DEVELOPMENT OF NEW STRATEGIES TO MODEL BOVINE FALLEN STOCK DATA FROM LARGE AND SMALL SUBPOPULATIONS FOR SYNDROMIC SURVEILLANCE USE *

Ana Alba-Casals[1,2], **Amanda Fernández-Fontelo**[3], **Crawford W. Revie**[4], **Fernanda C. Dórea**[5], **Javier Sánchez**[4], **Luis Romero**[6], **Germán Cáceres**[6], **Andrés Pérez**[1] and **Pere Puig**[3]

ABSTRACT

The continuous monitoring of fallen stock mortality in bovine farms has been demonstrated in different studies to have potential as an important component of veterinary syndromic surveillance. However, as far as we know, the usefulness of these systems to detect abnormal events in near real-time in the field has not been assessed. To implement this type of system, a number of challenges must be faced. The main difficulties are associated with the non-specific nature of fallen stock data, since multiple events may cause bovine mortality at farm level. Moreover, these data are originated from heterogeneous subpopulations that can be clustered and studied in accordance with different traits (e.g. production type, type of farm and/or individuals, husbandry and environmental conditions, or administrative level).

In this study, we present the main pillars of a syndromic system to collect continuous fallen stock data from a specific region and to model time series and detect abnormal events at large and small scale.

Keywords: Syndromic surveillance, Cattle, Fallen stock, Modelling, ARIMA, INAR.

RESUME

La mortalité enregistrée dans des élevages bovins a démontré dans différentes études avoir un potentiel important comme une composante de surveillance syndromique. Pourtant, l'usage de ces systèmes pour découvrir les événements anormaux en temps proche du réel n'a pas été évalué en pratique. Pour mettre en place ce type de système, un certain nombre de défis doivent être affrontés. Les difficultés principales sont associées à la nature non-spécifique de données, puisque de multiples événements peuvent provoquer de la mortalité bovine dans une ferme. De plus, ces données sont récoltées à partir de sous-populations hétérogènes qui peuvent être groupées et étudiées conformément à différents traits (par ex. le type de production, le type de ferme et/ou d'individus, les conditions d'élevage et de l'environnement, ou le niveau administratif).

.../..

---

* Texte de la communication orale présentée au cours de la Journée scientifique AEEMA, 20 mars 2015

[1] Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, St Paul, MN, USA
[2] Centre de Recerca en Sanitat Animal (CReSA), Fundació UAB-IRTA, 08193 Bellaterra, Barcelona, Spain
[3] Departament de Matemàtiques, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain
[4] Centre for Veterinary Epidemiological Research, AVC. University Prince Edward Island (UPEI), Canada
[5] Department of Disease Control and Epidemiology. National Veterinary Institute (SVA), Sweden
[6] MAGRAMA. Ministerio de Agricultura, Alimentación y Medio Ambiente, Spain

…/..

Dans cette étude, nous présentons les éléments essentiels d'un système syndromique pour recueillir des données de mortalité bovine dans une région, modéliser des séries temporelles et détecter des évènements anormaux à grande et petite échelle.

Mots-clés : surveillance syndromique, bovin, mortalité, modélisation, ARIMA, INAR.

I - INTRODUCTION

Fallen stock mortality in bovine farms has shown its potential as an important component of veterinary syndromic surveillance in different studies [Perrin *et al*., 2010, 2012; Dupuy *et al*., 2013; Alba *et al*., 2015]. The continuous monitoring of bovine fallen stock data for syndromic surveillance may not only be used for the early detection of abnormal events, but may also serve as an indicator of cattle health at the population level. In addition, this kind of system can assess the impact of events occurred throughout a time period and its implementation may help to substantiate freedom from disease, thus increasing the confidence of trading partners. Moreover, its implementation provides valuable information for risk based surveillance.

