Preventive Veterinary Medicine 153 (2018) 64–70

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

A time-series approach for clustering farms based on slaughterhouse health aberration data

B. Hulsegge, K.H. de Greef, Ina Hulsegge[*]

Animal Breeding and Genomics, Wageningen Livestock Research, P.O. Box 338, 6700 AH, Wageningen, The Netherlands

ARTICLE INFO

Keywords:
Meat inspection data
Time series
Characteristic-based clustering
Big data
Pneumonia
Pleuritis

ABSTRACT

A large amount of data is collected routinely in meat inspection in pig slaughterhouses. A time series clustering approach is presented and applied that groups farms based on similar statistical characteristics of meat inspection data over time. A three step characteristic-based clustering approach was used from the idea that the data contain more info than the incidence figures. A stratified subset containing 511,645 pigs was derived as a study set from 3.5 years of meat inspection data. The monthly averages of incidence of pleuritis and of pneumonia of 44 Dutch farms (delivering 5149 batches to 2 pig slaughterhouses) were subjected to 1) derivation of farm level data characteristics 2) factor analysis and 3) clustering into groups of farms. The characteristic-based clustering was able to cluster farms for both lung aberrations. Three groups of data characteristics were informative, describing incidence, time pattern and degree of autocorrelation. The consistency of clustering similar farms was confirmed by repetition of the analysis in a larger dataset. The robustness of the clustering was tested on a substantially extended dataset. This confirmed the earlier results, three data distribution aspects make up the majority of distinction between groups of farms and in these groups (clusters) the majority of the farms was allocated comparable to the earlier allocation (75% and 62% for pleuritis and pneumonia, respectively). The difference between pleuritis and pneumonia in their seasonal dependency was confirmed, supporting the biological relevance of the clustering. Comparison of the identified clusters of statistically comparable farms can be used to detect farm level risk factors causing the health aberrations beyond comparison on disease incidence and trend alone.

1. Introduction

According to legal regulations (European Community, 2004), all slaughtered pigs in the European Union are subject to a routine meat inspection at the slaughterhouses. Traditionally, meat inspection has been used to reduce food-borne risk to public health (Edwards et al., 1997). The meat inspection findings are also valuable indicators that can be used as a feedback system indicating animal health and to derive recommendations for improvement of farm management (Schuh et al., 2000). Meat inspection data can be used to inform farmers on the health status of their herd (benchmarking) since health aberrations indicate systems (housing, ventilation control) or management (treatment and prevention strategies) failures. Slaughterhouse data both reveal such problems and offer the opportunity to monitor effectivity of interventions. Current use of slaughterhouse health aberration data seems limited to periodic reporting of farm incidence averages. Understanding the data structure (such as temporal patterns) of aberrations in meat inspection data may provide important information beyond these average incidence figures.

One possible approach to analyse meat inspection data involves time series methods, such as exploratory methods (Sanchez-Vazquez et al., 2012; Alhaji et al., 2015) and autoregressive models (Neumann et al., 2014; Vial and Reist, 2014; Adachi and Makita, 2015). These methods however, require structured data, a sufficient number of observations that are fairly regularly measured over time, which is often not the case for data on batches of pigs delivered to slaughterhouses. Another possible approach is time series clustering directly on raw data. This method however does not account for the temporal sequences of the observed values and the autocorrelations structure of the data is ignored. Characteristic-based clustering has been developed to address the problem of clustering raw time series data (Hennig et al., 2015). This method has been proposed by several authors in various domains such as electricity (Räsänen and Kolehmainen, 2009), business (Davenport and Funk, 2015), and human health (Leffondré et al., 2004) (Niedermeyer et al., 2011). We applied this method to group farms based on similar statistical characteristics of meat inspection data, focussing on pneumonia and pleuritis.

The objective of this study was to explore whether an analysis which

[*] Corresponding author.
E-mail address: ina.hulsegge@wur.nl (I. Hulsegge).

https://doi.org/10.1016/j.prevetmed.2018.03.003
Received 18 August 2017; Received in revised form 20 February 2018; Accepted 5 March 2018

utilises more information from the data than incidence figures provides added value to make distinctions between individual farms. A comprehensive meat inspection dataset, collected over 3.5 years, was available for this. This more detailed farm characterisation may aid in finding risk factors for failures by comparing more uniform groups of farms.

2. Material and methods

2.1. Data source

Post mortem meat inspection data of carcass and organs are collected on every slaughtered pig in The Netherlands. The inspection procedures are described in detail in Regulation EC no. 854/2004 (European Community, 2004). Meat inspection data collected between January 2011 and August 2014 were provided by the major Dutch meat producer, one record for each slaughtered pig, with information on pneumonia and pleuritis and aberrations on legs, skin and liver. Respiratory disorders were chosen as study dataset as they are one of the major diseases affecting pigs worldwide (Brockmeier et al., 2002) and have reasonable incidences across farms and seasons and the repeatability of the slaughterhouse classification is adequate.

