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

Hulsegge, I., de Greef, K. H., & Hulsegge, I.

Hulsegge, I., de Greef, K. H., & Hulsegge, I. (2018). A time-series approach for clustering farms based on slaughterhouse health aberration data. Preventive Veterinary Medicine, 153, 64-70. DOI: 10.1016/j.prevetmed.2018.03.003

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

B. Hulsegge[a] and K.H. de Greef[a]

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

Corresponding Author:

Ina Hulsegge, Animal Breeding and Genomics, Wageningen Livestock Research, P.O. Box 338, 6700 AH, Wageningen, The Netherlands. Tel: +31-317-480513; E-mail: ina.hulsegge@wur.nl

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 5,149 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.

Highlight

- Characteristic-based clustering is able to cluster time series of meat inspection data of farms using a set of derived statistical characteristics.
- Seasonality and data dispersion characteristics such as autocorrelation have additional value to the conventional incidence figures of pneumonia and pleuritis.

- Farms were mainly clustered on: amount of variation in the data; distribution shape of the data and similarity between consecutive data points.

Keywords

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

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 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 5,149 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 (5,149 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.

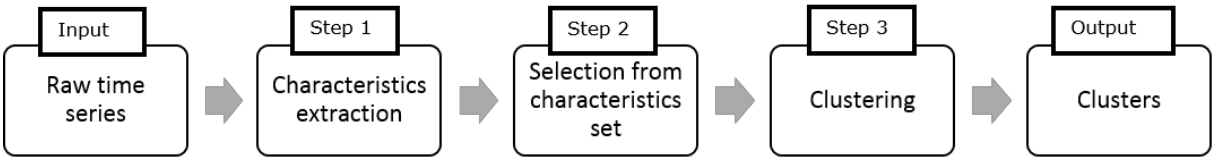

Fig. 1. Characteristics based clustering approach (after Wang et al. (2006)).

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).

Table 2. Summary of the used data characteristics, calculated from the raw batch data and on the detrended and deseasonalised data.

Characteristic	Definition	Acronym Raw data	Acronym Detrended and deseasonalised data
Mean	The average of the observations	'*mean*'	
Standard deviation	A measure of how spread out the data is.	'*sd*'	
Trend	A pattern found in time series; used to describe whether the data is showing an upward or downward movement for a part, or all of the time series.	*'trend'*	
Seasonality	A pattern of a time series in which the data experience regular and predictable changes that repeat every calendar year.	*'seasonality'*	
Remainder	The residuals of the time series after allocation into the seasonal and trends time series (also called "noise", "irregular" or "random").	*'remainder'*	
Hurst Exponent	A measure for longterm memory and fractality of a time series (an evaluation index of the self-similarity).	*'self.sim'*	
Autocorrelation	The correlation within a time series with its own past and future values (also called serial correlation)	*'autocorr'*	*'dc-autocorr'*
Skewness	A measure of how symmetrical a distribution is.	*'skewness'*	*'dc-skewness'*
Kurtosis	A measure which describes the distribution of the observed data around the mean. A measure of how peaked or flat a distribution is relative to the normal distribution	*'kurtosis'*	*'dc-kurtosis'*
Lyapunov Exponent	A measure of stability; Chaos	'*chaos*'	
Nonlinearity	A measure for not arranged in a straight line.	*'nonlin'*	*'dc-nonlin'*

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 cluster analysis, in order to identify the optimal number of clusters. Clusters were generated using the complete linkage method applied to Euclidean distances.

3. Results

3.1. Percentage aberrations at farm level.

The monthly percentage of aberrations for the farms in the periods January 2011 to July 2014 varied between farms and months (Fig. 2). Monthly farm pneumonia incidences in the study set varied between 0.8 and 25.1%, and pleuritis incidences varied between 1.4 - 24.0%. As the colouring in Figure 2 indicates, farms not only differ in monthly incidence, there is also considerable between farm variation in within-farm consistency in time. Consistent farms have either an entirely green coloured time-series (low incidence) or entirely purple coloured time-series (high incidence). Farms with alternating colours in their time series have low consistency in their incidences.