However, for such a system to achieve efficient operation, a number of challenges must be faced. Some difficulties are associated with the non-specific nature of fallen stock data, since multiple events may cause bovine mortality at farm level. Furthermore, fallen stock data originate from dynamic and heterogeneous subpopulations that can be clustered and studied in accordance with different traits (e.g. production type, type of farm and/or individuals, husbandry and environmental conditions, or administrative level) and this adds complexity to the analysis [Perrin *et al*., 2010, 2012; Alba *et al*., 2015]. Moreover, the analysis of fallen stock at large-scale is able to detect events which happen across a wide geographical region. Nevertheless, to put in place efficient preventive and control measures the decision making process has to be frequently conducted at smaller spatial scale. For example, if an abnormal event occurs in a small region, the preventive or control measures should initially be allocated in this specific subpopulation, and thus it is important to detect early this event at this level. Accordingly, to build an efficient surveillance system, the data analysis should be conducted in parallel at different grouping and geographical levels.

This paper proposes a novel strategy for the design of a syndromic surveillance system based on routinely collected data on fallen cattle, modelling and comparing time series at large and small scales.

II - MATERIAL AND METHODS

The system develops the following processes:

1. Design of a data warehouse to integrate continuous data registered from bovine populations and associated fallen stock;
2. Preliminary exploratory and descriptive analyses;
3. Analyses of time series at regional and provincial levels for different bovine production types, using classical time series models such as autoregressive integrated moving average models;

4. Analyses and plotting of hierarchical time series to explore simultaneously the mortality patterns at different geographical levels; and, finally,
5. Analyses of time series at finer spatial scales (such as county or municipality) based on integer-valued autoregressive time series models of different orders.

1. DATA USED

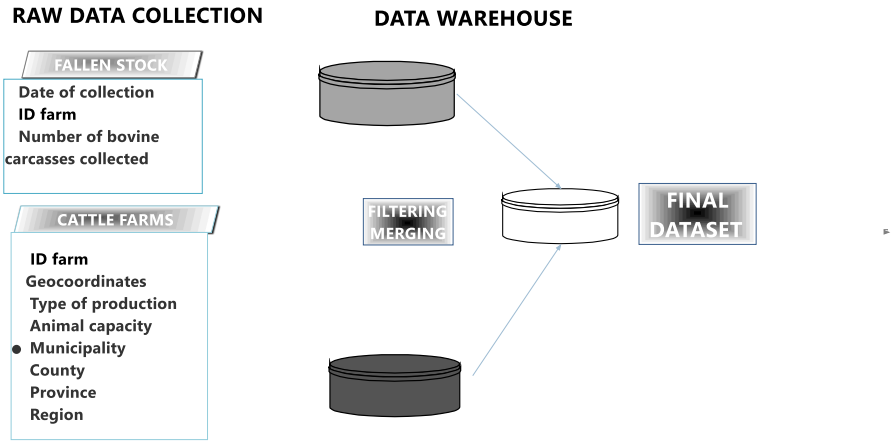
The system incorporates two types of raw data: cattle farm demographics and cattle fallen stock data. In the system used as an example, both types of data are gathered under an agreement of strict confidentiality from the official animal health services. The cattle population (updated yearly) is recorded in an Excel spreadsheet and contains the following variables: a unique identifier for each farm, its type of production, its animal capacity, and its location defined in terms of region, province, county and municipality. The data relating to fallen stock is stored in an Access database which records the following fields: unique identifier for each farm, date of collection, and number of bovine carcasses collected.

2. DATA WAREHOUSE STRUCTURE

Figure 1 shows a scheme of the basic structure of this data warehouse system.

Figure 1

Structure of data warehouse system



This structure is built using the base package of the software R [R Core Team, 2003] with *R Studio* as an integrated development environment [Van der Loo *et al*., 2012]. Two functions are combined to automate the processes. One function allows the importation and integration of the bovine population and fallen stock data into a unique dataset; while the other function filters, depurates and merges the data.