2.2. Study sample

Criteria were developed to derive a suitable sub-dataset for method development and analysis. August 2014 was excluded since it did not comprise the entire month, also batches with less than 10 animals were excluded. The two slaughterhouses with the largest number of records were selected. These slaughterhouses had complete datasets for the entire period and no obvious changes in inspection system. In this set, farms were selected that had delivered at least one batch with at least 10 pigs every month and at least 87 batches (more than 1 batch per 2 weeks on average). The resulting study sample contained information of 511,645 pigs submitted from 44 Dutch farms in 5149 batches.

Information on the percentage pneumonia and pleuritis in the batches is presented in Table 1. The analysis is principally batch based – records were created containing batch averages. The percentage of each aberration (pleuritis or pneumonia) in each batch was computed as number of pigs in that batch with the aberrations divided by total number of pigs in that batch multiplied by 100.

The study dataset is quite complete from a statistical point of view (no missing records, good distribution over the entire study period), but comprises a small part of the total dataset. For verification and validation reasons a second, larger, dataset was created. The selection criteria were released: all farms of the two slaughterhouses were included which met the criterion that the whole study period (all months) was reasonably covered: 6 month averages were allowed to be missing for each farm. This resulted in an three to almost fourfold size of the data: 163 farms delivering 15,276 batches comprising 1,829,762 slaughtered pigs. Table 1 illustrates that the characteristics of the validation set resemble those of the study set.

Table 1
Percentage pneumonia and pleuritis in the study sample (5149 batches) and validation sample (15,276 batches).

Aberration	# Batches with percentage 0% (%)	Mean percentage (95% CI) in a batch	Sd percentage	Max percentage
Pneumonia				
Study sample	615 (11.9%)	8.76 (8.51–9.01)%	9.10%	63.83%
Validation sample	1599 (10.5%)	9.12 (8.96–9.26)%	9.38%	78.15%
Pleuritis				
Study sample	375 (7.3%)	12.42 (12.12–12.72)%	10.91%	61.64%
Validation sample	1384 (9.06%)	10.04 (9.88–10.20)%	10.37%	82.58%

2.3. Time series visual explorations

For exploratory purpose, percentage aberrations were aggregated for each month of study. An exploratory analysis was conducted by plotting percentage aberrations of the study sample containing 44 farms in the period January 2011 to July 2014 in a multivariate time series plot using the R package mvtsplot (Peng, 2008). The mvtsplot method produces an adaptation of the multivariate time series plot which combines a heatmap with boxplot-like summaries and a basic line plot to provide a detailed overview of the data. The colours purple, grey and green in the heatmap correspond to low, medium and high values, respectively. The darker the shading the larger the value.

2.4. Time series clustering using global characteristics

We used a three step method to group farms with comparable statistical characteristics of health aberrations over time (Fig. 1). The first step of the method involved replacing the raw time series data with some global measures of time series characteristics, as described by Wang et al. (2006) and Räsänen and Kolehmainen (2009). The measures summarized information of the time series, to capture the 'global picture' of the data. The characteristics used in this study were: *mean, standard deviation, trend, seasonality, remainder, autocorrelation, skewness, kurtosis, chaos, nonlinearity,* and *self-similarity*. Table 2 describes the popularised interpretation of these characteristics and their acronym used below.

Trend and *seasonality* are common characteristics of time series, and it is natural to characterize a time series by its degree of trend and seasonality. In addition, once the trend and seasonality of a time series has been measured, the time series can be detrended and deseasonalised to enable additional features such as noise or chaos to be more easily detectable. The R function *stl* was used for detrending and deseasonaling the timeseries (Cleveland et al., 1990). For the validation sample (which contained missing values), the R package *stlplus* version 0.5.1 was used to detrend and deseasonlise the time series, applying a loess algorithm to handle missing values (Hafen, 2010).

To obtain a precise and comprehensive calibration, some measures are calculated on both the raw time series as well as the remaining time series after detrending and deseasonalising. All these characteristics (presented in a popular phrasing in Table 2) are thoroughly explained by Wang et al. (2009) and (Davenport and Funk, 2015).

In the second step a factor analysis, using the function *principal* from the R package *Pysch* version 1.5.8 (Revelle, 2015), was performed to select a subset of characteristics that condensed the information present in the characteristics and provided the best description. We only kept the factors with an eigen-value greater than 1 (Tabachnick and Fidell, 2006), those that are more informative than a single variable. The varimax rotation was used to facilitate the interpretation of results by maximising the loading of each individual variable on a single factor (i.e., its correlation with this factor). For each factor the measure that had the highest loadings (i.e. the highest correlation with a give factor) was selected.

Finally (third step), we used cluster analysis to identify clusters of farms with similar patterns of characteristics selected by the factor analysis. In order to weigh all characteristics equally, all characteristics were transformed to the same range (0,1). A measure near 0 for a certain time series indicates an absence of the characteristic while a measure near 1 indicates a strong presence of the characteristic (Wang et al., 2006). The measures were normalised with the function *SofMax* of the R package *DMwR* version 0.4.1 (Torgo, 2010). The R package *NbClust* version 3.0 (Charrad et al., 2014) was used to perform the