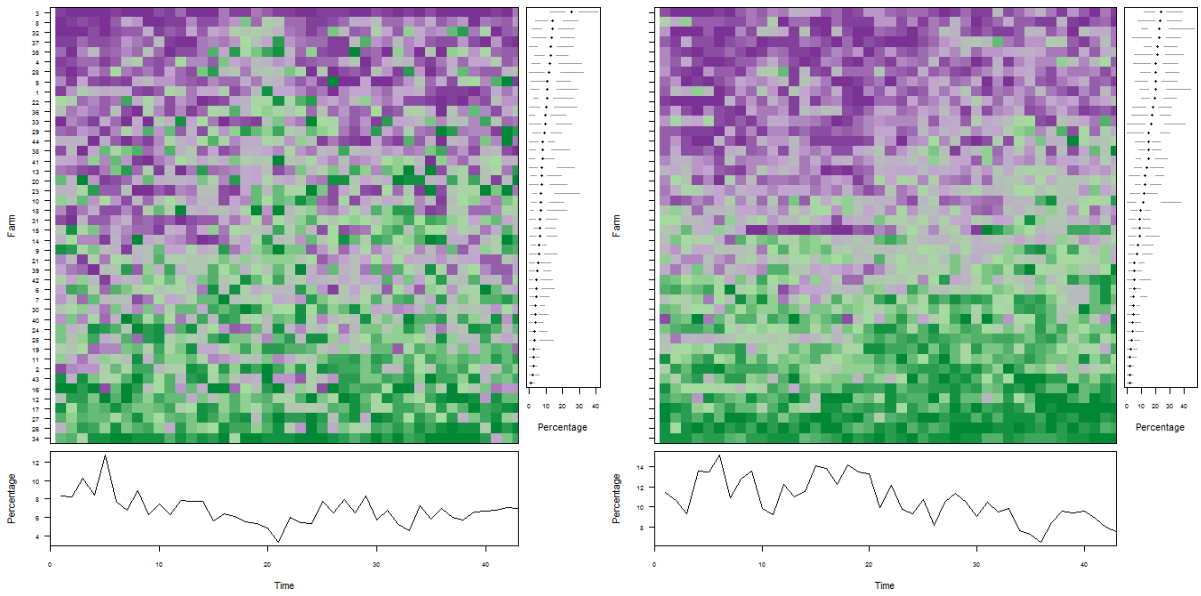

Fig. 2. Multivariate time series plot of percentage pneumonia and pleuritis for 44 farms. The purple to green palette represents variation in percentage aberrations (green represents low percentages; purple high percentages). The right panel presents summary statistics of percentage aberrations for each farm, the black dots denote the median while the horizontal lines represent the lower and upper quartiles. The lower panel shows the median values of percentage aberration across the time series of the 43 months (1= January 2011 and 43= July 2014) for each time point.

3.2. Time series clustering using global characteristics.

3.2.1. Pneumonia

Exploratory factor analysis of percentage pneumonia reduced the 15 global characteristics to four factors explaining 65% of the variance. The most informative global characteristics were: '*mean*', *'seasonality'*, *'autocorr'* and *'dc-kurtosis'* (Fig. 3). These four most informative global characteristics were used by cluster analysis and resulted in grouping the 44 farms into three clusters. Categorization of the three clusters data characteristics are shown in Figure 3.

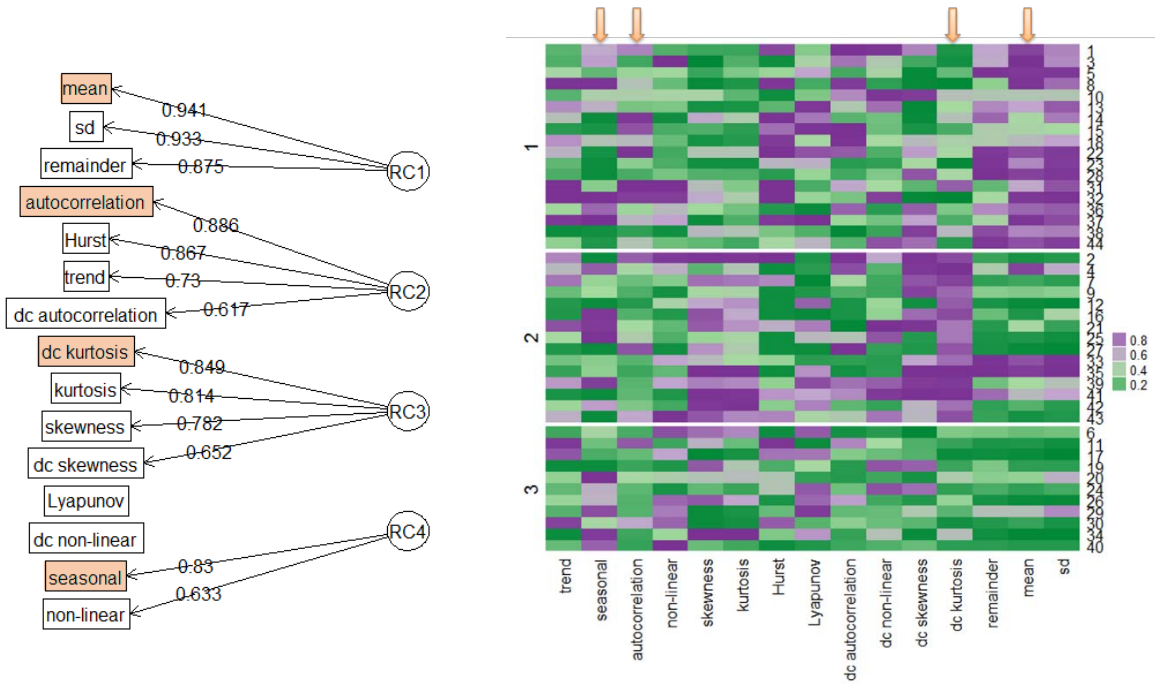

Fig. 3. Factor analysis path diagram of pneumonia (left pane) and allocation to three clusters (right pane). Left pane: The coloured square boxes are the characteristics of each principal component (PC) that are used in subsequent analysis. On the straight arrows, the loadings (correlation between the principal component and the characteristic) are presented. Only the largest loadings are shown. Right pane: Characteristics summary for the three identified clusters. Purple indicates high values, green indicates low values. The arrows at the top indicate the selected global characteristics; left axis: cluster number; right axis: farm number.

