Identification of Salmonella high risk pig-herds in Belgium by using semiparametric quantile regression

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

Kaatje Bollaerts[1], Marc Aerts[1], Stefaan Ribbens[2],

Yves Van der Stede[3], Ides Boone[3], Koen Mintiens[3]

[1] Hasselt University, Center for Statistics, Diepenbeek, Belgium

[2] Ghent University, Faculty of Veterinary Medicine, Merelbeke, Belgium

[3] Veterinary and Agrochemical Research Center, Brussels, Belgium

Author notes: All correspondence concerning this paper should be sent to Kaatje Bollaerts, Universiteit Hasselt, Center for Statistics, Agoralaan 1 gebouw D, B-3590 Diepenbeek (Belgium). [E-mail: kaatje.bollaerts@uhasselt.be]

Abstract

Consumption of pork contaminated with *Salmonella* is an important source of human Salmonellosis worldwide. To control and prevent Salmonellosis, Belgian pig herds with high *Salmonella* infection burden are encouraged to take part in a control programme supporting the implementation of control measures. The Belgian government decided that only the 10% pig herds with the highest *Salmonella* infection burden (shortly high risk herds) can participate. To identify these herds, serological data reported as SP-ratios are collected. However, SP-ratios have an extremely skewed distribution and are heavily subject to confounding seasonal and animal age effects. Therefore, we propose to identify the 10% high risk herds using semi-parametric quantile regression with P-splines. In particular, quantile curves of animal SP-ratios are estimated as a function of sampling time and animal age. Then, pigs are classified into low and high risk animals with high risk animals having an SP-ratio larger then the corresponding estimated upper quantile θ. Finally, for each herd, the number of high risk animals is calculated as well as the beta-binomial p-value reflecting the hypothesis that the *Salmonella* infection burden is higher in that herd compared to the other herds. The 10% pig herds with the lowest p-values are then identified as high risk herds. In addition, since high risk herds are supported to implement control measures, a risk factor analysis is conducted using binomial Generalized Linear and Generalized Linear Mixed Models to investigate factors associated with decreased or increased *Salmonella* infection burden. Finally, since the choice of a specific upper quantile is to a certain extent arbitrary, a sensitivity analysis is conducted comparing different choices of upper quantiles θ.

Keywords: *Salmonella*, pigs herds, risk factors, semi-parametric quantile regression, Generalized Linear Mixed Models

1 Introduction

Worldwide Salmonellosis, the illness from *Salmonella* infection, is the most frequently occuring zoonoses, which is an infectious disease directly or indirectly transmitted between animals and humans. Thereby Salmonellosis is a major concern in most industrialized countries having a significant economic impact (Mead et al, 1999). Global estimations vary between 14 and 120 cases of Salmonellosis per 100000 people (Thorns, 2000). Salmonellosis is characterized by fever, stomach cramps, and diarrhea and is mostly self-limiting. However, deaths due to *Salmonella* infections do occur, especially within the susceptible population of newborns, young children, pregnant women, elderly and immunocompromised persons. The majority of the cases of Salmonellosis is due to *Salmonella* Enteritidis and *Salmonella* Typhimurium infections, which comprised almost 80% of the total number of *Salmonella* infections in Belgium in 2005 (NRSS, 2005). Of the reported *Salmonella* infections 26 % of these are likely to be due to transmission of the pathogen via pork.

The recent establishment of the European Centre for Disease Prevention and Control (ECDC) has put prevention and control of zoonoses as one of its main priorities (de Jong & Ekdahl, 2006). Salmonellosis is one of the eight zoonoses listed in the European Union (EU) Zoonoses Monitoring Directive (2003/99/EG), for which continuous monitoring is mandatory. In the EU Regulation on the control of salmonella and other zoonotic agents (2160/2003/EG) it is stated that proper and effective measures are to be taken to detect and control salmonella and other zoonotic agents at all relevant stages of the food production chain, particularly at the level of primary production (that is, at herd level). In line with this regulation, the Belgian Federal Agency for the Safety of Food Chain (FASFC) started with a national *Salmonella* surveillance programme in professional pig herds in January 2005. Objective of this programme is to identify pig herds with high *Salmonella* infection burden, which are then (financially) encouraged to take part in a control programme supporting the implementation of control measures to reduce the infection burden. Because of financial and practical constraints, the Belgian government decided recently that only the 10% pig herds with the highest *Salmonella* infection burden can participate (to appear in Belgian Royal act on Salmonella Surveillance programme in pigs, june 2007). In the sequel, these 10% pig herds are shortly called high risk herds for reasons of convenience.

In order to be able to quantify *Salmonella* infection burden in pig herds, the Belgian FASFC collected serological data. In particular, blood samples of breeding and fattening

pigs are taken which are tested for *Salmonella*-specific sera with an indirect ELISA (Herdchek Swine, Idexx Laboratories). The tests are conducted following the test manufacturer's guidelines. This means that the optical density of the sample (OD_{sample}) is assessed as well as the mean optical density of the positive ($\overline{OD}_{pos.}$) and negative controls ($\overline{OD}_{neg.}$) to ensure a valid test run was performed. The obtained optical densities are normalized as Sample to Positive ratios (SP-ratios) being calculated as $(OD_{sample} - \overline{OD}_{neg.})/(\overline{OD}_{pos.} - \overline{OD}_{neg.})$. SP-ratios normally cover a range between 0 and 4, but even higher values can be observed. Furthermore, SP-ratios are extremely positively skewed. This follows since for negative samples and controls always low OD-values are observed whereas for positive samples and controls a wide range of OD-values can be observed going from moderately high to very high.

A first main objective of the current study is the selection of 10% high risk herds using the serological data collected by FASFC. However, it is not obvious how to use serological data at animal level in order to identify the 10% high risk herds. Currently, identification is based on a average SP-ratio which is calculated for each herd during consecutive sampling rounds. But SP-ratios have an extremely skewed distribution implying that the sample mean is not a good estimate of central tendency. Furthermore, SP-ratios are heavily subject to confounding effects. Seasonal effects is such a well known confounder with higher expected SP-ratios throughout the summer months. Animal age is a known confounder as well with the older the animal, the higher the expected SP-ratio. To deal with all these issues, we propose an alternative method to identify the 10% high risk herds. We propose to base identification on the number of high risk animals in a herd, taking into account the total number of pigs sampled in that herd as well as the intra-herd correlation. High risk animals are then defined as pigs having a very high SP-ratio being a SP-ratio larger than a specific upper quantile θ. To correct for confounding seasonal and animal age effects, quantile curves of animal SP-ratios are estimated as a function of sampling time and animal weight (proxy for animal age). In particular, seasonal effects are modeled in a flexible way using P-splines whereas age (weight) effects are modeled in an additive way.

