Development and application of statistical methodology for analysis of the phenomenon of multi-drug resistance in the EU: demonstration of analytical approaches using antimicrobial...

Article *in* EFSA Supporting Publications · September 2016

DOI: 10.2903/sp.efsa.2016.EN-1084

5 authors, including:

Stijn Jaspers
Hasselt University
14 PUBLICATIONS **152** CITATIONS
SEE PROFILE

Chellafe Ensoy
Hasselt University
12 PUBLICATIONS **120** CITATIONS
SEE PROFILE

Christel Faes
Hasselt University
194 PUBLICATIONS **2,177** CITATIONS
SEE PROFILE

Marc Aerts
Hasselt University
334 PUBLICATIONS **5,514** CITATIONS
SEE PROFILE

APPROVED: 25 August 2016 PUBLISHED: 09 September 2016

Development and application of statistical methodology for analysis of the phenomenon of multi-drug resistance in the EU: demonstration of analytical approaches using antimicrobial resistance isolate-based data

CenStat

Jaspers, S., Ganyani, T., Ensoy, C., Faes, C. and Aerts, M.

Abstract

Since antimicrobial resistance (AMR) has been one of the major public health burdens over the last decade, it is of great importance to appropriately monitor and analyse AMR data. Isolate-based data within the EU have been routinely collected since 2010 and reported to EFSA on a yearly basis. AMR data are collected for several bacterial species, tested for susceptibility against different antimicrobials and minimum inhibitory concentration (MIC) is reported. For analysis purposes, a dichotomised version of the MIC values based on the epidemiological cut-off is used to represent different resistance patterns. This report describes various methods to analyse multi-drug resistance data, including the identification of structure to construct groups of isolates with similar resistance patterns or with similar MIC values. Multivariate classification trees and hierarchical cluster analysis after application of principal components and multiple correspondence analyses are applied aiming at group discovering. Latent class analysis is presented as an alternative model-based approach. The generalised estimating equations method is presented handling univariate and multivariate binary outcomes. Bayesian network analysis provides the user with a graphical representation of the underlying associations in the data to identify new co-resistance patterns. Models that deal with spatial distribution of resistant isolates, in combination with their evolution over time, are constructed for univariate and bivariate outcomes. Finally, pattern and source attributions tools are presented, providing, in addition to exploratory analyses, a logistic model to assess variables influencing certain resistance patterns. Source attribution is used to attribute resistance cases in humans to resistance observed in animal, human food consumption patterns and antimicrobial usage data. For illustration purposes, these methods are applied to a subset of the AMR data using an application developed with the R package “shiny”.

Key words: Antimicrobial resistance; Bayesian networks; Classification trees; Generalised estimating equations; Hierarchical clustering; Latent class analysis; Pattern attribution; Source attribution; Spatio-temporal models

Question number: EFSA-Q-2014-00537

Correspondence: amu@efsa.europa.eu

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, 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 author.

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.

Acknowledgements: The authors would like to thank the EFSA staff members: José Cortiñas Abrahantes, Pierre-Alexandre Beloeil, Roberta Palumbo, Marios Georgiadis, Yves van der Stede and Jane Richardson for the support provided to this scientific output.

Suggested citation: Jaspers S, Ganyani T, Ensoy C, Faes C and Aerts M, 2016. Development and application of statistical methodology for analysis of the phenomenon of multi-drug resistance in the EU: demonstration of analytical approaches using antimicrobial resistance isolate-based data. EFSA supporting publication 2016:EN-1084. 54 pp.

Table of contents

1. Introduction

1.1. Background and Terms of Reference as provided by the requestor

In accordance with Decision 2013/652/EC, harmonisation of monitoring of antimicrobial resistance (AMR) in animals and food reporting will be further enhanced in the EU and reporting of AMR data at isolate level by MSs to EFSA will become mandatory from calendar year 2015 and onwards (2014 data and onwards). Based on AMR isolate-based data reported on a voluntary basis, the 2012 EU Summary Report on AMR summarises important information on multi-drug resistance (MDR) and already includes 'summary indicators' of MDR and the breakdown of the multi-/co-resistance patterns recorded. The isolate-based dataset allows the following to be reported: source of the sample (animal species, animal populations or food categories), the date of sampling, the country of origin, the bacterial species and subtype of the isolate tested and the susceptibility test results to a harmonised set of antimicrobial substances.

MDR is considered to be a major public health issue. It is important that EFSA can provide an evidence-based evaluation of the role of food production in the emergence and spread of multiple drug resistant micro-organisms. Further analytical and methodological preparatory work should be performed on the available 2010-2014 isolate-based data in order to have a more in-depth analysis of MDR, notably to investigate associations between resistance traits and to carry out tracing analyses of the geographical and temporal diffusion of MDR. This report aims at providing suitable analysis methods to address these questions and to identify areas for improvement in monitoring systems.

In order to develop appropriate statistical methodology to analyse the phenomenon of antimicrobial resistance, several primary and secondary objectives were addressed. Initially, focus was at investigating possible relationships between MDR patterns. More specifically, the aim was to identify possible groups or clusters while considering that specific combinations of resistance to antimicrobials may co-evolve plus additional resistance traits may be gained or lost. In addition, attention was also paid to the investigation of the spatial distribution of individual MDR patterns and groups of MDR patterns. Next to the spatial distribution, a time component was introduced as well into the models to identify possible evolutions over time. Finally, since antimicrobial resistance is not solely found in animals, the use of source attribution models, which aim at relating resistance in humans to resistance in distinct food sources was explored as well.

The statistical techniques introduced and discussed in this report are all applied to a specific subset of the data. For every method presented, a user-friendly application was developed to allow the later analysis of other subsets of interest and to aid in the analysis of future datasets. This application was created with the "shiny" R package. An accompanying tutorial was prepared as well, which should guide the user in performing the analyses appropriately.

1.2. Additional information

R and SAS were used as statistical software packages:

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. More information and download available at http://www.r-project.org.
- SAS (Statistical Analysis System) is a software suite developed by SAS Institute for advanced analytics, business intelligence, data management, and predictive analytics. More information and trial version download available at http://www.sas.com.

Since some of the analysis techniques could not be performed with SAS, focus was on creating a user-friendly interface with the “Shiny” R package. To guide the user in obtaining the results presented in this report, a tutorial was created, accompanying this report.

2. Data and Methodologies

2.1. Data

Data on antimicrobial resistance were collected yearly from 31 European Union (EU) Member States (MS) for the period 2010-2014. The available dataset was composed of isolate-based information on different bacteria subtypes tested for susceptibility against a common set of antimicrobials (depending on the bacteria of interest). Basic interest was in the outcomes from dilution experiments, from which the minimum inhibitory concentration (MIC) values were collected. Based on the Epidemiological cut-off (ECOFF) values, the MIC values were converted to a binary indicator of resistance (equal to 1 in case MIC>ECOFF). For illustration purposes, focus was on the analysis of these binary values, for one specific combination of bacteria and sample origin, namely indicator *E. coli* isolates collected from broilers. The table below gives an overview of the data at hand. The enhanced monitoring of AMR in bacteria from food and food-producing animals set out in the Commission Implementing Decision 2013/652/EU was successfully implemented in reporting MSs and non-MSs in 2014. Before, reporting in this format was not mandatory, explaining the lower numbers for the period 2010-2013.

Table 1: Overview of the number of isolates (*E. coli*) collected from broilers in different MS between 2010-2014

	Year					
Country	2010	2011	2012	2013	2014	**Total**
Austria	171	173	130	146	174	794
Belgium	0	410	0	232	145	787
Bulgaria	0	0	0	0	85	85
Croatia	0	0	0	0	169	169
Cyprus	0	0	0	0	52	52
Czech Republic	0	0	0	0	195	195
Denmark	0	131	115	125	191	562
Estonia	0	0	0	0	68	68
Finland	0	0	0	0	175	175
France	0	0	0	193	217	410
Germany	200	246	0	434	401	1281
Greece	0	0	0	0	167	167
Hungary	0	0	103	152	165	420
Ireland	0	0	0	0	160	160
Italy	0	0	0	0	403	403
Latvia	0	0	0	0	99	99
Lithuania	0	0	0	0	51	51
Malta	0	0	0	0	32	32
Netherlands	0	0	0	0	377	377
Norway	0	0	0	0	202	202
Poland	0	0	0	0	175	175
Portugal	0	0	0	0	190	190
Romania	0	0	0	0	844	844
Slovakia	0	0	0	0	70	70
Slovenia	0	0	0	0	77	77
Spain	0	101	0	170	145	416
Sweden	0	0	17	0	197	214
Switzerland	183	176	246	236	195	1036
United Kingdom	0	0	0	0	159	159
Total	554	1237	611	1688	5580	9670

In order to illustrate how classification trees could be of used, a dataset composed of 35509 isolates in total of which 17590 *E. coli* isolates and 17919 *Salmonella* isolates was used. These isolates were sampled in 18 Member States from 3 animal types (domestic fowl *Gallus gallus*, cattle, pigs) and 3 types of meat (meat from cattle, meat from pigs and meat from broilers).

2.2. Methodologies

Seven procedures, including multivariate classification trees, clustering methods (based on principal components and multiple correspondence analyses), generalised estimating equations, latent class analysis, spatio-temporal analysis and source attribution models have been applied to the AMR data. The notation used in this report is described below.

2.2.1. General Notation

The Response (MIC value and binary indicator)

For a particular combination of bacteria (sub)type (e.g. *E. coli*) and sample type (e.g. broilers), denote

$$\boldsymbol{z}_i = (z_{i1}, \dots, z_{ip}),$$

the MIC values of p antimicrobials, for isolate $i = 1, \dots, n$, where n is the number of isolates for that particular combination, i.e. the sample size. The MIC distribution of z_{ij} for a particular antimicrobial j can be considered as a mixture of the wild-type left component and the right resistant component, the latter component being typically another mixture distribution. The term wild-type refers to isolates that do not have acquired or mutational resistance mechanisms, while isolates that do have these mechanisms are referred to as resistant. Next to the marginal MIC distribution for one single antimicrobial, one can consider the joint distribution of all antimicrobials involved, or any particular subset of interest.

Using appropriate (harmonised) ECOFFs (epidemiological cut-off values), the MIC values are converted into resistance indicators

$$\boldsymbol{y}_i = (y_{i1}, \dots, y_{ip}),$$

where for $j = 1, \dots, p$,

$$y_{ij} = I(z_{ij} > \kappa_j),$$

With indicator $I(\text{true}) = 1$ and $I(\text{false}) = 0$, and κ_j the ECOFF used to dichotomise the MIC distribution of the j-th antimicrobial. Further denote

$$\pi_{ij} = P(y_{ij} = 1),$$

the probability for isolate i to be microbiologically resistant, i.e. to have reduced susceptibility to antimicrobial j. The probability for a particular multi-resistance pattern can be denoted as follows, for full resistance,

$$\pi_i(1 \dots 1) = P(y_{i1} = 1, \dots, y_{ip} = 1),$$

and full susceptibility

$$\pi_i(0 \dots 0) = P(y_{i1} = 0, \dots, y_{ip} = 0).$$

The general joint probability of interest is

$$\boldsymbol{\pi}_i(y_{i1}^* \dots y_{ip}^*) = P(y_{i1} = y_{i1}^*, \dots, y_{ip} = y_{ip}^*),$$

for all 2^pcombinations of values $y_{i1}^* \in \{0,1\}, \dots, y_{ip}^* \in \{0,1\}$.

The Covariates (explanatory variables)

The outcomes of interest (the responses), being the MIC values or the dichotomised resistance indicators, can be studied on their own, but also related to q covariates (one or more)

$$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq}),$$

including the animal population, the production stage and type, the sampling strategy, the country, etc. Two special "covariates" of interest are the sampling day/month/year (time) and area of sampling (spatial location), as they allow studying the temporal evolution/trends and the spatial relationships.

The main objective is to examine in which way covariates change the distribution of $\boldsymbol{z}_i$ or $\boldsymbol{y}_i$. So, focusing on the binary outcomes, how do the joint probabilities $\boldsymbol{\pi}_i(y_{i1}^* \ldots y_{ip}^*)$ change if particular covariate values are considered. The effect of a covariate on the distribution of the multivariate outcomes $\boldsymbol{z}_i$ or $\boldsymbol{y}_i$ can be studied essentially in two ways: (i) by splitting up the isolates according to the values (categories) of that covariate; (ii) by including the covariate into the statistical model, as a fixed or a random effect. Including the covariate in a model with a saturated fixed effect is essentially splitting up the sample, but option (ii) allows to simplify the model structure and to identify the simplest model that describes the relationships best (using goodness of fit criteria such as Akaike's information criterion, AIC). The effect of more covariates can be studied in the same way, but this might become cumbersome, as too few observations might be available for particular covariate combinations (sparseness).

2.2.2. Classification Trees

Univariate classification trees form a nonparametric, data-driven alternative to the classical logistic regression models (binary as well as multicategory models). Tree-based methods partition the covariate space into subspaces that are homogeneous in the response, in the current case being the resistance status as determined by the resistance indicators. This recursive-partitioning algorithm on which the partition is based, is fully data driven, making the method conceptually simple, yet powerful. It has its merits especially in high-dimensional cases (many covariates of mixed nature). The final constructed tree can be presented in a graphical way, which lends itself for easy interpretation. The estimated tree can be considered as a fit on its own or it can be used to guide a parametric modelling exercise. For further information on classification trees, see Hastie *et al.* (2009).

When studying MDR, interest goes more to multivariate classification trees, i.e. trees that take into account the outcome related to multiple antimicrobials simultaneously. As such, the multivariate classification trees form a similar alternative to the multivariate extensions of logistic regression such as generalised estimating equations (introduced in Section 2.2.4). They explain the variation of a multivariate categorical outcome using covariates that may be numeric and/or categorical. They do this by growing a tree structure that splits the dataset using covariates into non-overlapping clusters, each of which has similar values of the multivariate outcome.

In this report, a class of trees known as ***conditional inference trees*** (Hothorn *et al.*, 2006) is implemented. The method grows a tree by recursively applying a two-step algorithm. Starting with all data, represented by a single node at the top of the tree, the global null hypothesis of association between the multivariate outcome and any of the q covariates is tested. In case this hypothesis cannot be rejected, the algorithm stops; otherwise, the covariate with the strongest association with the multivariate outcome is selected. Once a covariate has been selected, a cut-point is chosen from all its values such that the resulting daughter nodes are as homogeneous as possible in terms of the multivariate outcome. These two steps are re-applied to each of the resulting daughter nodes until the global hypothesis cannot be rejected at a pre-specified nominal significance level α.

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, 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 author.

Benefits and disadvantages

Tree models are very appealing as they are largely data-driven and allow synthetic graphical presentations. It is however known that the resulting trees can be quite variable (from sample to sample from the same population) and are typically not the best predictive models. As for any other approach, sparseness (missing data) and separation issues (Ensoy et. al., 2015) in the data will hamper the performance of tree models.

Software

Analysis in R: ctree(.) function from partykit package
Analysis in SAS: /

2.2.3. Clustering

Kaufman and Rousseeuw (1990) define cluster analysis as the classification of similar objects into groups, where the number of groups, as well as their forms is unknown. The "form of a group" refers to the parameters of a cluster; that is, to its cluster-specific means, variances, and covariances that also have a geometrical interpretation. ***Cluster analysis*** is also called ***data segmentation***. In addition to the grouping or segmenting into subsets or clusters, the goal can be to arrange the clusters into a natural hierarchy, which involves successively grouping the clusters themselves such that at each level of the hierarchy, clusters within the same group are more similar to each other than those in different groups. For more details, see Hastie et al. (2009) and Johnson and Wichern (2002).

In this report, the aim is at detecting clusters after reducing the dimensionality of the data structure and hierarchical clustering will be used to construct clusters, as introduced in Section 2.2.3.3. The pre-processing of data-reduction depends on the nature of the employed data. For the continuous outcomes $\boldsymbol{z}_i$, the data reduction is performed by a principal components analysis, while multiple correspondence analysis is the data reduction tool for the categorical (including binary) values $\boldsymbol{y}_i$.

2.2.3.1 Principal Components Analysis

Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible, under the constraint that it is orthogonal to (i.e. uncorrelated with) the preceding components. More specifically, PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables. It then removes this variance and seeks a second linear combination, which explains the maximum proportion of the remaining variance, and so on. This is called the principal axis method and results in orthogonal (uncorrelated) factors. The resulting principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. Covariates $\boldsymbol{x}_i$ can be included as so-called supplementary variables.

Benefits and disadvantages

Ideally, one can limit oneself to two or at most three components, but depending on the data at hand, it might be necessary to incorporate more, and consequently data reduction might be rather limited.

Furthermore, the "Lowest (limit)" and "Highest (limit)" can vary across isolates (being for instance lab dependent). This affects the range of possible values of z_{ij}, especially the smallest and largest value. Moreover all MIC values need to be considered as rounded, interval-censored data; the smallest being left-censored and the largest being right-censored. The application of PCA on the values of $\boldsymbol{z}_i$

might be hampered by these very characteristics of the data provided by the 'dilution method' (experimental issues).

What the impact is of the rounding in combination with varying ranges of categorisations (as in our setting of MIC values) on principal component and cluster analysis has not been studied yet, to our knowledge. This does not apply to the binary values and the multiple correspondence analysis (introduced in the next section), but, on the other hand, available information about the MIC distribution is dramatically reduced by this dichotomised approach. Therefore, it is recommended to use both scales (ordinal multi-category and binary) and to compare both analyses for similarities and differences. A rigorous in-depth investigation of these issues would be a very interesting and relevant research project, but is beyond the scope of this report.

Software

Analysis in R: PCA(.) function from FactoMineR package
Analysis in SAS: proc princomp

2.2.3.2 Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is an extension of simple correspondence analysis, which allows one to analyse the pattern of relationships of several nominal categorical dependent variables. It can be considered as a generalisation of principal component analysis to *categorical variables*. It is also used to detect underlying structures by a representation in a low-dimensional Euclidean space. Similar to PCA, covariates $\boldsymbol{x}_i$ can be included as so-called supplementary variables.

Benefits and disadvantages

Being dichotomised, the binary values $\boldsymbol{y}_i$ do not suffer from the same complications as the values of $\boldsymbol{z}_i$ in their application of PCA. Binary values contain less information and revealing underlying structures by lower dimensional representations makes MCA to be disadvantageous as compared to PCA. Instead of dichotomisation, more complex categorisation into a higher number of categories is also possible and MCA could benefit from such practice, but lack of harmonisation when using the dilution method might hamper the analysis.

Software

Analysis in R: MCA(.) function from FactoMineR package
Analysis in SAS: proc corresp

2.2.3.3 Hierarchical Clustering

Once the dimensions of the data have been reduced, the construction of the clusters can be initiated. The hierarchical trees considered in this report use Ward's method. This criterion is based on the Huygens theorem, which allows decomposing the total inertia (total variance) in between and within-group inertia (variance). The within-group inertia characterises how homogeneous a cluster is. The total inertia can be decomposed as follows:

$$\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\left(x_{iqk}-\overline{x_k}\right)^2=\sum_{k=1}^{K}\sum_{q=1}^{Q}I_q\left(\overline{x_{qk}}-\overline{x_k}\right)^2+\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\left(x_{iqk}-\overline{x_{qk}}\right)^2$$

$$(\text{Total inertia} \quad = \quad \text{Between inertia} + \text{Within inertia})$$

with x_{iqk} the value of the variable k for the individual i of the cluster q, $\overline{x_{qk}}$ the mean of the variable k for cluster q, $\overline{x_k}$ the overall mean of variable k and I_q the number of individuals in cluster q. Ward's method consists in aggregating two clusters such that the growth of within-inertia is minimum, or, equivalently, minimising the reduction of the between-inertia, at each step of the algorithm. The hierarchical tree is represented by a dendrogram, which is indexed by the gain of within-inertia.

Choosing the number of clusters is a core issue and several approaches have been proposed throughout literature. Some of them rely on the hierarchical tree. Most frequently, one suggests a division into Q clusters when the increase of between-inertia between $Q-1$ and Q clusters is much greater than the one between Q and $Q+1$ clusters. An empirical criterion can formalise this idea. With $\Delta(Q)$ the between-inertia increase when moving from $Q-1$ to Q clusters, the proposed criterion is:

$$\frac{\Delta(Q)}{\Delta(Q+1)},$$

The minimum of which identifies the optimal Q. More information is provided in Husson *et al.* (2010).

Benefits and disadvantages

Hierarchical clustering provides an elegant way to detect subgroups in the data and to visualise them. Nevertheless, the technique is descriptive in nature and does not provide any inferential tools (confidence intervals or hypothesis tests). Not being model-based it also has its limitations to examine the effect of covariates. Moreover, using the results of PCA and MCA, hierarchical clustering shares the same risks and issues.

Software

Analysis in R: HCPC (.) function from FactoMineR package
Analysis in SAS: proc cluster

2.2.4. Generalised Estimating Equations

In a multivariate analysis, the resistance patterns $\boldsymbol{y}_i$ are modelled simultaneously. A generalized linear model consists of the following components:

- The linear component is defined exactly as it is for the traditional linear models, i.e.

$$\eta_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta}$$

- A monotonic differentiable link function $g(.)$, that describes how the expected value of y_{ij}, denoted by μ_{ij}, is related to the linear predictor, $\eta_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta}$:

$$g(\mu_{ij}) = \boldsymbol{x}_{ij}\boldsymbol{\beta}$$

- The response variables y_{ij} are independent for $i = 1,2, \dots$ and have a probability distribution from an exponential family.

Since interest is in binary data, the Bernoulli distribution is used, for which the mean is linked to the linear component using a logit link:

$$logit(\pi_{ij}) = log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \boldsymbol{x}_{ij}\boldsymbol{\beta},$$

which is the well-known logistic regression model.