Farms in cluster 1 are characterised by high incidence values ('*mean*') with large variability in pneumonia incidence and low trend and seasonally adjusted kurtosis ('*dc-kurtosis*'), having a flat top near the mean and produces fewer and less extreme outliers than does the normal distribution. Cluster 2 groups farms with the opposite: low '*mean*' and high '*dc-kurtosis*', having a distinct tall peak near the mean, decline rather rapidly and have fatter tails or more extreme values. Farms in cluster 3 share the low incidences with cluster 2, but combine this with low kurtosis, meaning that the trend and seasonally adjusted time series produces fewer and less extreme outliers than does a normal distribution. The factor analysis suggested '*seasonality'* as an informative characteristics,

the value for all clusters showed little recurring seasonal pattern, periods of above-average and below-average percentage pneumonia each year (Fig. 3). Farms belonging to cluster 1 fluctuated most with season, from -2.4% in September to 2.2% in December. For cluster 2 the lowest value of the seasonal component was observed in August (-1.1%) and the highest in May (1.2%). For cluster 3 these values varied from -1.4% in September to 1.5% in May.

3.2.2. Pleuritis

Factor analysis of percentage pleuritis reduced the 15 global characteristics to three factors with the most informative global characteristics being '*self.sim'*, '*chaos*' and '*mean*', explaining 62% of the variance (Fig. 4). Cluster 1 contains farms with high incidence values and a strong trend. These time series had also highly regular fluctuations over time (high '*self.sim'* exponent values; indicating a persistent time series) and showed no chaotic behaviour (low '*chaos*' values). Farms with low incidence figures were predominantly allocated to cluster 2, combined with high values for '*chaos*'. The time series showed no trend or seasonal effect and self-similarity was almost not present. Farms in cluster 3 are characterised by low levels of *'autocorr'*, '*chaos*' and '*self.sim'*, but differ mutually in their incidence figures.

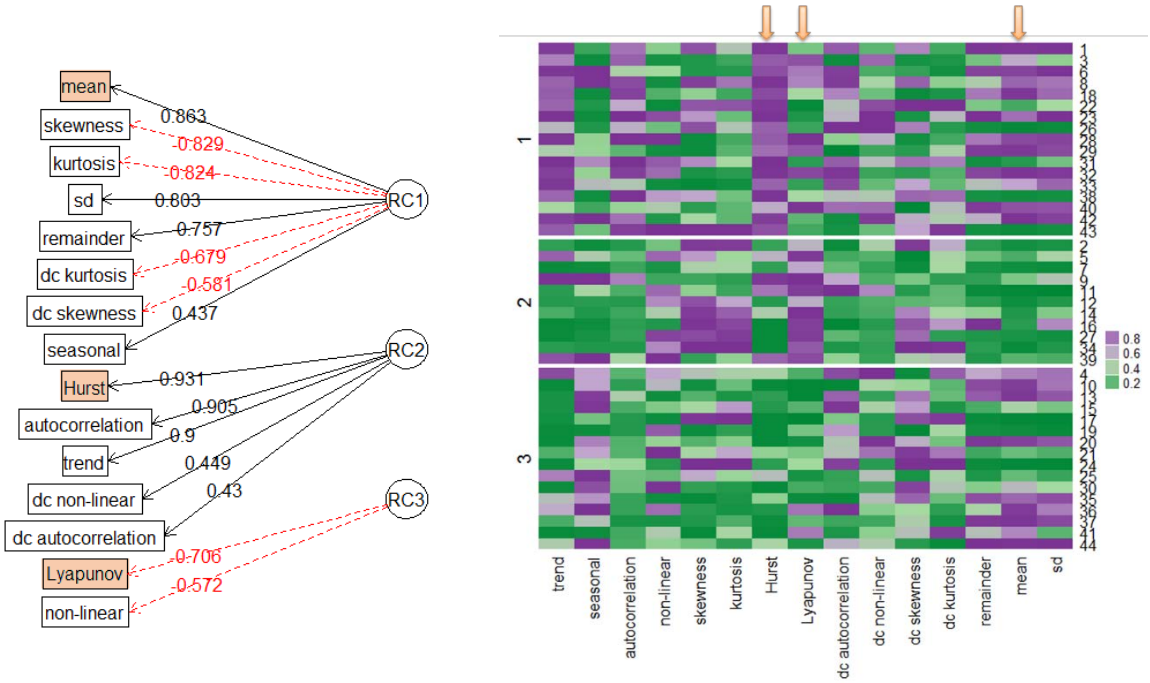

Fig. 4. Factor analysis path diagram of pleuritis (left pane) and allocation to three clusters (right pane). For explanation: see Figure 3.

3.2.3. Validation in the extended dataset

To test the robustness of the clustering, the analysis was repeated with a substantially extended dataset. By releasing 1) the criterion that farms have to have batches in all months of the study period and 2) the criterion that at least 87 batches were delivered, the study size was extended about fourfold. The only remaining criteria were that batches contained at least 10 pigs and that in most months a batch was available, maximally six missing. The *stlplus* allows handling of missing monthly values.

For both pneumonia and pleuritis aberration data, again the three archetypes of clusters evolve: one based on the variation characteristics, one on the distribution shape and one on the similarity of consecutive data points (Fig. A1 and A2). Specifically for pneumonia, the validation set, having missing data points in most farms, failed to identify the specific cluster indicating a group of farms with specific seasonal sensitivity. Similarly, the validation exercise on pleuritis figures was to some degree less distinctive in discriminating farms with regard to sequentiallity/chaos, but was stronger in its

separation between variance (‘*mean’*, ‘*sd’*) and shape characteristics (‘*skewness’*, ‘*kurtosis’*) with regard to pleuritis.

