Identification of *Salmonella* high risk pig farms in Belgium using semi-parametric quantile regression


*(1)*Center for Statistics, Hasselt University, Diepenbeek, Belgium

*(2)*Ghent University, Faculty of Veterinary Medicine, Merelbeke, Belgium

*(3)*Veterinary and Agrochemical Research Centre, Brussels, Belgium

*corresponding author: kaaje.bollaerts@uhasselt.be*

Abstract

Since consumption of pork contaminated with *Salmonella* is an important source of human *Salmonellosis* in Belgium, policy makers implement the identification of the 10% most *Salmonella* problematic pig herds. These herds are then encouraged to take control measures to reduce the *Salmonella* infection burden. To identify high risk herds, serological data were collected, reported as Sample to Positive ratios (SP-ratios).

Objectives of the current study are to identify the 10% highest risk herds and to investigate risk factors associated with high *Salmonella* prevalence. We propose to identify risk herds using semi-parametric quantile regression. The risk factor analysis is conducted using Generalized Linear Mixed Models. Finally, practical rules to identify risk farms are deduced.

Introduction

*Salmonellosis* is a disease affecting most livestock production worldwide. Nowadays national *Salmonella* surveillance programmes are implemented and encouraged within the European Union. The Belgian Federal Agency for the Safety of Food Chain (FASFC) started with a national *Salmonella* surveillance programme in pig herds. The surveillance programme started in January 2005 and consists of the collection of serological data to classify pig herds in *Salmonella* high risk herds. Then, these high risk herds are encouraged to take part in the surveillance programme. In this programme, high risk herds are (financially) supported to implement control measures.

Because of financial and practical constraints, the Belgian government decided that only 10% of the Belgian pig herds can participate in the surveillance programme. The identification of the 10% high risk herds is currently based on a mean SP-ratio which is calculated for each herd during consecutive sampling rounds. However, SP-ratios are heavily (positively) skewed. Therefore, least squares based measures and methods are not appropriate and the sample mean does not provide a good summary of the data.

As an alternative, we propose to select the 10% high risk farms based on the number of animals in that herd for which very high SP-ratios are observed. Very high SP-ratios are then defined as SP-ratios above a specific upper quantile. An additional complication is the presence of confounding factors such as seasonal effects with higher expected SP-ratios throughout the summer months and animal age with the older the animal, the higher the expected SP-ratio. To deal with all these issues, conditional quantile curves of animal SP-ratios are estimated while accounting for confounding seasonal and animal age effects. However, it is not easy to model seasonal effects and most parametric models do not provide a good fit. Therefore seasonal effects are modelled in a flexible way using P-splines (Eilers & Marx, 1996). P-splines belong to the family of non-parametric models that provide a data driven and smooth fit to the data.

Finally, since high risk farms are encouraged to implement control measures, a *Salmonella* risk factor analysis is conducted. To this end, the serological data are linked with data from a survey on biosecurity in Belgian pig herds conducted in 2005 (Ribbens et al, 2006). The risk factor analysis is conducted using Generalized Linear Mixed Models (Molenberghs & Verbeke, 2005). The use of mixed models is motivated by the presence of clustering in the data due to the sub-sampling of animals within herds.
Material and methods

Serological data from the *Salmonella* surveillance programme (FASFC) are used. Within this programme, blood samples are taken and analysed using an indirect ELISA. The results are reported as SP-ratios. Every herd is monitored 3 to 4 times a year. Each time, 10 or 12 samples from pigs of different weight categories are collected (within the frame of the eradication programme of Aujeszky disease). For every blood sample taken, the herd identification number, animal estimated weight (<40kg, 40-59kg, 60-80kg and >80kg) and sampling time are recorded. The serological data are linked to data from a survey on biosecurity (Ribbens et al., 2005) by herd identification number (Sanitel). The objective of the survey was to describe the degree of measures taken in Belgian pig herds to minimize the risk of introducing infectious agents into herds (external biosecurity measures) and of spreading an infectious agent within herds once it has been introduced (internal biosecurity measures). The combined dataset contains data on 314 pig herds. In total, 13649 serological observations sampled from January 2005 to July 2006 are retained for analysis.

The 10% high risk farms are identified based on the number of animals in that herd for which very high SP-ratios are observed (also called risk animals). It is natural to define very high SP-ratios by means of quantiles. To account for confounding seasonal and animal age effects, quantile curves of animal SP-ratios are estimated as a function of time and animal weight (proxy for animal age). In particular, the following semi-parametric model is used to estimate the $\theta \times 100\%$ quantile of animal SP-ratio for animal $i$ in function of sampling time and animal weight

$$\tilde{S}_{\theta,i} = h(\text{time})_i + I(\text{weight})_i$$

with $h$ being a smooth P-splines function and $I$ an indicator matrix. As such, seasonal effects are modelled in a very flexible way whereas the effect of age is assumed to be additive. Risk animals are defined as animals for which the observed SP-ratio is higher than the corresponding $\theta \times 100\%$ quantile or

$$R_{\theta,i} = \begin{cases} 1 & \text{if } S_{\theta,i} > \tilde{S}_{\theta,i} \\ 0 & \text{otherwise} \end{cases}$$