Of course, it has to be taken into account that observations from the same isolate are not independent. Rather, there could be a correlation between multiple antimicrobials (multi-drug resistance phenomenon). A generalised estimating equations (GEE) approach to estimate the parameters of the generalized linear model with a possible unknown correlation between outcomes can be employed. Through the specification of one of a variety of possible working correlation matrix structures to account for the within-subject correlations, the GEE method estimates model parameters by iteratively solving a system of equations based on quasi-likelihood distributional assumptions. GEE is not a likelihood based model, but a moment method, i.e. only the first and second moment are defined, which correspond to the mean and variance structure only. Therefore, the method only requires the specification of the marginal probabilities $\pi_{ij} = P(y_{ij} = 1)$ and a working assumption on the pairwise correlation of outcomes y_{ij}, y_{ik} of the same isolate i, being

$$\rho_{jk} = P(y_{ij} = 1, y_{ik} = 1) - P(y_{ij} = 1)P(y_{ik} = 1).$$

As there is no reason to assume any homogeneity or simplifying structure in these correlations, the focus in this report is on the unstructured working correlation matrix, which means that the correlations between any two responses are unknown and need to be estimated:

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{19} \\ \rho_{21} & 1 & \rho_{12} & \cdots & \rho_{29} \\ \rho_{31} & \rho_{12} & 1 & \cdots & \rho_{39} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{91} & \rho_{92} & \rho_{93} & \cdots & 1 \end{bmatrix}$$

Population based estimates for the effects of covariates are obtained as well, together with the adjusted standard errors.

Benefits and disadvantages

GEE is a semi-parametric method, requiring only the specification of the marginal probabilities and working assumptions for the intra-isolate correlation, and not requiring the full specification of the joint probabilities $\boldsymbol{\pi}_i(y_{i1}^* \dots y_{ip}^*) = P(y_{i1} = y_{i1}^*, \dots, y_{ip} = y_{ip}^*)$. This is known as its robustness property. This is though at the cost of efficiency (power, accuracy) as compared to fully parametric models. This loss is expected to be rather limited in our setting, because of the moderate correlations and cluster sizes. Hence, the use of full likelihood models would be unnecessarily complicated.

Software

Analysis in R: gee (.) function from gee package
Analysis in SAS: proc genmod

2.2.5. Latent Class Analysis

Latent class analysis is a statistical technique for the analysis of multivariate categorical data. When observed data take the form of a series of categorical responses, it is often of interest to investigate sources of confounding between the observed variables, identify and characterise clusters of similar cases, and approximate the distribution of observations across the many variables of interest. Latent class models are a useful tool for accomplishing these goals.

The latent class model seeks to stratify the cross-classification table of observed, or manifest variables by an unobserved, or latent, unordered categorical variable that eliminates all confounding between the manifest variables. Conditional upon values of this latent variable, responses to all of the manifest variables are assumed to be statistically independent. This assumption is typically referred to as conditional or local independence.

The model probabilistically groups each observation into a latent class, which in turn produces expectations about how that observation will respond on each manifest variable. Although the model does not automatically determine the number of latent classes in a given data set, it does offer a variety of parsimony and goodness of fit statistics that the researcher may use in order to make a theoretically and empirically sound assessment. For example, the user can fit the model while assuming several values for the number of latent classes and select the most optimal one based on the AIC criterion.

Because the unobserved latent variable is nominal (membership of a class), the latent class model is actually a type of finite mixture model. The parameters estimated by the latent class model are the proportion of observations in each latent class, and the probabilities of observing each response to each manifest variable, conditional on the latent class

Benefits and disadvantages

The latent class analysis allows for the detection of underlying latent structures in the data, while accounting for specific covariates. It provides a nice visual representation of these classes, which makes it easy to interpret for the user. A challenge with this approach is that it requires an appropriate model specification. Although rather exceptional, computational problems might occur when a large number of latent classes are assumed.

Software

Analysis in R: lca (.) function from poLCA package
Analysis in SAS: /

2.2.6. Bayesian Network Analysis

Bayesian network (BN) analysis is a form of graphical modelling in which the user attempts to find structure in the dataset by separating out indirect effects from direct associations. The basic objective is to perform a model search on the data to identify an optimal model.

An additive BN model for categorical data can be constructed by considering each individual variable as a logistic regression of the other variables in the data and hence the network model is composed of many combinations of local logistic regressions. In these models, the log marginal likelihood, or network score, is estimated using Laplace approximations at each node.

The key objective of the "abn" R package is to enable estimation of statistical dependencies in the data, which comprises multiple variables. In other words, the goal is to identify a Directed Acyclic Graph (DAG) which is robust and representative of the dependency structure of the stochastic system that generated the observed data. This search is performed in two big steps:

- In a first step, the aim is to identify the most probable DAG based on the observed data. In this respect, a large number of heuristic searches are run, for which different graphs are constructed and ranked according to their network scores. The term heuristic refers to a technique in which a, possibly approximate, solution is found in a reasonable time frame. As such, the solution might not be exactly optimal, but still valuable since finding it does not require a prohibitively long time.
- In order to avoid overfitting the data, the DAG found in step 1 is pruned. This trimming is performed using a parametric bootstrap analysis. Bootstrap datasets are generated, i.e. independent realisations from the model which can be used to generate a dataset of the same size of the observed dataset. Given this bootstrap data, the BN model search is repeated, treating the bootstrap datasets as the observed data. By generating many bootstrap datasets and conducting searches on each of them, this allows estimating the percentage support for each arc in the DAG of the highest scoring model. Arcs that are only present in less than half of the constructed DAGS (i.e. level of support <50%) are removed from the final DAG.

Benefits and disadvantages

The Bayesian network analysis can detect structures and patterns in complicated data settings. In the AMR situation, it can be very convenient to detect specific co-resistance patterns among the antimicrobials of interest. A disadvantage of the approach is the limited amount of inference that is possible. Nevertheless, simple association measures are computed to enable ease interpretation for the user.

Software

Analysis in R: "abn" package
Analysis in SAS: /

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, 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 author.

2.2.7. Spatio-Temporal Models

When data is collected across space (i.e. different countries) and possibly over time (i.e. different years), analysis should take into account the spatial and/or temporal dependence of the observations. The linear component of the spatio-temporal model for the binary data for a specific antimicrobial (isolate i, time t, location l) can be written as:

$$logit(\pi_{itl}) = log\left(\frac{\pi_{itl}}{1-\pi_{itl}}\right) = \beta_{\mathbf{0}} + r_t + s_l + u_{tl},$$

where r_t is the temporal effect, s_l is the location/spatial effect and u_{tl} is the spatio-temporal interaction term.

For the temporal effect, different choices can be made, depending on the data. Here, the following options are investigated: no time effect, a saturated time effect (time is treated as a factor), a linear time effect, a first-order random walk (RW1), a first-order autoregressive (AR1), and a second-order random walk (RW2). RW1, AR1 and RW2 are flexible smooth functions of time which assumes that the present observation is a function of the immediate past. Specifically,

- RW1: $r_t = r_{t-1} + \omega_t,$
- AR1: $r_t = \rho * r_{t-1} + \omega_t$
- RW2: $r_t = 2r_{t-1} - r_{t-2} + \omega_t,$

where ρ is a correlation parameter and $\omega_t \sim \mathrm{N}(0, \sigma_\omega^2)$. RW1 assumes that the current observation is equal to the immediate past observation whereas AR1 assumes that it is correlated to the immediate past. RW2, on the other hand, assumes a linear trend and penalizes for deviation from linearity.

The Besag, York and Mollie's (BYM) model was fitted to the spatial effect (s_l). The BYM model takes into account not only the spatial auto-correlation present in the data (structured spatial effect ($\boldsymbol{b}$)) but also assumes that the estimates obtained between areas are independent of each other (IID or unstructured effect ($\boldsymbol{c}$)). The spatial effect of the BYM model is an intrinsic Gaussian Markov random field (GMRF) model, also referred to as Besag model, which assumes that the expected value of each area depends on the values of the neighbouring areas (in this case, areas sharing boundaries). Thus, areas close together are more similar than areas that are far apart. In this application, it was assumed that the structured and unstructured effects are not independent of each other (Riebler et al, 2016). Thus, instead of the usual $\boldsymbol{s} = \boldsymbol{b} + \boldsymbol{c}$, here $\boldsymbol{s} = \frac{\mathbf{1}}{\sqrt{\tau_s}}\left(\sqrt{1-\phi}\boldsymbol{c} + \sqrt{\phi}\boldsymbol{b}\right)$. The model reduces to pure oversdispersion (unstructured) for $\phi = 0$ and to the ICAR/Besag model when $\phi = 1$. The marginal variance is $\sigma_s^2 = {\tau_s}^{-1}$, while ϕ is the proportion of the marginal variance explained by the spatial effect $\boldsymbol{b}$.

The spatio-temporal interaction u_{tl} models the relationship between the temporal and spatial trend. In the univariate model, different types of interaction were investigated: unstructured (type I), structured over time but unstructured over space (type II), and structured over time and space (type IV). While in the bivariate analysis, only the unstructured space-time interaction was used which was further assumed to be correlated between the two antimicrobials.

Weighting and incorporation of inter-country trade information in the model was also investigated at country-level. In this application, the weight was defined as the proportion of planned versus actual sample. For the inter-country trade information, this covariate is entered into the model as:

$$\mathrm{logit}(\pi_{itl}) = \beta_{\mathbf{0}} + \beta_{\mathbf{1}} \sum_{m=1}^{M} (a_{l,m} p_{m,t-1}) + r_t + s_l + u_{tl}$$

where l refers to the destination country (trade destination or the importing country), $m = 1, \ldots, M$ refers to the source country (trade origin), $a_{l,m}$ is the trade quantity (in tons) from country m to country l and $p_{m,t-1}$ is the proportion of resistant isolate in the source country at the previous time point. For this application, incorporation of trade information is only at the country- and year-level and here, the total trade times resistance ($\sum_{m=1}^{M}(a_{l,m}p_{m,t-1})$) information was log-transformed. Trade data can be downloaded from the Eurostat database website (http://ec.europa.eu/eurostat/data/database) under the international trade, EU trade since 1999 by HS2,4,6 and CN8 (DS-575274).

Since this is a Bayesian model, priors for the different hyperparameters had to be specified. For the precision of the flexible temporal effects ($1/\sigma_{\omega}^2$), spatial effect ($1/\sigma_s^2$), and spatio-temporal interaction ($1/\sigma_u^2$) different priors were used. Specifically, a gamma(1,0.01) was used for RW1, gamma(0.1,0.001) was used for AR1, gamma(1,0.001) was used for the IID precision parameter, and penalised complexity (PC) prior (Simpson et al., 2015) were used for both the RW2 and spatial effect precision. In all random effects (temporal, spatial and spatio-temporal), a sum-to-zero constraint was specified in order to avoid unidentifiability issues, especially with the intercept. Fitting of this model was done using the integrated nested laplace approximation by Rue et al. (2009).

Comparison of the different spatio-temporal models is done mainly using the deviance information criteria (DIC). DIC is a measure used to compare the fit of different Bayesian models. The model with smaller DIC is generally preferred. The mean deviance and effective number of parameters, which are also reported, is used to compute DIC. Other model-fit statistics are also reported here: logarithmic score and McFadden's R-squared. Logarithmic score is a proper scoring rule used to compare predictions. Models with high logarithmic score are generally preferred. McFadden's pseudo R-squared is the ratio of the deviance of the model being evaluated and the deviance of the null model. It is also used to compare model-fit. However, this should not be confused with the R-squared from linear regression as this is just a pseudo-R-squared and does not have the inherent interpretation of the linear regression R-squared.

Benefits and disadvantages

In order for the spatio-temporal model to be useful, a good temporal and spatial resolution is needed. A good spatial and temporal resolution would depend on the objective/question being asked. For instance, a country-level spatial resolution is not useful if the interest is on the spatial pattern of AMR in 3 countries, data at NUTS-2 level is more informative. Also, for studying the temporal trend, more flexible models can be investigated (and more insight can be gleamed) with monthly data as compared to yearly data. In the case where there is sparse data (no data in some areas), the method can suffer greatly. Furthermore, although the much faster integrated nested laplace approximation (INLA) is used instead of the MCMC estimation using WinBUGS or BRugs, computation can still take long, especially when many areas are involved in the analysis.

Software

Analysis in R: inla(.) function from INLA package (www.r-inla.org).
Analysis in SAS: /

2.2.8. Pattern Attribution Models

Pattern attribution models are constructed to investigate certain resistance patterns $\boldsymbol{y}_i$ into more detail. More specifically, interest is in how certain covariates influence the probability to observe a specific resistance pattern. A resistance pattern is defined as an array with length equal to the number of antimicrobials under investigation. In case an isolate shows resistance for an antimicrobial, the corresponding value in the array equals 1 (or R). Otherwise, when the isolate shows susceptibility against the antimicrobial, the corresponding entry is a 0 (or a blank). Next, after the user has specified a specific pattern of interest, an indicator variable is created which has the value 1 when the isolate shows the entire pattern, and a 0 otherwise. Firth-logistic regression is applied to investigate the effect of certain variables. More specifically, Firth (1993) suggested a correction to the standard logistic regression approach to render more appropriate standard errors in case of separation issues in the data.

Benefits and disadvantages

The exploratory graphs that are constructed provide a fast and nice overview of the probability to observe the selected resistance pattern. A more formal analysis is provided using the Firth model. The analysis is however descriptive in nature, and cannot be used for predicting evolutions in the future.

Software

Analysis in R: logistf(.) function from logistf package
Analysis in SAS: /

2.2.9. Source Attribution Models

In the source attribution section, interest is on understanding the contribution of different food-types to antimicrobial resistance in humans. It is assumed that antimicrobial resistance in humans depends on antimicrobial resistance in different food-types which they consume. The principle is to compare, at the country level, the proportion of human isolates resistant to a given antimicrobial with the proportion of food-type isolates resistant to that given antimicrobial as well as consumption of those food-types.

Firth-logistic regression is applied to model the probability (π) to be resistant to an antimicrobial as follows,

$$logit(\pi_c) = \beta_0 + \sum\beta_k X_{kc} + \sum\alpha_k Y_{kc} + \sum\gamma_k \boldsymbol{Z}_{kc}$$

where X_{kc} denotes proportion of the k^{th} food-type isolates, in country c which are resistant, Y_{kc} denotes consumption quantity of the k^{th} food-type, in country c and $\boldsymbol{Z}_{kc}$ denotes covariate(s) on antimicrobial usage in humans. Missing values for X_{kc}, Y_{kc} and $\boldsymbol{Z}_{kc}$ can be imputed using multiple imputation methods, but in order to perform such analysis, a dataset as complete as possible is strongly recommended. In the case of usage of imputation methods, analysis results should be interpreted with caution.

Benefits and disadvantages

The analysis offers a simple and quick way to understand, on average, which factors contribute to antimicrobial resistance in humans across all member states. However, the results must not be over interpreted since they are not country specific; the overall picture across all member states by no means reflects the situation within individual member states. In order to make valid inferences out of this analysis, data inputs should be as complete as possible.

Software

Analysis in R: logistf(.) function from logistf package. Multiple imputation packages: mice(.) from mice package, amelia(.) from amelia package, aregImpute(.) from Hmisc and rfImpute(.) from randomForest package.
Analysis in SAS: /

2.3. Overview and Summary of Statistical Methods

Table 2: Characteristics of statistical methods and models

	Method		Objective		Data-Driven		Benefits and Disadvantages	
	Name	Type	Primary	Secondary	Level	Use	Strength	Weakness
2.2.2	Classification Trees	Descriptive	Effect of covariate(s)	Detection of clusters, structures and patterns	Largely	No explicit model specification	Can handle complex and high-dimensional data	Variable, instable, no inference
2.2.3.1	Principal Component Analysis	Descriptive	Data reduction	Use in other methods and models	Largely	No explicit model specification	Reduction of dimension as preparatory step	No inference
2.2.3.2	Multiple Correspondence Analysis	Descriptive	Data reduction	Use in other methods and models	Largely	No explicit model specification	Reduction of dimension as preparatory step	No inference
2.2.3.3	Hierarchical Clustering	Descriptive	Detection of clusters	Effect of Covariates	Largely	No explicit model specification	Can detect clusters in complex data structures	No inference
2.2.4	Generalized Estimating Equations	Inferential	Effect of covariates Time trend	Detection of clusters	Partly	Needs model specification	Can model multivariate binary indicators as function of covariates	Needs model specification, computational problems
2.2.5	Latent Class Analysis	Inferential	Detection of Clusters Time trend	Effect of other covariates	Partly	Needs model specification	Can detect underlying latent structures while accounting for covariates	Needs model specification, computational problems
2.2.6	Bayesian Network Analysis	Descriptive	Detection of structures and patterns	Time trend	Largely	No explicit model specification	Can detect structures and patterns in complicated settings	No inference
2.2.7	Spatio-Temporal Models	Inferential	Effect of time and space	Effect of other covariates	Partly	Needs model specification	Can detect spatio-temporal patterns and trends and accommodate additional covariates.	Needs model specification, computational problems
2.2.8	Pattern Attribution Models	Descriptive	Effect of covariates on specific patterns	-	Partly	User selects pattern of interest	Can detect which patterns are mainly associated with certain values of covariates	-
2.2.9	Source Attribution Models	Descriptive	Contribution of food types to human AMR	-	Partly	Needs model specification	Can attribute human antimicrobial resistance to food types they consume	Requires different data sources such as human resistance, consumption, etc.

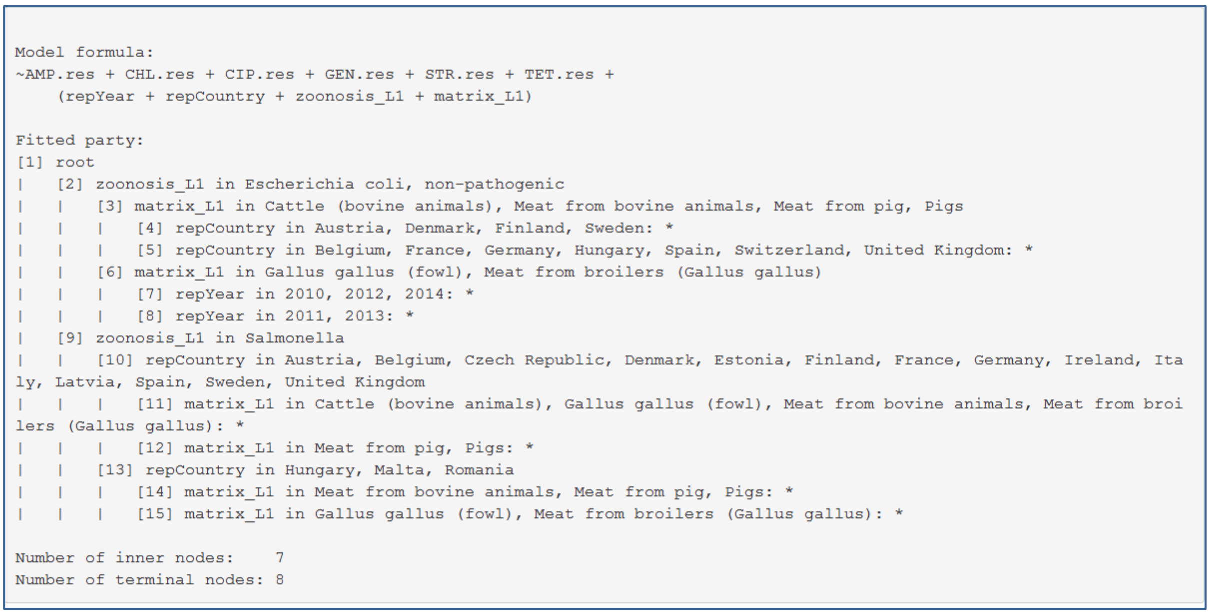
The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, awarded following a tender procedure. The present document is published 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 author.

3. Illustrative applications of the methods

In this section, the most important output of the discussed methods is presented, such that users can understand the use of each method for AMR analysis. All methods except classification trees are applied to the *E. coli* dataset obtained from broilers. The accompanying tutorial can guide the user to obtain these results using the shiny application (for more information, see appendix A).

3.1.1. Classification Trees

Figure 1 shows a tree structure obtained when Salmonella and E. coli data were collapsed over all reporting years (2010 - 2014), all reporting countries as well as the following sample types: cattle (bovine animals), Gallus gallus (fowl), and pigs, as well as meat from bovine animals, meat from broilers (Gallus gallus), meat from pig; the multivariate outcome was composed of binary outcomes for resistance to six antimicrobials namely, AMP, CHL, CIP, GEN, STR and TET. A graphical presentation of the tree is shown on Figure 2.



```
Model formula:
~AMP.res + CHL.res + CIP.res + GEN.res + STR.res + TET.res +
    (repYear + repCountry + zoonosis_L1 + matrix_L1)

Fitted party:
[1] root
|   [2] zoonosis_L1 in Escherichia coli, non-pathogenic
|   |   [3] matrix_L1 in Cattle (bovine animals), Meat from bovine animals, Meat from pig, Pigs
|   |   |   [4] repCountry in Austria, Denmark, Finland, Sweden: *
|   |   |   [5] repCountry in Belgium, France, Germany, Hungary, Spain, Switzerland, United Kingdom: *
|   |   [6] matrix_L1 in Gallus gallus (fowl), Meat from broilers (Gallus gallus)
|   |   |   [7] repYear in 2010, 2012, 2014: *
|   |   |   [8] repYear in 2011, 2013: *
|   [9] zoonosis_L1 in Salmonella
|   |   [10] repCountry in Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland, Ita
ly, Latvia, Spain, Sweden, United Kingdom
|   |   |   [11] matrix_L1 in Cattle (bovine animals), Gallus gallus (fowl), Meat from bovine animals, Meat from broi
lers (Gallus gallus): *
|   |   |   [12] matrix_L1 in Meat from pig, Pigs: *
|   |   [13] repCountry in Hungary, Malta, Romania
|   |   |   [14] matrix_L1 in Meat from bovine animals, Meat from pig, Pigs: *
|   |   |   [15] matrix_L1 in Gallus gallus (fowl), Meat from broilers (Gallus gallus): *

Number of inner nodes:    7
Number of terminal nodes: 8
```

Figure 1: Classification tree summary

The classification tree includes eight terminal nodes (clusters) with splits based on the bacterial type (zoonosis_L1), the food type (matrix_L1), the reporting country (repCountry) as well as the reporting year (repYear). At each stage of splitting, association between each covariate and the six antimicrobials jointly is tested, one covariate at a time.

The bacterial type determines the first split meaning that, for the root node (original data), it has the strongest association (smallest p-value) with resistance to the six antimicrobials jointly, compared to the other covariates.

Similarly, at the second generation split,

i. for Salmonella, reporting country has the strongest association. At the third generation split, both country nodes are further split based on food type meaning that it has the strongest association in either nodes;

ii. for E. Coli , food type has the strongest association. At the third generation split, one food-type node is further split based on reporting country meaning that, for that node it has the strongest association; the other food-type node is further split based on reporting year meaning that it has the strongest association.

