LARGE SCALE INTEGRATED ANALYSIS OF BEEF PRODUCTION AND QUALITY IN BRAZIL

by

Vera Cardoso Ferreira Aiken

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

(Animal Sciences)

at the

University of Wisconsin – Madison

2019

Date of final oral examination: 12/10/2019

The dissertation is approved by the following members of the final oral committee:

Guilherme J. M. Rosa, Professor, Animal Sciences

Brian W. Kirkpatrick, Professor, Animal Sciences

Anthony Gitter, Professor, Biostatistics and Medical Informatics

Doerte Doepfer, Professor, Medical Sciences

João R. R. Dórea, Assistant Professor, Dairy Science

Acknowledgments

Above all, I would like to thank God, for my faith and perseverance that always helped me keeping the focus to achieve my objectives.

I would like also to thank my advisor Dr. Guilherme Rosa for the wonderful opportunity of coming to grad school and improving my knowledge and also for his guidance, patience, and support throughout this journey. I learned a lot from his vast reservoir of knowledge and experience.

I also need to thank the members of my committee João Dórea, Doerte Doepfer, Anthony Gitter, and Brian Kirkpatrick not only for serving as members of my evaluation process but also for the appreciated feedback and suggestions in my research projects. I would like to thank the USDA for the financial support of my studies, the JBS and DSM companies for providing the data and the CHTC at the University of Wisconsin-Madison for providing computational resources. I need to express my gratitude to the staff of Animal Sciences, in particular Kathy Monson, Deb Schneider, Minh Ngo, and Cathy Rook for their support in the academic process.

I also feel the need to thank all members from the energetic group of the 4th and 5th floors of the Animal Sciences building for their support. Specially my grad school buddies Tiago Passafaro, Arthur Fernandes, Beth Lett, Allison Quick and Nicole Gross. We perseverated trough many courses together, and supported each other throughout the sometimes challenging inroads of grad school.

Lastly, my acknowledgment would be incomplete, without thanking the biggest source of my strength, my family. My parents, Leonor and Cleunicio for always believing on me and encouraging me to follow my dreams even if it meant I would be away and they would miss me.

To my siblings Vanice, Vanessa, Cleunicio Jr. (and Felipe) for being my partners and my biggest "cheerleaders". To my Aiken/Swenson family that made me part of their own and gave me a "home away from home". Also to our "friends that became family" and made this journey a more pleasant one, always there when I needed. Finally, the completion of this work would not be possible without my best half, Erik Aiken. My supportive husband that was always by my side and would take care of the day to day tasks so I could focus on my studies and achieve my goals.

Abstract

This dissertation focused on data analytics of large scale integrated beef cattle data from Brazil, being composed by three main research chapters. The first chapter provided a framework for data integration, developing the concept of "farm-matching" (i.e. identifying the same farms in different databases in the absence of a universal identifier). The efficiency of different deterministic, stochastic and machine learning approaches was compared for the task of "farm-matching". High levels of accuracy, precision, sensitivity and specificity were achieved with two fully automated machine learning methods for this task: Support Vector Machines and Bagged Clustering. The second research chapter evaluated forecasting models for beef production and quality at the national level using a unique integrated large data collection, including farm, market and environmental factors. More specifically, traditional approaches (regression) were compared to modern machine learning methods (random forests and neural networks) for the task of forecasting meat production and quality at the national level. Moderate-high levels of accuracy were achieved for the task of forecasting using both Random Forests and regression, but lower when neural networks were utilized. The third research chapter investigated how factors affecting meat production and quality vary across different regions of Brazil, utilizing the same previously mentioned large scale integrated data set. More specifically, linear regression was contrasted with Geographically Weighted linear regression to test the hypothesis that explanatory variables affect meat production and quality differently for different locations in Brazil. Results showed that after including climate and soil of the farms in the model, the effects do not seem to vary across regions. This indicates that those variables naturally capture spatial heterogeneity and after accounting for them, physiological mechanisms seem consistent across regions. Lastly, this dissertation accomplished the goal to show that there is value in extracting information from complex, messy,

large data collections in animal sciences. We suggest further implementations of data analytics in the field, as they can prove invaluable for addressing the challenge of increasing production in a sustainable manner in animal agriculture.

Table of Contents

List of tables

List of figures

Introduction

In the 21st century, the landscape of animal agriculture is changing. Improvements in data recording and availability have made it possible to generate an unprecedented abundance of data. These large data collections can be a valuable source of information, capturing aspects of real-world production systems that were never explored before. In fact, in the current scenario, where animal agriculture is prompted to increase production to feed a still growing human population, but at the same time remain sustainable, with low costs and environmental impact, extracting useful information from large data collections (i.e. data analytics) will be core to optimizing the whole food production chain. Studying past animal production trends grants understanding on how they vary under different conditions, allowing to predict the future and better prepare for different scenarios, optimizing allocation of resources at all levels of the production chain.

This thesis is focused on data analytics of beef cattle production and quality using large scale integrated data from Brazil. Chapter 1 provides a literature review of Brazilian beef production systems and how the Brazilian market is an important player globally. Additionally, the utilization of large data collections and data analytics in agriculture is reviewed. Chapter 2 comprises data integration, comparing different deterministic, probabilistic and machine learning methods to develop an efficient framework for "farm-matching" (i.e. identifying which farms are the same across different databases, in the absence of a "universal identifier"). In Chapter 3 we forecasted meat production and quality at the national level. To do so, an integrated large scale data set was used. It contained information on meat production and quality traits, nutrition used, participation in a technical consulting program, economic variables related to beef production, climate and soil of the farm. Lastly, in Chapter 4, utilizing similar data, we explore the spatial heterogeneity of beef production and quality, across different locations in Brazil.

Chapter 1: Literature review

1.1 Beef cattle production in Brazil in its position in the international scenario

Beef production represents an important sector of animal agriculture, being the third most produced meat in the world, after pork and poultry (FAO, 2014). With the largest commercial beef herd in the world of 214.7 million head of cattle (ABIEC, 2019), Brazil has been consistently the second largest producer of beef worldwide for the past decade, after the United States (Table 1.1).

	Production in 1,000 metric tones (Carcass weight equivalent)			
	2015	**2016**	**2017**	**2018**
United States	10,817	11,507	11,943	12,256
Brazil	9,425	9,284	9,550	9,900
European Union	7,684	7,880	7,869	8,003
China	6,169	6,169	6,346	6,440
India	4,100	4,200	4,250	4,265
Australia	2,547	2,125	2,149	2,306

Table 1.1: Beef and veal production for largest producing countries from 2015 to 2019. This table was created based on the information retrieved from USDA – Office of Global Analysis (2019).

Most of the Brazilian beef production is destined to fullfil the demands of the internal market, with 78.4% of the national production (8,003M out of the 10,210M metric tons produced) supplying domestic consumption (Table 1.2) in 2018. In Brazil, the beef consumption per capita for 2018 was 42.12 kg/year (ABIEC, 2019). Addittionaly, beef production plays an important role

in terms of generating employment and wealth to the Brazilian economy (Millen et al., 2011), being responsible for generating 8.7% of the Gross Domestic Product (GPD) for the country in 2018 (ABIEC, 2019).

	Total domestic consumption **1,000 metric tones (Carcass weight equivalent)**			
	2015	**2016**	**2017**	**2018**
United States	11,275	11,676	12,052	12,180
China	6,808	6,928	7,313	7,910
Brazil	7,781	7,652	7,750	7,865
European Union	7,742	7,899	7,750	7,865
India	2,294	2,436	2,401	2,709
Argentina	2,534	2,434	2,547	2,562
Mexico	1,797	1,809	1,841	1,872

Table 1.2: Beef and veal total domestic consumption for the largest consuming countries from 2015 to 2019. This table was created based on the information retrieved from USDA – Office of Global Analysis (2019).

Brazil has also been the biggest exporter of beef in the world since 2017 (Table 1.3) with almost 20% of global exports (2.1MM metric tons exported from the 9.9 produced) in 2018 (Zia et al., 2019). USDA predicts that Brazilian production will keep increasing in the near future, reaching 23% of the world's total exports by 2028 (Zia et al., 2019), remaining essential in the international agriculture market. The increase in exportation observed over the past years (Table 1.3) is considered a driver within the Brazilian market, regulating production aspects such as the

use of antibiotics and ionophores or even prohibiting the use of substances like growth promoters (Millen et al., 2011) .

Beef cattle production is developed in all Brazilian ecosystems (Oliveira, 2018), with the major producing states being Mato Grosso, Goiás, Minas Gerais, Mato Grosso do Sul, and Pará, which are responsible for 54.2% of the national production (Oliveira, 2019). Brazil has a mature beef cattle industry based on grass-fed cattle (Millen et al., 2011), in which cattle spend most of their lives grazing in pasture. This is appealing both from the standpoint point of animal welfare and sustainability, as well as consumer's preference and that is part of the reason for the expansion of Brazilian production systems in the past years.

	Total exports in 1,000 metric tones (Carcass weight equivalent)			
	2015	**2016**	**2017**	**2018**
Brazil	1,705	1,698	1,856	2,083
India	1,806	1,764	1,849	1,556
Australia	1,854	1,480	1,485	1,662
United States	1,028	1,160	1,297	1,434
Argentina	186	216	293	507
New Zealand	630	587	503	633
Canada	397	441	461	502

Table 1.3: Beef and veal total exports for the largest exporting countries from 2015 to 2019. This table was created based on the information retrieved from USDA – Office of Global Analysis (2019).

The Brazilian beef cattle industry is characterized by cow-calf, stocker, and feedlot operations (Millen et al., 2011).The vast majority of the production cycle in Brazil comprises

grazying systems (predominantly on pastures cultivated with *Brachiaria*), with only suplementation of minerals, leading to older animals in the market (Millen et al., 2011). This is due to the fact that while animals put on weight in the rainy season, when pasture quality is higher, they lose weight in the dry season with decreased availability of nutrients. Nonetheless, feedlot operations started being implemented in Brazil in early 2000's to finish animals, as a response to meat production demands (Millen, et. al., 2009), and such animals are mainly destined for the external market (Millen et al., 2011). However, as mentioned by Millen et al. (2011) the feedlot period is short (around 83 days for bulls). The primary source of grain utilized in feedlots is corn, with 51-65% of grains in the finishing diet. The main co-products used (level of inclusion = 15%) are whole cotton seed, citrus pulp, and soybean. The main source of roughage (level of inclusion = 28%) are fresh cropped sugar cane, corn sillage and sorghum sillage. Lastly, the main dry supplements utilized in feedlots are minerals and vitamins, and ionphores are most commonly used as an additive (Millen, et. al., 2009). Brazil produces mainly lower-value, leaner grass-fed beef. However, the Brazilian beef production has shown a trend of intensification over the course of the past years, with increasing prices being paid by meat packing plants and the implementation of differential payments based on carcass quality (Millen et al., 2011).

Regarding sanitary conditions of the beef production systems in Brazil, foot and mouth disease has historically been an issue to exports. However, in the past years Brazil has been making progress with vaccination programs and 24 out of the 27 Brazilian states became free of the disease with vaccination in 2014 (ABIEC, 2019). Regarding greenhouse gas production, as highlighted by Millen et al. (2011), despite Brazil showing the largest growth rates in methane emission between 2001-2011 (estimated to be 2.12% per year), the herd also increased considerably for the same period (4.01% per year), indicating that there was mostly a negative net increase in rate of methane

emissions per unit of product (-1.89% per year). Between 1990 and 2018, the productivity of the system increased 176% moving from 1.63 arroba/ha/year in 1990 to 4.5 arroba/ha/year in 2018 (ABIEC, 2019). Lastly, despite historically having an overall trend of deforestation to create pastures and allow beef production, the rates of deforestation towards the Amazon have decreased in the past years and the report produced by ABIEC (2019) highlights that 250.6 million ha (Brazilian metric for area) stopped being deforested with the intensification and use of technology in the past 28 years.

1.2 Big data and data analytics in animal agriculture

The 21st century presented agriculture with the challenge of augmenting food production to keep up with a growing human population worldwide. The Food and Agriculture Organization (FAO) of the United Nations has projected a 30% increase in population by 2050 (FAO, 2009). What makes this challenge even more pronounced is the fact that this necessary increase in production should be accomplished in a responsible way. For example, it cannot occur at the expense of drastic increases in the environmental footprint, animal welfare concerns or high cost, which would limit access for developing countries. It is imperative to develop and utilize approaches capable of sustainably optimizing the food chain. Fortunately, the 21st century also saw the rise of large data collections as a valuable source of information to solve challenges in multiple fields, and has been recently applied to agriculture (Rosa and Valente, 2013; Morota et al., 2018). Extracting useful information from complex “big data” will be a crucial task to efficiently address the current demand faced by agriculture (Kamilaris, et al., 2017; Liakos et al., 2018 and Morota et al., 2018).

The so called “Big Data” term became popular specifically due to the decreased cost of modern technologies for collection and storage of larger amounts of data (Morota et al., 2018). There are currently multiple definitions of big data across diferent fields. Nonetheless, “Big Data” usually encompasses data sets that have a large number of columns or rows, or both (volume), high speed of generation or short time window in which is relevant (velocity), multi-source, multi-temporal and/or different format (variety) and/or great complexity (De Mauro et al, 2016). Moreover, as highlighted by Morota et al. (2018), big data usually is not clean data and may present issues such as missing observations, typos, confouding data, or outliers. Dealing with such messy and noisy data increases the challenge of analyzing this data and often requires specific technology, analytical methods, and expert training for its transformation into value (De Mauro et al., 2016).

Successful analytics of large data collections involve multiple steps comprising data processing, cleaning, integration, and pattern extraction (Kamilaris, et al., 2017). More specifically, data processing includes transformation processes across different platforms of data collection. While the specific processes for these steps vary by field and application, common techniques utilized in data cleaning are standardizations across databases, correction of mistakes, and identifying and removing outliers (Morota et al., 2018). Integration can be performed in different ways, some are more trivial such as linking information collected in the same time frame or location and others are more complex, such as record matching or data fusion (Christen, 2012). After those steps are consistently performed, pattern extraction allows obtaining useful insights from large scale data, which might be relevant for different applications in animal agriculture (Kamilaris et al., 2017; Morota et al., 2018).

The fields of artificial intelligence and machine learning are core to pattern recognition and information extraction. Machine learning is a subfield of artificial inteligence that focuses on the

study of algorithms that can be used for prediction. As Morota et al. (2018) and Liakos et al. (2018) highlighted, machine learning is expected to be crucial for addressing current challenges in agriculture, as it presents new tools for predicting future outcomes (i.e. forecasting) with large scale data. Forecasting of animal production is important because it plays a significant role in production and sales planning. By projecting future trends, we can optimize allocation of resources, rendering the whole production chain more sustainable.

In fact, with the increased capability for data storing and processing, animal agriculture is developing to be a data driven field with "precision agriculture" (i.e. data analytics in farming) being central to support decision making at the farm level (Morota et al., 2018; Pham and Stack, 2018). An increase in the application of big data analytics in animal sciences has been observed in the past years, and some studies are described below.

As highlighted by Morota et al. (2018), genetics is the field in animal sciences that made the earliest use of large data collections to extract value. Genetic evaluations have been performed in animal breeding applications at a national or company-level for years, using millions of animals with massive amounts of molecular information such as Single Nucleotide Polymorphisms (SNP) or Ribonucleic Acid (RNA) information. In that field, multiple methods have been developed and utilized over the years to optimize computational calculations and deal with large data issues such as imputation of missing values or number of observations much smaller than the number of parameters (Gianola and Rosa, 2015).

More recently, artificial intelligence and machine learning approaches have been applied for prediction in other areas of animal sciences with the objective of optimizing economic efficiency of farm systems in many livestock species (Liakos et al., 2018). An interesting application is the utilization of artificial neural networks to model data obtained with automatic

milking machines for the detection of mastitis and monitoring of the herd health status (Sun et al., 2010). In this application, high levels of prediction accuracy were achieved, which is capable of providing farmers with an efficient diagnostic tool for managing animal health. Machine learning can also be useful to predict hard to measure traits in commercial operations. For example, in another recent study in dairy cattle Dórea et al. (2018) successfully applied neural networks to milk infrared spectroscopy data as a tool to predict feed intake.

Another area that has received attention lately is the use of digital image analysis for understanding complex animal behavior data (Valletta et al., 2017). Applications include using animal images to determine behavior that precedes disease, and understanding social networks and hierarchies amongst animals. This knowledge can aid animal behaviorists' and farmers to make decisions on when to make medical or culling interventions or separately feed animals that rank lower in hierarchy. Lastly, machine vision systems have also been applied to body weight determination in livestock. For example, Kongsro (2014) and Fernandes et al. (2019) utilized selected image sections obtained with a Microsoft Kinect camera to estimate pig volume, which was later correlated with body weight, achieving very small prediction error for pigs of different sizes and breeds.

As exemplified above, despite the challenges faced in the analysis of large data collections, it has a huge potential to provide value in animal sciences. As highlighted by Pham and Stack (2018), the digitalization of farming systems is expected to represent the third revolution of agriculture and animal scientists and livestock producers need to be prepared for it. In fact, Pham and Stack (2018) suggest that professionals involved in agriculture get training to deal with large scale data, as competition will depend on anyone and everyone that can capture value from data.

1.3 References

ABIEC. 2019. "Beef Report: Perfil da pecuária no Brasil". Retrieved from: http://www.abiec.com.br/controle/uploads/arquivos/sumario2019portugues.pdf

Christen, P. (First) 2012. Data matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Springer, London.

De Mauro, A., M. Greco, and M. Grimaldi. 2016. A formal definition of Big Data based on its essential features. Libr.Rev. 65(3), 122–135. doi: 10.1108/LR-06-2015-0061.

Dórea, J. R. R., G. J. M. Rosa, K. A. Weld and L. E. Armentano. 2018. Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. J Dairy Sci. 101(7): 5878-5889. doi: 10.3168/jds.2017-13997.

FAO. 2009. “How to feed the world in 2050.” Food and Agriculture Organization of the United Nations, Rome.

FAO. 2014. “Meat consumption” http://www.fao.org/ag/againfo/themes/en/meat/background.html (November, 2, 2019).

Fernandes, A. F. A, J. R. R. Dórea, R. Fitzgerald, H. Herring and G. J. M. Rosa. 2019. A novel automated system to acquire biometrical and morphologic measuramanets and predict body weight of pigs via 3D computer vision. J. Anim. Sci. 97(1): 496-508. doi: 10.1093/jas/sky418.

Gianola, D., and G. J. M., Rosa. 2015. One hundred years of statistical developments in animal breeding. Annu. Rev. Anim. Biosc. 3, 19–56. doi: 0.1146/annurev-animal-022114-110733.

Kamilaris, A., A. Kartakoullis, and F. X. Prenafeta-boldú. 2017. “A review on the practice of big data analysis in agriculture.” Comput. Eletron. Agr. 143: 23–37.

doi:10.1016/j.compag.2017.09.037.

Kongsro, J. 2014. Estimation of pig weight using a Microsoft Kinect prototype imaging system. Comput. Eletron. Agr. 109, 32–35. doi: 10.1016/j.compag.2014.08.008

Liakos, K. G., P. Busato, D. Moshou, S. Pearson, D. Bochtis. 2018. "Machine learning in agriculture: A review." Sensors 18(2674): 1–29. doi:10.3390/s18082674.

Millen, D. D., R. D. L. Pacheco, M. D. B. Arrigoni, M. L. Galyean, and J. T. Vasconcelos. 2009. "A snapshot of management practices and nutritional recommendations used by feedlot nutritionists in Brazil." J. Anim. Sci. 87(10): 3427–39. doi: 10.2527/jas.2009-1880.

Millen, D. D., R. D. L Pacheco, P. M. Meyer, P. H. M Rodrigues, and M. D. B. Arrigoni. 2011. "Current outlook and future perspectives of beef production in Brazil." Anim. Front. 1(2): 46–52. doi:10.2527/af.2011-0017.

Morota, G., R. V. Ventura, F. F. Silva, M. Koyama, and S. C. Fernando. 2018. "Machine learning and data mining advance predictive big data analysis in precision animal agriculture." J. Anim. Sci. 96: 1540–50. doi: 10.1093/jas/sky014.

Oliveira, M. 2018. "Contributions of Brazilian cattle." Pesquisa FAPESP (264). https://revistapesquisa.fapesp.br/en/2018/06/25/contributions-of-brazilian-cattle/.

Oliveira, M. 2019. "Produção da pecuária municipal 2018." Catalog of the Instituto Brasileiro de Geografia e Estatística 84(01014234): 1–8. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=784.

Pham, X., and M. Stack. 2018. "How data analytics is transforming agriculture." Bus. Horiz. 61(1): 125–33. doi:10.1016/j.bushor.2017.09.011.

Rosa, G. J. M., and B. D. Valente. 2013. Breeding and genetics symposium:: Inferring causal effects from observational data in livestock. J. Anim. Sci. 91: 553-564.

Sun, Z., S., Samarasinghe, and J. Jago, 2010. Detection of mastitis and its stage of progression by automatic milking systems using artificial neural networks. J. Dairy Res., 77(2), 168–175. doi: 10.1017/S0022029909990550

Valletta, J. J., C. Torney, M. Kings, A. Thornton, and J. Madden. 2017. Applications of machine learning in animal behaviour studies. Anim. Behav. 124, 203–220. doi:10.1016/j.anbehav.2016.12.005

Zia, M., J. Hansen, K. Hjort, and C. Valdes. 2019. “Brazil once again becomes the world’s largest beef exporter.” United States department of Agriculture - Economic Research Service. https://www.ers.usda.gov/amber-waves/2019/july/brazil-once-again-becomes-the-world-s-largest-beef-exporter/ (October 21, 2019).

Chapter 2: Record linkage for farm-level data analytics: Comparison of deterministic, stochastic and machine learning methods

2.1 Abstract

The advent of big data in agriculture increased the necessity of extracting useful information from large data collections. This knowledge is critical in optimizing production systems, while also addressing prevailing issues such as sustainability. One of the first, yet crucial, data analytics steps comprises integration. Integrated data from different sources can provide enhanced insight, as they may retain complementary information on the same entity. In the absence of a "unique universal identifier" to link entities (e.g. farms) from different databases, it is necessary to rely on their recorded attributes (e.g. farm name, owner). We propose a fully automated framework to match farms, across different datasets (i.e. farm matching) in a big data context. To assess performance, we used information on Brazilian beef cattle farms from two large datasets: 44,566 farms that made purchases at an animal nutrition company, and 32,776 that processed cattle at a meat packing company. Geographical search space reduction was implemented as an alternative to reduce the number of comparisons evaluated. To compare attributes between farm pairs, we contrasted two edit-based approaches, the Levenshtein and Jaro-Winkler metrics. We also compare deterministic, stochastic, and machine learning (ML) approaches, for classification of farm pairs as match or non-match. These techniques have been used in other record linkage domains. The deterministic approach requires all attributes to match exactly. The probabilistic approaches tested were Epi-Weights (CR) and Fellegi-Slunter (FS). Unsupervised ML approaches were k-means and bagged clustering (BC). Supervised methods

were recursive partitioning trees, bagging of decision trees, bootstrap based classification trees, stochastic boosting, support vector machines (SVM), single-layer neural networks and logistic regression. Labels were produced by specialist review for both a training set of 295,012 comparison pairs and a testing set of 32,780. All techniques were evaluated in terms of testing set quality (accuracy, precision, sensitivity, and specificity) and completeness (number of matches) as well as efficiency (run-time). ML approaches outperformed the deterministic matching, which was superior to probabilistic methods. Within ML approaches, supervised methods outperformed unsupervised (except for BC). The best string metric was the Levenshtein. The best classification method in terms of quality and completeness was SVM (accuracy = 99.9%, precision = 91.1%, sensitivity = 97.3%, specificity = 99.9%), followed by BC (accuracy = 99.9%, precision = 90.8%, sensitivity = 93.2%, specificity = 99.9%). Results indicate that both SVM and BC are suitable for farm matching in scenarios where training labels are available, or not, respectively.

