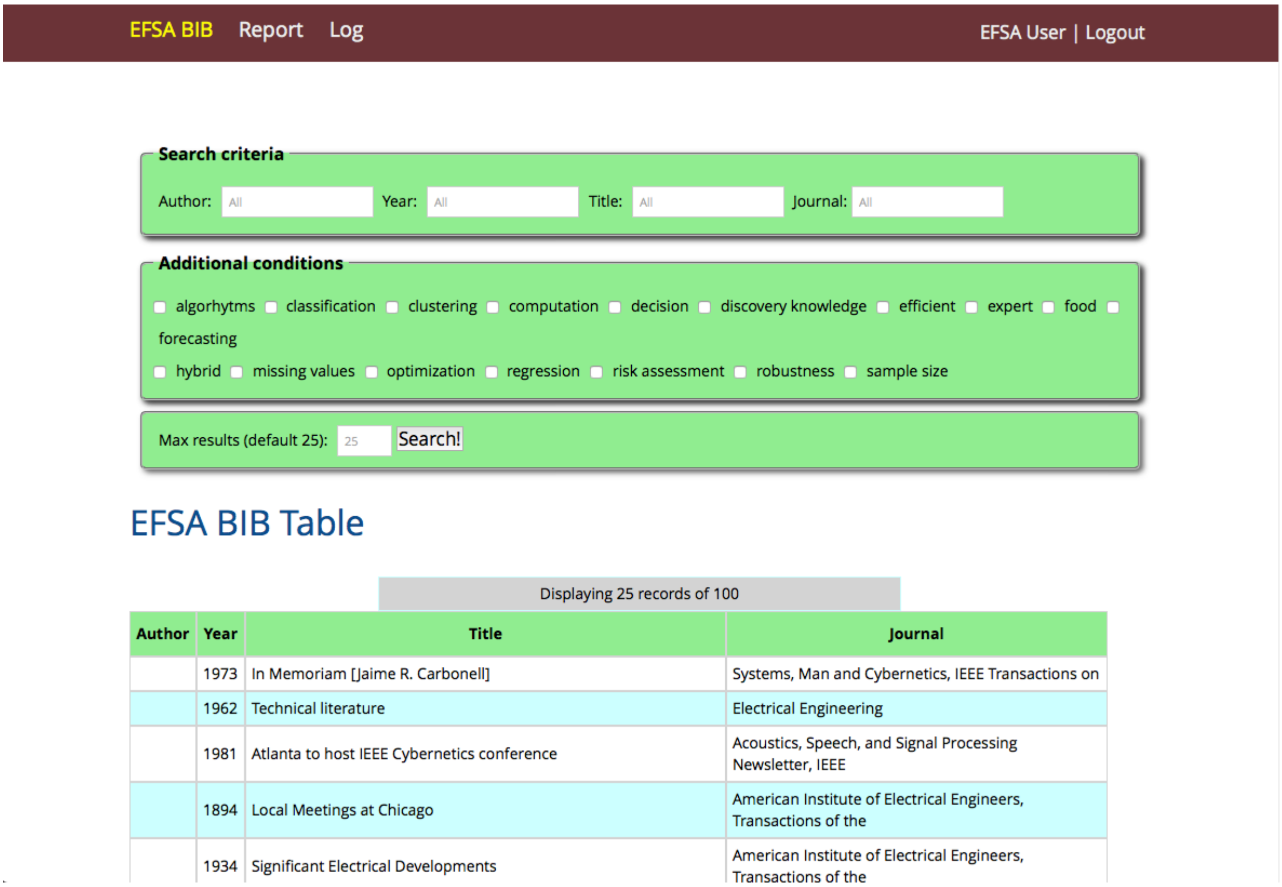
For any information or need, please write to it@zetaresearch.com

Powered by: ZetaResearch

C.2.3 The first page after login

Once in the system, the interface loads the data in the table “EFSA BIB Table”. Since there is a lot of records, the loading time of the table should be around 1-2 minutes with the default options (25 records per page). The table contains 2655365 records. After the loading of the page is done, the user will see a table with 25 rows per page by default.

The number of rows per page can be modified to a custom value.



Search criteria

Author: All Year: All Title: All Journal: All

Additional conditions

☐ algorhytms ☐ classification ☐ clustering ☐ computation ☐ decision ☐ discovery knowledge ☐ efficient ☐ expert ☐ food ☐ forecasting
☐ hybrid ☐ missing values ☐ optimization ☐ regression ☐ risk assessment ☐ robustness ☐ sample size

Max results (default 25): 25 Search!

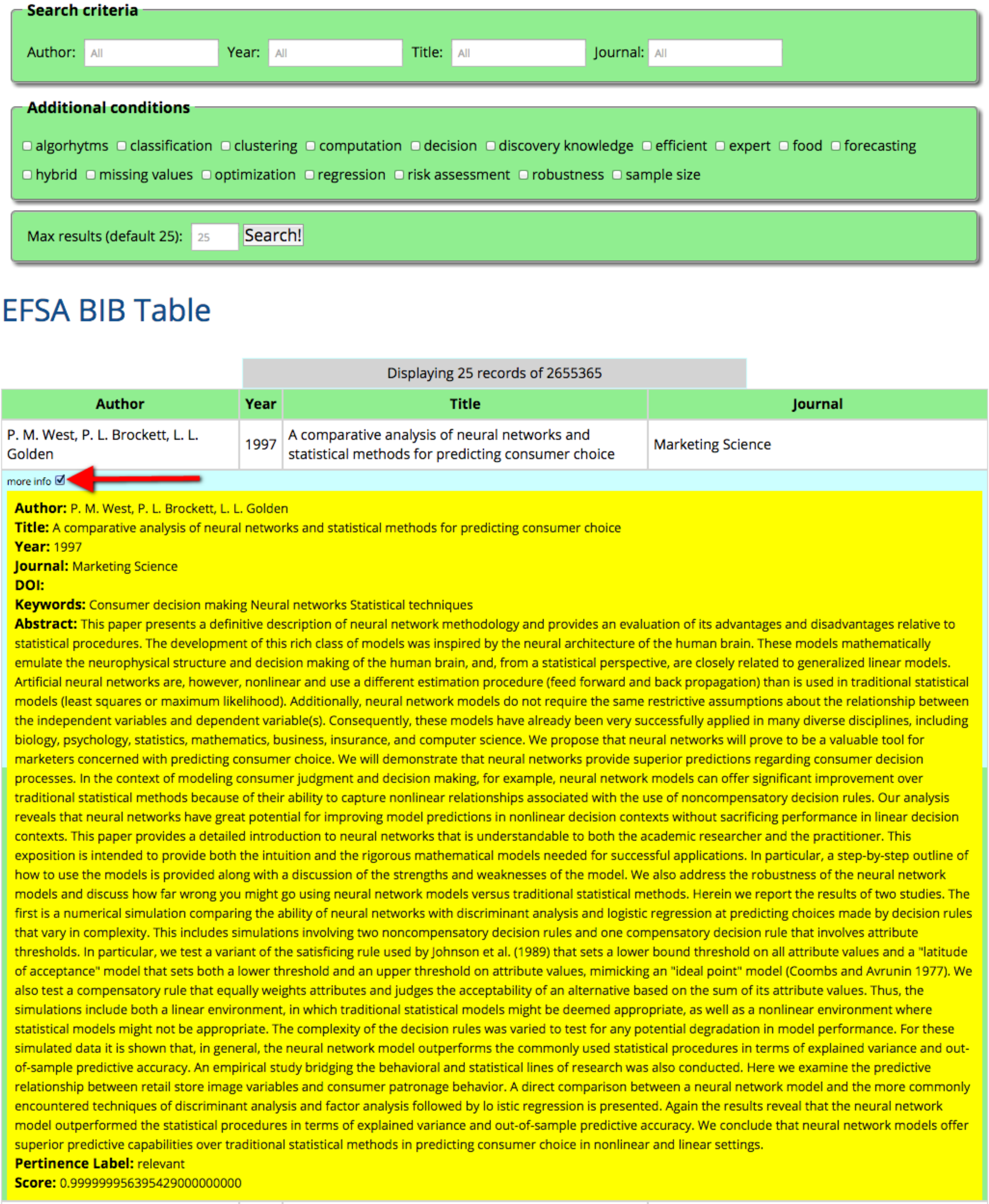
EFSA BIB Table

Displaying 25 records of 100

Author	Year	Title	Journal
	1973	In Memoriam [Jaime R. Carbonell]	Systems, Man and Cybernetics, IEEE Transactions on
	1962	Technical literature	Electrical Engineering
	1981	Atlanta to host IEEE Cybernetics conference	Acoustics, Speech, and Signal Processing Newsletter, IEEE
	1894	Local Meetings at Chicago	American Institute of Electrical Engineers, Transactions of the
	1934	Significant Electrical Developments	American Institute of Electrical Engineers, Transactions of the

When the table is loaded for the first time, is by default, sorted decreasingly by the value "Score".

Each table row of the table represent a record. Right now the variables shown are "Author", "Year", "Title" and "Journal". To see more info about a single row, we have to check "more info" box under a desired row. While doing this a yellow box will appear with more information regarding the selected record. "DOI", "Keywords", "Abstract", "Pertinence label" and "Score" are the additional informations we get by selecting "more info" box. In some records we will also see some additional conditions about the record in the box with more information.



Search criteria

Author: All Year: All Title: All Journal: All

Additional conditions

□ algorhytms □ classification □ clustering □ computation □ decision □ discovery knowledge □ efficient □ expert □ food □ forecasting
□ hybrid □ missing values □ optimization □ regression □ risk assessment □ robustness □ sample size

Max results (default 25): 25 Search!

EFSA BIB Table

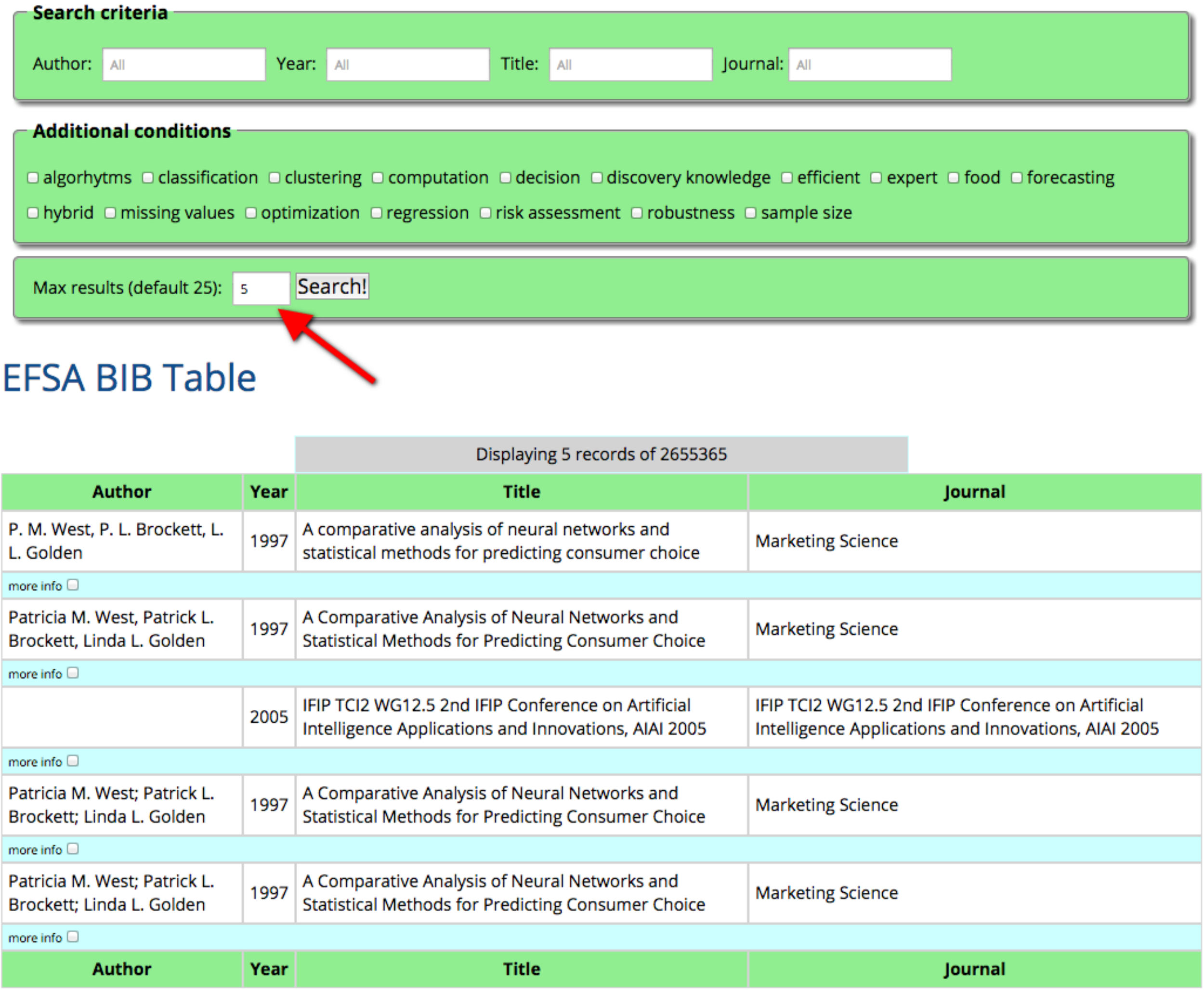
Displaying 25 records of 2655365

Author	Year	Title	Journal
P. M. West, P. L. Brockett, L. L. Golden	1997	A comparative analysis of neural networks and statistical methods for predicting consumer choice	Marketing Science

more info ☑

Author: P. M. West, P. L. Brockett, L. L. Golden
Title: A comparative analysis of neural networks and statistical methods for predicting consumer choice
Year: 1997
Journal: Marketing Science
DOI:
Keywords: Consumer decision making Neural networks Statistical techniques
Abstract: This paper presents a definitive description of neural network methodology and provides an evaluation of its advantages and disadvantages relative to statistical procedures. The development of this rich class of models was inspired by the neural architecture of the human brain. These models mathematically emulate the neurophysical structure and decision making of the human brain, and, from a statistical perspective, are closely related to generalized linear models. Artificial neural networks are, however, nonlinear and use a different estimation procedure (feed forward and back propagation) than is used in traditional statistical models (least squares or maximum likelihood). Additionally, neural network models do not require the same restrictive assumptions about the relationship between the independent variables and dependent variable(s). Consequently, these models have already been very successfully applied in many diverse disciplines, including biology, psychology, statistics, mathematics, business, insurance, and computer science. We propose that neural networks will prove to be a valuable tool for marketers concerned with predicting consumer choice. We will demonstrate that neural networks provide superior predictions regarding consumer decision processes. In the context of modeling consumer judgment and decision making, for example, neural network models can offer significant improvement over traditional statistical methods because of their ability to capture nonlinear relationships associated with the use of noncompensatory decision rules. Our analysis reveals that neural networks have great potential for improving model predictions in nonlinear decision contexts without sacrificing performance in linear decision contexts. This paper provides a detailed introduction to neural networks that is understandable to both the academic researcher and the practitioner. This exposition is intended to provide both the intuition and the rigorous mathematical models needed for successful applications. In particular, a step-by-step outline of how to use the models is provided along with a discussion of the strengths and weaknesses of the model. We also address the robustness of the neural network models and discuss how far wrong you might go using neural network models versus traditional statistical methods. Herein we report the results of two studies. The first is a numerical simulation comparing the ability of neural networks with discriminant analysis and logistic regression at predicting choices made by decision rules that vary in complexity. This includes simulations involving two noncompensatory decision rules and one compensatory decision rule that involves attribute thresholds. In particular, we test a variant of the satisficing rule used by Johnson et al. (1989) that sets a lower bound threshold on all attribute values and a "latitude of acceptance" model that sets both a lower threshold and an upper threshold on attribute values, mimicking an "ideal point" model (Coombs and Avrunin 1977). We also test a compensatory rule that equally weights attributes and judges the acceptability of an alternative based on the sum of its attribute values. Thus, the simulations include both a linear environment, in which traditional statistical models might be deemed appropriate, as well as a nonlinear environment where statistical models might not be appropriate. The complexity of the decision rules was varied to test for any potential degradation in model performance. For these simulated data it is shown that, in general, the neural network model outperforms the commonly used statistical procedures in terms of explained variance and out-of-sample predictive accuracy. An empirical study bridging the behavioral and statistical lines of research was also conducted. Here we examine the predictive relationship between retail store image variables and consumer patronage behavior. A direct comparison between a neural network model and the more commonly encountered techniques of discriminant analysis and factor analysis followed by lo istic regression is presented. Again the results reveal that the neural network model outperformed the statistical procedures in terms of explained variance and out-of-sample predictive accuracy. We conclude that neural network models offer superior predictive capabilities over traditional statistical methods in predicting consumer choice in nonlinear and linear settings.
Pertinence Label: relevant
Score: 0.999999995639542900000000

In the field "Max results (default 25)" we insert the number of results we want to visualize. After inserting the desired number we have to click the button "Search!" and wait for approximately 2 minutes for the table to refresh.



Search criteria

Author: All Year: All Title: All Journal: All

Additional conditions

□ algorhytms □ classification □ clustering □ computation □ decision □ discovery knowledge □ efficient □ expert □ food □ forecasting

□ hybrid □ missing values □ optimization □ regression □ risk assessment □ robustness □ sample size

Max results (default 25): 5

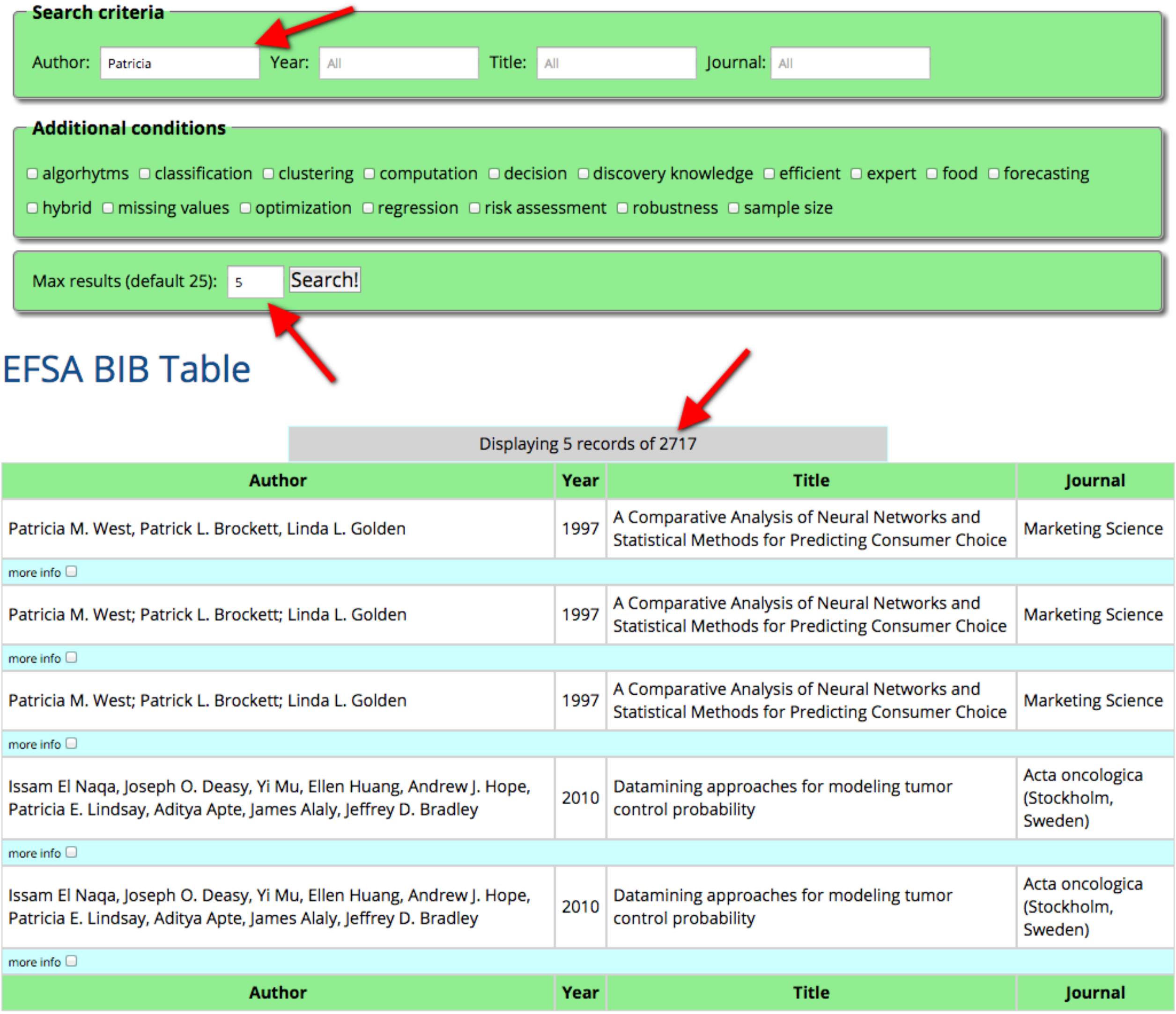
EFSA BIB Table

Displaying 5 records of 2655365

Author	Year	Title	Journal
P. M. West, P. L. Brockett, L. L. Golden	1997	A comparative analysis of neural networks and statistical methods for predicting consumer choice	Marketing Science
more info □			
Patricia M. West, Patrick L. Brockett, Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info □			
	2005	IFIP TCI2 WG12.5 2nd IFIP Conference on Artificial Intelligence Applications and Innovations, AIAI 2005	IFIP TCI2 WG12.5 2nd IFIP Conference on Artificial Intelligence Applications and Innovations, AIAI 2005
more info □			
Patricia M. West; Patrick L. Brockett; Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info □			
Patricia M. West; Patrick L. Brockett; Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info □			
Author	Year	Title	Journal

C.2.4 Individual column search

In the top of the table we have located the global search input "Search criteria" field. By inserting a search value for the desired column and clicking the "Search!" button, the table will refresh and show just the records containing the searched value in the selected column. In our case we searched for the word "Patricia" in the column "Author".



Search criteria

Author: Patricia Year: All Title: All Journal: All

Additional conditions

☐ algorhytms ☐ classification ☐ clustering ☐ computation ☐ decision ☐ discovery knowledge ☐ efficient ☐ expert ☐ food ☐ forecasting

☐ hybrid ☐ missing values ☐ optimization ☐ regression ☐ risk assessment ☐ robustness ☐ sample size

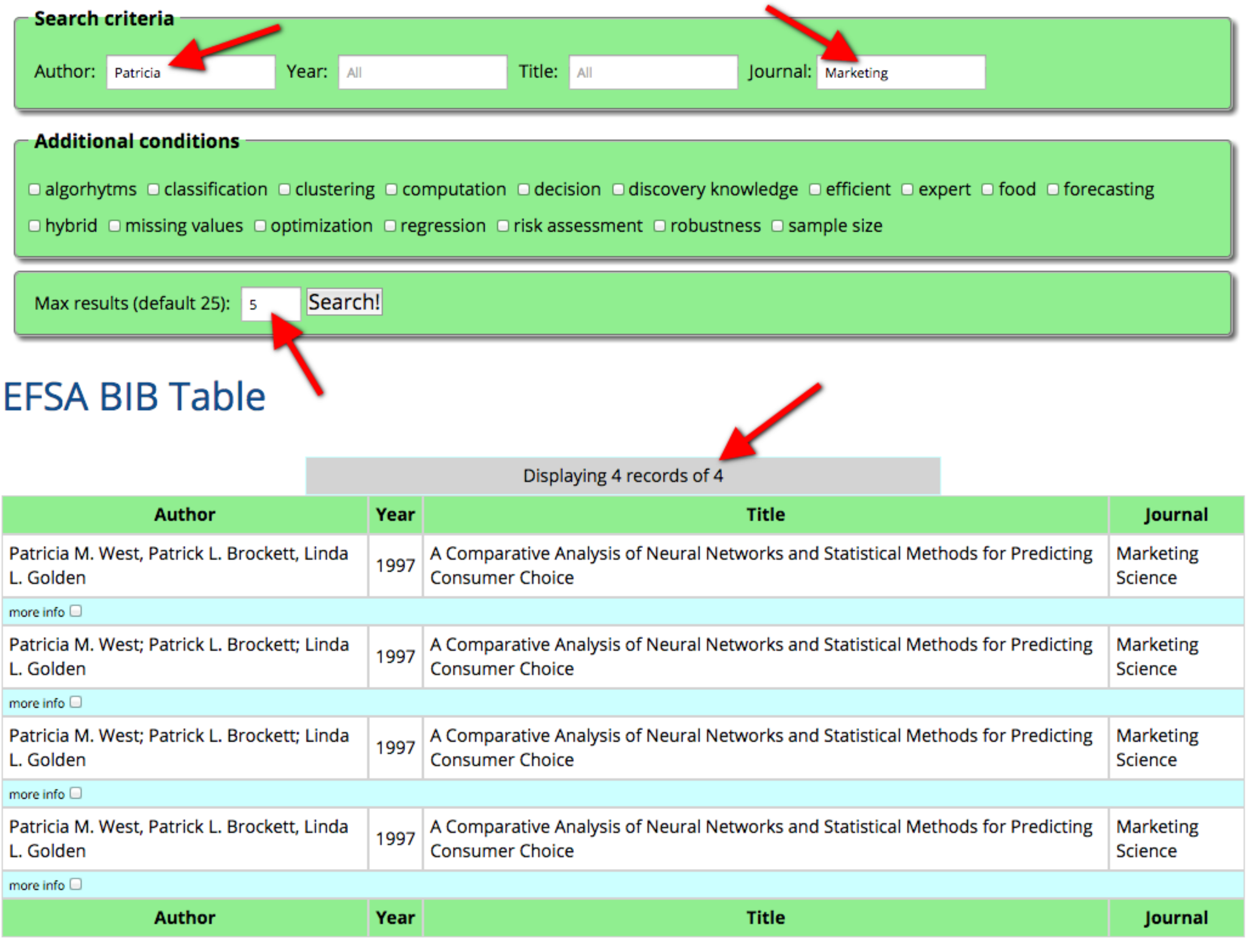
Max results (default 25): 5 Search!

