# Clustering of farms based on slaughterhouse health aberration data

Ina Hulsegge
Karel de Greef

WAGENINGEN
UNIVERSITY & RESEARCH

# Clustering of farms based on slaughterhouse health aberration data

Ina Hulsegge
Karel de Greef

This research was conducted by Wageningen Livestock Research, commissioned and funded by the Ministry of Economic Affairs, within the framework of Policy Support Research theme 'Feed4Foodure' – share program B nutrition, Intestinal health and Immunity (VDI) project number: BO-22.04-002-001
Breed4Food project number: BO-22.04-011-001

Wageningen Livestock Research
Wageningen, June 2017

Report 1023

WAGENINGEN
UNIVERSITY & RESEARCH

Ina Hulsegge, Karel de Greef, 2017. *Clustering of farms based on slaughterhouse health aberration data.* Wageningen Livestock Research, Report 1023.

Samenvatting

Clustering van bedrijven in slachtlijnafwijkingen

Monitoringsdata aan de slachtlijn kunnen gebruik worden om verschillen tussen bedrijven qua diergezondheid meer systematisch op te sporen. Van twee slachterijen zijn een half miljoen dieren geanalyseerd op de afwijkingen van longen, lever, huid en poten. De gegevens zijn afkomstig van 44 bedrijven en meer dan 5000 leveringen. Deze bedrijven zijn geclusterd op incidentie van afwijkingen, maar ook op bv seizoensgevoeligheid en voorspelbaarheid van het vóórkomen van afwijkingen. De resultaten wijzen op systematische variatie tussen bedrijven die tot nu toe nog niet beschikbaar was. Een clustering als deze kan helpen bij het onderkennen van riscofactoren voor slachtafwijkingen. Als de bedrijfskenmerken gekoppeld kunnen worden aan de bedrijfscategorisering, kan onderzocht worden welke bedrijfsfactoren kenmerkend zijn voor de betreffende clusters.

Summary

Clustering of farms based on slaughterhouse health aberration data.

Large amounts of data from meat inspections can be used as a tool to support interventions for improvement of herd health. We applied time-series analyses on 3.5 years of meat inspection data of two pig slaughterhouses to identify differences in health aberration patterns over time at farm-level. A negligibly evidence of seasonality and a substantial trend pattern in percentage aberrations over time were identified. Differences exist in percentage health aberrations between the farms and months. This distinction is more elaborate than just grouping farms on basis of aberration incidence.

This report can be downloaded for free at http://dx.doi.org/10.18174/415138 or at www.wur.nl/livestock-research (under Wageningen Livestock Research publications).


P.O. Box 338, 6700 AH Wageningen, The Netherlands, T +31 (0)317 48 39 53,
E info.livestockresearch@wur.nl, www.wur.nl/livestock-research. Wageningen Livestock Research is part of Wageningen University & Research.

# Table of contents

# Samenvatting

Clustering van bedrijven in slachtlijnafwijkingen

Vanuit de routinematige keuring van slachtvarkens aan de slachtlijn komen veel data beschikbaar. Deze monitoring wordt naar de individuele varkenshouder teruggekoppeld, maar is er meer rendement uit deze data te halen? Dit was de aanleiding om dit onderzoek te starten; kunnen deze data gebruikt worden om verschillen tussen bedrijven qua diergezondheid meer systematisch op te sporen?.

Van twee slachterijen zijn van 3,5 jaar slachthuisgegevens van ruim een half miljoen dieren geanalyseerd in een geanonimiseerde dataset. De nadruk lag hierbij op de routinematig verzamelde slachtafwijkingsdata: afwijkingen van longen, lever, huid en poten. De gegevens zijn afkomstig van 44 bedrijven en meer dan 5000 leveringen. Na een beschrijvende analyse van de data is met een 'Seasonal-trend-decomposition procedure' per kenmerk een onderscheid gemaakt tussen seizoen, meerjarige trend en restvariatie. Vervolgens is een set van 10 beschrijvende statistische kenmerken (waaronder gemiddelde, spreiding, autocorrelatie, seizoen effect, trend, skewness, kurtosis) met behulp van factoranalyse omgezet in een beperkt aantal onafhankelijke metavariabelen. De bedrijven zijn vervolgens met 'Characteristics based clustering' geclusterd in groepen van bedrijven met vergelijkbare kenmerken. Deze bedrijven zijn daarmee niet alleen geclusterd op incidentie van afwijkingen, maar ook op bv seizoensgevoeligheid en voorspelbaarheid van het vóórkomen van afwijkingen.

De resultaten wijzen op systematische variatie tussen bedrijven die tot nu toe nog niet beschikbaar was. Een clustering als deze kan helpen bij het onderkennen van riscofactoren voor slachtafwijkingen. Er kan verschil gemaakt worden tussen bedrijven die weliswaar vergelijkbare afwijkingspercentages hebben, maar waarvan de onderliggende factoren verschillend zijn. Deze onderlinge riscofactoren (bv (voeding)management of huisvestingsfactoren) kunnen met een meer uitgebreide analyse (waarin bedrijfskenmerken meegenomen worden) opgespoord worden. Als de bedrijfskenmerken gekoppeld kunnen worden aan de bedrijfscategorisering, kan onderzocht worden welke bedrijfsfactoren kenmerkend zijn voor de betreffende clusters.

# Summary

Clustering of farms based on slaughterhouse health aberration data.
A large amount of data is collected routinely in meat inspection in pig slaughterhouses. Meat inspection data can be used to inform farmers on the health status of their herd (benchmarking), so they can plan health management and monitor the effectiveness of treatment and prevention strategies helping farmers to make better decisions, improving performance and protecting pig welfare. This work shows an example of the utility of large amounts of data from meat inspections as a tool to support interventions for improvement of herd health. We applied time-series analyses on 3.5 years (January 2011 - July 2014) of meat inspection data to identify differences in health aberration patterns over time at farm-level.

A total of 511,645 pigs have been assessed across 2 pig slaughterhouses over the study period, submitted from 44 farms in 5,149 batches. A negligibly evidence of seasonality and a substantial trend pattern in percentage aberrations over time were identified. Differences exist in percentage health aberrations between the farms and months. Characteristic based clustering was able to cluster time series of meat inspection data of farms using just a set of derived statistical characteristics. This distinction is more elaborate than just grouping farms on basis of aberration incidence. The derived clusters can be analysed to detect similarities between the farms in these clusters in (nutritional) management and/or housing conditions causing the health aberrations.

# 1 Introduction

The proliferation of database management systems has contributed to recent massive gathering of all sorts of information in the animal production chain. Large amounts of data are collected frequently and automatically by these systems but most of them remain underemployed, whereas they could be of value. Some of these databases are implemented on basis of control-oriented programmes such as meat inspection. 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 the function of this meat inspection has been 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 the improvement of farm management (Schuh *et al.* 2000). Meat inspection data can be used to inform farmers on the health status of their herd. 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 farm averages. Understanding the data structure (such as temporal patterns) of aberrations in meat inspection data may provide important information beyond average incidence figures. Time series analysis aim to provide a concise description of data through time (Diggle 1990), usually by exploring both time trend and seasonal pattern.
The object of this study was to explore the potential of this long-term meat inspection data to cluster farms into groups with comparable health aberrations over time. 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

## 2.2 General

Post mortem meat inspection data of carcass and organs are recorded on every slaughtered pig in The Netherlands. The inspection procedures are described in considerable detail in Regulation EC no. 854/2004 (European Community 2004). Meat inspection data for this study were provided by 2 slaughterhouses and collected between January 2011 and August 2014. The dataset contained almost 5 million records with information on 5 aberrations: liver aberrations, pneumonia, pleurisy, skin- and leg aberrations.

## 2.3 Study sample

Criteria were developed to get a sub-dataset for method development and analysis. August 2014 was removed since it did not comprise the entire month. Farms with less than 87 batches (less than 1 batch per 2 weeks on average) in the entire data recording period were also removed, as were farms which did not deliver pigs every month. Batches with less than 10 pigs were also removed. The resulting study sample contained information of 511,645 pigs, submitted from 44 farms in 5,149 batches. Across the two slaughterhouses, the mean number of pigs assessed per month was 11,899 (95% CI: 11,617- 12,180), with a mean of 120 (95% CI: 117-122) batches. The mean number of delivered pigs per month per batch increased slightly over time.
A total of 56,190 pigs showed aberrations, which represent 11% of the pigs selected for further analysis.
The percentage of liver aberrations, pneumonia, pleurisy and skin- and legs aberrations in the batches are present in Table 1. The analysis is principally batch based – records were created containing batch averages. The percentage of each aberration in each record was computed as number of pigs in that batch with the aberrations divided by total number of pigs in that batch multiplied by 100.

