SYSTEMATIC PATTERN RECOGNITION AND MODELING WITH IMPERFECT DATA:

AN INTEGRATION OF DATA SCIENCE, DATA MINING, MACHINE LEARNING, AND EPIDEMIOLOGY

SYSTEMATIC PATTERN RECOGNITION AND MODELING WITH IMPERFECT DATA

MARLÈNE TREMBLAY

MARLÈNE TREMBLAY

SYSTEMATIC PATTERN RECOGNITION AND MODELING WITH IMPERFECT DATA

AN INTEGRATION OF DATA SCIENCE, DATA MINING, MACHINE LEARNING, AND EPIDEMIOLOGY

9 789463 803847

Systematic Pattern Recognition and Modeling with Imperfect Data:

An integration of data science, data mining, machine learning, and epidemiology

Marlène Tremblay

2019

Systematic Pattern Recognition and Modeling with Imperfect Data: An integration of data science, data mining, machine learning, and epidemiology
Marlène Tremblay, 2019

PhD dissertation, Utrecht University, The Netherlands
- with summaries in Dutch and English-

ISBN: 978-94-6380-384-7
Printed by: Proefschrift Maken, Utrecht, the Netherlands

Printing of this thesis was financially supported by the Department of Farm Animal Health of the Faculty of Veterinary Medicine, Utrecht University

Systematic Pattern Recognition and Modeling with Imperfect Data

An integration of data science, data mining, machine learning, and epidemiology

Systematische patroonherkenning en modellering met imperfecte gegevens: een integratie van data science, datamining, machine learning en epidemiologie

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 2 juli 2019 des ochtends te 10.30 uur

door

Marlène Tremblay

geboren op 27 mei 1987
te St-Jean-Sur-Richelieu, Canada

Promotor:
Prof. dr. J.A. Stegeman

Copromotor:
Dr. D. Döpfer

TABLE OF CONTENTS

SECTION 2- UNSUPERVISED LEARNING

CHAPTER 1- GENERAL INTRODUCTION

BACKGROUND

The accelerated rate of technological innovation in the 21st century has yielded many large data sets requiring new methods and approaches for analysis. In the past, data collection was mostly done after the development of a hypothesis-based experimental design and project plan (primary data based studies). In contrast, data from secondary data-based studies are large data sets, potentially big data, collected before the analysis' method or goal is determined (Sørensen, 1996; Olsen, 2008). Secondary data are often collected with the help of automated data collection and storage systems and they can lack a predefined experimental design and data management plan (Boslaugh, 2007; Van den Broeck, et al., 2013). Both types of data are widely used for parameter estimation, prediction modeling and pattern recognition.

It is common for secondary data sets to be used in veterinary epidemiological studies (Emanuelson and Egenvall 2014; Egenvall et al., 2009). Animal production data is a great example of secondary data with the potential to become big data (Kelton et al, 1997; Penell, 2009). Historic health records, and production data of poultry, aquaculture, swine and cattle are routinely extracted for epidemiological studies (Kelton et al, 1997; Penell, 2009). Other examples of secondary data in veterinary medicine include data from breed registries and associations, medical insurance data and medical records from veterinary clinics and hospitals (Egenvall et al., 2009; Penell, 2009).

Data from secondary data-based studies can be limited by the nature of their collection method, which can cause several types of bias (Sorensen et al., 1996; Terris et al., 2007; Olsen, 2008; Murakami, 2014). However, secondary data benefit from their large sample sizes, and the increased detail from larger numbers of variables. These benefits can yield to a significant increase in statistical power for inferences compared to primary data-based data sets (Emanuelson and Egenvall, 2014). Additional benefits of secondary data sets include how they are fast, inexpensive, allow the reuse of data for optimizing resources, and give access to historic data that can be used for large scale longitudinal studies (Van den Broeck, et al., 2013; Emanuelson and Egenvall, 2014). Secondary data sets also provide new opportunities for pattern discovery, data mining and hypothesis generation.

At the same time, large secondary data-based studies are more prone to result in imperfect data challenges (Emanuelson and Egenvall, 2014). Imperfect data challenges are those that require intensive data preprocessing steps before the data are ready for analysis. Challenges include, but are not limited to, imbalances in positive and negative outcomes, rare events, zero inflation, high-dimensionality, multicollinearity, missing data, multiple significant interactions, variety in structure, undefined outcomes, and having additional variables in the dataset not related to the question at hand (Parsons, 1996; Pearson, 2005; Kochanski et al., 2012). Given the data challenges, analytical methods need to be optimized to result in meaningful inferences.

Epidemiology has a strong foundation in using statistical methods for data analysis (Olsen, 2008; Pfeiffer and Stevens, 2015). However, epidemiology has recently followed trends in data science such as the emergence of data mining and machine learning (Pfeiffer and Stevens, 2015; Alkhamis

et al., 2018). Data mining and machine learning are commonly being used to prepare and analyze veterinary epidemiology data (Valdes-Donoso et al., 2017; Esener et al., 2018; Machado et al., 2019). One of the differences between statistics and data science can be highlighted by the disciplines' movement towards automation (Reid, 2016). Unlike applied statistics, there is the growing movement towards automation in data science (Gaber, 2009; Witten et al., 2016; Cearley, 2019). Several automatic data mining methods and several workflows for full data analysis have already been patented and developed (King et al., 2009; Minkin et al., 2018; Hermans et al., 2018; Norton and Berckmans, 2018). Automation of data analysis has the potential to increase the amount of output, and improve the resulting model performance (Gaber, 2009).

Veterinary epidemiological research, will continue to follow the trends in data sciences towards automation due to the growing use of imperfect, secondary, and potentially big data, and the desire to develop real-time surveillance and prediction capabilities (VanderWaal et al., 2017; Hermans et al., 2018). However, epidemiology will always need to be focused on biological relevance and meaningful interpretation of results. There is concern that data mining methods and automation will cause the field to move away from deductive reasoning and prohibit the incorporation of expert knowledge about biological relevance into the analysis (Dohoo et al., 2003; Faraway, 2016). Therefore, epidemiology needs to adapt or develop methods that can be automated in the future, and that do not remove the focus of research away from the biological relevance and interpretation of the results.

Consequently, there is a need to meet the challenges and special needs of imperfect data from secondary data-based studies for both supervised (i.e. parameter estimation, prediction modeling) and unsupervised learning (i.e. pattern recognition). Second, there is a need for systematic approaches to integrating and comparing of statistical analytical methods to streamline selection, and to prevent subjectivity and flawed outcomes when analyzing imperfect data from secondary data-based studies.

OBJECTIVES

In this dissertation, the goal was to develop solutions for the systematic integration for methods that address imperfect data. Additionally, the goal was to develop solutions for the systematic integration, comparison and selection of methods as a steppingstone towards automation while maintaining the focus on the biological relevance and interpretation of the results.

Given these two goals the following objectives were defined:

The first objective of this dissertation is to meet the challenges and special needs of imperfect data from secondary data-based studies for both supervised (i.e. parameter estimation, prediction modeling) and unsupervised learning (i.e. pattern recognition). This results in incorporating and adapting methods from data science, data mining, and machine learning or developing new methods for imputing missing values, modeling zero inflated data sets, systematically selecting interaction terms, variable selection, addressing imbalances in positive and negative outcomes, rare events, data with hierarchical structure, and using the example of principal component analysis (PCA) for variable reduction, and clustering for pattern recognition.

The second objective is to systematically combine, compare and select the most appropriate statistical methods for parameter estimation, prediction modeling, and pattern discovery in the face of large imperfect data sets, while maintaining the focus on the biological relevance and interpretation of the results.

OUTLINE

Section 1

In **Chapter 2** the focus is on solutions that address imperfect data and systematic supervised learning methods for parameter estimation. In three parts: **Chapter 2.1** describes a systematic approach to addressing imbalances in positive and negative outcomes (unbalanced data sets) for parameter estimation using zero augmented models while in search for the best fitting model. These methods are described using surveillance data from the Foodborne Diseases Active Surveillance Network (FoodNet) in the United States to build a descriptive model for *Campylobacter* infections. **Chapter 2.2** demonstrates a **s**ystematic approach to imputation of missing data, variable reduction and selection using data from the People, Animals and their Zoonoses (PAZ) project out of Kenya to build a description model for *Plasmodium falciparum* infection. **Chapter 2.3** focuses on systematic selection and interpretation of interaction terms for parameter estimation in search for the best fitting model while maintaining the focus on the biological relevance and interpretation of the results. This approach is demonstrated using production data from North American automated milking systems (AMS) with the goal of building a descriptive model for milk production outcomes.

Chapter 3 shows systematic approaches for supervised learning methods for prediction modeling in two parts: **Chapter 3.1** introduces a systematic approach to full model selection for prediction modeling using regression trees. The method is demonstrated using a data set including data from milk Fourier-transform infrared spectroscopy (FTIR), routine milk testing, and from automatic milking systems to predict blood nonesterified fatty acids (NEFA) and β-hydroxybutyrate acid (BHBA) in dairy cows during early lactation. **Chapter 3.2** illustrates the application of regression tree full model selection (rtFMS) methods for chapter 3.1 to a milk FTIR data set to predict rare events of antibiotic residue in bovine milk. In this chapter the problem of micro-macro multilevel modeling is addressed with the development of new method named "Extreme Value Micro-Macro" (EVMM) multilevel modeling. This method addressed the challenge of multilevel modeling when the central tendency of the micro level observations is not a good representation of the macro level outcome.

Section 2

In **Chapter 4** applied unsupervised learning methods for pattern recognition. The results were confirmed using post-hoc analyses and maintained focus on the biological relevance and interpretation of the results. Chapter 4.1 describes a novel classification of poor metabolic adaptation in dairy cows called poor metabolic adaptation syndrome (PMAS) discovered using cluster analysis. Clinical data, blood samples and milk testing data of Simmental cows in Bavaria were used for this purpose. Finally, Chapter 4.2 addresses how decision making processes and customized management advice can be facilitated by improved benchmarking within peer groups by means of clustering AMS data for a diverse set of locations in North America.

All the data set in this dissertation originate from secondary data-based studies.

REFERENCES

Alkhamis, M.A., Brookes, V.J. and VanderWaal, K., 2018. Applications of Novel Analytical Methods in Epidemiology. Frontiers in veterinary science, 5, p.243.

Boslaugh, S., 2007. Secondary data sources for public health: A practical guide. Cambridge University Press.

Cearley, D.W., Burke, B., Searle, S., Walker, M.J. and Claunch, C., 2019. The top 10 strategic technology trends for 2019. Gartner.

Dohoo, I.R., Martin, W. and Stryhn, H., 2003. Veterinary epidemiologic research (No. V413 DOHv). Charlottetown, Canada: AVC Incorporated.

Esener, N., Green, M.J., Emes, R.D., Jowett, B., Davies, P.L., Bradley, A.J. and Dottorini, T., 2018. Discrimination of contagious and environmental strains of Streptococcus uberis in dairy herds by means of mass spectrometry and machine-learning. Scientific reports, 8(1), p.17517.

Egenvall, A., Nødtvedt, A., Penell, J., Gunnarsson, L. and Bonnett, B.N., 2009. Insurance data for research in companion animals: benefits and limitations. Acta Veterinaria Scandinavica, 51(1), p.42.

Emanuelson, U. and Egenvall, A., 2014. The data–Sources and validation. Preventive veterinary medicine, 113(3), pp.298-303.

Faraway, J.J., 2016. Linear models with R. Chapman and Hall/CRC.

Gaber, M.M., 2009. Scientific data mining and knowledge discovery. Springer.

Hermans, K., Opsomer, G., Van Ranst, B., and Hostens, M., 2018. Promises and Challenges of Big Data Associated With Automated Dairy Cow Welfare Assessment. Animal Welfare in a Changing World, p.199.

Kelton, D.F., Bonnett, B.N. and Lissemore, K.D., 1997. Dairy cattle disease data from secondary databases-Use with caution!. Interbull Bulletin, (15), p.3.

King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N. and Sparkes, A., 2009. The automation of science. Science, 324(5923), pp.85-89.

Kochanski, A., Perzyk, M. and Klebczyk, M., 2012. Knowledge in imperfect data. In Advances in Knowledge Representation. IntechOpen.

Machado, G., Vilalta, C., Recamonde-Mendoza, M., Corzo, C., Torremorell, M., Perez, A. and VanderWaal, K., 2019. Identifying outbreaks of Porcine Epidemic Diarrhea virus through animal movements and spatial neighborhoods. Scientific reports, 9(1), p.457.

Minkin, A.M., McNally, M., Knight, W., Major, S., Lamoreaux, R. and Hernandez, L., U2 Science Labs A Montana, 2018. Systems and methods for automating data science machine learning analytical workflows. U.S. Patent Application 15/836,804.

Murakami, Y., 2014. Secondary data analysis of epidemiology in Asia. Journal of epidemiology, 24(5), pp.345-346.

Norton, T. and Berckmans, D., 2018. Precision Livestock Farming: the Future of Livestock Welfare Monitoring and Management?. Animal Welfare in a Changing World, p.130.

Olsen, J., 2008. Using secondary data. Modern epidemiology. Philadelphia PA: Lippincott, Williams & Wilkins, pp.481-91.

Parsons, S., 1996. Current approaches to handling imperfect information in data and knowledge bases. IEEE Transactions on knowledge and data engineering, 8(3), pp.353-372.

Pearson, R.K., 2005. Mining imperfect data: Dealing with contamination and incomplete records (Vol. 93). Siam.

Penell, J., 2009. Secondary databases in equine research (Vol. 2009, No. 59).

Pfeiffer, D.U. and Stevens, K.B., 2015. Spatial and temporal epidemiological analysis in the Big Data era. Preventive veterinary medicine, 122(1-2), pp.213-220.

Reid, D., 2016. Man vs. Machine: The Battle for the Soul of Data Science. In Big Data Challenges (pp. 11-22). Palgrave, London.

Sørensen, H.T., Sabroe, S. and Olsen, J., 1996. A framework for evaluation of secondary data sources for epidemiological research. International journal of epidemiology, 25(2), pp.435-442.

Terris, D.D., Litaker, D.G. and Koroukian, S.M., 2007. Health state information derived from secondary databases is affected by multiple sources of bias. Journal of clinical epidemiology, 60(7), pp.734-741.

Van den Broeck, J., Brestoff, J.R. and Kaulfuss, C., 2013. Epidemiology: principles and practical guidelines (p. 449). Springer.

Valdes-Donoso, P., VanderWaal, K., Jarvis, L.S., Wayne, S.R. and Perez, A.M., 2017. Using Machine learning to Predict swine Movements within a regional Program to improve control of infectious Diseases in the US. Frontiers in veterinary science, 4, p.2.

VanderWaal, K., Morrison, R.B., Neuhauser, C., Vilalta, C. and Perez, A.M., 2017. Translating big data into smart data for veterinary epidemiology. Frontiers in veterinary science, 4, p.110.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

CHAPTER 2- PARAMETER ESTIMATION INTRODUCTION

The starting point for statistical learning is quantifying associations between parameters and an outcome, quantifying statistical significance of associations, and interpreting associations given their biological relevance, direction and background. This is accomplished by means of parameter estimation modeling (James et al., 2013). Parameter estimation modeling is also referred to as inference, descriptive or explanatory modeling of an outcome, or dependent variable. For the purpose of this introduction the term "parameters" is used to mean the independent variables that drive the system. However, different disciplines use different terms including cofactors, risk factors, predictors, co-variate, independent variables, attributes, and features.

Assumptions of parameter estimation model

Parameter estimation models have several constraints. First, parameter estimation models assume linear relationships between parameter and the outcome (sometimes achieved using methods such as transformations, link functions on the outcome, offsets) and that the effects of multiple parameters are additive (Gelman and Hill, 2006). Second, parameter estimation models assume independence of errors (homoscedasticity) and normal distributions of residues (Pinheiro and bates, 2006; Gelman and Hill, 2006). Third, parameter estimation models are constrained with regards to the number of parameter estimates that can be included in the model (Bishop, 2006; Kuhn, 2013). Parameter estimation models assume that parameters are not highly correlated. Additionally, parameter estimation models rely on having parameters that are meaningful and biologically relevant to the associations under study (Gelman and Hill, 2006). The relevance of parameters in conjunction with the previously mentioned assumptions will determine the model's goodness of fit and therefore the model's reliability for representing the real-world observations. Representing the real-world situation is also accomplished by having measures of uncertainty for parameter estimates such as confidence intervals.

Secondary Imperfect Data Challenges

High dimensionality, high correlation and interactions among parameters, clustering of data and non-normal highly-skewed distributions of observations are some imperfect data challenges that violate parameter estimation modeling assumptions. These challenges are frequent aspects of large secondary data. High dimensionality refers to data sets that have a large number of parameters relative to the number of observations (Hastie, 2009). Including too many parameters can lead to overfitting when using the standard least squares or maximum likelihood estimation technique (Bishop, 2006; Kuhn, 2013). These data require the number of variables to be reduces as to not overfit a model (Kuhn, 2013). Many times, high dimensional data also have variables that are highly correlated amongst themselves. High correlation among parameters (multicollinearity) is another violation of parameter estimation modeling assumptions that can cause models to become unstable, and increase the variance and decrease the accuracy of parameter estimations (Matignon, 2007; Kuhn, 2013). The term interaction is used to describe when one parameter's effect on the outcome depends on the value of another parameter. Not accounting for interactions within a model will lead to non-additive effects in a model and will also cause a model's parameter estimations to become confounded (Hosmer et al., 2013; Faraway, 2016). Finally, clustering among observations is another challenge that violates the assumptions of parameter estimation

models (Zuur et al., 2009; Gelman and Hill, 2006). Overall, parameter estimation modeling is more constricting than prediction modeling. Prediction modeling will be discussed in chapter 3. Solutions for these imperfect data challenges given the constraints of parameter estimation models need to be accurate, reproducible, transparent, justifiable, and transferable.

Approaches to data imperfections

Before designing interventions and taking decisions based on the outcome of a parameter estimation models imperfect data challenges need to be solved as stated in **objective 1 and 2.**

In Chapter 2.1, a parameter estimation model for *Campylobacter* infections in the United States was developed using surveillance data from the Foodborne Diseases Active Surveillance Network (FoodNet). This surveillance data presented the potential challenge of zero-inflation. A case of zero-inflation is characterized by excess zero case counts compared to the positive case counts in a model (Dohoo et al., 2012). Zero inflation is associated with overdispersion, which is when the variance is larger than the mean of the data (Dobson and Barnett, 2018). The presence of overdisperson violates the assumption of a normal error distribution as discussed above (Zuur et al., 2009). Therefore, zero-augmented modeling methods including zero-inflated and hurdle models were used to address extra zeros in the FoodNet data. Zero-augmented models have two different model components: one binomial distribution to model zero case counts and a negative binomial distribution to model the positive case counts (Dohoo et al., 2012). To address the second objective, from the introduction (page 7), a systematic comparison of the zero-augmented and non-zero-augmented models was performed using the models' goodness of fit measures as guides. Although only 5 models were compared, **the systematic approach to comparing goodness of fit of parameter estimation models is a foundation for the automation of systematic comparisons.**

In Chapter 2.2 a parameter estimation model for *Plasmodium falciparum* infection was developed. The data used in this study originated from the People, Animals and their Zoonoses (PAZ) project out of Kenya. A data set that has highly correlated variables, large numbers of parameters and missing values is a good representation of large secondary data and their associated imperfections. Observations with missing values are normally removed from an analysis (Faraway, 2016). However, the same rate of missingness per parameter can have bigger consequence in a high dimensional data sets because of the larger number of parameters. Therefore, it is better to address missingness in a high dimensional data set. Consequently, imputation was used for this data set to address values that were missing at random. When using secondary data, one is more likely to come across data detailed at a different level than is desired. This was the case with the wealth parameters in the PAZ data set. Instead of having one or two parameters that represented overall wealth, this data set had thirty. To deal with variable extraction of disaggregated wealth variables, principal component analysis (PCA) was used. Finally, to address variable selection in a high dimensional data set an Elastic-Net regularized generalized linear model (glmnet) was used (Hastie et al., 2009). **The systematic approach to variable extraction, variable selection and missingness described in this chapter would make many large datasets more manageable and informative for decision-making processes avoiding modeling bias.**

In Chapter 2.3, a complex data set with many parameters collected from farms with automatic milking systems (AMS) was used. It had many potential interactions and confounding effects

among the many parameters when building a model to find associations with milk production. Interactions need to be addressed to meet the assumption that there is a linear relationships between each parameter and its outcome. Interactions are normally selected to be included in the model by using personal knowledge of biological causation involving interactions (Dohoo et al., 2012, Faraway, 2016). It has been suggested that automated selection of interactions should be avoided (Faraway, 2016). If automated selection is warranted, backwards elimination is commonly used (Mantel, 1970; Heinze et al., 2018). In the face of numerous interaction terms, the backwards step elimination procedure was not applicable for this data set since the model could not accommodate all interaction terms at once for the first step of backwards elimination. Parameter estimation models for many large high dimensional data sets suffer from similar limitations. Therefore forward selection was utilized for the selection of interactions in this AMS data set aimed at optimizing model fit. Although systematic selection of interactions with forward selection does not take into consideration the interactions' biological relevance, the biological relevance was emphasized during the interpretation of the selected interactions. **This work illustrates the potential and need for automated model selection. Finally, this work also illustrates that automation of variable selection still requires in-depth interpretation of the biological significant of the results.**

REFERENCES

Bishop, C.M., 2006. Pattern recognition and machine learning. springer.

Dobson, A.J. and Barnett, A.G., 2018. An introduction to generalized linear models. CRC press.

Dohoo, I.R., Martin, S.W. and Stryhn, H., 2012. Methods in epidemiologic research.

Faraway, J.J., 2016. Linear models with R. Chapman and Hall/CRC.

Gaber, M.M., 2009. Scientific data mining and knowledge discovery. Springer.

Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Hastie, T., Tibshirani, R. and Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction.

Heinze, G., Wallisch, C. and Dunkler, D., 2018. Variable selection–a review and recommendations for the practicing statistician. Biometrical Journal, 60(3), pp.431-449.

Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. Applied logistic regression (Vol. 398). John Wiley & Sons.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Kleinbaum, D.G. and Klein, M., 2010. Logistic regression: a self-learning text. Springer Science & Business Media.

Kuhn, M. and Johnson, K., 2013. Applied predictive modeling (Vol. 26). New York: Springer.

Mantel, N., 1970. Why stepdown procedures in variable selection. Technometrics, 12(3), pp.621-625.

Matignon, R., 2007. Data mining using SAS enterprise miner (Vol. 638). John Wiley & Sons.

Pinheiro, J. and Bates, D., 2006. Mixed-effects models in S and S-PLUS. Springer Science & Business Media.

Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A. and Smith, G.M., 2009. Mixed effects models and extensions in ecology with R. Springer Science & Business Media.

CHAPTER 2.1

Evaluation of the Use of Zero-Augmented Regression Techniques to Model Incidence of *Campylobacter* Infections in FoodNet

Foodborne pathogens and disease, 14(10), pp.587-592.

M. Tremblay[1], S.M. Crim[2], D.J. Cole[3], R.M. Hoekstra[2], O.L. Henao[2], D. Döpfer[1]

[1] Department of Medical Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, 2015 Linden Drive, Madison, WI 53706 USA
[2] National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333 USA
[3] USDA-APHIS-Veterinary Services, Centers for Epidemiology and Animal Health, 555 South Howes Street, Fort Collins, CO 80521

Corresponding author:
Marlène Tremblay, DVM
724-288-5159
mtremblay@wisc.edu
Department of Medical Sciences
School of Veterinary Medicine
University of Wisconsin-Madison
2015 Linden Drive
Madison, WI 53706

ABSTRACT
The Foodborne Diseases Active Surveillance Network (FoodNet) is currently using a negative binomial regression model to estimate temporal changes in the incidence of *Campylobacter* infection. FoodNet active surveillance in 483 counties collected data on 40212 *Campylobacter* cases between years 2004 and 2011. We explored models that disaggregated these data to allow us to account for demographic, geographic, and seasonal factors when examining changes in incidence of *Campylobacter* infection. We hypothesized that modeling structural zeros and including demographic variables would increase the fit of FoodNet's *Campylobacter* incidence regression models. Five different models were compared: negative binomial without demographic covariates, negative binomial with demographic covariates, hurdle negative binomial with covariates in the count component only, hurdle negative binomial with covariates in both zero and count components, and zero-inflated negative binomial with covariates in the count component only. Of the models evaluated, the non-zero-augmented negative binomial model with demographic variables provided the best fit. Results suggests that even though zero inflation was not present at this level, individualizing the level of aggregation and using different model structures and predictors per site might be required to correctly distinguish between structural and observational zeros and to account for risk factors that vary geographically.

INTRODUCTION
The Foodborne Diseases Active Surveillance Network (FoodNet) is a collaboration among the Centers for Disease Control and Prevention (CDC), 10 state health departments, the U.S. Department of Agriculture's Food Safety and Inspection Service (USDA-FSIS), and the Food and Drug Administration (FDA). FoodNet conducts active, population-based surveillance for laboratory-confirmed infections of nine bacterial and parasitic pathogens transmitted commonly through food. The FoodNet surveillance area includes the full states of Connecticut, Georgia, Maryland, Minnesota, New Mexico, Oregon, and Tennessee, and selected counties in California, Colorado, and New York. One aim of FoodNet is to track changes over time in the incidence of 9 enteric pathogens commonly transmitted through food. FoodNet is currently using a negative binomial regression model to estimate temporal changes (Henao *et al.*, 2010).

The FoodNet model is used on data aggregated by year and FoodNet site to account for the growth of the surveillance area from 5 sites in 1996 to 10 sites in 2004, and adjust for site to site variation in incidence. This level of aggregation limits one's ability to explore variations in incidence for smaller geographic areas or units of time, or demographic features of individual cases, such as patients' age and sex; all factors that have been shown to describe unique characteristics of *Campylobacter* epidemiology (Ailes *et al.*, 2008; Samuel *et al.*, 2004). Exploration of changes in incidence over time associated with specific subgroups may contribute to hypotheses regarding geographically- or time-varying sources of *Campylobacter* infection. However, disaggregating data can cause an increase in the proportion of case counts in each subgroup that are zero, because the total population in each group is decreased.

Zero-augmented models consist of two separate model components: one for modeling case counts (using a negative binomial distribution) and one for modeling the proportion of zeros (using a binomial distribution). The zero-inflated and hurdle models differ in whether their count model component can yield a count of zero. Zero-inflated models assume zeros can be either structural or true observational zeros and therefore zeros are estimated by both binary and count components

and have an additional mixing parameter not present in hurdle models. Hurdle models assume that all zeros are structural zeros and therefore only model the binary component and use conditionally specified versions of the negative binomial distribution which are truncated to begin at a count of one (Mullahy, 1986; Desjardins, 2013).

Consequently, zero-augmented models, hurdle and zero-inflated, may be useful to model *Campylobacter* case counts in FoodNet where the high proportion of observed zero counts may be attributed to factors that make it impossible to observe a case (structural zeros) as well as factors associated with the sampling (observational zeros) (Ridout *et al.*, 1998; Hu *et al.*, 2011). We hypothesized that factors such as diagnostic testing performance or population immunity may contribute to the presence of structural zeros, and that the size of the surveillance population contributes to observational zeros.

We examined zero-augmented modifications (zero-inflated, hurdle) of the regression model used by FoodNet to estimate changes over time and added predictors to account for additional sources of variation in incidence. We hypothesized that modeling structural zeros and including demographic variables would increase the fit of FoodNet's *Campylobacter* incidence regression models. The objectives were to explore modeling incidence at a finer geographic level, evaluate the effect of covariates that vary geographically, and examine the characteristics of zero counts in *Campylobacter* surveillance data.

MATERIALS AND METHODS

Dataset preparation

Data were available for 48088 cases of *Campylobacter* infection ascertained between 2004 and 2011 in the FoodNet surveillance system. The county, state, month, and year in which the *Campylobacter* cases were diagnosed and the age and sex of the patient were used for the analysis. Sixty-six cases with missing age or sex information were excluded.

Case-patients were classified by age group [Age_Group: less than 5 (1), 5-17 (2), 18-24 (3), 25-44 (4), 45-64 (5), and 65+ (6) years of age] using categories used in previous FoodNet publications and that represent different life stages: preschool age, school age, college age, younger working age, older working age, retirement age (Ailes *et al.*, 2008). Month of diagnosis was used to make a season variable (Season) which grouped the months into high (High) and low (Low) seasons with each season including 6 consecutive months with the highest or lowest case counts, respectively. The high season included May to October and the low season included November to April. The patients' sex remained a binary variable (Sex) with two levels: Male and Female.

Campylobacter cases were grouped into one of 24 possible subgroups per county and year arising from the total combinations of 6 age groups, 2 seasons, and 2 sex categories (6*2*2). Eight years of surveillance for each of 486 counties with 24 subgroups each generated 93312 subgroups (8*486*24). Population estimates by year, state, county, age, sex, and race were provided under a collaborative arrangement with the U. S. Census Bureau (US Census Bureau, 2011). The population data were used to calculate county level incidence by dividing the number of cases by the total population of each subgroup per county.

The distribution and basic statistics of case counts and incidence were examined for all subgroups. The annual observed incidences per county were divided into 4 quartiles. The quartiles were used to construct choropleth maps where counties were shaded by incidence quartile using qGIS version 1.8.0 (QGIS Development Team, 2013). Because California counties were the only surveillance area in FoodNet without any subgroup case counts of zero, all data from these 3 counties (information on 7810 case-patients) were removed from model analyses. The final dataset had 40212 observations and 92736 case count subgroups.

Model building and comparison

The data were evaluated for overdispersion by comparing the overall mean and variance of case counts for each subgroup (McCullagh and Nelder, 1989). Models of *Campylobacter* case counts in each subgroup were built using R version 3.1.2 and its MASS, stats and pscl libraries (R Core Team, 2013). A negative binomial distribution was assumed for the outcome variable in all the models. A histogram of case counts with a negative binomial fitted curve overlay was produced. The reference groups selected for State, Age_Group, and Sex were those that represented the largest proportion of the population: Georgia, 25-44 years old, and Male, respectively. For Year and Season, the earliest year (2004), and the low season (November to April) were used as reference groups.

The first model was a negative binomial (NB) that included Year and State as nominal categorical predictors. Season, Age_Group, and Sex were added as categorical predictors to produce the next model (NB.Plus). To focus on the mixture difference between the zero-inflated (ZINB) and hurdle models (Hurdle NB) and to facilitate comparison, the models were built without variables included in the models' component which models the proportion of zeros. This was followed by fitting a zero-inflated negative binomial and hurdle model using forward selection. Forward selection was used rather than backwards elimination since the saturated models did not converge or were overfit. Variables were added individually in both model components separately and any significant variables were used in the final combination model (ZINB Full, Hurdle NB Full) (Rao and Sumathi, 2011). Each model (NB, NB.Plus, Hurdle NB, Hurdle NB Full, ZINB, ZINB Full) was offset with the natural log of the population total in each subgroup (Gelman and Hill, 2006). To determine significance of covariates, all models used an error level, alpha, of 0.05.

The zero-augmented and non-zero-augmented models were estimated by a maximum likelihood algorithm. The *Akaike information criterion* (AIC), *Bayesian information criterion* (*BIC*), and -2 log-likelihood were computed for comparison. The BIC-corrected Vuong test was used to compare the fit of non-nested models and the likelihood ratio test was used to compare the fit of nested models (Vuong, 1989). The zero component intercepts in the zero-augmented models were evaluated as a large negative coefficient value does not support the idea of zero inflation in the data (Schwadel and Falci, 2012; Erdman *et al.*, 2008).

Model assessment was done by evaluating the mean absolute error using leave-one-out-cross-validation (Kuhn and Johnson, 2013). The difference between the predicted and observed zero case counts were compared for all models. Quantile-Quantile (Q-Q) plot and residual histogram for the best fitting model were inspected for normally distributed errors. The source code of all analysis steps are available by request.

RESULTS

Descriptive Statistics

On average 5027 (±SD 300) cases of *Campylobacter* infection were reported to FoodNet each year between 2004 and 2011 (Range: 4751 in 2004 to 5636 in 2011). The majority (63.0% ± 0.9%) were reported during the high season (May to October). The average annual incidence (all reported per 100000 persons) for all sites combined was 11.8 (± 0.5) and ranged between 11.3 in 2008 and 12.8 in 2011. The average state incidence was 13.4 (± 5.0) and varied from state to state (Range: 6.8 in Georgia to 19.5 in California). The average age group incidence was 14.2 (± 5.7) and was highest for children aged less than 5 years (25.4) and lowest among persons aged 5 to 17 years (9.0). Males had higher rates than females (14.6 vs. 11.5).

To provide a visual representation of geographic variation in incidence among counties, quartiles of annual county level incidence were mapped for Minnesota, Georgia, New Mexico, and Oregon as examples (Figure 1). The average annual incidence per county was 12.8 (± 10.0) per 100000. The wide standard deviation was a function of county incidence variation among and within states illustrated in Figure 1.

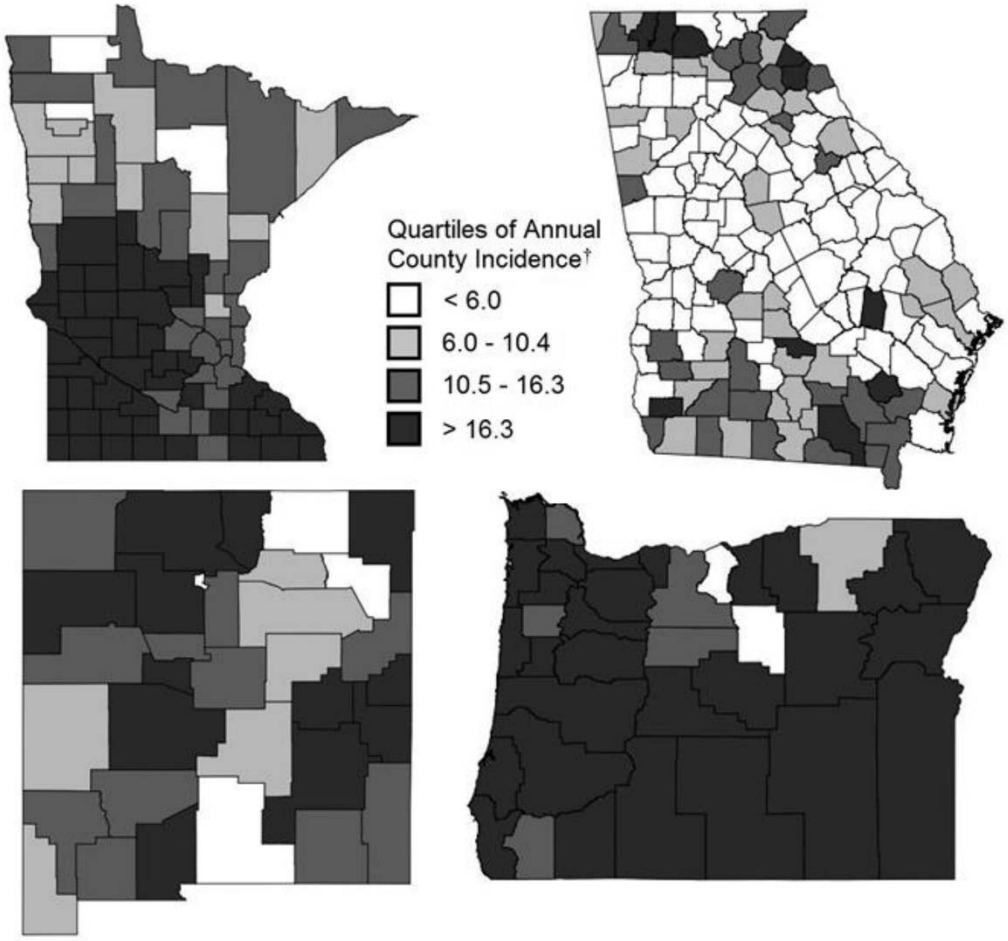

Figure1: Observed county incidence per 100000 in A) Minnesota, B) Georgia, C) New Mexico and D) Oregon in 2011. Counties are shaded based on the quartiles of county annual incidence per 100000.

Building Models

Variance (1.71) and mean (0.43) of all the *Campylobacter* counts in the final dataset were calculated. The large variance relative to the mean, suggested that the data were overdispersed (Rao and Sumathi, 2011). This was further supported by the negative binomial's estimated overdispersion parameter [log(theta)] which was significantly different from zero with a p-value less than 0.001 (Cameron and Trivedi, 2013). The histogram in Figure 2 shows the case count frequency with a normal negative binomial curve overlay (number of observations= 92736, mean = 0.434, theta = 0.213). Out of the 92736 total subgroups, 78.6% had a zero case count. The curve mirrors the observed values closely and zero inflation is not apparent.

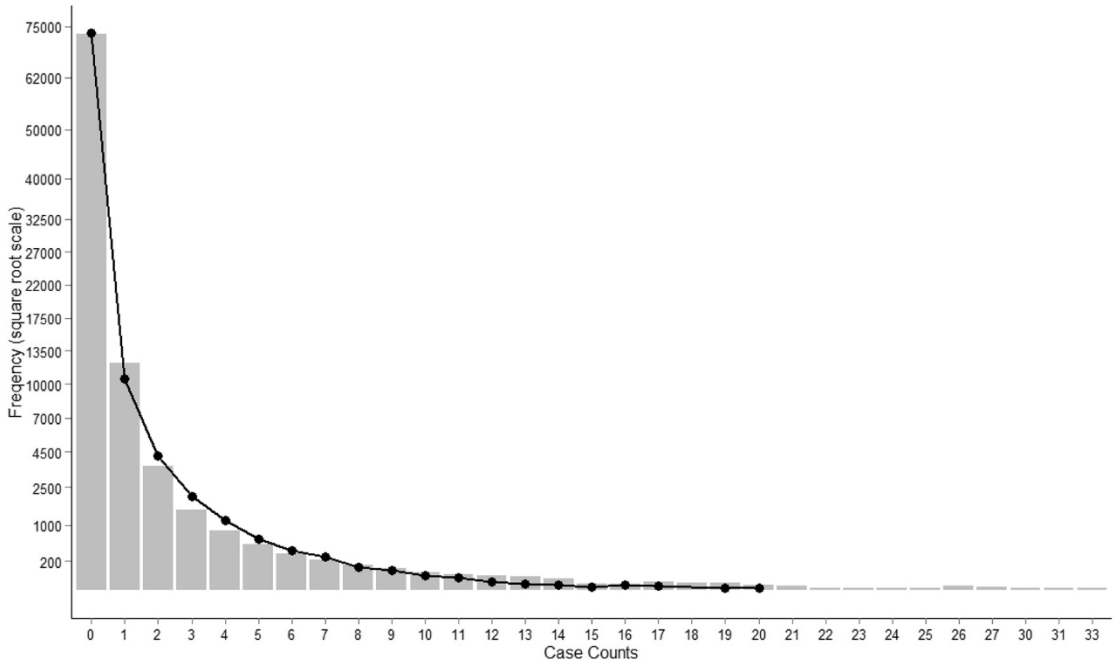

Figure 2: Count frequency of *Campylobacter* cases in FoodNet (bars) with normal negative binomial curve overlay (number of observations= 92736, mean = 0.434, theta = 0.213). Y axis is shown using a square root scale.

Model Results

All variables included in the non-zero-augmented models (NB, NB.Plus), both count and zero portions of the Hurdle models, and the count portion of the ZINB model were statistically significant predictors in the models. The ZINB Full was not included in the model comparison because none of the variables added by forward step selection were significant in the binary portion of the model. The individual model results are shown in Appendix.

The count components of all models (NB, NB.Plus, Hurdle NB, Hurdle NB Full, ZINB) had similar results in terms of coefficient direction, magnitude, and significance. However, Tennessee, year 2010, and age group 65+ were significant in the NB, NB.Plus and the ZINB models but not in the count components of the Hurdle NB and Hurdle NB Full models. The other difference was that the age group that includes 45-64 year olds was significant in the count component of the Hurdle NB and Hurdle NB Full models but not in the count component of the ZINB model.

Table 1: Goodness of fit and statistics comparison by model

		Model 1*				
Model 2		Hurdle NB	NB	Hurdle NB Full	ZINB	NB.Plus
M0†	LR		***			***
	V (BIC)	79.0, ***	79.9, ***	91.5, ***	89.5, ***	89.5, ***
Hurdle NB	LR			***		
	V (BIC)		9.2, ***	37.6, ***	40.4, ***	40.5, ***
NB	LR					***
	V (BIC)	(-9.2), ***		29.6, ***	33.4, ***	33.5, ***
Hurdle NB full	LR	***				
	V (BIC)	(-37.6), ***	(-29.6), ***		3.7, 0.0001	3.9, 5.1e-5
ZINB	LR					
	V (BIC)	(-40.4), ***	(-33.4), ***	(-3.7), 0.0001		174.2, ***
NB.Plus	LR		***			
	V (BIC)	(-40.5), ***	(-33.5), ***	(-3.9), 5.1e-5	(-174.2), ***	
-2 x log likelihood		-115539	-114403	-109482	-109525	-109525
‡		25	17	47	25	24
AIC		115589	114437	109576	109575	109573
BIC		115825	114597	110019	109811	109799
MAE		0.3963	0.4046	0.3809	0.3798	0.3798
Predicted no. zeros		72918	73540	72918	73403	73403

Models are listed from left to right and top to bottom as their fits improve; * Hurdle NB = Hurdle negative binomial with covariates in the count component only, NB = Negative binomial without demographic covariates, Hurdle NB Full = hurdle negative binomial with covariates in both zero and count components, ZINB = Zero-inflated negative binomial with covariates in the count component only, NB.Plus = Negative binomial with demographic covariates; † Null model; LR= Likelihood ratio test; V (BIC) = Vuong BIC corrected Non-Nested Hypothesis Test-Statistic; *** = p-value less than 2.2e-16 when testing model 1 versus model2 with alpha < 0.05; ‡Number of parameters estimated; AIC = *Akaike information criterion*; BIC = *Bayesian information criterion*; MAE = Mean absolute error

Model Assessment and Comparison

The zero component intercepts in the zero-augmented models all had large negative coefficient values which do not support the idea of zero inflation in the data. This is further supported by the goodness of fit evaluations summarized in Table 1. The likelihood ratio test led to the same results as the Vuong test when applied to nested models. Using the goodness of fit measures the NB.Plus model had the best fit. The ZINB and NB.Plus had the same log likelihood but different degrees of freedom. The Hurdle-NB model had the worst fit and the Hurdle NB Full had lower fit than both the ZINB and NB.Plus models. The residual histogram with a normal curve overlay is shown in Figure 3 for the NB.Plus model and displays deviation from homoscedasticity and normality.

Adding the demographic variables to the non-augmented models decreased the mean absolute error by 0.0249 (decreased the error). For the zero-augmented NB.Plus model the addition increased the mean absolute error by 2.726e-6 for the ZINB and by 0.0165 for the Hurdle NB model (increased the error). There were 72918 zero case counts in the dataset and the hurdle models predicted the exact number. When we rounded the predicted number of zeros to the nearest integer, both the ZINB and NB.Plus models predicted 73403 zeros or 485 more than the observed number of zero counts. The hurdle models were superior at predicting zero counts because of their truncated structure.

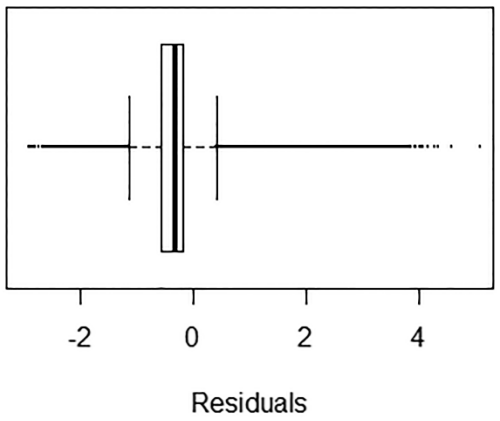

Figure 3: Residual boxplot of negative binomial model with demographic covariates (NB.Plus)

DISCUSSION

The aim of this analysis was to explore different methods to analyze campylobacteriosis case counts ascertained by FoodNet surveillance sites at a finer geographic level, to evaluate the effect on incidence of covariates that may vary geographically, and examine the characteristics of zero counts in FoodNet *Campylobacter* data. The subgroups selected for analysis represented demographic and geographic variables known to influence incidence of *Campylobacter* infections (Ailes *et al.*, 2008). Although a disproportionate number of observations were zero, zero inflation was not apparent, and the negative binomial model with inclusion of demographic and seasonal variables significantly increased the fit of the model (see Table 1, NB.Plus) compared with the model with only year and state included (NB in Table 1). Our findings suggest that the incidence of *Campylobacter* infection varies substantially among the FoodNet counties, making it worthwhile to explore differences in surveillance populations, exposures, laboratory practices, or other factors that differ among sites.

Zero-augmented modifications (zero-inflated, hurdle) of the regression models were used to examine a possible separation of observational and structural zeros. We anticipated that a significant proportion of zero case counts were observational; differences in county size and population demographics among the FoodNet surveillance sites result in very small subpopulation sizes among counties and a high probability that no cases will be observed among many counties. Our finding that the hurdle models did not fit the data well supports this assumption. Although we hypothesized that several surveillance and epidemiologic factors may contribute to structural zeros in the data, our analysis suggests that zero inflation is not apparent at the level of disaggregation of demographic covariates we studied; this finding is supported by the observation that inclusion of zero-augmentation mixing fractions did not improve the models' fit.

Although zero inflation was not present in the dataset, zero-augmented modeling techniques are likely to be important for future analyses including modeling of other pathogens under FoodNet surveillance. Our models included only data ascertained by FoodNet active surveillance activities, and it is likely that inclusion of data from sites conducting passive surveillance, as well as data obtained from other sources, such as household income and access to healthcare, would contribute to the presence of structural zeros in the modeled data. The differences in data collection associated

with different surveillance systems and data sources would likely result in excess zero case counts where at least a portion (structural zeros) arise from a process different from the positive counts. Although both hurdle and zero-inflated models may be used to model this type of data, it is likely best modeled by a zero-inflated model because the zeros are modeled as a mixture of both observational and structural zeros.

We removed the California observations because there were no zero case counts in any county subgroup, complicating our exploration of models for zero case counts. Removal of the California data eliminated convergence issues and allowed exploration of the effect of zero inflation. Removing the California data decreased the dataset's variance but overdispersion was still prominent. A negative binomial distribution helped in modeling the overdispersed data; however, there were still case counts that were outside the expected distribution. These case counts may be associated with undetected outbreaks (i.e., clusters of cases originating from a common exposure) which were not excluded from the analysis. Further exploration of these outliers, using compound distributions, would help better characterize them and might yield more information on risk factors of potential outbreaks (Hinde, 1982).

CONCLUSIONS

The addition of the demographic and seasonal variables when modeling *Campylobacter* counts accounted for more variability and resulted in improved goodness of fit compared with models that only included a state factor. However, the complexity and variation in the epidemiology of *Campylobacter* was still not fully addressed, suggesting that differences in surveillance populations among the FoodNet sites or other epidemiological factors vary geographically. For example, the models did not fully account for the incidence variation among counties and states as illustrated in Figure 1. County-level variation associated with differences in county geographic size, population and other unmeasured factors could result in additional sources of structural zeros in case counts. Although we investigated structural zeros at the state level, the possibility for structural zeros to vary by county was not examined. Potentially, the level of aggregation and the count distribution could be adjusted per site to better fit the data and further explore structural zeros. Therefore, future steps should focus on individual sites.

REFERENCES

Ailes E, Demma L, Hurd S, Hatch J, Jones TF, Vugia D, Cronquist A, Tobin-D'Angelo M, Larson K, Laine E, Edge K. Continued decline in the incidence of *Campylobacter* infections, FoodNet 1996-2006. Foodborne Pathog Dis 2008; 5:329-337.

Cameron AC, Trivedi PK. *Regression analysis of count data,* 2nd edition. New York, NY: Cambridge University Press, 2013.

Desjardins CD. Evaluating the performance of two competing models of school suspension under simulation-the zero-inflated negative binomial and the negative binomial hurdle. Diss. University of Minnesota, 2013.

Erdman D, Jackson L, Sinko A. Zero-inflated Poisson and zero-inflated negative binomial models using the COUNTREG procedure. In: Proceedings of the SAS Global Forum 2008 Conference, San Antonio, Texas, 2008, pp. 322-2008. Cary, NC: SAS Institute Inc., 2008.

Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press, 2006.

Henao OL, Scallan E, Mahon B, Hoekstra RM. Methods for monitoring trends in the incidence of foodborne diseases: Foodborne Diseases Active Surveillance Network 1996–2008. Foodborne Pathog Dis 2010; 7:1421-1426.

Hinde J. Compound Poisson regression models. In: *GLIM 82: Proceedings of the International Conference on Generalised Linear Models*. Gilchrist R (ed.). New York: Springer, 1982.

Hu MC, Pavlicova M, Nunes EV. Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. Am J Drug Alcohol Abuse 2011; 37:367-375.

Kuhn M, Johnson K. *Applied predictive modeling*. New York: Springer, 2013.

McCullagh P, Nelder JA. *Generalized linear models,* 2nd edition. Boca Raton, FL: Chapman & Hall/CRC press, 1989.

Mullahy J. Specification and testing of some modified count data models. J econometrics 1986; 33:341-365.

QGIS Development Team. QGIS Geographic Information System. Open Source Geospatial Foundation Project; 2013. Available from: http://qgis.osgeo.org (accessed 14 June, 2015)

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. Available from: http://www.R-project.org/ (accessed 14 June, 2015)

Rao A, Sumathi K. Selection of Variables in Regression Models Based on Inflated Distributions. Pakistan J Stat Oper Res 2011; 7:381-390.

Ridout M, Demetrio CGB, Hinde J. Models for count data with many zeros. In: Proceedings of the XIXth International Biometric Conference, Cape Town, 1998, pp. 179–192. Cape Town, South Africa: International Biometric Society, 1998.

Samuel MC, Vugia DJ, Shallow S, Marcus R, Segler S, McGivern T, Kassenborg H, Reilly K, Kennedy M, Angulo F, Tauxe RV. Epidemiology of sporadic *Campylobacter* infection in the United States and declining trend in incidence, FoodNet 1996–1999. Clin Infect Dis 2004; 38(Supplement_3):S165-S174.

Schwadel P, Falci CD. Interactive effects of church attendance and religious tradition on depressive symptoms and positive affect. Soc Ment Health 2012; 2:21-34.

U.S. Census Bureau. Population Estimates. Intercensal and postcensal estimates by year, state, county, age, sex, and race, prepared under a collaborative arrangement with the US Census Bureau (2004–2011). Washington, DC: U.S. Census Bureau, 2011.

Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 1989; 57:307-333.

APPENDIX: Model Results

Negative binomial without demographic covariates (NB)

Parameter	Estimate	SE†	95% Confidence Interval		P value‡	
(Intercept)	-9.600	0.023	-9.645	-9.555	< 2.00e-16	***
Year 2005 (ref: 2004)	-0.015	0.026	-0.066	0.036	5.59e-1	
Year 2006 (ref: 2004)	0.001	0.026	-0.050	0.052	9.62e-1	
Year 2007 (ref: 2004)	0.010	0.025	-0.039	0.059	6.88e-1	
Year 2008 (ref: 2004)	-0.002	0.026	-0.053	0.049	9.33e-1	
Year 2009 (ref: 2004)	0.040	0.025	-0.009	0.089	1.17e-1	
Year 2010 (ref: 2004)	0.064	0.025	0.015	0.113	1.10e-2	*
Year 2011 (ref: 2004)	0.107	0.025	0.058	0.156	1.75e-5	***
State CO (ref: GA)	0.868	0.029	0.811	0.925	< 2.00e-16	***
State CT (ref: GA)	0.795	0.028	0.740	0.850	< 2.00e-16	***
State MD (ref: GA)	0.115	0.027	0.062	0.168	1.44e-5	***
State MN (ref: GA)	0.981	0.021	0.940	1.022	< 2.00e-16	***
State NM (ref: GA)	1.005	0.027	0.952	1.058	< 2.00e-16	***
State NY (ref: GA)	0.664	0.024	0.617	0.711	< 2.00e-16	***
State OR (ref: GA)	1.023	0.024	0.976	1.070	< 2.00e-16	***
State TN (ref: GA)	0.064	0.024	0.017	0.111	9.00e-3	**

† Standard Error, ‡ Significant codes: '***' < 0.001, '**' < 0.01, '*' < 0.05, '.' < 0.1

CHAPTER 2.2

Shrinking a large dataset to identify variables associated with increased risk of *Plasmodium falciparum* infection in Western Kenya.

Epidemiol. Infect. (2015), 143, 3538–3545.
doi:10.1017/S0950268815000710

Tremblay, M.,[1] Dahm, J.S.,[1] Wamae, C.N.,[2,3] De Glanville, W.A.,[4,5] Fèvre, E.M.[5,6] Döpfer, D.[1]

[1] Departments of Medicine and Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, WI, USA
[2] Center for Microbiology Research, Kenya Medical Research Institute (KEMRI), Nairobi, Kenya
[3] School of Health Sciences, Mount Kenya University, Thika, Kenya
[4] Centre for Immunity, Infection and Evolution, Institute for Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, Edinburgh, UK
[5] International Livestock Research Institute, Nairobi, Kenya
[6] Institute of Infection and Global Health, University of Liverpool, Leahurst Campus, Neston, UK

ABSTRACT
Large datasets are often not amenable to analysis using traditional single-step approaches. Here, our general objective was to apply imputation techniques, principal component analysis (PCA), elastic net and generalized linear models to a large dataset in a systematic approach to extract the most meaningful predictors for a health outcome. We extracted predictors for Plasmodium falciparum infection, from a large covariate dataset while facing limited numbers of observations, using data from the People, Animals, and their Zoonoses (PAZ) project to demonstrate these techniques: data collected from 415 homesteads in western Kenya, contained over 1500 variables that describe the health, environment, and social factors of the humans, livestock, and the homesteads in which they reside. The wide, sparse dataset was simplified to 42 predictors of P. falciparum malaria infection and wealth rankings were produced for all homesteads. The 42 predictors make biological sense and are supported by previous studies. This systematic data-mining approach we used would make many large datasets more manageable and informative for decision-making processes and health policy prioritization.

INTRODUCTION
With the increasing production and availability of large amounts of data, it is common to have datasets that cannot be analysed using traditional single-step approaches. For example, it is not advisable to build simple regression models from datasets that have thousands of variables or those that have incomplete data. Many different data-mining and statistical techniques are commonly employed individually to address these issues, but a systematic approach has not been developed to take advantage of multiple methods' strengths and capacities. Our general objective is to apply imputation techniques, principal component analysis (PCA), elastic net and generalized linear models (GLM) in a systematic approach to extract the most meaningful predictors for a health outcome from a large covariate dataset while facing limited numbers of observations. The People, Animals, and their Zoonoses (PAZ) dataset will be used to demonstrate these techniques [1]. The PAZ project's goal is to explore the epidemiology and burden of a number of neglected zoonotic diseases in a sympatric population of animals and people. Currently, PAZ's only study site is in Western Kenya. The dataset contained variables that describe the health, environment, and social factors of the humans, livestock, and homesteads in which they reside. The specific aim of applying this protocol to the PAZ dataset is to develop and apply socioeconomic wealth indices and determine the best predictors of falciparum malaria infection exposure prevalence in individuals included in the PAZ dataset [2]. We hypothesize that these techniques can be used to develop a simplified dataset with the most meaningful predictors from a wide, sparse dataset. If successful, this systematic data-mining approach could make many large datasets more manageable and informative.

MATERIALS AND METHODS
Making a complete dataset
The dataset used in this study which originates from the PAZ project consist of questionnaire data from 416 rural homesteads and biological sampling data of 2113 humans and 983 cattle from these homesteads in the western Province of Kenya [1]. Homesteads determined to be outliers due to an extreme cattle–human ratio were excluded from the analysis.

All data analyses were performed using R version 3·0·1 [3]. A case of malaria was defined as a subject being positive for Plasmodium falciparum on thick or thin blood smears [4]. The homestead malaria prevalence was defined by:

Homestead Malaria Prevalence = (Number of Positive Malaria Cases)/(Total Human Subjects in a homestead)

To prepare the dataset for statistical analysis, all categorical variables were expanded into binary dummy variables and edited until missing values were all coded as 'NA'. The number of missing values was first calculated per dataset and frequency tables were used to examine the percent missingness per variable. Variables with >10% of values missing were removed from the dataset. This was important, because the deleted variables could not be determined to be 'missing at random (MAR)' due to the non-random approach to the data collection, and therefore keeping those variable in the dataset would have conflicted with the MAR prerequisite of multiple imputations [5].

After this new dataset was generated and further missingness was assumed to be at random, the remaining variables were subjected to piecewise multiple imputations by chained equations using the R package 'mice' [6, 7]. This package was selected due to its ability to handle both factor and continuous variables. After completing the imputation by 'mice', variables with missing values that could not be imputed were omitted from further statistical testing.

Frequency tables were created for all variables and data were analysed for uniformity. Variables where the most frequent value accounted for $\geqslant$99% of the observations were removed to avoid variables without contrasts in the dataset. A range of such cut-off percentages for uniformity was evaluated and the 99% cut-off resulted in the most consistent removal of variables without contrasts across the dataset.

Variables denoting the number of individuals per homestead for cattle and humans were created to serve as denominators for calculating prevalences. For each numeric variable in the human and cattle dataset, the mean value across each homestead was calculated to subsequently allow the dataset to be merged by homestead number.

Ethical considerations

Human data and samples collected in this study were collected following approval by the KEMRI Ethical Review Committee, SC#1701. Animal samples were collected following approval from the Roslin Institute Animal Welfare and Ethical Review Committee, AWA004. The Institutional Review Board (IRB) approved this study (IRB no. 2013-0072).

Creating wealth indices using PCA

Because wealth is often a predictor of disease prevalence, selected asset and livestock variables descriptive of wealth or socioeconomic status were shrunk into one wealth ranking value per homestead [2]. Historically, asset-based wealth indices have been based on household assets, but because wealth in rural areas is often dependent upon livestock ownership and the ability to call on human assistance, compared to urban areas, in which wealth is often expressed in material

possessions, two separate wealth rankings were created: one based on material assets (asset-based wealth ranking) and one based on a homestead's livestock (livestock-based wealth ranking) [2].

Both wealth indices were created using PCA, an ordination method commonly applied during wealth-indexing studies [8]. PCA converts a number of non-correlated variables into a number of orthogonal principal components (PCs) [9]. The first PC is the ordination of the variables that explains the most amount of variance, and each subsequent PC thereafter explains a decreasing amount of the variance. The starting subset of variables for each wealth index was selected from a previous study by Okell et al. that utilized a preliminary version of the same dataset with fewer homesteads [8]. All variables were formatted as numeric, and their respective minima were added to each variable set to assure non-negative values. The variables were scaled using the 'scale()' command in order to assure non-negative values in the dataset used for PCA, i.e. the overall minimum value of any observation was added to all values in the dataset.

Because highly correlated variables can skew a PCA analysis, a Pearson correlation matrix was used on both the asset-based and livestock-based variables to determine whether any two variables were highly correlated, in which case the biologically less relevant variable was removed. A correlation $\geqslant$90% was used as our limit [10]. The PCA was run on both the asset-based and livestock-based variables separately [11]. Based on the first six PCs of each of the two PCAs, it was determined which subset of variables contributed more than expected to the explanation of the overall variance in the respective datasets. The PCAs were repeated for the selected subset of covariates. The respective first PCs of the outcomes were taken as the livestock-based and asset-based wealth indices.

To explore the validity of the livestock wealth index, a third wealth index was created based on real-world valuation of livestock holdings. Current market value for each category of the livestock evaluated was based on interviews with market traders in the study region and subsequently multiplied by the number of livestock in the respective livestock categories of the dataset [8]. The summation of these values yielded the total livestock value (TLV) for each homestead, which was used as a real-world approximation estimate for livestock wealth [8]:

TLV = Σ (number of animals in a category * current market value of animal)

These wealth indices were merged with the final dataset by homestead. Since only 54% of the homesteads had cattle, the final dataset including the wealth indices was divided into two datasets for further analysis. Subset A was created from the homestead, human, and cattle variables containing only the 224 homesteads with cattle. Subset B was created using the homestead and human variables of all 415 homesteads only.

Selecting predictors with elastic net and GLM

Regularized regression models are a commonly accepted method for selecting predictors from large data. The elastic net was created by combining the penalties of the lasso and ridge regularized regression methods. This combination allows for better performance when the number of variables (p) is greater than the observation count (n) and when groups of variables exist that are highly correlated while still resulting in a parsimonious model [12]. The number of variables selected is

controlled by the alpha (α) parameter. The regression will more closely resemble a lasso regression or a ridge regression as α nears/approaches 1 or 0, respectively [12].

The glmnet package in R was used to fit the elastic-net regularization path for Poisson regression on homestead malaria prevalence for subsets A and B [13]. The model response was the count of malaria-positive cases in each homestead and an offset of the log of the total humans per homestead was used to model prevalence. A Poisson family was chosen since the response was a count. The cross-validation function (cv.glmnet) was used to find the best value of lambda (λ), the regularization parameter, and the number of folds was selected to be the number of observations (n) minus 1 (leave-one-out cross-validation). To select the best value of α, 50 iterations of 17 different α values between 0 and 1 were run and summarized. The α value that resulted in the lowest mean absolute error (MAE) was selected. The selected λ and α values were subsequently used for elastic-net variable selection using the glmnet function.

The variables selected by the elastic-net regularized penalized regression using non-zero coefficients were subsetted and included in a GLM using the glm package in R. Further variable selection was performed in a stepwise function based on Akaike's Information Criterion (AIC) using the step function. Both forward and backward directions were allowed [2]. To determine significance of covariates an error level, $\alpha = 0{\cdot}05$ was set. A model with only significant variables was desired so further backwards elimination was performed based on P value.

RESULTS

Making a complete dataset

Homestead 84 was considered an outlier due to a very high cattle–human ratio; therefore, all observations from homestead 84 (17 human subjects, 41 cattle) were excluded from the analysis. Eleven cattle and one human subject were removed because they did not have a homestead number recorded, 415 homesteads, 2095 humans and 931 cattle remained.

In the homestead dataset 2·81% (4753/168 905) of values were missing and there were 24/407 variables with >10% missingness. In the cattle dataset 16·95% (48 750/287 679) of values were missing and there were 78/309 variables with >10% missingness. In the human dataset 8·09% (111 810/1 382 700) of values were missing and there were 105/660 variables with >10% missingness. After the variables with >10% missing values were removed, 1169 variables remained. The number of variables left and removed per dataset is described in Table 1.

Table 1. Number of variables per dataset at each step

	Homestead	Human	Livestock
1. Starting number of variables	407	660	309
2. Number of variables removed due to >10% missingness	-24	-105	-78
3. Number of variables removed due to incomplete imputation	-18	-16	-2
4. Number of variables removed due to >99% uniform	-93	-188	-97
5. Final number of variables	272	351	132

There were 677 values still missing in the cattle dataset (0·32%, 677/215 061), 14 742 values still missing in the human dataset (1·27%, 14 742/1 164 820) and 1296 values still missing in the homestead dataset (0·82%, 1295/158 945) after removing variables with >10% missingness. The imputation of these missing values was unsuccessful for 36 variables which were removed from the analysis. On average the 36 variables were >99·9% (s.d. ± 0·32) uniform which explains the incomplete imputation.

The average percent uniformity for the remaining 1133 variables was 89·9%. The 278 variables with >99% uniformity were removed. The final variable count in each dataset is shown in Table 2. The total count of malaria-positive subjects was 621. The average count of malaria-positive cases per homestead was 1·50 cases and ranged from 0 to 8 with with >50% having zero positive cases. The average number of human subjects per homestead was 5·05 (s.d. ± 2·94) with a maximum of 21 people. Malaria prevalence per homestead averaged at 28·25% (s.d. ± 27·35) and the overall prevalence was 29·64% (621/2095) for the entire study.

Creating wealth indices with PCA

One variable in the asset data, 'number of mud walls', was found to correlate too highly with two other asset variables, 'number of dwellings' and 'number of earth floors', and was therefore omitted from the wealth-indexing PCA. The first six PCs were used to find the subsets of variables that explained more than average amount of variance in the data. The 11 and 30 variables selected for the livestock and asset subsets, respectively, are listed in Tables 2 and 3. The first PC generated using each subset of variables was used to create the wealth indices. The TLV and the livestock wealth index were determined to be collinear and therefore provided some evidence of its validity.

Table 2. List of asset wealth variables by variable type

Count (1-10)	**Count (11-20)**	**Binary**
Dwellings	Cooking fuel - Firewood	Radio
Iron roofs	Cooking fuel - Charcoal	Television
Thatch roofs	Cooking fuel - Gas stove	Cupboard
Unburnt brick walls	Cooking fuel - Paraffin stove	Sofa with cushions
Mud brick walls	Latrine on compound	Clock
Cement brick walls	Completely closed latrine	Wrist watch
Mud/cement walls	Partially closed latrine	Sewing Machine
Earth floors	Open pit latrine	Torch (flashlight)
Cement floors	Mobile phone charger	Bicycle
Electric solar	Mobile phone	Motorbike

Table 3. List of livestock wealth variables by variable type

Count	Binary
Weaned female calves	Chickens
Adult castrated male cattle	Ducks
Adult entire male cattle	
Adult female cattle	
Suckling pigs	
Weaned male pigs	
Weaned female pigs	
Sows	
Boars	
Chickens	

Selecting predictors with elastic net-regularized penalized regression and GLM

After a total of 50 iterations of cross-validation for each α level, the α values with the lowest MAE for subsets A and B were 0·05 and 0·2, respectively. The corresponding λ values used in the elastic-net modelling are listed in Table 4. There were 143 variables selected out of 757 from subset A and 105 out of 626 variables from subset B. The AICs of the starting GLMs with the subset of these non-zero coefficient variables are listed in Table 4. After stepwise selection of variables the models' AICs were reduced by 177 and 92 units for subsets A and B, respectively. Further backwards stepwise elimination based on P value was performed which reduced the amount of variables in the model to 22 for subset A and 25 for subset B. Five variables were found in both models. The final models' estimates are included in Tables 5 and 6.

Table 4. Cross-validation, elastic net and GLM parameters

Parameter	Subset A	Subset B
CV n-folds	223	414
Alpha (α)	0.05	0.2
Lambda	1.385	0.2464
Number of nonzero coefficients	143	105
AIC- at beginning of GLM	745	1123
AIC- after Step procedure	568	1031
AIC- after backwards elimination	578	1043

GLM, Generalized linear model.

Table 5. Subset A: Generalized linear model results*

	Estimate	Std. Error	RR [95% CI]	z value	Pr(>|z|)
(Intercept)	-0.3475	0.5563	0.7065 [0.2374 - 2.1019]	-0.62	0.5321
Keep chickens [yes vs. no]	-0.6002	0.1963	0.5487 [0.3735 - 0.8062]	-3.06	0.0022
Travel to medical facility by *Matatu* † [yes vs. no]	-0.7731	0.3183	0.4616 [0.2473 - 0.8614]	-2.43	0.0152
Last bought/acquired cattle 1 to 2 months age [yes vs. no]	-1.1209	0.4271	0.3260 [0.1411 - 0.7529]	-2.62	0.0087
Are cattle herded with goats or sheep [yes vs. no]	-0.4025	0.1337	0.6686 [0.5145 - 0.8690]	-3.01	0.0026
Control worms in cattle with drench (unknown drug) [yes vs. no]	-0.2855	0.1313	0.7516 [0.5811 - 0.9722]	-2.18	0.0296
Pigs- use a worm control product when they get thin [yes vs. no]	-1.6077	0.7212	0.2003 [0.0487 - 0.8235]	-2.23	0.0258
Number of houses with brick or cement walls	-0.7013	0.3057	0.4959 [0.2724 - 0.9029]	-2.29	0.0218
Own a bicycle for transportation [yes vs. no]	0.4330	0.1858	1.5419 [1.0713 - 2.2192]	2.33	0.0197
Number of individuals in the age group 5-9	1.4108	0.3342	4.0992 [2.1293 - 7.8918]	4.22	0.00002
Samia subgroup [yes vs. no]	0.5738	0.1889	1.7750 [1.2258 - 2.5703]	3.04	0.0024
Feeding livestock once a week [yes vs. no]	1.0625	0.2577	2.8936 [1.7462 - 4.7950]	4.12	0.00004
Used to but no longer involved with manure preparation [yes vs. no]	3.7715	1.5590	43.445 [2.0461 - 922.497]	2.42	0.0156
Human subject milks cow at least once a year [yes vs. no]	1.2721	0.6305	3.5683 [1.037 - 12.2786]	2.02	0.0436
Seek treatment for breathing problem at a hospital [yes vs. no]	-1.3600	0.5119	0.2567 [0.0941 - 0.7000]	-2.66	0.0079
Currently taking medications [yes vs. no]	-1.1713	0.4627	0.3100 [0.1252 - 0.7676]	-2.53	0.0114
Human fecal positive for Schisto-soma mansoni [yes vs. no]	-1.0352	0.4217	0.3552 [0.1554 - 0.8117]	-2.45	0.0141
Cattle fecal positive Trichuris (whipworm) [yes vs. no]	0.0874	0.0361	1.0913 [1.0168 - 1.1713]	2.42	0.0155
High-grade cattle breed, e.g. Friesian cross [yes vs. no]	-1.6162	0.7112	0.1987 [0.0493 - 0.8007]	-2.27	0.0231
Prophylactic treatment of cattle when ticks seen [yes vs. no]	0.4190	0.1559	1.5204 [1.1201 - 2.0638]	2.69	0.0072
Average cattle skin elasticity rating [yes vs. no]	-0.4189	0.1809	0.6578 [0.4614 - 0.9377]	-2.32	0.0206
Had fever but didn't seek treatment [yes vs. no]	0.6547	0.2636	1.9246 [1.1480 - 3.2263]	2.48	0.0130
Use Nambale cattle market [yes vs. no]	-0.6138	0.2423	0.5413 [0.3367 - 0.8703]	-2.53	0.0113

S.E., Standard error; RR, relative risk; CI, confidence interval.

* Number of observations = 224.

† Minibuses, station wagons, vans and pick-up trucks serve as matatus.

Table 6. Subset B: Generalized linear model results*

	Estimate	Std. Error	RR [95% CI]	z value	Pr(>|z|)
(Intercept)	0.0161	0.8778	1.0162 [0.1819 - 5.6778]	0.02	0.9854
Number of individuals in the age group 15-19	0.0849	0.0405	1.0886 [1.0055 - 1.1785]	2.09	0.0363
Keep ducks [yes vs. no]	-0.2538	0.1287	0.7758 [0.6029 - 0.9984]	-1.97	0.0487
Experienced drought in the last 6 months [yes vs. no]	0.3722	0.1151	1.4509 [1.1579 - 1.8181]	3.23	0.0012
Keep cattle to sell adult cattle [yes vs. no]	-0.2877	0.0996	0.7500 [0.6170 - 0.9117]	-2.89	0.0039
Use Nambale cattle market [yes vs. no]	-0.6991	0.2377	0.4970 [0.3119 - 0.7920]	-2.94	0.0033
Cattle's water collected from river-dry season [yes vs. no]	0.3053	0.1331	1.3570 [1.0454 - 1.7615]	2.29	0.0218
Pigs freeroam in the dry season [yes vs. no]	0.5482	0.2414	1.7301 [1.0780 - 2.7769]	2.27	0.0232
Waste is cooked prior to being fed to pigs [yes vs. no]	-0.3825	0.1595	0.6822 [0.4990 - 0.9325]	-2.40	0.0165
Number houses with cement floors	-0.2774	0.0777	0.7578 [0.6507 - 0.8824]	-3.57	0.0004
Own a bicycle for transportation [yes vs. no]	0.3894	0.1186	1.4761 [1.1699 - 1.8624]	3.28	0.0010
Altitude	-0.0015	0.0007	0.9985 [0.9971 - 0.9999]	-2.21	0.0273
Number of individuals in the age group 5-9	1.0692	0.2892	2.9130 [1.6526 - 5.1347]	3.70	0.0002
Number of individuals in the age group 10-15	1.0027	0.2760	2.7256 [1.5868 - 4.6816]	3.63	0.0003
Occupation- teacher [yes vs. no]	-4.3639	1.4921	0.0127 [0.0007 - 0.2371]	-2.92	0.0035
Occupation- fisherman [yes vs. no]	-3.7469	1.4198	0.0236 [0.0015 - 0.3813]	-2.64	0.0083
Occupation- none [yes vs. no]	1.2529	0.5319	3.5005 [1.2342 - 9.9285]	2.36	0.0185
Feeding livestock once a week [yes vs. no]	0.7506	0.2047	2.1183 [1.4182 - 3.1639]	3.67	0.0003
Pigs kept in buildings [yes vs. no]	0.8555	0.3267	2.3526 [1.2401 - 4.4630]	2.62	0.0088
Recent illness- abdominal pain [yes vs. no]	0.5050	0.2359	1.6570 [1.0436 - 2.6310]	2.14	0.0323
Recent illness- eye problems [yes vs. no]	-2.3010	0.8811	0.1002 [0.0178 - 0.5632]	-2.61	0.0090
Had fever and treated by chemist [yes vs. no]	-0.6691	0.2872	0.5122 [0.2917 - 0.8992]	-2.33	0.0198
Currently taking medications [yes vs. no]	-0.7147	0.3215	0.4893 [0.2606 - 0.9189]	-2.22	0.0262
Recent backache [yes vs. no]	-0.5276	0.2410	0.5900 [0.3679 - 0.9462]	-2.19	0.0286
Recent shortbreath [yes vs. no]	0.8706	0.3271	2.3883 [1.2580 - 4.5345]	2.66	0.0078
Recent adenitis [yes vs. no]	-1.2650	0.6213	0.2822 [0.0835 - 0.9538]	-2.04	0.0418

S.E., Standard error; RR, relative risk; CI, confidence interval.

* Number of observations = 415.

DISCUSSION

A well-defined protocol for shrinking large datasets to a manageable list of predictors has not yet been documented due to the difficultly in accommodating different needs and types of dataset. The PAZ data is a good representation of a dataset produced by many disciplines to which this methodology could be applied; it encompasses data from several different sources (biological sampling, questionnaires, direct observation), both binomial and categorical variables, many missing values, and highly correlated variables. The procedure described above successfully

reduced 1376 variables to 42 predictors of malaria and produced wealth rankings for all homesteads. We believe this protocol is simple and efficient while having enough flexibility in its method to accommodate different datasets.

The steps to make a complete dataset were effective and flexible. The original dataset had an average of 8·99% missing values and after the limit of 10% missingness was applied, 89·89% of those were eliminated from the analysis. This supported the use of the 10% limit and makes the imputations process less computationally taxing. This limit could be disregarded or increased with other datasets if they can meet the requirement of missing at random. Piecewise multiple imputations by chained equations (MICE) successfully imputed the majority of variables with only five iterations. The few variables that were not completely imputed were found to be uniform in nature and would have been eliminated in the next step, i.e. the elimination of highly uniform variables, even if full imputation would have been encouraged by increasing iterations. The number of MICE iterations and the uniform limit could be adjusted according to the needs of individual dataset.

PCA successfully grouped a subset of asset and livestock variables to create wealth indices. Even though the wealth indices were not part of the final models, because of lack of statistical significance, several wealth variables were found to be significant which supports the validity of the wealth indices. The step of choosing the best α level for the elastic net adds to the flexibility of this protocol and will accommodate other datasets that have different numbers of correlated variables. The final GLM also has options regarding how variables are eliminated from the model, i.e. forward, backward or both directions. Finally, depending on the study's needs, one could choose an end point as the model with the lowest AIC or one only having significant variables remaining.

In future editions of this protocol, other tools could be added such as Bayesian disease mapping and network analysis. Steps to determine if missing observations are missing at random could be incorporated in addition to other model types, such as zero-inflated models, which would also add variety to its application for outcomes with low prevalence. Elastic net is a good technique for data mining of large datasets but can struggle with highly correlated variables sometimes requiring correlated variables to be removed from the model in order for other significant predictors to emerge. Exploring possible correlations >89% between variables could be performed if highly correlated variables are expected and if there was an undesirable effect on the model's output.

The proposed systematic data-mining approach resulted in the selection of 42 risk factors, a portion of which were related to exposure, wealth, or age. Increased exposure variables are those that increase time spent outside or near water (e.g. ‘own a bicycle for transportation’, ‘feeding livestock once a week’, ‘water is collected from the river for cattle in the dry season’). Homesteads that ‘keep ducks’ and/or ‘keep chickens’ were associated with lower homestead malaria prevalence, which may be a result of decreased human exposure to malaria via zooprophylaxis, in which mosquitos might feed on animals in the area, making them less likely to feed on humans [14]. Cement floors and brick or cement walls were also associated with lower homestead malaria prevalence, which may be due to a decrease in the amount of mosquitoes in the home due to physical barriers. These homestead characteristics also represent a homestead's wealth which aligns with the correlation between wealth and decreased disease incidence [2]. Other variables

selected which might represent wealth include having high-grade cattle (e.g. Friesian cross) and having access to healthcare such as 'seek treatment for breathing problem at a hospital', 'currently taking medications' and 'had fever and treated by chemist' (in Kenya, a chemist is understood to be a healthcare professional that practises pharmacy). It has been well documented that children have the highest malaria prevalence [15]. Younger age groups (5–9, 10–14, 15–19 years) were found to be significant determinants of increased malaria diagnosis, along with variables related to being younger (e.g. 'occupation – none'). While some of these examples are supported by previously published associations, confounders and variables not measured in this study could be factors; therefore, this approach should be viewed as more of a hypothesis-generating tool.

In conclusion, the proposed approach in which a number of statistical techniques are used including multiple imputation of missing values, wealth indexing through PCA, elastic net, and generalized linear regression models was successful in reducing a wide, sparse dataset to a more useful, simplified set of predictors for falciparum malaria infection prevalence and producing socioeconomic wealth indices. The protocol's flexibility suggests that it may be applied to other areas of epidemiology and infectious diseases and it also may serve as a hypothesis-generating tool to guide more detailed studies. In addition, we can now prioritize variables associated with malaria prevalence in the area of study and this can help the Kenyan health policy-makers prioritize their resources.

ACKNOWLEDGEMENTS

The authors thank Dr Cécile Ané, Assistant Professor at the Department of Statistics at the University of Wisconsin–Madison, for her help with the statistical analysis. This paper has been published with the permission of the Director of KEMRI. This research received no specific grant from any funding agency, commercial or not-for-profit sectors. The PAZ project was supported by the Wellcome Trust (E.M.F., grant number 085 308). W.A.deG is supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC).

DECLARATION OF INTEREST

None.

REFERENCES

1. Doble L, Fèvre EM. Focusing on neglected zoonoses. Veterinary Record 2010; 166: 546–547.
2. Filmer D, Pritchett LH. Estimating wealth effects without expenditure data – or tears: an application to educational enrollments in states of India. Demography 2001; 38: 115–132.
3. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2013.
4. WHO. Basic Laboratory Methods in Medical Parasitology. Geneva, Switzerland: World Health Organization, 1991.
5. Sterne JAC, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. British Medical Journal 2009; 338: b2393.
6. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. Journal of Statistical Software 2011; 45: 1–67.
7. van Buuren S. Flexible Imputation of Missing Data. Boca Raton, FL: Chapman & Hall/CRC Press, 2012.

8. Okell CN. An analysis of the dynamics of livestock and asset ownership with human health in a rural population in West Kenya (MSc Project Report). London, United Kingdom: Royal Veterinary College, 2011, 7 pp.
9. Borcard D, Gillet F, Legendre P. Numerical Ecology with R. New York: Springer, 2011, pp. 117.
10. Field A. Discovering Statistics Using SPSS, 3rd edn. London: SAGE Publications Ltd, 2009, pp. 233.
11. Oksanen J, et al. Vegan: community ecology. R package version 2.0-8 (http://CRAN.R-project.org/package=vegan), 2011.
12. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2005; 67: 301–320.
13. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 2010; 33: 1–22.
14. Iwashita H, et al. Push by a net, pull by a cow: can zooprophylaxis enhance the impact of insecticide treated bed nets on malaria control? Parasites & Vectors 2014; 7: 52.
15. Laurent A, et al. Performance of HRP-2 based rapid diagnostic test for malaria and its variation with age in an area of intense malaria transmission in southern Tanzania. Malaria Journal 2010; 9: 294.

CHAPTER 2.3

Factors associated with increased milk production for automatic milking systems.

Journal of dairy science, 99(5), pp.3824-3837.
https://doi.org/10.3168/jds.2015-10152

Tremblay, M.,* Hess, J.P.,* Christenson, B.M.,* McIntyre, K.K.,* Smink, B.,†
van der Kamp, A.J., ‡ de Jong, L.G.‡ and Döpfer, D.*

*Departments of Medical Sciences, Section of Food Animal Production Medicine, School of Veterinary Medicine, University of Wisconsin,
2015 Linden Drive, Madison 53706
†Lely North America, 775 250th Avenue, Pella, IA 50219
‡Lely International N.V., Cornelis van der Lelylaan 1, 3147 PB, Maassluis, the Netherlands

ABSTRACT
Automatic milking systems (AMS) are increasingly popular throughout the world. Our objective was to analyze 635 North American dairy farms with AMS for (risk) factors associated with increased milk production per cow per day and milk production per robot per day. We used multivariable generalized mixed linear regressions, which identified several significant risk factors and interactions of risk factors associated with milk production. Free traffic was associated with increased production per cow and per robot per day compared with forced systems, and the presence of a single robot per pen was associated with decreased production per robot per day compared with pens using 2 robots. Retrofitted farms had significantly less production in the first 4 yr since installation compared with production after 4 yr of installation. In contrast, newly built farms did not see a significant change in production over time since installation. Overall, retrofitted farms did not produce significantly more or less milk than newly constructed farms. Detailed knowledge of factors associated with increased production of AMS will help guide future recommendations to producers looking to transition to an AMS and maximize their production.

INTRODUCTION
Automatic milking systems (AMS) are becoming increasingly popular throughout the world, especially in North America. A variety of recommendations have been made for AMS facility structure and management to maximize production, but few of these recommendations have been explored scientifically (as reviewed by Jacobs and Siegford, 2012). As AMS are integrated into farms with larger herds, facility details such as the number of robots per pen and traffic type (i.e., how cows move among the AMS, lying stalls, and feeding area) become increasingly important as minor effects on milk production in the short term can have major economic implications in the long term.

In free traffic barns, each cow decides when to enter the AMS and can move freely among the AMS, lying stalls, and feeding area. Non-free traffic (i.e., forced) may vary in the level of guidance that is applied during movement, but always directs movement from the lying stalls to the AMS before allowing access to the feeding alley. In strictly forced traffic situations, a cow is always milked before entering the feeding area, whereas alternative arrangements use selection gates (i.e., guided, semi-forced, or select) to select only those cows that have exceeded their milking interval (Melin et al., 2006).

Current literature does not give a clear consensus as to the ideal traffic type for maximizing production. The few studies published examining the relationship between traffic type and milk yield were limited by sample size. Hermans et al. (2003) and Bach et al. (2009) did not find a significant difference in milk yield between different traffic types but were limited to 85 cows and 130 cows, respectively. Similarly, Munksgaard et al. (2011) demonstrated slightly greater production with free traffic barns, but this was not a significant finding potentially due to their limited sample size (70 cows). Gygax et al. (2007) collected data from 20 cows per farm on 4 free traffic type farms and 4 forced traffic type farms each with either Brown Swiss or Holstein cows, but found no significant difference between traffic types.

The effect of the number of robots per pen of cows has also never been investigated in AMS herds. It has been suggested that producers keep group sizes under 100 cows to ensure that all cows recognize each other (Grant and Albright, 2001); however, this value has not been formally

evaluated in an AMS (Rodenburg, 2002). Although, no significant difference in milk production or behavior was found between group sizes of 6 or 12 cows (Telezhenko et al., 2012), no similar studies have examined larger groups.

To date, no large-scale data analyses are available comparing AMS facility structures that account for differences in management and specific environments (as reviewed by Jacobs and Siegford, 2012). The general aim of this study was to apply multivariable generalized mixed linear regression models to a data set from 635 North American dairy farms to identify risk factors and interaction terms significantly associated with milk production per cow per day and milk production per robot per day. Our hypothesis is that traffic type and the number of robots per pen are risk factors significantly associated with milk production per robot per day and per cow per day. Factors that significantly affect a herd's maximum production limit could be used to create benchmark comparison groups for producers in the future. Detailed knowledge about factors associated with increased production of AMS will help guide future recommendations to producers looking to transition to an AMS and maximize their production.

MATERIALS AND METHODS

We analyzed a data set collected from weekly observations collected over 4 yr (2011–2014) at 635 North American dairy farms with Lely Astronaut AMS (Lely Industries N.V., Maassluis, the Netherlands). These data included 71,213 weekly observations containing 21 AMS variables.

Of the 21 available variables, frequencies per category were computed for 9 categorical variables (Table 1). Traffic_Type was coded as "Free" or "Forced." "Forced" Traffic_Type included both strictly forced and guided traffic (i.e., semi-forced, select) as both use one-way traffic to guide the cows and they have the same effect on low-ranking cows (Thune et al., 2002; Melin et al., 2006). The Robots_per_Pen variable represented the number of robots per pen of cows. By default, this variable also represents the number of cows in a pen and the pen's physical dimensions. By design, each pen will have about 60 cows per robot. For example, Robots_per_Pen of "1" is designed with one robot in a pen of about 60 cows and Robots_per_Pen of "2" is designed with 2 robots in a pen of about 120 cows. Because the number of robots per pen was of more interest, the number of cows per pen was not included in the regression to avoid multicollinearity. The physical sizes of the farms' pens were not available for our analysis; however, the number of cows per robot was included to account for different ratios of cows to robots (Table 2). Observations that were labeled as having a Robots_per_Pen of "Unknown" or "0" were coded as missing values. Breed was categorized into 3 levels: "Holstein," "Jersey," and "Other." Breed "Other" represents all other breeds including Ayrshire, Brown Swiss, Guernsey, Red and White, Crosses, Mixed, and Unknown. Farm_Goal was either characterized by the "Quota" system for farms in Canada or "Max_Production" for farms in the United States that produce with the goal of maximum milk production. Grazing and organic farms (n = 3,768 observations) were not included in the analysis because they had relatively few observations. Year_Since_Install represented the time from the installation of the robots to the time of each observation. Observations from farms utilizing robots for more than 4 yr were grouped together as "> 4 yr." Robot_Free_Time is the average percentage of time per day the robot is unoccupied by a cow (this does not include the time per day the system is automatically cleaning the robot and the milk lines to the tank). Robot_Free_Time was broken down into 5 levels (Table 1). Record_Year was limited to 2011 to 2014. "Winter" was classified

as December through February, “Spring” as March through May, “Summer” as June through August, and “Fall” as September through November.

The 12 numeric variables were summarized using descriptive statistics. The names of the numeric variables and their explanation are listed in Table 2. Observations with missing values were omitted. Observations that had fewer than 10 Cows_per_Robot or greater than 90 Cows_per_Robot were removed as outliers. The histogram of Average_DIM showed outliers beginning at 365 d. After observations with an Average_DIM greater than 365 d were omitted, 54,065 observations remained representing 529 farms. The number of observations per categorical variable and their reference level are detailed in Table 1. Categorical levels were chosen as the level we were least interested in estimating an effect while still having a balanced amount of observations. The summary statistics of the numeric variables are shown in Table 2. All statistical analyses were performed in R version 3.0.1 (R Development Core Team, 2013).

All numeric variables were inspected for normality by creating histograms. Numeric variables were log-transformed when normality was not present upon visual inspection of the histogram or when the order of magnitude of the values was more than 3 logs higher than the other variables. All numeric variables were scaled and centered using the scale function in R (i.e., the mean of each variable was subtracted from all values per variable in the data set and then divided by the variable’s standard deviation). The correlations between each pair of numeric variables were examined. A threshold of 0.7 was used to determine if a pair of variables was too highly correlated as this would lead to multicollinearity (Dormann et al., 2013). Based on this threshold criterion, the variables Milk_Production_per_Robot_per_Day and Cows_per_Robot are too highly correlated and therefore could not be used in the same regression model.

The number of observations differed among the combinations of independent variable levels (Table 1). These unequal numbers of observations also led to unequal variances among groups which render ANOVA methods unsuitable (Quinn and Keough, 2002). Therefore, we used multivariable generalized mixed linear regression models to generate 2 models (Quinn and Keough, 2002). Model 1 evaluated Milk_Production_per_Cow_per_Day, whereas model 2 evaluated Milk_Production_per_Robot_per_Day. The variable Farm_ID was taken as the random effect to account for differences between farms and repeated measures between farms. For Milk_Production_per_Robot_per_Day regression, Cows_per_Robot was taken as an offset. To examine the effect of one predictor upon the other 2-way interactions were selected using forward selection and a t-value limit of 4 (Pasta, 2011). Backward elimination of simple main effects was performed based on an error level, α, of 0.05 only if the variable was also not a confounder. Confounding effects were determined using the change-in-estimate method (Greenland, 1989). This method compares the model estimates before and after removal of the potential confounder variable from the model and any change in the estimates greater than 10% would signify a possible confounding effect. Goodness-of-fit measures were examined for each regression model using normality plots of residuals and log-likelihood, Akaike information criterion, Bayesian information criterion, and deviance measures. Interaction plots were produced using the R library package “effects” (Fox, 2003).

Table 1. The number of observations per categorical variable and their reference level

Categorical Variable [1]	Levels	Number of observations[2]
Traffic_Type	Free	50,268
	Forced [3]	3,797
Robots_per_Pen	1	30,946
	2 [3]	20,522
	3+	2,597
Breed	Holstein [3]	49,124
	Jersey	1,131
	Other	3,810
New_or_Retro	New [3]	27,211
	Retro	26,854
Farm_Goal	Quota- CAN [3]	35,641
	Max_Production-USA	18,424
Years_Since_Install	0-1 yr	18,115
	1-2 yrs	13,643
	2-3 yrs	8,449
	3-4 yrs	4,896
	> 4 yrs [3]	8,962
Robot_Free_Time	0-5 %	11,694
	5-10 %	11,413
	10-15 % [3]	7,601
	15-20 %	6,282
	> 20 %	17,075
Record_Year	2011 [3]	4,680
	2012	20,450
	2013	16,079
	2014	12,856
Season	Winter [3]	12,318
	Spring	14,530
	Summer	15,991
	Fall	11,226

[1] Variable explanations: Traffic_Type= how cows are allowed to move among areas of a barn. "Free" refers to a system where cows can decide when to enter the AMS and can move freely between the AMS, lying stalls and the feeding area. "Forced" traffic type uses a one-way traffic system towards the AMS; Robots_per_Pen= number of AMS robots per pen; Breed= breed of cattle; New_or_Retro= newly built or retro fitted barn; Farm_Goal= Operate under the "Quota" system for farms in Canada or "Max_Production" for farms in the USA that produce with the goal of maximum milk production; Years_Since_Install= how recently (in years) the AMS was installed; Robot_Free_Time = percent of time per day the robot is not occupied; Record_Year= year at the time of record; Season= "Winter" was classified as December through February; "Spring" as March through May, "Summer" as June through August, and "Fall" was classified as September through November.

[2] 54,065 total observations

[3] reference level. Reference levels were chosen as the level we were least interested in estimating an effect while still having a balanced amount of observations.

Table 2. Numeric variables explanation and descriptive statistics

Numeric variables	Variable explanation	Mean[1]	SD[2]
Milk_Production_per_Cow_per_Day	Average kg of milk produced[3]	31.98	4.91
Milk_Production_per_Robot_per_Day	Average kg of milk produced[4]	1626.80	396.99
Cows_per_Robot	Number of cows per number of robots	50.53	9.54
Average_DIM	Average days in milk of the herd	177.70	27.87
Concentrates	Average concentrate (kg) consumed in robot or automatic feeder per 100 kg of milk yield	15.86	5.38
Rest_Feed	Average percent of concentrates from the cow's allowance that was not dispensed that day (%)[5]	7.73	7.38
Refusals[3]	Average number of non-milking visits[3]	1.86	1.38
Failures[4]	Average number of failed milkings[4]	5.49	3.46
Milkings[3]	Average number of successful milkings[3]	2.91	0.36
Milk_Speed	Average milk yield (kg) per milking time (minutes)	2.59	0.31
Boxtime[3]	Average minutes in the AMS[3] (milking time and treatment time)	6.84	0.70
Connection_Attempts[4]	Average number of failure where teats were detected, but a quarter was unable to be connected[4]	1.41	0.23

[1] 54,065 total observations
[2] standard deviation
[3] per cow per day
[4] per robot per day
[5] Possible causes include: a cow was not visiting the robot often enough or she was not able to finish her meal giving her milking time

RESULTS

The first multivariable generalized linear mixed regression model incorporated 18 main effects and 20 of their 2-way interactions. Farm_ID was kept as a random effect and backward elimination removed Robots_per_Pen. Robots_per_Pen was not significant as a main effect (P = 0.75) and was not a confounder as the unadjusted model estimates on average only differed by 0.11% (SD 0.286) compared with the adjusted estimates. The regression results and equation for model 1 are shown in Table 3. The second multivariable generalized linear mixed regression model incorporated 18 main effects, 22 two-way interactions, Farm_ID as a random effect, and Cows_per_Robot as an offset. None of the variables were dropped during backward elimination because all factors were involved in significant main or interaction effects. The regression results and equation are shown in Table 4. Ten of the interactions were shared by both models.

The results of both models were very similar in terms of the direction of effects of the estimates and significance of the variables. Thus, results are described as their effects on Milk_Production to allude to both Milk_Production_per_Cow_per_Day and Milk_Production_per_Robot_per_Day unless specified. Most of these interaction effects are illustrated using Milk_Production_per_Cow_per_Day only. Most variables are also included in interactions. In

these cases, the interpretation of interaction effects is considered more important than the main effect (Pasta, 2011). The model results are shown in the Tables 3 and 4.

Traffic_Type "Free" was associated with greater Milk_Production (both models $P < 0.001$) compared with "Forced" Traffic_Type. As a main effect, "Free" Traffic_Type produces on average 1.11 kg (CI: 0.79–1.43) more Milk_Production_per_Cow_per_Day and 67.2 kg (CI: 48.6–86.0) more Milk_Production_per_Robot_per_Day than "Forced."

On average, Robots_per_Pen "2" (2 robots per 120 cows) had greater ($P < 0.001$) Milk_Production_per_Robot_per_Day compared with Robots_per_Pen "1" (one robot per 60 cows) as a main effect. Robots_per_Pen was included in 2 interactions (Table 4). The interaction Robots_per_Pen:Record_Year describes how Robots_per_Pen "3+" and "2" had greater Milk_Production_per_Robot_per_Day compared with Robots_per_Pen "1" in Record_Year 2011 and 2012 (see nonoverlapping confidence intervals in Figure 1). Robots_per_Pen was also in an interaction with Milkings. The difference between Robots_per_Pen "1" and Robots_per_Pen "2" or "3+" becomes larger as Milkings decreases below its average of 2.91 (SD: 0.36) Milkings (Table 4).

We found associations between increased Milkings, Milk_Speed, or Boxtime and increased Milk_Production (all $P < 0.001$). The negative estimates for the interactions Milk_Speed:Boxtime and Milkings:Milk_Speed in model 1 and Milkings:Milk_Speed and Milkings:Boxtime in model 2 indicate that the positive effect of the variables on Milk_Production decreases (smaller increase in production per unit change) as the value of the other variable in the interaction increases (all $P < 0.001$). For example, the positive effect of Milk_Speed on Milk_Production decreases as Milkings increases. All 3 variables were also part of interactions with Connection_Attempts in both models that had negative estimates suggesting that the negative effect of Connection_Attempts on Milk_Production increases (all $P < 0.001$) as Milkings, Milk_Speed, or Boxtime increase (Tables 3 and 4).

"Jersey" Breed was associated ($P < 0.001$) with less Milk_Production than "Holstein" Breed, whereas the "Other" Breed category was not significantly different from "Holstein" Breed as a main effect (model 1, $P = 0.31$; model 2, $P = 0.08$). Although Breed was part of 2 interactions in both models, as a main effect "Jersey" produces on average 3.72 kg (CI: 3.29–4.14) less Milk_Production_per_Cow_per_Day and 216.71 kg (CI: 193.2–239.9) less Milk_Production_per_Robot_per_Day compared with "Holstein" (milk production not energy corrected). Tables 3 and 4 illustrate the increase in difference in milk production between Holstein and Jersey breeds as the number of milkings increases. (model 1 and 2, $P < 0.001$). When Refusals decreases, the difference in Milk_Production_per_Cow_per_Day between "Holstein" and "Jersey" increases ($P < 0.001$). When Connection_Attempts increases, the difference in Milk_Production_per_Robot_per_Day between "Holstein" and "Jersey" increases ($P < 0.001$).

Increases in Average_DIM, Failures, Concentrates, Refusals, and Connection_Attempts are all associated (all $P < 0.01$) with decreased Milk_Production (all of the above variables are also included in interactions), whereas an increase in Rest_Feed is associated with increased Milk_Production (models 1 and 2, $P < 0.01$). The Farm_Goal:Record_Year interaction implies

that "Max_Production" farms (United States) had more Milk_Production_per_Cow_per_Day compared with "Quota" farms (Canada) in all years except for 2011 (all $P < 0.001$).

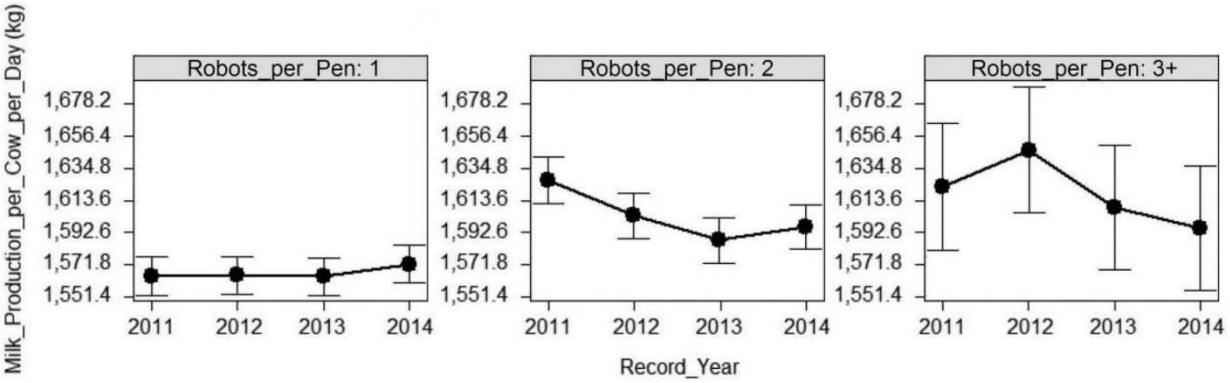

Figure 1. Interaction effects between predictor variables Robots_per_Pen and Record_Year with Milk_Production_per_Cow_per_Day as the response variable.

The interactions between Concentrates and 5 other numeric variables all have negative coefficients. Three out of the 5 interactions are included in both models as illustrated in Tables 3 and 4. This means that the positive effects of Milkings, Milk_Speed, and Boxtime diminish as Concentrates increases (all $P < 0.001$) and that the negative effects of Refusals and Connection_Attempts increases (all $P < 0.001$) as Concentrates increases (Figure 2).

Cows_per_Robot has a significant main effect ($P < 0.001$) in the Milk_Production_per_Cow_per_Day model and is included in 3 interactions with Farm_Goal, Milkings, and Boxtime (all $P < 0.001$). As Boxtime and Milkings decrease below 6.1 min or 2.4 milkings per cow per day, respectively, with other variables held at their mean, an increase in Cows_per_Robot will transition to having a negative effect on Milk_Production_per_Cow_per_Day (Figures 3 and 4).

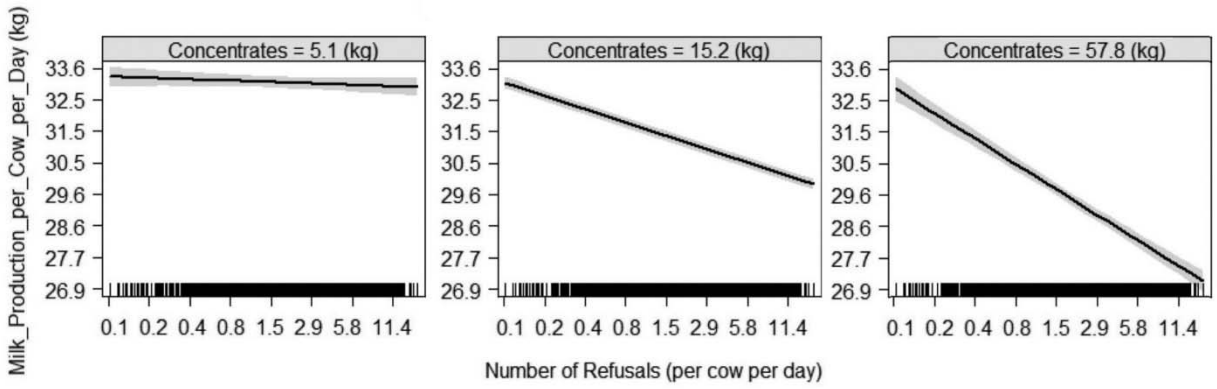

Figure 2. Interaction effects between predictor variables Concentrates and Refusals with Milk_Production_per_Cow_per_Day as the response variable.

