Faculty of Mathematics and Informatics

VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
MASTER'S STUDY PROGRAMME
MODELLING AND DATA ANALYSIS

FUNCTIONAL DATA: CASE STUDY

Master's thesis

Author: Nedas Brikas
VU email address: Nedas.Brikas@mif.stud.vu.lt

Supervisor: Dr. Jurgita Markevičiūtė

Vilnius
2023

Table of contents

1. ABSTRACT

This master's thesis concentrates on functional data analysis and how it can be applied in precision livestock farming. Interestingly published papers do not use functional data analysis, in spite the fact that well equipped farms collect perfectly suitable data of this kind of data analysis approach. Functional data is more useful and supplies more information to decision maker than any other model used in precision livestock farming to this day. This type of analysis is especially well fitting in describing feed and nutrition impacts on cow productivity. To prove the feasibility and describe impacts of total mix ration changes a dataset of daily milk yield is used containing information about milk yield and fodder composition. To provide evidence for previous claims regressions with functional responses and functional predictors is calculated. Those regressions show an impact of hay and corn on milk yield and how it changes over a period functional observation.

Key Words: Functional data, Precision livestock farming, Regression

2. Notations and abbreviations

A

P

T

3. Introduction

Functional data analysis is a statistical approach that involves analysing data that are observed over time or across a range of continuous variables. In the context of precision livestock farming, functional data could be used to analyse a wide range of factors that are important for optimizing the management of livestock.

One potential application of functional data analysis in precision livestock farming is the analysis of animal data. For example, an animal's weight or body condition score might be measured at regular intervals over time, and this data could be analysed using functional data analysis methods to understand how these variables change over time and how they are related to other factors such as feed type or disease status. This could help farmers identify trends or patterns in animal health that might indicate a need for changes in management practices, such as adjusting the type or amount of feed given to the animals.

Another area where functional data could be useful in precision livestock farming is in the analysis of data on feed and nutrition. Functional data analysis could be used to analyse data on the quality and nutrient content of different types of feed, and to study how these factors affect the health and productivity of livestock. For example, a farmer might use functional data analysis to compare the performance of a herd of cows fed on two different types of feed, and to identify any differences in milk production or weight gain that might be due to differences in the nutritional content of the feed.

Overall, functional data analysis has the potential to be a powerful tool for optimizing the management of livestock in precision farming. By analysing data on a wide range of factors such as animal health, feed and nutrition, functional data analysis can help farmers make informed decisions about how to allocate resources and optimize production. Additionally, functional data analysis can help farmers identify trends and patterns in data that might not be immediately apparent using other statistical approaches, allowing them to make more informed and nuanced decisions about how to manage their operations.

The objective of this master’s thesis is to confirm validity of functional data analysis usage in precision livestock farming and to prove that changes in fodder has an impact on variation in milk yield. These objectives will be accomplished using milk yield and total mix ration composition data. To test the influence of total mix ration on milk yield functional regressions with functional responses and functional predictors will be calculated.

4. Precision livestock farming literature review

4.1. What does precision livestock farming do?

The development of precision livestock farming applications for dairy farming started in the 1970s with the development of electronic cow identification. The possibility to identify individual animals led to the development of a range of possibilities to manage the individual cow. Besides the development of individual concentrate supplementation, PLF[1] applications were not implemented, although in the 1980's work was carried out into development of PLF applications. In the 1990's development of PLF applications for dairy farming was centred around automatic milking. Since the early 2000's development of PLF is reflected in European Conference for Precision Livestock Farming. During this conference initiatives that are potentially interesting for application on dairy farms often started from engineers and inventors. The development of hardware is, however, only a first step in the development of a PLF system. Hereafter, a large part of the work still has to be done and this work needs the involvement of data, veterinary and nutrition specialists in order to define gold standards and decision support tools. That work is costly and it is not sure whether the final PLF application will be a success. Decisions on which PLF system needs to be developed further are not made very consequently (H. Hogeveen and W. Steeneveld, 2013).

The big amount of data collected in dairy cattle farms equipped with PLF devices represents a source of radical innovation in improved animal health, welfare and productivity. In fact, proper data processing approaches can lead to implementation of early alert systems of alerts in animal behaviour or performances, this in consequent results in increased production efficiency and business sustainability. The increasing adoption of Automatic Milking Systems (AMS) in dairy farms has been continuously providing cow-specific data to decision makers with frequent updates. The impact of adopting AMS[2] on dairy farms with regards to changes in milking labour management, milk production and milk quality was documented (M. Zucali, V. Inzaghi, P. Thompson, J. Penry, P.S. Boloña and J. Upton ,2013) New devices and studies have exponentially increased the amount of data available on modern farms. Decision makers need to convert this data into useful, immediate and practical information for farm management and animal welfare (Bonora, F., Tassinari, P., Torreggiani, D. and Benni, S. ,2017). In recent years many mathematical models already used in other fields, such as neural networks (R. V. Sousa1 and L. S.

[1] Precision livestock farming
[2] Automatic milking system

Martello1), image processing (B. Meunier, E. Delval1, C. Cirié , M.M. Mialon , P. Pradel , Y. Gaudron , D. Ledoux , I. Veissier, 2017), regressions and clustering (F. Bonora, P. Tassinari, D. Torreggiani and S. Benni, 2017), have been applied to precision livestock farming. This technological advance now allows the farmer to monitor his herd in specific time intervals and to compare their state with past observations and other animals, thus accessing a big data set of variables. Some PLF technologies have been developed to precisely control the amount of food given to the animals. In herbivores, not only the amount of concentrates, but also that of total mix ration can be measured individually. Combined with an analysis of feeds quality and animal needs, this can serve to check whether animals are fed according to their needs. In addition, the time spent eating is now also available thanks collars with some localization device. These devices may provide indirect measures and can also be used to capture data on such variables as time spent at the trough or time spent chewing or ruminating. The water consumption can be measured by thermometer in cattle's stomach which is inserted in a form of bolus and stays in the animal for the rest of its life. All this information can be used to check whether all welfare needs are fulfilled.

Among the automated systems used to monitor animal behaviour in real time, those based on wearable inertial sensors are widely used for dairy cows housed in barns or grazing in the pastures. Among wearable sensors, accelerometers are widely used because of their low cost and ease of integration with other devices. As most new technologies firstly they were used in smart devices to monitor human health and behaviour, then they were passed down in automated systems for animal monitoring.

A framework with which the stages in development of PLF applications can be described is given in Figure 1. The figure is taken from H. Hogeveen and W. Steeneveld article, where they describe how data from sensor in question gets transformed into usable data for a decision maker. As shown in the plot and described in the article, "the first step in PLF development is the description of the technique and the hardware". It is not uncommon for some sensors to recycle data that they obtained into useful observations and only then send them to the main computer. The ways that this data generated is very diverse, as authors state, "the data algorithm produces a descriptive state such as lying down, walking, or standing". The cameras and machine learning algorithms proves to be cost effective measurement equipment for predicting cattle dry matter intake or health condition (M. Jorquera-Chavez, S. Fuentes, E.C. Jongman, R.D. Warner and F.R. Dunshea). The second stage authors called "data interpretation" and in this step measurements are changed in the sensor data to present information useful about changes. The authors suggest that in this level it is important to perform statistical testing on data i.e. stating

hypothesis then performing tests and lastly accept or reject the hypothesis in question. From a statistically tested relation, it is possible to build a predictive or descriptive model that classifies the cows' health status or describe impact of changes in total mix ration, shelter conditions or any other variable. For validation, a validation data must be set and used to assess the performance by comparing information with "the gold standard" as the authors have called it. This stage is a crucial step in development of PLF applications because in this stage, the data from a sensor is related to a physiological state of the animal that has a meaning for the decision maker. Moreover, this raw data interpretation can be very tedious and signs that something is amiss might be difficult to notice. Moreover, the decision makers should be able to compare performance of a PLF applications so that they would be able to determine whether one model is more advantageous to use. The models that farmer can compute depends solely on his education and may choose whatever model that seems fit for the data, it may be simple multilinear regression versus generalized autoregressive conditional heteroscedasticity with external variables (GARCH-X) versus functional data analysis. The third level of creating functioning PLF systems authors describe as "integration of the sensor information with other information such as economic information". The fact that every cog in the machine should complement each other is quite obvious, because with bunch of devices working in unison the data collection is more complete and as the authors suggest that "there might exist diminishing returns and marginal cost for extra litre of milk might be too high". Something that models might not incorporate so easily is biological barrier for how much milk can specific cow breed produce without affecting its' health. Furthermore, information of individual cows can be aggregated by a monitoring algorithm at the herd level. The output of this algorithm can be seen as either general information on the herd's health and productivity for the farmer or additional data input for the algorithm. In level IV the decision is eventually made either by the farmer or autonomously by the sensor system. If farmer truly trusts data generation and it's processing the farm might become automated to a significant level and because algorithms do not require any rest, they can manage the farm in a very delicate manner. Early detection systems may direct animal to a hospital if such alert arises, wearable sensors can detect lack of activity and assign an adjusted dosage of combined feed to compensate for lack of energy, cameras can automatically send data to feeding cart informing it how much dry matter is in the TMR[3] and consequently it can adjust how much feed cows need.

[3] Total mix ration

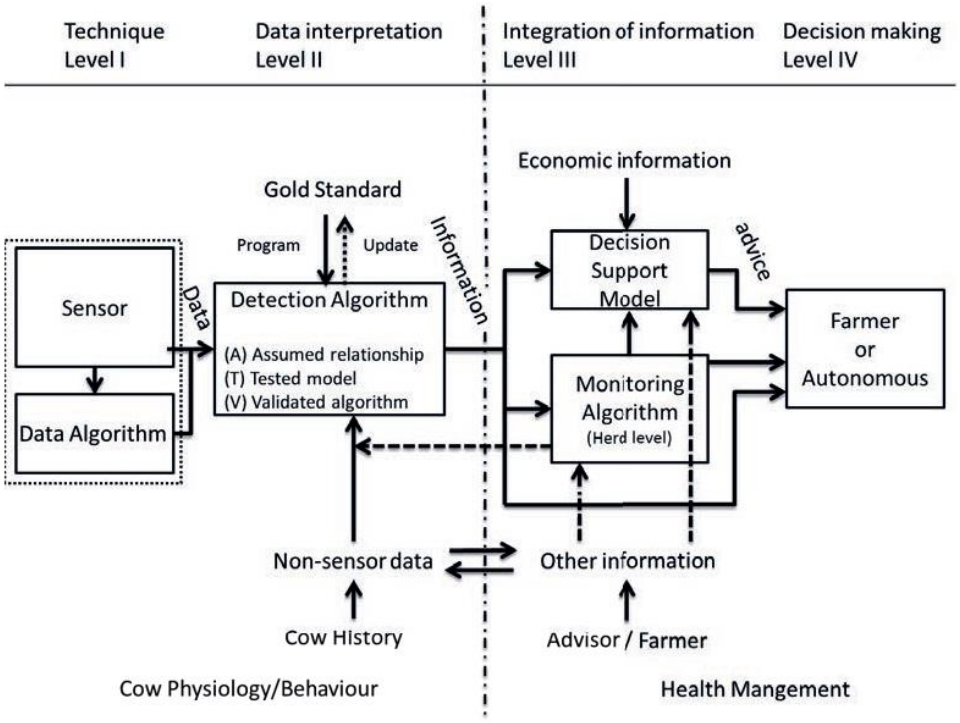

Figure 1. Framework to creating an effective PLF management system (source: "Essential steps in the development of PLF systems for the dairy sector ", H. Hogeveen).

4.2. What models does PLF use?

"Complex, individually different, time-varying and dynamic nature of living organisms" as author H. Hogeveen calls the living nature has an ultimate impact on the type of algorithms we need to develop. This implies that algorithms to monitor these time-varying individuals must continuously adapt, recognize new threats and tend the individual and in best case scenario use principles that can be used with real time data in the field applications. In this section a closer look will be taken in what models authors in field of PLF are using and what problems they are solving. Some of the most common models in PLF is cluster-graph approaches and random forest regressions.

One of the most famous authors that field tested clustering approach is Filippo Bonora and his colleagues. In article "A methodology for daily analysis of AMS data providing herd characterisation and segmentation". Here cow-specific data, early warning systems, cow performance and animal monitoring are discussed. The dataset was collected in Italian dairy farm. The variables for the implementation of the clustering procedure among the parameters available describing each cow's behaviour and health conditions were selected as follows:

- daily average of the activity score recorded by electronic collars;
- daily average of rumination time (minutes, recorded by electronic microphones);
- days in milk;
- daily average milk temperature (°C);
- daily average cow body mass (kg);
- daily average milk conductivity (cS/m);
- parity (#);

The variables expressing milk productivity, such as number of milkings and milk yield were considered dependent variables according to which the model was judged. A hierarchical clustering was selected as the approach. The author preferred to the alternative clustering algorithms based on k-means approach because “it does not require a starting point and definition of the number of subgroups”. The final output of a hierarchical clustering is not a unique clustering arrangement, but a dendrogram, appearing similar in shape to a tree. The level of grouping may be specific for each process, depending on conditions that can be stated on the basis of the study objectives. When the herd was divided into groups the averages for external variables were observed. Most common contrasting variables were number of parities and days in milk.

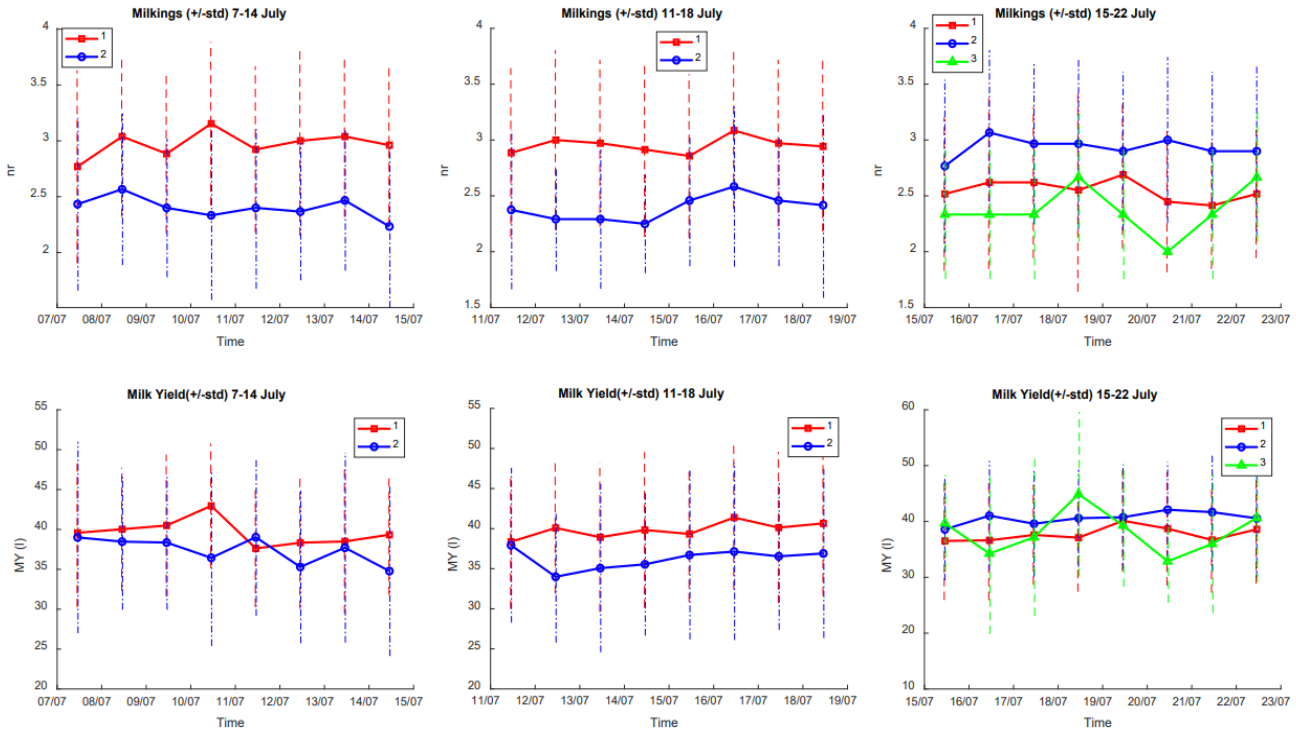

Figure 2 Trends of average number of daily milkings and milk yield in the clusters (source: "A methodology for daily analysis of AMS data providing herd characterisation and segmentation", S. Benni, F. Bonora, P. Tassinari, and D. Torreggiani).

During model calculations for 3 different 8 days' time periods 3 subgroups emerged. In the first subgroup about 2 years younger and 100 less days lactating cows were assigned than the second subgroup. In third time period third group of cows emerged showing that older, but not necessarily more days lactating cows a statistically different from previous two. Looking at graphs provided by the authors conclusions are hard to come by. Firstly, as time passes one group does not show consistently better results in milkings per day or milk yield. Secondly the confidence intervals are wide and overlapping.

As authors state this system would detect outlying cows from the herd and could serve as an early warning system and whilst carefully monitoring why anomalies occur health issues could be prevented. This kind of statistical tool is providing a more simplistic approach to monitoring the herd instead of individually looking for problematic cows the decision maker has an early heads up about impending illnesses. It is worth to mention that this kind of statistical tool is only applicable in highly advanced farms with an automatic milking parlour.

4.3. Are more complicated models better?

To explore the usefulness of statistical analytics for the estimation of cow individual feed intake, a study by C. Kamphuis, J.W. van Riel, R.F. Veerkamp, and R.M. de Mol Wageningen was published by the name of "Traditional mixed linear modelling versus modern machine learning to estimate cow individual feed intake". This research was conducted in Netherlands using data from experimental dairy farm and compared two different modelling approaches for estimating feed intake. The first model used traditional mixed linear regression to predict feed intake, and the third model was a machine learning algorithm called Boosted Regression Tree. In this research main variable was cows individual feed intake, as external variables milk yield and its qualitative indexes, live-weight, parity, outside temperature and humidity were chosen.

Experimental dataset was retrieved from ten feeding trails that took place in 2014 and 2015 years. The observations of feed intake were taken daily and each week averages for those 7 days were calculated for all variables. After excluding cows with not enough data, researchers obtained dataset with 407 cows and 3787 cow-weeks from seven feeding trails (APPENDIX 1). With the obtained data researchers calculated four new statistical models, two models were Mixed Linear Regressions with and without weather information and other two models were Boosted Regression Tree without and without weather information. To evaluate performance of each model in predicting feed intake, the Person's correlation between predicted feed intakes and actual feed intakes (APPENDEX 2).

Model	Training set	Test set		
	Correlation	Correlation	Mean difference (kg)	Range difference (kg)
MLR1	0.91	0.71	-1.23	-7.70 – 12.32
MLR2	0.91	0.72	-1.73	-8.24 – 11.76
BRT1	0.73	0.76	-0.35	-7.61 – 13.32
BRT2	0.73	0.76	-0.35	-7.61 – 13.32

Figure 3 This figure represents what models were used and what results were achieved (source: "Traditional mixed linear modelling versus modern machine learning to estimate cow individual feed intake", C. Kamphuis, J.W. van Riel, R.F. Veerkamp, and R.M. de Mol Wageningen, 2017)

In Figure 3 correlations between actual and estimated feed intakes are summarized. Both MLR[4] models have high correlations between actual and estimated values, this shows that models' outputs have a good statistical relationship, but it does not indicate a good fit. Considering that the average actual feed

[4] Mixed Linear Regression

intake of cows was 21.2 kg/day, mean differences are not high, furthermore the booted regression trees are significantly more accurate.

This research is an interesting because authors used milk qualitative statistics to calculate how much cow was going to eat, showing that if that kind of data is accessible daily via automated milking system farmer can predict how much fodder is going to be sufficient for upcoming day or forecast how much fodder will be needed for upcoming year. In conclusion, for feed intake prediction data mixed linear regression models were surpassed by more sophisticated machine learning model.

4.4. Literature review conclusions

All in all, during literature review a small number of articles were found related introduction of unordinary models and how changes fodder causes milk yield to fluctuate. A lot of research is concentrated on such subjects as early disease detection, smart ways on making observations how to measure dry matter intake, lowering electrical energy consumption or introducing new sensors. The lack of education provided to farmers about different points of view on data collected might even question the whole purpose of PLF.

The decline in the number of dairy farmers is not a new phenomenon, and not just in Lithuania. Annual decline in the number of milk producers has been present for some time and in many regions of the developed world (The Andersons Centre, 2013). As the number of market participants decreases, there is a noticeable increase in the average size of the herd, but the growth pace is decreases. There are many different production trends in different countries where some sectors of the economy are growing significantly despite the declining number of producers. It seems certain regions can exploit a competitive or comparative advantage, such as, parts of western United Kingdom can grow green fodder more cheaply, and climate conditions and topography can make the same area less suitable for alternative branches of agriculture. The analysis revealed that there are economic changes less influential in predicting the decisions of dairy producers. This suggests that decisions are more driven by faced personal or social factors. However, this is difficult to prove empirically, tests model producers' choices based on both economic and social affordances variables, explains only a very small part of the variation in producers' choices. There is no evidence that so that the big dairy farms contribute to the exit of the small ones. The segmentation of farms by size shows that there is no correct number of cows or production volume for profitable milk production. Large and small farms can be profitable or

unprofitable. Also, farms of all sizes can be efficient, progressive businesses. However larger farms have higher milk production volumes in consequence this gives a higher overall profit level and this is achieved through efficient business management (Popescu, Agatha, 2014).

5. Econometric models in dairy farming

5.1. Introduction

Dairy farms have constantly increasing number of dairy cows and their productivity is rising. At the same time, demand for dairy products remains rigid, leading to an imbalance between supply and demand. Such a market is not attractive to producers, so more and more of them are leaving the industry. Despite expanding economies there are very heterogeneous structural changes in countries in terms of technological progress, cost levels, and geographical location. The main forces behind increasing milk output are technological, efficiency, and scale changes. Structural changes usually mean a change in the size of the livestock herd or fluctuations in the level of milk production. Any quantitative methods can be useful in analysing dairy farms and their potential for development.

This case study is prepared using data from privately owned dairy farm. This farm is in the village of Dirvonenai in Siauliai district, situated on the border with Kelme and Telsiai districts. It has been operating since 2004. In 2022, about 180 dairy cows were milked twice per day. The farm harvests all the necessary fodder in the surrounding lands. Until the harvesting season of 2022, on average about 1000 tons of roughage was in one year. This roughage consists of fermented hay and corn silage. Also, about 150 tons of meal are fed per year. The meal consists of crushed grains and premixed flour. It often consists of ten or more ingredients.

The main objective of this study will be to determine whether functional data models are adequate for determining what impact do changes in TMR have on milk production quantity. To have a good background graphical analysis, descriptive statistics and a simple regression should be displayed.

5.2. Graphical analysis

Before performing detailed analysis, it is necessary to perform graphical analysis. This way it is easy to familiarize with the data at hand and draw some preliminary conclusions or take notes of points worth discussing. As mentioned, milk output will be the dependent variable, course feed components will be independent variables. Firstly, the research will concentrate on daily time series from 2019/06/01 to

2021/06/01, containing 720 unique observations. After analysing raw dataset, it will be segmented into months.

It is observable that the dependent variable has some chaotic fluctuations with a trend to grow in time. The main drops usually happen when there is no corn in the feed and grazing season comes to an end (Figure 4). Corn drops out of feed usually in the beginning of the summer, but output does not drop until late autumn. This drop might represent the fact that during the summer cows are grazing in pasture.

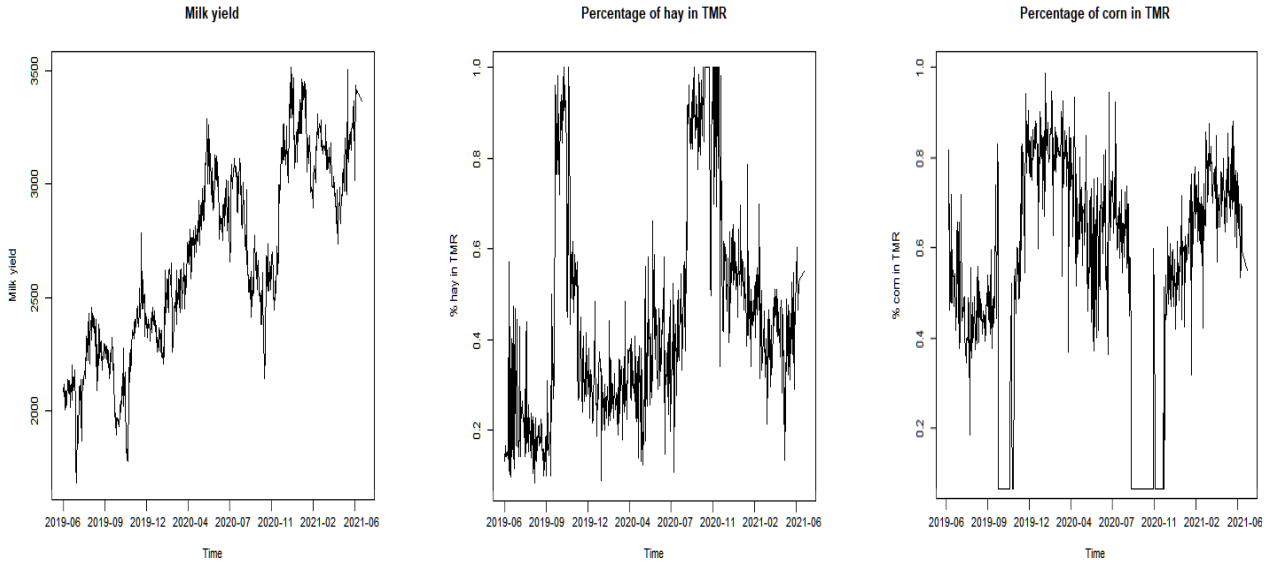

Figure 4. The first plot from the left is daily observations on milk yield. The middle graph reflects how much hay is in corresponding day's the total mix ration. The middle graph reflects how much corn is in corresponding day's the total mix ration.

The dependent regressor in theory and from earlier calculations can be explained by these two independent regressors, but since they are perfectly multicollinear, the researcher should be careful not to include both in any regression task. Linkage between what cows eat and their milk output is observable quite well (Figure 5). During times when corn is not supplied the milk yield drops significantly and consistently. In this data set it always happen during autumn months. This indicates incomplete and low nutrition diet. Fluctuations during 2020/01-2020/05 are influenced not necessarily by changes in TMR, but by growing number of milk producing cows.

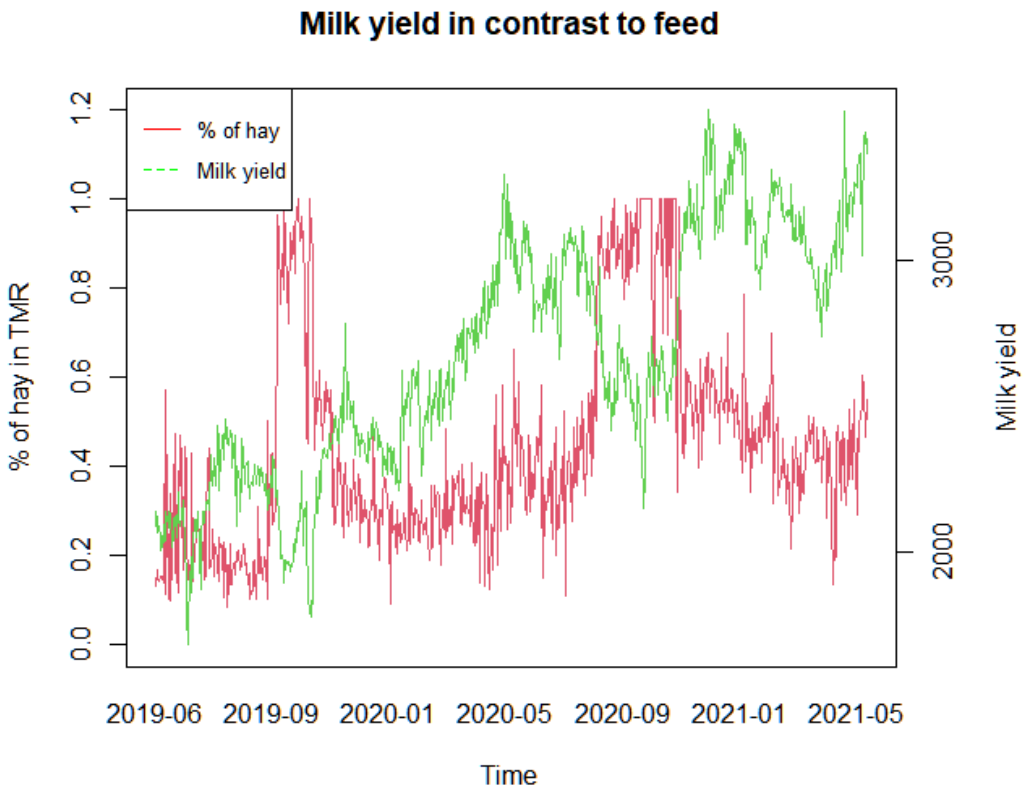

Figure 5 In this plot Milk yield is shown as green continuous line and its scale is located on the right side of the plot, percentage of hay in total mix ration is shown as red continuous line and its scale is located on the left side of the plot.

To make further insights using graphical analysis tools Tukey's honest significance test can be used. This test can be used to find means that are statistically significantly differ from each other. It compares all given averages of the groups to every other and identifies any difference between two means.

The Milk yield data was cut into 30-day months and grouped by each year's season (Figure 6). The grouping resulted in 2 groups of winter, summer and autumn seasons and 3 groups of autumn seasons there are 8 groups in total. These averages are not bunched up, but rather spread out. From economical point of view this is a good sign since milk yield has a trend to go up, but from analytical point of view this might cause some problems while trying to create descriptive models for data with time trend. 2019 winter average shows a big downwards motion, but 2020 spring average indicates a higher starting point and significantly higher end point. The difference is that cow herd is feeding on fresh frass in pastures. This sudden jump is surprising not only in sudden increase, but also that this is one-time occurrence and was not reproduced in this time frame.

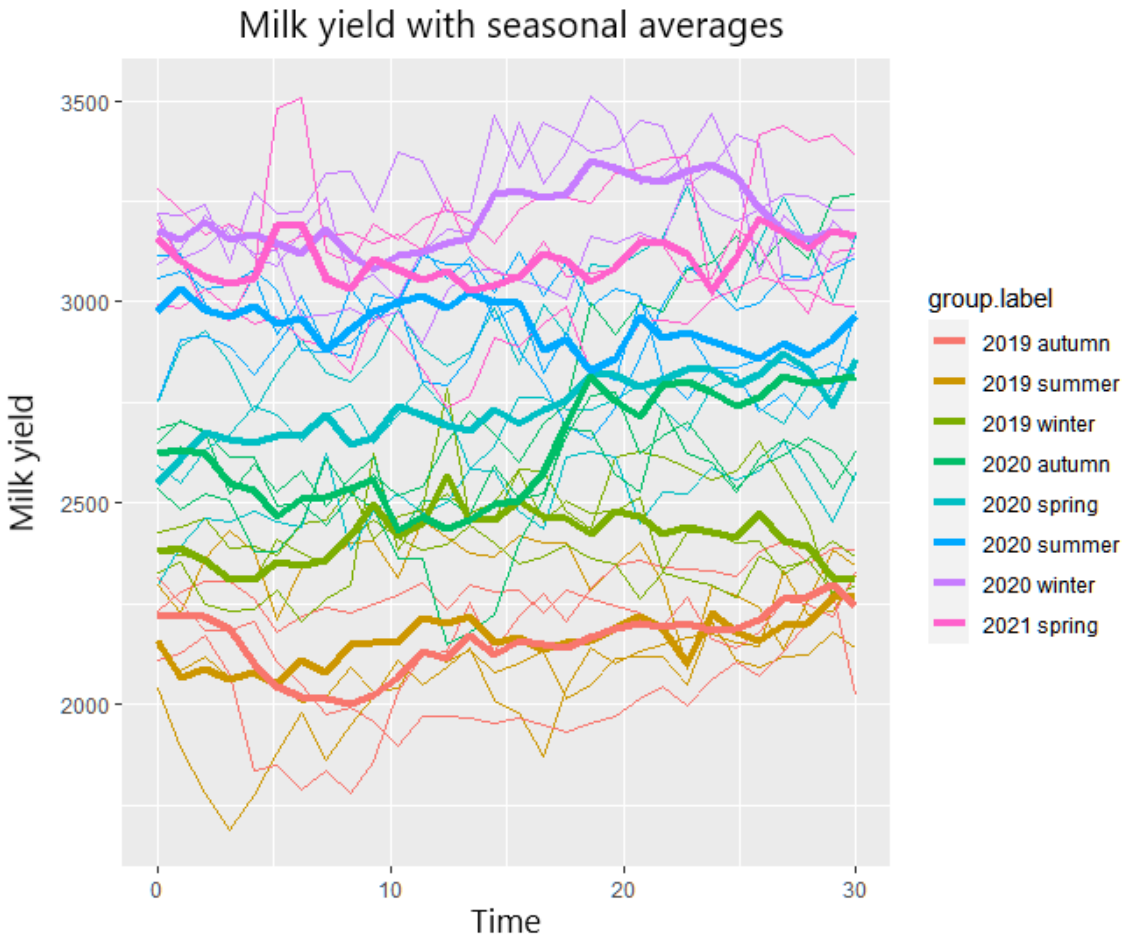

Figure 6 Raw segmented and grouped data, with thicker lines showing groups' means.

Pointwise Tukey's HSD [5] compilation of graphs (APPENDIX 3) performs pairwise comparisons of every group mean to every other group mean., this way it is much clearer what similarities exists in the data set. Tukey's HST test checks the following hypothesis:

$$H_0: \mu_0 \neq \mu_1$$
$$H_A: \mu_0 = \mu_1$$

The null hypothesis is rejected if $p > 0.5$, this threshold is marked by dotted blue lines and calculated p values are marked by solid black line (Figure 7). The significance test shows that not all groups are similar in their averages. These differences will produce an interesting variation patten in rotated principal components' scatterplot.

[5] Honest Significance Test

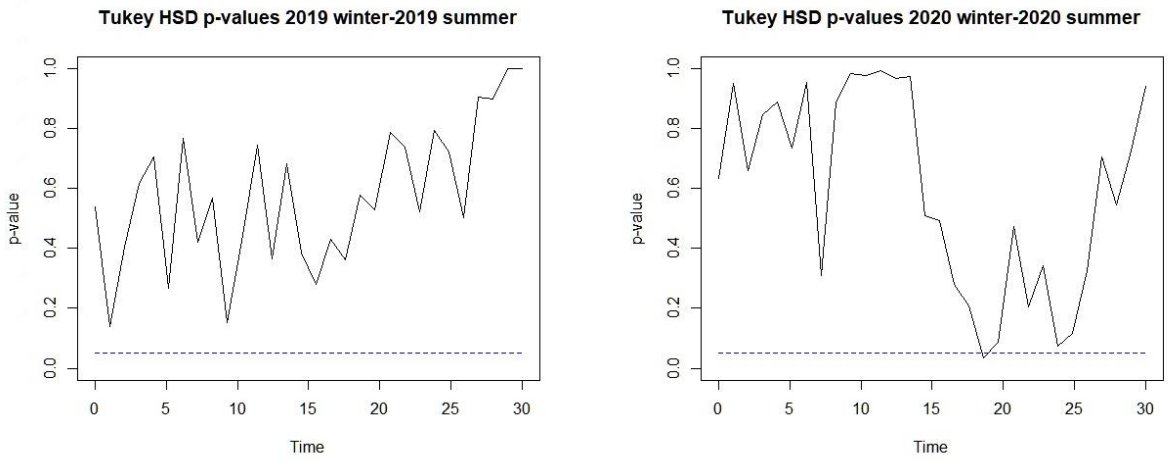

Figure 7. Tukey's HSD pairwise testing for similar means in seasons used with non-smoothed, segmented and grouped milk output data.

In this functional analysis raw data and smoothed data was analysed. Whilst inspecting milk yield as a continuous time series and plotting it in the same figure it became quite clear that milk production is depends on composition of TMR. The plot showed only hay percentage values in the feed, but since there are only two ingredients in TMR, so the lack of corn will cause milk production to stagnate.

To see whether seasons were significantly different from each other Tukey's HSD pairwise testing was performed and it calculated that these values usually do not tend to close up with each other.

5.3. Exploratory functional data analysis

Functional data analysis provides information about curves that a varying over a continuum. Many researches were conducted using FDA[6]. Systematic review on applications of functional analysis was conducted in 2013 in research article "Applications of functional data analysis: A systematic review" by Shahid Ullah and Caroline Finch. They compiled a list of 84 articles that in one way or another applied smoothing technique, functional principal component analysis, clustering adjustment, functional linear modelling or forecasting. In reviewed articles most, common spheres of application were biomedicine, biomechanics and linguistics, totalling 40% of revied papers. Since first found application of FDA was in year 1995 and more frequent adaptation of the analysis technique beginning in year 2005, it is worth to explore whether functional data analysis could be spread into precision livestock farming.

[6] Functional Data Analysis

The functions with features may be both unpredictable and complicated. Consequently, we need a strategy for constructing functions that works with parameters that are easy to estimate and that can accommodate nearly any curve feature.

Since the raw data has quite sharp changes, the smoothing is necessary. Since milk yield data is noncyclical and nonperiodic B-spline basis was chosen. Splines are more flexible and therefore more complex that finite Fourier basis. Splines are constructed by dividing a functional observation into subintervals. The subintervals are divided by break points or and data within these intervals take shape of polynomials. For smoothing this dataset B-spline basis was used.

Generalized cross-validation criterion is widely used to determine optimal value of a variable that is entering a regularization exercise. GCV is a straightforward score that is calculated by dividing sum squared residuals S_t by $t(1-\delta_t/t)^2$ value, here δ represents degrees of freedom. Full formula for calculating GCV is:

$$GCV_t = \frac{S_t}{t(1-\delta_t/t)^2} \qquad (1)$$

B-spline basis with fourteen knots and smoothing parameter of $\lambda = 0.1$, was chosen as it showed to minimize generalized cross-validation criterion (Figure 8).

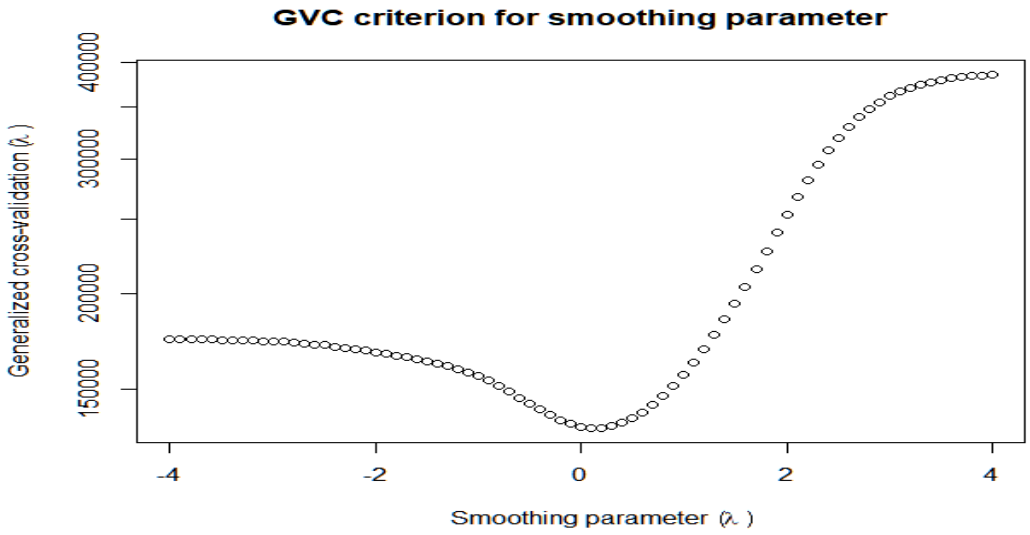

Figure 8 The values of the generalized cross-validation criterion for choosing the smoothing parameter λ for fitting milk yield curves.

The smoothing resulted in neatly looking functions, which will be easier to explain and fit (Figure 9). It is noticeable that some curves should be considered outliers since they are noticeably lower or higher than the average. Smoothed data does not show any sudden changes. Mean that is represented in

the figure seem to have slight upwards trend and shows a slight decrease ant the beginning of the months and slight increase at the end of the months.

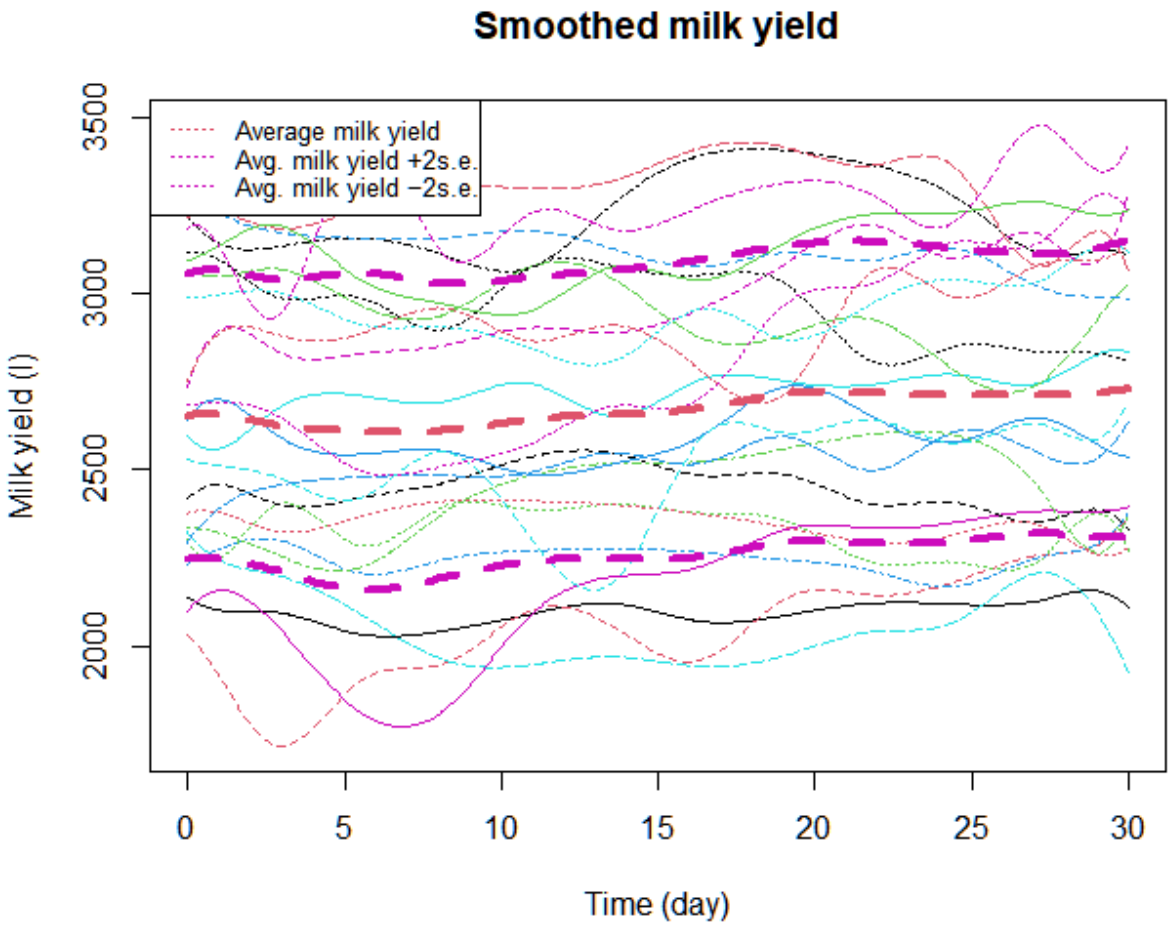

Figure 9 Segmented and smoothed milk output data. Dashed red line represents mean, dashed pink lines represents ± two standard errors.

To see how smoothed curves fit raw data it is advantageous to plot smoothed function, confidence intervals along on top of unsmoothed raw data (Figure 10). In the plot solid curve shows smoothed functional observation and dashed lines represent 95% pointwise confidence interval for smoothed curve. Some data points are outside the confidence interval, though it is not desirable, the smoothed curve captures the main trend, smoothing sharp changes and not overfits the raw data. Inside this figure are depicted 4 months, each representing different season. In top left plot data points dip quite significantly because during this period of time there was no corn inside TMR. Plot in top right corner may express effects for grazing or growth in lactating cows number. The bottom left visualization shows quite a steep dive of milk yield. The fall may be caused by combination of no good grazing grounds and no corn in

TMR. Finally, the bottom right plot displays a good growth period followed by a big decrease which might be caused by bad TMR management i.e., an increase in hay percentage in TMR. All in all, smoothed data is a good functional representation of nonsmoothed data.

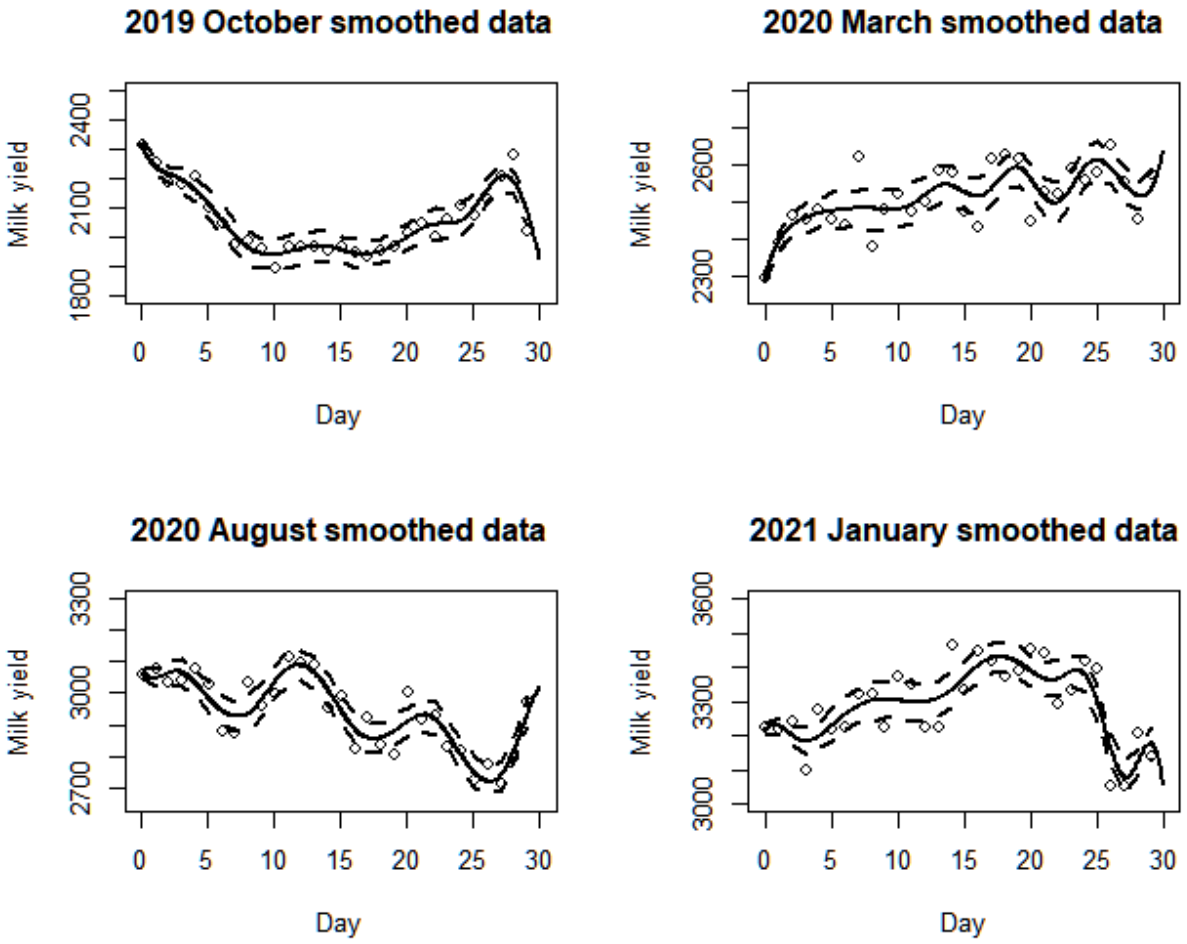

Figure 10 The solid curves are smoothed using B-spline basis with 14 break points. The dashed lines indicate 95% pointwise confidence limits for the smooth curve based on the raw data points shown as circles.

To see how residuals are changing fluctuating they are plotted (Figure 11). No month has large oscillations at the start, but around 5 days in a big spread happens. Smoothing does not appear whether it displays larger positive or negative residuals, but during 5-th to 10-th and 23-th to 26-th days the smoothing tends to have larger residuals. Residuals peak with $\pm$200 litres, and residual of 80 litres seems common. Considering how sharply and suddenly milk yield was changing during this time period these residuals are reasonable and at the peaks are about 10% of real value.

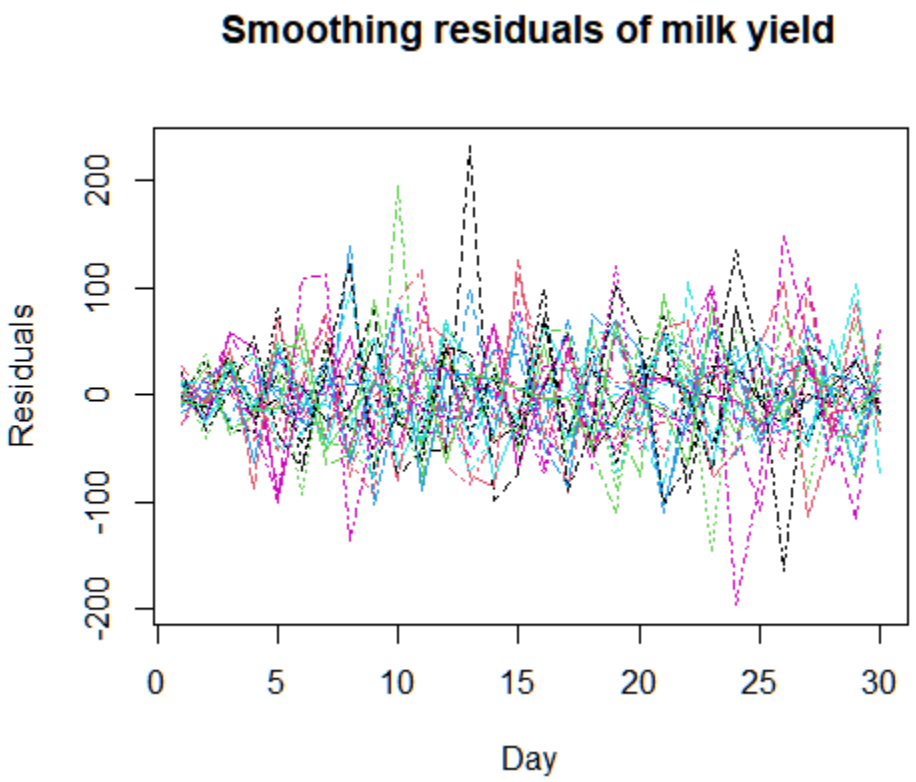

Figure 11 Residuals generated by smoothed milk yield data.

To better understand the variation during the months it is beneficial to observe variation surface and contour plots (Figure 12). The left variance-covariance surface shows a three-dimensional variance distribution of milk yield variance as the hight of the diagonal running from (0,0) to (30,30). There is much more variation in the first and middle part of the months.

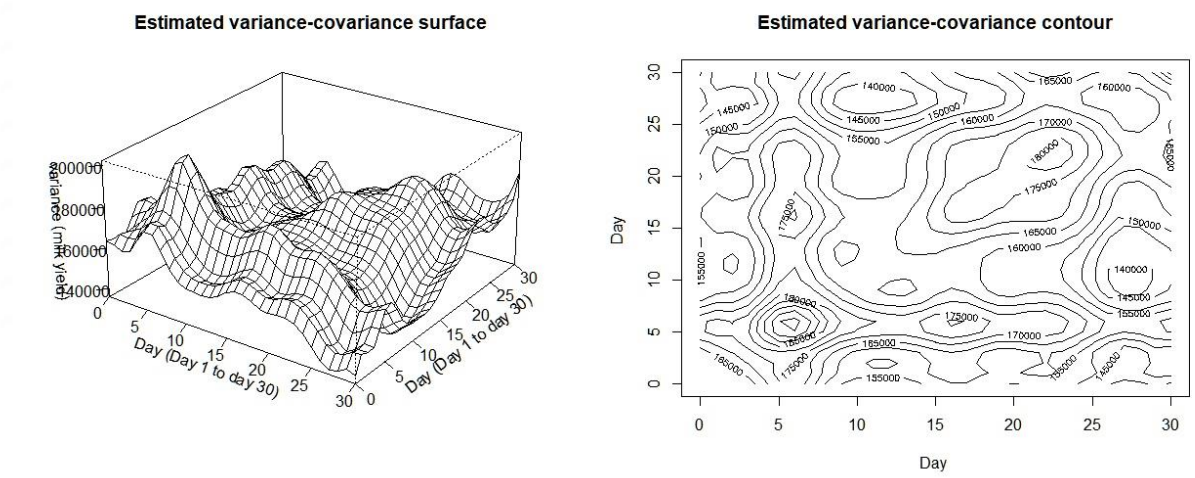

Figure 12 The estimated variance-covariance surface for the milk output data and contour plot for the bivariate variance-covariance surface for the milk output data.

Principal component analysis is often the first method that is used after descriptive statistics and plots. We want to see what primary modes of variation are in the data, and how many of them often seem to be substantial. The eigenvalues of the bivariate variance-covariance function are indicators of the importance of these principal components, and plotting eigenvalues is a method for determining how many principal components are required.

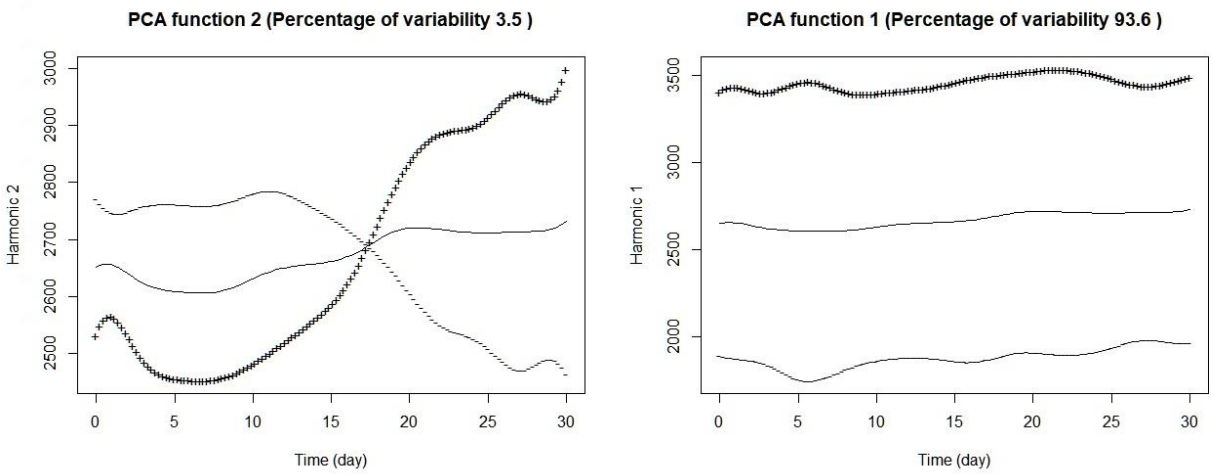

Figure 13 Two principal component functions or harmonics are shown as perturbations of the mean, which is the solid line. The +'s show what happens when a small amount of principal component is added to the mean, and the –'s show the effect of subtracting the component.

Calculations revealed that first two principal component functions have highest eigenvalues and account for total of 97.1% of variability (Figure 13). The two principal component functions by displaying the mean curve along +'s and -'s indicating the consequences of adding or subtracting a small amount of each principal component. The first harmonic, located on the right, is accounting for 93.6% of variation and represents a relatively constant mean and large impact interval. The second harmonic shows a contrast between the first half of the month and the second. The main take away from these two harmonics could be that do to first harmonic's wide range the output can change its direction quickly and unpredictably while the mean stays stable and the second harmonic suggests a decreasing trend in the first half of the month and a contrasting growing trend in the second half of the month.

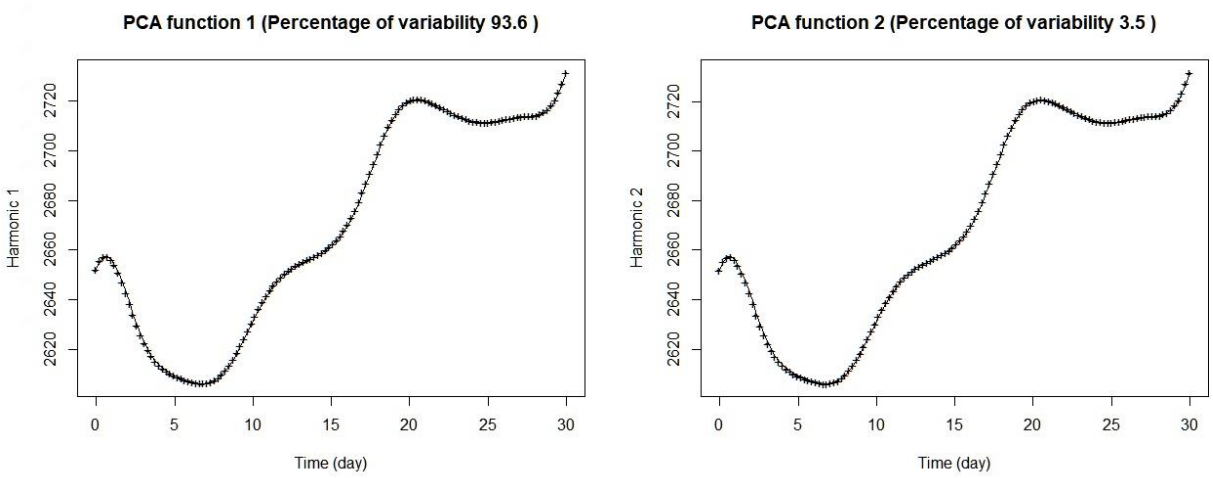

Figure 14 The two rotated principal component functions are shown as perturbations of the mean which is the solid line. Left panel contains the strongest component, but both components do not show any significant variation around the mean.

Since unrotated functional components do not reveal any evident variation, it is emphasized to investigate the rotated functional components (Figure 14). First two rotated harmonics are depicted in they account for 97.1% of variation. The two rotted principal components are shown as perturbations of the mean, which is follower by +' and –'s very closely indicating that the main variation originates in the mean. To figure out the months that are varying a lot rotated principal components are plotted as pairs (Figure 15). The pairs are presented as circles and are named according to which month they represent. This way of plotting reveals how variance is distributed along the diagonal. There seems to be about three clusters. The first cluster resigns at the bottom left corner and includes only summer and autumn months from year 2019, the second cluster is distributed just a bit higher and almost exclusively includes months that corn was not added to TMR, the third cluster of variations is located in the top right corner with months during which milk production was high. The model tends to overestimate time periods when production was lower than average and underestimate high yield months.

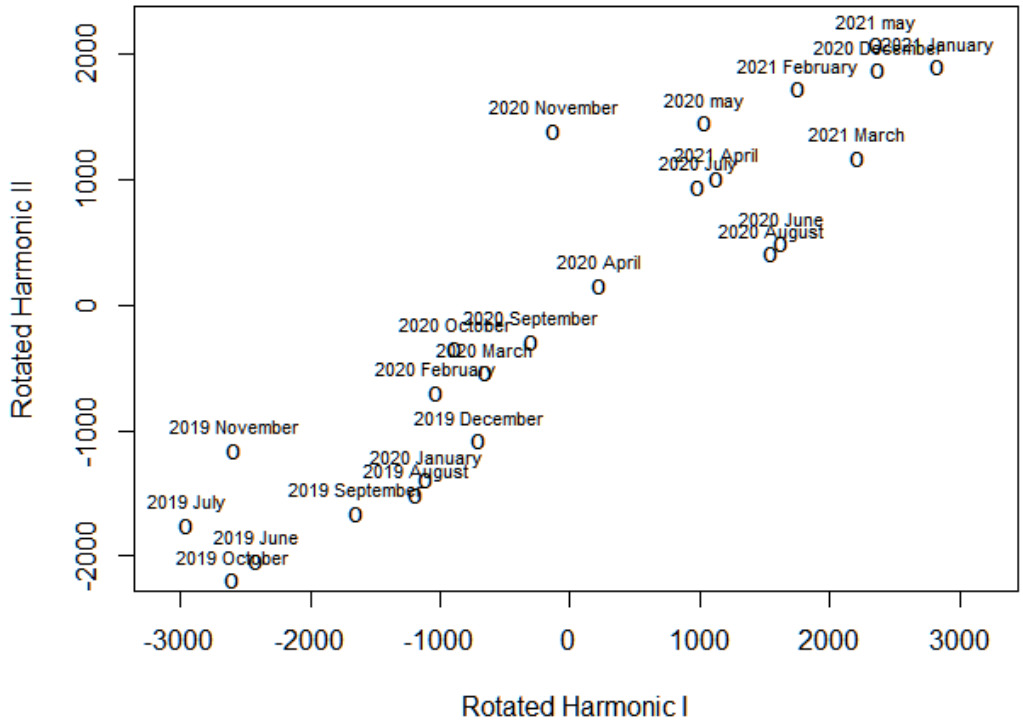

Figure 15 The scores for two rotated principal component functions are shown as circles. They correspond to years and months as they are labelled.

For data smoothing generalized cross-validation values were calculated to get an output of optimal $\lambda = 0.1$. Smoothed data fitted data points well while not overfitting. Rotated principal components plotted in scatter plot showed that smoothing model tends to overestimate time periods when production was lower than average and underestimate high yield months.

5.4. Functional Responses of milk yield to functional predictors

Functional data exploratory analysis is a great tool for getting familiar with data at hand, but it does not give any conclusions on how and by what the variable in question is influenced. The influence of total mix ration on milk yield is obvious, after all a cow eats what it gets. But does this influence change over time and how well can functional response of milk yield be described by functional predictor of hay or corn be described?

The following model is called concurrent. It relates the value of $y_i(t)$ to value of $x_{ij}(t)$ at the same time periods, meaning that past values of predictor do not influence value of response ahead or before in time. The concurrent model takes form off:

$$y_i(t) = \beta_0(t) + \sum_j^{q-1} x_{ij}(t)\beta_j(t) + \varepsilon_i(t) \quad (2)$$

Here: $\beta_0(t)$ effect of a constant at period t, $\sum_j^{q-1} x_{ij}(t)\beta_j(t)$ calculates the effect of external variable on regressand, with q equal to functional observation of regressor.

The first regression focuses on corn influence on milk yield (Figure 16). The top plot displays value of the intercept as solid black line and dotted line shows the average milk yield not influenced by external influences. The bottom plot represents corn coefficient as solid black line and $R^2$of the regression. As displayed in plots milk yield intercept effect dips when corn coefficient goes up. Thus, indicating that when more variance is explained by corn coefficient, prediction is less dependent on the average of milk yield. It is worth noticing that R^2 of equation is never rising far above 0.2 value. The lack of high squared multiple correlation suggests that the model is not good at giving explanations about milk yield. During all 30 days this coefficient stays above 0, thus indicating that it definitely increases milk yield whist it is included in total mix ration.

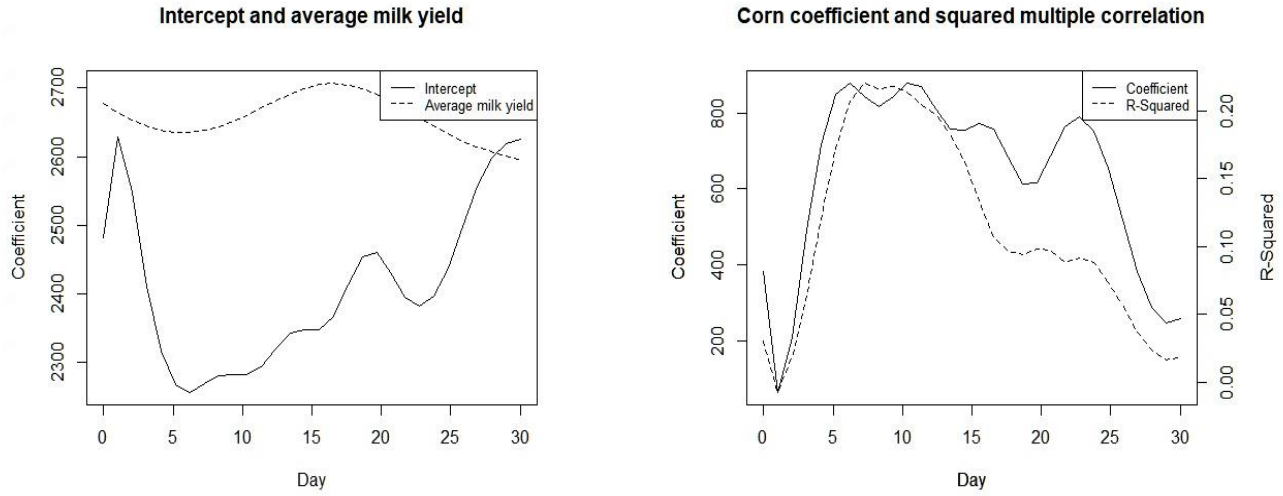

Figure 16 The top panel shows as a solid line the intercept term in predicting milk yield from corn percentage in TMR, the dashed line indicates the mean milk yield assuming no corn effect. The bottom panel shows as a solid line the functional regression coefficient in concurrent functional linear model and the dashed line shows the squared multiple correlation coefficient of the model.

The second regression focuses on hay influence on milk yield (Figure 17). The top plot displays value of the intercept as solid black line and dotted line shows the average milk yield not influenced by external

influences. The bottom plot represents corn coefficient as solid black line and $R^2$of the regression. As displayed in 16-th figure milk yield intercept effect dips when corn coefficient goes up, indicating that when more variance is explained by corn coefficient, prediction is less dependent on the average of milk yield. It is worth noticing that R^2 of equation is never rising far above 0.2 value and sometimes tips below zero. The lack of high *squared multiple correlation* suggests that the model is not good at giving explanations about milk yield, furthermore when it goes below zero it suggests that model does not follow trend and the straight line explains main variable better.

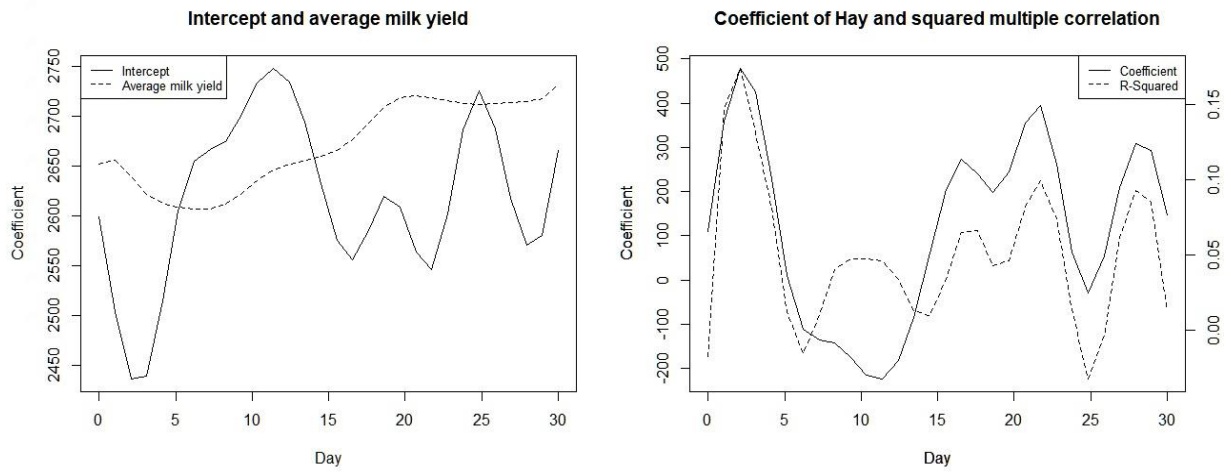

Figure 17 The top panel shows as a solid line the intercept term in predicting milk yield from hay percentage in TMR, the dashed line indicates the mean milk yield assuming no corn effect. The bottom panel shows as a solid line the functional regression coefficient in concurrent functional linear model and the dashed line shows the squared multiple correlation coefficient of the model.

The coefficients of hay and corn suggests reasonable conclusions that is when cows feed on less nutritious fodder they cannot produce the same levels of milk. Since this concurrent model does not pass past values' effects on to future responses it would be reasonable to omit months when corn was out of stock. Changing functional observation length to 90 days would help model to capture the trend better.

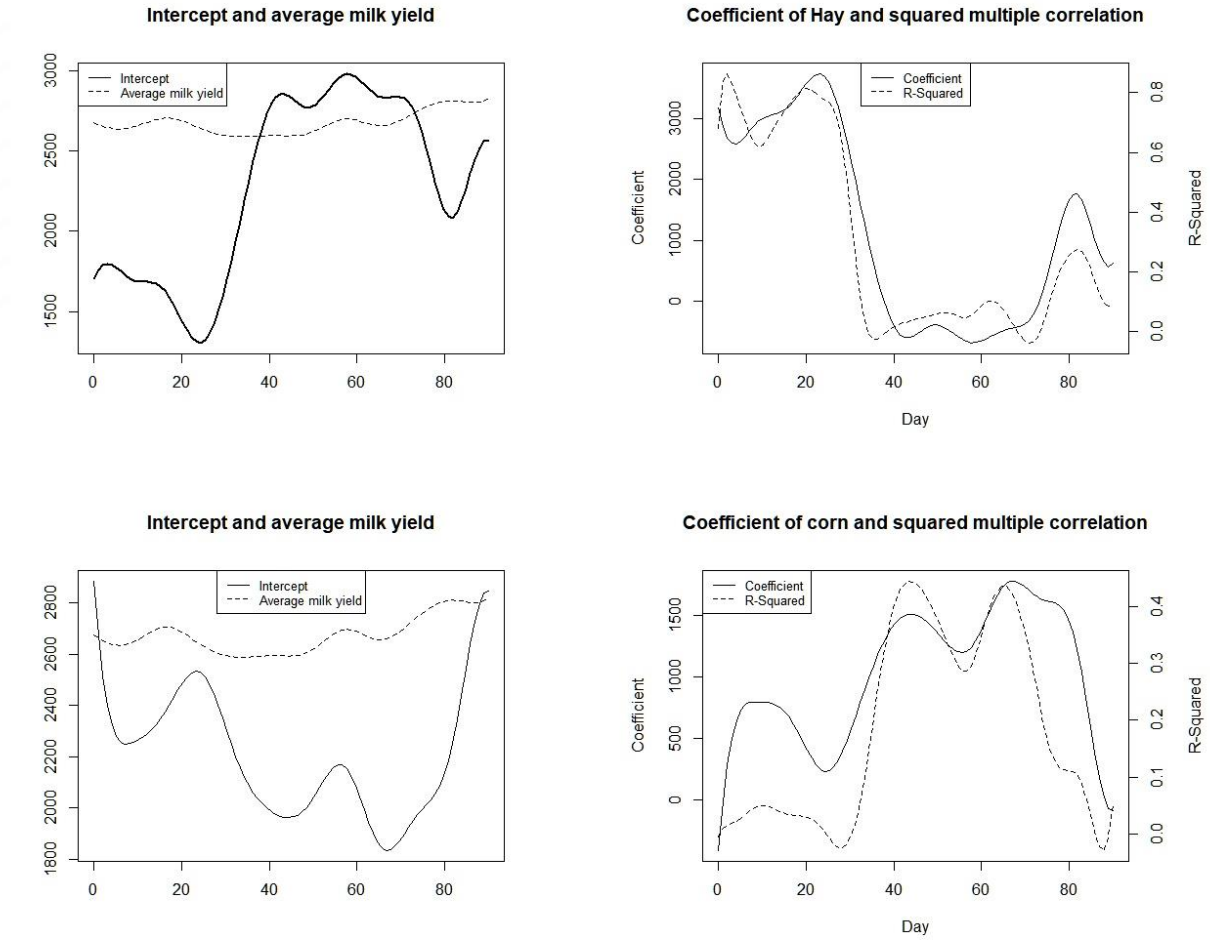

Figure 18 The top left panel shows as a solid line the intercept term in predicting milk yield from hay percentage in TMR, the dashed line indicates the mean milk yield assuming no external viable effect. The top right panel shows as a solid line the functional regression coefficient in concurrent functional linear model and the dashed line shows the squared multiple correlation coefficient of the model. Bottom panels represent intercept and corn coefficient for concurrent model where dependent variable is milk yield and percentage of corn in TMR is the regressor.

Concurrent functional models with adjusted observation length are shown to be more accurate than models with shorter functional observation length and includes zeros (Figure 18). The top row of this figure represents concurrent model with hay percentage in TMR as independent variable and bottom row of plots represent concurrent model with corn percentage in TMR as independent variable. As in previous figures these plots contain Intercepts, coefficients and squared multiple correlation coefficients. Differently than the previous models, these models have a lot higher R^2 values.

For these models' dataset was augmented by rearranging twenty four 30 day functional observations into seven 90 day functional observations. Since one of the new observations contained almost

exclusively zeros it was removed from the dataset. This augmentation resulted in new smoothing parameter $\lambda = 0.3$. This new smoothing parameter was the same for both models.

Both new models show essentially the same results as the two old ones:

- During some periods of time R^2 dips below zero,
- Models are ideologically correct, but statistically insignificant

The Values of coefficients and their importance in both regressions tend to fall when another rises. In short period functional regression, the coefficient of hay stays positive almost for whole duration of a functional observation, whilst in long period regression it is positive in first 30 days but drops in remaining 60. In long period regression after coefficient in a of hay drops, the coefficient of corn rises to a significant level, which is not observable in the short period. In both regressions when hay is the causing variable the intercepts are high, thus indicating that there are a big proportion of milk yield not explained by percentage of hay. The regressions of corn over milk yield results in differing results. In short period regression hay effect is more sporadic and whole regression is worse fitting than the long period one. Regression with observations lasting for 90 days reveals that corn is increasingly more significant and influential towards the later parts of an observation. In short observation functional regression at the peaks corn and hay coefficients are similar, but hay influence is more undefined. The coefficients of hay and corn during their peaks in long observation functional regression differ almost two times. Here coefficient of corn may reach 1500 and the coefficient of hay may reach 3000. Persistently higher coefficients of corn suggest that including this ingredient in the TMR is more impactful. Regressions with short and long period observations show that both variables create an impact on milk yield.

6. Conclusions

The composition of a total mixed ration can have a significant influence on milk yield in dairy cows. A TMR is a type of feed that is designed to provide cows with a balanced diet that meets their nutritional needs. The proportion of forages (such as hay or silage) to concentrates (such as grains or protein supplements) in the TMR can impact milk yield, as forages provide nutrients that are important for maintaining the health of the cow's rumen, while concentrates provide more energy and protein that can be used for milk production. Other factors that can influence milk yield include the quality and nutrient content of the feed ingredients, the overall energy content of the TMR, and the cow's stage of

lactation. To optimize milk yield, it is important to consider a wide range of factors and to work with a nutritionist or other expert to develop a TMR that meets the nutritional needs of the cows.

In graphical analysis of milk yield it is noticed that it has a strong upwards trend and functional observations are spread out over a wide interval of values. Scatterplot of rotated principal components showed that smoothing milk yield with B-spline basis produced a model that it overestimate time periods when production was lower than average and underestimate high yield months. The smoothing model for milk yield data worked. Means of ninety-day functional observations were checked using Tukey's honest significance test showed that not all seasons share statistically similar average during the length of observation.

During functional data analysis the milk yield and how it is influenced by composition of course fodder was under the spotlight. It was found that milk yield reacted to different composition of total mix ration. Functional data regression showed that influence of external variables changes over short and long periods of time. The regression analysis conveyed information that corn impacts milk yield with larger effect, though this effect was not always significant.

All in all, functional data analysis is still a relatively new way of analysing data and its usage should not be contained only in medical applications. As shown in this master's thesis it is a valid and informative approach to evaluate effects on dependent variables, it is showing an impact of external variables over period of time. For any future research using more individual data about cow performance would be advisable. Data used in this situation is additive of the production of a herd. In consequence it is influenced not only by hay or corn amount in total mix ration. Other factors that can influence milk yield include the quality and nutrient content of the feed ingredients, the overall energy content of the total mix ration, and the cow's stage of lactation (early lactation is typically associated with higher milk yield). Additionally, factors such as cow genetics, health, and management practices (such as milking frequency and herd size) can also have an impact on milk yield.

7. References

1. Popescu, Research on milk cost, return and profitability in dairy farming, *Economic Engineering in Agriculture and Rural Development, v. 14*, University of Agricultural Sciences and Veterinary Medicine Bucharest. 2014, p.p. 219-222.

2. Bonora, F., Tassinari, P., Torreggiani, D. and Benni, S., An innovative mathematical approach for a highly informative treatment of automatic milking system datasets: development and testing of enhanced clustering models, *8th European Conference on Precision Livestock Farming*, Nantes, September 2017, p.p. 692–700.

3. Kamphuis, J.W. van Riel, R.F. Veerkamp, R.M. de Mol Wageningen, Traditional mixed linear modelling versus modern machine learning to estimate cow individual feed intake, *8th European Conference on Precision Livestock Farming*, University & Research, Livestock Research, PO Box 338, the Netherlands, 2017, p.p. 366-376.

4. Reinemann, M. Zucali, V. Inzaghi, P. Thompson, J. Penry, P.S. Boloña and J. Upton, Prospects for precision milking management in automatic milking systems, University of Wisconsin, Madison, Wisconsin, USA; University of Milan, Italy; University of Melbourne, Veterinary School VCC, Werribee, Victoria, 3030, Australia; Teagasc, Animal & Grassland Research and Innovation Centre, Moorepark, Fermoy, Ireland, 2013, p.p. 212-219.

5. H. Hogeveen and W. Steeneveld, Essential steps in the development of PLF systems for the dairy sector Business Economics group, *6th European conference on Precision Livestock Farming* ,Wageningen University, Hollandseweg, 6706 KN Wageningen, The Netherland Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, Yalelaan, 2013, pp 47-55.

6. R. V. Sousa, L. S. Martello, An artificial neural network predictive model for assessing thermal stress in beef cattle using thermal radiation, *6th European conference on Precision Livestock Farming*, Faculty of Animal Science and Food Engineering, University of São Paulo, Av. Duque de Caxias Norte 225, Pirassununga, 2017, p.p. 415-423.

7. S. Benni, F. Bonora, P. Tassinari, D. Torreggiani, A methodology for daily analysis of AMS data providing herd characterisation and segmentation, *9th European Conference on Precision Livestock Farming*, Department of Agricultural and Food Sciences, University of Bologna, Bologna, 2019 p.p. 53-59.

8. The Andersons Centre, The structure of the GB dairy farming industry – what drives change?, Nottingham university, 2013, p.p. 53.

9. J.O. Ramsay, G. Hooker, S. Graves, Functional Data Analysis with R and MATLAB, Springer, 2009, p.p. 208.

8. Appendices

8.1. APPENDIX 1

Information about data used in regression analysis (source: "Traditional mixed linear modelling versus modern machine learning to estimate cow individual feed intake", C. Kamphuis, J.W. van Riel, R.F. Veerkamp, and R.M. de Mol Wageningen, 2017).

Table 1: Characteristics of trials used for training and testing including the number of cows, cow-days, and treatments (Treatm.) per trial (Trial), the range of parity and days in milk (DIM) per trial, and the year(s) in which the trial was conducted.

Trial	Treatm. (n)	Cows (n)	Cow-days (n)	Parity range	DIM range	Year
1	4	136	399	1 – 7	70 – 160	2014
2	3	52	75	1 – 6	42 – 207	2014–2015
3	3	96	2,594	1 – 7	7 – 364	2014–2015
4	1	10	65	3 – 5	6 – 56	2014
5	1	15	39	1 – 5	21 – 56	2014
6	3	39	177	2 – 9	14 – 140	2015
7	5	59	438	1 – 5	7 – 63	2014–2015
Total	20	407	3,787			

8.2. APPENDIX 2

Models their descriptions and external viables (source: “Traditional mixed linear modelling versus modern machine learning to estimate cow individual feed intake”, C. Kamphuis, J.W. van Riel, R.F. Veerkamp, and R.M. de Mol Wageningen, 2017).

Model	Description	Variables
MLR1	Mixed Linear Regression without weather info	Fixed effects: Parity, live-weight, fat and protein corrected milk Random effects: trail, treatment within trail, cowed, week in milk, month of the year
MLR2	Mixed Linear Regression with weather info	Fixed effects: Parity, live-weight, fat and protein corrected milk, temperature, humidity Random effects: trail, treatment within trail, cowed, week in milk, month of the year.
BRT1	Boosted Regression Tree without weather info	Parity, live-weight, fat and protein corrected milk, week in milk, month of the year
BRT2	Boosted Regression Tree with weather info	Parity, live-weight, fat and protein corrected milk, temperature, humidity week in milk, month of the year

8.3. APPENDIX 3

Tukey's HSD test for testing pointwise comparison of every 30 day functional day observations mean for similarity.