3. EXPLORATORY and DESCRIPTIVE ANALYSES

An initial exploratory analysis is conducted to identify a robust indicator of the cattle mortality from the available data. Moreover, this analysis aims to:

1. Describe the bovine population coverage by the system,
2. Extract the overall statistics of the bovine population and fallen stock for each production type at different spatial levels,
3. Identify those farms that will be part of the system, and
4. Represent the time series plots at larger scales.

In the current system it is proposed that, "the number of bovine carcasses collected by week for each production type at different spatial scale (i.e. region, province, county and municipality)" be selected as indicators of cattle mortality. In a previous study the same temporal and spatial units for each production type were used while the outcomes were the number of carcass disposal

visits conducted and the kilograms of carcasses collected at the farm level [Alba *et al*., 2015]

4. AUTOREGRESSIVE INTEGRATED MOVING AVERAGE MODELS (ARIMA) TO BE USED AT LARGE SCALE

ARIMA models have been demonstrated to provide good solutions to explore, predict events and provide evidence for the existence of irregularities in those time series that show regular patterns without including zeroes in the observations [Chatfield *et al*., 2004; Benshop *et al*., 2008; Cowpertwait *et al*., 2009; Alba *et al*., 2015]. In this sense, these models aim to identify and compare the historical baselines of bovine fallen stock for the main types of production at large spatial scales (e.g. at national, regional or province level). Moreover, using a previous adjustment of the model, it is easy to characterize the seasonality and trend of the series. This method, explained in more detail in Alba *et al*. [2015], comprises several steps. Firstly, the time series to be analysed is divided in two parts. One part of the data is used as training data to fit the model, and the rest is used as test data to validate the model. Secondly, to facilitate the fitting of the model, the possible seasonal patterns and/or trends are studied and adjusted in the training data by the least squares method using a multiple linear regression model. The coefficients of the seasonality are defined as trigonometric covariates in which the frequency ω is expressed as ω=2π/T and T is the value of the seasonal periodicity. This eventual trend is initially modeled as a linear pattern. The overall set of observations is expressed as X_t (1)

$$X_t = \mu + \alpha\cos(\omega t) + \beta\sin(\omega t) + \cdots + \delta(t) + Y_t \qquad (1)$$

X_t is composed by eventual seasonal components, expressed as trigonometric covariates such as $\alpha\cos(\omega t)$ or/and $\sin(\omega t)$, an eventual trend component $\delta(t)$, and Y_t that corresponds to the remaining ARIMA model. Thirdly, the respective orders of the components of the autoregressive/moving average processes are identified. The respective orders are determined based on corrected Akaike Information Criteria (AIC) and a diagnostic check that assesses the lack of autocorrelation and partial autocorrelation between the standardized residuals obtained. The remaining model Y_t is expressed as (2):

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \cdots + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \qquad (2)$$

Where p is the order of the autoregressive part of the model and q indicates the order of the moving average part.

Finally, using the model previously identified for each time series, a corresponding 95% confidence interval range is calculated for the next weeks, and a cross-validation process between the one-step ahead forecasts and the real observations is performed.

5. HIERARCHICAL TIME SERIES AT DIFFERENT GEOGRAPHICAL LEVELS

The method of hierarchical time series aims to explore simultaneously many series at different geographical levels, comparing baseline patterns for different subpopulations. This analysis also facilitates the identification of the spatial extend of irregular patterns previously detected at the regional level [Alba *et al*., 2015].

The hierarchical time series combine the information contained in two matrices. One matrix that contains the observations at the bottom-level (in our case: municipalities), and another matrix that contains information about the aggregation structure of the different spatial levels (called nodes) [Athanasopoulos *et al*., 2009; Hyndman *et al*., 2011, 2013] (see figure 2).

These hierarchical time series are built using the 'hts' package of R [Hyndman *et al*., 2011, 2013].

6. INTEGER-VALUED AUTORREGRESSIVE TIME SERIES MODELS OF DIFFERENT ORDERS TO MODEL TIME SERIES AT SMALLER SCALES

The final models proposed in the system are the integer-valued autoregressive time series models of different orders, abbreviated as INAR(k). INAR models, based on discrete or count time series techniques, can be understood to be an extension

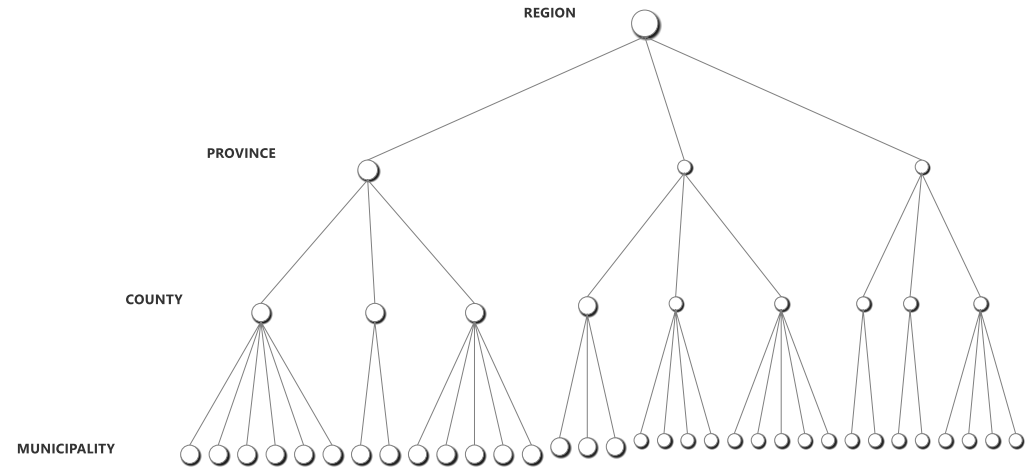
of the well-known AR models [Jung & Tremayne, 2006]. The diagram below shows in a simple way the similarities between the AR (1) and the INAR (1) models:

AR (1) model → $X_t = \rho X_{t-1} + Z_t \quad with \ Z_t \sim N(\mu, \sigma^2)$ (4)

INAR (1) model → $X_t = \rho o X_{t-1} + W_t \quad with\, W_t \sim Poisson\ (\lambda)$ (5)

Figure 2

Schematic diagram of the hierarchical time series structure proposed



.

In the AR(1) model, ρX_{t-1} corresponds to the autoregressive part (AR), that indicates the degree of dependency between the consecutive observations of the series, and Z_t is the random variable of the new events not explained by the previous observations. Whereas in the INAR(1) model, $\rho o X_{t-1}$ corresponds to a Binomial distribution with parameter $\rho \in (0, 1)$ linked to the previous number of observations by the binomial thinning operator, that is used to ensure the integer discreteness of the process. In our context $\rho o X_{t-1}$ would indicate the number of deaths that would be dependent on the previous events occurred at time t-1, and W_t would correspond to the new death counts that would not be explained by the past and that would depend on the data set context. W_t is fitted to a Poisson distribution.

In the event of a generalized INAR(*k*) process, the model would be defined by means of a similar equation but now this equation is *k* times recurrent,

$$X_t = \rho_1 o X_{t-1} + \rho_2 o X_{t-2} + \cdots + \rho_k o X_{t-k} + W_t \quad (6)$$

In this model the seasonal and trend behaviour is not covered. To solve this question it is also assumed that these components can be expressed as functions of time and, using suitable functions, the parameters are estimated and adjusted following a similar process to that used in ARIMA models. The model selection uses the AIC criteria, evaluates the statistical significance of the parameters and validates its analysis.

The model provides two types of forecasts: the future average behaviour of the series and the crude number of counts. The process followed for INAR is explained in detail by Fernandez-Fontelo *et al*. [2015].

III - RESULTS

To illustrate an example of the three modelling processes, time series of fallen stock data collected from dairy cattle between 2006 and 2013 have been modelled for a region, at different geographical levels.

The following plot shows the time series of fallen stock in dairy cattle at the regional level and the manner in which this series was divided into a training data set and a test data set (see figure 3).

Figure 3

Time series plot of fallen stock collected in dairy cattle between 2006 and 2014 at the regional level

The time series of dairy cattle at the regional level fits to a model ARIMA (1,0,1) and has an increased trend over time with half-yearly and yearly seasonality.

Its mathematical model may be expressed as:

$$X_t = 246.74 + 0.30t + 41.80\cos\left(\frac{2\pi t}{52}\right) - 22.39\sin\left(\frac{2\pi t}{52}\right) + 14.25\cos\left(\frac{2\pi t}{26}\right) + 28.95\sin\left(\frac{2\pi t}{26}\right) + Y_t \quad (7)$$

s.e. 6.26 0.03 3.73 3.80 2.99 3.01

$$Y_t = 0.86Y_{t-1} + Z_t - 0.72Z_{t-1} \quad (8)$$

s.e 0.07 0.10

(s.e.: standard errors)

The diagnostic checking indicates the lack of autocorrelation and partial autocorrelation of the standardized residuals (see figure 4):

The model predicts 374 carcasses on average during 2013 within a range of 352 and 501. All of the actual observations are included within this range (see figure 5).

The hierarchical time series plots of fallen stock for dairy cattle are shown in figure 6. These plots aims to compare the patterns observed at regional, provincial, county and municipality level.

Finally, an INAR model is used with a previous adjustment of the seasonality and trend to model the fallen stock pattern in an area that contains few dairy cattle farms (see figures 7 and 8).

In this case, the time series studies are fitted using an INAR (3) model expressed as:

$$X_t = 0.13oX_{t-1} + 0.10oX_{t-2} + 0.09oX_{t-3} + W_{\hat{a}1(t)} \quad (9)$$

$$\hat{a}1(t) = e^{-0.003-0.004t+0.38cos\left(\frac{2\pi t}{52}\right)} \quad (10)$$

Its residuals are distributed as white noise according to the ACF and PACF profiles.

Figure 4

Autocorrelation and partial autocorrelation plots to ensure the independency of the residuals

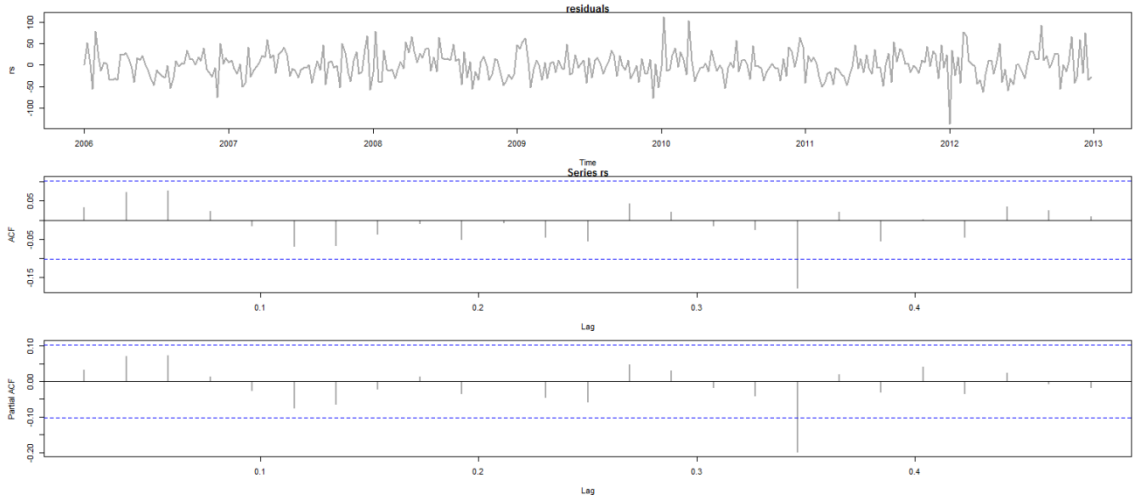


Figure 5

Time series plot of fallen stock collected in dairy cattle between 2006 and 2012 and forecasts (average value and 95% confidence interval) and observed values in 2013 using an ARIMA (1,0,1) with trend and seasonality

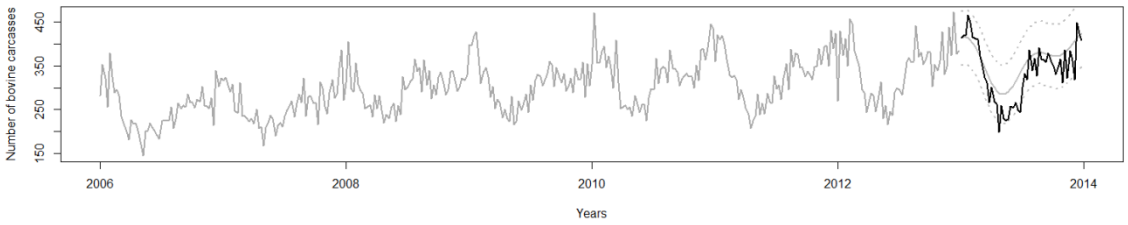


Figure 6

Hierarchical time series at four levels based on the number of carcasses collected by week in dairy cattle farms.

Level 0: region; Level 1: provinces; Level 2: Counties; Level 3: Municipalities.

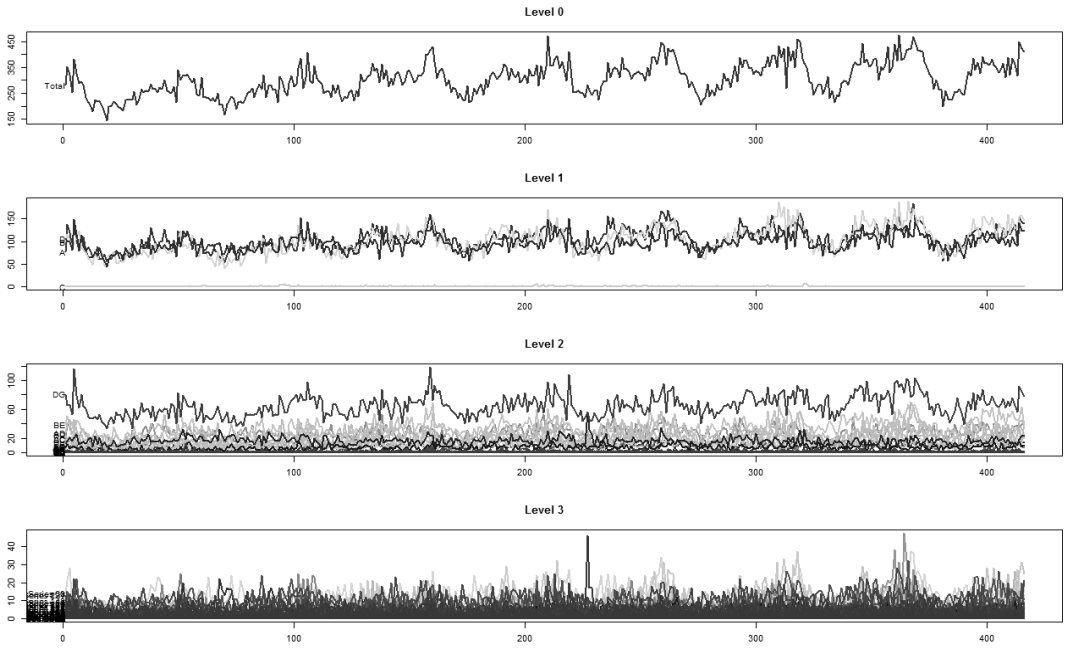


Figure 7

Time series plot of fallen stock for dairy cattle collected weekly in a small subpopulation

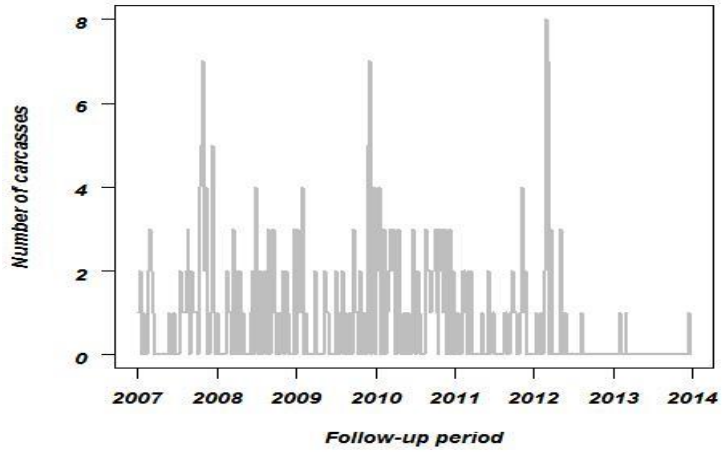
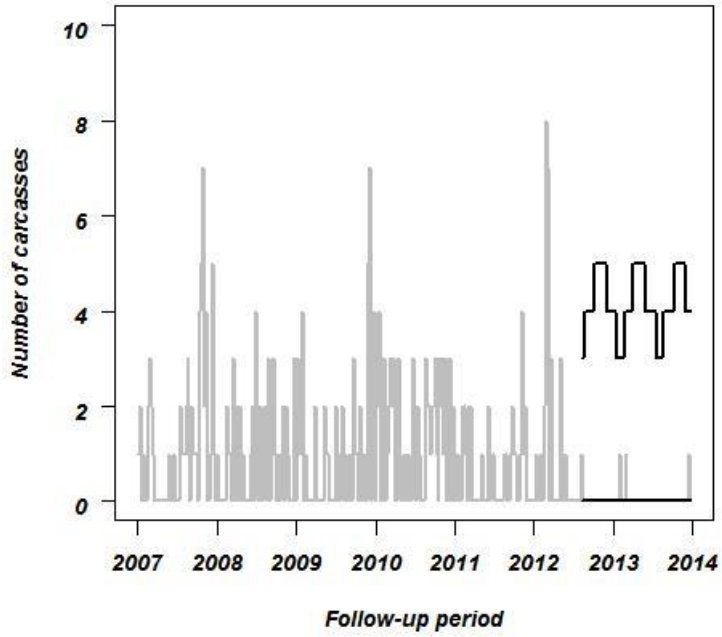


Figure 8

Time series plot of fallen stock collected in dairy cattle between 2007 and 2012 and forecasts (average value and 95% confidence interval) and observed values in 2013 using an INAR(3) with trend and seasonality



IV - DISCUSSION

Although previous studies have described different models that might be used to incorporate cattle fallen stock as a component of syndromic surveillance, as far as we know, none have shown an ordered strategy as to how this might be adapted to different spatial scales. In this sense the present work proposes a novel approach that combines the weak and strong points of different models to integrate the analyses of time series arising from large or small subpopulations in a unique system.

This study, in agreement with previous papers [Alba *et al*., 2015; Benshop *et al*., 2008], demonstrates the potential of ARIMA models to summarise historical patterns in a comprehensive manner in the event of regular series without zeros, which will tend to be the case for time series from large populations. These models are relatively easy to fit and may be useful to evidence the existence of irregularities. However, in the circumstance of small populations with irregular patters this approach is unlikely to be adequate.

To complement the information provided by ARIMA models at different scales, the use of hierarchical time series is proposed. This approach may be very useful for decision making. The method allows for an assessment of mortality data at different spatial levels, and in the event of an abnormal event, to determine the extent of the abnormality and adequate preventive and control measures according to its extension.

Finally, the system proposes the use of INAR (k) models to analyse fallen stock time series data with a high number of zeros or low counts. These models are applicable for small-scale analysis, and indeed, this is an important step forward for building a powerful syndromic surveillance system. However, it is important to state that fitting these models is challenging and further research should be conducted in order to automate their implementation [Fernandez-Fontelo *et al*., 2015; Moriña *et al*., 2011].