Inspection of the bar plots at each node in Figure 2 shows resistance, expressed as proportions (0 to 1), of that cluster to the six antimicrobials jointly; the codes '0' and '1' on each bar stand for 'non-resistant' and 'resistant', respectively. As an example, consider node 15, from isolates observed in Hungary, Malta and Romania. It is found that Salmonella from broilers of domestic fowl (Gallus gallus) and meat from broilers has high resistance to CIP, STR and TET, jointly. Other clusters can be interpreted in a similar way.

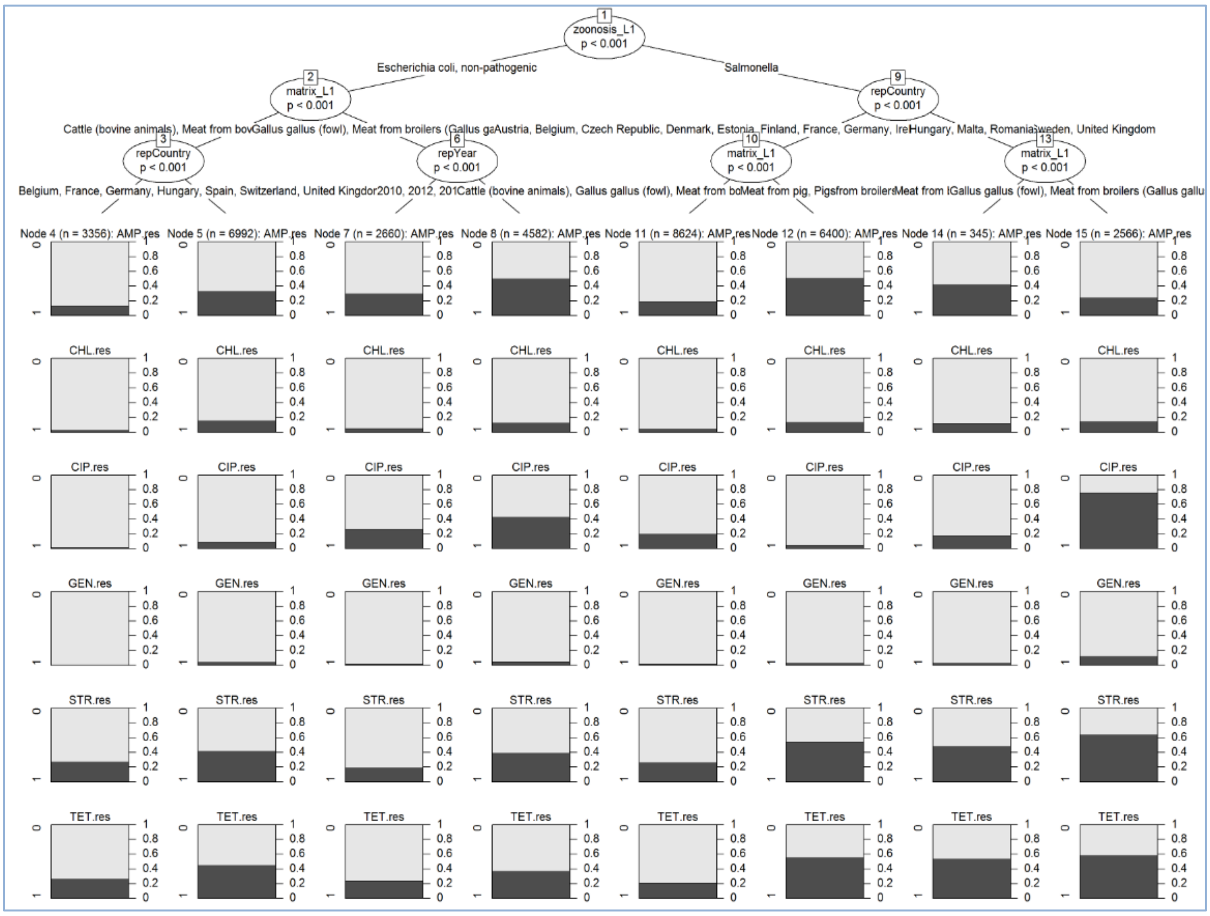


Figure 2: Graphical presentation of the classification tree

3.1.2. Clustering

Principal Components Analysis

The main goal of this PCA analysis is to reduce the number of variables of interest (data reduction). Indeed, 7 antimicrobials of interest are considered. Simply speaking, PCA considers combinations of these AMs, which can be subsequently used in additional analyses like the hierarchical clustering that will be discussed below. In this respect, one of the most important outputs from the PCA is presented on Figure 3. More specifically, the correlation circles show the influence that each of the original variables has on the constructed principal components (PC). The first dimension (principal component), explaining around 37% of the original variability, receives positive contributions from all original variables. This means that a higher MIC value for a certain antimicrobial results into a higher value for the first PC. Similarly, it can be observed that the second dimension (PC), explaining 14% of the total variability in the data, is mainly characterised by GEN and in a lesser extent, exhibits a contrast between CIP and CHL vs. TET, AMP and TMP.

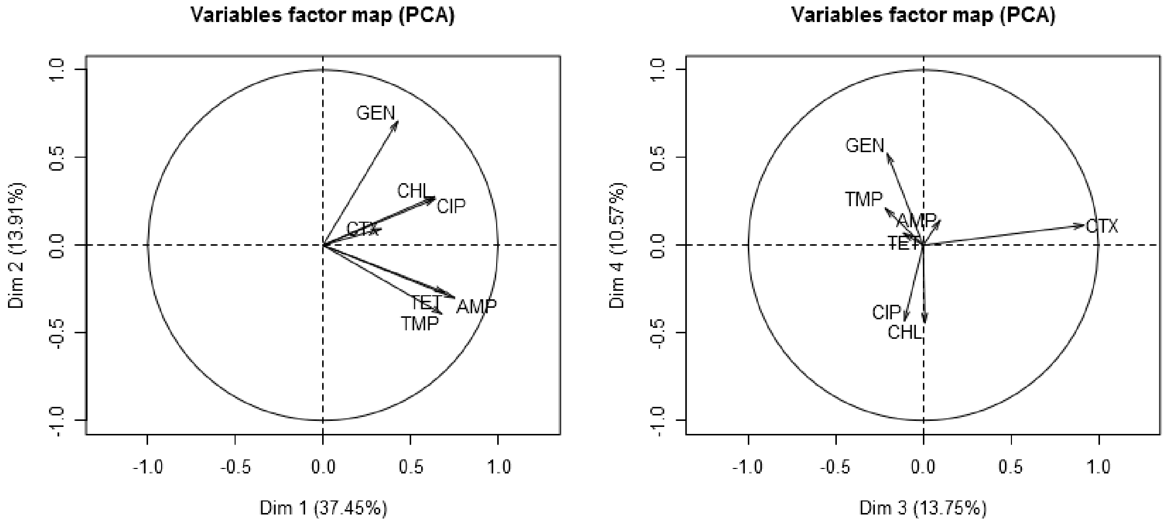
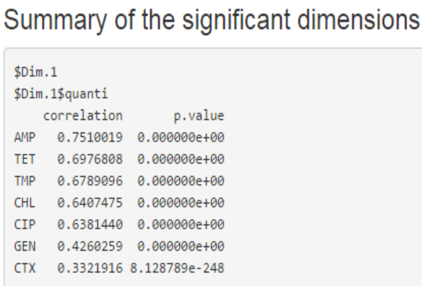


Figure 3: Variables factor map (correlations circle) in the first four dimensions

A more quantitative description of the associations is provided in Table 3. It is observed that all antimicrobials are significantly and positively correlated to the first dimension. Dimension 2 is also significantly correlated with all antimicrobials, but it is a contrast between two groups of antimicrobials. Finally, dimension three, explaining nearly 14% of variability, is significantly correlated with all AMs, except for CHL. This third dimension is a contrast between AMP and CTX on the one hand versus CIP, GEN, TET and TMP on the other. The correlation with the first group of antimicrobials is positive, meaning that a higher value for the third dimension corresponds to higher MIC values for that specific group, given that the MIC in the other group remains unchanged. None of the antimicrobials were significantly correlated with the fourth dimension. Next to the table, the corresponding output from the Shiny application is shown for the first dimension.

Table 3: Description of the first 3 dimensions/components of PCA: correlations between variables and dimensions/components

	Dimensions		
Antimicrobial	**Dim1**	**Dim2**	**Dim3**
AMP	***0.75 (<0.0001)***	-0.30 (<0.0001)	0.10 (<0.0001)
TET	***0.69 (<0.0001)***	-0.27 (<0.0001)	-0.11 (<0.0001)
TMP	***0.68 (<0.0001)***	-0.39 (<0.0001)	-0.21 (<0.0001)
CHL	***0.64 (<0.0001)***	0.26 (<0.0001)	
CIP	***0.64 (<0.0001)***	0.28 (<0.0001)	-0.11 (<0.0001)
GEN	0.43 (<0.0001)	***0.71 (<0.0001)***	-0.21 (<0.0001)
CTX	0.33 (<0.0001)	0.09 (<0.0001)	***0.92 (<0.0001)***



Summary of the significant dimensions

```
$Dim.1
$Dim.1$quanti
    correlation      p.value
AMP   0.7510019  0.000000e+00
TET   0.6976808  0.000000e+00
TMP   0.6789096  0.000000e+00
CHL   0.6407475  0.000000e+00
CIP   0.6381440  0.000000e+00
GEN   0.4260259  0.000000e+00
CTX   0.3321916 8.128789e-248
```

As mentioned in Section 2.2.3.1, the PCA can be extended with supplementary variables to check the influence of these variables on the MIC values of the isolates under investigation. Figure 4 shows the resulting factor map when including the year of monitoring (time). Due to the fact that reporting was only mandatory from 2014 onwards, not all countries provided data in all years. Therefore, this analysis is exemplary and should not be over interpreted.

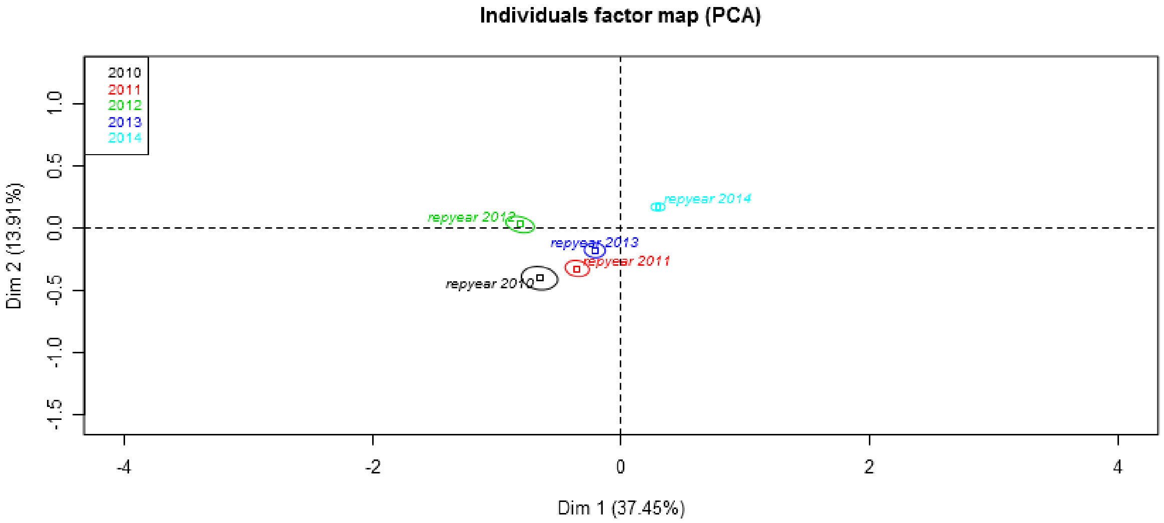


Figure 4: Individuals factor map (dimensions 1 and 2) with time as supplementary variable

Interpretation of Figure 4 above should be done with great care. The plot is mainly descriptive and one should not rush into statistical conclusions. It is seen that isolates from the year 2014 have a higher value for the first dimension. This means that isolates obtained in this year probably have higher MIC values compared to the other years. The lowest MIC values are probably located in 2010 and 2012, since the isolates from these years are located near the lower values of PC1. While the squares on the plot provide an idea of the mean value on both PC1 and PC2 for an isolate in the respective year, the ellipsoids give an indication of the associated variability around that mean. In case they do not overlap for two given years, this means that there is a significant difference between those years (considering the first two dimensions jointly).

Figure 5 shows the results of PCA with country as the supplementary variable. Notice, in particular, the location of the isolates from Bulgaria and Romania, who have the highest values for both dimension 1 and 2. These isolates were only sampled in the year 2014, so the effects of country and year might be blurred by each other in this case. It is very likely that these cases are very influential to the plot in Figure 4 as well. This is again an indication that one should be careful when interpreting these individuals' factor maps. Ideally, data for all countries over all years should be available. A goal that will be achieved as data collection in the current format is mandatory since the 2014 data.

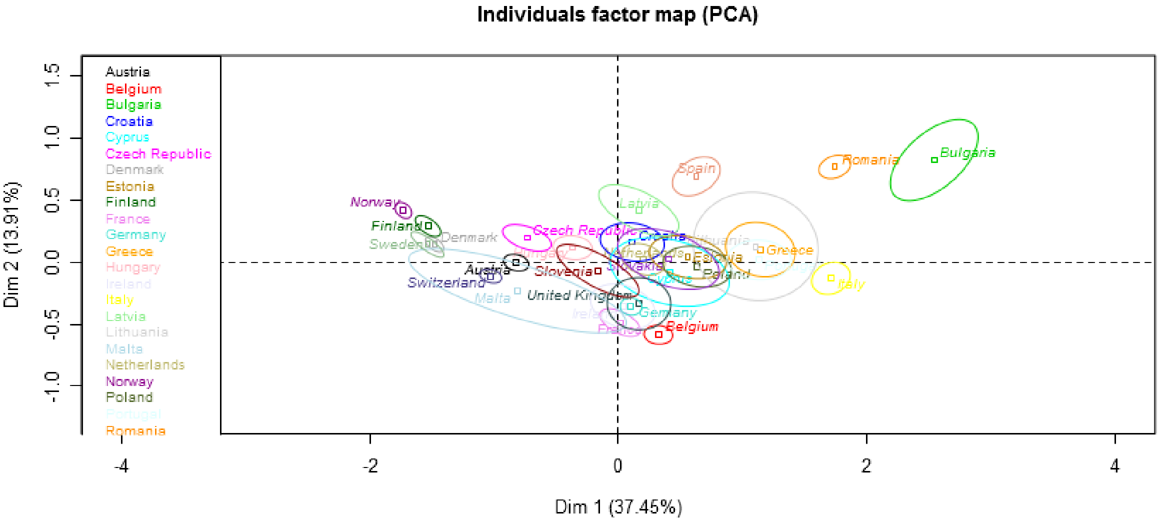


Figure 5: Individuals factor map (dimensions 1 and 2) with country as supplementary variable

Hierarchical clustering after PCA

Using the obtained principal components in the hierarchical clustering analysis, three relatively well separated clusters are obtained, containing respectively 4912, 4019 and 739 isolates. The resulting clusters are presented in Figure 6, where they are plotted on dimension 1 vs. dimension 2 and dimension 1 vs. dimension 3.

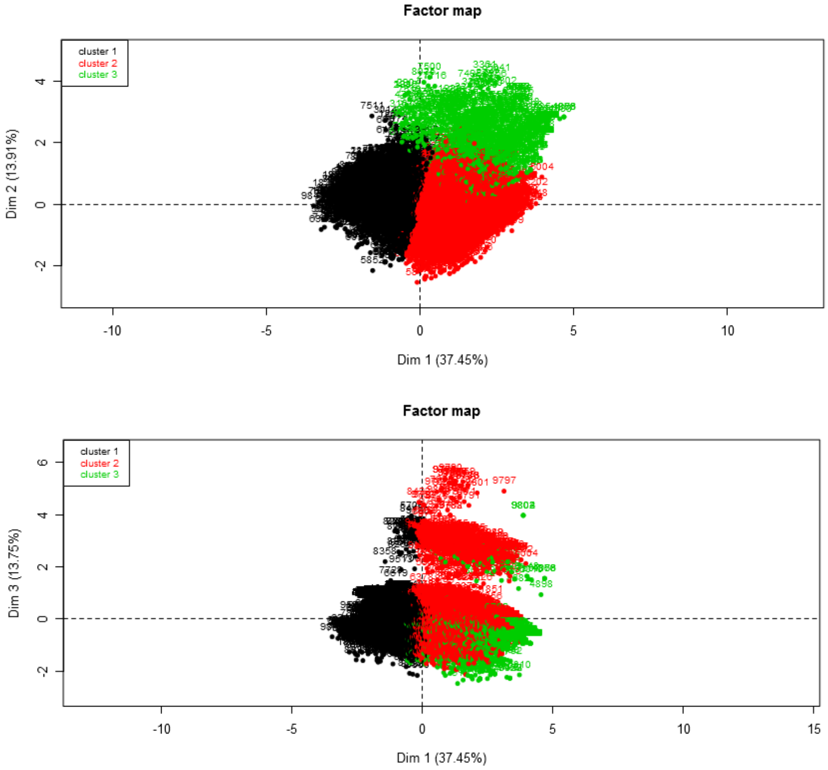


Figure 6: Hierarchical clustering analysis after PCA (top: dim1 vs. dim2; bottom: dim1 vs. dim3)

In addition, Table 4 shows the mean MIC structure of the resulting clusters. The first cluster is located near the lowest values of the first principal component. As a result, the isolates located in this first cluster have the lowest (or close to the lowest) mean MIC values, thereby suggesting that only a small proportion of isolates in this cluster show resistance. Indeed, this can be observed from Table 5, which summarises the cluster-specific proportions of resistance. Cluster 2 has higher values for PC1 compared to cluster 1. This is reflected in the increased mean MIC values for all antimicrobials but GEN. Cluster three has high values for both PC1 and PC2. It therefore has the highest mean MIC values for GEN, CHL and CIP (note the positive correlation with these AMs and PC 2 in Table 3).

Table 4: Mean MIC values (on the log-scale) for each of the clusters

Cluster	N	GEN	CHL	CTX	CIP	AMP	TET	TMP
1	4912	-0.27	-0.43	-0.28	-0.45	-0.71	-0.641	-0.63
2	4019	-0.23	0.36	0.30	0.38	0.76	0.67	0.67
3	739	3.07	0.91	0.18	0.92	0.61	0.64	0.56

Table 5: Cluster-specific proportions of resistance

Cluster	N	GEN	CHL	CTX	CIP	AMP	TET	TMP
1	4912	0.28	0.25	0.06	0.50	0.38	0.31	0.31
2	4019	0.12	0.45	0.33	0.55	0.73	0.55	0.57
3	739	1.00	0.45	0.28	0.57	0.65	0.55	0.50

Multiple Correspondence Analysis

While PCA focused on the continuous MIC values, MCA uses the binary patterns instead. Resistance to CIP and NAL was addressed simultaneously through a newly constructed indicator, CIPNAL, which takes the value 1 in case the isolate shows resistance to either CIP or NAL (or both). In Figure 7, the factor maps are shown for the first 4 dimensions.

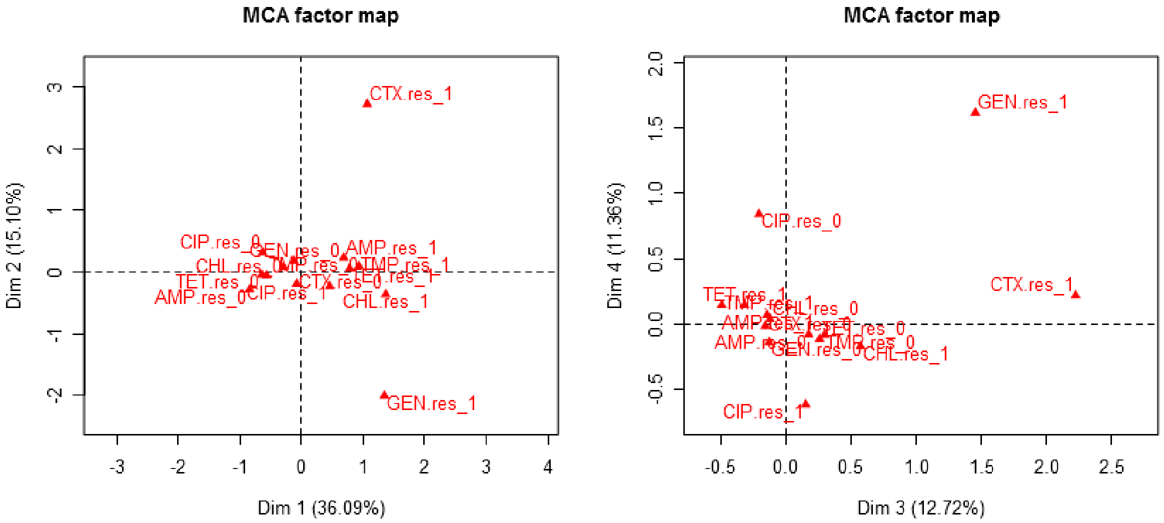


Figure 7: Factor map resulting from MCA

From the factor map, it is observed that higher values for the first dimension are related to resistance against CHL, CTX, GEN, AMP and TMP. Similarly, the highest contribution to dimension 2 is made by isolates resistant to CTX in contrast to GEN. A more quantitative description of the composition of the dimensions is provided in Table 6. The first dimension is composed of positive contributions of the isolates that show resistance against all antimicrobials. Hence, the higher the value of the first component, the more resistance the isolates exhibit. For the second dimension, resistance to the AMs CTX, AMP, TMP and TET contribute in a positive way, whereas susceptibility to the remaining AMs also contributes positively. For the third dimension, it is seen that resistance to CTX, GEN, CHL and CIPNAL (combination of ciprofloxacine and nalidixic acid) contributes positively to this third dimension as is the case for susceptibility against TMP, AMP and TET. Negative contributions are delivered with an equal quantity by the opposite of the variables mentioned here. E.g. the first dimension receives a negative contribution of 0.45 by susceptibility to TMP.