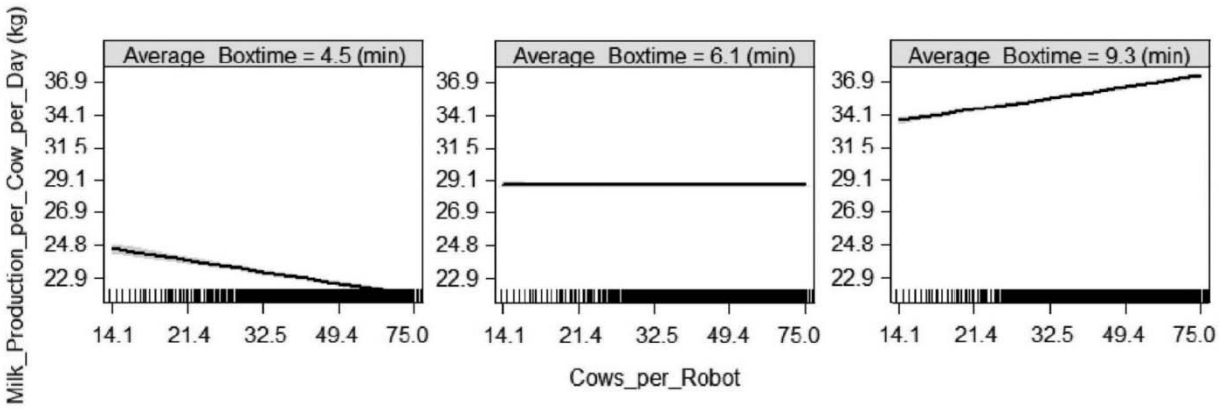

Figure 3. Interaction effects between predictor variables Cows_per_Robot and Boxtime with Milk_Production_per_Cow_per_Day as the response variable.

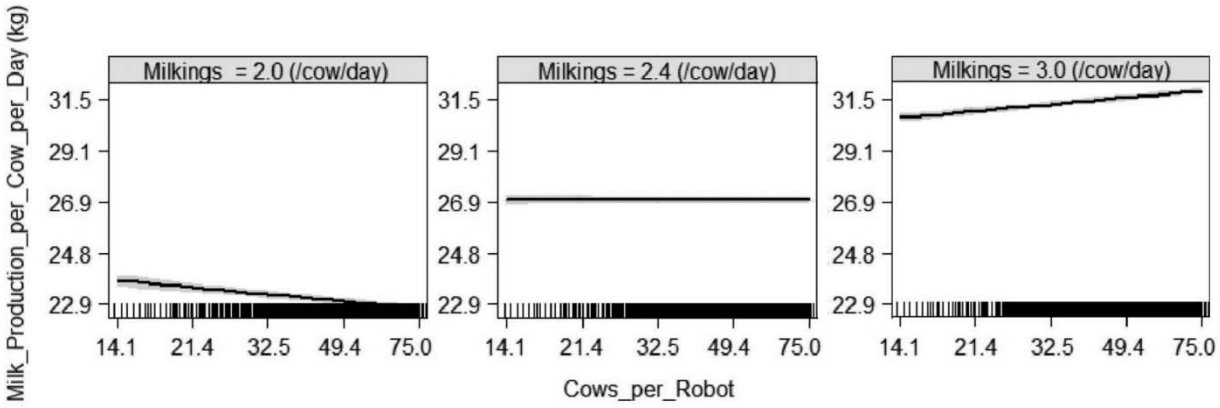

Figure 4. Interaction effects between predictor variables Cows_per_Robot and Milkings with Milk_Production_per_Cow_per_Day as the response variable.

Figure 5 illustrates the interaction New_or_Retro:Years_Since_Install for Milk_Production_per_Cow_per_Day, although the effect holds true for both models. The trend in Milk_Production_per_Cow_per_Day estimates and confidence intervals for the "New" farms, demonstrates that Milk_Production does not change significantly between Years_Since_Install groups (depicted by overlapping confidence intervals in Figure 5). In contrast, for "Retro" farms, the right-handed side of Figure 5 shows that Milk_Production is significantly greater in "> 4 yr" compared with all other categories of Years_Since_Install (depicted by nonoverlapping confidence intervals in Figure 5). On average, "Retro" farms do not produce significantly more or less than "New" farms, which is demonstrated by a nonsignificant P-value for the main effect (P = 0.06; Table 3). The confidence intervals for Milk_Production between "New" and "Retro" farms overlap in all groups except for the "1-2 yr" group. After 2 yr, the Milk_Production for "New" or "Retro" farms are not significantly different (Figure 5).

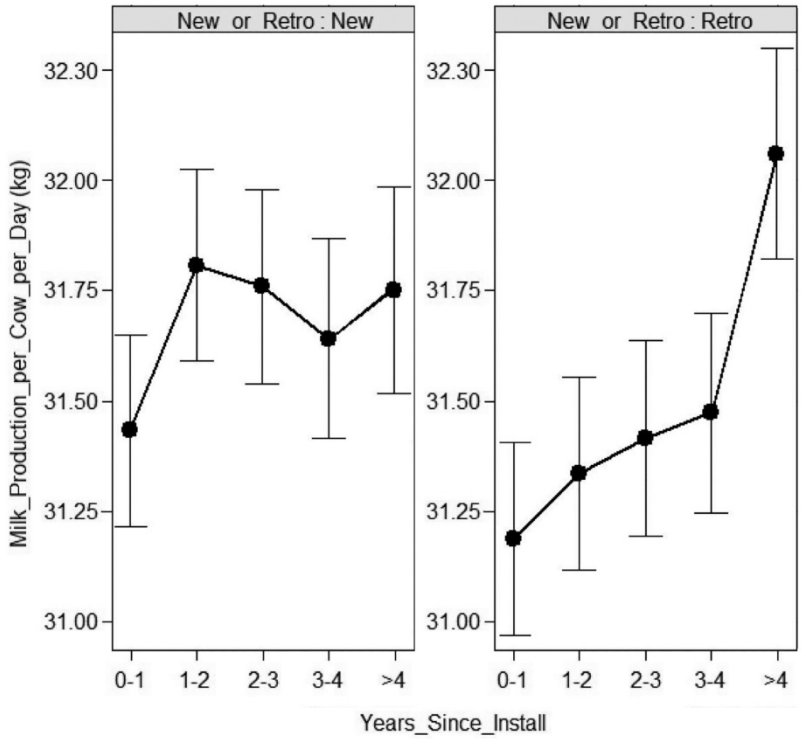

Figure 5. Interaction effects between predictor variables New_or_Retro and Years_Since_Install with Milk_Production_per_Cow_per_Day as the response variable.

The "Spring" Season had greater ($P < 0.001$) Milk_Production compared with Season "Winter." Season "Fall" was associated with less ($P < 0.001$) Milk_Production compared with "Winter." The effect of "Summer" on Milk_Production_per_Cow_per_Day was not significantly different from "Winter" ($P = 0.10$). In the Milk_Production_per_Robot_per_Day model, Season was part of an interaction with Farm_Goal wherein "Max_Production" had decreased seasonal effects (all $P < 0.001$) compared with Farm_Goal "Quota." This causes "Summer" to have on average greater milk production than "Winter" in "Quota" farms (Canada) but less than "Winter" in "Max_Production" farms (United States).

Robot_Free_Time was significant in both models (all levels $P < 0.001$) but depended highly on its interaction with Record_Year. There were relatively few observations with Robot_Free_Time greater than 15% in Record_Year 2011 and 2012 (480 observations from 43 farms). All Robot_Free_Time categories for year 2013 and 2014 had more than 2,000 observations except the 2014 "0-5%" category, which had 225 observations. Record_Year 2013 and 2014 did not have any significant differences in Milk_Production_per_Cow_per_Day by Robot_Free_Time category (non-overlapping confidence intervals). Significantly greater Milk_Production_per_Robot_per_Day is seen within the "> 20%" category in 2013 and 2014 compared with the reference group "10-15%" (Table 4).

Table 3. Milk_Production_per_Cow_per_Day Regression Model Results[1]

	Estimate	SE[2]	p-value
(Intercept)	-0.261	0.066	7.12E-05
Traffic_Type- Free (ref. Forced)	0.216	0.062	5.45E-04
Cows_per_Robot[3]	0.021	0.003	2.06E-13
Average_DIM[3]	-0.053	0.001	<2.00E-16
Rest_Feed	0.005	0.002	1.78E-03
Concentrates[3]	-0.083	0.003	<2.00E-16
Refusals[3]	-0.082	0.002	<2.00E-16
Failures[3]	-0.004	0.001	2.38E-03
Milkings[3]	0.671	0.002	<2.00E-16
Milk_Speed[3]	0.570	0.002	<2.00E-16
Boxtime[3]	0.508	0.002	<2.00E-16
Connection_Attempts[3]	-0.046	0.002	<2.00E-16
Farm_Goal- Max Production (ref. Quota)[3]	*-0.005*	*0.034*	*8.86E-01*
New_or_Retro- Retro (ref. New)[3]	*0.062*	*0.033*	*5.78E-02*
Breed- Jersey (ref. Holstein)[3]	-0.785	0.095	1.07E-16
Breed- Other (ref. Holstein)[3]	*-0.059*	*0.058*	*3.13E-01*
Years_Since_Install- 0-1yrs (ref. > 4 yrs)[3]	-0.060	0.012	3.03E-07
Years_Since_Install- 1-2 yrs (ref. > 4 yrs)[3]	*0.013*	*0.010*	*1.94E-01*
Years_Since_Install- 2-3 yrs (ref. > 4 yrs)[3]	*0.001*	*0.009*	*9.16E-01*
Years_Since_Install- 3-4 yrs (ref. > 4 yrs)[3]	-0.022	0.007	2.15E-03
Robot_Free_Time - 0-5% (ref.10-15%)[3]	0.085	0.015	3.64E-08
Robot_Free_Time - 5-10% (ref.10-15%)[3]	0.082	0.015	1.12E-07
Robot_Free_Time - 15-20% (ref.10-15%)[3]	-0.440	0.107	3.82E-05
Robot_Free_Time - >20% (ref.10-15%)[3]	-0.024	0.006	2.29E-05
Season- Spring (ref.Winter)	0.041	0.003	<2.00E-16
Season- Summer (ref.Winter)	*-0.005*	*0.003*	*1.03E-01*
Season- Fall (ref.Winter)	-0.044	0.003	<2.00E-16
Record_Year- 2012 (ref. 2011)[3]	0.078	0.016	1.29E-06
Record_Year- 2013 (ref. 2011)[3]	0.118	0.016	1.42E-13
Record_Year- 2014 (ref. 2011)[3]	0.163	0.017	<2.00E-16
Milk_Speed:Boxtime	-0.054	0.001	<2.00E-16
Milkings:Milk_Speed	-0.053	0.001	<2.00E-16
Concentrates:Boxtime	-0.034	0.001	<2.00E-16
Concentrates:Milkings	-0.014	0.001	<2.00E-16
Boxtime:Connection_Attempts	-0.030	0.001	<2.00E-16
Milkings: Connection_Attempts	-0.015	0.001	<2.00E-16
Farm_Goal: Record_Year			
Max_Production (ref.Quota): 2012 (ref. 2011)	0.051	0.008	1.66E-10
Max_Production (ref.Quota): 2013 (ref. 2011)	0.052	0.008	4.51E-11
Max_Production (ref.Quota): 2014 (ref. 2011)	0.087	0.008	<2.00E-16
Refusals:Breed- Jersey (ref. Holstein)	0.192	0.011	<2.00E-16
Refusals:Breed- Other (ref. Holstein)	0.030	0.007	2.00E-05
Milkings:Breed- Jersey (ref. Holstein)	-0.178	0.009	<2.00E-16
Milkings:Breed- Other (ref. Holstein)	-0.034	0.006	8.29E-08
Milk_Speed: Connection_Attempts	-0.023	0.002	<2.00E-16
Concentrates: Connection_Attempts	-0.028	0.002	<2.00E-16

Concentrates: Refusals	-0.019	0.002	<2.00E-16
Average_DIM: Refusals	0.019	0.001	<2.00E-16
Average_DIM:Boxtime	0.015	0.001	<2.00E-16
Robot_Free_Time :Record_Year			
0-5% (ref.10-15%): 2012 (ref. 2011)	-0.088	0.017	1.21E-07
5-10% (ref.10-15%): 2012 (ref. 2011)	-0.070	0.017	4.54E-05
15-20% (ref.10-15%): 2012 (ref. 2011)	0.624	0.109	1.06E-08
>20% (ref.10-15%): 2012 (ref. 2011)	*-0.032*	*0.017*	*6.57E-02*
0-5% (ref.10-15%): 2013 (ref. 2011)	-0.090	0.017	6.65E-08
5-10% (ref.10-15%): 2013 (ref. 2011)	-0.065	0.016	7.58E-05
15-20% (ref.10-15%): 2013 (ref. 2011)	0.428	0.107	6.42E-05
>20% (ref.10-15%): 2013 (ref. 2011)	*0.009*	*0.007*	*1.68E-01*
0-5% (ref.10-15%): 2014 (ref. 2011)	*-0.032*	*0.022*	*1.44E-01*
5-10% (ref.10-15%): 2014 (ref. 2011)	-0.049	0.017	3.16E-03
15-20% (ref.10-15%): 2014 (ref. 2011)	0.433	0.107	5.13E-05
Failures:Milkings	-0.003	0.001	2.72E-03
Cows_per_Robot:Boxtime	0.029	0.001	<2.00E-16
Cows_per_Robot:Milkings	0.023	0.001	<2.00E-16
Cows_per_Robot:Farm_Goal- Max Production (ref.Quota)	0.039	0.005	<2.00E-16
New_or_Retro:Year_Since_Install			
Retro (ref. New): 0-1yrs (ref. > 4yrs)	-0.112	0.012	<2.00E-16
Retro (ref. New): 1-2yrs (ref. > 4yrs)	-0.155	0.012	<2.00E-16
Retro (ref. New): 2-3yrs (ref. > 4yrs)	-0.128	0.011	<2.00E-16
Retro (ref. New): 3-4yrs (ref. > 4yrs)	-0.095	0.010	<2.00E-16

[1] Regression model equation: Milk_Production_per_Cow_per_Day ~ Traffic_Type + Cows_per_Robot + Average_DIM + Rest_Feed + Concentrates + Refusals + Failures + Milkings + Milk_Speed + Boxtime + Connection_Attempts + Farm_Goal + New_of_Retro + Breed + Years_Since_Install + Robot_Free_Time + Season + Record_Year + (1|Farm_ID) + Milk_Speed:Boxtime + Milkings: Milk_Speed + Concentrates: Boxtime + Concentrates:Milkings + Boxtime: Connection_Attempts + Milkings:Connection_Attempts + Farm_Goal: Record_Year + Refusals:Breed + Milkings:Breed + Milk_Speed:Connection_Attempts + Concentrates: Connection_Attempts + Concentrates:Refusals + Average_DIM: Refusals + Average_DIM:Boxtime + Robot_Free_Time : Record_Year + lognr_failures:Milkings + Cows_per_Robot:Boxtime + Cows_per_Robot: Milkings + Cows_per_Robot:Farm_Goal + New_or_Retro:Year_Since_Install)

[2] Standard error

[3] The variable is also included in an interaction

Italics and grey: Not a significant effect (P > 0.05)

Table 4: Milk_Production_per_Robot_per_Day Regression Model Results[1]

	Estimate	Std. Error[2]	p-value
(Intercept)	*0.086*	*0.049*	*7.63E-02*
Traffic_Type- Free (ref. Forced)	0.158	0.043	2.72E-04
Average_DIM	-0.027	0.001	<2.00E-16
Rest_Feed	0.023	0.001	<2.00E-16
Concentrates[3]	-0.073	0.002	<2.00E-16
Refusals[3]	-0.047	0.003	<2.00E-16
Failures	-0.007	0.001	7.52E-11
Milkings[3]	0.416	0.002	<2.00E-16
Milk_Speed[3]	0.310	0.003	<2.00E-16

Boxtime[3]	0.344	0.004	<2.00E-16
Connection_Attempts[3]	-0.043	0.002	<2.00E-16
Farm_Goal- Max Production (ref.Quota)[3]	*-0.016*	*0.024*	*4.90E-01*
New_or_Retro- Retro (ref. New)	*0.036*	*0.023*	*1.17E-01*
Breed- Jersey (ref. Holstein)[3]	-0.560	0.065	<2.00E-16
Breed- Other (ref. Holstein)[3]	*-0.071*	*0.040*	*7.91E-02*
Years_Since_Install- 0-1yrs (ref. > 4 yrs)[3]	-0.050	0.009	8.05E-09
Years_Since_Install- 1-2 yrs (ref. > 4 yrs)[3]	*-0.008*	*0.008*	*2.75E-01*
Years_Since_Install- 2-3 yrs (ref. > 4 yrs)[3]	-0.016	0.006	1.16E-02
Years_Since_Install- 3-4 yrs (ref. > 4 yrs)[3]	-0.038	0.005	1.57E-12
Robots_per_Pen- 1 (ref. 2)[3]	-0.146	0.023	4.54E-10
Robots_per_Pen- 3+ (ref. 2)[3]	*-0.009*	*0.053*	*8.65E-01*
Robot_Free_Time - 0-5% (ref.10-15%)[3]	-0.089	0.011	2.11E-15
Robot_Free_Time - 5-10% (ref.10-15%)[3]	-0.038	0.011	5.71E-04
Robot_Free_Time - 15-20% (ref.10-15%)[3]	0.502	0.076	5.18E-11
Robot_Free_Time - >20% (ref.10-15%)[3]	0.052	0.004	<2.00E-16
Season- Spring (ref.Winter)	0.034	0.002	<2.00E-16
Season- Summer (ref.Winter)	0.022	0.002	<2.00E-16
Season- Fall (ref.Winter)	-0.024	0.003	<2.00E-16
Record_Year- 2012 (ref. 2011)[3]	*-0.003*	*0.012*	*8.09E-01*
Record_Year- 2013 (ref. 2011)[3]	-0.148	0.011	<2.00E-16
Record_Year- 2014 (ref. 2011)[3]	-0.124	0.012	<2.00E-16
Milkings :Milk_Speed	-0.030	0.001	<2.00E-16
Milkings :Boxtime	-0.012	0.001	<2.00E-16
Boxtime: Connection_Attempts	-0.029	0.001	<2.00E-16
Milkings : Connection_Attempts	-0.018	0.001	<2.00E-16
Robot_Free_Time :Record_Year			
0-5% (ref.10-15%): 2012 (ref. 2011)	-0.068	0.012	3.03E-08
5-10% (ref.10-15%): 2012 (ref. 2011)	-0.048	0.012	8.95E-05
15-20% (ref.10-15%): 2012 (ref. 2011)	-0.381	0.078	1.03E-06
>20% (ref.10-15%): 2012 (ref. 2011)	-0.156	0.013	<2.00E-16
0-5% (ref.10-15%): 2013 (ref. 2011)	0.089	0.012	2.38E-13
5-10% (ref.10-15%): 2013 (ref. 2011)	0.058	0.012	9.92E-07
15-20% (ref.10-15%): 2013 (ref. 2011)	-0.499	0.077	6.82E-11
>20% (ref.10-15%): 2013 (ref. 2011)	*0.005*	*0.005*	*2.73E-01*
0-5% (ref.10-15%): 2014 (ref. 2011)	0.140	0.016	<2.00E-16
5-10% (ref.10-15%):: 2014 (ref. 2011)	0.055	0.012	6.14E-06
15-20% (ref.10-15%): 2014 (ref. 2011)	-0.496	0.077	9.30E-11
Milk_Speed: Connection_Attempts	-0.035	0.001	<2.00E-16
Concentrates: Connection_Attempts	-0.017	0.001	<2.00E-16
Robots_per_Pen: Record_Year			
1 (ref. 2): 2012 (ref. 2011)	0.059	0.006	<2.00E-16
3+ (ref. 2): 2012 (ref. 2011)	0.122	0.013	<2.00E-16
1 (ref. 2): 2013 (ref. 2011)	0.090	0.006	<2.00E-16
3+ (ref. 2): 2013 (ref. 2011)	0.060	0.013	3.86E-06
1 (ref. 2) 2014 (ref. 2011)	0.088	0.006	<2.00E-16
3+ (ref. 2): 2014 (ref. 2011)	*0.006*	*0.014*	*6.70E-01*
Concentrates :Boxtime	-0.021	0.001	<2.00E-16
Concentrates : Milkings	-0.015	0.001	<2.00E-16
Boxtime: Record_Year- 2012 (ref. 2011)	-0.014	0.003	3.74E-07

Boxtime : Record_Year- 2013 (ref. 2011)	-0.028	0.003	<2.00E-16
Boxtime : Record_Year- 2014 (ref. 2011)	-0.034	0.003	<2.00E-16
Milkings : Breed- Jersey (ref. Holstein)	-0.074	0.006	<2.00E-16
Milkings : Breed- Other (ref. Holstein)	*-0.007*	*0.004*	*7.87E-02*
Boxtime : Farm_Goal- Max Production (ref. Quota)	-0.046	0.002	<2.00E-16
Years_Since_Install: New_or_Retro			
0-1 yrs (ref. > 4 yrs) : Retro (ref. New)	-0.088	0.009	<2.00E-16
1-2 yrs (ref. > 4 yrs) : Retro (ref. New)	-0.113	0.009	<2.00E-16
2-3 yrs (ref. > 4 yrs) : Retro (ref. New)	-0.085	0.008	<2.00E-16
3-4 yrs (ref. > 4 yrs) : Retro (ref. New)	-0.051	0.008	1.59E-11
Milkings : Robots_per_Pen- 1 (ref. 2)	0.028	0.002	<2.00E-16
Milkings : Robots_per_Pen-3+ (ref. 2)	0.030	0.005	2.80E-08
Farm_Goal:Season			
Max Production (ref. Quota): Spring (ref. Winter)	-0.018	0.004	9.85E-06
Max Production (ref. Quota): Summer (ref. Winter)	-0.045	0.004	<2.00E-16
Max Production (ref. Quota): Fall (ref. Winter)	-0.022	0.004	5.22E-07
Refusals: Years_Since_Install- 0-1 yrs (ref. > 4 yrs)	0.057	0.004	<2.00E-16
Refusals : Years_Since_Install- 1-2 yrs (ref. > 4 yrs)	0.030	0.004	1.04E-16
Refusals : Years_Since_Install- 2-3 yrs (ref. > 4 yrs)	0.032	0.004	<2.00E-16
Refusals : Years_Since_Install- 3-4 yrs (ref. > 4 yrs)	0.021	0.004	7.11E-08
Milk_Speed: Years_Since_Install- 0-1 yrs (ref. > 4 yrs)	0.042	0.004	<2.00E-16
Milk_Speed : Years_Since_Install- 1-2 yrs (ref. > 4 yrs)	0.019	0.004	3.94E-08
Milk_Speed : Years_Since_Install- 2-3 yrs (ref. > 4 yrs)	0.020	0.004	2.68E-08
Milk_Speed : Years_Since_Install- 3-4 yrs (ref. > 4 yrs)	0.017	0.004	6.49E-06
Connection_Attempts : Breed- Jersey (ref. Holstein)	-0.087	0.007	<2.00E-16
Connection_Attempts : Breed- Other (ref. Holstein)	*-0.009*	*0.006*	*1.37E-01*
Concentrates:Milk_Speed	-0.012	0.001	<2.00E-16
Connection_Attempts: Farm_Goal- Max Production (ref. Quota)	0.045	0.004	<2.00E-16
Boxtime: Years_Since_Install- 0-1 yrs (ref. > 4 yrs)	0.037	0.003	<2.00E-16
Boxtime : Years_Since_Install- 1-2 yrs (ref. > 4 yrs)	0.031	0.003	<2.00E-16
Boxtime : Years_Since_Install- 2-3 yrs (ref. > 4 yrs)	0.022	0.004	4.40E-10
Boxtime : Years_Since_Install- 3-4 yrs (ref. > 4 yrs)	0.025	0.004	9.43E-11

[1] Regression model equation: Milk_Production_per_Robot_per_Day ~ Traffic_Type + Cows_per_Robot + Average_DIM + Rest_Feed + Concentrates + Refusals + Failures + Milkings + Milk_Speed + Boxtime + Connection_Attempts + Farm_Goal + New_or_Retro + Breed + Years_Since_Install + Robots_per_Pen + Robot_Free_Time + Season + Record_Year + (1|Farm_ID) + Milkings:Milk_Speed + Milkings: Boxtime + Boxtime: Connection_Attempts + Milkings: Connection_Attempts + Robot_Free_Time :Record_Year + Milk_Speed: Connection_Attempts + Concentrates: Connection_Attempts + Robots_per_Pen:Record_Year + Concentrates: Boxtime + Concentrates: Milkings + Boxtime:Record_Year + Milkings:Breed + Boxtime: Farm_Goal + New_or_Retro: Years_Since_Install + Milkings: Robots_per_Pen + Farm_Goal:Season + Refusals: Years_Since_Install + Milk_Speed: Years_Since_Install + Breed: Connection_Attempts + Concentrates: Milk_Speed + Connection_Attempts: Farm_Goal + Boxtime: Years_Since_Install, offset= Cows_per_Robot)

[2] Standard error

[3] The variable is also included in an interaction

Italics and grey: Not a significant effect

DISCUSSION

Goal-oriented production processes such as milk production in dairy herds can benefit from predictive models that correctly incorporate the effects of multiple risk factors and interactions of such factors simultaneously. Interactions facilitate the understanding of diverse relationships among management factors and improve our understanding of different management styles. Often the modeling approaches reported in literature are limited to the analysis of selected interactions. In contrast, the approach reported here systematically mined through all possible interactions to determine the most meaningful ones. Multivariable generalized mixed linear regression models identified sources of variation in Milk_Production for a large number of farms utilizing AMS across North America. Despite a large diversity of environments and management styles, several associations were consistent.

We found that "Free" Traffic_Type was associated with greater Milk_Production than "Forced" Traffic_Type in model 1 and model 2. A possible explanation for this is that forced traffic decreases the total feed a cow consumes, the total amount of time eating, and the number of times a cow visits the feed bunk (Ketelaar-de Lauwere et al., 2000; Harms et al., 2002; Melin et al., 2007). These changes in feeding behavior could potentially lead to rumen acidosis (Bach et al., 2009). Hermans et al. (2003) and Rodenburg (2012) suggested that forced traffic might negatively affect the behavior of timid cows more than dominant cows and Thune et al. (2002) found longer waiting periods for the AMS per cow in forced compared with free traffic type. Winter and Hillerton (1995) found that cows spent less time resting with forced traffic type although Munksgaard et al. (2011) did not find a significant difference in milking frequency, production, lying time, or feeding time between traffic type in a study with 70 cows.

Limited sample size has been a common problem in previous studies comparing traffic types. The largest sample size of any of the previous studies mentioned was 160 cows and most other studies included fewer than 100 cows. Furthermore, many of these studies did not correct for confounders and other risk factors in the applied modeling approaches. For example, Gygax et al. (2007) collected data on 160 cows from 2 different traffic type farms each with either Brown Swiss or Holstein cows. Although no significant difference between traffic types was found, the herds in the study were not matched by breed nor did the investigators correct for breed in their model. The majority of free traffic type herds (3/4 herds) were Brown Swiss herds, whereas 3 out of 4 forced traffic type herds were Holstein herds. Even though both Brown Swiss and Holstein breeds are high yielding, Holsteins still produce significantly more milk than Brown Swiss (De Marchi et al., 2008). Many studies also have low numbers of cows per robot or did not correct for differences in cows per robot among study groups. For example, the Munksgaard et al. (2011) study had only 35 cows per robot, and Gygax et al. (2007) did not correct for the variation in cows per robot in spite of a range from 30 to 56 cows per robot. Our study showed Cows_per_Robot had a significant effect on Milk_Production and was included in several significant interactions with other risk factors. The number of cows per robot could affect results because the hierarchic structure of the herd and its influence on timid cows is suspected to play a lesser role in small groups and to have a greater effect in large groups. This effect may further increase under normal farm circumstances with little Robot_Free_Time.

We found that Robots_per_Pen "2" was associated with greater Milk_Production_per_Robot_per_Day than Robots_per_Pen "1." Although this difference was

not significant for all years (Figure 1), the trend is consistent. A possible cause for this difference is that Robots_per_Pen "2" allows timid cows to have more opportunities for milking if dominant cows monopolize one of the robots. Timid cows have been shown to wait longer to use the robot than higher-ranking cows (Ketelaar-de Lauwere et al., 1996; Thune et al., 2002; Melin et al., 2006). Another benefit of having more than one robot per pen is that down time due to daily maintenance of a robot does not necessarily disrupt behavior because cows can still be milked in the other robot (Rodenburg, 2004). In addition, the effect of Robots_per_Pen could be representing the effect of the physical dimensions of the pens or the total group size, although these distinctions cannot be made based on this data set. Having 2 robots per pen will not be feasible for all farms especially those with low numbers of cows. As this is the first study examining the difference in number of robots per pen, subsequent studies are needed to understand the effects of pen dimensions and group size. It is important to consider additional arguments for or against increased Robots_per_Pen, which may reflect labor requirement, animal well-being and health, cow handling, as well as economics.

We found that increases in Boxtime and Milkings are associated with increased Milk_Production, but Boxtime and Milkings rarely increase simultaneously. As Cows_per_Robot increases, the Milkings will decrease and Boxtime will increase. Including interactions in the models allowed us to examine how cows with high Milk_Speed can help manage the conflicting goals of max production and cost efficiency. Cows with high Milk_Speed can be in a pen with greater Cows_per_Robot without negatively affecting production because cows with high Milk_Speed spend less time in the AMS and require fewer Milkings, while concurrently maintain greater production than cows with average Milk_Speed, Milkings, and Boxtime (Tables 3 and 4). Therefore, selection for faster Milk_Speed may balance milk production with efficiency and a farm's ability to increase their Cows_per_Robot will depend on the average Milk_Speed of their herd.

We found that an increase in Concentrates is associated with decreased Milk_Production. This association most likely reflects environments that do not support the production of high-quality feed such as corn silage. The use of low-energy basic forages in the feed bunk ration increases the amount of concentrate that needs to be supplemented in the robot. Farms that must feed high volumes of concentrate in their robots due to their geographic circumstance experience greater reductions in productivity due to increases in refusals or connection attempts (Figure 2). Another possible cause of the effect of Concentrates on Milk_Production may be the variation in milk yield among cows in the herd. A herd with high milk yield and low variation can successfully lessen the volume of concentrate fed in the AMS and maintain the robot attractiveness for the majority of the cows. A herd with a greater variation in milk production may require a higher volume of Concentrates to keep the low-yielding cows attracted to the robot.

Although it is suggested to maintain the refusals above 1 per cow per day (Kozlowska et al., 2013), the negative effect of refusals on Milk_Production should not be overlooked. Stefanowska et al. (2000) found refusals negatively affected behavior because cows went through fewer complete behavioral cycles (eating and lying down) after nonmilking visits (refusals) versus after milking visits. The number of refusals per cow per day is not a very good indication of overcrowding. Not only was the interaction Refusals:Cows_per_Robot not selected in the forward selection process, but refusals can be a positive sign of cows' interest and curiosity in coming to the robot. Therefore,

we suggest using the relationships between boxtime, milk speed, and number of milkings to develop overcrowding standards (Figure 3 and 4). This strategy could complement the evaluation of overcrowding based on number of refusals alone.

To our knowledge, this is the first attempt to assess the differences in the use of AMS between new and existing farms. We found that both scenarios can result in high milk production per cow or per robot, only that the "Retro" fitted farm need about 2 yr to reach the production levels of a "New" barn. Also, "Retro" fitted farms have significantly greater production in "> 4 yr" Years_Since_Install compared with all other categories of Years_Since_Install (Figure 5). The difference in "New" and "Retro" fitted barns might reflect differences in the average lactation numbers of the cows. When new barns are constructed, a sharp increase in herd size often occurs consisting of many replacement heifers that are introduced to the robots during their first lactation. Retrofitted barns often must adapt older cows to the robot after those cows were previously trained for in-parlor milking. Heifers have been shown to learn to use the AMS quicker than cows (Jago and Kerrisk, 2011). The population of cows that were originally trained on robots as heifers will slowly increase in relative numbers over the years on retro farms. In addition, retrofitted farms also reflect a more gradual genetic improvement in the herd through culling and replacement rather than a sudden expansion. These factors may explain why the "Retro" fitted barns, unlike the "New" barns, have a significantly greater production in "> 4 yr" Years_Since_Install compared with all other categories of Years_Since_Install.

The main effects of Robot_Free_Time are contrary to what would be expected. The low number of observations with greater than 15% Robot_Free_Time in 2011 and 2012 increases the uncertainty of the estimates and the data only represent relatively few farms, which increases the chance of a single farm creating bias in the estimates. Therefore, the years 2013 and 2014 should be examined as a more reliable representation of the differences among the Robot_Free_Time levels compared with 2011 and 2012. A likely reason for Milk_Production_per_Robot_per_Day increasing with greater Robot_Free_Time is the existence of an unknown confounder not included in the model. One such confounding effect might be automatic cleaning time. Farms trying to increase total production might decrease the number of times the system is cleaned to increase the time the robot is available for milking cows. Therefore, the robot is available more time per day to milk cows, but it still maintains a greater Robot_Free_Time. Another possible confounder is treatment time (time it takes the robot to clean, prep the udder, and postdip). Because Boxtime is made up of both milking time and treatment time, it is possible to maintain the same Boxtime and Milk_Speed by decreasing treatment time which yields longer milking time and more production. This is an example for a hypothesis-generating relationship that could be evaluated further in the future.

Although the 2 models included an extensive number of risk factors, they did not include every possible cause for variation, confounder, or type of farms. In addition to treatment time and cleaning time, variables such as lactation, milk components, herd size, cow health, robot model, nutrition, and other facility factors such as flooring, stall size, bedding, and ventilation were not part of the available data. The number of cows fetched into the AMS and cost efficiency were not included in the analysis, but could affect a farmer's decision between the 2 traffic systems, although the gain in milk per cow per day in free traffic type is potentially a substantial contributor to profit. Random effect of Farm_ID accounts for some differences between these unknown factors

including the level of genetic value in herds. Grazing and organic herds were not included as their analysis might require different risk factors examined compared with the current cohort of farms. Including additional interactions in the model would most likely improve the model fit, but would also make interpretation more difficult.

The current study, looking at differences in milk production across farms utilizing AMS, has benefitted from a large sample of farms in diverse locations applying very different management strategies, breeds, and additional associated risk factors. The resulting associations are relatively robust and should contribute to the benchmarking processes in the industry. The results could be used to advise the benchmarks regarding AMS farms, a subject that has not been fully addressed in literature where benchmarks are generated across AMS and conventional milking technology on farms alike. It remains to be seen whether AMS farms need their own set of benchmarks and decision support structures, because of their specialized production systems. The analysis of large cohorts of AMS farms for the association of risk factors with milk production is a step in the right direction of customizing advice for AMS farms based on their individual management characteristics that could cluster them into distinct groups.

ACKNOWLEDGMENTS

We thank Lely North America (Pella, IA) for their financial support of this study. We gratefully acknowledge the editing effort of Nick Robl and the expertise of Rik van der Tol (Lely Industries N.V., Maassluis, the Netherlands) for his assistance in the planning of this project.

REFERENCES

Bach, A., M. Devant, C. Igleasias, and A. Ferrer. 2009. Forced traffic in automatic milking systems effectively reduces the need to get cows, but alters eating behavior and does not improve milk yield of dairy cattle. J. Dairy Sci. 92:1272–1280. http://dx.doi.org/10.3168/jds.2008-1443.

De Marchi, M., G. Bittante, R. Dal Zotto, C. Dalvit, and M. Cassandro. 2008. Effect of Holstein Friesian and Brown Swiss breeds on quality of milk and cheese. J. Dairy Sci. 91:4092–4102. http://dx.doi.org/10.3168/jds.2007-0788.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. García Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach. 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27–46. http://dx.doi.org/10.1111/j.1600-0587.2012.07348.x.

Fox, J. 2003. Effect displays in R for generalised linear models. J. Stat.Softw. 8:1–27. http://www.jstatsoft.org/v08/i15/.

Grant, R. J., and J. L. Albright. 2001. Effect of animal grouping on feeding behavior and intake of dairy cattle. J. Dairy Sci. 84(E-Suppl.):E156–163. http://dx.doi.org/10.3168/jds.S0022-0302(01)70210-X.

Greenland, S. 1989. Modeling and variable selection in epidemiologic analysis. Am. J. Public Health 79:340–349.

Gygax, L., I. Neuffer, C. Kaufmann, R. Hauser, and B. Wechsler. 2007. Comparison of functional aspects in two automatic milking sys tems and auto-tandem milking parlors. J. Dairy Sci. 90:4265–4274. http://dx.doi.org/10.3168/jds.2007-0126.

Harms, J., G. Wendl, and H. Schön. 2002. Influence of cow traffic on milking and animal behaviour in a robotic milking system. Pages II8–II14 in First North Am. Conf. Robotic Milking. J. McLean, M. Sinclair, and B. West, ed. Wageningen Press, Wageningen, the Netherlands.

Hermans, G. G. N., A. H. Ipema, J. Stefanowska, and J. H. M. Metz. 2003. The effect of two traffic situations on the behavior and performance of cows in an automatic milking system. J. Dairy Sci. 86:1997–2004.

Jacobs, J. A., and J. M. Siegford. 2012. Invited review: The impact of automatic milking systems on dairy cow management, behavior, health, and welfare. J. Dairy Sci. 95:2227–2247. http://dx.doi.org/10.3168/jds.2011-4943.

Jago, J., and K. Kerrisk. 2011. Training methods for introducing cows to a pasture-based automatic milking system. Appl. Anim. Behav. Sci. 131:79–85. http://dx.doi.org/10.1016/j.applanim.2011.02.002.

Ketelaar-de Lauwere, C. C., S. Devir, and J. H. M. Metz. 1996. The influence of social hierarchy on the time budget of cows and their visits to an automatic milking system. Appl. Anim. Behav. Sci. 49:199–211. S0168–1591(96)01030–1.

Ketelaar-de Lauwere, C. C., M. M. W. B. Hendriks, J. Zondag, A. H. Ipema, J. H. M. Metz, and J. P. T. M. Noordhuizen. 2000. Influence of routing treatments on cows' visits to an automatic milking system, their time budget and other behaviour. Acta Agric. Scand. Anim. Sci. 50:174–183.

Kozlowska, H., A. Sawa, and W. Neja. 2013. Analysis of the number of cow visits to the milking robot. Acta Scientiarum Polonorum Zootechnica 12:37–47.

Melin, M., G. G. N. Hermans, G. Pettersson, and H. Wiktorsson. 2006. Cow traffic in relation to social rank and motivation of cows in an automatic milking system with control gates and an open waiting area. Appl. Anim. Behav. Sci. 96:201–214.

Melin, M., G. Pettersson, K. Svennersten-Sjaunja, and H. Wiktorsson. 2007. The effects of restricted feed access and social rank on feeding behavior, ruminating and intake for cows managed in automated milking systems. Appl. Anim. Behav. Sci. 107:13–21. http://dx.doi.org/10.1016/j.applanim.2006.09.026.

Munksgaard, L., J. Rushen, A. M. de Passillé, and C. C. Krohn. 2011. Forced versus free traffic in an automated milking system. Livest. Sci. 138:244–250. http://dx.doi.org/10.1016/j.livsci.2010.012.023.

Pasta, D. J. 2011. Those confounded interactions: Building and interpreting a model with many potential confounders and interactions. Paper 347–2011 in Proc. SAS Global Forum 2011. SAS Inst. Inc., Cary, NC.

Quinn, G. P., and M. J. Keough. 2002. Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge, UK.

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rodenburg, J. 2002. Robotic milkers: What, where . . . and how much!!?? Pages 1–18 in Proc. Ohio Dairy Management Conf., Columbus, OH. Ohio State University Extension, Columbus.

Rodenburg, J. 2004. Housing considerations for robotic milking. ASAE Paper No. 044189, American Society of Agricultural and Biological Engineers, St Joseph, MI.

Rodenburg, J. 2012. The impact of robotic milking on milk quality, cow comfort and labor issues. Pages 126–137 in Natl. Mastitis Counc. Annu. Meet. Proc. St. Pete Beach, FL, Natl. Mastitis Counc., Madison, WI.

Stefanowska, J., M. Plavsic, A. H. Ipema, and M. M. W. B. Hendriks. 2000. The effect of omitted milking on the behaviour of cows in the context of cluster attachment failure during automatic milking. Appl. Anim. Behav. Sci. 67:277–291. http://dx.doi.org/10.1016/S0168-1591(00)00087-3.

Telezhenko, E., M. A. G. von Keyserlingk, A. Talebi, and D. M. Weary. 2012. Effect of pen size, group size, and stocking density on activity in freestall-housed dairy cows. J. Dairy Sci. 95:3064–3069. http://dx.doi.org/10.3168/jds.2011-4953.

Thune, R. Ø., A. M. Berggren, L. Gravås, and H. Wiktorsson. 2002. Barn layout and cow traffic to optimize the capacity of an automatic milking system. Pages 45–50 in Proc. 1st North Am. Conf. Robotic Milking, Toronto, Canada. J. McLean, M. Sinclair, and B.

West, ed. Wageningen Pers, Wageningen, the Netherlands.

Winter, A., and J. E. Hillerton. 1995. Behaviour associated with feeding and milking of early lactation cows housed in an experimental automatic milking system. Appl. Anim. Behav. Sci. 46:1–15.

CHAPTER 3- PREDICTION MODELING INTRODUCTION

Goal

Prediction models are used for forecasting outcomes for new observations (i.e., external data) and obtaining measures of uncertainty for the forecasts. Prediction models are used in many realms including economics, business, politics, romantic partner searches, finance and health care. Now that a model's goal is to make predictions, the model's parameters are referred to as predictors.

Assumptions

Some assumptions previously discussed for parameter estimation modeling are relaxed for prediction modeling. For example, addressing multicollinearity and interactions are assumptions that prevent confounded estimates in a parameter estimation model. However, these assumptions are relaxed for prediction modeling since predictive models are not used to interpret associations (Kutner et al., 2005; Shmueli, 2010). Although these assumptions are relaxed for prediction modeling, addressing multicollinearity and interactions might lead to a significant improvement in the models' predictive performance. Moreover, prediction models generally require better model goodness of fit for improved performance compared to parameter estimation models (Shmueli, 2010). Prediction models also require more data partly due to the need for validation (Shmueli, 2010). In addition, when building prediction models one needs to consider that the model is developed with the goal of being applied to future data. Therefore, the predictors included in the model need to be those that will be or are routinely measured and are easily attainable. Finally, prediction models still depend on having predictors that are biologically relevant.

Challenges

Prediction modeling results in many challenges, among which: selecting among the large amount of methodological options available when building the prediction model, preventing model overfitting, variable selection, and using data from different levels of observations such as having the outcome at the population level (i.e., macro level) and the predictors at the individual level (i.e., micro level).

The goal in prediction modeling is to build the best performing model for the purpose of making predictions given the previously mentioned assumptions; however, all of the available choices of methods might perform slightly differently. In parameter estimation modeling even if two models have slightly different goodness of fit measures, most of the time both models estimates will have the same direction and order of magnitude, interpretation and translation into interventions. But with predictive modeling, a slightly better model performance can have significant effects on prediction performance leading to fewer false positives or false negative predictions. Therefore, the current approaches to model selection that rely on subjective preferences and decisions to select methods is prone to introduce bias and can lead to a lesser performing prediction model. In addition, this selection bias can prevent direct comparisons among studies.

The second challenge discussed in this dissertation are prediction models for multilevel data, which is data from different levels of observations. The majority of multilevel modeling is concerned with macro-micro modeling (Bennink, 2014). This is very commonly addressed with the addition of random effects in the models (Dohoo et al., 2003; Pinheiro and Bates, 2006). The lesser known multilevel modeling variation is the micro-macro model, which is faced with the outcome variable being at the population level (i.e., macro level) and the predictors being at the individual level (i.e., micro level). Although, the ideal situation would be to obtain data at the same level of observation, the cost or time to collect such data can be prohibitive. In addition, there are some outcomes that cannot be measured at the same level as the micro level observations. For example, when modeling the presence or absence of an outbreak of a specific disease per geographic region, data at the individual level of observations cannot be associated with an outcome value of "yes" or "no". This type of outcome variable is only associated with the macro level. There are methods available to address this type of problem called "micro-macro multilevel models" (Croon and Veldhoven, 2007). However, the current version of this method assumes that the central tendency of the micro level data (e.g., the mean) is a good representation of changes in the outcome variable at the macro level (Bennink, 2014). However, sometimes the extreme cases at the micro level are the best representation of the changes in the outcome variable at the macro level. For example, extreme cases of a variable quantifying international travel at the individual level might prove to be the best risk factor for a geographic region's disease outbreak outcome variable. Therefore, there is a need for a micro-macro multilevel modeling technique that accounts for the possible importance of extreme values at the micro level used to predict macro level outcomes.

Approaches to data imperfections

The challenge of selecting among the large amount of methodological options available when building a prediction model was addressed in Chapter 3.2. This was accomplished with regression tree full model selection (rtFMS). As stated in objective 2 (on page 9 in the introduction), the rtFMS method provides a method that systematically combines, compares and selects the most appropriate statistical methods for prediction modeling. When only two models are compared, the small sample size and their confidence intervals result in not enough statistical power to find significant differences between the models. But by building models for many combinations of options we gain the sample size and statistical power to find significant differences among method options. **The rtFMS method removed selection bias and will enable the selection of the best performing model by comparing all available options and combinations of method options. This method was built with the intention of becoming part of a future automated full model selection process that will help remove selection bias from prediction modeling as it selects the best performing model.**

The challenge of multilevel modeling when the central tendency of the micro level observations is not a good representation of the macro level outcome was addressed in Chapter 3.3. Chapter 3.3 described using extreme values of micro level observations to predict macro level outcomes. This method is called extreme value micro-macro (EVMM) multilevel modeling. The two micro-macro

modeling methods were compared using rtFMS analysis. In doing so, the rtFMS method illustrated the significantly better performance of the EVMM method compared to the current micro-macro multilevel modeling method that uses the mean of the micro level observations. **The EVMM method will allow more secondary data sets that combine data from multiple levels to be used to build better performing prediction models and could also be used for micro-macro parameter estimation models.**

References

Bennink, M., 2014. Micro-macro multilevel analysis for discrete data. Tilburg university.

Croon, M.A. and van Veldhoven, M.J., 2007. Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. Psychological methods, 12(1), p.45.

Dohoo, I.R., Martin, W. and Stryhn, H., 2003. Veterinary epidemiologic research (No. V413 DOHv). Charlottetown, Canada: AVC Incorporated.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W., 2005. Applied linear statistical models (Vol. 103). Boston: McGraw-Hill Irwin.

Pinheiro, J. and Bates, D., 2006. Mixed-effects models in S and S-PLUS. Springer Science & Business Media.

Shmueli, G., 2010. To explain or to predict?. Statistical science, 25(3), pp.289-310.

CHAPTER 3.1

Prediction Model Optimization using Full Model Selection with Regression Trees Demonstrated with FTIR Data from Bovine Milk

Preventive veterinary medicine, 163, pp.14-23.
https://doi.org/10.1016/j.prevetmed.2018.12.012

M. Tremblay,* M. Kammer,† H. Lange,‡# S. Plattner,‡# C. Baumgartner,‡ J.A. Stegeman,§ J. Duda,† R. Mansfeld,# D. Döpfer*

* Department of Medical Science, School of Veterinary Medicine, University of Wisconsin, 2015 Linden Dr., Madison 53706, United States of America

† LKV Bayern e.V., Landsberger Straße 282, 80687 München, Germany

‡ Milchprüfring Bayern e.V., Hochstatt 2, 85283 Wolnzach, Germany

§ Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, PO Box 80151, 3508 TD Utrecht, the Netherlands

Clinic for Ruminants with Ambulatory and Herd Health Services, Ludwig-Maximilians-Universität München, Sonnenstr. 16, D-85764 Oberschleissheim, Germany

ABSTRACT
Data preprocessing options and model algorithms are commonly selected empirically in epidemiological studies even though these decisions can significantly affect model performance. Accordingly, full model selection (**FMS**) methods were developed to provide a systematic approach to select predictive modeling methods; however, current limitations of FMS, such as its dependency on user-selected hyperparameters, have prevented their routine incorporation into analyses for model performance optimization.

Here we present the use of regression trees as an innovative method to apply FMS. Regression tree FMS (rtFMS) constructs a model for every combination of predictive modeling method options under consideration. The iterated, cross-validation performances of these models are then passed through a regression tree for selection of a final model. We demonstrate the benefits of rtFMS using a milk Fourier transform infrared spectroscopy dataset, wherein we build prediction models for two blood metabolic health parameters in dairy cows, nonesterified fatty acids (NEFA) and β-hydroxybutyrate acid (BHBA).

In contrast to previously reported FMS methods, rtFMS is not a black box, is simple to implement and interpret, it does not have hyperparameters, and it illustrates the relative importance of modeling options. Additionally, rtFMS allows for indirect comparisons among models developed using different datasets. Finally, rtFMS eliminates user bias due to personal preference for certain methods and rtFMS removes the dependency on published comparisons of methods. Thus, rtFMS provides clear benefits over the empirical selection of data preprocessing options and model algorithms.

INTRODUCTION
Currently, empirical selection is the standard method to select among predictive modeling method options including different preprocessing techniques, and model algorithms (Harrell et al., 1996; Kuhn and Johnson, 2013); however, these options and the order of decisions about predictive modeling methods can significantly influence model performance (Han et al., 2011; Horn et al., 2018; Rinnan, 2014; Shi and Yu, 2017; Weissenbacher et al., 2009). Consequently, full model selection (**FMS**) was developed to provide a systematic approach to eliminate bias in selecting predictive modeling method options for machine learning (Escalante et al., 2009). In short, FMS builds models for every combination of modeling methods under consideration (i.e., options), followed by comparisons of iterated cross-validated performances which yields a final optimized model. This system has been implemented in machine learning, but has largely been overlooked in predictive modeling in applied epidemiology. Applied epidemiology might benefit from incorporating FMS.