Keywords: big data; data integration; farm matching; probabilistic; machine learning

2.2 Introduction

The need for augmenting food production to feed an expanding human population, along with significant increases in data collection, warehousing, as well as availability have drastically changed the agriculture industry in the early 21st century. The Food and Agriculture Organization of the United Nations (FAO) estimates an increase of 30% in the global population by 2050 (FAO 2009). Agriculture must scale accordingly and the increases in production should be coupled with a minimal environmental footprint, improved animal welfare, and low enough costs to guarantee access for developing countries. The current situation emphasizes the need for approaches capable

of promptly evaluating food scenarios in order to sustainably optimize food supply. Large datasets have been extremely useful as a source of information to overcome challenges in many areas, including agriculture (Coble et al., 2018). Storing, processing, and analyzing large and complex data or "big data" (Christen, 2012b; Kamilaris et al., 2017) in a timely manner is critical for extracting useful information to address the current challenges of agriculture (Coble et al., 2018).

In fact, agriculture is rapidly becoming a data driven field with "precision agriculture" (i.e. data analytics in farming) being crucial for supporting decision making (Coble et al., 2018; Pham and Stack, 2018). Big data analytics in agriculture involves steps such as data processing, cleaning, integration, mining, and pattern extraction (Kamilaris et al., 2017). Data integration comprises connecting data from distinct sources into a unified collection (Lenzerini, 2002). This step is especially relevant to livestock systems because information from a specific farm, for example, recorded by different institutions/companies are generally disconnected across databases.

There are different ways to perform data integration. Some are more trivial such as linking information collected in the same time frame and others are more complex, such as record matching. Data matching involves identifying individual observations that belong to the same entity in different databases (or the same database, i.e. deduplication) in the absence of a "universal identifier" and connecting such information (i.e. data fusion) (Christen, 2012b).

Traditionally in the data matching field, those target entities are people, and the attributes one could rely on to make the linkage would be their names, addresses, etc. (Winkler and Thibaudeau, 1991). Christen (2012) highlights that entities can also be corporations, bibliographic citations, or commodities searched by users in shopping systems. Here we hypothesize that in the agricultural context, farms are one type of entity of interest to be matched. Common attributes of farms would be farm's name, location (city, state, address), state registration records, and

information about the owner. Defining a framework for farm matching will potentially allow connecting different sources of data.

Analyzing such integrated data can provide enhanced insight compared to independent analysis. This is because different data sources may contain complementary information on the same entity (Christen, 2012b), providing a more complete picture of the entire system, from production to the consumer (Coble et al., 2018). Results of such integrated analysis can, for example, point out the best management practices at the farm level for increased profit and productivity in different scenarios. However, farm matching is a very challenging task because usually the documentation of attributes for the entity to be linked (i.e. farm) is highly inconsistent across databases, due to misspellings, errors, or missing information, and manual data mining is not feasible to be implemented due to dataset size.

We suggest that farm matching can benefit from developments achieved for data linkage in the realm of statistics and health research (Fellegi and Sunter, 1969; Winkler and Thibaudeau 1991), and computer sciences (Winkler, 2006; Elmagarmid et al., 2007). Data linkage has been applied in many fields and different approaches are available for such task, ranging from more traditional stochastic (i.e. probabilistic) approaches to novel machine learning classification methods. It is worth noting that there is no "universal" best algorithm for data matching, as competing methods may perform differently across domains and match problems (Köpcke and Rahm, 2010). For this reason, when a new domain manifests the necessity for record linkage, such as agriculture, it is important to test different approaches and assess their performance.

The objectives of this study were to introduce the concept of "farm matching" in a big data livestock scenario and compare the performance of twelve different fully automated matching methods. The deterministic approach – DET (in which all attributes need to match exactly), was

compared to two probabilistic approaches (Epi-Weights – CR and Fellegi-Sunter weights – FS), two unsupervised machine learning approaches (k-means clustering – KC, and bagged clustering – BC), and seven supervised machine learning approaches (recursive partitioning trees – RPT, bagging of decision trees – BDT, bootstrap based classification trees – BCT, stochastic boosting – SB, support vector machines – SVM, neural networks – NN, and logistic regression – LR). For the probabilistic and machine learning approaches, both the Levenshtein metric and the Jaro-Winkler similarity criteria were contrasted as similarity functions. All methods were compared in terms of quality (accuracy, precision, sensitivity and specificity), completeness (number of correctly classified matches), and efficiency (run-time) in order to find an optimal approach for farm matching.

2.3 Material and Methods

The task of farm matching involves multiple steps, which are shown in Figure 2.1. Those steps are similar to any entity linkage process and they are well defined in domains such as statistic and epidemiology (Christen, 2012b) as well as computer sciences (Bilenko et al., 2005). In the case of two farm datasets to be matched (called here raw datasets I and II), the first step is the pre-processing of the variables to be utilized for matching. It consists of performing the same standardizations consistently in both datasets (such as transforming data to lower case, removing punctuation or special characters, expanding acronyms and correcting for common misspellings). The necessary standardizations are generally data specific.

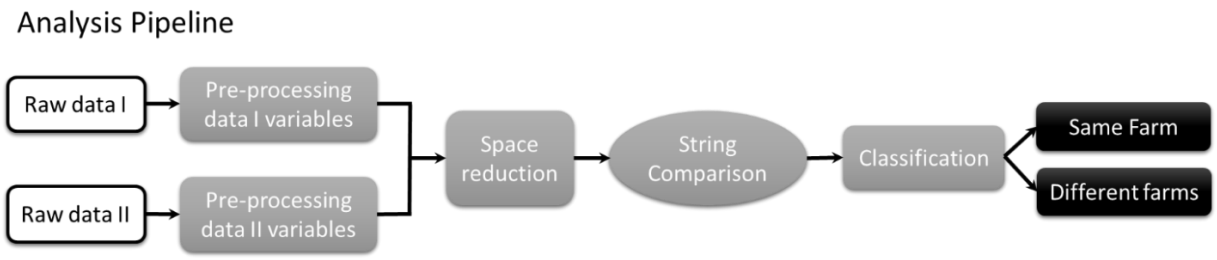

Figure 2.1: ***Pipeline of analysis for farm data linkage*****.** Lined rectangles on the left represent raw data from different sources to be matched. All four major steps of analysis are represented in gray: data pre-processing, search space reduction (or indexing), record pair comparison (string comparison utilized) and classification. Outputs of the classification method are presented in black.

The second step consists of a reduction of search space (also called indexing or blocking). This step is necessary because checking whether all possible pairs of farms are a match is costly. More precisely, the number of possible pairs is m x n, where m and n are the sizes of datasets I and II. Therefore, the number of pairs to be compared increases quadratically with the dataset sizes, making the search space reduction crucial. Especially for big data, the colossal number of possible comparisons to be evaluated constitutes one of the major challenges for linkage. Köpcke and Rahm (2010) suggested that it is necessary to reduce the search space in a way that produces likely matching pairs in order to achieve sufficiently fast execution times. Different approaches are available for this task (Christen, 2012a). The traditional approach is by applying blocking, where candidate pairs are required to share some attribute values, or "keys" (Jaro, 1989).

The definition of keys is critical. If the criteria are too broad, they may lead to over-selection of dissimilar pairs, decreasing efficiency. However, if the criteria are too restricted, they might sort out true matches, decreasing match quality (Köpcke and Rahm, 2010; Elfeky et al.,

2002). Christen (2012b) highlights the need for domain expertise when defining a key for blocking. For the specific issue of farm matching, we argue that a unique domain characteristic that is one of the most important attributes for identifying a farm is its geographical location, due to the impact of different environmental conditions on results. For this reason, performing disjoint blocking by geographical location becomes an efficient strategy to reduce the number of comparison pairs and guarantees that farms from the same region, for example only farms at the same city and state, will be compared. In scenarios where missing data on location is low and typos in such fields are rare (for example when location is selected from a category list), this type of space reduction would be feasible.

The third step consists of defining a record comparison function, such as a similarity function for comparing string (or numerical) pairs. Different string comparison functions allow 'fuzzy' appraisal of patterns and are available for this step, which allows for abbreviations and typos to still be matched. The Jaro-Winkler similarity criteria for names (Winkler, 1990) and Levenshtein metric for comparison of two strings (Levenshtein, 1965) are examples of such functions. Multiple attributes are generally compared for each candidate record pair, resulting in a comparison vector of similarity for each.

The fourth step in a fully automated pairwise approach comprises classification of pairs as matches (same farm) or non-matches (different farms) based on their full comparison vectors, where every single pair of candidate farms is classified independently from all others. Any binary classifier method can be employed for this task. Fully automated approaches with two categories (match or non-match) have the advantage of not depending on posterior clerical review (i.e. manual effort to label). They require less manual effort, which is important especially when dealing with large datasets, where those efforts can rapidly become unfeasible (Christen, 2008, Coble et al.,

2018). Köpcke and Rahm (2010) highlight that an ideal method should resolve tasks "automatically in a self-tuning manner". In the case of machine learning supervised methods however, the classifier must be trained using labeled data. Labeled data can be produced by curated clerical review (Elfeky et al., 2002) or when a unique identifier is available. Such a 'gold standard' can also be used to form a testing set, which allows assessing quality of the classification method of choice.

All previously discussed steps were implemented using a Brazilian beef cattle farm dataset, kindly provided by two major sources, an animal nutrition company (*DSM Produtos Nutricionais Brasil S.A.*) and a meat packing company (*JBS S.A.*). The DSM dataset (or dataset I) comprised of 44,566 farms that purchased beef cattle nutritional products from them and the JBS dataset (or dataset II) consisted of 32,776 farms that processed beef cattle in their facilities. Both sources collected data from January 2014 to December 2016. It is worth noting that the 32,776 farms in database II were responsible for the production of more than 19 million animals from 2014-2016, corresponding to 20.2% of the 94.2 million estimated Brazilian cattle production for the period (IBGE, 2018). The information available from both datasets to serve as attribute value matchers comprised location variables (city and state where the farm was located) as well as farm name, farm state ID, and name of the owner. Farm state ID was unique within states, but not between states. For protection of client privacy, all identifiers were mapped to positive integers prior to any processing, with the reverse map being held in a different secure location by the companies. Both databases were assigned unique identifiers within the dataset and did not contain duplicate records.

A brief description of the attributes as well as the amount of missing data associated to each of them is presented in Table 2.1. It is important to notice that the columns of city and state were selected from a pre-determined list in both datasets. It implies that no typos were allowed for

those fields, although there was a small chance that a wrong category had been selected. No missing data was associated with those columns in both datasets, and the previously mentioned technique of blocking by geographical reduction was implemented.

Table 2.1: Meta-data description of type of attribute, number of different attributes and frequency distribution of missing data.

Attribute	Attribute type	Different instances DSM	Different instances JBS	Missing data DSM	Missing data JBS
State	string	27	14	0%	0%
City	string	3,464	1,073	0%	0%
Name	string	21,278	12,501	11.36%	12.43%
State ID	string	39,260	27,915	0.54%	16.16%
Owner's name	string	39,437	26,929	0%	0%

A wide variety of methods were compared in this analysis for the task of farm matching (Figure 2.2). The first was a naïve single step DET, in which all fields needed to match exactly and no missing values are accepted. This is the most simplistic method that could be employed for the task of farm matching. The second group of methods comprised the stochastic framework developed by statisticians and epidemiologists where each entity pair is assigned a weight for the probability of match over non-match, and then an optimal threshold is adopted for classification. The stochastic methods applied here were CR; and FS (Fellegi and Sunter, 1969; Contiero et al.,

2005). The third group of methods encompassed modern machine learning techniques (both supervised and unsupervised) for the task of entity matching. More specifically, the unsupervised methods included KC (Hartigan and Wong, 1979) and BC (Leisch, 1999). While the supervised ones, contained RPT (Therneau, 1983), BDT (Breiman, 1996), BCT (Tibshirani and Knight, 1995), SB (Friedman et al., 2000), SVM (Cortes and Vapnik, 1995), NN (Ripley, 1996) and LR (Cox, 1958). We had three main objectives when choosing the appropriate machine learning algorithms. First, the chosen methods must span both unsupervised and supervised approaches. Second, we intended to take advantage of specific features of the methods that could be relevant for this paper (e.g. trees are known to be suitable in the presence of missing data and NN are robust to model non-linear relationships). Lastly, we aimed to contrast methods that have been successful for the task of entity matching in other fields (such as BC and SVM – Christen, 2008; Köpcke and Rahm, 2010) with methods that have not been much studied for such a task (e.g. NN and LR).

All stochastic and machine learning methods were implemented using two different types of string metrics for all attributes: Jaro-Winkler similarity criteria (Winkler, 1990) and Levenshtein edit distance metric (Levenshtein, 1965), both of which are continuous comparison values (i.e. comparison functions based on a distance metric between two values). This had the objective of assessing which string metric criterion provides better quality and completeness across different classification methods. More specifically, the Levenshtein metric can be defined as simply the minimum number of single character edits (insertion, deletion or substitutions) required to change a word into another. The Jaro-Winkler similarity criteria uses the Jaro distance, defined as the minimum number of single-character transpositions to change a word into another. This distance can be calculated using the number *m* of common characters that are within half the length of the longer string and the number of transpositions *t* as follows*:*

$$\text{Jaro distance} = \frac{1}{3}\left(\frac{\text{m}}{|\text{s}_1|} + \frac{\text{m}}{|\text{s}_2|} + \frac{\text{m} - \text{t}}{|\text{m}|}\right)$$

The Jaro-Winkler similarity criteria improves upon the Jaro algorithm by applying ideas based on empirical studies (fewer errors occur at the beginning of strings) and can be defined as:

$$\text{Jaro Winkler distance} = Jaro\ distance + l(1 - Jaro\ distance)$$

where l is the length of the common prefix at the start of the string up to a maximum of four characters.

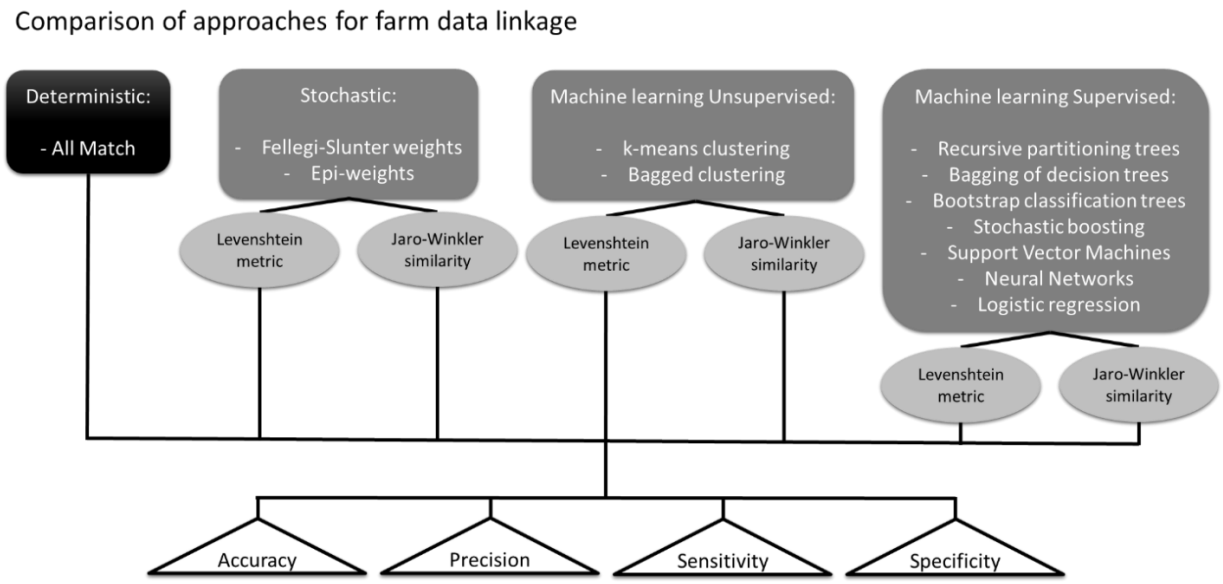

Figure 2.2: ***Schematic representation of model quality comparisons performed***. Rounded rectangles represent the classification approaches. The rectangle in black corresponds to the deterministic approach while the dark grey ones correspond to different classification methods. Light grey circles represent the string comparison method applied. Triangles represent the quality criteria used for comparison.

Evaluating all possible pairs for matching in this dataset would yield a total of approximately 1.4 billion comparisons. After applying space reduction using the geographical information of city and state in this dataset, we reduced this value over 500-fold, meaning the number of comparisons was reduced to 2.9 million pairs. The complexity of data matching is often characterized by the "reduction rate", defined as the number of record pairs generated by the blocking technique divided by all possible pairs (Christen, 2012b; Elfeky et al., 2002). The geographical space reduction suggested here was responsible for a complexity reduction of over 99%. However, it should be noted that it is still unfeasible to manually review all of the almost 3 million remaining pairs. Another measure of blocking quality is "pair completeness" (Elfeky et al., 2002; Köpcke and Rahm, 2010) , defined as the proportion of truly matching pairs preserved after blocking. As true farm matching is constrained within state and city, this value corresponded to 100% with our blocking criteria.

In order to implement the proposed approaches, both a training set and a testing set were defined. The training set was used for the supervised machine learning methods as well as to select an optimal threshold (Sariyar and Borg, 2010) for the probabilistic methods that aimed to minimize the number of misclassified record pairs. In order to choose a training set that would allow a manageable effort for this analysis we used the "minimal training set" approach proposed by Sariyar and Borg (2010). This approach consisted of a stratified sampling strategy in which pairs are chosen at random to span all feature regions present in the data set. Sariyar and Borg (2010) reported that supervised classification using this approach provides similar results to having much larger non-stratified random samples, making it a viable option for large collections. In this analysis, the minimal training set comprised 295,012 comparison pairs. In order to assess the efficiency of different methods, a testing set was chosen at random from all remaining pairs that

would compose 10% of all labeled observations (32,780 pairs). In this way, the desired proportions of 90% - 10% for training and testing sets, respectively, were achieved. Curated labels were produced by clerical review of domain experts, as suggested by Elfeky et al. (2002), for the 327,792 pairs. More specifically, reviewers evaluated farm attributes (farm name, state ID and owner's name) of each comparison pair, and based on such evidence deemed the pair as the same farm or not. A second domain expert checked every entry and, in the few disagreement cases, both reviewers discussed the proper label. Labels produced by review were used as gold standard both to train the models (semi-automatic selection) and to quantitatively assess the quality of the different approaches implemented. For the LR method, a Precision-Recall (PR) curve was used to determine the best probability threshold for classification.

All methods were compared in terms of quality for the labeled testing set (Figure 2.2). Quality was considered as measures of accuracy, defined as $\left(\frac{TP+TN}{TP+FP+TN+FN}\right)$; precision, defined as $\left(\frac{TP}{TP+FP}\right)$; sensitivity (or recall), defined as $\left(\frac{TP}{TP+FN}\right)$; and specificity, defined as $\left(\frac{TN}{TN+FP}\right)$; where TP stands for true positive, TN stands for true negative, FP stands for false positive, and FN stands for false negative. They were also compared in terms of completeness (Christen and Goiser, 2007), number of matches that each algorithm was able to determine compared to the gold standard, and execution time. The analyses were conducted using the "record linkage" R package (Sariyar and Borg, 2010) in a system with 8 CPUs and 126GB of RAM available.

When assessing matching/non-matching proportions in the training set, considerable imbalance (0.5% matches vs 99.5% non-matches) was noted. To address this issue, striving for improved model performance, the synthetic undersampling combined with oversampling approach (SMOTE) proposed by Chawla (2002) was implemented for all machine learning supervised

methods. Training sets with matching/non-matching proportions ranging from 10%-90%, 20%-80%, 30%-70%, 40%-60%, and 50%-50% were created, and the same measures of quality described above were used for the testing set.

2.4 Results

Results for the comparison of methods in terms of quality are shown in Table 2.2. As discussed in the previous section, two different types of string pair metric were utilized for all probabilistic and machine learning methods applied. The Jaro-Winkler string metric did not outperform the Levenshtein, while the latter performed consistently better for all machine learning (supervised and unsupervised) approaches, and provided similar results for probabilistic methods.

Regarding the comparison of classification approaches, overall, most machine learning methods outperformed both probabilistic and deterministic ones, with most supervised methods outperforming unsupervised (with the exception of BC). The FS probabilistic approach had worse performance than DET while CR performed very similarly.

The DET approach provided high accuracy and specificity along with perfect precision, meaning that all matches identified were TP. However, the sensitivity was small, indicating that it fails to identify a high percentage of pairs that correspond to a real match. For the probabilistic methods, the FS with both Levenshtein and Jaro-Winkler string metric provided nearly zero accuracy, precision, and specificity, while sensitivity for both was 100%, meaning that all TP were correctly classified as a match, but many FP were produced. The CR probabilistic method yielded virtually same results described above for the DET method with both string metrics.

Table 2.2: Quality of different predictive approaches used to perform farm data linkage.

Class	Method		String comparison	
			Levenshtein	Jaro-Winkler
Deterministic	All match	Accuracy (%)	99.9	
		Precision (%)	100.0	
		Sensitivity (%)	35.6	
		Specificity (%)	100.0	
Probabilistic	Fellegi-Slunter	Accuracy (%)	0.3	0.3
		Precision (%)	0.2	0.2
		Sensitivity (%)	100.0	100.00
		Specificity (%)	0.0	0.1
	Epi-Weights	Accuracy (%)	99.8	99.8
		Precision (%)	100.0	100.0
		Sensitivity (%)	35.1	35.1
		Specificity (%)	100.0	100.0
ML unsupervised	K-means clustering	Accuracy (%)	46.0	20. 8
		Precision (%)	0.3	0.2
		Sensitivity (%)	78.4	79.7
		Specificity (%)	45.9	20.6
	Bagged clustering	Accuracy (%)	99.9	14.9
		Precision (%)	90.8	0.2
		Sensitivity (%)	93.2	95.9
		Specificity (%)	99.9	14.7
ML supervised	Recursive partitioning trees	Accuracy (%)	99.9	99.2
		Precision (%)	86.7	21.0
		Sensitivity (%)	97.3	98.6
		Specificity (%)	99.9	99.2
	Bagging of decision trees	Accuracy (%)	99.9	89.7
		Precision (%)	83.9	02.1
		Sensitivity (%)	98.6	97.3
		Specificity (%)	99.9	89.7
	Bootstrap classification trees	Accuracy (%)	99.9	99.6
		Precision (%)	88.9	39.5
		Sensitivity (%)	97.3	98.6
		Specificity (%)	99.9	99.7
	Stochastic boosting	Accuracy (%)	99.9	81.5
		Precision (%)	85.7	01.0
		Sensitivity (%)	97.3	100.0
		Specificity (%)	99.9	81.4
	Support Vector Machines	Accuracy (%)	99.9	89.8
		Precision (%)	91.1	02.1
		Sensitivity (%)	97.3	100.0
		Specificity (%)	99.9	89.5
	Single-hidden-layer neural networks	Accuracy (%)	99.8	99.8
		Precision (%)	-	-
		Sensitivity (%)	0.0	0.0
		Specificity (%)	100.0	100.0
	Logistic regression	Accuracy (%)	99.92	99.91
		Precision (%)	88.89	90.74
		Sensitivity (%)	75.67	66.22
		Specificity (%)	99.98	100.00

*Deterministic class did not use either Levenshtein or Jaro-Winkler since it was manually linked.
- Symbol is used for indeterminate calculation results.

Regarding the machine learning unsupervised methods, KC yielded poor results in terms of accuracy and specificity, nearly zero precision, and moderate sensitivity for both string metrics. The BC approach, on the other hand, provided nearly perfect accuracy and specificity, and greater than 90% precision and sensitivity when combined with the Levenshtein string metric, proving to have a substantial ability to correctly classify matching pairs. Results for BC with Jaro-Winkler metric, however provided low accuracy, precision, and specificity, with high sensitivity.

Concerning machine learning supervised methods the decision tree-based methods RPT, BDT, BCT, and SB provided similar results amongst them. All had very high accuracy, sensitivity, and specificity but moderate precision for the Levenshtein metric and worse results for those combined with low precision when used with the Jaro-Winkler metric. That indicates that even though the overall model performs well, a fair amount of matching pairs were mistakenly classified as FP. SVM combined with the Levenshtein metric yielded the best results among all approaches, with nearly perfect accuracy, sensitivity, and specificity along with very high precision. In fact, the SVM model combined with the Levenshtein metric classified most pairs correctly, except for five FP. When combined with Jaro-Winkler, however, the performance of SVM was inferior. NN presented the worst results across all models. Even though accuracy and specificity were high, all pairs were classified as non-matches leading to an indeterminate precision value and zero specificity. LR combined with the Levenshtein metric yielded near-perfect accuracy and specificity, however sensitivity was lower than most supervised methods and precision was moderate-high. When combined with Jaro-Winkler, however, accuracy and specificity were virtually 100%, but with lower sensitivity and moderate-high precision. LR provided the best results for the Jaro-Winkler metric across all classifiers, but such results are still inferior than the

ones obtained with the Levenshtein metric due to lower sensitivity and considerably smaller area under the PR curves.

Regarding completeness, best results should have values close to the gold standard combined with very high precision, meaning that most of the pairs classified as match are TP and very few classification mistakes (FP) occur. Figure 2.3 indicates that the best results were obtained using SVM, and BC when combined with the Levenshtein metric, with SVM performing slightly better. The worst results were observed for the FS model, which mistakenly classified the vast majority of pairs as a match (highly inflated FP rate), and for NN, that in contrary classified all pairs as a non-match (highly inflated FN rate), as seen in Figure 2.3.

Results for computational efficiency of all probabilistic and machine learning methods are presented in Table 2.3. The FS model combined with Jaro-Winkler string comparison yielded the worst run-time of 43.5 min. The shortest run times were achieved by the CR model with Levenshtein (1.7 min) and Jaro-Winkler (1.6 min). It is noticeable that there is a considerable efficiency difference between approaches. When we compare the best and worst run-times, there is a 26-fold difference amongst them. However, none of the methods yielded unmanageable times in this analysis for a training plus testing set size of over 300,000 observations.

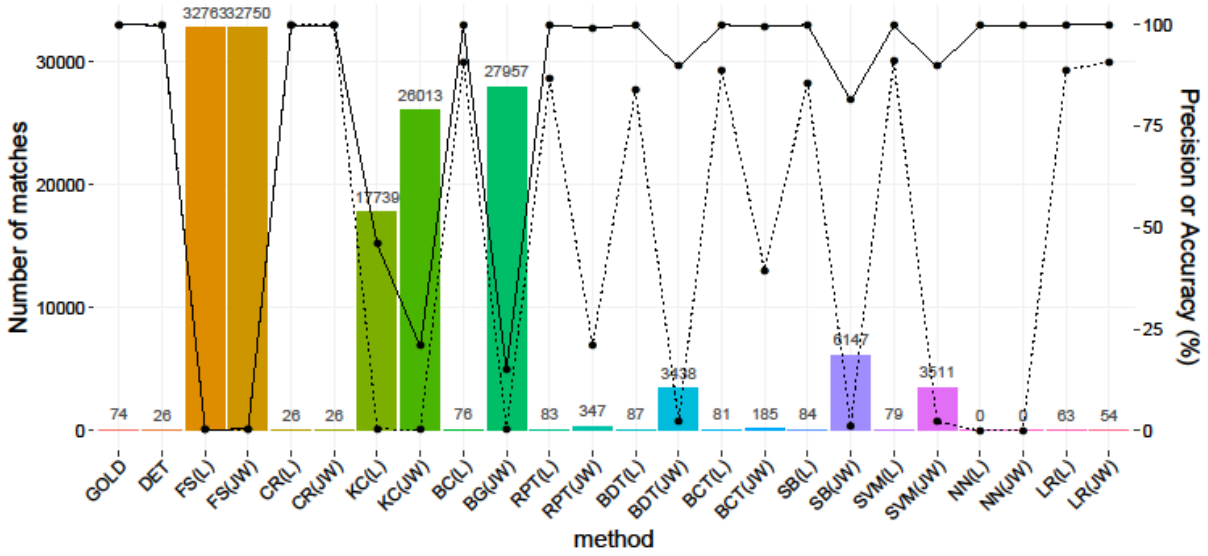

Figure 2.3. ***Comparative completeness for all methods assessed.*** Number of farm match pairs (colored bars), accuracy (–•–) and precision (···•···) achieved by each algorithm. Abbreviations are as following: GOLD – gold standard; DET – deterministic approach; FS – Fellegi-Sunter probabilistic method; CR – Epi-Weights probabilistic method; KC – K-means clustering; BC – bagged clustering; RPT – recursive partitioning trees; BDT – bagging of decision trees; BCT – bootstrap classification trees; SB – stochastic boosting; SVM – support vector machines; NN – single-hidden-layer neural networks; LR – logistic regression. For each probabilistic and machine learning approach, two string metrics were contrasted: Levenshtein (L) or Jaro-Winkler (JW).

Figure 2.4 presents a 3D representation of the classification performed by the two best approaches (SVM and BC with the Levenshtein metric, respectively). When comparing the matching status of pairs amongst the two methods, it is noticeable that results are very similar and pairs with similarity that is closer to one for all attributes tend to be consistently classified. The difference between the two approaches is that BC tends to mismatch pairs where one single

similarity axis is (very close to) zero, while SVM tends to miss, overall, fewer pairs for which all three attributes have values located in the center of the distribution.

Table 2.3: Computational efficiency defined in terms of execution time (min) of: search space reduction, string comparison and classification for all probabilistic and machine learning methods tested. For the supervised methods it included the time needed for training the algorithms.

Class	**Method**	**Run-time (min)**	
		String comparison	
		Levenshtein	Jaro-Winkler
Probabilistic	Fellegi-Slunter	9.3	43.5
	Epi-Weights	1.7	1.6
ML unsupervised	K-means clustering	3.8	3.8
	Bagged clustering	8.0	6.8
ML supervised	Recursive partitioning trees	3.4	4.2
	Bagging of decision trees	5.7	8.4
	Bootstrap classification trees	5.7	5.7
	Stochastic boosting	11.3	11.5
	Support Vector Machines	21.3	21.0
	Neural networks	4.2	4.2
	Logistic regression	4.0	4.0

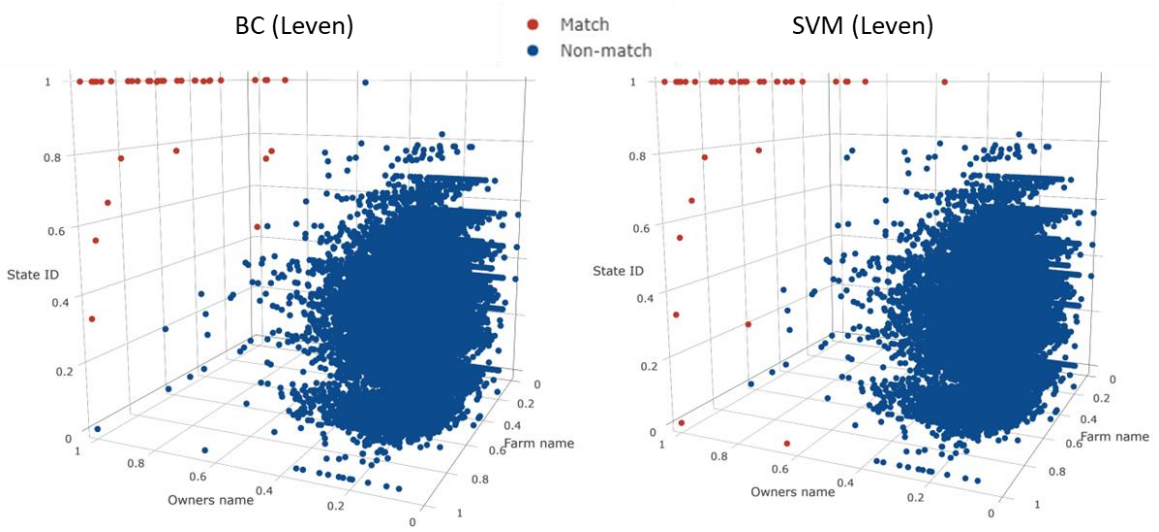

Figure 2.4. ***3-D plot of best testing set classification results.*** Farm match is represented in red and non-match in blue. The two methods with best results: bagged clustering – BC; and Support Vector Machines – SVM; both combined with the Levenshtein string metric – Leven are represented.

Results for the machine learning supervised methods obtained using synthetic training set matching/non-matching proportions with the Levenshtein metric are presented in the Appendix Table 2.A.1. We observe that with the exception of NN, no further improvement in quality was achieved for any of the methods at any of the levels of balancing by SMOTE tested, when compared to the results obtained with the original data presented in Table 2.2. For NN, gains in terms of precision and sensitivity were observed, nonetheless, despite such improvements this method still presented the worst quality across all supervised methods.

2.5 Discussion

When performing any type of entity matching it is important to consider several evaluation measures for making a decision on which approach is more suitable for a specific scenario (Bilenko et al., 2005), such as farm matching. It is also important to recognize that both quality and completeness are affected by all steps of the analysis (Christen, 2012b). When dealing with large datasets (i.e. big data), the interpretation of quality measures should be done carefully because such data will likely be highly unbalanced. This is because, as previously mentioned, while the number of matching pairs increases quadratically (or sub-quadratically after space reduction), the number of possible matches only increases linearly, therefore the proportion of matches to non-matches in large datasets tends to be very tiny (Christen, 2012b). For example, accuracy (or the proportion of correctly classified pairs) alone might not be a satisfactory evaluation measure in such a scenario because if a method completely fails to identify all TP (i.e. it labels all pairs as non-match), it will still yield high accuracy. For this reason, the examination below builds on combined evidence of different quality measures to assess and compare methods.

A relevant feature of any record linkage application is the choice of function used to measure similarity between records (Bilenko et al., 2005). Both Jaro-Winkler and Levenshtein are edit-base measures of string distance. In other words, they count the changes necessary to make two strings equal (Christen, 2012b), and are known to yield better results for record linkage than token-based measures (Köpcke and Rahm, 2010; Carreras et al., 2018). The aforementioned results for the string pair metrics suggest that the Levenshtein is consistently superior for farm matching than the Jaro-Winkler similarity, for all measures of quality. This differs from previous results using people as an entity to link (Elfeky et al., 2002). However, it is relevant to mention that the dataset used in this analysis was collected in Brazil, and attributes therefore had a Portuguese

origin, while previous results used English words. Portuguese names tend to be longer, and farm owners most times have multiple last names which led to lots of abbreviations in this dataset, imposing a challenge to record matching. The Levenshtein approach (Levenshtein, 1965) defines the difference between two string as the minimum number of single character edits (insertion, deletion or substitutions) required to change a word into another. For this reason, it functions well with abbreviations and might be better suited for the task of farm matching in Portuguese. The Jaro-Winkler similarity (Winkler, 1990) on the other hand, defines distance as the minimum number of single-character transpositions to change a word into another, which might be better suited for English phonetics applications.

With respect to the comparison between classification methods, overall results indicated that machine learning techniques performed better than deterministic, which in turn outperformed probabilistic approaches. Only a few previous publications in other domains have evaluated comparatively several techniques (Christen, 2012b), with most being between deterministic and probabilistic approaches, and a few between probabilistic and machine learning. For the former, with both simulated (Carreras et al., 2018) and real data (Campbell et al., 2007), there is a trend of superior performance of probabilistic approaches. However, in such applications the reasoning behind the choice of a threshold for determining a match was not clear and the cut-off was defined manually to achieve optimal results. This differs from the fully automated process applied here. In previous studies, when contrasting probabilistic and machine learning techniques, machine learning consistently outperformed probabilistic (e.g. Elfeky et al., 2002). Within the machine learning framework, our results showed that supervised methods outperformed unsupervised ones for farm matching. This is in agreement with applications of entity matching in other domains such

as computer sciences with both synthetic data, and real consumer matching databases (Elfeky et al., 2002; Christen, 2008).

We also observed a large difference in performance (quality, completeness and efficiency) of various methods within each class. DET results obtained in this study reinforced the belief that while this method can be effective in allocating non-matches, it tends to misclassify real matches as non-matches (FN) if any sort of discrepancy (such as a misspelling or abbreviation) was present in the entity attributes (Christen, 2012b). This is reflected in the sensitivity (i.e. proportion of real matches captured by the model) of 35.6%, indicating that only about one third of the real matches was identified with this approach.

Regarding probabilistic results, the low accuracy, precision, and specificity experienced by the FS contradicts some results in other domains (Campbell et al., 2007; Carreras et al., 2018). However, its high sensitivity has been previously observed when an appropriate threshold for match was chosen (Campbell et al., 2007). It should be noted that, as previously mentioned, in such studies the threshold for matching was determined manually, which differs from the fully automated approach implemented here. As defined above, the fully automated paradigm aims to minimize the rate of FP and FN in order to choose a suitable threshold (Sariyar and Borg, 2010). In this analysis it yielded a very low threshold for matching. While not a single TP was misclassified, it generated high numbers of FP matches. Elfeky et al. (2002) argue that the minimization of the probability of error might not be the most appropriate technique to select a matching threshold, since distinct incorrect classifications may have diverse implications. Campbell et al. (2007) even acknowledge that a program's capability will not be attained if an inappropriate cut-off is chosen, which may have occurred in this analysis. While the paradigm developed by Fellegi and Sunter (1969) has been successfully applied in epidemiology, authors

proved that a two-threshold procedure with clerical review would be ideal for their method. Yet, this it is unfeasible in a big data scenario.

Results for the CR model were virtually the same as for DET, contrasting the argument that probabilistic approaches should be able to detect matches that DET misses (Campbell et al., 2007; Christen, 2012b). The CR model could be experiencing the same threshold issues described for FS. Lastly, probabilistic approaches add complexity into the linkage process and can be time consuming (Carreras et al., 2018). In this analysis, the Expectation Maximization approach suggested by Fellegi and Sunter (1969) yielded the longest run-time across all methods, which should be considered when analyzing larger data or when a less effective reduction rate is achieved.

Machine learning algorithms produce a model that is able to predict the class of an unclassified object (Elfeky et al., 2002). In our application, the class is farm matching status and the object is the similarity score between attributes of a candidate farm pair. Unsupervised methods aim to arrive at a function by similarity between attributes (Elfeky et al., 2002). They have the strong advantage of not requiring labels, which is core in big data applications, especially when resources to create curated labels are limited or unavailable. For example, in order to produce the 295,012 training labels for this analysis, approximately 250 hours of field specialist's efforts were necessary.

Of the unsupervised machine learning algorithms tested in this study, the BC outperformed the KC by a large margin. Results for the KC method were poor regarding all quality measures. It is important to acknowledge that the choice of location of initial k centers is known to influence how classes are grouped. This is the case especially when a single random start is used, which is the set up in the Sariyar and Borg (2010) package used. BC bootstraps the data (sampling from data at random with replacement) to form new sets and then performs a hierarchical clustering on

the centers to choose the most appropriate ones. It appears to have overcome the KC problem, but at the cost of doubling the run-time. Clustering approaches had previously provided high quality measures in applications with different set ups (Elfeky et al., 2002; Bilenko et al., 2005). They even outperformed methods that build a training set using clustering assignment and then apply it to train a supervised approach for classification (Elfeky et al., 2002). The success of a clustering method however, is claimed to depend on the application (Sariyar and Borg, 2010). For the task of farm matching, the BC approach provided fairly reliable results, making very few classification mistakes and providing the second best results across all classifiers tested. In scenarios where labels for a training set are not available, this method can be considered as a suitable option.

Supervised machine learning methods rely on a training set to learn patterns in the data that are used for classifying objects. The dependence on a training set, which might not always be available, is a limitation of supervised approaches (Elfeky et al., 2002). Decision trees are one type of such method that will determine a set of rules for classifying an object. RPT is a binary tree that uses a recursive partitioning method with the Gini coefficient of inequality (Stuart and Ord, 1994) as a splitting criteria. A single tree is produced, hence results tend to be more unstable. BDT tries to overcome such volatility by bootstrapping the data and building many separate trees. The final model is then formed by a "majority vote" across all trees. BCT also bootstraps many trees, but instead of choosing the final model by majority vote, it maximizes a quality function (accuracy in the Sariyar and Borg (2010) set up) to elect the best tree. Lastly, SB is also bootstrap-based. After forming the first tree it calculates a quality measure (again accuracy in the Sariyar and Borg (2010) set up) and subsequently samples with replacement, assigning higher weights to values that obtained worse quality, until quality no longer improves. Trees are combined by the weighted average of different predictions.

Tree-based approaches are known to be significantly faster than SVMs (Köpcke and Rahm, 2010) for record linkage. This is consistent with the run-time results observed for all tree-based methods in this analysis (3 to 11 vs 21 min). Quality indicators among those were fairly similar for the Levenshtein metric. All indicators were above 90%, except for precision, which was in the 84-89% range. For BCT and SB, quality could be closely related with the working criterion of choice: accuracy alone in the Sariyar and Borg (2010) set up. It might not be the most suited for the unbalanced entity matching data, as previously discussed. Amongst the trees, the best precision results were attained by BCT. Previous research in other domains pinpoint that decision trees have better performance when large amounts of training data are available (Köpcke and Rahm, 2010).

The SVM approach with the Levenshtein metric yielded the best results regarding all quality measures across all methods. SVM draws a hyperplane in a n-dimensional space (where n is the number of attributes) and optimizes the distance between two support vectors and the hyperplane in the training set. The defined hyperplane is then used to classify objects in the testing set. SVM is known to be robust for record linkage in other domains. It outperformed cluster methods (Christen, 2008) and decision trees, even when training data was limited (Köpcke and Rahm, 2010). Our results suggested that it is also highly suitable for the task of farm matching. If training data is available, we suggest that this should be the method of choice. The efficiency of this method was the worst among the machine learning techniques, with a run-time at least twice as large as other methods. However, such run-time was not a limiting factor in this analysis.

Neural networks are based on how neurons work in the brain, consisting of layers of interconnected units (neurons). The single hidden-layer neural network applied here consists of an input layer (with the similarity values for the attributes), one middle (or hidden) layer and an output layer that provides the probability for each class (same farm or different farms). Information flows

from the input to the output layer, and each neuron computes a weighted sum of its inputs and applies a non-linear activation function (logistic used here) to predict the output. Backwards propagation was applied to obtain hyperparameters (weights and number of hidden units) that maximize entropy (or the likelihood of the Bernoulli distribution). NN produced the worst results for the task of farm matching across all methods, classifying all pairs as a non-match. This method could be suffering from data unbalancing. Also, it is very sensitive to the choice of model hyperparameters and we argue that the Sariyar and Borg (2010) set up might be not optimal, as it does not allow fine tuning of some parameters for farm matching. More precisely, one single layer might not be sufficient to suit the complexity of relationships and other hyperparameters (number of epochs, neurons, penalization and activation functions, etc.) could be poorly tuned.

LR is a generalized linear model in which a binary outcome (i.e. same or different farms) is converted to a linear scale using the log odds ratio of the probability of success (i.e. match). It is modeled as a function of explanatory variables (i.e. similarity values across attributes). Model parameters were estimated using an iterative re-weighted least squares approach. Estimated class probabilities were obtained by applying a reverse link function to linear scale results. The PR optimal cut-off aimed to maximize both precision and sensitivity. Under the Levenshtein metric, LR provided results that were competitive with the tree approaches but inferior to SVM and BC. When combined with the Jaro-Winkler metric, there was an 2% increase in precision but a 9% decrease in sensitivity as well as a 17% decrease in the area under the PR curve, indicating the Jaro-Winler metric performed worse than the Levenshtein's. As shown in Figure 2.3, with Jaro-Winkler, only 54 pairs were deemed as match (49 of which were TP), missing 20 real matches.

Regarding the utilization of synthetic training sets with balanced proportion on supervised methods, we highlighted that overall, no improvement was achieved. This is likely related to the

fact that the approach used for choosing a training set in this analysis was already optimal. Therefore, undersampling the majority class will likely lose important information for classification, while synthetically increasing the underrepresented class does not seem to improve the decision boundary for most models.

One important aspect of these results is that measures of quality reflect the degree of correspondence between the results obtained from manual review and respective linkage method. Albeit clerical review is not error free, it is a valid mechanism to resolve uncertainty in matching and has been successfully adopted in other linkage applications (Köpcke and Rahm, 2010).

Another important feature of the presented results is that, much like applications in other record linkage domains (Köpcke and Rahm, 2010), all analyses were performed using real data. Utilizing real information in comparison to simulated data has the benefit of being more realistic, as errors are domain specific and such collections embody a wide variety of errors that could occur in practice (Köpcke and Rahm, 2010). Even though we present a suitable solution for the described data, future applications should proceed carefully as the difficulties of farm matching may differ with the structure and quality of the data at hand. It is important to reiterate that with a lack of unique identifiers or poor-quality identifiers (such as those that are not accurate or change over time) matching must rely upon common attributes across different data sources. As those attributes are usually heterogeneous for the same entity (Elfeky et al., 2002), the task at hand is challenging. The choice of strategy (algorithm and string metrics) is case specific (Carreras et al., 2018) and we suggest that different approaches should be contrasted. Farm data, as any real data, is messy (Elfeky et al., 2002) and a number of factors can influence the result.

Unlike applications with smaller datasets in other domains (Campbell et al., 2007; Carreras et al., 2018) we defend a fully automated farm matching procedure to deal with large datasets in

agriculture. With the advent of big data, the demand for fully automated techniques is increasing as methods that require manual effort can rapidly become unfeasible (Christen, 2008).

Lastly, it is worth noting that all techniques described here can also be used to address deduplication in datasets consisting of an important data cleaning step (Köpcke and Rahm, 2010). In either case, it comprises a crucial step in the knowledge discovery process (Elfeky et al., 2002) that has recently become relevant to big data integration in agriculture. In fact, the third revolution of agriculture (Pham and Stack, 2018) envisions the use of data analytics to boost productivity while lowering inputs and reducing the environmental footprint (Kamilaris et al., 2017). Albeit on its early stages, this revolution is flourishing and efficiently performing data integration constitutes one of the early, yet crucial steps for achieving this goal. In such a scenario, adapting methods and proposing new solutions, such as farm matching, to the arising issues of big data analytics in agriculture becomes very relevant.

For many applications in agriculture the integration step constitutes a bottleneck for any further analysis. Not only enhanced results can be obtained by connecting information in different sources of data, but if the outcome variable (such as carcass quality) and predictors (such as animal nutrition) are contained in different databases (as here observed in the JBS and DSM data, respectively), it is not possible to even start the analysis before properly linking datasets.

While we are not aware of previous utilization of record matching techniques in agriculture, this analysis proposed a powerful framework and provided a comprehensive comparison of available methods for farm matching. The results are applicable for a multiplicity of operations that rely on efficiently connecting information of farms from different sources in the absence of a high quality, universal identifier to address the specific problems it tackles.

2.6 Conclusions

Results show that the geographical space reduction proposed, despite being a notably simplistic solution and capturing an inherent feature of farms, was effective in reducing the complexity of farm matching. We successfully introduced the concept of farm matching in a big data scenario in agriculture. Different deterministic, stochastic, and machine learning approaches applied in other fields of record linkage were contrasted for farm matching. The best string metric to compare pairs was the Levenshtein metric. Two fully automated machine learning classifiers yielded the best quality and completeness in allocating farm pairs into the categories of match and non-match. One was supervised (SVM) and another unsupervised (BC), both could address the challenge of farm matching, producing viable solutions in the presence or absence of labeled data. While SVM produced slightly superior results, it required reasonable amounts of labeled data.

2.7 References

Bilenko, M., Basu, S., Sahami, M., 2005. Adaptive product normalization : using online learning for record linkage in comparison shopping. In: Fifth IEEE International Conference on Data Mining 58–65.

Breiman, L., 1996. Bagging predictors. Machine learning 24(2): 123–40.

Campbell, K. M., Deck, D., Krupski, A., 2007. Record linkage software in the public domain: a comparison of link plus, the link king, and a ‘basic’ deterministic algorithm. Health Informatics Journal 14(1): 5–15.

Carreras, G., Simonetti, M., Cricelli, C., Lapi, F., 2018. Deterministic and probabilistic record

linkage: an application to primary care data. Journal of Medical Systems 42(82): 1–3.

Chawla, N. V., Bowyer, K. W., Hall, O. L., 2002. SMOTE : synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16: 321–57.

Christen, P., 2008. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining 151-159.

Christen, P., 2012a. A survey of indexing techniques for scalable record linkage and deduplication. In: IEEE Transactions on Knowledge and Data Enginieering, 1537–55.

Christen, P., (First) 2012b. Data matching: concepts and techniques for record linkage, entity resolution and duplicate detection. Springer, London.

Christen, P., Goiser, K., 2007. Quality and complexity measures for data linkage and deduplication. In: Quality measures in data mining, ed. Hamilton Guillert. Studies in computational inteligence. Bangkok, Thailand, pp. 507–14.

Coble, K. H., Ashok, K. M., Ferrel, S., Griffin, T., 2018. Big data in agriculture: a challenge for the future. Applied Economic Perspectives and Policy 40(1): 79-96.

Contiero, P., Tittarelli, A., Tagliabue, G., Maghini, A., Fabiano, S., Crosignani, P., Tessandori, R., 2005. The epilink record linkage software. Methods of Information in Medicine 44(1): 66–71.

Cortes, C., Vapnik, V. N., 1995. Support-vector networks. Machine learning 20(3): 273–97.

Cox, D. R., 1958. The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological) 20(2): 215–42.

Elfeky, M. G., Verykios, V. S., Elmagarmid, A. K., 2002. TAILOR : A record linkage toolbox. In: Proceedings of the 18th IInternational Conference in Data Engineering 17.

Elmagarmid, A. K., Ipeirotis, P. G., Verykios, V. S., 2007. Duplicate record detection : A survey. In: IEEE Transactions on Knowledge and Data Enginieering 1–16.

FAO, 2009. How to feed the world in 2050. Food and Agriculture Organization of the United Nations, Rome.

Fellegi, I. P., Sunter, A. B., 1969. A theory for record linkage. Journal of the American Statistical Association 64(328): 1183–1210.

Friedman, B. Y., Hastie, T. J., Tibshirani, R., 2000. Additive logistic regression : A statistical view of boosting. The Annals of Statistics 28(2): 337–407.

Hartigan, J. A., Wong, M. A., 1979. Algorithm AS 136: A k-means clustering algorithm. Applied Statistics 28: 100–108.

IBGE, 2018. Indicadores IBGE estatística da produção pecuária. Retrieved from URL: ftp://ftp.ibge.gov.br/Producao_Pecuaria/Fasciculo_Indicadores_IBGE/abate-leite-couro-ovos_201802caderno.pdf.

Jaro, M. A., 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association 84: 414–20.

Kamilaris, A., Kartakoullis, A., Prenafeta-boldú, F. X., 2017. A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture 143: 23–37.

Köpcke, H., Rahm, E., 2010. Frameworks for entity matching : A comparison. Data and Knowledge Engineering 69: 197–210.

Leisch, F., 1999. Bagged Clustering. Adaptative information systems and modeling in economics and management science, 51. WU University of Economics and Business, Vienna.

Lenzerini, M., 2002. Data integration : A theoretical perspective. In: ACM PODS, Madison, 233–46.

Levenshtein, V., 1965. Binary codes capable of correcting spurious insertions and deletion of ones. Problems of information Transmission 17: 1–8.

Pham, X., Martin S., 2018. How data analytics is transforming agriculture. Business Horizons 61(1): 125–33.

Ripley, B. D., 1996. Pattern recognition and neural networks. Cambridge, New York.

Sariyar, M., Borg, A., 2010. The record linkage package: detecting errors in data. The R journal 2(2): 61–67.

Stuart, A., Ord, J.K. (6th Eds), 1994. Advanced theory of statistics. Edward Arnold, London.

Therneau, T. M., 1983. A short introduction to recursive partitioning, Orion Technical Report 21.

Tibshirani, R., Knight, K., 1995. Model search and inference by bootstrap “Bumping.” Technical report, University of Toronto, Toronto.

Winkler, W. E., 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: Proceedings of the Section on Survey Research Methods, American Statistical Association 354–69.

Winkler, W. E., Thibaudeau, Y., 1991. An application of the fellegi-sunter model of record linkage to the 1990, U.S. decennial census.

Winkler, W.E,. 2006. Overview of Record Linkage and Current Research Directiond. Technical report RR2006/02, US Bureau of the census, Washington DC.

2.8 Appendix

Table 2.A.1: Dealing with unbalanced training sets in machine learning supervised methods; testing set performance under different levels of training data synthetic oversampling – S (number of extra cases from the minority class generated) and undersampling –U (number of extra cases from the majority classes selected for each case generated from the minority class).

Method	**Proportion**	**0**	**Performance measure**			
		1	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
Recursive partitioning trees	*Original (99.5%-0.5%)*	*293,551 1,461*	*99.9*	*86.7*	*97.3*	*99.9*
	S:2000/U:905 (90.0%-10.0%)	264,441 30,681	99.8	49.6	94.6	99.8
	S:4000/U:403 (80.0%-20.0%)	235,513 59,901	97.5	07.9	94.6	97.5
	S:6000/U:235 (70.0%-30.0%)	206,001 89,121	96.9	0.3	04.1	97.1
	S:8000/U:152 (60.0%-40.0%)	177,657 118,341	95.6	0.2	04.1	95.9
	S:10000/U:101 (50.0%-50.0%)	147,561 147,561	95.6	0.2	04.1	95.9
Bagging of decision trees	*Original (99.5%-0.5%)*	*293,551 1,461*	*99.9*	*83.9*	*98.6*	*99.9*
	S:2000/U:905 (90.0%-10.0%)	264,441 30,681	99.6	39.3	98.6	99.6
	S:4000/U:403 (80.0%-20.0%)	235,513 59,901	98.9	17.1	98.6	98.9
	S:6000/U:235 (70.0%-30.0%)	206,001 89,121	98.9	17.2	98.6	98.9
	S:8000/U:152 (60.0%-40.0%)	177,657 118,341	98.3	11.8	98.6	98.3
	S:10000/U:101 (50.0%-50.0%)	147,561 147,561	98.4	12.3	98.6	98.4
Bootstrap classification trees	*Original (99.5%-0.5%)*	*293,551 1,461*	*99.9*	*88.9*	*97.3*	*99.9*
	S:2000/U:905 (90.0%-10.0%)	264,441 30,681	99.8	50.0	94.6	99.8
	S:4000/U:403 (80.0%-20.0%)	235,513 59,901	97.6	08.3	94.6	97.6
	S:6000/U:235 (70.0%-30.0%)	206,001 89,121	95.6	0.2	04.1	95.9

	S:8000/U:152 (60.0%-40.0%)	177,657 118,341	95.6	0.2	04.1	95.9
	S:10000/U:101 (50.0%-50.0%)	147,561 147,561	95.6	0.2	04.1	95.9
Stochastic boosting	*Original (99.5%-0.5%)*	*293,551 1,461*	*99.9*	*85.7*	*97.3*	*99.9*
	S:2000/U:905 (90.0%-10.0%)	264,441 30,681	99.6	35.9	95.9	99.6
	S:4000/U:403 (80.0%-20.0%)	235,513 59,901	98.6	13.3	95.9	98.6
	S:6000/U:235 (70.0%-30.0%)	206,001 89,121	97.7	08.7	95.9	97.7
	S:8000/U:152 (60.0%-40.0%)	177,657 118,341	96.9	06.6	95.9	96.9
	S:10000/U:101 (50.0%-50.0%)	147,561 147,561	96.8	06.3	95.9	96.8
Support Vector Machines	*Original (99.5%-0.5%)*	*293,551 1,461*	*99.9*	*91.1*	*97.3*	*99.9*
	S:2000/U:905 (90.0%-10.0%)	264,441 30,681	99.8	53.3	97.3	99.8
	S:4000/U:403 (80.0%-20.0%)	235,513 59,901	99.4	27.3	97.3	99.4
	S:6000/U:235 (70.0%-30.0%)	206,001 89,121	98.9	17.1	97.3	98.9
	S:8000/U:152 (60.0%-40.0%)	177,657 118,341	98.2	10.7	95.9	98.2
	S:10000/U:101 (50.0%-50.0%)	147,561 147,561	96.6	06.2	97.3	96.6
Single-hidden-layer neural networks	Original (99.5%-0.5%)	293,551 1,461	99.8	-	0.0	100.0
	S:2000/U:905 (90.0%-10.0%)	264,441 30,681	99.7	40.6	78.4	99.7
	S:4000/U:403 (80.0%-20.0%)	235,513 59,901	99.1	21.6	97.3	99.2
	S:6000/U:235 (70.0%-30.0%)	206,001 89,121	97.5	08.1	97.3	97.5
	S:8000/U:152 (60.0%-40.0%)	177,657 118,341	97.8	07.7	81.1	97.8
	S:10000/U:101 (50.0%-50.0%)	147,561 147,561	95.8	05.1	98.6	95.9
Logistic regression	Original (99.5%-0.5%)	293,551 1,461	99.92	88.89	75.67	99.98
	S:2000/U:905 (90.0%-10.0%)	264,441 30,681	99.91	90.91	67.57	99.98
	S:4000/U:403 (80.0%-20.0%)	235,513 59,901	99.91	89.29	67.57	99.98
	S:6000/U:235 (70.0%-30.0%)	206,001 89,121	99.91	89.29	67.57	99.98
	S:8000/U:152 (60.0%-40.0%)	177,657 118,341	99.90	85.00	68.92	99.97
	S:10000/U:101 (50.0%-50.0%)	147,561 147,561	99.89	79.10	71.62	99.96

- Symbol is used for indeterminate calculation results.

Chapter 3: Forecasting beef production and quality using large scale integrated data from Brazil

3.1 Abstract

With agriculture rapidly becoming a data driven field it is imperative to extract useful information from large data collections to optimize the production systems. We compared the efficacy of regression (linear regression or generalized linear regression for continuous or categorical outcomes, respectively), random forest (RF) and multilayer neural networks (NN) to predict beef carcass weight (CW), age when finished (AS), fat deposition (FD), and carcass quality (CQ). The data analyzed contained information on over 4 million beef cattle from 5,204 farms, corresponding to 4.3% of Brazil's national production between 2014-2016. Explanatory variables were integrated from different data sources and encompassed animal traits, participation in a technical advising program, nutritional products sold to farms, economic variables related to beef production, month when finished, soil fertility, and climate in the location in which animals were raised. The training set was composed of information collected in 2014 and 2015, while the testing set had information recorded in 2016. After parameter tuning for each algorithm, models were used to predict the testing set. The best model to predict CW and AS was RF (CW: RMSEp = 0.65, R^2 = 0.61 and MAE = 0.49; AS: Accuracy = 28.7%, Kappa = 0.08). While the best approach for FD and CQ was generalized linear regression (Accuracy = 45.7%, Kappa = 0.05, and Accuracy = 58.7%, Kappa = 0.09, respectively). Across all models there was a tendency for better performance with RF and regression and worse with NN. Animal category, nutritional plan, cattle sales price, participation in a technical advising program and climate and soil in which animals were raised

were deemed important for prediction of meat production and quality with regression and RF. The development of strategies for prediction of livestock production using real-world large scale data will be core to projecting future trends and optimizing the allocation of resources at all levels of the production chain, rendering animal production more sustainable. Despite beef cattle production being a complex system, this analysis shows that by integrating different sources of data it is possible to forecast meat production and quality at the national level with moderate-high levels of accuracy.

Keywords: beef, forecasting, integrated, large scale data, machine learning, Brazil.

3.2 Introduction

In the 21st century, agriculture will have to scale up to feed a human population projected to increase 30% by 2050 (FAO, 2009). Nonetheless, it cannot happen at the expense of animal welfare, or an increase in the environmental footprint, or even at a higher cost, which would limit access for developing countries. Therefore, it is imperative to optimize the whole food chain to overcome such challenge. In this context, large data collections can be a valuable source of information to effectively address the current demand faced by agriculture (Kamilaris, et al., 2017; Liakos et al., 2018 and Morota et al., 2018) with data analytics being central to support decision making at the farm level (Morota et al., 2018; Pham and Stack, 2018). The emerging fields of artificial intelligence and machine learning are core to data analytics and present new tools for predicting (i.e. forecasting) outcomes such as yield and quality with large scale data. By projecting future trends, we can optimize allocation of resources rendering the whole production chain more efficient and sustainable.

Beef production represents an important sector of animal agriculture as the third most produced meat in the world, after pork and poultry (FAO, 2014). The largest commercial cattle population in the world (213.5 million head of cattle) is located in Brazil (Oliveira, 2019). The country is also the largest exporter in the world with almost 20% of global beef exports (2.1MM metric tons exported from the 9.9 MM produced) in 2018 (Zia et al., 2019). USDA predicts that Brazilian production will keep increasing in the near future, reaching 23% of the world's total exports by 2028 (Zia et al., 2019), remaining essential in the international agriculture market. Beef cattle production is developed in all Brazilian ecosystems (Oliveira, 2018), with the major producing states being Mato Grosso, Goiás, Minas Gerais, Mato Grosso do Sul, and Pará, which together are responsible for 54.2% of the national production (Oliveira, 2019). Brazil has a mature beef cattle industry based on grass-fed cattle (Millen et al., 2011), and cattle spend most of their lives grazing in pasture. The diversity in environments and conditions in which animals are raised combined with complex animal physiological mechanisms makes the prediction of meat production and quality at a national level a challenging task. In addition, historically the lack of consistent data collection and availability has made forecasting meat production and quality at a national level in Brazil a virtually imposible task.

This paper aimed to forecast beef cattle production and quality, using a large scale data set integrated from different sectors of industry in Brazil. We compared the efficacy of traditional methods (linear regression or generalized linear regression) and machine learning approaches (random forests – RF, and artificial neural networks – NN) to forecast beef cattle production traits (carcass weight – CW, age when finished – AS, fat deposition – FD, and carcass quality – CQ). Predictor variables included animal traits, farm participation in a technical advising program, nutritional products utilized by the farms, economic variables related to beef production, month

when finished, and soil fertility and climate classification in the location in which animals were raised. The data analyzed contain information on over 4 million animals, corresponding to 4.3% of the Brazilian national beef production between 2014-2016.

3.3 Material and Methods

Data acquisition and integration

The data set utilized in this analysis was integrated from different sources to provide a comprehensive view of the Brazilian beef cattle production context. The animal information utilized was pre-collected by the sources involved, such that procedures involving the use of animals in this study did not have to be approved. The data contained information on 828,292 observations (group of animals) from 5,204 farms comprising a total of 4,022,394 finished beef cattle between 2014 and 2016. It is worth noting that the data analyzed in this study corresponds to 4.3% of the Brazilian cattle national production of 94.2 million head of cattle for the 2014 to 2016 period (IBGE, 2018).

The data set contained information on 645 municipalities located in 12 of the 26 Brazilian states (Acre, Bahia, Goiás, Mato Grosso, Mato Grosso do Sul, Maranhão, Minas Gerais, Pará, Paraná, Rondônia, São Paulo and Tocantins). The distribution of farms and number of finished animals per state is presented in Figure 3.1.

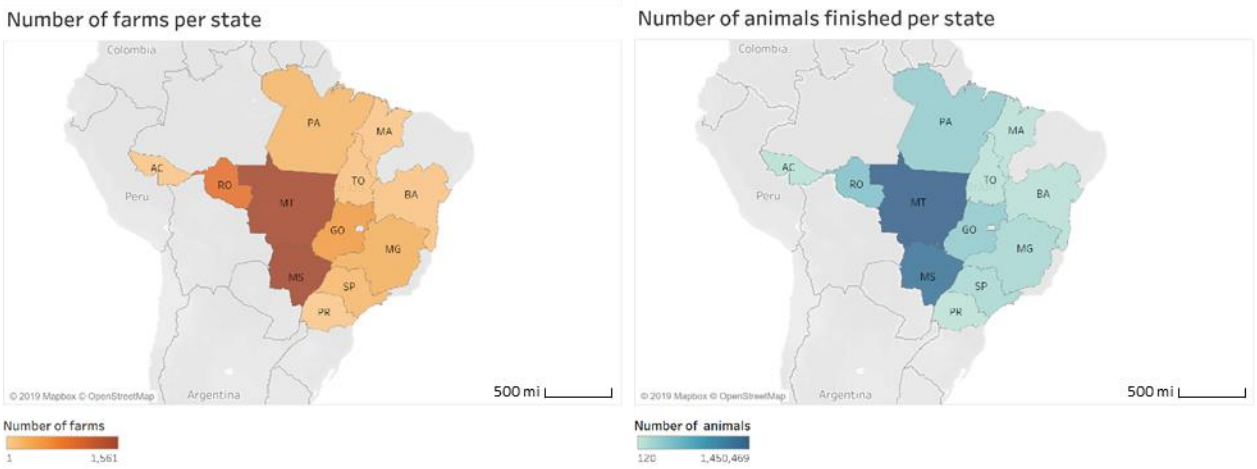

Figure 3.1: Distribution of farms (in the left) and finished animals (in the right) in the data set per state in Brazil.

The integrated data collection included five major sources of data encompassing animal traits, utilization of technology and nutritional products at the farm, economic variables related to beef production, soil fertility, and climate where animals were raised. Each data source and the data integration steps are described in detail below.

The animal and nutrition/technology data sets used for this study were kindly provided by a meat packing company (*JBS S.A., Brazil*) and an animal nutrition company (*DSM Nutritional products Brazil S.A.*), respectively. A farm matching integration procedure, described in detail by Aiken et al. (2019), was implemented to identify which farms in both databases were the same and to connect the information. In this study, results provided by the two best approaches for farm matching highlighted by Aiken et al. (2019), i.e. bagged clustering and support vector machines, were overlapped. When the two algorithms disagreed on a matching status, discrepancies were solved by expert clerical review to generate this data set containing 5,204 matched farms.

The animal traits obtained from the meat packing plant were: carcass weight (CW) – measured in kg, age when finished (AS) – obtained by carcass dental evaluation and divided in five categories (up to 20 mo; 20 to 24 mo; 24 to 36 mo; 36 to 48 mo; and above 48 mo old), fat deposition (FD) – obtained by visual evaluation of carcass fat coverage, and divided in five categories (absent – lower than 1 mm; low – 1 to 3mm; medium – 3 to 6mm; high – 6 to 10mm; and excessive – above 10mm), carcass quality (CQ) – which takes into account CW, AS, FD, gender of the animal, as well as body condition score, and is divided in three major categories (undesirable, acceptable, desirable), and animal category – defined as female, steer and bull. A distribution of the animal categorical traits is presented in Figure 3.2. For the continuous trait CW, the average was 252kg with a standard deviation of 61.6kg.

The information obtained from the nutritional company comprised two major parts. The first one contained information on farm participation in a technical advising program for improving results, as a binary variable. The second part had information on the amount of nutritional products utilized by farms where animals were raised in the year of slaughter. The amount of product used by farms was divided in three major categories: mineral premix for non-feedlot cattle – PNF, feedlot mineral premix – FP, and feedlot premix with additives – FA (all measured in kilograms). More specifically, PNF contained mainly minerals, while FP included feedlot concentrate and FA had concentrate with the additives of the following classes: essential oils, enzymes, ionophores, buffers, probiotics and/or yeast. For the three nutritional products previously mentioned, the total amount used (kilograms) was adjusted by the quantity of animals finished per farm in that year, generating a per animal value. Regarding the technical advising program, 921 farms participated in it, 3,835 did not, and the remaining farms had missing information for this variable. The average

FP per animal per year was 1.4kg (SD = 13.8kg), while PNF was 61.8kg (SD = 162.4kg), and FA was 3.1kg (SD = 28.2kg).

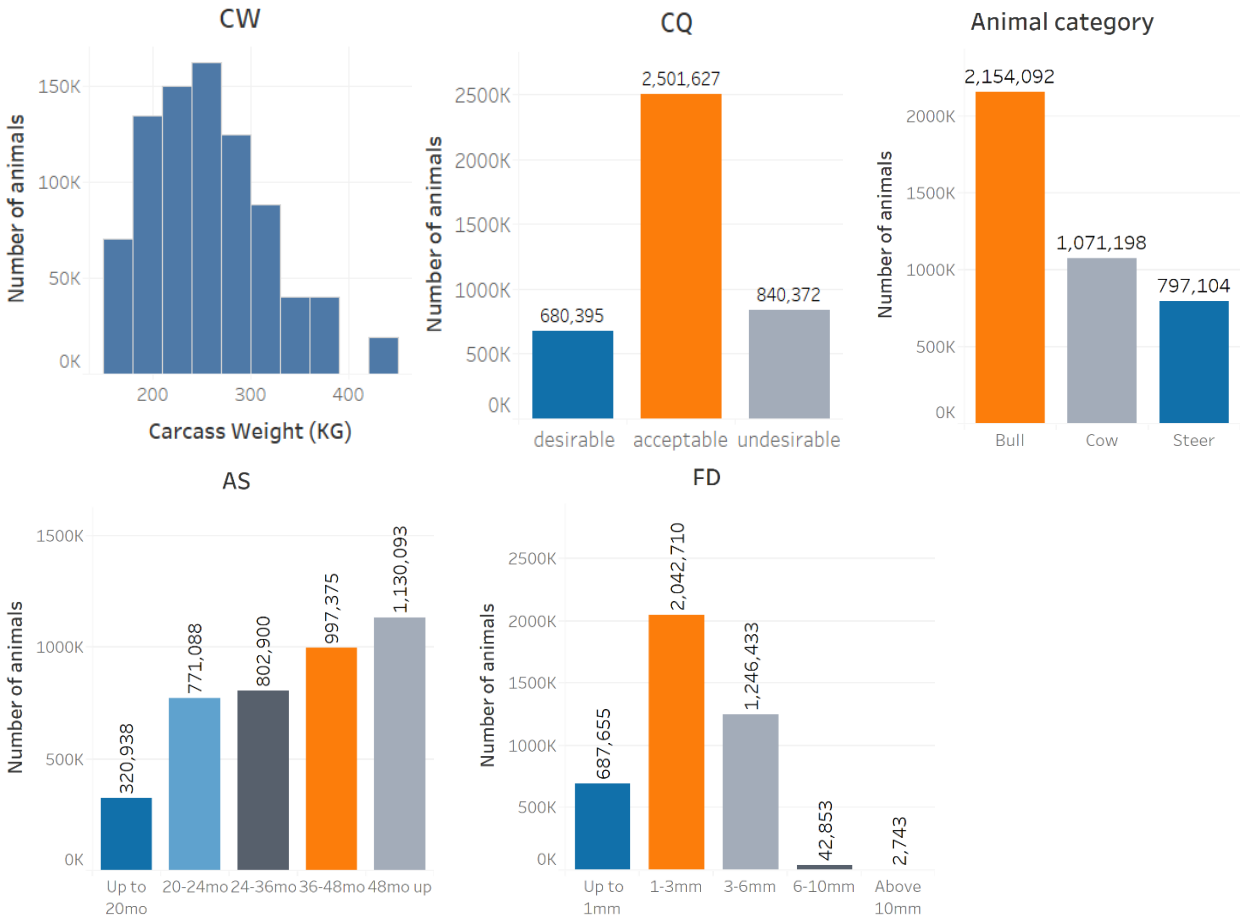

Figure 3.2. Distribution of animals finished according to carcass weight (CW), carcass quality (CQ), animal category, age at slaughter (AS) and fat deposition (FD).