EFSA BIB Table

Displaying 5 records of 2717

Author	Year	Title	Journal
Patricia M. West, Patrick L. Brockett, Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info ☐			
Patricia M. West; Patrick L. Brockett; Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info ☐			
Patricia M. West; Patrick L. Brockett; Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info ☐			
Issam El Naqa, Joseph O. Deasy, Yi Mu, Ellen Huang, Andrew J. Hope, Patricia E. Lindsay, Aditya Apte, James Alaly, Jeffrey D. Bradley	2010	Datamining approaches for modeling tumor control probability	Acta oncologica (Stockholm, Sweden)
more info ☐			
Issam El Naqa, Joseph O. Deasy, Yi Mu, Ellen Huang, Andrew J. Hope, Patricia E. Lindsay, Aditya Apte, James Alaly, Jeffrey D. Bradley	2010	Datamining approaches for modeling tumor control probability	Acta oncologica (Stockholm, Sweden)
more info ☐			
Author	Year	Title	Journal

In this case we know that there are 2717 records containing the word "Patricia" in the column "Author". We choose to show only 5 records as a result. Again, the rows are sorted decreasingly by the value "Score" which is the only and the default rule to show the results of our searches. In the example below we searched for a word "Patricia" in the column "Author" and for a word "Marketing" in the column "Journal". As a result we got 4 records.



Search criteria

Author: Patricia Year: All Title: All Journal: Marketing

Additional conditions

☐ algorhytms ☐ classification ☐ clustering ☐ computation ☐ decision ☐ discovery knowledge ☐ efficient ☐ expert ☐ food ☐ forecasting

☐ hybrid ☐ missing values ☐ optimization ☐ regression ☐ risk assessment ☐ robustness ☐ sample size

Max results (default 25): 5

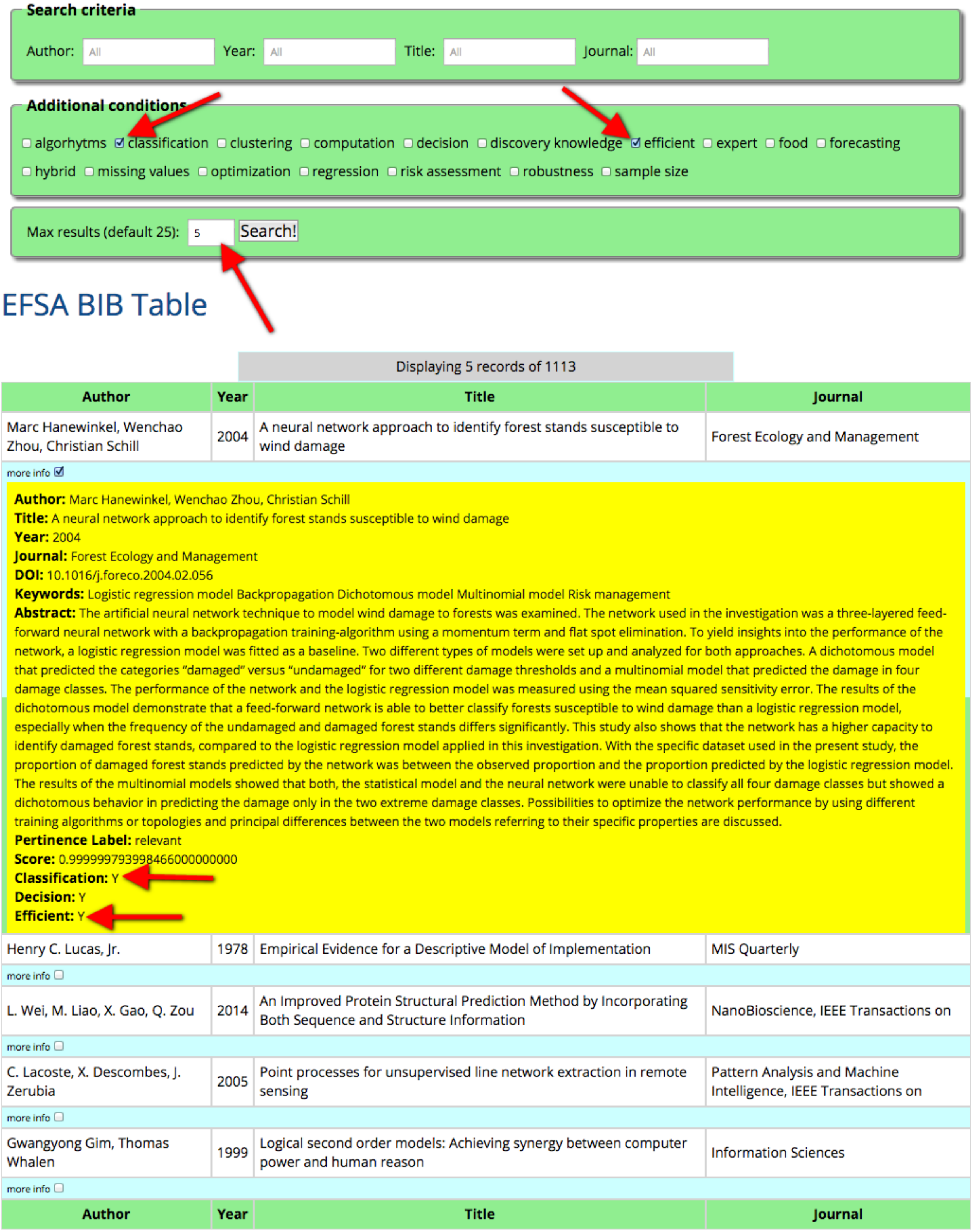
EFSA BIB Table

Displaying 4 records of 4

Author	Year	Title	Journal
Patricia M. West, Patrick L. Brockett, Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info ☐			
Patricia M. West; Patrick L. Brockett; Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info ☐			
Patricia M. West; Patrick L. Brockett; Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info ☐			
Patricia M. West, Patrick L. Brockett, Linda L. Golden	1997	A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice	Marketing Science
more info ☐			
Author	Year	Title	Journal

C.2.5 Additional conditions

In the “Search criteria” field we have also the “Additional conditions” filter box. By selecting an additional condition we can filter our results even more. In the example bellow we searched for records that have as an additional condition set “classification” and “efficent” as TRUE. In the addiational information box we can see at the bottom “Classification: Y”, “Decision: Y” and “Efficent: Y” which means that additional conditions “classification”, “decision” and “efficent” are set TRUE for the mentioned conditions.



Search criteria

Author: All Year: All Title: All Journal: All

Additional conditions

□ algorhytms ☑ classification □ clustering □ computation □ decision □ discovery knowledge ☑ efficient □ expert □ food □ forecasting □ hybrid □ missing values □ optimization □ regression □ risk assessment □ robustness □ sample size

Max results (default 25): 5

EFSA BIB Table

Displaying 5 records of 1113

Author	Year	Title	Journal
Marc Hanewinkel, Wenchao Zhou, Christian Schill	2004	A neural network approach to identify forest stands susceptible to wind damage	Forest Ecology and Management

more info ☑

Author: Marc Hanewinkel, Wenchao Zhou, Christian Schill
Title: A neural network approach to identify forest stands susceptible to wind damage
Year: 2004
Journal: Forest Ecology and Management
DOI: 10.1016/j.foreco.2004.02.056
Keywords: Logistic regression model Backpropagation Dichotomous model Multinomial model Risk management
Abstract: The artificial neural network technique to model wind damage to forests was examined. The network used in the investigation was a three-layered feed-forward neural network with a backpropagation training-algorithm using a momentum term and flat spot elimination. To yield insights into the performance of the network, a logistic regression model was fitted as a baseline. Two different types of models were set up and analyzed for both approaches. A dichotomous model that predicted the categories "damaged" versus "undamaged" for two different damage thresholds and a multinomial model that predicted the damage in four damage classes. The performance of the network and the logistic regression model was measured using the mean squared sensitivity error. The results of the dichotomous model demonstrate that a feed-forward network is able to better classify forests susceptible to wind damage than a logistic regression model, especially when the frequency of the undamaged and damaged forest stands differs significantly. This study also shows that the network has a higher capacity to identify damaged forest stands, compared to the logistic regression model applied in this investigation. With the specific dataset used in the present study, the proportion of damaged forest stands predicted by the network was between the observed proportion and the proportion predicted by the logistic regression model. The results of the multinomial models showed that both, the statistical model and the neural network were unable to classify all four damage classes but showed a dichotomous behavior in predicting the damage only in the two extreme damage classes. Possibilities to optimize the network performance by using different training algorithms or topologies and principal differences between the two models referring to their specific properties are discussed.
Pertinence Label: relevant
Score: 0.99999979399846600000000
Classification: Y
Decision: Y
Efficient: Y

Author	Year	Title	Journal
Henry C. Lucas, Jr.	1978	Empirical Evidence for a Descriptive Model of Implementation	MIS Quarterly
more info □			
L. Wei, M. Liao, X. Gao, Q. Zou	2014	An Improved Protein Structural Prediction Method by Incorporating Both Sequence and Structure Information	NanoBioscience, IEEE Transactions on
more info □			
C. Lacoste, X. Descombes, J. Zerubia	2005	Point processes for unsupervised line network extraction in remote sensing	Pattern Analysis and Machine Intelligence, IEEE Transactions on
more info □			
Gwangyong Gim, Thomas Whalen	1999	Logical second order models: Achieving synergy between computer power and human reason	Information Sciences
more info □			
Author	**Year**	**Title**	**Journal**

C.2.6 Table dictionary

Author: names of the authors;

Year: publication year;

Title: title;

Journal: name of the journal in which the work was released;

DOI: Digital Object Identifier;

Keyword: original keywords;

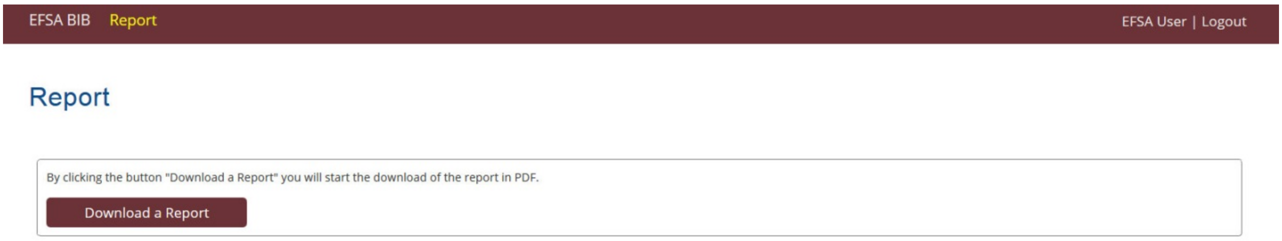
Abstract: Abstract;

Pertinence label: results of pertinence classification;

Score: score of pertinence.

C.2.7 Downloading the report

By clicking the option Report in the top menu, the user will be redirected to the page where he can download the desired version of report in PDF format.

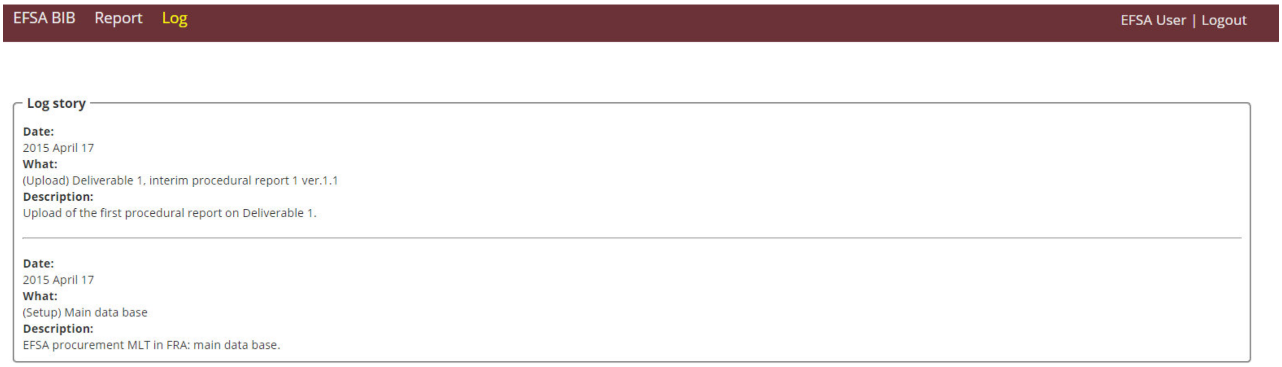


Report

By clicking the button "Download a Report" you will start the download of the report in PDF.

C.2.8 Log page

The log page is designed to give the user a brief presentation of the history of changes.



Log story

Date:
2015 April 17
What:
(Upload) Deliverable 1, interim procedural report 1 ver.1.1
Description:
Upload of the first procedural report on Deliverable 1.

Date:
2015 April 17
What:
(Setup) Main data base
Description:
EFSA procurement MLT in FRA: main data base.

C.2.9 Logout

By clicking the the caption "Logout" in the right top of the page, the user EFSA will log out from the system.

D Attachments

In following the list of the attachments for the present report. The files were uploaded in the Document Management System (DMS) EFSA.

D.1 Listings

D.1.1 Retrieval

ACM R-script to retrieve citation from ACM;

ArXiv R-script to retrieve citation from arχiv;

CIS R-script to retrieve citation from Current Index of Statistics (CIS);

CiteSeerX R-script to retrieve citation from CiteSeerX;

corefunct R-script core for CiteSeerX script;

DOAJ R-script to retrieve citation from DOAJ;

Ingenta R-script to retrieve citation from Ingenta Connect;

RePEc R-script to retrieve citation from REPEC;

glueAll R-script to create the export style format from `.RData` provided.

D.1.2 DB MySQL

20150513_EFSA_ImportBibliography R-script to manage the import BCs into MYSQL;

importbib MYSQL — import BCs into RT;

verifyInternalDuplicate MYSQL MYSQL — INDUP check into RT;

importToOverall MYSQL — import glsplbc from RT into OT;

verifyOverallDuplicate MYSQL — check for duplicate into OT;

D.1.3 Cleaning

clearFields MYSQL — cleaning data into DB;

Abstracts_Text_Cleaning R-script to pre-processing abstracts (selection of no-missing and english abstracts with more than 700 chars and text cleaning procedure).

D.1.4 SVM

ImportPertinence MYSQL — import score data;

SVM R-script providing AL and SVM classifier implementation.

D.1.5 NO

20150513_EFSA_ImportClassification R-script to manage the import classifications;

Importclass MYSQL — import the CSV of classification label into MYSQL;

importToClassification MYSQL — import labels in a single table;

NCO R-script for document annotation based on name co-occurrence analysis.

D.2 Documents

CitForVal.pdf Copy of the original pages taken from the book used for validation of BC;

E Data Dictionary MySQL DB

Below is the *Data Dictionary* of the fields into the DB provided and connected to the WEBi.

ID (int) Primary Key;

ref1 (int) EndNote field;

Reference_Type (varchar) EndNote field;

Author (text) EndNote field;

Year (int) EndNote field;

Title (text) EndNote field;

Secondary_Author (text) EndNote field;

Secondary_title (text) EndNote field;

Place_published (text) EndNote field;

Publisher (text) EndNote field;

Volume (varchar) EndNote field;

Number_of_Volumes (varchar) EndNote field;

Number (varchar) EndNote field;

Pages (varchar) EndNote field;

Section (varchar) EndNote field;

Tertiary_Author (text) EndNote field;

Tertiary_Title (text) EndNote field;

Edition (text) EndNote field;

Date (varchar) EndNote field;

Type_of_Work (text) EndNote field;

Subsidiary_Author (text) EndNote field;

Short_Title (varchar) EndNote field;

Alternative_Title (text) EndNote field;

ISBN_ISSN (varchar) EndNote field;

DOI (varchar) EndNote field;

Original_Publication (text) EndNote field;

Reprint_Edition (text) EndNote field;

Reviewed_Item (text) EndNote field;

Custom_1 (text) EndNote field;

Custom_2 (text) EndNote field;

Custom_3 (text) EndNote field;

Custom_4 (text) EndNote field;

Custom_5 (text) EndNote field;

Custom_6 (text) EndNote field;

Custom_7 (text) EndNote field;

Custom_8 (text) EndNote field;

Accession_Number (text) EndNote field;

Call_Number (varchar) EndNote field;

Label (text) EndNote field;

Keywords (text) EndNote field;

Abstract (text) EndNote field;

Notes (text) EndNote field;

Research_Notes (text) EndNote field;

URL (text) EndNote field;

File_Attachments (text) EndNote field;

Author_Address (text) EndNote field;

Figure (text) EndNote field;

Caption (text) EndNote field;

Access_Date (text) EndNote field;

Translated_Author (text) EndNote field;

Translated_Title (text) EndNote field;

Name_of_DataBase (text) EndNote field;

Database_Provider (text) EndNote field;

Language (text) EndNote field;

exported (varchar) Utility (record exported to R);

ieeexplore (varchar) Record present in resource (string 1);

jstor (varchar) Record present in resource (string 1);

sciencedirect (varchar) Record present in resource (string 1);

acm (varchar) Record present in resource (string 1);

arxiv (varchar) Record present in resource (string 1);

citeseerx (varchar) Record present in resource (string 1);

doaj (varchar) Record present in resource (string 1);

psycinfo (varchar) Record present in resource (string 1);

wos_ssci (varchar) Record present in resource (string 1);

pubmed (varchar) Record present in resource (string 1);

medline (varchar) Record present in resource (string 1);

wos_sci (varchar) Record present in resource (string 1);

wos_core (varchar) Record present in resource (string 1);

scopus (varchar) Record present in resource (string 1);

repec (varchar) Record present in resource (string 1);

mathscinet (varchar) Record present in resource (string 1);

wos_ahci (varchar) Record present in resource (string 1);

cochrane (varchar) Record present in resource (string 1);

cis (varchar) Record present in resource (string 1);

cinahl (varchar) Record present in resource (string 1);

econlit (varchar) Record present in resource (string 1);

ingentaconnect (varchar) Record present in resource (string 1);

woscore_import_str02 (varchar) Record present in resource (string 2);

wossci_import_str02 (varchar) Record present in resource (string 2);

ingenta_import_str02 (varchar) Record present in resource (string 2);

str01 (varchar) Record present in string 1;

str02 (varchar) Record present in string 2;

arXiv_import_str02 (varchar) Record present in resource (string 2);

CIS_import_str02 (varchar) Record present in resource (string 2);

DOAJ_import_str02 (varchar) Record present in resource (string 2);

MedLine_import_str02 (varchar) Record present in resource (string 2);

PsycInfo_import_str02 (varchar) Record present in resource (string 2);

PubMed_import_str02 (varchar) Record present in resource (string 2);

RePEc_import_str02 (varchar) Record present in resource (string 2);

ACM_import_str02 (varchar) Record present in resource (string 2);

Econlit_import_str02 (varchar) Record present in resource (string 2);

CiteSeerX_import_str02 (varchar) Record present in resource (string 2);

arXiv_OV (varchar) Record present in resource (string 1/2);

CIS_OV (varchar) Record present in resource (string 1/2);

CitseerX_OV (varchar) Record present in resource (string 1/2);

DOAJ_OV (varchar) Record present in resource (string 1/2);

EconLit_OV (varchar) Record present in resource (string 1/2);

Ingenta_OV (varchar) Record present in resource (string 1/2);

MedLine_OV (varchar) Record present in resource (string 1/2);

PsycInfo_OV (varchar) Record present in resource (string 1/2);

PubMed_OV (varchar) Record present in resource (string 1/2);

RePEC_OV (varchar) Record present in resource (string 1/2);

WoSCORE_OV (varchar) Record present in resource (string 1/2);

WoSSCI_OV (varchar) Record present in resource (string 1/2);

ACM_OV (varchar) Record present in resource (string 1/2);

n_resources (int) Record presence (number of resources);

pert_label (int) Pertinence (1: not pertinent; 2: pertinent);

Score (decimal) Pertinence score;

algorithms (varchar) Pertinence label;

classification (varchar) Pertinence label;

clustering (varchar) Pertinence label;

computation (varchar) Pertinence label;

decision (varchar) Pertinence label;

discovery_knw (varchar) Pertinence label;

efficient (varchar) Pertinence label;

expert (varchar) Pertinence label;

food (varchar) Pertinence label;

forecasting (varchar) Pertinence label;

hybrid (varchar) Pertinence label;

missing_vl (varchar) Pertinence label;

optimization (varchar) Pertinence label;

regression (varchar) Pertinence label;

risk_assmnt (varchar) Pertinence label;

robustness (varchar) Pertinence label;

sample_size (varchar) Pertinence label;

labelled (varchar) Record labelled.