Then, high risk herds are defined as herds 'k' having a large proportion of risk animals or

$$P_{\theta,k} = \frac{1}{n_k} \sum_i R_{\theta,i}$$

However, proportions do not take the total number of observations $n_k$ into account nor the correlated nature of the data due to the subsampling of animals within herds. As an alternative, beta-binomial $p$-values can be calculated. Under the null hypothesis that high levels of SP-ratios are equally likely in all herds, the number of risk animals $Y_{\theta,k}$ is beta-binomially distributed and the $p$-value corresponding to the alternative hypothesis that high levels of SP-ratios are more likely in herd 'k' compared to the other herds equals

$$p_{\theta,k} = P\{Y_{\theta,k} \geq y_{\theta,k} | Y_{\theta,k} \sim \text{BB} (n_k,1-\theta,\rho)\}$$

with $\rho$ being the intra-herd correlation coefficient under the null hypothesis.

A risk factor analysis using Generalized Linear Mixed Models is conducted with the number of risk animals in the herd being the response variable of interest. Random intercepts are included in the model to account for clustering due to subsampling of animals within herds. In total, 20 potential risk factors coming from a study on biosecurity (Ribbens et al., 2006) are investigated using forward selection.
Results

A conservative choice is made to define risk animals by choosing $\theta = 0.90$. These 90% quantile curves of animal SP-ratios in function of sampling time and animal weight are graphically represented in Figure 1. By means of comparison, the conditional median functions are given as well. Clearly, strong seasonal effects are observed with higher expected values of SP-ratios during the summer months. Furthermore, weight (age) effects are observed as well. Except for the weight categories 40-59kg and 60-79kg, for which the 90% quantile curves are identical, higher SP-ratios are observed for higher weight categories. The proportion risk animals as well as the corresponding beta-binomial p-values are displayed in Figure 2. In this figure, herds are ordered following increasing proportion of risk animals. Clearly, large differences in proportion risk animals between herds exist with a large number of herds having a proportion of zero whereas for one herd the proportion equals one. Similarly, large differences between herds can be observed when looking at the beta-binomial p-values. Selection of exactly 10% most problematic pig herds corresponds to selecting herds with a proportion risk animals $> 0.30$ or with beta-binomial p-values $< 0.0005$.

Based on the estimated quantile curves, practical identification rules can be deduced. For each month by weight category, the average estimated quantile SP-ratios are calculated. By means of example, the average SP-ratios for the months January and February are displayed in Table 1. This can be used to identify risk animals. Then, high risk herds are defined as herds having a proportion risk animals $> 0.30$.

Finally, from the risk factor analysis it can be concluded that the number of risk animals increases with farm size (OR 95%CI [1:1.004]). Furthermore, nose contact between pigs from different pens (OR 95%CI [1.14:3.45]) is an important risk factor whereas systematic insect control (OR 95%CI [0.28:0.74]) and regularly cleaning the stables (OR 95%CI [0.13:0.89]) are found to be remedial measures.

<table>
<thead>
<tr>
<th>Weight Category</th>
<th>January</th>
<th>February</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40kg</td>
<td>0.97</td>
<td>1.01</td>
</tr>
<tr>
<td>40-79kg</td>
<td>1.19</td>
<td>1.23</td>
</tr>
<tr>
<td>&gt;80kg</td>
<td>1.49</td>
<td>1.52</td>
</tr>
</tbody>
</table>

Tab 1. Average estimated quantile SP-ratios ($\theta = 0.90$) by month and weight category.

Discussion and Conclusion

Currently, the identification of the 10% high risk herds as recommended by the Belgian government is based on a mean SP-ratio. However, SP-ratios are extremely skew implying that the mean does not provide a good summary of the data. Furthermore, the current approach does not account for confounding seasonal and animal age effects. Therefore, we propose to identify high risk herds using estimated quantile curves of SP-ratios conditional on sampling time and
animal age. This method does not suffer from the aforementioned shortcomings. However, the choice of upper quantiles $\theta$ is to a certain extent arbitrary. Therefore, a sensitivity analysis is conducted (results not shown here), from which it can be concluded that $\theta = 0.80$, $\theta = 0.90$ and $\theta = 0.95$ yield similar results. From the risk factor analysis, it is concluded that nose contact is an important risk factor and regularly cleaning and insect control are remedial measures. Finally, in the current study, the longitudinal aspect of the data has not been investigated as such. This way, increasing or decreasing trends of SP-ratios could be investigated. However, analyzing longitudinal data poses important new challenges for quantile regression.

References


Acknowledgement

This study has been carried out with the financial support of the Belgian Federal Public Service of Health, Food Chain Safety, and environment research programme (R-04:003-Metzoon) ‘Development of a methodology for quantitative assessment of zoonotic risks in Belgium applied to ‘Salmonella in pork’ model’.