MASTER THESIS IN MICRO DATA ANALYSIS

Master of Science in Business Intelligence

Analyzing automatic cow recordings to detect the presence of outliers in feed intake data recorded from dairy cows in Lovsta farm

Author: Gloria Kogo
Supervisor: Lars Ronnegard
Examiner: Siril Yella
Subject/main field of study: Business intelligence program
Course code: MI4001
Credits: 30 ECTS
Date of examination: 19/08/2016

At Dalarna University it is possible to publish the student thesis in full text in DiVA. The publishing is open access, which means the work will be freely accessible to read and download on the internet. This will significantly increase the dissemination and visibility of the student thesis.

Open access is becoming the standard route for spreading scientific and academic information on the internet. Dalarna University recommends that both researchers as well as students publish their work open access.

I give my/we give our consent for full text publishing (freely accessible on the internet, open access):

Yes ☒

No ☐

Dalarna University– SE-791 88 Falun– Phone +4623-77 80 00
Abstract:
Outliers are a major concern in data quality as it limits the reliability of any data. The objective of our investigation was to examine the presence and cause of outliers in the system for controlling and recording the feed intake of dairy cows in Lovsta farm, Uppsala Sweden. The analyses were made on data recorded as a timestamp of each visit of the cows to the feeding troughs from the period of August 2015 to January 2016. A three step methodology was applied to this data. The first step was fitting a mixed model to the data then the resulting residuals was used in the second step to fit a model based clustering for Gaussian mixture distribution which resulted in clusters of which 2.5% of the observations were in the outlier cluster. Finally, as the third step, a logistic regression was then fit modelling the presence of outliers versus the non-outlier clusters. It appeared that on early hours of the morning between 6am to 11.59am, there is a high possibility of recorded values to be outliers with odds ratio of 1.1227 and this is also the same time frame noted to have the least activity in feed consumption of the cows with a decrease of 0.027 kilograms as compared to the other timeframes. These findings provide a basis for further investigation to more specifically narrow down the causes of the outliers.

Keywords: Outlier detection, Anomaly, Feed, Forage, Silage, Trough
TABLE OF CONTENTS

1. INTRODUCTION ..............................................................................................................1
   1.1 BACKGROUND ........................................................................................................ 3

2. MATERIALS AND METHODS .........................................................................................3
   2.1 DATA PREPROCESSING .......................................................................................... 5
      2.1.1 Data Exploration ............................................................................................ 5
      2.1.2 Data Handling ............................................................................................... 5
      2.1.3 Data Quality .................................................................................................. 5
   2.2 STATISTICAL ANALYSES ....................................................................................... 6
      2.2.1 Step 1: Fitting the generalized linear mixed model ....................................... 7
      2.2.2 Step 2: Fitting a mixture distribution ............................................................. 8
      2.2.3 Step 3: Analysis of the relationship between the outliers and the non-outliers .... 10

3. RESULTS ........................................................................................................................ 11
   3.1 FITTING THE LINEAR MIXED MODEL .................................................................. 11
   3.2 FITTING THE MIXTURE OF GAUSSIANS MODEL ............................................... 12
   3.3 FITTING A LOGISTIC REGRESSION ON PRESENCE OF OUTLIERS AND NON-OUTLIERS IN THE DATA ....................................................... 13

4. DISCUSSION ................................................................................................................... 14
   4.1 DISCUSSION ........................................................................................................... 14
   4.2 ALTERNATIVE METHODS ..................................................................................... 15
      4.2.1 Neural networks ............................................................................................ 15
      4.2.2 Support vector machine based method ......................................................... 16
   4.3 RECOMMENDATIONS ........................................................................................... 16

5. CONCLUSION .................................................................................................................. 18

6. ACKNOWLEDGEMENTS ............................................................................................ 18

REFERENCES ...................................................................................................................... 19

APPENDIX A ....................................................................................................................... 21
   PLOTS ............................................................................................................................ 21
   R CODE ......................................................................................................................... 23
1. INTRODUCTION

Dairy farmers around the world endeavor to find ways to optimize their cows’ milk production capacity and also, optimize their feed intake. This would help these farmers to be efficient whilst saving costs and thus gain more profits from their milk produce.

A new management concept that has emerged in recent decades is precision livestock farming (PLF). The main goal of PLF is to optimize livestock production by online monitoring and control of the production process, utilizing the ever evolving technical possibilities of automation (Cox 2005). PLF is still in its developmental stages but it gives great hopes to farmers and business men alike. It still requires extensive research and development before it kicks off fully (Wathes et al. 2008). Wathes(2008) proposes that, “the new technology to be developed should consist of integrated monitoring and control systems for biological processes”. Compared to industrial processes, biological processes are harder to monitor and control. Industrial monitoring and control systems are already successfully implemented and can thus be controlled effectively. They do not have life in them and are easily predictable leading to precise definition of targets which are set independently of time and weather. In contrast, biological processes are more difficult to control because they naturally vary due to immense differences between individuals, and dynamic changes attributable to age, the environment and reproduction.

Within the dairy farming industry, the use of automated animal feeders and milking systems is increasing, enabling the implementation and possibility of setting the daily concentrate intake and milking interval (Zom, André, & van Vuuren, 2012). Although the current settings are based on knowledge about energy and nutrient requirements in relation to milk production, they do not account for variation between and within individual dairy cows. André et al.(2010) found considerable variation in milk yield in response to concentrate intake and milking interval length among individual dairy cows and concluded that individual variation in response can be exploited to improve economic profitability of dairy farming by optimization of individual feeding and enhancing utilization of automatic milking systems (AMSSs). They recommended an individual dynamic approach to utilize the individual variation in response within management decision support systems for dairy farming.

The data from these systems that record the feed intake need to be of utmost quality to enable the immense predictions and to facilitate continuous research leading to vital recommendations to the users and farmers for increased efficiency.
Outlier detection has been used for a long time to detect and where deemed fitting, removes anomalous observations from data (Evans, Love, & Thurston, 2015). Barney and Lewis (1994) defines an outlier as, “An outlying observation, or an outlier, is defined as one that appears to deviate markedly from other members of the sample in which it occurs.” A further definition by Hodge and Austin (2004) states that an outlier is defined as an observation or a subset of observations which appear to be inconsistent with the remainder of that set of data. Hodge and Austin further continue to state that outliers can be attributed to many different causes: outliers could arise due to mechanical faults in the system, changes in the system behavior, human error in cases where data was entered manually, fraudulent behavior seen in credit card fraud cases, faulty machines resulting in an error or simply through natural deviations in populations.

The aim of this study is to identify the outliers in the feed intake data and to analyze the cause of these outliers. The main problems seen in the feed intake data used in this study is the presence of anomalies in the data which distorts the accuracy of the true values. The users of the data easily identify values which are a total deviation from sensible amounts of feed intake that an individual cow can consume at a particular visit to the feeding troughs. This makes it impossible for the users to trust the data. This then causes a limitation in the use of this data in any further analyses, research or predictions.

In our attempt to fulfill this objective (identify outliers and their causes), we apply a three step methodology which is detailed in the statistical analyses section. The first step is fitting a mixed model to the data using the cows as a random variable. The second step is using the residuals to fit a model based clustering; Gaussian mixture model resulting in a formation of clusters. One of the clusters is identified containing the most number of observations seen as outliers in the data. The third step of the methodology involves modeling a logistic regression showing the likelihood of presence of outliers in the data versus the presence of non-outliers in the data. The analyses of the results in this paper is furthered in a discussion section and concluding remarks of this study stated. An appendix is also attached containing more information gathered during the study.
1.1 Background
One of the biggest challenges in business automation is data quality. Be it in research or business/corporate context. The idea of automation must be adopted with a focus on maintaining accuracy. If there are duplicates, invalid records or missing data, the whole automation process may be undermined. Some of these issues are observed in the system for recording feed intake used in this study.

The main focus of this research project is to solve the outlier problem faced in the system for recording cow feed intake in Lovsta farm. The data was collected in Lovsta - the Swedish Livestock Research Centre- located seven km south-east of Uppsala, an experimental farm which offers housing for research and teaching where new animal production technologies can be researched and tested (http://www.slu.se/en/faculties/vh/departments/the-swedish-livestock-research-center). Study research visits were made to this farm where the dairy cows and their system for control and recording of feed intake was seen in action. A Doctoral Degree student (PHD) in animal science from the Swedish School of Agricultural Sciences (SLU) was a focal point of connection throughout the visits and ensuring accessibility of the required data. Assistance was available throughout the entire data analysis process of this project.

2. MATERIALS AND METHODS
The study was conducted using data recorded from 28th of August 2015 to 20th of January 2016. This was the preferred time period because the cows are indoors in the farm the entire duration therefore their feed intake is closely monitored as opposed to the summer season when they are allowed to graze outdoors. Their feed intake is not recorded when the cows graze outside therefore resulting in inaccurate measure of their feed consumption during this time.

Lovsta farm employs the Bio Control system technology which is a system for controlling and recording feed intake (CRFI). It gives individual cows’ access to specified mangers that are placed on weighing cells. Cows are identified by means of neck transponders or ear tags, and the purpose of the system is to record their feed intake.

When the cow approaches the manger and moves past the transponder reader, the system checks whether the cow should be given access to the trough. If the cow should have access to the trough, the access gate opens to allow the cow access to the feed. The individual cow information is then registered and the feed consumption recorded.
This information is transferred to the Central processor which is connected to the users’ PC and allows authorized users to extract and modify how the central processor controls the different station controllers, readers, weighing cells and air valves. The central processor collects the data on the feed intake of the individual cow.

![Diagram](image.png)

**Fig1.** A cycle showing the general functioning of the system for recording feed intake. It starts from when a new cow is registered in the system and the cows’ feed type is assigned to the manger, then access control measures applied to the feed when the cow visits the troughs and finally the feed consumption recorded and sent to the central processor. (BioC, 2013)

Every visit in any of the roughage feed stations is registered with respect to cow number, time of the visit, feed type, Date of visit, gate number, feed consumed, duration of the visit. This data is stored as daily log files which contain multiple visits of all the cows with access to the feed from the system. In this dataset, there are 437,856 observations with a record of 206 individual cows.

During this study period, the cows were consuming Silage (feed type 1) from the feeding troughs where the system for recording the feed intake is installed. They were allowed to consume as much as they can with no restrictions set on time and the feed quantity.
2.1 Data preprocessing

2.1.1 Data Exploration
The data is limited to what is available. Information about the method of data collection gives clues as to the likelihood of the data being already highly structured—a very important feature as this indicates the operations needed to handle the data. (Sapsford & Jupp, 2006) In a highly structured data a great deal will already have been determined. The data in this study is highly structured.

The method of recording of the data in this study is even more automated. Everything is computer controlled: the access to the feeding trough, the registration of the animal, the feed consumption record, there is real time recording of the visits to the troughs and at the end of the process, a log file is sent to the central processor on a daily basis containing all this information. The log files are available for download as text files.

The data sets comprising detailed information of daily feed intake records of the 206 dairy cows open up the possibilities for treating the data quantitatively and doing tests of statistical significance.

2.1.2 Data Handling
The data was shaped to enable analysis of the data more readily. All the text files of the daily log files for the 6-month period was merged into one text file as they all had the same structure and variables recorded. This formed one data matrix containing the variables.

2.1.3 Data Quality
Data exploration of the feed intake dataset containing the feed intake revealed that many of the animals have no records of their feed intake on many of the days. Too many missing observations will complicate the analysis when it comes to the problem of the outlier identification in the feed intake data as mentioned in the introduction. Therefore, it is crucial to work with complete data which consists of daily feed intake records of all the days in the study period of each of the animals registered. There are a number of reasons as to why some days may have no records which include; malfunctioning of the trough gates causing a mishap resulting in no records sent to the central processor therefore no records of feed intake in those days until the problem is resolved.

In addition, cows that fall ill are secluded from the barn resulting in no records taken of their forage intake. They later join the rest of the animals if they get well but this could be after months at a time. Some of the animals die.
The feed intake system which is the system for controlling and recording the feed intake (CRFI) as mentioned earlier is implemented in two barn areas in the farm and is mostly used for cows currently undergoing experiments. If the cows have missing records of their feed intake for a substantial amount of time it strongly suggests that they have been moved to different barns hence no possible way of recording their forage intake. With respect to this, it was observed that many of the cows have no records of their feed intake on a number of days. To handle the missing data, cows which had more than one third (1/3) of their total daily feed intake observations missing were removed from the dataset. The resulting data set contained 78 animals with a total of 263743 observations recorded. Only silage (feed type 1) was used as this was the feed given in the troughs during this period. This resulted in a total of 180,372 observations.

2.2 Statistical analyses
This section describes the statistical methods used in the analysis of the data. It also gives a description of the variables used as independent and dependent variables. The statistical analyses are carried out in a three step methodology which is described in detail in each section below. The first step in the three step methodology is fitting a generalized linear mixed model to our data. The residuals of the first step are used in the second step which is fitting a mixture model based cluster analysis. The identified clusters will then be used in the third step which is fitting a logistic regression to model the presence of outliers versus non outliers.
Definitions of dependent and independent variables used in statistical models.

**Dependent(Y) variable**

**Feed Quantity** is the response variable in grams which captures the amount of feed intake consumption of the dairy cows. This feed is known as Silage.

**Independent(X) variables**

**Hour of day** is a variable that records the time of each visit the cow makes to the feeding troughs. This was split to form four groups therefore a factor variable that consists of 6 hour time frames in a 24 hour day schedule. Hour A to D, with hour A starting from after 00:00hrs to 05:59hrs, B from after 6:00hrs to 11.59hrs, C from after 12:00hrs to 17.59hrs and D from after 18:00hrs to 23:59. This would be used to note and compare which hours of the day the cows’ consume the most feed.

**Day of year** is another variable used which captures the daily observations recorded in the selected period. It is used in the models as a numeric data type in the form of the number of the day in the year (1-365+) starting from the first day of the year of the data recorded in the dataset as day 1. The day of year variable is used as a linear trend to make the analysis simpler in the interpretation of its effects in the results.

**Animal ID**, this is the identifier for the animals. It is also a factor variable and is used a random variable in the models since all the individual cows have different feeding behavior and hence a better method leading to better results in better analyses.

**2.2.1 Step 1: Fitting the generalized linear mixed model**

Dependent and independent variables were defined in Table 1. We analyzed three different dependent variables. These variables were hour of the day (hour), a categorical variable of four different groups of 6 hour intervals, the day of the year the feed was recorded (Day of year) as a numeric data type signifying the day as number of the year and the unique identifier of the cows (Animal ID). The analysis was carried out using lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in the R programming software.

The response variable was log transformed so as to make the distribution more normally distributed. Visual inspection of the histogram plot of the response variable (feed quantity) showed a left skew in the data. As a result of this transformation, the variable was more normally distributed. A generalized linear mixed effects model was fit to the data:

\[ Y_i = X_i \beta + Z_i b_i + \epsilon_i \]  

(1)
Where $X_i\beta$ contains the fixed terms and $Z_i b_i$ contains the random explanatory variable hence a mixed effect model. Variables Hour and day of year were modeled as fixed effects and animal ID was modeled as a random effect. Animal ID was modeled as a random variable mainly because the feeding behaviors of cows vary. It is quite expected that some animals would elicit different feed intake behavior at different hours of the day. Therefore, emphasizing that the effect of hour on feed intake may be different for different animals. A random intercept model was used to model the outcome where the Animals showed differing intercepts (Winter, 2013)

### 2.2.2 Step 2: Fitting a mixture distribution

Cluster analysis is a classification problem in which the number and properties of groups within the data are unknown. The main goal of clustering methods is to discover a hidden pattern or an essential structure in the data which normally consists of distinct groups of observations, such that the observations within a group are similar to one another in some way (Gnanadesikan and Gnanadesikan, 1997).

The histogram and density plot of the residuals seen from the output (See Figure A1) in approach one of fitting the GLMM shows that the data comes from a mixture of two distributions. This finding led to this second approach of fitting a mixture model to the residuals of the regression result.

In mixture models there are a group of data points which for instance is suspected that they came from J different distributions but there is no certainty as to which points come from which distribution. The parameters mean and variance is not known in our dataset, therefore we are not able to figure out which point is from which distribution. Gaussian mixtures can emulate all kinds of distributional shapes (Hennig, 2010)

In our dataset containing the feed intake variable, each value is either an outlier or a true value of the cow feed intake recorded. Therefore, for each point $x_i$ the mixture model will figure out which distribution it came from and hence make it possible to identify the outcome of the point (outlier of true value). This process will be done by the EM algorithm detailed below. In this case, we know that $y_i$ comes from one of the J populations but we do not know which one. We define the complete data $x= (x_1, \ldots, x_n)$, where $x_i = (y_i, Z_i)$. The marginal probability of $Z_i$ is $P(Z_i = j) = \pi_j$; conditional on $Z_i = j$, assume $y_i$ has density $P_j(u | \theta_j)$. Now let

$$\log L(\theta; x) = \sum_i \log L(\theta; x_i)$$
EM algorithm computes the probability that it came from one distribution and or it came from the other distribution then it will use the numbers to compute the mean and the variances. This process keeps iterating until it converges.

EM algorithm assigns data to cluster with some probability. It is a probabilistically sound way of doing soft clustering. Each cluster corresponds to a probability distribution. The mean and the covariance are unknown and EM algorithm allows you to infer those parameter values. In each iteration stage, there are two steps, E (for expectation) and M (for maximization).

**E-step:** Estimate the distribution of the hidden variable given the data and the current value of the parameters. The E-step consists of finding the conditional probabilities (defined as the posterior probabilities)

\[
p_{ij} = E(I(Z_i = j)|y, \theta) = P(Z_i = j|y, \theta) = \frac{\pi_j p_j(y_i|\theta_j^0)}{p(\theta^0)}
\]

This is the estimated probability of \(y_i\) coming from population \(j\). The posterior probabilities can be used to segment data by assigning each observation to the class with maximum posterior probability. \((\theta_k)\) will be referred to as mixture components or classes, and the groups in the data induced by these components as clusters.

**M-step:** Maximize the joint distribution of the data and the hidden variable. Compute the maximum likelihood parameters for each component, where each data point is weighted by the cluster’s responsibility

\[
\sum_i p_{ij} log p_j(y_i|\theta_k)
\]

The E- and M-steps are repeated until the likelihood improvement falls under a pre-specified threshold or a maximum number of iterations are reached.

Mclust (Fraley & Raftery, 2002) is a package in R programming language which allows model-based clustering with noise, namely outlying observations that do not belong to any cluster.
The Mclust package was used in this approach. Mclust consists of 9 vectors (because 9 clusters is the default) of the mean values of each of the variables. A plot is generated and a visual clue is given of the number of clusters that appear in the data.

Mclust usually assumes a normal or Gaussian mixture model. Gaussian mixture ML is sensitive towards outliers. (Hennig c. , 2009).

\[ \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \varphi_k(x_i \mid \mu_k, \Sigma_k), \]  

(5)

Where x represents the data, G is the number of components, \( \tau_k \) is the probability that an observation belongs to the kth component \( (\tau_k \geq 0; \sum_{k=1}^{G} \tau_k = 1) \). (Fraley & Raftery, 2003)

### 2.2.3 Step 3: Analysis of the relationship between the outliers and the non-outliers

A logistic regression will be used to identify the likelihood of the presence of outliers and the presence of non-outliers in the observation. The outlier cluster identified in approach 2 will be used to fit a logistic regression which would be used to analyze the occurrence of outliers and non outliers in our dataset.

The results would give an idea as to when the outliers are most likely to occur hence recommend this to the users of the system. The outcome was modeled as a binomial model.

The non outlier clusters will be merged as one in comparison to the outlier cluster. The response variable in our logit model will be the presence of outliers versus the presence of non-outliers in the feed intake data and the explanatory variables hour of the day, day of year of the feed intake and the Animal ID modeled as a random variable

A generalized linear mixed model (GLMM) was fit with the variable Animal Id modeled as a random variable using the function in lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R programming software. The following model was applied on the data:

\[
\text{Logit}(P_{ij}) = \alpha + \beta_1 \ast \text{Hour}_{ij} + \beta_2 \ast \text{DayofYear}_{ij} + b_i \ast \text{AnimalID}_{ij}
\]  

(6)

The notation logit stands for the logistic link. \( P_{ij} \) is the probability that the feed intake \( j \) on animal \( i \) is an outlier.

\( \text{Hour}_{ij} \) is the hour of the day, \( \text{Day of Year}_{ij} \) is the day of the year both modeled as predictor variables \( \beta_1 \) and \( \beta_2 \)

\( \text{Animal}_{ij} \) is the unique identifier of the animals modeled as a random variable \( b_i \). (Zuur, Ieno, & Walker, 2009).
3. RESULTS
This section details the results of the three-step methodology applied in this paper: the
generalized linear mixed model analysis, the model based clustering of the residuals as a
mixture of Gaussians and the relationship between the presence of outliers and non outliers
modeled using a logistic regression.

3.1 Fitting of the linear mixed model
The lme4 package (Bates, Mächler, Bolker, & Walker, 2015) in R was used to perform a
linear mixed effect analysis of the relationship between the feed intake and the Hour. The
variables Hour and Day of year were used as fixed effects and for random effects, the
variable Animal Id was used. P-values were obtained by likelihood ratio tests of the full
model.

Table 1. Random effects

<table>
<thead>
<tr>
<th>Groups</th>
<th>Names</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal Id</td>
<td>(Intercept)</td>
<td>0.1454</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>0.6064</td>
</tr>
</tbody>
</table>

The intercept variance and the residual variance are the key components here. The intercept
variance is less than the residual variance. The intercept variance shows the individual
variation between the cows, this shows how much variation of their feeding behavior exists
amongst the cows across all hours of the day and the day of the year. The residual variance
shows the level of variance within the cows, this shows how the feeding behavior per every
feed recorded of the cow varies during the entire period.

This means that the same amount of variance is seen between cows at each level, but the
individuals no longer vary consistently across the hour of the day and day of year.
Table 2. Fixed effects

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std.error</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.8110</td>
<td>0.0234</td>
<td>157.92</td>
</tr>
<tr>
<td>Hour B</td>
<td>-0.0271</td>
<td>0.0046</td>
<td>-5.87</td>
</tr>
<tr>
<td>Hour C</td>
<td>0.0286</td>
<td>0.0045</td>
<td>6.35</td>
</tr>
<tr>
<td>Hour D</td>
<td>0.0518</td>
<td>0.0044</td>
<td>11.70</td>
</tr>
<tr>
<td>Day of year</td>
<td>0.0001</td>
<td>0.00001</td>
<td>10.56</td>
</tr>
</tbody>
</table>

The results of the coefficients of the fixed effects are statistically significant with values less than (Pr<0.05). The estimate in hour B shows that a 1 unit increase of feed intake causes a change in feed intake in hour B to decrease by 0.02712 g as compared to hour A. 1 unit increase of feed intake on the other hand causes an increase in feed intake in hour C and D respectively by 0.0286g and 0.05178g as compared to hour A. As the day of year increases, an increase in feed intake by 0.0001 grams is noted.

3.2 Fitting the mixture of Gaussians model

The residuals of the model in the mixed effect analysis were used to fit a clustering model to identify the various clusters available and to mainly help in identifying the outliers in the model. Mclust package in R (Fraley & Raftery, 2002) was used to run this analysis. The Mclust package consists of 9 (because 9 clusters is the default) vectors of the mean values of each of the variables. A plot is generated and displayed and a visual clue is given of the number of clusters that appear in the data. This model resulted in 9 clusters as this is the default number of clusters produced in this model. The 9th cluster showed a huge presence of outliers. Clusters 1 to 8 were merged and modeled as non-outliers in the logistic regression. The outlier cluster (9) consisted of 2.5% of the total observations.
3.3 Fitting a logistic regression on presence of outliers and non outliers in the data

Table 3. Coefficient estimates of the random logistic model

<table>
<thead>
<tr>
<th>Groups</th>
<th>Names</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random Effects:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Animal Id</td>
<td>(Intercept)</td>
<td>0.8132</td>
</tr>
<tr>
<td><strong>Fixed Effects:</strong></td>
<td><strong>Estimate</strong></td>
<td><strong>Std. error</strong></td>
</tr>
<tr>
<td>(Intercept)</td>
<td>-3.7775</td>
<td>0.1161</td>
</tr>
<tr>
<td>Day of year</td>
<td>-0.0012</td>
<td>0.0001</td>
</tr>
<tr>
<td>Hour B</td>
<td>0.1158</td>
<td>0.0472</td>
</tr>
<tr>
<td>Hour C</td>
<td>-0.1027</td>
<td>0.0480</td>
</tr>
<tr>
<td>Hour D</td>
<td>-0.1002</td>
<td>0.0472</td>
</tr>
</tbody>
</table>

The random intercept variance of 0.8132 is not far off from 0 showing that the cows feeding behavior is randomized and the variation there is a huge variation between the individual cows. All the coefficients of the fixed effects were statistically significant.

An increase in 1 unit of feed intake in hour B increases the odds of presence of an outlier by 12.27% as compared to hour A. A 1 gram increase in feed intake in hour B increases the odds of having an outlier by 1.1227. The odds of hour C and D are less than 1 therefore the odds of hour C and D having outliers is less than the odds of hour A having outliers.

An Odds Ratio of 1.00 means that the two groups were equally likely to have presence of outliers, therefore any day of the year is equally likely to have presence of outliers.
4. DISCUSSION

4.1 Discussion
The relationship in Table 3 from the linear mixed model analysis was found to be significant. The feed intake seemed to increase as the hour of the day increased. The cows consume the least amount of feed during hour B which is between 7am to midday, followed by hour A which is from 1am to 6am. The cows are noticed to consume more as the day progresses in hour C and D with the most consumption taking place during hour D which is between 7pm to midnight. As the year progresses, it implies that the cows’ feed intake increases by the day. In Table 4, from the logistic regression analysis of outliers, all the variables were found to be statistically significant. Hour B shows a higher likelihood of presence of outliers as compared to hour A whereas Hour C and hour D show a lower presence of outliers with hour D as the least likely hour to list outliers as compared to hour A. As the day of year progresses, there seems to be a decrease in the presence of outliers.

The two tables discussed above imply that the outliers are more present in the data during the hour of least consumption, hour B. The hours with the most consumption which is hour C and D respectively imply a minimum number of outliers. The timing between midday and noon when the cows consume the least shows the most presence of outliers in the feed intake data.

The feeding behavior of the animals should be further observed during the time period hour D (00.00-12.00) to observe if there is any peculiar activity occurring at this time. An assumption given was the cows lean their heads on the weighing scales attached to the feeding trough hence skewing the record of the amount of feed consumed (BioC, 2013b).

In this particular study, the users in Lovsta farm want to find out what causes the outliers in their system. For this reason we have undertaken the further step of running the logistic regression on the outliers. This study is only able to identify the time period that is seen to have a higher likelihood of occurrence of the outliers. Also, the coefficients show that the cows are randomized and no visible pattern is seen in their feeding behavior, therefore, it is not possible to pinpoint any particular cow as having a higher likelihood of outliers or being the cause of the outliers in our feed intake data.

There are some alternative machine learning algorithms which could be used to conduct our analyses. Some of these methods are discussed in the coming section.
4.2 Alternative methods
Supervised networks require a pre-classified data set to permit learning. If this pre-classification is unavailable then an unsupervised neural network is desirable. (Hodge & Austin, 2004)

There is no single universally applicable or generic outlier detection approach. In outlier detection, the analyst should select an algorithm that is suitable for their data set in terms of the correct distribution model, the correct attribute types, the scalability, the speed, any incremental capabilities to allow new exemplars to be stored and the modelling accuracy (Yamanishi, Takeuchi, Williams, & Milne, 2004).

Thus said, two unsupervised learning methods that could be applied to find the outliers in our dataset are detailed in the following section. The methods were not implemented in our study.

4.2.1 Neural networks
Unsupervised neural networks can be applied to find outliers in the data. Neural networks do not require pre-labelled data to permit learning and can identify those data points that are not reproduced well at the output layer as outliers. (Zhang, Meratnia, & Havinga, 2010) Outliers can be detected by assigning a confidence measure to network decisions. Confidence should be high in regions well represented in the training data and low on other areas. (Bullen, Cornford, & Nabney, 2003) Different methods can be used to assign this confidence. The reconstruction error can be used as the measure of outliers seen in data points. (Zimek, Schubert, & Kriegel, 2012)

Neural network based approaches, strictly belong to a semi-parametric method and are trained to model the underlying data distribution without any assumption of a standard data distribution for the data. The data is mapped into a trained network or a feature space to identify if these points deviate from the trained network model. They are used to effectively identify outliers and automatically reduce the input features based on the key attributes. However, they are susceptible to high dimensional data sets, where neural networks are harder to be trained well. High dimension data sets contain a large number of data and each data points also have a large number of attributes. Moreover, they need extra training time and are also sensitive to model parameters.
4.2.2 Support vector machine based method
Unsupervised support vector machine methods (SVM) can also be used for outlier identification. Similar to neural networks, Unsupervised SVM-based approaches do not have any assumption on data distribution and can effectively identify outliers without pre-labelled data. They identify outliers by using the kernel functions (Schölkopf et al, 2001). This approach uses the kernel function which efficiently maps the data into a vector space (Schölkopf et al, 2001). Outlier detection is then performed dependent on the position of the points in the featured vector space. The points that are distant from most of the other points or are in relatively sparse regions are then labeled as outliers.
However, the computation of the kernel functions is a computationally expensive task. Also, it is not easy to determine appropriate parameters to control the size of boundary region. Both of the two methods described above follow a similar approach on the data set which involves first splitting the dataset. Data splitting involves partitioning the data into an explicit training dataset used to prepare the model and an unseen test dataset used to evaluate the model's performance on unseen data. It is useful when you have a very large dataset so that the test dataset can provide a meaningful estimation of performance, or for when you are using slow methods and need a quick approximation of performance (Hodge & Austin, 2004). Model accuracy can also be evaluated using a gold standard method such as Cross validation.

4.3 Recommendations
We recommend that Lovsta farm consider using this three step approach to help identify the occurrence of outliers in their data this approach enables the users to have a simpler understanding of the analyses methods applied as they visualize what outcome ensues from each step. More variables could be used to analyze the feeding behavior of each of the animals. For instance, to analyze the effect of the day of the year on the feed intake, the link between the insemination data that contains the date when the animals are on heat could be used to analyze the feed behavior of the animals as it is assumed that when cows are on heat, they eat less (BioC, 2013b).
For further research, the next step would require added resources in monitoring the weighing scales and the feeding troughs at the suggested hours. This would help determine the causes of the outliers which once resolved would increase confidence in data quality making it viable for predictions and other uses. The data can be linked with different datasets for instance, the milk production data of the dairy cows and analysis of maximizing the milk production can be conducted, helping the dairy farmer to achieve their main goal of maximizing their cows’ milk production.

Additionally, the increased confidence in our findings could be achieved by applying alternative methods like unsupervised learning methods discussed to the feed intake data. It is then possible to do a comparison of the methods for purposes of determining the model with the best performance.
5. CONCLUSION

Addressing the problem of outliers is crucial in automated systems. Further analyses cannot be conducted with inaccurate data. Researchers and analysts need to ensure the data they work with meets the data quality standard before proceeding with further analyses. Fitting the logistic regression to the data described as step three of the methods is best suited for determining the presence and causes of outliers in the feed intake data at Lovsta farm as the results pinpoints instances of the occurrence of the outliers in our data. As the main objective was to identify the cause of the outliers in the system, the findings of our analysis has enabled us to get one step closer by narrowing it down to the time of the day most likely to record the feed intake amount as an outlier which is between 7am to midday which happens, according to our analysis, to be the time of the day where the least amount of feed intake is consumed. Based on this we recommend that further research be undertaken at the stated times to more specifically pinpoint the causes of the outliers for purposes of resolving the outlier issue and improving data quality.

6. ACKNOWLEDGEMENTS

I am grateful to the Swedish Livestock Research Centre (Lovsta farm) which provided the feed intake data analyzed. Thanks are also due to J. Karlsson for the assistance in making available the data and for valuable discussion on the subject matter. Finally, my acknowledgement goes to the reviewers of this paper. I wish to thank the reviewers for their insightful comments that allowed me to improve this paper.
REFERENCES


APPENDIX A

Plots

Figure A1. Histogram of the residuals of the lme4 regression model.

Figure A2. Resulting clusters of the mixture distribution.
Figure A3. Box plots of the feed intake data over the given period

Figure A4. Random effects on intercept model of the logistic regression. The animal IDs on the y axis and their intercept values
**Figure A5.** QQ plot of random effects model of the logistic regression.

**R code**

```r
# fitting the glm model and getting residuals
Library(lme4)
regmodel<-lmer(logfeedquantity~(1|AnimalId)+hour + dayofyear,REML=TRUE,data = subfeed1)
ress<-resid(regmodel)
# using mclust
## outliers
Library(mclust)
# ressx<-rbind(ress,c(7,30),c(3,80))
mfeed <- mclustBIC(ress,prior=priorControl())
smfeed <- summary(mfeed,ress)
y<-plot(ress,col=smfeed$classification)
# noise
library(prabclus)
library(fpc)
# nnc <- as.logical(1-NNclean(chevron[,2:3],15,plot=TRUE)$z)
nnc<- as.logical(1-NNclean(ress,k=4)$z)

mfeedxn <- mclustBIC(ress,
                       initialization=list(noise=initnoise))
```

23
smfeedxn <- summary(feedaboot, ress)
plot.new(ress, col=smfeedxn$classification+1)

faithfulnn <- NNclean(faithfulx, k=4, plot=TRUE)
plot(faithfulx, col=1+faithfulnn$z)
plot(faithfulx, col=smfaithfulxn$classification+1)

# logistic regression
# setting the outlier cluster as dummy variable
classes <- cbind(smfeed$classification, ress, subfeed1)
va <- as.vector(classes[,1])
test <- va[1:180372]
newclass <- rep(0, 180372)
nineindex <- which(test==9)
new <- replace(newclass, nineindex, 1)
logistreg <- glm(new~hour+dayofyear, family = "binomial", data =subfeed1)
# random logistic model
Library(lme4)
logisticfull <- glmer(new~dayofyear+hour+(1|AnimalId), family= binomial, data =subfeed1)
# odds ratio
exp(OR=coef(logisticfull))
# odds ratio and 95% CI
exp(cbind(OR=coef(logisticfull), confint(logisticfull)))
se <- sqrt(diag(vcov(logisticfull)))
# table of estimates with 95% CI
(tab <- cbind(Est = fixef(logisticfull), LL = fixef(logisticfull) - 1.96 * se, UL = fixef(logisticfull) + 1.96 * 
             se))
# log odds ratio of fixed effects
exp(tab)
# correlation matrix of fixed effects
# installing sjplot and sjmisc
install.packages("sjmisc", dependencies=TRUE, repos=c("http://rstudio.org/_packages","http://cran.rstudio.com"))
library(sjPlot)
sjp.glmer(logisticfull, type = "fe.cor")  
# qq plot of random effects
sjp.glmer(logisticfull, type = "re.qq")  
# plot probability curve of fixed effects
sjp.glmer(logisticfull, type = "fe.pc")  
# plots probability curves for each covariate
# grouped by random intercepts
sjp.glmer(logisticfull, type = "ri.pc")
sjp.glmer(logisticfull, type = "ri.slope", facet.grid = FALSE)  
#sjp.glmer(logisticfull, type = "fe.slope", show.ci = TRUE)
# plots the faceted plots of the random effects intercept
sjp.glmer(logisticfull, y.offset = .4)  
# define an 80%/20% train/test split of the dataset
library(klaR)
trainIndex <- createDataPartition(subfeed1$logfeedquantity, p=0.8, list=FALSE)
dataTrain <- subfeed1[ trainIndex,]
dataTest <- subfeed1[-trainIndex,]  
# 10fold cross validation
trainControl <- trainControl(method="repeatedcv", number=10, repeats=3)
metric <- "RMSE"
set.seed(7)
fit.knn <- train(logfeedquantity~., data = dataTrain, method="knn", trControl=trainControl)
print(fit.knn)
print(fit.knn$finalModel)  
# estimate skill on validation dataset
predictions <- predict(fit.knn, newdata=dataTrain)