V - CONCLUSION

This system provides information to identify populations at high risk, define the historical patterns of fallen stock in the populations studied, and evaluate changes in those patterns at the regional level. In addition, this work assesses the utility of alternative and novel methods to model and forecast the patterns of subpopulations at smaller scales, facilitating local intervention and the allocation of resources at this level.

This system may provide useful information for decision making to allocate resources at different spatial levels and facilitate both central and local intervention.

Further research should be conducted to:

1. *Identify specific causes of mortality peaks, and* adequate thresholds for alarms,
2. Remove abnormal events from basal patterns,
3. Re-test and validate the novel algorithms,
4. Transfer all this information within the veterinary services,
5. Determine which decisions should best be made based on the information provided by the system.

REFERENCES

Alba A., Dórea F.C., Arinero L., Sanchez J., Cordón R., Puig P., Revie C. - Exploring the Surveillance Potential of Mortality Data: Nine Years of Bovine Fallen Stock Data Collected in Catalonia (Spain) 2015. PLoS ONE 10(4): e0122547. doi:10.1371/journal.pone.0122547.

Athanasopoulos G., Ahmed R.A., Hyndman R.H. - Hierarchical forecasts for Australian domestic tourism. *Int. J. Forecasting*, 2009, **25**, 146-166.

Benshop I., Stevenson M.A., Dahl J., Morris R.S., French N.P. Temporal and longitudinal analysis of Danish Swine Salmonellosis Control Programme data: implications for surveillance. *Epidemiol. Infect*., 2008, **136**(11), 1511-1520.

Core Team R. - A Language and environmental for Statistical Computing R Foundation for Statistical Computing Vienna, Austria. 2013. Available: http://www.R-project.org.

Cowpertwait P.S.P., Metcalfe A.V.. - Introductory Time Series with R. Springer Dordrecht Heidelberg London New York, 2009.

Chatfield Ch. - The Analysis of Time Series An Introduction. Sixth Edition. Chapman&Hall/CRC. Boca Raton, 2004, 33-167.

Dupuy C., Bronner A., Watson E., Wuyckhuise-Sjouke L., Reist M., Fouillet A. *et al*. - Inventory of veterinary syndromic surveillance initiatives in Europe (Triple-S project): Current situation and perspectives. *Prev. Vet. Med*., 2013, **111**, 220-229.

Fernández-Fontelo A., Puig P., Alba A. - Generalized INAR models with trend and seasonality for veterinary syndromic surveillance. Proceedings of the International Workshop of Statistical modeling. Johannes Kepler University, Linz, Austria, 2015, Volume 1.

Hyndman R.J., Ahmed R.A., Athanasopoulos G., Shang H.L. - Optimal combination forecasts for hierarchical time series. *Comput. Stat. Data An*, 2011, **55**(9), 2579-2589.

Hyndman R.J., Ahmed R.A., Shang H.L. - hts: An R Package for Forecasting Hierarchical or Grouped Time Series. R package version 3.00, 2013.Available: http://CRAN.R-project.org/ package=hts.

Moriña D., Puig P., Ríos J., Vilella A., Trilla A. - A statistical model for hospital admissions caused by seasonal diseases. *Statmed*.,2011, **30**, 3125-3136.

Perrin J.B., Ducrot C., Vinard J.L., Morignat E., Gauffier A., Calavas D. *et al*. - Using the National Cattle Register to estimate the excess mortality during an epidemic: Application to an outbreak of Bluetongue serotype 8. *Epidemics*, 2010, **2**, 207-214.

Perrin J.B., Ducrot C., Vinard J.L., Morignat E., Calavas D., Hendrick P. - Assessment of the utility of routinely collected cattle census and disposal data for syndromic surveillance. *Prev. Vet. Med*., 2012, **105**, 244-252.

Van der Loo M.P.J., de Jonge E. - Learning R Studio for R Statistical Computing. Packt Publishing. open source Birmingham-UK, 2012.

Van der Loo M.P.J, de Jonge E. - Learning R Studio for R Statistical Computing. Packt Publishing. open source Birmingham-UK, 2012.