Current FMS methods in machine learning use evolutionary algorithms and swarm intelligence algorithms, most notably particle swarm optimization (PSO) (Escalante et al., 2009). PSO is a black box method in which one final model is selected. However, the options' influence on making this selection is not visible to the user. In addition, PSO has hyperparameters, those are parameters of a prior distribution (e.g. inertia weight, acceleration coefficients, velocity clamping), that can change the output and be difficult to select, and these methods are not easily applied without advanced machine learning experience. These facts have slowed the incorporation of FMS into applied epidemiology.

In this paper, we describe the use of regression trees as an innovative FMS method (rtFMS) in applied epidemiology predictive modeling during supervised learning. Regression trees use recursive partitioning to repeatedly separate data into subsets most distant from each other, which results in a decision tree (Hothorn et al., 2017). We propose the use of regression trees for the separation of model performance measures to optimize a final combination of options that results in the best final model. Unlike PSO, rtFMS is not a black box, it is easy to implement, does not have hyperparameters, and is straightforward to interpret.

Our objective was to illustrate that rtFMS results in an optimized prediction model and in addition provides the following information and benefits: (i) rtFMS illustrates the relative importance of modeling options. (ii) It allows indirect comparisons among models that were fit to different datasets by examining their terminal node location in the regression tree. (iii) rtFMS allows for the comparison of a much larger number of preprocessing and model algorithm options simultaneously than would be feasible without FMS. (iv) Finally, it also removes user bias due to familiarity or personal preference for certain methods on prediction performance.

Our aim was to demonstrate rtFMS by applying it to a Fourier-transform infrared spectra (FTIR) dataset to optimize prediction models, because spectrometry research has many preprocessing options that are commonly chosen empirically (Belay et al., 2017; Botelho et al., 2015; Dehareng et al., 2012; Etzion et al., 2004).

MATERIALS AND METHODS

A methods overview is available in Table 1. All data analyses were done in R 3.4.2 (R Development Core Team, 2017).

Step 1 Data preparation:
The FMS approach is described using the example of a data set previously reported by Tremblay et al. (2018). Briefly, a total of 1505 observations were collected from 381 predominantly Simmental cows located on 26 Bavarian robotic milking farms. Farm and cow identification numbers, date, days in milk, breed, lactation number, and milk production records were collected. Blood samples were collected, and nonesterified fatty acids (NEFA) and β-hydroxybutyrate acid (BHBA) were measured using the Cobas c311-Analyzer (Roche Diagnostics, Rotkreuz, Switzerland). Milk fat and protein percent, urea, and lactose were measured using the MilkoScan FT-6000 (FOSS GmbH, Hamburg, Germany), and milk somatic cell count was determined using the Fossomatic 5000 (FOSS GmbH, Hamburg, Germany). The milk samples were subjected to Fourier Transform Infrared (FTIR) spectrometry on 12 different MilkoScan FT-6000 machines calibrated using FOSS standards (FOSS GmbH, Hamburg, Germany). Fatty acid predictions were produced by Qlip N.V. (Leusden, the Netherlands). FTIR data was used to produce FOSS's ketosis screening tool predictions of BHBA and acetate (FOSS Analytical A/S, Hillerød, Denmark).

Table 1. Overview of regression tree full model selection (rtFMS) methods

Step	Method
1	**Data preparation:** A dataset is prepared by formatting variables, and removing outliers, repeats, and errors.
2	**Outcome selection:** An outcome variable is selected according to the project's goal.
3	**Standard methods:** A set of "standard methods", which are methods that will be applied to all models and will not benefit from comparisons of different options, are selected reflective of the data. For example, if a dataset had missing data, applying imputation would be selected as a standard method, although different imputation functions could be compared in step 4.
4	**Comparison categories:** Categories relating prediction model methods and multiple options within each category are selected for comparison. Examples of categories include input data subsets, feature extraction and model algorithms (see Table 2).
5	**Modeling:** A model is run for every combination of options per category described in step 4. All models are run with same fold index for validation.
6	**Performance measure:** A performance measure is selected depending on a dataset's characteristics and the use of the final prediction model. For example, if a dataset is imbalanced, balanced accuracy or kappa coefficient would be the preferred performance of the final model.
7	**Regression tree:** Then the models' performance measures are run through a regression tree to visualize the best combination of options and which selections made significant differences and which order these selections were prioritized. A nonparametric regression tree was selected to be the most inclusive in cases where the outcome variable does not follow a normal distribution (Hothorn et al., 2017).
8	**Final model:** A final model is selected based on the regression tree selections. If there were no significant difference among options, then personal preference can be used and justified in making a selection.

Only the milk sample collected nearest to the time of the blood collection within the previous 24 hours was used. Therefore, 254 observations representing earlier milk samples were removed. Thirty observations were removed due to missing milk production data from the robot. The final dataset for the BHBA model contained 1035 observations. One observation had a missing NEFA value leaving 1034 observations in the final dataset for the NEFA outcome.

Step 2

Outcome selection NEFA: Blood NEFA $\geq$ 0.7 mmol/L served as the case definition for the prediction models (Andrews et al., 2008; Tremblay et al., 2018). This would allow the detection of poor metabolic adaptation syndrome (PMAS) and also conditions such as displaced abomasum (NEFA $\geq$ 1.0 mmol/L) and ketosis (> 1.5 mmol/L) where cows are off-feed or have decreased feed intake, and increased fat mobilization (Andrews et al., 2008; LeBlanc et al., 2005; Tremblay et al., 2018). In the final dataset, 210 observations had blood NEFA $\geq$ 0.7 mmol/L, and 824 observations had blood NEFA < 0.7 mmol/L.

Outcome selection BHBA: Blood BHBA $\geq$ 1.2 mmol/L was used as the case definition for the prediction models (McArt et al., 2012; Overton et al., 2017; Suthar et al., 2013). In the final dataset,

105 observations had blood BHBA ≥ 1.2 mmol/L, and 930 observations had blood BHBA < 1.2 mmol/L.

Step 3 Standard methods: Standard methods are methods that will be applied to all models and will not benefit from comparisons of different options as done in step 4. For the data presented here, standard methods included removing wavenumbers representing water: the O-H bending region 1615-1692 cm^{-1}, the O-H stretching region 3057-3689 cm^{-1} (Afseth et al., 2010). Also, observations were flagged for potential FTIR equipment errors if they did not have a max absorbance within the instrument's working range of 0.1-1.0 absorbance units (Beleites and Sergo, 2017). No error observations were identified in this dataset. Variables with zero or near zero variances needed to be removed from the analysis, but none were present in this dataset (Kuhn and Johnson, 2013).

The groupKFold function within the caret library was used to make sure cross-validation folds were split by farm (Kuhn, 2008). Out of 1034 total observations, cross-validation folds for the NEFA model averaged 930.6 (SD 23.6) observations in the training sets. Out of 1035 total observations, cross-validation folds for the BHBA model averaged 931.5 (SD 24.3) observations in the training sets. Since the variables were on different scales, auto-scaling was used to obtain zero mean values and standard deviations equal to one (Gelman and Hill, 2006).

Our datasets were faced with class imbalance due to the low prevalence within the outcome classes, only 20.3% and 10.1% of observations being in the NEFA and BHBA minority class, respectively (He and Ma, 2013). To address the class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was used to balance the number of observations in the training set (Chawla et al., 2002). The minority classes were over-sampled by 200% as suggested by Chawla et al. (2002), and the majority classes were under-sampled by 150% to obtain a one-to-one ratio between the majority and minority classes' observations.

Step 4

All of the options per categories of predictive modeling methods are listed in Table 2. Categories of predictive modeling methods were separated into 3 areas: (4.1) input subsets, (4.2) preprocessing methods, and (4.3) algorithms (Table 2).

4.1. Input subset

The milk data subset category (4.1.A) includes 4 options: the component (COMP), FTIR, fatty acid predictions (FA), and FOSS's ketosis screening tool predictions (FOSS) subset. The selection of these 4 options was guided by how milk data are generated and their availability for future model application; Milk testing agencies and automatic milking systems generate the COMP subset, the milk analyzers produce the FTIR data, and Qlip N.V. (Leusden, the Netherlands) and FOSS Analytical A/S (Hillerød, Denmark) calibration models using FTIR data produce the FA and FOSS subsets, respectively.

Table 2. Options per corresponding category and area (step 4) selected for comparison using regression tree full model selection when applied to a milk Fourier transform infrared spectroscopy dataset to build prediction models for two blood metabolic health parameters in dairy cows: nonesterified fatty acids (NEFA) and β-hydroxybutyrate acid (BHBA)

Area	Category	Options
1. Input Subset	A. Milk Data Subset	Component (COMP) Fatty acid predictions (FA) Fourier transform infrared spectroscopy (FTIR) FOSS's ketosis screening tool predictions (FOSS)*
	B. Cow Information	Include cow information (+CowInfo) Exclusion of cow information (-CowInfo)
2. Pre-processing	A. Standardization	Raw absorbance values (Raw-FTIR) 1st derivative (FD) 2nd derivative (SD)
	B. Feature Extraction	Performed a PCA (+PCA) Did not perform a PCA (-PCA)
	C. Wavenumber Subset	Removed "no-information" wavenumbers (excl.WN) Included all wavenumbers (AllWN)
	D. High Correlation	Removed highly correlated predictors (excl.HighCorr) Did not remove highly correlated predictors (incl.HighCorr)
3. Algorithm	Algorithm	generalized linear models (GLM) lasso and elastic-net regularized generalized linear models (GLMNET) linear discriminant analysis (LDA) linear support vector machines (SVM) nearest neighbor methods (KNN) naive Bayes (NB) classification trees (RPART) neural networks (NNET) gradient boosting machine (GBM) random forests (RF) multivariate adaptive regression splines (MARS)

* BHBA models only

The COMP subset included the following variables: milk fat (%), protein (%), urea (mg/dL), somatic cell count (1000 cells/mL), and lactose (%). FTIR subset included 874 wavenumbers between 925.2-1611.39 cm^{-1}, 1696.2-3053.16 cm^{-1}, and 3693.09-5007.645 cm^{-1}. The FA subset included prediction of blood BHBA (mmol/L), three ratios of blood fatty acids [C16:0/C16:1, (C16:0 + C18:0)/C18:1, and C18:0/C18:1], and the following blood fatty acids measured in μmol/L: C13:0, C14:0, C14:1, C15:0 result 1, C15:0 result 2, C15:0_total, C16:0, C16:1, C16:2, C17:0, C17:1 result 1, C17:1 result 2, C17:1_total, C17 to C24, C18:0, C18:1, C18:2, C18:3, C18:4, C19:0, C19:1 result 1, C19:1 result 2, C19:1_total, C19:2, C19:3, C20:0, C20:1, C20:2, C20:3, C20:4, C20:5, C21:0, C21:1, C21:3, C21:4, C22:0, C22:1, C22:3, C22:4, C22:5, C22:6, C23:0, C23:1, C24:0, C24:1, C24:5, C24:6, C25:0, C25:1 result 1, C25:1 result 2, C25:1_total,

C25_3, C25:5, C26:0, C26:1, C26:2, C27:0, C27:1, C27:3, C28:0, C28:1, C29:0, C29:4 result 1, C29:4 result 2, C29:4_total, C29:6 result 1, C29:6 result 2, C29:6_total, C30:1, total long chain (≥ C18), mono-unsaturated, poly-unsaturated, total saturated, total NEFA, and total unsaturated. The FOSS subset included predictions of milk BHBA (mmol/L) and milk acetone (mmol/L).

Finally, including (+CowInfo) or not including cow information (-CowInfo) were compared within the cow information category (4.1.B). Cow information includes the following: days in milk (DIM), milk production (kg/day), and lactation number.

4.2. Preprocessing

A standardization method is necessary to adjust for instrumental differences since the data in this dataset come from 12 different machines, and the goal is to apply the model to external data that will also be from different machines calibrated at different times. The standardization category (4.2.A) compared the raw absorbance FTIR values (raw-IR) with two baseline corrections: first derivative (FD) and second derivative transformations (SD) (Baker et al., 2014; Beleites and Sergo, 2012; Duckworth, 2004; Smith et al., 2018). The FD is very effective for removing baseline offset and SD is very effective for both the baseline offset and linear trends in spectra (Duckworth, 2004; Rinnan, 2014).

Three categories of dimension reduction methods, also a pre-processing method, were included for comparision using FMS as part of pre-processing: feature extraction (4.2.A), wavenumber subset (4.2.B), and high correlation (4.2.C).

A feature extraction category (4.2.B) compared performing principal component analysis (PCA) (+PCA) or not performing PCA (-PCA). When PCA was applied, the number of components representing 95% of the features' variance were selected (Kuhn and Johnson, 2013). As a feature selection step, a wavenumber subset category (4.2.C) was used to compare performance with (AllWN) or without wavenumber variables (excl.no-infoWN) that are thought to not represent any information ("no information regions"). These are regions from 1800.285 cm^{-1} to 2798.73 cm^{-1} and 3693.09 cm^{-1} to 5007.645 cm^{-1} (Andersen et al., 2002; Dagnachew et al., 2013; Iñón et al., 2004). A high correlation category (4.2.D) was also included that compared including (incl.HighCorr) or excluding highly correlated variables (excl.HighCorr). A high correlation filter was applied using the findCorrelation function within caret (Kuhn, 2008) with a tolerance set to 0.1 (limit at 0.9), which corresponds to a VIF of 10 (Hair, 2007).

4.3. Algorithms

The *algorithm* category included 11 algorithms to compare: generalized linear models (GLM), lasso and elastic-net regularized generalized linear models (GLMNET), linear discriminant analysis (LDA), linear support vector machines (SVM), nearest neighbor methods (KNN), naive Bayes (NB), classification trees (RPART), neural networks (NNET), gradient boosting machine (GBM), random forests (RF), and multivariate adaptive regression splines (MARS). These algorithms were run using the caret model methods "glm", "glmnet", "lda", "svmLinear", "knn", "nb", "rpart", "nnet", "gbm", "rf", and "earth", respectively (Kuhn, 2008). A random grid search with a tune length = 10 was applied for hyperparameter tuning related to the algorithms (and not the selection method) for 7 out of 11 model algorithm: GLMNET, SVM, GBM, NB, RF, NNET,

KNN. (Bergstra and Bengio, 2012; Kuhn, 2008). The default convergence criteria for each model algorithm were used (Kuhn, 2008).

Step 5 & 6. A total of 660 and 704 models for NEFA and BHBA models, respectively, were run for every combination of options per category described in step 4 (Table 2). We performed 10 repeated iterations of 10-fold cross-validation (Bali and Sarkar, 2016). Balanced accuracy was the selected performance parameter because it performs well when the data sets are imbalanced (Japkowicz and Stephen, 2002). See Table 3 for a list of possible performance measures that were available. The average of the 100 cross-validation folds' balanced accuracies were used as the models' point estimate to be used in the regression tree.

Table 3. Final models' performance measures with 95% confidence intervals

	Blood nonesterified fatty acids final model		Blood β-hydroxybutyrate acid final model	
Performance measure	estimate	95% CI	estimate	95% CI
Apparent prevalence, %	33.7	(30.8-36.6)	29.2	(26.4-32.1)
True prevalence, %	20.3	(17.9-22.9)	10.1	(8.4-12.1)
Sensitivity, %	77.1	(70.9-82.6)	84.8	(76.4-91)
Specificity, %	77.4	(74.4-80.2)	77.1	(74.3-79.8)
Diagnostic accuracy, %	77.4	(74.7-79.9)	77.9	(75.2-80.4)
Balanced accuracy, %	77.3	(72.6-81.4)	80.9	(75.3-85.4)
Positive predictive value, %	46.6	(41.2-51.9)	29.5	(24.4-35)
Negative predictive value, %	93.0	(90.8-94.8)	97.8	(96.5-98.7)
Likelihood ratio of a positive test	3.42	(2.95-3.96)	3.70	(3.21-4.27)
Likelihood ratio of a negative test	0.295	(0.230-0.380)	0.198	(0.126-0.311)
Kappa	0.438	(0.381-0.496)	0.338	(0.288-0.388)

CI= confidence interval

Step 7. The models' performance measures were run through a nonparametric regression tree (Equation 1), available through the party R package, using equation 1 (Hothorn et al., 2017).

Equation 1. *Balanced Accuracy ~ Milk Data Subset + Cow Information + Standardization + Feature Extraction + Wavenumber Subset + High Correlation + Algorithm*

A bonferonni correction for multiple comparisons of means was applied. A p-value of 0.05 was used as the limit to visualize branching.

Step 8. The regression tree was inspected to locate the terminal node with the best performance, i.e. highest balance accuracy. The decision nodes leading to the best performing terminal node were described. If a category did not have an option selected by the regression tree then personal preference was justified in making those decisions since they would not make a statistically significant difference in model performance. The selected final model was applied to the entire original dataset for final performance measures and measures of uncertainty. The 20 most influential predictors were extracted and ranked for each final model using the varImp function in

caret (Kuhn, 2008). Their importances were scaled to 100 so that the most influential predictor had a value of 100 and the least influential had a value near zero.

RESULTS

NEFA

Step 5 & 6. Nine NEFA models did not converge according to each model's convergence criteria (Kuhn, 2008). The remaining 651 NEFA models had a mean balanced accuracy of 66.48 (SD 4.84).

Step 7. The NEFA FMS regression tree had 25 decision nodes and 26 terminal nodes (Figure 1). The average number of models per terminal node was 25.0 (SD 16.4). The 25th terminal node had the highest performance. It contained 8 models with an average balanced accuracy of 74.5 (SD 0.56). The final model was selected after 3 decision nodes. (i) The first decision node selected the FA subset (p-value < 0.001). (ii) The second decision node selected the following model algorithms (p-value < 0.001): GLMNET, SVM and NNET. (iii) The final decision node selected the GLMNET model algorithm (p-value = 0.017) (Figure 1).

If the FA subset had not been available, as is often the case in practice, then the models represented in the fifth terminal node would have resulted in the best performance. It contained 16 models with an average balanced accuracy of 73.20 (SD 0.79). This model would have been selected after 4 decision nodes: (i) the following model algorithms were selected (p-value < 0.001): MARS, GBM, GLMNET, LDA, NNET, SVM. (ii) The derivative-transformed (FD, SD) FTIR input subsets were selected (p-value < 0.001). (iii) The next decision node selected the GLMNET model algorithm (p-value < 0.001). (iv) Finally, not performing a PCA (-PCA) was selected (p-value < 0.001) (Figure 1).

Step 8. Options within the cow information, feature extraction and high correlation categories were not selected by the NEFA FMS regression tree. This leaves room to make these decisions empirically. It was decided to include cow information, to not remove highly correlated variables, and to not perform a PCA. The final NEFA model (options: FA, +CowInfo, -PCA, incl.HighCorr, GLMNET) had a final balanced accuracy of 77.3 (95% CI: 72.6 – 81.4), sensitivity of 77.1 (95% CI: 70.9 – 82.6), specificity of 77.4 (95% CI: 74.4 – 80.2) and diagnostic accuracy of 77.4 (95% CI: 74.7 – 79.9) (Table 3). The final hyperparameter values used in the model were alpha = 0.4 and lambda = 0.0117. The most influential predictors in the final NEFA model were ranked and listed in Table 4.

Figure 1. Blood nonesterified fatty acids (NEFA) rtFMS regression tree results

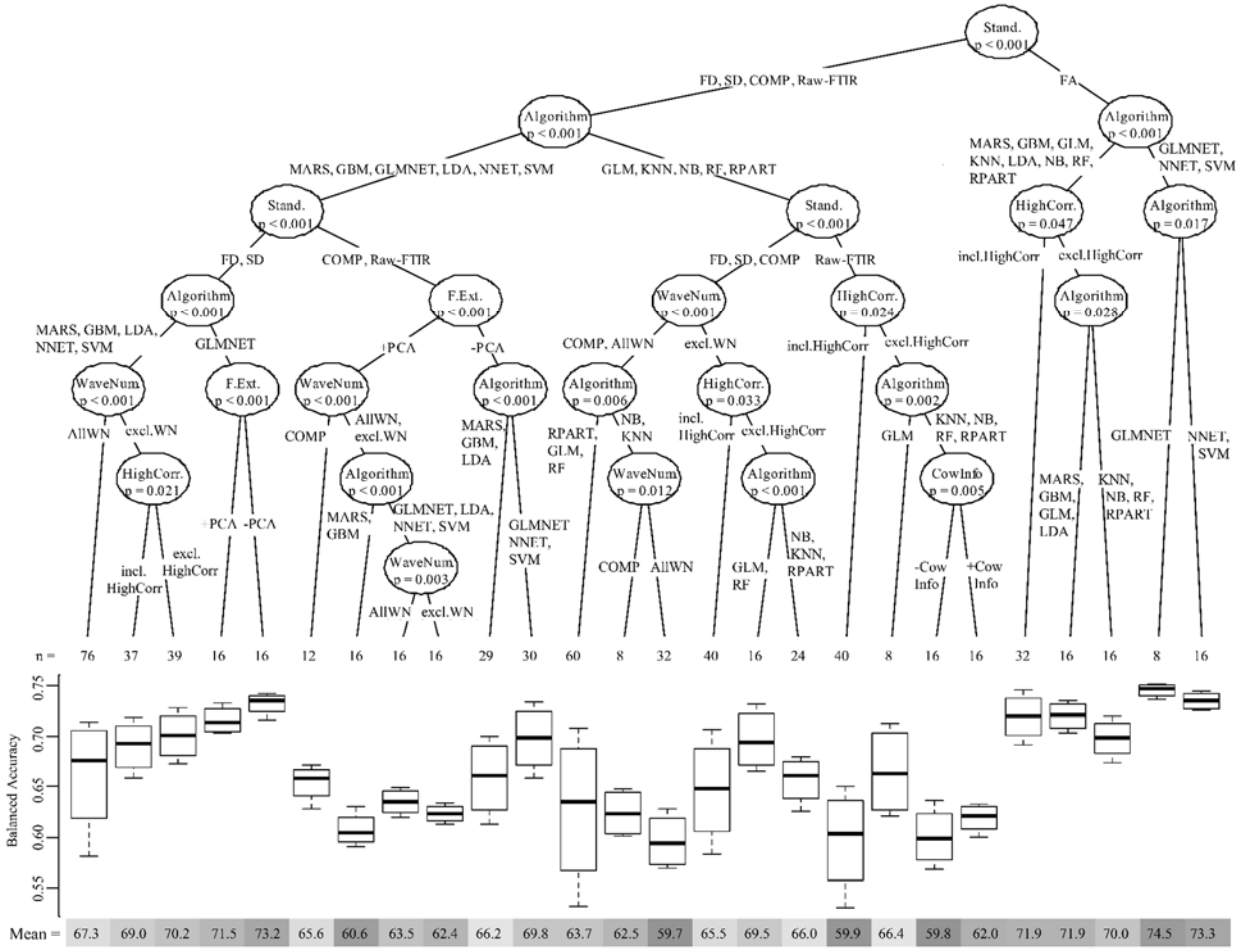

Figure 1 legend: n = number of models in each terminal node. boxplot showing mean and quartile of balanced accuracy of models per terminal model; Subset= Milk Data Subset category; COMP= component; FA= fatty acid predictions; FTIR= Fourier transform infrared spectroscopy; FOSS= FOSS's ketosis screening tool predictions; Cow Info= Cow information category; +CowInfo= Include cow information; -CowInfo= Exclusion of cow information; Stand. = Standardization category; Raw-FTIR= Raw absorbance values; FD= 1st derivative; SD= 2nd derivative; F.Ext.= Feature Extraction category; +PCA= Performed a PCA; -PCA = Did not perform a PCA;WaveNum. = Wavenumber Subset category; excl.WN = Removed "no-information" wavenumbers; AllWN = Included all wavenumbers; HighCorr = High Correlation category; excl.HC= Removed highly correlated predictors; incl.HighCorr= Did not remove highly correlated predictors; Algorithm= Algorithm category; GLM= generalized linear models algorithm; GLMNET= lasso and elastic-net regularized generalized linear models algorithm; LDA= linear discriminant analysis algorithm; LDA= linear discriminant analysis algorithm; SVM= linear support vector machines algorithm; KNN= nearest neighbor methods algorithm; NB= naive Bayes algorithm; RPART = classification trees algorithm; NNET= neural networks algorithm; GBM= gradient boosting machine algorithm; RF= random forests algorithm; MARS= multivariate adaptive regression splines algorithm

Figure 2. Blood β-hydroxybutyrate acid (BHBA) rtFMS regression tree results

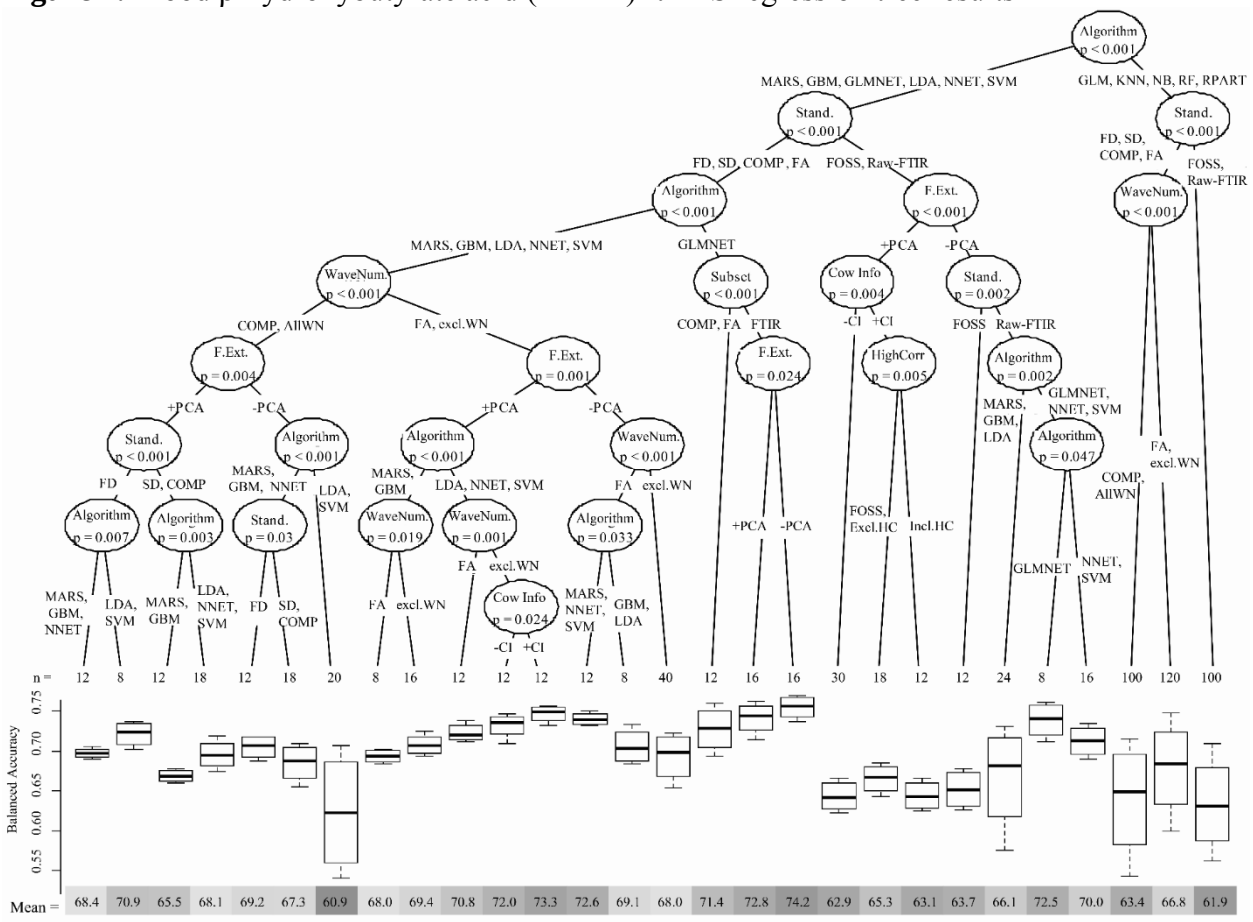

Figure 2 legend: n = number of models in each terminal node. boxplot showing mean and quartile of balanced accuracy of models per terminal model; Subset= Milk Data Subset category; COMP= component; FA= fatty acid predictions; FTIR= Fourier transform infrared spectroscopy; FOSS= FOSS's ketosis screening tool predictions; Cow Info= Cow information category; +CowInfo= Include cow information; -CowInfo= Exclusion of cow information; Stand. = Standardization category; Raw-FTIR= Raw absorbance values; FD= 1st derivative; SD= 2nd derivative; F.Ext.= Feature Extraction category; +PCA= Performed a PCA; -PCA = Did not perform a PCA;WaveNum. = Wavenumber Subset category; excl.WN = Removed "no-information" wavenumbers; AllWN = Included all wavenumbers; HighCorr = High Correlation category; excl.HC= Removed highly correlated predictors; incl.HighCorr= Did not remove highly correlated predictors; Algorithm= Algorithm category; GLM= generalized linear models algorithm; GLMNET= lasso and elastic-net regularized generalized linear models algorithm; LDA= linear discriminant analysis algorithm; LDA= linear discriminant analysis algorithm; SVM= linear support vector machines algorithm; KNN= nearest neighbor methods algorithm; NB= naive Bayes algorithm; RPART = classification trees algorithm; NNET= neural networks algorithm; GBM= gradient boosting machine algorithm; RF= random forests algorithm; MARS= multivariate adaptive regression splines algorithm

Table 4: The 20 most important predictors in the blood nonesterified fatty acids (NEFA) final prediction model (options: FA, +CowInfo, -PCA, incl.HighCorr, GLMNET) and their relative importance

Predictor	Importance*	Predictor	Importance*
C14:0, µmol/L	100	C25:0, µmol/L	22.30
C30:1, µmol/L	83.35	C17:0, µmol/L	18.54
Milk production, kg	71.94	C29:4 result 1, µmol/L	14.81
Lactation number	66.08	C24:5, µmol/L	14.71
C20:4, µmol/L	61.14	C25:1 total, umol/L	11.54
BHBA, mmol/L	54.83	C22:6, µmol/L	11.19
C16:0, µmol/L	47.72	C19:1 total, umol/L	10.56
C22:5, µmol/L	40.42	C15:0 result 1, µmol/L	9.94
Days in milk (DIM)	35.73	C25:3, µmol/L	9.30
C25:1 result 1, µmol/L	24.89	C23:0, µmol/L	9.25

* Importance scaled to 100; FA= fatty acid predictions; +CowInfo= Include cow information; -PCA = Did not perform a PCA; incl.HighCorr= Did not remove highly correlated predictors; GLMNET= lasso and elastic-net regularized generalized linear models algorithm

BHBA

Step 5 & 6. The 704 BHBA models had a mean balanced accuracy of 66.31 (SD 4.58).

Step 7. The BHBA FMS regression tree had 27 decision nodes and 28 terminal nodes (Figure 2). The average number of models per terminal node was 25.1 (SD 29.7). The eighteenth node had the highest performance. It contained 8 models with an average balanced accuracy of 74.2 (SD 1.03). The final model was selected after 5 decision nodes. (i) The first decision node selected the following model algorithms (p-value < 0.001): MARS, GBM, GLMNET, LDA, NNET, SVM. (ii) The second decision node selected the derivative-transformed (FD, SD) FTIR, COMP and FA subset (p-value < 0.001). (iii) The third decision node selected the GLMNET model algorithm (p-value < 0.001). (iv) Next, the derivative-transformed (FD, SD) FTIR subsets were selected over the COMP and FA subsets (p-value < 0.001). (v) Finally, not performing a PCA (-PCA) was selected (p-value < 0.024).

Step 8. Options within the cow information, wavenumber subset and high correlation criteria were not selected by the BHBA FMS regression tree. Therefore, it was appropriate to make these decisions empirically. It was decided to include cow information, to not subset the no-information wavenumbers, and to not remove highly correlated variables. The BHBA FMS regression tree did not discern between the FD and SD FTIR standardizations. Therefore, we empirically decided to select the FD standardization.

The final BHBA model (options: FTIR, +CowInfo, FD, -PCA, incl.HighCorr, AllWN, GLMNET) had a final balanced accuracy of 80.9 (95% CI: 75.3 – 85.4), sensitivity of 84.8 (95% CI: 76.4 – 91.0), specificity of 77.1 (95% CI: 74.3 – 79.8) and diagnostic accuracy of 77.9 (95% CI: 75.2 – 80.4) (Table 3). The final hyperparameter values used in the model were alpha = 0.3 and lambda = 0.0735. The most influential predictors in the final BHBA model were ranked and listed in Table 5.

Table 5: The 20 most important predictors in the blood β-hydroxybutyrate acid (BHBA) final prediction model (options: FTIR, +CowInfo, FD, -PCA, incl.HighCorr, allWN, GLMNET) and their relative importance

Predictor	Importance*	Predictor	Importance*
1549.71 cm^{-1}	100	1125.66 cm^{-1}	29.72
1214.325 cm^{-1}	69.72	4668.405 cm^{-1}	29.59
Lactation number	63.46	1372.38 cm^{-1}	26.45
1333.83 cm^{-1}	49.45	1545.855 cm^{-1}	25.64
2629.11 cm^{-1}	40.93	1299.135 cm^{-1}	24.26
1210.47 cm^{-1}	39.61	4336.875 cm^{-1}	22.30
1491.885 cm^{-1}	35.16	4888.14 cm^{-1}	19.97
Milk production, kg	33.38	2270.595 cm^{-1}	15.85
2043.15 cm^{-1}	30.93	1164.21 cm^{-1}	15.81
4891.995 cm^{-1}	29.96	2764.035 cm^{-1}	15.25

* Importance scaled to 100; FTIR= Fourier transform infrared spectroscopy; +CowInfo= Include cow information; FD= 1st derivative; -PCA = Did not perform a PCA; incl.HighCorr= Did not remove highly correlated predictors; AllWN = Included all wavenumbers; GLMNET= lasso and elastic-net regularized generalized linear models algorithm

DISCUSSION

FMS

Our proposed rtFMS method provides a systematic and unbiased approach to optimizing prediction model performance given many possible options for algorithms and preprocessing methods. Our method demonstrated how different combinations of decisions led to statistically significant differences in model performance. The rtFMS selected different preprocessing options for different model outcomes (NEFA, BHBA) within the same dataset, which illustrates the importance of incorporating this technique into all prediction modeling efforts.

Unlike PSO-FMS, rtFMS does not contain hyperparameters and user-friendliness is further improved by the visual representation of the results. In addition, the rtFMS method provides information about the relative importance of options when selecting the final model selection. The relative location of nodes in the tree reflects the importance of the decision on the performance of the prediction model. Our method also allows indirect comparisons among models developed using different datasets, by examining the terminal node location of different option combinations in a regression tree. The ability to eliminate bias by performing these indirect comparisons is important when teams with different personal preferences and experiences are collaborating. This information is key when developing future study designs, and when determining future exploration of additional modeling methods.

FMS removes user bias due to familiarity and personal preference with regards to certain prediction models methods. In contrast to PSO-FMS, rtFMS allows for empirical decisions when appropriate; however, it removes this source of bias on performance when a significantly superior performing model would be possible. We expect that the benefits and flexibility of rtFMS will accelerate its incorporation into the field of applied epidemiology.

We performed 10 repeated iterations of 10-fold cross-validation to obtain accurate estimates of model performance. This method is most appropriate for small dataset (Chollet, 2017). A one-time hold-out test set could have been used to estimate model performance; however, a hold-out test set is only appropriate for large data sets to assure the test set has enough data to minimize the confidence intervals of performance measures (Chollet, 2017).

rtFMS only depends on being supplied an outcome variable such as a goodness of fit or performance measure; therefore, any type of model can be optimized using rtFMS. rtFMS can be applied to longitudinal models, multinomial models, and even unsupervised learning. When applied to prediction models, the performance measure used as the outcome variable in the regression tree can be selected according to the user's needs regarding performance (e.g. high sensitivity, specificity, accuracy, positive predictive value, or negative predictive value). Categorical performance outcomes could also be accommodated by using classification trees (Therneau et al., 2015).

The current rtFMS method optimized a single performance measure but multiple performance measures could be optimized simultaneously using multi-target regression tree (Aho et al., 2012; Osojnik et al., 2015). In some cases, the ease of applying a model to new data or having a transparent model that is easily interpretable can be just as important as its performance. A user could chose to select a final model based on both optimal accuracy and computational time or interpretability. The resulting performance landscapes of multi-target regression tree reflect the many facets of model preferences, selection and application.

Final models

A general overview of rtFMS findings

Glmnet was consistently one of the best-performing model algorithms. This is most likely because both the FTIR and FA input subsets have many highly correlated variables, which glmnet addresses with the *elastic-net penalty, alpha* (James et al., 2013; Zou and Hastie, 2005). The final selection of a glmnet algorithm is in contrast to Fernandez-Delgado et al. (2014) who found the random forest algorithm performed the best when applied to over 100 datasets. However, this study assumed that preprocessing would affect all algorithms similarly and that algorithms would be ranked similarly for all dataset. The differences between this study and ours suggests that findings from published non-FMS comparisons of model algorithms or preprocessing options cannot be applied to other datasets without FMS comparisons.

Discussion of findings for the FTIR data sets and their application in practice

There are two broad categories of FTIR standardization methods (Wise, 1996; Wise et al., 2007). First, direct standardization uses a transformation that maps the response of a 'slave' FTIR instrument onto a 'master' instrument as done by Grelet et al. (2017). Second, preprocessing standardization applies the same preprocessing methods, such as a derivative transformation, to data from all FTIR instruments such that any shifts and differences in calibrations due to instrumental differences are no longer an issue. For a standardization method to be suitable for this study, it needed to be rapid, simple, outcome dependent, require no additional samples, applicable to data already collected, and applicable to new observations individually without depending on the remaining dataset. Therefore, only standardization by preprocessing was appropriate because

direct standardization methods require many additional samples for mapping, such methods are not continuous over time, and they do not easily expand to data already collected (Feudale et al., 2002). First or 2nd derivative transformations were consistently favored over the raw FTIR data in both models. This suggests that standardization is needed to adjust for changes in calibration over time and differences among instruments.

The next step for the prediction modeling of FTIR data sets is to perform a comparison among more standardization methods including the piece-wise direct standardization method reported by Grelet et al. (Grelet et al., 2017) using rtFMS. In addition, rtFMS could be used to compare SMOTE to other methods of balancing the minority and majority classes of observations.

The component dataset did not perform as well as the FTIR and fatty acid subsets in both the NEFA and BHBA models. This is most likely because fatty acid and FTIR data provide more detailed information, such as wavenumbers representing the fatty acid composition of milk fat, compared to component data that only reports total milk fat. This finding supports that it is necessary to invest in incorporating FTIR data and fatty acid calibrations for routine milk analysis. Selection of a model utilizing the FTIR subset would require farms to produce in-line FTIR measurements. In contrast, the fatty acid prediction subset (FA) would require the additional step of sending FTIR data to Qlip N.V. (Leusden, the Netherlands) to produce the fatty acid panel prior to its use in a prediction model.

NEFA model

The NEFA regression tree selected the fatty acid input subset for the final model. This indicated that the additional calibrations for more than 60 different fatty acids by Qlip NV (Leusden, the Netherlands) improved the information gathered by FTIR. We hypothesize that these calibrations are acting as a targeted feature extraction step. Fatty acids that are synthesized *de novo* from ketones in mammary epithelial cells and are distinguished by the presence of fewer than 16 carbon atoms (Bauman and Davis, 2013). Pre-formed fatty acids on the other hand, have more than 16 carbons and originate from NEFA or lipoproteins in the circulation (Barber et al., 1997; Neville and Picciano, 1997). Thirdly, mixed fatty acids have 16 carbons and can be pre-formed or synthesized *de novo*. Blood NEFA has been found to be highly correlated with milk C18:1 cis-9, and also inversely correlated with the proportion of *de novo* fatty acids in milk (Bell, 1995; Friedrichs et al., 2015; Jorjong et al., 2014). The use of ratios between the different fatty acids in milk has been shown to perform better than measurements of single fatty acid in predicting blood NEFA (Dórea et al., 2017; Mann et al., 2016). These findings support our results that the most important predictors in our NEFA model represent all types of fatty acids including *de novo*, preformed, and mixed fatty acids.

BHBA model

Some of the most important predictors in the final BHBA model are located in the acetone region of the FTIR spectra between 1450 and 1200 cm^{-1} (Hansen, 1999; Heuer et al., 2001). This is in line with the previous finding of a high correlation between milk acetone and blood BHBA (Steger et al., 1972). Acetone information was not available in the fatty acid or component datasets, which could explain why the rtFMS selected the FTIR input subset to predict blood BHBA. In addition to acetone, the other highly important predictor of blood BHBA was wavenumber 1542 cm^{-1} that represents milk protein. Other important predictors in the final model were wavenumbers in the

"no information regions" of the spectra. Fatty acids have been shown to increase the baseline of the spectra in the "no information" spectral regions, (e.g. wavenumbers greater than 4000 cm^{-1}) (Grabska et al., 2017). Taken together, this suggests it is necessary to further investigate FTIR patterns in the so-called 'non-information" regions and to study how those regions of the FTIR spectrum relate to milk composition. In contrast, the FTIR wavenumbers associated with milk fat (i.e., 2927, 2862, 1743, 1454 and 1390 cm^{-1}) were not among the important predictors in our model (Socrates, 1980). This finding suggests that fatty acid information is more important than overall fat composition when predicting blood BHBA. These findings will improve the recommendations for cow health and well-being that can be made based on milk testing data in the near future.

Previous BHBA prediction models based on FTIR data used datasets with different breeds of cattle, geographic regions, sampling structure (DIM), and cross-validation methods, wherein direct comparisons were not possible. However, our rtFMS method allows for indirect comparisons of models using different datasets. Our final BHBA model performance measures overlapped or were significantly better than those of previously published prediction models of blood BHBA that used FTIR data including van Knegsel et al. (2010) and Chandler et al. (2017) that used FOSS milk acetone and milk BHBA predictions in their model. The FOSS input subset did not perform as well in our analysis compared to the FA, COMP, and derivative-transformed IR input subsets. We suspect that this is the case because FOSS calibrations were developed for milk BHBA as the outcome variable rather than blood BHBA used for the current analysis, wherein the correlation between milk and blood BHBA varies widely from 0 to 0.88 (Geishauser et al., 1998). Belay (2017), reported a regression prediction model for blood BHBA that applied feature extraction via partial least squares regression (PLS) regression, akin to PCA. Our rtFMS results showed that eliminating feature extraction using PCA yielded a better performing BHBA model. Therefore, we hypothesize that the use of rtFMS would benefit Belay's model performance for these data through the inclusion of additional comparisons of preprocessing and model algorithm options. Most recently, Pralle et al. (2018) compared 3 model algorithms and 2 data inputs subsets to predict blood hyperketonemia (BHBA $\geq$ 1.2 mmol/L). Based on our results, improved predictive performance could be achieved for this dataset by the addition of a derivative transformation of the spectral data and the use of the glmnet algorithm without variable reduction.

Outlook

The benefit and robustness of rtFMS should be evaluated with additional types of data including those with various dimensionalities and different dataset sizes. We foresee additional applications of rtFMS to deep learning networks, which are popular for modeling outcome predictions from large data sets. Indeed, the optimization of deep learning models through multiple iterations of k-fold cross-validation and extensive hyperparameter tuning could benefit from the rtFMS approach described in this manuscript. We recognize that access to the computational capacity necessary to apply FMS can be limited; however, we suspect that this issue will be resolved in the near future. Automation of these methods would also be beneficial for incorporating rtFMS in standard prediction modeling efforts in applied epidemiology.

CONCLUSION

In conclusion, rtFMS will allow for the consistent application of FMS to applied epidemiology to improve and optimize prediction model performance and rtFMS will eliminate the bias associated with empirical selection of method options. Other research areas depending on prediction models

such as diagnostic imaging, spatial analyses, surveillance, single-nucleotide polymorphism, and microbiome analyses will greatly benefit from applying rtFMS. In the future, rtFMS will continue to provide simplicity and structure to FMS during prediction modeling.

ACKNOWLEDGEMENTS

The authors acknowledge the Bayerisches Staatsministerium für Ernährung, Landwirtschaft und Forsten (i.e. the Bavarian Ministry for Nutrition, Agriculture and Forests) for supporting the collection of the data. The project was supported by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme.

REFERENCES

Afseth, N.K., Martens, H., Randby, A., Gidskehaug, L., Narum, B., Jørgensen, K., Lien, S., Kohler, A., 2010. Predicting the fatty acid composition of milk: A comparison of two Fourier transform infrared sampling techniques. Applied spectroscopy 64, 700–707.

Aho, T., Ženko, B., Džeroski, S., Elomaa, T., 2012. Multi-target regression with rule ensembles. Journal of Machine Learning Research 13, 2367–2407.

Andersen, S.K. er, Hansen, P.W., Andersen, H.V., 2002. Vibrational spectroscopy in the analysis of dairy products and wine. Handbook of vibrational spectroscopy.

Andrews, A.H., Blowey, R.W., Boyd, H., Eddy, R.G., 2008. Bovine Medicine: Diseases and Husbandry of Cattle. John Wiley & Sons.

Baker, M.J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H.J., Dorling, K.M., Fielden, P.R., Fogarty, S.W., Fullwood, N.J., Heys, K.A., Hughes, C., Lasch, P., Martin-Hirsch, P.L., Obinaju, B., Sockalingum, G.D., Sulé-Suso, J., Strong, R.J., Walsh, M.J., Wood, B.R., Gardner, P., Martin, F.L., 2014. Using Fourier transform IR spectroscopy to analyze biological materials. Nature Protocols 9, 1771–1791. https://doi.org/10.1038/nprot.2014.110

Bali, R., Sarkar, D., 2016. R Machine Learning By Example. Packt Publishing Ltd.

Barber, M.C., Clegg, R.A., Travers, M.T., Vernon, R.G., 1997. Lipid metabolism in the lactating mammary gland. Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism 1347, 101–126. https://doi.org/10.1016/S0005-2760(97)00079-9

Bauman, D.E., Davis, C.L., 2013. Biosynthesis of milk fat. Lactation: a comprehensive treatise 2, 31–75.

Belay, T.K., Dagnachew, B.S., Kowalski, Z.M., Adnøy, T., 2017. An attempt at predicting blood β-hydroxybutyrate from Fourier-transform mid-infrared spectra of milk using multivariate mixed models in Polish dairy cattle. Journal of dairy science 100, 6312–6326.

Beleites, C., Sergo, V., 2012. hyperSpec: a package to handle hyperspectral data sets in R. Rpackage version 0.98-20120224, J. Stat. Software, http://hyperspec. r-forge. r-project. org, in preparation.

Bell, A.W., 1995. Regulation of organic nutrient metabolism during transition from late pregnancy to early lactation. Journal of animal science 73, 2804–2819.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Journal of Machine Learning Research 13, 281–305.

Botelho, B.G., Reis, N., Oliveira, L.S., Sena, M.M., 2015. Development and analytical validation of a screening method for simultaneous detection of five adulterants in raw milk using mid-infrared spectroscopy and PLS-DA. Food chemistry 181, 31–37.

Chandler, T.L., Pralle, R.S., Dórea, J.R.R., Poock, S.E., Oetzel, G.R., Fourdraine, R.H., White, H.M., 2017. Predicting hyperketonemia by logistic and linear regression using test-day milk and performance variables in early-lactation Holstein and Jersey cows. Journal of dairy science 101, 2476-2491.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. 1 16, 321–357. https://doi.org/10.1613/jair.953

Chollet, F., 2017. Deep learning with python. Manning Publications Co.

Dagnachew, B.S., Kohler, A., Ådnøy, T., 2013. Genetic and environmental information in goat milk Fourier transform infrared spectra. Journal of Dairy Science 96, 3973–3985. https://doi.org/10.3168/jds.2012-5972

Dehareng, F., Delfosse, C., Froidmont, E., Soyeurt, H., Martin, C., Gengler, N., Vanlierde, A., Dardenne, P., 2012. Potential use of milk mid-infrared spectra to predict individual methane emission of dairy cows. Animal 6, 1694–1701.

Dórea, J.R.R., French, E.A., Armentano, L.E., 2017. Use of milk fatty acids to estimate plasma nonesterified fatty acid concentrations as an indicator of animal energy balance. Journal of dairy science 100, 6164–6176.

Duckworth, J., 2004. Mathematical data preprocessing. Near-infrared spectroscopy in agriculture 115–132.

Escalante, H.J., Montes, M., Sucar, L.E., 2009. Particle swarm model selection. Journal of Machine Learning Research 10, 405–440.

Etzion, Y., Linker, R., Cogan, U., Shmulevich, I., 2004. Determination of protein concentration in raw milk by mid-infrared Fourier transform infrared/attenuated total reflectance spectroscopy. Journal of dairy science 87, 2779–2788.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems. J. Mach. Learn. Res 15, 3133–3181.

Feudale, R.N., Woody, N.A., Tan, H., Myles, A.J., Brown, S.D., Ferré, J., 2002. Transfer of multivariate calibration models: a review. Chemometrics and Intelligent Laboratory Systems 64, 181–192. https://doi.org/10.1016/S0169-7439(02)00085-0

Friedrichs, P., Bastin, C., Dehareng, F., Wickham, B., Massart, X., 2015. Final OptiMIR Scientific and Expert Meeting: From milk analysis to advisory tools (Palais des Congrès, Namur, Belgium, 16-17 April 2015): optimir-a project aiming the development of novel mid-infrared based management tools for dairy herds. Biotechnologie, Agronomie, Société et Environnement 19, 97.

Geishauser, T., Leslie, K., Kelton, D., Duffield, T., 1998. Evaluation of five cowside tests for use with milk to detect subclinical ketosis in dairy cows. Journal of Dairy Science 81, 438–443.

Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Grabska, J., Beć, K.B., Ishigaki, M., Wójcik, M.J., Ozaki, Y., 2017. Spectra-structure correlations of saturated and unsaturated medium-chain fatty acids. Near-infrared and anharmonic DFT study of hexanoic acid and sorbic acid. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 185, 35–44.

Grelet, C., Pierna, J.F., Dardenne, P., Soyeurt, H., Vanlierde, A., Colinet, F., Bastin, C., Gengler, N., Baeten, V., Dehareng, F., 2017. Standardization of milk mid-infrared spectrometers for the transfer and use of multiple models. Journal of dairy science 100, 7910–7921.

Hair, 2007. Multivariate Data Analysis. Pearson Education.

Han, J., Pei, J., Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.

Hansen, P.W., 1999. Screening of dairy cows for ketosis by use of infrared spectroscopy and multivariate calibration. Journal of dairy science 82, 2005–2010.

Harrell, F.E., Lee, K.L., Mark, D.B., 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 15, 361–387.