A second main objective of the current study is to investigate factors associated with an increased or decreased *Salmonella* infection burden at farm level. To this end, the serological data are linked with data from a survey on biosecurity in Belgian pig farms conducted in 2005 (Ribbens et al, 2005). A risk factor analysis is conducted using binomial Generalized Linear Models (GLMs) and Generalized Linear Mixed Models (GLMMs) with the use of mixed models being motivated by the presence of clustering in the data due to the sub-sampling

of pigs within herds. The results of these analysis can then be used to motivate control measures to be taken by high risk herds in order to reduce the *Salmonella* infection burden. Finally, since the choice of a specific upper quantile θ used to identify high risk animals is to a certain extent arbitrary, a sensitivity analysis is conducted comparing different choices of upper quantiles θ.

The remainder of this paper is organised as follows: in Section 2, the data, both the serological data as well as the data from the survey on biosecurity, are introduced; in Section 3, the identification of *Salmonella* high risk animals is discussed starting with methodological subsections on ordinary least squares regression with P-splines and quantile regression with P-splines; in Section 4, the identification of *Salmonella* high risk herds is discussed; in Section 5, the risk factor analyses are described. Finally, some concluding remarks and suggestions for further research are given in Section 6.

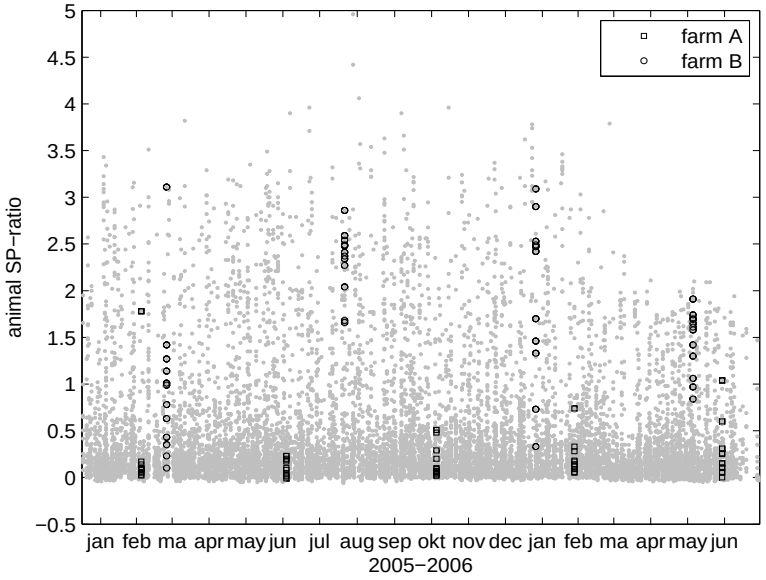


Figure 1: Serological data: animal SP-ratio by sampling time.

2 Data

Serological Data

A serological *Salmonella* surveillance programme, organized by the FASFC, has been in place since January 2005 in Belgium. In particular, *Salmonella*-specific antibodies are determined by an indirect ELISA in blood samples of breeding and fattening pigs. These samples are collected within the frame of the eradication programme of pseudorabies virus (Aujeszky disease). Veterinarians are obliged to collect 10 or 12 samples (depending on the size of the herd) on each professional meat producing pig herd (having at least 31 pigs) from pigs of different weight categories every 3 to 4 months per year. The results of the tests are reported as SP-ratios. For every blood sample taken, the herd identification number, animal estimated weight (<40kg, 40-59kg, 60-80kg and >80kg) and sampling time are recorded. For the current analysis, only data from January 2005 to July 2006 were considered.

Survey on Biosecurity

A postal survey on biosecurity in Belgian pig herds (both professional and non-professional) has been conducted in 2005 by Ribbens et al (2006). 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). To this end, written questionnaires were sent to a sample of 609 pig herds randomly chosen from the Belgian herd Identification & Registration (I&R) database (Sanitel-Pigs, 2005). The sample was stratified by province using proportional allocation. In total, 436 questionnaires were properly filled in and returned, yielding a response rate as high as 71.6 %. It was evaluated whether statistical differences were found between responders and non-responders based on external information available in the I&R database. The results indicated that the mean number of pigs in the non-responding herds was significantly lower than in the responding herds (Ribbens et al, 2005). Of the filled in questionnaires, 332 were originating from professional meat producing pig herds, which were considered for further analysis.

Combined Dataset

The serological and the biosecurity datasets were linked by the herd identification number. Only data on pig herds that were included in both datasets ($n = 314$) were retained for the current analysis. This resulted in 14301 observations to be retained from the serological dataset. Of these observations, 652 observations ($< 5\%$) were not complete and were discarded. Hence, 13649 observations remain, yielding an average of 43.47 observations made per pig herd from January 2005 to July 2006. The serological data used in the current study are graphically represented by means of a scatter plot of animal SP-ratios by sampling time (Figure 1). By means of illustration, the observations are marked for two herds, herd A and B. Clearly, large differences in SP-ratios between the two herds can be observed, with, in general, much larger SP-ratios for herd B compared to herd A.

3 *Salmonella* high risk animals

The number of high risk animals within a herd will be used to select the 10% high risk herds. High risk animals are defined as pigs having a SP-ratio above a specific upper quantile. To correct for confounding seasonal and animal age effects, quantile curves of animal SP-ratios are estimated as a function of sampling time and animal weight (proxy for animal age). In this section, methodological discussions on ordinary least squares regression with P-splines and quantile regression with P-splines are given first. More elaborated discussions can be found in Eilers and Marx (1996) and Bollaerts et al (2006). Finally, quantile regression with P-splines is applied to estimate quantile curves of animal SP-ratios as a function of sampling

time and animal weight.

