DTU Library

Dynamic generalized linear models for monitoring endemic diseases
moving beyond univariate process monitoring control algorithms

Lopes Antunes, Ana Carolina; Jensen, Dan; Hisham Beshara Halasa, Tariq; Toft, Nils

Published in:
SVEPM 2016 conference proceedings

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

DYNAMIC GENERALIZED LINEAR MODELS FOR MONITORING ENDEMIC DISEASES: MOVING BEYOND UNIVARIATE PROCESS MONITORING CONTROL ALGORITHMS

A.C. LOPES ANTUNES[*], D. JENSEN, T. HALASA, N. TOFT

SUMMARY

The objective was to use a Dynamic Generalized Linear Model (DGLM) based on a binomial distribution with a linear trend, for monitoring the PRRS (Porcine Reproductive and Respiratory Syndrome sero-prevalence in Danish swine herds. The DGLM was described and its performance for monitoring control and eradication programmes based on changes in PRRS sero-prevalence was explored. Results showed a declining trend in PRRS sero-prevalence between 2007 and 2014 suggesting that Danish herds are slowly eradicating PRRS. The simulation study demonstrated the flexibility of DGLMs in adapting to changes in trends in sero-prevalence. Based on this, it was possible to detect variations in the growth model component. This study is a proof-of-concept, demonstrating the use of DGLMs for monitoring endemic diseases. In addition, the principles stated might be useful in general research on monitoring and surveillance of endemic and (re-)emerging diseases.

INTRODUCTION

New methods for monitoring animal diseases continue to be an active area of research. In the past decade, several studies applied statistical quality control methods for syndromic surveillance in human and veterinary medicine (Buckeridge et al., 2005; Jackson et al., 2007; Dórea et al., 2013). Many of these studies applied univariate process monitoring control algorithms to detect outbreaks of re-emerging diseases. In these cases, control and/or eradication measures are implemented whenever certain threshold levels related to the infection or disease status have been exceeded. However, the term "monitoring" can also be used to describe actions, where a continuous process of collecting data on animal diseases is ongoing, but without any instant control activities (Salman, 2003).

For endemic diseases, it is common to implement control and eradication programmes at herd and regional levels to reduce the economic impact of diseases. Often, these programmes are based on laboratory diagnostics. One example is the Danish monitoring programme for Porcine Reproductive and Respiratory Syndrome (PRRS).

---

[*] Ana Carolina Lopes Antunes, National Veterinary Institute, Technical University of Denmark, Bülowsvej 27, 1870 Frederiksberg C, Denmark. Email: aclan@vet.dtu.dk

Despite disease control efforts in Denmark, PRRS continues to contribute towards the economic losses of the industry since its first diagnosis in 1992. PRRS monitoring is primarily based on serological testing performed on regular basis from herds that have the Specific Pathogen Free System (SPF) certificate (Specific Pathogen Free System (SPF-SuS), 2015). The frequency of testing depends on the SPF herd type, being performed once a month for breeding herds and once a year for finisher herds. The SPF herds represent about 40% of all Danish swine (SPF-SuS, 2015). For non-SPF herds, PRRS diagnostic test are not mandatory and different reasons might explain the variation in frequency of laboratory testing. Thus, diagnostic laboratory submissions of PRRS are collected based on different purposes and frequencies in Denmark.

For disease monitoring, the resulting time series are characterized by observational noise as a result of the variation in the disease prevalence and of the number of samples and herds tested over time. Furthermore, its randomness and non-stationary nature are difficult to model. In these cases, it is necessary to use models with a more dynamic structure, where it is possible to add trends, cyclic patterns and also allow the parameters to change over time. State space models are one possible approach in which relevant prior knowledge and current information are combined. While state space models have been adopted in herd management (Jensen et al., 2015; Madsen & Kristensen, 2005; Ostersen et al., 2010), their use has been underutilized in veterinary sciences for diseases surveillance purposes. In the literature, there are few studies using these type of models for disease monitoring and surveillance in humans (Cao et al., 2014; Cowling et al., 2006).