Table 6: Description of dimensions of MCA

Dimension 1		Dimension 2		Dimension 3	
Variable	**Estimate (p-value)**	**Variable**	**Estimate (p-value)**	**Variable**	**Estimate (p-value)**
CIPNAL.res_1	0.33 (<0.0001)	GEN.res_0	0.43 (<0.0001)	TMP.res_0	0.14 (<0.0001)
TMP.res_1	0.45 (<0.0001)	CTX.res_1	0.57 (<0.0001)	GEN.res_1	0.28 (<0.0001)
TET.res_1	0.43 (<0.0001)	CIPNAL.res_0	0.11 (<0.0001)	CTX.res_1	0.42 (<0.0001)
GEN.res_1	0.44 (<0.0001)	AMP.res_1	0.10 (<0.0001)	TET.res_0	0.10 (<0.0001)
CHL.res_1	0.50 (<0.0001)	CHL.res_0	0.09 (<0.0001)	CHL.res_1	0.12 (<0.0001)
AMP.res_1	0.46 (<0.0001)	TMP.res_1	0.03 (<0.0001)	CIPNAL.res_1	0.06 (<0.0001)
CTX.res_1	0.34 (<0.0001)	TET.res_1	0.02 (<0.0001)	AMP.res_0	0.06 (<0.0001)

Again here, the analysis can be extended with supplementary variables. The resulting graph for supplementary variable time is shown in Figure 8. The results obtained from this MCA closely resemble the conclusions from the PCA.

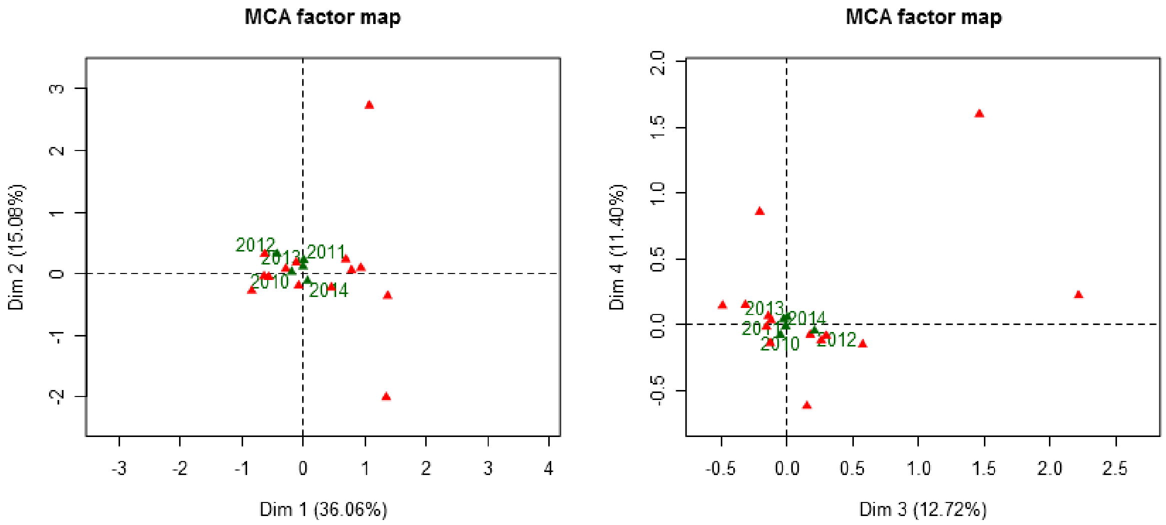


Figure 8: Factor map resulting from MCA, with time as supplementary variable

Also when including the covariate country, very similar results compared to the PCA are obtained, with Bulgaria located near high values for the first dimension (see Figure 9). The two plots on the top reflect the locations of the countries on dimensions 1-2 and 3-4, respectively. In addition, confidence ellipses are included in the bottom plot. Interpretations of these ellipses correspond with results shown from PCA. More specifically, in case the ellipses of two distinct countries do not overlap, those countries can be considered to differ significantly.

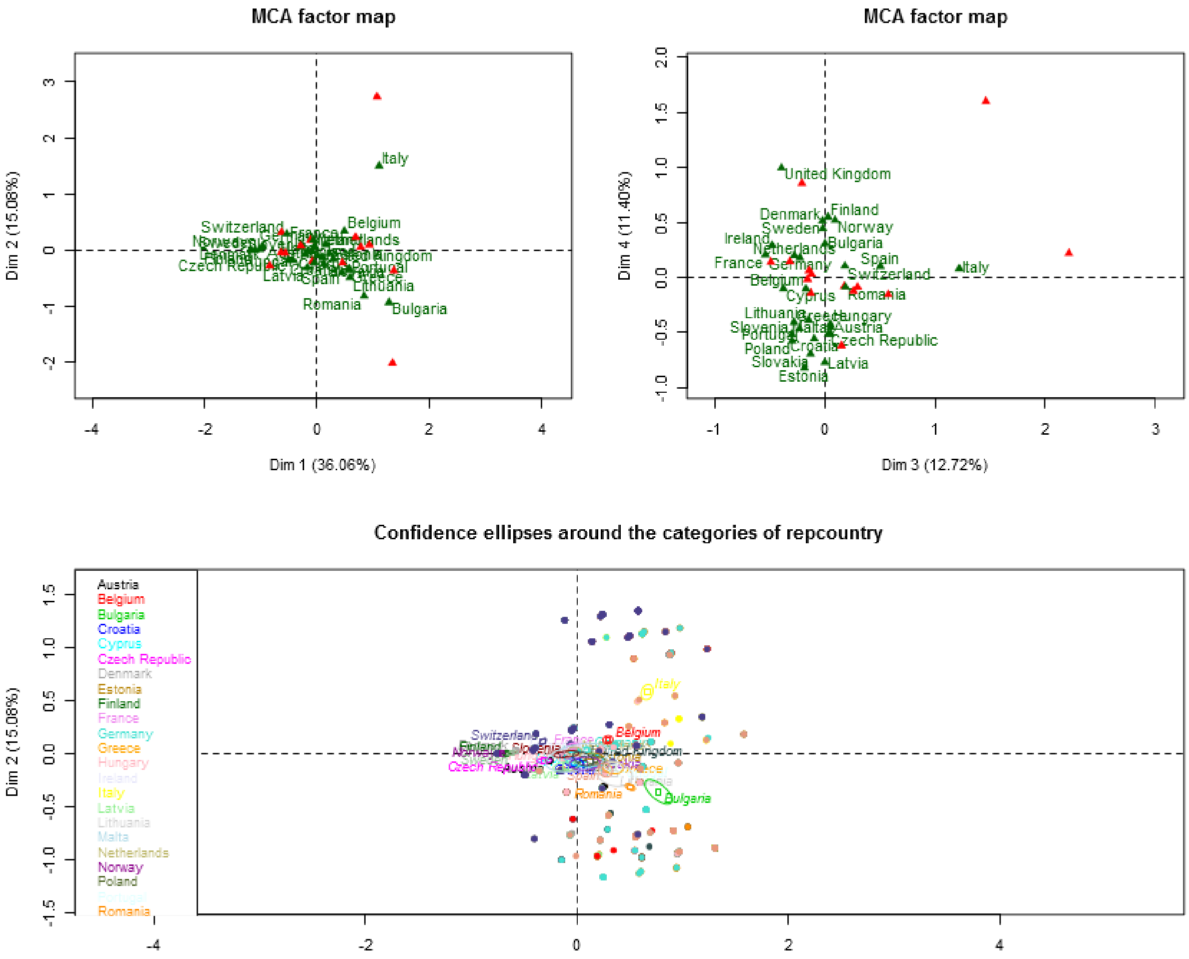


Figure 9: Factor map resulting from MCA, with country as supplementary variable

Hierarchical clustering after MCA

Applying the hierarchical clustering algorithm to the constructed dimensions, 6 clusters are identified, containing 2682, 2048, 2491, 642, 1076 and 731 isolates, respectively. They are represented in the factor map in Figure 10.

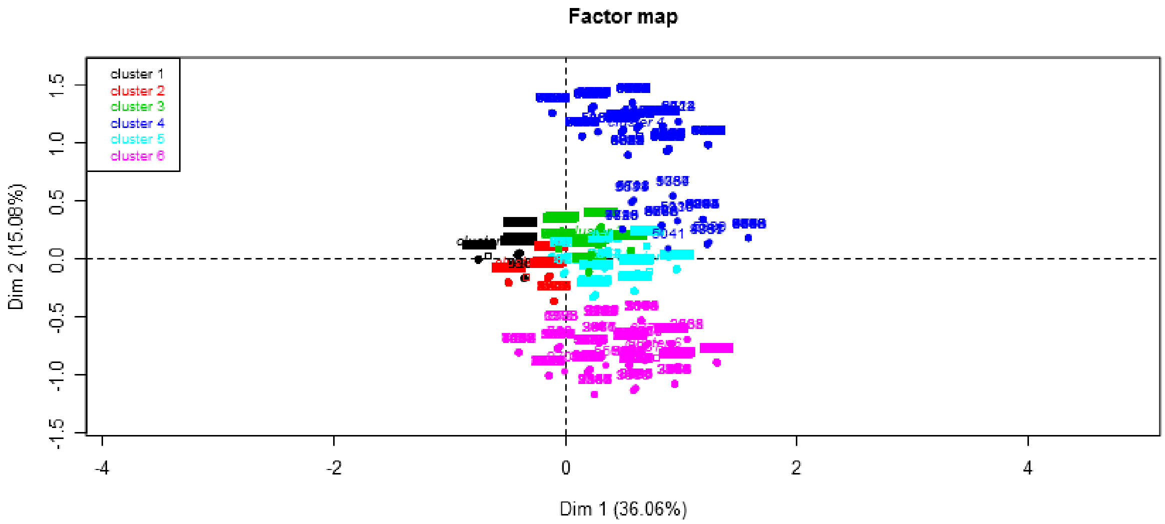


Figure 10: Factor map resulting from hierarchical clustering after MCA

Clusters 6 and 4 have similar values for the first dimension, but differ greatly in their value for dimension 2. The higher values for the first dimension indicate higher proportions of resistant isolates in these clusters. It is seen that cluster 6 has negative values for dimension two, which indicates that the isolates in this cluster have lower resistance against CTX and AMP compared to the isolates in cluster 4. Moreover, all isolates that are resistant to CTX are located in cluster 4. This can be seen in Table 7, which shows the cluster-specific proportions of resistance. Cluster 1 has the smallest value for the first dimension and therefore corresponds to isolates which show very little resistance against all of the AMs. Cluster 2 is closely related to the first cluster, but has slightly higher values for the first PC. This is reflected in the proportion of isolates resistant to CIPNAL, which equals 1 in cluster 2 and 0 in cluster 1. Note that Table 6 showed that CIPNAL is most significantly related to the first dimension. Isolates in clusters 3 and 5 show slightly elevated levels of resistance.

Table 7: Cluster-specific proportions of resistance

Cluster	N	GEN	CHL	CTX	AMP	TET	TMP	CIPNAL
1	2682	0.00	0.20	0	0.20	0.20	0.20	0.00
2	2048	0.00	0.20	0	0.20	0.20	0.20	1.00
3	2491	0.00	0.00	0	0.75	0.75	0.75	0.50
4	642	0.41	0.48	1	1.00	0.56	0.48	0.56
5	1076	0.00	1.00	0	0.57	0.57	0.57	0.50
6	731	1.00	0.47	0	0.50	0.50	0.53	0.53

3.1.3. Generalised Estimating Equations

In this particular data example, a generalised linear model relating the binary outcomes to time was fitted. Table 8 summarises the parameter estimates, with the robust standard errors and p-values for the slope parameters. It is observed that the slope for CTX is significantly negative, meaning that the odds of being resistant to CTX decrease over time. All remaining slopes, are significantly positive, meaning that the odds of being resistant to any of the other antimicrobials increases over time. For example, for AMP, the odds of being resistant in 2011 is exp(0.06) = 1.06 times the odds of being resistant in 2010.

Table 8: Parameter estimates from GEE

	Intercept		Slope		
Variable	**Estimate**	**S.e.[(a)]**	**Estimate**	**S.e.[(a)]**	**p-value**
AMP	-0.05	0.07	0.06	0.02	0.0001
CHL	-2.34	0.11	0.19	0.02	<0.0001
CIPNAL	-0.10	0.07	0.10	0.02	<0.0001
CTX	-1.88	0.11	-0.19	0.03	<0.0001
GEN	-3.79	0.18	0.32	0.04	<0.0001
TET	-0.59	0.07	0.09	0.02	<0.0001
TMP	-0.70	0,07	0.05	0.02	0.0036

(a): Robust standard errors

Finally, a heat-map representation of the estimated working correlation matrix is presented in Figure 11. The plot was constructed in such a way that the two AMs with the highest estimated correlation among them are located in the top-right. There seems to be an elevated correlation between AMP, TET and TMP, while the remaining correlations are more moderate. CTX does not seem to be correlated to GEN and only very marginal with the other AMs.

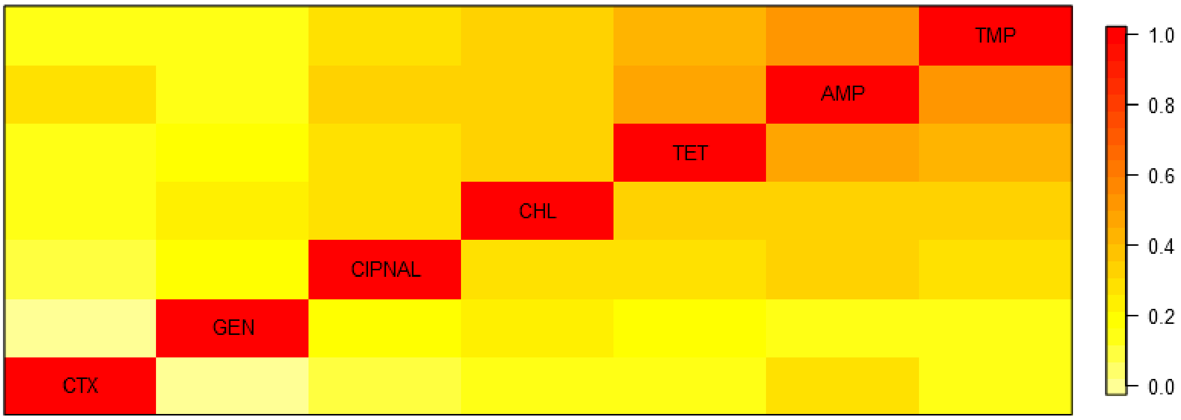


Figure 11: Estimated working correlation

3.1.4. Latent Class Analysis

A latent class analysis was performed using three latent classes (i.e. number of clusters found when performing hierarchical clustering after PCA). From the plot in Figure 12, the population shares of the constructed classes can be observed. A large class, containing 46% of the population, shows only minor resistance. The isolates contained in this latent class show practically no resistance to CHL, CTX, GEN and TMP. The probability to be resistant against AMP and TET is also very low, while resistance to CIPNAL is slightly elevated. On the other hand, the class on the right of Figure 12 (23% population share) contains isolates that have a higher probability to show resistance against all AMs of interest, except for CTX. Note that the probability to show resistance against CTX is low in all classes. The class on the left can be termed intermediate, with higher probability to be resistant against AMP, TET, TMP and CIPNAL and lower probabilities for the remaining AMs.

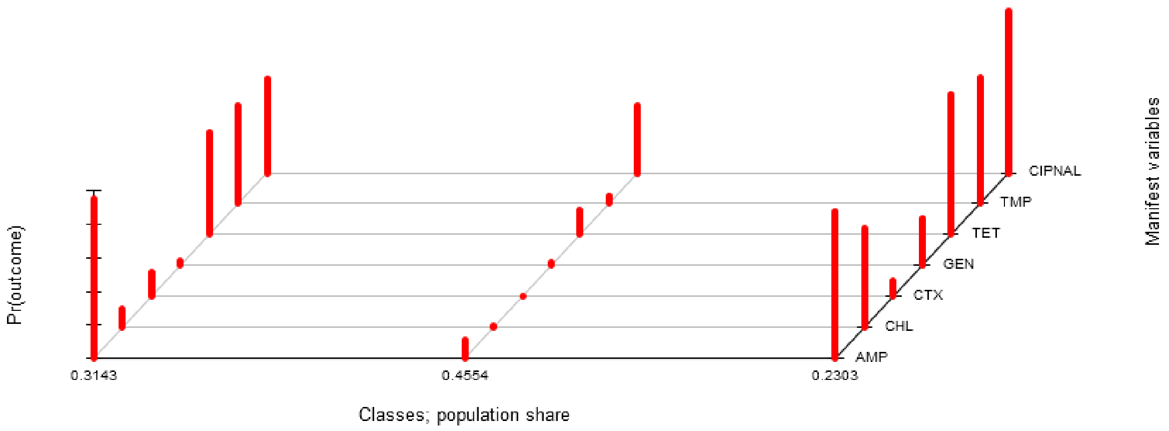


Figure 12: The three latent classes underlying the structure of the dataset under investigation

The covariate time was included as well to see how the odds to be within a specific latent class changes over the years. These results are presented in Table 9 below. Classes are numbered from left to right in Figure 12 above (class 1 is located on the left, followed by classes 2 and 3). It is seen that the odds of being in class 2 compared to class 1 is slightly increasing over time, while the odds of being in class 3 compared to class 1 is slightly decreasing.

Table 9: Parameter estimates from GEE

Classes	Parameter	Estimate (se)	p-value
2/1	Intercept	0.00 (1e-5)	0.99
	year	0.00018 (1e-5)	<0.0001
3/1	Intercept	0.00 (1e-5)	0.99
	year	-0.00015 (1e-5)	<0.0001

A model with six latent classes (i.e. number of clusters found when performing hierarchical clustering after MCA) was also fitted. The resulting classes are shown in Figure 13. The class on the left, containing 35% of the population is the class with the least resistance. Classes three, four and five could be seen as having isolates that show intermediate resistance, while classes two and six show relatively high resistance. There were some model fitting issues (inherent to the underlying latent allocations) when trying to include the time component with the six latent classes. These results could therefore not be presented. On the other hand, when 5 latent classes were selected, the model converged without any problems. A possible explanation is that 6 latent classes, in combination with a time component introduce classes that are sparse, resulting into errors in the model fitting process. In case such errors appear, the number of assumed latent classes should be decreased.

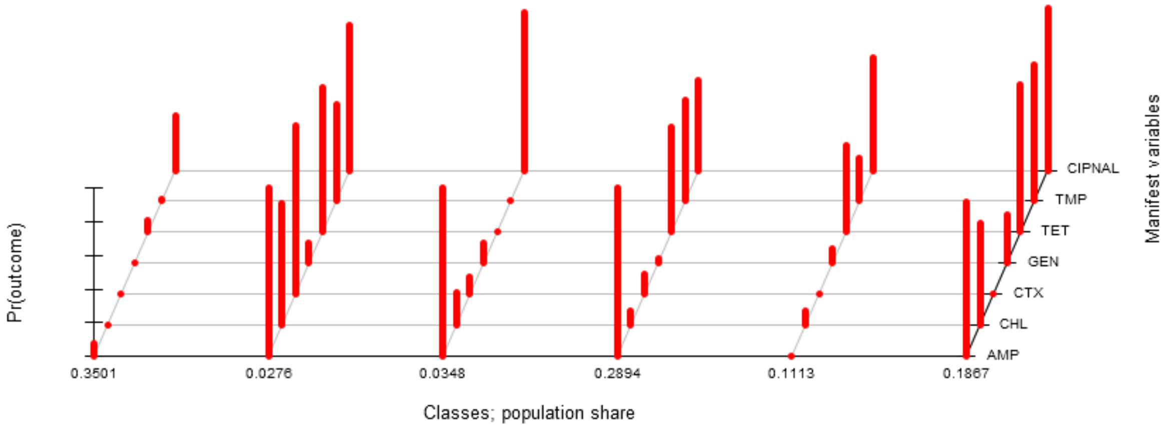


Figure 13: The six latent classes underlying the structure of the dataset under investigation

3.1.5. Bayesian Network Analysis

A DAG including the 7 antimicrobials of interest, together with the variable year was constructed. After creating an initial graph, a bootstrap procedure was performed for trimming of edges that were in the initial DAG due to chance. 50 new datasets were sampled based on the initially constructed DAG and an edge was retained in case it was present in more than 50% of the created bootstraps DAGs. In general, the more bootstrap datasets are considered, the more stable the final results are to be expected. The choice of the amount of bootstrap samples is mainly a consideration of time. In most cases, 50 new datasets should be a sufficient lower limit. The final result is shown in Figure 14. On the edges, the odds ratios comparing the two variables the respective edge connects are shown.

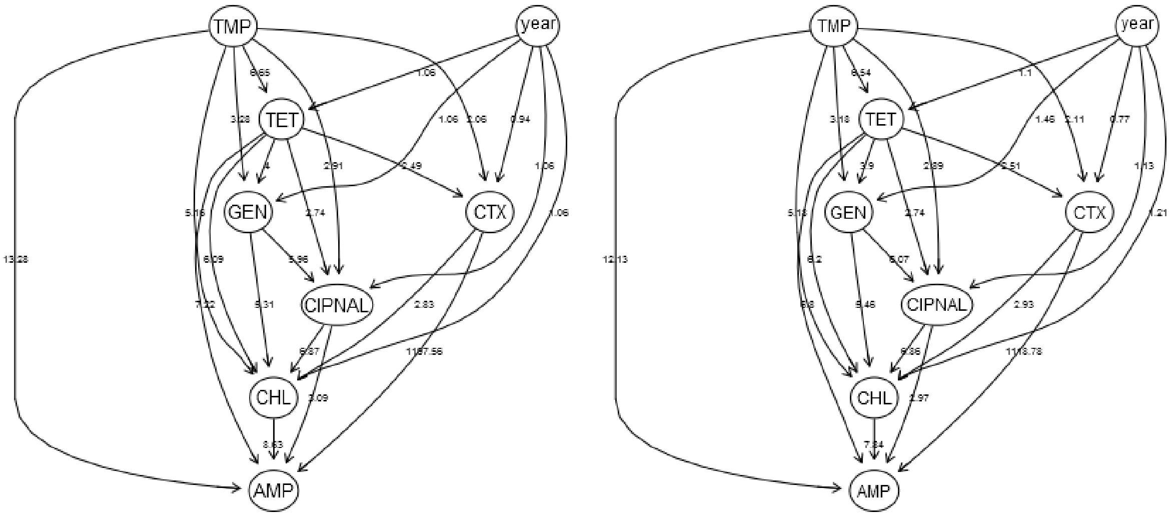
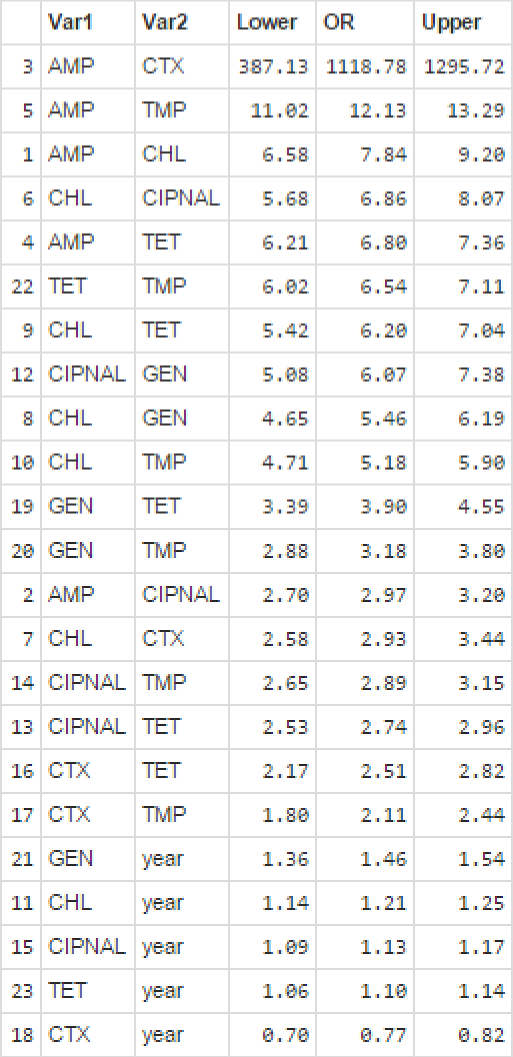


Figure 14: Final DAG relating the underlying relations amongst antimicrobials and between antimicrobials and the time component.