3.1 Ordinary least squares regression using P-splines

In a target article, Eilers and Marx (1996) introduced P-splines regression within an ordinary least squares framework. This is essentially ordinary least squares regression with an excessive number of equally spaced B-splines (De Boor, 1978; Dierckx, 1993) and an additional discrete penalty to correct for overfitting. In simple regression using B-splines with predictor variable $\mathbf{x}$, a basis of r overlapping B-splines is constructed, which is such that

$$\forall x : \sum_{j=1}^{r} B_j(x, q) = 1, \tag{3.1}$$

with $B_j(x, q)$ denoting a B-spline of degree q with left most knot j. In Figure 2a, one can see an example of a basis of B-splines of the third degree, which is the most commonly used degree in B-splines regression. Then, the B-splines of a B-splines basis act as predictors in spline regression. With, for m observations (x_i, y_i),

$$\widehat{y}_{(\boldsymbol{\alpha})_i} = \sum_{j=1}^{r} \alpha_j B_j(x_i, q), \text{ i=1,}\ldots\text{ m,} \tag{3.2}$$

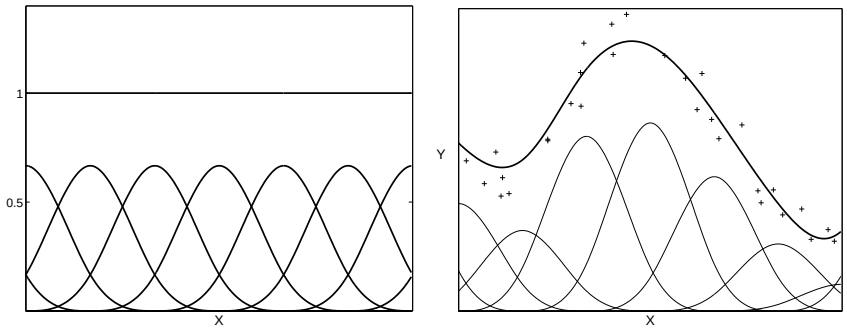
and with α_j being the coefficient of the corresponding B-spline. The vector $\boldsymbol{\alpha}$ is commonly estimated using the Least Squares loss function or L_2-norm:

$$S_2 = \sum_{i=1}^{m} (y_i - \widehat{y}_{(\boldsymbol{\alpha})_i})^2, \tag{3.3}$$

with the conditional mean function being the minimand. An example of B-splines regression using splines of degree 3 is given in Figure 2b. However, a major problem in B-splines regression is the choice of the optimal number of B-splines. An insufficient number of B-splines leads to underfitting, whereas too many B-splines leads to overfitting. To regularize smoothness, Eilers and Marx (1996) proposed to use an excessive number of equally spaced B-splines with, in order to correct for overfitting, a smoothness penalty based on differences of the coefficients of adjacent B-splines. They called this approach P-splines regression. The corresponding loss function based on the L_2-norm equals

$$S_2 = \sum_{i=1}^{m} (y_i - \widehat{y}_{(\boldsymbol{\alpha})_i})^2 + \lambda \sum_{j=d+1}^{r} (\triangle^d \alpha_j)^2, \tag{3.4}$$

with $\triangle^d \alpha_j$ being the d^{th} order differences, that is $\triangle^d \alpha_j = \triangle^1(\triangle^{d-1}\alpha_j)$ with $\triangle^1 \alpha_j = \alpha_j - \alpha_{j-1}$ and with λ being a smoothness parameter. Mostly, a penalty on second order differences is



(a) B-splines basis (b) B-splines regression

Figure 2: Spline regression with B-splines of third degree.

used. Smoothness can be controlled for by means of λ. When λ is small, the smoothness penalty weakly influences the fit and the fit is mainly governed by closeness to the observed data. On the contrary, when λ is large, the smoothness penalty highly influences the fit and the fit might be too coarse. In general, if $\lambda \to \inf$, then, for a regression with a smoothness penalty on d^{th} order differences, the fitted function will approach a polynomial of degree $d-1$. In order to optimally select λ, Aikaike's Information criterion (AIC), K-fold cross-validation or generalized cross-validation can be used (Eilers & Marx, 1996).

3.2 Quantile regression using P-splines

Koenker and Bassett (1978) introduced quantile regression as an alternative to ordinary least squares regression. Recently, an extended monograph on quantile regression appeared (Koenker, 2005). In quantile regression, $\boldsymbol{\alpha}$ is estimated by minimizing the asymmetric least absolute deviations loss function or asymmetric L_1-norm:

$$S_1 = \theta \sum_{y_i \geq \widehat{y}_{(\boldsymbol{\alpha})_i}} |y_i - \widehat{y}_{(\boldsymbol{\alpha})_i}| + (1-\theta) \sum_{y_i < \widehat{y}_{(\boldsymbol{\alpha})_i}} |y_i - \widehat{y}_{(\boldsymbol{\alpha})_i}|. \tag{3.5}$$

Clearly, positive values are weighted with a factor θ, whereas (strictly) negative values are weighted with a factor $1-\theta$, yielding the $\theta \times 100\%$ conditional quantile curve. An alternative formulation of the L_1-norm is

$$S_1 = \sum_{i=1}^{m} \rho_\theta(y_i - \widehat{y}_{(\boldsymbol{\alpha})_i}), \tag{3.6}$$

where ρ_θ is called the 'check function', which is defined as

$$\rho_\theta(\tau) = \begin{cases} \theta\tau & \text{if } \tau \geq 0 \\ (\theta-1)\tau & \text{otherwise.} \end{cases}$$

Then, in order to fit conditional quantile functions using P-splines, the asymmetric least absolute deviations loss function given in (3.6) is extended as

$$S_1 = \sum_{i=1}^{m} \rho_\theta(y_i - \widehat{y}_{(\boldsymbol{\alpha})_i}) + \lambda \sum_{j=d+1}^{r} |\triangle^d \alpha_j|, \tag{3.7}$$

where $\widehat{y}_{(\boldsymbol{\alpha})_i}$ is defined as in (3.2), where $\triangle^d \alpha_j$ are the d^{th} order differences and where λ is a smoothness parameter, which can be chosen optimally by, e.g. K-fold cross-validation (Bollaerts et al, 2006).

Solving an L_1 regression problem relies on reformulating the corresponding loss function as a linear programming problem, which can be solved using a different types of algorithms (Vanderbei, 2001). For the current application, we adopt the approach proposed by Portnoy and Koenker (1997). They proposed, as an alternative to simplex based methods, interior point optimization in combination with statistical preprocessing for L_1-type of problems. The latter approach has the advantage over simplex based methods of being computationally less demanding, especially in large data sets. Matlab code to solve L_1-type of problems using Portnoy and Koenker's approach as well as quantile regression sofware in R can be found on Koenker's home-page at the University of Illinois (http://www.econ.uiuc.edu/∼roger).

3.3 Quantile curves of SP-ratios

In order to identify high risk animals, $\theta \times 100\%$ quantile curves of animal SP-ratios are estimated while accounting for confounding seasonal and animal age effects. In particular, the following semi-parametric model is used for each pig i of herd k measured at time j

$$\widehat{SP}_{\theta.ijk} = h(time)_{ijk} + I(weight)_{ijk}, \quad \text{i} = 1,\ldots, n_{jk}; \text{ j} = 1,\ldots, r_k; \text{ k} = 1,\ldots, 314, \tag{3.8}$$

with $h(.)$ being a smooth P-splines function and I being an indicator matrix. As such, seasonal effects are modelled in a very flexible way whereas the effect of age (using animal

weight as proxy variable) is assumed to be additive. High risk animals are defined as pigs for which the observed SP-ratio is higher than the corresponding $\theta \times 100\%$ quantile or

$$Z_{ijk} = \begin{cases} 1, & SP_{ijk} > \widehat{SP}_{\theta.ijk} \\ 0, & SP_{ijk} \leq \widehat{SP}_{\theta.ijk}. \end{cases} \tag{3.9}$$

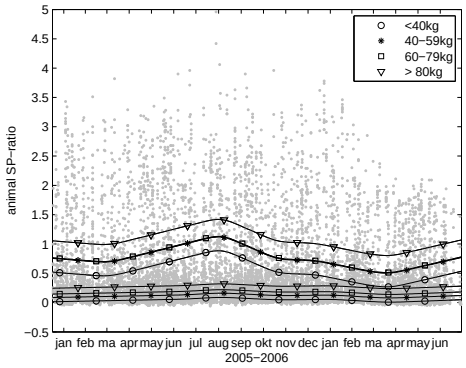
A conservative choice is made by selecting upper quantiles for θ. In particular, the conditional quantile functions $\theta = 0.85$, $\theta = 0.90$ and $\theta = 0.95$ are investigated. All these quantile functions are estimated by means of P-splines regression. Since the rationale behind P-splines regression is to use an excessive number of B-splines in combination with a smoothness penalty which 'automatically' corrects for overfitting, we opt to use a basis of as much as 25 B-splines. The degree of B-splines is chosen to be $d = 3$ being the most common used degree d because of its good trade-off between model flexibility, model smoothness and complexity. The optimal value for the smoothness parameter λ is determined using K-fold cross-validation, with each subsample containing data from one specific herd ($K = 314$). The candidate values λ are chosen from an approximately geometric grid. The variability of the estimated quantile functions $f_\theta(\hat{\boldsymbol{\alpha}}, X)$ is assessed using residual bootstrap. To maximally reflect the structure of the data, residuals are resampled within each combination of herd and sampling time. Let $\hat{F}_{n_{jk}}(u)$ denote the empirical distribution of the residuals corresponding to the measurements taken at time j for herd k with $j = 1, ..., r_k$ and $k = 1, ..., 314$. In particular, the residuals are defined as $u_{ijk} = y_{ijk} - \hat{y}_{ijk}$ with $\hat{y}_{ijk}$ being derived from the model given in (3.8) for $\theta = 0.50$. Of course, other values θ could be chosen as well to generate bootstrap samples that reflect the structure of the data. However, the median $\theta = 0.50$ has the best robustness properties, that is the highest break-down point, amongst all other quantiles θ. Then, for each combination jk, a subsample $\mathbf{y}^*_{jk} = [y^*_{1jk}; ...; y_{n_{jk}jk^*}]$ is generated by drawing residuals $u^*_1, ..., u^*_{n_{jk}}$ with replacement from $\hat{F}_{n_{jk}}(u)$ and setting $y^*_{ijk} = \hat{y}_{ijk} + u^*_i$. The different subsamples $\mathbf{y}^*_{jk}$ are combined in $\mathbf{y}^*$ to generate one bootstrap sample. Then, for each bootstrap sample $\mathbf{y}^*$ the model f_θ is fitted or $\widehat{\mathbf{y}}^* = f^*_\theta(\hat{\boldsymbol{\alpha}}, X)$. This process is repeated $B = 1000$ times, yielding $B = 1000$ different bootstrap estimates of $\boldsymbol{\alpha}$. In order to estimate the $100(1-2\alpha)\%$ pointwise confidence interval for $f_\theta(\hat{\boldsymbol{\alpha}}, X)$, percentile intervals are calculated conditional on X. The latter are defined as $[f^*_\theta(\hat{\boldsymbol{\alpha}}, X)_{[(B+1)\alpha]}; f^*_\theta(\hat{\boldsymbol{\alpha}}, X)_{[(B+1)(1-\alpha)]}]$ with $f^*_\theta(\hat{\boldsymbol{\alpha}}, X)_{[(B+1)\alpha]}$ being the $[(B+1)\alpha]^{th}$ order statistic of $f^*_\theta(\hat{\boldsymbol{\alpha}}, X)$.

The results of the K-fold cross-validation are shown in Table 1, with optimal values $\lambda_{0.50} =$

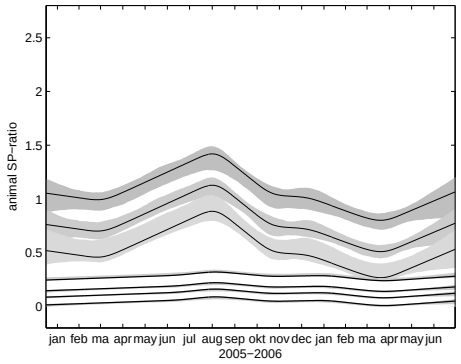
Table 1 *Optimal values for smoothness parameter λ based on K-fold Cross-Validation.*

λ	0.1	1	5	10	50	100	500	1000	10000
$\theta = 0.50$	7.4282	7.4212	7.4196	7.4199	7,4169	7.4155	**7.4118**	7.4161	7.4134
$\theta = 0.85$	8.4026	8.3877	9.3238	8.2904	8.2697	8.2537	**8.2390**	8.2644	8.2517
$\theta = 0.90$	7.1557	7.0725	7.0488	7.0192	6.9863	6.9970	**6.9717**	6.9931	7.0024
$\theta = 0.95$	4.4660	4.4418	4.4322	4.4167	4.3963	**4.3757**	4.3832	4.390	4.3870

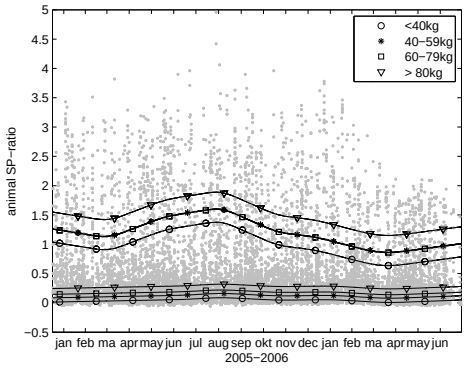
$\lambda_{0.85} = \lambda_{0.90} = 500$ and $\lambda_{0.95} = 100$. A graphical representation of the fitted conditional quantile curves $\theta = 0.85$, $\theta = 0.90$ and $\theta = 0.95$ are given in Figure 3. The left most figures display the conditional quantile functions together with a scatter plot of the data. The right most figures display the 95% residual bootstrap confidence intervals. For reasons of comparison, in each figure, the conditional median functions are displayed as well. As can be seen, strong seasonal effects are observed for the upper quantiles $\theta = 0.85$, $\theta = 0.90$ and $\theta = 0.95$ with higher expected values of SP-ratios during the summer months. Furthermore, weight effects can be observed as well with, in general, the higher animal weight, the higher the conditional quantile curve. For $\theta = 0.50$, the 95% bootstrap confidence intervals of the quantile curves for the four different weight categories are nicely separated. However, this is not the case for the upper quantile functions with the quantile curves of the weight categories $40 - 59$kg and $60 - 79$kg being virtually identical. As a consequence, these weight categories are merged for further analysis of the upper quantiles. Furthermore, the 95% bootstrap confidence intervals of the quantile curves for the different weight categories are overlapping. For $\theta = 0.85$, the overlap is small whereas the overlap is much larger for $\theta = 0.90$ and $\theta = 0.95$. Finally, as indicated by the width of the confidence intervals, there is most variability in estimating the quantile curves $\theta = 0.90$ and $\theta = 0.95$, less variability in estimating the quantile curves $\theta = 0.85$ and little variability in estimating the median curves. This is as expected since it generally holds that the more extreme the quantile curve, the larger the variability in estimating the curve.



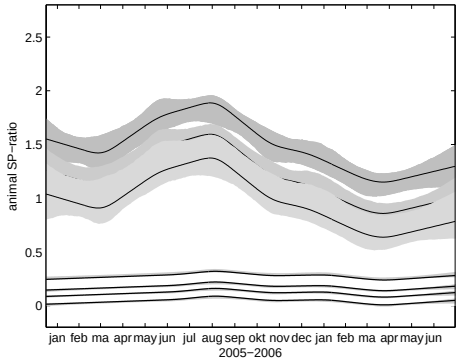
(a) $\theta = 0.85$



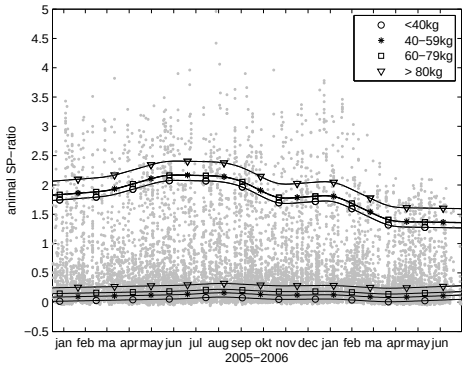
(b) $\theta = 0.85$



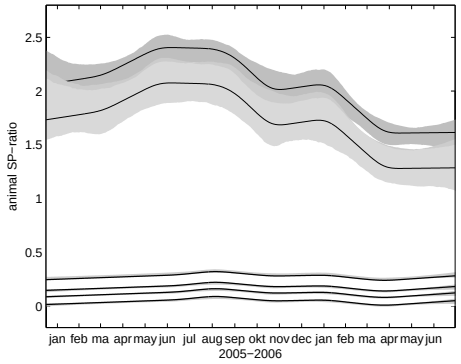
(c) $\theta = 0.90$



(d) $\theta = 0.90$



(e) $\theta = 0.95$



(f) $\theta = 0.95$

Figure 3: Estimated quantile curves $\theta = 0.85$, $\theta = 0.90$ and $\theta = 0.95$ = of animal SP-ratios conditional on sampling time and animal weight + 95% confidence intervals. The results for $\theta = 0.50$ are given by means of comparison.

4 *Salmonella* high risk herds

To identify the 10% high risk herds, the proportion of high risk animals P_k is calculated for each herd k or

$$P_k = \frac{1}{n_k} \sum_{j=1}^{r_k} \sum_{i=1}^{n_{jk}} Z_{ijk} \tag{4.1}$$

with $n_k = \sum_{j=1}^{r_k} n_{jk}$ and with Z_{ijk} being defined as in (3.9). Then, using the beta-binomial distribution as a natural choice for correlated binary data, it follows that, under the null hypothesis that the *Salmonella* infection burden is equal in all pig herds, the number of high risk animals Y_k in herd k is beta-binomially distributed as

$$Y_k \sim BB(n_k, 1-\theta, \rho)$$

with n_k being the total number of sampled pigs in herd k, with $1-\theta$ the probability of being a high risk animal and with ρ being the intra-herd correlation. The p-value corresponding to the alternative hypothesis that *Salmonella* infection burden is higher in herd k compared to the other herds is then equal to

$$p_k = P\{Y_k \geq y_k | Y_k \sim BB(n_k, 1-\theta, \rho)\}.$$

These p-values can then be used to select the 10% high risk herds, which are then the herds with the lowest p-values or

$$R_k = \begin{cases} 1, & p_k \leq \delta \\ 0, & p_k > \delta \end{cases} \tag{4.2}$$

with δ being the $0.10 \times 100\%$ quantile of p-values.

We will focus discussion on $\theta = 0.90$ meaning that high risk animals are defined as pigs for which an SP-ratio above the 90% quantile is observed. Then, under the null hypothesis that the *Salmonella* infection burden is equal in all pig herds, the number of high risk animals Y_k in herd k is beta-binomially distributed as

$$Y_k \sim BB(n_k, 0.10, \rho).$$

An additional complication is that it is not clear how the intra-herd correlation ρ can be estimated under the null hypothesis. The latter is done by fitting a beta-binomial distribution with fixed $\pi = 0.10$ to data from herds for which the null hypothesis is most likely to hold. These are herds having a proportion of high risk animals close to the expected proportion

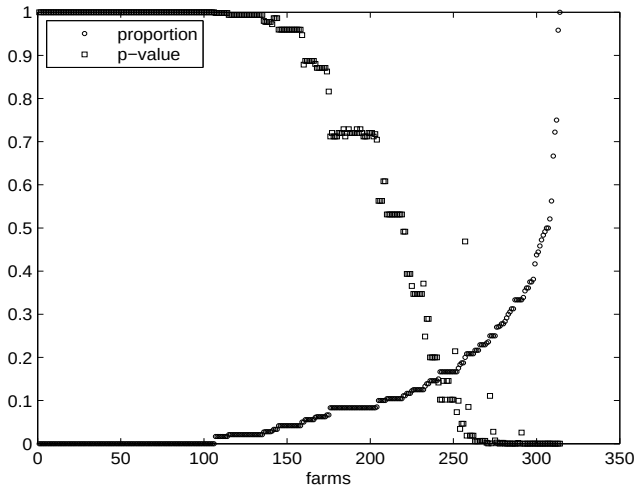


Figure 4: Proportion risk animals per herd and corresponding beta-binomial p-values.

under the null hypothesis, $E(Y_k/n_k) = E(P_k) = 0.10$. We decided to use the data from the 84 herds with $0.05 \leq P_k \leq 0.15$. Fitting a beta-binomial with fixed $\pi = 0.10$ to the selected data yields $\widehat{\rho} = 0.00037$ (so essentially uncorrelated).

Then, for each herd, the beta-binomial p-values are calculated with fixed $\pi = 0.10$ and $\widehat{\rho} = 0.00037$, which are graphically displayed in Figure 4 together with the observed proportions of high risk animals. In this figure, herds are ordered along the X-axis following increasing proportion of high risk animals with the 'circles' representing the proportions and the 'squares' the p-values. Clearly, large differences in proportion high risk animals between herds exist with a large number of herds having a proportion high risk animals of zero whereas for one herd, the proportion is equal to one. However, proportions are misleading since they do not take the number of observations n_k (that is, the total number of sampled pigs) into account. Therefore, it is better to look at the beta-binomial p-values. Again, large differences between herds can be observed with the percentage of herds for which $p_k < \delta = 0.001$ being equal to 11.8%. Selection of exactly 10% herds with the lowest p-values corresponds to selecting herds with $p_k < \delta = 0.0005$.

In the above mentioned analysis, $\theta = 0.90$ is used to identify high risk animals. Of course, other upper conditional quantile curves could be considered as well. To investigate the effect of the choice of θ on the selection of the 10% high risk herds, a small sensitivity analysis is conducted comparing the results for $\theta = 0.85$, $\theta = 0.90$ and $\theta = 0.95$. For each of

these values of θ, the corresponding beta-binomial p-values are calculated with the intra-herd correlation as estimated before. Selection of the 10% pig herds having the lowest p-values for $\theta = 0.85$, $\theta = 0.90$ and $\theta = 0.95$ corresponds to selecting herds having p-values smaller than $\delta_{.85} = 0.00007$, $\delta_{.90} = 0.0005$ and $\delta_{.95} = 0.0193$, respectively. It is investigated whether the same herds are identified as high risk herds using the different choices of upper quantiles θ. By means of comparison, average SP-ratios are also used to identify high risk herds. The results are summarized in Table 2. In total, 85.67% (resp. 5.41%) of the herds are selected as low risk herds (resp. high risk herds) by all four methods, yielding an overall agreement of 91.08%. If we restrict attention to the selection criteria based on quantile regression, we see that the results are similar for $\theta = 85$ and $\theta = 90$ whereas the results for $\theta = 0.95$ are more different. Out of the 22 herds for which there is no overall agreement for the three methods under consideration, 7 (31.82%) herds are identified as high risk herds using both $\theta = 85$ and $\theta = 90$ but not using $\theta = 95$ and for 9 (40.91%) herds the inverse is true. This indicates that, as one might expect, the choice of $\theta = 0.95$ is too extreme, in some way also indicated by the large threshold value $\delta_{.95}$. If we compare the identified high risk herds using average SP-ratios with the ones identified using quantile regression with $\theta = 0.85$ and $\theta = 0.90$, we also see differences. Out of the 17 herds for which there is no overall agreement for the three methods under consideration, 6 (35.29%) herds are identified as high risk herds using both $\theta = 85$ and $\theta = 90$ but not using average SP-ratios and for 5 (29.41%) herds the inverse is true.

5 Risk factor analysis

So far, we discussed the selection of the 10% high risk herds. However, especially since high risk herds are encouraged to implement control measures, it is important to know which factors affect the probability of being a high risk herd. Therefore, a risk factor analysis is conducted using logistic regression with

$$\begin{aligned} R_k &\sim \text{Bernouilli}(\pi_k) \\ \log\left(\frac{\pi_k}{1-\pi_k}\right) &= X_k\boldsymbol{\beta} \end{aligned} \tag{5.1}$$

where R_k is as defined in (4.2), X is a $(n \times p)$-dimensional matrix of known risk factors and $\boldsymbol{\beta}$ is a $(p \times 1)$-dimensional vector of regression parameters. In total, 20 potential risk factors

Table 2 *Herds identified as high risk herds using average SP-ratio and quantile regression with* $\theta = 85$, $\theta = 80$ *and* $\theta = 95$ *(0 indicates that the herd is not identified, 1 otherwise).*

	$\overline{SP}$	$\theta = 85$	$\theta = 90$	$\theta = 95$	frequency	percentage
	0	0	0	0	269	85.67
	0	0	0	1	6	1.91
	0	0	1	0	1	0.32
	0	1	0	0	1	0.32
	0	1	1	0	2	0.64
	0	1	1	1	4	1.27
	1	0	0	0	2	0.64
	1	0	0	1	3	0.96
	1	0	1	0	2	0.64
	1	1	0	0	1	0.32
	1	1	0	1	1	0.32
	1	1	1	0	5	1.59
	1	1	1	1	17	5.41
tot.	31	31	31	31	314	100

X are investigated. A list of these risk factors together with a brief description is given in the Appendix. For more detailed information, the reader is referred to Ribbens et al (2006).

Notice the analogy of the current application with case-control studies. In case-control studies, the marginal distribution of the response variable of interest Y is fixed by design as in the current application. Here, the marginal probability of being a high risk herd $P(Y = 1)$ is a priori determined to be 0.10. As such, it is impossible to estimate the conditional probability $P(Y = y|X = x)$, although being of primary interest. However, the inverse conditional probability $P(X = x|Y = y)$ can be estimated as well as the odds ratio. The latter follows since the odds ratio is symmetric and invariant to the selection of X or Y as response variable. This also means that, as opposed to other binary response models, a logistic regression model is adequate for case-control studies since its exponentiated parameters $\exp(\beta)$ are to be interpreted as odds ratios.

From the sensitivity analysis described in the previous section, it was concluded that $\theta = 0.85$ and $\theta = 0.90$ are equally good choices to define high risk animals. For both choices of θ, a risk factor analysis is conducted yielding the same single significant risk factor, namely nose contact. For $\theta = 0.85$ ($\theta = 0.90$), the odds of being a risk herd is estimated to be 5.39 (3.40) times larger when pigs of different pens can have nose contact as compared to when they can not have nose contact. The corresponding 95% confidence intervals are $[1.26; 23.14]$ and $[1.009; 11.54]$ for $\theta = 0.85$ and $\theta = 0.90$, respectively.

The identification of the 10% herds with the highest *Salmonella* infection burden is not necessarily compatible with an optimal risk factor analysis. Indeed, there might be herds just above and below the artificial 10% level, with essentially the same values for the risk factors, and therefore obscuring the identification of discriminating risk factors. As an alternative to the approach taken above, one can also conduct a risk factor analysis with the number of high risk animals in a herd being the response variable of interest. However, pigs are clustered within herds, which should be accounted for in the analysis. Different types of models that take clustering into account exist and can be grouped in conditional, marginal and cluster-specific models (Verbeke & Molenberghs, 2000; Molenberghs & Verbeke, 2005). The latter approach is adopted. In particular, logistic mixed models are used with herd-specific random intercepts. This model has an interpretation conditional on the random effects and its model

formulation equals

$$Y_k|u_k \sim \text{Binomial}(n_k, \pi)$$
$$\log\left(\frac{\pi}{1-\pi}\right) = X\boldsymbol{\beta} + u_k \tag{5.2}$$

with Y_k being the number of high risk herds in herd k and with n_k being the total number of sampled pigs. Again, X is a $(n \times p)$-dimensional matrix of known risk factors and $\boldsymbol{\beta}$ is a $(p \times 1)$-dimensional vector of regression parameters, also called fixed effects common to all subjects. The random intercepts u_k reflect the between-herds heterogeneity in the population with respect to Y. In addition, it is assumed that the random intercepts are sampled from a normal distribution with mean 0 and variance σ^2. The same 20 potential risk factors X are investigated (Ribbens et al, 2006).

For both $\theta = 0.85$ and $\theta = 0.90$, we start building the model with inclusion of a measurement of herd size since herd size is considered to be an important *Salmonella* risk factor with, in general, the larger herd size, the higher the infection pressure implying higher *Salmonella* infection burden. However, measuring herd size is not straightforward. Merely counting all pigs, including piglets before weaning age, will overestimate herd size. Not counting piglets will underestimate herd size and piglets herds will even have a nonsensical herd size of zero. Therefore, two variables are added to the null model containing only intercepts in order to properly reflect herd size, namely (a) the total number of pigs including piglets (N) and (b) a binary variable indicating the presence or absence of piglets (PIGLETS = 1 if piglets are present and 0 otherwise). Then, forward selection is used to investigate the other potential risk factors. All main effects as well as second-order interactions are considered for inclusion. Aikaike's Information Criterion (AIC) is used to decide upon model selection.

The final models are summarized in Table 2. The effect of herd size (N and PIGLETS) is as expected. First, the predicted number of high risk animals increases with the total number of pigs N. Second, the overestimation of herd size due to counting piglets is properly corrected with the predicted number of high risk animals being lower if piglets are present. For $\theta = 0.85$, the model contains three binary variables in addition to herd size. These are systematic insect control (INSECT), nose contact between pigs of different pens (NOSE) and systematic rodent control (RODENT). For $\theta = 0.90$, the additional binary variables are systematic insect control (INSECT), nose contact between pigs of different pens (NOSE) and regularly cleaning the stables (CLEANING). For both choices of θ, systematic insect control is a remedial measure whereas nose contact between pigs of different pens is found to be a

Table 2 *Summary of final logistic mixed models for* $\theta = 0.85$ *and* $\theta = 0.90$.

	$\theta = 0.85$			$\theta = 0.90$		
Odds Ratio	est.	95% CI	p-value	est.	95% CI	p-value
N	1.0002	[1;1.0004]	0.027	1.0002	[1;1.0004]	0.026
PIGLETS	0.66	[0.42;1.020]	0.061	0.58	[0.35;0.96]	0.033
INSECT	0.39	[0.25;0.62]	<0.0001	0.46	[0.28;0.74]	0.0018
NOSE	1.73	[1.067;2.81]	0.026	1.98	[1.14;3.45]	0.016
RODENT	3.12	[1.39;7.014]	0.006			
CLEANING				0.3382	[0.13;0.89]	0.028
$\sigma^2(u_k)$	2.34	[1.82;2.87]	0.005	2.79	[2.10;3.48]	0.005

risk factor. For $\theta = 0.85$, systematic rodent control is also found to be a risk factor. This sounds contra-intuitive since rodent control is expected to decrease the *Salmonella* infection burden. However, it might be that measures to control rodents are only taken whenever problems with rodents occur. Furthermore, regularly cleaning the stables is found to be a remedial measure for $\theta = 0.90$. Finally, note that the estimates of the variance of the random intercepts $\sigma^2(u_k)$ are significant as well indicating substantial between-herd heterogeneity.

6 Concluding remarks

In the current study, an alternative method is proposed to identify pig herds with high *Salmonella* infection burden based on serological data reported as SP-ratios. To deal with confounding effects and issues of skewness, semi-parametric quantile regression with P-splines is used. In particular, quantile curves of animal SP-ratios are estimated as a function of sampling time and animal age. Then, pigs are classified into low and high risk animals with high risk animals having an SP-ratio larger then the corresponding estimated upper quantile θ. Finally, for each herd, the number of high risk animals is calculated as well as the beta-binomial p-value reflecting the hypothesis that the *Salmonella* infection burden is higher in that herd compared to the other herds with the lower the p-value the higher the infection burden. As such, herds with high *Salmonella* infection burden can be identified within a proper inferential framework.

In addition, since these herds are encouraged to implement control measures, a risk factor analysis is conducted as well indicating that nose contact between pigs of different pens is the most persistent *Salmonella* risk factor. Although the obtained response rate of the biosecurity survey (Ribbens et al, 2005) of 71.6% is reasonably high (Thrusfield, 2005), potential bias might result due to non-response. Analysis based on external data indicate that the non-response herds are most likely to be small herds (Ribbens et al, 2005). Although we did not find any indication that the relationship between potential risk factors and the *Salmonella* infection burden would be different in small herds compared to large herds (no significant interactions found between herd size and other risk factors in the risk factor analyses described in Section 5), the possibility of bias can not be ruled out. A cautious practice would be to generalize the findings to large herds only and avoid generalization to the whole population of pig herds. Also from a pragmatic point of view, this would be sensible since the FASFC *Salmonella* surveillance programme is restricted to professional herds only having at least 31 pigs.

