SCIENTIFIC REPORT

Statistical analysis of temporal and spatial trends of zoonotic agents in animals and food

Part I: Critical review of the statistical analysis carried out on the Community Summary Report 2006 data [1]

Scientific Report of the Task Force on Zoonoses Data Collection

(Question N° EFSA-Q-2008-264)

Adopted on 31 March 2009

SUMMARY

The European Union (EU) Member States (MSs) collect data on zoonoses, zoonotic agents, antimicrobial resistance and food-borne outbreaks. The European Food Safety Authority (EFSA) is assigned the tasks of examining and analysing the data collected and preparing the Community Summary Report (CSR).

Statistical analyses were carried out for the first time on the 2006 data. The objective was to evaluate the significance of temporal variations in the EU-level prevalence of zoonotic agents in animals and in the EU-level proportion of positive food units. Indeed, trend analysis provides information on developments in the Community and Member States and it may give information on the effects of Community or national control measures to reduce the occurrence of zoonotic agents.

[1] For citation purposes: Report of the Task Force on Zoonoses Data Collection on the statistical analysis of temporal and spatial trends of zoonotic agents in animals and food. Part I: critical review of the statistical analysis carried out on the Community Summary Report 2006 data. *The EFSA Journal* (2009) 253, 1-77.

A critical review of the statistical analysis of time trends in 2006 was carried out. Two critical characteristics of the data, which can affect the validity of the analysis, were identified: 1) varying sampling probability of epidemiological units among MSs and across years in the same MS, corresponding to disproportionate stratified sampling, where MSs are considered as strata; 2) correlation among observations on zoonotic agents in the same MS in subsequent years.

The SURVEYLOGISTIC procedure in the SAS system, which was used in the 2006 analysis, is specifically suited to analyse survey data taking into account sampling design. In order to account for disproportionate stratified sampling, weights were calculated, per MS and year, as the reciprocal of the sampling ratio (number of units in the MS population, divided by number of sampled units). The choice of the population to be used to calculate weights is critical since weighting may strongly affect parameter estimates and statistical significance of trends. Moreover, finite population correction could be applied if considered as appropriate (design-based approach).

The CLUSTER statement was used to take into account correlation among observations from the same MS. Consequently, the standard error for the parameter corresponding to the effect of year on the probability of a positive result is inflated and the probability of statistical significance reduced. In SURVEYLOGISTIC, the CLUSTER statement is specifically aimed to deal with a design aspect, namely the random selection of clusters. Therefore, in the absence of a specific sampling design, other techniques, such as generalised estimating equations (GEE, REPEATED statement in the GENMOD procedure) would be more appropriate. However, from a practical point of view, the approach to the analysis of time trends used by EFSA was valid, and results are the same or very similar to those obtained by using GEE. By using techniques for correlated data, the effects of between MS harmonisation problems were reduced, and the analysis was valid as long as harmonisation was constant within MSs, across years.

Further developments of trend analysis are introduced. The Bayesian approach, for example, is a flexible and powerful set of tools in the analysis of time trends, where model parameters are not assumed to be a fixed unknown constant to be estimated, but instead they are seen as random variables. The distribution of parameters (called posterior distribution) quantifies all uncertainty and information present in the model and the data, but also possibly from other prior sources of knowledge available (defining prior distribution of parameters).

Investigating the spatial and temporal distribution of zoonotic agents and diseases, performing specific spatial and temporal analyses is a powerful approach in the study of temporal and spatial trends of zoonoses. Therefore, spatial and space-time clustering analyses of zoonoses is an important field of further development carried out by EFSA. A review of the most commonly used statistical methods in spatial epidemiology is included, together with data needs for meaningful spatial analysis at EU level.

Results of complex statistical analysis and certain details on the methods need to be communicated to risk managers and the general public. In a section of this report, basic principles on how to communicate in a comprehensible way are presented, and a worked-out example shown.

In a part II report the use of different software and statistical packages (including open source) will be compared in the analysis of temporal trends. Moreover, the analysis of spatial trends in zoonotic agents will be developed through a case study.

Keywords: zoonoses, statistical analysis, trend analysis, spatial analysis, data analysis

TABLE OF CONTENTS

BACKGROUND

Directive 2003/99/EC on the monitoring of zoonoses and zoonotic agents lays down the Community procedure for the monitoring and collection of information on zoonoses, which obliges Member States (MSs) to collect relevant and, where applicable, comparable data of zoonoses, zoonotic agents, antimicrobial resistance and food-borne outbreaks. The European Food Safety Authority (EFSA) is assigned the tasks of examining and analysing the data collected and preparing the Community Summary Report (CSR).

Data collected in the framework of Directive 2003/99/EC relate to the occurrence of zoonotic agents isolated from animals, food, and feed, as well as to antimicrobial resistance in these agents. Information concerning zoonoses cases in humans and related antimicrobial resistance is derived from Communicable Disease Networks laid down by Council Decision No 2119/98/EC that are currently coordinated by the European Centre for Disease Prevention and Control (ECDC). EFSA, in close collaboration with ECDC, has published three CSRs on Trends and Sources of Zoonoses, Zoonotic Agents, Antimicrobial Resistance and Food-borne Outbreaks in the European Union, the latest from data for 2007. EFSA is assisted by the Zoonosis Collaboration Centre (ZCC) in the Danish Technical University in the drafting of the CSR.

In the CSR, data submitted by MSs are summarised and presented in tables, graphs, and thematic maps. Moreover, for the first time, statistical analyses were carried out on 2006 data to estimate the significance of temporal variations in the EU-level prevalence of zoonotic agents in animals and in the proportion of positive food units.

Difficulty in comparing country data has been acknowledged in the annual zoonoses data collection. In fact, in certain cases, sampling strategies and laboratory methods may be not fully harmonised. However, it is reasonable to assume that, within countries, data are relatively comparable between reporting years (unless the monitoring systems have been changed considerably). Further harmonisation of data received from MSs would improve the possibilities of analysing and interpreting information at Community level. Moreover, detailed information on geographic locations of tested epidemiological units (animals, herds, food) and of underlying populations would allow the analysis of spatial trends.

The following of trends, both at Community and MS specific levels, may provide information on the effects of Community or national control measures to reduce the occurrence of zoonotic agents. Moreover, hypotheses can be generated on the effects of potential risk factors, either natural or associated with human activities. Finally, the emergence or re-emergence of pathogens can be detected by analysing trends.

A valid statistical analysis is necessary to estimate epidemiological parameters and to obtain useful Community level information from the data provided by MSs. Furthermore, testing statistical significance is necessary to draw conclusions on whether observed trends are due to chance alone or represent real changes in population. Data peculiarities, associated with sampling design and correlation among epidemiological units, need to be taken into account in order to avoid misleading conclusions. In addition, different statistical techniques, as well as software for statistical analysis can be considered, and results obtained by different approaches can be compared and discussed.

Each year, MSs report a substantial amount of data on zoonotic agents in a wide range of different animal species and food categories, according to the framework of the zoonoses Directive 2003/99/EC. As data collection and management result from considerable efforts made by MSs and all involved institutions, further development in data analysis should be considered in order to make the best possible use of available data and of the existing data collection system. Changes in surveillance and sampling systems, including the adoption of targeted and risk-based sampling, should also be considered for their effect on data harmonisation and in view of the application of appropriate statistical techniques.

TERMS OF REFERENCE

Directive 2003/99/EC sets out the Community procedure for the monitoring and collection of information on zoonoses, which obliges Member States to collect relevant, and where applicable, comparable data on zoonoses, zoonotic agents, antimicrobial resistance and food-borne outbreaks. EFSA is assigned the tasks of examining the data collected and preparing the Community Summary Report.

EFSA applies epidemiological and statistical methods in the examination of data and, in particular, seeks to analyse trends over subsequent years as well as geographical distributions. The statistical techniques that can be used to estimate changes in time and space in the frequency of zoonotic agents in animals and food should be carefully checked/evaluated in terms of their validity for available data and for specific objectives.

Therefore, EFSA asked its Task Force on Zoonoses Data Collection to consider the statistical methods to be used in the analyses of the annual zoonoses data.

The Task Force on Zoonoses Data Collection is asked to consider and issue reports on:

- the validity of statistical methods that are currently used to analyse time and spatial trends of zoonotic agents in animals and food in the Community Summary Report;
- the identification of the most appropriate statistical methods for analysing temporal trends taking into account that data collection is not fully harmonised for different zoonotic agents and that information from some Member States is unavailable;
- the identification of the most appropriate statistical methods for analysing the geographical distribution of zoonotic agents taking into account the above-mentioned factors;
- the perspectives for the harmonisation of data collection and the monitoring of zoonotic agents in MSs in view of improving data;
- the improved communication of statistical analysis results.

This report is the part I report that covers the critical review of the statistical analyses carried out, preliminary considerations of the further developments in trend and spatial analyses and communication of the methods and results from statistical analyses. The report part II will include a comparison of different software and statistical packages (including open source) in the analysis of temporal and spatial trends. Moreover, the analysis of spatial trends in zoonotic agents will be developed through a case study.

ACKNOWLEDGEMENTS

The Task Force wishes to thank the members of the Working Group for the preparation of this report: Marc Aerts, Giusi Amore, Billy Amzal, Pierre-Alexandre Beloeil, Frank Boelaert, Annamaria Conte, José Cortiñas, Renata Del Rosario, Hanne-Dorthe Emborg, Klemens Fuchs, Uriel Kitron, Alessandro Mannelli, Giovanna Nicolini, Monica Pratesi, Francesca Riolo, Robin Sayers, Tobias Verbeke, and Therese Westrell.

CONSIDERATION

1. Current data collection system

The information on zoonoses, zoonotic agents and antimicrobial resistance has to be submitted to the European Commission (EC) each year in accordance with Community legislation. Since 2005, this has been done through a web-based reporting system, and information is then stored in a large database, hosted and maintained by EFSA.

The web-based reporting system allows reporting officers in each MS to enter data concerning zoonotic agents in foodstuffs, animals and feedstuffs collected through a variety of different systems, i.e. official monitoring and control programmes, HACCP and own-check programmes, surveys and baseline studies. Data are entered into the database using a number of pre-defined tables, pick-lists for variables, and text templates. However, the reporting officer may add new lines and footnotes to the tables, if relevant.

Based on public health priorities[2,3,4] the Directive 2003/99/EC lists eight zoonoses and zoonotic agents to be monitored, and data are collected on a mandatory basis for the following zoonotic agents: *Salmonella*, thermophilic *Campylobacter*, *Listeria monocytogenes*, verotoxigenic *E. coli*, *Mycobacterium bovis*, *Brucella*, *Trichinella* and *Echinococcus*. Furthermore, mandatory reported data includes antimicrobial resistance in *Salmonella* and *Campylobacter* isolates, food-borne outbreaks and susceptible animal populations. Also, based on epidemiological situations in MSs, data have been reported on the following zoonotic agents: *Yersinia*, *Lyssavirus* (the cause of rabies), *Toxoplasma*, *Cysticerci*, *Sarcocystis*, *Coxiella burnetti* (the cause of Q fever), *Chlamydia psittaci,* (the cause of psittacosis) and *Leptospira* spp. Data on antimicrobial resistance in indicator *E. coli* and *Enterococci* isolates are also reported. Since 2006, MSs have also submitted data concerning other microbiological contaminants such as histamine, Staphylococcal enterotoxins and *Enterobacter sakazakii*, for which food safety criteria are set out in the Community legislation 1441/2007/EC. In 2007, data were submitted by 27 MSs and two non-MSs (Norway and Switzerland).

Decision 2119/98/EC on setting up a network for epidemiological surveillance and the control of communicable diseases in the Community, as complemented by Decision 2000/96/EC on the diseases to be progressively covered by the network, established data collection on human communicable diseases from MSs. Since 2005, ECDC has provided data on zoonotic infections in humans, in conjunction with their analyses, for the CSR. In 2007, the data used for analysis were derived from several disease networks: the new European Surveillance System (TESSy), which has been implemented and is maintained by ECDC, and one Dedicated Surveillance Network (DSN): Euro-TB. In 2007, data on human zoonoses cases were received through the Communicable Disease Networks from all 27 MSs and additionally from four non-MSs: Iceland, Liechtenstein, Norway and Switzerland.

Once the annual data has been submitted by MSs (deadline 31 May), data are collated, analysed and presented in the CSR. The CSR is prepared by EFSA in collaboration with ECDC and ZCC and published on EFSA's homepage each year in December.

[2] Opinion on Zoonoses from the Scientific Committee on Veterinary Measures relating to Public Health (Adopted 12 April 2000)

[3] Opinion on Proposal for a directive on the monitoring of zoonoses and zoonotic agents from the Economic and Social Committee (OJ C 94, 18.4.2002 p.18)

[4] Opinion on the Resistance to antibiotics as a threat to public health from the Economic and Social Committee (OJ C 407, 28.12.1998 p 7)

2. Objectives of this report

The primary objective of report part I is to examine carefully the statistical analysis of temporal trends of zoonotic agents in animals and food that was carried out on data that were collected in 2006 for the CSR. The statistical analysis was reviewed regarding methods which were used to take into account the two critical characteristics of the data:

1. the disproportionate sampling of units, such as animals, groups of animals, or food, from different MSs;
2. the potential correlation among the infection status of units sampled in the same MS, during subsequent years.

To aid clarity and transparency, worked examples of different approaches are presented and discussed.

Another important objective of this report is to present basic principles of survey methodology relevant for the interpretation and critical review of statistical analysis. Moreover, needs for the harmonisation of current data collection can be identified. A small section of report part I is devoted to the communication of statistical analyses among professionals with different levels of statistical knowledge.

An additional objective is to introduce issues that will be developed further in report part II, including spatial analysis and Bayesian statistics.

3. Statistical analysis

3.1 Introduction to survey methodology

In this section, basic concepts and principles of survey methodology and related statistical techniques are set out which are useful in the following sections on the statistical analysis of temporal trends. This section is not meant to be exhaustive, and each concept and method can be studied further. Guidance on field surveys within EFSA's remit, and a glossary of terms can be found in a specific EFSA report[5].

Target population and sampling

A sample survey is a study where a sample of elements or elementary units (herds, animals, food batches) or, in this case, of epidemiological units, is selected from a larger, well-defined population. Observations are then carried out on elements of the sample, in order to draw conclusions on the population at large (target population) (Levy and Lemeshow, 1992; Boelaert, 2003).

In surveys, a sample is selected from the sampling frame (also referred to as 'frame population'), that is the set of all elements that could possibly be drawn. In other words it is the part of the target population (the one to be generalised to) which is available for sampling. As an example, should a sample of broiler carcasses be selected from all of those which are produced in a country during a year (target population), access may only be given to those broiler carcasses which are processed in certain slaughterhouses, during certain periods of the year (sampling frame).

In epidemiology, surveys are carried out both for estimating descriptive measures of disease frequency, such as incidence and prevalence, and for estimating the effects of risk factors on diseases, by relative risk, odds ratio, and other measures of association.

Statistical inference and error

In a survey, a parameter of interest is measured in the sample, with the objective of drawing conclusions on the value of the parameter in the target population. Therefore, the value that is obtained for the sample is considered an estimate of the "real" value in the target population. For example, the proportion of units in a sample which are infected with a zoonotic agent (corresponding to prevalence for animals or proportion positive for food units) is calculated as an estimate of the proportion of infected units in the target population.

The difference between the estimate and the real value - the error - can have two components (causes):

1. random selection of individuals in the sample, in the presence of variability of probability of infection in the population (random error or precision); or

[5] Report of Task Force on Zoonoses data collection on guidance document on good practices for design of field surveys, The EFSA Journal (2006), 93, 1-24.

2. selection of individuals which are characterised by a greater or smaller probability of infection in comparison with the probability in the population (systematic error or bias).

Random error

This is largely affected by the size of the sample relative to the target population and by the sampling design (or plan). It is an error attributed to the fact that our investigation is limited to a random sample of the population, rather than to the whole population. In fact, it is often called sampling error. A useful way to express the extent of the sampling error is to report the confidence interval (CI) of a parameter estimate. In a practical definition, the CI corresponds to the interval covering the real value of the parameter in the population at a pre-assigned level of confidence.

Systematic error

This can be due to one or more groups of units which are characterised by a particularly small or great risk of disease, being over- or under-represented in the sample, in comparison with the target population. For example, in the collection of data on *Salmonella* spp. in flocks of laying hens, different MSs may select samples from flocks corresponding to different fractions of their populations. In this way, the MS composition of the sample of flocks would be different from the composition of the EU population. Consequently, if the MSs that are over-represented in the sample are characterised by a relatively high prevalence of *Salmonella* spp. in flocks of laying hens, the estimate of EU level prevalence will be systematically over-estimated. The issue of systematic error is discussed further below in the section on stratified sampling.

Coverage error is a specific type of systematic error which is likely to happen when the sampling frame (frame population) is not identical to the target population. Two types of coverage errors generally arise when such correspondence does not exist: under-coverage and over-coverage. Units that are in the target population but not in the frame population (sampling frame) constitute under-coverage. On the other hand, units that are in the frame population but not in the target population constitute over-coverage. It follows that under-coverage units have zero probability of being selected for any sample drawn from the population. If the risk of disease in these units differs systematically from the risk in other units in the target population, the estimate of prevalence will be biased (Särndal et al. 1992).

Assumptions in the statistical analysis of survey data and confidence intervals

In the majority of cases, data that are submitted to EFSA within the framework of the CSR derive from analytical tests for the detection of infection by zoonotic agents in animals and food. Results of these tests can be either positive or negative and, therefore, are suitable for analysis by statistical methods for binary outcomes. Statistical details on the estimation of proportions and confidence intervals are reported in Appendix B at the end of this report.

When estimating the proportion of units that are infected with a zoonotic agent, the CI can be calculated by using, for example, the *binom.test* function in the R package (http://cran.r-project.org). This is based on the binomial distribution (exact binomial CI) and it is valid even for low (<0.2) and high (>0.8) proportions. Contrary to the asymptotic CI for a proportion, the exact CI for proportion holds for any sample size *n* (whereas the asymptotic CI requires a minimum-sized sample).

The construction of a CI for a proportion is based on two basic binomial assumptions:

1. the sample consists of *n* independent copies of the population variable of interest (infection/contamination indicator); and

2. the probability of the event of interest (infection/contamination) is identical for all sample units (no heterogeneity in infection probability).

It is, therefore, assumed that samples of independent and identically distributed (i.i.d.) random variables (observations) are drawn from an infinite population. When the above assumptions are not met, specific statistical techniques have to be adopted in order to avoid systematic error and to obtain valid CIs. The same assumptions must be taken into account even in the analysis of risk factors for diseases, by, for example, **logistic regression**, which is the most commonly used statistical model for binary outcomes (Hosmer and Lemeshow, 2000). Details on statistical methodology are reported in Appendix B at the end of this report.

Finite *vs* infinite target population

In a survey, if the sampling ratio is not small (i.e., if the sample is a relevant part of the target population), **finite population corrections** can be applied in estimating parameters, and the resulting CIs are relatively narrow in comparison to the situation when the population can be considered as infinite. It is important to note that this correction limits the inference to a specific target population which is, therefore, the only interest of the investigator. In this approach, also referred to as “design-based” inference, the values of the variable of interest in the population units are considered fixed values, even if not all are known. The randomness from sample to sample derives from the random selection of any of the other possible samples from the finite population of interest. This is commonly adopted in official control plans where the objective is, for example, to estimate prevalence or to declare freedom of infection (or a certain level of infection) in a well-specified population.