On the right, the averaged bootstrap odds ratios are included as labels, while on the left, the odds ratios resulting from the original data is shown. When interpreting, focus should be on the right plot.

For example, it is seen that there is an edge between TET and CIPNAL. The corresponding odds ratio is 2.74. This mean that the odds of observing resistance to CIPNAL when an isolate shows resistance to TET is 2.74 times the odds of observing resistance to CIPNAL when an isolate shows susceptibility to TET. Similar interpretations follow for the other edges. In order to see whether these ORs are significant, Figure 15, obtained from the Shiny application, can be regarded.



	Var1	Var2	Lower	OR	Upper
3	AMP	CTX	387.13	1118.78	1295.72
5	AMP	TMP	11.02	12.13	13.29
1	AMP	CHL	6.58	7.84	9.20
6	CHL	CIPNAL	5.68	6.86	8.07
4	AMP	TET	6.21	6.80	7.36
22	TET	TMP	6.02	6.54	7.11
9	CHL	TET	5.42	6.20	7.04
12	CIPNAL	GEN	5.08	6.07	7.38
8	CHL	GEN	4.65	5.46	6.19
10	CHL	TMP	4.71	5.18	5.90
19	GEN	TET	3.39	3.90	4.55
20	GEN	TMP	2.88	3.18	3.80
2	AMP	CIPNAL	2.70	2.97	3.20
7	CHL	CTX	2.58	2.93	3.44
14	CIPNAL	TMP	2.65	2.89	3.15
13	CIPNAL	TET	2.53	2.74	2.96
16	CTX	TET	2.17	2.51	2.82
17	CTX	TMP	1.80	2.11	2.44
21	GEN	year	1.36	1.46	1.54
11	CHL	year	1.14	1.21	1.25
15	CIPNAL	year	1.09	1.13	1.17
23	TET	year	1.06	1.10	1.14
18	CTX	year	0.70	0.77	0.82

Figure 15: Averaged bootstrap ORs for the final DAG.

The bootstrap averaged lower and upper limits of the confidence interval for the odds ratio are shown. In case 1 is not contained within the interval, the OR is significant. For TET and CIPNAL, this is the case (CI = [2.53-2.96]).

Finally, while the covariate year was included directly into the DAG above, another approach can be followed as well. Indeed, when interest is solely in identifying structure between the antimicrobials, it is possible to account for possible covariates (time, country, animal species,…) beforehand. More specifically, one can build a logistic regression model, where the response is taken to be resistance to a certain AM and the covariates are (functions of) the variables of interest. Consecutively, one can use the residuals from these logistic regression models as input values for the Bayesian network (instead of working with the raw binary data). In case of the *E. coli* example considered above, logistic regression models with a linear effect of time were fitted and the resulting residuals were used as input to the Bayesian networks analysis. As a result, the following DAG was found.

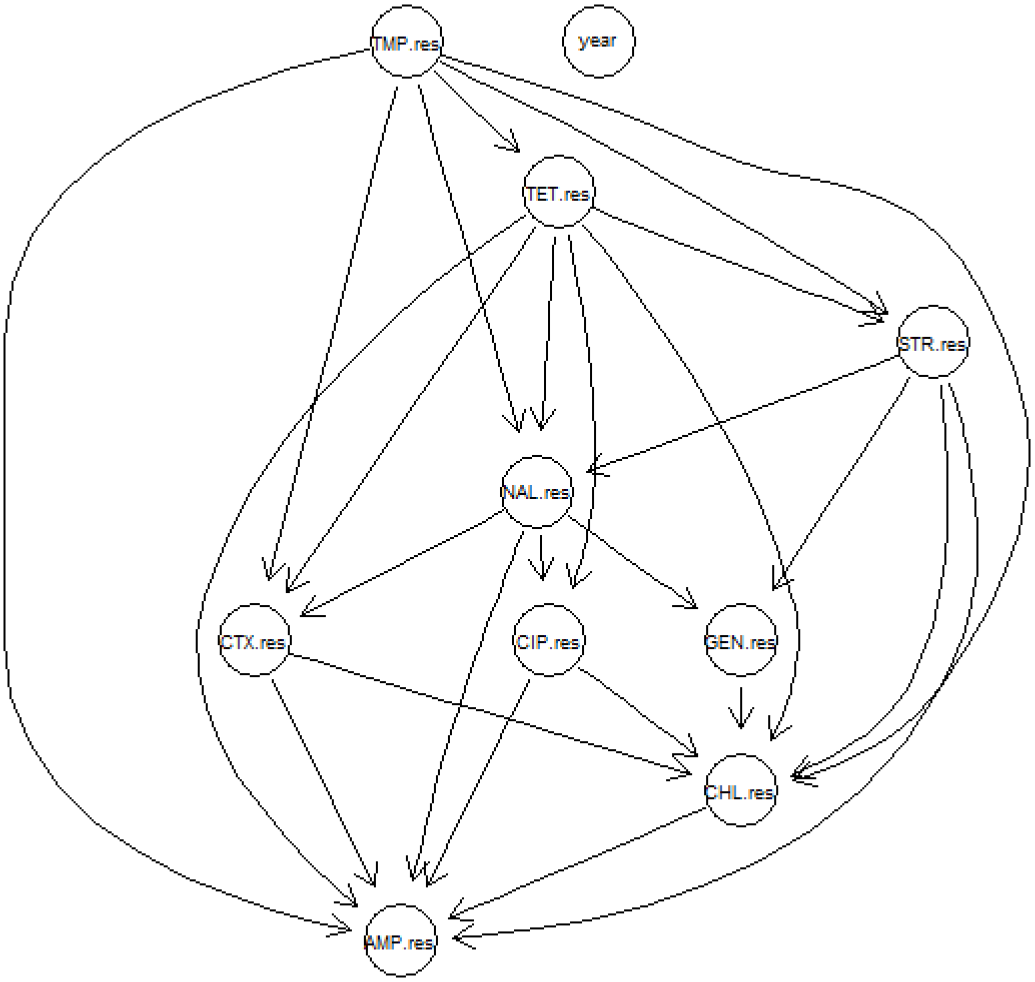


Figure 16: DAG based on the Pearson residuals from the logistic regression models

It can be observed that the DAG is consistent with the one obtained using the raw binary data, except that the variable year is no longer connected to any AM. Nevertheless, this latter option should be considered with caution at this point. Indeed, one of the underlying assumptions is that the residuals that are substituted into the model should be Gaussian. This is often not the case with logistic regression models, as it is shown in Figure 17 below.

Moreover, additional research in this direction is required. Indeed, it should be investigated whether the DAG build with the raw binary data is always consistent with the DAG build using the residuals and, which residuals should be preferred (at this point, Pearson residuals were employed). For these reasons, the option of accounting for covariates beforehand was not yet included in the application.

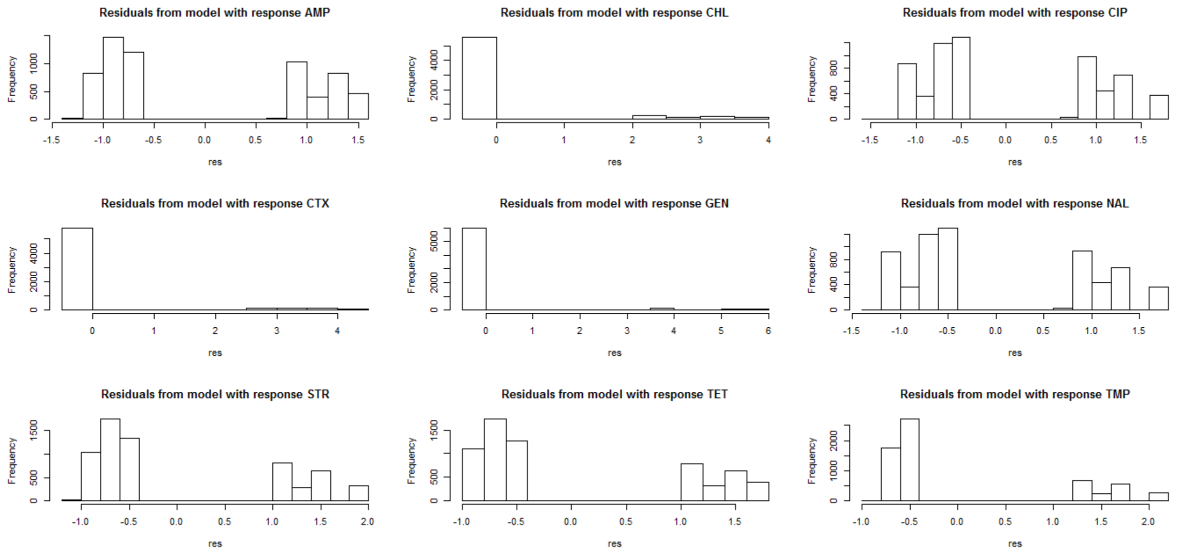
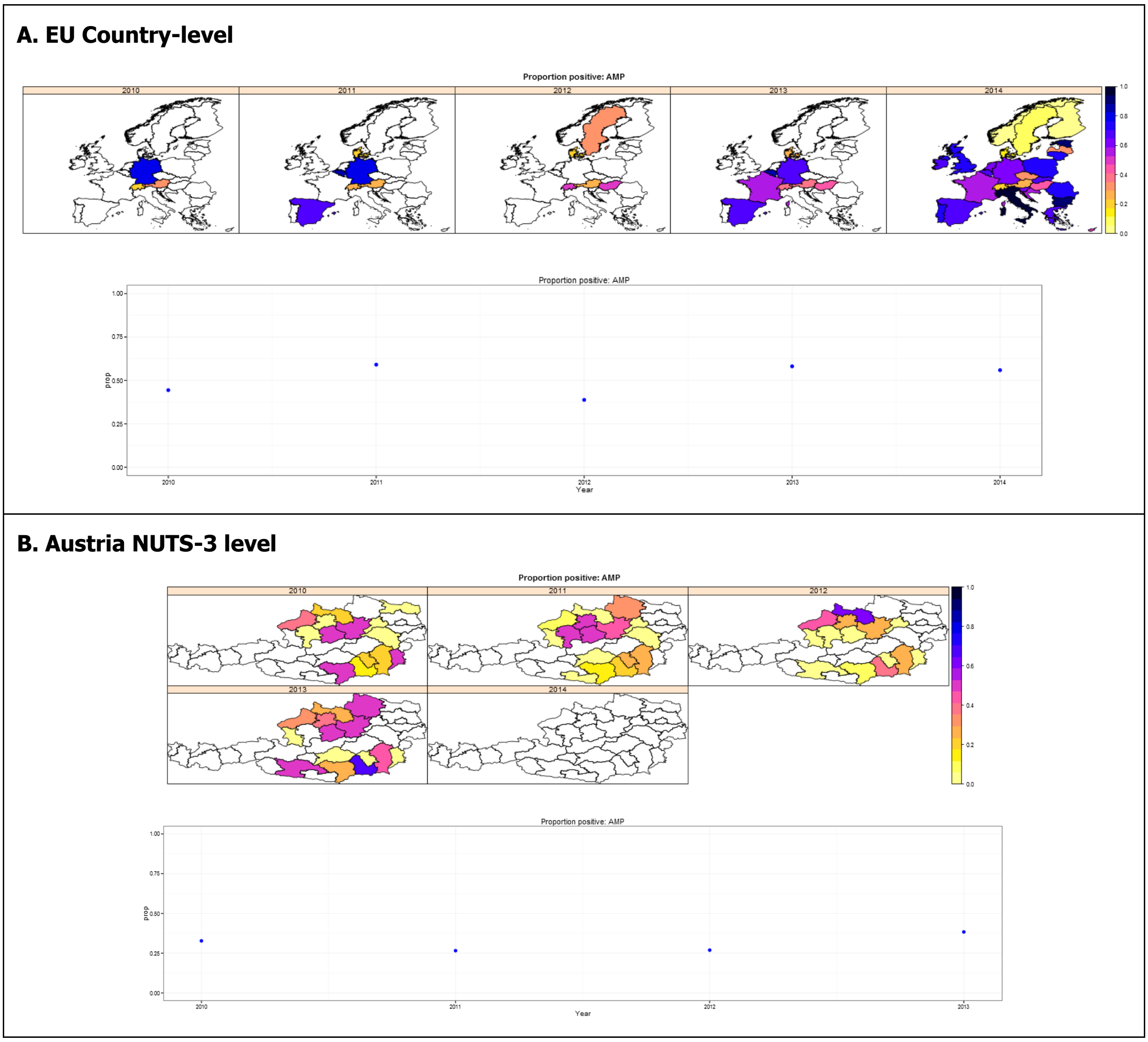


Figure 17: Histograms of the Pearson residuals from the logistic regression models

3.1.6. Spatio-Temporal Models

3.1.6.1 Univariate Analysis – Country-level

A univariate spatio-temporal model was fitted to the *E.coli* data for broilers at the country, NUTS-1, NUTS-2 and NUTS-3 level. The Nomenclature of Units for Territorial Statistics (NUTS) 3 level refers to the city or municipality per country. It was observed that in the country level, data was not available for all countries, while in the NUTS-1 to NUTS-3 level, data were available for only specific areas (see Figure 18). Although countries/areas without information were included, it was treated as 'NA' in the analysis, hence gives no contribution to the likelihood.


A. EU Country-level

B. Austria NUTS-3 level

Figure 18: Observed proportion of isolates with AMP resistance at A) country-level and B) Austria NUTS-3 level. The maps show the spatial pattern of resistance over the years while the dot plots show the over-all yearly proportion of resistance.

Table 10: Best temporal trend (time = year) for Ampicillin, based on the DIC criterion.

Space-time Interaction	Best temporal trend			
	Country-level	Austria		
		NUTS1	NUTS2	NUTS3
None	RW1	AR1	AR1	RW2
Unstructured (I)	RW2	AR1	None	RW2
Structured Time-Unstructured Space (II)	RW1	AR1	AR1	AR1
Structured Space-Time (IV)	Linear	AR1	AR1	AR1

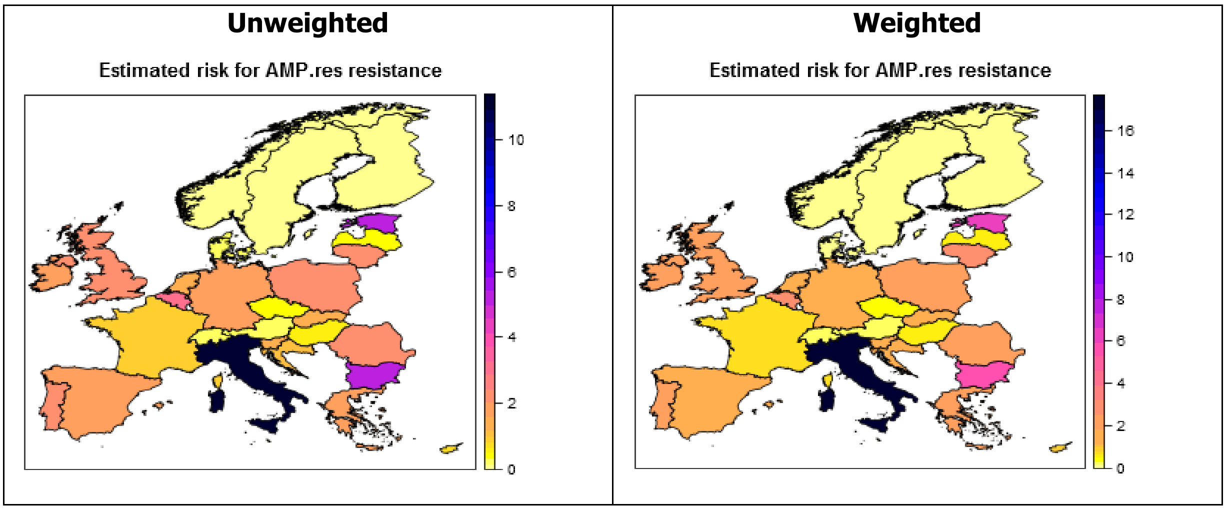
The best model (in terms of the time trend) for Ampicillin, using different space-time interaction, is given in Table 10. Note that in order for one to decide which space-time interaction to use, one needs to take note and compare the DIC between the respective best models for the different interaction terms.

The possibility of doing a weighted analysis and/or incorporating trade information as covariate was also investigated. For this application, the weights were calculated based on the 2014 EFSA report "Technical specifications on randomised sampling for harmonised monitoring of antimicrobial resistance in zoonotic and commensal bacteria" (EFSA, 2014). An input file (CountryProd.csv) is needed indicating whether each EU country has less than 100 000 tons of poultry or pig meat production or less than 50 000 tons of bovine meat production (coded as 1 in the input file, 0 otherwise, see Appendix B on how the file should look like). The production information used in this report was obtained from the Eurostat database (Eurostat, 2016a). For the trade information, 3 files (within one zipped file) are needed: Country_abbrev.csv, period.csv, and TradeData.csv. The first two files contain the corresponding country names/abbreviation and the period/time information. The TradeData file contains the import data of a declarant country (the country which imports or the destination country) from the partner country (the source country). This information can be downloaded from the Eurostat database (Eurostat, 2016b). Two ways of incorporating this (inter-country) trade information is possible in this application: a yearly trade information, or an average (fixed over the years) trade information. The latter is quite sensible in the case where there is little trade information (large number of missing data) in some years. See Appendix B for a screenshot on how each file should look like.

Results shown below (Table 11) are the estimates from the weighted and unweighted model (with poultry trade information, RW1 time component and unstructured spatio-temporal interaction) for resistance in AMP in all EU countries. Table 11: shows the parameter estimates while Figures 19-21 shows the plots of the estimated odds of AMP resistance for each country. Difference between the estimates from the unweighted and weighted model can be observed. In the unweighted analysis, poultry trade information has no significant effect on the AMP resistance, while in the weighted analysis, a significant effect was observed. A small estimate of the time variance parameter indicates a rather smooth temporal effect. With respect to the spatial component, an increase/decrease in odds can be observed in countries where information was available. The marginal spatial variance is estimated around 1.4/1.5 with only a small proportion (0.06) of this variance explained by the (structured) spatial variation. Thus, the spatial effect is mostly dominated by the overdispersion or the unstructured effect. Figures 20 and 21 shows the estimated probability of AMP resistance, including a prediction for 2015. Italy, Bulgaria and Estonia showed higher odds of AMP resistance, while the Nordic countries: Norway, Sweden, Finland and Denmark showed lower odds of AMP resistance. Looking back however to Figure 18, it is obvious that the observed spatial pattern is determined by the 2014 data since it is this year that information for most of the countries is available. Furthermore, a high or low spatial risk is predicted depending on the observed data in a particular area and also of the neighbouring areas. So, an area with high proportion of positives can have low predicted spatial risk if the neighbouring areas have low proportion of positives, and vice versa.

Table 11: Parameter estimates from the (unweighted and weighted) RW1 with unstructured space-time interaction model for AMP resistance in EU countries. Yearly poultry trade information between the EU countries is included as fixed effect. Est refers to the estimate, s.e. refers to the standard error, while C.I. refers to the credible interval.