F Opinion classification

Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
topic_1	-	NDA	3408	Scientific Opinion	Provide advice on nutritional requirements, their role and composition	5	0		LDA
topic_1	-	NDA	3760	Scientific Opinion	Provide advice on nutritional requirements, their role and composition	5	0		LDA
topic_1	-	NDA	3845	Scientific Opinion	Provide advice on energy, macronutrients, fibre. Provide Guidance	3	0		LDA
topic_2	topic_12	CEF	1921	Scientific Opinion	Evaluations of flavourings and needed for further investigation	2	0		LDA
topic_2	topic_12	CEF	2178	Scientific Opinion	Evaluations of flavourings and needed for further investigation	2	0		LDA
topic_2	topic_12	CEF	4335	Scientific Opinion	Safety assessment on 3 substances	1	0		LDA
topic_3	topic_1	PLH	3857	Scientific Opinion	Pest Risk Assessment	3	0		LDA
topic_3	topic_1	PLH	3923	Scientific Opinion	Pest Risk Assessment	3	0		LDA
topic_3	topic_1	PLH	3988	Scientific Opinion	Pest Risk Assessment	3	0		LDA
topic_4	topic_6	EFSA	2325	Reasoned Opinion	Risks to the consumer by modification of MRL	1	1	descriptive stat	LDA
topic_4	topic_6	CONTAM	2985	Scientific Opinion	Risks to human health Hg in food. Exp assessment	5	1	descriptive stat	LDA
topic_4	topic_6	NDA	3761	Scientific Opinion	Risk/benefit of fish consumption related to metilHg	2	1	descriptive stat	LDA
topic_4	topic_6	EFSA SC	3982	Scientific Opinion	Second step: benefits of fish consumption	2	1	descriptive stat	LDA
topic_5	topic_7	PPR	2668	Scientific Opinion	Risk Assessment on PPP on bees	4	1	descriptive stat	LDA

Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
topic_5	topic_7	PPR	3800	Scientific Opinion	Guidance on RA for non-target terrestrial plants	2	0		LDA
topic_5	topic_7	PPR	3996	Scientific Opinion	Guidance on RA for non-target arthropod	2	0		LDA
topic_6	-	NDA	253	Scientific Opinion	Review on Omega 3 Fatty acids	1	0		LDA
topic_6	-	NDA	1461	Scientific Opinion	Advice on macronutrients, fibre and energy	3	0		LDA
topic_6	-	ANS	1512	Scientific Opinion	Safety of sucrose esters	2	0		LDA
topic_6	-	NDA	2168	Scientific Opinion	Evaluation of scientific fatty acids	1	0		LDA
topic_7	topic_14	EFSA/ECDC	3590	Scientific Report	European Union Summary Report	1	1	prevalence estimate spatial analysis	LDA
topic_7	topic_14	EFSA/ECDC	4036	Scientific Report	European Union Summary Report	1	1	prevalence estimate spatial analysis	LDA
topic_7	topic_14	EFSA/ECDC	4380	Scientific Report	European Union Summary Report	1	1	prevalence estimate spatial analysis	LDA
topic_8	-	CONTAM	1570	Scientific Opinion	Risk to human health related to the presence of lead in foodstuffs	1	1	Margin of exposure approach Uncertainty analysis	LDA
topic_8	-	CONTAM	1627	Scientific Opinion	To assess the currents EU limits of various marien biotoxins	1	1	descriptive stat	LDA

Control dataset
(91 EFSA documents)

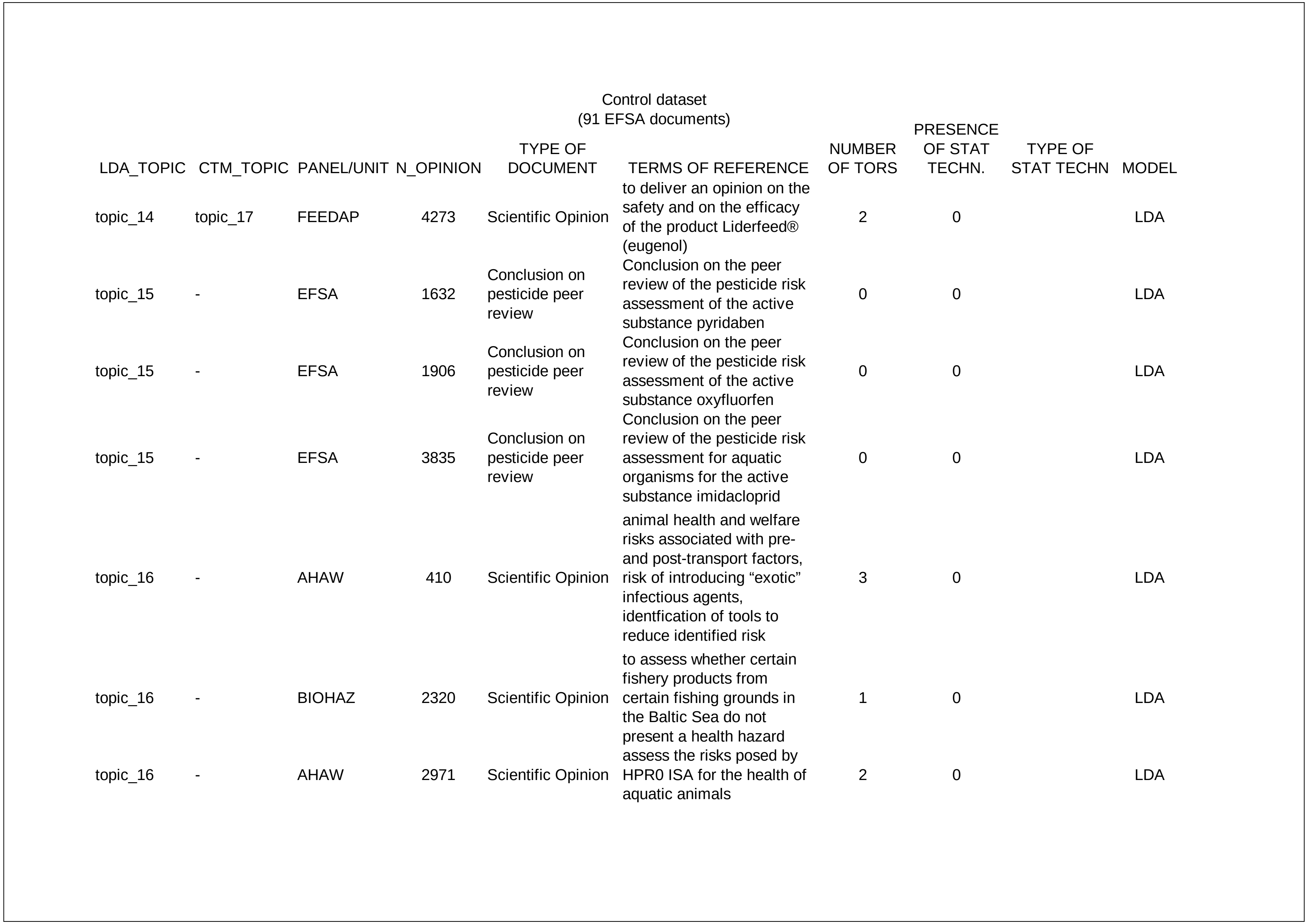
LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
topic_8	-	CONTAM	3597	Scientific Opinion	To provide a report on exposure to organic arsenic	2	1	descriptive stat	LDA
topic_9	topic_11	EFSA	3247	Reasoned Opinion	To provide reasoned opinion on the modification of the existing MRLs for indoxacarb	1	0		LDA
topic_9	topic_11	EFSA	4076	Reasoned Opinion	Combined review	1	0		LDA
topic_9	topic_11	EFSA	4381	Reasoned Opinion	Review on Omega 3 Fatty acids	1	0		LDA
topic_10	-	BIOHAZ	12	Scientific Opinion	To asses if surveillance data allow an estimate of prevalence on TSE	3	0		LDA
topic_10	-	BIOHAZ	442	Scientific Opinion	update on the risks posed by tissues of sheep to human health	1	1	prevalence estimate	LDA
topic_10	-	BIOHAZ	1875	Scientific Opinion	to update TSE infectivity distribution in ruminant tissues	2	0		LDA
topic_10	-	BIOHAZ	3781	Scientific Opinion	Trend on classical scrapie situation	5	1	Adjusted prevalence Spatio-temporal analysis linear function negative binomial model adjusted	LDA

Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
topic_11	topic_5	BIOHAZ	1503	Scientific Opinion	to analyse the results of the baseline survey on on Campylobacter spp. in broiler flocks	2	1	prevalence estimate, risk assessment	LDA
topic_11	topic_5	BIOHAZ	2351	Scientific Opinion	to Identify and rank the main risks for public health that should be addressed by meat inspection	4	1	risk assessment: risk ranking	LDA
topic_11	topic_5	BIOHAZ	2741	Scientific Opinion	to Identify and rank the main risks for public health that should be addressed by meat inspection	4	1	risk assessment: risk ranking	LDA
topic_11	topic_5	BIOHAZ	3601	Scientific Opinion	to assess if it is possible to apply alternative core temperatures for the transport of meat	4	1	predictive microbiology growth models application	LDA
topic_12	topic_4	GMO	470	Scientific Opinion	scientific assessment of the genetically modified maize 59122 for food and feed uses	1	0		LDA
topic_12	topic_4	GMO	524	Scientific Opinion	to carry out a scientific assessment of the genetically modified soybean A2704-12 for food and feed uses	1	0		LDA

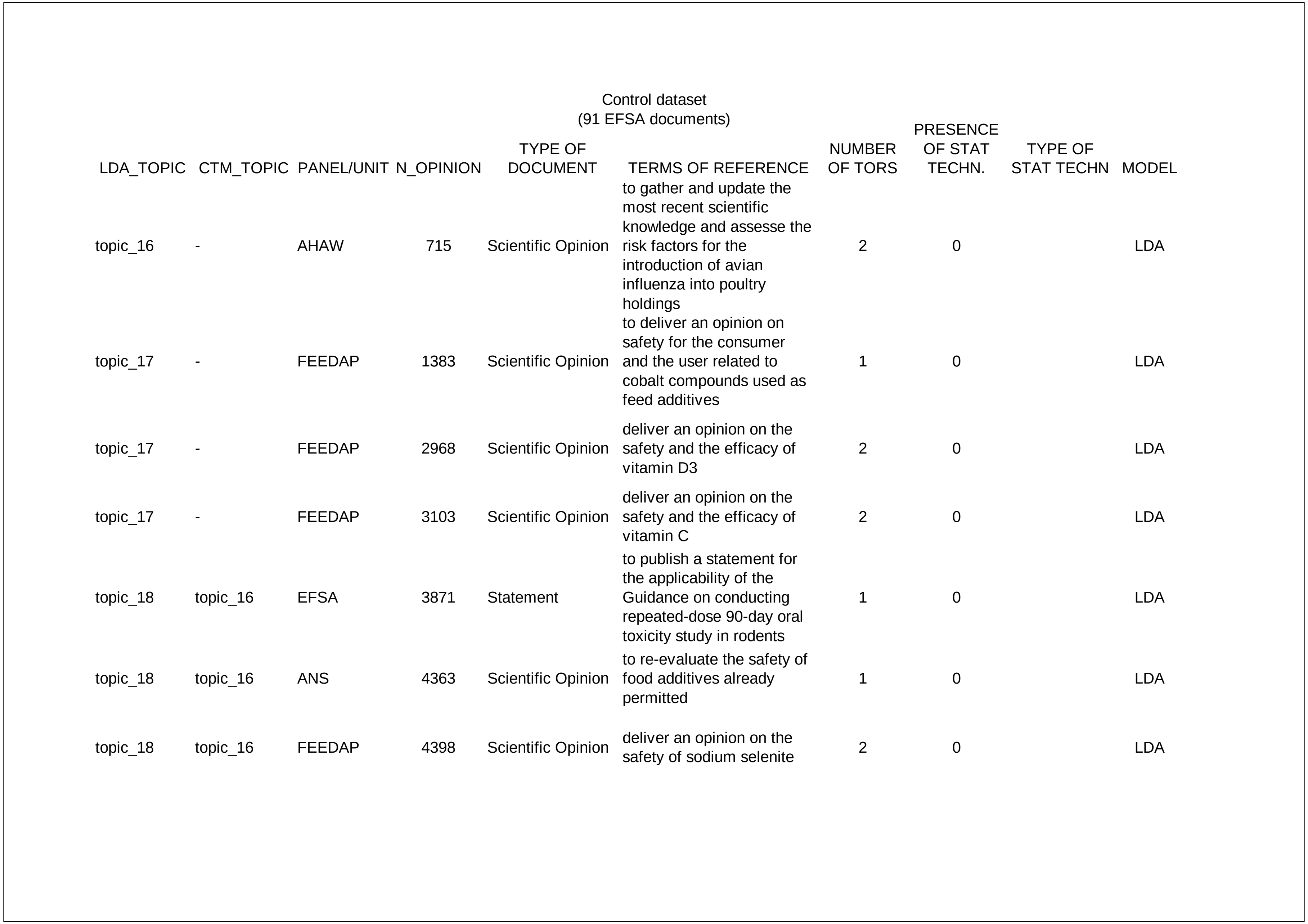
Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
topic_12	topic_4	GMO	4167	Scientific Opinion	to carry out a scientific assessment of soybean FG72 for food and feed uses	1	0		LDA
topic_13	-	AHAW	45	Scientific Opinion	to report on the welfare aspects of the main systems of stunning and killing	3	0		LDA
topic_13	-	AHAW	326	Scientific Opinion	to issue a scientific opinion on the main systems of stunning and killing	1	0		LDA
topic_13	-	AHAW	1966	Scientific Opinion	to assess the scientific information available on the welfare of animals during transport	3	0		LDA
topic_13	-	FEEDAP	4394	Scientific Opinion	to deliver an opinion on the safety and on the efficacy of the product GAA	2	0		LDA
topic_14	topic_17	FEEDAP	2670	Scientific Opinion	to deliver an opinion on the safety and the efficacy of the product AviPlus®	1	0		LDA
topic_14	topic_17	FEEDAP	2924	Scientific Opinion	to deliver an opinion on the safety and the efficacy of the product Toyocerin®	1	0		LDA
topic_14	topic_17	FEEDAP	3167	Scientific Opinion	to deliver an opinion on the safety and the efficacy of the product Bonvital (Enterococcus faecium)	1	0		LDA



Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
topic_14	topic_17	FEEDAP	4273	Scientific Opinion	to deliver an opinion on the safety and on the efficacy of the product Liderfeed® (eugenol)	2	0		LDA
topic_15	-	EFSA	1632	Conclusion on pesticide peer review	Conclusion on the peer review of the pesticide risk assessment of the active substance pyridaben	0	0		LDA
topic_15	-	EFSA	1906	Conclusion on pesticide peer review	Conclusion on the peer review of the pesticide risk assessment of the active substance oxyfluorfen	0	0		LDA
topic_15	-	EFSA	3835	Conclusion on pesticide peer review	Conclusion on the peer review of the pesticide risk assessment for aquatic organisms for the active substance imidacloprid	0	0		LDA
topic_16	-	AHAW	410	Scientific Opinion	animal health and welfare risks associated with pre- and post-transport factors, risk of introducing “exotic” infectious agents, identfication of tools to reduce identified risk	3	0		LDA
topic_16	-	BIOHAZ	2320	Scientific Opinion	to assess whether certain fishery products from certain fishing grounds in the Baltic Sea do not present a health hazard	1	0		LDA
topic_16	-	AHAW	2971	Scientific Opinion	assess the risks posed by HPR0 ISA for the health of aquatic animals	2	0		LDA



Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
topic_16	-	AHAW	715	Scientific Opinion	to gather and update the most recent scientific knowledge and assesse the risk factors for the introduction of avian influenza into poultry holdings	2	0		LDA
topic_17	-	FEEDAP	1383	Scientific Opinion	to deliver an opinion on safety for the consumer and the user related to cobalt compounds used as feed additives	1	0		LDA
topic_17	-	FEEDAP	2968	Scientific Opinion	deliver an opinion on the safety and the efficacy of vitamin D3	2	0		LDA
topic_17	-	FEEDAP	3103	Scientific Opinion	deliver an opinion on the safety and the efficacy of vitamin C	2	0		LDA
topic_18	topic_16	EFSA	3871	Statement	to publish a statement for the applicability of the Guidance on conducting repeated-dose 90-day oral toxicity study in rodents	1	0		LDA
topic_18	topic_16	ANS	4363	Scientific Opinion	to re-evaluate the safety of food additives already permitted	1	0		LDA
topic_18	topic_16	FEEDAP	4398	Scientific Opinion	deliver an opinion on the safety of sodium selenite	2	0		LDA

Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
topic_18	topic_16	AFC	414	Scientific Opinion	to issue an opinion on the safety in use of polyethylene glycol as a film coating agent for use in food supplement products	1	0		LDA
topic_19	topic_13	NDA	1252	Scientific Opinion	to provide advice on adequate information is provided on the characteristics of the food pertinent to the beneficial effect	2	0		LDA
topic_19	topic_13	NDA	2262	Scientific Opinion	to provide advice on adequate information is provided on the characteristics of the food pertinent to the beneficial effect	2	0		LDA
topic_19	topic_13	NDA	3415	Scientific Opinion	issue an opinion on the scientific substantiation of a health claim related to a combination of cabbages and maintenance of normal blood LDL-cholesterol concentration	1	0		LDA
topic_20	topic_8	EFSA	1897	Scientific Opinion	Conclusion on the peer review of the pesticide risk assessment of the active substance cyproconazole	0	0		LDA

Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
topic_20	topic_8	EFSA	2797	Scientific Opinion	Conclusion on the peer review of the pesticide risk assessment of the active substance kieselgur	0	0		LDA
topic_20	topic_8	EFSA	3166	Scientific Opinion	Conclusion on the peer review of the pesticide risk assessment of the active substance fenazaquin	0	0		LDA
topic_20	topic_8	PPR	922	Scientific Opinion	opinion regarding the relative utility of total concentration and pore water concentration as exposure metrics in the assessment of ecotoxicological risks from pesticides	1	1	risk assessment risk model design	LDA
-	topic_2	PLH	3468	Scientific Opinion	Pest Risk Assessment	3	0		CTM
-	topic_2	PLH	3850	Scientific Opinion	Pest Risk Assessment	1	0		CTM
-	topic_2	PLH	3923	Scientific Opinion	Pest Risk Assessment	1	0		CTM
-	topic_2	PLH	3988	Scientific Opinion	Pest Risk Assessment	1	0		CTM
-	topic_3	NDA	1463	Scientific Opinion	to provide advice on adequate information is provided on the characteristics of the food pertinent to the beneficial effect	3	0		CTM
-	topic_3	EFSA	2098	Reasoned Opinon	to provide a view of setting temporary MRLs	1	1	descriptive stat linear regression	CTM
-	topic_3	ANS	3467	Scientific Opinion	Provide advice on a tolerable upper intake level (UL)	1	0		CTM

Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
-	topic_3	EFSA SC	3593	Scientific Opinion	to develop a generic assessment system allowing for priority setting among the botanicals	1	0		CTM
-	topic_9	AHAW	4	Scientific Opinion	to report on the welfare of animals during transport	1	0		CTM
-	topic_9	AHAW	783	Scientific Opinion	To report on main welfare risks related to the farming of sheep	1	0		CTM
-	topic_9	AHAW	4373	Scientific Opinion	to describe E. multilocularis infection, surveillance, risk factors and laboratory testing in EU and adjacent countries	5	1	descriptive stat Bayesian approach deterministic mathematical model	CTM
-	topic_9	AHAW	584	Scientific Opinion	to give scientific advice on fish diseases	2	1	risk assessment risk model design	CTM
-	topic_10	NDA	3408	Scientific Opinion	Provide advice on the nutritional requirements of infants and young children	2	0		CTM
-	topic_10	NDA	3760	Scientific Opinion	Provide advice on nutritional requirements, their role and composition	5	0		CTM
-	topic_10	NDA	3845	Scientific Opinion	Provide advice on energy, macronutrients, fibre. Provide Guidance	3	0		CTM
-	topic_10	NDA	3957	Scientific Opinion	Provide advice on the essential compositional requirements for total diet replacements	1	0		CTM

Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
-	topic_15	NDA	1462	Scientific Opinion	to provide advice on energy, macronutrients and dietary fibre.	2	0		CTM
-	topic_15	NDA	1924	Scientific Opinion	provide scientific substantiation of health claims in relation to sugar beet fibre and reduction of post-prandial glycemia	1	0		CTM
-	topic_15	NDA	3837	Scientific Opinion	provide scientific substantiation of health claims in relation to rye bread and reduction of post-prandial glycemia	1	0		CTM
-	topic_15	NDA	4098	Scientific Opinion	provide scientific substantiation of health claims in relation to FRUIT UP and reduction of post-prandial glycemia	1	0		CTM
-	topic_18	NDA	184	Scientific Opinion	to consider the likelihood of adverse reactions triggered in susceptible individuals by the consumption (wine)	1	0		CTM
-	topic_18	NDA	534	Scientific Opinion	to consider the likelihood of adverse reactions triggered in susceptible individuals by the consumption (wine)	1	0		CTM

Control dataset
(91 EFSA documents)

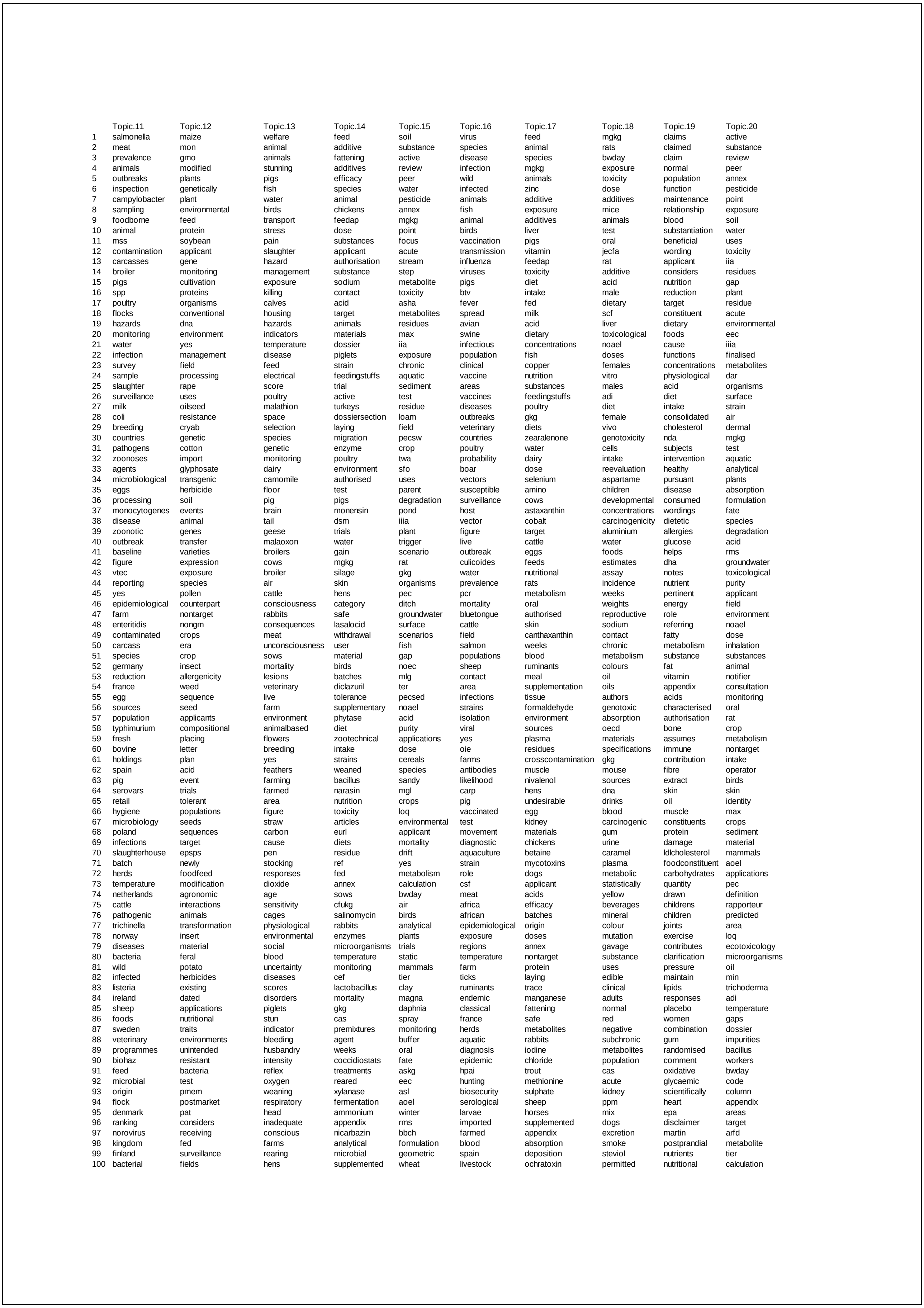
LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
-	topic_18	NDA	3894	Scientific Opinion	Recommendations for threshold concentrations of each allergen in food that would provide an acceptable level of protection for at-risk consumers;	3	1	mathematical modelling: benchmark dose approach	CTM
-	topic_18	NDA	768	Scientific Opinion	assessment for 'Ice Structuring Protein (ISP)" as food ingredient	1	0		CTM
-	topic_19	NDA	253	Scientific Opinion	scientific substantiation of nutrition claims relating to omega-3 fatty acids, mono-unsaturated fat, polyunsaturated…	1	0		CTM
-	topic_19	NDA	1461	Scientific Opinion	to advise on population reference intakes of micronutrients in the diet	3	0		CTM
-	topic_19	NDA	3408	Scientific Opinion	Provide advice on the importance of the role that growing-up milks' may have as a liquid element in the diet of young children	5	0		CTM
-	topic_20	AFC	243	Scientific Opinion	re-evaluate di-(2-ethylhexyl) phthalate (DEHP) for use in the manufacture of food contact materials.	1	0		CTM

Control dataset
(91 EFSA documents)

LDA_TOPIC	CTM_TOPIC	PANEL/UNIT	N_OPINION	TYPE OF DOCUMENT	TERMS OF REFERENCE	NUMBER OF TORS	PRESENCE OF STAT TECHN.	TYPE OF STAT TECHN	MODEL
-	topic_20	GMO	2438	Scientific Opinion	to develop principles and guidance for the establishment of protocols for 90-day feeding studies in rodents with whole food/feed.	1	1	Power analysis,descriptive statistics, hypothesis testing	CTM
-	topic_20	GMO	3871	Scientific Opinion	to provide an explanatory statement for the establishment of protocols for 90-day feeding studies in rodents	1	1	Power analysis	CTM
-	topic_20	CONTAM	3907	Scientific Opinion	the risks to human and animal health related to the presence of chloramphenicol in food and feed	5	1	Descriptive statistics, risk assessment	CTM