Pneumonia was not observed in 11.9% of the batches. The mean prevalence of pneumonia in a batch was 8.8% (95% CI: 8.5%-9.0%), with a maximum up to 63.8%. The mean prevalence of pleurisy in a batch was 12.4% (95% CI: 12.1%-12.7%), with a maximum up to 61.6%. About 50% of the batches showed liver aberrations. Legs and skin aberrations were not observed in approximately 70% and 91% of the batches, respectively.

***Table 1*** *Percentage of 5 aberrations in the 5,149 batches.*

| Aberration | # Batches with percentage 0% (%) | Mean percentage (95% CI) in a batch | Sd percentage | Max percentage |
|---|---|---|---|---|
| Pneumonia | 615 (11.9%) | 8.76 (8.51–9.01)% | 9.10% | 63.83% |
| Liver | 2587 (50.2%) | 2.19 (2.06-2.33)% | 5.08% | 70.37% |
| Pleurisy | 375 (7.3%) | 12.42 (12.12-12.72)% | 10.91% | 61.64% |
| Legs | 3607 (70.1%) | 0.51 (0.48-0.54)% | 1.08% | 15.00% |
| Skin | 4662 (90.5%) | 0.19 (0.15-0.23)% | 1.33% | 44.90% |

# 3 Exploratory data analyses

## 3.1 Descriptive analyses

To visually summarise and compare percentage aberrations per batch by farm, boxplots were used.

## 3.2 Time series visual explorations

For exploratory purpose, batch percentage aberrations were aggregated for each month of study. The seasonal-trend decomposition procedure based on locally weighted regression (LOESS), known as "STL" (Cleveland *et al.* 1990), was used to decompose the time series in order to visualise temporal patterns. STL is a filtering procedure based on decomposing the time series into 3 additive components of variation: trend (Tt), seasonal (St), and remainder (Rt).

$$Yt = Tt + St + Rt$$

In this study, Yt specifically stands for monthly prevalence of pneumonia, liver aberrations, pleurisy, legs aberrations, skin aberrations or number of delivered pigs per month per batch; t is time in unit of month. The seasonal component is found by LOESS smoothing the seasonal sub-series of the overall time series. The seasonal values are removed, and the deseasonalised remainder smoothed to find the trend. The remainder component is the residuals from the seasonal plus trend fit. STL has six parameters that determine the degree of smoothing in trend and seasonal components:

- $n_p$ = the number of observations in each cycle of the seasonal component;
- $n_i$ = the number of passes through the inner loop;
- $n_o$ = the number of robustness iterations of the outer loop;
- $n_l$ = the smoothing parameter for the low-pass filter;
- $n_s$ = the smoothing parameter for the seasonal component; and
- $n_t$ = the smoothing parameter for the trend component.

To explore potential inter-annual variation of the seasonal component, the seasonal LOESS smoothing parameter ($n_s$) was set to estimate the seasonal effect based on 7 months (Cleveland *et al.* 1990) (Sanchez-Vazquez *et al.* 2012b). In this study for $n_i$ and $n_o$ a robust option was chosen being 1 and 5 respectively as also suggested by Sanchez-Vazquez (2012b). The other three parameters were predefined following the recommendations from Cleveland et al. (1990), being $n_p$ = 12 months, $n_l$ = 13 months and $n_t$ = 23 months. Further information on the method and parameters can be found in the original paper describing the STL method (Cleveland *et al.* 1990).

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 mvtsplot package (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.

## 3.3 Time series clustering using global characteristics

We used a three step method to group farms with comparable percentage of health aberrations over time (Figure 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 & Kolehmainen (2009). The measures summarized information of the time series, which capture the 'global picture' of the data. The measures used in this study were: *mean, standard deviation, trend, seasonality, serial correlation, skewness, kurtosis, chaos, nonlinearity,* and *self-similarity*. *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 *trend* is identified as a long term change in the means level, upward or downward. *Seasonality* is a repeating pattern over certain intervals of time. *Serial correlation* is a measure of the relationship between a point and itself over various time intervals. *Skewness* is a measure of the lack of symmetry. *Kurtosis* is a measure of whether the data are peaked or flat, relative to a normal distribution. A data set with high kurtosis tends to have a distinct peak near the mean, decline rather rapidly. Distributions having high kurtosis have fatter tails or more extreme values. Data sets with low kurtosis tend to have a flat top near the mean. Distributions having low kurtosis produces fewer and less extreme outliers than the normal distribution does. *Nonlinearity* measures the nonlinear behaviour. The presence of *chaos* is measured by Lyapunov exponent and the *self-similarity* of time series is also considered and measured using Hurst exponent. 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 are thoroughly explained by Wang et al. (2009) and the script in R provided by them is used in this study (Hyndman 2012).

In the second step, a factor analysis, using the function "principal" from the package "Pysch version: 1.5.8", (Revelle 2015), was performed to select a subset of measures that condensed the information present in the measures and provided the best description. We only kept the factors with an eigen-value greater than 1 (Tabachnick & 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, we used cluster analysis to identify clusters of similar patterns of measures selected by the factor analysis. In order to weight all characteristics equally, all characteristics values 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.

All the analyses were performed by R 3.2.3 (Team 2015).


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

# 4 Results

## 4.1 STL

Figure 2 shows the STL decomposition plots of the time series of percentage pneumonia, liver aberrations, pleurisy, leg- and skin aberrations, respectively. The raw data is fitted in the top of each panel. The STL seasonal, trend and remainder components are shown in the second, third and fourth layers of each of the five panels, respectively. The length of the bars on each side of the panels are an indication of the magnitude of the individual components. Each bar has the same length, but plotted on different scales. The grey bar in the trend layer is much larger than the one in the data layer, indicating that the variation in the trend component is only a small part of the variation in the data layer. In other words, if we shrunk the trend such that the grey bar became the same size as that of the data, the range of variation on the shrunk trend would be similar to but much smaller than that on the data layer. The range of variation of the seasonal layer is smaller than that on the data layer, but larger than that on the trend layer.

Overall, especially the relatively small magnitude of the trends is clear. Furthermore, the seasonality of pleurisy incidence is considerably smaller than that of pleurisy and leg abnormalities.
A monotonic downward trend from January 2011 to October 2012, followed by an upward trend was observed for percentage pneumonia, liver aberrations and legs aberrations. For percentage pleurisy the largest decrease in the trend was observed between May 2012 and July 2013. The seasonal component observed showed an increase in percentage pneumonia in autumn and winter, and a decline in spring and summer. For percentage liver aberrations a peak was observed in August. Pleurisy incidence had a peak in spring and a clear decline in summer.