The information on economic variables related to beef production was extracted from the Agrolink public database (Agrolink, 2019). Two variables were included in this analysis: the finished cattle sales price at the state the farm was located, for the month and year each animal was

harvested, and the price for the corn at the state the farm was located three months before the harvesting date. For example, if an animal was harvested in December, the sales price of corn in September for the same state was utilized. All prices were in the Brazilian currency (R$, Reais). The defined time window of three months approximates the average time in Brazil (83 days) that beef animals are finished on feedlots before slaughter (Millen et al., 2011). The average sales price and corn price per state per month is shown in Figure 3.3.

The soil fertility classification at the farm in which animals were raised was accessed utilizing the interactive geographic mapping platform, available from the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística - IBGE, 2019). The national digital atlas of Brazil for agricultural potential of soils in terms of fertility and characteristics was overlaid to the geopositioning (latitude and longitude) of the farm to determine the soil type where animals were raised. From the ten soil classifications defined at the atlas, nine were present in the data set: light green, cream, orange, yellow, purple, dark green, pink, light blue and gray. A full description of the classification for those soils is presented in Table 3.1. The number of observations per soil type was as follows: light green = 59,686, cream = 27,206, orange = 527,431, yellow = 370, purple = 10,269, dark green = 9,795, pink = 13,154, light blue = 12,790 and gray = 167,591.

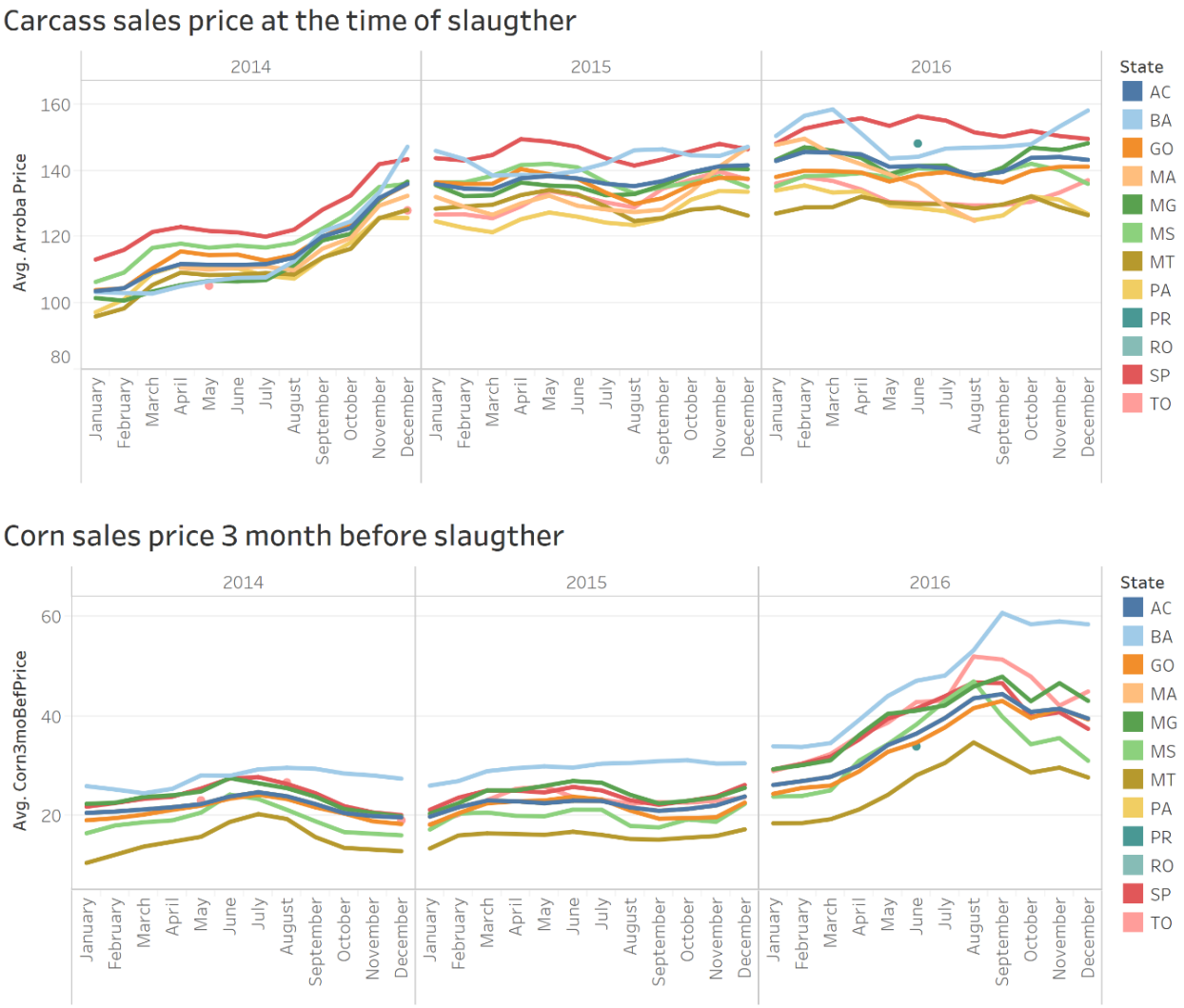

Figure 3.3. Average cattle sales price per state (top) in Brazilian currency (R$, Reais), and corn sales price per state per month (bottom). Points represent states that contained information on single months. Source: adapted from Agrolink.

Lastly, we considered the climate at the municipality where cattle were raised. The Köppen's climate classification for Brazil (Álvares et al., 2014) was chosen as it is considered the most widely used classification method across geographical and climatologic societies in the world. This classification utilized historical information on monthly temperature and rainfall to

produce a climate map with high spatial resolution that allows detection of climatic variations at the landscape level. From the 12 climate classifications identified in Brazil (Álvares et al., 2014), nine were present at the locations in this data set: Af – tropical zone, without dry season; Am – tropical zone, monsoon; Aw – tropical zone, with dry winter; As – tropical zone, with dry summer; BSh – dry zone, semi-arid, low latitude and longitude; Cfa – Humid subtropical zone, oceanic climate without dry season, with hot summer; Cfb – Humid subtropical zone, oceanic climate without dry season, with temperate summer; Cwa - Humid subtropical zone, with dry winter, and hot summer; Cwb - Humid subtropical zone, with dry winter, and temperate summer. The number of observations for each climate was: Af – 29,551; Am – 392,180; As – 659; Aw – 362,603; BSh – 417; Cfa 35,349; Cfb – 253; Cwa – 5,899 and Cwb – 1,381.

Table 3.1. Classification of soil agricultural potential of Directory of Geosciences, Coordination of natural resources and environmental studies (IBGE, 2019) for the nine soil types in this data.

Soil class	Fertility	Attributes	Relief	Major limitations
Light green	High	Good	Flat and slightly undulating	No major limitations
Cream	Mean	Good	Flat and slightly undulating	Medium to low availability of nutrients
Orange	Low	Good	Flat and slightly undulating	Low availability of nutrients, excess of aluminum
Yellow	Low	Regular	Flat and slightly undulating	Low availability of nutrients
Purple	Mean-High	Regular	Flat to undulating	Steep slopes, shallow depth, rough texture
Dark green	Mean-High	Good	Highly undulating	Steep slopes
Pink	Low	Regular	Undulating to mountainous	Steep slopes, restricted drainage, aluminum excess
Light blue	Low	Regular	Flat and slightly undulating	Sodium excess, restricted drainage, flooding risk
Gray	Not recommended to the agricultural activity			

Data Analysis

After comprehensive data integration of the previously mentioned sources, the data was pre-processed for the prediction of CW, AS, FD and CQ. From all variables considered in the models, only the binary variable for participation in the technical advising program had missing information. More specifically, 27% of the farms had missing information (in at least one year). This variable was imputed using bagged trees with the R package "missForest" (Stekhoven, 2013). Bagged trees were created using all other variables in the training set, such that when a sample had a missing value for a predictor, the bagged model was used to predict this value. The estimated error of the imputation (out of bag proportion of falsely classified samples) was very low (0.0074%). The data set presented no major issues with near-zero variance and there was no high correlation amongst predictors (all correlations below 0.5). All continuous variables were centered and scaled (mean = 0; SD = 1) prior to analysis.

Three algorithms were contrasted for forecasting CW, AS, FD and CQ. Those algorithms were: 1) linear regression (LR) for the continuous trait CW or generalized linear regression (GLR) for ordered categorical traits AS, FD and CQ, 2) random forests (RF), and 3) multilayer perceptron neural networks (NN). The choice of algorithms to analyze the data covers different types of algorithms from methods more traditionally used in animal sciences, such as regression, along with modern machine learning methods, which have been successfully used for the task of forecasting in other fields (Biau and Scornet, 2016; Kuhn and Johnson, 2016). Also, the specific machine learning algorithms were chosen to explore strengths of the methods for the prediction task. For example, RF is known to be robust to noise in the predictor variables (Biau and Scornet, 2016; Kuhn and Johnson, 2016), which is a likely occurrence with data collected in non-experimental settings such as the farm data analyzed here. On the other hand, NN is known for its

ability to properly model complex non-linear relationships (Kuhn and Johnson, 2016). The data set utilized in this study aims to use complex relationships between environmental and physiological variables for the task of prediction of meat production and quality. For this reason, NN could be a good algorithm to model such complex relationships. All methods were implemented in the R environment using the "caret" package (Kuhn, 2019). All analyses were performed utilizing the capabilities of the Center for High Throughput Computing (CHTC) at the University of Wisconsin-Madison.

The explanatory variables used to predict each outcome (CW, AS, FD and CQ) with regression, RF and NN are detailed in Table 3.2. The training sets had 542,935 observations obtained from 2014 and 2015, while the remaining 285,357 observations from 2016 were used as an independent testing set. A 10-fold cross-validation scheme within the training set was implemented in which for each cross validation run, the model was trained in 9-folds and the tenth fold was used to validate tuning parameters (for models that required parameter tuning). The model that produced the best results across the 10-fold cross validation was selected and further utilized in the testing set. For GLR, different link functions were tested (logistic, probit, cloglog, loglog and cauchit) and the one that provided the highest accuracy was chosen. For RF, the only parameter requiring tuning was the number of explanatory variables included in the model at a time (mtry), which was done using exhaustive search (testing 1 to all available explanatory variables) in each model. For NN, three parameters were tuned with a grid search: the number of hidden layers (1 to 3), the number units per hidden layer (1, 5, 10, 50 or 100), and the rate of weigth decay utilized in the training backpropagation procedure (0, 0.0001 or 0.1). A grid search was chosen as a reasonable tuning method due to major constrains of run time and memory related to large size of the data set analyzed and complexity of analyses. The activation function utilized at each hidden

layer was the logistic (i.e. sigmoidal) function. The maximum number of iterations allowed to train the model was 100 (with early stopping criteria), and one hot encoding was applied to all categorical explanatory variables.

Table 3.2: Models for forecasting carcass weight (CW); age when finished (AS); fat deposition (FD) and carcass quality (CQ). Explanatory variables to models included: animal category, participation in a technical advising program (PTAP); kg of premix for non-feedlot per beef animal (PNF); kg of feedlot premix per beef animal (FP); kg of feedlot premix with additive products per beef animal (FA); finished cattle sales price (FCSP); corn price 3mo before finished (CP3B); soil fertility classification (SOIL); climate classification (CLIM); and month when finished (MO).

Outcome	**Predictors**
CW	AS; animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO
AS	animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO
FD	AS; animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO
CQ	animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO

The predictive ability of each model (regression, RF and NN) was calculated for each outcome variable. For continuous traits (CW), the predictive ability was assessed in terms of predicted Root Mean Square Error (RMSEp) of the testing set, coefficient of determination (R^2), and Mean Absolute Error (MAE). For categorical traits (AS, FD and CQ), the predictive ability was evaluated in terms of accuracy and the Cohen's kappa coefficient (Kappa). For all outcomes the respective predictive ability metric is presented with the standard deviation of the resamples.

Amongst all methodologies tested, two produce simple and intuitive metric for variable importance: regression and RF. For regression, variable importance was assessed using the absolute value of the t-statistic for each explanatory variable used in the best training set. For RF, variable importance was determined by recording the prediction accuracy of the out-of-bag sample when each tree was formed. This was repeated after permuting each of the explanatory variables. The difference between the two accuracies was then averaged across all trees and normalized by the standard error. All measures of importance were scaled to have a maximum value of 100.

3.4 Results

Results for the choice of GLR link function for the categorical variables AS, FD and CQ are presented in Table 3.3. More specifically, 10-fold cross validation results (out of bag accuracy average and SD) for the training set using different link functions are presented. For each categorical variable, the link function that provided the highest accuracy was chosen and later fitted to the independent test set. The link function that provided the highest accuracy for AS (30.6%) was cloglog, for FD was loglog (45.0%), and for CQ was the cauchit (57.7%).

Results for the parameter tuning of RF using exhaustive search (with number of variables included in the model at a time from 1 to all explanatory variables) are presented in Figure 3.4. The best results for the 10-fold cross validation performed in the training set, in terms of maximum accuracy for categorical variables and minimum RMSEp for continuous variables were chosen. For the variables CW, AS, FD, CQ, the best results were with number of variables included in the model equal to 11, 10, 5 and 9, respectively.

Table 3.3: Accuracy results from10-fold cross-validation for the training set of generalized linear regression. Results are presented as the average accuracy (converted to original scale) across the 10 out of bag folds, followed by the ±SD (in parenthesis) for the three categorical variables: age when finished (AS), carcass fat deposition (FD) and carcass quality (CQ). The highest accuracy across different link functions is highlighted for each trait

	Generalized linear model link function				
Variable	Logistic	Probit	Cloglog	Loglog	Cauchit
AS	0.2924 (±0.0019)	0.2906 (±0.0017)	**0.3056 (±0.0011)**	0.2704 (±0.0011)	0.2991 (±0.0016)
FD	0.4477 (±0.0013)	0.4479 (±0.0013)	0.4301 (±0.0017)	**0.4500 (±0.0022)**	0.4478 (±0.0017)
CQ	0.5768 (±0.0017)	0.5761 (±0.0019)	0.5617 (±0.0012)	0.5735 (±0.0018)	**0.5772 (±0.0037)**

NN results for parameter tuning of number of layers (1 to 3), number of nodes per layer (1, 5, 10, 50 and 100) and rate of decay (0, 0.0001 and 0.1) for the four explanatory variables CW, AS, FD, CQ are presented in Figures 5 – 8. Best results for CW (Figure 3.5) were: layer = 3, nodes per layer = 100 in the first, 100 in the second and 100 in the third, and decay = 0. For AS (Figure 3.6) were: layer = 3, nodes per layer = 100 in the first, 100 in the second and 100 in the third, and decay = 0. For FD (Figure 3.7) were: layer = 3, nodes per layer = 50 in the first, 50 in the second and 50 in the third, and decay = 0. Lastly for CQ (Figure 3.8) were: layer = 3, nodes per layer = 100 in the first, 100 in the second and 100 in the third, and decay = 0.

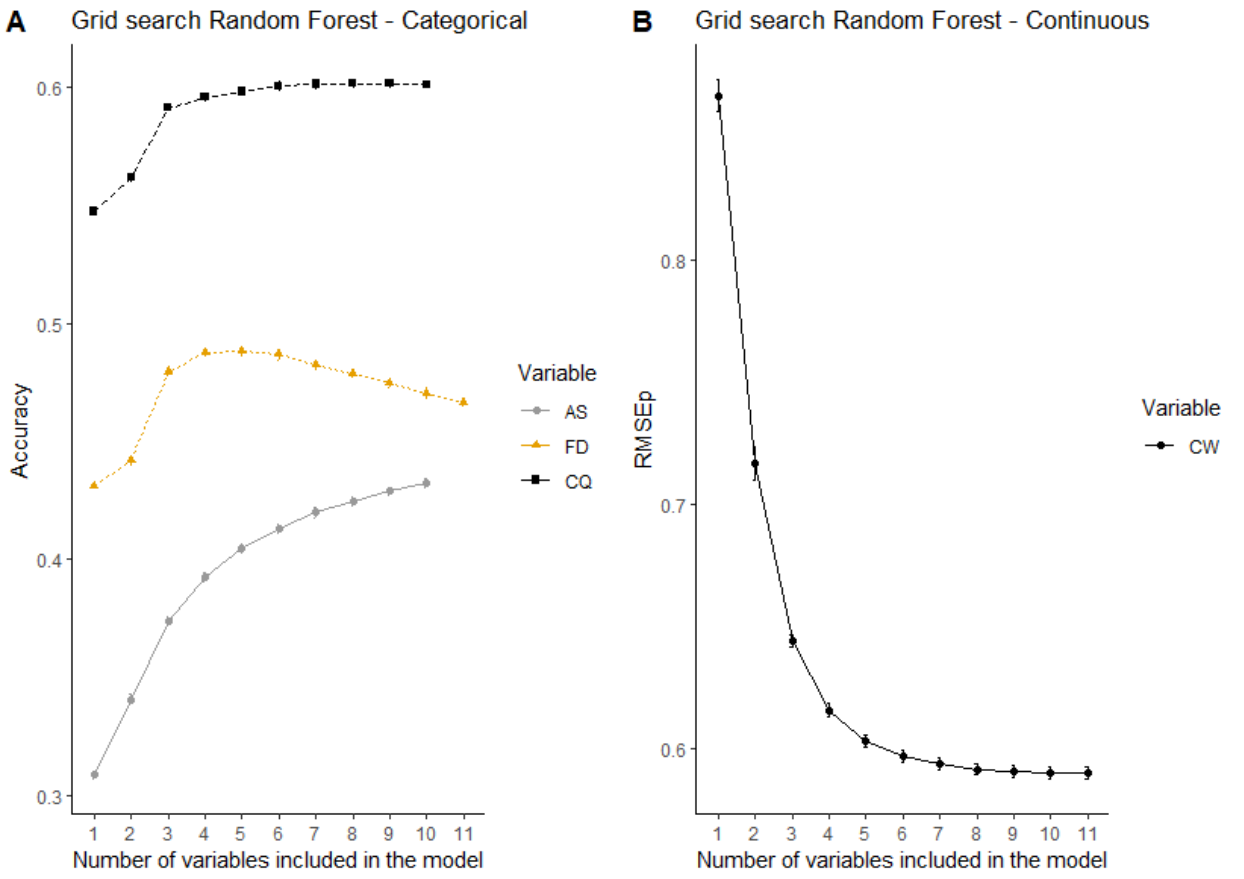

Figure 3.4. Results for exhaustive grid search, performed with 10-fold cross validation to test for different numbers of explanatory variables included in the Random Forest model at a time. Mean predictive accuracy and SD (vertical line for each point) across the 10 folds are presented for the categorical variables: age at when finished (AS), carcass fat deposition (FD) and carcass quality (CQ) is presented in panel A. Mean predictive Root Mean Square Error (RMSEp) and SD (horizontal line for each point) across the 10 folds are presented for the continuous variable carcass weight (CW) in panel B.

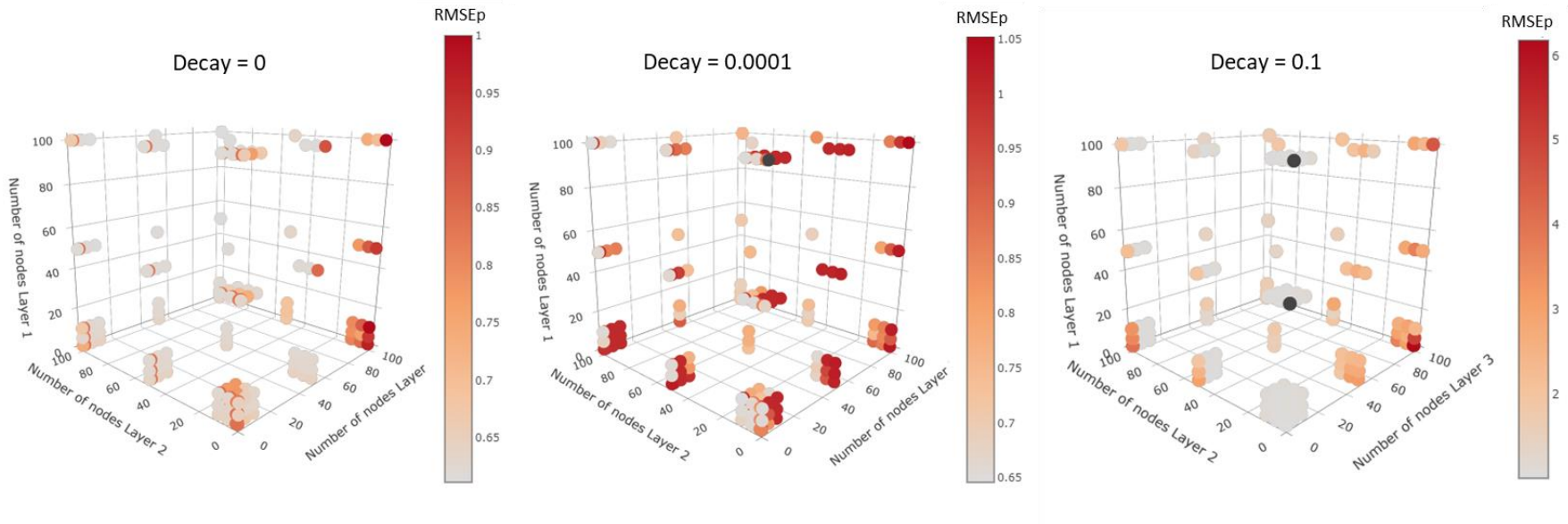

Figure 3.5. Results for grid search parameter tuning on the outcome variable carcass weight (CW) performed with 10-fold cross validation on the training set. Each node represents the mean predictive Root Mean Square Error (RMSEp) across the 10 folds for all possible combinations of different number of layers (1 to 3), nodes per layer (1, 5, 10, 50, 100) and rate of decay (0, 0.0001 and 0.1).