G LDA Classification

	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
1	intake	substances	pest	fish	exposure	acid	resistance	exposure	residue	bse
2	calcium	flavouring	plant	exposure	species	acids	isolates	gkg	mgkg	sheep
3	dietary	flno	plants	mercury	field	fatty	antimicrobial	dietary	residues	animals
4	vitamin	jecfa	area	mgkg	approach	lactobacillus	animals	foods	mrl	cattle
5	iron	class	citrus	concentrations	plant	oil	salmonella	toxins	mrls	test
6	foods	intake	fruit	pcb	soil	esters	mss	sample	existing	bovine
7	children	candidate	spread	dietary	pesticides	methyl	res	arsenic	commodities	tse
8	chromium	substance	host	fat	ppr	strain	bacteria	children	trials	scrapie
9	infants	test	species	environmental	test	fats	coli	shellfish	review	animal
10	acid	negative	virus	liver	bees	intake	reporting	survey	others	goats
11	protein	mgkg	disease	intake	toxicity	bifidobacterium	gallus	irradiation	crops	infectivity
12	nutrition	msdi	probability	thyroid	crop	oils	qps	dose	median	age
13	intakes	threshold	pet	meat	uncertainty	cla	resistant	population	adi	sensitivity
14	supplements	structural	establishment	milk	tier	ethyl	countries	meat	outdoor	surveillance
15	milk	approach	areas	bde	models	methacrylate	chloramphenicol	cadmium	exposure	infected
16	allergy	toxicity	citri	bpa	water	fat	mrsa	toxin	gap	atypical
17	clinical	acid	hosts	dioxins	appendix	saturated	antimicrobials	water	lettuce	tissue
18	zinc	fge	options	rats	birds	cncm	species	countries	wheat	pool
19	nutritional	genotoxicity	planting	methylmercury	scenarios	plantarum	acid	alkaloids	appendix	prevalence
20	fluoride	effa	material	age	figure	microorganisms	spp	bound	enforcement	ruminants
21	mgday	rat	pra	congeners	substances	intestinal	pigs	analytical	definition	disease
22	selenium	assay	management	dose	scenario	sulphide	indicator	sampling	grapes	infection
23	age	dose	organisms	pcbs	environmental	strains	ciprofloxacin	monitoring	pesticide	population
24	adults	register	uncertainty	children	pesticide	amino	strains	concentrations	leaves	deer
25	water	typhimurium	pathogen	ngg	chemicals	mixture	netherlands	lod	apples	goat
26	lycopene	category	spp	feed	crops	disulfide	meat	vegetables	milk	material
27	diet	revision	recycling	gkg	step	thiols	genes	contam	seed	classical
28	sources	rats	countries	seafood	mammals	linoleic	susceptibility	grains	tomatoes	transmission
29	scf	scf	annex	blood	acute	gastrointestinal	denmark	figure	commodity	negative
30	women	oral	infected	toxicity	ground	branched	spain	estimates	arfd	tuberculosis
31	gday	additives	french	contam	residues	combination	wheat	uncertainty	fruit	monitoring
32	lutein	specification	eppo	serum	modelling	ester	nalidixic	nitrate	beans	specificity
33	allergic	cells	fastidiosa	maternal	spray	trans	cattle	age	plant	milk
34	proteins	gml	phytosanitary	pbdes	population	rhamnosus	zoonotic	fish	loq	sample
35	magnesium	ames	tomato	countries	plants	lactis	animal	element	animal	exposure
36	reactions	gplate	viruses	water	bcs	collection	ampicillin	feed	vegetables	agent
37	yeast	noael	tabaci	perchlorate	concentrations	unpublished	germany	intake	diet	healthy
38	absorption	industry	categorisation	environment	surface	sucrose	austria	infants	intake	spongiform
39	patients	flavourings	symptoms	median	bbch	species	teliospores	collection	potatoes	pos
40	energy	fgerev	france	species	endpoints	sulphides	campylobacter	acute	modification	slaughtered
41	nutrient	subgroup	temperature	poppy	mortality	longum	cefotaxime	bread	active	protein
42	allergens	vitro	infection	animals	area	glycerol	clinical	cereal	foliar	countries
43	aged	metabolic	recycled	population	bee	oleic	infections	upper	acute	prion
44	bioavailability	mutation	leaf	contamination	cumulative	probiotic	mic	categories	fresh	ruminant
45	population	unpublished	step	contaminants	melamine	considers	monitoring	loq	liver	tses
46	concentrations	vivo	consequences	infants	chapter	propyl	bacillus	vegetable	origin	clinical
47	sodium	aliphatic	reduction	cancer	herbicide	lmg	antibiotic	adults	meat	prp
48	potassium	metabolism	effectiveness	morphine	seeds	diallyl	gentamicin	wheat	substance	annex
49	serum	mtamdi	crops	women	applications	straightchain	france	code	metabolism	byproducts
50	infant	insoluble	harmful	sum	substance	conjugated	tetracyclines	category	eec	prpsc
51	plasma	mouse	vectors	animal	oecd	material	bacterial	fruit	fat	category
52	young	esters	xanthomonas	male	dose	sulfide	tetracycline	cereals	sum	birth
53	supplementation	alcohols	trees	ngkg	focus	culture	trends	coffee	iesti	diagnostic
54	individuals	intakes	materials	dioxinlike	diet	alcohols	figure	median	peer	slaughter
55	folate	flavour	environmental	tissue	hazard	dsm	areas	contaminants	toxicological	hazards
56	formulae	cho	plh	mice	ecosystem	butyl	probability	elicitation	neu	estimates
57	subjects	soluble	banana	diet	seed	allyl	escherichia	chronic	cxl	species
58	iodine	structurally	ispm	oil	selection	animalis	update	marine	code	mice
59	applicant	flavis	flakes	pfos	ter	dimethyl	typhimurium	beauvericin	pods	protocol
60	upper	gavage	quarantine	ambrosia	active	paracasei	erythromycin	elderly	sugar	feed
61	petitioner	alcohol	crop	adults	consultation	isomers	broilers	nitrite	analytical	meat
62	manganese	methyl	organism	brominated	pollen	characterised	appendix	sum	processing	brain
63	allergen	derivatives	fruits	substances	endpoint	bacterial	dispersed	acid	spraying	herd
64	fish	bwday	decontamination	polychlorinated	existing	bacteria	agents	enniatins	proposal	encephalopathy
65	picolinate	concentrations	contact	pbde	mammal	disulphide	foodproducing	foodex	cabbage	srm
66	page	metabolites	reduce	bound	focal	hydrolysis	jejuni	reporting	wine	infectious
67	amino	activation	absent	hbcdd	spraying	subsp	faecium	approach	calculation	crl
68	disease	metabolised	citricarpa	pfoa	aquatic	substances	microbiol	processing	tentative	resistance
69	blood	acetate	strawberry	hbcdds	page	acidophilus	amr	toddlers	aiha	blood
70	allergenic	step	leaves	pbbs	sensitivity	cas	streptomycin	doses	kidney	ssc
71	deficiency	annex	protected	page	default	trisulfide	cutoff	toxicity	root	veterinary
72	men	faowho	populations	dlpcbs	reproductive	casei	ireland	milk	dietary	agents
73	doses	ethyl	soil	analytical	text	pathogenic	sweden	scenario	crop	active
74	ige	mice	potato	tbbpa	organisms	lipid	aureus	grain	livestock	review
75	chromiumiii	normal	inspection	oral	chronic	oxidation	mgl	bmd	seu	probability
76	tolerable	urine	susceptible	flame	sample	unsaturated	poland	bmdl	trr	mss
77	oral	secretariat	spain	doses	probability	polysulphides	microorganisms	animal	child	cohort
78	cancer	materials	processes	female	spatial	branchedchain	gene	citrinin	oranges	biohaz
79	dose	coe	seeds	cohort	services	rapeseed	bunted	biotoxins	indoor	neg
80	countries	benzyl	import	iodine	mouse	tri	spilled	irradiated	reasoned	revision
81	peanut	hydrolysis	efficiency	rat	combination	rosin	hungary	cells	arfdadi	breeding
82	allergies	mutagenicity	canker	bisphenol	nectar	diet	belgium	rice	jmpr	processing
83	bone	specifications	figure	air	invertebrates	specification	epidemiological	fruits	plants	herds
84	chloride	acute	transmission	environ	tiered	isomer	antibiotics	consumed	chronic	approach
85	exposure	liver	articles	pnd	scheme	lcysteine	enterococcus	contamination	metabolites	scenario
86	healthy	dna	origin	biphenyls	fields	identification	republic	maize	edible	genotype
87	acids	microgrampersonday	infested	receptor	definition	chs	czech	contribution	rotational	analytical
88	novel	validity	forest	breast	probabilistic	stearic	estonia	drinking	bean	france
89	silicon	categories	seed	seeds	residue	helveticus	infection	germany	peas	selection
90	carotene	aldehydes	climate	pregnancy	management	thermophilus	finland	don	beet	semen
91	mgkg	hamster	wood	developmental	drift	atcc	harmonised	faowho	annex	transmissible
92	egg	ntp	imported	polybrominated	refinement	ssp	sulfonamides	inorganic	uses	casings
93	annex	commerce	usda	cells	arthropods	insulin	breakpoints	diet	input	eradication
94	notes	male	input	seed	areas	cultures	lactobacillus	models	frozen	lymphoid
95	metabolism	acids	rating	monitoring	parameter	acetate	cereus	mice	spinach	brucellosis
96	nda	reverse	territory	toxicological	processes	pufa	susceptible	adolescents	rms	flocks
97	cells	chromosomal	feasibility	fatty	lines	hydrogen	slovakia	acrylamide	parent	gbr
98	allergenicity	unsaturated	vector	males	methodologies	salts	substances	contaminated	acid	figure
99	mgl	ketones	damage	weeks	review	smethyl	plant	radiation	gaps	oral
100	sulphate	exposure	field	Risk benefit	mixture	tht	veterinary	animals	metabolite	epidemiological



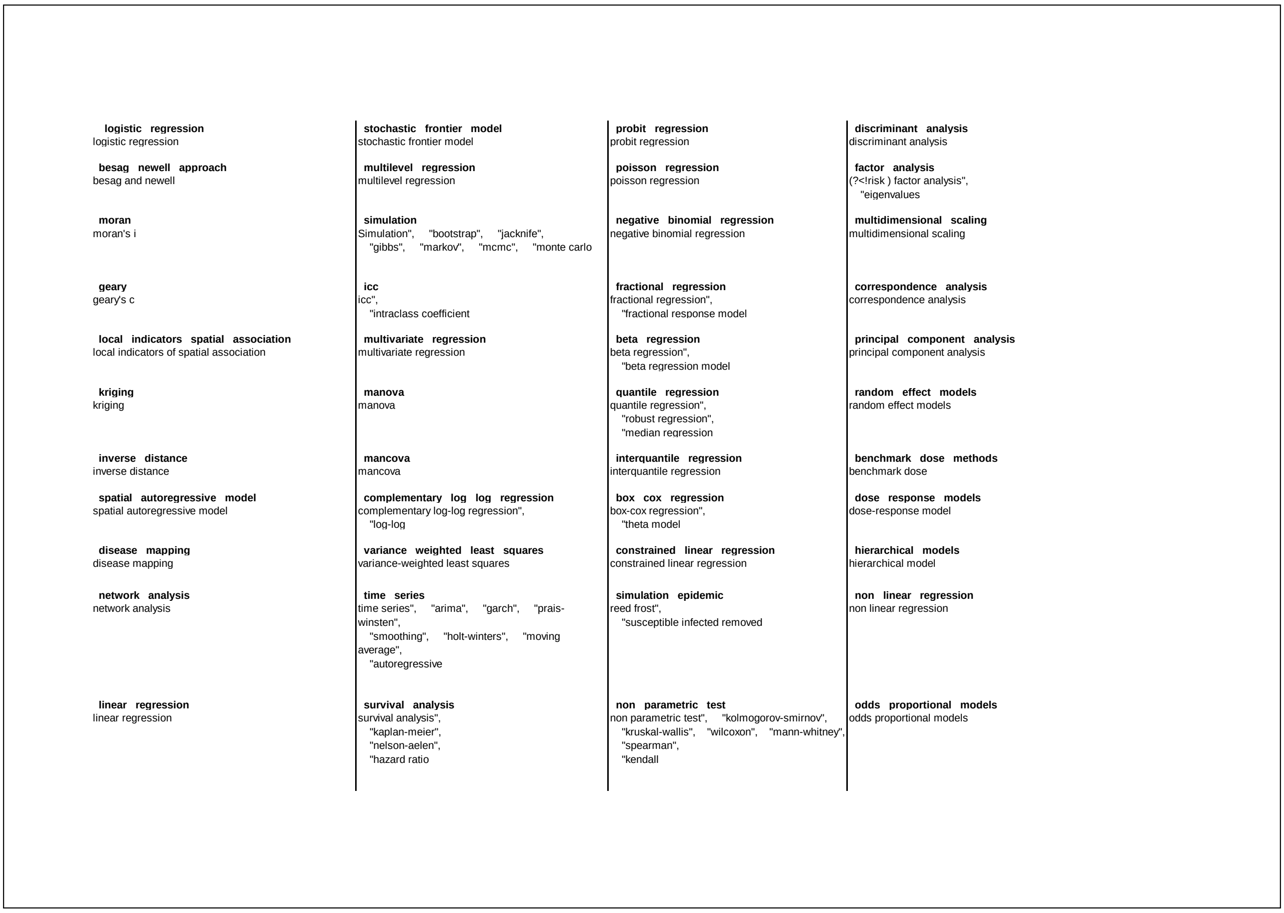
	Topic.11	Topic.12	Topic.13	Topic.14	Topic.15	Topic.16	Topic.17	Topic.18	Topic.19	Topic.20
1	salmonella	maize	welfare	feed	soil	virus	feed	mgkg	claims	active
2	meat	mon	animal	additive	substance	species	animal	rats	claimed	substance
3	prevalence	gmo	animals	fattening	active	disease	species	bwday	claim	review
4	animals	modified	stunning	additives	review	infection	mgkg	exposure	normal	peer
5	outbreaks	plants	pigs	efficacy	peer	wild	animals	toxicity	population	annex
6	inspection	genetically	fish	species	water	infected	zinc	dose	function	pesticide
7	campylobacter	plant	water	animal	pesticide	animals	additive	additives	maintenance	point
8	sampling	environmental	birds	chickens	annex	fish	exposure	mice	relationship	exposure
9	foodborne	feed	transport	feedap	mgkg	animal	additives	animals	blood	soil
10	animal	protein	stress	dose	point	birds	liver	test	substantiation	water
11	mss	soybean	pain	substances	focus	vaccination	pigs	oral	beneficial	uses
12	contamination	applicant	slaughter	applicant	acute	transmission	vitamin	jecfa	wording	toxicity
13	carcasses	gene	hazard	authorisation	stream	influenza	feedap	rat	applicant	iia
14	broiler	monitoring	management	substance	step	viruses	toxicity	additive	considers	residues
15	pigs	cultivation	exposure	sodium	metabolite	pigs	diet	acid	nutrition	gap
16	spp	proteins	killing	contact	toxicity	btv	intake	male	reduction	plant
17	poultry	organisms	calves	acid	asha	fever	fed	dietary	target	residue
18	flocks	conventional	housing	target	metabolites	spread	milk	scf	constituent	acute
19	hazards	dna	hazards	animals	residues	avian	acid	liver	dietary	environmental
20	monitoring	environment	indicators	materials	max	swine	dietary	toxicological	foods	eec
21	water	yes	temperature	dossier	iia	infectious	concentrations	noael	cause	iiia
22	infection	management	disease	piglets	exposure	population	fish	doses	functions	finalised
23	survey	field	feed	strain	chronic	clinical	copper	females	concentrations	metabolites
24	sample	processing	electrical	feedingstuffs	aquatic	vaccine	nutrition	vitro	physiological	dar
25	slaughter	rape	score	trial	sediment	areas	substances	males	acid	organisms
26	surveillance	uses	poultry	active	test	vaccines	feedingstuffs	adi	diet	surface
27	milk	oilseed	malathion	turkeys	residue	diseases	poultry	diet	intake	strain
28	coli	resistance	space	dossiersection	loam	outbreaks	gkg	female	consolidated	air
29	breeding	cryab	selection	laying	field	veterinary	diets	vivo	cholesterol	dermal
30	countries	genetic	species	migration	pecsw	countries	zearalenone	genotoxicity	nda	mgkg
31	pathogens	cotton	genetic	enzyme	crop	poultry	water	cells	subjects	test
32	zoonoses	import	monitoring	poultry	twa	probability	dairy	intake	intervention	aquatic
33	agents	glyphosate	dairy	environment	sfo	boar	dose	reevaluation	healthy	analytical
34	microbiological	transgenic	camomile	authorised	uses	vectors	selenium	aspartame	pursuant	plants
35	eggs	herbicide	floor	test	parent	susceptible	amino	children	disease	absorption
36	processing	soil	pig	pigs	degradation	surveillance	cows	developmental	consumed	formulation
37	monocytogenes	events	brain	monensin	pond	host	astaxanthin	concentrations	wordings	fate
38	disease	animal	tail	dsm	iiia	vector	cobalt	carcinogenicity	dietetic	species
39	zoonotic	genes	geese	trials	plant	figure	target	aluminium	allergies	degradation
40	outbreak	transfer	malaoxon	water	trigger	live	cattle	water	glucose	acid
41	baseline	varieties	broilers	gain	scenario	outbreak	eggs	foods	helps	rms
42	figure	expression	cows	mgkg	rat	culicoides	feeds	estimates	dha	groundwater
43	vtec	exposure	broiler	silage	gkg	water	nutritional	assay	notes	toxicological
44	reporting	species	air	skin	organisms	prevalence	rats	incidence	nutrient	purity
45	yes	pollen	cattle	hens	pec	pcr	metabolism	weeks	pertinent	applicant
46	epidemiological	counterpart	consciousness	category	ditch	mortality	oral	weights	energy	field
47	farm	nontarget	rabbits	safe	groundwater	bluetongue	authorised	reproductive	role	environment
48	enteritidis	nongm	consequences	lasalocid	surface	cattle	skin	sodium	referring	noael
49	contaminated	crops	meat	withdrawal	scenarios	field	canthaxanthin	contact	fatty	dose
50	carcass	era	unconsciousness	user	fish	salmon	weeks	chronic	metabolism	inhalation
51	species	crop	sows	material	gap	populations	blood	metabolism	substance	substances
52	germany	insect	mortality	birds	noec	sheep	ruminants	colours	fat	animal
53	reduction	allergenicity	lesions	batches	mlg	contact	meal	oil	vitamin	notifier
54	france	weed	veterinary	diclazuril	ter	area	supplementation	oils	appendix	consultation
55	egg	sequence	live	tolerance	pecsed	infections	tissue	authors	acids	monitoring
56	sources	seed	farm	supplementary	noael	strains	formaldehyde	genotoxic	characterised	oral
57	population	applicants	environment	phytase	acid	isolation	environment	absorption	authorisation	rat
58	typhimurium	compositional	animalbased	diet	purity	viral	sources	oecd	bone	crop
59	fresh	placing	flowers	zootechnical	applications	yes	plasma	materials	assumes	metabolism
60	bovine	letter	breeding	intake	dose	oie	residues	specifications	immune	nontarget
61	holdings	plan	yes	strains	cereals	farms	crosscontamination	gkg	contribution	intake
62	spain	acid	feathers	weaned	species	antibodies	muscle	mouse	fibre	operator
63	pig	event	farming	bacillus	sandy	likelihood	nivalenol	sources	extract	birds
64	serovars	trials	farmed	narasin	mgl	carp	hens	dna	skin	skin
65	retail	tolerant	area	nutrition	crops	pig	undesirable	drinks	oil	identity
66	hygiene	populations	figure	toxicity	loq	vaccinated	egg	blood	muscle	max
67	microbiology	seeds	straw	articles	environmental	test	kidney	carcinogenic	constituents	crops
68	poland	sequences	carbon	eurl	applicant	movement	materials	gum	protein	sediment
69	infections	target	cause	diets	mortality	diagnostic	chickens	urine	damage	material
70	slaughterhouse	epsps	pen	residue	drift	aquaculture	betaine	caramel	ldlcholesterol	mammals
71	batch	newly	stocking	ref	yes	strain	mycotoxins	plasma	foodconstituent	aoel
72	herds	foodfeed	responses	fed	metabolism	role	dogs	metabolic	carbohydrates	applications
73	temperature	modification	dioxide	annex	calculation	csf	applicant	statistically	quantity	pec
74	netherlands	agronomic	age	sows	bwday	meat	acids	yellow	drawn	definition
75	cattle	interactions	sensitivity	cfukg	air	africa	efficacy	beverages	childrens	rapporteur
76	pathogenic	animals	cages	salinomycin	birds	african	batches	mineral	children	predicted
77	trichinella	transformation	physiological	rabbits	analytical	epidemiological	origin	colour	joints	area
78	norway	insert	environmental	enzymes	plants	exposure	doses	mutation	exercise	loq
79	diseases	material	social	microorganisms	trials	regions	annex	gavage	contributes	ecotoxicology
80	bacteria	feral	blood	temperature	static	temperature	nontarget	substance	clarification	microorganisms
81	wild	potato	uncertainty	monitoring	mammals	farm	protein	uses	pressure	oil
82	infected	herbicides	diseases	cef	tier	ticks	laying	edible	maintain	min
83	listeria	existing	scores	lactobacillus	clay	ruminants	trace	clinical	lipids	trichoderma
84	ireland	dated	disorders	mortality	magna	endemic	manganese	adults	responses	adi
85	sheep	applications	piglets	gkg	daphnia	classical	fattening	normal	placebo	temperature
86	foods	nutritional	stun	cas	spray	france	safe	red	women	gaps
87	sweden	traits	indicator	premixtures	monitoring	herds	metabolites	negative	combination	dossier
88	veterinary	environments	bleeding	agent	buffer	aquatic	rabbits	subchronic	gum	impurities
89	programmes	unintended	husbandry	weeks	oral	diagnosis	iodine	metabolites	randomised	bacillus
90	biohaz	resistant	intensity	coccidiostats	fate	epidemic	chloride	population	comment	workers
91	feed	bacteria	reflex	treatments	askg	hpai	trout	cas	oxidative	bwday
92	microbial	test	oxygen	reared	eec	hunting	methionine	acute	glycaemic	code
93	origin	pmem	weaning	xylanase	asl	biosecurity	sulphate	kidney	scientifically	column
94	flock	postmarket	respiratory	fermentation	aoel	serological	sheep	ppm	heart	appendix
95	denmark	pat	head	ammonium	winter	larvae	horses	mix	epa	areas
96	ranking	considers	inadequate	appendix	rms	imported	supplemented	dogs	disclaimer	target
97	norovirus	receiving	conscious	nicarbazin	bbch	farmed	appendix	excretion	martin	arfd
98	kingdom	fed	farms	analytical	formulation	blood	absorption	smoke	postprandial	metabolite
99	finland	surveillance	rearing	microbial	geometric	spain	deposition	steviol	nutrients	tier
100	bacterial	fields	hens	supplemented	wheat	livestock	ochratoxin	permitted	nutritional	calculation