He, H., Ma, Y. (Eds.), 2013. Imbalanced Learning: Foundations, Algorithms, and Applications, 1 edition. ed. Wiley-IEEE Press.

Heuer, C., Luinge, H.J., Lutz, E.T.G., Schukken, Y.H., Van Der Maas, J.H., Wilmink, H., Noordhuizen, J., 2001. Determination of acetone in cow milk by Fourier transform infrared spectroscopy for the detection of subclinical ketosis. Journal of dairy science 84, 575–582.

Horn, B., Esslinger, S., Pfister, M., Fauhl-Hassek, C., Riedl, J., 2018. Non-targeted detection of paprika adulteration using mid-infrared spectroscopy and one-class classification – Is it data preprocessing that makes the performance? Food Chemistry 257, 112–119. https://doi.org/10.1016/j.foodchem.2018.03.007

Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., 2017. party: A Laboratory for Recursive Partytioning.

Iñón, F.A., Garrigues, S., de la Guardia, M., 2004. Nutritional parameters of commercially available milk samples by FTIR and chemometric techniques. Analytica Chimica Acta 513, 401–412. https://doi.org/10.1016/j.aca.2004.03.014

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, Springer Texts in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study. Intelligent data analysis 6, 429–449.

Jorjong, S., Van Knegsel, A.T.M., Verwaeren, J., Lahoz, M.V., Bruckmaier, R.M., De Baets, B., Kemp, B., Fievez, V., 2014. Milk fatty acids as possible biomarkers to early diagnose elevated concentrations of blood plasma nonesterified fatty acids in dairy cows. Journal of dairy science 97, 7054–7064.

Kuhn, M., 2008. Caret package. Journal of statistical software 28, 1–26.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer Science & Business Media.

LeBlanc, S.J., Leslie, K.E., Duffield, T.F., 2005. Metabolic Predictors of Displaced Abomasum in Dairy Cattle. Journal of Dairy Science 88, 159–170. https://doi.org/10.3168/jds.S0022-0302(05)72674-6

Mann, S., Yepes, F.L., Duplessis, M., Wakshlag, J.J., Overton, T.R., Cummings, B.P., Nydam, D.V., 2016. Dry period plane of energy: Effects on glucose tolerance in transition dairy cows. Journal of dairy science 99, 701–717.

McArt, J.A.A., Nydam, D.V., Oetzel, G.R., 2012. Epidemiology of subclinical ketosis in early lactation dairy cattle. Journal of dairy science 95, 5056–5066.

Neville, M.C., Picciano, M.F., 1997. Regulation of milk lipid secretion and composition. Annu. Rev. Nutr. 17, 159–183. https://doi.org/10.1146/annurev.nutr.17.1.159

Osojnik, A., Panov, P., Džeroski, S., 2015. Comparison of tree-based methods for multi-target regression on data streams, in: International Workshop on New Frontiers in Mining Complex Patterns. Springer, pp. 17–31.

Overton, T.R., McArt, J.A.A., Nydam, D.V., 2017. A 100-Year Review: Metabolic health indicators and management of dairy cattle. Journal of dairy science 100, 10398–10417.

Pralle, R.S., Weigel, K.W., White, H.M., 2018. Predicting blood β-hydroxybutyrate using milk Fourier transform infrared spectrum, milk composition, and producer-reported variables with multiple linear regression, partial least squares regression, and artificial neural network. Journal of dairy science 101, 4378–4387.

Rinnan, Å., 2014. Pre-processing in vibrational spectroscopy – when, why and how. Anal. Methods 6, 7124–7129. https://doi.org/10.1039/C3AY42270D

Shi, H., Yu, P., 2017. Comparison of grating-based near-infrared (NIR) and Fourier transform mid-infrared (ATR-FT/MIR) spectroscopy based on spectral preprocessing and wavelength selection for the determination of crude protein and moisture content in wheat. Food Control 82, 57–65. https://doi.org/10.1016/j.foodcont.2017.06.015

Smith, B.R., Baker, M.J., Palmer, D.S., 2018. PRFFECT: A versatile tool for spectroscopists. Chemometrics and Intelligent Laboratory Systems 172, 33–42. https://doi.org/10.1016/j.chemolab.2017.10.024

Socrates, G., 1980. Infrared Characteristic Group Frequencies, J. Wilay and Sons, New York 87.

Steger, H., Girschewski, H., Voigt, G., Piatkowski, B., 1972. Die Beurteilung des Ketosisstatus laktierender Rinder aus der Konzentration der Ketokörper im Blut und des Azetons in der Milch. Archives of Animal Nutrition 22, 157–162.

Suthar, V.S., Canelas-Raposo, J., Deniz, A., Heuwieser, W., 2013. Prevalence of subclinical ketosis and relationships with postpartum diseases in European dairy cows. Journal of dairy science 96, 2925–2938.

Team, R., 2013. R development core team. RA Lang Environ Stat Comput 55, 275–286.

Therneau, T., Atkinson, B., Ripley, B., 2015. rpart: Recursive Partitioning and Regression Trees. R package version 4.1–10.

Tremblay, M., Kammer, M., Lange, H., Plattner, S., Baumgartner, C., Stegeman, J.A., Duda, J., Mansfeld, R., Döpfer, D., 2018. Identifying poor metabolic adaptation during early lactation in dairy cows using cluster analysis. Journal of Dairy Science. https://doi.org/10.3168/jds.2017-13582

Van Knegsel, A.T.M., Van Der Drift, S.G.A., Horneman, M., De Roos, A.P.W., Kemp, B., Graat, E.A.M., 2010. Ketone body concentration in milk determined by Fourier transform infrared spectroscopy: Value for the detection of hyperketonemia in dairy cows. Journal of dairy science 93, 3065–3069.

Weissenbacher, A., Kasess, C., Gerstl, F., Lanzenberger, R., Moser, E., Windischberger, C., 2009. Correlations and anticorrelations in resting-state functional connectivity MRI: A quantitative comparison of preprocessing strategies. NeuroImage 47, 1408–1416. https://doi.org/10.1016/j.neuroimage.2009.05.005

Wise, B.M., 1996. Introduction to instrument standardization and calibration transfer. Shedding New Light on Disease: Optical Diaganostics for a New Millenium, Winnipeg, CA, June 2000.

Wise, B.M., Gallagher, N.B., Bro, R., Shaver, J., Windig, W., Koch, R.S., 2007. PLS Toolbox 4.0. Eigenvector Research Incorporated: Wenatchee, WA, USA.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

CHAPTER 3.2

A Novel Micro-Macro Multilevel Modeling Approach using Extreme Values of Cow Level FT-MIR Spectrometry Data to Predict Antimicrobial Residues in Raw Bulk Milk

M. Tremblay,* M. Kammer,† C. Baumgartner,‡ J.A. Stegeman,§ J. Duda,† D. Döpfer*

* Department of Medical Science, School of Veterinary Medicine, University of Wisconsin, 2015 Linden Dr., Madison 53706, United States of America

† LKV Bayern e.V., Landsberger Straße 282, 80687 München, Germany

‡ Milchprüfring Bayern e.V., Hochstatt 2, 85283 Wolnzach, Germany

§ Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, PO Box 80151, 3508 TD Utrecht, the Netherlands

INTRODUCTION

Antimicrobials are used in the dairy industry to treat a number of infections. Antimicrobial residues can be found in cows' milk during and after receiving such treatments. The most common antimicrobial used in dairy cattle is penicillin, a beta-lactam (Sundlof et al., 1995; USDA, 2008; Andrew et al., 2009; De Briyne et al., 2014). Testing for antimicrobial residues in milk is an established standard in many countries to prevent contaminated milk from entering the food chain. This testing is of great importance for public health, as antimicrobial residues pose health risks to consumers such as the potential for allergic reactions (Beyene, 2016). It is also important to prevent inhibition of bacterial growth needed during the production of fermented milk products (Albright et al., 1961; Kebede et al., 2014).

FTIR

Fourier transform mid-infrared (**FT-MIR**) spectrometry is currently used to estimate milk components, such as fat and protein percentages, during routine raw milk analyses. Routine milk testing using FT-MIR, commonly performed monthly, is done for individual cows by milk testing laboratories. Since FT-MIR data are already being routinely collected, there is the opportunity for their use as a proxy for additional milk characteristics. FT-MIR has not yet been used to predict penicillin antimicrobial residues in milk; However, Sivakesava and Irudayaraj (2002) were able to successfully use FT-MIR to predict tetracycline residues in milk. **The aim of this study was to evaluate the potential use of cow level FT-MIR spectral measurements to predict herds at high risk for having samples positive for antimicrobial residues, specifically penicillin.** However, unlike FT-MIR data being collected at the cow level, testing for antimicrobial residues is performed at the herd level by testing bulk tank milk samples.

Micro-Macro

Using individual level data (micro level) to predict population level outcomes (macro level) requires regression modeling methods called multilevel modeling (Dohoo et al., 2003). Multilevel modeling, or micro-macro models, were developed to analyze structurally hierarchical data (Croon & Van Veldhoven, 2007). The current micro-macro model method uses the mean, or other measures of central tendency, of the micro level observations as predictors in the model for a macro level outcome (Gelman and Hill, 2006 Bennink, 2014). This method will be referred to in this study as the "mean micro-macro" (**MMM**) multilevel modeling method. However, this method is only appropriate when changes in the outcome that is being predicted is associated with changes in the mean of the micro-level observations. In cases such as antimicrobial residues in bulk milk, only one cow with antimicrobial residues in her milk can change the outcome from a negative (no antimicrobial residues present) to a positive outcome (tested positive for antimicrobial residues). However, the FT-MIR value of one cow positive for antimicrobial residues from an entire herd of cows would not likely significantly change the herd's mean value of FT-MIR wavenumbers. Therefore, the FT-MIR characteristics of negative cows would overshadow the possibly unique FT-MIR characteristics of positive cows when cow level data are averaged together in the usual micro-macro models. **Therefore, a micro-macro method is necessary that uses extreme observations for each predictor at the micro level as predictors for the macro level outcome.** According to the authors, the application of extreme values as predictors in micro-macro multilevel models have not been reported in literature.

In this study, the isolation of the unique FT-MIR risk factors of milk samples positive for antimicrobial residues was attempted by assuming that positive samples would have the extreme values, such as the maximum or minimum value, of at least one FT-MIR wavenumber. Therefore, the **first objective** in this study was to develop a new method of micro-macro modeling that would use extreme values in addition to mean values of the micro-level data as predictors in the model for a macro-level outcome. In this study, this method will be referenced to as the **"extreme value micro-macro"** (**EVMM**) multilevel modeling method. It was **hypothesized** that the EVMM multilevel modeling method would perform significantly better compared to the currently available MMM modeling methods.

rtFMS

With the availability of many machine learning algorithms, data mining and preprocessing methods, comparing and selecting a final model optimized for performance in this study needs to be done systematically. Tremblay et al. (2018) illustrated the benefit of using the regression tree full model selection method (rtFMS) to compare and select among different classification algorithms and preprocessing techniques. rtFMS allowed the comparison of many method options simultaneously and systematically to remove user bias when selecting the best final model. Therefore, the **second objective** of this study was to compare and select the best final model from many possible combinations of methods by means of rtFMS.

We **hypothesized** that using a systematic approach, rtFMS, to compare multiple optimized machine learning methods would lead to the development of a meaningful model able to predict herds at high risk of having milk sample positive for antimicrobial residues. Also, it was hypothesized that the comparisons of many modeling options would lead to the discovery of significant differences in performances among those method options.

Impact

A model able to predict herds at high risk for antimicrobial residues in bulk milk using cow level data would benefit producers, testing agencies and public health. It would aid testing agencies in decisions relating to test scheduling and could justify higher or lower frequency sampling and testing for some locations. In addition, such a model that uses routinely generated data would not be cost prohibitive for producers and testing agencies. Finally, an EVMM modeling method able to distinguish extreme values' influence on a macro outcome would improve the analysis of many multilevel data sets in the future where the central tendency values of the micro level observations do not accurately reflect changes in macro level outcomes.

MATERIALS AND METHODS

Data analyses were performed in R 3.5.0 (R Core Team, 2018) using Amazon Web Services (AWS) cloud computing service Amazon Elastic Compute Cloud (Amazon EC2) (http://aws.amazon.com/ec2/). The study used the following R packages: DMwR, MLmetric, party, partykit, glmnet, randomForest, gbm, earth, klaR, epiR, fastICA, caret (Liaw and Wiener, 2002; Weihs et al., 2005; Hothorn et al., 2006; Friedman et al., 2010; Torgo, 2010; Hothorn and Zeileis, 2015; Yachen, 2016; Marchini et al., 2017; Ridgeway, 2017; Kuhn, 2018; Milborrow, 2018; Stevenson et al., 2018).The following functions available within the caret package were used: preprocess, groupKFold, findCorrelation, sbf, trainControl, and train. The functions ctree

and SMOTE were available within the party and DMwR packages, respectively. The rtFMS protocol is described in more detail in Tremblay et al. (2018).

STEP 1 & 2: Data Preparation and Outcome Selection

For 5 years, between July 2013 and August 2018, all bulk milk collected from Bavarian farms were tested for antimicrobial residues (i.e., penicillin) randomly 4 to 6 times per month using a brilliant black reduction test (BRT) (AIM-Analytik in Milch Produktions- und Vertriebs GmbH, Munich, Germany) and cow level FT-MIR data were collected during monthly routine milk testing at those same farms. The current maximum residues limit (MRL) for penicillin is 4 μg kg (Commission Regulation (EU) No. 37/2010, 2010). During that time, a total of 1,165,609 herd level BRT penicillin residues results were collected and a total of 11,025,962 cow level FT-MIR observations were collected from 4824 farms. In this study, the term antimicrobial residues is used interchangeably with penicillin residues.

The herd level antimicrobial residues testing data and cow level FT-MIR data were matched according to two limits: 1) when they occurred within 7 days of each other and 2) when they occurred within 1 day of each other (Table 1). The first dataset included 108 samples positive for antimicrobial residues and 220,002 negative herd level samples when matched within 7 days (7d). The second dataset included 50 samples positive for antimicrobial residues and 99,657 negative herd level samples when matched within 1 day (1d).

Cow level FT-MIR measurements included 1060 wavenumbers between 925.2 nm and 5007.645 nm. The application of a standardization method is necessary to adjust for instrumental and calibration differences since the aim of this study is to develop a model able to be used with external data also challenged with calibration differences. In this study the raw absorbance FT-MIR values (Raw) were compared with standardized FT-MIR data (Table 1). Standardization was accomplished by applying a derivative calculation to the spectra. The first derivative (FD) of spectra helps address baseline offsets and the second derivative transformations (SD) effectively addresses baseline offsets and linear trends in spectra (Duckworth, 2004; Beleites and Sergo, 2012; Baker et al., 2014; Smith et al., 2018; Rinnan, 2014).

We attempted to isolate the unique FT-MIR characteristics of milk samples from cows positive for antimicrobial residues by assuming that a positive cow would be extreme with respect to at least one FT-MIR wavenumber on a farm positive for antimicrobial residues. Therefore, FT-MIR results were summarized for all cows on a farm for each sample by calculating the mean, minimum and maximum of each of the 1060 FT-MIR wavenumbers for each herd per test date. When building MMM models, only the variables representing the means of the FT-MIR wavenumbers were used as predictors. However, when building EVMM models, the variables representing the mean, maximum and minimum values of the FT-MIR wavenumbers were used as predictors together in the same model. The means as predictors were also included in the EVMM models as offsets or contrasts to the extreme values to serve as a baseline correction for the maximum and minimum values.

STEP 3: Standard Methods

Highly correlated variables were removed using the findCorrelation function within the caret package with a tolerance set to 0.1 which is equivalent to a limit of 0.9 (Kuhn, 2018). Data were

centered and scaled (auto-scaling) to obtain mean values of zero and standard deviations of one for all variables (Gelman and Hill, 2006). As described in more detail in Tremblay et al. (2018), 10 repeated iterations of 10-fold cross-validation was performed by specifying method = "repeatedcv", number = 10, and repeats = 10 in the trainControl command within the caret package (Bali and Sarkar, 2016; Kuhn, 2018). In addition, the groupKFold function within the caret package was applied to make sure the models were not biased due to observations from the same farm being included in both the training and test sets in a cross-validation fold (Kuhn, 2018).

Table 1. The predictive modeling method options per corresponding category and area (step 4) selected for comparison using regression tree full model selection when applied to a cow level milk Fourier-transform infrared spectroscopy dataset to build prediction models for herd level penicillin antimicrobial residues in bulk tank milk samples.

Area	Category	Options
4.1. Input data	A. Data Subset	+- 7 days (7d) +- 1 day (1 d)
	B. Micro-Macro Modeling Method	Mean micro macro (MMM) Extreme value micro macro (EVMM)
4.2. Pre-processing	A. Standardization	Raw absorbance values (Raw) 1st derivative (FD) 2nd derivative (SD)
	B. Feature Selection	Sbf univariate filter (SBF) Included all wavenumbers (AllWN)
	C. Feature Extraction	None (none) Performed a PCA (PCA) Performed a ICA (ICA)
	D. Outcome Class Balancing with SMOTE	200% up and 150% down sampling (200) 500% up and 120% down sampling (500)
4.3. Algorithm		generalized linear models (GLM) lasso and elastic-net regularized generalized linear models (GLMNET) linear discriminant analysis (LDA) linear support vector machines (SVM) nearest neighbor methods (KNN) naive Bayes (NB) classification trees (RPART) neural networks (NNET) gradient boosting machine (GBM) random forests (RF) multivariate adaptive regression splines (MARS)

STEP 4: Comparison categories

There were many input data set, preprocessing and algorithm options to compare when building a prediction model for milk samples positive for antimicrobial residues at the herd level using cow level FT-MIR data. All of the predictive modeling method options compared in this study are listed

in Table 1. The predictive modeling method options were separated into 4 main areas: (4.1) input subsets, (4.2) preprocessing methods, and (4.3) algorithms (Table 1).

The input subset section (4.1) compared the use of two different input data sets (4.1.A) and two different modeling methods (4.1.B). The input data set category (4.1.A) compared the previously described 7 days (7d) or 1 day (1d) matched data set. The modeling method category (4.1.B) compared the EVMM and MMM modeling method (Table 1).

The preprocessing section (4.2) included categories for standardization (4.2.A), feature selection (4.2.B), feature extraction (4.2.D), and for data set balancing using SMOTE (4.2.D) (Table 1). As previously described, the three standardization options were compared for possible standardization methods: raw-IR, FD, SD (4.2.A). As a feature selection step, step 4.2.B compared modeling results when all wavenumber where used (AllWN) after removal of highly correlated variables or when using a wavenumber subset selected by the selection by filter (SBF) univariate filter available in the caret package in R (Kuhn and Johnson, 2013). The SBF filter is applied at each cross-validation fold. For each cross-validation fold, the variables are individually evaluated as a predictor for the outcome using a univariate ANOVA model. Only the variables with an ANOVA model p-value less than 0.05 are retained for the multivariable modeling for that specific cross-validation fold (Kuhn and Johnson, 2013). A feature extraction category (4.2.C) compared 3 options: performing principal component analysis (PCA) (+PCA), independent component analysis (ICA) (+ICA), or not performing any feature extraction (none). When applying PCA, the default 95% variance threshold was used to select the number of resulting components (Kuhn and Johnson, 2013). When applying ICA, the number of independent components used was equal to the number of predictors in the dataset (Kuhn and Johnson, 2013; Marchini et al., 2017).

Due to the low prevalence of observations positive for antimicrobial residues, this data set is considered imbalanced. Depending on the classification algorithm used, data sets will be considered imbalanced when the ratio between the two outcome classes reach 1:2 to 1:10 (Sun et al., 2009; He and Ma, 2013). Not only do rare event cause imbalances between the number of observations per outcome classes (i.e., positive and negative) but this is also often associated with a small number of observations in the minority class. Common classification algorithms are not well suited to handle small sample sizes per outcome class and imbalanced data (Sun et al., 2009). However, sampling techniques such as the synthetic minority over-sampling technique (SMOTE) have been used successfully to address these issues (Sun et al., 2009; Batista, 2004). The synthetic minority over-sampling technique (SMOTE) generates new observations of the minority class (positive observations) and under samples the majority class to obtain a balanced dataset for training (Chawla et al., 2002). The balancing step (4.2.B) compared the use of two different amount of up sampling amounts when applying SMOTE. During each iteration of the model training, SMOTE was applied by over-sampling the minority classes by 500% and 200% and under-sampling the majority class by 120% and 150%, respectively. The up and down sampling percentages resulted in one-to-one ratios between the number of observations in the majority and minority classes.

The algorithm section (4.3) included the comparison of the following eleven machine learning algorithms available through the caret library (Kuhn, 2008) were compared: logistic generalized linear models (GLM), lasso and elastic-net regularized generalized linear models (GLMNET),

linear discriminant analysis (LDA), linear support vector machines (SVM), nearest neighbor methods (KNN), naive Bayes (NB), classification trees (RPART), neural networks (NNET), gradient boosting machine (GBM), random forests (RF), and multivariate adaptive regression splines (MARS). These algorithms were run using the following commands in the caret package and default convergence criteria were used: "glm", "glmnet", "lda", "svmLinear", "knn", "nb", "rpart", "nnet", "gbm", "rf", and "earth", respectively (Kuhn, 2018). The hyperparameters for following 7 model algorithms were automatically selected (i.e., fine-tuned) using a grid search: GLMNET, SVM, GBM, NB, RF, NNET, KNN. The grid search compares 10 hyperparameter values that span a meaningful range for each hyperparameter (Bergstra and Bengio, 2012; Kuhn, 2018). A probability threshold of 0.5 was used to differentiate between positive and negative predictions (Hastie et al., 2009).

Step 5 & 6: Modeling and Performance Measure. A model was run for every combination of options described in step 4 and in Table 1. A total of 1584 models were run. Since the goal of this model was to predict a herd's overall risk of having a sample positive for residues and due to the imblanced nature of the data set, balanced accuracy was selected as the performance parameter for the regression tree (Japkowicz and Stephen, 2002). The 100 cross-validation folds' balanced accuracies were average together for each model. This value was used as the models' performance estimate in the regression tree.

Step 7: Regression Tree. As described in step 7 of Tremblay et al. (2018), the models' performance measures were used as the outcome variable in a nonparametric regression tree. The regression tree was building using the ctree function from the party package in R (Hothorn et al., 2006). The categories described in step 4 became the predictors in the regression tree (Equation 1). The factor levels of the variables used in equation 1 are described in Table 1.

Equation 1. Balanced Accuracy ~ Data Subset + Micro-Macro Modeling Method + Standardization + Feature Selection + Feature Extraction + Outcome Class Balancing using SMOTE + Algorithm

The regression tree discovers the independent variable most associated with the dependent outcome variable (i.e., Balanced Accuracy) with a p-value less than 0.05. If such an association exists, the data is split into two according to the selected variable. These subsets are represented by branches and nodes in the regression tree. This is repeated until no more significant differences are found between independent variable and the outcome variable. As described by Tremblay et al. (2018), a bonferonni correction for multiple comparisons of means was used and pruning of the tree was not performed (Hothorn et al., 2006).

Due to the regression tree's large size, only the half of the tree with the better performance was shown. The half of the tree with the lesser performance (i.e., representing the models having used the AllWN option in step 4.2.B) was not included. The resulting regression tree was shown in Figure 1.

Step 8: Final Model. The terminal node with the best performance, i.e. highest balance accuracy was located and the options selected by the regression tree that led to this final node were noted and described. The number of models remaining in the final node are those that were not separated

by significant differences in modeling options listed in Table 1. Personal preference was used to select the final model from the models represented by the final node with the best performance. The selected final model was applied to the entire original dataset for final performance measures and measures of uncertainty. The final model's performance measure were then listed in Table 2.

RESULTS

Step 5-7. The first half of the regression tree (32 terminal nodes) is shown in (Figure 1). This half of the regression tree represents the half with the better performing branches due to the selection of the SBF option (p-value < 0.001). Within this half of the regression tree, the average number of models per terminal node was 23.3 (SD 20.4), and the average balanced accuracy was 54.6% (SD 2.91). The 1st terminal node had the highest performance. It contained 10 models with an average balanced accuracy of 60.9 (SD 0.59). The final model was selected after 6 decision nodes. (i) The first decision node selected the SBF feature selection option over using all wavenumbers (p-value < 0.001) (not shown in Figure 1). (ii) The second decision node selected the following model algorithms (p-value < 0.001): GLM, GLMNET, LDA, NNET, and SVM (p-value < 0.001). (iii) The third decision node selected the ICA feature extraction option over the PCA and no feature extraction options (p-value < 0.001). (iv) The fourth decision node selected the FD option over the SD and Raw standardization options (p-value < 0.001). (v) The fifth decision node selected the EVMM modeling method over the MMM modeling method (p-value < 0.001). (vi) The final decision node selected the 1 day data subset option over the 7 day data subset option (p-value < 0.02) (Figure 1).

Step 8. Options within the Outcome Class Balancing with SMOTE (4.2.D) categories were not selected by the regression tree. No significant differentiation was made among the following model algorithms: GLM, GLMNET, LDA, NNET, and SVM. The final model selected had the following options: 1 day, EVMM, FD, SBF, ICA, 200%, and GLMNET. The initial dataset for this model had 99707 observations and 3177 predictors. Each cross-validation fold had on average 45 (SD 1.9) positive observations and 89691 (SD 465.3) negative observations. Applying a 200% up-sampling and 150% down-sampling SMOTE percentages resulted in a balanced dataset with an average of 270 (SD 11.5) total observations. On average, SBF used 122 (SD 42.4) predictors per fold. The 10 most commonly selected predictors include: the minimum of wavenumbers 1048.56, 1052.415, 1056.27, and 1060.125; the mean of wavenumber 2679.225, 2683.08, 2686.935, 1403.22, and 2675.37; and the maximum of wavenumber 2683.08. The final model cross-validated performances when predicting herd level outcomes are listed in Table 2. The herd level data set had a true prevalence of 0.05%, while the final model had a balanced accuracy of 62.5 (95% CI: 54.9 – 69.3), sensitivity of 62.0 (95% CI: 47.2 – 75.4), and specificity of 62.9 (95% CI: 62.6 – 63.2) (Table 2). The final model's hyperparameter values were alpha = 0.7 and lambda = 0.0001862347.

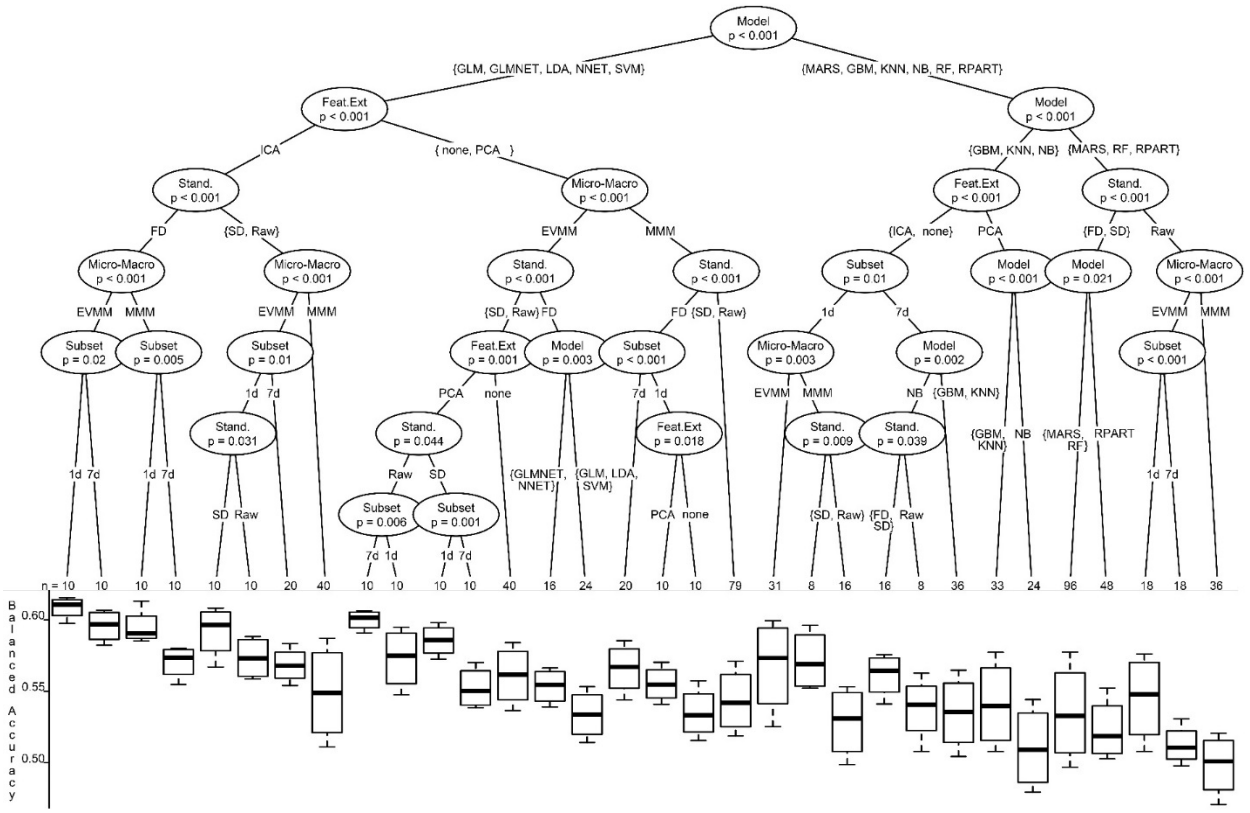

Figure 1. rtFMS regression tree results. Due to space, only the half of the regression tree with the better performing models was shown (only the models using SBF feature selection option); n=number of models in each terminal node; boxplots visually represent the balanced accuracy of models per terminal model. The bottom and top of the box represent the 25th and 75th percentiles, respectively, and the horizontal line inside the box is the median; Subset= Data Subset category; 7d= +- 7 days; 1d = +- 1 day; Micro-Macro = Micro-Macro Modeling Method category; MMM = Mean micro macro modeling method; EVMM= Extreme value micro macro modeling method; Stand. = Standardization category; Raw= Raw absorbance values; FD= 1st derivative; SD= 2nd derivative; Feat.Ext = Feature Extraction category; none = no feature extraction performed; PCA= Performed a PCA; ICA = Performed a ICA; SMOTE= Outcome Class Balancing with SMOTE category; 200= 200% up and 150% down-sampling; 500 = 500% up and 120% down-sampling; Algorithm=Algorithm category; GLM=logistic generalized linear models algorithm; GLMNET=lasso and elastic-net regularized generalized linear models algorithm; LDA=linear discriminant analysis algorithm; LDA=linear discriminant analysis algorithm; SVM=linear support vector machines algorithm; KNN=nearest neighbor methods algorithm; NB=naive Bayes algorithm; RPART=classification trees algorithm; NNET=neural networks algorithm; GBM=gradient boosting machine algorithm; RF=random forests algorithm; MARS=multivariate adaptive regression splines algorithm.

Table 2. Final models' performance measures with 95% confidence intervals.

Performance Measure	Estimate	95% Confidence Interval	
		Lower	Upper
Apparent prevalence, %	37.12	36.82	37.42
True prevalence, %	0.050	0.037	0.066
Sensitivity, %	62.00	47.17	75.35
Specificity, %	62.90	62.59	63.20
Diagnostic accuracy, %	62.90	62.59	63.20
Balanced accuracy, %	62.45	54.88	69.27
Positive predictive value, %	0.084	0.057	0.119
Negative predictive value, %	99.97	99.95	99.98
Likelihood ratio of a positive test	1.67	1.34	2.08
Likelihood ratio of a negative test	0.60	0.42	0.86
Kappa	0.0007	0.0003	0.0010

DISCUSSION

In this study a new method of micro-macro modeling, EVMM, was developed. EVMM was demonstrated by the development of a model for the prediction of herds' risk for antimicrobial residues (penicillin) in raw bulk milk using cow level FT-MIR spectrometry data. The rtFMS method was successful in systematically comparing and selecting modeling methods.

EVMM

For this study, EVMM was found to perform consistently better compared to the current MMM multilevel modeling approach. These results suggest that EVMM is an appropriate alternative to current MMM methods specifically when a small proportion or even just one of the micro level (individual level) observations can cause a change in the macro level (population level) outcome.

This data set included repeated measurements at the cow and farm level. However, random effects or corrections for within cluster correlations and repeated measures per farm were not used in this study. This is because if the models would have corrected for repeated measures per farm and cow, the effect of the extreme values would have been mitigated by the random effects (Gelman and Hill, 2006; Clark and Linzer, 2015). An antimicrobial residues violation by a single cow on an otherwise average farm with regards to FT-MIR wavenumbers is a single, short lived exception that could be represented by an extreme value. Such a unique occurrence would be lost if it was averaged across the farm and cow's other data using random effects. However, not correcting for repeated measures can overfit a model by overinflating parameter estimates and overestimating effects' statistical significance and underestimating standard errors (Dohoo et al., 2003; Gelman and Hill, 2006; Crawley, 2013). Nonetheless, since this study built prediction models and not parameter estimation models the focus was more on the performance of predictions, and not so much on the significance of estimated effects. Additionally, the cross-validation method employed did not use data from the same farm in both the training and test sets per cross-validation fold. This guaranteed that the repeated measures did not cause model overfitting. Finally, as a prediction model, it is not advisable to be restricted to only applying the model to data from the farms that were used in building the model. Therefore, bypassed mixed effect modeling were bypassed or the

sake of developing an optimized prediction model able to make meaningful predictions of external data for single, short-lived exceptions like antimicrobial residues violations.

An example of when the EVMM method would be appropriate is when modeling the presence or absence of a disease at the population level with individual level data. In some scenarios where the transmission rate is high, the population is susceptible and the contact rate is high, it only takes one or very few transmission events or infectious individuals to introduce a disease into a new population (Anderson and May, 1992). The individuals responsible for the introduction of a disease might have extreme values of the number of past international travel, contacts with livestock, or might have been mingling with large groups of people. The EVMM modeling technique might find an association with these parameters' extreme values and the introduction of a disease, while the mean of the individual level of the same variables might not be significantly associated with the outcome. Additionally, in finance, one company or industry's extreme event can be the best predictor of significant changes in the overall financial market (Embrechts et al., 2013). These examples of extreme events might be better predictors for the macro level outcome compared to the central tendency of all the data. In such cases using an EVMM model could help identify the characteristics of the unique event, individual or mutation that correspond to an outcome change at the macro level.

In the future, EVMM could be investigated as a first step toward developing micro level models when micro level outcomes are not available. For example, in this study the same model built for herd level outcomes could be applied to cow level observations to predict which individual cows are most likely contributing to a herd sample positive for antimicrobial residues. To accomplish this, a cow's single FT-MIR data value per FT-MIR wavenumber would serve both as an extreme value and as a mean value per FT-MIR wavenumber. Future on-farm measurements of FT-MIR data could facilitate this cow flagging in real time. This jump between levels of observation from cow to herd level and back is at the core of applications for EVMM.

In future studies, other combinations or other types of summary statistics could be used for the purpose of identifying extreme events for the analysis. Possible examples for alternative parameters from summary statistics that could represent extreme events include quartiles, standard deviation and other measures of variation, and the mode or median as well as other measures of central tendency. Additionally, methods that are usually employed to identify outliers such as a limit of plus or minus 3 standard deviations could be used in future EVMM modeling methods to identify extreme values (Ben-Gal, 2005). Prediction models using different combinations or other types of summary statistics warrants systematic comparisons of their performance.

Final Model

The use of rtFMS for the comparison and selection of modeling methods resulted in a final model for predicting herds' risk for penicillin residues in herds' raw bulk milk. Given the extremely low prevalence of the event of samples positive for antimicrobial residues (0.05%), a sensitivity of 61.4% and a specificity of 61.8% was notable given the challenges belonging to these data, among which: a rare event with imbalanced data, multilevel data that included mismatched data in time (cow and herd level data were not collected nor tested on the same day). Although, the final model had suboptimal performance, this study suggests that there is meaningful information present in milk FT-MIR data relating to the presence of antimicrobial residues (specifically penicillin).

The MMM models were able to reach 59% (CI- 58.5 - 60.6) balanced accuracy; Although not notable, it did signify that there was some information available for predictions by averaging the cow level data together. This information is most likely associated with the farm-level management, such as overall lesser total milk solids and greater SCC at the herd level. These characteristics have been shown to be associated with antimicrobial residues in bulk milk (Althaus et al., 2003). However, the EVMM models consistently performed significantly better than the MMM models. This signifies that working towards isolating the extreme values from the cow(s) responsible for the residue positive result performs better than finding the average herd-level FTIR characteristics associated with antimicrobial residues. The EVMM models could also be finding associations with milk characteristics associated with residues among which total solids or greater SCC but now at the cow level (Althaus et al., 2003). However, EVMM models could also be representing the true chemical effects of the residue on the spectra. These results will initiate further research aimed at discovering such patterns and information in more detail.

One of the reasons for the relatively low predictive performance of the current model could be associated with the fact that most herd and cow level samples were not taken or analyzed on the same day. Significant changes in the presence of antimicrobial residues among several consecutive days is likely. However, given the low prevalence of samples positive for antimicrobial residues, it was impossible to limit the data used to data where the herd and cow level were collected on the same day. Therefore, there is a chance that the FT-MIR data did not always correctly represent the herd level testing result. Since violations of antimicrobial residues in milk are short-lived, cow level data analyzed for FT-MIR spectra would not be representative for the herd level test results when measurements are too far apart. In future studies, it is advisable to take cow level FT-MIR samples on same day as the herd level bulk tank residues testing.

The current BRT diagnostic test does not have 100% specificity. BRT's false positive rates were found to be 3.75% in sheep's milk and 2.5% in goats' milk (Althaus et al., 2003; Romero et al., 2016). Although this could influence the final EVMM model's performance, it most likely does not account for all of the final false negative sample results. Additionally, the detection level of the current diagnostic test for penicillin residues is 2 µg/kg (Fejzić et al., 2014). Therefore, there is the potential for the milk of one residues positive cow to be diluted in the bulk tank past the detection level of the current diagnostic tests. If this is the case, it could suggest that some of the final model's false positives could in fact be correctly identifying residues at a level lower than the MRL of 4 µg/kg. Especially, since the FT-MIR data are available at the cow level, the FT-MIR technique might prove more sensitive than bulk tank testing. If this is the case, a quantitative test would be needed as a second step to determine if the level of the residues is above the MRL at the herd level.

rtFMS and FT-MIR

The first, and therefore the most important, decision selected by the regression tree was the use of a univariate filter, SBF, to select variables to go into the model instead of using all wavenumbers after the highly correlated variables were removed. This ensures that subsequent feature extraction methods extract signals associated with the outcome and not other information present in FT-MIR data such as fat and protein content. This is the first published comparison of FT-MIR results when using all available wavenumbers versus using a SBF filter. The significant improvement from the

resulting model is an argument in favor of applying this method more commonly for the analyses of big, high-dimensional data in the future.

Similarly to the SBF results, ICA performed significantly better compared to both PCA and not performing feature extraction. The use of ICA for analyzing FT-MIR data has been described previously by Hahn and Yoon (2006). The results suggest that ICA is able to separate out distinct signals from the complex FT-MIR spectrum; On the other hand, PCA, which can be seen as compressing the many data signals from FT-MIR data (Sun, 2012), was not selected as the better performing approach to feature reduction by rtFMS. To the best of the author's knowledge, this is the first comparison between ICA and PCA preprocessing techniques during FT-MIR data modeling. Significance of the three options for feature extraction were not always the same among the branches of the rtFMS. Therefore, the results of the current study highlight the importance of a comparison and systematic approach to full model selection.

Tremblay et al., 2018 found that using either FD or SD transformations were preferred over raw FT-MIR data. However, in this study, the FD transformation was preferred over both the SD transformation and the raw data for improved model performance. Similar finding were reported by Soyeurt et al. (2011), Dal Zotto et al. (2008), and De Marchi et al. (2014). This suggests that correcting for baseline shifts by using a FD transformation significantly improves the model performance, but that removing linear trends using a SD actually removes information that is important when analyzing for AMR signals. In addition, many of the linear models tended to perform significantly better compared to many non-linear models. This suggests the presence of a linear and additive correlation between the most influential predictors and the outcome variable as discussed in literature (Gelman and Hill, 2006).

CONCLUSIONS

In this study, the EVMM modeling method was shown to perform significantly better compared to the MMM modeling approach. This suggests that extreme observations at the micro level were better predictors for the macro level outcome compared to the predictors' mean value when modeling a single, short-lived exception (i.e., outlier event) like antimicrobial residues in milk events. This finding, among many others, was made possible by the use of rtFMS that allowed systematic comparisons of many modeling options. Finally, this study provided evidence that cow level FT-MIR spectral data hold information that could be used to predict herds' risk for positive penicillin residues at the herd level.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Rainer Lang at MPR (Milchprüfring Bayern e.V., Wolnzach, Germany)

REFERENCES

Albright, J.L., Tuckey, S.L. and Woods, G.T., 1961. Antibiotics in milk—a review. Journal of Dairy Science, 44(5), pp.779-807.

Althaus, R., Torres, A., Peris, C., BELTRAN, M.C., Fernandez, N. and MOLINA, M.P., 2003. Accuracy of BRT and Delvotest microbial inhibition tests as affected by composition of ewe's milk. Journal of food protection, 66(3), pp.473-478.

Anderson, R.M. and May, R.M., 1992. Infectious diseases of humans: dynamics and control. Oxford university press.

Andrew, S.M, K.M. Moyes, A.A. Borm, L.K. Fox, K.E. Leslie, J.S. Hogan, S.P. Oliver, Y.H. Schukken, W.E. Owens, and C. Norman. 2009. Factors associated with the risk of antibiotic residues and intramammary pathogen presence in milk from heifers administered prepartum intramammary antibiotic therapy. Vet. Microbiol. 134:150-156.

Bali, R., Sarkar, D., 2016. R Machine Learning By Example. Packt Publishing Ltd.

Baker, M.J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H.J., Dorling, K.M., Fielden, P.R., Fogarty, S.W., Fullwood, N.J., Heys, K.A., Hughes, C., Lasch, P., Martin-Hirsch, P.L., Obinaju, B., Sockalingum, G.D., Sule-Suso, J., Strong, R.J., Walsh, M.J., Wood, B.R., Gardner, P., Martin, F.L., 2014. Using Fourier transform IR spectroscopy to analyze biological materials. Nat. Protoc. 9, 1771–1791. https://doi.org/10.1038/nprot.2014.110.

Batista, G.E., Prati, R.C. and Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter, 6(1), pp.20-29.

Beleites, C., Sergo, V., 2018. hyperSpec: a package to handle hyperspectral data sets in R. R package version 0.99-20180627. http://hyperspec.r-forge.r-project.org.

Ben-Gal, I., 2005. Outlier detection. In Data mining and knowledge discovery handbook (pp. 131-146). Springer, Boston, MA.

Bennink, M., 2014. Micro-macro multilevel analysis for discrete data. Tilburg university.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach.Learn. Res. 13, 281–305.

Beyene, T., 2016. Veterinary drug residues in food-animal products: its risk factors and potential effects on public health. J Vet Sci Technol, 7(1), pp.1-7

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 1 (16), 321–357. https://doi.org/10.1613/jair.953.

Clark, T.S. and Linzer, D.A., 2015. Should I use fixed or random effects?. Political Science Research and Methods, 3(2), pp.399-408.

Crawley, M.J., 2013. The R book second edition. John Wiley & Sons.

Croon, M.A. and van Veldhoven, M.J., 2007. Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. Psychological methods, 12(1), p.45.

Commission Regulation (EU) No. 37/2010 (2010). Official Journal of the European Union 20.1.2010, L 15/1.

Dal Zotto, R., M. De Marchi, A. Cecchinato, M. Penasa, M. Cdro, P. Carnier, L. Gallo, and G. Bittante. 2008. Reproducibility and repeatability of measures of milk coagulation properties and predictive ability of mid-infrared reflectance spectroscopy. J. Dairy Sci. 91:4103–4112

De Briyne, N., Atkinson, J., Pokludová, L. and Borriello, S.P., 2014. Antibiotics used most commonly to treat animals in Europe. The Veterinary Record, 175(13), p.325.

De Marchi, M., Toffanin, V., Cassandro, M. and Penasa, M., 2014. Invited review: Mid-infrared spectroscopy as phenotyping tool for milk traits. Journal of Dairy Science, 97(3), pp.1171-1186.

Dohoo, I.R., Martin, W. and Stryhn, H., 2003. Veterinary epidemiologic research (No. V413 DOHv). Charlottetown, Canada: AVC Incorporated.

Duckworth, J., 2004. Mathematical data preprocessing. Near-Infrared Spectrosc. Agric.115–132.

Embrechts, P., Klüppelberg, C. and Mikosch, T., 2013. *Modelling extremal events: for insurance and finance* (Vol. 33). Springer Science & Business Media.

Fejzić, N., Begagić, M., Šerić-Haračić, S. and Smajlović, M., 2014. Beta lactam antibiotics residues in cow's milk: comparison of efficacy of three screening tests used in Bosnia and Herzegovina. Bosnian journal of basic medical sciences, 14(3), p.155.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22. http://www.jstatsoft.org/v33/i01/.

Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Hahn, S. and Yoon, G., 2006. Identification of pure component spectra by independent component analysis in glucose prediction based on mid-infrared spectroscopy. Applied optics, 45(32), pp.8374-8380.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning: Data mining, inference and prediction, Volume second. New York: Springer-Verlag.

He, H., Ma, Y. (Eds.), 2013. ImbAlanced Learning: Foundations, Algorithms, and Applications, 1 edition. Wiley-IEEE Press ed.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. 15 (3), 651–674.

Hothorn, T., Zeileis, A., 2015. Partykit: a modular toolkit for recursive partytioning in R. J. Mach. Learn. Res. 16, 3905–3909. http://jmlr.org/papers/v16/hothorn15a.html.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intell.Data Anal. 6, 429–449.

Kebede, G., Zenebe, T., Disassa, H. and Tolosa, T., 2014. Review on detection of antimicrobial residues in raw bulk milk in dairy farms. AJBAS, 6(4), pp.87-97.

Kuhn, M., Contributions from Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., and Hunt, T., 2018. caret: Classification and Regression Training. R package version 6.0-80. https://CRAN.R-project.org/package=caret.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer Science & Business Media.

Kyle, R., Popp, J., and Jay, M., 2018. epiR: tools for the analysis of epidemiological data. R package version 0.9–79. https://cran.r-project.org/package=epiR.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.

Marchini, J.L., Heaton, C., Ripley, M.B. and Suggests, M.A.S.S., 2017. Package 'fastICA'.

Milborrow, S., 2018. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. earth: Multivariate Adaptive Regression Splines. R package version 4.6.3. https://CRAN.Rproject.org/package=earth.

R Core Team, 2018. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ridgeway G., 2017. gbm: Generalized Boosted Regression Models. R package version 2.1.3. https://CRAN.R-project.org/package=gbm.

Rinnan, A., 2014. Pre-processing in vibrational spectroscopy – when, why and how. Anal.Methods 6, 7124–7129. https://doi.org/10.1039/C3AY42270D.

Romero, T., Van Weyenberg, S., Molina, M.P. and Reybroeck, W., 2016. Detection of antibiotics in goats' milk: Comparison of different commercial microbial inhibitor tests developed for the testing of cows' milk. International Dairy Journal, 62, pp.39-42.

Sivakesava, S. and Irudayaraj, J., 2002. Rapid determination of tetracycline in milk by FT-MIR and FT-NIR spectroscopy. Journal of Dairy Science, 85(3), pp.487-493.

Smith, B.R., Baker, M.J., Palmer, D.S., 2018. PRFFECT: a versatile tool for spectroscopists. Chemom. Intell. Lab. Syst. 172, 33–42. https://doi.org/10.1016/j.chemolab.2017.10.024.

Soyeurt, H., F. Dehareng, N. Gengler, S. McParland, E. Wall, D. P. Berry, M. Coffey, and P. Dardenne. 2011. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production sys-tems, and countries. J. Dairy Sci. 94:1657–1667

Stevenson, M., Nunes, T., Heuer, C., Marshall, J., Sanchez, J., Thornton, R., Reiczigel, J., Robison-Cox, J., Sebastiani, P., Solymos, P., Yoshida, K., Jones, G., Pirikahu, S., and Firestone, S., 2018. Kyle R. epiR: Tools for the Analysis of Epidemiological Data. R package version 0.9-93.

Sun, D.W. ed., 2012. Computer vision technology in the food and beverage industries. Elsevier.

Sun, Q., 2014. Meta-learning and the Full Model Selection Problem. Doctoral dissertation.
University of Waikato.

Sundlof, S.F. 1995. Human health risks associated with drug residues in animal-derived foods. J. Agromedicine. 1(2):5-20.

Torgo, L., 2010. Data Mining With R, Learning With Case Studies. Chapman and Hall/CRC. http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

Tremblay, M., Kammer, M., Lange, H., Plattner, S., Baumgartner, C., Stegeman, J.A., Duda, J., Mansfeld, R. and Döpfer, D., 2019. Prediction model optimization using full model selection with regression trees demonstrated with FTIR data from bovine milk. Preventive veterinary medicine, 163, pp.14-23.

U.S. Department of Agriculture (USDA), APHIS., Veterinary Services , National Animal Health Monitoring System (NAHMS). 2008. Dairy 2007: Part III: Reference of dairy cows health and management practices in the United States, 2007. Available at: http://www.aphis.usda.gov/animal_health/nahms/dairy/downloads/dairy07/Dairy07_dr_PartIII _rev.pdf Accessed Dec. 2013.

Weihs, C., Ligges, U., Luebke, K., Raabe, N., 2005. klaR analyzing German business cycles. In: Baier, D., Decker, R., Schmidt-Thieme, L. (Eds.), Data Analysis and Decision Support. Springer-Verlag, Berlin, pp. 335–343.

Yachen Y., 2016. MLmetrics: Machine Learning Evaluation Metrics. R package version 1.1.1. https://CRAN.R-project.org/package=MLmetrics.

CHAPTER 4- UNSUPERVISED LEARNING INTRODUCTION

Goal

In the supervised learning section (Chapters 2 and 3), there was still the buttress of associations between risk factors and an outcome. However, for the pattern recognition section, an outcome variable was not available and so unsupervised learning methods needed to be employed. The goal of unsupervised learning is to discover grouping of observations based on the variance explained by the input parameters and the patterns among the input parameters (Bishop, 2006; Hastie et al., 2009). Unsupervised learning techniques can be invaluable and necessary during the initial process of understanding and grasping the different aspects of a dataset before moving on to a supervised learning task (Chollet, 2018).