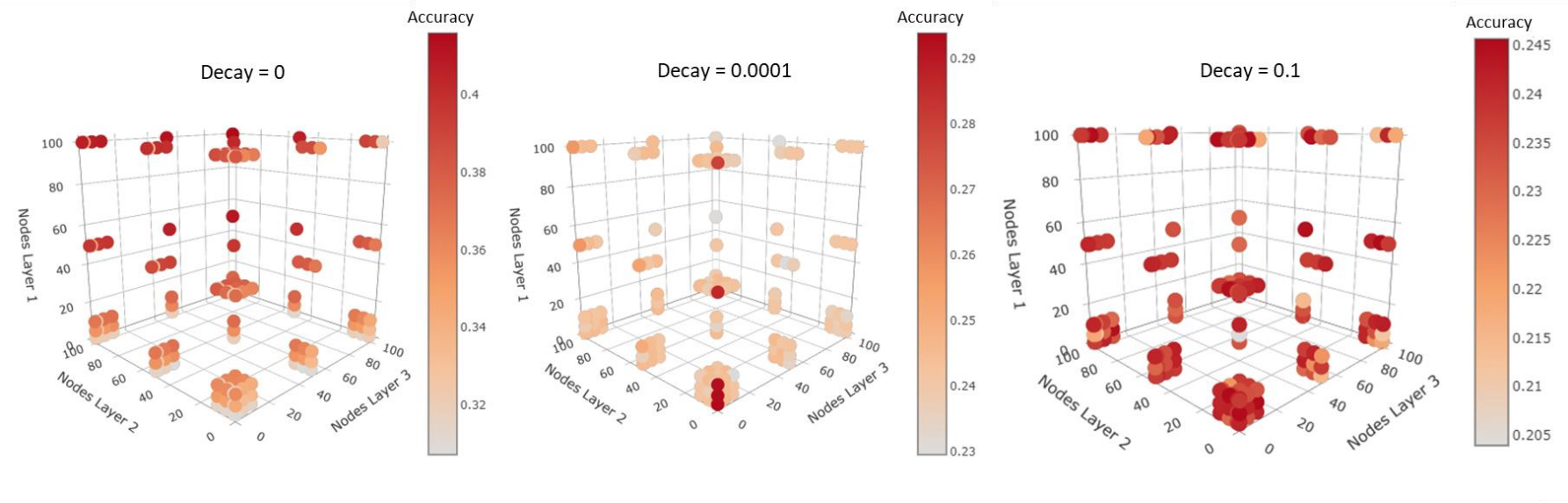

Figure 3.6. Results for grid search parameter tuning on the outcome variable age at slaughter (AS) performed with 10-fold cross validation on the training set. Each node represents the mean predictive accuracy across the 10 folds for all possible combinations of different number of layers (1 to 3), nodes per layer (1, 5, 10, 50, 100) and rate of decay (0, 0.0001 and 0.1).

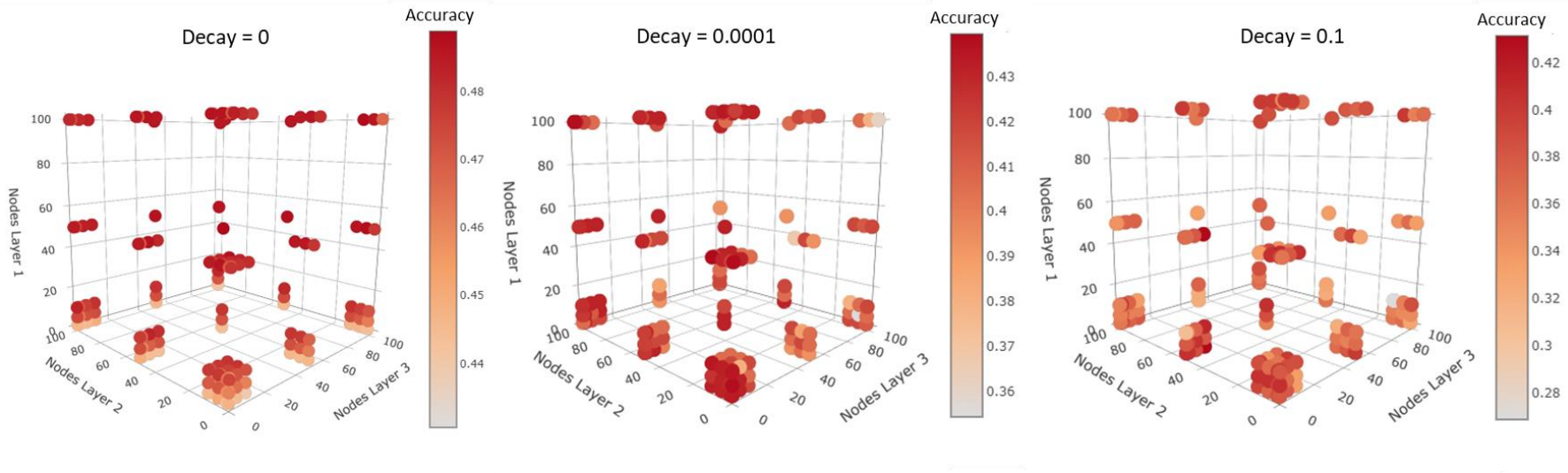

Figure 3.7. Results for grid search parameter tuning on the outcome variable fat deposition (FD) performed with 10-fold cross validation on the training set. Each node represents the mean predictive accuracy across the 10 folds for all possible combinations of different number of layers (1 to 3), nodes per layer (1, 5, 10, 50, 100) and rate of decay (0, 0.0001 and 0.1).

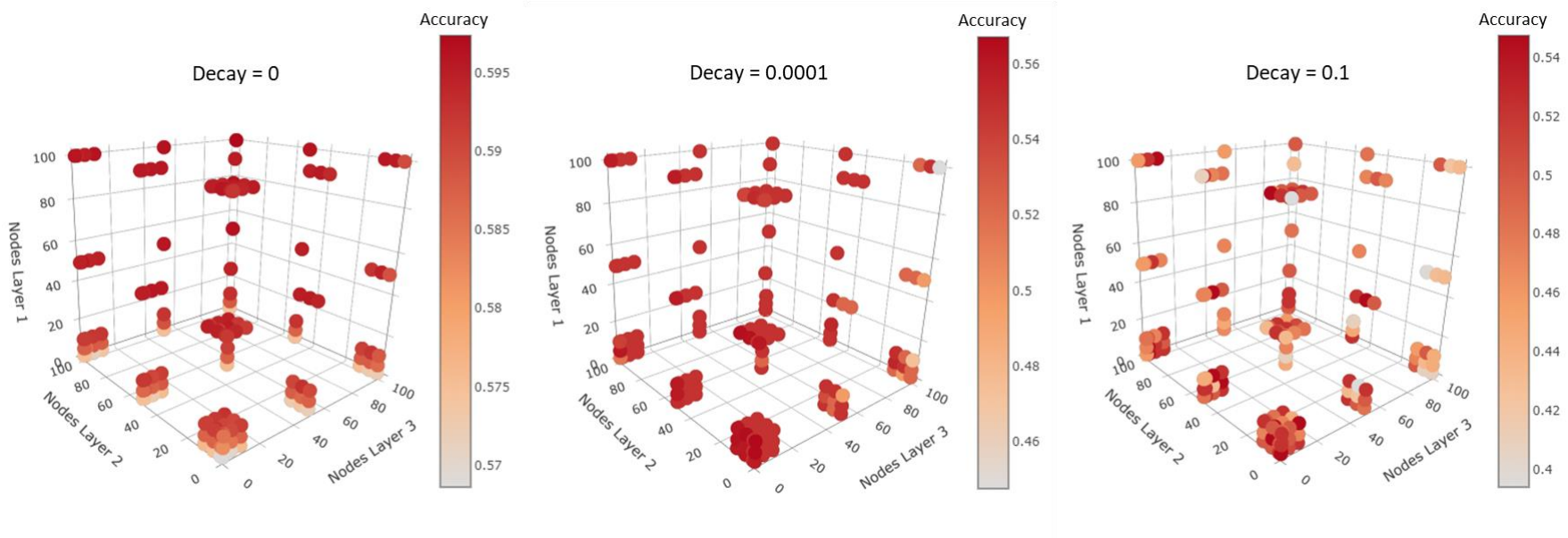

Figure 3.8. Results for grid search parameter tuning on the outcome variable carcass quality (CQ) performed with 10-fold cross validation on the training set. Each node represents the mean predictive accuracy across the 10 folds for all possible combinations of different number of layers (1 to 3), nodes per layer (1, 5, 10, 50, 100) and rate of decay (0, 0.0001 and 0.1).

After parameter tuning of all models using 10-fold cross validation in the training set, the best results in terms of maximum accuracy for categorical variables and minimum RMSEp for continuous variables were chosen to be fitted to the independent test set for each outcome variable. Results for testing set predictive ability in terms RMSEp, R^2 and MAE, for continuous variables; and accuracy and Kappa for categorical variables are presented in Table 3.4. The testing set predictive ability measure and respective standard deviation (estimated with 10-fold cross-validation procedure performed in the training set) are presented for CW, AS, FD and CQ.

The best model for CW was RF (RMSEp = 0.66, R^2 = 0.59 and MAE = 0.50), with results very similar to regression (RMSEp = 0.67, R^2 = 0.60 and MAE = 0.51). The best model for AS was also RF (Accuracy = 28.7%, Kappa = 0.08) with very similar performance presented by regression (Accuracy = 28.7%, Kappa = 0.07) and slightly lower performance presented by NN (Accuracy = 25.3%, Kappa = 0.02). The same pattern was observed for FD (RF: Accuracy = 45.7%, Kappa = 0.05 and regression: Accuracy = 44.9%, Kappa = 0.05), however the lower performance of NN (Accuracy = 37.4%, Kappa = 0.05) was more pronounced for this variable. Regarding CQ, the best predictions were obtained using regression (Accuracy = 58.7%, Kappa = 0.09), followed by RF (Accuracy = 53.9%, Kappa = 0.09) and NN (Accuracy = 46.4%, Kappa = 0.07), with a considerable drop in performance of the two models, compared to regression.

Table 3.4: Models predictive ability for carcass weight (CW), age when finished (AS), fat deposition (FD) and quality (CQ). For continuous traits (CW), testing set predictive ability was measured in terms of predicted Root Mean Square Error (RMSEp), coefficient of determination (R^2), and Mean Absolute error (MAE). For categorical traits (AS, FD and CQ), it was assessed in terms of Accuracy and the Cohen's kappa coefficient (Kappa). The testing set predictive ability is presented along with ±SD (in parenthesis) obtained in the training set 10-fold cross validation.

		Outcome variable		
		Categorical		
Model	**Measure**	**AS**	**FD**	**CQ**
Generalized linear regression	Accuracy	0.2867 (±0.0011)	**0.4576 (±0.0022)**	**0.5867 (±0.0019)**
	Kappa	0.0666 (±0.0015)	**0.0476 (±0.0037)**	**0.0862 (±0.0037)**
Random Forest	Accuracy	**0.2871 (±0.0019)**	0.4494 (±0.0020)	0.5390 (±0.0016)
	Kappa	**0.0759 (±0.0026)**	0.0523 (±0.0032)	0.0930 (±0.0032)
Multilayer perceptron neural networks	Accuracy	0.2536 (±0.0028)	0.3742 (±0.0019)	0.4640 (±0.1999)
	Kappa	0.0237 (±0.0034)	0.0501 (±0.0160)	0.0670 (±0.0017)
		Continuous		
		CW (centered and scaled)	**CW (original scale)**	
Linear regression	RMSEp	0.6765 (±0.0027)	41.2697 kg	
	R^2	0.6017 (±0.0017)	0.6017	
	MAE	0.5097 (±0.0017)	31.0941 kg	
Random Forest	RMSEp	**0.6626 (±0.0025)**	**40.4217 kg**	
	R^2	**0.5920 (±0.0024)**	**0.5920**	
	MAE	**0.5018 (±0.0013)**	**30.6122 kg**	
Multilayer perceptron neural networks	RMSEp	0.8073 (±0.0030)	49.2491 kg	
	R^2	0.4657 (±0.0037)	0.4657	
	MAE	0.5905 (±0.0045)	36.0233 kg	

Variable importance results (for regression and RF) for CW, AS, FD and CQ are presented in Figures 9 – 12, respectively. For CW (Figure 3.9), sales price for cattle and corn, as well as

technical consulting were important for both RF and regression. However, regression assigned heavier weights to animal category while RF deemed the nutrition given to the animal, month, climate and soil as important variables. For AS (Figure 3.10), variable importance was consistent between regression and RF. The most important predictors were animal category, animal nutrition, cattle sales price, and use of technical consulting. Corn price, climate, and soil at the location animals were raised had smaller importance. For FD (Figure 3.11), results were somewhat consistent between regression and RF models. Animal category was the most important predictor, followed by cattle sales price and use of technical consulting. However, when compared to regression, RF assigned higher importance to corn price, use of additives in the nutrition, climate, and soil in which the animal was raised. Unlike RF, the regression model assigned higher importance to AS as a predictor of FD. Lastly, for CQ (Figure 3.12) results were mostly consistent between regression and RF. The most important predictors across models were: animal category, sales price, and use of technology. However, RF assigned higher importance to the nutrition, corn sales price, soil, and climate in which animals were raised, as well as the month of the year.

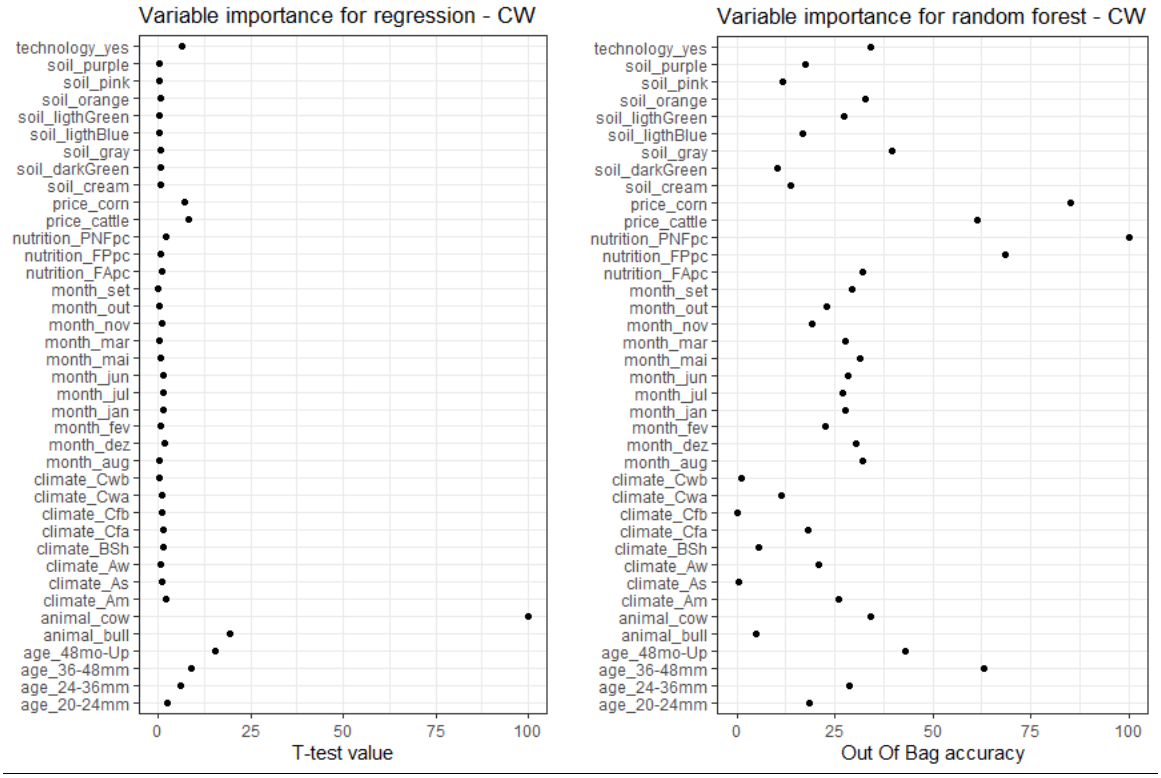

Figure 3.9. Variable importance results for the prediction of carcass weight (CW) with regression and random forests (RF). For regression, variable importance was assessed using the T-test value of the regression fitted to the test set while for RF it was estimated as the out of bag accuracy of permuting each explanatory variable.

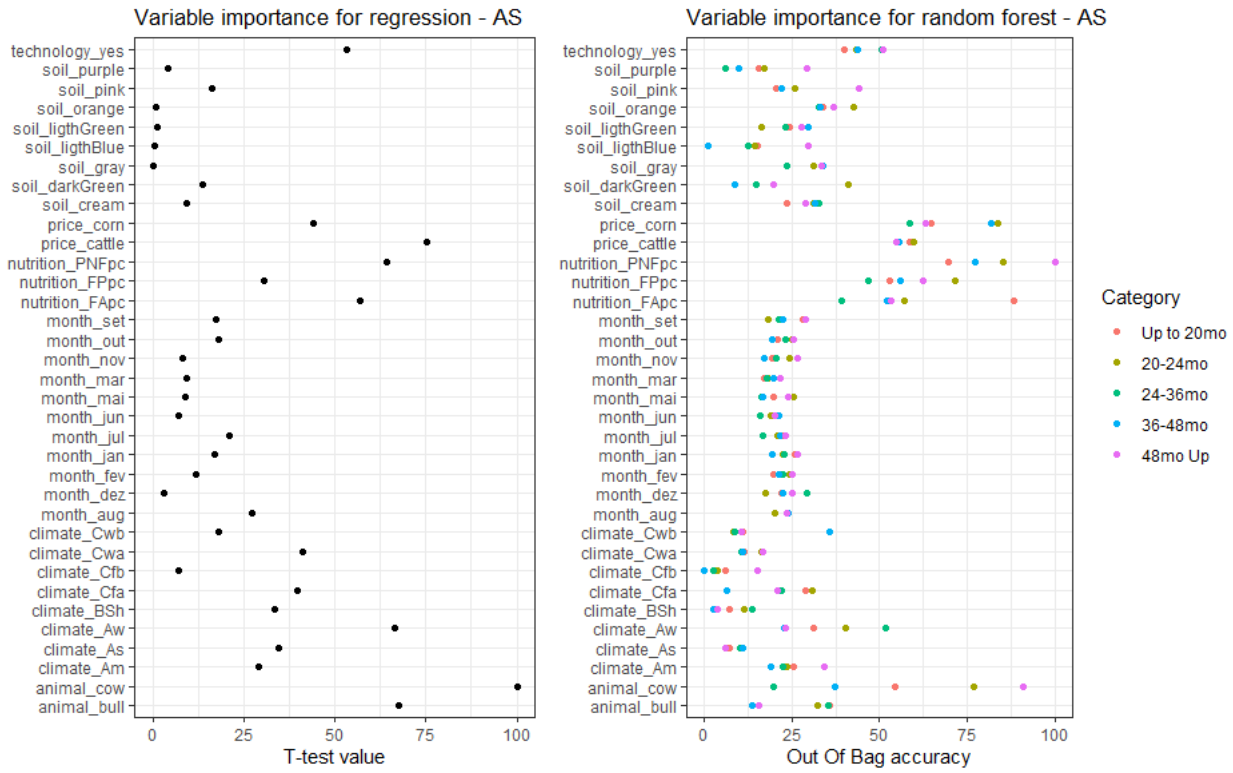

Figure 3.10. Variable importance results for the prediction of age at slaughter (AS) with regression and random forests (RF). For regression, variable importance was assessed using the T-test value of the regression fitted to the test set while for RF it was estimated as the out of bag accuracy of permuting each explanatory variable.

The run-time and disk space required by all models varied greatly. Regression models were considerably less demanding than the other methods (i.e. RF and NN) with 6 computing hours on 4 CPUs, requiring a total of 40 GB of memory for all outcome variables. A total of 2,370.5 computing hours on 109 CPUs, and 8 TB total memory were needed for the RF analysis and 15,482.02 computing hours on 5,580 CPUs, requiring 223.2 TB total memory were needed for the NN analysis.

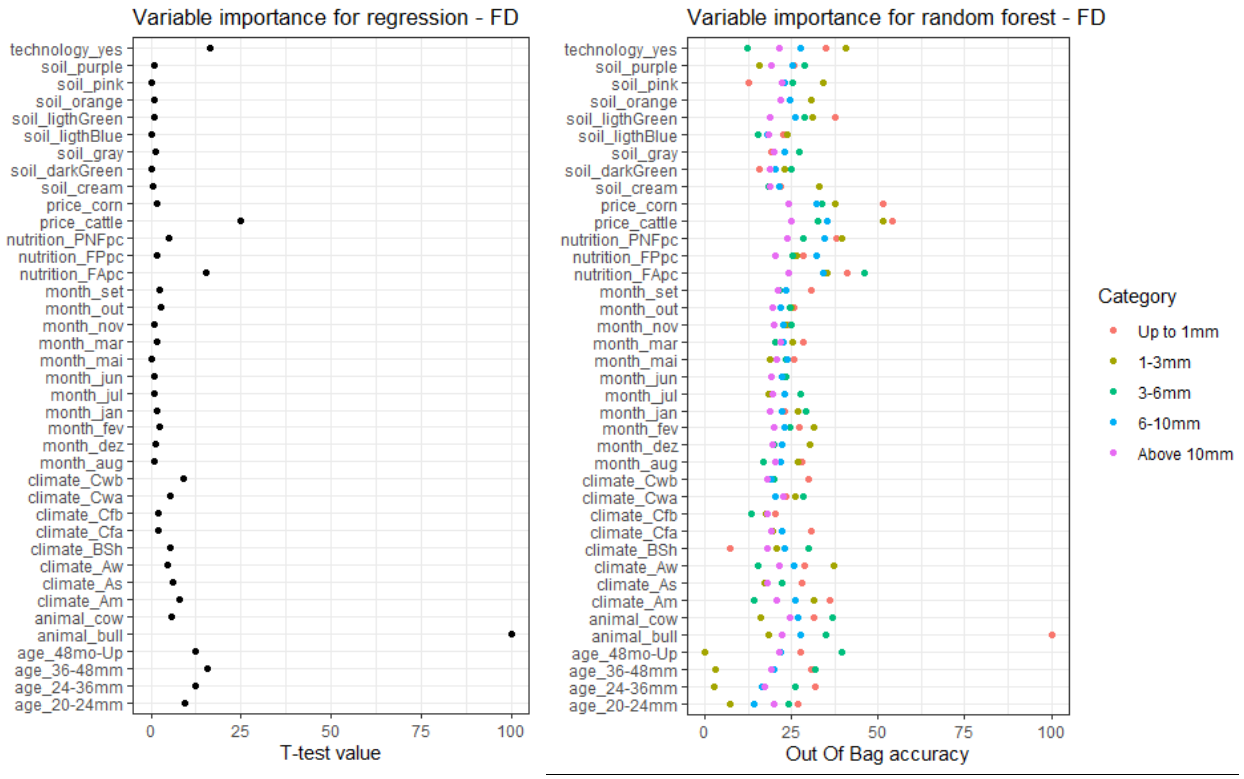

Figure 3.11. Variable importance results for the prediction of fat deposition (FD) with regression and random forests (RF). For regression, variable importance was assessed using the T-test value of the regression fitted to the test set while for RF it was estimated as the out of bag accuracy of permuting each explanatory variable.