H CTM Classification

	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
1	fruit	pest	seeds	maize	water	exposure	exposure	substance	animals	vitamin
2	plant	virus	tea	mon	salmonella	gkg	species	active	animal	calcium
3	citrus	species	extract	gmo	contamination	fish	approach	soil	welfare	intake
4	plants	plant	alkaloids	modified	pathogens	dietary	field	review	virus	zinc
5	pest	plants	seed	plants	microbiological	feed	test	peer	disease	iron
6	area	host	feed	genetically	meat	concentrations	plant	annex	fish	chromium
7	areas	area	plant	environmental	eggs	meat	pesticides	pesticide	pigs	acid
8	spread	spread	herbal	plant	milk	contam	ppr	water	infection	selenium
9	citri	disease	species	feed	hazards	children	soil	point	bse	dietary
10	probability	vectors	caffeine	soybean	processing	milk	bees	mgkg	sheep	supplements
11	options	france	green	protein	temperature	toxins	uncertainty	exposure	birds	nutritional
12	host	hosts	leaves	gene	foods	arsenic	toxicity	acute	cattle	sources
13	flowers	vector	animal	applicant	egg	mercury	models	focus	infected	nutrition
14	management	spp	oil	monitoring	bacteria	intake	tier	toxicity	species	copper
15	species	organisms	cocoa	cultivation	coli	animal	figure	residues	wild	foods
16	tabaci	french	extracts	dna	monocytogenes	bound	yes	iia	stunning	mgday
17	establishment	pra	poppy	organisms	foodborne	pcb	substances	metabolites	test	lycopene
18	malathion	areas	plants	conventional	irradiation	analytical	appendix	stream	veterinary	potassium
19	uncertainty	probability	morphine	environment	microbial	liver	crop	uses	transmission	magnesium
20	wheat	categorisation	pas	proteins	contaminated	countries	pesticide	step	population	scf
21	countries	ticks	substances	management	hygiene	cadmium	sample	residue	water	supplementation
22	disease	viruses	glucosamine	resistance	carcasses	population	scenarios	metabolite	poultry	diet
23	tomato	diseases	botanical	field	reduction	contaminants	population	aquatic	exposure	yeast
24	effectiveness	infected	gossypol	uses	fresh	age	scenario	asha	slaughter	absorption
25	camomile	pathogen	leaf	rape	spp	shellfish	birds	test	vaccination	lutein
26	crops	establishment	ergot	processing	pathogenic	fat	environmental	max	age	sodium
27	infection	eppo	intake	oilseed	outbreak	foods	step	sediment	hazard	bioavailability
28	planting	annex	infusions	cryab	poultry	environmental	water	plant	surveillance	manganese
29	reduction	banana	bean	genetic	microbiology	lod	chemicals	gap	bovine	chloride
30	crop	citrus	alkaloid	genes	vegetables	water	modelling	field	clinical	mgkg
31	material	harmful	mgkg	cotton	bacterial	toxin	residues	iiia	sensitivity	water
32	pathogen	consequences	meal	import	biohaz	mgkg	acute	chronic	influenza	exposure
33	infected	transmission	theobromine	glyphosate	microbiol	sum	mammals	crop	transport	intakes
34	plh	symptoms	dried	herbicide	fish	median	ground	organisms	scrapie	sulphate
35	fruits	infection	medicinal	transgenic	chloramphenicol	grains	concentrations	degradation	tse	plasma
36	protected	quarantine	undesirable	events	norovirus	contamination	bcs	surface	meat	gday
37	temperature	ispm	acid	transfer	listeria	species	area	loam	goats	nutrient
38	reduce	environmental	exposure	animal	infection	infants	spray	pecsw	infectious	vitamins
39	usda	countries	coumarin	soil	nitrate	zearalenone	cumulative	environmental	hazards	concentrations
40	xanthomonas	spain	botanicals	varieties	strains	bde	endpoints	groundwater	management	adults
41	inspection	organism	glycosides	yes	log	sample	bbch	parent	swine	folate
42	potato	planting	supplements	expression	carcass	monitoring	bee	twa	diseases	petitioner
43	trees	pcr	pyrrolizidine	nontarget	microorganisms	loq	crops	sfo	feed	astaxanthin
44	canker	susceptible	doses	counterpart	surface	survey	chapter	pec	pig	upper
45	phytosanitary	material	herbs	pollen	bacillus	dioxins	sampling	rat	figure	milk
46	import	absent	water	nongm	outbreaks	estimates	monitoring	trigger	stress	additives
47	dispersed	phytosanitary	flowers	crops	ranking	adults	mortality	acid	live	children
48	imported	tick	fruits	exposure	heat	wheat	hazard	air	farmed	metabolism
49	appendix	genus	fresh	species	beef	methylmercury	applications	purity	pain	cobalt
50	consignments	uncertainty	dose	crop	efficacy	marine	selection	gkg	monitoring	carotene
51	leaves	strawberry	acute	insect	pathogen	vegetable	herbicide	noael	deer	serum
52	feasibility	southern	root	acid	reduce	figure	surface	eec	countries	amino
53	forest	departments	cyanide	era	nitrite	congeners	plants	scenario	btv	picolinate
54	soil	africa	spp	weed	disease	categories	element	species	avian	trace
55	spain	poland	livestock	letter	treatments	animals	probability	finalised	prevalence	bone
56	leaf	leaf	herb	compositional	origin	upper	consultation	pond	boar	niacin
57	figure	guadeloupe	tas	applicants	environmental	mycotoxins	oecd	dar	probability	carbonate
58	populations	verticillium	black	seed	escherichia	pcbs	focus	applicant	fever	liver
59	viruses	insect	sinensis	plan	sources	cereals	diet	fish	genetic	blood
60	environmental	martinique	toxicity	sequence	frozen	ngkg	parameter	analytical	calves	canthaxanthin
61	fresh	fruit	carvone	placing	berries	ngg	text	applications	temperature	thiamine
62	orchards	larvae	alfalfa	trials	decontamination	cereal	seeds	scenarios	mortality	tolerable
63	susceptible	ambrosia	vulgaris	tolerant	temperatures	inorganic	estimates	fate	farm	carotenoids
64	field	overseas	saponins	event	illness	category	ecosystem	plants	vaccine	minerals
65	symptoms	trees	sources	sequences	cereus	uncertainty	collection	dose	ruminants	chromiumiii
66	chrysanthemum	management	urinary	populations	infections	seafood	code	crops	indicators	population
67	treatments	populations	camellia	foodfeed	washing	contaminated	existing	noec	killing	phosphate
68	infested	greece	fruit	epsps	environment	elderly	sensitivity	ter	infectivity	salts
69	wood	portugal	gkg	seeds	inactivation	acid	substance	ditch	farms	phosphorus
70	cut	records	constituents	target	animal	foodex	default	loq	housing	methionine
71	annex	impacts	datura	bacteria	leafy	vegetables	dose	birds	vaccines	silicon
72	greenhouse	yellow	glucosinolates	allergenicity	hazard	contribution	definition	formulation	viruses	animal
73	greenhouses	isolates	hydrochloride	agronomic	hepatitis	bread	ter	rms	tissue	betaine
74	hosts	pollen	clinical	interactions	crosscontamination	substances	elicitation	metabolism	brain	zeaxanthin
75	climate	crops	patients	thuringiensis	clostridium	blood	active	mgl	blood	riboflavin
76	rating	prunus	materials	modification	viruses	acute	review	mammals	yes	women
77	water	wild	genus	insert	irradiated	toddlers	management	monitoring	classical	ohd
78	ornamental	movement	beans	newly	spores	maize	page	mortality	dairy	deficiency
79	origin	valley	fed	nutritional	acid	germany	probabilistic	mlg	breeding	substances
80	transport	transmitted	oral	feral	dairy	bmdl	pollen	pecsed	negative	mineral
81	spilled	damage	roots	herbicides	cheese	sampling	endpoint	oral	salmon	annex
82	seed	mosaic	flavanols	transformation	tomatoes	nivalenol	chronic	calculation	contact	normal
83	africa	soil	infusion	dated	cause	don	mammal	aoel	lesions	synthetic
84	bemisia	cause	diet	animals	stec	coffee	organisms	toxicological	selection	criii
85	irrigation	germany	derivatives	material	retail	scenario	focal	tier	areas	chelate
86	dispersal	runion	rats	environments	cfug	chronic	seed	trials	area	taurine
87	climatic	grapevine	aqueous	applications	radiation	toxicity	combination	bwday	outbreaks	betacarotene
88	spp	invasive	ingredients	existing	appl	environment	spraying	cereals	spread	nicotinamide
89	eppo	strains	material	traits	transport	pbdes	processes	drift	field	folic
90	territory	wilt	chocolate	resistant	botulinum	eggs	fields	yes	likelihood	acids
91	infestation	losses	catechins	unintended	contact	page	ssd	sandy	electrical	toxicity
92	grain	xanthomonas	var	potato	oysters	dairy	reproductive	static	environment	uses
93	transfer	review	plantes	pmem	hot	diet	point	spray	diagnostic	supplemented
94	reducing	republic	dietary	postmarket	chilling	fruit	aquatic	animal	specificity	diets
95	nurseries	regions	medicines	pat	diseases	materials	methodologies	definition	space	oxide
96	pinus	seeds	cake	fed	kgy	sheep	services	consultation	cows	dossier
97	population	origin	china	receiving	agents	pigs	examples	environment	farming	inorganic
98	resistance	pathogens	tec	surveillance	dose	origin	formulation	substances	review	nicotinic
99	netherlands	field	pigs	considers	organisms	oil	please	daphnia	annex	molybdenum
100	cultivation	identity	poisoning	btmaize	exposure	spain	scheme	adi	atypical	excretion

	Topic.11	Topic.12	Topic.13	Topic.14	Topic.15	Topic.16	Topic.17	Topic.18	Topic.19	Topic.20
1	residue	substances	claims	resistance	fluoride	mgkg	feed	protein	dietary	mgkg
2	mgkg	flavouring	claimed	salmonella	glucose	acid	additive	allergy	intake	rats
3	residues	flno	claim	meat	gum	exposure	animal	proteins	fatty	dose
4	mrl	jecfa	normal	animals	aspartame	additives	species	allergic	acids	toxicity
5	mrls	class	population	mss	water	contact	fattening	clinical	infants	mice
6	existing	intake	maintenance	isolates	fibre	bwday	additives	reactions	acid	bwday
7	commodities	candidate	function	prevalence	carbohydrates	substance	feedap	allergens	children	animals
8	trials	substance	relationship	campylobacter	glycaemic	materials	efficacy	foods	pet	exposure
9	review	test	substantiation	antimicrobial	sugar	additive	chickens	fish	nutrition	oral
10	others	acid	beneficial	pigs	postprandial	substances	strain	applicant	intakes	male
11	crops	negative	wording	reporting	chewing	jecfa	animals	milk	fat	rat
12	median	msdi	considers	sampling	responses	scf	applicant	patients	age	liver
13	adi	threshold	blood	countries	dietary	migration	feedingstuffs	plant	dha	doses
14	outdoor	structural	target	inspection	reduction	aluminium	dose	novel	iodine	females
15	exposure	approach	constituent	outbreaks	sucrose	adi	target	allergen	energy	males
16	gap	fge	applicant	monitoring	blood	foods	authorisation	allergenic	oil	diet
17	lettuce	genotoxicity	reduction	broiler	intake	reevaluation	substances	children	women	cells
18	wheat	toxicity	nutrition	animal	applicant	sodium	piglets	allergenicity	recycling	vitro
19	appendix	mgkg	cause	flocks	beverages	oils	qps	individuals	diet	female
20	enforcement	effa	functions	gallus	starch	cas	pigs	ige	milk	test
21	definition	assay	physiological	coli	drinks	toxicological	laying	soy	foods	noael
22	grapes	rat	consolidated	spp	dental	children	poultry	placebo	recycled	concentrations
23	pesticide	register	nda	zoonotic	tooth	flavourings	environment	oil	step	weeks
24	milk	category	lactobacillus	res	sugars	toxicity	acid	peanut	aged	vivo
25	apples	revision	concentrations	survey	material	estimates	dossiersection	lactose	infant	developmental
26	arfd	typhimurium	pursuant	sample	cassia	colours	sodium	page	protein	metabolism
27	leaves	scf	foods	surveillance	juice	dietary	turkeys	mix	epa	dna
28	commodity	specification	wordings	figure	foods	specifications	trial	ingredient	population	carcinogenicity
29	beans	additives	helps	germany	meal	oil	dossier	symptoms	adults	genotoxicity
30	seed	dose	characterised	spain	category	test	authorised	egg	flakes	incidence
31	fruit	oral	intervention	netherlands	methanol	intake	active	test	concentrations	reproductive
32	tomatoes	gml	pertinent	breeding	byproducts	esters	enzyme	fermented	formulae	urine
33	loq	ames	referring	slaughter	caries	melamine	hens	amino	decontamination	absorption
34	plant	gplate	dietetic	cattle	diet	caramel	skin	wine	disease	assay
35	animal	methyl	substance	zoonoses	insulin	processing	strains	ingredients	nutrient	blood
36	vegetables	industry	allergies	foodborne	bone	material	batches	wheat	young	chronic
37	diet	flavourings	consumed	poultry	phenylalanine	articles	bacillus	allergies	materials	mouse
38	intake	subgroup	healthy	agents	nutrition	cef	safe	index	cla	toxicological
39	potatoes	noael	acid	denmark	carbohydrate	acids	dsm	sterols	clinical	metabolites
40	active	fgerev	role	typhimurium	subjects	yellow	gain	immunology	fish	weights
41	modification	cells	strain	epidemiological	animal	edible	diets	annex	men	plasma
42	foliar	esters	subjects	bacteria	concentrations	colour	lactobacillus	acid	subjects	authors
43	acute	rats	notes	enteritidis	betaglucans	mineral	mgkg	prevalence	processes	water
44	substance	unpublished	diet	france	fruit	fats	fed	disease	gday	metabolic
45	liver	aliphatic	appendix	resistant	glycosides	formaldehyde	monensin	challenge	cholesterol	clinical
46	origin	metabolic	authorisation	indicator	sweeteners	aids	nutrition	subjects	articles	acute
47	meat	alcohols	assumes	baseline	plaque	applicant	tolerance	nutrition	tfa	kidney
48	fresh	insoluble	cholesterol	infection	oral	uses	category	mustard	efficiency	excretion
49	eec	mtamdi	disease	carcasses	barley	gkg	silage	elisa	pufa	gavage
50	metabolism	vitro	dietary	vtec	sugarfree	drinks	trials	crossreactivity	diets	dietary
51	fat	mutation	immune	mrsa	mgl	water	birds	dose	pregnancy	tissue
52	sum	cho	combination	austria	energy	authorised	supplementary	skin	healthy	genotoxic
53	peer	metabolism	metabolism	serovars	population	wax	user	clin	cardiovascular	cancer
54	iesti	intakes	nutrient	bovine	guar	beverages	water	oral	survey	oecd
55	toxicological	flavour	contribution	holdings	children	fatty	eurl	peptides	contact	statistically
56	neu	vivo	skin	antimicrobials	abp	sources	diet	analytical	countries	negative
57	code	mouse	bone	yes	beet	smoke	residue	doses	polyunsaturated	fed
58	cxl	soluble	muscle	trends	tallow	permitted	withdrawal	processing	linoleic	activation
59	pods	acids	foodconstituent	population	gbr	copolymer	microorganisms	gluten	input	mutation
60	analytical	structurally	damage	species	fat	dossier	premixtures	wines	blood	intake
61	sugar	ethyl	joints	ireland	acacia	petitioner	annex	gelatine	temperature	bpa
62	processing	flavis	drawn	poland	fructose	ans	lasalocid	mgkg	saturated	carcinogenic
63	spraying	alcohol	quantity	infections	healthy	carbon	zootechnical	casein	nutrients	maternal
64	proposal	derivatives	ldlcholesterol	clinical	wheat	categories	trichoderma	soybean	breast	dogs
65	cabbage	gavage	constituents	pig	xylitol	faowho	test	dietetic	fats	brain
66	wine	acetate	exercise	ciprofloxacin	demineralisation	ammonia	material	immunol	nutr	ppm
67	calculation	metabolised	intestinal	norway	resistant	category	appendix	severe	test	skin
68	tentative	step	clarification	acid	qra	ref	phytase	antibodies	omega	ntp
69	aiha	annex	contributes	slaughterhouse	diabetes	ammonium	diclazuril	nut	trans	radioactivity
70	dietary	faowho	gastrointestinal	farm	clinical	fcf	residues	population	challenge	tumours
71	seu	activation	maintain	sweden	tolerance	hydrocarbons	resistance	sensitisation	kcal	rabbits
72	livestock	normal	childrens	batch	plasma	genotoxicity	intake	digestion	heart	animal
73	kidney	metabolites	energy	belgium	bran	active	rabbits	sera	record	gain
74	crop	secretariat	scientifically	finland	adults	adults	dairy	cause	girls	feed
75	root	concentrations	disclaimer	herds	acids	ethyl	monitoring	peptide	cancer	serum
76	trr	hydrolysis	oxidative	kingdom	soft	come	weaned	sequence	ala	lesions
77	child	benzyl	microorganisms	hungary	oat	population	supplemented	animal	boys	thyroid
78	oranges	coe	bifidobacterium	programmes	hydrolysis	rats	sows	species	followon	urinary
79	indoor	specifications	pressure	republic	consumed	flavoured	substance	esters	sfa	toxicol
80	reasoned	bwday	intake	trichinella	residual	red	microorganism	nda	text	subchronic
81	arfdadi	materials	martin	broilers	fiber	tdi	narasin	phytosterols	deficiency	drinking
82	jmpr	aldehydes	antioxidant	czech	exposure	alcohol	analytical	lupin	serum	gestation
83	chronic	microgrampersonday	ambroise	slovakia	mineralisation	analytical	cfukg	vitro	applicant	damage
84	plants	categories	randomised	susceptibility	sweetener	industry	fermentation	histamine	lipid	species
85	metabolites	validity	lipids	harmonised	sweetened	ethanol	feeds	intake	technology	metabolite
86	edible	mutagenicity	tetens	appendix	drinking	afc	microbial	intolerance	hour	induction
87	rotational	urine	inge	flock	considers	acetic	cattle	asthma	breastfed	lung
88	annex	acute	sean	nalidixic	hydrolysed	enzymes	dogs	adults	max	bone
89	peas	unsaturated	verhagen	disease	faecal	authorisation	mortality	serum	coronary	excreted
90	uses	commerce	balanced	sources	bones	methyl	muscle	sensitivity	plasma	reduction
91	bean	hamster	weighing	estonia	syrup	blue	amino	age	thyroid	mutagenicity
92	input	cas	hans	jejuni	individuals	tier	cows	celery	review	gkg
93	beet	reverse	bowel	romania	mbm	salt	respiratory	binding	supplementation	gene
94	spinach	carboxylic	hendrik	substances	enamel	manufacturing	weeks	kda	weeks	chromosomal
95	frozen	ketones	placebo	veterinary	calcium	polyethylene	salinomycin	exposure	relationship	chemicals
96	rms	saturated	improvement	slovenia	bread	brilliant	coccidiostats	biogenic	pellets	normal
97	parent	innocuous	loveren	portugal	pap	anticipated	faecium	atopic	plastic	highdose
98	acid	ntp	bresson	retail	normal	impurities	experiment	simplex	benefits	neurotoxicity
99	gaps	solubility	jeanlouis	ampicillin	soluble	ester	crosscontamination	immune	bottles	wistar
100	metabolite	gcapitaday	seppo	strains	nondigestible	noael	age	cells	brain	aberrations

I Statistical techniques vocabulary



logistic regression
logistic regression

besag newell approach
besag and newell

moran
moran's i

geary
geary's c

local indicators spatial association
local indicators of spatial association

kriging
kriging

inverse distance
inverse distance

spatial autoregressive model
spatial autoregressive model

disease mapping
disease mapping

network analysis
network analysis

linear regression
linear regression

stochastic frontier model
stochastic frontier model

multilevel regression
multilevel regression

simulation
Simulation", "bootstrap", "jacknife",
"gibbs", "markov", "mcmc", "monte carlo

icc
icc",
"intraclass coefficient

multivariate regression
multivariate regression

manova
manova

mancova
mancova

complementary log log regression
complementary log-log regression",
"log-log

variance weighted least squares
variance-weighted least squares

time series
time series", "arima", "garch", "prais-winsten",
"smoothing", "holt-winters", "moving average",
"autoregressive

survival analysis
survival analysis",
"kaplan-meier",
"nelson-aelen",
"hazard ratio

probit regression
probit regression

poisson regression
poisson regression

negative binomial regression
negative binomial regression

fractional regression
fractional regression",
"fractional response model

beta regression
beta regression",
"beta regression model

quantile regression
quantile regression",
"robust regression",
"median regression

interquantile regression
interquantile regression

box cox regression
box-cox regression",
"theta model

constrained linear regression
constrained linear regression

simulation epidemic
reed frost",
"susceptible infected removed

non parametric test
non parametric test", "kolmogorov-smirnov",
"kruskal-wallis", "wilcoxon", "mann-whitney",
"spearman",
"kendall

discriminant analysis
discriminant analysis

factor analysis
(?<!risk) factor analysis",
"eigenvalues

multidimensional scaling
multidimensional scaling

correspondence analysis
correspondence analysis

principal component analysis
principal component analysis

random effect models
random effect models

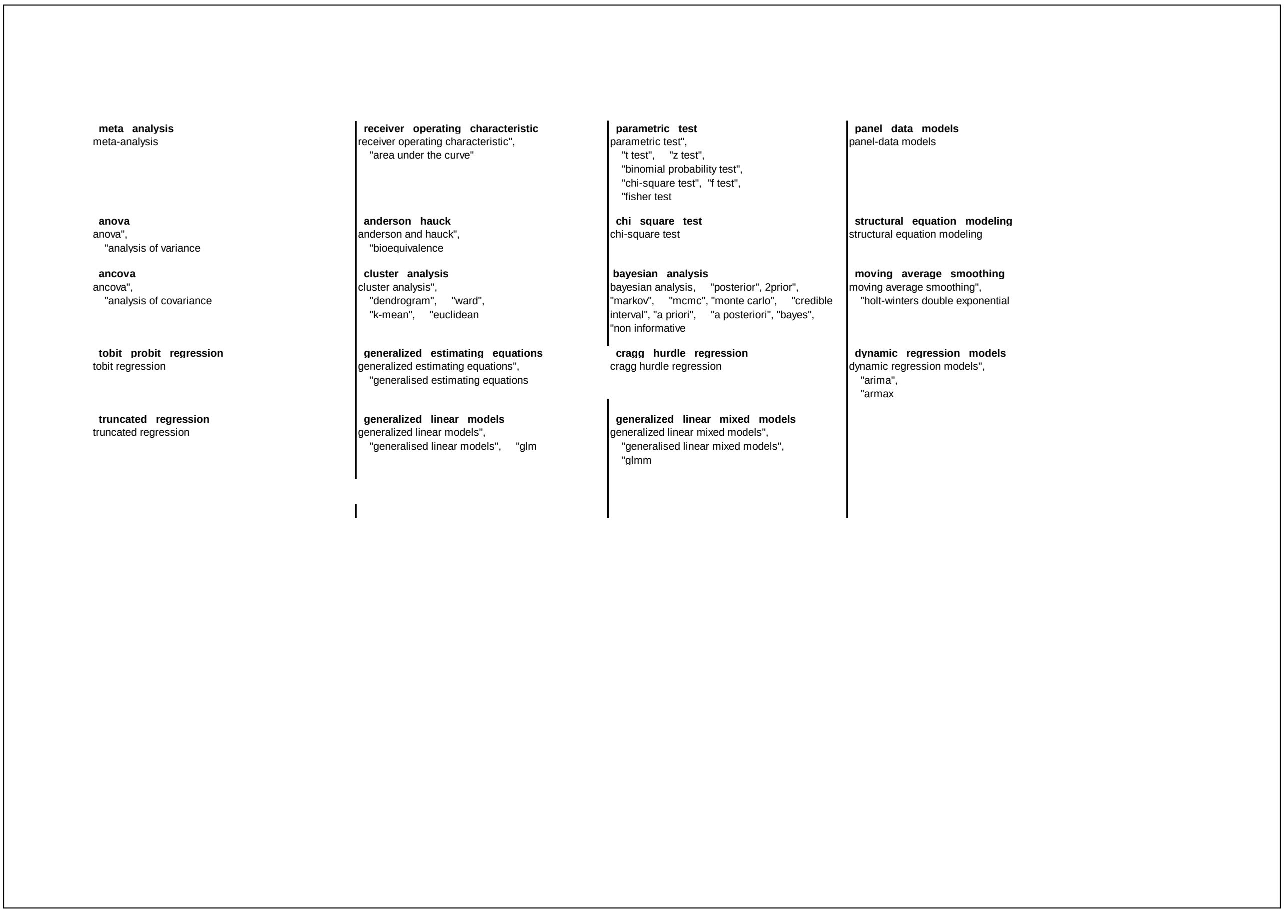
benchmark dose methods
benchmark dose

dose response models
dose-response model

hierarchical models
hierarchical model

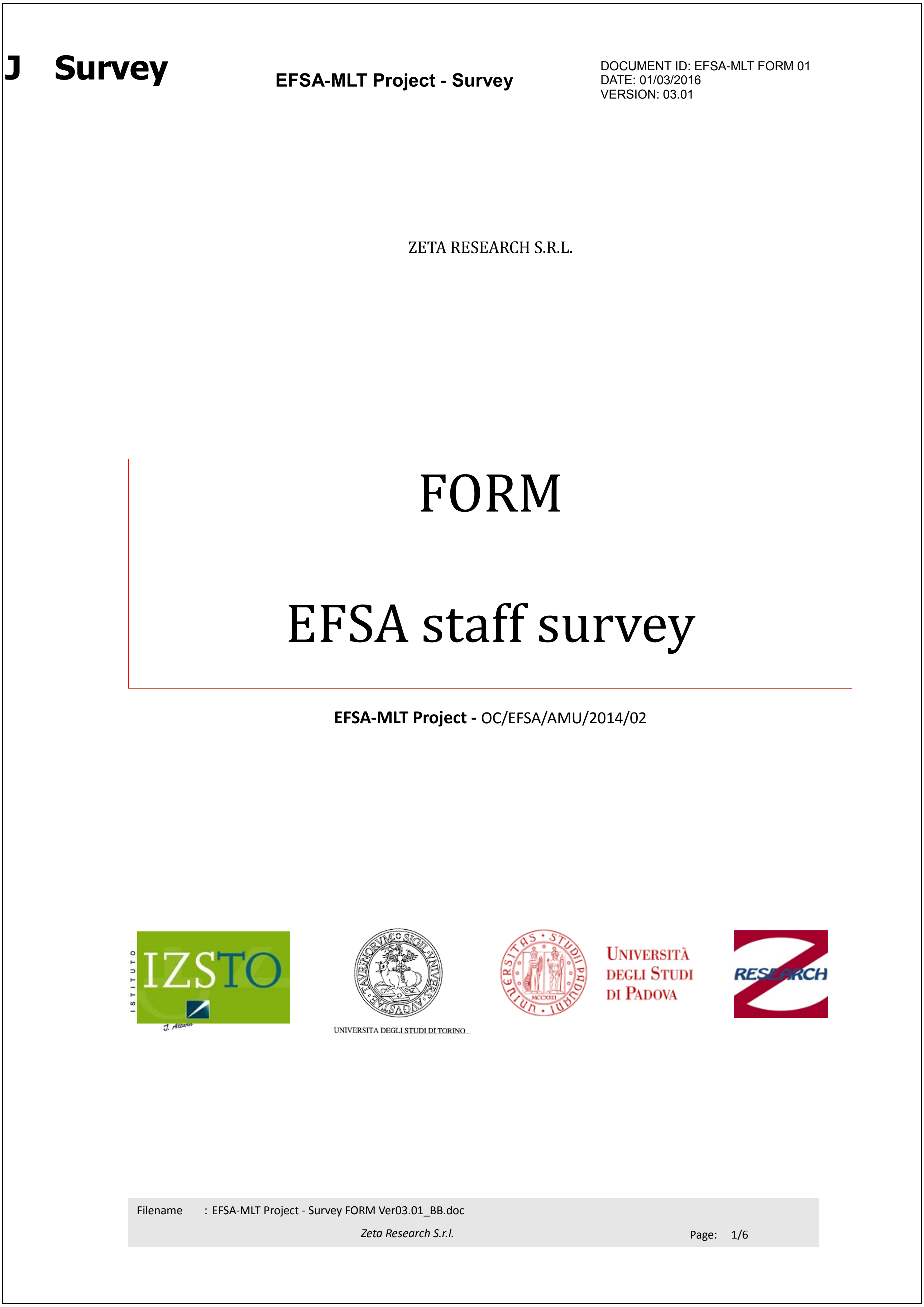
non linear regression
non linear regression

odds proportional models
odds proportional models



meta analysis meta-analysis	**receiver operating characteristic** receiver operating characteristic", "area under the curve"	**parametric test** parametric test", "t test", "z test", "binomial probability test", "chi-square test", "f test", "fisher test	**panel data models** panel-data models
anova anova", "analysis of variance	**anderson hauck** anderson and hauck", "bioequivalence	**chi square test** chi-square test	**structural equation modeling** structural equation modeling
ancova ancova", "analysis of covariance	**cluster analysis** cluster analysis", "dendrogram", "ward", "k-mean", "euclidean	**bayesian analysis** bayesian analysis, "posterior", 2prior", "markov", "mcmc", "monte carlo", "credible interval", "a priori", "a posteriori", "bayes", "non informative	**moving average smoothing** moving average smoothing", "holt-winters double exponential
tobit probit regression tobit regression	**generalized estimating equations** generalized estimating equations", "generalised estimating equations	**cragg hurdle regression** cragg hurdle regression	**dynamic regression models** dynamic regression models", "arima", "armax
truncated regression truncated regression	**generalized linear models** generalized linear models", "generalised linear models", "glm	**generalized linear mixed models** generalized linear mixed models", "generalised linear mixed models", "glmm	

J Survey



EFSA-MLT Project - Survey

DOCUMENT ID: EFSA-MLT FORM 01
DATE: 01/03/2016
VERSION: 03.01

ZETA RESEARCH S.R.L.