Assumptions

Unsupervised learning methods work under the assumption that there is an underlying grouping pattern to find, that the data set is representative of the population, that collinearity is not present among the input parameters, that they are normally distributed and that the variables are meaningful for explaining the underlying pattern (Jain and Dubes, 1988; Hair et al., 2006; Hastie et al., 2009).

Challenges

The major challenge with unsupervised learning is how to validate the results especially since the methods will always produce results. Since an outcome variable is not available, testing the biological relevance of the results is not straight-forward (Hastie et al., 2009; Hair et al., 2006). In addition, unsupervised learning methods can be limited in their ability to be generalized because the results are dependent on the parameters used in the analysis (Jain and Dubes, 1988). Finally, challenges associated with unsupervised learning techniques such as cluster analysis include selecting the number of clusters, quantifying the degree of misclassifications, and selecting the input parameters (Hair et al., 2006; Jain and Dubes, 1988).

Approaches

In Chapter 4.1 a cluster analysis was performed on blood metabolic parameters of early lactation Simmental cows in Bavaria, Germany. To validate the findings of the cluster analysis, post-hoc regression models were used. The post-hoc regression analysis examined associations between the observations' cluster classification and other clinical, milk and blood parameters that were not used in the cluster analysis. The post-hoc findings supported and aided in the interpretation of the clustering results. The cluster analysis with the addition of these post-hoc steps lead to the description of a novel syndrome for poor metabolic adaptation in dairy cows called "Poor Metabolic Adaptation Syndrome" (PMAS).

In Chapter 4.2, a data set was analyzed that included data from 529 dairy farms with automatic milking systems (AMS) in North America. Only the variables that had been shown to be associated with farms' performance in Chapter 2.3 were used as input parameters. This assured that the cluster analysis would focus on the pattern of interest related to farms' performance in the data. The common concern about how generalizable the results of clustering are to a larger population was

mitigated in two ways: 1) by only suggesting for these results be applied to the farms that were used in the cluster analysis, 2) by using 80% of all the farms with the same brand of AMS to perform the cluster analysis. To validate the findings of the cluster analysis, a comparison was made between the resulting cluster classification and the current method that is used to classify these farms for benchmarking. The cluster analysis classification was a better predictor for a farm's milk production than was the current benchmarking method.

Unsupervised learning is a key step for the understanding of a data set's characteristics before engaging in a supervised learning task. Therefore, it should be included in systematic approaches to data analysis. However, post-hoc analyses of unsupervised learning results and the careful interpretation of the results also need to be included in the systematic approach to unsupervised learning methods.

Reference

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer.

Chollet, F., 2018. Deep learning with Python. Manning Publications Co.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L., 2006. Multivariate data analysis . Uppersaddle River.

Hastie, T., Tibshirani, R. and Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction.

Jain, A.K. and Dubes, R.C., 1988. Algorithms for clustering data.

CHAPTER 4.1

Identifying poor metabolic adaptation during early lactation in dairy cows using cluster analysis

Journal of dairy science, 101 (8), pp. 7311-7321.
https://doi.org/10.3168/jds.2017-13582

M. Tremblay,*[1] M. Kammer,† H. Lange,‡# S. Plattner,‡# C. Baumgartner,‡ J.A. Stegeman,§ J. Duda,† R. Mansfeld,# D. Döpfer*

* Department of Medical Science, School of Veterinary Medicine, University of Wisconsin, 2015 Linden Dr., Madison 53706, United States of America
† LKV Bayerne.V., Landsberger Straße 282, 80687 München, Germany
‡ Milchprüfring Bayern e.V., Hochstatt 2, 85283 Wolnzach, Germany
§ Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, PO Box 80151, 3508 TD Utrecht, the Netherlands
# Clinic for Ruminants, Ludwig-Maximilians-Universität Munich, Sonnenstr. 16, D-85764 Oberschleissheim, Germany

[1]Corresponding author:
Marlène Tremblay, DVM
Department of Medical Sciences
School of Veterinary Medicine
University of Wisconsin-Madison
2015 Linden Dr., Madison, WI 53706
724-288-5159, mtremblay@wisc.edu

ABSTRACT
Currently, detection of cows with poor metabolic adaptation during early lactation, or poor metabolic adaptation syndrome (PMAS), are often identified based on detection of hyperketonemia. Unfortunately, elevated blood ketones do not manifest consistently with indications of PMAS. Expected indicators of PMAS include elevated liver enzymes and bilirubin, decreased rumen fill, reduced rumen contractions, and a drop in milk production. Expected characteristics of cows with PMAS are higher producing, older cows earlier during lactation with greater BCS at the start of lactation.

It was our aim to evaluate commonly used measures of metabolic health (input variables) that were available (i.e., blood beta-hydroxybutyrate acid, milk fat to protein ratio, blood non-esterified fatty acids (**NEFA**)) to characterize PMAS. Bavarian farms (n=26) with robotic milking systems were enrolled for weekly visits for on average of 6.7 weeks. Physical examinations of the cows (5 to 50 days in milk) were performed by veterinarians during each visit and blood and milk samples were collected. Resulting data included 790 observations from 312 cows (309 Simmental). Principal component analysis (PCA) was conducted on the three input variables, followed by K-means cluster analysis of the first two orthogonal components. The five resulting clusters were then ascribed to "Low", "Intermediate" or "High" PMAS classes based on their degree of agreement with expected PMAS indicators and characteristics in comparison with other clusters.

Results revealed that PMAS classes were most significantly associated with blood NEFA levels. Next, we evaluated NEFA values that classify observations into appropriate PMAS classes in this dataset, which we called separation values. Our resulting NEFA separation values (< 0.39 [0.360 - 0.410] mmol/L to identify Low PMAS observations, and ≥ 0.7 [0.650 - 0.775] mmol/L to identify High PMAS observations) were similar to values determined for Holsteins in conventional milking settings diagnosed with hyperketonemia and clinical symptoms such as anorexia and a reduction in milk yield as reported in literature. Data from additional locations, breeds, and milking systems should validate the findings. Future studies evaluating additional clinical and laboratory data are needed to validate these finding. The aim of future studies would be to build a PMAS prediction model to alert producers of cows needing attention and help evaluate on-farm metabolic health management at the herd level.

INTRODUCTION

Detection of cows with poor metabolic adaptation during early lactation, or poor metabolic adaptation syndrome (PMAS), are often identified based on detection of hyperketonemia (blood BHBA ≥ 1.2 mmol/L). In spite of initial observations (Sjollema and Van der Zande, 1923; Shaw, 1956), elevated blood ketone levels do not manifest consistently with indications of poor metabolic adaptation during early lactation (Andersson, 1984; Simensen et al., 1990; Duffield et al., 2009). The indications for poor metabolic adaptation to negative energy balance (**NEB**) during early lactation are secondary to the high energy demands of milk production (Baird, 1982). Expected indications of PMAS include elevated liver enzymes and bilirubin, decreased rumen fill, reduced rumen contractions, and a drop in milk production (Sevinç et al., 1998; Sahinduran et al. 2010; Issi et al., 2016; Ghanem et al. 2016; Cao et al., 2017). Expected characteristics of cows with PMAS are higher producing, older cows, earlier during lactation, and with greater BCS at the start of lactation (Baird, 1982; Rukkwamsuk et al., 1999; Andrews et al., 2004; Ghanem et al. 2016).

The need for an accurate measurement associated with PMAS has not been addressed. It was our aim to re-evaluate the commonly-used measures of metabolic health (input variables) that were available (i.e., blood beta-hydroxybutyrate acid, milk fat to protein ratio, blood non-esterified fatty acids (**NEFA**)) to characterize patterns of PMAS. Unlike some infectious diseases with clear case definitions (present or absent), cases of metabolic disease are more defined as syndromes observed on a spectrum of signs. A strictly binomial outcome variable such as "diseased or healthy" can be difficult to define for the purpose of prediction models. Principal Component Analysis (PCA) and cluster analysis do not require an outcome variable. A PCA detects important patterns among cases by generating linear combinations of meaningful potential predictors that represent the data's variance associated with disease. The PCA is followed by a cluster analysis that systematically groups the most similar observations into clusters that best explain the data's variance and therefore disease states (Brocard et al, 2011).

We hypothesized that performing a PCA and a cluster analysis using the input variables would differentiate groups of cattle with regards to patterns of PMAS. A clear understanding of PMAS is needed to further study the underlying mechanisms, possible prevention and treatment options, and to provide better indicators of genetic selection for metabolic health.

MATERIALS AND METHODS

Data Collection

Sixty farms equipped with Lely or Lemmer-Fullwood automatic milking systems (AMS) up to 70 kilometers from Munich were asked to participate in the study (Lely Industries N.V., Maassluis, the Netherlands; Lemmer-Fullwood GmbH, 53790 Lohmar, Germany). Twenty-six Bavarian farms (10 Lely, 16 Lemmer-Fullwood) were enrolled between May 2015 and December 2015. Data were collected as farms were enrolled between May 2015 and February 2016. On average, farms were visited for 6.65 (SD 1.16) consecutive weeks (range: 3 to 10).

Up to 8 early lactation cows between 5 and 50 days in milk (DIM) were evaluated during each visit. If more than 8 cows were between 5 and 50 DIM, the 8 cows earliest in their lactation were sampled. There was no minimum number of cows sampled to be included in the analysis. Milk samples were collected from all milkings on the day before the visit using an automatic sample collecting system attached to the automatic milking system for a minimum duration of 12 hours (7:00-19:00 hrs or 8:00-20:00 hrs). Milk collection had to be from voluntary milkings (sample are not to be collected by hand, and cows are not to be fetched into milking robot for collection).

Physical exams of the cows and blood sample collection were performed by the same two veterinarians (SP and HL). To screen animals for negative health conditions other than PMAS, physical exams included evaluating behavior, hygiene, and conformation, measuring internal body temperature, heart rate, and respiration rate, and performing heart auscultation, lung auscultation, complete udder examination, abdominal auscultation, percussion, and rectal palpation. Farm and cow identification numbers, date, DIM, breed and lactation number were recorded.

Clinical information documented for use in the analysis was the frequency of rumen contractions as described by Dirksen (1979), milk reduction compared to the day before, back fat measured by ultrasound as described by Staufenbiel (1992), change in back fat in one week, and rumen fill was

scored between 1 and 5 with 5 representing the most fill (Zaaijer and Noordhuizen, 2003; Appendix).

Blood samples were analyzed using the Cobas c311-Analyzer (Roche Diagnostics, Rotkreuz, Switzerland) for total blood protein, albumin, cortisol, bilirubin, aspartate aminotransferase (AST), gamma-glutamyl transferase (GGT), glutamate dehydrogenase (GLDH), creatine kinase (CK), beta-hydroxybutyrate (BHBA), and non-esterified fatty acids (NEFA). Milk production (kg) was calculated using the AMS mid-24 hour milk production measurement. Corresponding milk samples were analyzed for milk fat and protein percent, urea, and lactose using the MilkoScan FT-6000 (FOSS GmbH, Hamburg, Germany), and somatic cell count was measured using the Fossomatic 5000 (FOSS GmbH, Hamburg, Germany).

Data Editing

Several criteria were used to select data for the analysis. Observations were removed if any non-PMAS related health event was suspected or diagnosed at the time of the physical exam and if milk data were not collected from the robot. The earlier observations were removed if multiple milk samples were collected from a cow within the previous 12 hour period. Outliers, most likely due to data entry errors, were identified by visual inspection of each variables' histogram. Finally, observations were removed if it had a missing value for an input variable. The descriptive statistics (mean, standard deviation, and number of missing values) and variable descriptions of the final dataset were examined.

PCA and Cluster Analysis

All analyses were performed using the program R version 3.0.1 (R Development Core Team, 2013). The princomp and kmeans functions were used to perform the PCA and cluster analysis, respectively. The assumption of PCA is that input variables are normally distributed and that they have linear relationships (Borcard et al., 2011). The statistical assumption about the independence of observations can be relaxed with heuristic procedures (non-inference methods) such as PCA and cluster analysis (Jolliffie, 2002). The three input variables, those are: NEFA, BHBA, and FPR, were scaled and centered to standardize the data using the scale function in R. The scale function subtracts the mean of each variable from all the variable's values and then divides each value by the variable's standard deviation. Furthermore, scatter plots of the input variables were inspected for non-linear relationships.

A principal component analysis (PCA) was performed to transform the data into a number of orthogonal principal components (PCs) (Borcard et al., 2011). The PCs are ordered in descending order based on the amount of the variance they explain. The PCA results were examined by means of a scree plot that shows the decreasing amount of variance explained by PCs sorted by the amount of variance explained. The "elbow rule" was applied to determine how many PCs would be used in the cluster analysis. Briefly, the "elbow rule" selects PCs up until the elbow of the plot that is where the slope between PCs begins to increase most prominently (Johnson and Wichern, 2002; Jackson, 1993).

A cluster analysis was performed using K-means, a least-squares method. K-means is a linear method and as such requires normally distributed input variables which are not highly correlated (Borcard et al., 2011); therefore, the resulting PCs were visually inspected for normality by

creating histograms and pairwise Pearson correlations were calculated. The wrapper cascadeKM() calculated the simple structure index (ssi) criterion 1000 times per cluster number between 2 and 10 clusters (Borcard et al., 2011). The final number of clusters was selected by applying the elbow rule to the ssi plot. This was done to balance the minimum number of clusters with the maximum ssi criteria (Hothorn and Everitt, 2014). The silhouette plot was used to identify misclassifications, those are any observations with negative silhouette widths, and to evaluate the distribution of observations among clusters.

Comparison of External Variables per Cluster or PMAS Class

External variables are all the variables available that were not used as input variables for the PCA and cluster analysis: DIM, lactation, clinical information and blood and milk data excluding BHBA, FPR, and NEFA. Linear mixed-effect regression models were used to test for statistically significant associations between each of the external variables and the clusters with an alpha of 0.05. Cow ID and Farm ID were included as random effects on the intercept. Because there was no within-cow variation in lactation number, duplicate cow-cluster observations were removed and only farm ID was included as a random effect when modeling lactation number (see footnote in Table 2 and Table 3). A fixed effect of DIM and an interaction between DIM and cluster number were added if they significantly improved the models' goodness of fit using a log-likelihood ratio test. Goodness of fit was evaluated using diagnostic plots of the residuals among which the predicted versus fitted values. External variables were log transformed to normalize residuals, but the model estimates were transformed into the original scale for reporting the results. Results were presented as least-squares means and standard errors per cluster and post hoc comparisons among clusters' estimates were adjusted for multiple comparisons using Tukey's HSD method (Gelman and Hill, 2006). The significance of cluster number as a fixed effect was based on type III sum of squares and used an alpha level of 0.05 to determine significance. Post-hoc estimates for Back Fat were also reported at the beginning of lactation (DIM =5) when the interaction between cluster number and DIM was significant. Linear mixed-effect regression models were used again to quantify associations among each of the external variables and the 3 PMAS classes described in the next paragraph.

Classification of Clusters to PMAS Classes

The clusters' external variable characteristics were compared to expected indicators and characteristics of PMAS including: elevated liver enzymes and bilirubin, decreased rumen fill, reduced rumen contractions, and a drop in milk production (Sevinç et al., 1998; Sahinduran et al. 2010; Issi et al., 2016; Ghanem et al. 2016; Cao et al., 2017). Expected characteristics of cows with PMAS are higher producing, older cows, earlier during lactation, and with greater BCS at the start of lactation (Baird, 1982; Rukkwamsuk et al., 1999; Andrews et al., 2004; Ghanem et al. 2016). The clusters were then ascribed to "Low", "Intermediate" or "High" PMAS classes based on their degree of agreement with expected PMAS indicators in comparison with other clusters.

Separation of PMAS Classes

The PCA biplot was examined to identify how the input variables influenced the cluster separation, and how clusters separated into the new PMAS classifications. The most influential input variable(s) was selected as the PMAS measure to be used to identify values that classify observations into appropriate PMAS classes in this dataset, which we called separation values. Separation values that maximized the accuracy of classification were selected. Accuracy is the

proportion of correctly classified observations out of all observations (Dohoo et al., 2012). First, separation values of the selected PMAS measure were evaluated for correctly predicting the PMAS classifications of Intermediate PMAS observations compared to Low PMAS observations in this dataset. Second, separation values of the PMAS measure were evaluated for correctly predicting the PMAS classifications of the High PMAS observations compared to Intermediate PMAS observations in this dataset.

RESULTS

Data Collection

On average, there were 14.65 (SD 3.68) cows sampled per farm (range: 9 to 21). A total of 381 cows were evaluated. There was an average of 57.88 (SD 20.50) observations collected per farm (range: 22 to 116). Each cow was evaluated on average 3.95 times (SD 2.50).

Data Editing

The starting dataset contained 1505 observations. Four hundred and twenty-seven observations were removed due to a negative health condition other than PMAS having been suspected. Examples of such conditions include mastitis, retained placenta, milk fever, and displaced abomasum. In addition, 254 observations were removed because of multiple milk samples corresponding to a blood sample, and 30 observations were removed due to missing milk data from the robot. Outlier observations were removed including two observations with CK values above 12,000 U/l, and one outlier sample with a blood protein less than 5 g/l. One sample was removed due to a missing NEFA value.

The resulting data set contained 790 observations from 26 farms and represented 312 cows of which 309 were German Simmental cows, 1 was a Red Holstein cow and 2 were Holstein cows. On average, there were 12 (SD 2.99) cows sampled per farm (range: 8 to 19). There were on average of 30.38 (SD 7.81) observations collected per farm (range: 13 to 42). Each cow was evaluated on average 2.53 times (SD 1.32). Of those, there were 67 cows in their first lactations, 81 cows in their second lactations, and 164 cows in their third or later lactations. There was 260 missing Change in Back Fat values because the calculation of this value depended on having two consecutive measurements.

The descriptive statistics (mean, standard deviation, and number of missing values) of the final dataset are shown in Table 1. On average, cows in this study were in 27.51 DIM (SD 12.01) and produced 32.02 kg of milk per day (SD 7.10). Mean FPR was 1.28 (SD 0.25), BHBA mean was 0.80 mmol /L (SD 0.38), and NEFA mean was 0.45 mmol /L (SD 0.35).

PCA and Cluster Analysis

The standardized input variables (i.e., BHBA, FPR, NEFA) met the linearity assumption and were then transformed into PCs by means of a PCA to be used in the cluster analysis. The first and second component (PC1, PC2) explained 76.5% of the variance in the data and the second component was identified as the elbow in the scree plot. The loadings of NEFA, BHBA, and FPR in PC1 were -0.55, -0.59 and -0.59, respectively. The loadings of NEFA, BHBA, and FPR in PC2 were 0.84, -0.38 and -0.40, respectively.

Table 1: Descriptive statistics of all variables in a data set of n= 790 observations originating from 312 cows and 26 of Bavarian herds sampled between 5 and 50 DIM

Variable	Description (units)	Mean	SD[1]	#NA[2]
Lactation	Lactation number	3.00	1.60	0
DIM	Days in milk	27.5	12.0	0
Milk Production	Mid-24 hour milk calculated from robot data (kg)	32.0	7.1	0
Milk Fat	Fat content (%)	4.16	0.83	0
Milk Protein	Protein content (%)	3.27	0.32	0
FPR	Milk fat protein ratio	1.28	0.25	0
SCC	Somatic cell count (1000 cells/mL)	158.8	488.4	0
Urea	Urea content (mg/dL)	23.8	8.7	0
Lactose	Lactose content (%)	4.83	0.17	0
Blood Protein	(g/L)	71.2	5.1	0
Albumin	(g/L)	36.5	2.8	0
Bilirubin	(µmol/L)	1.21	1.08	0
AST	Aspartate aminotransferase (U/L)	84.2	25.1	0
GGT	Gamma-glutamyl transferase (U/L)	19.8	6.1	0
GLDH	Glutamate dehydrogenase (U/L)	12.4	11.2	0
CK	Creatine kinase (U/L)	281	452	0
BHBA	Beta-hydroxybutyric acid (mmol/L)	0.80	0.38	0
NEFA	Non-esterified fatty acids (mmol/L)	0.45	0.35	0
Cortisol	(ng/mL)	26.0	20.2	1
Rumen Contractions	Number of rumen contractions in two minutes	2.02	0.33	0
Rumen Fill[3]	Diagnostic rumen fill score (TR[4]: 1-5)	3.08	0.68	1
Back Fat	Back fat measured by ultrasound (mm)	12.1	3.9	15
Milk Production Reduction	Milk production reduction in one day (kg)	0.012	0.055	15
Change in Back Fat	Difference in back fat in one week (mm)	-0.63	2.37	260

[1] SD= standard deviation

[2] #NA= number of missing values (total number of observations= 790)

[3] Scoring system described in the Appendix

[4] TR = Theoretical Range

A feature of PCA is that the resulting orthogonal PCs are normally distributed and not correlated (Borcard et al., 2011); therefore, PC1 and PC2 met the assumptions for cluster analysis. The cluster analysis results were visualized by means of an ssi plot. Based on the elbow rule, the elbow in the ssi plot was identified at five clusters (ssi =1.21). Therefore, five clusters were selected for our final clustering results. No misclassifications were recognized in the silhouette plot, and the number of observations and silhouette widths were similar among clusters. Cluster 1 included 234 observations, Cluster 2 included 157 observations, Cluster 3 included 137 observations, Cluster 4 included 142 observations and Cluster 5 included 120 observations. Boxplots of the input variables per cluster number are described in Table 2. On average, a cow had observations in 1.776 difference clusters (SD 0.838).

Table 2: Results of the linear mixed-effects regression models including least-squares means and standard errors by cluster number and type III sum of squares P-values. Multiple comparisons among cluster numbers are adjusted using Tukey's HSD method; the data set originated from 312 cows and 26 Bavarian herds sampled 5 to 50 DIM (n=790)

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	P-value
Lactation [1]	2.70[a] (0.17)	2.84[a] (0.19)	3.11[a] (0.20)	3.09[a] (0.19)	2.87[a] (0.19)	0.16
DIM	30.6[a] (1.03)	30.5[a] (1.17)	23.2[b] (1.24)	23.5[b] (1.20)	26.3[b] (1.26)	< 0.001
Milk Prod., kg[2]	31.2[bc] (0.73)	31.4[bc] (0.74)	32.7[a] (0.76)	32.1[ab] (0.75)	30.9[c] (0.75)	< 0.001
Milk Fat, %	3.76[d] (0.053)	4.44[b] (0.060)	4.23[c] (0.064)	4.89[a] (0.062)	3.56[e] (0.064)	< 0.001
Milk Protein, %[3]	3.34[a] (0.031)	3.24[b] (0.033)	3.18[bc] (0.035)	3.11[c] (0.034)	3.36[a] (0.034)	< 0.001
FPR[4,5]	1.12[c] (0.012)	1.36[b] (0.014)	1.33[b] (0.015)	1.57[a] (0.015)	1.06[d] (0.016)	< 0.001
SCC, 1000 cells/mL	66.7[a] (6.70)	81.5[a] (8.87)	66.0[a] (7.59)	72.3[a] (8.04)	74.6[a] (8.45)	0.17
Urea, mg/dL	23.9[a] (1.11)	23.5[a] (1.16)	23.7[a] (1.20)	23.7[a] (1.18)	22.8[a] (1.20)	0.68
Lactose, %	4.85[a] (0.013)	4.82[ab] (0.015)	4.82[ab] (0.016)	4.80[b] (0.016)	4.83[ab] (0.016)	0.058
Blood Protein, g/L	71.2[a] (0.47)	70.7[a] (0.50)	71.5[a] (0.52)	70.7[a] (0.51)	71.8[a] (0.52)	0.061
Albumin, g/L	36.1[a] (0.25)	36.1[a] (0.27)	36.6[a] (0.28)	36.3[a] (0.27)	36.2[a] (0.27)	0.15
Bilirubin, µmol/L[3]	0.86[c] (0.071)	0.77[c] (0.086)	1.90[a] (0.092)	1.33[b] (0.091)	1.38[b] (0.092)	< 0.001
AST, U/L [3,6]	81.7[b] (1.97)	79.9[b] (2.24)	89.8[a] (2.41)	89.2[a] (2.31)	83.8[ab] (2.38)	< 0.001
GGT, U/L [7]	20.0[a] (0.38)	19.3[a] (0.42)	20.3[a] (0.44)	19.9[a] (0.43)	20.0[a] (0.43)	0.15
GLDH, U/L[3,8]	9.28[a] (0.47)	9.48[a] (0.53)	10.24[a] (0.60)	10.50[a] (0.60)	9.71[a] (0.55)	< 0.001
CK, U/L [9]	179[a] (9.7)	175[a] (11.3)	205[a] (14.4)	205[a] (13.8)	186[a] (13.2)	0.24
BHBA, mmol/L [10,5]	0.68[c] (0.029)	0.86[b] (0.032)	0.79[b] (0.034)	1.11[a] (0.033)	0.60[c] (0.034)	< 0.001
NEFA, mmol/L[3,5,11]	0.264[c] (0.016)	0.242[c] (0.020)	0.889[a] (0.021)	0.516[b] (0.021)	0.490[b] (0.021)	< 0.001
Cortisol, ng/mL	18.6[a] (1.51)	17.8[ab] (1.62)	23.1[a] (2.24)	14.2[b] (1.33)	19.5[a] (1.89)	< 0.001
Rumen Contractions [12]	2.00[a] (0.024)	2.02[a] (0.029)	1.99[a] (0.031)	2.01[a] (0.030)	2.06[a] (0.032)	0.52
Rumen Fill [13]	3.17[a] (0.058)	3.21[a] (0.065)	2.92[b] (0.070)	3.09[ab] (0.067)	2.95[b] (0.069)	< 0.001
Back Fat, mm[3]	12.2[a] (0.41)	12.3[a] (0.43)	12.5[a] (0.44)	12.1[a] (0.43)	12.1[a] (0.44)	< 0.001
DIM= 5	13.2[bc] (0.54)	12.8[c] (0.64)	15.1[a] (0.57)	14.7[ab] (0.57)	13.8[abc] (0.56)	
MPR, kg [14]	0.019[a] (0.004)	0.005[a] (0.004)	0.010[a] (0.005)	0.012[a] (0.005)	0.010[a] (0.005)	0.18
Change in Back Fat, mm	-0.18[b] (0.189)	-0.51[ab] (0.217)	-0.87[ab] (0.244)	-1.25[a] (0.248)	-0.74[ab] (0.273)	0.010

[a–d] Means within a row with different superscripts differ ($P < 0.05$); [1] duplicate cluster-cow combinations removed due to a lack in variance per cow, n=554; [2] Prod.= Production; [3] Significant interaction between cluster and DIM ($P < 0.05$); [4] FPR = milk fat to protein ratio; [5] These variables were used as input variables for the cluster analysis and are therefore expected to be significantly associated among clusters; [6] AST = aspartate aminotransferase; [7] GGT = gamma-glutamyl transferase; [8] GLDH = glutamate dehydrogenase; [9] CK = creatine kinase; [10] BHBA = blood beta-hydroxybutyric acid; [11] NEFA = blood non-esterified fatty acids; [12] number of rumen contractions in two minutes; [13] The description of the scoring system is available in Table 1 and the Appendix; [14] MPR = Milk Production Reduction

Comparison of External Variables per Cluster

SCC, GLDH, CK, and Cortisol were log transformed to normalize residuals. All regression models of the external variables include DIM as a fixed effect except DIM, FPR, BHBA, Rumen Fill, and Change in Back Fat. The only regression model that included an interaction between cluster number and DIM were Milk Protein, Bilirubin, AST, GLDH, NEFA and Back Fat. All external variables, with the exception of Urea, SCC, Albumin, GGT, CK, Rumen Contractions, and Milk Production Reduction were significantly associated with cluster assignment (p-value < 0.05) (Table 2). The input variables' linear mixed-effects regression model results were also reported for comparison (Table 2), although it is to be expected that they would be significantly associated with the cluster classifications (Legendre and Legendre, 2012).

Classification of Clusters to PMAS Classes

Cluster 1 and 2 had the greatest rumen fill, and were younger cows and cows later in lactation compared to the other clusters (Table 2). Cluster 1 and 2 had low bilirubin, AST, GLDH, CK and NEFA. These characteristics align with characteristics of healthy cows. Cluster 3 had greater milk production, greater back fat at the beginning of lactation (DIM=5) and earlier DIM compared to Clusters 1 and 2 (Table 2). These risk factors in addition to decreased rumen fill, and elevated bilirubin, AST, GLDH, CK and NEFA align with expected characteristics of cows with PMAS. Cluster 4 and 5 had intermediate back fat at the beginning of lactation (DIM=5), rumen fill, bilirubin and NEFA (Table 2). These intermediate levels of liver values and clinical results during early lactation placed Cluster 4 and 5 between the levels of agreement of the other clusters. Therefore, Clusters 1 and 2 were classified together as "Low", Clusters 4 and 5 as "Intermediate" and Cluster 3 was redefined as the only cluster with "High" agreement with expected PMAS indicators.

On average, cows had observations in 1.532 PMAS classes (SD 0.641). Eighty-seven cows had at least one observation classified in the High PMAS class. Thirty-one cows had more than one observation classified in the High PMAS class.

Comparison of External Variables per PMAS Class

SCC, GLDH, CK, and Cortisol were log transformed to normalize residuals. All regression models of the external variables include DIM as a fixed effect except DIM, Rumen Fill, and Change in Back Fat. The only regression model that included an interaction between cluster number and DIM were Milk Protein, Bilirubin, AST, GLDH, BHBA, NEFA and Back Fat. All external variables, with the exception of Lactation, urea, SCC, lactose, blood protein, GGT, CK, rumen contractions, milk production reductions were significantly associated with the PMAS classifications (p-value < 0.05) (Table 3). The input variables' linear mixed-effects regression model results were also reported for comparison, although it is to be expected that they would be significantly associated with the PMAS classifications (Legendre and Legendre, 2012). The Low PMAS class had significantly lower average FPR, bilirubin, AST, and NEFA compared to the Intermediate and High PMAS classes (Table 3). The Low PMAS class also had significantly greater DIM, rumen fill and milk protein compared to the Intermediate and High PMAS classes (Table 3). Although not significantly different, the Low PMAS class had lower average lactation number, milk production, and albumin compared to the Intermediate and High PMAS classes. The High PMAS class had significantly lower BHBA and greater milk production, bilirubin, NEFA, and cortisol compared to the Intermediate PMAS class (Table 3). Although not significantly different, the High

PMAS class had the lowest rumen fill, and greatest back fat at beginning of lactation compared to Low and Intermediate PMAS classes. NEFA and bilirubin were the only variables significantly different among all three PMAS classifications.

Table 3: Results of the linear mixed-effects regression models including least-squares means and standard errors by poor metabolic adaptation syndrome (PMAS) classification and type III sum of squares P-values. Multiple comparisons among PMAS classification are adjusted using Tukey's HSD method; the data set originated from 312 cows and 26 Bavarian herds sampled 5 to 50 DIM (n=790)

	PMAS Classification [1]			
Variable	Low	Intermediate	High	P-value
Lactation [2]	2.76[a] (0.158)	2.97[a] (0.164)	3.12[a] (0.202)	0.15
DIM	30.5[a] (0.93)	24.8[b] (1.00)	23.2[b] (1.24)	< 0.001
Milk Prod., kg [3]	31.3[b] (0.72)	31.5[b] (0.72)	32.7[a] (0.76)	< 0.001
Milk Fat, %	4.03[b] (0.064)	4.27[a] (0.068)	4.20[ab] (0.084)	< 0.001
Milk Protein, % [4]	3.30[a] (0.030)	3.24[b] (0.031)	3.18[b] (0.035)	0.028
FPR[5,6]	1.22[b] (0.018)	1.33[a] (0.019)	1.32[a] (0.025)	< 0.001
SCC, 1000 cells/mL	72.0[a] (6.71)	73.4[a] (7.13)	65.8[a] (7.56)	0.51
Urea, mg/dL	23.8[a] (1.08)	23.3[a] (1.10)	23.7[a] (1.20)	0.69
Lactose, %	4.84[a] (0.012)	4.81[a] (0.013)	4.82[a] (0.016)	0.12
Blood Protein, g/L	71.0[a] (0.45)	71.3[a] (0.46)	71.5[a] (0.53)	0.41
Albumin, g/L	36.1[b] (0.24)	36.3[ab] (0.25)	36.6[a] (0.28)	0.037
Bilirubin, µmol/L[4]	0.83[c] (0.060)	1.38[b] (0.068)	1.90[a] (0.093)	< 0.001
AST, U/L [4,7]	80.4[b] (1.76)	86.3[a] (1.90)	87.9[a] (2.44)	< 0.001
GGT, U/L [8]	19.7[a] (0.35)	20.0[a] (0.37)	20.3[a] (0.44)	0.31
GLDH, U/L[4,9]	9.38[a] (0.456)	10.10[a] (0.508)	10.21[a] (0.602)	< 0.001
CK, U/L [10]	177[a] (8.1)	196[a] (10.1)	204[a] (14.3)	0.12
BHBA, mmol/L [11,6]	0.761[b] (0.032)	0.847[a] (0.033)	0.771[ab] (0.041)	< 0.001
NEFA, mmol/L[4,6,12]	0.256[c] (0.014)	0.507[b] (0.015)	0.889[a] (0.021)	< 0.001
Cortisol, ng/mL	18.1[b] (1.35)	16.5[b] (1.31)	23.3[a] (2.28)	< 0.001
Rumen Contractions [13]	2.01[a] (0.020)	2.03[a] (0.023)	1.99[a] (0.031)	0.46
Rumen Fill [14]	3.19[a] (0.054)	3.02[b] (0.057)	2.92[b] (0.070)	< 0.001
Back Fat, mm[4]	12.2[a] (0.40)	12.0[a] (0.40)	12.4[a] (0.44)	< 0.001
DIM= 5	13.1[b] (0.49)	14.1[a] (0.48)	15.1[a] (0.56)	
MPR, kg [15]	0.013[a] (0.003)	0.011[a] (0.003)	0.010[a] (0.005)	0.77
Change in Back Fat, mm	-0.32[b] (0.143)	-1.02[a] (0.184)	-0.87[ab] (0.244)	0.007

[a–d] Means within a row with different superscripts differ (P < 0.05); [1] PMAS Classification= degree of agreement with expected PMAS indicators in comparison with other clusters; [2] duplicate cluster-cow combinations removed due to a lack in variance per cow, n=478; [3] Prod.= Production; [4] Significant interaction (P < 0.05); [5] FPR = milk fat to protein ratio; [6] These variables were used as input variables for the cluster analysis and are therefore expected to be significantly associated among clusters; [7] AST = aspartate aminotransferase; [8] GGT = gamma-glutamyl transferase; [9] GLDH = glutamate dehydrogenase; [10] CK = creatine kinase; [11] BHBA = blood beta-hydroxybutyric acid; [12] NEFA = blood non-esterified fatty acids; [13] number of rumen contractions in two minutes; [14] The description of the scoring system is available in Table 1 and the Appendix; [15] MPR = Milk Production Reduction

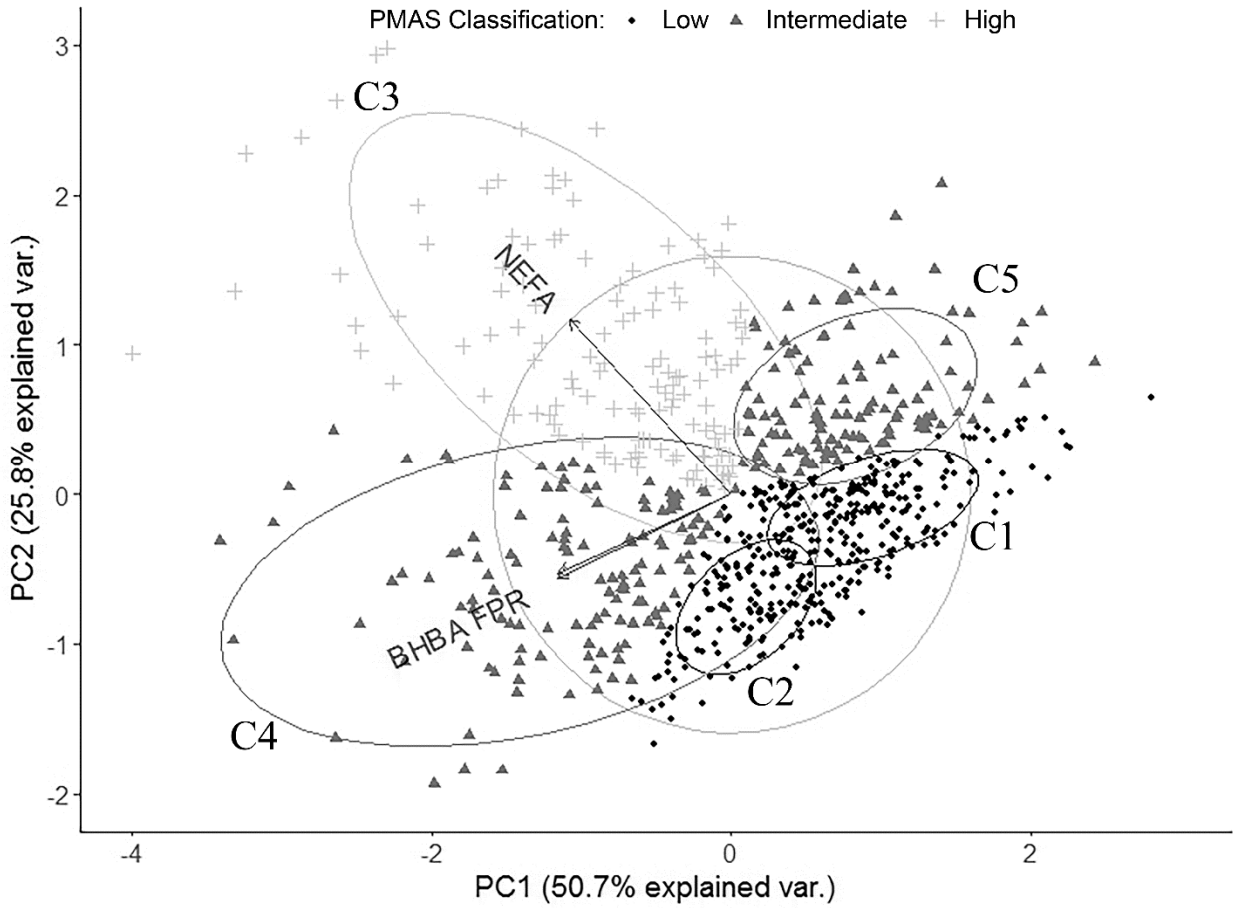

Figure 1: Biplot of the principal components analysis results by cluster number and assigned PMAS classification; PC = principal component; var. = variance; PMAS= poor metabolic adaptation syndrome; PMAS Classification= degree of agreement with expected PMAS indicators in comparison with other clusters; Cluster number: C1= Cluster 1, C2= Cluster 2, C3= Cluster 3, C4= Cluster 4, C5= Cluster 5; the data set originated from 312 cows and 26 Bavarian herds sampled 5 to 50 DIM (n=790)

Separation of PMAS classes

Examining the biplot of the PCA, it is apparent that NEFA's direction of influence is what separated out the three PMAS classifications in our dataset (Figure 1). The influence of BHBA was in the same direction as the one of FPR (arrows overlap in Figure 1). BHBA and FPR's direction of influence separated out Cluster 1 from Cluster 2, and Cluster 4 from Cluster 5 within their own classification of Low and Intermediate PMAS, respectively. NEFA was selected as the PMAS measure for this dataset because NEFA's direction of influence in the biplot was responsible for separating out Low, Intermediate and High PMAS classifications, and NEFA was the only input variable significantly different among all three PMAS classifications. The greatest accuracy of separation between Low and Intermediate PMAS observations was at a value of 0.390 [0.360 - 0.410] mmol/L NEFA (Figure 2). The greatest accuracy of separation between Intermediate and High PMAS observations in this dataset was at a value of 0.700 [0.650 - 0.775] mmol/L NEFA (Figure 3).

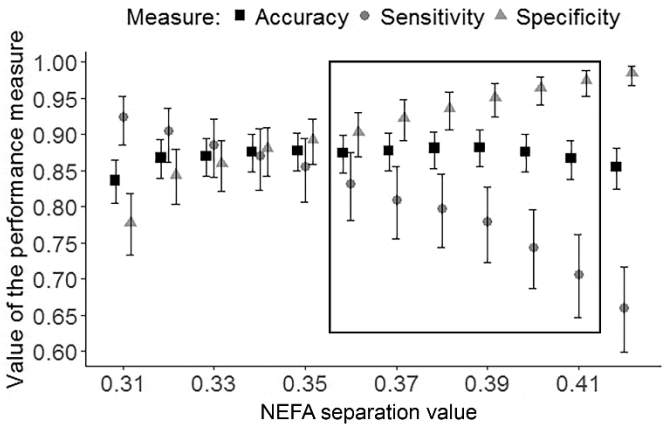

Figure 2: Classification performance measures (accuracy, sensitivity and specificity) for classifying Low PMAS and Intermediate PMAS observations by NEFA value. The box surrounds values that have overlapping confidence intervals with the separation value that has greatest accuracy; the data set originated from 312 cows and 26 Bavarian herds sampled 5 to 50 DIM (n=790); PMAS= poor metabolic adaptation syndrome.

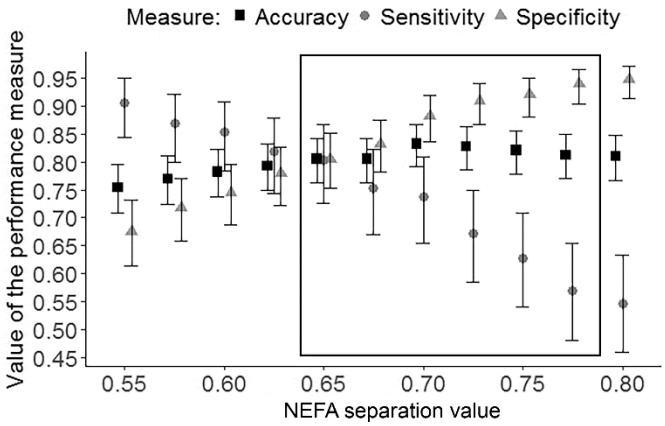

Figure 3: Classification performance measures (accuracy, sensitivity and specificity) for classifying Intermediate PMAS and High PMAS observations by NEFA values. The box surrounds separation values that have overlapping confidence intervals with point that has greatest accuracy; the data set originated from 312 cows and 26 Bavarian herds sampled 5 to 50 DIM (n=790); PMAS= poor metabolic adaptation syndrome

DISCUSSION

Metabolic Adaptation to NEB

The three levels of agreement with expected PMAS indicators did not follow differences in BHBA levels. This was highlighted by the differences between Cluster 3 and Cluster 4 wherein Cluster 3 had the highest agreement with expected PMAS indicators, while Cluster 4 had the highest BHBA values. The contrast between PMAS classes and BHBA measurements may be due to the fact that the majority of cows experience NEB during the first months post-partum due to the demands of high milk production. Ketogenesis, and resulting ketonemia, are a normal physiological response to compensate for NEB, and do not necessarily reflect pathological

changes. Indeed, keto-adaptation is a well-known phenomenon; in humans, ketones become the major fuel source following a period of adaptation to low carbohydrate intake. Furthermore, endurance athletes have been shown to be in nearly a constant state of ketonemia during NEB (Volek et al., 2016). As ketonemia does not necessarily reflect pathology, it becomes important for veterinary clinicians to be able to distinguish between appropriate and inappropriate responses to NEB.

Klein et al (2012) propose that cows may compensate for NEB in one of two ways: either by reducing fat in milk or by increasing fat mobilization from adipose tissue. Only the latter group consistently developed hyperketonemia (Klein et al., 2012). Our data support this hypothesis. Cluster 5, a group with intermediate NEFA levels had low milk fat but no elevation in BHBA compared to Cluster 4 that had similar NEFA levels. This suggests that Cluster 5 adapts to NEB by either limiting milk fat, or by being limited in ketogenesis, which in turn limits milk fat (Baumgard et al., 2000). Cluster 4 had the highest BHBA level as well as the highest milk fat of any cluster. This suggests that Cluster 4 adapts to NEB by increasing ketogenesis and not by limiting milk fat. Cluster 3 had the highest agreement with expected PMAS indicators. These observations did not have decreased milk fat like Cluster 5, or mobilized ketones like Cluster 4, which suggests that they did not adapt appropriately to NEB. At the same time, Cluster 3 exhibited higher NEFA values than either Cluster 4 or 5.

NEFA Separation Values

NEFA values are currently used during the pre-partum period to indicate the success of transition cow management programs (Oetzel, 2007). The majority of studies have focused on the use of NEFA values to predict negative sequellae during lactation (e.g., displaced abomasum, retained placenta, metritis, culling, reduced reproduction performance etc). These outcomes can result from elevated NEFA which can impair immune, liver and ovarian function (Adewuyi et al., 2005). Furthermore, NEFA values above 0.4 mmol/L during the pre-partum period are associated with negative outcomes during the subsequent lactation (Whitaker, 2004; McArt et al., 2013). When measured during the post-partum period, the NEFA cut-off value used to predict negative outcomes is > 0.7 mmol/L (Whitaker, 2004; McArt et al., 2013). The separation values we determined for these data (NEFA < 0.39 [0.360 - 0.410] mmol/L to identify Low PMAS observations, and ≥ 0.7 [0.650 - 0.775] mmol/L to identify High PMAS observations) were similar to those values used to predict negative health outcomes later during lactation.

Cao et al. (2017) suggest NEFA values greater than 0.82 mmol/L as the cut-off for diagnosing cows with BHBA greater than 1.2 mmol/L and clinical symptoms such as anorexia and a reduction in milk yield. Considering that Cao et al. (2017) examined Holsteins exclusively, and used a case definition of cows with BHBA greater than 1.2 mmol/L and clinical symptoms, their reported cut-off values for NEFA were surprisingly similar to the High PMAS separation value determined in our study that examined predominately Simmentals. However, 0.82 mmol/L is not included in our separation value's confidence intervals of 0.650 - 0.775 mmol/L. In addition to differences in breed and case definitions, the difference in NEFA separation values between Cao et al (2017) and our study could be due to the difference in ability to identify subtle indications of PMAS of the individual performing the exam.

Outlook

The number of rumen contractions was not significantly associated with the clusters in our study. This finding was surprising and could be caused by several factors including large individual variation among cows, differences in time between feeding and sampling, as well as differences in nutrition. The most likely reason for the lack of detectable difference in rumen contractions among clusters is due to short intervals of measurements of 2 minutes as described by Dirksen (1975) versus 5 minutes used by Issi et al. (2016) used when describing a significant difference in rumen contractions. Reduced milk production was not significantly associated with PMAS classifications in our study, although it was an expected indication of PMAS (Ghanem et al., 2016). The lack of an association between PMAS and reductions in milk production in our study may be due to fluctuations in milk production that were not detected during weekly visits, or because the differences in milk production were not adjusted for the expected milk production of each cow. To better characterize the clusters, future studies should count the number of rumen contractions for at least 5 minutes, and record milk production every day to improve the ability to detect reduced milk production.

Our study was limited by the fact that we did not include observations from cows experiencing negative health conditions other than PMAS. It is possible that other health events could also cause elevated NEFA, in which case the NEFA values from these cows could affect the accuracy of the chosen separation values to identify PMAS cows. In our final dataset, all cows were Simmental cows except three, and these data were only from AMS herds. Thus, it is possible that our findings are particular to this breed and milking system. In this analysis we did not consider feed intake; time between feeding and sampling of cows; or previous treatments, interventions, and health events because these data were not available in the provided dataset. These missing variables would have been useful to characterize the clusters in more detail and could have a significant influence on cluster classification.

It is necessary to further investigate the effects of genetics on the development of PMAS as well as the various physiological mechanisms by which cows compensate for NEB in order to develop selection criteria against cows that are predisposed to developing PMAS. The most appropriate management strategy may vary depending on the physiological compensation mechanism. Our resulting NEFA separation values for are similar to those determined for Holsteins with BHBA greater than 1.2 mmol/L and clinical symptoms in conventional milking settings, but follow-up analyses are required to determine if these separation values should be adjusted further to account for additional variables such as location, DIM, breed, milking system and season. Further adjustments may also be necessary to differentiate PMAS from other health conditions. The selection of separation values should result in a balance between the needs for high sensitivity or high specificity or both. Finally, future studies are needed to validate these findings in different populations, breeds, seasons, and locations. Since NEFA is expensive to measure, future studies could also evaluate milk Fourier-transform infrared spectroscopy data for its ability to distinguish PMAS classes. This would allow routine in-line measurements to be used for PMAS prediction. Beyond individual cow detection, these separation values should be tested at the herd detection level as well to determine a herd prevalence alarm level.

CONCLUSION

A cluster analysis was able to differentiate groups of cattle in terms of NEB compensation mechanisms and PMAS classifications: Low, Intermediate, and High. NEFA was the best indicator

of PMAS classifications for these data and separation values were selected at < 0.39 [0.360 - 0.410] mmol/L to identify Low PMAS observations, and ≥ 0.7 [0.650 - 0.775] mmol/L to identify High PMAS observations. Future prospective studies are needed to validate these findings and to evaluate other possible predictors for metabolic health, such as FTIR data from milk. The aim of future studies would be to build a prediction model for PMAS to alert producers of cows needing attention in addition to helping evaluate on-farm metabolic health management (e.g., transition cow management, nutrition).

ACKNOWLEDGEMENTS

The authors acknowledge the Bayerisches Staatsministerium für Ernährung, Landwirtschaft und Forsten (i.e. the Bavarian Ministry for Nutrition, Agriculture and Forests) for supporting the collection of the data. The project was supported by funds of the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme. We gratefully acknowledge the editing effort of Nick Robl.

REFERENCES

Adewuyi, A. A., E. Gruys, and F. van Eerdenburg. 2005. Non esterified fatty acids (NEFA) in dairy cattle. A review. Vet. Q. 27:117–126.

Andersson, L. 1984. Concentrations of blood and milk ketone bodies, blood isopropanol and plasma glucose in dairy cows in relation to the degree of hyperketonaemia and clinical signs. Zbl. Vet. Med. A 31:683–693.

Andrews, A. H., R. Blowey, H. Boyd, and R. Eddy. 2004. Bovine Medicine: Diseases and Husbandry of Cattle. Blackwell Publishing Company, Hoboken, NJ.

Baird, G. D. 1982. Primary ketosis in the high-producing dairy cow: Clinical and subclinical disorders, treatment, prevention, and outlook. J. Dairy Sci. 65:1–10.

Baumgard, L. H., B. A. Corl, D. A. Dwyer, A. Sæbø, and D. E. Bauman. 2000. Identification of the conjugated linoleic acid isomer that inhibits milk fat synthesis. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology. 278:R179-R184.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical Ecology with R. Springer. New York, NY.

Cao, Y., J. Zhang, W. Yang, C. Xia, H. Y. Zhang, Y. H. Wang, and C. Xu. 2017. Predictive value of plasma parameters in the Agreement of postpartum ketosis in dairy cows. Journal of Veterinary Research. 61:91-95. https://dx.doi.org/10.1515/jvetres-2017-0011.

Dirksen, G. 1979. Digestive system. Pages 184-258 in Clinical examination of cattle. 2nd ed. G. Rosenberger, ed. Verlag Paul Parey, Berlin and Hamburg, Germany.

Dohoo, I., W. Martin, and H. Stryhn. 2012. Methods in Epidemiologic Research. Ver. Inc., Charlottetown, Prince Edward Island, Canada.

Duffield, T. F., K. D. Lissemore, B. W. McBride, and K. E. Leslie. 2009. Impact of hyperketonemia in early lactation dairy cows on health and production. J. Dairy Sci. 92:571–580.

Gelman, A. and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, New York, NY.

Ghanem, M.M., M. E. Mahmoud, Y. M. Abd El-Raof, and H. M. El-Attar. 2016. Alterations in biochemical parameters and hepatic ultrasonography with reference to oxidant injury in ketotic dairy cows. Benha Veterinary Medical Journal. 31:231-240.

Hothorn, T., and B. S. Everitt. 2014. A handbook of statistical analyses using R. 3rd ed. CRC press, Boca Raton, FL.

Issi, M., Y. Gül, and O. Başbuğ. 2016. Evaluation of renal and hepatic functions in cattle with subclinical and clinical ketosis. Turkish Journal of Veterinary and Animal Sciences, 40:47-52.

Jackson, D.A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology. 74:2204-2214.

Johnson, R. A., and D. W. Wichern. 2002. Applied Multivariate Statistical Analysis. 5th ed. Prentice-Hall International, Upper Saddle River, NJ.

Jolliffe, I. T. 2002. Principal Component Analysis. 2nd ed. Springer-Verlag New York Inc., New York, NY.

Klein, M. S., N. Buttchereit, S. P. Miemczyk, A. K. Immervoll, C. Louis, S. Wiedemann, W. Junge, G. Thaller, P. J. Oefner, and W. Gronwald. 2012. NMR metabolomic analysis of dairy cows reveals milk glycerophosphocholine to phosphocholine ratio as prognostic biomarker for Agreement of ketosis. J. Proteome Res. 11:1373–1381.

Legendre, P., and L. F. Legendre. 2012. Numerical Ecology. Vol. 24. Elsevier, Oxford, UK.

McArt, J. A. A., D. V. Nydam, G. R. Oetzel, T. R. Overton, and P. A. Ospina. 2013. Elevated non-esterified fatty acids and β-hydroxybutyrate and their association with transition dairy cow performance. Vet. J. 198:560–570.

Oetzel, G. R. 2007. Herd-level ketosis: Diagnosis and Agreement factors. Pages 67–91 in Proc. Am. Assoc. Bov. Pract., Vancouver, BC, Canada. Am. Assoc. Bovine Pract., Opelika, AL.

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rukkwamsuk, T., T.A.M. Kruip, and T. Wensing. 1999. Relationship between overfeeding and overconditioning in the dry period and the problems of high producing dairy cows during the postparturient period. Vet. Q. 21:71-77.

Sahinduran, S., K. Sezer, T. Buyukoglu, M. K. Albay, and M. C. Karakurum. 2010. Evaluation of some haematological and biochemical parameters before and after treatment in cows with ketosis and comparison of different treatment methods. J. Anim. Vet. Adv. 9:266-271.

Sevinç, M., A. Başoğlu, I. Öztok, M. Sandikçi, and F. Birdane. 1998. The clinical-chemical parameters, serum lipoproteins and fatty infiltration of the liver in ketotic cows. Turk. J. Vet. Anim. Sci. 22:443-448.Shaw, J.C. 1956. Ketosis in dairy cattle. A review. J. Dairy Sci. 39:402-434.

Shaw, J.C. 1956. Ketosis in dairy cattle. A review. J. Dairy Sci. 39:402-434.

Simensen, E., K. Halse, P. Gillund, and B. Lutnaes. 1990. Ketosis treatment and milk yield in dairy cows related to milk acetoacetate levels. Acta Vet. Scand. 31:433–440.

Sjollema, B., J. E. Van der Zande. 1923. Metabolism in Acetonemia of Milch Cows. J. Metab. Res. 4:525–533.

Staufenbiel, R. 1992. Energie- und Fettstoffwechsel des Rindes. Untersuchungskonzept und Messung der Rückenfettdicke. Mh. Vet. Med. 47:467–474.

Volek, J. S., D. J. Freidenreich, C. Saenz, L. J. Kunces, B. C. Creighton, J. M. Bartley, P. M. Davitt, C. X. Munoz, J. M. Anderson, C. M. Maresh, and E. C. Lee. 2016. Metabolic characteristics of keto-adapted ultra-endurance runners. Metabolism. 65:100-110.

Whitaker, D. A. 2004. Metabolic profiles. Pages 804–819 in Bovine Medicine, Diseases and Husbandry of Cattle. 2nd ed. A. H. Andrews with R. W. Blowey, H. Boyd, and R. G. Eddy, ed. Blackwell Publishing, Oxford, UK.

Zaaijer, D., and J. P. Noordhuizen. 2003. A novel scoring system for monitoring the relationship between nutritional efficiency and fertility in dairy cows. Ir. Vet. J. 56:145–151.

APPENDIX

Appendix. Description of the Rumen Fill scoring system quoted from Zaajer and Noordhuizen (2003)

Rumen Fill Score	Description[1]
1	The para lumbar fossa[2] cavitates more than a hand's width behind the last rib and also a hand's width inside under the transversal processes.
2	The para lumbar fossa cavitates a hand's width behind the last rib and to a lesser extent inside under the transversal processes.
3	The para lumbar fossa cavitates less than a hand's width behind the last rib and falls about a hand's width vertically downwards from the transversal processes and then bulges out.
4	The para lumbar fossa skin is covering the area behind the last rib and arches immediately outside below the transversal processes due to an extended rumen.
5	The rumen is quite distended and nearly obliterates the fossa; the last rib and the transversal processes are not visible.

The rumen fill scoring system was developed and described by Zaajer and Noordhuizen (2003); Scoring was performed when standing at the left hind side of the cow.

[1] Please refer to Zaajer and Noordhuizen (2003) for more information and example photographs

[2] The para lumbar fossa is between the last rib, the transversal processes and the hipbone

CHAPTER 4.2

Customized recommendations for production management clusters of North American automatic milking systems.

Journal of dairy science, 99(7), pp.5671-5680.
https://doi.org/10.3168/jds.2015-10153

Tremblay, M.,[*] Hess, J.P.,[*] Christenson, B.M.,[*] McIntyre, K.K.,[*] Smink, B.,[†] van der Kamp, A.J.,[‡] de Jong, L.G.[‡] and Döpfer, D.[*]

*Department of Medical Sciences, Food Animal Production Medicine Section, School of Veterinary Medicine, University of Wisconsin-Madison,
2015 Linden Drive, Madison 53706
†Lely North America, 775 250th Avenue, Pella, IA 50219
‡Lely International N.V., Cornelis van der Lelylaan 1, 3147 PB, Maassluis, the Netherlands

ABSTRACT
Automatic milking systems (AMS) are implemented in a variety of situations and environments. Consequently, there is a need to characterize individual farming practices and regional challenges to streamline management advice and objectives for producers. Benchmarking is often used in the dairy industry to compare farms by computing percentile ranks of the production values of groups of farms. Grouping for conventional benchmarking is commonly limited to the use of a few factors such as farms' geographic region or breed of cattle. We hypothesized that herds' production data and management information could be clustered in a meaningful way using cluster analysis and that this clustering approach would yield better peer groups of farms than benchmarking methods based on criteria such as country, region, breed, or breed and region. By applying mixed latent-class model-based cluster analysis to 529 North American AMS dairy farms with respect to 18 significant risk factors, 6 clusters were identified. Each cluster (i.e., peer group) represented unique management styles, challenges, and production patterns. When compared with peer groups based on criteria similar to the conventional benchmarking standards, the 6 clusters better predicted milk produced (kilograms) per robot per day. Each cluster represented a unique management and production pattern that requires specialized advice. For example, cluster 1 farms were those that recently installed AMS robots, whereas cluster 3 farms (the most northern farms) fed high amounts of concentrates through the robot to compensate for low-energy feed in the bunk. In addition to general recommendations for farms within a cluster, individual farms can generate their own specific goals by comparing themselves to farms within their cluster. This is very comparable to benchmarking but adds the specific characteristics of the peer group, resulting in better farm management advice. The improvement that cluster analysis allows for is characterized by the multivariable approach and the fact that comparisons between production units can be accomplished within a cluster and between clusters as a choice.

INTRODUCTION
Automatic milking systems (AMS) are increasing in popularity and number around the world (de Koning, 2010). As systems become more advanced under constraints of well-being, technical improvements, and economic feasibility, the variety in dairy management systems increases—from organic grazing to standard herds, from tie stalls to AMS, and from small family farms to large freestall herds. Even with the best technology in place, it is necessary to know one's strengths and weaknesses to make continuous improvements and set appropriate management and production goals. The dairy industry is similar to other production systems in which benchmarking is used to compare herds and motivate producers to set goals for their farm (Khade and Metlen, 1996; Boda, 2006; von Keyserlingk et al., 2012), but it is important for benchmarking to be based on the correct comparison group given the wide variety in the dairy industry.

Many dairy record systems, benchmarking programs, and benchmarking results have been published in non-peer-reviewed publications that enable producers to compare themselves with others and monitor their production progress. Benchmarking uses percentile ranks of the production values of groups of farms to compare farms within peer groups. However, grouping for conventional benchmarking is commonly limited to the use of a few factors such as farms' geographic region or breed of cattle. For example, the USDA's National Agricultural Statistics Service (NASS) summarizes yearly production by region or by herd size (USDA NASS, 2014). Similarly, the DHI's executive analysis "Udder Health Monitor" report compares a herd's SCC with that of herds of a similar size broken down into 3 groups: 1–199, 200–999, and >999 cows

(Dairy Records Management Systems, 2014). In addition, DHI's "Herd Management Comparison" report uses breed averages (Holstein or Jersey) by region and the industry's standard goals (Dairy Records Management Systems, 2014). More advanced programs, such as DairyMetrics, can be used to select smaller comparison groups but with the additional restriction items being limited to data found in DHI reports such as SCC and milking frequency (Dairy Records Management Systems, 2012). Specific to AMS, the social network "Benchmark" (Lely Industries N.V., Maassluis, the Netherlands) allows farmers to compare performance variables to that of others in their social network or from selecting others in the same region or with the same farm size. However, these benchmarking methods rely on personal judgment to create peer groups, and the restrictions used (e.g., country, breed, region, or breed and region) do not account for the wide range of systems and conditions in today's dairy industry.

In contrast to the previously mentioned methods, cluster analysis is used to make groups of similar observations that can be based on many different variables (Borcard et al., 2011). Brotzman et al. (2015)) used 16 performance values to cluster large Wisconsin dairies into 6 groups that were then characterized into best, good, and poor performance. In a similar industry, the dairy goat farming systems in Italy was successfully characterized into 3 major groups separated into 5 clusters using a cluster analysis of a variety of performance, facility, and management data (Usai et al., 2006). Clusters define neighbors not necessarily as geographic neighbors but neighbors in "similarity of farm characteristics."

Given the wide range in conditions in the dairy industry, to make comparison groups, many factors that significantly affect a herd's production ability need to be assessed simultaneously. In Brotzman et al. (2015)), many other limiting factors exist, although herd size was limited to those with at least 200 cows and some environmental variation was limited by only examining Midwestern US dairy herds. In addition, many factors unique to AMS that might affect production are not included in these aforementioned benchmarking and clustering methods. For example, traffic type and the number of robots per pen have been shown to significantly affect milk production in AMS farms (Tremblay et al., 2016). Also, some criteria, such as milking frequency (2 or 3 times per day), do not apply to AMS because cows in an AMS are free to regulate their milking frequency individually. In addition, most benchmarking tools are based on data collected via DHI databases, which is based on measurements taken only once every 3 to 4 wk. Automatic milking systems or parlor systems and sensor technology provides an opportunity to use results collected on a daily basis.

There is a need to compare AMS farms based on relevant variables in an unbiased fashion, which is not currently being provided for these specialized farms. The goals of this study were to characterize farming patterns of AMS herds to prioritize and customize advice for producers regarding their farm management. We hypothesized that herds' production data and management information could be grouped into meaningful multivariable clusters and that this clustering approach would produce better peer groups than conventional benchmarking methods that create peer groups based on criteria such as country, region, breed, or breed and region alone. The specific aim was to perform a cluster analysis of hundreds of North American AMS dairy farms with respect to significant risk factors identified by a generalized mixed linear model. Identifying a farm's nearest neighbor in terms of production patterns and management limitations would allow advice to be tailored to these modern specialized producers.

MATERIALS AND METHODS

A total of 529 North American dairy farms with Lely Astronaut AMS (Lely Industries N.V., Maassluis, the Netherlands) had weekly data collections for 4 yr (2011–2014), which produced 54,065 observations. A previous study found 20 variables from this data set to be significantly associated with changes in milk production (kg) using a generalized linear mixed regression model (Tremblay et al., 2016).

Of the 20 available variables, 5 were categorical variables. The numbers of farms per categorical variable levels and variable explanations are detailed in Table 1. Traffic type (i.e., how cows move through the pen among the AMS, freestalls, and feed fence) can be free or forced. With free cow traffic, cows decide when to enter the AMS, whereas with forced cow traffic, the producer creates one-way traffic toward the AMS. The variable Traffic_Type was coded as "free" or "forced." The Robots_per_Pen variable represented the number of robots per pen of cows. By default, this variable also represents the number of cows in a pen and the pen's physical dimensions. By design, each pen will have about 60 cows per robot. For example, Robots_per_Pen of "1" is designed with 1 robot in a pen of about 60 cows and Robots_per_Pen of "2" is designed with 2 robots in a pen of about 120 cows. Breed was categorized into 3 levels: "Holstein," "Jersey" and "other." Breed "other" represents all other breeds: Ayrshire, Brown Swiss, Guernsey, Red and White, crosses, mixed, and unknown. Farm_Goal was characterized either by the "quota" system for farms in Canada or "maximum production" for farms in the United States that produce with the goal of maximum milk production. Grazing and organic farms were not used in the previous analysis of this data set because they had relatively few observations. The New_Retro variable was either "new" for AMS robots that were installed in newly built barns or "retro" for AMS robots that were retrofitted in existing barns.

Table 1. The number of farms per categorical variable

Categorical Variable[1]	Levels	Number of farms[2]
Traffic_Type	Free	493
	Forced	36
Robots_per_Robots	1	295
	2	208
	3+	26
Breed	Holstein	473
	Jersey	15
	Other	41
Farm_Goal (country)	Quota (Canada)	350
	Max_Production (USA)	179
New_Retro (newly built barn or retro fitted)	New	266
	Retro	263

[1] Variable explanations: Traffic_Type= how cows are allowed to move among areas of a barn. "Free" refers to a system where cows can decide when to enter the AMS and can move freely between the AMS, free stalls and the feeding area. "Forced" traffic type uses a one-way traffic system towards the AMS; Robots_per_Pen= number of AMS robots per pen; Breed= breed of cattle; Farm_Goal= Operate under the "Quota" system for farms in Canada or "Max_Production" for farms in the USA that produce with the goal of maximum milk production; New_or_Retro= newly built or robots retro-fitted in an existing barn; [2] 529 total observations

Thirteen numeric AMS variables were available: Milk_Production_per_Cow_per_Day, Cows_per_Robot, Average_DIM, Concentrates, Rest_Feed, Refusals, Failures, Milkings, Milk_Speed, Boxtime, Connection_Attempts, Robot_Free_Time, and Days_Since_Installation. The variable explanations are presented in Table 2. The data set was previously limited to observations that had >10 Cows_per_Robot and <90 Cows_per_Robot. Observations with an Average_DIM greater than 365 d were also removed as outliers. All weekly numeric observations were averaged per farm to produce a final data set for clustering with one observation per farm (n = 529). The summary statistics and explanations of the numeric variables are shown in Table 2. All statistical analyses were done using the program R version 3.0.1 (R Development Core Team, 2013).

Table 2. Numeric variables explanation and descriptive statistics

Numeric variables	Variable explanation	Mean[1]	SD[2]
Milk_Production_per_Cow_per_Day	Average kg of milk produced[3]	31.82	4.37
Cows_per_Robot	Number of cows per robot	49.80	8.91
Average_DIM	Average days in milk of the herd	176.98	19.94
Concentrates	Average concentrate (kg) consumed in robot or automatic feeder per 100 kg of milk yield	15.64	4.93
Rest_Feed	Average percent of concentrates from the cow's allowance that was not dispensed that day (%)[4]	7.85	5.40
Refusals[3]	Average number of non-milking visits[3]	1.94	1.17
Failures[5]	Average number of failed milkings[5]	5.87	2.63
Milkings[3]	Average number of successful milkings[3]	2.90	0.28
Milk_Speed	Average speed of milk flow during the milking (kilogram per minute)	2.61	0.27
Boxtime[3]	Average minutes in the AMS[3] (milking time and treatment time)	6.78	0.55
Connection_Attempts[5]	Average number of times the robot arm moved up to get connect teats per milking[5]	1.42	0.20
Days_Since_Installation	how many days ago the automatic milking system was installed	683.98	839.81
Robot_Free_Time	percent of time per day the robot is not occupied by a cow[6]	19.42	11.23

[1] 529 total observations; [2] standard deviation; [3] per cow per day; [4] Possible causes include: a cow was not visiting the robot often enough or she was not able to finish her meal giving her milking time; [5] per robot per day; [6] The denominator does not include the time per day the system is automatically cleaning the robot and the milk lines to the tank

The variables Season and Record_Year were not included as they were not meaningful when working with farm as the unit of observation. A cluster analysis was performed using the 18 variables. Due to the mixture of continuous and categorical variables, 2 of which had more than 2 factor levels, a mixed latent-class model-based approach was chosen (Hennig, 2010). Another benefit of model-based clustering is that normalization and scale differences among variables do not affect the outcome (Vermunt and Magidson, 2002). The method was computed by the function flexmixedruns in the R package fpc (Hennig, 2010). Maximum likelihood estimation was used to determine the best model and the Bayesian information criterion (BIC) determined the best number of clusters. One hundred starts of the expectation-maximization (EM) algorithm with random initialization were compared during a sensitivity analysis to optimize the model for each number

of clusters between 2 and 20 (Hennig, 2010). Because only variables previously selected from Tremblay et al. (2016)) were used in the cluster analysis, and the number of variables was not larger than the number of observations, the mixed latent-class model-based approach did not require further variable selection (Dean and Raftery, 2010; Poon et al., 2010).

Categorical variables were examined per cluster, and the $\chi 2$ test was used to test for significant changes in proportions between variable levels in each cluster compared with the entire population. Averages of the numeric production variables were calculated from all farms in each cluster. Testing for significant differences between clusters based on variables that were used in the clustering is inappropriate because cluster analysis separates observations based on these variables (Legendre and Legendre, 2012). Therefore, we did not perform significance testing of the variables' means per cluster. Average values were ranked across clusters and color-coded for visual identification in a cross table. All of the farms represented in the final data set were mapped using a heat map to keep individual farm identity anonymous. Farms in each cluster were mapped to look for geographic patterns.

To simulate commonly applied grouping methods in the industry, 4 other grouping classification were assigned based on commonly used criteria: country, breed, region and a combination of breed and region. Regions were defined as Midwest, East, Northeast and Northwest. Midwest states included Iowa, Illinois, Indiana, Michigan, Minnesota, Missouri, North Dakota, Ohio, South Dakota, and Wisconsin. East states included Massachusetts, New York, Pennsylvania, Virginia, and Vermont. Northeast provinces included New Brunswick, Nova Scotia, Ontario, and Quebec. Northwest provinces included Saskatchewan, Manitoba, Alberta, and British Columbia. There were no Jersey farms in the Northwest region and only one Jersey farm in the Eastern region, which was reassigned to the Jersey Midwest group for the breed and region classification. In addition, there was only one "other" breed farm in the East region; therefore, it was reassigned to the "other" Midwest group. In the end, there were 9 breed and region groups.

The amount of milk produced per robot per day was not used for the cluster analysis; however, it is one of the major determinants of income for dairy farms including AMS farms. Thus, the external variable Milk_Production_per_Robot_per_Day was used for validation (Aldenderfer and Blashfield, 1984; Yang, 2012) when comparing the following grouping variables for predictions: farm clusters generated by cluster analysis and groups of farms based on the conventional benchmarking criteria (i.e., country, region, breed, and a combination of region and breed). The 5 grouping variables (see Table 5) were used to predict milk production per robot per day by means of a generalized linear regression model with the number of farms per group as an offset. The fit of the regression models and their predictive ability were compared among the 5 grouping methods in addition to the null model using these criteria: log-likelihood, BIC, Akaike information criterion (AIC), mean absolute error, and root mean square error.

RESULTS

The 5 categorical variables are described in Table 1. The 13 numeric variables used in the cluster analysis are described in Table 2. The cluster analysis resulted in a latent-class model with a log-likelihood of −9,817.368 with 203 df, an AIC of 20,040.74, and a BIC of 20,907.75. This model resulted in 6 clusters with an average of 88 farms per cluster (range: 50–124).

Table 3. Number of farms per cluster by categorical variable (n=529)

			Cluster					
Variable	Level[1]	Total	1	2	3	4	5	6
Traffic_ Type	Forced	36 (6.81%)[2]	7 (5.65%)[3]	11 [4] (22.00%)	1 [4] (1.12%)	7 (9.33%)	8 (6.67%)	2 (2.82%)
	Free	493 (93.19%)	117 (94.35%)	39 [4] (78.00%)	88 [4] (98.88%)	68 (90.67%)	112 (93.33%)	69 (97.18%)
Robots_ per_Pen	1	295 (55.77%)	41 [4] (33.06%)	31 (62.00%)	60 [4] (67.42%)	55 [4] (73.33%)	74 (61.67%)	34 (47.89%)
	2	208 (39.32%)	71 [4] (57.26%)	14 (28.00%)	27 (30.43%)	18 [4] (24.00%)	43 (35.83%)	35 (49.30%)
	3+	26 (4.91%)	12 [4] (9.68%)	5 (10.00%)	2 (2.25%)	2 (2.67%)	3 (2.50%)	2 (2.82%)
Breed	Holstein	473 (89.41%)	98 [4] (79.03%)	38 [4] (76.00%)	84 (94.38%)	68 (90.67%)	115 [4] (95.83%)	70 [4] (98.59%)
	Jersey	15 (2.84%)	5 (4.03%)	10 [4] (20.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
	Other	41 (7.75%)	21 [4] (16.94%)	2 (4.00%)	5 (5.62%)	7 (9.33%)	5 (4.17%)	1 [4] (1.41%)
Farm_ Goal	Quota	350 (66.16%)	64 [4] (51.61%)	45 [4] (90.00%)	89 [4] (100.0%)	57 (76.00%)	31 [4] (25.83%)	64 [4] (90.14%)
	Max_ Production	179 (33.84)	60 [4] (48.39%)	5 [4] (10.00%)	0 [4] (0.00%)	18 (24.00%)	89 [4] (74.17%)	7 [4] (9.86%)
New_ Retro	New	266 (50.28%)	55 (44.35%)	18 (36.00%)	56 [4] (62.92%)	34 (45.33%)	57 (47.50%)	46 [4] (64.79%)
	Retro	263 (49.72%)	69 (55.65%)	32 (64.00%)	33 [4] (37.08%)	41 (54.67%)	63 (52.50%)	25 [4] (35.21%)
Total		529	124	50	89	75	120	71

[1] Variable explanations: Traffic_Type= how cows are allowed to move among areas of a barn. "Free" refers to a system where cows can decide when to enter the AMS and can move freely between the AMS, lying stalls and the feeding area. "Forced" traffic type uses a one-way traffic system towards the AMS; Robots_per_Pen= number of AMS robots per pen; Breed= breed of cattle; Farm_Goal= Operate under the "Quota" system for farms in Canada or "Max_Production" for farms in the USA that produce with the goal of maximum milk production; New_or_Retro= newly built or retro fitted barn; Years_Since_Install= how recently (in years) the AMS was installed; Robot_Free_Time = percent of time per day the robot is not occupied; Record_Year= year at the time of record; Season= "Winter" was classified as December through February; "Spring" as March through May, "Summer" as June through August, and "Fall" was classified as September through November; [2] (%) Percent of total farms; [3] (%) Percent of farms per cluster; [4] significantly different compared to the total population

The distribution of herds among each categorical variable level was examined per cluster (Table 3). Several differences were found in the proportions of farms per variable level of each cluster compared with the overall population of 529 farms (Table 3). Compared with all the farms, cluster 1 had a higher proportion of farms with 2 or more robots per pen and farms with breeds other than Jerseys and Holsteins. Cluster 2 had a higher proportion of farms with forced traffic and Jerseys under the quota system compared with the entire population. Cluster 3 farms were exclusively under the quota system (Canadian). Cluster 4 had a high proportion of farms with 1 robot per pen

compared with all farms. Cluster 5 had a high proportion of Holstein and maximum production farms in the United States, whereas cluster 6 had a high proportion of newly built barns and Holstein farms under the quota system compared with the entire population. See Table 3 for all results.

Table 4. Averages and standard deviation (SD) of 13 numeric variables for all farms in each cluster

Variable	Cluster 1 (n= 124)[1]	2 (n=50)	3 (n=89)	4 (n=75)	5 (n=120)	6 (n=71)
Milk_Production_per_Cow_per_Day	31.10 ± 3.49	25.04 ± 3.85	32.21 ± 2.9	29.71 ± 3.15	35.52 ± 2.65	33.30 ± 3.88
Cows_per_Robot	49.89 ± 6.88	48.54 ± 11.09	48.84 ± 5.62	55.01 ± 6.86	55.82 ± 4.11	36.01 ± 4.98
Boxtime	6.61 ± 0.42	6.57 ± 0.64	6.54 ± 0.37	7.24 ± 0.51	7.07 ± 0.44	6.49 ± 0.54
Robot_ Free_Time	21.55 ± 8.09	31.04 ± 12.61	17.02 ± 5.61	9.90 ± 4.69	11.56 ± 4.82	33.90 ± 9.77
Milk_Speed	2.63 ± 0.24	2.32 ± 0.27	2.55 ± 0.19	2.44 ± 0.28	2.79 ± 0.20	2.68 ± 0.23
Milkings	2.89 ± 0.24	2.68 ± 0.36	3.11 ± 0.20	2.70 ± 0.20	2.82 ± 0.20	3.15 ± 0.23
Refusals	1.84 ± 0.67	2.51 ± 1.33	2.64 ± 0.92	1.48 ± 0.71	1.01 ± 0.35	2.89 ± 1.71
Failures	7.73 ± 2.26	9.09 ± 3.70	5.00 ± 1.28	5.07 ± 2.21	4.83 ± 1.57	4.03 ± 1.32
Connection_ Attempts	1.51 ± 0.08	1.28 ± 0.58	1.40 ± 0.07	1.40 ± 0.10	1.42 ± 0.09	1.41 ± 0.10
Concentrates	14.77 ± 2.24	17.37 ± 5.98	18.93 ± 7.18	17.53 ± 5.78	13.34 ± 1.86	13.75 ± 3.14
Rest_Feed	6.98 ± 3.19	11.44 ± 8.32	8.17 ± 3.77	10.63 ± 7.22	5.84 ± 2.77	6.91 ± 6.35
Average_DIM	179.84 ± 20.65	187.42 ± 24.56	174.11 ± 16.81	181.98 ± 24.79	170.04 ± 14.18	174.69 ± 16.20
Days_Since_ Installation	305.51 ± 344.63	1807.64 ± 1781.79	521.60 ± 445.28	1207.32 ± 720.89	543.10 ± 380.02	442.51 ± 518.57

Darker gray or white shading indicate extremely high or low values per column. In general, lighter shading means preferred averages, and darker shading means less preferred averages but these typical conventions might not be true for each clusters' situation. [1] total number of observations per cluster (one observation per farm)

The average numeric variables values of each farm are shown per cluster in Table 4. Cluster 1 had the greatest Connection_Attempts and the most recent installation of AMS robots. Cluster 2 had the lowest average of Milk_Production_per_Cow_per_Day, Milkings, Milk_Speed, and Connection_Attempts. The greatest average of Rest_Feed, Failures, Days_Since_Installation, and Average_DIM were present in cluster 2. Cluster 3 had the greatest average of Concentrates, and

cluster 4 had the greatest average of Boxtime and lowest average of Robot_Free_Time. Milk_Production_per_Cow_per_Day, Milk_Speed, and Cows_per_Robot had the greatest average in cluster 5. Cluster 5 herds had the lowest average of Concentrates, Rest_Feed, Refusals, and Average_DIM. Cluster 6 had the lowest Cows_per_Robot, Boxtime, and Failures; Robot_Free_Time, Milkings, and Refusals had lowest average in cluster 6.

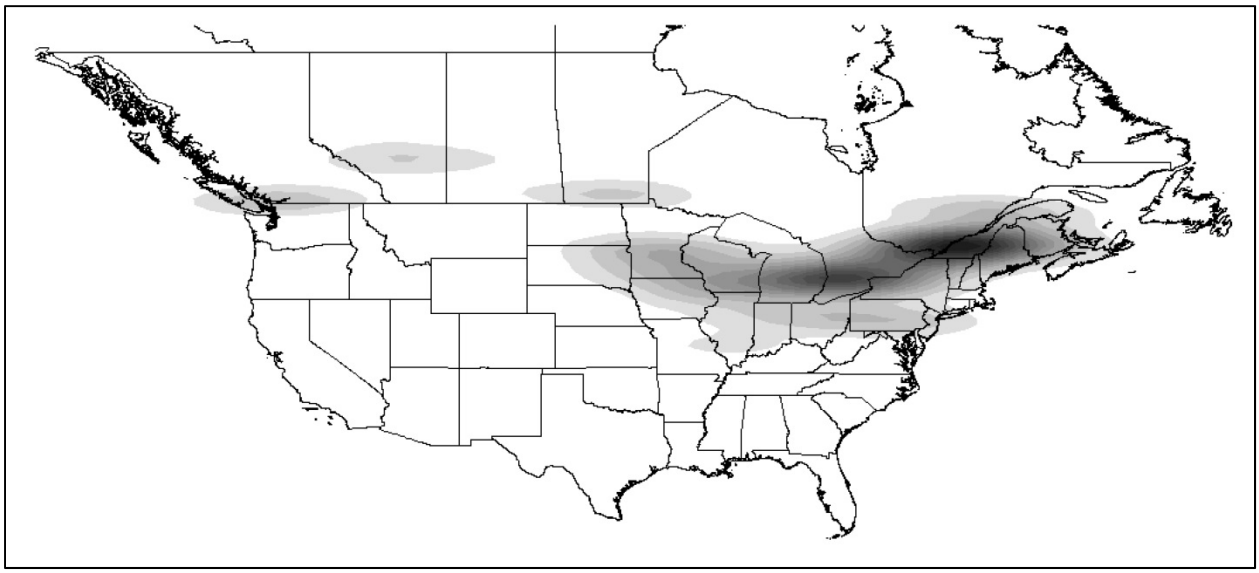

Figure 1. Heat map of all farm locations (n = 529)

A map of the 529 farms is shown in Figure 1. The highest concentrations of farms were in the Midwest United States (Minnesota and Wisconsin), southern Ontario (between Detroit and Toronto), and lower Quebec (between Ottawa and Quebec City). Other concentrations of farms were located outside Vancouver, outskirts of Winnipeg, between Edmonton and Calgary, and eastern Pennsylvania.

Cluster 5 herds were mainly located in the Midwest (see Figure 2E). Cluster 2 and 6 farms were centered in the east, whereas cluster 1 and 4 farms had an even distribution in all concentrations of farm locations (see Figure 2A and D). The most northern cluster of the study group was cluster 3 (see Figure 2C).

Table 5 describes the results of the 5 regression models predicting production per robot per day using the grouping methods as independent variables. The cluster analysis external validation model had the best fit with the largest log-likelihood, and lowest residual deviance, AIC, BIC, mean absolute error, and root mean square error.

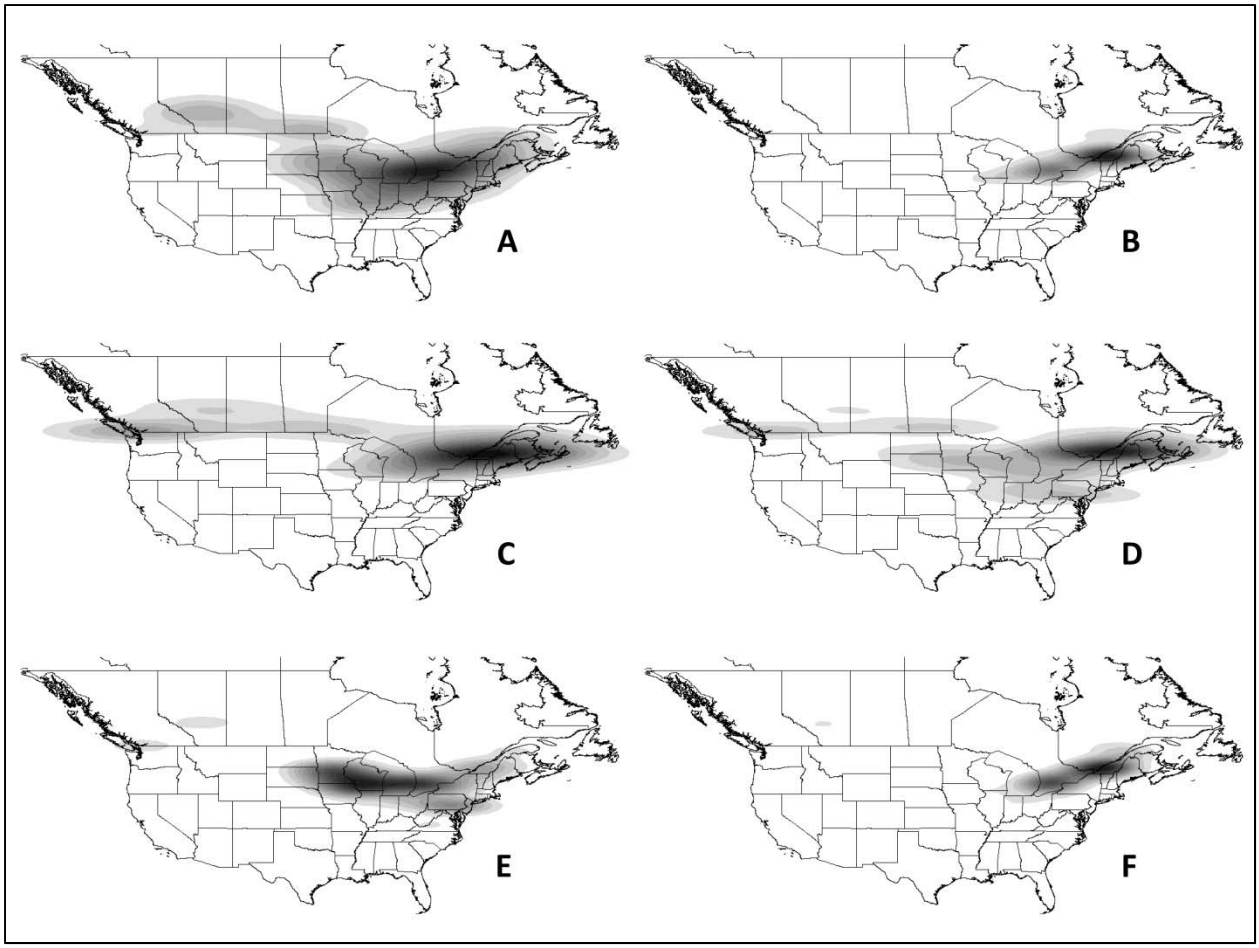

Figure 2. Heat map of farm location by cluster: (A) Cluster 1 (n=124), (B) Cluster 2 (n=50), (C) Cluster 3 (n=89), (D) Cluster 4 (n=75), (E) Cluster 5 (n=120), (F) Cluster 6 (n=71)

Table 5. Goodness of fit parameters and predictive ability of the linear regression models

Model [1]	df [2]	logLik [3]	Res. Dev. [4]	AIC [5]	BIC [6]	MAE [7]	RMSE [8]
Cluster	523	-3659.46	31594133	7332.9	7362.8	192.5	244.4
Breed and Region	520	-3806.23	55030856	7632.5	7675.2	262.7	322.5
Region	525	-3816.24	57153025	7642.5	7665.2	268.5	328.7
Breed	526	-3861.71	67872296	7731.4	7748.5	293.2	358.2
Country	527	-3827.56	59651230	7661.1	7673.9	271.9	335.8
Null Model [9]	528	-3867.35	69336252	7738.7	7747.3	297.2	362.0

[1] Linear regression model with Milk_Production_per_Robot_per_Day (kilograms) as the dependent variable, comparison groups (Cluster, Breed and Region, Region, Breed, Country) as a discrete independent variable, and an offset as the number of farms per comparison group; [2] Degrees of freedom; [3] -2 x log likelihood; [4] residual deviance; [5] AIC = *Akaike information criterion;* [5] BIC = *Bayesian information criterion;* [7] MAE = Mean absolute error, [8] RMSE= Root mean square error; [9] Milk_Production_per_Robot_per_Day ~ 1.

DISCUSSION

The need to compare industry standards for AMS is great because of the investments made in these specialized systems. Current benchmarking tools do not compare AMS based on all relevant variables or lend specialized advice based on multiple variables simultaneously. Eighteen variables significantly associated with milk production per day (Tremblay et al., 2016) were used to cluster 529 automatic milking farms. We compared the predictive ability of 6 farm clusters generated by cluster analysis to groups of farms based on the conventional benchmarking criteria (i.e., country, region, breed, and a combination of region and breed). Cluster analysis comparison groups were better at predicting milk production per robot than benchmarking comparison groups. Better AMS peer groups allow for improved comparison within groups because farms within clusters are more similar compared with the general average across all farms. Allowing farms to set appropriate goals according to individual situations (e.g., recent robot installation, environmental, facilities constraints) minimizes the potential for goals to be unrealistic. Each of the 6 clusters had different farm characteristics and therefore the clusters can benefit from different recommendations following on the priorities of the cluster member herds. Next, we will discuss the characteristics and recommendations per cluster.

Cluster 1

On average, cluster 1 farms had the most recent robot installations (see Table 4). Recent installation has been shown to significantly decrease milk production compared with systems that have been in place for >4 yr (Tremblay et al., 2016). As cluster 1 herds become established during their start-up period, they will need to continue selecting cows that are best for AMS milking.

Cluster 1 farms represented breeds other than Holstein (21/41 "other" breed farms, 5/15 Jersey breed farms) with high average Failures and Connection_Attempts (see Table 4). Although a previous analysis of this data set did not find a significant difference in Milk_Production_per_Cow_per_Day between the "other" and "Holstein" breed groups, this might be a result of all the other breeds having been grouped together (Ayrshire, Brown Swiss, Holstein Crosses, Guernsey, Red and White, mixed; Tremblay et al., 2016). When breeds were examined separately, average milk yield varied among breeds (Cerbulis and Farrell, 1975; VanRaden and Sanders, 2003; USDA-AIPL, 2013). Ayrshire, Jersey, and Holstein breeds were also found to vary in milking speed and milking temperament (Sewalem et al., 2010). In addition, the difference between small breeds (Jersey, Guernsey, Ayrshire) and large breeds (Holstein, Brown Swiss) could affect how cows fit and align in a milking robot, and their difference in udder conformation could make one breed, such as a large or small breed, more prone to failed connections (Norman et al., 1988; Rodenburg, 2002; Capper et al., 2009). Therefore, it might not be advisable to formulate production goals for all "other" breed herds based on subsets that include "Holstein" breed herds. It is recommended that cluster 1 herds examine whether their settings are correctly adjusted according to their non-Holstein cows.

Cluster 2

Cluster 2 consisted of farms with the lowest Milk_Production_per_Cow_per_Day and Milk_Speed. Cluster 2 also had the highest average Failures (see Table 4). This cluster had 10 of the 15 Jersey breed farms. Jersey farms most likely fit in best with low-producing Holstein herds because of their lower milk yield on an equal boxtime, which results in lower average milk speed (Prendiville et al., 2010; Tremblay et al., 2016). These low-production Holstein farms in cluster 2

could be characterized as farms with mediocre management, farms that have made a clear choice for low input, or farms where the major income is from something other than dairy production.

Cluster 2 farms had the highest proportion of forced Traffic_Type. Forced traffic type has been shown to result in significantly less milk per cow and per robot per day compared with free traffic type (Tremblay et al., 2016). The recommendation for these farms is to check with their robot consultant whether it is feasible to open up all space to the robots and change the gates in the barn to move into a free cow traffic system.

If a cluster 2 farm is faced with high Rest_Feed, they should assess feed allowance settings to ensure that each cow is allotted sufficient milkings per day. This is important to allow enough time to finish their concentrate in the robot. These farms may also need to increase the density of the concentrate feed (for faster energy intake) or add additional feed dispensers in the robot for high-density concentrates.

Cluster 3

Cluster 3 farms fed high levels of concentrates and were generally the most northern farms of the study group (see Figure 2C). Their environment and feed availability offer unique challenges, as they can offer only basic forages at the feed bunk compensated with additional higher concentrate feed in the robot; therefore, they should be compared with other farms facing similar challenges. It is especially important for cows in these farms to receive all of their allotted concentrate. These farms could add separate automatic feeding stations outside of the robot so that cows can finish their allowance in between milkings.

Cluster 4

A high proportion of cluster 4 farms had only 1 robot per pen. These farms need to make sure that the downtime or inaccessibility of the robot is minimized, whereas farms with 2 or more robots per pen can shut down one robot for daily maintenance and keep milking cows in the other robot (Tremblay et al., 2016). Also, a single robot per pen has a greater effect on timid cows compared with 2 robots because a single robot does not allow timid cows additional opportunities for milking when dominant cows crowd the single robot. Timid cows have been shown to wait longer to go to the robot compared with higher-ranking cows (Ketelaar-de Lauwere et al., 1996; Thune et al., 2002; Melin et al., 2006), and the presence of a single robot does not allow these cows alternative opportunities for milking. This effect on a small number of cows can severely affect the overall average of the herd's AMS variable values. Advice for cluster 4 farms includes identifying individual cows that are not suited for AMS milking (e.g., too timid for milking in a pen with a single robot or need to be fetched often).

Cluster 5

Compared with the entire population of 529 farms, cluster 5 had a higher proportion of Holstein breed farms, farms located in the Midwest, and farms with a Farm_Goal of maximum production (see Table 3 and Figure 2E). They had, on average, the highest Milk_Production_per_Cow_per_Day of all the clusters (see Table 4). This cluster was the most intense in terms of Cows_per_Robot but continued to meet the recommendations for average milkings of >2.6 milkings per cow per day (Sitkowska et al., 2015). As the highest producing farms, cluster 5 farms had well-run AMS in place. Customized advice to cluster 5 should consist

of small adjustments; for example, decreasing Failures, because failures can disturb a cow's time budget (Stefanowska et al., 2000). Milking failures lead to interrupted milkings and have been shown to cause milk leakage, a potential risk factor for mastitis (Elbers et al., 1998; Stefanowska et al., 2000). In addition, these farms will need to select cows for high milk speed and cow-robot efficiency to gain milk production.

Cluster 6

Clusters 6 farms had high Milk_Production_per_Cow_per_Day and Milk_Speed averages but had a low average number of Cows_per_Robot (see Table 4). The low number of Cows_per_Robot in cluster 6 could reflect low available milk quotas, given that cluster 6 had a higher proportion of farms located in Canada compared with the entire population (Table 3). A low ratio of Cows_per_Robot negatively affects Milk_Production_per_Robot_per_Day and Boxtime while increasing Refusals, because there is a large surplus of robot capacity (Tremblay et al., 2016). Although extreme values exist, secondary to their values of Cows_per_Robot, all other values were well controlled and they had the lowest average Failures of all clusters. Advice for cluster 5 and cluster 6, given the limited available quota, would focus on optimizing the efficiency of milk production by producing more milk with less cows and resources such as feed, cost of operation, and reproduction.

Although the 18 variables used for the clustering of farms were readily available through the Lely T4C (Time for Cows) herd management system (Lely Industries N.V.) and represent important performance indicators for dairy farms, other variables could be included in a larger scale study. For instance, the many breed variations mentioned above suggest that breaking down the "Breed" variable would add to the analysis. Variables pertaining to animal health, herd genetics, farm economics, reproduction, facilities, and feeding management could be examined as potential risk factors for different production levels. Including the grazing and organic farms and data from other AMS companies and countries would also broaden the impact of a future data analysis project. In addition to increasing the number of risk factors included in the cluster analysis, follow-up surveys could be used to analyze each cluster in more depth for the sake of validating the outcomes of the current study.

Adding these techniques to current benchmarking and management tools available to individual farms would greatly benefit the farms' management but would also raise some challenges. For example, missing data would need to be addressed to ensure quality without eliminating participants from the analysis. Data imputation techniques could be used to estimate values for missing data. Some degree of misclassifications (i.e., assigning a farm to an inappropriate cluster) will always be present when clustering is applied, irrespective of the choice of technique. Also, as data sets evolve with time, the most appropriate clustering technique might change. The clusters describe the average farm within a cluster, and caution must be taken in interpreting the results of this study to one individual farm. The current results regarding characteristics of the clusters are meant as decision aids and orientation for customized expert advice in the field of AMS dairy herds.

CONCLUSIONS

A cluster analysis of 529 North American AMS herds with respect to significant predictors for milk production identified 6 clusters of production patterns and management characteristics.

Unlike current benchmarking grouping techniques, cluster analysis produces more appropriate peer groups among diverse farms. Each cluster exhibited a unique multivariable production pattern and management style that can result in distinct recommendations per farm. In addition, farms can set realistic goals according to comparisons within each cluster.

ACKNOWLEDGMENTS

We thank Lely North America (Pella, IA) for their financial support of this study. We gratefully acknowledge the expertise of Rik van der Tol (Lely Industries N.V., Maassluis, the Netherlands) for his assistance in the planning of this project.

REFERENCES

Aldenderfer, M. S., and R. K. Blashfield. 1984. Cluster Analysis. Quantitative Applications in the Social Sciences. Vol. 44. Sage Publications, Beverly Hills, CA.

Boda, G. 2006. Benchmarking dairy information using interactive visualization for dairy farm decision making. MSc Thesis. Department of Animal Science, McGill Univ., Canada.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical Ecology with R. Springer, New York, NY.

Brotzman, R. L., N. B. Cook, K. Nordlund, T. B. Bennett, A. Gomez Rivas, and D. Döper. 2015. Cluster analysis of Dairy Herd Improvement data to discover trends in performance characteristics in large Upper Midwest dairy herds. J. Dairy Sci. 98:3059–3070. http://dx.doi.org/10.3168/jds.2014-8369.

Capper, J. L., R. A. Cady, and D. E. Bauman. 2009. The environmental impact of dairy production: 1944 compared with 2007. J. Anim.Sci. 87:2160–2167. http://dx.doi.org/10.2527/jas.2009-1781.

Cerbulis, J., and H. M. Farrell Jr.. 1975. Composition of milks of dairy cattle. I. Protein, lactose, and fat contents and distribution of protein fraction. J. Dairy Sci. 58:817–827. http://dx.doi.org/10.3168/jds.S0022-0302(75)84644-3.

Dairy Records Management Systems. 2012. Dairy Metrics. Accessed May 2, 2015. http:// www.drms.org/dairymetricsinfo.aspx?node_id=Dflt6.

Dairy Records Management Systems. 2014. DHI Report Options. Accessed July 16, 2015. http://www.drms.org/pdf/materials/optman.pdf.

de Koning, C. J. A. M. 2010. Automatic milking—A common practice on dairy farms. Pages 52–67 in Proc. First North American Conference on Precision Dairy Management, Toronto, Canada. Omnipress, Madison, WI.

Dean, N., and A. E. Raftery. 2010. Latent class analysis variable selection. Ann. Inst. Stat. Math. 62:11–35. http://dx.doi.org/10.1007/s10463-009-0258-9.

Elbers, A. R. W., J. D. Miltenburg, D. De Lange, A. P. P. Crauwels, H. W. Barkema, and Y. H. Schukken. 1998. Risk factors for clinical mastitis in a random sample of dairy herds in the southern part of the Netherlands. J. Dairy Sci. 81:420–426. http://dx.doi.org/10.3168/jds.S0022-0302(98)75592-4.

Hennig, C. 2010. fpc: Flexible Procedures for Clustering. Version 2.0.3. https://cran.r-project.org/web/packages/fpc/index.html.

Ketelaar-de Lauwere, C. C., S. Devir, and J. H. M. Metz. 1996. The influence of social hierarchy on the time budget of cows and their visits to an automatic milking system. Appl. Anim. Behav. Sci. 49:199–211.

Khade, A. S., and S. K. Metlen. 1996. An application of benchmarking in the dairy industry. Benchmark. Qual. Manag. Technol. 3:34–41. Legendre, P., and L. F. Legendre. 2012. Numerical Ecology. Vol. 24. Elsevier, Oxford, UK.

Melin, M., G. G. N. Hermans, G. Pettersson, and H. Wiktorsson. 2006. Cow traffic in relation to social rank and motivation of cows in an automatic milking system with control gates and an open waiting area. Appl. Anim. Behav. Sci. 96:201–214.

Norman, H. D., R. L. Powell, J. R. Wright, and B. G. Cassell. 1988. Phenotypic and genetic relationship between linear functional type traits and milk yield for five breeds. J. Dairy Sci. 71:1880–1896. http://dx.doi.org/10.3168/jds.S0022-0302(88)79758-1.

Poon, L. K. M., N. L. Zhang, T. Chen, and Y. Wang. 2010. Variable selection in model-based clustering: To do or to facilitate. Pages 887–894 in Proc. 27th International Conference on Machine Learning. ACM, New York, NY.

Prendiville, R., K. M. Pierce, and F. Buckley. 2010. A comparison between Holstein-Friesian and Jersey dairy cows and their F1 cross with regard to milk yield, somatic cell score, mastitis, and milking characteristics under grazing conditions. J. Dairy Sci. 93:2741– 2750. http://dx.doi.org/10.3168/jds.2009-2791.

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rodenburg, J. 2002. Robotic milkers: What, where . . . and how much!!?? Pages 1–18 in Proc. Ohio Dairy Management Conf., Columbus, OH. Ohio State University Extension, Columbus.

Sewalem, A., F. Miglior, and G. J. Kistemaker. 2010. Analysis of the relationship between workability traits and functional longevity in Canadian dairy breeds. J. Dairy Sci. 93:4359–4365. http://dx.doi.org/10.3168/jds.2009-2969.

Sitkowska, B., D. PiwczyĔski, J. Aerts, and M. WaĞkowicz. 2015. Changes in milking parameters with robotic milking. Arch. Anim. Breed. 58:137–143. http://dx.doi.org/10.5194/aab-58-137-2015.

Stefanowska, J., M. Plavsic, A. H. Ipema, and M. M. W. B. Hendriks. 2000. The effect of omitted milking on the behaviour of cows in the context of cluster attachment failure during automatic milking. Appl. Anim. Behav. Sci. 67:277–291. http://dx.doi.org/10.1016/S0168-1591(00)00087-3.

Thune, R. Ø., A. M. Berggren, L. Gravås, and H. Wiktorsson. 2002. Barn layout and cow traffic to optimise the capacity of an automatic milking system. Pages 45–50 in Proc. 1st N. Am. Conf. Robotic Milking, Toronto, Canada. J. McLean, M. Sinclair, and B. West, ed. Wageningen Press, Wageningen, the Netherlands.

Tremblay, M., J. P. Hess, B. M. Christenson, K. K. McIntyre, B. Smink, A. J. van der Kamp, L. G. de Jong, and D. Döpfer. 2016. Factors associated with increased milk production for automatic milking systems. J. Dairy Sci. 99:3824–3837. http://dx.doi.org/10.3168/jds.2015-10152.

Usai, M. G., S. Casu, G. Molle, M. Decandia, S. Ligios, and A. Carta. 2006. Using cluster analysis to characterize the goat farming system in Sardinia. Livest. Sci. 104:63–76. http://dx.doi.org/10.1016/j.livsci.2006.03.013.

USDA-AIPL (Animal Improvement Program Laboratories). 2013. USDA summary of 2013 herd averages (DHI report K-3). Accessed July 16, 2015. https://www.cdcb.us/publish/dhi/dhi14/hax.html.

USDA-NASS (National Agricultural Statistics Service). 2014. Quick Stats Database. Accessed July 16, 2015. http://quickstats.nass.usda.gov/.

VanRaden, P. M., and A. H. Sanders. 2003. Economic merit of crossbred and purebred US dairy cattle. J. Dairy Sci. 86:1036–1044. http://dx.doi.org/10.3168/jds.S0022-0302(03)73687-X.

Vermunt, J. K., and J. Magidson. 2002. Latent class cluster analysis. Pages 89–106 in Applied Latent Class Analysis. Vol 11. J. A. Hagenaars and A. L. McCutcheon. Cambridge University Press, New York, NY.

von Keyserlingk, M. A. G., A. Barrientos, K. Ito, E. Galo, and D. M. Weary. 2012. Benchmarking cow comfort on North American freestall dairies: Lameness, leg injuries, lying time, facility design and management, for high producing Holstein dairy cows. J. Dairy Sci. 95:7399–7408. http://dx.doi.org/10.3168/jds.2012-5807.

Yang, R. 2012. A hierarchical clustering and validity index for mixed data. PhD Thesis. Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames.

CHAPTER 5- GENERAL DISCUSSION

SUMMARY

This dissertation addresses how analytical methods should be optimized for analyzing imperfect data to address their associated challenges, while still benefitting from the use of secondary data. This work developed methods for systematically comparing and selecting the most appropriate statistical methods available to optimize selected performance measures, streamline selection, and to prevent subjectivity, bias and flawed outcomes. Collectively, these methods provide an unbiased framework in which to approach the analysis of large data, without losing the emphasis on the biological relevance and sound interpretation of the results. This work is a steppingstone towards automation and towards ensuring that future data analyses are not hindered by large data imperfections. Instead, this work ensures that having many possible methods available from statistics and data sciences becomes an advantage and not a hurdle. Finally, this work suggests that automation of data analysis can coincide with a focus on biological relevance and sound interpretation.

APPROACH

A large variety of data sets from secondary data-based studies that were available for analysis were used to reach the objectives. A range of methods from various disciplines were applied, and new methods and protocols were developed to address imperfect data challenges in these large, imperfect data sets from secondary data-based studies. The studies had diverse goals that fit into three main areas: parameter estimation (descriptive modeling), prediction modeling, and pattern discovery. The data sets used, the methods applied and the goals for each study were representative of what is needed by epidemiology.

MAJOR FINDINGS

Large epidemiological data sets from secondary data-based studies share many similarities among each other and with primary data-based studies in terms of imperfect data challenges and goals. This work is a foundation for a future systematic approach that addresses many potential data imperfections with methods addressed in this work including: imputing missing values, modeling zero inflated data sets, systematically selecting interaction terms, variable selection, addressing imbalances in positive and negative outcomes, rare events, data with hierarchical structure, and using the example of principal component analysis (PCA) for variable reduction, and clustering for pattern recognition. A systematic approach to comparing goodness of fit of parameter estimation models would make many large datasets more manageable and informative for decision-making processes avoiding modeling bias. This work illustrates the potential and need for automated data preparation and model selection. However, this work also illustrates that automation of data analysis would still require in-depth interpretation of the biological significance of the results.

Different methods developed to address the same need can have significantly different performances. This work illustrates the importance of comparing performance and model fit among different methods to obtain the best and improved results when analyzing imperfect data. The rtFMS method removed selection bias and enabled the selection of the best performing model by comparing all available options and combinations of method options. This method was built

with the intention of becoming part of a future automated full model selection process that will help remove selection bias from prediction modeling as it selects the best performing model. This work left no doubt that the current method of empirical method selection is deficient and should be replaced by a more objective and systematic approach as shown in this dissertation.

Finally, this work illustrated how unsupervised learning can be a key step for the understanding of a data set's characteristics before engaging in a supervised learning task. Therefore, unsupervised learning could be included in systematic approaches to data analysis. However, post-hoc analyses of unsupervised learning results and the careful interpretation of the results also need to be included in the systematic approach to unsupervised learning methods.

MAIN FINDINGS PER CHAPTER

In Chapter 2.1 a study was presented that applied a systematic approach to zero augmented models. This was done to address possible zero inflation in a surveillance data set from the Foodborne Diseases Active Surveillance Network (FoodNet) in the United States. The goal was to build a descriptive model for *Campylobacter* infections. Several common types of zero-augmented models (i.e., Hurdle, zero-inflated models) were compared to each other and to a nonzero-augmented negative binomial model. The 5 models compared for this study were ranked and compared using a likelihood ratio test and Vuong BIC non-nested hypothesis test statistic. The results showed that the nonzero-augmented negative binomial model was the better fitting model. The systematic approach to dealing with a data sets with zero-inflation in this study was able to rank models in terms of fit and also demonstrate the lack of zero-inflation. This approach addressed the possibility of zero inflation without neglecting other modeling options. This was demonstrated using a dataset that had not been analyzed previously using routine methods due to the fear of zero-inflation. Although only 5 models were compared, the systematic approach to comparing goodness of fit of parameter estimation models is a foundation for the automation of systematic comparisons.

In Chapter 2.2 a study was presented that developed the first systematic approach to addressing missing data, high dimensionality and high correlation among variables in the same data. The data used in this study originated from the People, Animals and their Zoonoses (PAZ) project out of Kenya. These data were a good representation of data produced by many disciplines to which this methodology could be applied. The goal for these data was to build a description model for *Plasmodium falciparum* infection. The systematic approach developed combined multiple methods needed to address the challenges faced in this data set including imputation of missing data, variable extraction using PCA, and variable reduction and selection using an elastic-net regularized generalized linear model (glmnet). The sequential and parallel application of methods was successful in reducing a wide, sparse dataset with 1376 variables to a more useful, simplified set of 42 predictors for *Plasmodium falciparum* infection prevalence and producing socioeconomic wealth indices from many highly correlated variables. The protocol's flexibility and ability to accommodate other additional methods within its approach suggests that it may be easily applied to a variety of other imperfect data. This approach addressed several imperfect data challenges while still benefitting from the large amount of data in this data set.

In Chapter 2.3 a study was presented that applied a systematic selection process and systematic interpretation of interaction terms. The methods addressed an overwhelming amount of significant interactions while in search for the best fitting parameter estimation model. This was demonstrated

using production data from North American automated milking systems with the goal of building two descriptive models of milk production. The 2-way interactions were selected using forward selection with a t-value limit of 4. A total of 20 and 22 interactions were included in the final models. The interactions significantly increased the fit of the model and led to a very meaningful discussion on the interactions among variables. This method of selecting interactions could be beneficial to many projects to improve their models' fit and improve their understanding of the complex relationships in the data. This work illustrated the potential and need for automated model selection. Finally, this work also illustrated that automation of variable selection still requires in-depth interpretation of the biological significance of the results.

In Chapter 3.1 a study was presented that described a novel systematic approach to full model selection for prediction modeling using regression trees. The method was demonstrated using a data set comprised of data from milk Fourier-transform infrared spectroscopy (FTIR), routine milk testing, and from automatic milking systems to predict blood nonesterified fatty acids (NEFA) and β-hydroxybutyrate acid (BHBA) in dairy cows during early lactation. Regression tree full model selection (rtFMS) constructs a model for every combination of predictive modeling method options under consideration. The iterated, cross-validation performances of these models are then passed through a regression tree for selection of a final model. Using rtFMS, a significantly better performing model was obtained compared to what would have been obtained if method options had been selected subjectively. In addition, rtFMS provides simplicity and structure to FMS to improve and optimize prediction model performance. Finally, rtFMS eliminates the bias associated with empirical selection of method options. This method was built with the intention of becoming part of a future automated full model selection process that will help remove selection bias from prediction modeling as it selects the best performing model.

In Chapter 3.2 a study was presented that described a novel micro-macro multilevel modeling method using a data set of antibiotic residues in bovine milk. Micro-macro modeling methods are applicable when the data set has predictors on the individual (micro) level while the outcome variable is on the population (macro) level. The new method presented in Chapter 3.2 is called extreme value micro-macro (EVMM) multilevel modeling. It was developed to address when the central tendency of the micro level observations is not a good representation of the macro level outcome. In the case of EVMM, extreme values of the micro level observations are used to predict macro level outcomes. Two micro-macro modeling methods were compared using rtFMS. For the antimicrobial residues in milk data set, the EVMM method performed significantly better compared to the current micro-macro multilevel modeling method that uses the mean of the micro level observations. The EVMM method will allow more secondary data sets that combine data from multiple levels to be used to build better performing prediction models.

In Chapter 4.1 a study was presented that described how to gain meaningful results without the benefit of an outcome variable. It was demonstrated using clinical data, blood samples and milk testing data of Simmental cows in Bavaria. The goal was to reexamine the classification of metabolic adaptation in dairy cows. To validate the findings of the cluster analysis, post-hoc regression models were used. The post-hoc regression analysis examined associations between the observations' cluster classification and other clinical, milk and blood parameters that were not used in the cluster analysis. The post-hoc findings supported and aided in the interpretation of the clustering results. The cluster analysis with the addition of these post-hoc steps led to the

description of a novel syndrome for poor metabolic adaptation in dairy cows called "Poor Metabolic Adaptation Syndrome" (PMAS). These methods were able to describe a novel classification of data without the benefit of an outcome variable. In addition, an indicator blood parameter was able to be selected to classify future observations into the aforementioned metabolic health classifications. Finally, this work illustrated how unsupervised learning can be a key step for the understanding of a data set's characteristics before engaging in a supervised learning task.

When having access to large data, it might be instinctual to group all the data together during the analysis to gain extra statistical power. However, it might be more beneficial to separate data into different groups. Again, without a defined outcome variable for grouping data unsupervised learning is necessary. In Chapter 4.2 a study is presented that describes how decision making processes and customized management advice can be facilitated by improved benchmarking within peer groups through clustering. By applying mixed latent-class model-based cluster analysis to 529 North American automated milking systems (AMS) dairy farms with respect to 18 significant risk factors, 6 clusters were identified. The resulting clusters of data represented groups of farms with unique management styles, challenges, and production patterns. When compared with peer groups based on criteria similar to the conventional benchmarking standards, the 6 clusters better predicted milk production per robot per day. Therefore, separating data into subsets made for better results and subsequent recommendations to the producers. This study highlights that large data do not always benefit from their large size. This study also illustrates the benefit of comparing results when using all data and when first separating the data into smaller subsets.

SIMILARITIES NEEDS AMONG DATA SETS

The same dataset will be considered imperfect to different degrees for different projects, goals, methods, and disciplines resulting in personal subjectivity. No matter which method is applied to imperfect data challenges, they should be able to significantly improve analytical results. This improvement should be quantified by means of systematic comparisons of model performance and fit. After performance and fit comparisons have been exhausted, model complexity and computation time are the decision guides for selection of methods that solve imperfect data challenges.

An extreme degree of missingness is the lack of an outcome variable. And unsupervised learning is the way of pattern recognition when the outcome variable is missing. Unsupervised learning has great opportunity for cross-over and has great potential for becoming part of an integrated approach that resolves imperfect data challenges. Unsupervised learning, or cluster analysis, could have been applied to the data in Chapters 2.1, 2.2, and 2.3. Unsupervised learning in Chapter 2.1 could have been used to discover different patterns in *Campylobacter* cases among states that might have different risk factors per region. When examining *Plasmodium falciparum* infection risk factors in Chapter 2.2, clustering could have led to more targeted recommendations and outreach implementation for groups of homesteads based on each groups' particular risk factors for infection. Finally, different parameter estimation models could have been built for each AMS cluster described in Chapter 4.2 resulting in different risk factors of milk production per cluster.

In addition to missingness and unbalanced data sets, multilevel modeling is a commonly faced challenge because data are merged from many different sources collected at different time points or at different levels of observations. This challenge was evident in the datasets used in this

dissertation. The *Campylobacter* data set in Chapter 2.1 had census data at the county level and *Campylobacter* cases at the individual patient level. PAZ questionnaire data in chapter 2.2 were at the homestead level while the biological sampling data of humans and cattle were collected at the individual level. In addition, the original AMS dataset in chapter 2.3 had robot milk production at the robot or farm level while specific robot milking data, such as milking speed and boxtime, were at the cow level. In such complex systems, the more extreme observations per variable at the micro level can be a better reflection of the characteristics at the macro level. Averaging of the observation at the micro can result in loss of information. The different types of micro-macro modeling could be used in these cases to identify whether extreme observations, or averaging of the observations per variable, leads to better model outcome.

Choices for modeling with regards to levels of observations and clustering are numerous. Taking this to the extreme, one can foresee many opportunities for the applications of rtFMS. Additional applications of rtFMS, as described in Chapter 3.1, can be applied to a variety of statistical learning projects including the following examples in this dissertation. Chapter 2.1 includes a comparison among hurdle and zero-inflation model types, and Poisson and negative binomial model distributions. However, rtFMS would allow the comparison of other options such as performing clustering (unsupervised learning) before modeling to select states or regions that require separate models. In Chapter 2.2, rtFMS would help to compare results for using data imputation or not, and for using PCA to make wealth indices compared to backwards elimination of highly correlated wealth variables. Additionally, rtFMS could compare different limits for removing highly-correlated variables, and among different hyperparameters value in the glmnet model.

Another chapter that could have benefitted from rtFMS is chapter 2.3. rtFMS would allow the comparison of correlation limits and t-value limit when forward selection of interaction terms is performed. In addition, the forward selection method could have been compared to other selection methods such as backwards selection of interactions and to performing a network analysis instead of a generalized linear model. Ideally, interactions would have been also included in all models in this dissertation, including for *Campylobacter* and *Plasmodium falciparum* infections, to increase the models' goodness of fit and therefore also increase the accuracy of the risk factor estimations.

The flexibility of rtFMS is reflected in the fact that a multitude of performance measures can be chosen. Alternative to the currently used balance accuracies in the prediction modeling section (see chapter 3.1), akaike information criterion (AIC), bayesian information criterion (BIC), mean absolute error (MAE), and -2xlog likelihood could be used for parameter estimation models.

It is commonly noted that secondary data will be more prone to biases including selection, random and misclassification biases (Fan et al. 2014; Harford 2014; Haine et al., 2018). It is sometimes mentioned that big data could compensate for some biases through sheer numbers of observations (Seely-Gant and Frehill, 2015). However, that has not been shown to be the case (Lazer 2014). In this work, certain biases were encountered and mitigated. Selection bias is sometimes referred to as availability bias in data mining. In Chapter 4.1, selection bias was present since only data from Simmental cows were available. This bias was acknowledged and the potential for the results to also apply to Holsteins was discussed. Second, in Chapter 4.2, the farms included in the study only originated from one brand of AMS. This bias was mitigated since the results were not generalized to other brands of AMS or non-AMS farms. Future studies will collect data from different breeds

and milking systems. Additionally, the results were only used to provide peer groups for the farms used in the study. In Chapter 2.2, random error was present due to the many missing values. The missing data was deemed to be missing at random and were therefore imputed. However, if the missingness had been systematic and not random, attempting a solution by imputation would have only introduced more bias. Finally, misclassification bias was encountered in Chapter 3.2 when prediction models for antimicrobial residues were developed. The outcome values in the training data were determined using an imperfect test (the brilliant black reduction test) that could lead to misclassification of the observations. Epidemiology training includes identifying, acknowledging and addressing bias. This emphasis needs to continue to be applied to studies where data science methods and secondary data are used.

In summary, the many examples of secondary data sets in this dissertation illustrate the similarities for imperfect data challenges including having imbalances in positive and negative outcomes, rare events, zero inflation, high-dimensionality, multicollinearity, missing data, multiple significant interactions, variety in structure, and undefined outcomes.

SIMILARITIES AND DIFFERENCES AMONG DATA DISCIPLINES

Within each discipline, researchers will always be dependent on having data that are relevant to their research question. This leads to many similarities within divergent data disciplines such as data sciences and statistics. One of the differences between statistics and data science can be highlighted by the disciplines' movement towards automation (Reid, 2016). Unlike applied statistics, there is the growing movement towards automation in data science (Gaber, 2009; Witten et al., 2016; Cearley, 2019). Data science is also more associated with "big data" than is mathematical statistics (Reid, 2016).

The term "big data" is often defined by three "V" terms: Volume, Variety and Velocity (Mooney et al., 2015). Epidemiology has many overlapping features with big data which are highlighted in this thesis. In term of volume, several data sets used in this dissertations' studies had over 40,000 observations. In addition, variety is a big component of these studies. These chapters covered many outcomes from human *Campylobacter* infections in the USA, *Plasmodium falciparum* infection in Kenya, to metabolic health and milk production in cattle. Other than outcome, there was a lot of variety seen among the studies in this dissertation in terms of data size, location, goal, study design, time frame, disease, goal, data type, outcome, technique, and population. Finally, although epidemiology has most commonly dealt with historical data instead of real-time data, real time data is also used. This could be the case in the future with the AMS data in Chapter 2.3. AMS production and milk data are routinely collected and so prediction models could be incorporated into normal real-time routines.

Although epidemiology overlaps with data sciences in many facets, epidemiology has more emphasis on the biological relevance and interpretation of results. In addition, epidemiology has a stronger foundation in study designs and is more hypothesis driven, which originate from the influence of statistics (Dohoo et al., 2003). For example, there is more emphasis in epidemiology for evaluating if data meet statistical test assumptions such as independence of observations, normal distributions, and addressing potential autocorrelation and random effects. These strengths of epidemiology can offer a lot in interdisciplinary collaborations. However, there is concern that data mining methods and automation will cause the field to move away from deductive reasoning

and prohibit the incorporation of expert knowledge about biological relevance into analyses (Dohoo et al., 2003; Faraway, 2016).

Although data mining methods do not originate from deductive reasoning, the unsupervised methods can expose investigators to new ways of looking at a research area. In addition, the use of data science methods does not inhibit biological relevance from coming into play at other stages of a study. This fact was illustrated in many chapters of this work. First, although the automated selection of interactions was used in Chapter 2.3, biological relevance was key in selecting which interactions were discussed and explored further in the study. Secondly, Chapter 4 used unsupervised methods to find patterns in large sets of data without specific outcome variables. Although deductive reasoning did not initiate the studies, the results were subjected to post-hoc analyses. The post-hoc analyses were pivotal in how the results were interpreted and allowed the biological relevance to be incorporated. At this stage of interpreting results, the possible biases and errors, that were previously discussed (page 136), are also taken into consideration. Although data mining does not originate from deductive reasoning, its results initiate many new hypotheses. These new hypotheses can then be investigated in a more formal hypothesis driven process in follow-up primary studies.

Veterinary epidemiological research, will continue to follow the trends in data sciences towards automation due to the growing use of imperfect, secondary, and potentially big data, and the desire to develop real-time surveillance and prediction capabilities (VanderWaal et al., 2017; Hermans et al., 2018). This work suggests that systematic methods can lead to automation of epidemiological data analyses, and can coincide with a focus on biological relevance and sound interpretation.

MOVING TOWARDS A UNIFIED AUTOMATED METHOD

The idea of systematic protocols for method selection was consistent throughout the dissertation. One can imagine the future development of a comprehensive rtFMS procedure to embrace and compare all potential data imperfection challenges. Alternatively, a new algorithm could be developed that encompasses many imperfections discussed in this work, similar to glmnet accounting for high-dimensionality and highly correlated variables at the same time. In this case, automation of such an all-encompassing rtFMS procedure or algorithm would be the obvious next step. The ideal situation of such a procedure or algorithm would only require the user to select the outcome variable for optimization and the goal of the data analysis such as parameter estimation, prediction modeling, forecasting, pattern recognition, or survival analysis. The automation of these methods would make many diverse methods from a variety of disciplines available to many more users. Especially with cloud computing becoming more accessible, even computationally taxing methods will not hold users back.

The rtFMS method prevents overfitting by means of systematic iterated cross-validation. Also, the method incorporates multiple comparison corrections to adjust the results for the rate of false discoveries. However, there are some potential down-sides of using rtFMS for selection modeling methods. First, the performance measures that are currently used to optimize models and make selections (e.g. balanced accuracy) are based only on model performance in terms of predictive ability. However, as methods continue to develop, the best performing model might not be an easily interpretable model. Models that are not easily interpretable are sometimes referred to as

"black box" methods because direct associations between predictors and outcomes are not easily extracted from the model (Chollet, 2018). Algorithms difficult to interpret include deep learning and neural networks. When this occurs, researchers would be faced with a decision between performance and interpretability. However, rtFMS is very flexible in term of the outcomes that it can use for optimization. In the future, a new performance parameter could be developed that balances performance and simplicity. This could be accomplished similarly to how the Bayesian information criterion (BIC) is used in parameter estimation models to balance model goodness of fit and simplicity.

LIMITATIONS

Several limitations need to be examined when considering this work. One of the major limitations of this work is that although the aim was to cover the most common types of data imperfection, there are other types of imperfections that were not addressed. For example, working with data coming from different sources that do not have a common identifier for merging is a challenge currently faced in data mining. It is sometimes referred to as the data fusion problem (Bareinboim and Pearl, 2016). Secondly, not all available methods for dealing with specific data imperfections were addressed within this work. For example, when faced with large number of significant interactions, a network analysis might be more appropriate than a generalized linear model. Thirdly, there are more areas of statistical learning that were not addressed in this dissertation such as forecasting, survival and simulation, in addition to deep learning and reinforcement learning. Although this work could not exhaust all the possible types or combinations of data imperfections one could face, or give examples in all types of statistical learning areas, the tools provided in Chapter 3, rtFMS, could be used for all of these statistical learning areas and combination of data imperfection challenges.

Although this work provide the tools to make such comparisons among methods in the prediction modeling (Chapter 3), rtFMS was not applied to methods described in the parameter estimation (Chapter 2) or unsupervised learning (Chapter 4). Due to the succession of the work, this method was not available at the time. Data sciences are more commonly used for the goal of prediction modeling rather than parameter estimation. However, applying machine learning for parameter estimation is an obvious extension of machine learning methods that is done by optimizing model fit instead of optimizing for predictive ability. The potential for overfitting still needs to be addressed, but this could be done using a limit of the number of parameters that could be included as a ratio to the number of observations in the dataset. Recommended limits lie between a 1:5 and a 1:25 ratio (Peduzzi et al., 1996; Babyak, M.A., 2004; Hair et al., 2009; Harrell, 2015). For parameter estimation models, interactions can be important additions to a model. Interactions are not commonly included in machine learning models since some algorithms such as the neural network already account for such relationships (Chollet, 2018). In addition, machine learning methods do not currently include a step that uses unsupervised learning to disaggregate the data into more representative groups before modeling. It would be worth investigating if including certain parameters estimation aspects such as interactions and unsupervised learning prior to modeling would benefit the performances of some types of machine learning algorithms. The resulting performances could be easily compared using the rtFMS method presented in this dissertation.

Additionally, this work did not include comparisons between results with and without applying specific data imperfection methods in all chapters. For example, results with and without applying SMOTE were not compared in Chapter 3.1, and the results with and without imputing missing data were not compared in Chapter 2.2. These comparisons are important to put into perspective at what point these methods became necessary. The rtFMS method described in Chapter 3 should include these comparisons in the future. Readers should embrace the idea that imperfections of data can be dealt with in many ways, but that a formal comparison of methods is needed to obtain the best results.

FUTURE PERSPECTIVES

It is to be expected that the amount of technological innovation to continue to accelerate in the future. As new technologies are embraced and become ubiquitous, the amount of data collected and analyzed will only increase and will become a rich source of potential data for epidemiologists in the future. As such, imperfect data challenges will also rise in quantity and variety. Therefore, it is necessary to further develop and compare modeling and pattern recognition methods that address imperfect data challenges commonly faced in this discipline. In addition, preparations are needed for new types of imperfect data challenges that will arise in the future such as the data fusion problem (Bareinboim and Pearl, 2016). However, this challenge is not yet commonly seen in epidemiological studies. It should become apparent that much more research is needed in this discipline to improve imperfect data methods, as these data and challenges become more common and complex in the future.

Future studies are needed to determine thresholds of data imperfection in which the use of specific analytical techniques yield significant improvements in the performance of the resulting model. In addition, new methods should determine if they have individual limits of how much imperfection it can handle. For example, the method of multiple imputation by chained equations (MICE) has determined a limit of 10% missingness or less for the method to be used successfully (in addition to a requirement from missingness to be random). Other methods should aim to reach similar understanding of their limitations. Balancing methods such as SMOTE might have distinct limits as well. Pairing the degrees of imperfection with the data set's intended use should be considered more systematically. This would also be helpful to limit the amount of methods that are applied and compared in an rtFMS procedure to reduce the complexity of the analysis.

With the similarities in the data disciplines' needs and with a rise in data and their corresponding imperfect data challenges in the future, interdisciplinary collaborations will become more common. This can already be observed with secondary data-based studies in human medicine and epidemiology. For example, citizen research has been used to monitor air quality and social media have been used to better monitor and detect influenza infections (Broniatowski et al., 2013, Snik et al., 2014). These studies have routinely incorporating data mining methods. However, 21st century tools are also part of primary data-based studies. For example, Genome-wide association studies (GWAS) are commonly employed in epidemiology. Genetics statistics have been merged with human epidemiology to better understand disease (McGrath et al., 2013). Microbiome studies are being performed across all of veterinary medicine and require intensive bioinformatics support (Barko et al., 2018). Sensor technologies in precision farming that were developed for their ability to improve welfare and production are being adopted throughout the world (Berckmans, 2014; Norton and Berckmans, 2018). Finally, epidemiology's future is heading in the direction of

exposomology, which examines individuals' exposures over a lifetime to get the most accurate estimates on risks and associations for disease (Niedzwiecki et al., 2019). The goal of exposomes will dependent on approaches from data disciplines such as data science, data mining, machine learning and the development of new methods such as advanced multilevel modeling techniques. Citizen research, GWAS, microbiomes, sensor technologies in animals and exposome studies are perfect examples for the future of epidemiology that will require regular large interdisciplinary team efforts. This degree of interdisciplinary collaboration has large consequences for teaching of epidemiology.

Informed decisions have to be made if classic epidemiologist will lead the way during the development of methods from data sciences or whether they will adopt automated models from the cloud or recruit the help of data scientists for analyses. Furthermore, epidemiology leaders should discuss if topics such as data mining, data science, and machine learning need to be added to the epidemiology curriculum. If this does not occur, the future will most likely include the automation of a unifying method for such analyses or the farming-out of big imperfect data to data scientists. If the future of epidemiology does not focus on methods as described in this dissertation, the discipline could focus on the meaningful interpretation and biological relevance of the results.

PRACTICAL IMPLICATIONS

This dissertation provides a framework to address common imperfect data challenges in parameter estimation, predictive modeling, and pattern recognition. These systematic procedures and methods are comprehensive and flexible enough to account for different data sets, data imperfections, and goals for analysis. This dissertation guides those faced with imperfect large data sets into optimized data analyses despite imperfect data challenges. Consequently, data sets from primary or secondary data-based studies are equally usable for the primary goals of modeling and pattern recognition based on this work.

Finally, this dissertation provides systematic procedures for comparing a multitude of methods available within and across disciplines such as data science, data mining and machine learning to optimize selected performance measures. Applying rtFMS facilitates the incorporation of more advanced statistical analytical methods resulting in improved models and outcomes. Future models' performance and fit will benefit from the widespread of rtFMS in epidemiology.

CONCLUSIONS

- This work illustrates and describes the similar needs and challenges across large imperfect data sets. This work illustrated several new methods for addressing the challenges of large imperfect data sets which were applied to parameter estimation, predictive modeling, and pattern recognition.

- This work developed methods for systematically combining, comparing and selecting the most appropriate statistical methods available to optimize selected performance measures, streamline selection, and to prevent subjectivity, bias and flawed outcomes. Systematic approaches to analysis make the presented methods and working with large, imperfect, secondary data more accessible.

- This work is a steppingstone towards ensuring that future data analyses are not hindered by large data imperfections and that the many possible methods available become an advantage and not a hurdle.

- This work is an example for how utilizing methods from discipline such as data science and machine learning can improve models' performance and fit.

- Finally, as large data sets from secondary data-based studies become more widely available, interdisciplinary methods will prove crucial for the discipline of epidemiology to maintain forward momentum.

- This thesis has shown that imperfect data challenges can be solved, and in the name of improved model performance and fit, systematic model choices are crucial.

- Collectively, these methods provide an unbiased framework in which to approach the analysis of large data, without losing the emphasis on the biological relevance and sound interpretation of the results.

- This work suggests that systematic methods can lead to automation of epidemiological data analyses, and can coincide with a focus on biological relevance and sound interpretation.

IMPACT

The challenges of comparing and selecting pattern recognition, descriptive or prediction models and methods using large, imperfect data sets are realities seen in many disciplines. This is also true for the need to make inferences for decision-making processes based on these models. In addition, the use of large, imperfect data from secondary data-based studies will only become more common in the future. Finally, the number of different methodological options available for each type of data imperfection will also increase. Therefore, the methods described and used in this dissertation will allow more access to data for analysis while gaining the benefit of a large data set.

REFERENCES

Babyak, M.A., 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic medicine, 66(3), pp.411-421.

Bareinboim, E. and Pearl, J., 2016. Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences, 113(27), pp.7345-7352.

Barko, P.C., McMichael, M.A., Swanson, K.S. and Williams, D.A., 2018. The gastrointestinal microbiome: a review. Journal of veterinary internal medicine, 32(1), pp.9-25.

Broniatowski, D.A., Paul, M.J. and Dredze, M., 2013. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PloS one, 8(12), p.e83672.

Cearley, D.W., Burke, B., Searle, S., Walker, M.J. and Claunch, C., 2019. The top 10 strategic technology trends for 2019. Gartner.

Chollet, F., 2018. Deep learning with Python. Manning Publications Co.

Dohoo, I.R., Martin, W. and Stryhn, H., 2003. Veterinary epidemiologic research (No. V413 DOHv). Charlottetown, Canada: AVC Incorporated.

Fan, Jianqing, Fang Han, and Han Liu. 2014. “Challenges of Big Data Analysis”. National Science Review, 1(2), 293-314.

Faraway, J.J., 2016. Linear models with R. Chapman and Hall/CRC.

Gaber, M.M., 2009. Scientific data mining and knowledge discovery. Springer.

Haine, D., Dohoo, I. and Dufour, S., 2018. Selection and misclassification biases in longitudinal studies. Frontiers in veterinary science, 5, p.99.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. and Tatham, R.L., 2009. Análise multivariada de dados. Bookman Editora.

Harford, Tim. 2014. “Big data: A big mistake?” Financial Times, 11(5), 14-19.

Harrell Jr, F.E., 2015. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer.

Hermans, K., Opsomer, G., Van Ranst, B., and Hostens, M., 2018. Promises and Challenges of Big Data Associated With Automated Dairy Cow Welfare Assessment. Animal Welfare in a Changing World, p.199.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. “The Parable of Google Flu: Traps in Big Data Analysis”. Science 343(6176), 1203-1205.

McGrath, J.J., Mortensen, P.B., Visscher, P.M. and Wray, N.R., 2013. Where GWAS and epidemiology meet: opportunities for the simultaneous study of genetic and environmental risk factors in schizophrenia. Schizophrenia bulletin, 39(5), pp.955-959.

Mooney, S.J., Westreich, D.J. and El-Sayed, A.M., 2015. Epidemiology in the era of big data. Epidemiology (Cambridge, Mass.), 26(3), p.390.

Niedzwiecki, M.M., Walker, D.I., Vermeulen, R., Chadeau-Hyam, M., Jones, D.P. and Miller, G.W., 2019. The exposome: molecules to populations. Annual review of pharmacology and toxicology, 59, pp.107-127.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R., 1996. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology, 49(12), pp.1373-1379.

Reid, D., 2016. Man vs. Machine: The Battle for the Soul of Data Science. In Big Data Challenges (pp. 11-22). Palgrave, London.

Seely-Gant, K. and Frehill, L.M., 2015. Exploring Bias and Error in Big Data Research. Journal of the Washington Academy of Sciences, 101(3), pp.29-38.

Snik, F., Rietjens, J.H., Apituley, A., Volten, H., Mijling, B., Di Noia, A., Heikamp, S., Heinsbroek, R.C., Hasekamp, O.P., Smit, J.M. and Vonk, J., 2014. Mapping atmospheric aerosols with a citizen science network of smartphone spectropolarimeters. Geophysical Research Letters, 41(20), pp.7351-7358.

VanderWaal, K., Morrison, R.B., Neuhauser, C., Vilalta, C. and Perez, A.M., 2017. Translating big data into smart data for veterinary epidemiology. Frontiers in veterinary science, 4, p.110.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

SUMMARY

The accelerated rate of technological innovation in the 21st century has yielded many large data sets requiring new methods and approaches for analysis. In the past, data collection was mostly done after the development of a hypothesis-based experimental design and project plan ("primary data-based studies"). In contrast, data from secondary data-based studies are large data sets, potentially big data, collected before the analysis' method or goal is determined. While data from secondary data-based studies can be limited by the nature of their collection method, they can compensate with their large sample sizes, and the increased detail from larger numbers of variables. At the same time, large secondary data-based studies are more prone to result in imperfect data challenges including: rare events, high-dimensionality, missing data, multilevel data, and undefined outcomes. Routinely collected animal health and production data are good examples of secondary data commonly used in veterinary epidemiology.

Epidemiology has a strong foundation in using statistical methods for data analysis. However, epidemiology has recently followed trends in data science such as the emergence of data mining and machine learning. Unlike statistics, there is the growing movement towards automation in data science. Automation of data analysis has the potential to increase the amount of output, and improve the resulting model performance. We can assume that some portions of veterinary epidemiological research will continue to follow the trends in data sciences towards automation. This is due to the growing use of imperfect, secondary, and potentially big data, and the desire to develop real-time surveillance and prediction capabilities in epidemiology. However, epidemiology will always need to be focused on biological relevance and meaningful interpretation of results. Therefore, epidemiology needs to prepare for the trend towards automatic data analyses. This could be accomplished by adapting or developing methods that can be automated in the future and that do not remove the focus of the research from the biological relevance and interpretation of the results. In this dissertation, the goal was to develop methods for analyzing imperfect data and solutions for the systematic integration, comparison and selection of methods to streamline selection, and to prevent subjectivity when analyzing imperfect data.

A large variety of data sets were used from large, secondary data-based studies that were available for analysis. A range of methods to address imperfect data challenges were applied, and new methods and systematic protocols were developed all while being focused on biological relevance and sound interpretation of the results. The studies had diverse goals that fit into three main areas: parameter estimation modeling, prediction modeling, and pattern discovery.

In Chapter 2 the focus was on systematic supervised learning methods for parameter estimation while addressing data imperfections. Chapter 2.1 described a systematic approach to addressing the excess of zero case counts for parameter estimation using zero augmented models while in search for the best fitting model. Chapter 2.2 demonstrated a systematic approach to imputation of missing data, variable reduction and selection in a high-dimensional data set. The sequential and parallel application of methods was successful in reducing a wide, sparse dataset to a more useful, simplified set of predictors while still benefitting from the large amount of data in this data set. Chapter 2.3 focused on systematic selection and interpretation of interaction terms for parameter

estimation in search for the best fitting model. The interactions significantly increased the fit of the regression model and did not prevent meaningful discussion on the biological relevance and interpretation of interactions among parameters.

In Chapter 3 a systematic approach for supervised learning methods for prediction modeling was developed that allows systematic comparisons of many modeling options. Chapter 3.1 introduced the systematic approach to full model selection for prediction modeling using regression trees: regression tree full model selection (rtFMS). A significantly better performing model was obtained using rtFMS compared to what would have been obtained if method options had been selected subjectively. In addition, rtFMS eliminated the bias associated with empirical selection of method options. Chapter 3.2 described the development of new method of micro-macro multi-level modeling developed named "Extreme Value Micro-Macro" (EVMM) multilevel modeling. This method addressed the challenge of multilevel modeling when the central tendency of the micro level observations is not a good representation of the macro level outcome. The rtFMS method illustrated the significantly better performance of the EVMM method compared to the current micro-macro multilevel modeling method that uses the mean of the micro level observations.

In Chapter 4 unsupervised learning methods for pattern recognition were applied while the results' biological relevance were confirmed with post-hoc methods. In Chapter 4.1, to validate the findings of the cluster analysis, post-hoc regression models were used to examine associations between the cluster classifications and other parameters that were not used in the cluster analysis. The post-hoc findings supported and aided in the interpretation of the clustering results. In Chapter 4.2, to validate the findings of a cluster analysis of dairy farm observations, a comparison was made between the predictive performance of the resulting cluster classification and the current method that is used to classify these observations. The cluster analysis classification was a better predictor for a farm's milk production than was the current benchmarking method.

This work illustrated several new methods for addressing the challenges of large imperfect data sets which were applied to parameter estimation, predictive modeling, and pattern recognition. Collectively, these methods provide an unbiased framework in which to approach the analysis of large data, without losing the emphasis on the biological relevance and sound interpretation of the results. This work developed methods for systematically combining, comparing and selecting the most appropriate statistical methods available to optimize selected performance measures, streamline selection, and to prevent subjectivity, bias and flawed outcomes. This work is a steppingstone towards ensuring that future data analyses are not hindered by large data imperfections and that the many possible methods available become an advantage and not a hurdle.

The challenges of comparing and selecting pattern recognition, descriptive or prediction models and methods when using large, imperfect data sets are realities seen in many disciplines. This is also true for the need to make inferences for decision-making processes based on these models. In addition, the use of large, imperfect data from secondary data-based studies will only become more common in the future. Therefore, the methods described and used in this dissertation aimed at addressing the challenges associated with imperfect data sets in a systematic way will allow access to more data for analysis while gaining the benefit of a large data set. This work suggests that automation of data analysis can coincide with a focus on biological relevance and sound interpretation.

SUMMARY- DUTCH

Het versnelde tempo van technologische innovatie in de 21e eeuw heeft geleid tot veel grote gegevenssets die nieuwe methoden en benaderingen voor analyse vereisen. In het verleden werd het verzamelen van gegevens meestal gedaan na de ontwikkeling van een op hypothesen gebaseerd experimenteel ontwerp en projectplan ("primaire op gegevens gebaseerde studies"). Gegevens van secundaire, op gegevens gebaseerde studies daarentegen zijn grote gegevenssets, mogelijk big data, verzameld voordat de methode of het doel van de analyse is bepaald. Hoewel gegevens van secundaire, op gegevens gebaseerde onderzoeken kunnen worden beperkt door de aard van hun verzamelmethode, kunnen ze compenseren met hun grote steekproefgroottes en de toegenomen details van grotere aantallen variabelen. Tegelijkertijd zijn grote secundaire, op gegevens gebaseerde studies meer vatbaar voor imperfecte gegevensuitdagingen, waaronder: zeldzame gebeurtenissen, hoge dimensies, ontbrekende gegevens, multiniveau-gegevens en niet-gedefinieerde resultaten. Routinematig verzamelde diergezondheids- en productiegegevens zijn goede voorbeelden van secundaire gegevens die gewoonlijk worden gebruikt in de veterinaire epidemiologie.

Epidemiologie heeft een sterke basis in het gebruik van statistische methoden voor data-analyse. De epidemiologie heeft echter recent trends in de gegevenswetenschap gevolgd, zoals de opkomst van datamining en machine learning. In tegenstelling tot statistieken is er de groeiende beweging naar automatisering in de gegevenswetenschap. Automatisering van data-analyse heeft het potentieel om de hoeveelheid output te vergroten en de resulterende modelprestaties te verbeteren. We kunnen aannemen dat sommige delen van veterinair epidemiologisch onderzoek de trends in data-wetenschappen richting automatisering zullen blijven volgen. Dit komt door het toenemende gebruik van imperfecte, secundaire en potentieel big data en de wens om real-time surveillance- en voorspellingsmogelijkheden in de epidemiologie te ontwikkelen. De epidemiologie zal echter altijd gericht moeten zijn op biologische relevantie en zinvolle interpretatie van de resultaten. Daarom moet de epidemiologie zich voorbereiden op de trend naar automatische gegevensanalyses. Dit kan worden bereikt door methoden aan te passen of te ontwikkelen die in de toekomst kunnen worden geautomatiseerd en die de focus van het onderzoek niet verwijderen van de biologische relevantie en interpretatie van de resultaten. In dit proefschrift was het doel om methoden te ontwikkelen voor het analyseren van imperfecte gegevens en oplossingen voor de systematische integratie, vergelijking en selectie van methoden om selectie te stroomlijnen en om subjectiviteit te voorkomen bij het analyseren van imperfecte gegevens.

Een grote verscheidenheid aan gegevenssets werd gebruikt van grote, secundaire, op gegevens gebaseerde onderzoeken die beschikbaar waren voor analyse. Een reeks methoden om onvolmaakte gegevensuitdagingen aan te pakken werden toegepast en nieuwe methoden en systematische protocollen werden allemaal ontwikkeld, terwijl ze gericht waren op biologische relevantie en een goede interpretatie van de resultaten. De studies hadden verschillende doelen die in drie hoofdgebieden passen: parameterschattingmodellering, voorspellingsmodellering en patroonontdekking.

In Hoofdstuk 2 lag de focus op systematische gesuperviseerde leermethoden voor parameterschatting bij het aanpakken van onvolkomenheden in de gegevens. Hoofdstuk 2.1 beschreef een systematische aanpak voor het aanpakken van de overschrijding van nul casus-tellingen voor parameterschatting met behulp van nul-vergrote modellen terwijl op zoek was naar het best passende model. Hoofdstuk 2.2 toonde een systematische aanpak van de imputatie van ontbrekende gegevens, variabele reductie en selectie in een hoogdimensionale gegevensverzameling. De sequentiële en parallelle toepassing van methoden was succesvol in het reduceren van een brede, schaarse dataset tot een meer bruikbare, vereenvoudigde set van voorspellers, terwijl toch profiteerde van de grote hoeveelheid gegevens in deze dataset. Hoofdstuk 2.3 was gericht op systematische selectie en interpretatie van interactietermen voor parameterschatting bij het zoeken naar het best passende model. De interacties verhoogden significant de fit van het regressiemodel en verhinderden geen zinvolle discussie over de biologische relevantie en interpretatie van interacties tussen parameters.

In Hoofdstuk 3 is een systematische aanpak voor begeleidende leermethoden voor voorspellingsmodellering ontwikkeld die systematische vergelijkingen van vele modelleeropties mogelijk maakt. Hoofdstuk 3.1 introduceerde de systematische benadering van volledige modelselectie voor voorspellingsmodellering met behulp van regressiebomen: regressieboom volledige modelselectie (rtFMS). Een significant beter presterende model werd verkregen met behulp van rtFMS in vergelijking met wat zou zijn verkregen als de methode-opties subjectief waren geselecteerd. Bovendien elimineerde rtFMS de bias die samenhangt met empirische selectie van methode-opties. Hoofdstuk 3.2 beschreef de ontwikkeling van een nieuwe methode van micro-macro multi-level modeling ontwikkeld met de naam "Extreme Value Micro-Macro" (EVMM) multilevel modellering. Deze methode ging over de uitdaging van multilevel modellering wanneer de centrale tendens van de microniveau-waarnemingen geen goede weergave is van de uitkomst op macroniveau. De rtFMS-methode illustreerde de aanzienlijk betere prestaties van de EVMM-methode in vergelijking met de huidige micro-macro multilevel modelleringsmethode die het gemiddelde van de microniveau-waarnemingen gebruikt.

In hoofdstuk 4 werden niet-gecontroleerde leermethoden voor patroonherkenning toegepast, terwijl de biologische relevantie van de resultaten werd bevestigd met post-hoc-methoden. In Hoofdstuk 4.1, om de bevindingen van de clusteranalyse te valideren, werden post-hoc regressiemodellen gebruikt om associaties te onderzoeken tussen de clusterclassificaties en andere parameters die niet werden gebruikt in de clusteranalyse. De post-hocbevindingen ondersteunden en ondersteunden de interpretatie van de clusteringresultaten. In Hoofdstuk 4.2, om de bevindingen van een clusteranalyse van observaties van melkveebedrijven te valideren, werd een vergelijking gemaakt tussen de voorspellende prestaties van de resulterende clusterclassificatie en de huidige methode die wordt gebruikt om deze waarnemingen te classificeren. De clusteranalyseclassificatie was een betere voorspeller voor de productie van boerderijmelk dan de huidige benchmarkmethode.

Dit werk illustreerde verschillende nieuwe methoden voor het aanpakken van de uitdagingen van grote imperfecte gegevenssets die werden toegepast op parameterschatting, voorspellende modellering en patroonherkenning. Gezamenlijk bieden deze methoden een onbevooroordeeld kader om de analyse van grote gegevens te benaderen, zonder de nadruk te verliezen op de biologische relevantie en de correcte interpretatie van de resultaten. Dit werk ontwikkelde

methoden voor het systematisch combineren, vergelijken en selecteren van de meest geschikte beschikbare statistische methoden om geselecteerde prestatiemetingen te optimaliseren, selectie te stroomlijnen en subjectiviteit, vertekening en gebrekkige resultaten te voorkomen. Dit werk is een springplank om ervoor te zorgen dat toekomstige gegevensanalyses niet gehinderd worden door grote onvolkomenheden in de gegevens en dat de vele mogelijke methoden een voordeel worden en geen hindernis.

De uitdagingen van het vergelijken en selecteren van patroonherkenning, beschrijvende of voorspellingsmodellen en methoden bij het gebruik van grote, imperfecte gegevenssets zijn realiteiten die in veel disciplines worden gezien. Dit geldt ook voor de noodzaak om conclusies te trekken voor besluitvormingsprocessen op basis van deze modellen. Bovendien zal het gebruik van grote, imperfecte gegevens uit secundaire, op gegevens gebaseerde studies pas in de toekomst meer voorkomen. Daarom zullen de methoden die in dit proefschrift worden beschreven en gebruikt om de uitdagingen van imperfecte gegevenssets op een systematische manier aan te pakken, toegang bieden tot meer gegevens voor analyse, terwijl het voordeel van een grote gegevensset wordt behaald. Dit werk suggereert dat automatisering van data-analyse kan samenvallen met een focus op biologische relevantie en correcte interpretatie.

SHORT CURRICULUM VITAE

Dr. Marlène Tremblay is from Saint-Georges-de-Clarenceville, Quebec, Canada. She received her Bachelor of Science (B.S.) degree in Animal Sciences from the University of Kentucky in 2009 and her Doctor of Veterinary Medicine degree (D.V.M.) from the University of Wisconsin-Madison in 2013. After an epidemiology and public health research internship, she continued as an assistant researcher in the Food Animal Production Medicine Department at the University of Wisconsin-Madison, School of Veterinary Medicine. Concurrently, she pursued her Ph.D. degree at Utrecht University.

LIST OF PUBLICATIONS

Tremblay, M., Kammer, M., Lange, H., Plattner, S., Baumgartner, C., Stegeman, J.A., Duda, J., Mansfeld, R. and Döpfer, D., 2018. Prediction Model Optimization using Full Model Selection with Regression Trees Demonstrated with FTIR Data from Bovine Milk. Preventive Veterinary Medicine.

Tremblay, M., Kammer, M., Lange, H., Plattner, S., Baumgartner, C., Stegeman, J.A., Duda, J., Mansfeld, R. and Döpfer, D., 2018. Identifying poor metabolic adaptation during early lactation in dairy cows using cluster analysis. Journal of dairy science, 101 (8), pp. 7311-7321.

Tremblay, M., Crim, S.M., Cole, D.J., Hoekstra, R.M., Henao, O.L. and Döpfer, D., 2017. Evaluation of the Use of Zero-Augmented Regression Techniques to Model Incidence of Campylobacter Infections in FoodNet. Foodborne pathogens and disease, 14(10), pp.587-592.

Tremblay, M., Hess, J.P., Christenson, B.M., McIntyre, K.K., Smink, B., van der Kamp, A.J., de Jong, L.G. and Döpfer, D., 2016. Customized recommendations for production management clusters of North American automatic milking systems. Journal of dairy science, 99(7), pp.5671-5680.

Tremblay, M., Hess, J.P., Christenson, B.M., McIntyre, K.K., Smink, B., van der Kamp, A.J., de Jong, L.G. and Döpfer, D., 2016. Factors associated with increased milk production for automatic milking systems. Journal of dairy science, 99(5), pp.3824-3837.

Tremblay, M., Dahm, J.S., Wamae, C.N., De Glanville, W.A., Fèvre, E.M. and Döpfer, D., 2015. Shrinking a large dataset to identify variables associated with increased risk of Plasmodium falciparum infection in Western Kenya. Epidemiology & Infection, 143(16), pp.3538-3545.