When the population can be considered as **infinite**, the so-called “model-based” inference is referred to, which is based on the randomness of the i.i.d. random variables constituting the population units. According to this approach, it can be considered that even the prevalence of infection in all animals in a country at a certain time is just an estimate of the risk of infection in the country and it is subject to random variation. Consequently, all broiler carcasses produced in a country during a one-year period are but one of the potentially infinite “outcomes” of the broiler production and slaughtering process taking place in that country and in that year. Accordingly, the proportion of *Camplylobacter* spp. contaminated broiler carcasses, is but one of the possible results of a complex combination of factors acting during all phases of the poultry meat production. This type of approach is most common in research.

PROC SURVEYLOGISTIC (SAS, 2009) was specifically developed for the analysis of data from surveys including finite population correction. For a worked example, see Appendix A: Worked examples of the SAS procedure application, Example 2. SURVEYLOGISTIC is very similar to PROC LOGISTIC (used for ordinary logistic regression) as it also allows to build a predictor model and to select a link function, but the inference is design-based rather than model-based.

Effect of the type of sampling

When describing random error and CI calculation, it was assumed that in this situation all units in the target population (if it coincides with the population frame) have the same probability of being selected to be part of the sample. This corresponds to a **simple random sampling**. Only two other types of sampling will be referred to below which are most relevant for this report: stratified and cluster sampling (Särndal et al. 1992; Levy and Lemeshow, 1992).

Stratified sampling

In stratified sampling, the units are classified into distinct groups, or strata, and a simple random sample is then drawn from each stratum. Stratified sampling is usually carried out when strata are expected to differ in the parameter of interest and, in this case, it can provide more precise estimates (narrower CI) than those that can be obtained by simple random sampling. For example, prevalence of an infection may differ in different types of farming such as total confinement of animals, or free-range farms. Also differences among MSs in animal production and occurrence of the agent often justify the MSs to be considered as different strata. Separate, stratified sampling can be carried out in each farming type and results from the sample are used to obtain a more precise prevalence estimate. Moreover, stratum specific estimates can be, of course, obtained by stratified sampling.

When each stratum is equally represented in the sample as in the population, the stratified sampling is defined as proportionate (**proportionate stratified sampling**). In this case, the calculation of prevalence of infection can be obtained, as in the case of simple random sampling, as the ratio between positive and tested units. Conversely, if, for example, a certain type of farm constitutes 10% of the population and it is over-represented in the sample, where it accounts for 50% of the units (**disproportion stratified sampling**), the prevalence estimate can be affected by systematic error if prevalence is calculated as in random sampling.

Stratified sampling may have an impact on the point estimates of parameters measuring the effects of risk factors as well as on their standard errors. Therefore, disproportionate stratified sampling may also lead to complications in risk factor analysis. A common way to account for disproportionate stratified sampling in the phase of statistical analysis (after the sample is collected) is the application of **weights** to each observation in the sample. The role of weighting is to reconstruct the composition of the target population in the sample.

For instance, if the same number of flocks of broilers was sampled in MSs with different broiler population sizes, MSs with small populations would be over-represented in the sample, whereas large MSs would be under-represented. A common and simple weight is the reciprocal of the sampling proportion for flocks, which can be calculated as the number of flocks constituting the population of a MS divided by the number of sampled flocks in the same MS. Flocks from large MSs, where a smaller fraction of the population was sampled, would, therefore, have a greater weight than flocks from MSs with smaller populations. In this way, MSs would be represented, in the sample, in the same way as in the population. Another example, including graphical representation and step by step calculations, is set out in Section 6 Communication of methods and results of statistical analysis.

In the SAS system, procedures such as SURVEYMEANS, SURVEYFREQ, SURVEYREG and SURVEYLOGISTIC were specifically developed to analyse data from stratified sampling. In R, the *survey* package is available. Other procedures, such as GENMOD, also allow the inclusion of weights to account for disproportionate stratified sampling. Worked examples of weighting, including SAS programming codes, are presented in Appendix A: Worked examples of the SAS procedure application, Examples 3 and 6.

Cluster sampling

In cluster sampling, units are not individually selected from a sampling frame, but groups, (clusters) of units are primarily sampled. Individual units are subsequently examined from each cluster. Units belonging to the same cluster cannot be considered as independent, e.g. cluster sampling can be carried out to detect zoonotic agents in poultry flocks (epidemiological units) that are grouped in holdings (clusters). Poultry flocks belonging to the same holding share risk factors for the introduction of *Salmonella*. Moreover, transmission of *Salmonella* spp. is more likely to occur among flocks belonging to the same holding than among flock of different holdings. Ignoring such non-independence, or correlation, would lead to a relatively narrow CI of prevalence of infected flocks. However, such a very precise estimate would be non-valid.

Cluster sampling in risk factor analysis

A further example of non-independent observations arising from cluster sampling can be seen in pigs slaughtered in the same slaughterhouse. The pigs share a similar (correlated) risk for the contamination of carcasses with *Salmonella* spp., due to common conditions including, among others, the bacterial contamination of slaughterhouse environment, the hygiene during slaughter and subsequent processing, the slaughter techniques (EFSA, 2008).

When estimating the effects of risk factors for *Salmonella* spp., if individual pig carcasses were used as the units of statistical analysis without taking into account non-independence, some risk factors might result as having a significant effect. Such a conclusion may, however, be non-valid due to an under-estimate of the standard error and, consequently, of the confidence interval of the parameter measuring the association between the risk factor and the outcome: *Salmonella* spp. infection. In fact, the effective sample size of the study was reduced because of the non-independence of observations.

Non-independence of observations may arise in situations other than cluster sampling as discussed so far. **Repeated observations in time** on the same units, for example, also result in correlation among outcomes (Diggle et al. 2002).

There are several ways to take into account clustering or correlation among observations when estimating prevalence and in risk factor analysis. One way of correcting for correlation, which is used in SURVEYLOGISTIC, is by means of computing a so-called **design effect**. Generally, the design effect is a factor comparing the precision under simple random sampling with the precision of the actual design. Standard errors, computed as if the design had been simple random sampling, can then be inflated using the design effect. As a consequence, the probability of finding significant results might be reduced.

In contrast to the previous viewpoint, correlation can be investigated. In fact, correlation among observations can be of scientific interest, and hypotheses can be generated on underlying risk factors and epidemiological processes. Moreover, in the case of repeated observations in time, correlation may vary based on time intervals separating data collection. There are two important families of models which can be used for investigating correlation: **marginal models** and **random-effects models**. In a marginal model, the effect of each risk factor on the outcome is obtained at population level, while there is no cluster specific parameter. For binary data, one possible approach is to use **generalised estimating equations** (GEE) (Diggle et al. 2002; Aerts et al. 2002; Molenberghs and Verbeke, 2005). In this approach, instead of specifying the full distribution for the correlated binary response, assumptions are made about the mean, variance and correlation. In general, GEE yields the same parameter estimates as an ordinary logistic regression, but the corresponding standard errors are inflated as a result of correlation among observations (which is also estimated). Statistical significance is, therefore, more difficult to attain by using GEE than ordinary logistic regression.

A variety of possible working correlation structures can be used in GEE. Some of the more common choices are:

a) independence: the simplest choice is the working independence model;

b) exchangeable: it may be most appropriate when there is no logical ordering for the observations within a cluster;

c) autoregressive: when repeated samples are taken at the same cluster, it can sometimes be assumed that the correlation between samples depends on the time lag between samples;

d) unstructured: a totally unspecified correlation matrix.

Worked examples of the analysis of non-independent observations by SURVEYLOGISTIC and by GEE (GENMOD), including SAS programming codes, are presented in Appendix A: Worked Examples of the SAS procedure application, Examples 4 and 5.

Alternatively, in a random-effects model, the correlation among non-independent observation is taken into account by explicitly estimating a measure of the cluster effect as an additional parameter in the model. This is based on the assumption that clusters (slaughterhouses, holdings) were randomly selected from a population. In addition to estimating population level parameters, random effect models allow the estimation of **cluster-specific parameters** (i.e. baseline level of prevalence and the effect of risk factors) and are most useful when inference to individual clusters is of interest (not available by using GEE) (Brown and Prescott, 1999; Littell et al. 1996). The application of random-effects models on zoonotic agents will be developed in a subsequent report.

3.2 Description of the analysis carried out on 2006 data

Data that were used for the statistical analysis of temporal trends in the CSR for 2006 consisted of the number of tested units, the number of positive units, and the total number of units - population size - in each MS, for three years from 2004 to 2006.

Graphical visualisation was carried out by diagrams using the lattice package in the R software (Sarkar, 2008). In the graphs, the proportion of tested units that were positive for zoonotic agents during each year was plotted for each MS. In Figure 1, the prevalence of *Salmonella* spp. in flocks of laying hens, by year and MS is represented. The width of 95% exact, binomial confidence intervals, represented by vertical bars, is associated with the precision of the prevalence estimate. Potential non-independence of observations from the same MS was not taken into account in estimating 95% CIs. Therefore, in these graphs, CIs are reported in order to represent the effect on the number of units that were tested. In fact, large sample sizes are associated with narrow CI's and relatively precise estimates, conversely, estimates obtained using small samples are characterised by low precision and large CIs.

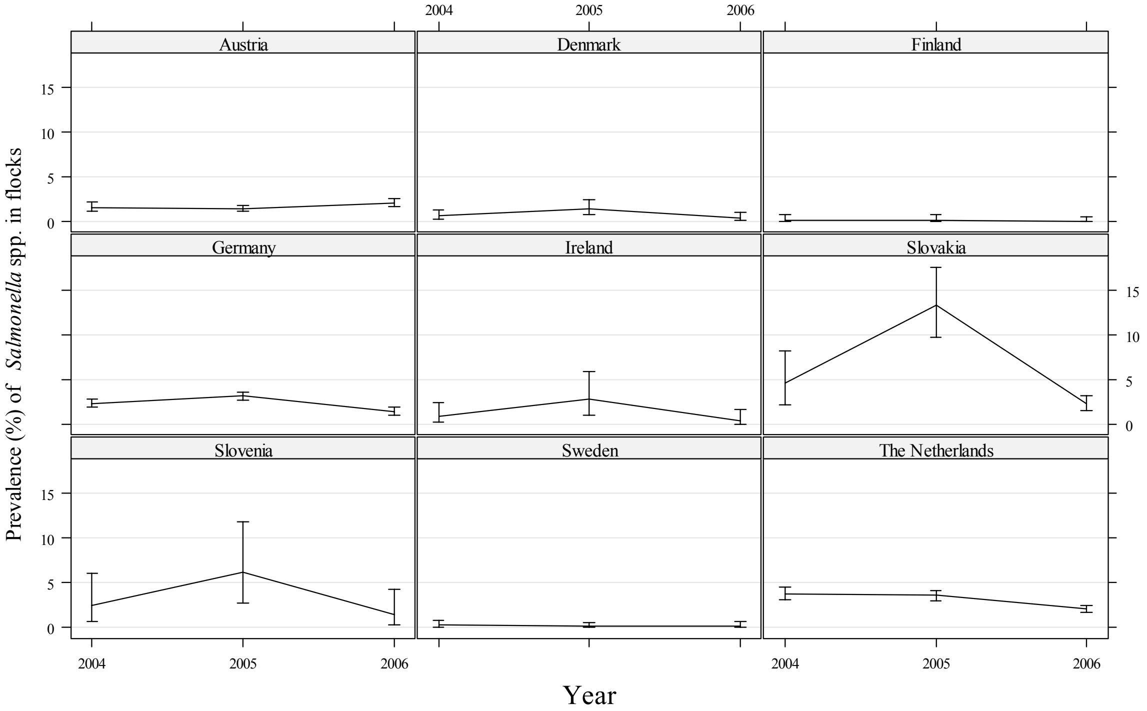

Figure 1. **Prevalence of *Salmonella* spp. in laying hen flocks in nine MSs, from 2004 to 2006. Vertical bars represent 95% exact binomial confidence intervals.**

In order to obtain yearly estimates of the ratios between positive and tested samples, for groups of examined MSs, the SURVEYMEANS procedure in the SAS System was used. A weight was applied for each observation, corresponding to the reciprocal of the sampling fraction in each MS (number of units in the population / tested units), to take into account disproportionate sampling at MS level (Figure 2). Estimated numbers of laying hens in each MS were used as the number of units in the populations to calculate weights in the analysis of *Salmonella* spp. in laying hen flocks.

In the following SAS statement the yearly, weighted ratio of positive units, out of those that were tested, is obtained for all MSs providing data.

```
proc surveymeans;
        by year;
        ratio positives / tested;
                weight wt;
run;
```

The weight, *wt*, was obtained by dividing the number of units in the population by the number of tested units.

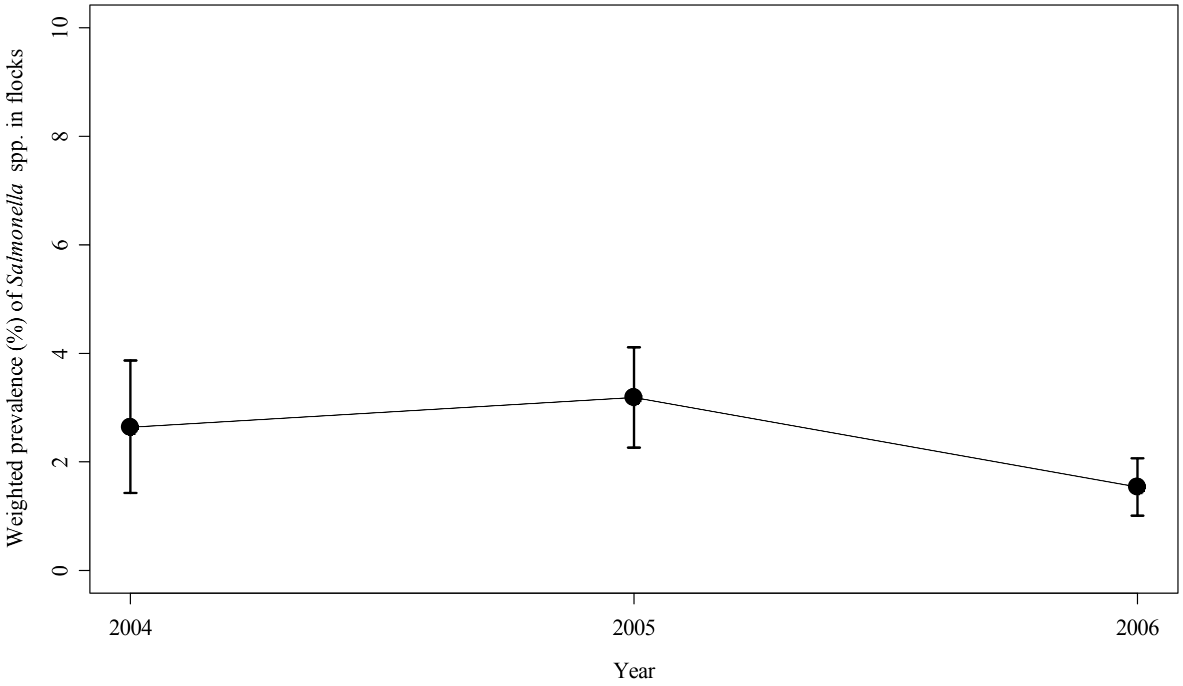

Figure 2. **Weighted prevalence of *Salmonella* spp. in flocks of laying hens for each year from 2004 to 2006. Data from nine countries were weighted by the reciprocals of the sampling fraction.**

Statistical significance of a three-year linear trend was tested by a weighted logistic regression for binomial data, using the SURVEYLOGISTIC procedure.

In the following SAS statement MSs are identified as clusters of epidemiological units.

```
proc surveylogistic;
        cluster MS;
        model positives/tested = year;
                weight wt;
run;
```

In SURVEYLOGISTIC, the cluster statement is used to deal with a design aspect, namely the random selection of cluster which leads to an additional source of variability in most practical cases. In the zoonoses trend analyses application, the cluster statement is used to take into account correlation among observations from the same MS. Consequently, the standard error for the parameter corresponding to the effect of year on the probability of a positive result is inflated and the probability of statistical significance reduced.

Different approaches to the analysis were adopted. The objectives were to compare results by using the SURVEYLOGISTIC and the GENMOD procedures, by taking account, or not, weighting and correlation among observations. More in-depth comparisons among different approaches to the analysis are presented in the following section on the critical review of the statistical analysis carried out on 2006 data.

3.3 Critical review of statistical techniques used for the 2006 report

In this section, statistical techniques used on data for 2006 were examined. A specific evaluation was made of the appropriateness of methods taking into account disproportionate sampling across MSs together with the non-independence of observations in logistic regression. In fact, the fractions of the populations of animals and food which were sampled in different MSs varied with the year, leading to disproportionate stratified sampling. From this point of view, MSs can be considered as strata. Moreover, it can be reasonably assumed that correlation among observations from the same MS in subsequent years was present. In this context, MSs would correspond to clusters. Another issue related to sampling design in a survey, which will be considered in this section, is the finite population correction in logistic regression analysis, even though this was not included in the 2006 analysis.

In order to carry out a review of statistical analysis, detailed information on the sampling process would be needed. Specifically, it would be critical to know if data can be considered if collected through a random selection of units, or according to a specific survey design (such as stratified sampling based, for example, on geographical areas; cluster sampling, or multi-stage sampling). In this section, such details are considered as not available. However, in general, it is acknowledged that sufficient information is available on data on zoonotic agents in animals which were analysed for time trends. On the other hand, data on food are more problematic, and a major effort is needed to improve related information.

Data on foodstuffs, animals and feedingstuffs

The EU weighted ratio of positive out of tested units was estimated by weighting the MS-specific proportion of positive units with the reciprocal of the sample fraction. There is an important issue on the non-availability of the total number of units in the population and the use of “reliable proxies”. For example, in the case of broiler meat, broiler population was used to calculate weights. Estimated numbers of laying hens in each MS were used as populations to calculate weights in the analysis of *Salmonella* spp. in laying hens flocks (see the following examples). This choice, which was dictated by data availability, should be reconsidered. The use of flock population sizes would, in fact, be more appropriate for laying hens. Moreover, it would allow considering the inclusion of reasonable finite population correction. In general, sensitivity analysis would be useful to assess the impact of the choice of weights on results at EU level.

Statistical analysis of trends over time

As for the estimation of proportions of positive units out of those that were sampled, weights were used in the analysis of trends over time to account for varying sampling fractions across MSs and in different years in the same MS. Population estimates (or proxies) for the year 2006 were used, but this was not always specified.

The SURVEYLOGISTIC procedure in the SAS System was used to examine any trend over time for the 2006 CSR. The effect of year of sampling was examined through a linear trend (with only two parameters, namely, the intercept and the effect of year), assuming that the frequency of zoonotic agents changed by the same amount for every year, namely the same change in prevalence occurred from 2004 to 2005 and from 2005 to 2006. This was due to the fact that interest was in the overall trend at EU level.

In SURVEYLOGISTIC, the CLUSTER statement names variables that identify the clusters in a clustered sample design. It is also assumed that clusters are randomly sampled from a sampling frame. In the context of the CSR data, no information is available on cluster sampling or on the random selection of clusters from a frame. However, the cluster statement was used to account for non-independence of observations from the same MS.

The issue of non-independence as mentioned in the report is related to the concept of "over-dispersion" of binomial data, where frequency counts of positive and of tested units are available in an aggregate fashion. An alternative data format would be the binary format, where one observation is available for each unit which is attributed a Bernoulli type outcome (positive or negative in the case of diagnostic tests).

SURVEYLOGISTIC was specifically developed for design-based analyses. Therefore, it is not the most appropriate procedure to account for correlated data in the absence of a specific sampling design. In this context, a model-based approach would be more appropriate and generalised estimating equations (GEE, procedure GENMOD in SAS) or generalised linear mixed models could be used (SAS procedures NLMIXED, GLIMMIX). Moreover, such models also allow the incorporation of serial correlation structures in addition to a time trend. However, from a practical point of view, the ultimate effect of using the CLUSTER statement in PROC SURVEYLOGISTIC is to account for non-independence of observations belonging to the same cluster, and the results are almost identical to those obtained using PROC GENMOD for repeated observations.