In Table 3 and 4, the degree of similarity in allocation to the clusters between the study analysis and the validation analysis is presented for the 44 farms that were involved in the study sample and again allocated to new clusters during from the validation analysis.

Table 3. Clustering of farms for pneumonia characteristics: comparison of the coherence in allocation to clusters in the test analysis (vertical: clusters 1, 2, 3) and to clusters in the validation analysis (horizontal: clusters A, B, C, D).

Validation cluster → Original cluster ↓	A	B	C	D	*Total*
1	3	15			*18*
2	3	4	1	7	*15*
3	3	5		3	*11*
Total	*9*	*24*	*1*	*10*	*44*

Table 4. Coherence in clustering of farms for pleuritis characteristics: number of farms allocated to clusters in the test analysis (vertical: clusters 1, 2, 3) and to clusters in the validation analysis (horizontal: clusters A, B, C).

Validation cluster → Original cluster ↓	A	B	C	*Total*
1	4	13		*17*
2	4	1	6	*11*
3	14		2	*16*
Total	*22*	*14*	*8*	*44*

For pleuritis, the overall correspondence of the two clustering analyses amounts

75%. 14, 6 and 13 farms (from cluster 1, 2 and 3 respectively) which were grouped together in the original analysis were again allocated together into the new clusters. For pneumonia, the overall correspondence is less. It amounts 61% and an extra cluster is formed. The largest new cluster (B, containing 24 farms) comprises the majority of two of the original clusters (1 and 3). Overall, most of the farms that were grouped together in the study analyses were allocated into joint clusters again in the validation study, both for pleuritis and pneumonia.

4. Discussion

Meat inspection generates a large amount of time series data that are used to only a limited extend for animal health surveillance purposes. And if so, use is generally limited to the average incidence ('*mean'*), and its change in time, solely on farm level. Current exercise enriches this by combining data across farms. Understanding the underlying information and interpretation of the results for meaningful purposes (such as management support or detection of risk factors) is an opportunity, but also a challenge due to the high diversity between farms, batches and underlying factors.

A dataset containing more than 3½ years of historical meat inspection data was available to explore the potential of a data analysis to cluster farms into groups with comparable health aberration patterns over time. In this dataset two respiratory disorders, pneumonia and pleuritis, were chosen as study objects as they are among the major diseases affecting pigs worldwide (Brockmeier et al., 2002), (Merialdi et al., 2012) (Eze et al., 2015) and the most common slaughter aberrations found in pigs (Sanchez-Vazquez, 2013). Also, the diversity of incidences and the reasonable repeatability of slaughterhouse pleuritis and pneumonia classification is helpful from a statistical point of view to develop the proposed method.

4.1. The method

Clustering is among the most widely used method in the analysis of time series data (Fidaner et al., 2015) (Chen et al., 2017) and for our casus, it offers the opportunity to

identify farms with similar patterns of percentage pneumonia or pleuritis over time discerning similarities between those farms beyond obvious characteristics such as incidence figures (Fidaner et al., 2015).

Characteristic-based clustering first converts raw time series data into a characteristic vector of lower dimension, after which clustering is applied. Characteristic-based clustering, in the literature also called Feature based clustering or Statistical measures based clustering, has been proposed by several authors across science for clustering time series. For example, Leffondré et al.(2004) used this method for identifying patters of change in quantitative human health indicators and Räsänen and Kolehmainen (2009) for electricity use time series data. We applied this method to group farms based on similar statistical characteristics of meat inspection data over time. Our approach consists of three distinct steps: 1) computation of the data characteristics on farm level from monthly farm averages; 2) factor analysis to identify the major explaining variation among farms; and 3) cluster analysis to group farms on basis of similarity in their data characteristics. Ad 1), we used the set of characteristics as proposed by Wang et al. (2005; 2006) that contains measures of '*trend'*, '*seasonality'*, *'autocorr'*, '*skewness'*, '*kurtosis'*, '*chaos'*, '*nonlin'*, and *'self.sim'* to represent time series. The proposed statistical characteristics were selected because they are simple and easy to compute. Ad 2), for selection of the most relevant characteristics of the data set, various approaches can be used. We used factor analysis as search mechanism to find the best selection from the characteristics set as suggested by Leffondré et al. (2004). This is an easy and widely accepted method to identify common patterns in data with diverse correlations structures. Ad 3), we chose the clustering method according to Leffondré et al. (2004) as it seems to fit our ambition well.

The correlation matrix between the characteristics illustrates that several characteristics were highly correlated; e.g. mean and standard deviation for percentage pneumonia as well as percentage pleuritis (data Fig. A3 and A4). From a methodological point of view, this correlation structure implies that some features are interchangeable. Having two

highly correlated characteristics makes one virtually redundant – in this case it may be useful to select the one which is easiest to interpret, as suggested by Leffondré et al. (2004).

4.2. The clustering of farms on basis of statistical characteristics

The results showed that the applied approach is able to discriminate between farms with regard to their meat inspection data. Both the data on pneumonia and on pleuritis resulted in three clusters. A closer look into the composition of these clusters reveals that both in the pleuritis data and in the pneumonia data farms were clustered mainly on: 1) amount of variation in the data; 2) distribution shape of the data and 3) similarity between consecutive data points. Both the consistent distinction between groups of characteristics and the consistency between the study results and the validation results confirm that the method is able to make distinction between farms beyond grouping them on the conventional way: incidence (percentage of pigs), possibly grouped in categories like high, moderate and low incidence.

4.3. Study set versus validation set.

The dataset on 44 farms (511,645 animals) was the ideal set to develop the method. But, regarding the small sample size, quite distant from the data as a whole. Extension to a larger (163 farms, 1,829,762 animals), but less optimal (less data points per farm, some missing month averages) set is a feasible model to verify whether the method holds for in a less ideal situation. 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).

Switching to less structured data also revealed a trade off between accuracy (a small but precise data set) and volume (a larger but more rough dataset). The study sample

revealed a specific vector for seasonal sensitivity for pneumonia, which was not detected in the larger dataset which had missing datapoints.

4.4. Relevance

Classically, farms are compared on basis of the incidence of lung problems. Obvious first next level comparisons comprise the variability and change in time of individual farm health performance. The high correlation between mean, standard deviation and coefficient of variation of the aberrations in both pneumonia and pleuritis indicates that variability between batches is not a valuable extra trait in itself, as it does not add substantial information additional to the average level of aberrations. On the other hand, other characteristics, such as repeatability patterns in time do aid in making distinctions between farms.

In literature, farm factors that affect problems like pleuritis and pneumonia are often assessed by comparing farming systems factors such as organic versus conventional farms (e.g. Alban et al. (2015) or comparing large versus small scale farms (e.g. (Enoe et al., 2002; Fablet et al., 2012). Data analysis offers an additional entry: the statistical grouping of farms may point at similarities in farm characteristics within the groups or differences between the groups (clusters) of farms that do not vary between for example organic and conventional systems, but rather are underlying factors in both systems that are causally related to the incidence of health aberrations. The clustering approach thus goes beyond comparing farms on basis of systems characteristics (size, type) or performance (incidences of aberrations) alone and bears the promise to reveal relevant risk factors from data of seemingly similar farms

A real practical validation requires insight of the farm characteristics. Relating farm characteristics (farm size, housing characteristics etc.) to the clusters is the next step to

utilize its relevance for enhancing health performance. A promising approach to identify risk factors for lung aberrations is to study whether the farms in different clusters also structurally deviate at farm-level either in (nutritional) management practices or in environmental (housing and ventilation) factors. Our dataset was unique in its size and consistency, but it contains only slaughter data, farms were coded, implying that no farm characteristics were available in the analysis. On availability of adequate data, comparing the clusters with regard to farm characteristics is an obvious next step in studying added value of these clustered slaughterhouse data. Do farms that are clustered on statistical grounds also resemble in farm characteristic? And which farm characteristics? If so, this is a signal that these characteristics may be closely related to real risk factors. Further developments in data sharing and in data analytics (*big data*, *machine learning*) are likely to further develop such opportunities.

4.5. Biological interpretation/ relevance

Pleuritis and pneumonia are both disorders of the respirational system, but have different aetiology. Present paper is not intended to elaborate on this, but lines towards biological interpretation can be drawn. Patterns and trends in incidences of pleuritis and pneumonia are readily discernable in massive slaughterhouse data, but are difficult to quantify in detail on the individual farm level. Current method identifies these patterns on individual farms, making use of the trends and patterns of related farms. And it subsequently groups farms with similar aberration characteristics. This grouping is considerably different for pneumonia and for pleuritis. Comparison of the clusters of farms for pneumonia and for pleuritis reveals that only a minority of the farms shares the same clusters (data not shown). This illustrates the different underlying factors affecting the (slaughterhouse detected pathological indicators of) these two respiratory disorders. Furthermore, the analysis identified within- and between farm variation related to seasonality for pneumonia, rather than for pleuritis. This is in line with earlier studies (for example by Fablet et a. (2012) who identified distinct risk factors for pneumonia and

pleuritis (ventilation and seasonality versus temperature and barn climate) in slaughtered pigs.

Interpretation of the parameters from a biological point of view is possible, but speculative. For example, the parameters for seriality may indicate that farms with high figures for this have the characteristic that they are quite consistent between months in their aberration performance. Also, high or low levels of kurtosis could be interpreted as relatively long or short problem periods. However, such interpretations are speculative, and to our knowledge, such interpretations have not been made in literature.

Results like those presented here confirm the biological ground under the identified clusters. Also, they support the expectation that data analysis points at less obvious underlying phenomena, which is helpful in further understanding the farm level aetiology of these disorders. Also, management opportunities can be strengthened by combing farm characteristics to the wealth of routinely collected data in slaughter houses, primarily in detecting husbandry related risk factors.

Current work has higher relevance for practical application (such as identification of farm factors affecting incidence levels) than for enhanced understanding of the underlying biology of the diseases involved.

5. Conclusion

Characteristic-based clustering was able to cluster time series of meat inspection data of farms using a set of derived statistical characteristics. The stepwise analysis of the slaughterhouse dataset reveals structured variation among farms in incidence of pneumonia and pleuritis. The applied method groups them into clusters of ‘similar’ farms beyond clustering them just on basis of observed incidence of aberrations. Seasonality and data dispersion characteristics such as autocorrelation had additional value to the conventional disease incidence figures. The differences between the clusters likely point

458 at systematic differences between individual farms. Validation on a substantially
459 extended dataset confirmed the results of the study dataset.

460

461 **Acknowledgements**

462 This research was commissioned and partly funded by The Feed4Foodure program line
463 “Nutrition, Intestinal Health, and Immunity” (Feed4Foodure; BO-22.04-002-001). The
464 authors would like to thank Vion Food Group for providing the data.

465

466 **References**

467 Adachi, Y., Makita, K., 2015. Real time detection of farm-level swine mycobacteriosis
468 outbreak using time series modeling of the number of condemned intestines in
469 abattoirs. The Journal of Veterinary Medical Science 77, 1129-1136.
470 Alban, L., Petersen, J.V., Busch, M.E., 2015. A comparison between lesions found during
471 meat inspection of finishing pigs raised under organic/free-range conditions and
472 conventional, indoor conditions. Porcine Health Management 1, 4.
473 Alhaji, N.B., Odetokun, I.A., Shittu, A., Onyango, J., Chafe, U.M., Abubakar, M.S.,
474 Muraina, I.A., Fasina, F.O., Lee, H.S., 2015. Time-series analysis of ruminant
475 foetal wastage at a slaughterhouse in North Central Nigeria between 2001 and
476 2012.
477 Brockmeier, S.L., Halbur, P.G., Thacker, E.L., 2002. Porcine Respiratory Disease
478 Complex. Polymicrobial Diseases. American Society of Microbiology.
479 Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. Nbclust: An R package for
480 determining the relevant number of clusters in a data set. Journal of Statistical
481 Software 61, 1-36.
482 Chen, Y., Wang, L., Li, F., Du, B., Choo, K.K.R., Hassan, H., Qin, W., 2017. Air quality
483 data clustering using EPLS method. Information Fusion 36, 225-232.

484 Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: A seasonal-
485 trend decomposition procedure based on loess. Journal of Official Statistics 6, 3-
486 73.
487 Davenport, F., Funk, C., 2015. Using time series structural characteristics to analyze
488 grain prices in food insecure countries. Food Security 7, 1055-1070.
489 Edwards, D.S., Johnston, A.M., Mead, G.C., 1997. Meat inspection: An overview of
490 present practices and future trends. Veterinary Journal 154, 135-147.
491 Enoe, C., Mousing, J., Schirmer, A.L., Willeberg, P., 2002. Infectious and rearing-system
492 related risk factors for chronic pleuritis in slaughter pigs. Preventive Veterinary
493 Medicine 54, 337-349.
494 European Community, 2004. Regulation (EC) No 854/2004 of the European Parliament
495 and of the Council of 29 April 2004 laying down specific rules for the organization
496 of official controls on products of animal origin intended for human consumption.
497 OJ L 139,, 206–320.
498 Eze, J.I., Correia-Gomes, C., Borobia-Belsué, J., Tucker, A.W., Sparrow, D., Strachan,
499 D.W., Gunn, G.J., 2015. Comparison of respiratory disease prevalence among
500 voluntary monitoring systems for pig health and welfare in the UK. PLoS ONE 10.
501 Fablet, C., Dorenlor, V., Eono, F., Eveno, E., Jolly, J.P., Portier, F., Bidan, F., Madec, F.,
502 Rose, N., 2012. Noninfectious factors associated with pneumonia and pleuritis in
503 slaughtered pigs from 143 farrow-to-finish pig farms. Preventive Veterinary
504 Medicine 104, 271-280.
505 Fidaner, I.B., Cankorur-Cetinkaya, A., Dikicioglu, D., Kirdar, B., Cemgil, A.T., Oliver,
506 S.G., 2015. CLUSTERnGO: A user-defined modelling platform for two-stage
507 clustering of time-series data. Bioinformatics 32, 388-397.
508 Hafen, R.P., 2010. Local regression models: Advancements, applications, and new
509 methods. Purdue University.
510 Hennig, C., Meila, M., Murtagh, F., Rocci, R., 2015. Handbook of Cluster Analysis. CRC
511 Press.

512 Leffondré, K., Abrahamowicz, M., Regeasse, A., Hawker, G.A., Badley, E.M., McCusker,
513 J., Belzile, E., 2004. Statistical measures were proposed for identifying
514 longitudinal patterns of change in quantitative health indicators. Journal of Clinical
515 Epidemiology 57, 1049-1062.
516 Merialdi, G., Dottori, M., Bonilauri, P., Luppi, A., Gozio, S., Pozzi, P., Spaggiari, B.,
517 Martelli, P., 2012. Survey of pleuritis and pulmonary lesions in pigs at abattoir
518 with a focus on the extent of the condition and herd risk factors. Veterinary
519 Journal 193, 234-239.
520 Neumann, E.J., Hall, W.F., Stevenson, M.A., Morris, R.S., Ling Min Than, J., 2014.
521 Descriptive and temporal analysis of post-mortem lesions recorded in slaughtered
522 pigs in New Zealand from 2000 to 2010. New Zealand Veterinary Journal 62, 110-
523 116.
524 Niedermeyer, E., Schomer, D.L., da Silva, F.H.L., 2011. Niedermeyer's
525 Electroencephalography: Basic Principles, Clinical Applications, and Related Fields.
526 Wolters Kluwer Health/Lippincott Williams & Wilkins.
527 Peng, R., 2008. A Method for Visualizing Multivariate Time Series Data. Journal of
528 Statistical Software 25, (Code Snippet) 1-17.
529 Räsänen, T., Kolehmainen, M., 2009. Feature-based clustering for electricity use time
530 series data. Lecture Notes in Computer Science (including subseries Lecture Notes
531 in Artificial Intelligence and Lecture Notes in Bioinformatics). Kuopio, 401-412.
532 Revelle, W., 2015. psych: Procedures for Psychological, Psychometric, and Personality
533 Research. Northwestern University, Evanston, Illinois, USA,.
534 Sanchez-Vazquez, M.J., 2013. Epidemiological Investigations Utilizing Industry Abattoir
535 Data – A study in Finishing Pigs. Utrecht University, Utrecht, The Netherlands.
536 Sanchez-Vazquez, M.J., Nielen, M., Gunn, G.J., Lewis, F.I., 2012. Using seasonal-trend
537 decomposition based on loess (STL) to explore temporal patterns of pneumonic
538 lesions in finishing pigs slaughtered in England, 2005-2011. Preventive Veterinary
539 Medicine 104, 65-73.

Schuh, M., Köfer, J., Fuchs, K., 2000. Installation of an information feedback system for control of animal health - Frequency and economical effects of oraan lesions in slaughter pigs. Wiener Tierarztliche Monatsschrift 87, 40-48.

Tabachnick, B.G., Fidell, L.S., 2006. Using Multivariate Statistics (5th Edition).

Torgo, R., 2010. Data Mining with R: Learning with Case Studies. Chapman \& Hall/CRC.

Vial, F., Reist, M., 2014. Evaluation of Swiss slaughterhouse data for integration in a syndromic surveillance system. BMC Veterinary Research 10.

Wang, X., Smith-Miles, K., Hyndman, R., 2009. Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. Neurocomputing 72, 2581-2594.

Wang, X., Smith, K., Hyndman, R., 2006. Characteristic-based clustering for time series data. Data Mining and Knowledge Discovery 13, 335-364.

Wang, X., Smith, K.A., Hyndman, R.J., 2005. Dimension reduction for clustering time series using global characteristics. In: Sunderam, V.S., Albada, G.D., Sloot, P.M.A., Dongarra, J.J. (Eds.), 5th International Conference on Computational Science - ICCS 2005, Atlanta, GA, 792-795.

Appendix A. Supplementary data

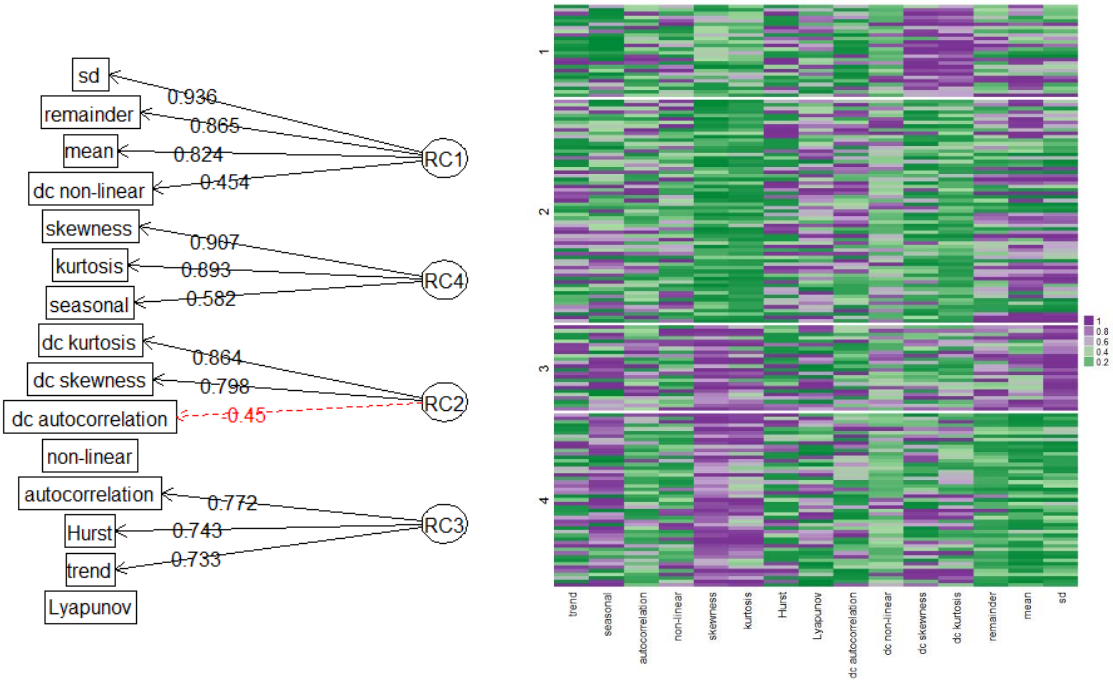

Fig. A1. Factor analysis path diagram of pneumonia (left pane) and allocation to four clusters (right pane) in the validation exercise (163 farms). For explanation: see Figure 3.