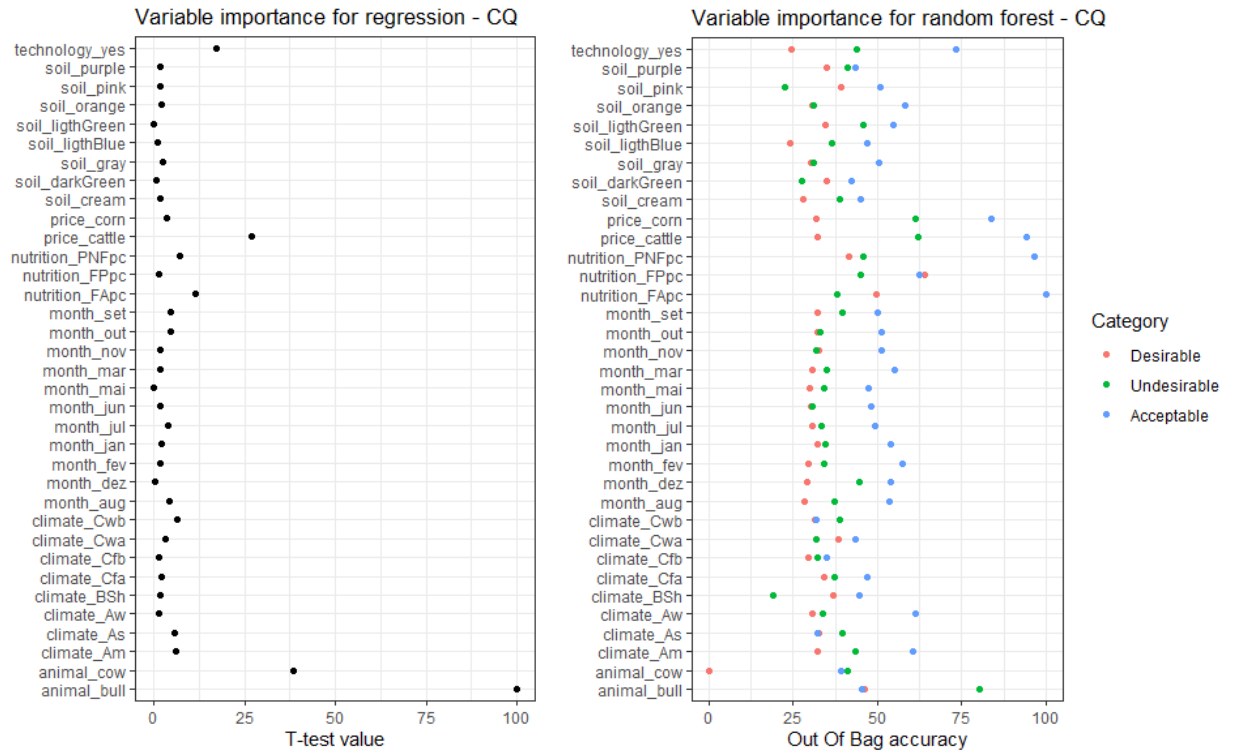

Figure 3.12. Variable importance results for the prediction of carcass quality (CQ) with regression and random forests (RF). For regression, variable importance was assessed using the T-test value of the regression fitted to the test set while for RF it was estimated as the out of bag accuracy of permuting each explanatory variable.

3.5 Discussion

When applying statistical and machine learning approaches, a very important step is to choose optimal model hyperparameters. For the GLR, when predicting the class of multi-class ordered variables, the model could be suboptimal if the chosen link function is not appropriate. The selection of a link function usually depends on knowledge of the response distribution (in terms of type and parameters). In cases where the distribution is not known, as in this study,

empirical tests can be implemented with the Kuhn (2019) approach to make an educated guess on the best fit available to a data set. As explained in the materials and methods, we compared five different link functions (logistic, probit, cloglog, loglog and cauchit) to choose the most appropriate one for each of the outcome variables. The best link function was different for each of the categorical variables tested (i.e cloglog for AS, loglog for FD and cauchit for CQ). However, the difference in performance across link functions was very small (below 4% accuracy difference for all variables).

Regarding the results for hyperparameter tuning of RF, the best number of explanatory variables available for splitting at each tree node (i.e. mtry) ranged from half of the available number (5 for FD) to all explanatory variables (10 for AS and 11 for CW) (Figure 3.4). As reviewed by Biau and Scornet (2016) there is no consensus in the literature on the effect of adopting different values of mtry (i.e. number of variables included in the model at a time). Some authors claim that this tuning parameter has little impact on the performance of the method while others recommend using values as large as possible (if possible equal to the number of all explanatory variables). In this analysis, the small number of available explanatory variables allowed us to perform an exhaustive search on mtry. In other words, all possible values of mtry were tested to choose the best one. Results across analysis of all outcome variables indicate that the model performance benefits from having larger values of mtry, in some cases equal to all available explanatory variables.

For the hyperparameter tuning of NN, the grid search method was used. Grid search is a general approach in which a set of candidate values is defined, then reliable estimates of model performance across candidate values are produced to determine the optimal setting (Kuhn and

Johnson, 2016). The candidate values were chosen to span a wide search range, while exploiting the computing capabilities available at a Center for High Throughput Computing where the analysis was performed. The NN grid search results tended to favor higher number of hidden layers with higher number of nodes per layer and smaller values of decay (Figures 5 – 8). Both higher number of hidden units and layers enable expressing more complicated non-linear functions, extending the classification capability (Liakos et al., 2018). Results may suggest that the saturation of the networks was not yet reached, meaning that fitting more complex networks (in terms of larger number of layers and nodes per layer) could yield better performance. However, due to the large scale of the data set in this study, the effort conduced in the NN analysis already exceeded 15 thousand computing hours. Therefore, it would be unfeasible to make further tests for increased complexity (and consequently run-time) with the computational capabilities available. Lastly, it is relevant to understand that even when broader search spaces for the grid are utilized, solutions for best tuning parameters are not guaranteed to be the "global" solution (Kuhn and Johnson, 2016).

Regarding model quality assessment, across all variables there was a tendency of superior performance of regression and RF methods, while NN tended to present the poorest performance. Results suggests that methods that intrinsically model non-linear relationships (such as RF and NN) did not perform better for the outcome variables studied (CW, AS, FD and CQ). It is important to mention that linear models can be adapted to nonlinear trends in the data by manually adding higher order model terms. However, to do this, one must know the specific nature of the nonlinearity in the data (Kuhn and Johnson, 2016), for example interaction among specific variables or quadratic effects. The knowledge on such non-linear relationships was not available for this data set. Inherently non-linear in nature models have the advantage that the exact form of the nonlinearity does not need to be known explicitly or specified prior to model training (Kuhn

and Johnson, 2016). Lastly, it could be argued that the better performance of RF for CW and AS could be related to the fact of the method being robust to noise in the predictor variables (Biau and Scornet, 2016).

Results for explanatory variable importance for the prediction of CW, AS, FD and CQ highlighted patterns that interestingly were mostly in agreement with conclusions reached in experimental settings and field observations. One of those patterns is the importance of nutrition used by the farm, and the price of corn 3mo before animals were finished, followed by soil quality and climate to predict AS. As described by Millen et al. (2011), production cycles carried out solely on grazing systems with only mineral supplementation lead to older animals at market. This is due to animals putting on weight during the rainy season when the grass quality is higher but losing body weight in the dry season. A big reduction of AS can be achieved when animals are finished in feedlots (Millen et al., 2009). The same nutritional variables also showed importance as predictors of FD, which is in agreement with the observation that feedlot operations are often times utilized just to finish animals and achieve a minimum of 4mm fat cover as demanded by the Brazilian market (Millen et al., 2009, 2011). Lastly, nutritional variables ranged from moderate to important (for regression and RF) to predict CQ. Even though the effects of nutrition on specific quality parameters, such as FD and CW, are well studied this can be due to the fact that CQ takes into account several other variables such as AS, gender and body condition scores.

Participation in a technical advising program showed moderate to high importance for the prediction of all beef production and quality variables. This indicates that expert knowledge is important to aid farmers making management and production decisions that can improve production outcomes. Another variable that showed moderate to high relevance across all outcome

variables was the finished cattle sales price. This implies that to a certain degree, the beef market conditions are also relevant for the prediction of meat traits. Lastly, the previously mentioned variable importance results point to an important feature to predict carcass production and quality at the national level: not only physiological variables (such as animal category and AS) are important predictors, but also environmental and external variables play an important role as well. Our results highlight that for real-world prediction these factors should not be ignored.

Machine learning approaches are currently being applied for prediction with the objective of optimizing economic efficiency of farm systems in many livestock species (Liakos et al., 2018; Passafaro et al., 2019). Common algorithms in such applications are decision trees and NN. In fact, the use of machine learning techniques to predict beef cattle traits is not new. For example, Alonso et al. (2007) used machine learning to predict beef cattle conformity scores and growth with 91 animals. Additionally, Alonso et al. (2013) applied support vector machines to predict CW in advance to slaughter for the Asturiana de los Valles cattle breed based on zoometric measurement features with 144 animals. The novelty of the application presented here is not only the utilization of machine learning methods to predict beef production and quality, but more importantly, the capability of utilizing real-world large scale integrated data, representative of a diverse national context, to do so. Knowledge on which methods and variables are necessary to forecast beef production and quality at the national level can be a valuable tool to predict the future of the Brazilian market. Such projections can be useful in the following years to better allocate resources, with the hope of improving the sustainability of beef production. Lastly, forecasting can also be a tool to aid decision making, allowing farmers to prepare for changes ahead of time.

It is arguable that the analysis performed here could be improved. For instance, the data set might be missing important explanatory variables for the prediction of meat production and quality, such as the genetic merit, breed composition, and health of the animals. Unfortunately, no other variables were available from the sources used. Recording those variables in different sectors of the market, such that they can be included in future applications, could increase the accuracy of prediction models. Another important point is that despite the fact that this analysis aims to provide a national snapshot of the Brazilian production, the production system is quite heterogeneous (Millen et al., 2009 and Oliveira, 2018). This means that regional variations not accounted for in this analysis are possible, and this could be explored in future studies. Lastly, we acknowledge that the objective of this analysis was solely prediction of future trends, in other words, accurately projecting the chances that something will (or not) happen. The focus of this type of method is to optimize prediction accuracy (Kuhn and Johnson, 2016). Therefore, no causal claim can be made from these results (Rosa and Valente, 2013; Bello et. al., 2018).

In the years to come it will be essential to address the current challenge of augmenting production to nourish a growing human population without increasing the environmental footprint. With greater awareness of the need to preserve natural resources, methods with sustainable perspectives become more appealing (Millen et al., 2011). Understanding how to predict the future of livestock production using large scale data will be core to projecting future trends and optimizing the allocation of resources at all levels of the production chain, rendering animal production more sustainable. In this analysis we were capable of predicting future beef production and quality with information on over 4 million head of cattle, corresponding to 4.3% of the Brazilian national production. Despite beef cattle production being a complex system, many times influenced by the farmer's personal interests, meat market regulators, and sanitary issues (such as

spread of diseases) this analysis shows that by integrating different sources of data, it is possible to forecast meat production and quality at the national level with moderate-high levels of accuracy.

3.6 References

"Agrolink." 2019. https://www.agrolink.com.br/cotacoes (October 20, 2019).

Aiken, V. C. F., J. R. R. Dórea, J. S. Acedo, F. G. Sousa, F. G. Dias, and G. J. M. Rosa. 2019. "Record linkage for farm-level data analytics : comparison of deterministic, stochastic and machine learning Methods." Comput. Eletron. Agr. 163: 104857. doi:10.1016/j.compag.2019.104857.

Alonso, J., A. Bahamonde, A. Villa, and A. R. Castañón. 2007. "Morphological assessment of beef cattle according to carcass value." Livest. Sci. 107: 265–73. doi:10.1016/h.livsci.2006.09.027.

Alonso, J., A. R. Castañón, and A. Bahamonde. 2013. "Support vector regression to predict carcass weight in beef cattle in advance of the slaughter." Comput. Eletron. Agr. 91: 116–20. doi:10.1016/j.compag.2012.08.009.

Álvares, C. A., J. L. Stape, P. C. Sentelhas, J. L. de Moraes Gonçalves, and S. Gerd. 2014. "Köppen' s climate classification map for Brazil." Meteorol. Z. 22(6): 711–28. doi:10.1127/0941-2948/2013/0507.

Bello, N. M., V. C. Ferreira, D. Gianola and G. J. M. Rosa. 2018. "Conceptual framework for investigating causal effects from observational data in livestock". J. Anim. Sci. 96:4045-4062. doi: 10.1093/jas/sky277.

Biau, G., and E. Scornet. 2016. “A random forest guided tour.” Test 25(2): 197–227. doi:10.1007/s11749-016-0481-7.

FAO. 2009. “How to feed the world in 2050.” Food and Agriculture Organization of the United Nations, Rome.

FAO. 2014. “Meat consumption” http://www.fao.org/ag/againfo/themes/en/meat/background.html (November, 2, 2019).

IBGE. 2018. “Indicadores IBGE estatística da produção pecuária.”: 37–40. ftp://ftp.ibge.gov.br/Producao_Pecuaria/Fasciculo_Indicadores_IBGE/abate-leite-couro-ovos_201802caderno.pdf.

“Instituto Brasileiro de Geografia e Estatística - IBGE.” 2019. https://www.ibge.gov.br/apps/atlas_nacional/ (September 10, 2019).

Kamilaris, A., A. Kartakoullis, and F. X. Prenafeta-boldú. 2017. “A review on the practice of big data analysis in agriculture.” Comput. Eletron. Agr. 143: 23–37. doi:10.1016/j.compag.2017.09.037.

Kuhn, M. 2019. “CARET: Classification And REgression Training.” R package version 6.0-84.

Kuhn, M. and K. Johnson. 2016. Applied predictive modeling. 1th ed. Springer, Saline, Michingan, USA.

Liakos, K. G., P. Busato, D. Moshou, S. Pearson, D. Bochtis. 2018. “Machine learning in agriculture: A review.” Sensors 18(2674): 1–29. doi:10.3390/s18082674.

Millen, D. D., R. D. L. Pacheco, M. D. B. Arrigoni, M. L. Galyean, and J. T. Vasconcelos. 2009.

"A snapshot of management practices and nutritional recommendations used by feedlot nutritionists in Brazil." J. Anim. Sci. 87(10): 3427–39. doi: 10.2527/jas.2009-1880.

Millen, D. D., R. D. L Pacheco, P. M. Meyer, P. H. M Rodrigues, and M. D. B. Arrigoni. 2011. "Current outlook and future perspectives of beef production in Brazil." Anim. Front. 1(2): 46–52. doi:10.2527/af.2011-0017.

Morota, G., R. V. Ventura, F. F. Silva, M. Koyama, and S. C. Fernando. 2018. "Machine learning and data mining advance predictive big data analysis in precision animal agriculture." J. Anim. Sci. 96: 1540–50. doi: 10.1093/jas/sky014.

Oliveira, M. 2018. "Contributions of Brazilian cattle." *Pesquisa FAPESP* (264). https://revistapesquisa.fapesp.br/en/2018/06/25/contributions-of-brazilian-cattle/.

Oliveira, M. 2019. "Produção da pecuária municipal 2018." Catalog of the Instituto Brasileiro de Geografia e Estatística 84(01014234): 1–8. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=784.

Passafaro, T. L., D. V. de Stroet, N. M. Bello, N. H. Williams and G. J. M. Rosa. 2019. "Generalized additive mixed model on the analysis of total transport losses of market-weight pigs." J. Anim. Sci. 97(5): 2025–34. doi: 10.1093/jas/skz087.

Pham, X., and M. Stack. 2018. "How data analytics is transforming agriculture." Bus. Horiz. 61(1): 125–33. doi:10.1016/j.bushor.2017.09.011.

Rosa, G. J. M., and B. D. Valente. 2013. "Breeeding and genetics symposium: Inferring causal effects from observational data in livestock". J. Anim. Sci.. 91: 553-564.

Stekhoven, D. J. 2013. "MissForest: nonparametric missing value imputation using random

forest." R package version 1.4.

Zia, M., J. Hansen, K. Hjort, and C. Valdes. 2019. "Brazil once again becomes the world's largest beef exporter." United States department of Agriculture - Economic Research Service. https://www.ers.usda.gov/amber-waves/2019/july/brazil-once-again-becomes-the-world-s-largest-beef-exporter/ (October 21, 2019).

Chapter 4: Exploring spatial heterogeneity of beef production and quality using large scale integrated data from Brazil

4.1 Abstract

This study used real, large scale beef cattle data integrated from different sources to explore spatial heterogeneity of beef production in Brazil by contrasting Linear Regression (LR) and Geographically Weighted Regression (GWR) techniques. The aims were to verify whether the factors that influence meat production and quality differ according to the farm's geographical location, and to compare a set of models estimated using GWR with the global model estimated via LR, in terms of model fit. More specifically, outcome variables to study meat production and quality at the national level included carcass weight when finished (CW), age when finished (AS), carcass fat deposition (FD) and carcass quality (CQ). Outcome variables were obtained from a meat packing company with facilities located in most Brazilian beef producing states. Explanatory variables included: animal category, nutritional products used, adoption of technical consulting, economic variables – such as corn and carcass prices, climate, and soil fertility at the location animals were raised. Models were evaluated in terms of goodness of fit using the Akaike Information Criterion (AIC) and the adjusted-R^2. The GWR models presented very little variation in the estimated coefficients for the variables that influence all outcomes (CW, AS, FD and CQ) across different farm locations in the country. Also such coefficients were very close to the "global" values estimated with LR for all outcomes. With further testing (presented in the Appendix), we concluded that after accounting for variables that naturally capture spatial heterogeneity (such as climate and soil at the farm) meat production and quality were not

influenced differently by the explanatory variables in each of the regions in the study. Regarding model fit, the LR and GWR presented similar results with virtually the same adjusted-R^2 and AIC for most outcomes. For CW, AIC = 173204.20 for both models and adjusted-R^2 of 76.99% for LR and of 76.98% for GWR. For AS, AIC = 140561.10 and adjusted-R^2 = 09.31% for both models. For FD, AIC = 45568.68 and adjusted-R^2 = 35.72% for both models. Lastly, for CQ, AIC was slightly better for GWR (-12585.38), when compared to LR (-12585.36), while adjusted-R^2 was 35.48% for both. These results indicate that when inferring meat quality and production in Brazil, the inclusion of explanatory variables, such as climate and soil of the farms where animals were raised, naturally captures spatial heterogeneity. After accounting for those variables, animal physiology seems to be consistent across regions and very little spatial heterogeneity is captured by GWR models, presenting mostly no advantage when compared to LR.

Keywords: beef cattle, geographically weighted regression, integrated, large scale data, spatial heterogeneity.

4.2 Introduction

Animal agriculture plays an important role as a source of high-quality nutrients that can improve human diets around the world, and beef is one of the most important sectors of animal agriculture. It is the third most produced livestock meat in the world after pork and poultry (FAO, 2014). Beef production is a common practice in many countries, and Brazil has the largest commercial herd (213.5 million heads of cattle) and is the largest exporter of beef in the world (Oliveira, 2019; Zia et al., 2019), with 2.1MM metric tons exported from the 9.9 produced in 2018, corresponding to 20% of total beef exports around the world. The largets production states in Brazil

are Mato Grosso, Goiás, Minas Gerais, Mato Grosso do Sul, and Pará, which together are responsible for 54.2% of the national production (Oliveira, 2019). The Brazilian beef market has shown a trend of expansion over the past years, and the USDA predicts that this trend will continue with the country reaching 23% of the world's total exports by 2028 (Zia et al., 2019). An aspect that recently has been contributing to the expansion of the Brazylian meat production is the fact that the mature beef cattle industry in Brazil is based mostly on grass-fed cattle (Millen et al., 2011). Cattle spend most of their lives in a free-range grazing pasture and this is appealing from an animal welfare and sustainability standpoints.

Beef cattle production is an activity developed in all Brazilian ecosystems with great variability in terms of climate where animals are raised, breed compositions used, quality of pastures and type of nutrients avaialable to cattle, among other components (Millen et al., 2009; Oliveira, 2018). The combination of the diversity of conditions in which beef production takes place with complex biological mechanisms that define meat production and quality makes studying these traits at a national level a challenging task.

Regression models have traditionally been used in animal sciences for the inference of many traits (Bello and Renter, 2018). Linear regression (LR) is a multivariate analysis technique that aims to explain the relationship between a continuous outcome and a set of explanatory variables. It utilizes a "global" approach in which a single common relationship between explanatory variables and the outcome of study is defined for all observations. This is well suited when the mechanisms that influence the outcome do not change across the regions where information was collected.

When analyzing beef cattle production at the national level, having variability in the production systems raises a concern about the utilization of a "global" approach. Specifically, if beef production and quality is a non-stationary process, then the "global" approach with traditional regression is not appropriate. The Geographically Weighted Regression (GWR) method (Brunsdon et. al, 1996) is a method that allows for modeling spatially heterogeneous (non-stationary) processes. Those are processes that may change (in mean, median, variance, etc.) across different regions. GWR adjusts the regression model to each region in the data set using geographical location of the other observations to weight parameter estimates (Albuquerque et al., 2016). GWR has been used to model non-stationary processes in many fields such as Economics (Albuquerque et al., 2016) and Health (Mena et al., 2018).

This study used large scale data integrated from different sources to explore the spatial heterogeneity of beef cattle production systems in Brazil. The data contained information on almost 4 million finished animals, corresponding to 4.2% of the Brazilian national beef production between 2014-2016. The outcome variables to study meat production and quality at the national level included carcass weight when finished (CW), age when finished (AS), carcass fat deposition (FD), and carcass quality (CQ). Explanatory variables included: animal category, nutritional products used by farms where animals were raised, adoption of technical consulting, economic variables – such as price of corn and price paid per carcass when finished, and climate and soil fertility of the farms were animals were raised. The objectives were to verify whether the factors that influence meat production and quality in Brazil differ according to the farm's geographical location, and to compare a set of models estimated using GWR with the global model estimated via LR.

4.3 Material and Methods

Integrated database

The data set utilized in this analysis followed the same integration procedures with the same sources, including the same variables described in detail in chapter 3. The data utilized in this chapter contained information on 816,254 observations (group of animals) from 5,185 farms comprising a total of 3,949,446 beef cattle between 2014 and 2016. The data analyzed in this study corresponds to 4.2% of the Brazilian cattle national production of 94.2 million heads of cattle for the 2014 to 2016 period (IBGE, 2018).

The data set contained information on 642 municipalities located in 12 of the 26 Brazilian states (Acre, Bahia, Goiás, Mato Grosso, Mato Grosso do Sul, Maranhão, Minas Gerais, Pará, Paraná, Rondônia, São Paulo and Tocantins). The distribution of farms and number of animals finished per state is presented in Figure 4.1.

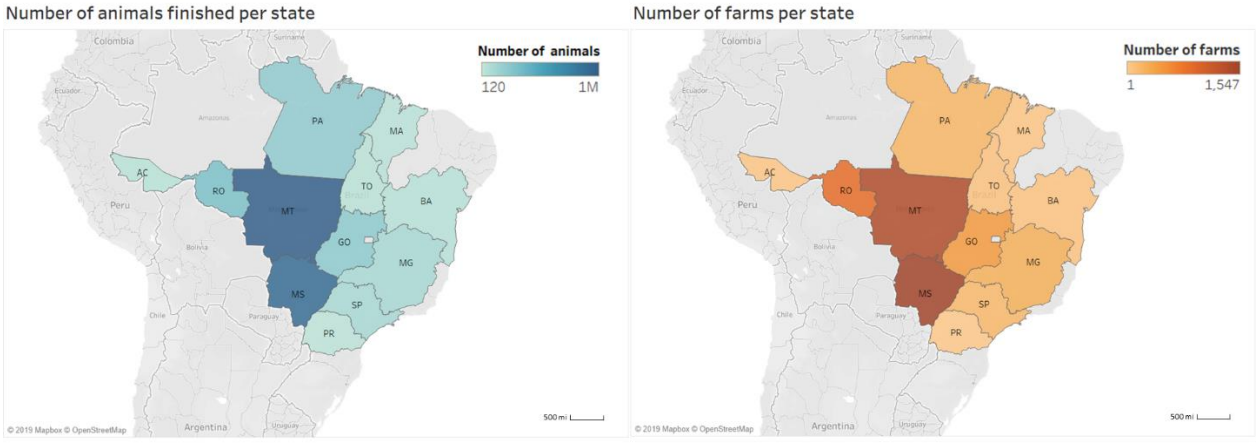

Figure 4.1: Distribution of observations for farms (in the left) and finished animals (in the right) per state in Brazil.