	Unweighted			**Weighted**		
Parameter	**Est**	**s.e.**	**95% C.I.**	**Est**	**s.e.**	**95% C.I.**
Intercept	0.297	0.240	-0.1760, 0.7710	0.267	0.239	-0.2010, 0.7350
Trade Information	0.005	0.016	-0.0270, 0.0370	0.032	0.014	0.0040, 0.0600
Variance for t.ID *(time – RW1 varaince)*	0.034	0.042	0.0004, 0.1410	0.075	0.081	0.0060, 0.2810
Variance for s.ID *(Marginal Variance)*	1.439	0.366	0.8720, 2.2840	1.543	0.372	0.9540, 2.3890
Phi for s.ID *(space - proportion of the spatial marginal variance that can be attributed to the structured spatial effect)*	0.059	0.059	0.0030, 0.2220	0.062	0.061	0.0040, 0.2270
Variance for st.ID.t (*Spatio-temporal interaction variance*)	0.098	0.050	0.0350, 0.2220	0.014	0.025	0.0001, 0.0380

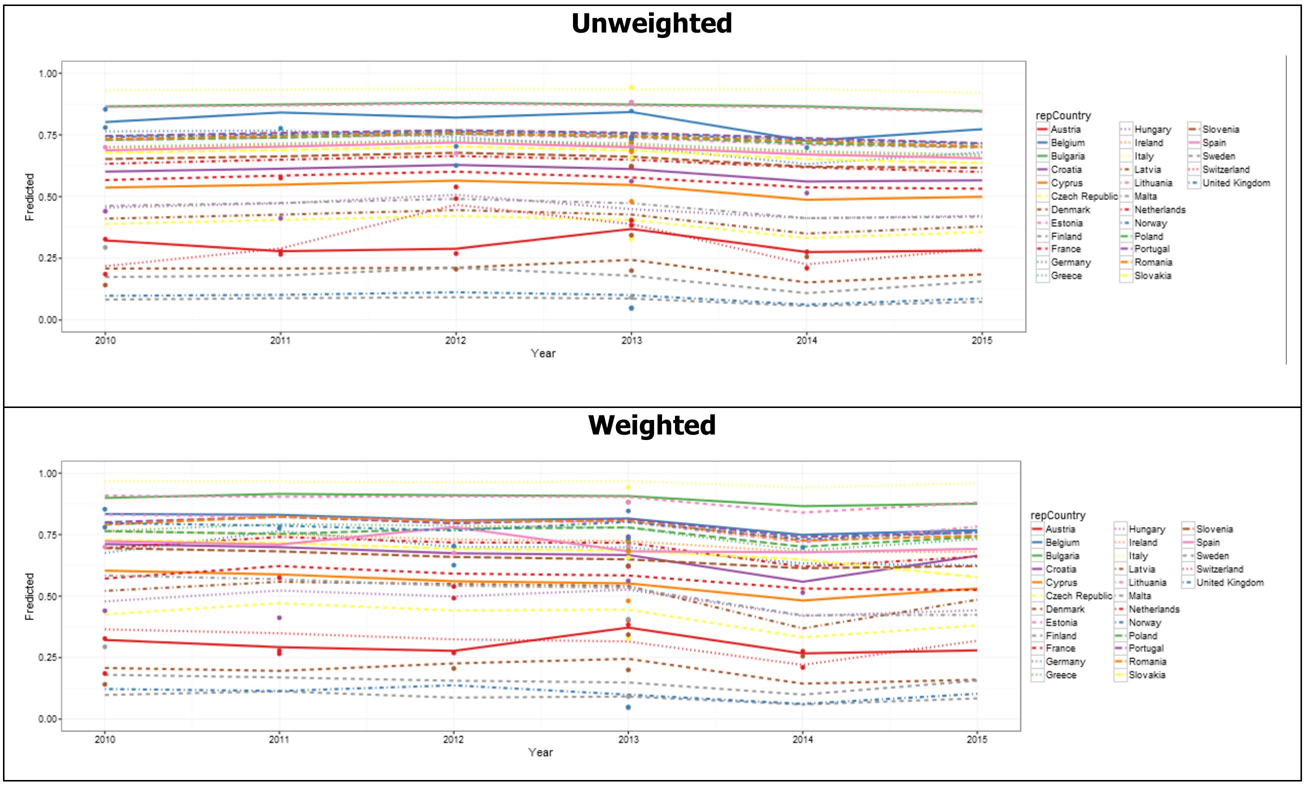


Unweighted

Weighted

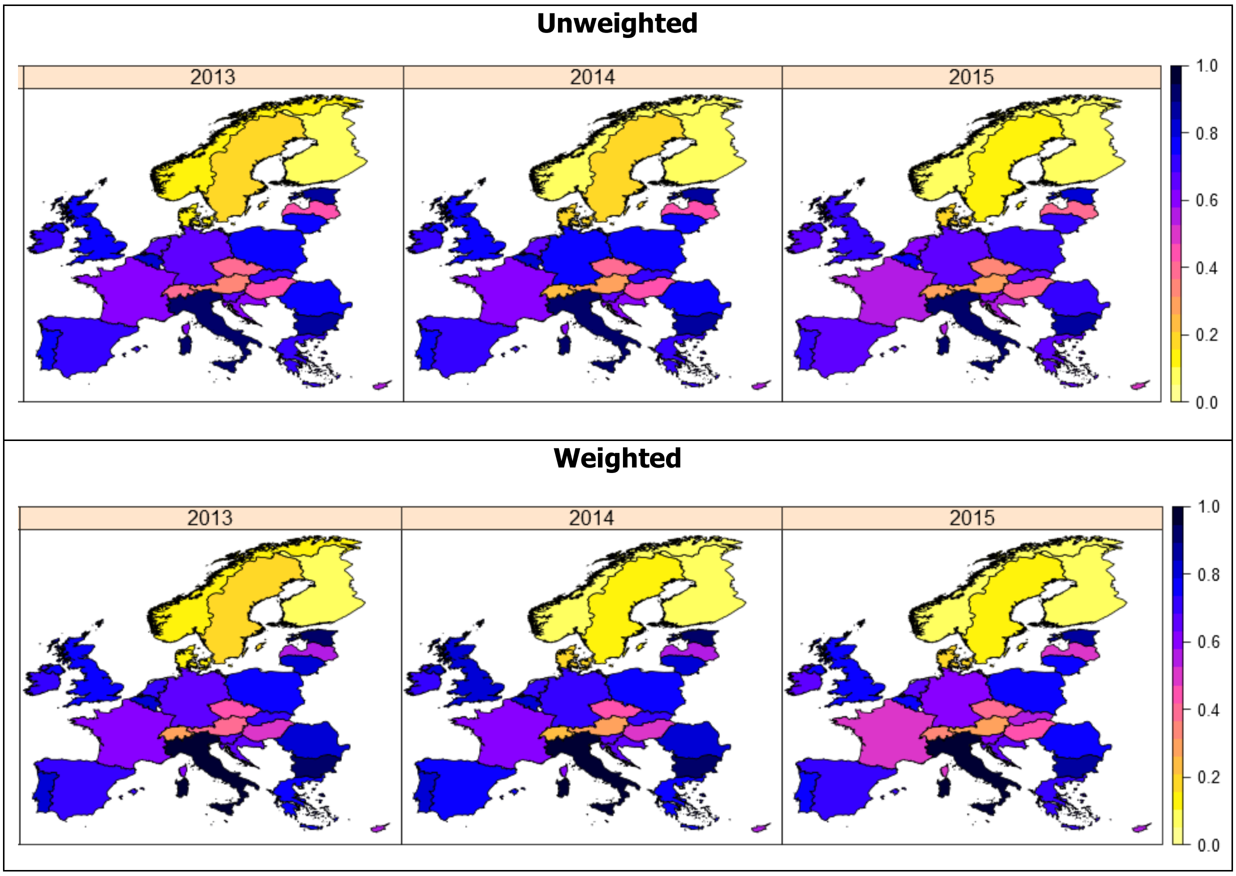
Estimated risk for AMP.res resistance
0
2
4
6
8
10
Estimated risk for AMP.res resistance
0
2
4
6
8
10
12
14
16

Figure 19: Estimated risk of AMP resistance ($\exp(s_l)$) based on the RW1+unstructured model at country-level.


Unweighted

Weighted

Figure 20: Estimated (overall) probability of AMP resistance for all spatial units, based on the RW1+unstructured model at country-level.



Unweighted

Weighted

Figure 21: Estimated (overall) probability of AMP resistance for 2013-2014 and prediction for 2015, based on the RW1+unstructured model at country-level.

In general, interpretation of parameter estimates and some statistics from the model-fit is fairly straightforward. The output gives the fixed parameters, random parameters and some model-fit statistics. In Table 11 for instance, under the fixed parameters:

- An intercept estimate $(\widehat{\beta_0})$ is given. This gives the log-odds at baseline.
- An estimate of the effect of the previous-year trade information is given. This effect reflects not only the contribution of previous-year trading but also trading between countries with AMP resistance at the previous year. Hence, a (significant) positive effect means that an increase of trading (import) or proportion of AMP resistance of the source countries at the previous year would lead to an increase in AMP resistance of a particular country at the current year.
- Since a flexible temporal effect is assumed, there is no estimate for it under the fixed parameters. However, if a linear or saturated time effect is chosen,
    - for the linear time effect (t.ID), an estimate for the linear trend will be given. This will reflect the increase or decrease of the log-odds of AMP resistance every year (or month or week).
    - For the saturated time effect (ctime), estimates corresponding to each time point will be given (the first time point is taken as the reference time). In the case of the yearly AMP resistance, a yearly estimate is given, where the first year (2010 in this

case) is taken as the reference year. So, ctime2 refers to the 2011 estimate and ctime6 refers to the 2015 estimate, all in reference to 2010.

Under the random parameters, estimate for the variance ($\hat{\sigma}_{\omega}^{2}$) of the flexible temporal trends (RW1, RW2 or AR1) is given, along with the correlation parameter Rho (ρ) if the AR1 effect is chosen, marginal variance ($\hat{\sigma}_{s}^{2}$) of the spatial effect, Phi (ϕ), which is the proportion of marginal variance attributed to spatial variation, and variance ($\hat{\sigma}_{u}^{2}$ or Var for st.ID.t) for the space-time interaction if it is chosen.

With respect to the model-fit statistics, the deviance information criteria (DIC), mean deviance, effective number of parameters, logarithmic score and McFadden's pseudo R-squared value are shown in the output. A short description of these statistics is given in Section 2.2.7. Generally, models with smaller DIC and higher logarithmic score and McFadden's pseudo R-squared are preferred. It should be noted that care should be taken in interpreting the pseudo R-squared as this cannot be interpreted the same way as in the normal linear regression case.

3.1.6.2 Univariate Analysis – NUTS-3-level

Results shown below are the estimates from the RW2 model with unstructured spatio-temporal interaction for resistance in AMP in Austria, at NUTS 3 level. Table 12 shows the parameter estimates while Figures 22-23 show the plots of the estimated risk/probability of AMP resistance per NUTS 3 areas. Estimate for the variance of the RW2 time effect shows a small value which implies a rather smooth temporal trend. Figure 23 confirms this observation where it can be observed that a smooth curved predicted trend. For the spatial component, a small marginal variance can also be observed which similarly implies a smooth trend. And of this marginal variance, only around 28% of the variability can be explained by the structured spatial effect. The marginal variance of the spatial component refers to the variance of the iid + structured spatial effect.

Figure 22 shows a spatial trend of resistance, although looking closely at the range of values, it seems not to be so different from each other. Figure 23 on the other hand, shows the estimated probability of AMP resistance for each area and time. A one-year-ahead prediction is also added.

Table 12: Parameter estimates from the RW2 + unstructured spatio-temporal interaction model for AMP resistance in Austria.

Parameter	**Estimate**	**s.e.**	**95% C.I.**
Intercept	-0.728	0.177	-1.080, -0.372
Variance for the time (RW2 component)	0.040	0.071	0.0005, 0.209
Marginal Variance for the Spatial component	0.054	0.0630	0.002, 0.218
Phi for the Spatial effect	0.275	0.262	0.006, 0.895
Variance for st.ID.s	0.036	0.049	0.0001, 0.019

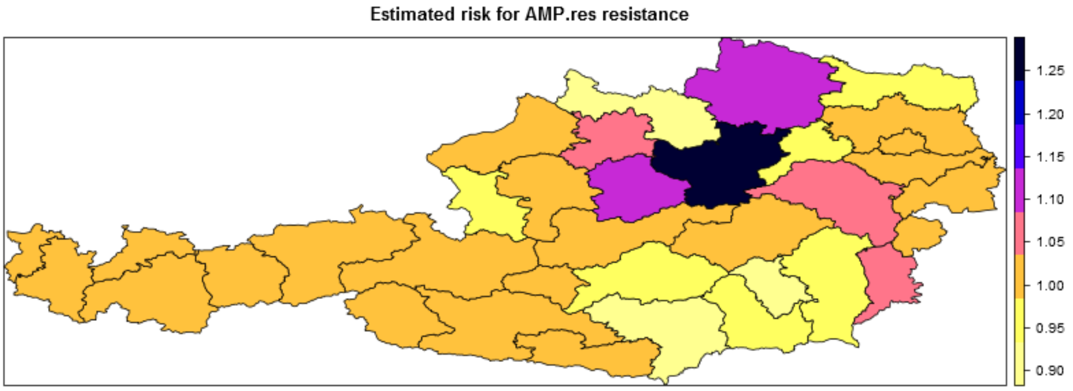
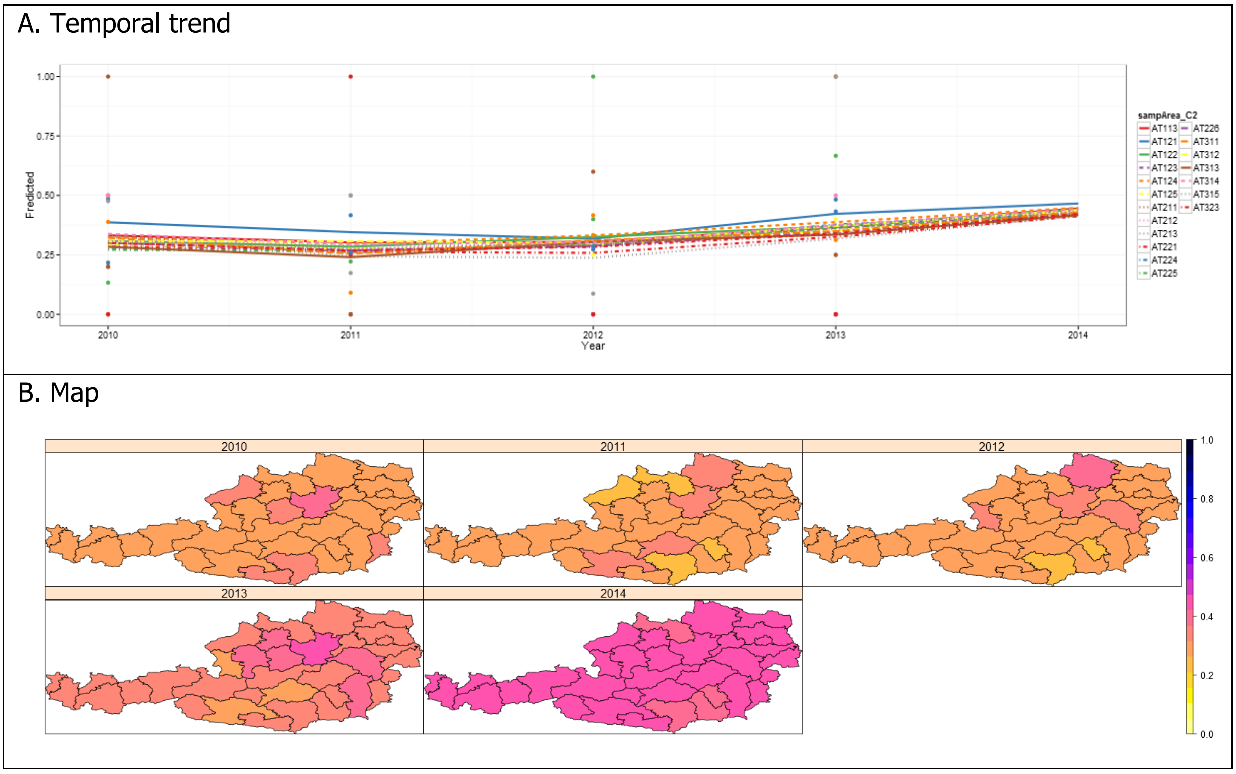


Figure 22: Estimated spatial risk of AMP resistance ($\exp(s_l)$) based on the RW2 + unstructured spatio-temporal interaction model at NUTS3-level.


A. Temporal trend

B. Map

Figure 23: Estimated probability of AMP resistance (year 2010 – 2014) for all NUTS-3 municipalities in Austria based on the RW2 + unstructured spatio-temporal interaction model. A) shows the temporal trend while B) shows the map or the spatial trend over the years.

3.1.6.3 Bivariate Analysis

A bivariate spatio-temporal model was also fitted to the *E.coli* data for broilers at the EU country-level. Output for the antimicrobials AMP and TET are given below. Table 13 shows the parameter estimates from the model where a separate intercept term is estimated for the two antimicrobials. If a time-effect was also added into the model, this is also estimated separately for each antimicrobial.

With respect to the spatial effect (s.ID), a marginal variance of 1.3 was estimated, where only around 7% of this variation is due to the structured effect, the majority of this variation is due to the unstructured effect or overdispersion. This was also observed in the univariate case, for AMP. A high correlation between the spatial effect of the two antimicrobials can also be seen (see Figure 24), which is estimated to be around 0.84. This means that there is a high degree of similarity between the spatial pattern of AMP and TET resistance. With respect to the space-time interaction, a very low variance of the unstructured effect was observed. Correlation between the space-time interaction of the two antimicrobials, was estimated around 0 with a high estimated standard error. This almost zero correlation and high standard error means that the model does not support a correlated space-time interaction between the two antimicrobials but rather, the space-time interaction between the two antimicrobials seems to be independent.

Figure 24 shows the different EU countries with increased risk of antimicrobial (AMP and TET) resistance, based on the spatial effect. Italy and Bulgaria have remarkably higher risk compared to other countries. Note that these are the countries with high risk of AMP resistance in the univariate analysis. It is also important to note that the observed spatial trend is mostly due to the 2014 data (see Figure 18), where there is information for these countries (and other countries as well) but not in other years.

Figures 25 and 26 give the predicted probability of antimicrobial resistance over the years and across different countries. A one-year-ahead prediction is also given. The figures show that certain countries have high predicted probabilities of AMP resistance (such as Italy, Bulgaria, Germany, Poland and Romania) and TET resistance (such as Italy, Bulgaria and Romania). Although a low risk based on the spatial effect can be observed for some of these countries, the risk is still greater than 1.

Table 13: Parameter estimates from the intercept-only with unstructured space-time interaction model for AMP and TET resistance in Europe.

Parameter	Estimate	s.e.	95% C.I.
Intercept (AMP)	0.203	0.212	-0.216, 0.620
Intercept (TET)	-0.256	0.212	-0.675, 0.162
Marginal variance for s.ID	1.318	0.303	0.826, 1.997
Phi for s.ID	0.074	0.065	0.006, 0.246
GroupRho for s.ID: Correlation of the Spatio-temporal effects of AMP and TET	0.844	0.072	0.670, 0.946
Variance for st.ID.t: Space-time interaction	0.101	0.030	0.056, 0.171
GroupRho for st.ID.t: Correlation of the space-time interaction between AMP and TET	-0.016	0.659	-0.976, 0.975

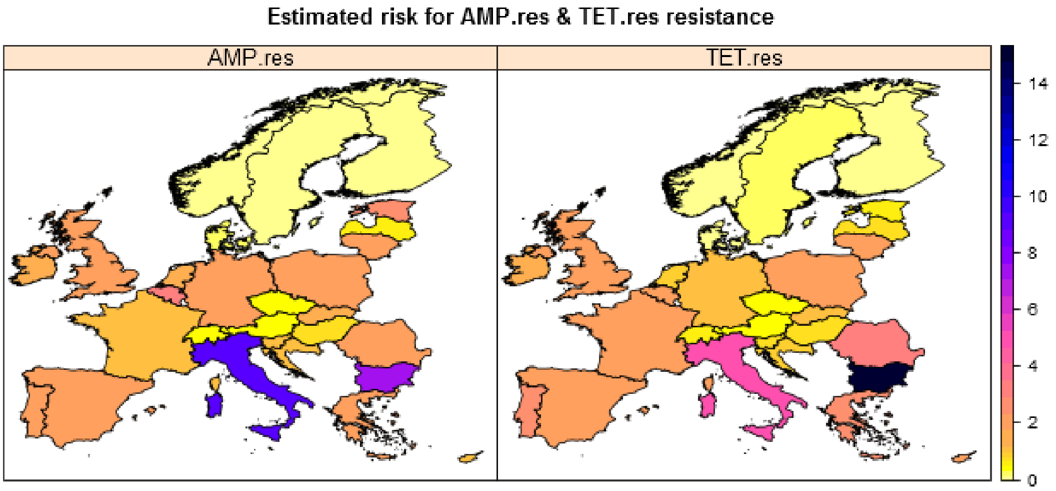


Figure 24: Estimated risk (spatial effect) of AMP and TET resistance ($\exp(s_l)$) based on the no-intercept model at country level.

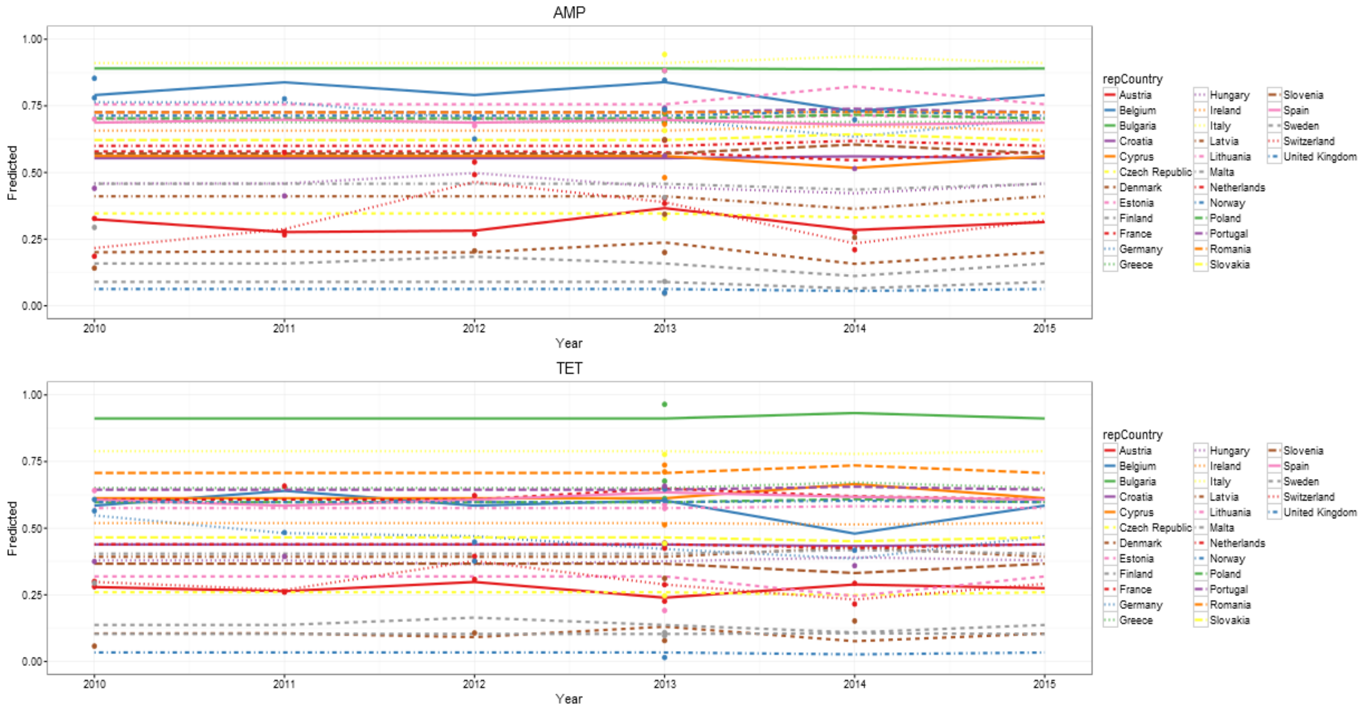


Figure 25: Estimated overall probability (temporal trend) of observing AMP and TET resistance based on the no-intercept model at country level.