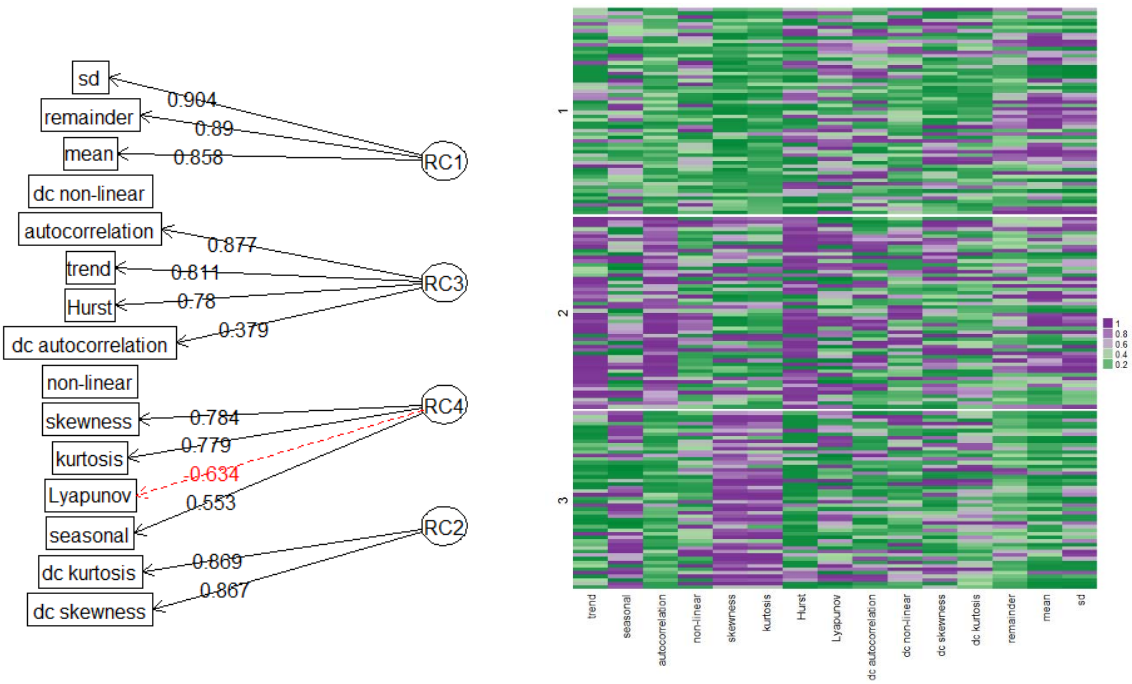

Fig. A2. Factor analysis path diagram of pleuritis (left pane) and allocation to three clusters (right pane) in the validation exercise (163 farms). For explanation: see Figure 3.

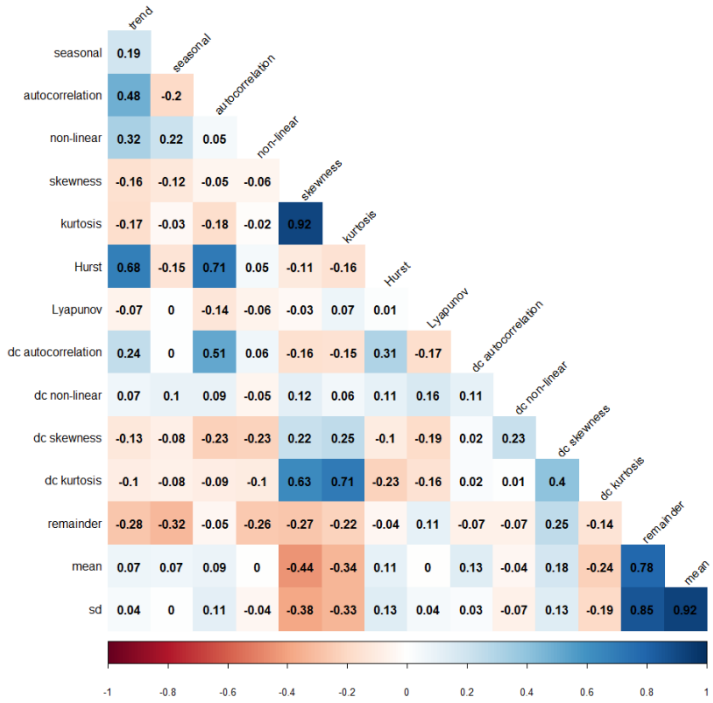

Fig. A3. Correlation matrix for the characteristics of percentage pneumonia

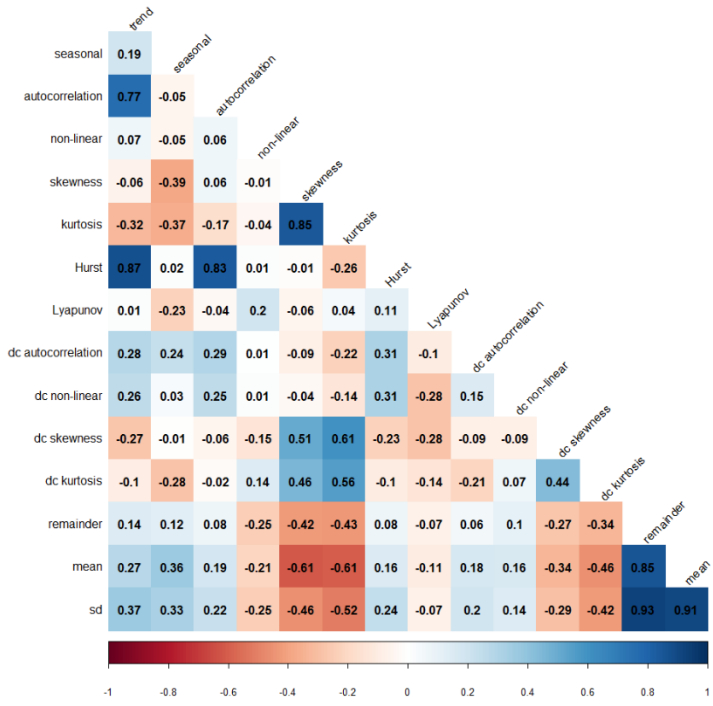

Fig. A4. Correlation matrix for the characteristics of percentage pleuritis