The objective was to use a state space model for monitoring the PRRS sero-prevalence in Danish swine herds. The binomial DGLM with a linear growth was described and its performance for monitoring control and eradication programmes based on changes in PRRS sero-prevalence was explored. This study is a proof of concept, demonstrating the use of DGLMs for monitoring endemic disease, but the principles stated might also be useful in general research on monitoring and surveillance of endemic and (re-)emerging diseases.

MATERIALS AND METHODS

Data source

Laboratory submission data stored in the National Veterinary Institute – Technical University of Denmark (DTU Vet) information management system and in the Laboratory for Swine Diseases-SEGES Pig Research Centre (VSP-SEGES) were used to determine the weekly PRRS sero-prevalence in Danish swine herds from January 2007 to December 2014.

Each laboratory submission consisted of individual blood samples collected from the same herd on the same day from different animals. Only submissions where at least 2 individual blood samples were tested by serological tests including Blocking Enzyme-Linked Immunosorbent Assay (ELISA) and/or Immunoperoxidase monolayer assay (IPMA) for one or both PRRSV (Porcine Reproductive and Respiratory Syndrome Virus)strains were included in the analysis. These serological tests used were a DTU Vet "in-house" ELISA (Sørensen et. al, 1997) and IPMA (Bøtner et al., 1994). Furthermore, diagnostic test results performed at VSP-SEGES were based on IDEXX PRRS X3 Ab ELISA test (IDEXX, Ludwigsburg, Germany). Results from experimental studies were excluded from the analysis.

Herds were classified as PRRS sero-positive when at least 2 individual blood samples in each submission tested PRRS positive, independently of the PRRS strain. The between-herd PRRS sero-prevalence was calculated weekly as the proportion of PRRS positive herds from the total number of herds tested for PRRS.

<u>Modelling</u>

A binomial DGLM with a linear growth as described by West and Harrison (1997) was used to model the data. The general purpose of the DGLM is to estimate the underlying parameter vector from the observed data (θ) combined with any prior information available at time 0 (D_0) before any observation is made. This can be achieved sequentially where the estimated value is updated each time a new value (PRRS sero-prevalence) is obtained. In this case, the conditional distribution of θ_t given by D_t ($\theta_t|D_t$) was estimated. These models can be used to estimate a one-step forecast of the mean, allowing for a comparison with the actual observed PRRS sero-prevalence. Moreover, the linear growth component includes a time-varying slope (or local linear trend), allowing the system to adapt to a possible positive or negative growth for each t.

In general, the DGLM consists of an observation equation (Eq. 1) and a system equation (Eq. 2):

$$g(p_t) = F_t'\theta_t \tag{1}$$

$$\theta_t = G_t\theta_{t-1} + W_t \tag{2}$$

Equation 1 describes how the values of an observation (PRRS sero-prevalence) derive from $g(p_t)$, depends on an unobservable parameter vector (θ) for time t based on a linear function. For the model specification, $g()$ is the identity function. Equation 2 describes the dynamic properties of the unobservable parameter vector θ. In this study, the transposed design matrix (F_t') has the structure presented in Table 1, in order to estimate underlying values of PRRSV sero-prevalence according to Eq 1. The system matrix (G) used to update the mean of the PRRSV sero-prevalence for each time step taking into account the trend. Both matrix structures were constant for each t (week). The variance-covariance matrix (W_t) describes the evolution of variance and covariance of each parameter for each time step. Rather than estimating (W_t), the system variance was modelled using a discount factor (see Eq. 4).

<p align="center">Table 1. Matrices structure used in Eq. 1 and 2.</p>

F_t'	G_t
$[1 \quad 0]$	$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$

The DGLM update for each time step t was performed as follows:

a) the posterior distribution for θ_{t-1} was expressed by a prior mean (m_{t-1}) and a variance (C_{t-1}), $(\theta_{t-1} \mid D_{t-1}) \sim [m_{t-1}, C_{t-1}]$;

b) the prior distribution for θ_t $(\theta_t \mid D_{t-1}) \sim [a_t, R_t]$ was made based on the prior mean (a_t) and prior variance (R_t) which were calculated as described in Eq. 3 and 4. The specification of the variance components was specified using a discount factor (δ);

$$a_t = G_t m_{t-1} \tag{3}$$

$$R_t = \frac{1}{\delta} G_t C_{t-1} G_t' \tag{4}$$

c) the prior distribution for Y_t $(Y_t \mid D_{t-1}) \sim [f_t, q_t]$ was calculated based on the forecast mean (f_t) and forecast variance (q_t) (Eq. 5 and 6);

$$f_t = F_t' a_t \tag{5}$$

$$q_t = F_t' R_t F_t \tag{6}$$

d) the posterior mean (f_t^*) and variance (q_t^*) were calculated as described in Eq. 7 and 8. In this case, it was assumed that the prior probability p (PRRS sero-prevalence) of a binomial distribution was Beta(α, β). If κ successes (PRRS positive herds) out of n trials (number of herds tested for PRRS) were observed, the posterior p, given the new observation was Beta$(\alpha_t + \kappa_t, \beta_t + n_t - \kappa_t)$. The parameters α_t and β_t were calculated according to Eq. 9 and 10.

$$f_t^* = \frac{\alpha_t + \kappa_t}{\alpha_t + \beta_t + n_t} \tag{7}$$

$$q_t^* = \frac{f_t^*(1 - f_t^*)}{\alpha_t + \beta_t + n_t + 1} \tag{8}$$

$$\alpha_t = f_t \left(\frac{f_t(1 - f_t)}{q_t} - 1 \right) \tag{9}$$

$$\beta_t = (1 - f_t) \left(\frac{f_t(1 - f_t)}{q_t} - 1 \right) \tag{10}$$

e) the posterior distribution for θ_{t-1} in a) was calculated based on its mean matrix m_t and its variance-covariance matrix C_t as demonstrated in Eq. 11 and 12.

$$m_t = a_t + R_t F_t (f_t^* - f_t)/q_t \tag{11}$$

$$C_t = R_t - R_t F_t F_t' R_t (1 - q_t^*/q_t)/q_t \tag{12}$$

Model initialization: Reference analysis was used to estimate the initial parameters $D_0 \sim [m_0, C_0]$ as described by West and Harrison (1997). We defined the matrices K_t and H_t and the vectors k_t and h_t for the first two observations $p_{1:2}$.

For $t = 1$, the initial parameters were defined as $H_1 = \underline{0}$, $h_1 = \underline{0}$, $K_1 = H_1 + F_1 F_1'$ and $k_1 = h_1 + F_1 p_1$. For $t = 2$, the vectors and matrices were updated as described in Eq. 13 to Eq. 16.

$$H_2 = G_2^{-1'} K_1 G_2^{-1} \tag{13}$$

$$h_2 = G_2^{-1'} k_1 \tag{14}$$

$$K_2 = H_2 + F_2 F_2' \tag{15}$$

$$k_2 = h_2 + F_2 p_2 \tag{16}$$

Then, the prior distribution for $t = 3$ was calculated according to Eq. 17 and 18.

$$m_2 = K_2^{-1} k_2 \tag{17}$$

$$C_2 = K_2^{-1} \tag{18}$$

System variance: The DGLM model was run based on different discount factors (δ) ranging from 0.1 up to 1 by increments of 0.01. The discount factor which minimized the sum of the squared forecast errors based on the first two years of the data was chosen for the analysis.

Monitoring model components: The values obtained from the m vector for each time step t were used to decompose the time series and obtain the model growth (PRRS sero-prevalence trend). The variance on the growth parameter was calculated from the C matrix and used to calculate 95% confidence intervals (CI).

Simulated scenarios: PRRS sero-prevalence baseline was simulated for 8 years, in which the number of positive herds (X) per week was drawn from a binomial distribution ($X \sim bin(\text{n}, \text{p})$) with a probability (p) (PRRS sero-prevalence) and a sample size (n) equal to the number of Danish herds tested for PRRS per week between 2007 and 2014. The weekly sero-prevalence was calculated as the simulated number of sero-positive herds divided by the total weekly number of herds tested. The first 104 weeks were simulated with a constant initial prevalence of 0.24, corresponding to the average PRRS sero-prevalence in Danish herds observed based on the laboratory diagnostic data from 2007 to 2014. In the first scenario (Scenario A), a constant decrease from p=0.24 to p=0.10 during 4 years followed by constant sero-prevalence was simulated. The second scenario (Scenario B) represented a decrease in the sero-prevalence from p=0.24 to p=0.10 during 2 years, followed by an increase to p=0.18 during the subsequent 2 years.

The sensitivity (Se) and timeliness were used to evaluate the performance of the DGLM to detect significant changes in the simulated scenarios. The Se was defined as the proportion of

simulations in which significant changes in the model growth component from zero were found. Timeliness was defined as the number of weeks between a change in the PRRS sero-prevalence (decrease, increase, constant) was simulated and detected.

Convergence rate: A total of 20,000 simulations of weekly PRRS sero-prevalence with a constant decrease from 0.24 to 0.05 over 5 years were carried out. The number of iterations needed to reach a stable variance in the average time to detect significant changes (convergence) was determined visually by plotting the variance of the average timeliness with a stepwise increase of 100 iterations up to 20,000 iterations against the number of iterations. Stable results were observed when using only 10,000 iterations and hence all further simulations were run with 10,000 iterations.

All analyses were performed using R (version 3.1.1) (R Core Team, 2014).

RESULTS

Data description

A total of 56,341 laboratory submissions from 5,390 Danish swine were included in the analysis. The average weekly number of herds tested for PRRSV was 130 (min=9, max=206); the mean weekly number of PRRS positive herds was 31 herds (min=0, max= 60). The weekly average PRRS sero-prevalence was 0.24 (min=0, max=0.38). The yearly average of PRRS sero-prevalence declined from 0.28 in 2007 to 0.20 in 2014, with an average decrease of 0.01 per year.

Model initialization and discount factor

Table 2 shows the posterior C_2 and m_2 matrices obtained from the reference analysis and used as priors for the DGLM model for $t = 3$. The discount factor which minimized the sum of forecast errors for the data was δ=0.98.

Table 2. Priors for $t = 3$ obtained from the reference analysis.

m_2	C_2
$\begin{bmatrix} 0.30 \\ 0.07 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

Modelling and decomposing DGLM

Results show a declining trend of PRRS sero-prevalence between 2007 and 2014. Significant decreases (95% CI excluding zero) were detected mainly in the last 6 months of 2007; end of 2008 to the first semester of 2010 and from the last quarter of 2010 until the beginning of 2013 (Fig. 1). No significant increases in PRRS sero-prevalence were observed and all values for the growth component were below 0.
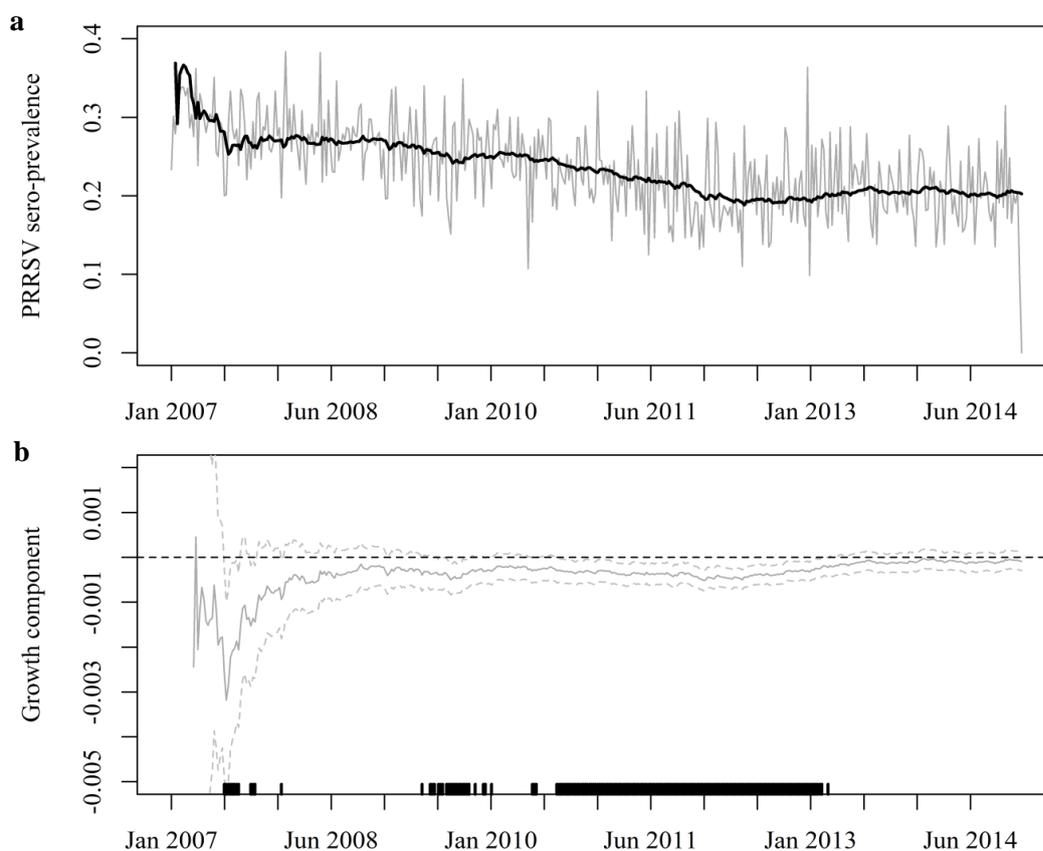
Fig. 1 Using a DGLM to monitor PRRS sero-prevalence in Danish swine herds from 2007 to 2014. Results show the weekly PRRS sero-prevalence and the filtered mean (black) (a) and the corresponding DGLM growth component (b). The black rugs indicate were the growth component is significantly different from zero.

Simulated scenarios

The simulated scenarios are represented in Fig. 2 and 3. The results for the simulation study are presented in Table 3. Significant changes in the model growth component from zero were found in both scenarios. However, the DGLM detected changes in the growth with a higher sensitivity for decreasing changes when compared to constant growth in the time series. The lowest sensitivity was found for Scenario A when the PRRS sero-prevalence became constant after the decrease, with the DGLM growth component being non-significantly different from zero in 39.02% of the simulations.

Table 3. Timeliness (weeks) and Se for the simulated scenarios.

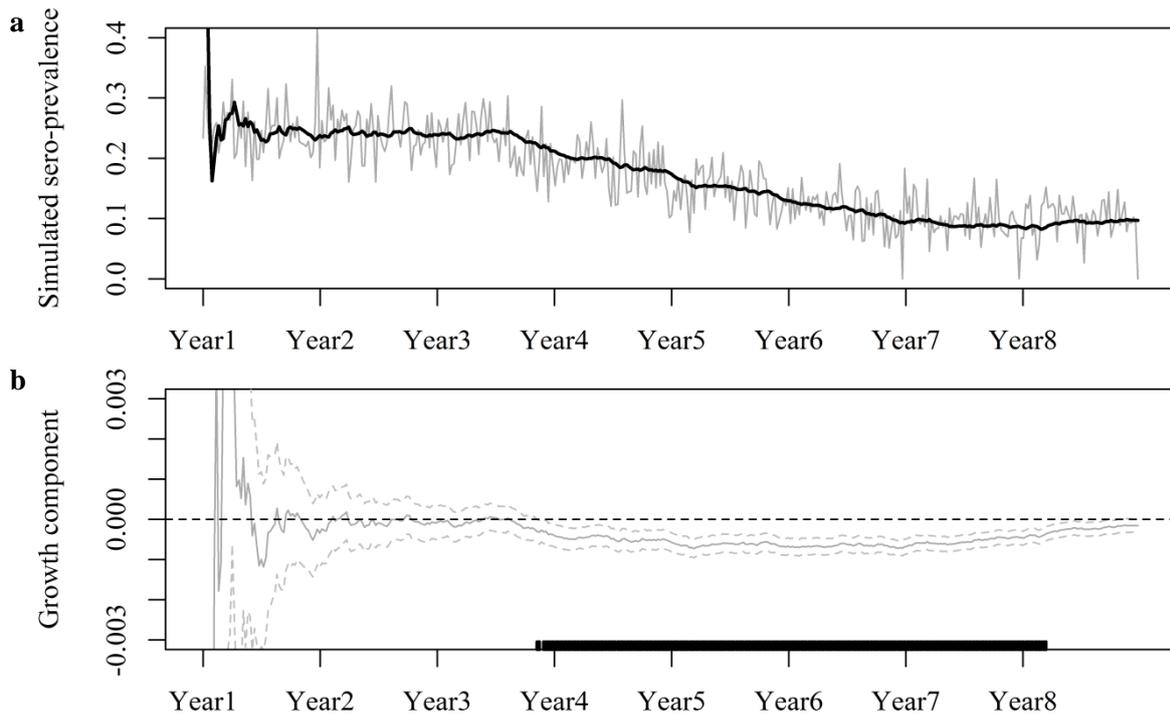|	Scenario A		Scenario B	
Intervention	Decrease	Constant	Decrease	Increase
Timeliness (median)	47	96	27	146
(min-max)	(0-89)	(57-106)	(0-56)	(110-257)
Se (%)	100	39.02	100	99.64

Fig. 2 Simulated control program Scenario A. PRRS sero-prevalence was constant during 104 weeks, followed by a decrease to 0.10 during 208 weeks and then a constant prevalence. The DGLM filtered mean (black line) (a) and the corresponding DGLM growth component (b) (grey lines) are presented. The black rugs indicate a significant negative the growth component based on 95% CI.
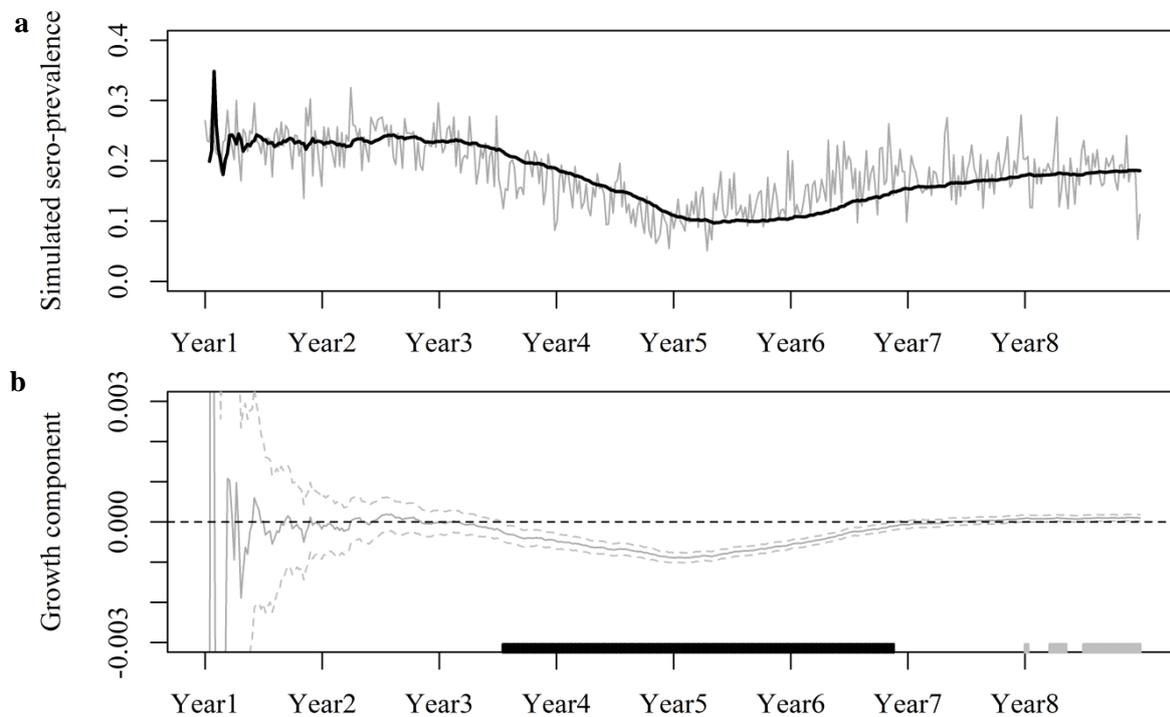

Fig. 3 Simulated control program Scenario B. PRRS sero-prevalence was constant during 104 weeks, followed by a decrease to 0.10 during 104 weeks and an increase up to 0.18 during 104 weeks. The DGLM filtered mean (black line) (a) and the corresponding DGLM growth component (b) (grey lines) are presented. The black and grey rugs indicate significant declines and increase in the growth component based on 95% CI, respectively.

DISCUSSION

The objective of this study was to use a binomial DGLM with a linear growth component for monitoring PRRS sero-prevalence in similar contexts to the Danish Pig Industry. The same model can be used for monitoring other prevalence data. These types of models can also be derived for Poisson distribution for monitoring count data, such as the number of samples submitted for analysis etc. Moreover, an ordinary Dynamic Linear Model (DLM) can be used if the data are normally distributed. They also allow for modelling interventions as well as changes in level shift through multi-process models (Thysen, 1993). The DGLM provide a flexible framework in which it is possible to include different data sources in a multivariate process as shown by (Jensen et al., 2015). Moreover, the use of this method allows monitoring of trends and also other components of time series such as seasonal, regression and autoregressive effects components which have a wide interest in biomedical time series applications (West & Harrison, 1997).

As no information on PRRS outbreaks and eradication programmes is available for Danish swine herds, a simulation study was conducted. One limitation of this study is related to the simulation approach used; the simulated sero-prevalence was based on a binomial distribution. The variation in the number of herds tested had an impact on the simulated prevalence contributing to the variation (noise in the baseline). As a consequence, the timeliness to detect interventions showed a wide range of values and the sensitivity was not similar for all interventions. One approach to overcome this issue could be to aggregate the data on a monthly basis, thus reducing the noise in the baseline and possibly improve the performance of the model to adapt to changes in the trend.

The DGLM model was able to detect changes in both scenarios. However, it is important to notice that decreases were larger compared to the increases, corresponding to an absolute decay in sero-prevalence of 0.145 and absolute increase of 0.08. For scenario B, significant positive changes in the model growth component were found after a period in which non-significant changes were found. These justify the longer time needed to detect increases. The variation in the growth parameter was monitored based on 95% CI's. Different approaches could be, *e.g.* Shewart control charts, cumulative sensitivities, V-mask (Montgomery, 2013) or target values, which might yield improved the performances.

In a Bayesian framework the choice of priors is critical for making inference. Reference analysis was used to initiate the DGLM model. From a practical point of view, when a system is set up, the number of observations is low to make the influence of the priors significant. In this case, the use of "non-informative" priors can be used. This method offers an easily applied default analysis (West & Harrison, 1997) when running a DGLM. However, it can be seen from the simulated scenarios that the DGLM takes 3 months to adapt to the data. For this reason, it is important to have historical data (retrospective analysis) to train the model when setting up a monitoring system.

The systems variance was defined based on a discount factor, expressing the decay of information in the system. Defining δ=0.98 implies a small systems variance with a very slow adaptation to new observations. This value was defined using the same method described in Kristensen et al. (2010), where δ should optimized for the performance of the model in making forecasts, *i.e.,* minimizing the forecast errors for the first two years of data (retrospective analysis). In recent literature (Bono et al., 2012; Jensen et al., 2015), the Expectation-Maximization algorithm (Dempster et al., 1977) was used to define the W variance-covariance matrix. This approach offers a general approach to iterative computation

of maximum-likelihood estimates when the observations can be viewed as incomplete data. The use of a discount factor provides a parsimonious approach when compared to the full estimation of W.

In summary, results show a declining trend on PRRS sero-prevalence between 2007 and 2014 suggesting more Danish herds are eradicating PRRS. The simulation study highlighted that DGLM are flexible models able to adapt to changes in the time series. It was possible to detect variations in the growth component of simulated scenarios. This study is a proof of concept, demonstrating the use of DGLMs for monitoring endemic disease, but the principles stated might also be useful in general modelling, monitoring and surveillance of (re)emerging diseases. Further analysis to compare the performance of the DGLM, including different components, to other models will be investigated in future studies.