FORM

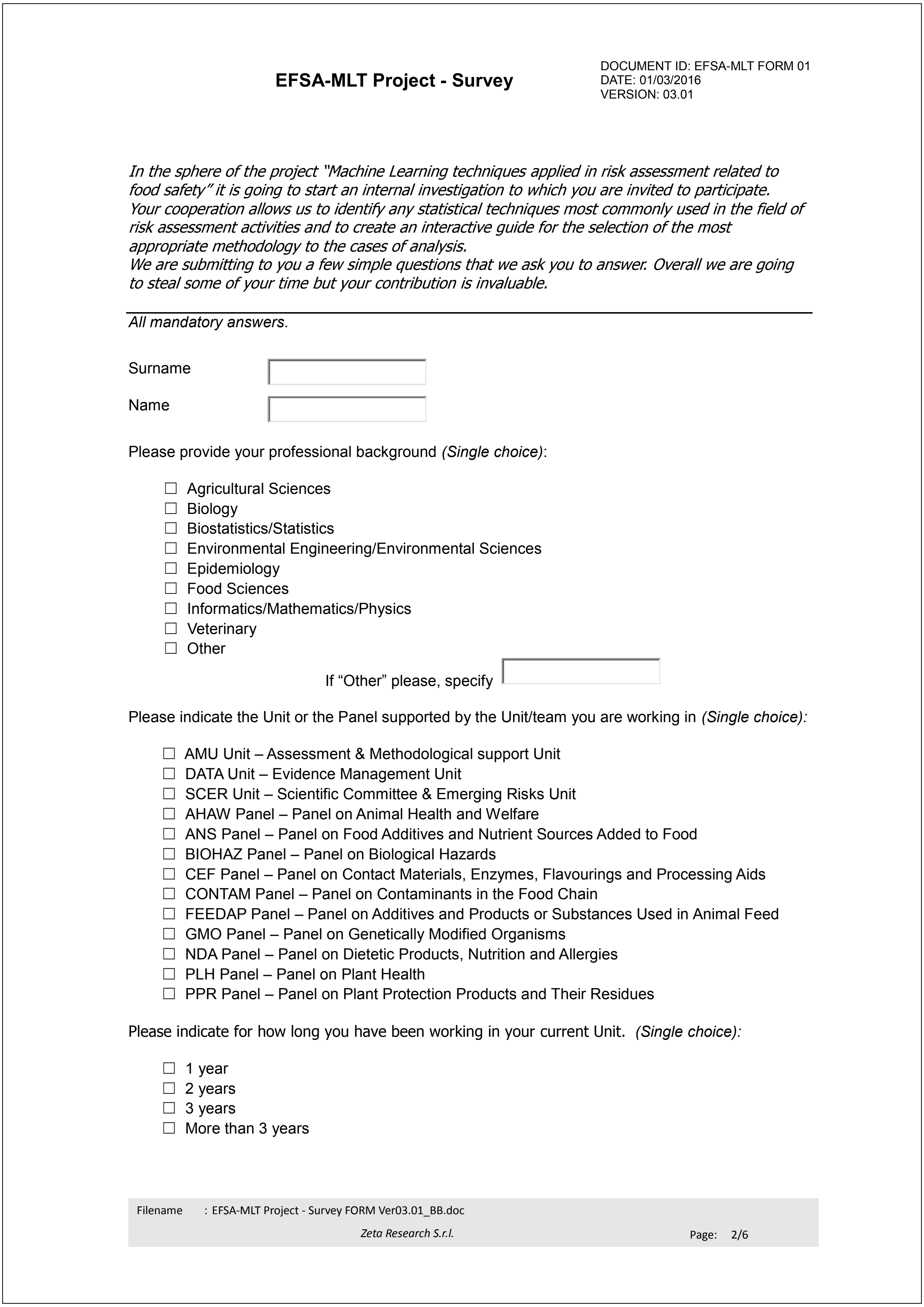
EFSA staff survey

EFSA-MLT Project - OC/EFSA/AMU/2014/02

ISTITUTO IZSTO
UNIVERSITA DEGLI STUDI DI TORINO
UNIVERSITÀ DEGLI STUDI DI PADOVA
RESEARCH

Filename : EFSA-MLT Project - Survey FORM Ver03.01_BB.doc

Zeta Research S.r.l. Page: 1/6



EFSA-MLT Project - Survey

DOCUMENT ID: EFSA-MLT FORM 01
DATE: 01/03/2016
VERSION: 03.01

In the sphere of the project "Machine Learning techniques applied in risk assessment related to food safety" it is going to start an internal investigation to which you are invited to participate.
Your cooperation allows us to identify any statistical techniques most commonly used in the field of risk assessment activities and to create an interactive guide for the selection of the most appropriate methodology to the cases of analysis.
We are submitting to you a few simple questions that we ask you to answer. Overall we are going to steal some of your time but your contribution is invaluable.

All mandatory answers.

Surname

Name

Please provide your professional background *(Single choice)*:

- □ Agricultural Sciences
- □ Biology
- □ Biostatistics/Statistics
- □ Environmental Engineering/Environmental Sciences
- □ Epidemiology
- □ Food Sciences
- □ Informatics/Mathematics/Physics
- □ Veterinary
- □ Other

If "Other" please, specify

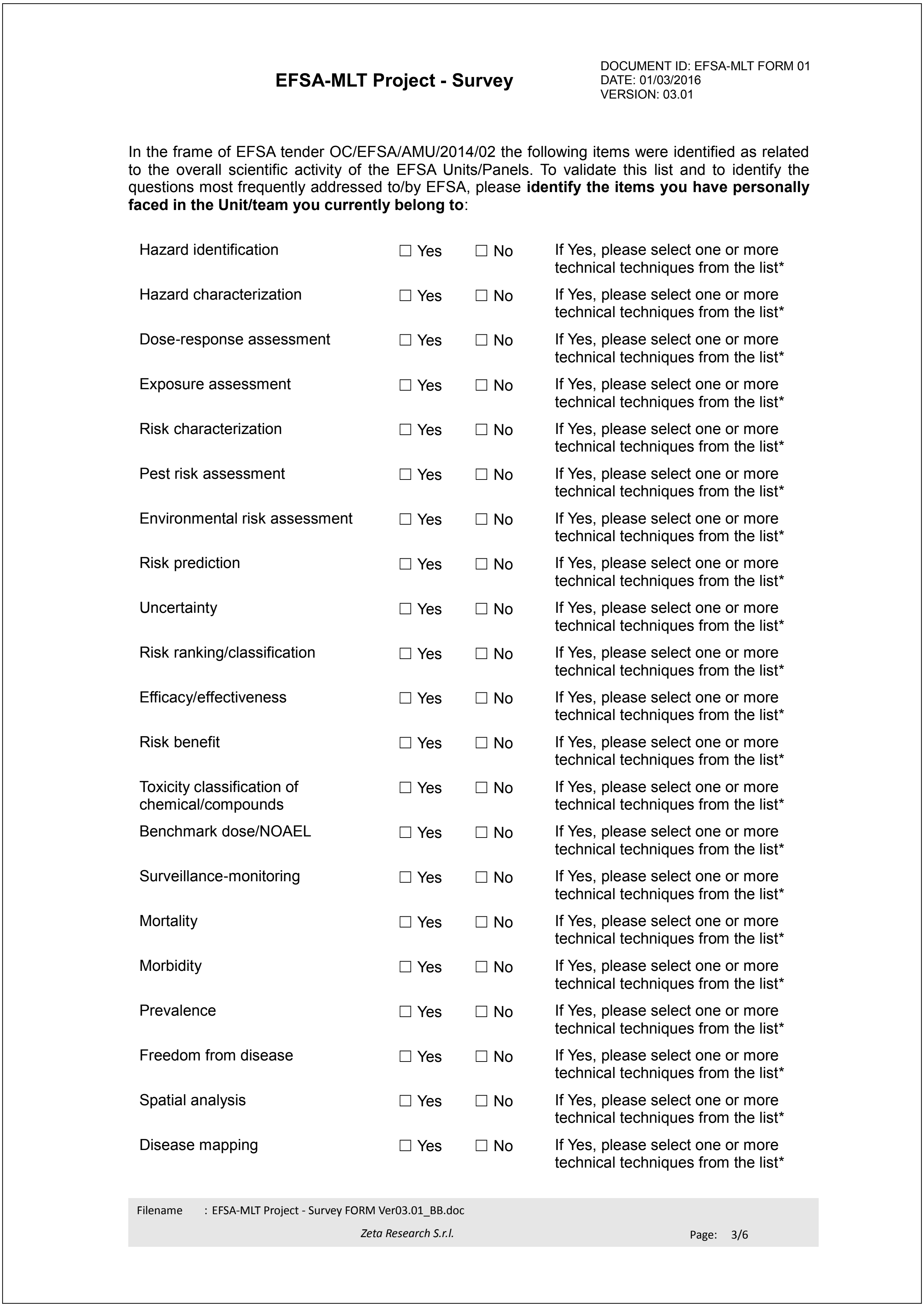
Please indicate the Unit or the Panel supported by the Unit/team you are working in *(Single choice):*

- □ AMU Unit – Assessment & Methodological support Unit
- □ DATA Unit – Evidence Management Unit
- □ SCER Unit – Scientific Committee & Emerging Risks Unit
- □ AHAW Panel – Panel on Animal Health and Welfare
- □ ANS Panel – Panel on Food Additives and Nutrient Sources Added to Food
- □ BIOHAZ Panel – Panel on Biological Hazards
- □ CEF Panel – Panel on Contact Materials, Enzymes, Flavourings and Processing Aids
- □ CONTAM Panel – Panel on Contaminants in the Food Chain
- □ FEEDAP Panel – Panel on Additives and Products or Substances Used in Animal Feed
- □ GMO Panel – Panel on Genetically Modified Organisms
- □ NDA Panel – Panel on Dietetic Products, Nutrition and Allergies
- □ PLH Panel – Panel on Plant Health
- □ PPR Panel – Panel on Plant Protection Products and Their Residues

Please indicate for how long you have been working in your current Unit. *(Single choice):*

- □ 1 year
- □ 2 years
- □ 3 years
- □ More than 3 years

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

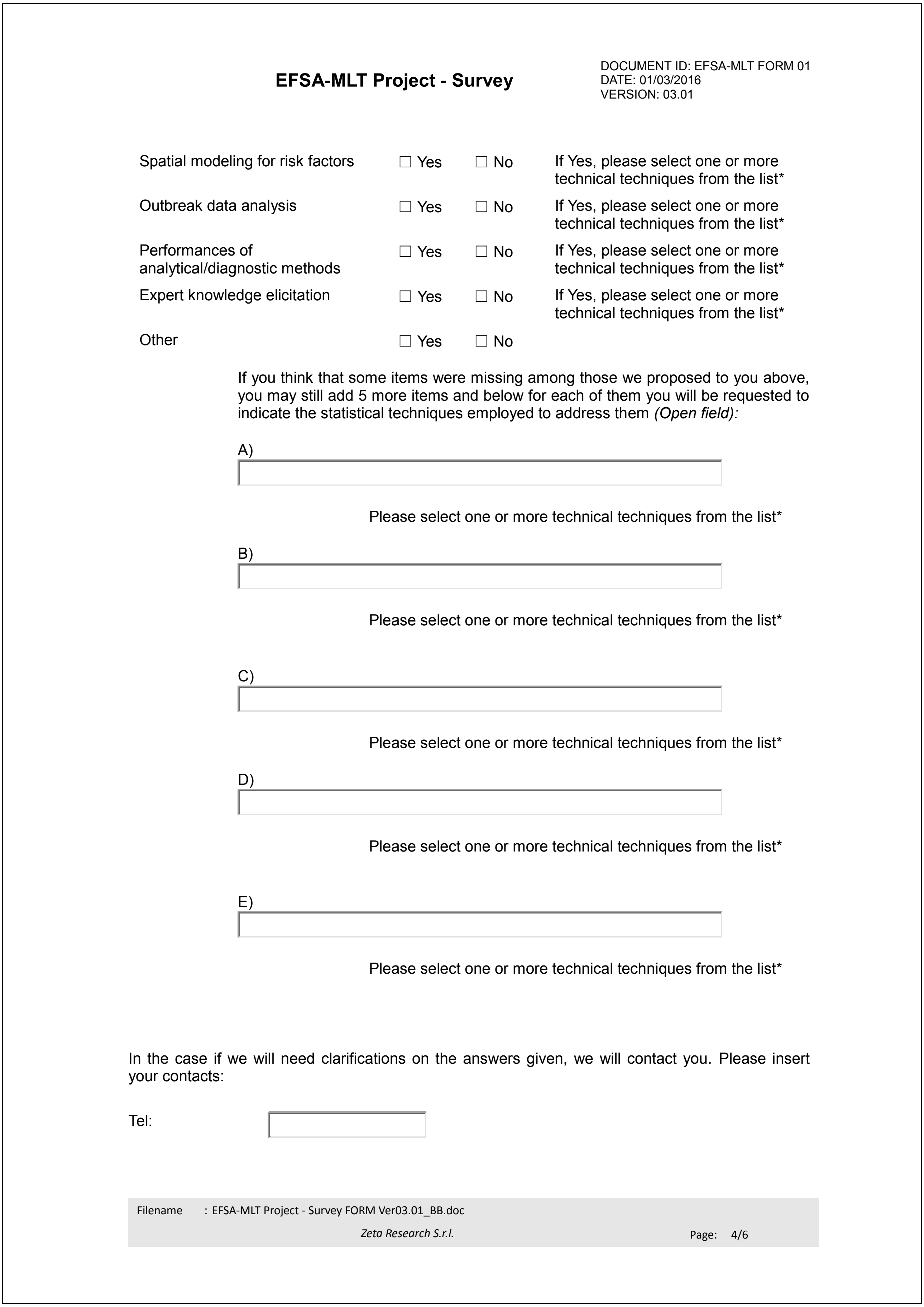


EFSA-MLT Project - Survey

DOCUMENT ID: EFSA-MLT FORM 01
DATE: 01/03/2016
VERSION: 03.01

In the frame of EFSA tender OC/EFSA/AMU/2014/02 the following items were identified as related to the overall scientific activity of the EFSA Units/Panels. To validate this list and to identify the questions most frequently addressed to/by EFSA, please **identify the items you have personally faced in the Unit/team you currently belong to**:

Hazard identification	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Hazard characterization	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Dose-response assessment	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Exposure assessment	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Risk characterization	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Pest risk assessment	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Environmental risk assessment	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Risk prediction	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Uncertainty	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Risk ranking/classification	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Efficacy/effectiveness	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Risk benefit	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Toxicity classification of chemical/compounds	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Benchmark dose/NOAEL	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Surveillance-monitoring	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Mortality	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Morbidity	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Prevalence	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Freedom from disease	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Spatial analysis	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Disease mapping	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*



EFSA-MLT Project - Survey

DOCUMENT ID: EFSA-MLT FORM 01
DATE: 01/03/2016
VERSION: 03.01

Spatial modeling for risk factors	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Outbreak data analysis	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Performances of analytical/diagnostic methods	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Expert knowledge elicitation	☐ Yes	☐ No	If Yes, please select one or more technical techniques from the list*
Other	☐ Yes	☐ No	

If you think that some items were missing among those we proposed to you above, you may still add 5 more items and below for each of them you will be requested to indicate the statistical techniques employed to address them *(Open field):*

A)

Please select one or more technical techniques from the list*

B)

Please select one or more technical techniques from the list*

C)

Please select one or more technical techniques from the list*

D)

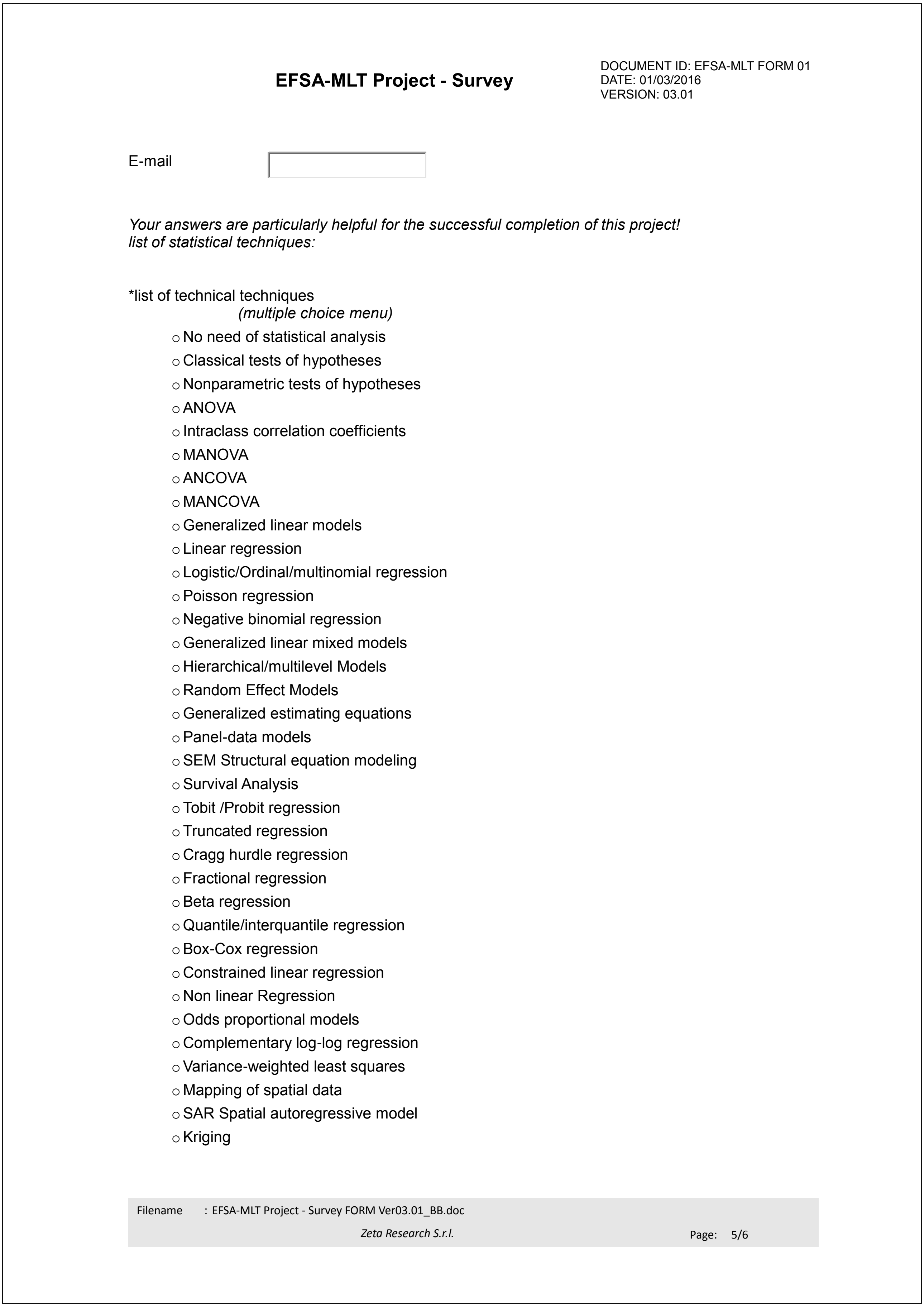
Please select one or more technical techniques from the list*

E)

Please select one or more technical techniques from the list*

In the case if we will need clarifications on the answers given, we will contact you. Please insert your contacts:

Tel:



EFSA-MLT Project - Survey

DOCUMENT ID: EFSA-MLT FORM 01
DATE: 01/03/2016
VERSION: 03.01

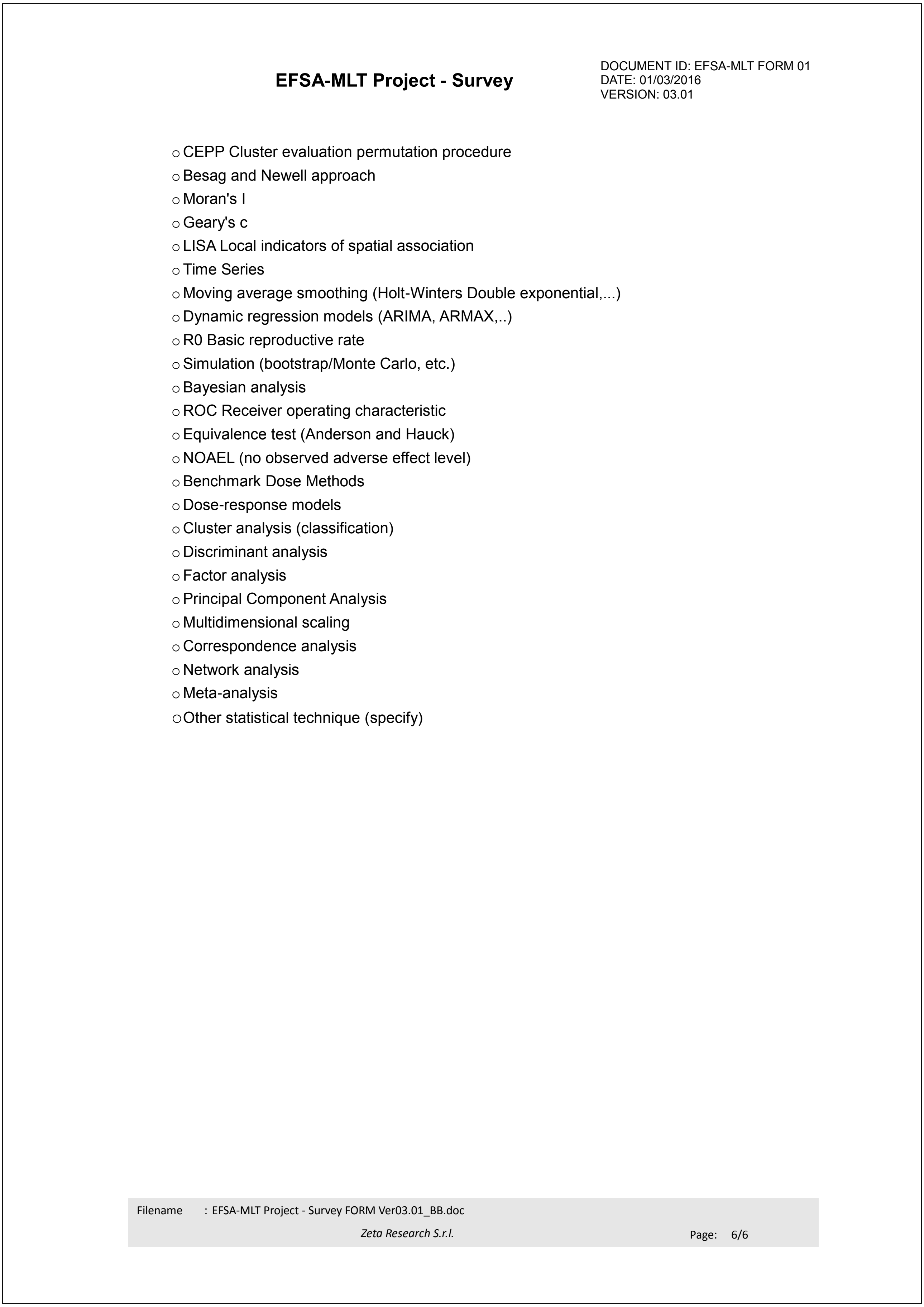
E-mail ______

Your answers are particularly helpful for the successful completion of this project!
list of statistical techniques:

*list of technical techniques

(multiple choice menu)

○ No need of statistical analysis
○ Classical tests of hypotheses
○ Nonparametric tests of hypotheses
○ ANOVA
○ Intraclass correlation coefficients
○ MANOVA
○ ANCOVA
○ MANCOVA
○ Generalized linear models
○ Linear regression
○ Logistic/Ordinal/multinomial regression
○ Poisson regression
○ Negative binomial regression
○ Generalized linear mixed models
○ Hierarchical/multilevel Models
○ Random Effect Models
○ Generalized estimating equations
○ Panel-data models
○ SEM Structural equation modeling
○ Survival Analysis
○ Tobit /Probit regression
○ Truncated regression
○ Cragg hurdle regression
○ Fractional regression
○ Beta regression
○ Quantile/interquantile regression
○ Box-Cox regression
○ Constrained linear regression
○ Non linear Regression
○ Odds proportional models
○ Complementary log-log regression
○ Variance-weighted least squares
○ Mapping of spatial data
○ SAR Spatial autoregressive model
○ Kriging



EFSA-MLT Project - Survey

DOCUMENT ID: EFSA-MLT FORM 01
DATE: 01/03/2016
VERSION: 03.01

○ CEPP Cluster evaluation permutation procedure
○ Besag and Newell approach
○ Moran's I
○ Geary's c
○ LISA Local indicators of spatial association
○ Time Series
○ Moving average smoothing (Holt-Winters Double exponential,...)
○ Dynamic regression models (ARIMA, ARMAX,..)
○ R0 Basic reproductive rate
○ Simulation (bootstrap/Monte Carlo, etc.)
○ Bayesian analysis
○ ROC Receiver operating characteristic
○ Equivalence test (Anderson and Hauck)
○ NOAEL (no observed adverse effect level)
○ Benchmark Dose Methods
○ Dose-response models
○ Cluster analysis (classification)
○ Discriminant analysis
○ Factor analysis
○ Principal Component Analysis
○ Multidimensional scaling
○ Correspondence analysis
○ Network analysis
○ Meta-analysis
○ Other statistical technique (specify)

Filename : EFSA-MLT Project - Survey FORM Ver03.01_BB.doc
Zeta Research S.r.l. Page: 6/6

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.

Recruitment of participants

From EFSA a complete list of the staff was obtained. From the list a subset of EFSA staff was selected in the following way:

- restriction to the staff from the two operational scientific directorates – Risk Assessment and Scientific Assistance (RASA) and Scientific Evaluation of Regulated Products (REPRO);
- involvement of all the scientific officers and senior officers after exclusion of the support staff and of Unit/team managers, After this procedure a list of 160 participants was defined and a mailing list was prepared.

Implementation of the online Computer Assisted Web Interviewing (CAWI) investigation

Type of data collection The mentioned questionnaire served to develop an online interview. To maximize the participation rate, participants were sent an initial email from EFSA to anticipate and officially present the survey. The email text was agreed with the internal manager of the Project and it is included below:

> *Dear Colleague, in the sphere of the project "Machine Learning techniques applied in risk assessment related to food safety" we are inviting you to participate in a survey in which your cooperation allows us to identify any statistical techniques most commonly used in the field of risk assessment activities with the objective to create an interactive guide for the selection of the most appropriate methodology for each analysis case.*
>
> *In the coming days you will receive an email from Zeta Research with a link which will redirect you to the online survey. The survey contain simple yes no questions which refers to specific topics potentially addressed by the Panel/Unit where you work. Being the reply to a question " yes" i.e. the topic was actually addressed within your activity, the statistical methods used will be asked through a multiple choice menu.*
>
> *The overall completion of the survey should not take more than 30 minutes but your contribution will be extremely useful.*

Afterwards, participants were sent a second e-mail from the Consortium with the web link to access the survey online as show in Figure 61. Once connected to the link, the participants had the opportunity to answer the questions, with the possibility to continue later or complete immediately the survey. Once questionnaire is successfully completed, the system sends an automatic e-mail as shown in Figure 62. Finally up to 4 kind reminders were sent over the following month to the employees who have not yet accessed and completed the survey.
The mailing list has been uploaded into the Consortium database and the delivery of the communications is managed by an automated mail server.

Privacy The survey is not anonymous as the participants are requested to provide their personal contact data. In order to meet the regulatory requirements regarding the protection of personal data (European General Data Protection Regulation) in collaboration with the Legal EFSA staff, it has been prepared the mentioned "Note on the processing of personal data in the context of survey" which is available in the first page of the survey, directly to the participant and reported in Figure 63

Data Collection and Quality Controls Preliminary, a short pilot survey has been carried out in order to assess the timing and comprehension of the questionnaire. The pilot survey was shared with EFSA's project staff who helped in circulating the draft questionnaire among one referent by EFSA Panel team. On the basis of the collected comments it was decided implement a second release to clarify or replace part of the initially included items and by eliminating as much as possible open fields, with the aim to increase the quality of the survey.

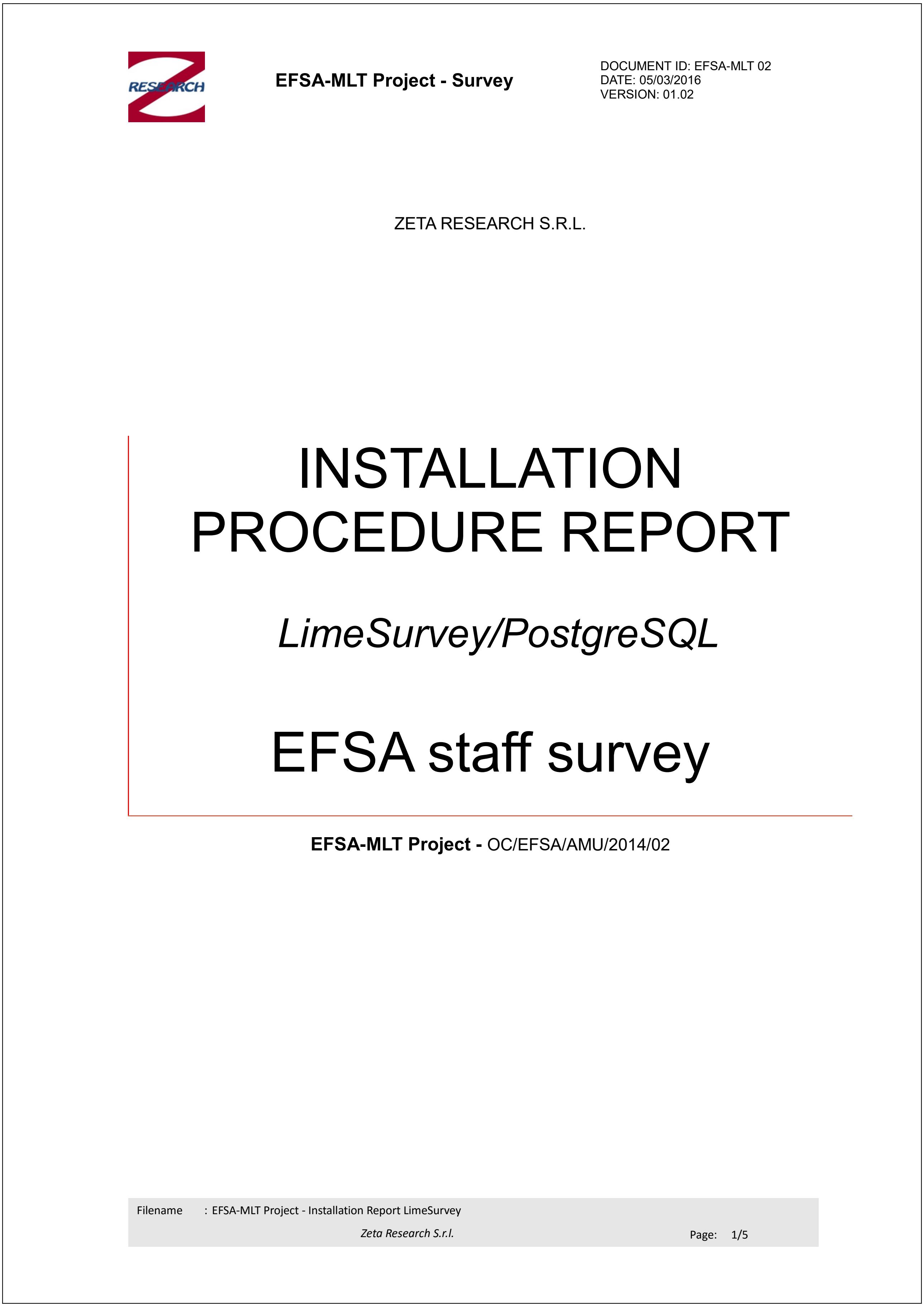
The use of the CAWI system allows quality control and coherence of the answers, as the software sets in advance the "rules" that must be followed in filling out the questionnaire. The data of the survey on the web platform are available in real time in the ZETA data server. In this way the Company guarantees both the "protection" of the information collected and the control on the survey trend. To protect EFSA and the quality accuracy of the information collected, the system used by ZETA makes impossible to change the data collected during the interviews.

Each participant was associated with an electronic ID in order to monitor the survey compliance and the delivery of personalized reminders. The following preliminary quality assurance checks/trials were performed:

- Upload of the participants mailing list
- Check of the individual ID and of the matching
- Correct automated delivery of communications (first and second email)
- Test on response times
- Test on correct on line page visualization from EFSA personal computers

Database set up Due to the extensive number of variables to be associated with multiple choice menus (i.e.60 statistical techniques), it was not possible to use classical data collection tools like LimeSurvey/MySQL,for its limitation to 1000-column. Therefore was developed an ad hoc tool allowing a 1800-column database (http://surveys.zetafield.eu/largesurvey/). .

Installing procedure



EFSA-MLT Project - Survey

DOCUMENT ID: EFSA-MLT 02
DATE: 05/03/2016
VERSION: 01.02

ZETA RESEARCH S.R.L.

INSTALLATION PROCEDURE REPORT

LimeSurvey/PostgreSQL

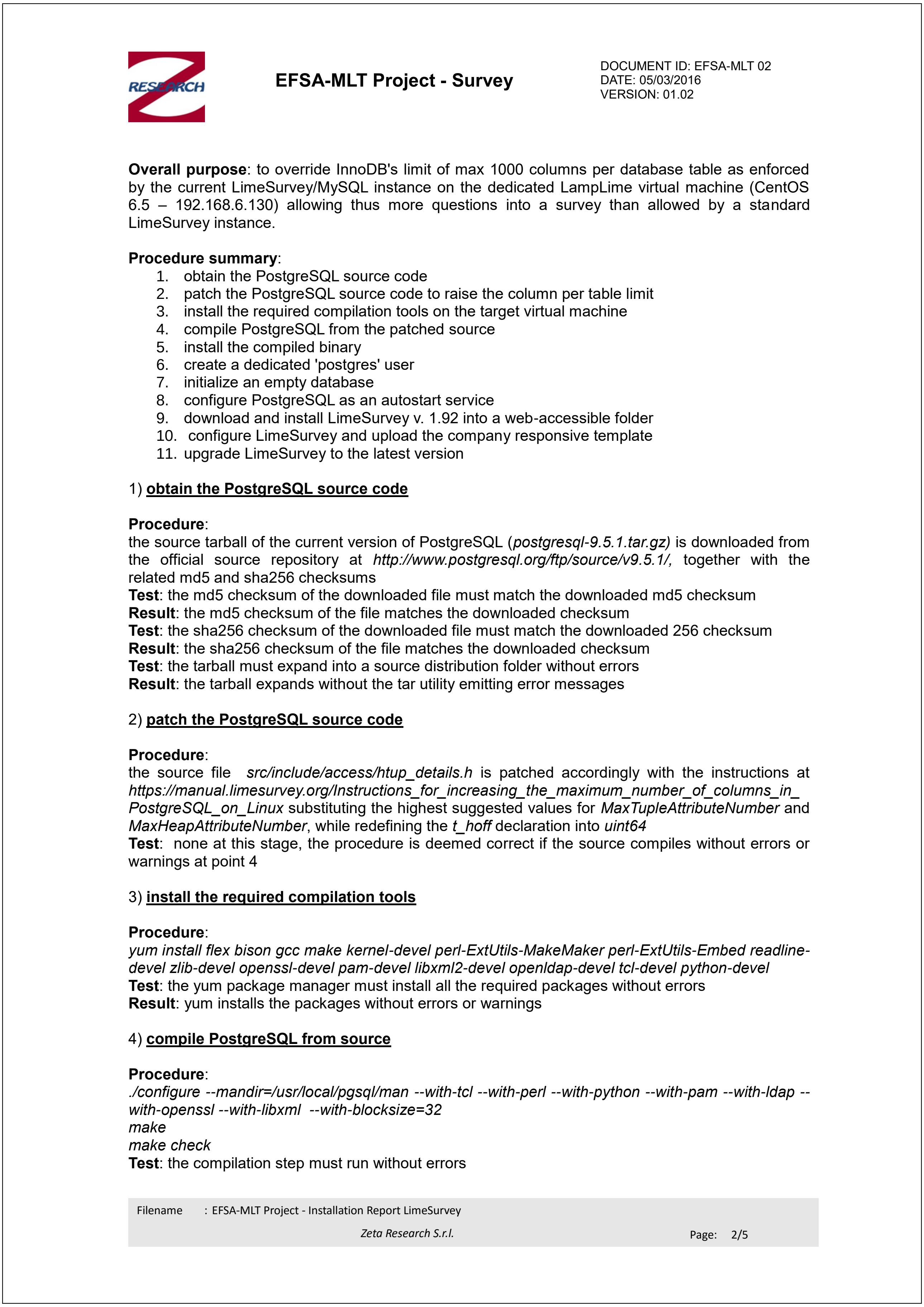
EFSA staff survey

EFSA-MLT Project - OC/EFSA/AMU/2014/02

Filename : EFSA-MLT Project - Installation Report LimeSurvey

Zeta Research S.r.l. Page: 1/5

The present document has been produced and adopted by the bodies identified above as author(s). This task has been carried out exclusively by the author(s) in the context of a contract between the European Food Safety Authority and the author(s), awarded following a tender procedure. The present document is published complying with the transparency principle to which the Authority is subject. It may not be considered as an output adopted by the Authority. The European Food Safety Authority reserves its rights, view and position as regards the issues addressed and the conclusions reached in the present document, without prejudice to the rights of the authors.



EFSA-MLT Project - Survey

DOCUMENT ID: EFSA-MLT 02
DATE: 05/03/2016
VERSION: 01.02

Overall purpose: to override InnoDB's limit of max 1000 columns per database table as enforced by the current LimeSurvey/MySQL instance on the dedicated LampLime virtual machine (CentOS 6.5 – 192.168.6.130) allowing thus more questions into a survey than allowed by a standard LimeSurvey instance.

Procedure summary:

1. obtain the PostgreSQL source code
2. patch the PostgreSQL source code to raise the column per table limit
3. install the required compilation tools on the target virtual machine
4. compile PostgreSQL from the patched source
5. install the compiled binary
6. create a dedicated 'postgres' user
7. initialize an empty database
8. configure PostgreSQL as an autostart service
9. download and install LimeSurvey v. 1.92 into a web-accessible folder
10. configure LimeSurvey and upload the company responsive template
11. upgrade LimeSurvey to the latest version

1) **obtain the PostgreSQL source code**

Procedure:
the source tarball of the current version of PostgreSQL (*postgresql-9.5.1.tar.gz)* is downloaded from the official source repository at *http://www.postgresql.org/ftp/source/v9.5.1/,* together with the related md5 and sha256 checksums
Test: the md5 checksum of the downloaded file must match the downloaded md5 checksum
Result: the md5 checksum of the file matches the downloaded checksum
Test: the sha256 checksum of the downloaded file must match the downloaded 256 checksum
Result: the sha256 checksum of the file matches the downloaded checksum
Test: the tarball must expand into a source distribution folder without errors
Result: the tarball expands without the tar utility emitting error messages

2) **patch the PostgreSQL source code**

Procedure:
the source file *src/include/access/htup_details.h* is patched accordingly with the instructions at *https://manual.limesurvey.org/Instructions_for_increasing_the_maximum_number_of_columns_in_PostgreSQL_on_Linux* substituting the highest suggested values for *MaxTupleAttributeNumber* and *MaxHeapAttributeNumber*, while redefining the *t_hoff* declaration into *uint64*
Test: none at this stage, the procedure is deemed correct if the source compiles without errors or warnings at point 4

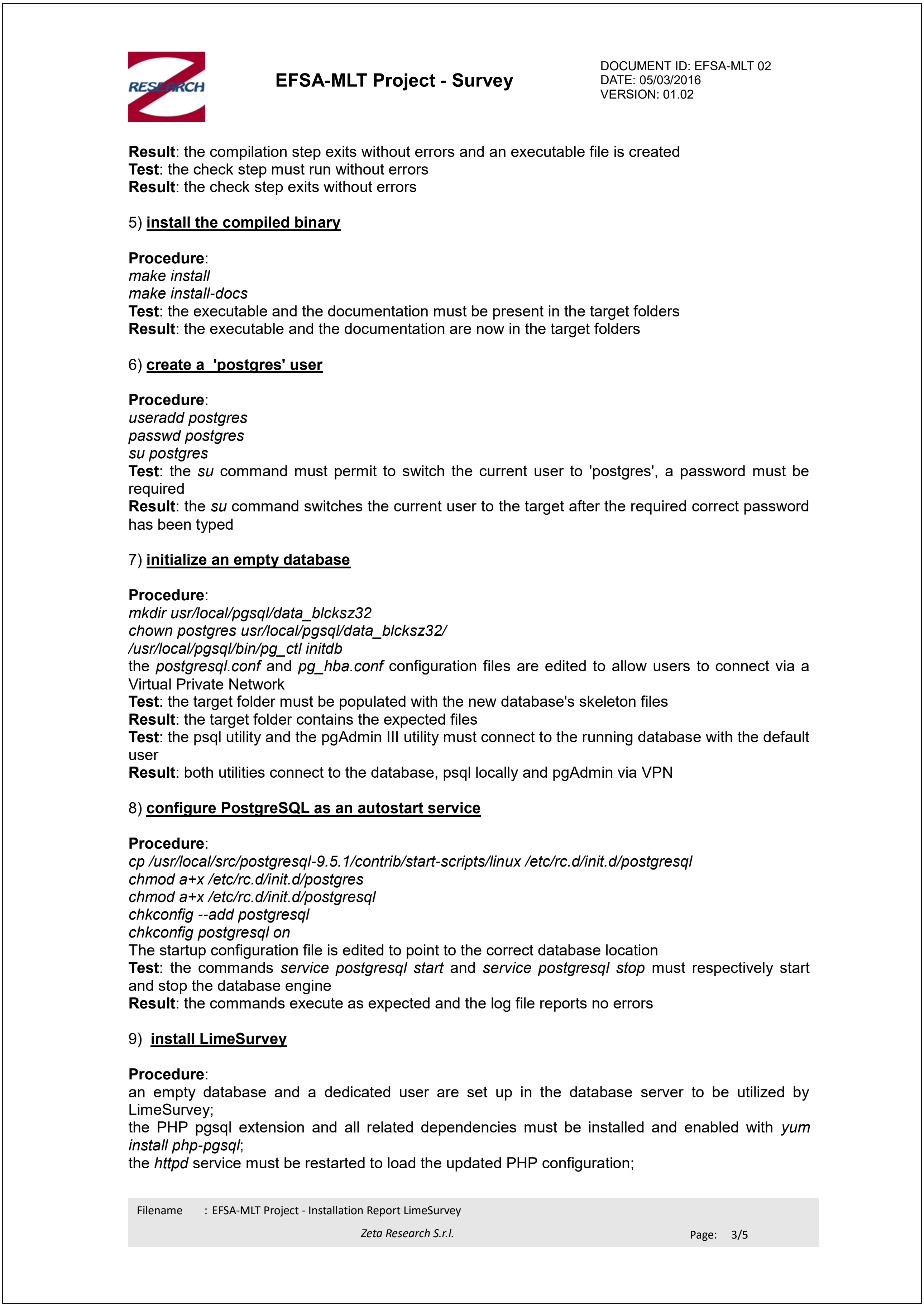
3) **install the required compilation tools**

Procedure:
yum install flex bison gcc make kernel-devel perl-ExtUtils-MakeMaker perl-ExtUtils-Embed readline-devel zlib-devel openssl-devel pam-devel libxml2-devel openldap-devel tcl-devel python-devel
Test: the yum package manager must install all the required packages without errors
Result: yum installs the packages without errors or warnings

4) **compile PostgreSQL from source**

Procedure:
./configure --mandir=/usr/local/pgsql/man --with-tcl --with-perl --with-python --with-pam --with-ldap --with-openssl --with-libxml --with-blocksize=32
make
make check
Test: the compilation step must run without errors

Filename : EFSA-MLT Project - Installation Report LimeSurvey
Zeta Research S.r.l. Page: 2/5



Result: the compilation step exits without errors and an executable file is created
Test: the check step must run without errors
Result: the check step exits without errors

5) **install the compiled binary**

Procedure:

```
make install
make install-docs
```

Test: the executable and the documentation must be present in the target folders
Result: the executable and the documentation are now in the target folders

6) **create a 'postgres' user**

Procedure:

```
useradd postgres
passwd postgres
su postgres
```

Test: the *su* command must permit to switch the current user to 'postgres', a password must be required
Result: the *su* command switches the current user to the target after the required correct password has been typed

7) **initialize an empty database**

Procedure:

```
mkdir usr/local/pgsql/data_blcksz32
chown postgres usr/local/pgsql/data_blcksz32/
/usr/local/pgsql/bin/pg_ctl initdb
```

the *postgresql.conf* and *pg_hba.conf* configuration files are edited to allow users to connect via a Virtual Private Network
Test: the target folder must be populated with the new database's skeleton files
Result: the target folder contains the expected files
Test: the psql utility and the pgAdmin III utility must connect to the running database with the default user
Result: both utilities connect to the database, psql locally and pgAdmin via VPN