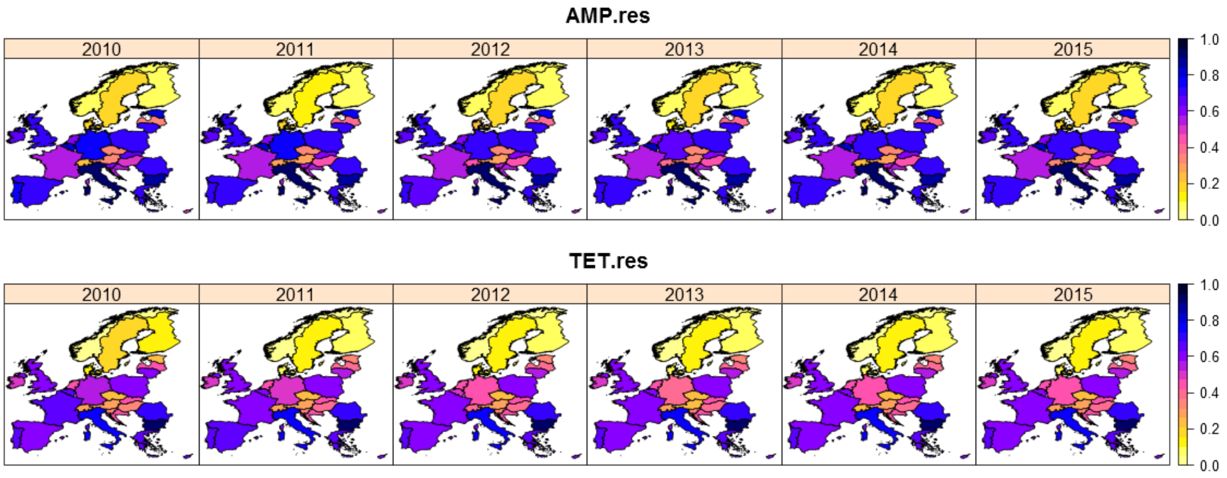


Figure 26: Estimated overall probability (spatial trend) of observing AMP and TET resistance based on the no-intercept model at country level. A one-year-ahead prediction (2015) is given.

3.1.7. Pattern Attribution Models

Pattern attribution models are used when interest is in describing the different MDR patterns and how these patterns depend on certain variables of interest, including the sampling year, country, animal origin and zoonosis type. There are 89 different resistance patterns in the dataset under investigation. The pattern in which isolates show resistance against AMP, TET and TMP and susceptibility against GEN, CHL, CTX and CIPNAL was investigated. A total of 419 isolates, sampled in 25 different countries, show the pattern of interest.

For the current subset, it can be shown graphically how the probability to show this resistance pattern evolves over the considered time period and between the reporting countries. In case different animal subtypes or bacteria subtypes were selected, these could be investigated as well. In this regard, the following plots can be investigated.

Figure 27 shows the evolution over time. It is observed that the overall prevalence of this pattern is rather low. The highest prevalence is observed in 2011, but the corresponding confidence limits overlap with those obtained for 2010 and 2013. There seems to be a significant drop in prevalence between 2011 and 2012 and the prevalence in 2014 is on the same low level compared to 2010 and 2012.

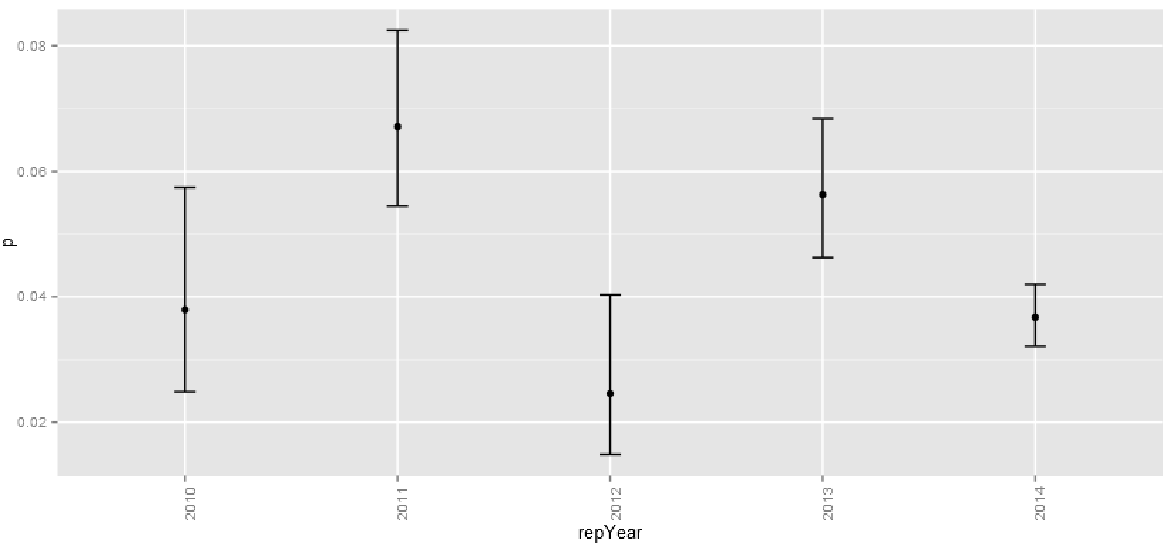


Figure 27: Observed prevalence of the selected pattern over the distinct years.

Figure 28 shows a similar plot, but now compares the prevalence between the distinct EU member states. The overall prevalences are plotted, disregarding the years at this point. Higher prevalences are observed in the UK and Ireland, but confidence limits are rather wide.

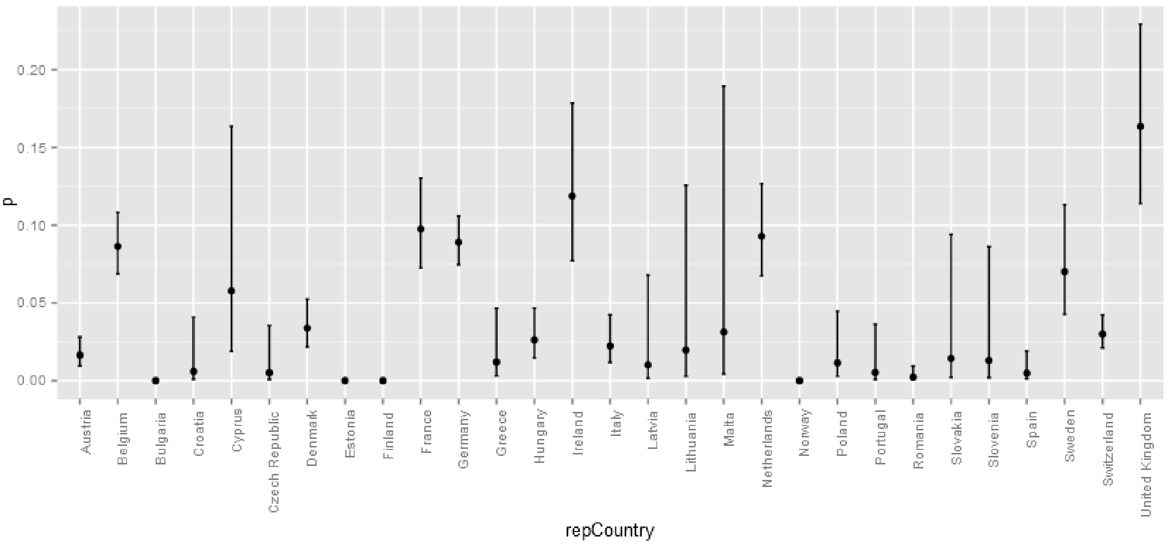


Figure 28: Observed prevalence of the selected pattern over the distinct countries.

In order to see how the prevalence evolves over the years, dependent on the country, both variables can be selected and plotted simultaneously in the Shiny application. In order to keep the plot readable, only two countries were selected to be shown in Figure 29. In Austria (shown in pink), the proportion of isolates that show the specific resistance pattern increases between 2010 and 2011. Next, after a small drop in 2012, it gradually increases again. The confidence limits are overlapping, so there is no significant increase. A similar behaviour is observed for Germany (shown in blue). Finally, it is seen that the probability to show the pattern is significantly higher in Germany compared to Austria (as the confidence intervals within the years do not overlap between Germany and Austria).

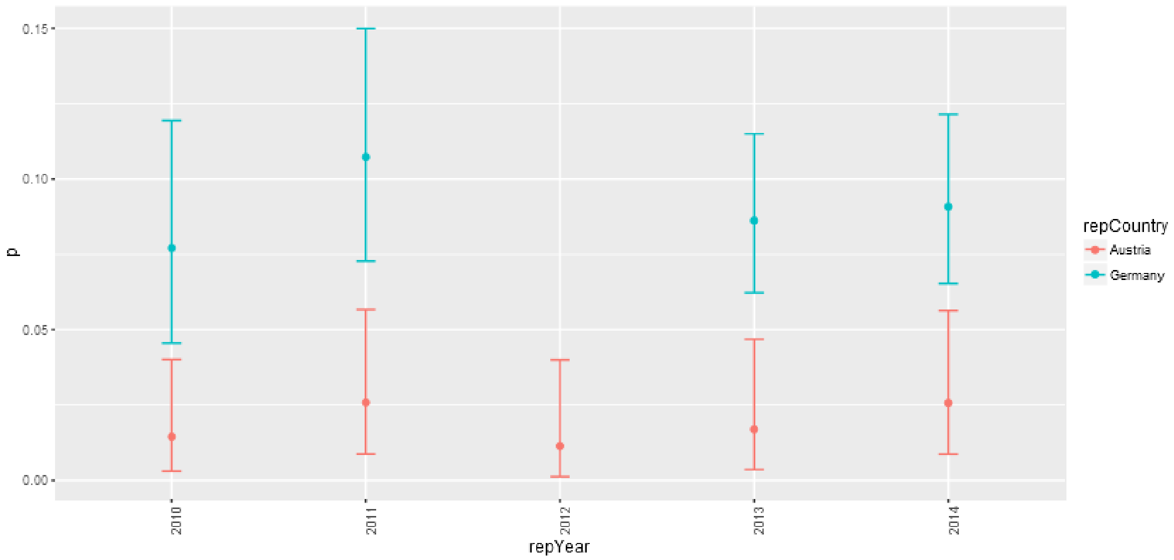


Figure 29: Observed prevalence of the selected pattern over the distinct years, per country.

In order to formally test whether there are differences between the levels of certain variables, Table 14 presents the result from the Firth logistic regression. It is seen that, for the years 2010, 2011, 2013 and 2014, the odds to observe the selected pattern in Austria is significantly lower than the odds in Germany. Similarly, in 2014, the odds of observing the pattern are higher in Austria as compared to Romania.

Table 14: Overview of significant differences in odds to observe the selected pattern.

Coeficient 1	Coeficient 2	OR (lower-upper)	Interpretation
Austria-2010	Germany-2010	0.18 (0.05-0.68)	Significant lower odds to show the selected pattern for Austria-2010 compared to Germany-2010
Austria-2011	Germany-2011	0.22 (0.08-0.61)	Significant lower odds to show the selected pattern for Austria-2011 compared to Germany-2011
Austria-2013	Germany-2013	0.18 (0.05-0.67)	Significant lower odds to show the selected pattern for Austria-2013 compared to Germany-2013
Austria-2014	Germany-2014	0.26 (0.10-0.72)	Significant lower odds to show the selected pattern for Austria-2014 compared to Germany-2014
Austria-2014	Romania-2014	8.89 (1.87-42.20)	Significant higher odds to show the selected pattern for Austria-2014 compared to Romania-2014

3.1.8. Source Attribution Models

In order to be able to apply the source attribution models, data on both resistance in humans as well as on data in animals are required. At this point, the available information is too scarce to provide the reader with a meaningful output and interpretation. For this reason, results are not included in this report. Nevertheless, the reader is referred to the accompanying tutorial for an exemplary analysis. In addition, it can be seen there that additional information on consumption and antimicrobial usage can be included in the model as well. Further exploration of the model with future data is required in order to determine the performance of the developed method.

4. Conclusions

This report introduces several methods for analysing multi-drug resistance data. First of all, attention was paid to classification trees and hierarchical clustering, which aim at identifying homogeneous subgroups of the data. In case of the former, a dataset was considered that contained information on both *E. coli* and *Salmonella* isolates, sampled from several animal or food species in different European member states. The final result of the analysis is a summary of subgroups which show similar resistance patterns. The subgroups were obtained by dividing the original dataset into respectively smaller groups, based on the value of certain covariates. As such, the "trees" analysis is a relatively straightforward method with a fairly simple interpretation of the results. Likewise, the hierarchical clustering (both based on the principal components as well as on the multiple correspondence analyses) approach provides a nice way to identify subgroups (corresponding to similar resistance patterns) in the data. From this analysis onwards, only *E. coli* isolates in broilers were considered. The PCA uses the original continuous MIC values as input. Three clusters in the data were observed, in which isolates showed similar values for both the continuous MIC values and the binary resistance data. It was discussed above that results from the PCA might be hampered by different dilution ranges across the reporting member states. For this reason, it can be recommended to compare the results from the clustering after PCA with those obtained after MCA. Indeed, the latter analysis is based on the binary profiles and does not suffer from different dilution ranges. On the other hand, dichotomising the continuous data could lead to the loss of valuable information since all isolates above the ECOFF are considered to be resistant, but no distinction is made on how large the actually observed MIC value is. Nevertheless, the clustering after MCA provides a nice insight into the underlying subgroups in the dataset. It was observed that the analysis resulted into 6 clusters, which show a more homogeneous resistance profile compared to the PCA analysis. Based on both cluster analyses, a latent class analysis was also performed, assuming respectively 3 and 6 latent classes. This approach is a model-based alternative, in which the effect of additional covariates can be quantified as well. In contrast, descriptive plots for the supplementary variables were provided for the hierarchical clustering after PCA and MCA.

Generalised estimating equations method was also employed to model multivariate binary data while accounting for an underlying correlation structure. A heat map of the underlying correlation structure was provided, to give the user an idea on the associations between the antimicrobials. However, these results should not be over-interpreted, but should trigger further in-depth analyses of specific co-resistance patterns (e.g. using the proposed pattern attribution models). Similarly, the Bayesian networks show the underlying associations between the distinct antimicrobials. A visual representation of the constructed Directed Acyclic Graph (DAG) is provided, which consists of nodes and edges. The nodes, containing the antimicrobials and possible time component, were connected, in case a significant association was observed. Again here, the resulting observations should be further investigated before jumping to conclusions. In this respect, an interactive discussion with experts in the field of microbiology is recommended.

Spatio-temporal models to account for the spatial structure of the data were also fitted. Both weighted/unweighted univariate (with and without the inclusion of trade data) and bivariate analyses can be performed, where the user is guided to select the appropriate spatial and temporal structures through the DIC criterion. Predictions for the upcoming year were shown, which could warn the user for possible shifts in the future.

Finally, the results from the pattern attribution models were discussed. Using these models, the user can obtain more information about the occurrence of a specific resistance pattern and how it depends on certain covariates like time, region and sampling source. First, descriptive plots were shown to provide some initial ideas on the observed proportion of the selected pattern for different values of certain covariates. For the pattern under consideration in this report, it was seen that higher proportions were observed in Germany compared to Austria. A more comprehensive result was provided based on Firth logistic regression. In this way, the user is informed on possible significant differences in observing the selected pattern between countries or over the years. The tool developed also foresees the potential use of source attribution models in which human resistance can be attributed to different sources, such as resistance reported in food in combination with the consumption patterns for each country as well as their usage within the human population.

5. Recommendations

The methods introduced in this report provided useful insights into the underlying structure of resistance in the antimicrobial population as well as into the spatio-temporal distribution of resistant isolates. Nevertheless, some improvements can still be made as there remains some room for further research. For example, the use of the Bayesian network models could be explored in more depth, especially the residuals method that was hinted upon at the end of Section 3.5.1. Indeed, incorporating time and other covariates beforehand might be a nice alternative for future analyses. In addition, different parametric and non-parametric functions of time could be investigated as well for the GEE and pattern attribution models. This can be of interest in future applications, when the included time span is much longer, thereby allowing for the estimation of more advanced functions over time.

The application of the methods was sometimes hampered by several data flaws. Some of them were mentioned throughout the report, but focus is on **survey and data recommendations** below.

In the light of the hierarchical clustering after PCA and MCA, it was discussed that the results of the PCA might be hampered by the fact that different member states employ different dilution ranges when determining MIC values. Hence, it might be advisable to **harmonise the data collection** routine across the member states **with relation to well defined dilution ranges** for reporting purposes.

With regard to the spatial/spatio-temporal analysis, it was apparent in the results that more data is needed for a more meaningful analysis. As it stands now, there are several areas (country-level or NUTS-level) without information and areas with information at only one or two time points (i.e. years). This means that any effect (spatial or temporal) incorporated in the model is estimated from only a few areas or time points. This results to possible instability in the analysis and hinders investigation/use of a more flexible time trend or spatio-temporal trend. It is therefore important to **obtain sufficient spatial and temporal coverage** in future data collections.

Furthermore, in order to get meaningful results, the way these these data are collected is quite important (sampling design). Consider for instance the NUTS-3 level. In case data is collected only at a particular area, it is not quite correct to model the spatial pattern of resistance at this level. In addition, if the analysis is to be done anyway, the obtained results need to be interpreted carefully, keeping in mind that the existence of areas without information or with zero resistance could be due to the sampling design. Hence, for the spatio-temporal analysis, collection of data needs to be improved. In case an analysis at a finer level, i.e. NUTS-3 level and weekly-level, is foreseen, then data needs to be collected at these levels. However, (complete) monthly data at NUTS-2 level is already seen to be quite informative and would allow the investigation of more flexible trends. It would also allow proper investigation of spatio-temporal risk within 1 or 2 countries as compared to having data only at country or NUTS-1 level or at yearly level.

Similarly, it was discussed that results for the source attribution models were not included due to serious lacks in the available data. Currently, 4 imputation packages were included to impute missing data. Of course, in order to be able to perform appropriate analyses, efforts should be made to reduce this amount of missing data to a minimum in future data collection procedures. In addition, these models could be extended to also include information on e.g. trade patterns, contact structures and observed prevalence of certain bacterial zoonotic agents in different species.

References

EFSA (2014). Technical specifications on randomised sampling for harmonised monitoring of antimicrobial resistance in zoonotic and commensal bacteria. EFSA Journal 2014; 12(5)3686.

Ensoy C, Rakhmawati TW, Faes C, Aerts M, 2015. Separation Issues and Possible Solutions: Part II – Comparison of different methods for separation in logistic regression. EFSA supporting publication 2015:EN-869. 176 pp.

Eurostat (2016a). Agricultural production, http://ec.europa.eu/eurostat/data/database. Accessed on 25 May 2016.

Eurostat (2016b). EU trade since 1999 by HS2,4,6 and CN8 (DS-575274), http://ec.europa.eu/eurostat/data/database. Accessed on 25 May 2016.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Downloadable from http://www-stat.stanford.edu/ tibs/ElemStatLearn/.

Husson, F. Josse, J. and Pagès, J. (2010). Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data? Technical Report – Agrocampus.

Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651-674.

Johnson, R.A. and Wichern, D.W. (2002). Applied Multivariate Statistical Analysis. 5th Edition, Prentice-Hall.

Kaufman, L., and Rousseeuw, P.J. (1990). Finding groups in data: An introduction to cluster analysis. New York: John Wiley and Sons, Inc.

Riebler, A., Sørbye, S., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. arXiv:1601.01180 [stat.ME].

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B,71(2):319-392, 2009.

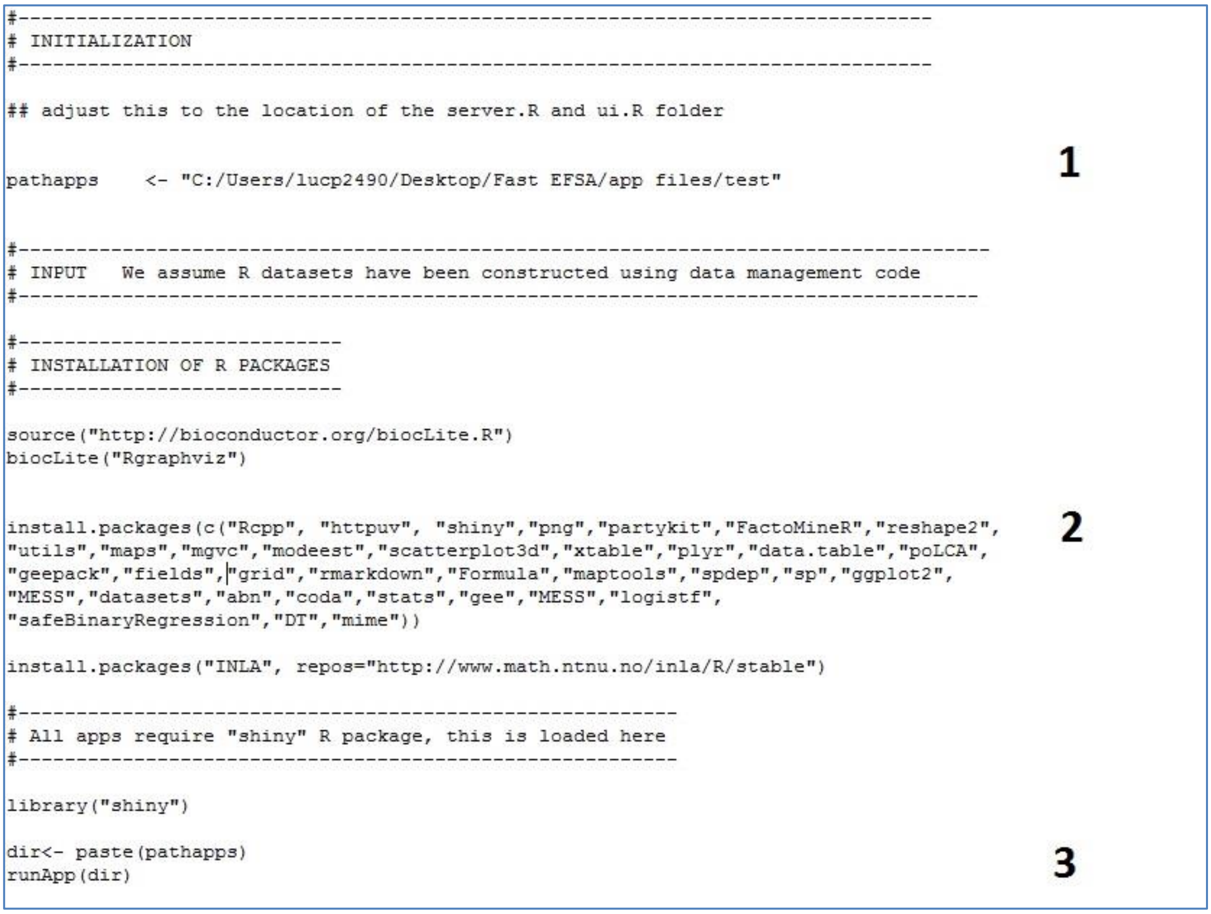
Simpson, D., Harvard, R., Martins, T., Riebler, A., and Sørbye, S. (2015). Penalising model component complexity: A principled practical approach to constructing priors. arXiv:1403.4630 [stat.ME].