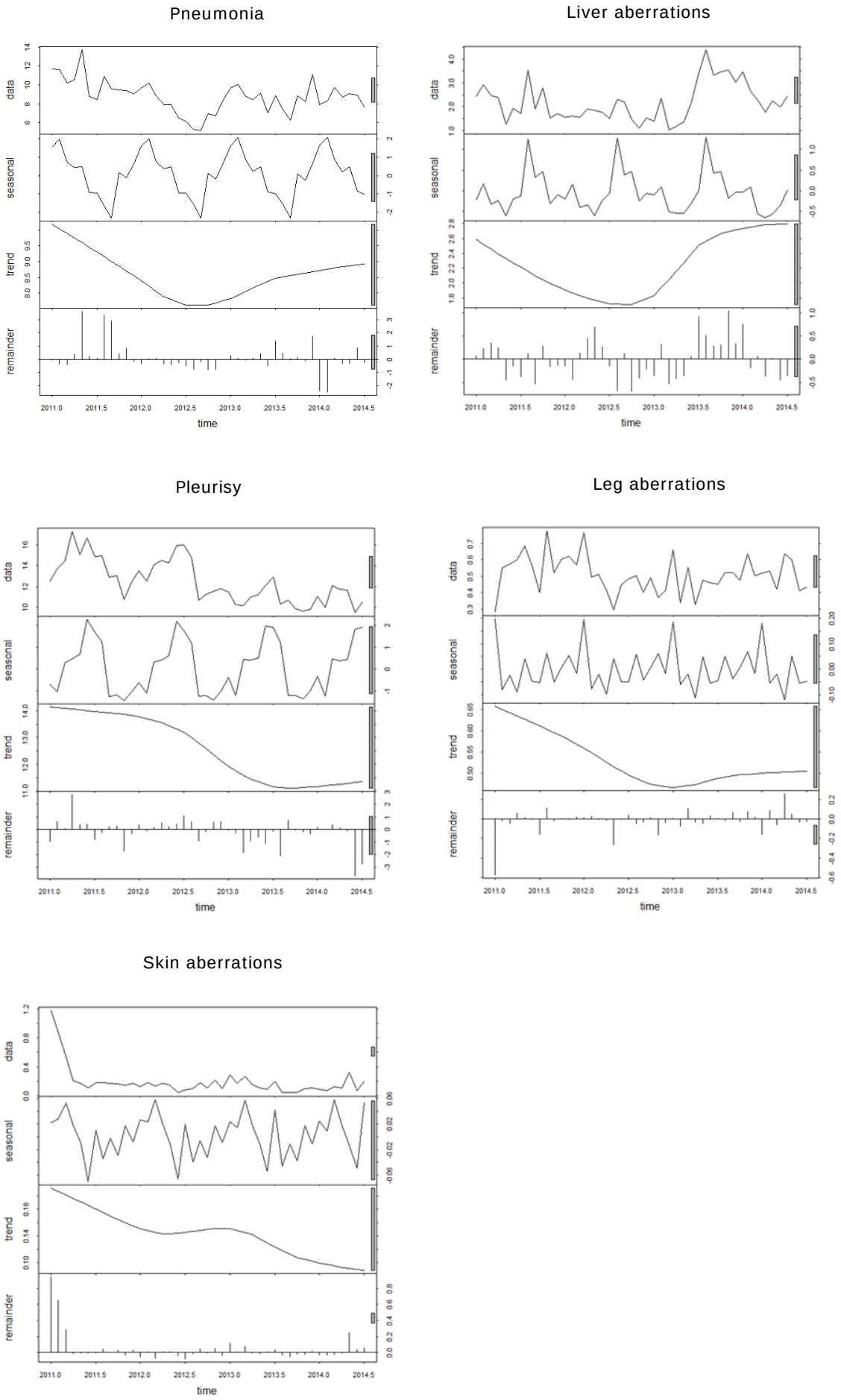



***Figure 2*** *STL decomposition of percentage pneumonia, liver aberrations, pleurisy, legs- and skin aberrations by month (January 2011 - July 2014), into seasonal, trend and remainder components (long bar at the right site indicates a small effect).*

## 4.2 Percentage aberrations at farm level

Data from January 2011 to July 2014 for the 44 farms is depicted as an multivariate time series (MVTS) plot of monthly averages in Figure 3. The monthly percentage of aberrations for the farms varied by farm and months. The box at the right side of each panel shows that differences exist in percentage aberrations between farms, especially for pneumonia (mean 0.8 to 25.1%), pleurisy (mean 1.4 to 24.0%) and liver aberrations (mean 0.0 to 13.0%). These aberrations also vary from month to month over the 44 farm considered (bottom panel). There were farms with the percentage aberrations consistent over time (nearly all months shows only green or only purple colour). The flat line on the bottom panel of skin aberrations indicates that the median values across all the time series for each time point was zero. Table 1 shows that skin aberrations were not observed in approximately 91% of the batches.

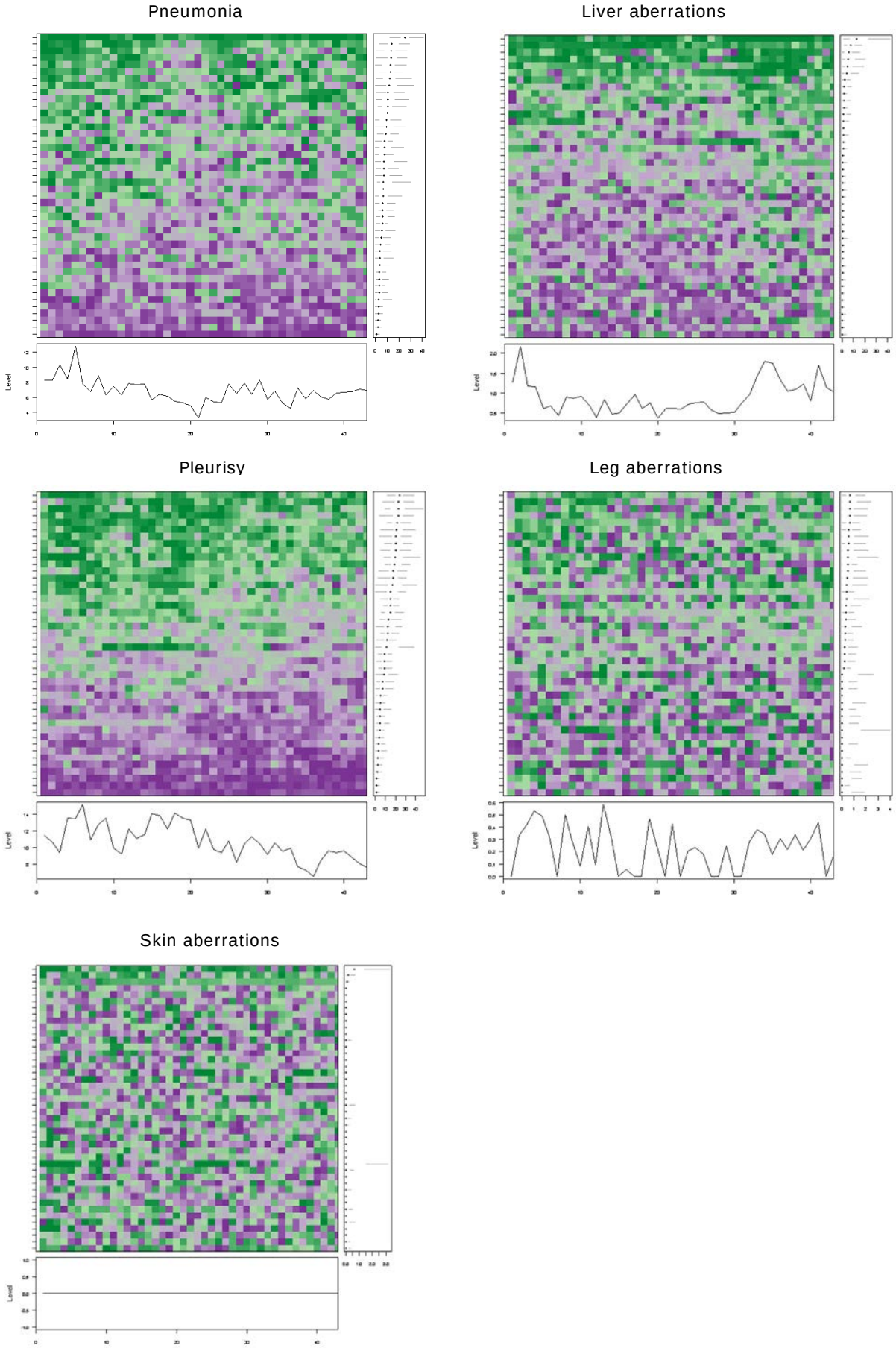



***Figure 3*** *Multivariate time series plot of percentage aberrations for 44 farms. The purple to green palette represents variation in percentage aberrations (purple represents low percentages; green high percentages). The right panel presents summary statistics of percentage aberrations for each 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.*

## 4.3 Time series clustering using global characteristics

### 4.3.1 Pneumonia

Exploratory factor analysis of percentage pneumonia reduced the 15 global characteristics to four factors that explained 65% of the variance. The four most informative characteristics were: "*mean*", "*seasonal*", "*autocorrelation*" and "*dc kurtosis*" (Fig. 4). Cluster analysis grouped the 44 farms into three clusters based on these four most informative characteristics. Categorization of the three clusters data characteristics are shown in Figure 4. Farms in cluster 1 showed high mean values (mostly green values) with large variability. The trend and seasonally adjusted time series measurements showed low values for kurtosis ("*dc kurtosis*"; purple values), having a flat top near the mean and produces fewer and less extreme outliers than does the normal distribution. Cluster 2 has low mean with little variability. The "*dc kurtosis*" of this cluster was high, having a distinct tall peak near the mean, decline rather rapidly and have fatter tails or more extreme values. Cluster 3 had very low mean and variability. The "*dc kurtosis*" value for this cluster was low meaning that the trend and seasonally adjusted time series produces fewer and less extreme outliers than does the normal distribution. The factor analysis suggested "*seasonal*" 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 Figure 4. Farms belonging to cluster 1 fluctuated the most with season, from -2.4% in September to 2.2% in December. For cluster 2 the lowest value of the STL 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.

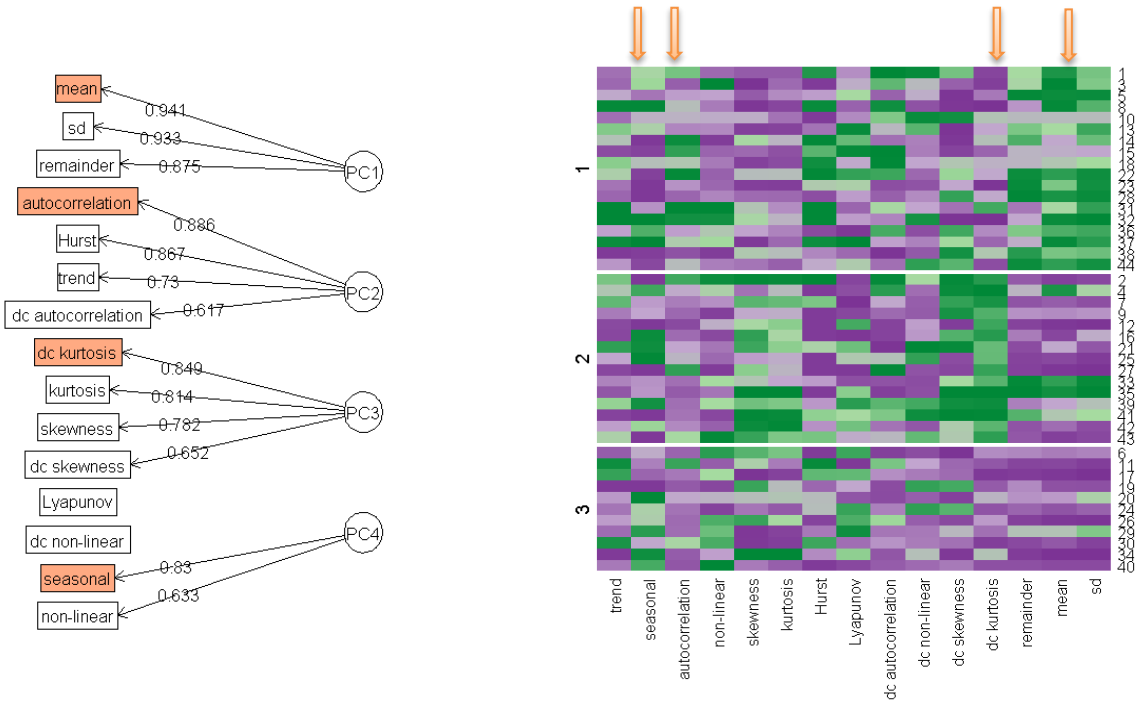



***Figure 4*** *Factor analysis path diagram (left). The square boxes are the observed variables, and the ovals are the unobserved factors. The straight arrows are the loadings, the correlation between the factor and the observed variable(s). Only the biggest loadings per item are shown. Characteristics summary on cluster basis (right) for percentage pneumonia. Green indicates high values and purple indicates low values (scaled values). (Arrows: the selected characteristics; dc: remaining time series after detrending and deseasonalising; left axis: cluster number; right: axis farm number).*

### 4.3.2 Liver aberrations

Dimension reduction of the global characteristics showed that four characteristics were sufficient to describe the percentage liver aberrations of the 44 farms. The four most informative characteristics were: "*standard deviation*", "*dc non-linear*", "*dc autocorrelation*" and "*kurtosis*" and together they explained 72% of the total variance of the 15 global characteristics (Fig. 5). Using these four characteristics the farms were classified into five clusters (Fig. 5).
Cluster 1 regrouped farms with medium values for mean and standard deviation. This cluster had low (purple) values for "*kurtosis*", indicating a flat distribution and high (green) values for "*dc autocorrelation*", indicating that the percentage liver aberrations in a month is highly related to the percentage in previous month. This cluster had low value for *nonlinearity*. Cluster 2 had very low level of mean percentage liver aberrations with low variability and almost no relation between aberrations inconsecutive months ("*low autocorrelation*"). The time series in cluster 3 were skewed, had high kurtosis and were nonlinear. High kurtosis indicates that the distribution of time series were more peaked than a normal distribution. This cluster had medium mean values with some high variability. The time series in cluster 4 were skewed and had high kurtosis. The values for *mean* and *standard deviation* were low and *self-similarity* was almost not present in this cluster (purple values for the Hurst exponent). This cluster had low values for "*standard deviation*", "*dc non-linear*", "*dc autocorrelation*". Cluster 5 regrouped farm with high mean values, large variability and general excessive volatility.

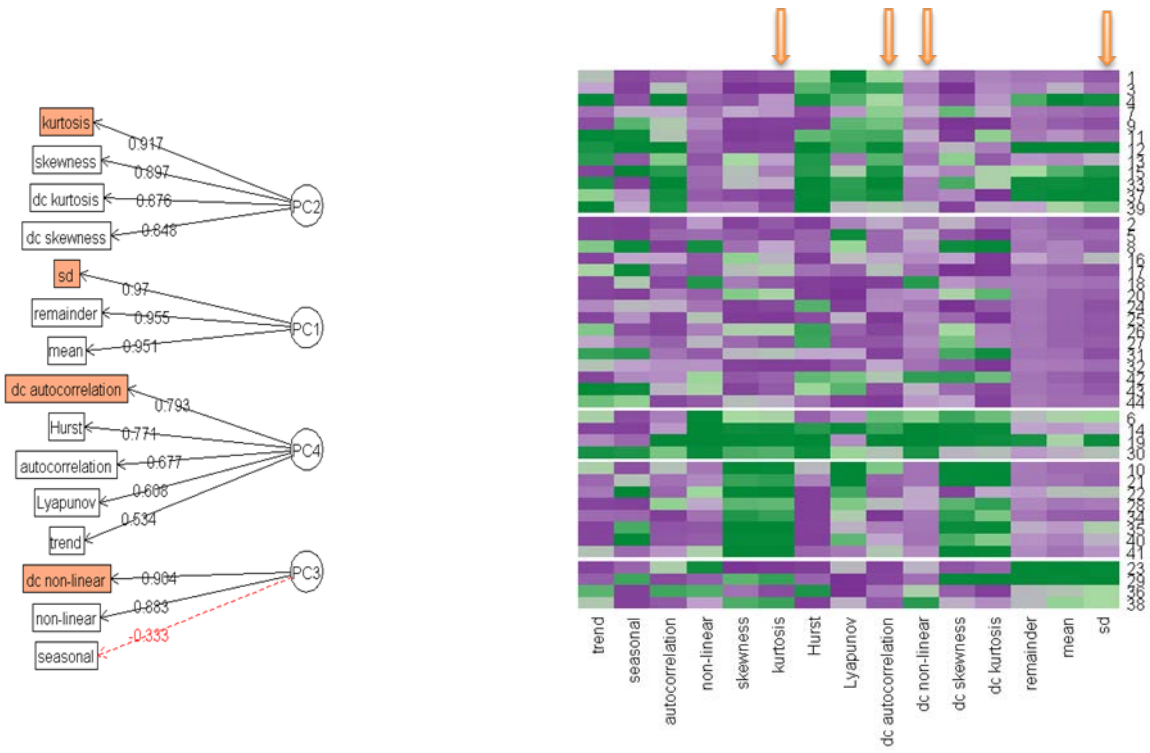



***Figure 5*** *Factor analysis path diagram of liver aberrations (left). The square boxes are the observed variables, and the ovals are the unobserved factors. The straight arrows are the loadings, the correlation between the factor and the observed variable(s). Only the biggest loadings per item are shown. Characteristics summary on cluster basis (right) for percentage liver aberrations. Green indicates high values and purple indicates low values. (scaled values). (Arrows: the selected characteristics; dc: remaining time series after detrending and deseasonalising; left axis: cluster number; right: axis farm number).*

### 4.3.3 Pleurisy

Exploratory factor analysis of percentage pleurisy reduced the 15 global characteristics to three factors ("*hurst*", "*lyaponov*" and "*mean*") that explained 62% of the variance (Fig. 6). Cluster 1 (Fig. 6) contains time series with medium mean values and variability and strong trend. These time series had also highly regular fluctuations over time (high self-similarity) since the Hurst exponent was extremely high (green colour) and showed no chaotic behaviour (extremely low "lyaponov" values). Cluster 2 contain farms with low percentage pleurisy and variability and high values for skewness and kurtosis. The time series showed no trend or seasonal effect and self-similarity was almost not present. Farms in cluster 3 showed high mean values (mostly green values) with large variability. The Lyapunov exponent quantified large amount of chaos in the time series.

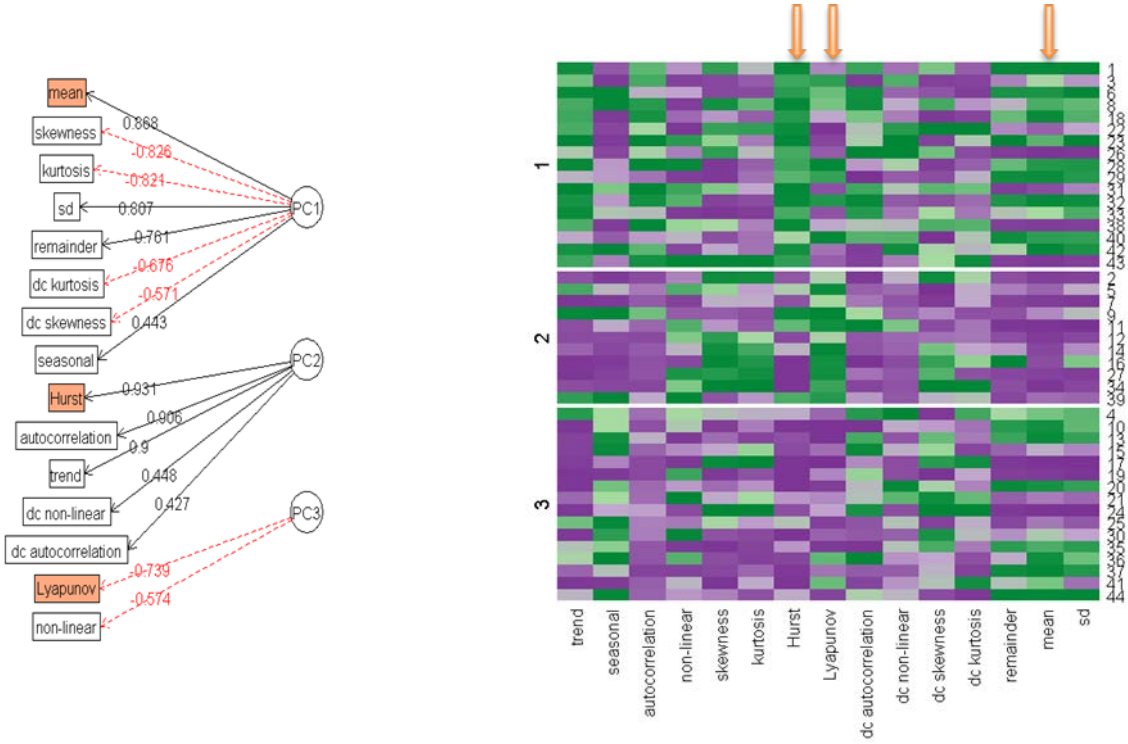



***Figure 6*** *Factor analysis path diagram of pleurisy (left). The square boxes are the observed variables, and the ovals are the unobserved factors. The straight arrows are the loadings, the correlation between the factor and the observed variable(s). Only the biggest loadings per item are shown. Characteristics summary on cluster basis (right) for percentage pleurisy. Green indicates high values and purple indicates low values. (scaled values). (Arrows: the selected characteristics; dc: remaining time series after detrending and deseasonalising; left axis: cluster number; right: axis farm number).*

### 4.3.4 Skin aberrations

The Lyaponov exponent could not be calculated for 20 farms. Therefore this characteristics was excluded for further analysis. The most informative characteristics for percentage skin aberrations were “non-linear”, “dc skewness” and “sd”, explaining 69% of the variance (Fig. 7). Based on these three factors the farms were grouped into two clusters (Fig. 7). The first cluster had low values for almost all informative characteristics. The second cluster contain very irregular time series data (high (green) values for “dc skewness”. The kurtosis in this cluster is also high.

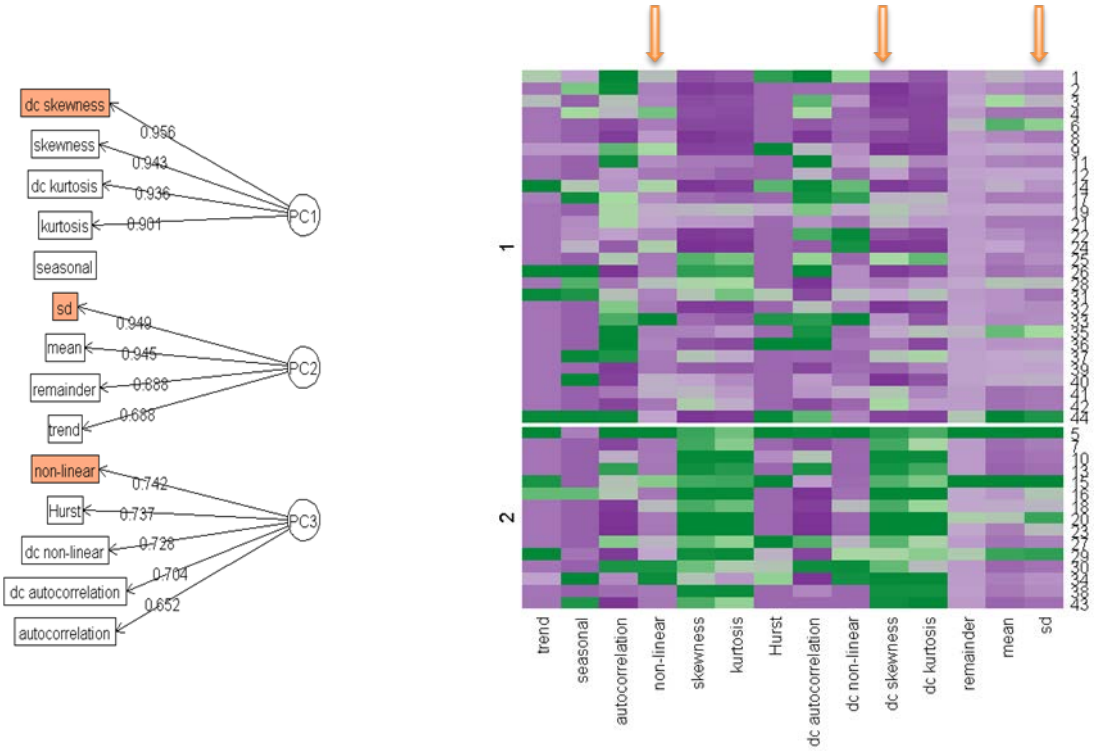



***Figure 7*** *Factor analysis path diagram of skin aberrations (left). The square boxes are the observed variables, and the ovals are the unobserved factors. The straight arrows are the loadings, the correlation between the factor and the observed variable(s). Only the biggest loadings per item are shown. Characteristics summary on cluster basis (right) for percentage skin aberrations. Green indicates high values and purple indicates low values. (scaled values). (Arrows: the selected characteristics; dc: remaining time series after detrending and deseasonalising; left axis: cluster number; right: axis farm number).*

### 4.3.5 Leg

The most informative characteristics for percentage skin aberrations were “seasonal”, “dc skewness”, “dc auto correlation” and “sd”, explaining 69% of the variance. These four characteristics grouped the time series of 44 farms into 4 clusters (Fig. 8). Cluster 1 showed low to medium values for most characteristics, but very high autocorrelation value. This is also applicable to cluster 2. However the autocorrelation value of this cluster is low. The time series in cluster 3 showed no trend or seasonal effect. The farms in this cluster had high percentage of leg aberration. The time series data in this cluster were highly skewed and had high kurtosis. Cluster 4 showed medium values for most characteristics. This cluster showed a seasonal pattern and an high variability.

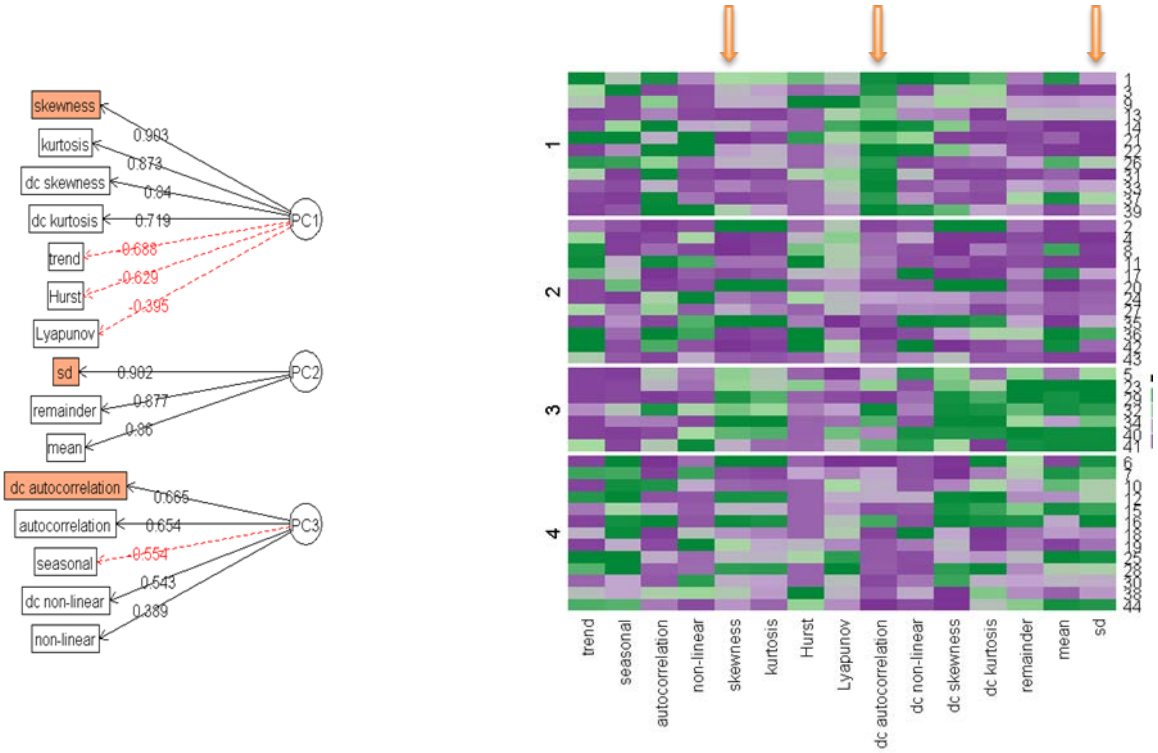



***Figure 8*** *Factor analysis path diagram of leg aberrations (left). The square boxes are the observed variables, and the ovals are the unobserved factors. The straight arrows are the loadings, the correlation between the factor and the observed variable(s). Only the biggest loadings per item are shown.* Characteristics summary on cluster basis (right) for percentage leg aberrations. Green indicates high values *and purple indicates low values. (scaled values). (Arrows: the selected characteristics; dc: remaining time series after detrending and deseasonaling; left axis: cluster number; right: axis farm number).*

# 5 Discussion

This work shows an example of the utility of large amount of data from meat inspections as a supporting tool to identify underlying factors of health aberrations. More than 3½ years of historical meat inspection data were used to explore the potential of such data to cluster farms into groups with comparable health aberrations over time. The data were analysed using time series analysis.
Time series analyses aim to provide concise description of data correlated through time (Diggle 1990). Visualisation and exploratory analyses are important to understanding the complexity of serially correlated data. In this study we used the exploratory method STL as filtering procedure for decomposing time series into additive components of variation (trend, seasonal and the remainder) (Cleveland *et al.* 1990). The STL has good visualization capabilities (Hafen *et al.* 2009). This method was also used by Sanchez-Vazquez *et al*. to analyses temporal variations of pneumonic lesions present in slaughtered pigs in England (Sanchez-Vazquez *et al.* 2012b) and to monitoring Ascaris suum related liver pathologies in English abattoirs (Sanchez-Vazquez *et al.* 2012a). The multivariate time series tool mvtsplot was used for visualisation the multivariate time series data (Peng 2008). Mvtsplot become useful when several, more than five or six, time series data need to be compared. This plot provides quick insight if values were high or low, before using other analysis.

*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 example, Leffondré et al.(2004) used this method for identifying patters of change in quantitative human health indicators. Räsänen and Kolehmainen (2009) used feature based clustering for electricity use time series data. We applied this method to group farms based on similar statistical characteristics of meat inspection data. In this, we used the set of characteristics as proposed by Wang et al. (2005; 2006) that contains measures of *trend, seasonality, periodicity, serial correlation, skewness, kurtosis, chaos, nonlinearity, and self-similarity* to represent time series. Various criteria can be used to select the most relevant characteristics of the data set. We used factor analysis as search mechanism to find the best selection from the characteristics set that were used as the clustering inputs, as also was suggested by Leffondré et al. (2004). The correlation matrix between the characteristics indicated that several characteristics are highly correlated (data in Appendix). The latter 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).

Present study shows that Characteristic based clustering is able to cluster time series of meat inspection data of farms using a set of derived statistical characteristics. The main advantages of this method are its ability to reduce the dimensionality of original time series, the low sensitivity to missing values and the capability to handle different lengths of time series (Räsänen & Kolehmainen 2009). Furthermore, the results are interpretable (Fulcher & Jones 2016). The performance of the Characteristics clustering approach reported here is based on an idealised dataset (no missing values and considerable frequency of batches). The study should be validated on the whole dataset verify the feasibility and effectiveness of the approach. A closer look on the results of this test-set reveals that 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.

From the perspective of practical application of the method, questions arise as to whether farms in different clusters also structurally deviate at farm-level. Either in (nutritional) management practices of in environmental (housing and ventilation) factors. Alban et al. (2015) reported significant differences in the prevalence of the various lesions – measured as recordings by meat inspectors – between organic/free-range and conventionally raise pig production systems in Denmark. Present method may offer the reverse situation: the statistical grouping of farms may point at similarities within groups or differences between groups that are causally related to the incidence of health aberrations. It is known that increased prevalence of disease conditions e.g. pleurisy, are correlated

with decreased carcass weight (Maes *et al.* 2001; Martínez *et al.* 2007). Increased herd size was found to be a risk factor for pleurisy in several studies (Enoe *et al.* 2002; Fablet *et al.* 2012). On availability of adequate data, comparing the clusters with regard to their housing and management characteristics is an obvious next step in studying added value of these clustered slaughterhouse data. The cluster information bears the promise to reveal more data derived contrasts between seemingly similar farms, beyond comparing farms on basis of averages alone. Current analysis confirms the view that there are large differences in the health status of animals (based on slaughterhouse detected health aberrations) between individual farms. The challenge is to identify the structural factors causing this.

# 6 Conclusions

The stepwise analysis of the optimised (cropped) slaughterhouse dataset reveals structured variation among farms in incidence of health abnormalities. The applied method groups them into clusters of ‘similar’ farms beyond clustering them just on basis of observed incidence of aberrations. The differences between these clusters likely point at systematic differences between individual farms. Analysis of these differences potentially could result in more effective mitigation strategies. This should encourage increasing use of meat inspection data, even though these datasets are large and complex.

Relating the identified clusters to (in this analysis not available) farm characteristics is a key step in this. Before that, the developed method needs verification in a less ideal dataset

# 7 Implication

Statistical clustering of farms based om slaughterhouse derived health aberration data into groups of seemingly comparable farms promises an aid in the identification of risk factors for especially lung health.

# 8 References


Alban L., Petersen J.V. & Busch M.E. (2015) A comparison between lesions found during meat inspection of finishing pigs raised under organic/free-range conditions and conventional, indoor conditions. Porcine Health Management 1, 4.

Charrad M., Ghazzali N., Boiteau V. & Niknafs A. (2014) Nbclust: An R package for determining the relevant number of clusters in a data set. Journal of Statistical Software 61, 1-36.

Cleveland R.B., Cleveland W.S., McRae J.E. & Terpenning I. (1990) STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics 6, 3-73.

Diggle P. (1990) Time series: a biostatistical introduction.

Edwards D.S., Johnston A.M. & Mead G.C. (1997) Meat inspection: An overview of present practices and future trends. Veterinary Journal 154, 135-47.

Enoe C., Mousing J., Schirmer A.L. & Willeberg P. (2002) Infectious and rearing-system related risk factors for chronic pleuritis in slaughter pigs. Preventive Veterinary Medicine 54, 337-49.

European Community (2004) Regulation (EC) No 854/2004 of the European Parliament and of the Council of 29 April 2004 laying down specific rules for the organization of official controls on products of animal origin intended for human consumption. OJ L 139,, 206–320.

Fablet C., Dorenlor V., Eono F., Eveno E., Jolly J.P., Portier F., Bidan F., Madec F. & Rose N. (2012) Noninfectious factors associated with pneumonia and pleuritis in slaughtered pigs from 143 farrow-to-finish pig farms. Preventive Veterinary Medicine 104, 271-80.

Fulcher B.D. & Jones N.S. (2016) Automatic time-series phenotyping using massive feature extraction. bioRxiv.

Hafen R.P., Anderson D.E., Cleveland W.S., MacIejewski R., Ebert D.S., Abusalah A., Yakout M., Ouzzani M. & Grannis S.J. (2009) Syndromic surveillance: STL for modeling, visualizing, and monitoring disease counts. BMC Medical Informatics and Decision Making 9.

Hyndman R.J. (2012) Measuring time series characteristics. URL http://robjhyndman.com/hyndsight/tscharacteristics/.

Leffondré K., Abrahamowicz M., Regeasse A., Hawker G.A., Badley E.M., McCusker J. & Belzile E. (2004) Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. Journal of Clinical Epidemiology 57, 1049-62.

Maes D.G., Deluyker H., Verdonck M., De Kruif A., Ducatelle R., Castryck F., Miry C. & Vrijens B. (2001) Non-infectious factors associated with macroscopic and microscopic lung lesions in slaughter pigs from farrow-to-finish herds. Veterinary Record 148, 41-6.

Martínez J., Jaro P.J., Aduriz G., Gómez E.A., Peris B. & Corpa J.M. (2007) Carcass condemnation causes of growth retarded pigs at slaughter. Veterinary Journal 174, 160-4.

Peng R. (2008) A Method for Visualizing Multivariate Time Series Data. Journal of Statistical Software 25, (Code Snippet) 1-17.

Räsänen T. & Kolehmainen M. (2009) Feature-based clustering for electricity use time series data. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 401-12, Kuopio.

Revelle W. (2015) psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois, USA,.

Sanchez-Vazquez M.J., Nielen M., Gunn G.J. & Lewis F.I. (2012a) National monitoring of Ascaris suum related liver pathologies in English abattoirs: A time-series analysis, 2005-2010. Veterinary Parasitology 184, 83-7.

Sanchez-Vazquez M.J., Nielen M., Gunn G.J. & Lewis F.I. (2012b) Using seasonal-trend decomposition based on loess (STL) to explore temporal patterns of pneumonic lesions in finishing pigs slaughtered in England, 2005-2011. Preventive Veterinary 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, 40-8.

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

Team R.D.C. (2015) R: A Language and Environment for Statistical Computing.

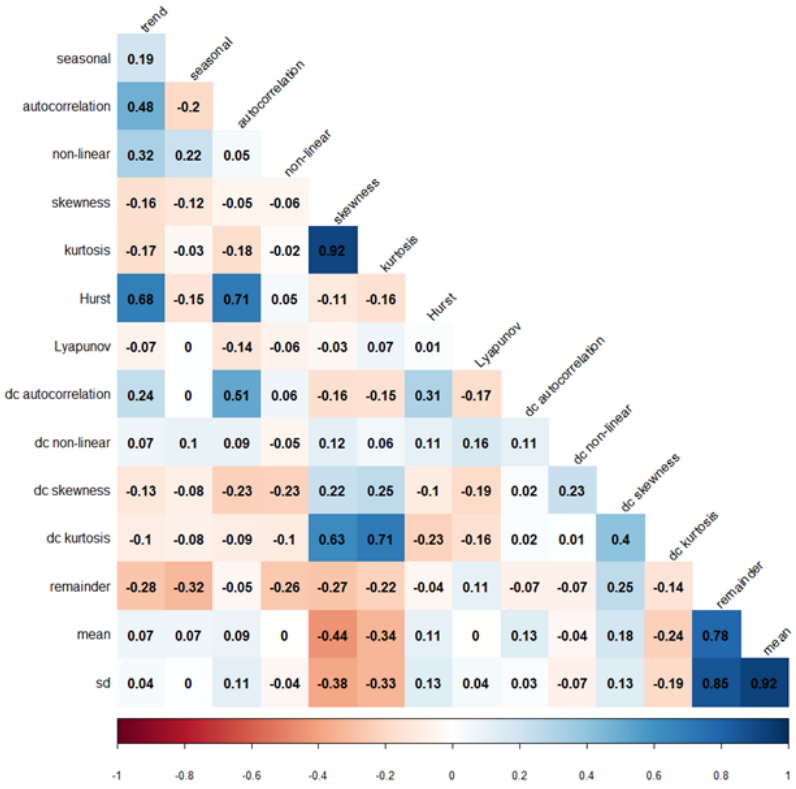
Torgo R. (2010) Data Mining with R: Learning with Case Studies. Chapman \& Hall/CRC.

Wang X., Smith-Miles K. & Hyndman R. (2009) Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. Neurocomputing 72, 2581-94.

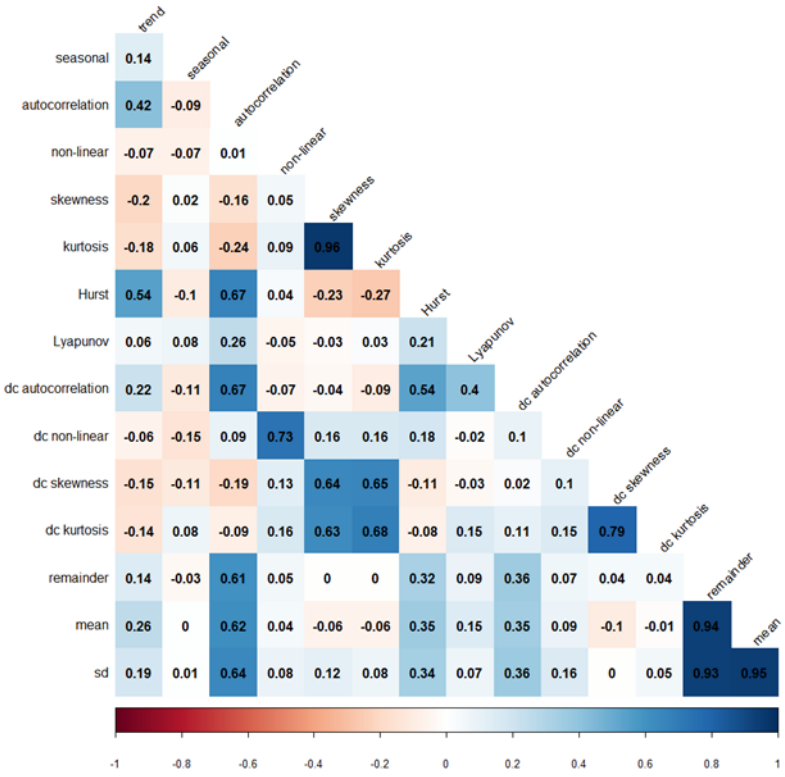
Wang X., Smith K. & Hyndman R. (2006) Characteristic-based clustering for time series data. Data Mining and Knowledge Discovery 13, 335-64.