8) **configure PostgreSQL as an autostart service**

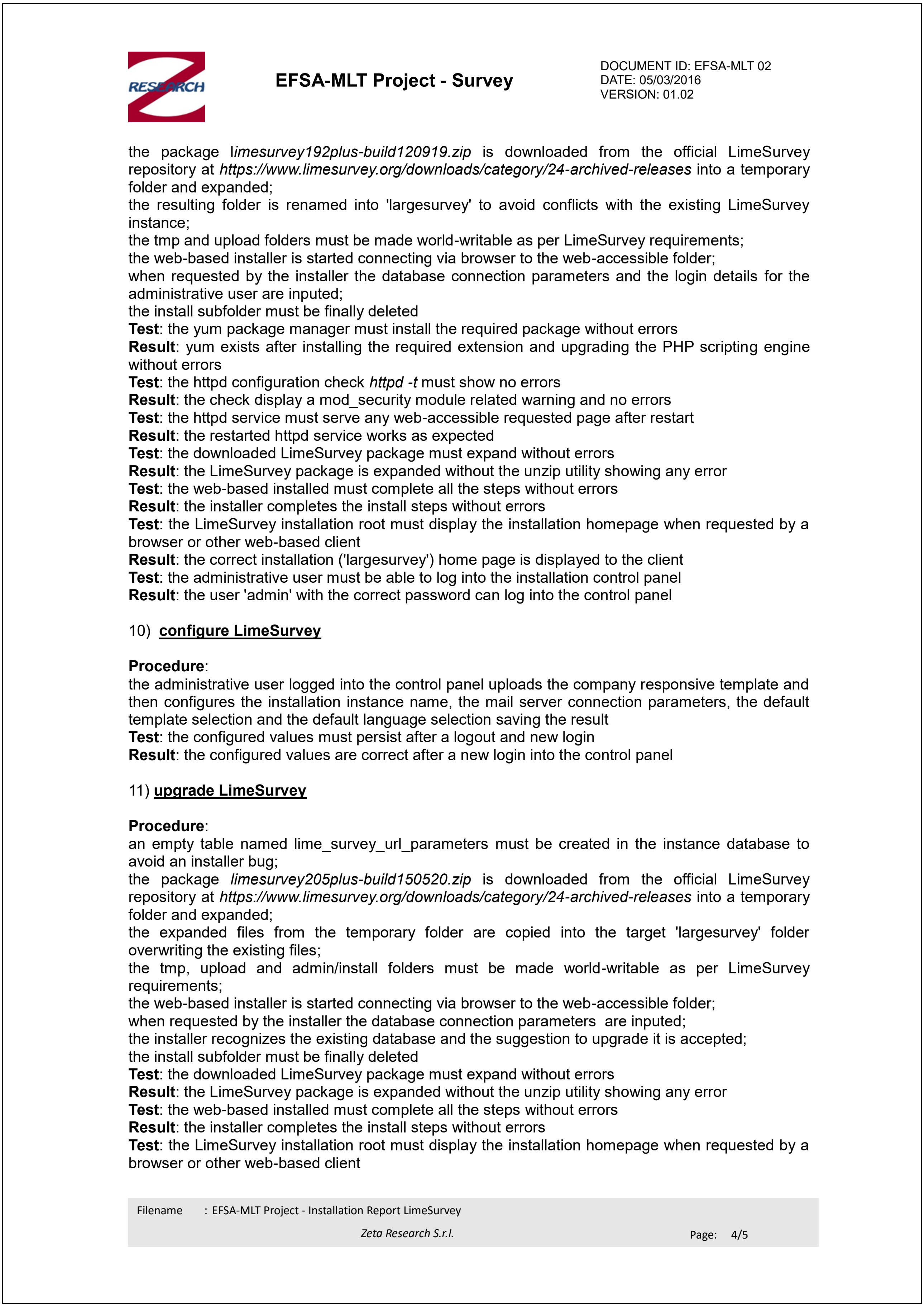
Procedure:

```
cp /usr/local/src/postgresql-9.5.1/contrib/start-scripts/linux /etc/rc.d/init.d/postgresql
chmod a+x /etc/rc.d/init.d/postgres
chmod a+x /etc/rc.d/init.d/postgresql
chkconfig --add postgresql
chkconfig postgresql on
```

The startup configuration file is edited to point to the correct database location
Test: the commands *service postgresql start* and *service postgresql stop* must respectively start and stop the database engine
Result: the commands execute as expected and the log file reports no errors

9) **install LimeSurvey**

Procedure:
an empty database and a dedicated user are set up in the database server to be utilized by LimeSurvey;
the PHP pgsql extension and all related dependencies must be installed and enabled with *yum install php-pgsql*;
the *httpd* service must be restarted to load the updated PHP configuration;



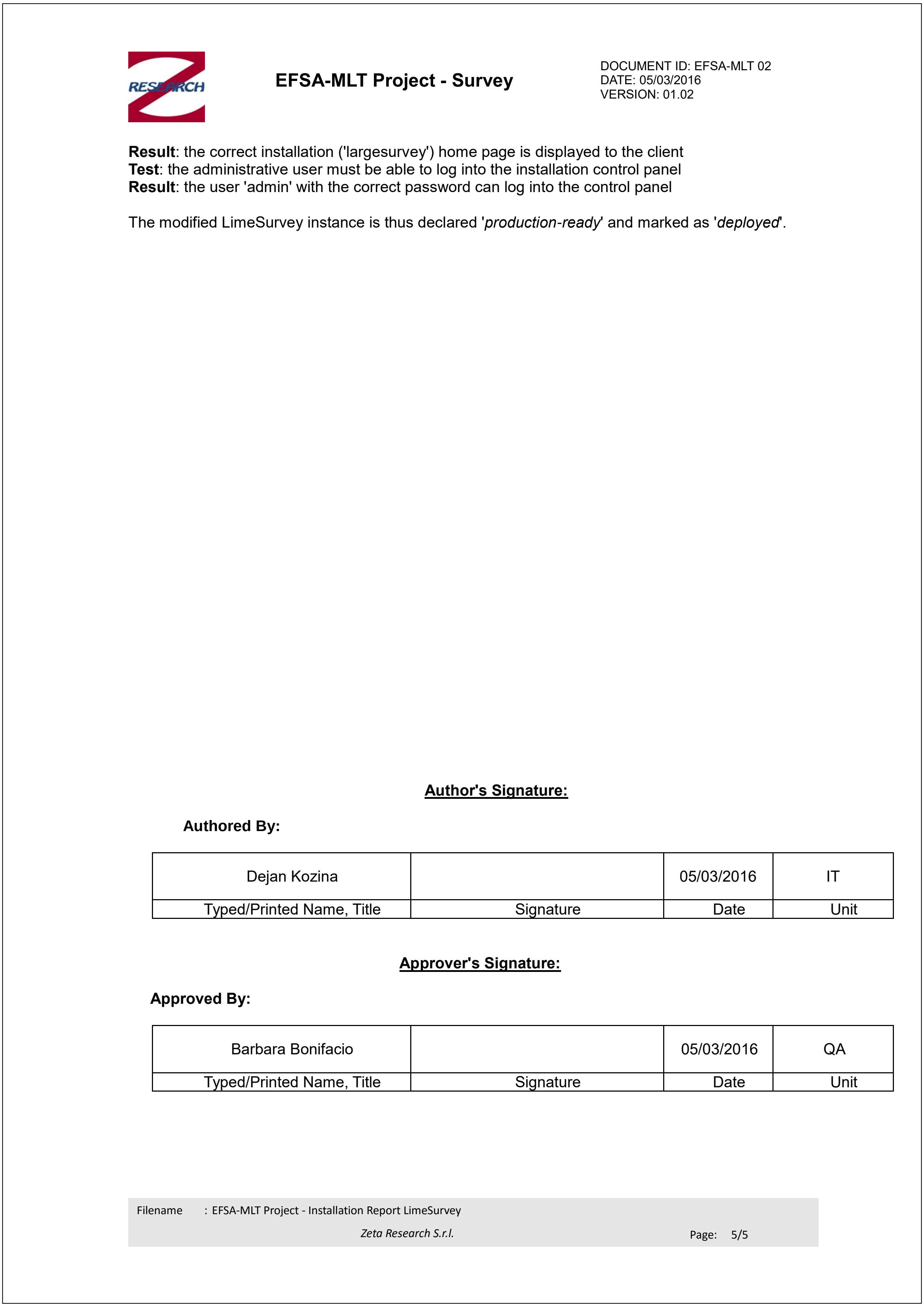
the package l*imesurvey192plus-build120919.zip* is downloaded from the official LimeSurvey repository at *https://www.limesurvey.org/downloads/category/24-archived-releases* into a temporary folder and expanded;
the resulting folder is renamed into 'largesurvey' to avoid conflicts with the existing LimeSurvey instance;
the tmp and upload folders must be made world-writable as per LimeSurvey requirements;
the web-based installer is started connecting via browser to the web-accessible folder;
when requested by the installer the database connection parameters and the login details for the administrative user are inputed;
the install subfolder must be finally deleted
Test: the yum package manager must install the required package without errors
Result: yum exists after installing the required extension and upgrading the PHP scripting engine without errors
Test: the httpd configuration check *httpd -t* must show no errors
Result: the check display a mod_security module related warning and no errors
Test: the httpd service must serve any web-accessible requested page after restart
Result: the restarted httpd service works as expected
Test: the downloaded LimeSurvey package must expand without errors
Result: the LimeSurvey package is expanded without the unzip utility showing any error
Test: the web-based installed must complete all the steps without errors
Result: the installer completes the install steps without errors
Test: the LimeSurvey installation root must display the installation homepage when requested by a browser or other web-based client
Result: the correct installation ('largesurvey') home page is displayed to the client
Test: the administrative user must be able to log into the installation control panel
Result: the user 'admin' with the correct password can log into the control panel

10) **configure LimeSurvey**

Procedure:
the administrative user logged into the control panel uploads the company responsive template and then configures the installation instance name, the mail server connection parameters, the default template selection and the default language selection saving the result
Test: the configured values must persist after a logout and new login
Result: the configured values are correct after a new login into the control panel

11) **upgrade LimeSurvey**

Procedure:
an empty table named lime_survey_url_parameters must be created in the instance database to avoid an installer bug;
the package *limesurvey205plus-build150520.zip* is downloaded from the official LimeSurvey repository at *https://www.limesurvey.org/downloads/category/24-archived-releases* into a temporary folder and expanded;
the expanded files from the temporary folder are copied into the target 'largesurvey' folder overwriting the existing files;
the tmp, upload and admin/install folders must be made world-writable as per LimeSurvey requirements;
the web-based installer is started connecting via browser to the web-accessible folder;
when requested by the installer the database connection parameters are inputed;
the installer recognizes the existing database and the suggestion to upgrade it is accepted;
the install subfolder must be finally deleted
Test: the downloaded LimeSurvey package must expand without errors
Result: the LimeSurvey package is expanded without the unzip utility showing any error
Test: the web-based installed must complete all the steps without errors
Result: the installer completes the install steps without errors
Test: the LimeSurvey installation root must display the installation homepage when requested by a browser or other web-based client



Result: the correct installation ('largesurvey') home page is displayed to the client
Test: the administrative user must be able to log into the installation control panel
Result: the user 'admin' with the correct password can log into the control panel

The modified LimeSurvey instance is thus declared '*production-ready*' and marked as '*deployed*'.

Author's Signature:

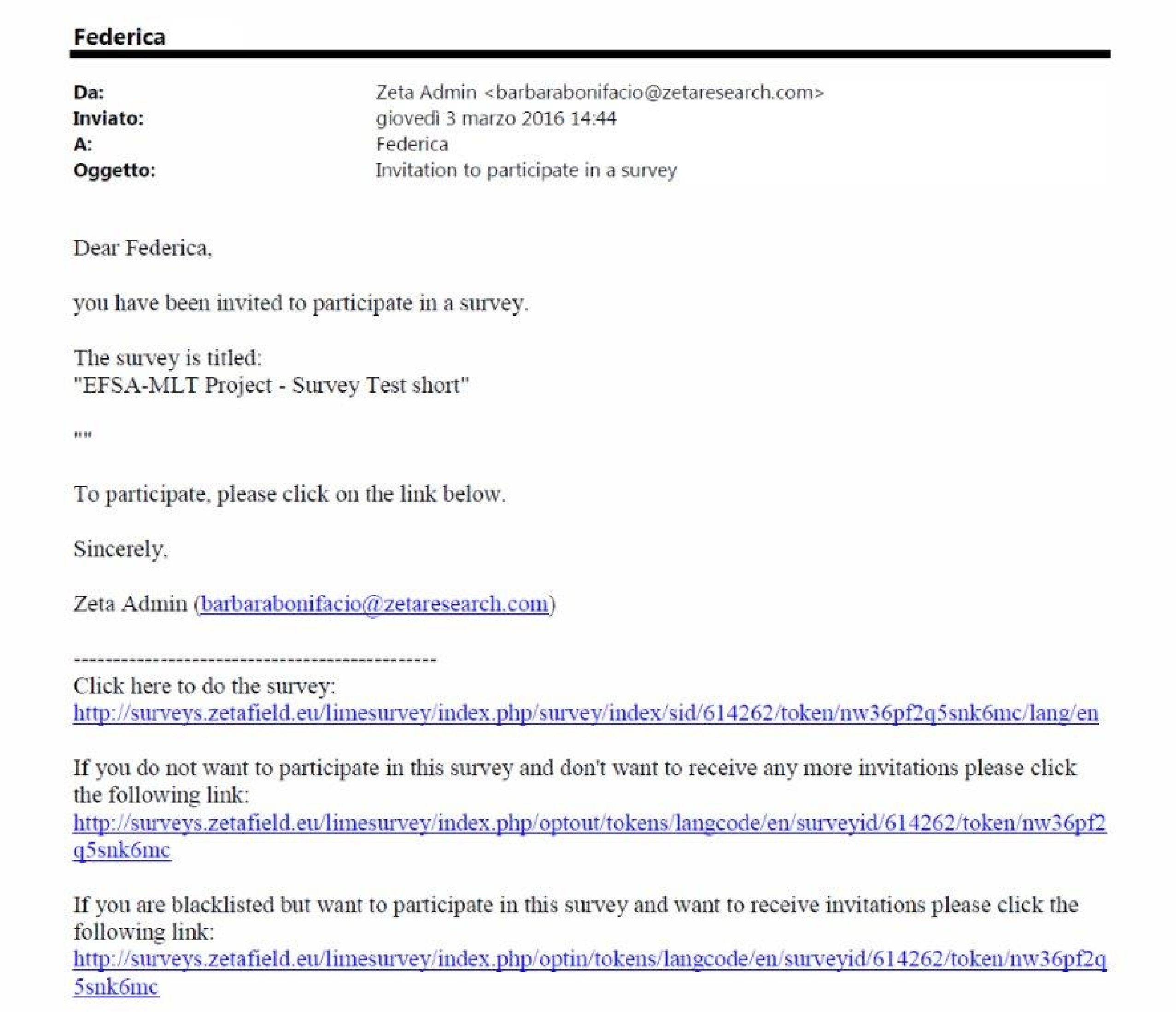
Authored By:

Dejan Kozina		05/03/2016	IT
Typed/Printed Name, Title	Signature	Date	Unit

Approver's Signature:

Approved By:

Barbara Bonifacio		05/03/2016	QA
Typed/Printed Name, Title	Signature	Date	Unit



Federica

Da:	Zeta Admin <barbarabonifacio@zetaresearch.com>
Inviato:	giovedì 3 marzo 2016 14:44
A:	Federica
Oggetto:	Invitation to participate in a survey

Dear Federica,

you have been invited to participate in a survey.

The survey is titled:
"EFSA-MLT Project - Survey Test short"

""

To participate, please click on the link below.

Sincerely,

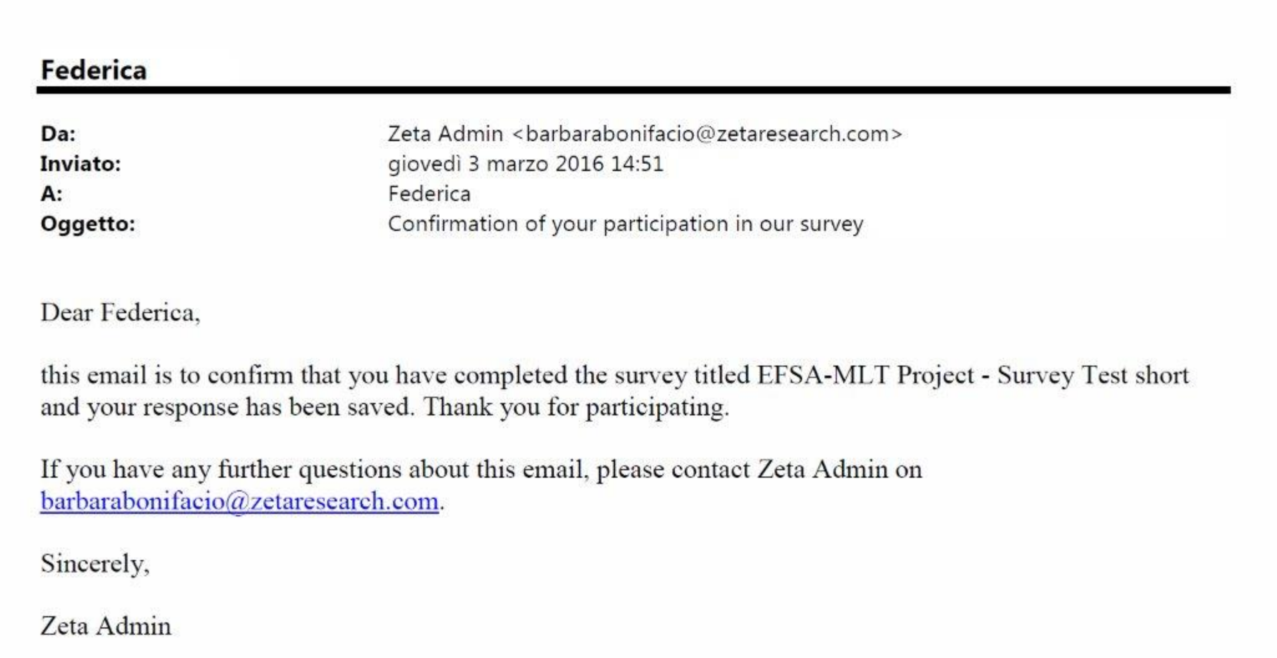
Zeta Admin (barbarabonifacio@zetaresearch.com)

Click here to do the survey:
http://surveys.zetafield.eu/limesurvey/index.php/survey/index/sid/614262/token/nw36pf2q5snk6mc/lang/en

If you do not want to participate in this survey and don't want to receive any more invitations please click the following link:
http://surveys.zetafield.eu/limesurvey/index.php/optout/tokens/langcode/en/surveyid/614262/token/nw36pf2q5snk6mc

If you are blacklisted but want to participate in this survey and want to receive invitations please click the following link:
http://surveys.zetafield.eu/limesurvey/index.php/optin/tokens/langcode/en/surveyid/614262/token/nw36pf2q5snk6mc

Figure 61: Invitation to partecipate in a survey.



Federica

Da: Zeta Admin <barbarabonifacio@zetaresearch.com>
Inviato: giovedì 3 marzo 2016 14:51
A: Federica
Oggetto: Confirmation of your participation in our survey

Dear Federica,

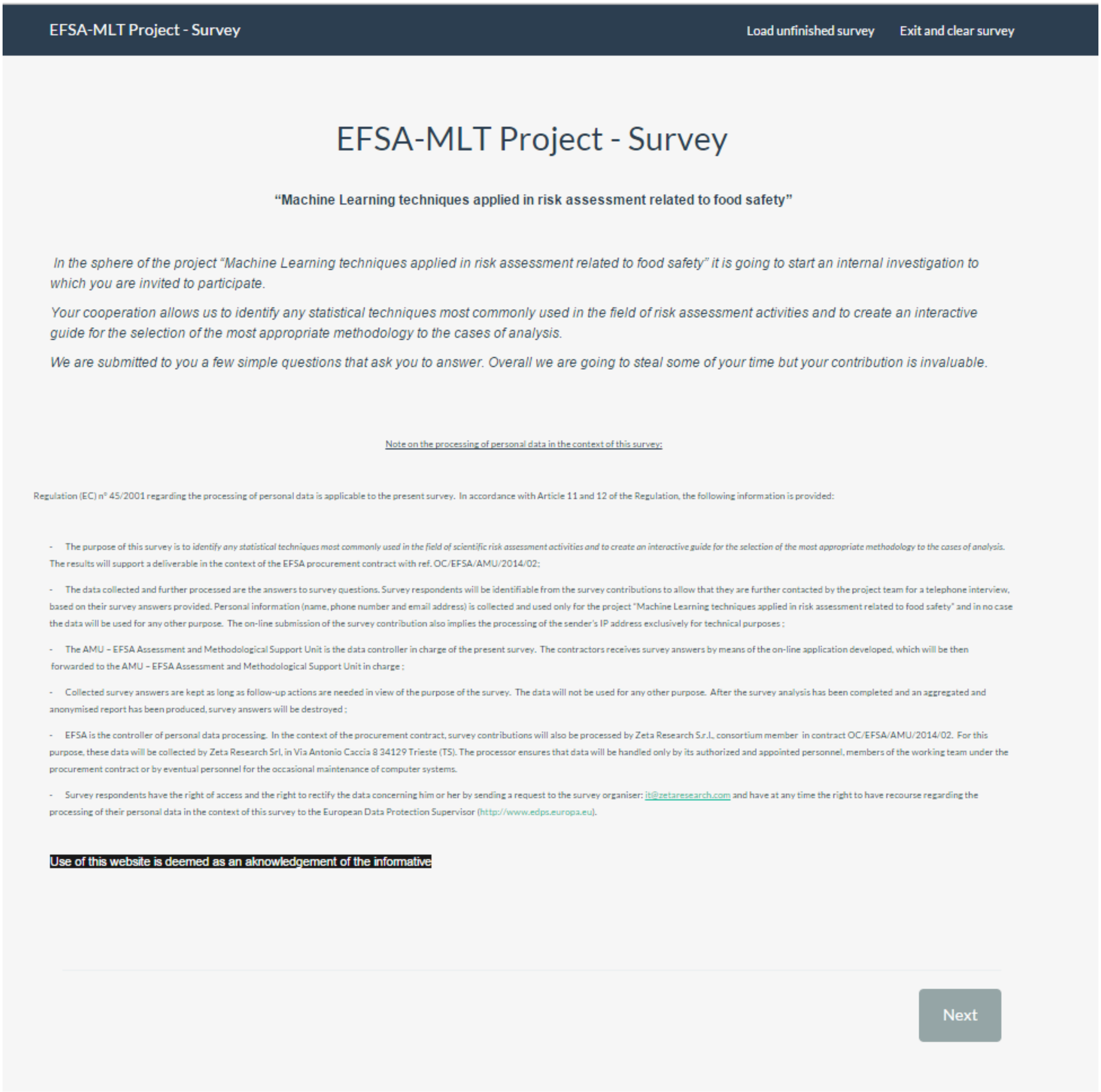
this email is to confirm that you have completed the survey titled EFSA-MLT Project - Survey Test short and your response has been saved. Thank you for participating.

If you have any further questions about this email, please contact Zeta Admin on barbarabonifacio@zetaresearch.com.

Sincerely,

Zeta Admin

Figure 62: Confirmation of your participation in our survey.



EFSA-MLT Project - Survey

"Machine Learning techniques applied in risk assessment related to food safety"

In the sphere of the project "Machine Learning techniques applied in risk assessment related to food safety" it is going to start an internal investigation to which you are invited to participate.

Your cooperation allows us to identify any statistical techniques most commonly used in the field of risk assessment activities and to create an interactive guide for the selection of the most appropriate methodology to the cases of analysis.

We are submitted to you a few simple questions that ask you to answer. Overall we are going to steal some of your time but your contribution is invaluable.

Note on the processing of personal data in the context of this survey:

Regulation (EC) n° 45/2001 regarding the processing of personal data is applicable to the present survey. In accordance with Article 11 and 12 of the Regulation, the following information is provided:

- The purpose of this survey is to *identify any statistical techniques most commonly used in the field of scientific risk assessment activities and to create an interactive guide for the selection of the most appropriate methodology to the cases of analysis.* The results will support a deliverable in the context of the EFSA procurement contract with ref. OC/EFSA/AMU/2014/02;
- The data collected and further processed are the answers to survey questions. Survey respondents will be identifiable from the survey contributions to allow that they are further contacted by the project team for a telephone interview, based on their survey answers provided. Personal information (name, phone number and email address) is collected and used only for the project "Machine Learning techniques applied in risk assessment related to food safety" and in no case the data will be used for any other purpose. The on-line submission of the survey contribution also implies the processing of the sender's IP address exclusively for technical purposes ;
- The AMU – EFSA Assessment and Methodological Support Unit is the data controller in charge of the present survey. The contractors receives survey answers by means of the on-line application developed, which will be then forwarded to the AMU – EFSA Assessment and Methodological Support Unit in charge ;
- Collected survey answers are kept as long as follow-up actions are needed in view of the purpose of the survey. The data will not be used for any other purpose. After the survey analysis has been completed and an aggregated and anonymised report has been produced, survey answers will be destroyed ;
- EFSA is the controller of personal data processing. In the context of the procurement contract, survey contributions will also be processed by Zeta Research S.r.l., consortium member in contract OC/EFSA/AMU/2014/02. For this purpose, these data will be collected by Zeta Research Srl, in Via Antonio Caccia 8 34129 Trieste (TS). The processor ensures that data will be handled only by its authorized and appointed personnel, members of the working team under the procurement contract or by eventual personnel for the occasional maintenance of computer systems.
- Survey respondents have the right of access and the right to rectify the data concerning him or her by sending a request to the survey organiser: it@zetaresearch.com and have at any time the right to have recourse regarding the processing of their personal data in the context of this survey to the European Data Protection Supervisor (http://www.edps.europa.eu).

Use of this website is deemed as an aknowledgement of the informative

Figure 63: Open page.