Another important and general concern about the "trend over time" analysis is the very limited time related information in the data, as data for only three consecutive years were available for the analysis at hand (2004, 2005 and 2006).

In order to carry out a thorough review of the analyses of time trends which were carried out on the 2006 data, worked examples of possible approaches to the analysis were carried out on a selected data set on *Salmonella* spp. in flocks of laying hens. Please refer to Appendix A for worked examples of the SAS procedure application.

Summary of results of the review of the analysis by worked examples

In this section, a critical review was carried out of the statistical analysis of time trends on data which were provided by MSs to EFSA within the context of the CSR. Data were provided in an aggregated way, as frequency counts of tested and positive epidemiological units. Therefore, no information on individual units, such as, for example, exposure to potential risk factors, was available. Nevertheless, it was decided to carry out statistical analyses on certain zoonotic agents, animals and food.

In some cases, the level of harmonisation was believed to be sufficiently high to warrant valid conclusions. Since it is reasonable to believe that data collection was sufficiently harmonised in different years within the same MSs, statistical methods for repeated observations at MS level provided valid results even if harmonisation among MSs was not perfect. In general, statistical analysis of time trends was carried out as an incentive and a guide to an improved data collection process.

The purpose of these analyses was to give as correct and accurate as possible estimates of time trends. Therefore, statistical techniques were applied to take into account two main characteristics of the data collection, namely, the disproportionate sampling from MS populations, and the potential correlation among probabilities of infection in animals and food from the same MS (non-independent observations).

The SAS procedure SURVEYLOGISTIC was used in the analysis of data for 2006. In order to carry out a careful review of the analyses, examples are given of different approaches to the analysis, based on different options available in the SURVEYLOGISTIC and the GENMOD procedures, which were used in the same data set. Both procedures can be used to deal with disproportionate sampling and non-independent observations. There are, however, differences between the so-called design-based analysis, performed by SURVEYLOGISTIC (allowing, for example, for finite population correction), and the model-based analysis by GENMOD, in addition to different technical details.

In order to account for non-independence of observations, the cluster statement was used in SURVEYLOGISTIC whereas, in GENMOD, generalised estimating equations (GEE) were fitted by using the REPEATED statement. The two procedures and statements both resulted in increased standard errors of the parameter estimates. This is because non-independence of observations generally corresponds to a reduction of the effective sample size and consequently the probability of finding significant time trends is reduced.

Although the CLUSTER statement in SURVEYLOGISTIC was developed to deal with a specific design where clusters were the primary sampling units, its use to account for potential correlation among outcomes can be considered as valid even in the absence of detailed information on survey design. In the analysis of data of zoonotic agents, MSs were considered as clusters to account for repeated observations across years. This was a pragmatic use of the SURVEYLOGISTIC procedure, leading to a cautious approach to testing for significance of trends.

When accounting for disproportionate sampling in SURVEYLOGISTIC, weights are automatically standardised and, therefore, sample size is not inflated. When using GENMOD, weights need to be standardised so that their sum is equal to the sample size.

Standardisation is, however, automatically performed in GENMOD if the repeated statement is used (GEE).

In the analysis of data on *Salmonella* spp. in laying hens, standard errors of parameter estimates are inflated after accounting for non-independence of observations, in comparison with standard errors from ordinary logistic regression (that assumes independence of observations). On the other hand, it is of interest to note that, on this specific data set, standard errors are not inflated when simultaneously taking into account non-independence and disproportionate sampling, with the inclusion of weights (set out in Appendix A: Worked examples of the SAS procedure application, Example 5). This was not the case in the analysis of *Salmonella* spp. in broiler meat (analyses not shown) where both the cluster statement in SURVEYLOGISTICS and the repeated statement in GENMOD yielded inflated standard errors of the parameter estimates, even in the presence of weights. In certain data sets, the inclusion of weights may affect the correlation among observations within the same clusters therefore reducing the effect of using specific statistical techniques for non-independent observation. This does not, however, reduce the validity of the general effect of increased standard errors after taking into account clustering.

Several options to account for survey design are available in PROC SURVEYLOGISTIC. MSs can be considered as strata, rather than clusters, using the STRATA statement. In general, stratified and cluster sampling are two distinct designs. In the analysis of zoonotic agents, it was, cautiously, decided to use the CLUSTER statement to account for repeated measurement from the same MS over three years. It is important to note that the same variable (i.e. MS) cannot be used in both the CLUSTER and STRATA statement.

In SURVEYLOGISTIC, finite population corrections can be applied at population level and for specific strata. As expected, standard error estimates were reduced with increasing sampled fraction (as set out in Appendix A: Worked examples of the SAS procedure application, Example 2) and smaller population. In the data used in our example, the sample was a very small fraction of the population and, therefore, the finite population correction had a negligible effect on the results. Using the number of laying hen flocks instead of the number of laying hens as population sizes would have lead to different results.

In PROC GENMOD, different types of correlation among observations can be chosen, based upon hypotheses on the mechanism giving rise to the correlation. If, for example, certain permanent characteristics of individual MSs can be considered as associated with the outcome of interest (i.e. *Salmonella* spp. infection in laying hens) the exchangeable correlation type can be selected. This is based on the assumption that correlation among observations from a specific MS in different years is the same regardless of the time period between observations. If, on the other hand, observations close in time can be considered as more strongly correlated than observations separated by longer time periods, then autoregressive correlation can be used. In Example 4, Appendix A, the choice of the correlation structure did not affect parameter estimates.

Moreover, in GENMOD, the scale option is available to correct over-dispersion of counts in those cases where variability is greater than expected according to the binomial distribution. This is another way to account for correlated observations (giving rise to over-dispersion) which can be used for very heterogeneous proportions. The outcome is generally very

cautious and statistical significance is more rarely attained in comparison with the other approaches that are presented (CLUSTER in SURVEYLOGISTIC and GEE in GENMOD).
In a further analysis, year was regarded as a categorical variable. The effect of year was coded in such a way as to compare the year 2006 against 2004 and 2005 respectively; 2006 was, therefore, used as the baseline year. Prevalence of *Salmonella* spp. in laying hens in both 2004 and 2005 was greater than in 2006, confirming the decreasing time trend. However, the parameter comparing 2005 and 2006 (0.74, as set out in Appendix A: Worked examples of the SAS procedure application, Example 8) was greater than the parameter comparing 2004 and 2006 (0.55), indicating that prevalence increased between 2004 and 2005, but then decreased to its lowest level in 2006. That is, there was a non-linear effect of year on prevalence. The original analysis, which was used for the CSR on 2006 data, disregarded such a non-linear effect and only considered a linear trend. This was due to the particular interest in the overall trend of prevalence. Nevertheless, it is advisable to explore and test for non-linear trends.

Further developments

This report was limited to the discussion of techniques that were used in the 2006 CSR, using PROC SURVEYLOGISTIC, and related options in PROC GENMOD, in the SAS System. In a subsequent report, analyses will be carried out on the same data by using alternative packages, such as R (www.r-project.org) and STATA (www.stata.com). Results from the three packages will be compared and possible differences and discrepancies will be discussed. In addition, random-effects models will be shown as means for analysis of the same data sets. The advantage over the technique currently used would be the estimation of MS-specific parameters, which is an important objective for both EU institutions and individual MSs. For the same objective, the Bayesian approach will be described and considered as a valid alternative in the analysis of data from non-independent observations.

3.4 Conclusions and recommendations of the statistical trend analyses

- The most important characteristics of the data available on zoonotic agents in animals and food, which were provided by MSs to EFSA, from 2004 to 2006, were taken into account in the analysis of time trends, namely:

  a) disproportionate sampling in reporting MSs; and

  b) correlation among observations in the same MSs in subsequent years.

- The use of two SAS procedures based upon different conceptual statistical approaches (SURVEYLOGISTIC: design-based approach; GENMOD: model-based approach) provided the same or very similar results when options for taking into account disproportionate sampling and correlation among observations.
- The CLUSTER statement in SURVEYLOGISTIC was used to take into account the correlation of observations in subsequent years. This corresponded to a cautious approach to the test of the hypothesis of significant temporal trends in the proportion of epidemiological units that were infected by zoonotic agents.
- In the data set that was used in the worked examples in this report (*Salmonella* spp. in laying hens), flocks were the unit of analysis. Nevertheless, the number of laying hens in

each MS was used in the calculation of weights to account for disproportionate sampling. An effort should be made to obtain reliable information on the number of flocks in MSs since it would be the most appropriate population estimate. Such information would allow the application of finite population correction if considered as appropriate (design-based approach).

- If a model-based approach to the analysis would be chosen, therefore, interest would be in the risk of infection in MSs, rather than in the particular epidemiological units that were tested at a specific time and GENMOD would be more appropriate than SURVEYLOGISTIC. In fact, in GENMOD, different types of correlations among observations in different years could be used. It could be hypothesised, for example, that observation separated by only one year would be more correlated than observations separated by two or more years.
- In the example presented in this report (*Salmonella* spp. in laying hens), the effect of year on prevalence of infection was clearly non-linear. However, in the published analysis, year was included as a continuous variable, therefore assuming linearity of effect. Non-linearity could be anticipated by exploratory and graphical analysis and verified through appropriate techniques. However, in the analysis carried out by EFSA, interest was on general trends and not on comparison among individual years. Nevertheless, different options should be tested.
- Other options for the analysis would be available. In the examples, the use of the scale parameter in GENMOD was included to take into account over-dispersion of binomial counts. This approach should be considered only in the case of extreme heterogeneity of counts with consequent difficulties in model-fitting, since standard errors of parameters are greatly inflated and the result is very conservative and statistical significance is more rarely attained.

4. Alternative approaches to statistical analysis: the Bayesian approach

Although 200 years old, Bayesian statistics have been increasingly developed and used in the last few decades, especially in the field of biological applications. As a matter of fact, since the 90s, all the theory and toolboxes have become mature, robust and sufficiently advertised to be widely implemented by statisticians.

Bayesian statistics are fundamentally based on a different paradigm as the one from "frequentist" statistics. In a nutshell, model parameters are not assumed to be a fixed unknown constant to be estimated, but instead they are seen as random variables. The point-estimate problem of frequentist statistics is turned into a more general problem of finding the distribution over parameters. Such a distribution (called **posterior distribution**) quantifies all uncertainty and information present in the model and the data, but also possibly from other prior sources of knowledge available (defining **prior distribution** of parameters).

Once evaluated, the posterior distribution over parameters can of course provide all the common quantities of interest such as estimates and confidence intervals, but also much more at no or little cost. For example, in the case of trend analysis, posterior distribution of trend parameters would also easily provide outputs such as "the posterior probability that the trend is more than 10% per year is above 60%" that could helpful for risk managers to appreciate the information/uncertainty gathered.

The evaluation of posterior distributions is generally made using specific Monte Carlo simulation algorithms called Monte Carlo Markov chains (MCMC). Even in dedicated software packages such as WinBUGS (freeware, see Lunn et al., 2000), the convergence assessment of those MCMC is usually left to the end-user, based on diagnostic plots and convergence tests provided by the software. In WinBUGS, this can be done using the built-in Gelman-Rubin test statistics which should be close to one at convergence, and with other diagnostic plots such as MCMC trajectories and posterior densities visualisation. Details, theory and examples of Bayesian methodology could be found e.g. in Gelman, 2002 or Robert, 2007.

Generally speaking, Bayesian statistics are particularly suitable in the following cases:

- when fitting random-effects (multi-level hierarchical) models;
- when the information is regularly updated;
- when information from different sources of various levels of reliability has to be combined together (e.g. EU information combined with industry information and expert knowledge and local surveys etc...);
- in the context of decision analysis with uncertainty;
- when making predictions or extrapolation to the future, with uncertainty; and
- when the sample size is low and classical (asymptotic) tests are no longer reliable.

The “prices” to pay for using Bayesian approaches include:

- minimal expertise in statistical modelling and Bayesian algorithms are needed as each model needs to be hard-coded, and convergence of algorithm checked and monitored by the end-user;
- computation time is generally larger than for non-Bayesian approaches.

4.1 Possible benefits of using Bayesian approaches for trend analysis

In the context of trend analysis for zoonotic agents, Bayesian analysis could be beneficial for the following reasons:

- to facilitate statistical inference:
  - Bayesian inference algorithms are much more “decorrelated” from modelling assumptions, e.g. there is no need to make assumptions on residual distributions, on linearity etc.;
  - by enabling easy updates of results once new information comes up (e.g. new survey results);
  - by enabling complex model-based inference e.g. using mixed-effects models, nested random-effects, missing values imputations etc.; and
  - Bayesian analysis offers suitable standard model selection procedures for mixed-effect models.
- to use or weight all information available:
  - data on zoonotic agents in animals and food may come from other various/heterogeneous sources including EU baseline surveys, national surveys or monitoring programmes etc.; and
  - to deal with sample sizes, missing data and uncertainty in a natural and straightforward way.
- to ease results communicability and exploitability for risk managers:
  - Bayesian outputs can be displayed in terms of probability distributions which are generally richer and more intuitive than *p*-values from the frequentist tests; and
  - the Bayesian set-up is by nature suitable for decision-making with uncertainty and for predictions for decision-makers (e.g. by deriving the probability that a given target is achieved within a given time frame).

Please refer to Appendix C for an illustrative example.

4.2 Conclusions and perspectives of Bayesian approaches for trend analysis

Bayesian approaches are flexible and powerful tools to fit *ad hoc* models and to integrate heterogeneous information. The simplistic example presented in Appendix C, also illustrates how the use of additional information such as population sizes of MSs with missing prevalence information could be used. Models that could benefit from such inference techniques include, for example, mixed-effects models (especially with several hierarchical levels), latent class variable models and Markov models. Many applications have been implemented, mostly in the areas of economy, marketing and ecology (for example a free software package is available for ecological applications, called BEAST (Bayesian Ecological

Analysis of Statistical Trends) (www.beastsoftware.org), for Bayesian estimation of population trends, which requires the user to do no programming). For longer term purposes, time and space trend analyses could also be investigated within a Bayesian set-up.

The main software packages for Bayesian inference include WinBUGS for general models and GeoBUGS for spatial data, possibly plugged into an R procedure. Alternatively, specific *ad hoc* MCMC algorithms can be coded in Matlab, S Plus, R and even in SAS using PROC IML. Such *ad hoc* programming should be left to the most complicated cases, as a high level of expertise in stochastic algorithms is required for robust analysis. In the simplest examples or for results validation, SAS PROC GENMOD could also be used with SAS version 9.2.

5. Spatial epidemiology

The objectives of spatial epidemiology are the description of spatial patterns, identification of disease clusters and the explanation and prediction of disease risk. In order to achieve these objectives, georeferenced data (points or areas) need to be used in addition to classical attribute data to describe the characteristics of the entity studied.

With recent advances in spatial analysis tools - geographic information systems (GIS), satellite imagery, spatial statistics and models - spatial analysis is becoming a core component of epidemiological research and training. For zoonotic diseases, with their complex transmission systems and the dependence on environmental factors of both reservoir hosts, pathogens and human exposure, spatial epidemiology is particularly important. Simultaneous consideration of both space and time, following separate spatial and temporal analyses is a particularly powerful approach to study temporal and spatial trends of zoonoses.

EFSA, with its mandate for surveillance, reporting and control of food-borne zoonoses, is in a unique position to take the lead in the identification of the most appropriate statistical methods for analysing the geographical distribution of zoonotic agents, as well as in addressing the limitations and constraints associated with spatial data and their interpretation.

This section aims to give an introduction to spatial epidemiology, illustrating the main components of the **spatial analysis framework**, such as:

- data management of both attribute and georeferenced data;
- spatial analysis methods: visualisation, exploration and modelling.

First, a brief description of data management is set out with particular reference to EFSA's geodatabase. A concise overview of the spatial analysis methods most commonly used in veterinary epidemiology is then set out. Current needs and required actions for developing spatial analysis approaches of zoonoses will be described in the conclusions of this section.

5.1 Data management: EFSA's Geodatabase

Management of georeferenced data is performed using geographic information systems (GIS) and database management systems (DBMS), and is of relevance throughout all the phases of spatial data analysis.

A geodatabase is available at EFSA and currently populated with the most common baseline data, such as classified administrative boundaries based on the "Nomenclature of territorial units for statistics" (NUTS). Currently, the 2007 CSR data for a number of selected variables (mostly country level prevalence data for selected zoonoses and host species) were extracted and stored in the geodatabase and were used for the production of the maps included in the 2007 CSR. The GIS software available at EFSA is ESRI ArcGIS Desktop (ArcEditor license) as well as the web-based ArcGIS Server. A GIS Community site was created as an aggregation point for GIS users and projects at EFSA and to facilitate access to geodata and geoprocessing resources.

5.2 Spatial analysis methods

Spatial analysis methods can be divided into three groups: visualisation, exploration and modelling:

1. visualisation is probably the most commonly used spatial analysis method, resulting in maps that describe spatial patterns and which are useful for both stimulating more complex analyses and for the communication of their results (Pfeiffer et al., 2008);
2. exploration of spatial data involves the use of statistical methods to verify whether observed patterns are random in space or not; and
3. modelling introduces the concept of cause-effect relationships using both spatial and non-spatial data sources to explain or predict spatial patterns (Pfeiffer et al., 2008).

In the following paragraphs a concise overview of the visualisation and exploration methods most commonly used in epidemiology is set out. Modelling methods will not be covered in this report.

5.2.1 Visualisation of spatial data

Visualisation techniques are represented by charts, graphs, diagrams, and maps. The goal of visualisation is to help assessment of the range and the general nature of data. Using different visualisation techniques it is possible to answer the following questions: what is the distribution of attributes (e.g. population, disease cases, farms, prevalence, etc.)? What is the arrangement of values across space? Where are the outliers? In spatial epidemiology, visualisation methods, resulting especially in maps, are particularly important to visualise the spatial pattern of the data and investigate the presence of disease clusters. Visualisation also represents a key tool for the successful communication of epidemiological data. Indeed, it is particularly important in communication with administrators, policy-makers, the public and other non-scientific audiences.

The advantages of using maps to convey research findings and to record and guide surveillance and control efforts are obvious. What also needs to be considered is the risk of misinformation through maps. Maps are pictorial models. Like all models they are selective in the type, amount and accuracy of information they present and cannot, at the same time, be general, precise and realistic. Like pictures, the selection of shapes, lines, and colours is crucial. Moreover, every map should include a scale bar, to let people understand the distance measurements, an indication of north, a legend, a listing symbol and attribute classification used on the map.

In addition to maps, other tools of visualisation, such as graphs, diagrams and tables can be used independently, but also linked to maps to optimise the visualisation of findings and recommendations.

The choice between different types of maps depends on the nature of the data to be visualised and on the objectives of the study. Examples of spatial data are point data and aggregated data.

Point data. The most frequently used method for visualising point data (such as single cases or outbreaks of disease) is to plot the location of the study subject using point maps. These types of maps use dots to show the occurrence of events in space, representing it with different symbols. Dots can either represent the simple location of the event and a quantitative attribute (using symbols of different sizes). Point maps provide a means of presenting the data in its raw format. This can be particularly useful for communication, allowing the appreciation of the spatial pattern without being burdened by the technical details of any analysis carried out to facilitate data display.

One of the possible problems encountered when using point maps if there are many events at the same location is when resulting point maps tend to be cluttered, making it difficult to appreciate the density of events. In order to visualise better high densities of events at the same location, point data can, for example, be summarised by areas (e.g. administrative units, regions, etc.) and visualised using maps specific for aggregated data, such as choropleth maps.