ACKNOWLEDGEMENTS

REFERENCES

Bono, C., Cornou, C. and Kristensen, A.R. (2012). Dynamic production monitoring in pig herds I: Modeling and monitoring litter size at herd and sow level. Livestock Science 149, 289–300. doi:10.1016/j.livsci.2012.07.023

Buckeridge, D. L., Switzer, P., Owens, D., Siegrist, D., Pavlin, J. and Musen, M. (2005). An evaluation model for syndromic surveillance: assessing the performance of a temporal algorithm. MMWR. Morbidity and Mortality Weekly Report 54 Suppl, 109–115. doi:su5401a18

Bøtner, A., Nielsen, J. and Bille-Hansen, V. (1994). Isolation of porcine reproductive and respiratory syndrome (PRRS) virus in a Danish swine herd and experimental infection of pregnant gilts with the virus. Veterinary Microbiology 40(3-4), 351–360. doi:10.1016/0378-1135(94)90122-8

Cao, P.-H., Wang, X., Fang, S.-S., Cheng, X.-W., Chan, K.-P., Wang, X.-L., Lu, X.,Wu, C.-L.,Tang, X.-J., Zhang, R.-L., Ma, H.-W., Cheng, J.-Q., Wong, C.-M. and Yang, L. (2014). Forecasting influenza epidemics from multi-stream surveillance data in a subtropical city of China. PloS One 9(3), e92945. doi:10.1371/journal.pone.0092945

Cowling, B.J., Wong, I. O.L., Ho, L.-M., Riley, S. and Leung, G.M. (2006). Methods for monitoring influenza surveillance data. International Journal of Epidemiology 35(5), 1314–1321. doi:10.1093/ije/dyl162

Dempster, A.P.A., Laird, N.M.N. and Rubin, D.D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B Methodological, 39(1), 1–38. doi:http://dx.doi.org/10.2307/2984875

Dórea, F.C., McEwen, B.J., Mcnab, W.B., Revie, C.W. and Sanchez, J. (2013). Syndromic surveillance using veterinary laboratory data: data pre-processing and algorithm

performance evaluation. Journal of the Royal Society, Interface / the Royal Society 10(83), 20130114. doi:10.1098/rsif.2013.0114

Dupuy, C., Morignat, E., Dorea, F., Ducrot, C., Calavas, D. and Gay, E. (2015). Pilot simulation study using meat inspection data for syndromic surveillance: use of whole carcass condemnation of adult cattle to assess the performance of several algorithms for outbreak detection. Epidemiology and Infection 143(12), 2559–2569. doi:10.1017/S0950268814003495

Jackson, M.L., Baer, A., Painter, I. and Duchin, J. (2007). A simulation study comparing aberration detection algorithms for syndromic surveillance. BMC Medical Informatics and Decision Making 7(1), 6. doi:10.1186/1472-6947-7-6

Jensen, D.B., Cornou, C., Toft, N. and Kristensen, A.R. (2015). A multi-dimensional dynamic linear model for monitoring slaughter pig production. In 7th European Conference on Precision Livestock Farming (pp. 503–512). Milan. doi:10.1017/CBO9781107415324.004

Kristensen, A.R., Jørgensen, E. and Toft, N. (2010). Herd Management Science II. Advanced topics. p92

Madsen, T.N. and Kristensen, A.R. (2005). A model for monitoring the condition of young pigs by their drinking behaviour. Computers and Electronics in Agriculture 48(2), 138–154. doi:10.1016/j.compag.2005.02.014

Montgomery, D. (2013). Introduction to statistical quality control. (7th Ed.). Arizona State University: John Wiley and Sons, Inc. 234-276,416-447p

Ostersen, T., Cornou, C. and Kristensen, R. (2010). Detecting oestrus by monitoring sows' visits to a boar. Computers and Electronics in Agriculture 74(1), 51–58. doi:10.1016/j.compag.2010.06.003

R Core Team. (2014). R. A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org

Specific Pathogen Free System (SPF-SuS). (2015). Specific Pathogen Free System. Retrieved May 12, 2015, from https://www.spf.dk/

Sørensen, K. J., Bøtner, A., Smedegaard Madsen, E., Strandbygaard, B. and Nielsen, J. (1997). Evaluation of a blocking Elisa for screening of antibodies against porcine reproductive and respiratory syndrome (PRRS) virus. Veterinary Microbiology 56(1-2), 1–8. doi:10.1016/S0378-1135(96)01345-4

Thysen, I. (1993). Monitoring Bulk Tank Somatic Cell Counts by a Multi-Process Kalman Filter. Acta Agriculture Scandinavia (43), 58 – 64

West, M. and Harrison, J. (1997). Bayesian Forecasting and Dynamic Models (2nd Ed.). New York, USA: Springer. 136-137,p516-560