Appendix A – Organization of software codes

R-code

For the R users, an interface was created in order to facilitate the analysis. This interface was constructed using the shiny package in R. A tutorial was prepared to guide the user how to use these interfaces. Below, it is briefly explained how to launch the application. For the different analyses, the reader is referred to the tutorial.

The app consists of three files, i.e. script.R, ui.R and server.R, from which the user only needs to use the script.R file (the application will also be available under the R4EU platform https://efsa.openanalytics.eu/app/MDR, for which there will not be necessary to interact with R and all steps provided below will no longer be needed, instruction manual will be accessible from the R4EU platform using the About link). When the user opens this file in the R console, the following codes appears, with a numbering (1-3) added on the figure below for additional explanations:

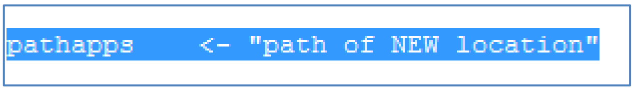


```
#------------------------------------------------------------------------------
# INITIALIZATION
#------------------------------------------------------------------------------

## adjust this to the location of the server.R and ui.R folder


pathapps    <- "C:/Users/lucp2490/Desktop/Fast EFSA/app files/test"


#-----------------------------------------------------------------------------------
# INPUT   We assume R datasets have been constructed using data management code
#----------------------------------------------------------------------------------

#----------------------------
# INSTALLATION OF R PACKAGES
#----------------------------

source("http://bioconductor.org/biocLite.R")
biocLite("Rgraphviz")


install.packages(c("Rcpp", "httpuv", "shiny","png","partykit","FactoMineR","reshape2",
"utils","maps","mgvc","modeest","scatterplot3d","xtable","plyr","data.table","poLCA",
"geepack","fields","grid","rmarkdown","Formula","maptools","spdep","sp","ggplot2",
"MESS","datasets","abn","coda","stats","gee","MESS","logistf",
"safeBinaryRegression","DT","mime"))

install.packages("INLA", repos="http://www.math.ntnu.no/inla/R/stable")

#---------------------------------------------------------
# All apps require "shiny" R package, this is loaded here
#---------------------------------------------------------

library("shiny")

dir<- paste(pathapps)
runApp(dir)
```

In part 1, the user needs to specify the location where the ui.R and the server.R files are saved. More specifically, in the example above, the ui.R and server.R files are stored in a folder which has location with path *"C:/Users/lucp2490/Desktop/Fast EFSA/app files/test".* This path needs to be changed into the location where the specific user has saved the files. Hence, the user needs to select this path and overwrite it with the new one by manually typing the correct new path (or simply by copy-pasting).

Important note: the path needs to be specified using forward slashes.

After the location has been adjusted, the user needs to select the line

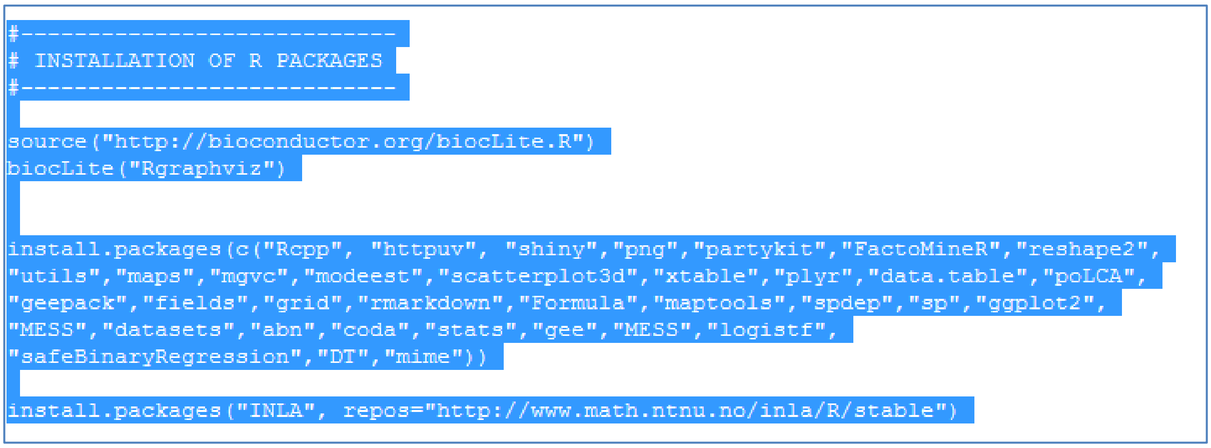


```
pathapps     <- "path of NEW location"
```

The present document has been produced and adopted by the bodies identified above as author. This task has been carried out exclusively by the author in the context of a contract between the European Food Safety Authority and the author, 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 author.

and press "F5" to run this part of the code.

While running the app, some important R packages are used internally to provide the desired output of the different analyses. Therefore, it is required to make sure that these packages are installed on the user's PC. This is taken care of in part 2. This part of the code can be executed by selecting all lines beneath the "Installation of R packages" title

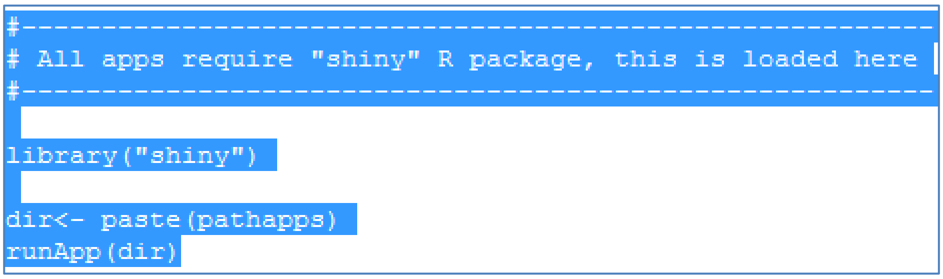


```
#----------------------------
# INSTALLATION OF R PACKAGES
#----------------------------

source("http://bioconductor.org/biocLite.R")
biocLite("Rgraphviz")


install.packages(c("Rcpp", "httpuv", "shiny","png","partykit","FactoMineR","reshape2",
"utils","maps","mgvc","modeest","scatterplot3d","xtable","plyr","data.table","poLCA",
"geepack","fields","grid","rmarkdown","Formula","maptools","spdep","sp","ggplot2",
"MESS","datasets","abn","coda","stats","gee","MESS","logistf",
"safeBinaryRegression","DT","mime"))

install.packages("INLA", repos="http://www.math.ntnu.no/inla/R/stable")
```

and pressing "F5" to run the selected lines.

Running this part of the code will guide the user through the installation process. First, a screen will pop up, asking to specify the CRAN mirror. Chose option 2: 0-Cloud and press "OK". This will install the required packages. It is possible that some packages were already installed before and an older version of the package is already available. In this case, R will pop a question "Update all/some/none? [a/s/n]:". In this case, the user is advised to type "a" after the colon and press "Enter", after which those already installed packages are updated to the most recent version.

Finally, in part 3, the shiny R package is loaded and the app is launched. Select the corresponding lines



```
#------------------------------------------------------
# All apps require "shiny" R package, this is loaded here
#------------------------------------------------------

library("shiny")

dir<- paste(pathapps)
runApp(dir)
```

and press "F5". The app will launch in a tab of the default browser of the user's PC.

Sas-code

For the SAS users, macros were created for the generalised estimating equations and the hierarchical clustering after PCA. In these macro's, the user needs to specify which antimicrobial and which sample type are to be selected and the according analysis will be performed.

The user will receive a working directory containing the following items:

Contents of the working directory are:

- The SAS macros (GEE_heatmap.sas PCA.sas): this file should basically never be touched unless for editing purposes.
- The main analysis file (Analysis_PCA_GEE.sas): contains the calls of the macro for different bacteria and antibiotics. This is the program you need to run
- AnaData folder is the analysis folder containing the SAS datasets imported from CSV format
- CSVData folder is the folder containing the would be raw CSV files exported from R at this point, where data manipulation was performed.

For example, the code that needs to be run for the GEE analysis looks as follows:

```
%global root;
%let root=%qsubstr(%sysget(SAS_EXECFILEPATH), 1,
%length(%sysget(SAS_EXECFILEPATH)) -
(%length(%sysget(SAS_EXECFILEname)) + 1)); * capture root directory
;
%put NOTE: Current Working Directory is &root;
libname anadata "&root\AnaData";
options minoperator mindelimiter=",";
%inc "&root\GEE_heatmap.sas";              ***** running this and
above lines once is sufficient ;;



%GEE_heatmap(import=0,bacterium=E_coli,samptype4=broilers,
imagefmt=tiff, lowerCorr=0);
```

SAS will write the output (i.e. figures) in the same working directory where the programs are resident. Make sure the listing option in SAS is indicated.

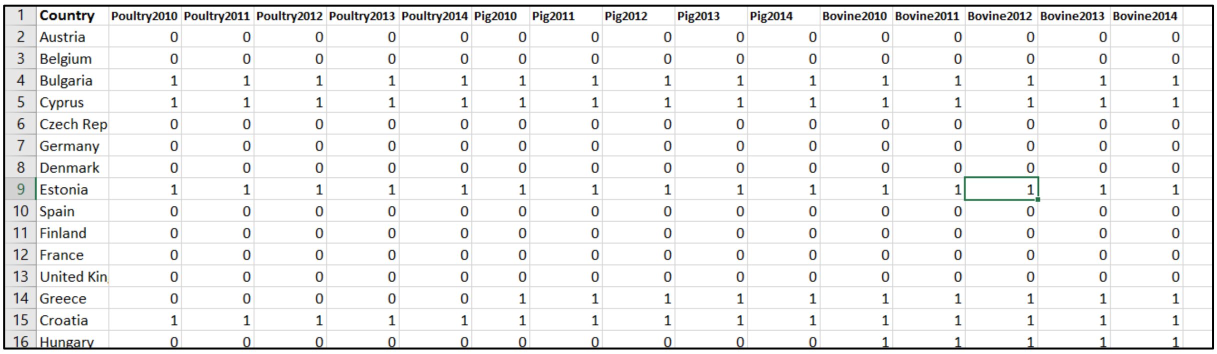
The latter can be done as follows: Tools->Options->Preferences, select tab "Results" and highlight the "Create Listing" tick box.

Appendix B – Input files

Spatio-temporal analysis

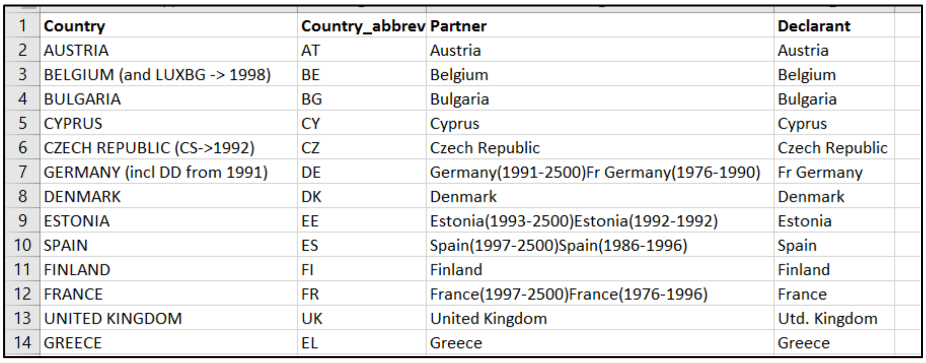
Below are screenshots of how the input files for the weighted analysis and analysis with trade data looks like:

CountryProd.csv



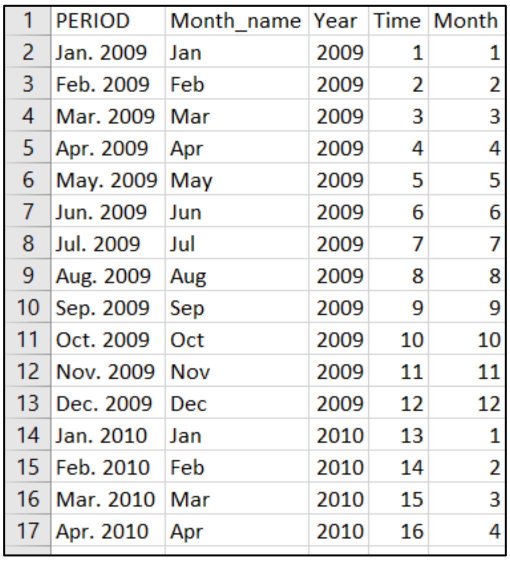
1	Country	Poultry2010	Poultry2011	Poultry2012	Poultry2013	Poultry2014	Pig2010	Pig2011	Pig2012	Pig2013	Pig2014	Bovine2010	Bovine2011	Bovine2012	Bovine2013	Bovine2014
2	Austria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Belgium	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Bulgaria	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	Cyprus	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	Czech Rep	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Germany	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Denmark	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Estonia	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	Spain	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Finland	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	France	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	United Kin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Greece	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	Croatia	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	Hungary	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Country_abbrev.csv



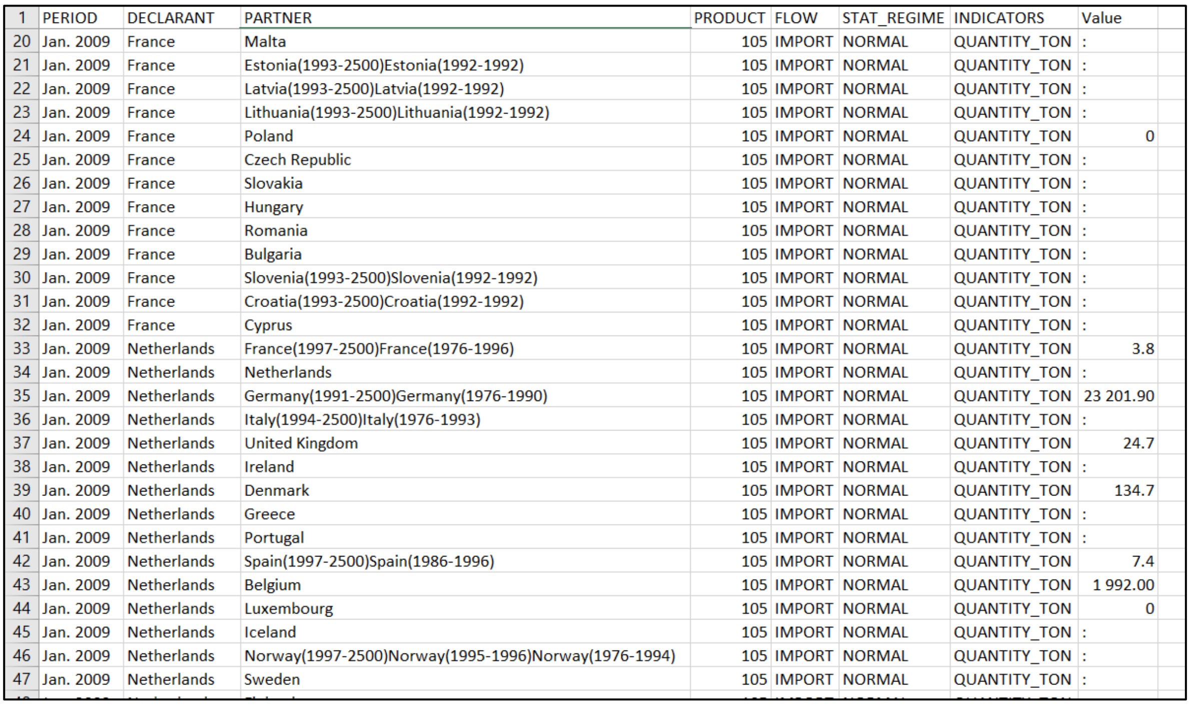
1	Country	Country_abbrev	Partner	Declarant
2	AUSTRIA	AT	Austria	Austria
3	BELGIUM (and LUXBG -> 1998)	BE	Belgium	Belgium
4	BULGARIA	BG	Bulgaria	Bulgaria
5	CYPRUS	CY	Cyprus	Cyprus
6	CZECH REPUBLIC (CS->1992)	CZ	Czech Republic	Czech Republic
7	GERMANY (incl DD from 1991)	DE	Germany(1991-2500)Fr Germany(1976-1990)	Fr Germany
8	DENMARK	DK	Denmark	Denmark
9	ESTONIA	EE	Estonia(1993-2500)Estonia(1992-1992)	Estonia
10	SPAIN	ES	Spain(1997-2500)Spain(1986-1996)	Spain
11	FINLAND	FI	Finland	Finland
12	FRANCE	FR	France(1997-2500)France(1976-1996)	France
13	UNITED KINGDOM	UK	United Kingdom	Utd. Kingdom
14	GREECE	EL	Greece	Greece

Period.csv



1	PERIOD	Month_name	Year	Time	Month
2	Jan. 2009	Jan	2009	1	1
3	Feb. 2009	Feb	2009	2	2
4	Mar. 2009	Mar	2009	3	3
5	Apr. 2009	Apr	2009	4	4
6	May. 2009	May	2009	5	5
7	Jun. 2009	Jun	2009	6	6
8	Jul. 2009	Jul	2009	7	7
9	Aug. 2009	Aug	2009	8	8
10	Sep. 2009	Sep	2009	9	9
11	Oct. 2009	Oct	2009	10	10
12	Nov. 2009	Nov	2009	11	11
13	Dec. 2009	Dec	2009	12	12
14	Jan. 2010	Jan	2010	13	1
15	Feb. 2010	Feb	2010	14	2
16	Mar. 2010	Mar	2010	15	3
17	Apr. 2010	Apr	2010	16	4

TradData.csv



1	PERIOD	DECLARANT	PARTNER	PRODUCT	FLOW	STAT_REGIME	INDICATORS	Value
20	Jan. 2009	France	Malta	105	IMPORT	NORMAL	QUANTITY_TON	:
21	Jan. 2009	France	Estonia(1993-2500)Estonia(1992-1992)	105	IMPORT	NORMAL	QUANTITY_TON	:
22	Jan. 2009	France	Latvia(1993-2500)Latvia(1992-1992)	105	IMPORT	NORMAL	QUANTITY_TON	:
23	Jan. 2009	France	Lithuania(1993-2500)Lithuania(1992-1992)	105	IMPORT	NORMAL	QUANTITY_TON	:
24	Jan. 2009	France	Poland	105	IMPORT	NORMAL	QUANTITY_TON	0
25	Jan. 2009	France	Czech Republic	105	IMPORT	NORMAL	QUANTITY_TON	:
26	Jan. 2009	France	Slovakia	105	IMPORT	NORMAL	QUANTITY_TON	:
27	Jan. 2009	France	Hungary	105	IMPORT	NORMAL	QUANTITY_TON	:
28	Jan. 2009	France	Romania	105	IMPORT	NORMAL	QUANTITY_TON	:
29	Jan. 2009	France	Bulgaria	105	IMPORT	NORMAL	QUANTITY_TON	:
30	Jan. 2009	France	Slovenia(1993-2500)Slovenia(1992-1992)	105	IMPORT	NORMAL	QUANTITY_TON	:
31	Jan. 2009	France	Croatia(1993-2500)Croatia(1992-1992)	105	IMPORT	NORMAL	QUANTITY_TON	:
32	Jan. 2009	France	Cyprus	105	IMPORT	NORMAL	QUANTITY_TON	:
33	Jan. 2009	Netherlands	France(1997-2500)France(1976-1996)	105	IMPORT	NORMAL	QUANTITY_TON	3.8
34	Jan. 2009	Netherlands	Netherlands	105	IMPORT	NORMAL	QUANTITY_TON	:
35	Jan. 2009	Netherlands	Germany(1991-2500)Germany(1976-1990)	105	IMPORT	NORMAL	QUANTITY_TON	23 201.90
36	Jan. 2009	Netherlands	Italy(1994-2500)Italy(1976-1993)	105	IMPORT	NORMAL	QUANTITY_TON	:
37	Jan. 2009	Netherlands	United Kingdom	105	IMPORT	NORMAL	QUANTITY_TON	24.7
38	Jan. 2009	Netherlands	Ireland	105	IMPORT	NORMAL	QUANTITY_TON	:
39	Jan. 2009	Netherlands	Denmark	105	IMPORT	NORMAL	QUANTITY_TON	134.7
40	Jan. 2009	Netherlands	Greece	105	IMPORT	NORMAL	QUANTITY_TON	:
41	Jan. 2009	Netherlands	Portugal	105	IMPORT	NORMAL	QUANTITY_TON	:
42	Jan. 2009	Netherlands	Spain(1997-2500)Spain(1986-1996)	105	IMPORT	NORMAL	QUANTITY_TON	7.4
43	Jan. 2009	Netherlands	Belgium	105	IMPORT	NORMAL	QUANTITY_TON	1 992.00
44	Jan. 2009	Netherlands	Luxembourg	105	IMPORT	NORMAL	QUANTITY_TON	0
45	Jan. 2009	Netherlands	Iceland	105	IMPORT	NORMAL	QUANTITY_TON	:
46	Jan. 2009	Netherlands	Norway(1997-2500)Norway(1995-1996)Norway(1976-1994)	105	IMPORT	NORMAL	QUANTITY_TON	:
47	Jan. 2009	Netherlands	Sweden	105	IMPORT	NORMAL	QUANTITY_TON	: