REPORT
"Development and application of text-mining tools for the use of pathology data in epidemiological surveillance of animal diseases"
- Retrospective time series analysis-

Céline Faverjon[1], Susanne Küker [1], Lenz Furrer [2], John Berezowski [1], Flavie Vial [1], Horst Posthaus [3], Fabio Rinaldi [2], Gertraud Schüppach [1]

[1] Veterinary Public Health Institute, University of Bern, CH-3097, Liebefeld, [2] Institute of Computational Linguistics, University of Zurich, CH-8050 Zürich, [3] Institute of Animal Pathology, University of Bern, CH-3001 Bern

EXECUTIVE SUMMARY

This report was prepared for the project "Development and application of text-mining tools for the use of pathology data in epidemiological surveillance of animal diseases"(Project:1.14.03). It focuses on the last part of the study: performing a retrospective analysis of the syndrome time-series and identification of possible risk factors.

First, we identified the most frequent syndromes for each species (cattle and swine) and age class. They were the same in both species: gastrointestinal system, serous membranes, and respiratory system. However, some syndromes were much more prevalent in certain species and/or age classes. Second, we performed a visual examination of each syndrome time series. We identified abnormalities, some of which had not been noticed by pathologists. When we looked at these changes in more detail, we were able to determine that some of them were linked to non-standardized reporting procedures, but most of them remained unexplained. The unexplained abnormalities were increased numbers of cases, which could have been due to increases in disease in the population. However we did not have any information about epidemics in the populations that were served by the diagnostic laboratory, and therefore we considered these abnormalities to be unexplained. Third, we modelled the time series using generalized linear models. Clear differences in temporal patterns were observed between age classes. Our models also showed that there were different temporal patterns between syndromes affecting the same age classes. Fourth, the time series models were used to implement temporal and spatio-temporal aberration detection algorithms on two syndrome time series. We were able to identify abnormal clusters of cases comparing nine different algorithms. However, we were unable to choose the best algorithms due to the lack of information about the true epidemic time periods in the data.

We believe that text mining the free text final reports of pathologists produces data and information that could be used for multiple purposes and by different actors in Switzerland. Pathologists could use reports from this data to identify and investigate in detail pathological lesions (i.e., lesions association, risk factors). Diagnostic

pathology laboratory managers will find information that is of value for managing the laboratory and ensuring they provide consistently high quality service to their clients. Pathology and clinical medicine instructors will find information about seasonal and other diseases patterns that can be used to improve the education of veterinary undergraduates and pathology residents. These data could be also used at the national level to better understand farmers' practices regarding carcass submission, and the general health status of the Swiss farmed animal population. Most importantly these data could contribute to national animal health surveillance by detecting changes in endemic diseases, identifying emerging diseases and demonstrating freedom from disease. There are many well-known biases in veterinary diagnostic laboratory data. It is our opinion from this study that the benefits outweigh the biases, but appropriate use of the data will always be essential to avoid making errors. It is likely that data from pathology reports will not make a stand alone surveillance system, but it will provide data that can contribute valuable if not essential information to a larger integrated national surveillance system. However, even if these data are promising, they are not yet ready to use. In particular more data would be needed to make further analysis (e.g., adding data from other diagnostic laboratories to improve data representativeness, providing more information on true abnormal events occurring in the field to compare aberration detection algorithms). Indeed, as a reminder, we only used the data collected at ITPA. Developing more standardized procedures for data collection could also help improving data quality. In conclusion the results of this study provide evidence that there is benefit in pursuing the further development of text mining tools with the ultimate goal of implementing a real-time application that can create data from pathology records, automatically analyze it, develop user specific reports and communicate them to the appropriate stakeholders.

INTRODUCTION

This study was performed to: 1) describe, using descriptive statistics, post-mortem data collected by the institute of animal pathology (ITPA) between 2000 and 2012; 2) develop and evaluate an automated text-mining and syndrome-classifying tool; 3) perform a retrospective analysis of the time-series (TS) of syndromes and identify possible risk factors.

A complete description of post-mortem data collected by ITPA is available in (Küker et al. 2016 (in prep)). As a reminder, a total of 6031 pigs and 2911 cattle were necropsied at the ITPA during the years 2000 and 2011 which were included in this study.

A complete description and evaluation of the text mining tool developed during this project is available in (Lenz et al. 2015), in the code provided and in the manual written by Susanne Küker. As a reminder, we built a set of 13 syndrome categories primarily based on the topography of organ systems: “gastrointestinal” (GASTRO), “respiratory” (RESPI), “heart” (HEART), “lymphatic” (LYMPH), “nervous” (NEURO), “musculoskeletal” (MUSKO), “urinary” (URI), “reproduction” (REPRO) and “serous membranes” (SEROSA). For the monitoring of congenital malformations and neoplasia, two specific categories were created: “malformations” (MISS) and “neoplasia” (NEO). For diagnoses referring to a specific pathologic pattern that could not be classified into one of the categories above, the category “other” (OTHER) was created. OTHER includes for example intoxications, skin lesions and septicemia. Some reports didn’t refer to any specific pathologic pattern, had no clear pathologic findings or had an unclear cause of death, and these were grouped into a category named “non-classified” (NON-CLASS). All the animals included in this study were classified in at least one of these categories as defined above.

This report focuses on the last part of the study: performing a retrospective analysis of the syndrome time-series and identification of possible risk factors. However, we also report some of the most important results obtained by S. Küker on the general data descriptive analysis.

MATERIAL & METHODS

1. Data description

We briefly reported in this document the main results obtained by S. Küker during her doctoral thesis on general data description. The data quality was assessed looking at the number of missing and unusable entries. Sex, age, breed and reason of death of animals were also assessed.

As diseases may affect young animals and adults indifferent ways, we defined 5 age classes for cattle:1) neonatal calves (<1month), 2) calves (1-6month), 3) young (6-12month), 4) young adults (1-2years) and 5) adults (>2years)) and 3 age classes for pigs: 1) piglets (0-5weeks), 2) weaner pigs (6-12weeks), and 3) adult pigs (>12weeks)). Adult pigs included fatteners but also breeding sows and boars. The main syndrome categories were identified for each species and age class. Associations between categories were also considered.

The descriptive analysis previously performed by S. Küker showed that 27% of entries for animal age were missing or invalid in the dataset (see Table 1). In order to reduce the number of missing age values, we implemented an age correction for reports on abortion and/or stillbirth. 580 reports were classified as "Abort/Totgeburt" according to the ITPA variable named "productionstatus". Most of these reports (567 in total: 399 cattle, 169 pigs) did not contain information regarding the age of the necropsied animals. This can be explained by the fact that, in the case of an abortion, only the fetus and/or placenta are sent to ITPA. We choose to consider that these reports came from a pathological problem in sexual mature animals. These reports were thus assigned to the age class of adult cattle (>2 years) or fattening/adult pigs (> 12 weeks) depending on the species considered.

Each syndrome category was aggregated into a weekly time series using the ISO 8601 date and time standard for week definition (ISO 2004). A visual description of the main temporal patterns was performed and abnormal patterns found into the dataset were identified and explained. A full spatial description of the syndromes is also available in the report prepared by Margarida Arede, VPHI intern (see appendix).

2. Data modelling

The total number of animals necropsied and the selected syndrome categories were modelled. Data from 2000 to 2005 were used to train our models, and data from 2006 to 2011 were used to validate them.

Our modelling approach was to use generalized linear regression models (GLMs) that were appropriate for count data (Poisson and negative binomial (NB) regression). Models included different types of seasonality (sinusoidal functions with 1, 2, or 3 periods/year and season or month) as factorial variables. To account for differences between years, we also calculated the average counts over 53 consecutive weeks (*histmean*). To ensure that an ongoing outbreak would not influence the estimate, we used a 10-week guard band for the calculation of *histmean*. A list of tested variables is available in the appendix (table 1).

Different generalized linear regression models were evaluated on the training data from 2000 to 2005 using the Akaike information criterion (AIC) (Bozdogan 1987). The best models were then evaluated and compared using the auto-correlation (ACF) and partial auto-correlation (PACF) functions of the residuals and the root-mean-squared error (RMSE). ACF is the linear dependence of a variable on itself at two points in time and PACF is the auto-correlation between two points in time after removing any linear dependence between them (Box, Jenkins, and Reinsel 2008). ACF and PACF were used to find repeating patterns (e.g., seasons) in a dataset. RMSE is a measure of the difference between the values predicted by a model and the values actually observed from the environment that is being modeled (Chai and Draxler 2014). This criterion was calculated for the differences between the observations and the predicted values within both the calibration period ($RMSE_c$) from 2000 to 2005 and the validation period ($RMSE_v$) from 2006 to 2011. In either case, the lower the criterion, the better the predictive performance of the model.

Models were implemented in R x64 version 3.0.2 (R Development Core Team 2008). Dynamic regression was performed with the functions *glm* (package {stats}), and *glm.nb* (package {MASS} (Ripley et al. 2015)). The expected numbers of counts at time *t* were estimated with the *predict* functions of the respective packages.

3. Aberration detection algorithms

Our objective was to perform a retrospective analysis of the syndromes in order to compare and optimize algorithms detection based on abnormal patterns already present in the dataset. However, our data did not contain information on true epidemic periods. To compare and optimize algorithms detection performance it would thus have been necessary to simulate outbreaks. However, outbreak simulation was not included in the project proposal and only makes sense when working with the final data streams, which are selected to meet specific predefined surveillance objectives. In agreement with BLV, we thus choose to only use aberration detection algorithms as an illustration of the expected outputs if more information were available.

We tested 7 temporal detection algorithms and 2 spatiotemporal algorithms in order to see if they were able to identify abnormal patterns in the data. We used as an example the RESPI syndrome category in cattle (one of the most prevalent categories), and the NEURO syndrome category in cattle (one of the syndrome categories with very few cases).

The algorithms tested for temporal analysis where EARS (including C1, C2 and C3 algorithms), Farrington, CUSUM, Shewhart and EWMA. These algorithms were selected because they are the most frequently used algorithms for event detection in syndromic surveillance and because they are able to detect different types of abnormal events: CUSUM and EWMA are especially adapted for detecting small but repeated differences between observed and expected values (Mandl et al. 2004)(Hunter 1986)(Dórea et al. 2013). The Shewhart chart, its derivatives (i.e., EARS) and regression techniques (i.e., Farrington (Farrington et al. 1996)) are less well suitable when outbreaks do not occur instantaneously and are transient. They are more effective for detecting large outbreaks with a sudden increasing of cases (Unkel et al. 2012). For EARS and Farrington, functions available in the R package “surveillance” were used with their default optimized values. For CUSUM, EWMA and Shewhart, we used the functions available in the R-package “vetsyn” with their default optimized values. All algorithms were applied on GLM residuals if the time series were not stationary except for Farrington algorithm which is able to deal with data with trend and season.

The spatiotemporal analyses were implemented by Margarida Arede, a summer intern at the VPHI. The necropsies were aggregated by month due to the fact that weekly analysis would be very time-consuming to run. According to the descriptive spatial analysis, the vast majority of animals necropsied at ITPA came from the west of Switzerland and Bern surroundings. In order to reduce biases in the analysis, we decided to remove all Swiss Cantons to the East of Zurich including: Zurich, Schaffhausen, Thurgau, Appenzell Ausserrhoden, Appenzell Innerrhoden, Saint Gallen, Zug, Schwyz, Glarus, Nidwalden, Obwalden, Uri, Ticino and Graubünden. Local space-time clusters were then estimated with scan statistics. This technique is characterized by a circular window, which is moved in space for each possible geographic location and time point. A space-time permutation and a space-time Poisson model were applied using SatScan™ (version 9.4.1, Martin Kulldorff and Information Management Services Inc.; http://www.satscan.org/). The most likely space-time clusters for higher risk were identified by likelihood ratio testing, and their significance estimated through a Monte Carlo simulation consisting of 999 random replications of the dataset. A significance level of 5% was used.

4. Data selection

In the project proposal, we planned to only describe two syndrome time series. However, the text mining tool was able to provide much more information and 13 syndrome categories were created. This represents a very large amount of data and analyses as these syndromes should be separated into different species classes and also into separate age classes for monitoring purposes. There was a total of 65 (13*5) different time series for cattle and 39 (13*3) for pigs.

With the limited time available and in order to prioritize our work, we decided during a meeting with BLV (July 20th, 2016. VPHI – Annette Nigsch, Margarida Arede, Céline Faverjon) to focus our analysis only on the following syndromes in both species: GI, RESPI, SEROSA, REPRO, NEURO and NONCLASS. We also wanted to look at pigs with both RESPI and REPRO syndromes because they could be linked with PRRS. However only 7 pigs in the entire dataset had these 2 categories associated. The choice was based on the main syndrome categories identified during the descriptive analysis and on the main diseases of interest for early detection by the BLV. The list of the datasets used for each stage of the project (i.e., data description (D), data modelling (M) and implementation of aberration detection algorithms (A)) is available in tables 1 and 2.

Table 1: Cattle data used for each stage of the project (i.e., data description (D), data modelling (M) and implementation of aberration detection algorithms (A)). Datasets are defined according to the age class and the syndrome category considered. Grey cells indicate that a specific dataset has been used for a specific project stage. Numbers represent the page numbers in which readers can find the results of a specific project stage applied to a specific dataset.

	Total cattle necropsied			**Neonatal calves** <1mth			**Calves** 1-6mth			**Young** 6-12mth			**young adults** 1-2yrs			**Adults** >2yrs		
	D	**M**	**A**	**D**	**M**	**A**	**D**	**M**	**A**	**D**	**M**	**A**	**D**	**M**	**A**	**D**	**M**	**A**
All syndromes	8-16, 32-33	11		9-10	11		9-10	11		9-10	11		9-10	11		9-10	11	
GI	8-16, 32-33	16		10	16		10	16		10	16		10	16		10	16	
RESPI	8-16, 32-33	17	22	10	17		10	17		10	17		10	17		10	17	
SEROSA	8-16, 32-33	17		10			10			10			10			10		
OTHER	8-16, 32-33			10			10			10			10			10		
REPRO	8-16, 32-33	17		10			10			10			10			10		
MUSKO	8-16, 32-33			10			10			10			10			10		
NEURO	8-16, 32-33	18	24	10			10			10			10			10		
HEARTH	8-16, 32-33			10			10			10			10			10		
URI	8-16, 32-33			10			10			10			10			10		
LYMPH	8-16, 32-33			10			10			10			10			10		
NEO	8-16, 32-33			10			10			10			10			10		
MISS	8-16, 32-33			10			10			10			10			10		
NONCLASS	8-16, 32-33	19		10			10			10			10			10		

Table 2: Pig data used for each stage of the project (i.e., data description (D), data modelling (M) and implementation of aberration detection algorithms (A)). Datasets are defined according to the age class and the syndrome category considered. Grey cells indicate that a specific dataset has been used for a specific project stage. Numbers represent the page number in which reader can find the results of a specific age stage applied to a specific dataset.

	Total pig necropsied			**Piglets** 0-5weeks			**Weaners** 6-12weeks			**Adults** >12weeks		
	D	**M**	**A**	**D**	**M**	**A**	**D**	**M**	**A**	**D**	**M**	**A**
All syndromes	8-16, 34-35	12		9-10	12		9-10	12		9-10	12	
GI	8-16, 34-35	19		10	19		10	19		10	19	
RESPI	8-16, 34-35	19		10	19		10	19		10	19	
SEROSA	8-16, 34-35	20		10			10			10		
OTHER	8-16, 34-35			10			10			10		
REPRO	8-16, 34-35	20		10			10			10		
MUSKO	8-16, 34-35			10			10			10		
NEURO	8-16, 34-35	21		10			10			10		
HEARTH	8-16, 34-35			10			10			10		
URI	8-16, 34-35			10			10			10		
LYMPH	8-16, 34-35			10			10			10		
NEO	8-16, 34-35			10			10			10		
MISS	8-16, 34-35			10			10			10		
NONCLASS	8-16, 34-35	22		10			10			10		

RESULTS

1. Descriptive analysis

- ***General data description***

The general data description was performed by S. Küker during her doctoral thesis and we only report her main results here.

Regarding data quality, the data export from the ITPA mistakenly horizontally shifted columns for all sub-reports excluding the macroscopic findings. The entries for the other values not belonging to the final necropsy report were exported correctly. One important consequence of the shift in fields was that the laboratory results obtained on animals necropsied at ITPA were not available. The details of data quality assessment are available in table 3.

A higher percentage of cattle submissions were from animals that had been euthanized (52% versus 32% in pigs) while 68% of the pigs necropsied had been found dead (versus 48% of cattle). Very few animals (total of 31) had been emergency slaughtered. The most common cattle breeds were Holstein Friesian (n=573), Simmental Fleckvieh (n=371), Red Holstein (n=365) and crosses between Simmental Fleckvieh and Red Holstein (n=250). These represent the most common breeds in Switzerland. The greatest proportion of cattle necropsied fell into the "2 years and above" age class (30.2%). These animals were almost exclusively females. However, the large parts of cattle submissions were of animals younger than 6 months (37.5%). For pigs, the most prevalent age class was piglets (35.3%) followed by weaners (21.7%) and adults animals (14.2%).

Table 3: Evaluation of data quality (completeness and credibility) for the 13 descriptive variables extracted from the post-mortem reports for pig and cattle between 2000 and 2011. Columns shaded in grey correspond to variables which were only present in cattle submissions. The TVD number is the unique cattle identification number. (S. Küker)

		Submission number	**Sex**	**Age (days)**	**Weight (kg)**	**How animal died**	**Owner name**	**Owner address (street)**	**Owner address (postcode)**	**Breed**	**TVD number**
Cattle	*Complete*	2862	2753	2114	2742	2786	2862	2611	2847	2279	1604
	Missing	-	109	549	1	76	-	251	15	583	9
	Invalid	-	-	199	119	-	-	-	-	-	1284
	% unusable entries	0	4	27	4	3	0	9	<1	21	47
Pig	*Complete*	5997	5823	4361	5888	5901	5996	5809	5983	NA	NA
	Missing	-	174	1471	1	96	1	188	14	NA	NA
	Invalid	-	-	165	108	-	-	-	-	NA	NA
	% unusable entries	0	3	27	2	2	<1	3	<1	NA	NA

- ***Pathological patterns found in necropsy reports***

A relatively large proportion of animals were classified into more than one syndrome category (36% of pigs and 45% of cattle) (see figure 1). However, approximately 10% of animals in both species were not classified into a specific syndrome category and were put in the category NON-CLASS. These animals had an unclear cause of death with no specific pathological pattern, or they were misclassified (i.e., they should have been classified into a specific syndrome category but the text mining tool was not able to correctly classify them).

The four most frequent pathological syndrome classifications were the same in both species: gastrointestinal system (pigs 55%, cattle 41%), serous membranes (pigs 16%, cattle 28%), respiratory system (pigs 18%, cattle 22%) and "other" (pigs 15%, cattle 14%). Reproduction system syndrome represented a significant proportion of cattle necropsies (19%) but this syndrome was less common in pigs (3.8%) (Figure 2).

Important differences were observed between age classes (Figure 2). For example, reproduction system syndromes were mainly found in adult cattle and pigs, and respiratory syndromes mainly in young cattle. Piglets were more commonly classified into the category "OTHER" compared to older age classes of pigs. These results are consistent with what is known about the main diseases in these species and age classes: only adult animals reproduce and they are thus the ones being send in for reproductive problems, respiratory diseases are well known to be most common in calves, and piglets tend to have quite many diseases and to die from unclear reasons when the sow is sick.

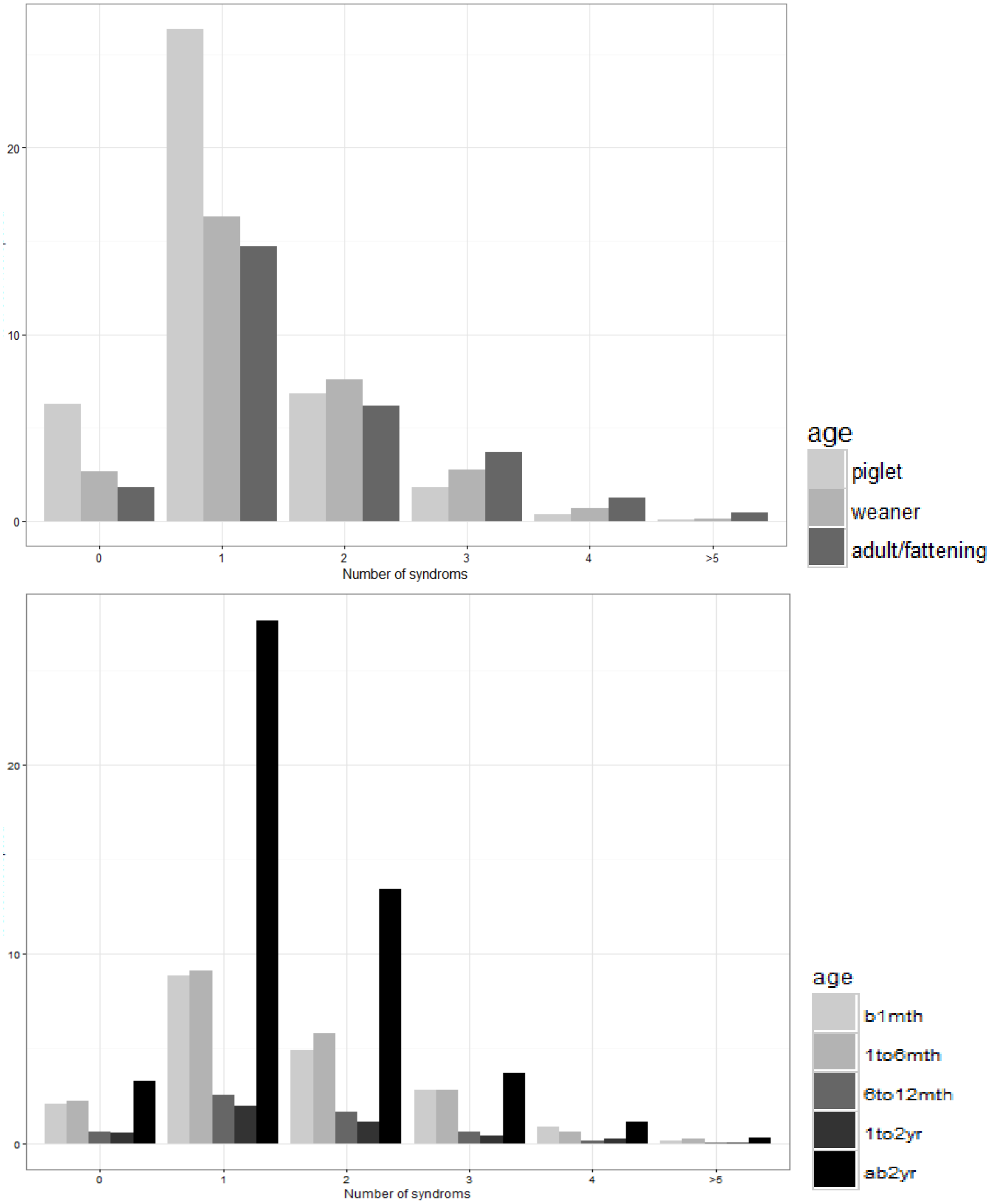


Figure 1: Percentage of animals necropsied according to the number of syndromes (or syndrome categories) associated (pigs: upper graph, cattle: lower graph)

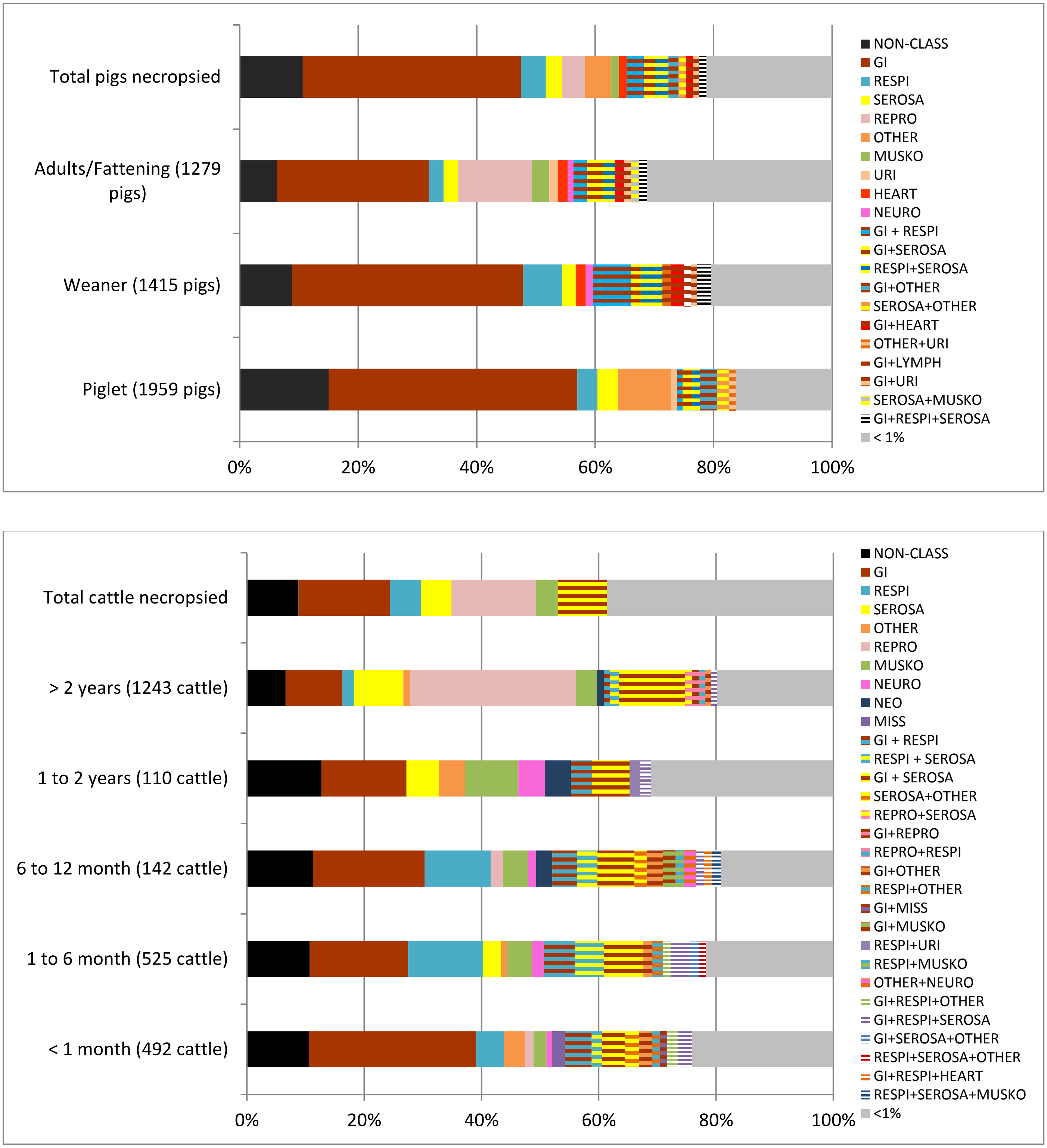


Figure 2: Main syndromes or groups of syndrome found in animals necropsied at ITPA between 2000 and 2011. Syndromes or groups of syndromes with less than 1% of animals in each age class are not shown (e.g., in the upper graph, category NEO does not appear, because less than 1% of pigs belong to the category NEO).

2. Time Series of the total number of submissions

- ***Cattle***

The total number of submissions for cattle was fairly stable over the study period. Necropsied animals came mostly from the western part of Switzerland (figure 3).

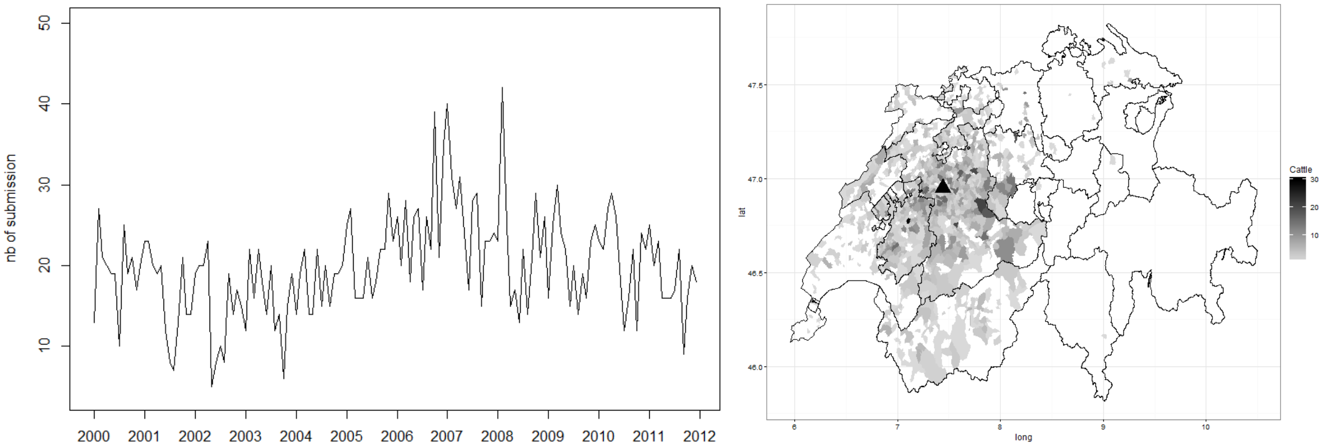


Figure 3: temporal and spatial distribution of cattle necropsied at ITPA between 2000 and 2011 (weekly data)

Models tested did not properly fit the data when all cattle age classes were merged into one TS. Age classes were thus separated into individual TS, according to the number of animals reported in the age classes previously defined. Young cattle from 6 months to 2 years were combined into one TS.

Submissions from neonatal calves and calves between 1 and 6 months had no trend and the same seasonal pattern (peak in winter) and thus these age classes were combined into one TS for rest of this report. Important differences were observed among the other age classes: adult cattle (greater than 2 years of age) had 2 annual peaks in the number of submission, one in spring and one in autumn; whereas younger animals only had one peak, in winter or autumn depending on the age class (see Figure 4 and appendix (table 2)). In adult cattle, an increasing number of reports were observed over the study period. This increase was mainly due to an increase in the submissions of adult cows, but it was also observed to a lesser extent in the submissions of abortion/stillbirth (which belong to adults cows). The number of calves submitted was stable over time, and the number of young cattle submitted decreased slightly over the study period.

The ACF and PACF plots of the model residuals, shows that autocorrelations remained in the residuals which indicate that the best fitting models were not able to model all of the variation present in these TS. The remaining autocorrelations might be explained by the presence of some big peaks in the number of submissions in the TS (e.g., mid 2007 in young cattle, winter 2010 in adult cattle). Removing them from the data using automatic procedures is possible and could help to improve our models. However, because we don’t know if these peaks represent normal extreme variations in the data or some abnormal patterns (like disease outbreaks), we choose to keep them in the data for the rest of the analysis.

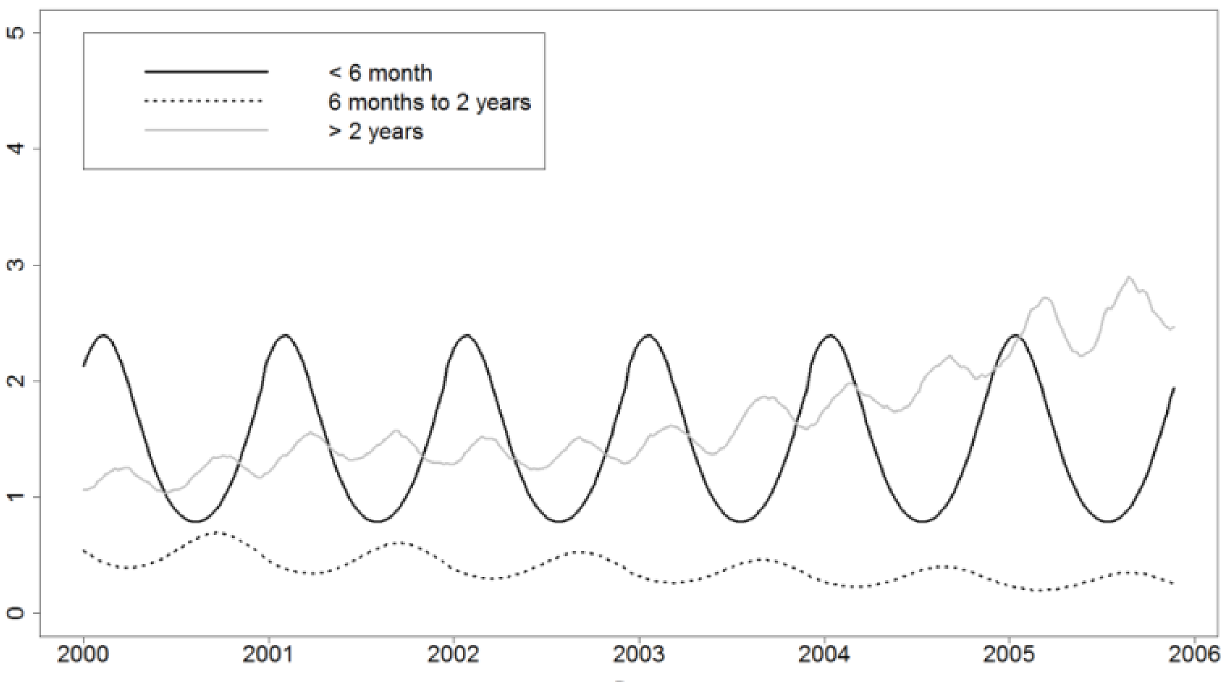


Figure 4: Mean of the fitted values for each age class in cows (data after 2006 not shown). The number of calves < 6month necropsied follows a Negative binomial distribution (Y ~ mycos+ mysin); The number of young animals from 6 months to 2 years necropsied follows a Poisson distribution (Y ~ mysin+ time); The number of adult animals > 2 years necropsied follows a Negative binomial distribution (Y ~ time + histmean + sinX2.shift)

- ***Pigs***

The total number of pig submissions decreased over time. Animals came only from the area surrounding Bern (see Figure 5).

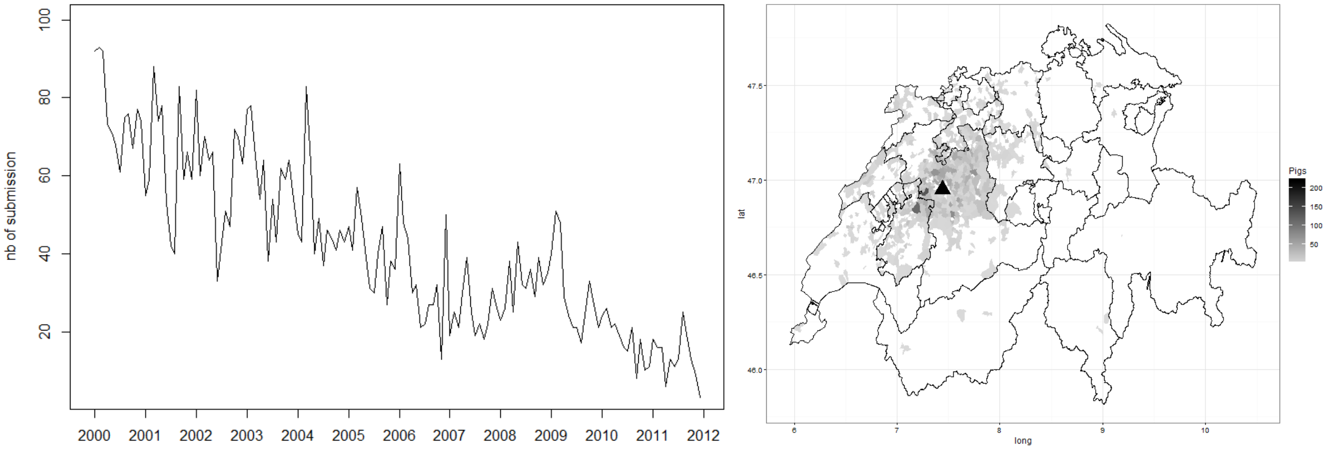


Figure 5: temporal and spatial distribution of pigs necropsied at ITPA between 2000 and 2011

Similar to cattle, data for the different age classes were analyzed separately. A similar seasonal pattern was found across all age classes, with fewer pigs being necropsied in summer and more in winter (see Figure 6 and appendix (table 2)). However, seasonality in piglets was slightly different: 2 peaks were observed in autumn and winter and the decrease in submissions was more prominent in the spring than in summer.

Similarly to cattle, the ACF and PACF plots of the model residuals, show small autocorrelations remaining in the residuals which means that our models were not able to model all of the variation present in the TS. These remaining autocorrelations may also be explained by the presence of some big peaks in the dataset (e.g., winter 2008 in weaners, summer 2004 in fattening/adults pigs). Removing them is also possible using automatic procedures, which could help to improve our models. However, as we don't know if these peaks represent normal extreme variations in the data or some abnormal patterns (like disease outbreaks), we choose to keep them in the data for the remainder of the analysis.

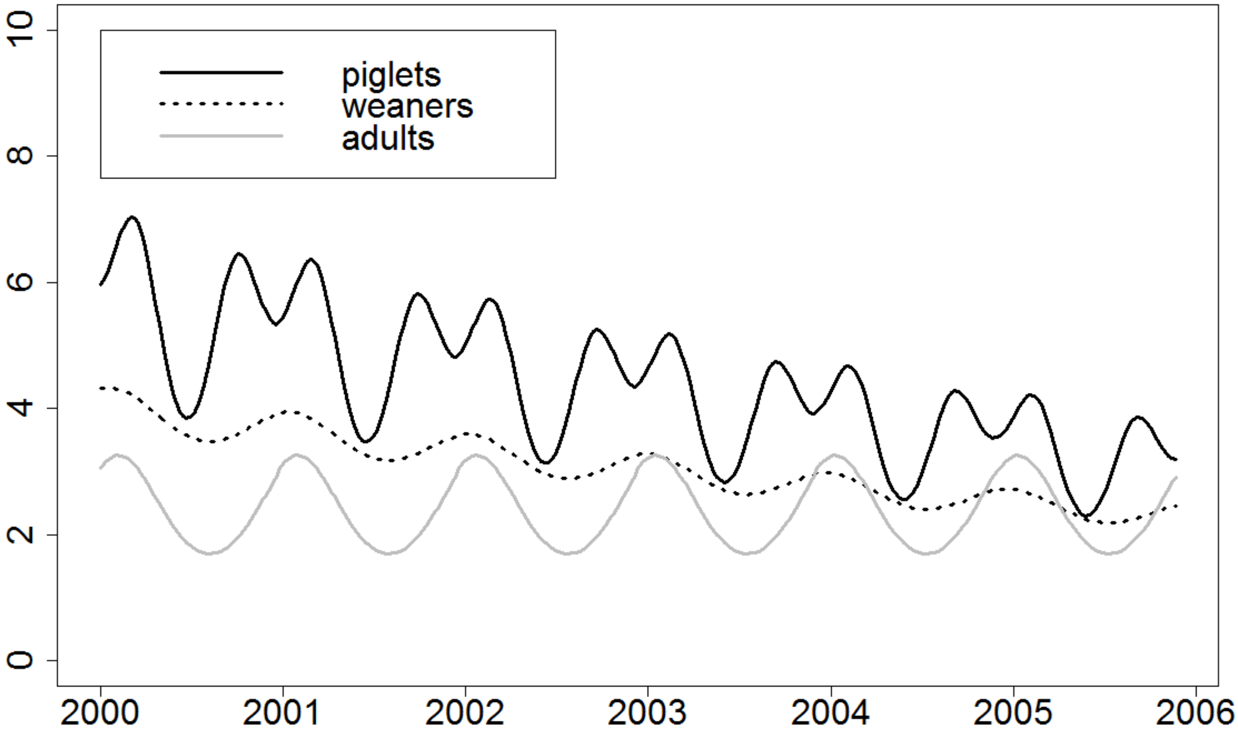


Figure 6: Mean of the fitted values for each age class of pigs (data after 2006 not shown). The number of piglets necropsied followed a Negative binomial distribution (Y ~ time + mycos + sinX2.shift); The number of weaner pigs necropsied followed a Negative binomial distribution (Y ~ time + sinminus18); The number of adults pigs necropsied followed a Negative binomial distribution (Y ~ mycos + mysin)

3. Temporal description of syndrome categories all age classes together

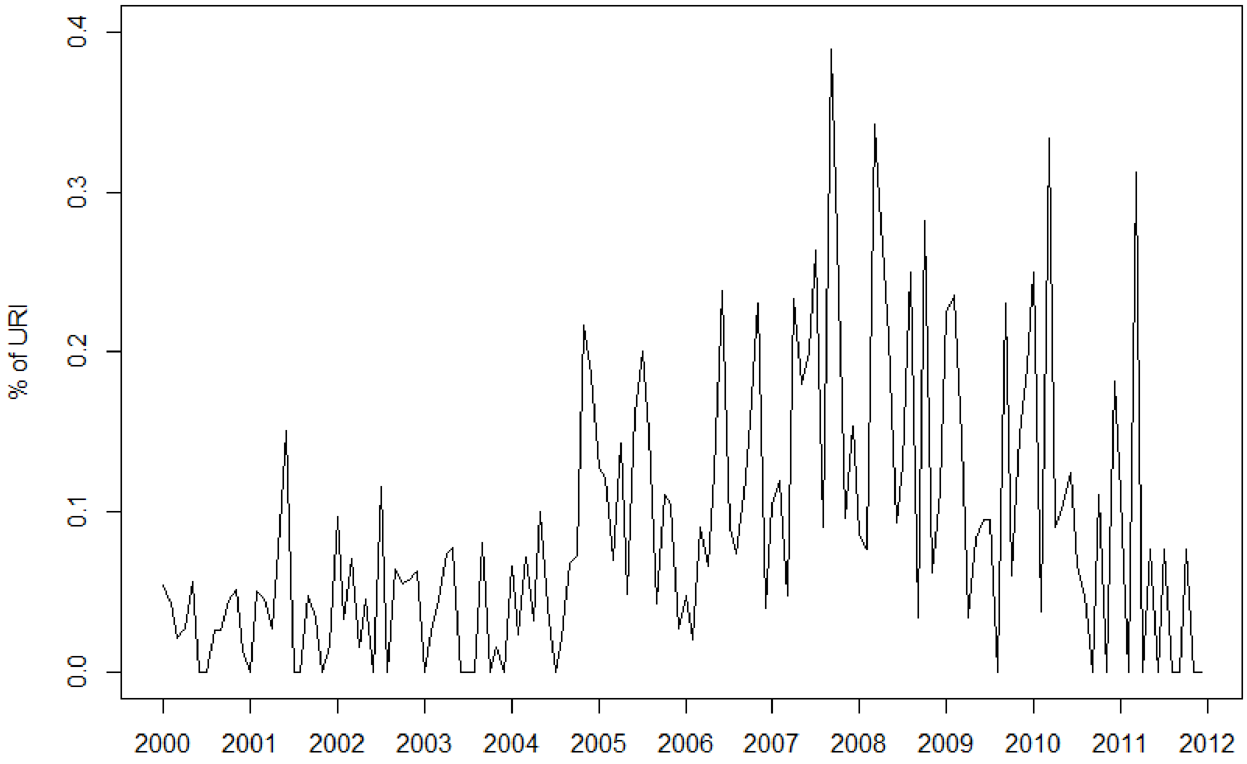
In the first step of the analysis, visual examination of the TS created for each syndrome category was performed. All TS are presented in the Appendix (Figures 1 and 2).

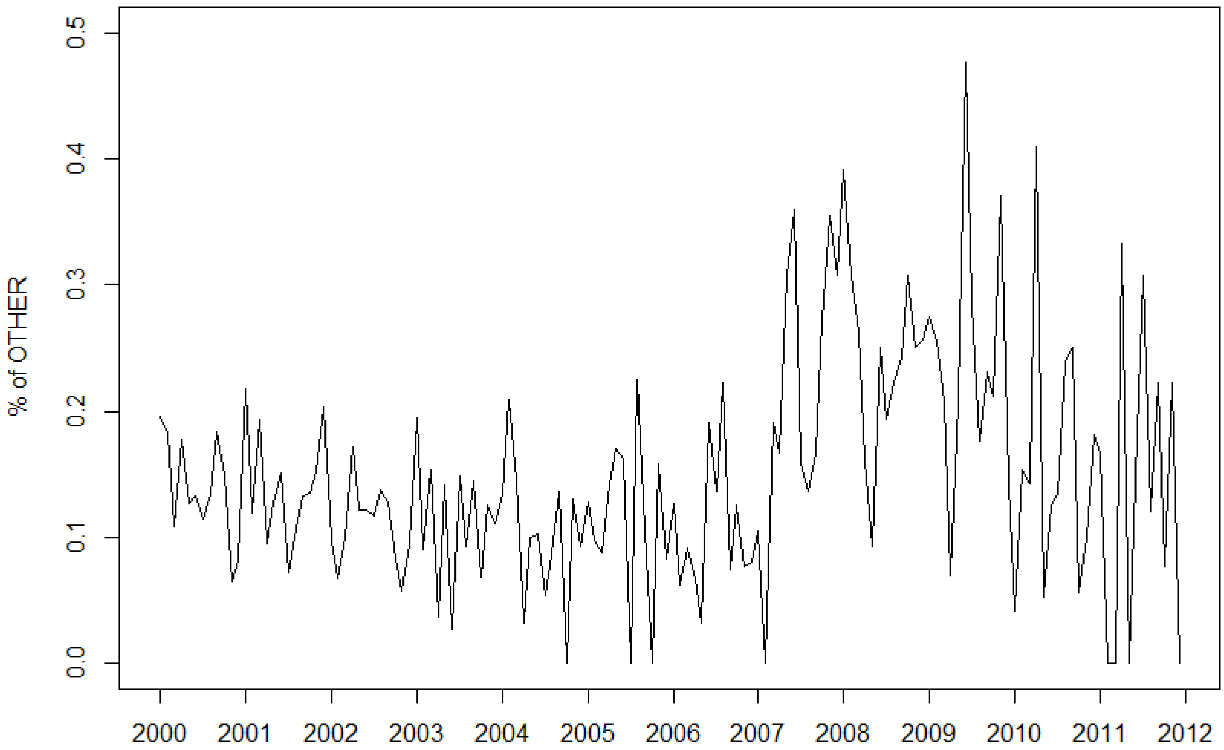
- ***Cattle***

An increasing number of submissions were observed in cattle between the years 2006 and 2008. The increase was seen mainly in cattle between 1 and 6 months and in the RESPI, GI and SEROSA syndromes. No other large peak was observed in cattle TS.

- ***Pigs***

There was an increase in the number of pigs presenting with URI, LYMPH and OTHER syndromes between the years 2007 and 2010 (see figure 8 for proportion data and appendix (figure 2) for count data).





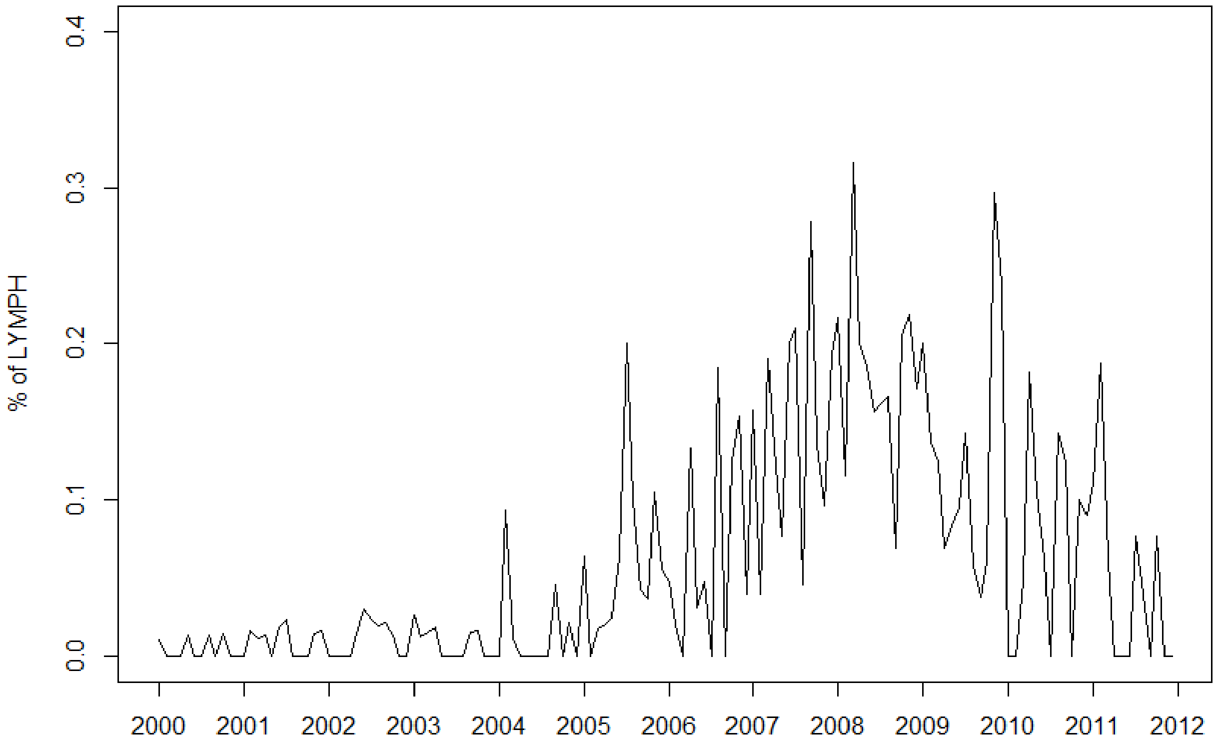


Figure 8: The percent of cases (%) of pigs of all ages combined in syndrome categories URI, OTHER and LYMPH between 2004 and 2010.

An increase in the number of animals in the NON-CLASS category was also observed between 2003 and 2005 with a second peak during the winter of 2005/2006 (see figure 9 for proportion data and appendix (figure 2) for count data). The increase in the number of cases from 2003 to 2005 affected all age classes. However, the specific peak in the winter of 2005/2006 mainly affected piglets. Other important variations in the number of cases are also observed in 2010 and mid 2011 in this syndrome category.

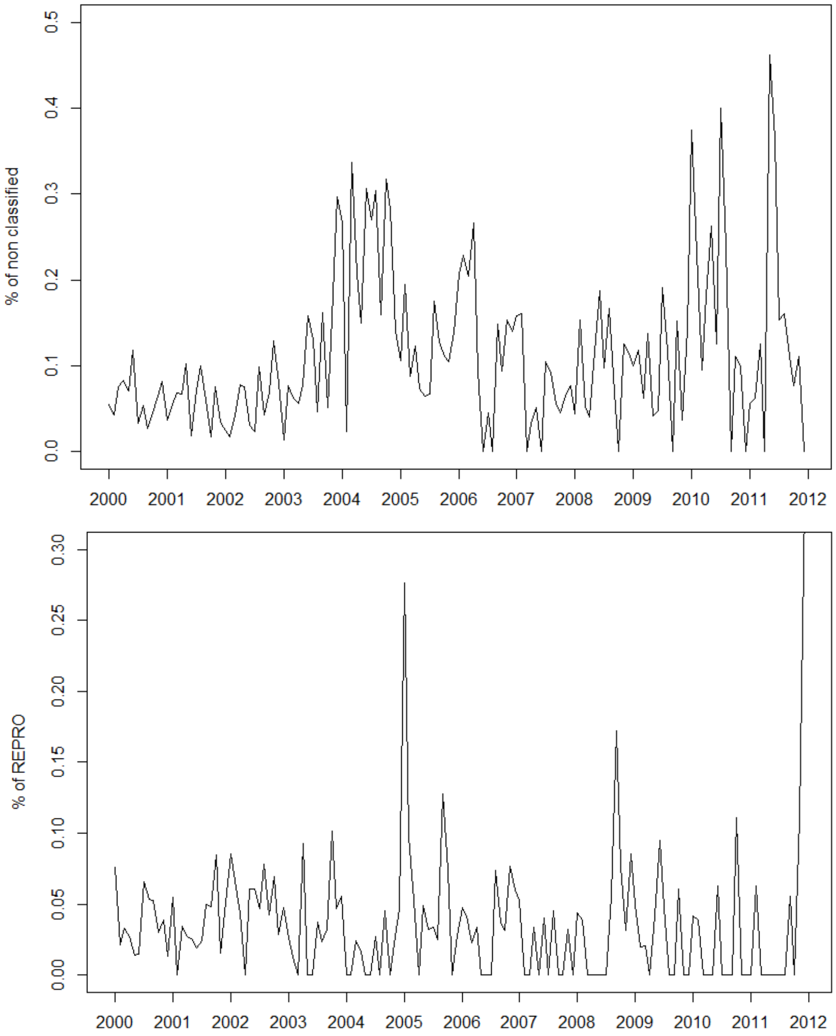


Figure 9: The percent of cases (%) of pigs of all ages combined in syndrome categories NONCLASS and REPRO.

A single peak in the REPRO syndrome category was also observed in 2005 (see figure 9 for proportion data and appendix (figure 2) for count data). The 19 reports causing this peak were investigated in more detail. They came mainly from Bern area. However, the big peak observed is mainly due to the fact that one report was written for each fetus submitted even though some of them came from the same litter. No diagnosis was made and the findings in all fetuses were the same, which is common in pigs. Writing one report for each fetus in the same litter is not a common practice. Thus this peak was most likely due to a real increase in the number of abortions in the population.

After the preliminary visual inspection of the data, the systematic variations present in the syndrome categories where investigated with all age classes combined into one TS. Only a few syndrome categories showed a monthly effect: GI, RESPI, SEROSA and NON-CLASS for cattle, and GI, RESPI, SEROSA, REPRO and URI for pigs. However, as reported in the previous section, seasonal effects may be different in different age classes. To address this question, we focused on the syndromes selected in consultation with BLV.

4. Modelling of syndrome TS by age class in cattle

We analyzed only the syndrome categories previously selected (i.e., GI, RESPI, REPRO, SEROSA, NEURO, NONCLASS). We first modeled these TS with all age classes combined into one TS and then we modelled individual age classes for the most important syndrome categories (RESPI and GI).

a. GI syndrome in cattle

The model that best fit the GI syndrome TS with all age groups combined into one TS was obtained using a Negative Binomial distribution (GI ~ season + sinX2) (See figure 10). The TS had peaks in autumn and winter corresponding to the peaks observed in calves and young cattle.

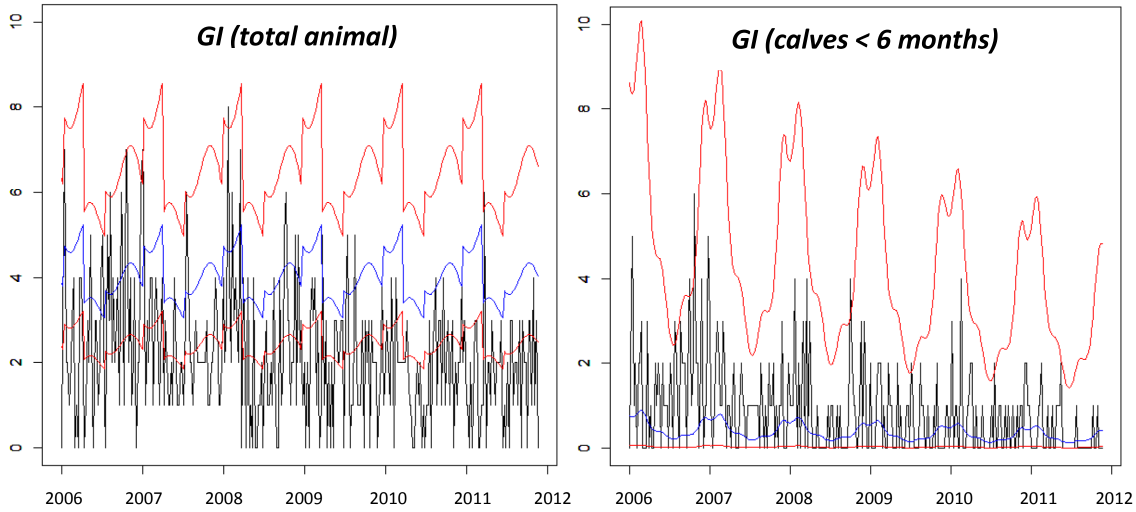


Figure 10: TS models of cattle with GI syndrome. Blue line=mean fitted value, red lines=95% Prediction Interval

Calves (<6month): The best model was obtained with a negative binomial distribution (GI ~ sinminus18 + time + sinX4). See figure 10. The seasonal pattern found in this TS is similar to the seasonal pattern found in the TS with the total number of calves submitted.

Young (6month-2years): The best model was obtained with a Negative binomial distribution (GI ~ mysin + sinX4.shift + sinX3). More submissions were seen in summer and autumn which is similar to the seasonal pattern found in the total number of young cattle submitted.

Adults (>2years): The best model was obtained with a Poisson distribution (GI ~ time). There was no clear seasonal pattern except for a positive trend as seen in the TS of the total number of adult cattle necropsied at ITPA.

b. *RESPI syndrome in cattle*

The best fitting model for RSEPI syndrome for the TS with cattle of all ages combined was obtained using a Negative Binomial distribution (RESPI ~sinminus18 + sinX3). In this TS there were more cases during the winter season (see figure 11). The seasonal and trend component are similar to those observed in young animals. This is consistent with the fact that RESPI is a syndrome category mainly present in young animals.

Calves (<6month): The best model was obtained with a Poisson distribution (RESPI ~ mycos + mysin). Again, the seasonal pattern found in the TS is similar to the seasonal pattern found in the TS of the total number of calves submitted.

Young cattle (6month-2years): The best model was obtained with a negative binomial distribution (RESPI ~ sinX3 + sinX3.shift). Even though it was the best model, it did not fit the data very well. Autocorrelations remained in the residuals. This may be due to the small number of reports that were available in this age class and the poor fitting model.

Adults (>2years): The best model was obtained with a Poisson distribution (GI ~ time + sinX2.shift). The seasonal pattern found in this TS was one with two annual peaks, one in spring and one in autumn. This is similar to the seasonal pattern found in the TS of the total number of adult cattle necropsied.

c. SEROSA syndrome in cattle

The best model was obtained using a Negative Binomial distribution (SEROSA~ season + time). This TS had one peak in winter/spring and a positive (increasing) trend over time (see figure 11). This increasing trend is consistent with the fact that most cattle submissions with SEROSA were adults and that the number of adult cattle necropsied at ITPA had an increasing trend over time.

d. REPRO syndrome in cattle

The best model was obtained using a Negative Binomial distribution (SEROSA ~ time + sinminus6 + sinX4.shift). This TS included abortions, stillbirth and animals with reproduction problems. More cases were observed in autumn and fewer cases in spring with a positive trend over time (see figure 11). This is consistent with the fact that almost all cattle with REPRO syndrome were adults and that the number of adults cattle necropsied at ITPA has an increasing trend over time. However, the TS for cattle with REPRO had a seasonal pattern which was different from the seasonal pattern observed in the TS of the total number of adult cattle, which has two annual peaks, one in autumn and one in spring. The TS of adult cattle with REPRO syndrome thus has a specific seasonal pattern among adult cattle. This may be due to the inclusion of abortions in this syndrome category. Depending on the objective of surveillance, it could be interesting to create a specific syndrome category for abortions and/or stillbirth separate from REPRO syndrome.

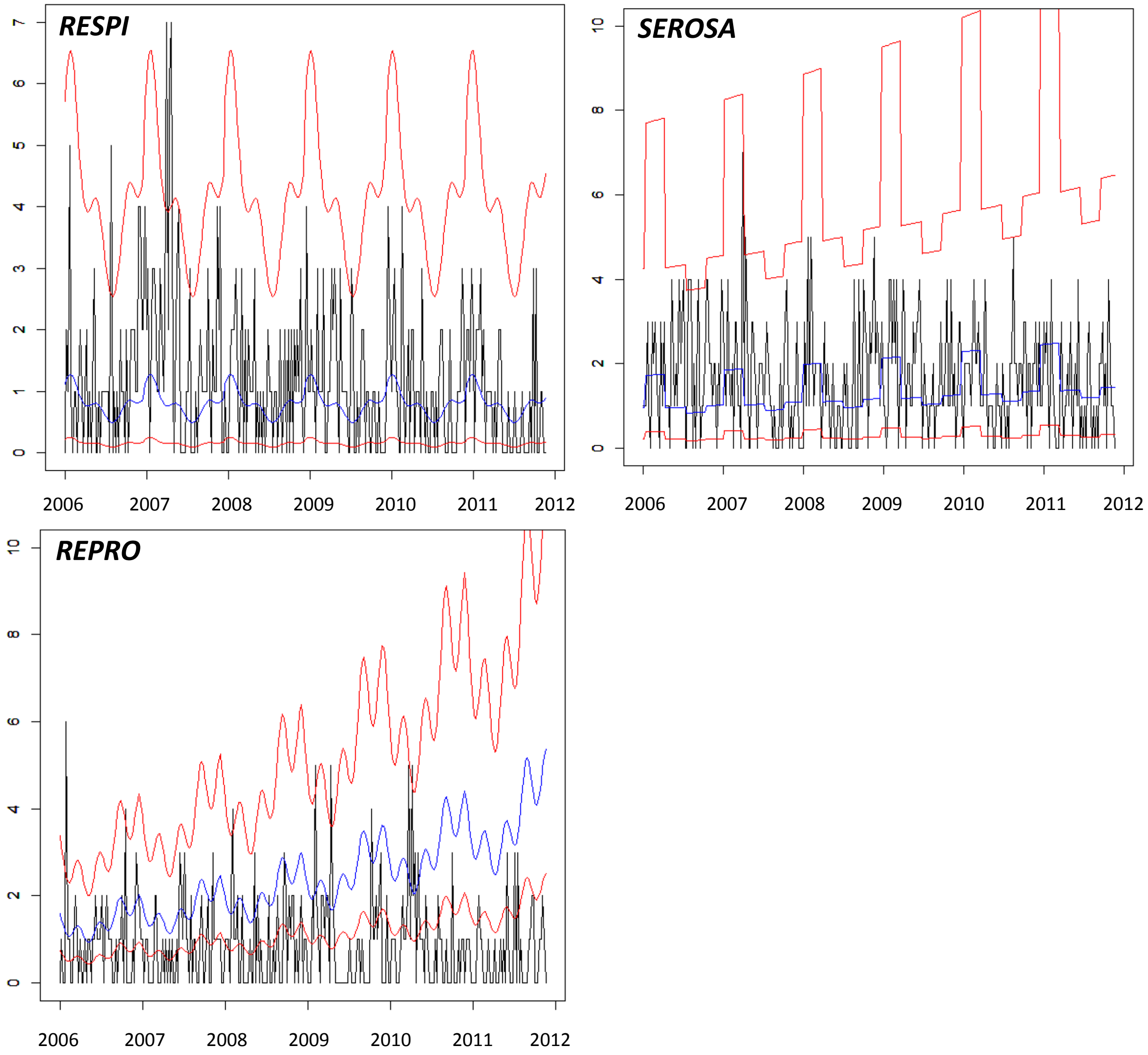


Figure 11: TS models of the total number of cattle for 3 syndrome categories. Blue line=mean fitted value, red lines=95% Prediction Interval

e. NEURO syndrome in cattle

The null model (i.e., the model without any explanatory variables) was the best fitting model for this time series. The TS was stationary, which means that the distribution of the TS does not change with time (no seasonality and no trend). The number of cattle necropsied at ITPA that had nervous lesions was thus constant over time.

f. NON-CLASS syndrome in cattle

The null model was the best model for this TS. It was also stationary which means that the marginal distribution of the process does not change with time (no seasonality and no trend). The number of cattle necropsied at ITPA and not classified into a syndrome category was thus also constant over time.

5. Modelling of syndrome categories by age class in pigs

We worked only with the syndrome categories previously selected (i.e., GI, RESPI, REPRO, SEROSA, NEURO, NON-CLASS). Pigs with both RESPI and REPRO syndromes were not investigated because only 7 pigs in the entire dataset presented with these 2 categories. Our first step was to model the TS with all age classes combined into one TS, and then to create specific TS for each age class.

a. GI syndrome in pigs

The model that best fit the GI syndrome for all pigs was a Negative Binomial distribution (GI ~ time + mycos + sinX2.shift) (See figure 12). The TS had a peak in autumn/winter similar to the peak identified in the TS of all piglets. This could be explained by the fact that GI syndrome is mainly present in piglets.

Piglets: The best model for piglets with GI syndrome was obtained with a negative binomial distribution (GI ~ time + sinX2.shift + sinX3.shift). See figure 11. However, some autocorrelation remained in the residuals of this TS, which was mainly due to the increased number of submissions in 2004 in this age class.

Weaners: The best model was obtained with a Negative binomial distribution (GI ~ time). No seasonality was found and the main temporal component present in this TS was a decreasing trend (see figure 12).

Adults/Fattening pigs: The best model was obtained with a Negative binomial distribution (GI ~ mycos + mysin + time) with more cases reported during the winter season.

b. RESPI syndrome in pigs

The best model for the TS of pigs of all ages combined with RESPI syndrome was obtained using a Negative Binomial distribution (RESPI ~ time + mycos) (See figure 13). The TS had a peak in winter and spring.

Piglets: The best model was obtained with a negative binomial distribution (RESPI ~ season + sinX3.shift + sinminus18) with more cases observed in winter and spring.

Weaners: The best model was obtained with a Negative binomial distribution (RESPI ~ mysin + histmean + time + sinX3.shift) with more cases observed in winter and spring.

Adults/Fattening pigs: The best model was obtained with a Negative binomial distribution (RESPI ~ sinminus18 + time + mycos) with more cases reported in winter and spring.

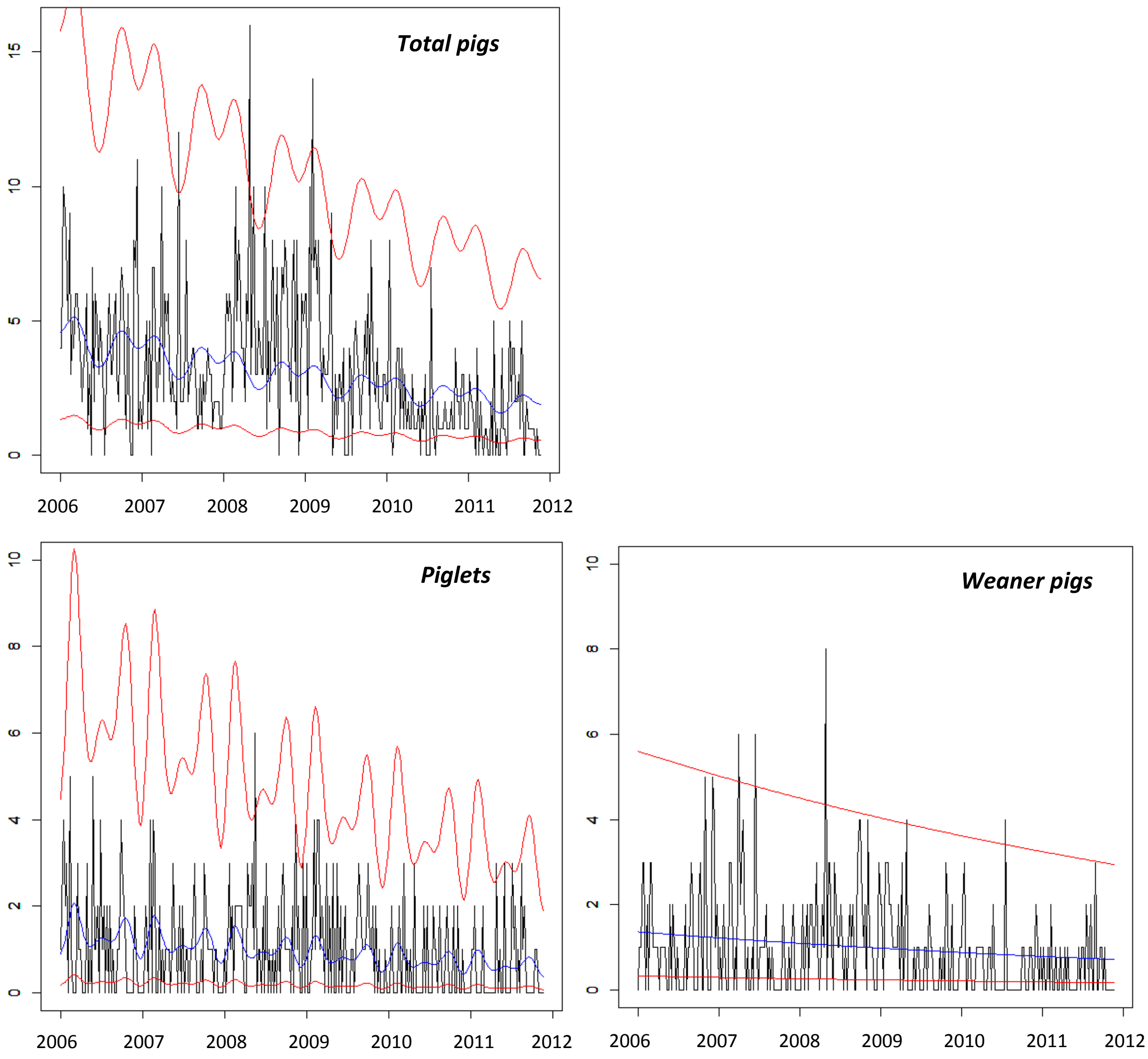


Figure 12: TS models of pigs with GI syndrome. Blue line=mean fitted value, red lines=95%Prediction interval

c. SEROSA syndrome in pigs

The best model for pigs of all ages with SEROSA syndrome was obtained using a Negative Binomial distribution (SEROSA ~ mycos + mysin + sinX2.shift) (See figure 13). However significant autocorrelations remained in the residuals probably partially due to some peaks present in the TS. The TS had a peak in winter plus an increase in the number of cases in autumn.

d. REPRO syndrome in pigs

The best fitting model for REPRO syndrome with all ages combined was obtained using a Negative Binomial distribution (REPRO ~ season + sinX3 + time + histmean) (See figure 13). However autocorrelations remained in the residuals, probably partially due to a peak in the number of cases in 2005. In the REPRO TS there were fewer cases in the spring than in the rest of the year.

e. NEURO in pigs

The best model for all pigs with NEURO syndrome was obtained using a Negative Binomial distribution (NEURO ~ time + season.shift2) (See figure 13). However important autocorrelations remained in the residuals which may have been due to a peak in submissions in 2004. In the NEURO syndrome TS, there were more cases during the winter than other times of the year.

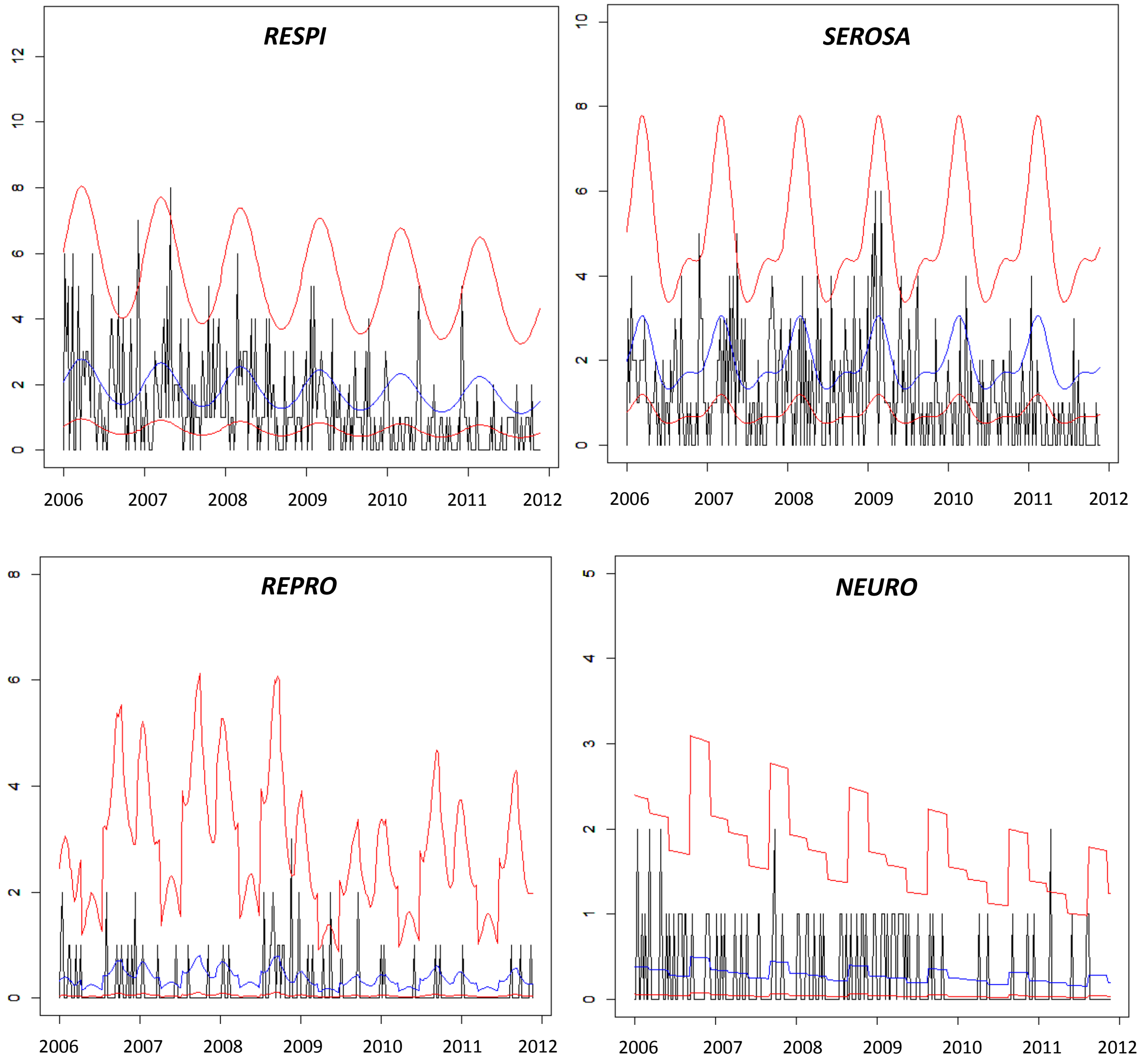


Figure 13: modelling of the total number of pigs for 4 syndrome categories. Blue line=mean fitted value, red lines=95%prediction Interval

f. NON-CLASS in pigs

We were not able to model this TS because of the abnormal patterns previously described.

6. Aberration detection algorithms in cows

We tested different aberration detection algorithms and we were able to identify abnormal events in the TS using temporal early detection algorithms and spatio-temporal detection algorithms. Only the full results obtained with RESPI and NEURO in cattle are reported below.

- **RESPI in cattle (all age classes combined into one TS)**

We compared 7 different temporal event detections algorithms. The raw TS were not stationary and for this reason we applied all algorithms, except Farrington's on model residuals. The model used was the one defined in previous section, a Negative binomial distribution model with the equation: RESPI ~sinminus18 + sinX3. The results obtained with each algorithm are presented in figure 14.

As already explained in the Material and Methods section, it was impossible to identify the best algorithm for detecting an epidemic as true epidemic periods were not present in the data. However, more than four algorithms produced an alarm on week 24 in 2005, week 31 in 2006, weeks 15, 17 and 18 in 2007 which may indicate that there were true abnormal events occurring during these time periods. These results are also consistent with those obtained with the spatio-temporal analysis (see figure 15): clusters 3 and 4 in purple, clusters 2 and 4 in blue also cover the same time periods. Other time periods identified as "abnormal" in the spatio-temporal analysis were also identified by temporal algorithms. In particular, the first weeks of January 2010 was identified as abnormal by cluster 1 in purple, cluster 3 in blue and by the 3 EARS algorithms C1, C2 and C3. However, we cannot make any further comparisons between the spatio-temporal and temporal analysis because the latter works on a weekly basis and the former on a monthly basis. The work done by M. Arede was an extra and not foreseen part of the initial project proposal. This additional part can be regarded as a promising pilot study which needs further follow up.

There were also differences in the aberrations detected by the various algorithms. The EARS algorithms tended to produce more alarms than the CUSUM, EWMA or Shewhart chart algorithms. The differences might be explained by the fact that, even if RESPI is one of the most prevalent syndrome categories, the RESPI TS may contain too many zeros to obtain valid results from some of the algorithms. CUSUM, Shewhart chart and EARS do not perform well when there are too many zeros. EWMA and Farrington are reported to perform better on sparse data and the outputs of these algorithms may be more reliable on the RESPI TS. Differences may also be explained by the fact that the algorithms do not detect exactly the same type of abnormal events. However, we cannot make further algorithm comparison because of the absence of true epidemics in the data.

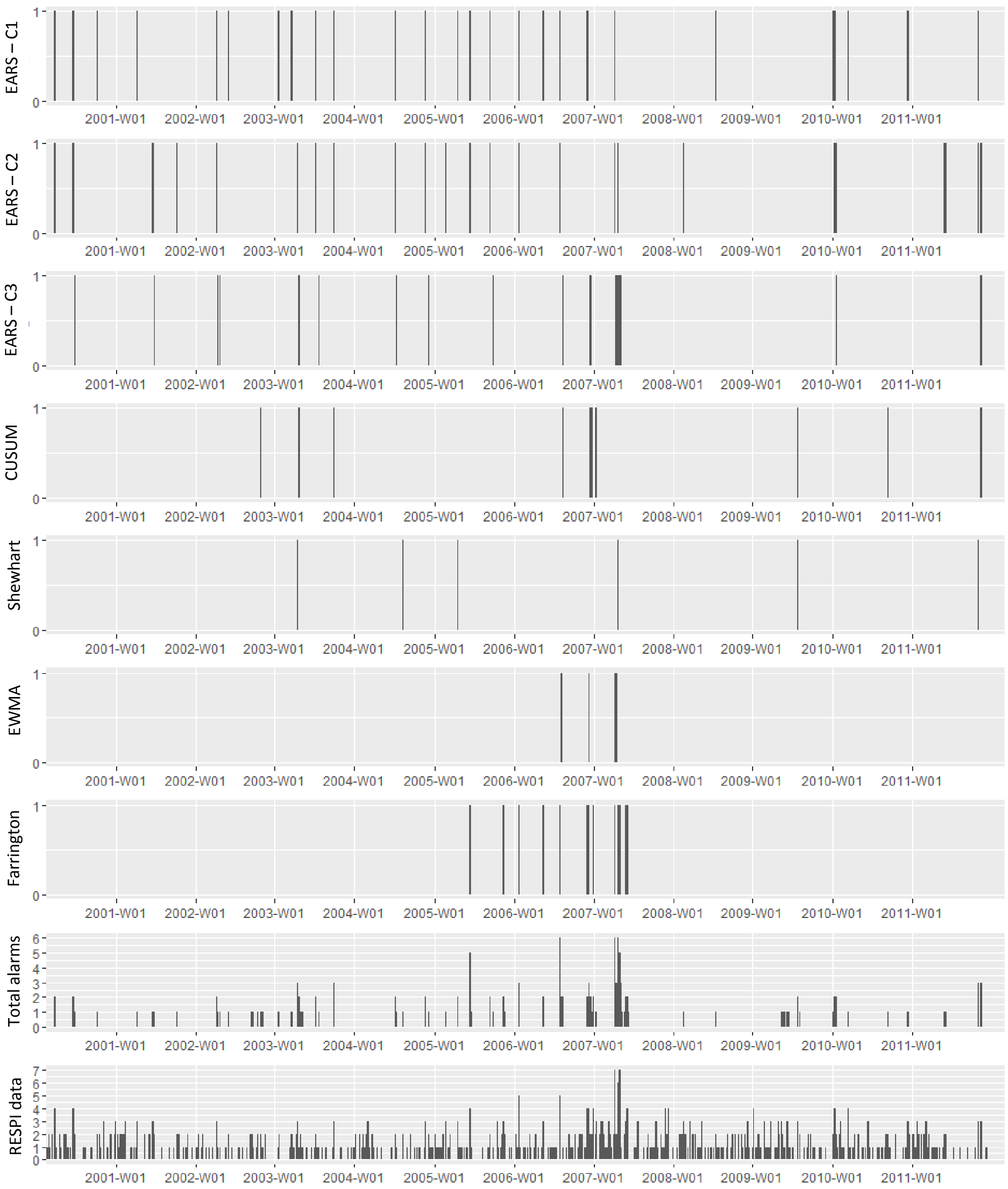
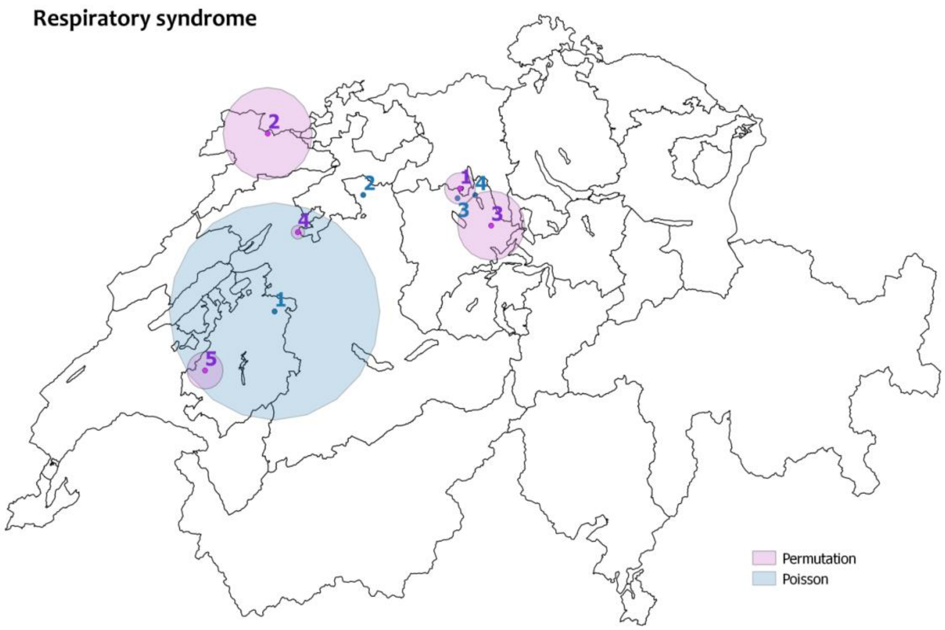


Figure 14: Alarms produced by 7 different aberration detection algorithms (0: no alarm, 1: alarm) applied to RESPI weekly data. Cumulative number of alarms obtained with different algorithms is summarized in graph named "total alarms" and the raw data RESPI are available in the lower graph.



Cluster	Start date	End date
1	01.08.2009	28.02.2010
2	01.11.2000	30.11.2000
3	01.12.2006	31.07.2007
4	01.10.2006	30.11.2006
5	01.08.2010	30.11.2011
1	01.10.2005	30.04.2011
2	01.02.2007	31.05.2007
3	01.08.2009	28.02.2010
4	01.12.2006	31.01.2007

Figure 15: spatio-temporal analysis of RESPI monthly data in cattle. For more details, please look at the report written by Margarida Arede, VPHI intern.

- **NEURO in cattle (all age classes combined into one TS)**

Since the underlying data for the time series was quite sparse we only compared the 2 different detection algorithms adapted for count data with many zeros: EWMA and Farrington. The TS was stationary and thus the EWMA algorithm was directly applied to the raw data. The results are presented in figure 16.

As already explained in the Material and Methods section, it was impossible to identify the best algorithm as the true epidemic periods were unknown. However, the two algorithms both produced an alarm on week 8 in 2008 which may indicate that there was an abnormal event occurring during this time period.

The Farrington algorithm only raised one alarm (i.e., week 8 of 2008). This is consistent with what we would expect: this algorithm is not good for detecting a transient single peak, but is better for detecting large outbreaks with rapid increases in the number of submissions. This might explain why this algorithm did not raise an alarm for the single peaks occurring in week 18 in 2007 or in week 44 in 2008, even if 3 cases with neurological signs were reported. However, the Farrington algorithm raised an alarm in 2008 because many cases occurred in repeated weeks and were then followed by an increasing number of cases. EWMA generated more alarms than Farrington. Alarms were raised when cases occurred in consecutive weeks (e.g., in 2006 or 2003). Again, this is consistent with how this algorithm works: EWMA is mainly able to detect small but repeated changes in the data.

Once again, we cannot make further algorithms comparison given the absence of true positive cases in the data.

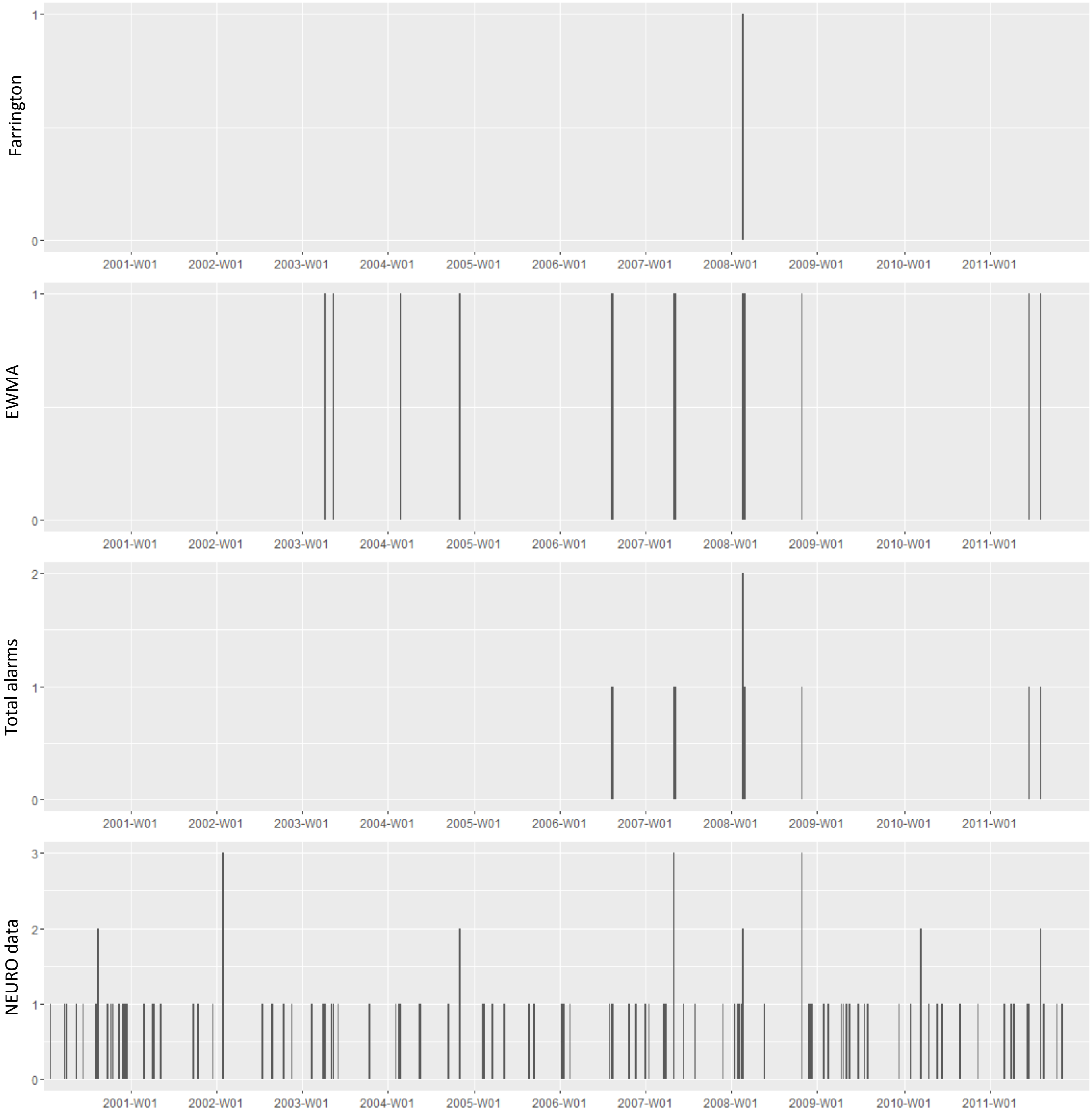


Figure 16: Alarms produced by 2 different aberration detection algorithms (0: no alarm, 1: alarm) applied to NEURO weekly data. Cumulative number of alarms obtained with different algorithms is summarized in graph named "total alarms" and the raw data NEURO are available in the lower graph.

DISCUSSION

In this study, clear differences between species and age classes were identified in terms of pathological and temporal (i.e., trend and season) patterns. Some syndrome categories were much more frequent in certain age classes (e.g., RESPI syndromes more frequent in young animals) and/or during certain time periods (e.g., in adults cattle RESPI syndromes occur mainly in spring and autumn but GI syndromes are present throughout the year). As the syndrome categories chosen covered a wide range of syndromes and situations, we were thus able to identify the main causes of death in animal submissions to the ITPA for each species and age class. Using syndrome categories, we were also able to identify by visual inspection and/or with aberration detection algorithms some temporal patterns not previously known by pathologists from ITPA. For example, the increase in the number of URI, LYMPH and OTHER syndromes between 2004 and 2010 in pigs was previously unknown. In addition, pathologists communicated an increase in the number of gastro intestinal syndromes in pigs during the past years which was not confirmed by our analysis (Horst Posthaus personal communication). These observations confirm that the data collected by ITPA and classified using the text mining tool can be used to better understand some of the diseases that are ongoing in Swiss animal populations. In particular we propose that these data can be used to:

- Define the normal behavior of animal populations in the absence of a specific epidemic (e.g., season with highest number of RESPI cases in young cattle). This information may have value for setting national benchmarks against which the livestock industries can measure improvements and also for national surveillance by providing some evidence for the absence of certain diseases (for example PRRS or BVD). In addition, having good estimates of the normal disease distribution is essential for identifying the emergence of new previously unseen diseases.
- Identify the main trends in the diseases causing cattle and pig submissions from their respective populations. For example we observed no increase in the number of pigs with GI syndromes but there was an increase in the number of pigs with URI syndrome. This information may have value for national surveillance programs by providing some evidence that an endemic disease may be increasing in importance in the population.
- Better understand farmers practices (e.g., increased number of submission for adult cattle but decreased number of submission for calves, global decrease of submission for pigs, preferences towards submitting females when farmers submit piglets for necropsy (Küker et al. 2016 (in prep)). This is important information for identifying gaps in surveillance. For example the downward trend in the number of swine submissions is an indication of farmers choosing to not use the Bern diagnostic lab for diagnostic information. The reason for this decline was not known. However, if the number of submissions was decreasing during a period when the swine population was constant, there would be a concern that routine diagnostic pathology would be less effective for detecting emerging diseases.
- Identify changes in disease surveillance programs or in the way that necropsies results are reported that may not be associated with a change in the disease status of the population. An example in this study was the creation of multiple reports for GI for piglets from the same litter. This information will be of value for quality assurance. Administrators who manage diagnostic laboratories could investigate the causes of abnormalities in submission reporting to ensure that the cause was not an unusual reporting method.
- Better understand and define associations with pathological lesions. Pathologists could use these data to identify risk factors and specific pathological presentations.

- Teaching at veterinary schools. Having specific data about diseases presentations, and seasonal or regional variation in diseases in the country may help future pathologists and clinicians to better understand livestock disease patterns which may improve their diagnostic abilities.
- Perform continuing education for veterinary practitioners and producers about common diseases and trends in disease prevalence and the importance of veterinary diagnostics for the producer over longer time periods. Using this data for example the effectiveness of vaccination or eradication programs could be documented and communicated in an easily understandable way.

These data can be used for many purposes but, it is important to point out that the ITPA data have limited representativeness of the Swiss animal population. For pigs, the data collected by ITPA only represent a small fraction of the pig population in Switzerland. For cattle, the representativeness seems to be a bit better (i.e., animals come from half of Switzerland) but is still incomplete coverage of the total population. For future development it will be important to incorporate data from additional pathology diagnostic laboratories to achieve a coverage of most if not all farm animal necropsies in Switzerland. Our study also shows that variations in reporting necropsy results within one academic institution can vary greatly over time and creates major difficulties for the automated categorization of pathology findings. This will be even more challenging when data from different laboratories and language regions in Switzerland will be combined. Therefore, to achieve Swiss wide coverage for automated screening of pathology data a harmonized reporting system between all participating laboratories should be implemented.

Pathologists at the ITPA have reported an increasing number of colitis submissions in pigs during the past years. This pathological finding belongs to the syndrome category GI, but no abnormal pattern was observed in the TS for this syndrome category. However when a new syndrome category focusing on the word “colitis” was created, the increasing number of cases of “colitis” was observed (data not shown). This example raises the question of the validity of the syndrome categories chosen. The syndrome categories used in this study were primarily based on the topography of organ systems. However, they could be easily modified for any specific surveillance need. Indeed, the text mining tool used in this study is very flexible and we could thus easily build a new category focusing for example on pigs with RESPI and/or REPRO syndrome combined in order to monitor Porcine Reproductive & Respiratory Syndrome (PRRS). We believe that the categories chosen are suitable for a wide range of situations and especially for non-specific health monitoring. These categories cover the most frequent syndromes seen in animals. They are the same syndrome definitions that were used in other studies allowing the results to be compared to those in other studies. There is no theoretical limit to the number of syndromes that care monitored. It would be possible to monitor the syndrome chosen for this study, and in addition create and monitor a number of additional syndromes aimed at identifying specific diseases.

In this study only two syndrome categories were problematic: OTHER and REPRO. The syndrome category OTHER includes very diverse pathological lesions (e.g., skin lesions, intoxication, and septicemia) and the difference between OTHER and NONCLASS was often hard to explain and understand. We would thus suggest redefining this category to either build new categories based on more specific pathological patterns, or combine OTHER with NONCLASS in order to have a unique category with all animals that were not classified into one of the main syndrome categories. The REPRO syndrome category currently includes data on abortion and stillbirth. We suggest building a specific category for abortion and stillbirth as they represent a specific pathological pattern which is widely used in animal disease surveillance (e.g., (Bronner et al., 2015)). Such a syndrome category would be very valuable in a surveillance system even without any specific disease focus.

Ten percent of the submissions were not classified into a specific syndrome category and belonged to the category NONCLASS. There are two types of animals in this category: animals with an unclear cause of death and animals that were misclassified (i.e., they should have been classified into at least one syndrome category but the text mining tool failed to correctly classify them). The proportion of misclassified reports has been estimated to be low in this data (Lenz et al. 2015). However, the number of misclassified reports should be reduced as much as possible to ensure that data are a valid representation of the pathologists' reports. This is especially important for syndrome classes that are sparse in submission numbers. More standardized procedures for data collection could help to improve the classification performance of the text mining tool and reduce bias in the data. An example is the REPRO syndrome in pigs in January 2005, when non-standardized procedures for data reporting created an artificial peak in the data. We can also imagine that, in the same way, non-standardized reporting could also hide real peaks. Overall, in case a harmonized reporting system would be implemented, these categories and the terminology leading to categorization should be refined to meet all requirements for a Swiss wide surveillance system.

We were able to model the main temporal components of each TS investigated using Poisson or negative binomial distributions. Only 3 TS did not have any temporal pattern and were stationary. The models proposed for cattle were better than those for pigs (i.e., they had no residual autocorrelations). The "bad" models obtained for pigs can be explained by the presence of some large peaks in pig data, especially during the years used for model training. These abnormal peaks make data modelling more complicated. Information on the true epidemic periods were missing in the data used in this study. As an example, the laboratory results obtained from animals necropsied at ITPA were not available. We were thus not able to see if these peaks were part of the normal behavior of the data (e.g., because of the reporting system or some other reason) or, if they were part of a true abnormal event (e.g., an epidemic). If this information were available, it may be possible to manually remove the peaks linked with true abnormal events in order to improve the models. Improving our models is important to ensure good detection performance of aberration detection algorithms. Automatic procedures could have been used to remove abnormal patterns present in a dataset. However, most of the abnormal patterns found in the data were not single peaks but long "aberrations" covering many months or even years and there is currently no automatic procedure available to deal with such long abnormal patterns. We tried to explain the reason behind the long "aberrations" found in the data. The increasing number of submission in cattle between 2006 and 2008 could be linked with the outbreak of bluetongue virus in Europe in 2006. Although few positive BTV cases were found in Switzerland during the BTV outbreak (i.e., 4 cases in 2007, 23 in 2008 and 35 in 2009), the awareness of farmers and veterinarians may have increased and this may explain the increase in the number of cattle submission during this period. Similarly for pigs, the increasing number of URI, LYMPH and OTHER submission between 2004 and 2010 could be linked with the spread of Circovirus and PDNS (porcine dermatitis and nephropathy syndrome) in Switzerland which occurred during this time period. However, due to the lack of additional information, we cannot formally demonstrate that there is a link between abnormal patterns in the ITPA data and the spread of these diseases in the field. As the ITPA primarily serves as an academic institution, variations can easily occur e.g. due to research projects which specifically attract necropsies or even clinical investigations (and thus increased submissions of animals for necropsies) of certain animals or age classes. Expanding our approach to all pathology laboratories in Switzerland would certainly reduce this confounding effect of academic projects. Having better information about the true abnormal events present in the data would be very useful to prove that the data collected by pathology laboratories are sensitive to health changes in Swiss animal population. It would also help to choose the most appropriate detection algorithms. Providing information on true abnormal events occurring in the population is thus critically important in order to draw further conclusions on the feasibility of a real-time disease surveillance and early warning system based on post-mortem data.

CONCLUSION & RECOMMANDATIONS

We believe that the data obtained from professional necropsy reports written in free text can be used for multiple purposes and by different actors. First, pathologists can use them to better understand and define association of pathological lesions with submission trends. These data could thus improve diagnostics approaches and could be also used for educational purposes. Data collected by pathology laboratories can be used for monitoring changes in syndrome and disease prevalence and incidence in a country or even region which are not necessarily be associated with changes in the disease status of the entire population. It could help decision makers at various levels (BLV, state veterinarians, breeding or producer organizations) to design new or adapt existing programs for disease control and prevention, education, etc. Pathology data can also be used at the national level for disease surveillance, specifically for monitoring presence and changes in animal diseases in Switzerland. They could be used to better understand disease distribution in the Swiss animal population (e.g., long trends ongoing in Swiss animal population, seasonality of endemic diseases and help to identify abnormal increases in the number of cases that have not yet been noticed by clinicians, diagnosticians or pathologists. It is important to notice, that we propose that pathology data should be integrated into a national (and potentially international) surveillance system but cannot serve as an independent stand-alone syndromic surveillance system. To ensure full effectiveness, these data have to be combined with all other data used for syndromic surveillance. In addition, even if these data are promising, they are not yet ready to use for an early warning system.

Our main recommendations are:

- Improve data representativeness by adding data from other diagnostic laboratories;
- Based on the text mining tool developed in this project, refine pathology terminology and categorization towards a Swiss wide harmonized reporting system. The syndrome categories should be design according to the objective of surveillance;
- Improve data collection and data standardization by developing more standardized procedures for data entry, such as entering the predefined concerned organ systems / syndrome categories (or using a coding system) and developing a data entry protocol taking into account pathologists and epidemiologists needs;
- Combine these data with clinical data from veterinarians (or other sources) to increase the amount and variety of information generated from surveillance and also to improve the timeliness of event detection;

- Provide more information on true abnormal events occurring on the field (e.g., by extracting the confirmatory diagnosis obtained for animals necropsied at ITPA, combine ITPA data with other information like official disease declaration or suspicion).

Implementing these recommendations will require a broad and very diverse set of skills (e.g., text mining methods, big data management, disease surveillance priorities, abnormal/normal pathological patterns, time series analysis, aberration detection algorithms, etc.). For this reason, we believe that they could not be implemented by a single institution disconnected from the expectations and expertise of other stakeholders. We thus recommend that the next steps for implementing these recommendations should be undertaken in a new collaborative and transdisciplinary project which includes the BLV, pathologists, researchers working on text mining and researchers working in the field of animal health surveillance. Furthermore, it is our opinion that no single Institution involved in this project has the technical or other capacity to operationalize these tools in a working surveillance or other type of knowledge creation system. Therefore, we also recommend that if these tools are operationalized into a working system, it should be done as a multi-Institute collaboration. Such collaboration may have additional advantages such as spreading the infrastructure and human resource requirements across multiple Institutions, making it a more economically feasible project. More importantly transdisciplinary collaborations bring together people with different interests, training and backgrounds and often result in creative new research approaches and creative solutions to technical and societal problems.

The goal of this project was to develop and pilot text mining tools in animal health surveillance in Switzerland. The project was successful in adapting text mining tools to pathology data in a proof of concept. The project clearly demonstrated the value of using pathology data for many purposes and has also opened the door for adapting the text mining tool to many other animal health data written in free text (e.g., clinical data).

REFERENCES

Box, George E P, Gwilym M Jenkins, and Gregory C Reinsel. 2008. *Time Series Analysis: Forecasting & Control*. 4th ed. Hoboken, N.J: Wiley. BOOK. http://tocs.ulb.tu-darmstadt.de/176871241.pdf.

Bozdogan, Hamparsum. 1987. “Model Selection and Akaike’s Information Criterion (AIC): The General Theory and Its Analytical Extensions.” *Psychometrika* 52(3): 345–70. http://link.springer.com/article/10.1007/BF02294361.

Bronner, Anne et al. 2015. “Devising an Indicator to Detect Mid-Term Abortions in Dairy Cattle: A First Step towards Syndromic Surveillance of Abortive Diseases.” 10(3).

Chai, T, and R R Draxler. 2014. “Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against Avoiding RMSE in the Literature.” *Geoscientific Model Development* 7(3): 1247–50. http://www.geosci-model-dev.net/7/1247/2014/.

Dórea, Fernanda C et al. 2013. “Syndromic Surveillance Using Veterinary Laboratory Data : Data Pre-Processing and Algorithm Performance Evaluation.” *Journal of The Royal Society Interface* 10(April): 20130114. http://rsif.royalsocietypublishing.org/content/10/83/20130114.

Farrington, C P, N J Andrews, A D Beale, and M a Catchpole. 1996. “A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease.” *Journal of the Royal Statistical Society. Series A* 159(3): 547–63. http://www.jstor.org/stable/2983331.

Hunter, J Stuart. 1986. “The Exponentially Weighted Moving Average.” *Journal of Quality Technology* 18(4): 203–10.

ISO. 2004. “ISO 8601:2004 Data Elements and Interchange Formats — Information Interchange — Representation of Dates and Times.” https://www.iso.org/obp/ui/#iso:std:iso:8601:ed-3:v1:en.

Lenz, Furrer et al. 2015. “Constructing a Syndromic Terminology Resource for Veterinary Text Mining.” In *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence*,.

Mandl, Kenneth D et al. 2004. “Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience.” *Journal of the American Medical Informatics Association : JAMIA* 11(2): 141–50. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC353021/.

R Development Core Team. 2008. “R: A Language and Environment for Statistical Computing.” http://www.r-project.org.

Ripley, Brian et al. 2015. *MASS: Support Functions and Datasets for Venables and Ripley’s MASS.*

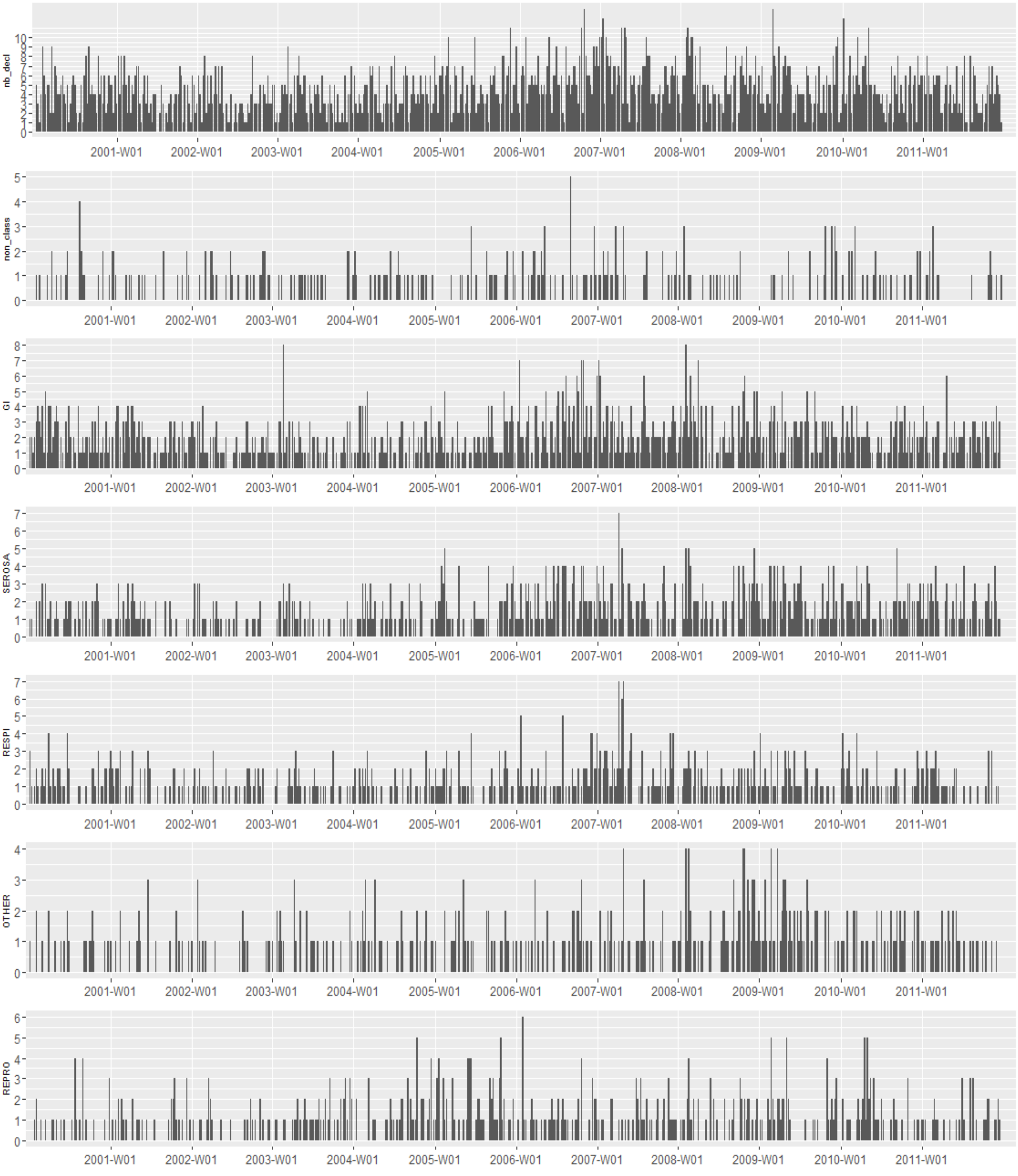
Unkel, Steffen et al. 2012. “Statistical Methods for the Prospective Detection of Infectious Disease Outbreaks: A Review.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175(1): 49–82. http://dx.doi.org/10.1111/j.1467-985X.2011.00714.x\nhttp://onlinelibrary.wiley.com/store/10.1111/j.1467-985X.2011.00714.x/asset/j.1467-985X.2011.00714.x.pdf?v=1&t=i8q04f5b&s=4a2cc2c5a326fc1245c320c065d3325e9f3957e7\nhttp://onlinelibrary.wiley.com/store/10.

APPENDIX

Table 1: Variables tested for each time series and for Poisson and negative binomial distributions

Variable name	Variable description
year	Year considered
week	Week of the year considered
time	Week number according to the total number of weeks available in the dataset
season	Season of the year considered
month	Month of the year considered
mysin	sin(2*pi*(week/53))
mycos	cos(2*pi*(week/53))
period8	Round(week*8/52)
period8shift	Round((3+week)*8/52)
season.shift2	Round((week-4)*4/52)
sinX2	sin (2*pi* week/26.5)
sinX2.shift	sin (2*pi*(week-6)/26.5)
sinminus6	sin (2*pi*(week-6)/53)
sinminus12	sin (2*pi*(week-12)/53)
sinminus18	sin (2*pi*(week-18)/53)
sinX4.shift	sin (2*pi*(week-3)/13.25)
sinX4	sin (2*pi*(week)/13.25)
sinX3.shift	sin (2*pi*(week-4)/18.33)
sinX3	sin (2*pi*(week)/18.33)
histmean	mean of the 53 previous weeks (and guard band of 10 weeks)

Figure 1: number of cattle necropsied in each syndromic category (weekly data). All age classes considered together.



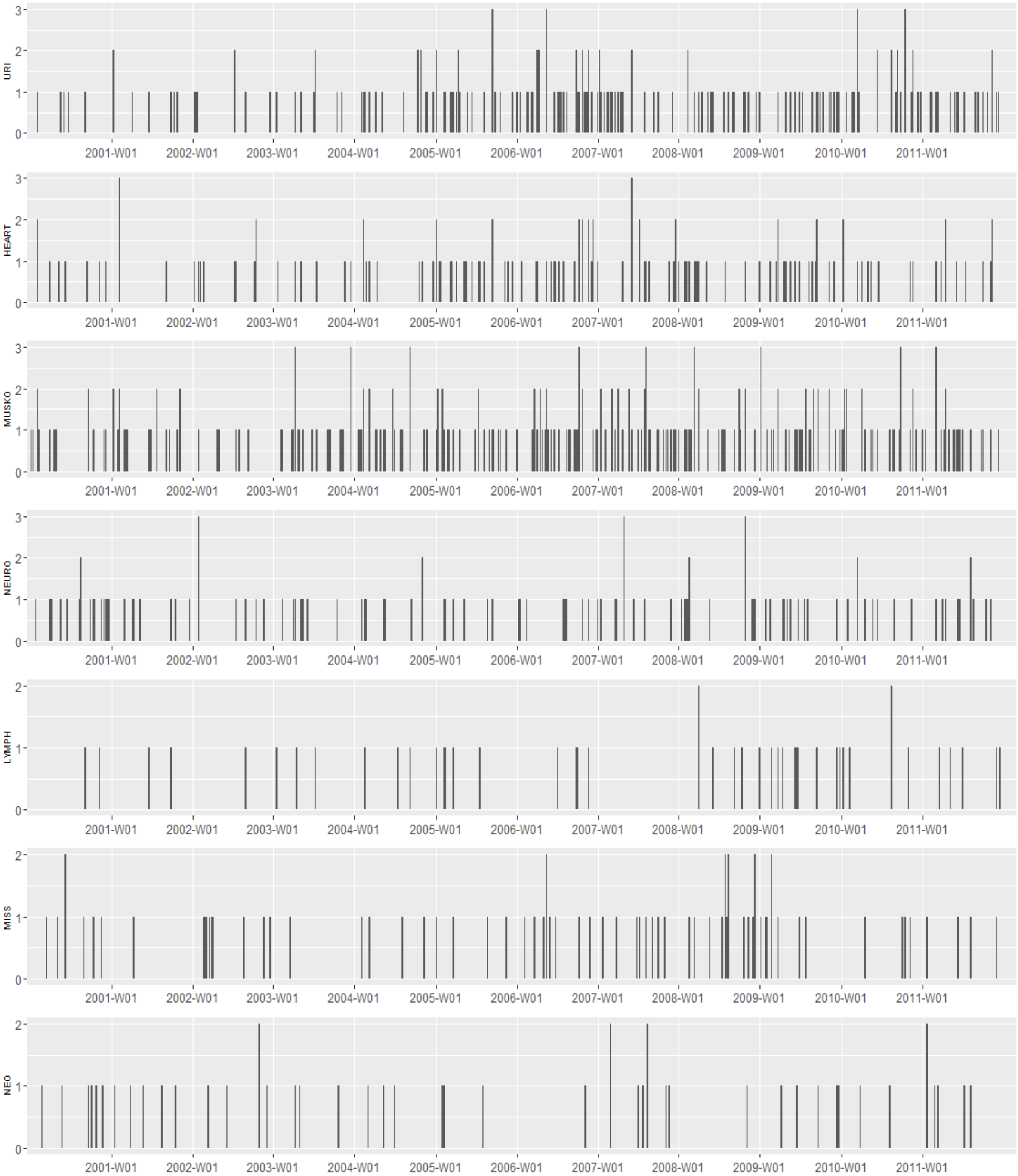
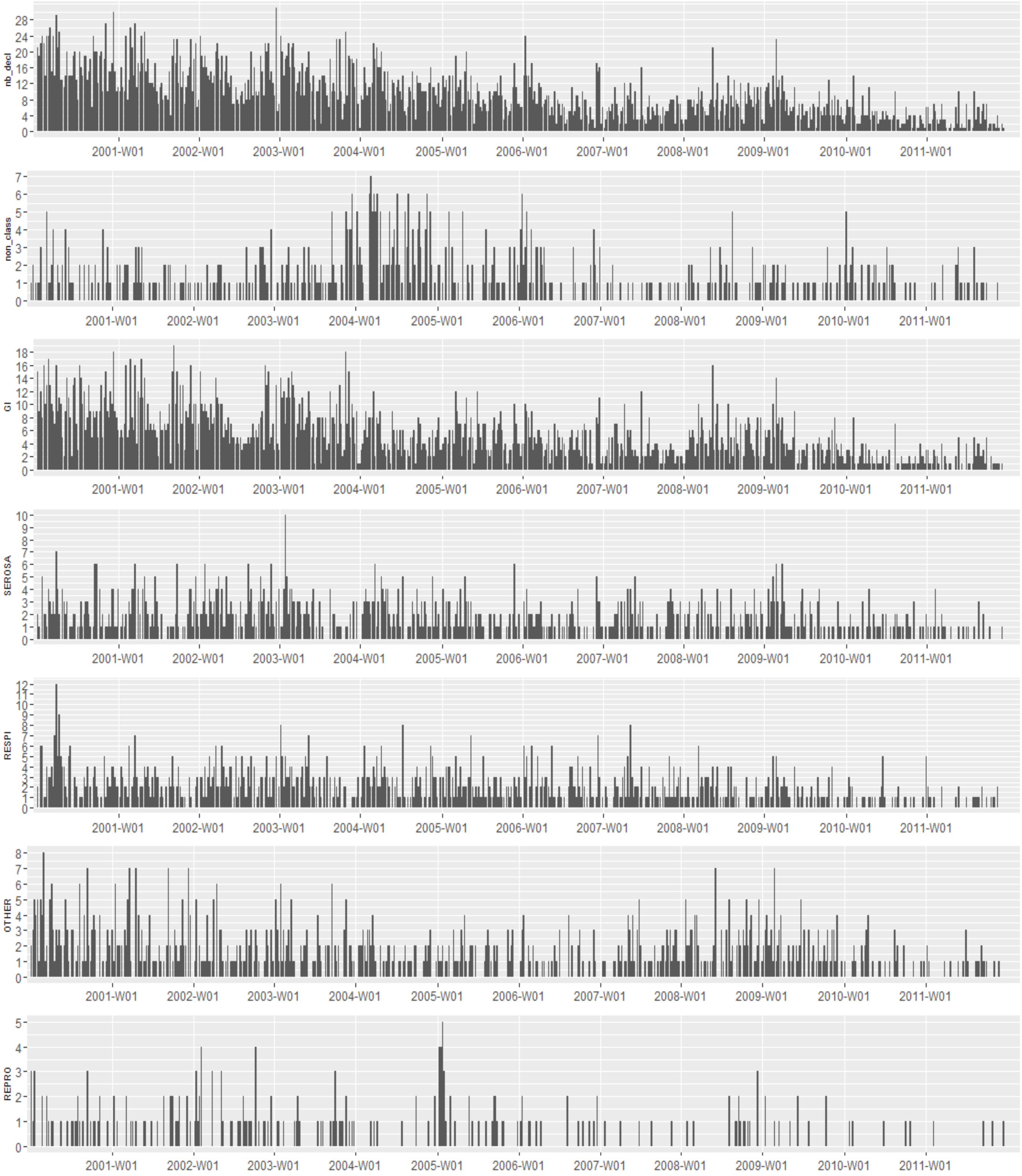

URI
HEART
MUSKO
NEURO
LYMPH
MISS
NEO
0
1
2
3
2001-W01
2002-W01
2003-W01
2004-W01
2005-W01
2006-W01
2007-W01
2008-W01
2009-W01
2010-W01
2011-W01

Figure 2: number of pigs necropsied in each syndrome category (weekly data). All age classes considered together.



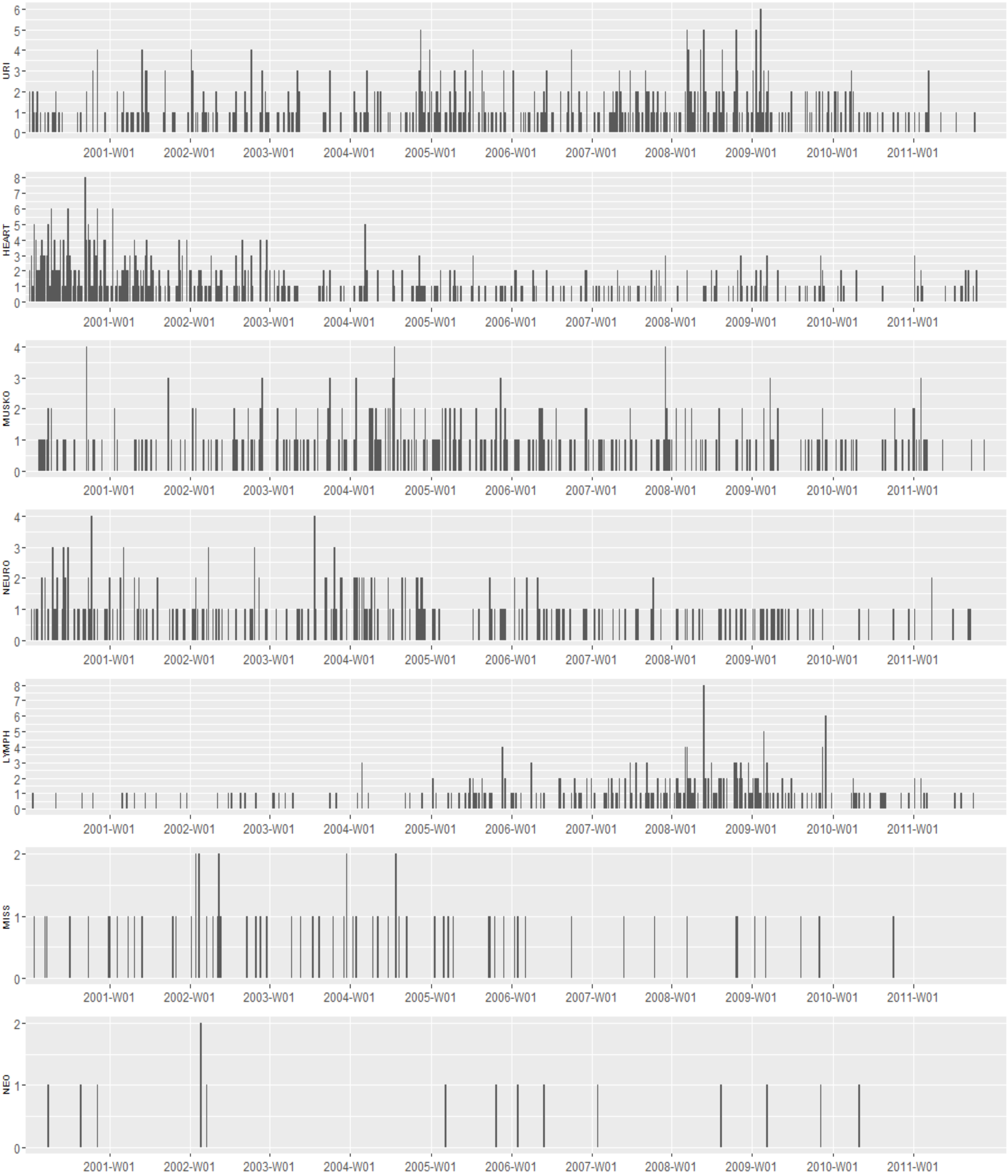
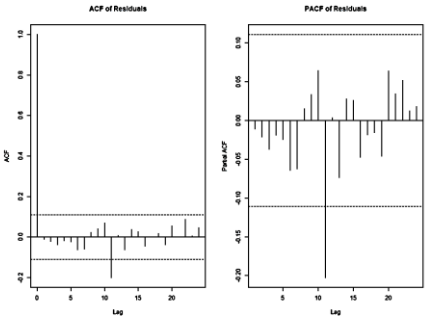
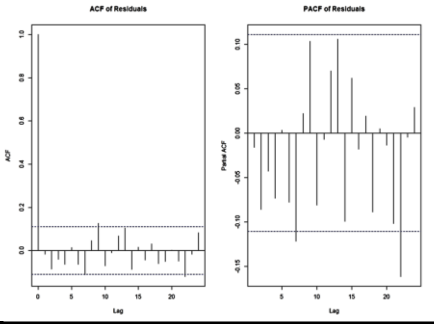
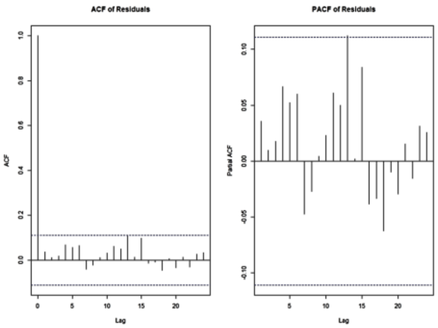
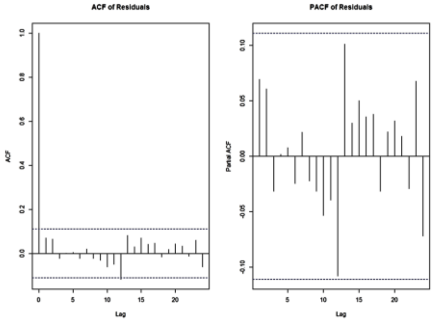
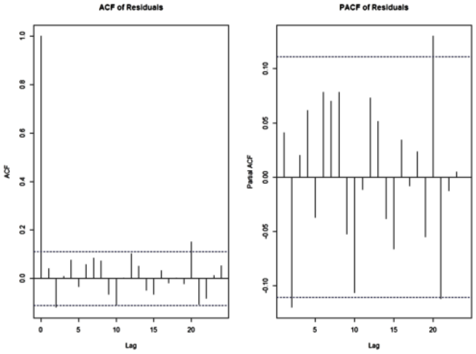

URI
0
1
2
3
4
5
6
HEART
0
1
2
3
4
5
6
7
8
MUSKO
0
1
2
3
4
NEURO
0
1
2
3
4
LYMPH
0
1
2
3
4
5
6
7
8
MISS
0
1
2
NEO
0
1
2
2001-W01
2002-W01
2003-W01
2004-W01
2005-W01
2006-W01
2007-W01
2008-W01
2009-W01
2010-W01
2011-W01

Table 2: Models details for total number of submission in cattle and pigs.

Alternative generalized linear regression models were evaluated on the training data from 2000 to 2005 using the AIC. The best models were then evaluated and compared using the auto-correlation and partial auto-correlation functions of the residuals (ACF and PACF, respectively) and the root-mean-squared error (RMSE).

- *AIC is the Akaike information criterion and, the lower the AIC, the better the model.*
- *RMSE is a measure of the difference between the values predicted by a model and the values actually observed from the environment that is being modeled. This criterion was calculated for the differences between the observations and the predicted values within both the calibration period ($RMSE_c$) from 2000 to 2005 and the validation period ($RMSE_v$) from 2006 to 2011. In either cases, the lower the criterion, the better the predictive performance of the model.*
- *ACF is the linear dependence of a variable on itself at two points in time and PACF is the auto-correlation between two points in time after removing any linear dependence between them. ACF and PACF are used to find repeating patterns (e.g., seasons) in a dataset. When verticals lines go above the significant levels (horizontal lines in ACF and PACF graph), it means that there is significant autocorrelations in the data analyze. If such autocorrelations are present in model residuals, it means that the model is not very good and that some explainable patterns are not well take into account.*

Age classes	model	AIC	$RMSE_c$	$RMSE_v$	Residuals (ACF and PACF)
Calves <6 months	Neg binomial nb_decl ~ mycos+ mysin	960	1.27	1,56	
Young cattle 6 months to 2 years	Poisson nb_decl ~ mysin+ time	494	0,58	0,76	

Adult cattle >2years	Neg binomial nb_decl ~ time + histmean + sinX2.shift	1009	1,32	2,13	
Piglets	Neg binomial nb_decl ~ time + mycos + sinX2.shift	1496	2.91	1.87	
Weaner pigs	Neg Binomial nb_decl ~ time + sinminus18	1298	2.11	1.54	
Adults/fattening pigs	Neg Binomial nb_decl ~ mycos + mysin	1175	1.71	1.73