Aggregated data. The most common form of data aggregation occurs when counts of disease events within a defined area are summed to yield the total number of disease cases in each area. Disease counts can be expressed as a function of the population size to provide estimates of prevalence or incidence risk per unit area. Data aggregated over previously defined regions (such as countries, provinces) can be displayed using choropleth maps.

A choropleth map is a map in which areas are coloured or shaded in proportion to the measurement of the variable being displayed on the map, such as population density or prevalence of disease. In a choropleth map, the data values that fall within a specific class interval are assigned a unique colour, shade or pattern.

A key issue in choropleth mapping is the choice of the classification schemes used to divide continuous attribute data into categories. Indeed, changing the class interval scheme (e.g. natural breaks, quantile breaks, equal-interval breaks, etc.) can fundamentally change how the map looks and the message it sends. The most common classification schemes used to visualise continuous data in choropleth maps, are:

- natural breaks: where classes are defined according to apparently natural grouping of data values;
- quantile breaks: where data are divided into pre-determined numbers of classes which contain an equal number of observations;
- equal-interval breaks: in which the difference between the highest and the lowest attribute value is divided into evenly spaced steps;
- standard deviation classification: when using this method, the GIS calculates the mean value and then generates class breaks in standard deviation measures above and below it.

In order to show the impact of changing class interval schemes, examples of choropleth maps using natural breaks, quantile breaks and equal-interval breaks are reported in Appendix D.

The nature of areas themselves presents several challenges for mapping. The map's appearance and the message it conveys vary depending on size, number, and configuration of area units, leading to the so-called "*modifiable area unit problem*", MAUP (Openshaw, 1984). One important aspect of this problem is the location of area boundaries in relation to the distribution of health events and population. Indeed, depending on where boundaries are located, the areas can divide clusters of disease or concentrate them in a single zone. Choropleth maps vary with the number and sizes of area units. Small areas are more likely to capture the underlying pattern of health events, showing finer variation over space. Conversely, large areas conceal local differences, increasing the impact of MAUP. When mapping data aggregated by area, it is therefore essential to analyse the effects of the modifiable area unit problem both on maps and results. One of the possible solutions to MAUP is using small-area data, which allow mapping more detailed spatial patterns than with large-area information.

Another issue to be considered when dealing with area-data is the problem of very high rates associated with a small population, also known as the "small numbers problems" (Cromley and Mclafferty, 2002). Indeed, when using choropleth maps of disease incidence or prevalence with areas differing in population size, the calculated rates of disease for those areas have different degrees of reliability. Rates for areas with small populations vary more and are less reliable than those for large areas. For small areas, a difference of one or two cases can make a huge difference in incidence or prevalence rates. Different methods can be used for addressing the small numbers problem, such as for example the empirical Bayes smoothing. In empirical Bayes smoothing, rates are adjusted according to the size of the population on which they are based. The smoothing process brings raw rates closer to the mean rate across all areas (e.g. national or regional rate) making the rates more stable and less variable. The rates for small areas are smoothed more than those for large areas, reflecting differences in reliability linked to population size.

5.2.2 Exploration of spatial data

One of the main objectives of explorative analysis is to identify clustering of disease occurrence. If mapping the occurrence of a disease may provide a general idea of where the disease is, explorative analysis will prove, through statistical tests that the pattern visualised is not random, but caused by underlying factors. Determination of true clustering versus apparent clustering is an integral part of epidemiologic surveillance and interpretation.

Spatial clustering of a disease provides useful information on the possible causes of the disease. Methods to investigate spatial clustering could be used before the application of any spatial model, and, of course before recommendations for interventions and control measures are determined and applied.

A variety of statistical methods to detect and describe spatial and temporal clustering have been developed from descriptive spatial statistics to tests for spatial clustering, tests for temporal clustering and tests for time-space clustering. Applications have been most common to vector-borne and zoonotic diseases as well as to cancer and health care delivery and access.

Several problems exist in studying clusters:

a. disease events are often rare and case occurrence may occur over a long period;

b. information on the population at risk may be unavailable;

c. rates of disease expected in the absence of clustering may be unknown;

d. the choice of spatial and temporal units may influence the identification of clusters; and

e. different weight matrices can be taken into account (not only real distances, but also for example intensity of trading between countries).

An important issue to be considered when investigating clustering is to define the geographical extent, or *scale*, to which clustering occurs. Scale critically affects the kinds of inferences that can be drawn (Cromley and Mclafferty, 2002). For example, clustering within small areas or communities reflects localised factors such as point of source of environmental contamination. Conversely, elevated disease rates for states or regions result from region-wide factors like climate, international animal and foodstuff trade policies, etc. The scale to which a health problem is studied should reflect an understanding of the disease process and likely causative factors (Cromley and Mclafferty, 2002). It is important to keep in mind that patterns at one geographical scale can conceal patterns at other scales. For example, a country's "average" rate of a certain zoonosis in animal population may result from having some administrative areas with unusually high rates and others with unusually low rates. Such high rate administrative areas are lost in the state-wide average, and can be revealed only using analyses below the country scale.

Analysing clustering requires a set of criteria to assess how much cluster exists and then to define "significant" clusters. In general, spatial and space-time methods rely on statistical criteria that describe the likelihood of clusters arising by chance in a given population. Such criteria may use a known probability distribution such as the Poisson distribution, or they may use the Monte Carlo simulation methods which involve generating a large number of random possible outcomes (Cromley and Mclafferty, 2002).

The techniques adopted for analysing spatial patterns of disease depend on the type of data available. These methods are usually divided in two groups: global and local clustering methods. Space-time clustering methods are furthermore applied to investigate the space-time pattern of disease spread.

Some of the most common methods for investigating spatial and space-time clustering of diseases are mentioned in the following paragraphs, and they are described in more details in Appendix E Spatial Statistics.

Spatial methods for investigating global clustering

Global clustering methods are used to test for spatial clustering throughout the study region without the ability of locating specific clusters sites. Their results provide a single statistic that measures the degree of spatial clustering, the statistical significance of which can also be assessed. The null hypothesis of global clustering methods is simply that "clustering does not exist" (Pfeiffer et al., 2008).

Examples of global clustering tests for aggregated data used in veterinary epidemiology include indices of autocorrelation, such as Moran's I and Geary's C. Autocorrelation measures the extent to which the occurrence of an event in a geographical unit (points or polygon) makes the occurrence of an event in a neighbouring geographical unit more probable. The most commonly used global clustering methods for point data include the Cuzick and Edwards test and the global second-order K-functions (Ripley's K-function and weighted K-function), all of which are described in Appendix E Spatial Statistics.

Spatial methods for investigating local clustering

Local clustering methods define the location and the extent of the identified spatial clusters and can be divided into focussed and non-focussed tests. Non-focussed statistics, or simply local methods, identify the location of all likely clusters in the study region, while focussed statistics investigate whether there is an increased risk of disease around a specific location or focus. These tests are particularly useful for exploring possible clusters of disease near potential sources of infection (Pfeiffer et al., 2008).

An important class of methods for investigating local spatial clustering includes localised measures of spatial dependence such as the most commonly used Local Moran test and Getis and Ord G_i and G_i^* statistics. Alternative tests for identifying local spatial clustering are based on scanning local rates, such as the spatial scan statistic. The above mentioned examples of local clustering methods are described in Appendix E Spatial Statistics.

Space-time clustering methods

An important issue in monitoring communicable diseases and planning interventions is to understand the space-time pattern of spread. Investigating space-time clustering consists in evaluating whether, for example, cases that are close in space are also close in time and *vice versa*, adjusting for any purely spatial or temporal clustering. Various space-time clustering tests have been developed for this purpose; among them one of the most used is the Jacquez's *k* nearest neighbours test (see also Appendix E Spatial Statistics).

5.3 Conclusion and further developments

This section set out an introduction to spatial epidemiology, illustrating the main components of the spatial analysis framework as well as a brief description of data management with particular reference to EFSA's geodatabase. Successively, point maps and choropleth maps have been reported as an example of visualisation methods for point and aggregated data, respectively. Possible problems of visualisation techniques, such as MAUP and "small numbers problems" have been also briefly addressed. A schema of the spatial statistics described in this report is given below:

- Global measures of spatial clustering:
  - Aggregated data: Moran's I and Geary's C; and
  - Point data: Cuzick and Edwards test, Ripley's K-function and weighted K-function.

- Local measures of spatial clustering:
  - Aggregated data: Local Moran test and Getis and Ord G_i and G_i^* statistics;
  - Point data: spatial scan statistic.
- Space-time statistics: Jacquez's *k* nearest neighbours test.

Spatial statistics mentioned above represent an example of the methods most commonly used for studying the epidemiology of zoonoses. New methods for detecting local clustering are continuously being developed, such as several Bayesian approaches.

Identifying spatial clustering of diseases can provide useful information on the possible causes of diseases and can help better understand their epidemiology. Investigating possible disease clustering requires a systematic approach. When performing explorative spatial analyses it is good practise to apply first global spatial statistics in order to identify the presence of significant clusters throughout the study region. Once significant spatial clustering at global level is identified, local spatial statistics can be applied to detect the location of disease clusters. In general, it is recommended that a combination of statistics be used when studying local clustering to ensure that different aspects of spatial patterns are identified and to check whether the results from different analyses are consistent.

An important issue to be addressed when investigating clustering of diseases is to define the geographical extent, or *scale*, to which clustering occurs, since *scale* critically affects the kinds of inferences that can be drawn. In order to evaluate how the choice of different scales may affect the outcome of epidemiology investigations carried out by EFSA and the kinds of inferences that can be drawn, data at different geographic scales are needed. Based on the last considerations, the working group on statistical analysis of temporal and spatial trends considers it important to identify a specific case study to serve as a prototype on the use of spatial analysis methods to first describe, then analyse and potentially explain and predict patterns of zoonoses within the EU, based on available data.

Carrying out a **case study on spatial analysis** can help:

- perform a focussed literature review;
- suggest the most appropriate statistical methods for analysing the geographical distribution of zoonotic agents;
- address the limitations and constraints associated with spatial data and their interpretation;
- identify potential additional variables to be included in the data management system not currently or readily available (e.g. environmental, demographic variables); and
- put in place a workflow from data gathering to output production in which technical aspects will also be taken into account, such as geodatabase management, the development of web-based analytical tools and result sharing/visualisation capabilities.

Zoonoses data are currently aggregated and reported through the EFSA data collection system only at country level with the exception of voluntary reporting at regional level for brucellosis and tuberculosis data. The availability of the latter data is not consistent among MSs or years and linkage to geographic information makes use of codes that have not been standardised well.

Ways and possibilities of collecting better geographic information through the zoonoses data collection system shall be investigated and proposed. In particular, the working group on the statistical analysis of temporal and spatial trends will investigate the possibility of collecting finer geographic information through the zoonoses data collection system and explore MS availability in providing such data. Obtaining data at a finer scale (e.g. regional data) from some MSs will enable the application of more detailed spatial analysis methods and compare the results with those obtained performing spatial analyses at country level. Moreover, as previously underlined, data on different geographic scales will allow the evaluation of how the scale of analysis may affect the outcome of epidemiology investigations carried out by EFSA and the kinds of inferences that can be drawn.

6. Communication of methods and results of statistical analysis

..."data are not just numbers, they are numbers with a context" …
(*George Cobb and David Moore - 1997*)

Interpretation and clear communication are fundamental aspects of data analysis. Scientific information is more useful to the audience and greater success in communication is achieved if the information provided is relevant and easily understood.

The importance of communicating results is a fundamental issue in standard documentation of International Institutions. For example the World Organisation for Animal Health (OIE) dedicates Article 2.2.7 to define the principles of Risk Communication (OIE (2008) Terrestrial Animal Health Code) and the FAO in 1999 published "The application of risk communication to food standards and safety matters", a report of a Joint FAO/WHO Expert Consultation.

Usually, articles/reports contain impenetrable statistical jargon and unfamiliar mathematical expressions. Epidemiologists and statisticians have a responsibility to present and communicate their results in ways that are transparent to everyone and have an "obligation" to make the data they collect useful to the public. Terms meaningful to a statistician may be foreign not only to a layperson, but even for public health officials and decision-makers. In too much research, understanding even the substantive conclusions of sophisticated quantitative models can be challenging at best and impossible at worst.

Therefore, different audiences need different forms of communication; interpreting and disseminating information to a variety of audiences that may not be familiar with statistics is not an easy task. To this end, a few useful steps are set out below as a guide to attain understandable and meaningful communication.

6.1 Steps to improve communication

Understandable communication can be summarised in five points:

1. define communication objectives;
2. identify the audience;
3. determine key point messages;
4. organise communication of statistical results;
5. draw up a conclusion.

1. Defining communication objectives: why are results to be disseminated?

Communication of epidemiological research results may include the following:

- raise awareness about new research findings;
- clarify a controversial issue;
- provide information to help policy-makers make decisions;
- redirect programme priorities;
- seek support; and
- introduce a new strategy.

2. Identifying the audience

Part of the communication strategy is to identify the audience (i.e. risk-managers, general public, consumers, risk analysts, legal analysts, etc.) and to establish the bases for communication, taking into consideration the characteristics of each audience. In general, approaches to policy and academic audiences are different:

- Risk-managers want to hear statements of problems and their solutions. So focus should be on key messages, data linked to the audience's concerns, recommendations without excessive qualifiers; and
- Researchers want to know the theoretical framework, the methodology (for credibility) and results and discussion.

Therefore, it is important to focus on what the audience or the readership needs to know, not on what the scientists know. It is important to consider the audience's level of technical knowledge, motivations, and interests. If it is a mixed audience, the best is to identify common interests and needs and what it is hoped the audience will do as a result of the analyses which in turn will guide the lay-out of communication.

3. Determining the key point messages

People often fail to communicate effectively due to a lack of clear communication goals and key messages to support them. Setting such goals and identifying support messages are decisions that should be made prior to the issuing of any public comment. Identify a few points that you want the audience to remember, and build the communication around these points, tailoring them to the audience's technical level, information needs, and interests. Avoid technical jargon.

4. Organising the communication of statistical results

Content to include:

- two sentences as an introduction to the problem, linked to audience concerns;
- key messages (what the audience should remember);
- objectives of the analyses;
- a brief description of the methodology (depending on the audience);
- major findings and implications; and
- recommendations, if appropriate.

How to display results

Graphs

A picture is indeed worth a thousand words, or a thousand data points. Graphs can be extremely effective in expressing key results presenting a clear, visual message, with an analytical heading. A complicated graph is difficult to decipher and at worst could be misleading.

Tables

The purpose of a table is to summarise and present in a concise, well-organised way information that cannot be presented simply in the text. Tables should be readable and understandable without consulting the text. The title of a table should communicate what is being presented, how variables are classified, when and where data were obtained.

Maps

Well-designed maps can be used to illustrate differences or similarities across geographical areas and often clarify local or regional patterns, which may be hidden within tables or charts. Maps are a rapidly expanding area of data presentation, with methods of geographic analysis and presentation becoming more accessible and easier to use.

Producing statistical maps can be a simple process. The audience should be focussed on so that the map produced is easy to understand.

5. Draw up a conclusion

Design an effective conclusion: what should the audience remember?

The audience should be informed of the possible consequences of taking or not taking an action on a particular issue.

6.2 Communication within EFSA's mission

Risk communication is one of the three components of risk analysis, together with risk assessment and risk management. According to Regulation 178/2002, which defines EFSA's mission and activities, communication is one of the Authority's main tasks. In addition to the communication of scientific findings to the public, EFSA is responsible for communicating effectively with the European Commission, MSs, and other institutions that use EFSA's advice to make decisions in the field of food safety and, more in general, for the protection of public health.

Sophisticated statistical techniques are often employed in data analyses carried out by EFSA's Unit on zoonoses. These techniques are necessary to take into account the complexity of sampling procedures which are, in turn, associated with complex epidemiological and field situations. Results of the analyses should, however, be communicated to risk-managers and public health officials in a clear and univocal fashion. Specifically, key issues which justify the adoption of complex methods must be clearly explained.

The application of weights when estimating the frequency of zoonotic agents in animals and food, and in the analysis of temporal and spatial trends, is an example of a statistical technique which is a prerequisite for the validity of results but that requires an effort in communication. In fact, weighting (to account for disproportionate stratified sampling, see sections on statistics) often produces parameter estimates which are different from the (biased) estimates which would be obtained without the application of weights. These differences might generate confusion and need careful explanation.

In the following example, an attempt is made to present the need of weighting for the zoonoses dataset when estimating prevalence at EU level. The objective is the acceptance of results by public health officials and decision-makers following the clear understanding of theoretical and practical support for the application of specific statistical techniques.

In a graphical representation (Figure 3), two squares can refer to geographic areas or different types (strata) of epidemiological units. The units (for example, holdings where animals are raised) are represented as points which, in turn, are assigned a colour based upon the value of a binary variable, such as presence or absence of a zoonotic agent: white for negatives or uninfected, red for positives or infected (cases). Therefore, in this simulation, the real status of the total, target population is known. Consequently, when estimating prevalence of infection though the analysis of a sample, the biased results of disproportionate stratified sampling can be demonstrated, and the application of weights, in order to reconstruct the composition of the target population in the sample, is clearly justified. The weighted prevalence in the sample is obtained through simple calculation, and it is shown to be an unbiased, valid estimate of the prevalence in the target population. For simplicity, no random error is included in this simulation.

If there are two different strata with the following characteristics:

Stratum 1: Population (N_1) =100 holdings, cases (s_1) = 10, prevalence (p_1) = 10%
Stratum 2: Population (N_2) = 500, cases (s_2) = 125, prevalence (p_2) = 25%

If the two strata are summed up the **real** situation of the target population is obtained, ie: Total: Population (N) = 600, cases (s) = 135, prevalence (p) = 22.5%

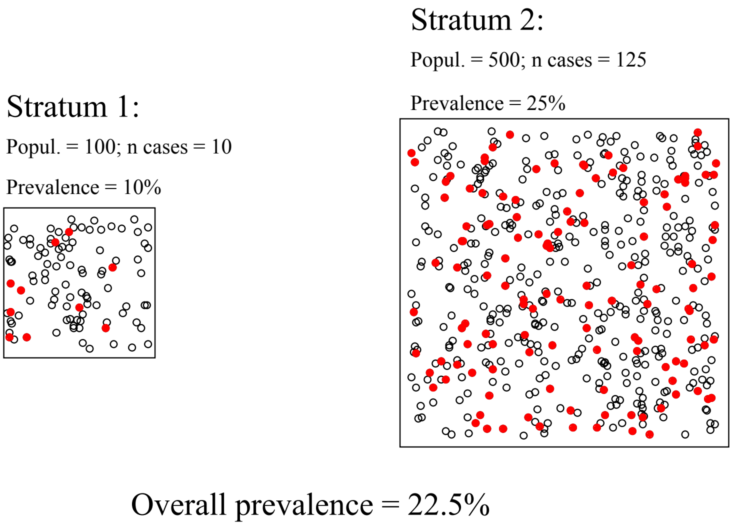
Stratum 1:
Popul. = 100; n cases = 10
Prevalence = 10%
Stratum 2:
Popul. = 500; n cases = 125
Prevalence = 25%
Overall prevalence = 22.5%

Figure 3: **Representation of two strata of epidemiological units. Red points represent units that are positive for infection by zoonotic agents. White points represent negative units**

If during sampling (Figure 4) the same number of units were selected from both strata: 100 units were tested in stratum 1 (corresponding to the entire population for stratum 1), ten of which resulted as positive, and 100 units were tested in stratum 2 (corresponding to 20% of the population of stratum 2) of which 25 were positive.

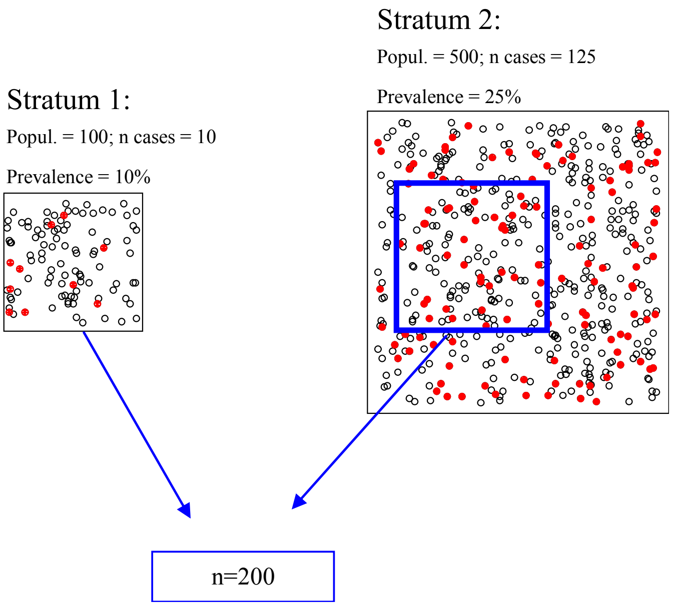

Figure 4: **Disproportionate stratified sampling from two strata: 100 units are sampled from each stratum in spite of different population sizes (stratum 1, N = 100; stratum 2, N = 500)**

The sample is, therefore constituted by 200 units. Raw prevalence of infection can be obtained as follows:

Stratum	Sample	Positive*	Population
1	**100**	**10**	**100**
2	**100**	**25**	**500**

* expected number of positive units in the absence of random error.

Through the calculation (10+25)/(100+100) = 35/200 = 0.175 is obtained as an estimate. This is lower than the known population prevalence and it is, therefore biased due to systematic error. In fact, stratum 1, characterised as a relatively low prevalence, is **over-represented** in the sample, where it constitutes 50% of the units, as compared with the target population, where it constitutes 16.7% of all units.

6.3 Application of weights to obtain a valid prevalence estimate for the population

Each observation is applied a weight corresponding to the reciprocal of the sampling fraction in the two strata, calculated as the ratio between sample size and population size:

Sampling fraction stratum 1: sample (n1) /population (N1) = 100/100 = 1
Sampling fraction stratum 2: sample (n2)/population(N2) = 100/500 = 0.20

Weight stratum 1: W1 = 1/1
Weight stratum 2: W2 = 1/0.2 = 5

The weighted, population prevalence estimated can, therefore, be obtained based upon the following values:

Stratum	Sample	Positive	Population	wt	NposWt
1	**100**	**10**	**100**	**1**	**10**
2	**100**	**25**	**500**	**5**	**125**

The weighted prevalence estimate is, therefore, obtained as:
(10*1+25*5)/(100*1+100*5) = 135/600 = 22.5%
that, in the absence of random error, exactly corresponds to the true population prevalence.

The same simulation could be extended to explain the subsequent logical process of standardisation of weights, whose objective is to avoid the inflation of sample size (see example). The proposed style of presentation, including graphics, and step by step calculations, should be used when epidemiologists and statisticians are presenting a plan of statistical analysis or its results to public health professionals and decision-makers. In this way, decisions in the analysis can be discussed in a transparent way, in light of both scientific and policy related implications. Moreover, through clear communication, epidemiologists can explicitly acknowledge potential biases and limitation of their work, which are sometimes intrinsic to the study design and impossible to adjust (Bhopal, 2008). In this way, survey results on large and complex populations can be taken into proper account in public health decision-making processes.

GENERAL CONCLUSIONS

The study of temporal and spatial trends of zoonotic agents in animals and food is a major objective of the statistical analyses that are carried out by EFSA within the framework of the CSR.

- It is particularly important that these analyses are carried out at EU level, since MSs belong to the same geographic area, share a common market, and are subject to common policies in disease control and prevention. At the same time, major differences in the natural environment and in management exist among EU MSs, and appropriate statistical analyses can be used to take this into account.

- In the analysis of trends in the frequency of zoonotic agents in animals and food in the EU, the first general question to be asked is whether observed changes in frequency are real or an artefact (Bhopal, 2008). In synthesis, trends may be affected by random error associated with small sample size. However, systematic error yielding biased results of the analysis is most insidious in many aspects. Changes in data collection methods or in the laboratory diagnosis across MSs, or within the same MS in different years, for example, would lead to artefact trends. The application of statistical methods can provide confidence intervals as measures of random error of parameter estimates. Moreover, the characteristics of the sampling design, when is it known, can be taken into account in the analysis to reduce the risk of systematic error.

- The level of harmonisation of zoonoses data collection is still to be improved and, in certain cases, not enough information is available on such a key issue. It is, however, reasonable to believe that, for certain animal categories, the data collection process is harmonised or, at least, well-known in the different MSs. Food data collection is probably where more work is required to improve harmonisation.

- Statistical analyses of temporal trends were carried out, for the first time, for the CSR on data collected from 2004 to 2006. As data collection may be not fully harmonised throughout different MSs, this source of heterogeneity should be recognised by the statistical models. As data over years from the same MS are also expected to be correlated (being more similar as they refer to the same country with its country specific latent characteristics, one of which is the data collection procedure), a repeated measures model with MS as subject-indicator was used for the trend analysis. Therefore, the analysis was valid as long as harmonisation was constant inside MSs, across years, whereas the potential effects of between MS harmonisation problems were reduced.

- Another important characteristic of the data was that different MSs provided data corresponding to different fractions of their populations, therefore, the composition of the sample, in terms of MS contributions, was different from the MS composition of the EU population. Statistical techniques were consequently used to reconstruct the EU population in the analysed sample, to reduce the risk of biased results. This was achieved through the use of weights corresponding to the reciprocal of the sampling fraction. The validity of such weights is greatly affected by the availability of reliable population estimates for epidemiological units.

- In conclusion, statistical analyses carried out on temporal trends for the 2006 CSR were generally valid. Major recommendations for the improvement of the analysis of temporal and spatial trends are relative to the availability of more detailed information on data collection processes to favour interpretation of results or to use appropriate statistical methods.

- Information on potential risk factors would be used to adjust confounding when estimating trends. Specifically, data on season of data collection would allow adjustment for potential confounding due to seasonal effects on certain zoonotic agents. Moreover, the availability of data for geographic areas within each MS, if not geographic coordinates of holdings, would allow more refined spatial and spatial temporal analysis. Objectives would be the detection of changing temporal trends at different locations.

- In the part II report of the activities of this working group, the use of different software and statistical packages will be compared in the analysis of temporal trends. Further developments in statistical analysis will include a description and examples of random effect models and Bayesian models to obtain MS-specific trend analyses. A specific section of the part II report will be devoted to the development of spatial analysis with examples of application to data from certain MSs.

TASK FORCE ON ZOONOSES DATA COLLECTION MEMBERS

Andrea Ammon, Alenka Babusek, Marta Bedriova, Karin Camilleri, Marianne Chriel, Georgi Chobanov, Ingrid Dan, Jürg Danuser, Kris De Smet, Sylvie Francart, Matthias Hartung, Birgitte Helwigh, Merete Hofshagen, Patrícia Inácio, Sarolta Idei, Elina Lahti, Lesley Larkin, Peter Much, Edith Nagy, Lisa O'Connor, Rob Van Oosterom, Jacek Osek, José Luis Paramio Lucas, Antonio Petrini, Melanie Picherot, Christodoulos Pipis, Saara Raulo, Antonia Ricci, Petr Šatrán, Joseph Schon, Jelena Sõgel, Snieguole Sceponaviciene, Ana María Troncoso González, Kilian Unger, Luc Vanholme, Dimitris Vourvidis, Nicole Werner-Keišs.

LIST OF APPENDICES

REFERENCES

Aerts, M., Geys, H., Molenberghs, G. and Ryan, L., 2002. Topics in Modelling of Clustered Data. Chapman & Hall, London, pp. 336.

Anselin, L., 1995. Local Indicators of Spatial Association - LISA. *Geographical Analysis* 27: 93–115.

Bailey, T.C. and Gatrell, A.C., 1995. Interactive spatial data analysis. Longman Higher Education, Harlow, pp.432.

Bhopal, R., 2008. Concepts of epidemiology, 2nd ed. Oxford University Press, pp. 417.

Boelaert, F., 2003. Elements of survey design and analysis regarding endemic infections in livestock. PhD thesis, Faculty of Veterinary Medicine, Ghent University, pp. 214. ISBN 90-5864-045-0

Brown, H. and Prescott, R., 1999. Applied Mixed Models in Medicine. Wiley, pp. 408.

Cobb, G.W., and Moore, D.S., 1997. Mathematics, statistics, and teaching, *American Mathematics Monthly*, *104*, 801-823.

Cromley, E.K. and Mclafferty, S.L., 2002. Gis and Public Health. The Guilford Press, pp. 339.

Cuzick, J., Edwards, R., 1990. Spatial clustering for inhomogeneous populations. J. R. Statist. Soc. Ser. B 52, 73-104.

Diggle, P., Heagerty P., Liang, K.-Y. and S. Zeger. 2002. Analysis of Longitudinal Data. Oxford University Press, pp. 379.

EC (European Community), 1998. Decision No 2119/98/EC of the European Parliament and of the Council of 24 September 1998 setting up a network for the epidemiological surveillance and control of communicable diseases in the Community, 1998 (OJ L 268, 3.10.1998, p.1). <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31998D2119:EN:HTML>

EC (European Community), 1998. Opinion of 28 December 1998 of the Economic and Social Committee on the 'Resistance to antibiotics as a threat to public health' (OJ C 407, 28.12.98, p. 7). <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:1998:407:0007:0017:EN:PDF>

EC (European Community), 2000. Commission Decision No 2000/96/EC of 22 December 1999 on the communicable diseases to be progressively covered by the Community network under Decision No 2119/98/EC of the European Parliament and of the Council, (OJ L 28, 3.2.2000, p. 50-53) <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2000:028:0050:0053:EN:PDF>

EC (European Community), 2000. Opinion of 12 April 2000 on food-borne zoonoses from the Scientific Committee on Veterinary Measures relating to Public Health. <http://ec.europa.eu/food/fs/sc/scv/out32_en.pdf>

EC (European Community), 2002. Opinion of 18 April 2002 of the Economic and Social Committee on: the Proposal for a directive of the European Parliament and of the Council on the monitoring of zoonoses and zoonotic agents, amending Council Decision 90/424/EEC and repealing Council Directive 92/117/EEC, and the Proposal for a regulation of the European Parliament and of the Council on the control of *Salmonella* and other food-borne zoonotic agents and amending Council Directives 64/432/EEC, 72/462/EEC and 90/539/EEC, (OJ C 94, 18.4.2002, p.18). <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2002:094:0018:0022:EN:PDF>

EC (European Community), 2003. Directive 2003/99/EC of the European Parliament and of the Council of 17 November 2003 on the monitoring of zoonoses and zoonotic agents, amending Council Decision 90/424/EEC and repealing Council Directive 92/117/EEC, (OJ L 325, 12.12.2003 p.31). <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:325:0031:0040:EN:PDF>

EC (European Community), 2007. Commission Regulation (EC) No 1441/2007 of 5 December 2007 amending Regulation (EC) No 2073/2005 on microbiological criteria for foodstuffs, (OJ L 322, 7.12.2007, p.12). <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:322:0012:0029:EN:PDF>

EFSA (European Food Safety Authority), 2009. The Community Summary Report on Trends and Sources of Zoonoses, Zoonotic Agents, Antimicrobial Resistance and Foodborne Outbreaks in the European Union in 2007, *The EFSA Journal* (2009), 223.

EFSA (European Food Safety Authority), 2008. Report of the Task Force on Zoonoses Data Collection on the Analysis of the baseline survey on the prevalence of *Salmonella* in slaughter pigs, Part B, *The EFSA Journal* (2008), 206, 1-111.

EFSA (European Food Safety Authority), 2006. Report of Task Force on Zoonoses data collection on guidance document on good practices for design of field surveys, *The EFSA Journal* (2006), 93, 1-24.

Elliott, P. and Wartenberg, D., 2004. Spatial Epidemiology: Current Approaches and Future Challenges. *Environmental Health Perspectives*. Volume 112, number 9: 998-1006.

FAO (Food and Agriculture Organization), 1999. The application of risk communication to food standards and safety matters. Report of a Joint FAO/WHO Expert Consultation. FAO Food and Nutrition Paper No. 70. Rome. 46 pp. http://www.fao.org/docrep/005/x1271e/X1271E00.HTM#TOC, accessed the 7th of November 2008.

Geary, R., 1954. The contiguity ratio and statistical mapping. *The incorporated statistician,* 5, 115-145.

Gelman, A., 2002. Bayesian Data Analysis. Chapman-Hill, pp. 696.

Getis, A. and Ord J.K., 1996. Local spatial statistics: an overview. Spatial analysis: modelling in a GIS environment. P. Longley and M. Batty (eds). Geoinformation, Cambridge, pp. 261-277.

Getis, A. and Ord. J.K., 1992. The Analysis of Spatial Association by Use of Distance Statistics, *Geographical Analysis*, 24: 189-206.

Getis, A., 1984. Interaction modelling using second-order analysis. *Environ. Plan A,* 16: 173-183.

Haining, R., 2003. Spatial Data Analysis – Theory and Practice. Cambridge University Press, Cambridge, pp. 452.

Hosmer, D.W., Lemeshow, S., 2000. Applied logistic regression, 2nd ed. Wiley Series in probability and statistics, pp. 307.

Jacquez, G.M., 1996. A k-nearest neighbor test for space-time interaction. *Statistics in Medicine*, 15: 1935-1949.

Kitron, U., Jones, C.J., Bouseman, J.K., Nelson, J.A. and Baumgarter, D.L., 1992. Spatial analysis of the distribution of *Ixodes dammini* (Acari: Ixodidae) on white-tailed deer in Ogle County, Illinois. *J. Med. Entomol*. 29, 259–266.

Kulldorff, M., 1997. A spatial scan statistic. *Communications in Statistics: Theory and Methods.* 25(6): 1481-1496.

Levy, P., Lemeshow, S., 1999. Sampling of Populations: Methods and Applications. 3rd ed. Wiley, pp.525.

Littell, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger, R.D., 1996. SAS System for mixed models. SAS Institute Inc, pp. 633.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS - a Bayesian modeling framework: concepts, structure and extensibility Statistics and Computing 10: 325-337.

Mannelli, A., Mandola, M.L., Pedri, P., Tripoli, M. and Nebbia, P., 2003. Associations between dogs that were serologically positive for *Rickettsia conorii* relative to the residences of two human cases of Mediterranean spotted fever in Piemonte (Italy). *Preventive Veterinary Medicine*, 60, 13-26.

Molenberghs, G. and Verbeke, G., 2005. Models for Discrete Longitudinal Data. Springer, pp. 683.

Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. Biometrika 37, 17-23.

OIE (World Organisation for Animal Health), 2008. Terrestrial Animal Health Code World Organisation for Animal Health (OIE), accessed the 7th November 2008. <http://www.oie.int/eng/normes/mcode/en_chapitre_1.2.2.htm>

Openshaw, S., 1984. The modifiable areal unit problem (Concepts and techniques in Modern Geography, No. 38). Geo Books, pp. 41.

Pfeiffer, D.U., Robinson, T.P., Stevenson, M., Stevens, K.B., Rogers, D.J. and Clements A.C.A., 2008. Spatial Analysis in Epidemiology. Oxford University Press, pp. 160.

Ripley BD, 1976. The second-order analysis of stationary point processes. J Appl Prob 13: 255-266.

Robert, C.P., 2007. The Bayesian Choice. Springer, pp. 602.

Sarkar, D., 2008. Lattice. Multivariate Data Visualization with R. Springer, pp. 280.

Särndal, C.E, Swensson, B. and Wretman, J., 1992. Model Assisted Survey Sampling, Springer Verlag, NY.

SAS (Statistical Analysis System), 2009. <http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/surveylogistic_toc.htm>

APPENDIX A: WORKED EXAMPLES OF THE SAS PROCEDURE APPLICATION

In order to carry out a thorough review of the analyses of time trends which were carried out on data for 2006, examples are presented of possible approaches to the analysis on a selected data set on *Salmonella* spp. in flocks of laying hens (for details on data collection, see the 2006 CSR), by using mostly SAS SURVEYLOGISTIC and GENMOD procedures. Parameter estimates obtained by these approaches, including the one that was actually used for the 2006 data, are compared and discussed.

Selected data

Data on *Salmonella* spp. in flocks of laying hens are in binomial format, the number of positive flocks (*pos*) out of the number of tested flocks (*N*) is available for each MS reporting data, and for each year from 2004 to 2006. Names of the nine reporting MSs were replaced by letters (A to I). The number of laying hens that were raised in each country in 2006 is used as the sampled population (*pop*). Weight (*wt*) is calculated by year and by MS as the ratio *pop*/*N* - the reciprocal of the sampling fraction. The following data step is used to create the data set.

```
data ms;
input id ms $ year N pos pop;
wt = pop/N;
datalines;
1     A     2004     2649      41       5450000
2     B     2004     1009       6       3099504
3     C     2004      815       1       3103333
4     D     2004     4916     112      36157100
5     E     2004      355       3       1988500
6     F     2004     3148     117      41641960
7     G     2004      219      10       2709000
8     H     2004      167       4       1119666
9     I     2004      909       2       5065260
10    A     2005     4735      66       5450000
11    B     2005      913      13       3099504
12    C     2005      817       1       3103333
13    D     2005     5331     166      36157100
14    E     2005      217       6       1988500
15    F     2005     4117     145      41641960
16    G     2005      309      41       2709000
17    H     2005      130       8       1119666
18    I     2005     1109       1       5065260
19    A     2006     4359      88       5450000
20    B     2006      854       3       3099504
21    C     2006      749       0       3103333
22    D     2006     2764      39      36157100
23    E     2006      340       1       1988500
24    F     2006     5008     100      41641960
25    G     2006     1298      29       2709000
26    H     2006      205       3       1119666
27    I     2006      913       1       5065260
;
run;
```

Example 1. Ordinary logistic regression

In the first example, ordinary logistic regressions (OLR), ignoring any of the design characteristics, were fitted using first **SURVEYLOGISTIC** and then **GENMOD**.

```
proc surveylogistic data=ms;
        model pos/N = year;
run;

proc genmod data=ms;
        model pos/N = year / dist=binomial link=logit;
run;
```

PROC SURVEYLOGISTIC automatically implements the binomial distribution and the *logit* link, whereas in PROC GENMOD, distribution and link must be specified since this procedure can be used to fit different generalised linear models. In both procedures, binomial frequency counts are used and, in the SAS manual this format is referred to as *event* / *trials* format (SAS, 2009). Results of the analysis are shown in Table 1.

Table 1. **Parameter estimates obtained by ordinary logistic regression analysis of the effect of year of sampling on the prevalence of *Salmonella* spp. infection in flocks of laying hens, in nine EU MSs, from 2004 to 2006**

Procedure	Parameter estimate (standard error)	
	Intercept	Year
SURVEYLOGISTIC	**252.9 (75.3)**	**-0.128 (0.038)**
GENMOD	**252.9 (80.3)**	**-0.128 (0.040)**

Point estimates of regression parameters which were obtained by using the two procedures are exactly the same, whereas some minor differences are observed in the standard errors, which can be explained by different estimation methods. In fact, in SURVEYLOGISTIC, the design-based variances of parameters are estimated using a Taylor series expansion approximation (SAS, 2009). In GENMOD, on the other hand, the maximum likelihood theory is used, and asymptotic standard errors can be derived from the second derivatives of the likelihood function.

Example 2. Sampling from finite populations

In a survey, sampling can be carried out from a finite population (as opposed to an infinite or very large population). The consequence of a finite population size is that parameters can be estimated with a relatively high degree of precision, as expressed by reduced standard errors of estimates. No finite population correction was taken into account in the 2006 analysis. Finite population correction can only be incorporated in a design-based approach, as available in SURVEYLOGISTIC. This design feature is illustrated in this example by hypothetically assuming that 25%, 50% and 100% of the population were sampled. Moreover, in the following SAS statement, an input data set was created in which the _TOTAL_ variable corresponds to the finite population size.

```
data population;
input id _TOTAL_;
datalines;
1           5450000
2           3099504
3           3103333
4          36157100
5           1988500
6          41641960
7           2709000
8           1119666
9           5065260
10          5450000
11          3099504
12          3103333
13         36157100
14          1988500
15         41641960
16          2709000
17          1119666
18          5065260
19          5450000
20          3099504
21          3103333
22         36157100
23          1988500
24         41641960
25          2709000
26          1119666
27          5065260
;
```

In the following SAS statements, r= indicates the sampling rate, whereas total= indicates the input data set (population) where the population size is expressed as _TOTAL_. Results are shown in Table 2.

```
proc surveylogistic data=ms r=25;
title ‘Sampling 25% of the population’;
        model pos/N = year;
run;

proc surveylogistic data=ms r=50;
title ‘Sampling 50% of the population’;
        model pos/N = year;
run;

proc surveylogistic data=ms r=100;
title ‘Sampling 100% of the population’;
          model pos/N = year;
run;

proc surveylogistic data=ms total=population;
title ‘Sampling from _TOTAL_‘;
        model pos/N = year;
run;
```

Table 2. **Parameter estimates obtained by logistic regression analysis of the effect of year of sampling on the prevalence of *Salmonella* spp. infection in flocks of laying hens, in nine EU MSs, from 2004 to 2006. Varying sampling fractions-rates (from 25% to 100%) were simulated. Moreover, finite population size was set equal to the laying hen population size for each MS (Sampling from _TOTAL_)**

Sampling rate	Parameter estimate (standard error)	
	Intercept	Year
25%	**252.9 (65.2)**	**-0.128 (0.033)**
50%	**252.9 (53.2)**	**-0.128 (0.027)**
100%	**252.9 (0.0)**	**-0.128 (0.000)**
Sampling from _TOTAL_	**252.9 (74.9)**	**-0.128 (0.037)**

Parameter estimates are not affected by the sampling fraction (Table 2), whereas, as expected, the standard error is smaller when larger fractions of the populations are sampled. When the sample covers the full population (sampling fraction = 100%), the estimate equals the population value with standard error equal to zero. When the finite population was set equal to the laying hen population for each MS (sampling from _TOTAL_) the finite population correction has only a small effect on standard errors, since the sample is only a limited fraction of the full population (the standard error is only slightly smaller than in ordinary logistic regression, see Table 1). This last result would probably have been different if the flock population would have been used, instead.

Example 3. Accounting for disproportionate sampling

In the following analyses, some of the design characteristics will be incorporated when fitting the model, starting with accounting for the disproportionate sampling scheme, using weights to correct estimates. First SURVEYLOGISTIC will be used and then the GENMOD procedure.

```
proc surveylogistic data=ms;
        model pos/N = year;
        weight wt;
run;

proc genmod data=ms;
        model pos/N = year / dist=binomial link=logit;
        weight wt;
run;
```

In the above codes, the weight statement identifies a numerical variable (wt) which is used as the sampling weight for each observation in the data set.

Table 3. **Parameter estimates obtained by weighted logistic regression analysis of the effect of year of sampling on the prevalence of *Salmonella* spp. infection in flocks of laying hens, in nine EU MSs, from 2004 to 2006. Sampling weights were calculated as the reciprocals of the sampling fractions: laying hen population divided by number of tested laying hen flocks**

Procedure	Parameter estimate (standard error)	
	Intercept	Year
SURVEYLOGISTIC	**460.1 (84.2)**	**-0.231 (0.042)**
GENMOD	**460.2 (0.94)**	**-0.231 (0.0005)**

As in example 1, SURVEYLOGISTIC and GENMOD procedures produce exactly the same point estimates. These are different from the estimates from Example 1 as they are now corrected for disproportionate sampling across the countries through the weights. The standard errors reported by SURVEYLOGISTIC are similar to those of Example 1, but procedure GENMOD seems to produce unrealistically small standard errors. The reason for this erroneous reporting is that the weights need to be standardised. If not standardised, the procedure acts as if the sample size was much larger, as indicated by the output (not mentioned above) whereas the sum of weights should remain 27. This can be rectified by **standardising the weight** through the following SAS code.

```
data msb;
        set ms;
        wts=wt*27/165485.25;
run;

proc genmod data=msb;
        model pos/N = year / dist=binomial link=logit;
        weight wts;
run;
```

Standardised weight for observations belonging to stratum *s* is obtained by the equation:

$$SW_s = \frac{W_s \cdot \sum_{s=1}^{m} n_s}{\sum_{s=1}^{m} W_s}$$

Where *Ws* = weight for stratum *s* (wt in the SAS statement); $\sum_{s=1}^{m} n_s$ = sum of all observations across all strata (27 in the example); $\sum_{s=1}^{m} W_s$ = sum of all weights across all strata (165,485.25 in the example).

Table 3 bis. **Parameter estimates obtained by weighted logistic regression analysis of the effect of year of sampling on the prevalence of *Salmonella* spp. infection in flocks of laying hens, in nine EU MSs, from 2004 to 2006. Sampling weights were calculated as the reciprocals of the sampling fractions and subsequently standardised**

Procedure	Parameter estimate (standard error)	
	Intercept	Year
GENMOD	**460.2 (72.4)**	**-0.231 (0.036)**

As shown in Table 3 bis, the GENMOD analysis with standardised weights leads to standard errors in line with (but somewhat smaller) than those reported by SURVEYLOGISTIC.

Example 4. Taking into account correlation among observations from the same MS

In this example, it is assumed that data from the same MS over consecutive years are correlated, as they are expected to share common environmental, hygienic and other conditions. There are several modelling approaches available to take into account non-independence among observations, and the subsequent correlation among outcomes. However, this example is limited to options available in SURVEYLOGISTIC and GENMOD procedures. In SURVEYLOGISTIC, the sampling design can be specified by adding the cluster statement. In this case, when there are clusters, or primary sampling units (PSUs), in the sample design, the procedure estimates the variance from the variation **among** the PSUs (http://support.sas.com/rnd/app/da/new/dasurvey.html).

The GENMOD procedure, using the repeated statement, implements GEE, where the quasi-likelihood approach assumes only a relationship between the mean (μ) and the variance Var(Y), rather than a specific probability distribution for Y. It allows for departures from the usual assumptions, such as over-dispersion caused by correlated observations or unobserved explanatory variables (Agresti, 2002). In GEE, the structure of the correlation among repeated observations form the same subjects (in our example, the member stated) can be chosen. Here the exchangeable correlation structure is shown, which is based upon the assumption that correlation is the same among any two observations from the same subject, regardless of the time interval separating them. Alternatively, the autoregressive structure, AR(1), is based upon the assumption that correlation decreases with increasing time between observations.

```
proc surveylogistic data=ms;
        cluster MS;
        model pos/N = year;
run;

proc genmod data=ms;
        class MS;
        model pos/N = year / dist=binomial link=logit;
        REPEATED SUBJECT=MS / TYPE=EXCH;
run;

proc genmod data=ms;
        class MS;
        model pos/N = year / dist=binomial link=logit;
        REPEATED SUBJECT=MS / TYPE=AR(1);
run;
```

Table 4. **Parameter estimates obtained by logistic regression analysis of the effect of year of sampling on the prevalence of *Salmonella* spp. infection in flocks of laying hens, in nine EU MSs, from 2004 to 2006. Potential correlation among observations from the same MS is taken into account**

Procedure	Parameter estimate (standard error)		
	Intercept	Year	ρ
SURVEYLOGISTIC cluster	**252.9 (146.6)**	**-0.128 (0.073)**	**–**
GENMOD, GEE Exch	**301.48 (149.41)**	**-0.1523 (0.0745)**	**0.45**
GENMOD, GEE AR(1)	**301.01 (170.14)**	**-0.1521 (0.0848)**	**one year: 0.55** **two years: 0.30**

ρ = working correlation; Exch = Exchangeable correlation structure; AR(1) = first order autoregressive correlation structure.

Since no weighting has been used, the point estimates are again equal to or close to those of Example 1 (ordinary logistic regression). By comparing the standard errors of the estimates it can be observed that these are now almost doubled, as compared to Example 1. This is a typical phenomenon when accounting for the non-independent nature of the data, as the effect of clustered or correlated data can be interpreted as an implicit sample size reduction (a reduction in effective independent units of information). Ignoring non-independence typically leads to overly optimistic standard errors, and consequently to erroneous and misleading inference. Empirical corrected standard error estimates, which are calculated based on the working correlation, are reported in Table 4.

It is worth noting that also the point estimates change slightly in the GEE approach, compared with the ordinary logistic regression, not taking into account non-independence of observations (Example 1). This is a result of the formulation of the GEE, in which the estimation of the correlation parameter also affects the estimation of the probability parameter.

The GENMOD procedure allows different working correlation structures to be specified. Here two plausible choices have been shown and the results obtained for each of them. Similar results were obtained for both working correlation assumptions.

Below are set out a description of analyses taking both complications (weighting together with intra-MS) correlation, into account.

Example 5. Taking simultaneously into account disproportionate sampling (with weighting) and non-independence of observations

Here, both complications of data characteristics, weighting and clustering, are combined in both SAS procedures.

```
proc surveylogistic data=ms;
        cluster MS;
        model pos/N = year;
        weight wt;
run;

proc genmod data=ms;
        class MS;
        model pos/N = year / dist=binomial link=logit;
        weight wt;
        REPEATED SUBJECT=MS / TYPE=EXCH;
run;

proc genmod data=ms;
      class MS;
      model pos/N = year / dist=binomial link=logit;
      weight wt;
      REPEATED SUBJECT=MS / TYPE=AR(1);
run;
```

Table 5. **Parameter estimates obtained by logistic regression analysis of the effect of year of sampling on the prevalence of *Salmonella* spp. infection in flocks of laying hens, in nine EU MSs, from 2004 to 2006. Disproportionate sampling and potential correlation among observations from the same MS are taken into account**

Procedure	Parameter estimate (standard error)		
	Intercept	Year	ρ
SURVEYLOGISTIC cluster & weight	**460.1 (76.9)**	**-0.230 (0.038)**	**–**
GENMOD Exch weight	**459.9 (72.5)**	**-0.230 (0.036)**	**0.36**
GENMOD AR(1)	**449.21 (78.3)**	**-0.226 (0.039)**	**one year: 0.39** **two years: 0.15**

ρ = working correlation; Exch = Exchangeable correlation structure; AR(1) = first order autoregressive correlation structure.

The analysis of SURVEYLOGISTIC and GENMOD are now strikingly similar, although they originate from different paradigms (design-based SURVEYLOGISTIC versus model-based GENMOD). The point estimates changed again through the weighting to the same value as in Example 3 for SURVEYLOGISTIC or to a value very close to GENMOD. When looking at the standard errors of the year parameter, the effect is observed of combining weighting and clustering which leads to standard errors not far from their values without clustering (see Table 3 for SURVEYLOGISTIC, and Table 3 bis for GENMOD with standardised weights). Therefore, in this data, after the inclusion of weights, taking into account non-independence does not affect standard errors of estimates. This unexpected observation might be due to the fact that varying weights (possibly associated with varying sample sizes) may affect correlation among observations from the same MS.

Standardisation of weights in GENMOD is no longer necessary since it leads to exactly the same results. The reason is that the computation of standard errors for GEE parameter estimates implicitly standardised weights. Also here several working correlation assumptions have been used. The results obtained point again to the fact that model fits are similar.

Example 6. MSs as strata

Another feature of SURVEYLOGISTIC is that stratification can be incorporated in the model. If the sample design has stratification at multiple stages, only the first-stage strata in the STRATA statement can be identified. The analysis can be done also using MS as stratum, first without using weights, and then incorporating the disproportionate sampling into the model. When the design is stratified, the procedures pool stratum variance estimates to compute the overall SAS variance estimate.

```
proc surveylogistic data=ms;
        model pos/N = year;
        strata MS;
run;

proc surveylogistic data=ms;
        model pos/N = year;
        strata MS;
        weight wt;
run;
```

Table 6. **Parameter estimates obtained by logistic regression analysis of the effect of year of sampling on the prevalence of *Salmonella* spp. infection in flocks of laying hens, in nine EU MSs, from 2004 to 2006. MSs are considered as strata. Moreover, disproportionate sampling and potential correlation among observations from the same MS are alternatively taken into account**

Procedure	Parameter estimate (standard error)	
	Intercept	Year
SURVEYLOGISTIC strata	**252.9 (75.3)**	**-0.1281 (0.0375)**
SURVEYLOGISTIC strata & weight	**460.1 (84.2)**	**-0.2313 (0.042)**

SURVEYLOGISTIC with strata produces the same parameter estimates and standard errors as in Example 1 (ordinary logistic regression), while when weights are used, then the results are the same as the one obtained from Example 3 (inclusion of weight, without cluster statement). Therefore, in the absence of weights, the strata MS statement yields smaller standard errors than the cluster MS statement. The strata statement allows getting strata specific information and can be used to incorporate finite strata population corrections (which have no effect in our example).

Example 7. Accounting for over-dispersion of binomial counts through the SCALE parameter (GENMOD)

This example is limited to PROC GENMOD and deals with the general concept of "over-dispersion" as a greater than expected variability of binomial counts. Over-dispersion can be measured by the ratio of the deviance or Pearson goodness-of-fit measures, expected to be about one if model assumptions, being independence and constant probability, are met, and (much) larger than one if such assumptions are not met. In GENMOD, using the scale=D or the scale=P statement leads to the computation of corrected standard errors to deal with over-dispersion. This is another approach to take into account non-independence of observations since it can lead to over-dispersion of binomial counts. Again, as with GEE, this approach does not require the weights to be standardised, as the correction of standard errors includes an implicit standardisation.

```
proc genmod data=ms;
        class MS;
        model pos/N = year / dist=binomial link=logit scale=D;
        weight wt;
run;
```

Table 7. **Parameter estimates obtained by logistic regression analysis of the effect of year of sampling on the prevalence of *Salmonella* spp. infection in flocks of laying hens, in nine EU MSs, from 2004 to 2006. Over-dispersion of binomial counts is taken into account**

Procedure	Parameter estimate (standard error)		
	Intercept	Year	Scale
GENMOD scale=D; Weight;	**460.15 (293.8)**	**-0.231 (0.147)**	**317.7**

The output shows the same point estimates as in earlier GENMOD analyses (including weights). The standard errors are however larger than any other former analysis. This is a typical phenomenon as this approach intends to correct for any source of over-disperion in a conservative way.

Example 8. Dummy coding of year effect

In this example, year is not included in the model as a continuous predictor (assuming linear effect) but as a categorical variable. The year 2006 is used as the reference. This is automatically accomplished in GENMOD after including year in the class statement. In SURVEYLOGISTIC, two dummy variables were created representing years 2004 and 2005, meaning that these variables take the value '1' if year is equal to one of them, and are equal to '0' otherwise. It is in, fact, important to recognise that parameterisation of the models in SURVEYLOGISTIC and GENMOD are different. Here only the model that accounts for the cluster will be focussed on as well as the disproportionate sampling.

```
proc surveylogistic data=ms;
        cluster MS;
        model pos/N = year1 year2;
        weight wt;
run;

proc genmod data=ms;
        class MS year;
        model pos/N = year / dist=binomial link=logit;
        weight wt;
        REPEATED SUBJECT=MS / TYPE=EXCH COVB CORRW MODELSE;
run;

proc genmod data=ms;
        class MS year;
        model pos/N = year / dist=binomial link=logit;
        weight wt;
          REPEATED SUBJECT=MS / TYPE=AR(1) COVB CORRW MODELSE;
run;
```

Table 8. **Parameter estimates obtained by logistic regression analysis of year of sampling on the prevalence of *Salmonella* spp. infection in flocks of laying hens, in nine EU MSs, from 2004 to 2006. The non-linear effect of year is analysed after dummy coding**

Procedure	Parameter estimate (standard error)			
	Intercept	2004 vs 2006	2005 vs 2006	ρ
SURVEYLOGISTIC Cluster & Weight	**-4.1541 (0.1526)**	**0.5494 (0.07)**	**0.7425 (0.13)**	**–**
GENMOD Exch weight	**-4.1541 (0.1439)**	**0.5494 (0.07)**	**0.7425 (0.12)**	**0.57**
GENMOD AR(1)	**-4.1541 (0.1439)**	**0.5494 (0.07)**	**0.7425 (0.12)**	**one year: 0.66** **two years: 0.43**

ρ = working correlation; Exch = Exchangeable correlation structure; AR(1) = first order autoregressive correlation structure.

In Table 8, the parameter comparing the risk of *Salmonella* spp. in laying hens in 2005 and in 2006 is greater than the parameter for 2004 vs 2006, indicating that the effect of year on *Salmonella* spp. was not linear. Therefore, these results indicate that the inclusion of year in the model, as a numerical variable is not ideal. On the other hand, the main interest of the 2006 analysis was on a linear trend at EU level, rather than on a non-linear trend. The results obtained from SURVEYLOGISTIC and GENMOD procedures are comparable. The choices of the working correlation structure do not have any impact on the empirical estimates. Little changes were observed for the model-based outputs.

APPENDIX B: STATISTICAL DETAILS

In the case of a laboratory test detecting infection by a zoonotic agent, if the test is carried out only on a single unit, the outcome variable will assume the value of one in the case of a positive test result, while it will correspond to zero in the case of a negative test. The purpose of this variable is to observe if after drawing a sample of *n*=1 unit from a finite population the presence/absence of the characteristic in which interest is shown can be detected (for example an infection). In a formal way:

$$y_k = \begin{cases} 1 \text{ if animal k is infected} \\ 0 \text{ if animal k is non - infected} \end{cases}$$

y_k is well known as the Bernoulli random variable (r.v.) with probability of success equal to: $P(y_k = 1) = \pi_k = \pi$ and defined by : $y_k \approx Bernoulli(\pi)$. The probability π_k is called prevalence of infection, and could be either constant $\pi_k = \pi$ or dependent on the specific characteristics of the unit (age, weight, etc.) as well as more global characteristics (such as environmental conditions, hygienic conditions, farm or flock conditions).

If the experiment is performed on the entirety of the *N* units of the finite population, the population results will be divided into two separated and complementary classes:

- the first one is *D* which includes those units with modality 1 (for instance infected animals) and is equal to $\sum_{k=1}^{N} y_k = N_D$
- while the second one is $\overline{D}$ including instead those units with modality 0 and equals: $N_{\overline{D}} = N - N_D$ (for example non-infected animals).

Therefore N_D and $N_{\overline{D}}$ represent the absolute frequency of the animals, respectively infected and not infected. It will follow that the proportion or prevalence of infected units among the population is $\pi = N_D / N$ coinciding with the probability of extracting from the population a unit possessing the characteristic under observation while the proportion of non-infected units is:

$$1 - \pi = 1 - N_D / N = N_{\overline{D}} / N .$$

In the case of extraction from the finite population of a simple random sample of *n* units with replacement (*srswr*) to estimate π, in the sample *y* infected units are noticed and *n-y* non-infected; *y* represents the number of successes out of *n* independent trials with the same probability. In a sample of *n* units the variable can assume the values *y*= 0,1,2,…,*n* with the probability $P(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$. Considering the way the experiment has been built *y* represents the sum of *n* independent and equally distributed Bernoulli r.v. and it is well known as the Binomial r.v.: $y \approx Bi(n;\pi)$. The correct estimation for π is $p = y/n$; when the sample changes, according to the design-based formulation, p changes and describes the *Sample Proportion* r.v. which is obtained multiplying the Binomial times $1/n$. The variance of this r.v. is estimated by the quantity $v(p) = p(1-p)/(n-1)$.

In case of a simple random sample without replacement (*srs*) the estimation for π is still $p = y/n$ while the estimation for the variance is $v_{wr}(p) = 1 - f\frac{p(1-p)}{n-1}$, where $f = n/N$ is the finite population correction. The *srswr* can be considered a sample drawn from an infinite population while the *srs* is a sample drawn from a finite population. In general, in real cases, populations are finite and there is no interest in detecting the same unit more than once, therefore the sample *srs* is the one normally used. Moreover when the population is large and the sample size is very small, compared to the population, the fraction of survey tends to zero and the variance for the estimators in the two designs tends to be the same; so treating finite population as infinite population is a very reasonable mathematical simplification.

As a consequence of the Central Limit Theorem the Sample Proportion *r.v.* is approximately distributed normally; therefore the confidence interval for π can be defined for a confidence level of $(1-\alpha)$:

$$p \pm \left[z_{\alpha/2}\sqrt{(1-f)\frac{p(1-p)}{n-1}} + \frac{1}{2n} \right]$$

where $1/2n$ is the correction it should be brought to improve the approximation of a discrete variable as a continuous variable and $z_{\alpha/2}$ is the percentile of order $\alpha/2$ of the Standard Normal distribution. For small (<0.1) or large (>0.9) values of proportions, the exact binomial confidence interval is more appropriate.

Analysis of binary dependence variable

It may be interesting to analyse the dependence of the study variable from one or more "covariates" (explanatory variables). When y_k is a Bernoulli r.v. with success probability π_k (e.g. the animal is infected), the dependence of variable y_k on a covariate is modelled through *logistic regression*. Let x_k denote the value of such a covariate for subject k (for simplicity only one covariate is assumed).

The logistic function or logit is

$$\text{logit}(\pi_k) = \log\left(\frac{\pi_k}{1-\pi_k}\right) = \beta_0 + \beta_1 x_k$$

where the ratio between success probability and failure probability $\frac{\pi_k}{1-\pi_k}$ is called *odds* and $\pi_k = \frac{\exp(\beta_0 + \beta_1 x_k)}{1+\exp(\beta_0 + \beta_1 x_k)}$ is the probability that y_k =1 given the model; in other words it is the probability that the animal is infected conditionally to the covariate x_k. The logit link has a major advantage that it leads to the interpretation of a log odds ratio for the regression slope β_1, which allows the method to be applied for retrospectively collected case-control data, and for cross-sectional surveys

Assuming the model is correctly specified (meaning that it approximates the truth satisfactorily), the regression coefficients can be estimated by the simple random sample $\{(x_k, y_k)\}_{k=1}^{n}$ through the maximum likelihood function (see e.g. McCullagh and Nelder 1989, Agresti 2002).

Note that in many cases, subjects are sharing the same covariate information, such as herd characteristics. In case the above Bernoulli model is formulated with all information aggregated at this higher level of structure, it can be equivalently formulated as a "binomial" model (see below).

Cluster and stratified sampling

Often the data structure is hierarchical and as a consequence, data are correlated within natural clusters. Let π_i be the probability for a study object to be positive in country *i*, and let n_{ij} be the number of study objects in cluster *j* from country *i*. The starting point for inference on prevalence is the binomial distribution (model) for the number of positive y_{ij} in cluster *j* in country *i*:

$$y_{ij} \sim \mathrm{Bin}(n_{ij}, \pi_i)$$

In a case of a fully random sample the y_{ij} could be combined in a straightforward way to estimate the prevalence for country *i*. Typical complications however are the following:

- the sample is not a random sample, but typically the sample is the result of a complex survey design, such as stratified sampling;
- the independence assumption on the binomial distribution is violated because outcomes from the same cluster are expected to be more alike (correlated) as compared to outcomes from different clusters (hierarchical correlation structure);
- the constant probability assumption is violated because samples, even from the same cluster or even from the same sampling unit, might have different probabilities e.g. to be infected (heterogeneity of probability).

APPENDIX C: ILLUSTRATED EXAMPLE OF THE BAYESIAN APPROACH TO STATISTICAL ANALYSIS

Hereafter is reported an illustrative example based on the Zoonoses Report data from 2006. Prevalence of *Salmonella* spp. in flocks of laying hens was available for only 11 countries in the EU in the following format: per country, the total number of flocks tested, total number of positives and the population sizes for three consecutive years (2004, 2005 and 2006). In addition, information may be available for all 27 countries, such as population sizes. Such additional information, usually discarded, can be of value and easily incorporated within a Bayesian framework, as shown below.

The model chosen is a binomial/logistic random effect model. For a given MS, the number of positive flocks (*n.positives*) is assumed to be a binomial draw from the total number tested (*n.tested*) with prevalence *P*. Total population sizes (*n.tot*) are seen to be large enough to validate this assumption.

$$n.positives_{(MS)} \sim Bin\left\lfloor n.tested_{(MS)}, P_{(MS)}\right\rfloor$$

Then, prevalence probabilities need to be converted into the logistic scale. To describe the time effect, a logistic regression model is assumed, with random effects on the parameters:

$$\text{Logit}\ [\text{P}_{(MS)}] = A_{(MS)} + B_{(MS)}\text{*year}$$

Random effects describe here inter-country variability, to account for the fact that 16 MSs out of the 27 are missing. Random effects are assumed to be normal but any other relevant choice could be made without much incidence on implementation.

At MS level, the parameter *B* stands for the trend parameter in the logit scale. A similar parameter needs to be also defined at EU level, accounting for the respective weights of each country proportional to their population sizes. For this purpose, and for the sake of this short example, one could make the following approximation, valid for small prevalence:

$$P_{(MS)} \approx P_{(MS)}^{year=0}\left(1 + B_{(MS)} * year\right)$$

This allows a simple expression for the time slope at EU level:

$$EU\_slope = \frac{\sum_{MS=1}^{11} n.tot_{(MS)} * B_{(MS)} * P_{(MS)}^{year=0}}{\sum_{MS=1}^{11} n.tot_{(MS)} * P_{(MS)}^{year=0}}$$

The purpose of statistical evaluation is now to assess how likely it is that this slope is negative, and to estimate its size.

The Bayesian evaluation of this model is made using WinBUGS 1.4, with standard non-informative priors. The code is reported at the end of this appendix.

The average (posterior mean) yearly decrease in prevalence was estimated to be about -18% at EU level, with a relatively wide 95%-confidence (credibility) interval [-27%; -10%]. The posterior distribution for the EU slope parameter is represented in Figure 5 below.

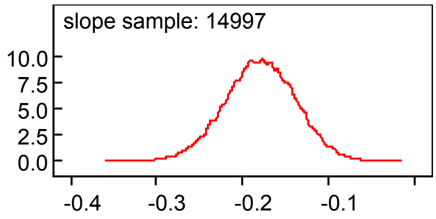

Figure 5: **WinBUGS output showing posterior distribution of the average slope, using a random-effect model, without accounting for the population sizes of countries with missing prevalence**

The posterior probability of the slope being negative was greater than 99%, hence showing a high level of confidence in negative trend. Subsequently, the same model was fitted on the 27 MSs, using the available population sizes for each of them. The procedure now accounts for the relative weights (total population sizes) of all the 27 countries and the Bayesian scheme propagates uncertainty in a sound manner. Within this set-up, the EU slope was now estimated to be -22% (see Figure 6). But as expected, uncertainty of this estimate has now increased, as shown by Figure 6. The 95%-credibility interval is now estimated to be [-53%,+6%] and the posterior probability for the EU slope being negative was about 95.3% so just above the "classical level" of statistical significance (5%).

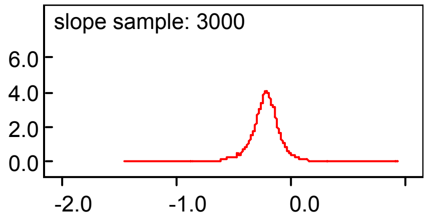

Figure 6: **WinBUGS output showing posterior distribution of the average slope, using a random-effect model, when accounting for the population sizes of countries with missing prevalence**

It is interesting to notice that an attempt to reproduce the results using the same model with SAS proc GLIMMIX (with maximum likelihood inference) was unsuccessful (results not shown), even in simple cases without accounting for weights for missing countries. However, the same model may be coded in SAS using newly added Bayesian features. As a matter of fact, SAS 9.2 now offers a Bayesian module in proc GENMOD, which could be convenient in the simplest cases. This module was tested here on the same example as above (without accounting for population sizes of countries with missing prevalence). The output was very consistent with the WinBUGS output as illustrated by the posterior simulations shown in Figure 7, together with examples of diagnostic plots of MCMC convergence.

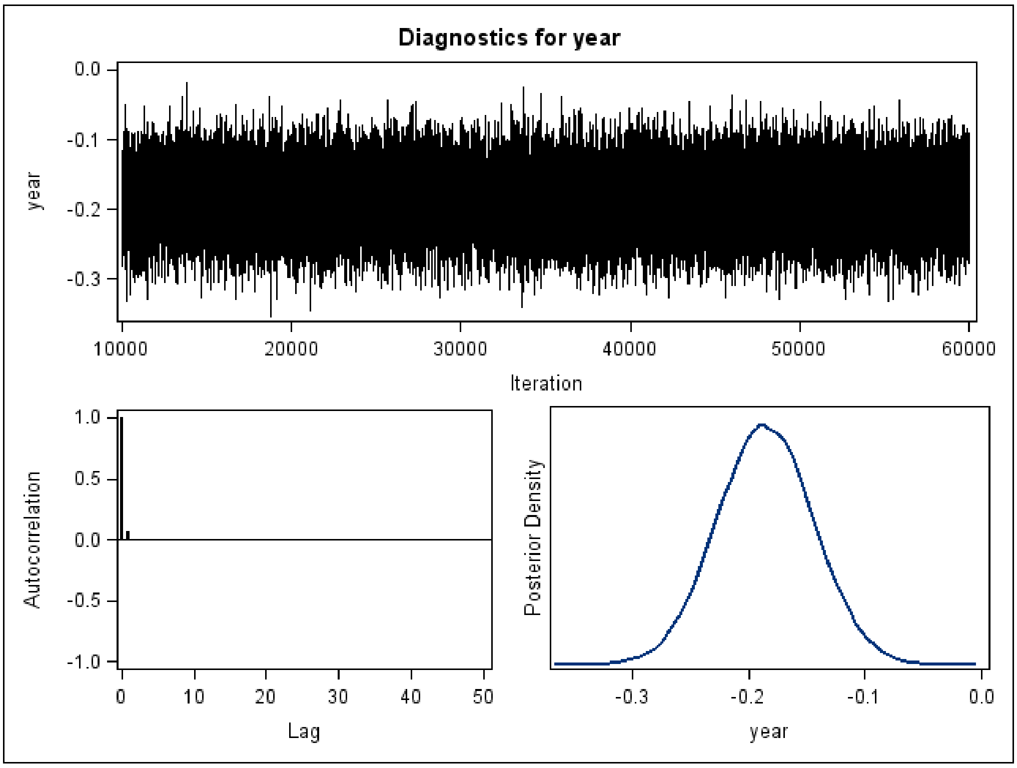

Figure 7: **SAS proc GENMOD output showing 3 diagnostic plots: the posterior distribution of the average slope (bottom right), Markov chain simulations (top) and auto-correlation of simulations (bottom left)**

The posterior mean of the yearly decrease of prevalence was estimated to be -18.5% with a 95% credibility interval of [-10%; -27%]. Although proc GENMOD does not fully estimate a random effect model, outputs using Bayesian inference are quasi identical with the WinBUGS approach in this simplistic case.

Bayesian code to fit a random effect Binomial-Logistic model for trend analysis

```
model
        {
        for( ctr in 1:11 ) {
                for( yr in 1:3 ) {
                  #Binomial/logistic model
                  np[ctr,yr] ~ dbin(p[ctr,yr],nn[ctr,yr])
                p[ctr,yr] <- exp(b[ctr]+(yr-1)*a[ctr])/(1+exp(b[ctr]+(yr-1)*a[ctr]))
                }
                #random effect of the regression
                a[ctr] ~ dnorm(m.a,tau.a)
                b[ctr] ~ dnorm(m.b,tau.b)

                #needed for the final calculation of the EU slope
                p1[ctr]<- N[ctr]*exp(b[ctr] )*a[ctr]
                p2[ctr]<- N[ctr]*exp(b[ctr] )
        }

        #non-informative priors
        tau.a ~ dgamma(0.001,0.001)
        tau.b ~ dgamma(0.001,0.001)
        m.a ~ dnorm(0.0,1.0E-6)
        m.b ~ dnorm(0.0,1.0E-6)

        #final EU slope
        slope<-sum(p1[])/sum(p2[])
}
```

APPENDIX D: Examples of choropleth maps using different class interval schemes

In order to show the impact of changing class interval schemes, examples of choropleth maps using equal-interval breaks, quantile breaks and natural breaks are reported in this appendix.

All the three following maps show the prevalence of the five targeted serovars *Salmonella* Enteritidis, *S*. Typhimurium, *S*. Infantis, *S*. Virchow or *S*. Hadar in *Gallus gallus* breeding flock during the production period, 2007 (see Figure SA15 in the CSR 2007).

As it is possible to note from the following choropleth maps, the same data can be visualised in different ways, depending on the classification scheme used to divide continuous prevalence data into categories.

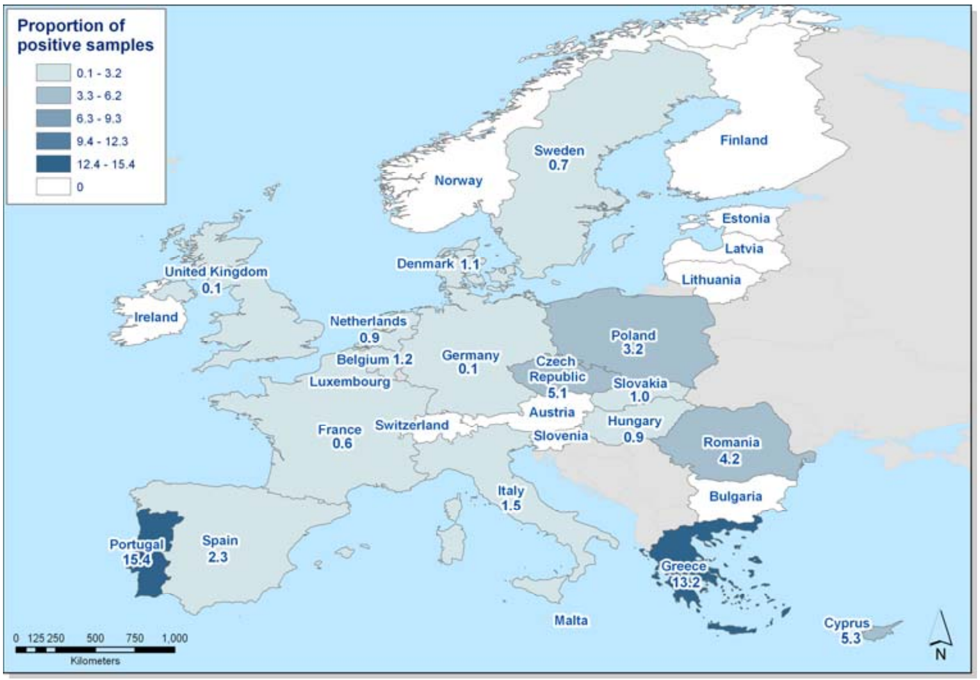

Figure 8. **Choropleth map showing the prevalence of the five targeted serovars *Salmonella* Enteritidis, *S*. Typhimurium, *S*. Infantis, *S*. Virchow or *S*. Hadar in *Gallus gallus* breeding flocks during the 2007 production period. Equal interval breaks are used as a classification scheme.**

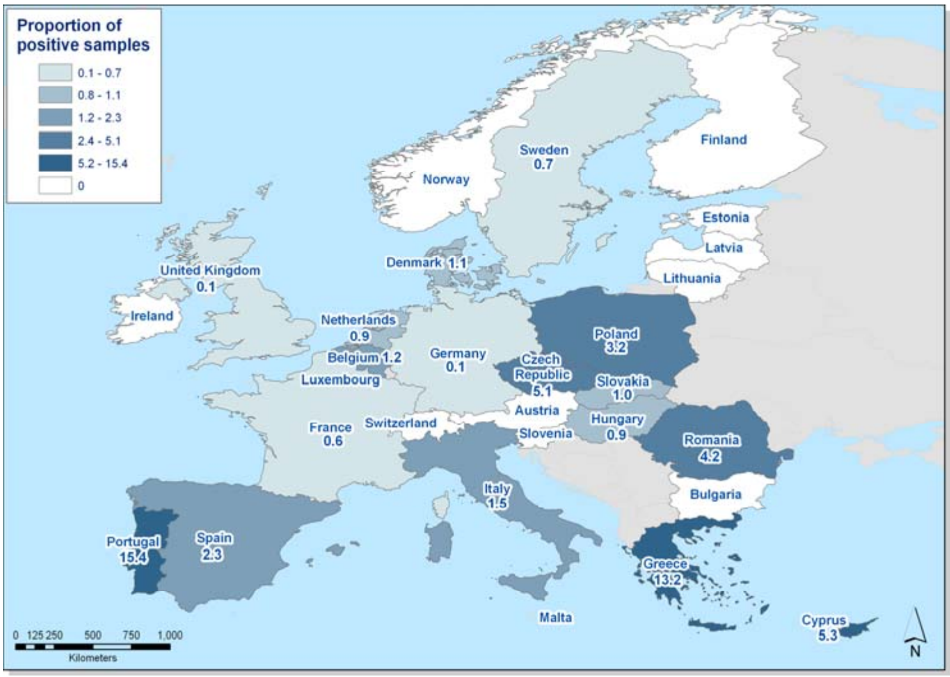

Figure 9. **Choropleth map showing the prevalence of the five targeted serovars *Salmonella* Enteritidis, *S*. Typhimurium, *S*. Infantis, *S*. Virchow or *S*. Hadar in *Gallus gallus* breeding flocks during the 2007 production period. Quantile breaks are used as a classification scheme.**

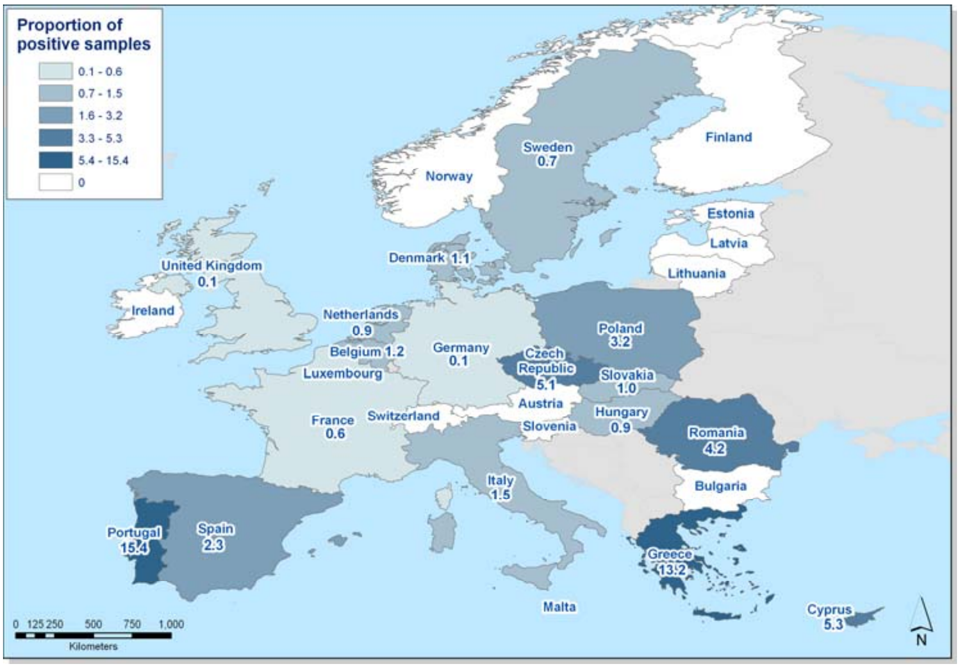

Figure 10. **Choropleth map showing the prevalence of the five targeted serovars *Salmonella* Enteritidis, *S*. Typhimurium, *S*. Infantis, *S*. Virchow or *S*. Hadar in *Gallus gallus* breeding flocks during the 2007 production period. Natural breaks are used as a classification scheme.**

APPENDIX E: SPATIAL STATISTICS

E.1 SPATIAL METHODS FOR INVESTIGATING GLOBAL CLUSTERING

MORAN'S I

The Moran's I statistic is one of the oldest indicators of spatial autocorrelation (Moran, 1950). It is considered a standard for determining spatial autocorrelation and it is applied to both point and area data.

Moran's I is similar to Pearson's correlation coefficient, and quantifies the similarity of an outcome variable among areas that are defined as spatially related. A weight matrix is used to define the spatial relationships so that regions close in space are given a greater weight when calculating the statistic than those that are distant. Therefore, Moran's I can be defined as a weighted correlation coefficient that seeks to determine deviations from spatial randomness of values across locations.

When no correlation exists between neighbouring values (e.g. disease incidence rates of neighbouring regions) the expected value of Moran's I approaches zero. Therefore, a Moran's I close to zero indicates the null hypothesis of no clustering, which means random distribution of the values across space. Moran's I varies between -1 and +1. In general, a Moran's I value greater than zero (near +1) indicates clustering (e.g. high disease incidence rates are found in neighbouring regions) while an index value below zero (near -1) indicates dispersion (e.g. high incidence regions are close to low incidence regions).

Moran's I is calculated as follows: $I = \dfrac{N\sum_i \sum_j W_{i,j}(X_i - \bar{X})(X_j - \bar{X})}{(\sum_i \sum_j W_{i,j})\sum_i (X_i - \bar{X})^2}$

where: N is the number of regions

X_i is the variable value at a particular location

X_j is the variable value at another location

$\bar{X}$ is the mean of the variable

W_{ij} is a weight applied to the comparison between location i and location j (often it is a simple contiguity matrix. If zone j is adjacent to zone i, the interaction receives a weight of 1; at other times it can be a distance-based weight which is the inverse distance between locations i and j ($1/dij$))

For significance testing, a z-score under the assumption of randomisation is calculated, as follows: $Z(I) = \dfrac{I - E(I)}{SD_{E(I)}}$ E(I) is the expected value of Moran's I for a random distribution, this is calculated as -1/(n-1), where n is the number of features in the whole study area. $SD_{E(I)}$ is the standard deviation of I for the expected distribution.

GEARY'S C

The Geary's contiguity ratio (Geary, R.C. 1954), or Geary's C, is another weighted measure of spatial autocorrelation, but, differently to Moran's I, it emphasises the differences in values between pairs of observations, rather than the co-variation between the pairs. Its value ranges between zero and two. If values of a zone are spatially unrelated to any other zone, the expected value of C will be one, no spatial autocorrelation. C smaller (greater) than one means positive (negative) spatial autocorrelation.

Geary's C is given by: $C = \frac{(N-1)\sum_i \sum_j W_{i,j} (X_i - X_j)^2}{2(\sum_i (X_i - \bar{X})^2)(\sum_i \sum_j W_{ij})}$

where W_{ij} are elements of a weight matrix for which a value of one indicates that a pair of two observations, X_i and X_j, are in the same distance class d and a value of zero indicates all other cases. $\bar{X}$ is the mean of the variable of interest.

CUZICK AND EDWARDS TEST

When exact locations of events are known, it is possible to use the Cuzick and Edwards test (1990), developed for spatial clustering and taking into account the inhomogeneous distribution of populations. The test is based on the comparison of the spatial distribution of the case with control locations (controls are selected from the same source population) in a study region.

The method examines the *k* nearest neighbours to each case. If cases are geographically clustered, most of those *k* nearest neighbours will also be cases. The test statistic T_k is the number of cases that are nearest neighbours (nearest neighbours = *k*) to each individual case. When cases are clustered, a case will have as a nearest neighbour another case and T_k will be large. Therefore, a large value of T_k indicates that cases are spatially clustered in relation to the underlying geographical distribution of controls. Conversely, when all cases have controls as nearest neighbours, T_k will be zero. The test allows the choice of the number of nearest neighbours (*k*) to be included in the analysis.

The method includes a significant test for T_k that determines if the actual level of clustering is significantly greater than would be expected if cases and controls were randomly assigned.

The significance of the test can be assessed using a *z*-statistic: $z = \frac{T_k - E(T_k)}{\sqrt{\mathrm{var}(T_k)}}$

E[*Tk*] is calculated as *pkn*, where *n* is the sample population size and *k* is the number of nearest neighbours being considered and *p*= (c/n)[(c-1)/(n-1)], where *c*= number of cases and *n*= total sample size.

When k>1, a combined P-value for all tests performed at one initial alpha level is provided. This is calculated through Bonferroni and Simes adjustments.

Bonferroni Pc= j[min(Pi)]
Simes Pc= min(n +1-i)Pi

In this case, *Pc* denotes the combined P-value for all tests, *Pi* the value for an individual test, *j* is the number of comparisons, and *i* is the sequential index for the individual test considered.

GLOBAL SECOND-ORDER SPATIAL ANALYSES

Global second-order spatial analyses are represented by the global Ripley's *K*-function (Ripley, 1976) and the global weighted *K*-function (Getis, 1984).

K function and weighted *K* function are among the most commonly used methods for identifying the distance at which clustering of point data occurs throughout the study area. These methods are called second-order analyses indicating that the focus is on the variance, the second-moment, of pairs of inter-event distances. In general, both the *K* function and the weighted *K* function consider all combinations of pairs of points, comparing the observed pattern of points at all distances with the one expected based on a random spatial distribution of points. Therefore, these methods can identify distances where clustering occurs and where departure from randomness is most pronounced. In particular, the *K* function is used to classify spatial patterns of point data (e.g. disease case locations, locations of farms, etc.) over the entire study area as being random, clustered, or uniformly dispersed, with significance determined using Monte Carlo simulations. Once determined the overall spatial pattern of points (e.g. locations of pig farms) using the *K*-function, the weighted *K*-function can be applied to evaluate the spatial patterns of values (e.g. disease incidence) within the spatial patterns of points previously detected. Practically, using the weighted *K*-function it is possible to investigate, for example, if cluster of *Salmonella* spp. incidences exists all over the study area, beyond the spatial pattern of pig farms detected using the non-weighted *K*-function.

For an isotropic process with an intensity of λ points per unit area, the *K* function at a distance *s* can be defined as *K*(s) using the following formula: $K(s) = \frac{1}{\lambda^2 A} \sum_{i \neq j} \sum I_s(d_{ij})$

where *A* is the area of interest, d_{ij} is the distance between the *i*th and the *j*th events in A, and $I_s(d_{ij})$ is an indicator function which equals one if $d_{ij} \leq d$ and zero otherwise. In the presence of spatial clustering, each event is likely to be in close proximity to other events of the same type and, for small s, *K*(s) will be large. When performing this analysis it is recommended to restrict the range of *s* to not greater than 0.5 times the length of the shorter side of a rectangular study area.

The *K* function is used to classify spatial patterns of point data over the entire study area as being random, clustered, or uniformly dispersed, with significance determined using Monte Carlo simulations. Under the assumption of complete spatial randomness *m* independent simulations of *n* events in the study region may be performed. For each simulated point pattern *K*(*s*) can be estimated and then the maximum and minimum of these functions used for the simulated patterns to define the upper and lower simulation envelope. If the estimated

K(*s*) lies above the upper envelope, this could show significant spatial clustering, if it lies below the lower limit, this is evidence of regularity in the arrangements of events (Bailey and Gatrell, 1995).

A modified version of the global Ripley's *K*-function has been also applied for investigating clustering around specific foci of disease (Kitron et al. 1992, Mannelli et al. 2003).

E.2 SPATIAL METHODS FOR INVESTIGATING LOCAL CLUSTERING

LOCAL MORAN'S I

The local Moran test (Anselin, 1995) detects local spatial autocorrelation in data aggregated by area. It is related to Moran's I for global spatial autocorrelation. In essence, the local Moran decomposes Moran's I into contributions for each area within a study region, defined as Local Indicators of Spatial Association (LISA). These indicators detect clusters of either similar or dissimilar disease frequency values around a given observation. The sum of LISA for all observations is proportional to the global Moran's I. There can be two interpretations of LISA statistics, as indicators of local spatial clusters (regions where adjacent areas have similar values) and as diagnostic for spatial outliers (areas distinct from their neighbours).

It is given by: $I_i = (X_i - \overline{X})\sum_{j=1}^{n} w_{ij}(X_j - \overline{X})$

where: X_i is the variable value at a particular location

X_j is the variable value at another location

$\overline{X}$ is the mean of the variable

w_{ij} is a spatial weights matrix

The Local Moran's I can only be interpreted within the context of the computed *Z* score.

The *Z* score represents the statistical significance of the computed I_i value. It indicates whether the apparent similarity (or dissimilarity) in values between the feature and its neighbours is greater than one would expect simply by chance. A high positive *Z* score for a feature indicates that the surrounding features have similar values (either high or low). A group of adjacent features having high *Z* scores indicates a cluster of similarly high or low values. A low negative *Z* score for a feature indicates the feature is surrounded by dissimilar values - that is, if a feature gets a negative *Z* score, it has a different value than its neighbours (a high value relative to a neighbourhood that has low values or a low value relative to a neighbourhood that has high values).

Getis-Ord $G_i(d)$ and $G_i^*(d)$

$G_i(d)$ and $G_i^*(d)$ were introduced by Getis and Ord (1992, 1996) for the study of local patterns in spatial data. These statistics indicate the extent to which a location is surrounded by a cluster of high or low values (e.g. incidence of a disease).

The difference between $G_i(d)$ and $G_i^*(d)$ is that the Gi(d) statistic investigates disease clustering only around an area (*i*) without including the value of *i* in the calculation. Therefore, $G_i(d)$ is a "*focused*" statistic and it is mostly used to study the spread or diffusion of diseases around potential disease foci. Conversely, the $G_i^*(d)$ statistic tests whether the area (*i*) together with surrounding areas, within a distance d, form a cluster of values higher (or lower) than average. $G_i^*(d)$ is then a more appropriate statistic to study local clustering of diseases, compared to $G_i(d)$.

The null hypothesis states that there is no association between the values of *X*, for example the disease incidence, at a site *i* and its neighbours, the *j*s, up to and including a distance *d*, measured from site *i* in all directions. The null hypothesis appropriate for the G_i statistic requires that x_i be excluded from the summation, while the null hypothesis appropriate for G_i^* statistic requires that the value at *i* itself be summed together with the *j* values within d of *i*. If a spatial autocorrelation exists, it will be exhibited by a spatial clustering of high or low values of disease incidence. When there is a cluster of high disease incidence values, the resulting $G_i(d)$ and $G_i^*(d)$ will be positive. Clustering of low values yields a negative $G_i(d)$ and $G_i^*(d)$.

The $G_i(d)$ statistic is calculated as:

$$G_i(d) = \frac{\sum_{j,j\neq i}^{N} w_{ij}(d)x_j - \bar{x}_i \sum_{j,j\neq i}^{N} w_{ij}(d)}{S_{(i)}\sqrt{\left[(N-1)\sum_{j,j\neq i}^{N} w_{ij}^{\,2}(d) - \left(\sum_{j,j\neq i}^{N} w_{ij}(d)\right)^2\right] \Big/ (N-2)}}$$

where: x_i is the observed value at location *i*, $\bar{x}_i = \frac{\sum_{j,j\neq i}^{N} x_j}{N-1}$, w_{ij} is a symmetric binary spatial weight matrix, and $S_{(i)} = \sqrt{\frac{\sum_{j,j\neq i}^{N} x_j^{\,2}}{N-1} - (\bar{x}_i)^2}$.

The G_i*(d) statistic is calculated as:

$$G_i^{*}(d) = \frac{\sum_{j}^{N} w_{ij}(d)x_j - \bar{x}\sum_{j}^{N} w_{ij}(d)}{S_{(i)}\sqrt{\left[N\sum_{j}^{N} w_{ij}^{\,2}(d) - \left(\sum_{j}^{N} w_{ij}(d)\right)^2\right] \Big/ (N-1)}}$$

where: $\bar{x} = \frac{\sum_{j}^{N} x_j}{N}$, and $S_{(i)} = \sqrt{\frac{\sum_{j}^{N} {x_j}^2}{N} - (\bar{x})^2}$.

$G_i(d)$ and $G_i^*(d)$ are assumed to be normally distributed and can be transformed into a *Z* statistic to test for significance. The local sum for a feature and its neighbours is compared proportionally to the sum of all features; when the local sum is very different to the expected local sum, and that difference is too large to be the result of random chance, a statistically significant *Z* score is the result. For statistically significant positive *Z* scores, the larger the *Z* score, the more intense the clustering of high values. For statistically significant negative *Z* scores, the smaller the *Z* score, the more intense the clustering of low values.

TEST FOR LOCAL SPATIAL CLUSTERING BASED ON SCANNING LOCAL RATES: SPATIAL SCAN STATISTIC

The spatial scan statistic is a test for investigating local spatial clustering based on scanning local rates. The general statistical theory behind the spatial scan statistic is described in detail by Kulldorff (1997). The spatial scan statistic uses two different probabilistic models, based on the Bernoulli and Poisson distributions respectively. With either model, the scan statistic adjusts for the uneven population density present in almost all populations, and the analysis is conditioned on the total number of cases observed. These statistics are performed with the software SATScan, which uses a circular window of different sizes to scan the study area. For each location and size of the scanning window, the alternative hypothesis is that there is an elevated rate (e.g. risk of disease infection) within the window as compared to outside. The likelihood function of the chosen probability model is maximised over all windows, identifying the window that constitutes the most likely cluster. This is the cluster that is least likely to have occurred by chance.

The likelihood ratio for this window is noted and constitutes the maximum likelihood ratio test statistic. Its distribution under the null-hypothesis and its corresponding p-value is obtained by repeating the same analytical exercise on a large random number of replications of the data set generated under the null hypothesis, in a Monte Carlo simulation.

THE JACQUEZ'S K NEAREST NEIGHBOURS TEST (JACQUEZ, 1996)

This test compares spatial distance to temporal distance and it is used for investigating space-time clustering of point data (e.g. disease case locations and dates). The *k* nearest neighbour statistic is the number of case pairs that are nearest neighbours in both space and time, and it is evaluated under the null hypothesis of independent space and time nearest neighbour relationships (i.e. that the probability of two events being nearest neighbours in space is independent of the probability of their being nearest neighbours in time). The statistical theory behind Jacquez's nearest neighbours test is described in detail by Jacquez (1996).

APPENDIX F: DEFINITIONS OF CERTAIN TERMS USED IN SECTION 5: SPATIAL EPIDEMIOLOGY AND APPENDIX E: SPATIAL STATISTICS

Spatial and space-time clusters: unusual concentrations of disease cases or other health-related events in time and space. In particular, **spatial clusters** refer to the aggregation of health events in a certain geographic location, while **space-time clusters** refer to the concentration of health events both in time and space. The number of cases in the cluster may or may not exceed the expected number. This is determined by spatial and space-time cluster analysis, a set of statistical methods used to analyse clusters.

Complete Spatial Randomness: a situation when an event is equally likely to occur at any location within a study area, regardless of the locations of other events.

Database Management System (DBMS): software that allows a computer to perform database functions of storing, retrieving, adding, deleting and modifying data. Relational database management systems (RDBMS) implement the relational model of tables and relationships.

Geographic Information System (GIS): system that allows the capturing, storing, manipulating, analysing and visualising of georeferenced data. It is built on five major components including hardware, software, data, people, and procedures.

Georeferenced data: data that can be referred to a location on the Earth's surface by means of geographic or projected coordinates.

Monte Carlo (MC) simulation: is one of the methods for assessing if an observed pattern of event locations appears to differ significantly from complete spatial randomness. Many tests for clustering use the MC simulation in order to determine the statistical significance of the cluster. In general, the test statistic value is calculated first based on the data observed and then, using the MC simulation, calculate the same statistic for a large number of datasets (e.g. 99, 999, etc.) simulated independently under the null hypothesis of interest. The proportion of test statistic values based on simulated data exceeding the value of the test statistic observed for the actual dataset provides a MC estimate of the p-value for a one-sided test. Using more simulations to estimate the distribution of the null-hypothesis (e.g. 999 versus 99), means that smaller and more stable p-values can be calculated.

P-value: the *p*-value for a hypothesis test is the probability of obtaining a result at least as extreme as the one that was actually observed, given that the null hypothesis is true. The p-value expresses the probability that the observed differences could be due to chance, and not due to the presence of the factor being evaluated. *P* is the probability of making a type I (α) error. The *p*-value for a test may also be defined as the smallest value of α for which the null hypothesis can be rejected. A *p*-value is therefore a measure of how much evidence is possessed against the null hypothesis. A **Type I (α) error** is the error committed when a true null hypothesis is rejected. A **Type II (β) error** is the error committed when a false null hypothesis is not rejected.

Scale: the ratio or fraction between the distance on a map, chart, or photograph and the corresponding distance on the surface of the Earth.

Spatial epidemiology: description and analysis of geographic variations in disease with respect to demographic, environmental, behavioural, socio-economic, genetic, and infectious risk factors (Elliott and Wartenberg, 2004).

Spatial weights matrix: describe how observations in a dataset are spatially related to each other. Spatial weights matrices are required in statistical methods to take into account spatial dependence of the studied objects. Different types of matrices can be calculated. A binary contiguity matrix describes whether or not spatial objects (e.g. farms, regions) are neighbours. The information stored in matrices can also be more complex, for example, parameters such as distance, or length of a common border (Haining, 2003).

ABBREVIATIONS

BEAST	Bayesian Ecological Analysis of Statistical Trends
CI	Confidence interval
CSR	Community Summary Report
DBMS	Database management systems
DSN	Dedicate Surveillance Network
EC	European Commission
ECDC	European Centre for Disease Prevention and Control
EFSA	European Food Safety Authority
EU	European Union
FAO	Food and Agriculture Organization
GEE	Generalised estimating equation
GIS	Geographic information systems
HACCP	Hazard Analysis and Critical Control Point
i.i.d.	Independent and identically distributed
LISA	Local Indicators of Spatial Association
MAUP	Modifiable area unit problem
MC simulation	Monte Carlo situation
MCMC	Monte Carlo Markov chain
MS(s)	Member State(s)
N	Number
NUTS	Nomenclature of territorial units for statistics
OIE	World Organisation for Animal Health
OLR	Ordinary logistic regression
pop	Population
pos	Positive flocks
PSU	Primary sampling unit
TESSy	The European Surveillance System
VTEC	Verotoxigenic *Escherichia* coli
WHO	World Health Organization
wt	Weight
ZCC	Zoonosis Collaboration Centre