Wang X., Smith K.A. & Hyndman R.J. (2005) Dimension reduction for clustering time series using global characteristics. In: 5th International Conference on Computational Science - ICCS 2005 (eds. by Sunderam VS, Albada GD, Sloot PMA & Dongarra JJ), pp. 792-5, Atlanta, GA.

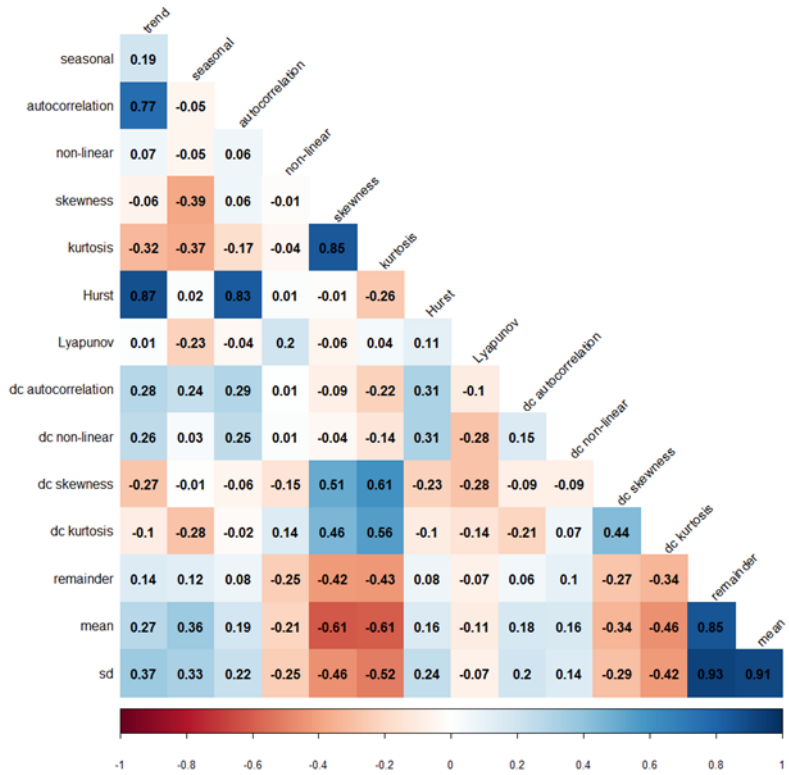
# Appendix 1 Correlation matrix for the characteristics of health aberrations



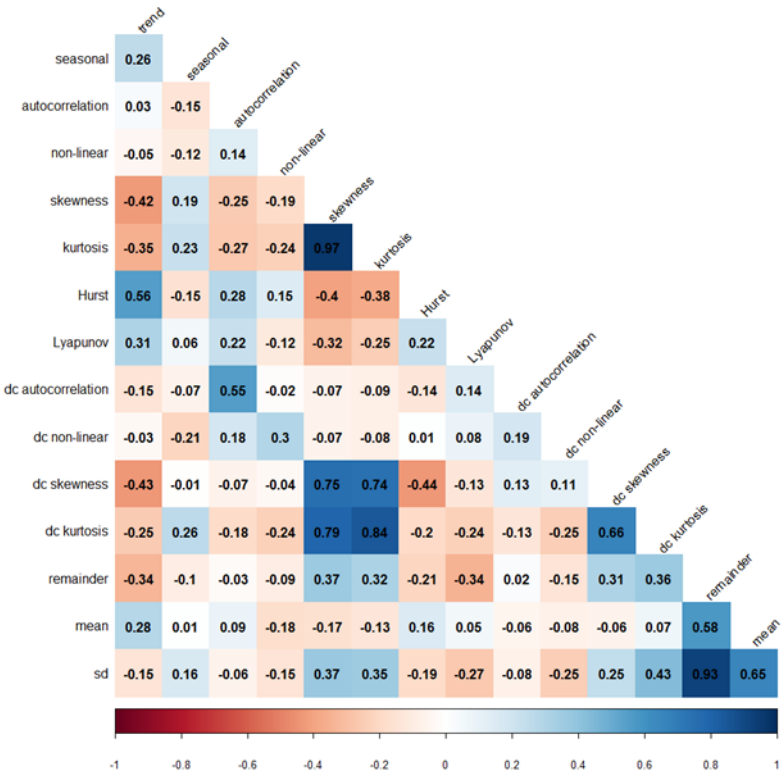

Correlation matrix for the characteristics of percentage pneumonia



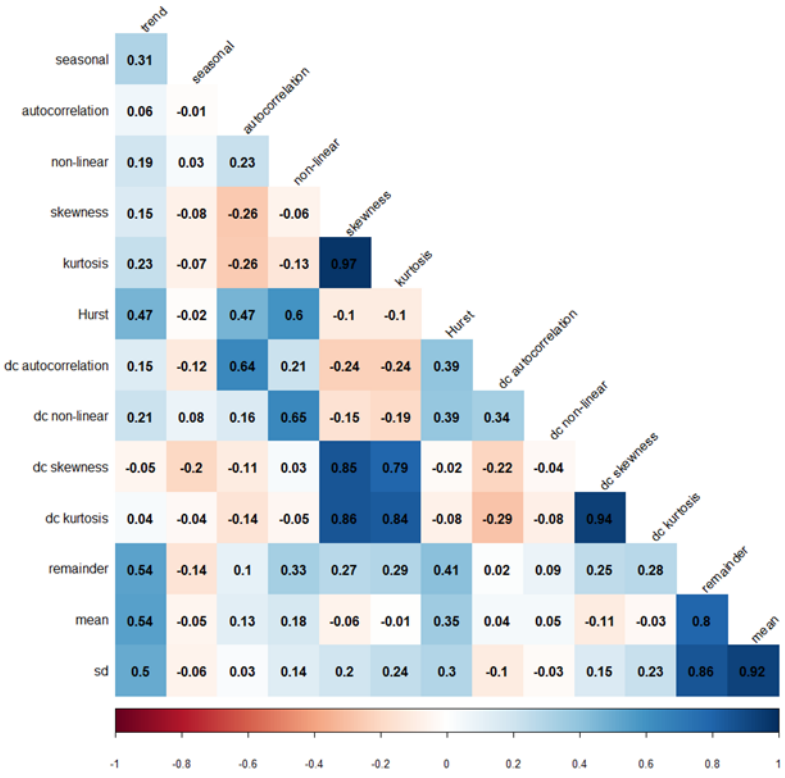

Correlation matrix for the characteristics of percentage liver aberrations




Correlation matrix for the characteristics of percentage pleurisy




Correlation matrix for the characteristics of percentage leg aberrations




Correlation matrix for the characteristics of percentage skin aberrations

To explore the potential of nature to improve the quality of life

Wageningen Livestock Research
P.O. Box 338
6700 AH Wageningen
The Netherlands
T +31 (0)317 48 39 53
E info.livestockresearch@wur.nl
www.wur.nl/livestock-research

Together with our clients, we integrate scientific know-how and practical experience to develop livestock concepts for the 21st century. With our expertise on innovative livestock systems, nutrition, welfare, genetics and environmental impact of livestock farming and our state-of-the art research facilities, such as Dairy Campus and Swine Innovation Centre Sterksel, we support our customers to find solutions for current and future challenges.

The mission of Wageningen University & Research is 'To explore the potential of nature to improve the quality of life'. Within Wageningen University, nine specialised research institutes of the DLO Foundation have joined forces with Wageningen University to help answer the most important questions in the domain of healthy food and living environment. With approximately 30 locations, 6,000 members of staff and 10,000 students, Wageningen University is one of the leading organisations in its domain worldwide. The integral approach to problems and the cooperation between the various disciplines are at the heart of the unique Wageningen Approach.