The approach we presented models seasonal effects on quantile curves of animal SP-ratios using P-splines (Eilers & Marx, 1996, Bollaerts et al, 2006). This smoothing method estimates quantile curves in a flexible way, allowing the estimation of sudden changes and varying seasonal trends. Of course, other quantile smoothing methods could be considered as well, most of which fit within the locally polynomial or splines framework. Locally polynomial quantile regression is explored and applied by, amongst others, Chaudhuri (1991) and Welsh (1996).

Koenker et al (1994) introduce quantile smoothing splines, where the smoothing penalty is a function of the first or higher order derivative of the fitted function. In the P-splines approach we adopted, the smoothness penalty is based upon finite differences of the coefficients of adjacent B-splines. The latter approach has the advantage of being computationally less demanding and easier to implement.

Finally, the longitudinal aspect of the data has not been investigated as such in the current study. It would be interesting as well to investigate the dynamics of animal SP-ratios within herds over time. This way, increasing or decreasing trends of the *Salmonella* infection burden within herds could be detected and as such, the impact of the implementation of control measures could be evaluated. However, analyzing longitudinal data poses important new challenges for quantile regression and proper statistical methods need to be developed further (Koenker, 2005).

Acknowledgements

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 the 'Salmonella in pork' model'. Partners in this programme are Hasselt University, Liège University, Ghent University, the Institute for Agricultural and Fisheries Research (Melle), the Veterinary and Agrochemical Research Center (Brussels), and the Institute for Public Health (Brussels). The work of Marc Aerts is partly funded by the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy).

References

Bollaerts, K., Eilers, P.H.C. & Aerts, M. (2006). Quantile regression with monotonicity constraints using P-splines and the L1-norm. *Statistical Modelling, 6,* 189-207.

De Boor, C. (1978). *A practical guide to splines.* Berlin: Springer.

Chaudhuri, P. (1991). Global nonparametric estimation of conditional quantile functions and their derivatives. *Journal of multivariate statistics, 39,* 246-269.

de Jong, B. & Ekdahl, K. (2006). The comparative burden of Salmonellosis in the European Union member states assocated and candidate countries). *BMC Public Health, 6(4)*

Dierckx, P. (1993). *Curve and surface fitting with splines.* Oxford: Clarendon.

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. *Official Journal of the European Union 2003; L 325/31: 12.12.2003. (http://europa.eu.int/eur-lex/pri/en/oj/dat/2003/l_325/l_32520031212en00310040.pdf)*

Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statistical Science, 11(2)*,89-121.

Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica, 46,* 33-50.

Koenker, R., Ng, P. & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika, 81,* 676-680.

Koenker, R. (2005). *Quantile Regression.* New York: Cambridge University Press.

Mead, P. S., Slutsker, L., Dietz, V., McCaig, L. F., Bresee, J. S., Shapiro, C., Griffin, P.

M. & Tauxe, R. V. (1999). Food-related illness and death in the United States. Emerging Infectious Diseases. *Emerging Infectious Diseases, 5 (5),* 607-625*,* submitted.

Molenberghs, G. & Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

NRRS (2005). National Reference Centre for Salmonella and Shigella: Annual report. (`http://www.iph.fgov.be/bacterio/iframes/rapports/2004/Salm_2004_NL_cover.pdf`)

Portnoy, S. & Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: computability of squared-error vs. absolute-error estimators (with discussion). *Statistical Science, 12,* 279-296.

Regulation (EC) No 2160/2003 of the European Parliament and of the Council of 17 November 2003 on the control of salmonella and other specified food-borne zoonotic agents. *Official Journal of the European Union 2003; L 325/1: 12.12.2003. (http://europa.eu.int/eur-lex/pri/en/oj/dat/2003/l_325/l_32520031212en00010015.pdf)*

Ribbens, S., Dewulf, J., Maes, D., Koenen, F., Mintiens, K., Desadeleer, L. & Kruif, A. (2006). A survey on biosecurity in Belgian pig herds. *Preventive Veterinary Medicine,* submitted.

Sanitel-Pigs (2005). Identification and Registration data of registered Belgium pig herds: Federal Agency of the Safety of the Food Chain (FASFC).

Thorns, C. J. (2000). Bacterial food-borne zoonoses. *Review scientific et technique (International Office of Epizootics), 19 (5),* 226-239.

Thrusfield, M. (2005). *Veterinary Epidemiology.* London: Blackwell Science.

Vanderbei, R.J. (2001). *Linear Programming: Foundations and Extensions, Second edition.* Dordrecht: Kluwer, Academic Publishers.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data* New York: Springer.

Welsh, A.H. (1996). Robust estimation of smooth regression and spread functions and their derivatives. *Statistica Sinica, 6,* 347-366.

Appendix

Description of risk factors investigated in the study.

1. **N** = total number of pigs in the herd, including boars, sows, gilts, finishers and piglets.
2. **piglets*** = piglets are kept in the herd.
3. **cleaning stables*** = finishing stables are cleaned more than three times a year.
4. **insect control*** = systematic measures are taken to control insects.
5. **nose contact*** = finishing pigs from different pens can have nose contact.
6. **disinfection baths*** = use of disinfection baths at the entrance of the location.
7. **hygiene entry*** = facility where visitors put on herd clothing.
8. **quarantine*** = use of a quarantine period for animals entering the herd.
9. **number of transports** = number of trucks entering the herd per month
10. **acid*** = use of acidified feed and/or drinking water.
11. **disinfect transport*** = livestock trucks have been disinfected before they may enter the herd
12. **dogs and cat enter*** = pet animals can enter the stables.
13. **moving pigs***= finishing pigs are moved during finishing period
14. **empty period*** = pens stay empty for a certain time period between different rounds.
15. **purchase** = introduction of piglets to the herd
16. **birds enter*** = birds do have access to the stables.
17. **rodent control*** = systematic measures are taken to control rodents.
18. **exclusively pigs*** = besides pigs, no other animal species are commercially kept on the herd.
19. **pigs go outside*** = pigs are allowed to go outside/ 'free range' husbandry.
20. **separated material*** = use of separated material between different pens.
21. **grid floor** = % grid floor in stables

* binary variable with 0 = no and 1 = yes