The distribution of the animal traits obtained from the meat packing plant for the whole country is presented in Figure 4.2. For this study, 914 farms participated on a technical advising program, 3,823 did not, and the remaining farms had missing information for this variable. Regarding nutrients used by the farm, the average animal consumption of premix for non-feedlot animals (PNF) was 62.1kg (SD = 163.5kg), while for feedlot premix (FP) was 1.4kg (SD = 13.8kg), and for feedlot premix with additives (FA) was 3.1kg (SD = 28.4kg). The number of observations per soil type was as follow: light green = 59,686, cream = 27,206, orange = 517,972, yellow = 370, purple = 10,269, dark green = 9,795, pink = 13,154, light blue = 12,291 and gray = 165,511. The number of observations for each climate was: tropical zone, without dry season (Af) – 29,052; – tropical zone, monsoon (Am) – 387,252; tropical zone, with dry summer (As) – 659; tropical zone, with dry winter (Aw) – 355,992; dry zone, semi-arid, low latitude and longitude (BSh) – 417; humid subtropical zone, oceanic climate without dry season, with hot summer (Cfa) 35,349; humid subtropical zone, oceanic climate without dry season, with temperate summer (Cfb) – 253; humid subtropical zone, with dry winter, and hot summer (Cwa) – 5,899; humid subtropical zone, with dry winter, and temperate summer (Cwb) – 1,381.

Table 4.1 presents the distribution of outcome variables per Brazilian state (Acre, Bahia, Goiás, Mato Grosso, Mato Grosso do Sul, Maranhão, Minas Gerais, Pará, Paraná, Rondônia, São Paulo and Tocantins). The average CW per state ranged from 244.15kg (±60.55) in Bahia to 290.45kg (±59.57) in Paraná, indicating that CW varies substantially across states. The distribution of AS per state presented drastic variation, with Paraná having all animals finished before 36mo of age, while Bahia had over 60% of the animals finished after 36mo of age. FD also varies considerably across Brazil, with states like Bahia having 68% of animals under 1mm, which is below the minimum of 4mm desired by the Brazilian market (Millen et al., 2011), while states like

Goiás, Mato Grosso, Mato Grosso do Sul, Paraná and São Paulo have most animals between the 1-10mm. Likewise, CQ varies greatly across Brazilian states, with Maranhão having only 0.1% of finished animals in the desirable range while Paraná had 57% of animals finished in the desirable range.

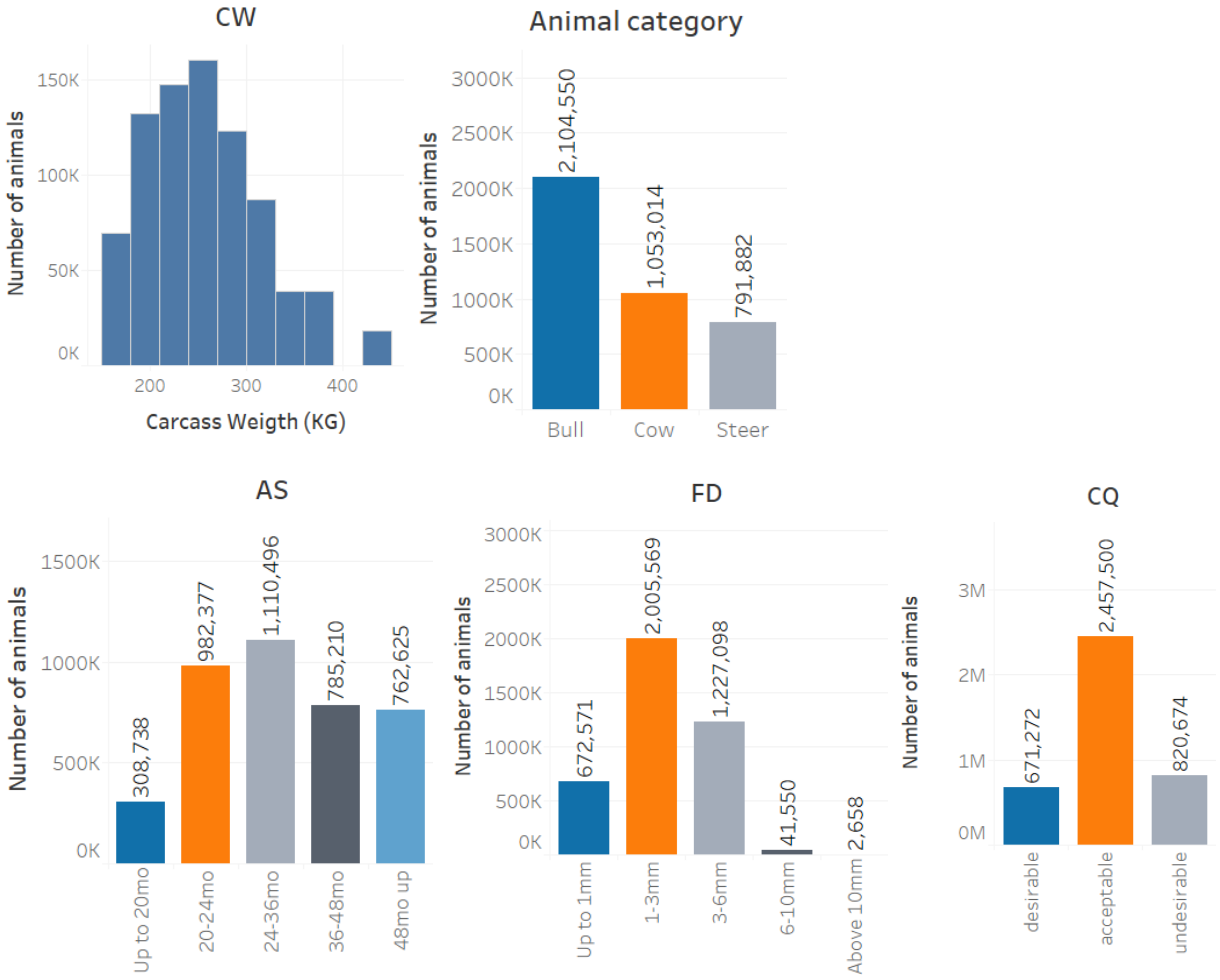

Figure 4.2. Distribution of animals finished according to of carcass weight (CW), carcass quality (CQ), animal category, age at slaughter (AS) and fat deposition (FD).

Table 4.1: Distribution of outcome variables per state. The mean (kg) and SD (±) is presented for the continuous variable carcass weight (CW), while the percentage in each category is presented for age at slaughter (AS), fat deposition (FD) and carcass quality (CQ).

		State											
		Acre	**Bahia**	**Goiás**	**Mato Grosso**	**Mato Grosso do Sul**	**Maranhão**	**Minas Gerais**	**Pará**	**Paraná**	**Rondônia**	**São Paulo**	**Tocantins**
CW	Mean	252.01	251.45	265.03	248.03	257.28	258.24	261.73	244.15	290.45	235.82	272.55	265.75
	(SD)	(54.26)	(60.14)	(61.43)	(65.04)	(57.38)	(55.81)	(59.57)	(60.55)	(22.52)	(59.38)	(59.82)	(56.59)
AS	< 20mo	0.01%	0.20%	10.87%	07.91%	11.17%	0.01%	3.56%	0.23%	0.83%	0.91%	15.21%	0.03%
	20-24mo	6.88%	9.33%	29.34%	27.68%	23.90%	0.00%	16.42%	12.23%	82.50%	24.29%	37.18%	53.58%
	24-36mo	92.98%	29.79%	35.25%	25.41%	24.35%	30.36%	23.06%	48.26%	16.67%	25.30%	30.84%	39.69%
	36-48mo	0.13%	43.15%	14.05%	22.94%	16.44%	69.58%	50.29%	18.46%	0.00%	17.17%	09.01%	03.98%
	> 48mo	0.00%	17.53%	10.49%	16.06%	24.14%	0.05%	06.67%	20.82%	0.00%	32.33%	07.76%	02.72%
FD	< 1mm	48.63%	67.79%	14.13%	20.60%	04.91%	58.45%	39.76%	26.36%	0.00%	27.56%	03.26%	11.67%
	1-3mm	15.61%	21.99%	66.47%	47.94%	49.75%	41.34%	43.13%	59.99%	43.33%	48.78%	62.95%	74.74%
	3-6mm	35.67%	09.47%	18.07%	29.94%	44.18%	0.21%	16.99%	13.27%	56.67%	23.06%	33.14%	13.16%
	6-10mm	0.09%	0.71%	01.26%	01.41%	01.11%	0.00%	0.10%	0.36%	0.00%	0.56%	0.61%	0.43%
	> 10mm	0.00%	0.04%	0.07%	0.11%	0.05%	0.00%	0.02%	0.02%	0.00%	0.04%	0.04%	0.00%
CQ	Desirable	27.84%	3.01%	9.33%	16.40%	25.39%	0.15%	6.74%	6.10%	56.67%	8.99%	21.52%	10.32%
	Acceptable	21.33%	25.82%	72.11%	58.98%	67.09%	38.86%	49.56%	63.93%	43.33%	57.57%	70.30%	85.01%
	Undesirable	50.83%	71.17%	18.56%	24.62%	7.52%	60.99%	43.70%	29.97%	0.00%	33.44%	8.18%	4.67%
Number of animals		*30,274*	*33,749*	*290,560*	*1,378,724*	*1,254,468*	*17,737*	*130,183*	*269,859*	*120*	*391,460*	*126,329*	*25,983*

Data manipulation and analysis

After data integration (as described in Chapter 3), the data was pre-processed for the analysis of CW, AS, FD and CQ. From all explanatory variables considered in the models, only the binary variable for participation in the technical advising program had missing data, with 27% of the farms having missing information for at least one year. This variable was imputed using bagged trees with the R package "missForest" (Stekhoven, 2013). Bagged trees were created using all other variables in the training set, such that when a sample had a missing value for a predictor, the bagged model was used to predict this value. The estimated error of the imputation (out of bag proportion of falsely classified samples) was very low (0.0077%).

Due to computing memory constraints imposed by the software used to implement the approach described below, the data was compressed (i.e. averaged) from 816,254 observations on 5,185 farms to 19,673 observations of 5,185 farms. All traits were considered as a summary of the season per farm per year. The season was defined in two categories: first half of the year (January through June) and second half of the year (July through December), for all three years (2014, 2015 and 2016). The animal traits CW, AS and FD were considered as the average of the farm for the specific period. For AS, each category's value was considered as 18mo, 22mo, 30mo, 42mo and 56mo (midpoint for the category range) and then averaged. For FD, each category was also summarized by its midpoint, i.e. the categories "absent", "low", "medium", "high and "excessive" were set to 0.5mm, 2mm, 4.5mm, 8mm, 13mm, respectively. Observations were then averaged for farm and period. CQ was grouped as a binary variable (acceptable or undesirable) and the proportion of acceptable outcomes for the season per farm was used. Animal category was hot encoded (i.e. each category was divided in separate columns) containing the percentage of animals

in that category. The economic variables price of corn three months before slaughter and sales price for the finished animals per month were averaged across the six months of the season for the state each farm was located. The participation in a technical consulting program and nutrients sold to farms were already collected in a yearly basis per farm, such that no data processing was necessary for this variable. The soil and climate class of each farm remained constant over the studied period, therefore data processing was also not necessary for these variables. The distribution of outcome variables CW, AS, FD and CQ in terms of averages, SD, minimum and maximum after data compression are presented in Table 4.2.

Table 4.2: Distribution outcome variables carcass weight (CW), age at slaughter (AS), Fat deposition (FD) and carcass quality (CQ) after data compression per farm per season per year.

	Mean	**SD**	**Minimum**	**Maximum**
CW (kg)	247.40	41.47	150.00	420.00
AS (mo)	37.87	9.10	18.00	56.00
FD (mm)	2.62	0.96	0.50	10.17
CQ (%)	0.73	0.22	0.00	100.00

To test the hypothesis that effects influencing meat production and quality in Brazil differ according to the farm's geographical location, a "global" regression model was contrasted with a "local" model utilizing GWR (Brunsdon et al., 1996). The explanatory variables considered for each outcome in both models were exactly the same, and are presented in Table 4.3.

Table 4.3: Predictor variables utilized in the analysis of carcass weight (CW); age when finished (AS); fat deposition (FD) and carcass quality (CQ). Explanatory variables to models included: animal category, participation in a technical advising program (PTAP); kg of premix nor non-feedlot per animal (PNF); kg of feedlot premix per beef animal (FP); kg of feedlot premix with additive products per beef animal (FA); finished cattle sales price (FCSP); corn price 3mo before finished (CP3B); soil fertility classification (SOIL); climate classification (CLIM); and month when finished (MO).

Outcome	**Predictors**
CW	AS; animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO
AS	animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO
FD	AS; animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO
CQ	animal category; PTAP; PNF, FP, FA; FCSP; CP3B; SOIL; CLIM; and MO

The "global" model refers to a standard multiple linear regression (LR) which can be described as:

$$y_i = \beta_0 + \sum_{p=1,\ldots,q} \beta_p X_{ip} + \varepsilon_i \quad (1)$$

where y_i is the ith observation of the continuous outcome variable; X_{ip} is the ith observation of the pth explanatory variable, and ε_i are independent normally distributed error terms with zero mean and common variance σ_ε^2. β_0 represents the overall intercept and β_1 to β_p are the coefficients relating each of the p explanatory variables to the outcome. The maximum likelihood solution for the model parameters $\beta = [\beta_0, \beta_1, \ldots, \beta_p]^t$ can be expressed in matrix notation as:

$$\hat{\beta} = (X^t X)^{-1}(X^t Y)$$

where the outcome variable is the single column vector $Y = [y_1, y_2, \dots, y_n]^t$ and the explanatory variables are represented in the incidence matrix $X = [1, x_1, \dots, x_p]$, where 1 is a column vector of ones, and $x_j = [x_{1j}, x_{2j}, \dots, x_{nj}]$ represents the vector for each of the explanatory variables.

The equivalent expression for the GWR, proposed by Brunsdon et al. (1996) is:

$$y_i = \beta_{i0} + \sum_{p=1,\dots,q} \beta_{ip} X_{ip} + \varepsilon_i$$

where β_{i0} and β_{ip} are parameters specific for each location i, with all other variables defined as in model (1). All assumptions of a classical linear regression remain for GWR, i.e. independent and identically distributed residuals, linear relation between outcome and explanatory variables and normal distribution of outcomes.

In the case of the GWR model, the maximum likelihood solution can be represented in matrix notation as:

$$\hat{\beta}(i) = (X^t W_i X)^{-1} (X^t W_i Y),$$

in which the diagonal matrix W_i can be defined as:

$$W_i = \begin{bmatrix} w_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{im} \end{bmatrix}$$

W_i differs for each point of latitude and longitude defined for i. The weights w_{im} presented in the main diagonal are obtained via a weighting kernel function.

As shown above, the GWR model requires the choice of a spatial weighting function that quantifies the spatial relationship or dependency between data points. There are three key elements for building the weighting matrix: the kernel function, the type of distance, and the bandwidth

(Brunsdon et al., 1996). The natural choice and one of the most common weighting functions found in the literature that has been successful for GWR in other applications is the Gaussian function (Albuquerque et al., 2016), which was the function of choice in this analysis and is described in detail below. Also, the traditional way of defining the distance between two data points is by using the Euclidean distance, which was the approach utilized here (described below). Lastly, GWR requires choosing the bandwidth parameter as fixed or variable. The bandwidth controls the variance in the weighting function (Brunsdon et al., 1996). In situations where the data is not equally distributed between regions, it is recommended that an adaptive bandwidth is used (Albuquerque et al., 2016) and that was then the approach used here. When using adaptive bandwidths, a 'fixed quantity that reflects local sample size', or number of neighbors, k, must be defined to ensure sufficient (and constant) local information for each local calibration (Brunsdon et al., 1996). In this analysis, optimal bandwidths were automatically selected using a leave-one-out cross-validation procedure to determine k using the *bw.gwr* function in GWmodel package in R (Binbin et al., 2019). More specifically, the best bandwidth was found by testing different k values using cross-validation and choosing the one that minimized the Root Mean Square Error (RMSE).

Formally, the adaptive Gaussian Kernel can be described as:

$$w_{im} = exp\left\{-\frac{1}{2}\,(d_{im}/b_{i(k)})^2\right\}$$

where d_{im} represents the Euclidean distance between two data points, defined as:

$$d_{im} = \sqrt{(i_{latitude} - m_{latitude})^2 + \left(i_{longitude} - m_{longitude}\right)^2}$$

and $b_{i(k)}$ represents the adaptive bandwidth with k representing the number of neighbors closest to point i, chosen to determine the bandwidth.

The model fitness of the LR and GWR approaches was evaluated via two techniques, the Akaike Information Criteria (AIC) and the proportion of the variability in the outcome variable explained by the explanatory variables (adjusted- R^2). All analyses were performed using the GWmodel R package (Binbin et al., 2019).

4.4 Results

Results for both global (i.e. LR) and local (i.e. GWR) models are presented in Tables 4-11 for each of the outcome variables. The number of closest neighbors chosen by cross-validation to determine the adaptive bandwidths in the Gaussian model was 2,589 for all models.

For the global model fitted for CW, results presented in Table 4.4 indicate that important variables for the inference of CW were: animal category (coefficient estimates equal to -95.40 – P-value < 0.001, and 10.38 with P-value < 0.001, for the categories cow and steer, respectively, compared to bull), technical consulting (estimated coefficient equal to 5.21 with P-value < 0.001 for use compared to non-use) and AS (estimate of 0.46 with P-value < 0.001). Nutrition and sales price had significant but smaller effects. Significant differences were also observed between some levels of climate and soil (lower level of significance for soil). The coefficients estimated for CW using GWR are presented in Table 4.5. It is noted that both the intercept and coefficients presented small variation between different locations. Results for model comparison are presented in Table

4.12. The AIC of the global model was 173176.20 and adjusted-R^2 = 76.99%, while the GWR model had the same AIC and virtually the same R^2 (76.98%).

Table 4.4: Global model estimation of the outcome carcass weight (CW) and respective coefficients for explanatory variables.

Variables	Coefficients	Standard Error	T value	Pr (>|t|)
Intercept	234.2000	6.369	36.775	< 2e-16 ***
Age	0.4637	0.01639	28.286	< 2e-16 ***
Animal category_Steer	-10.4000	0.5245	-19.822	< 2e-16 ***
Animal category_Cow	-95.3800	0.4170	-228.726	< 2e-16 ***
Technical consulting_Yes	5.2120	0.3610	14.436	< 2e-16 ***
PNFpc	0.005096	0.0003966	12.848	< 2e-16 ***
FPpc	0.01818	0.007163	2.538	0.0112 *
FApc	0.006990	0.003532	1.979	0.0478 *
Cattle sales price	0.2350	0.01455	16.151	< 2e-16 ***
Corn 3mo before price	-0.1403	0.02026	-6.927	4.42e-12 ***
Climate_Am[1]	-4.241	0.9951	-4.262	2.04e-05 ***
Climate_As[1]	-0.1832	4.470	-4.098	4.18e-05 ***
Climate_Aw[1]	0.1780	1.007	0.177	0.8596
Climate_BSh[1]	20.019	11.61	1.739	0.0820 .
Climate_Cfa[1]	-2.313	1.246	-1.856	0.0634 .
Climate_Cfb[1]	-5.841	8.194	-0.713	0.4759
Climate_Cwa[1]	-1.552	2.085	-0.744	0.4567
Climate_Cwb[1]	-4.459	3.307	-1.348	0.1775
Soil_Light blue[2]	6.374	6.170	1.033	0.3016
Soil_Gray[2]	13.36	6.010	2.223	0.0262 *
Soil_Cream[2]	10.92	6.067	1.800	0.0719 .
Soil_Orange[2]	12.93	6.005	2.152	0.0314 *
Soil_Pink[2]	5.255	6.120	0.859	0.3905
Soil_Purple[2]	5.106	6.151	0.830	0.4065
Soil_Light green[2]	6.446	6.026	1.070	0.2847
Soil_Dark green[2]	11.29	6.151	1.836	0.0664 .

Significance (P-value) codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1]In relation to the base level Af.;

[2]In relation to the base level yellow.

Table 4.5: Adaptive Gaussian Geographically Weighted model estimation of the outcome carcass weight (CW) and respective coefficients for explanatory variables.

Variables	**Minimum**	**Q1**	**Median**	**Q3**	**Maximum**
Intercept	234.202226	234.202450	234.202547	234.202707	234.202900
Age	0.463731	0.463734	0.463735	0.463736	0.463700
Animal category_Steer	-10.397003	-10.396995	-10.396992	-10.396986	-10.397000
Animal category_Cow	-95.381349	-95.381338	-95.381332	-95.381324	-95.381300
Technical consulting_Yes	5.212028	5.212042	5.212045	5.212049	5.212100
PNFpc	0.005096	0.005096	0.005096	0.005096	0.005096
FPpc	0.018181	0.018181	0.018181	0.018182	0.018200
FApc	0.006990	0.006990	0.006990	0.006990	0.007000
Cattle sales price	0.235001	0.235002	0.235003	0.235003	0.235004
Corn 3mo before price	-0.140344	-0.140343	-0.140343	-0.140343	-0.140342
Climate_Am[1]	-4.241286	-4.241230	-4.241199	-4.241185	-4.241100
Climate_As[1]	-18.317143	-18.317076	-18.317054	-18.317033	-18.317034
Climate_Aw[1]	0.177986	0.177999	0.178006	0.178009	0.178009
Climate_BSh[1]	20.194087	20.194180	20.194214	20.194238	20.194300
Climate_Cfa[1]	-2.312639	-2.312622	-2.312616	-2.312610	-2.312609
Climate_Cfb[1]	-5.841457	-5.841442	-5.841429	-5.841415	-5.841400
Climate_Cwa[1]	-1.551861	-1.551841	-1.551829	-1.551823	-1.551820
Climate_Cwb[1]	-4.459308	-4.459298	-4.459295	-4.459292	-4.459300
Soil_Light blue[2]	6.373762	6.373779	6.373783	6.373787	6.373800
Soil_Gray[2]	13.363348	13.363370	13.363376	13.363379	13.363400
Soil_Cream[2]	10.920162	10.920192	10.920200	10.920207	10.920208
Soil_Orange[2]	12.925198	12.925213	12.925223	12.925234	12.925236
Soil_Pink[2]	5.255112	5.255158	5.255170	5.255181	5.255200
Soil_Purple[2]	5.105609	5.105705	5.105729	5.105753	5.105800
Soil_Light green[2]	6.446042	6.446059	6.446066	6.446073	6.446100
Soil_Dark green[2]	11.294329	11.294369	11.294376	11.294390	11.294400

[1]In relation to the base level Af.;

[2]In relation to the base level yellow.

For the global model fitted for AS, results presented in Table 4.6 indicate that important variables were: animal category (estimates of -3.02 with P-value < 0.001, and 5.21 with P-value < 0.001, for the categories cow and steer, respectively, compared to bull), climate where animals were raised (differences of up to -5.20, with P-value < 0.001, when compared to the baseline level "yellow") and technical consulting (estimate -0.99 with P-value < 0.001 for use compared to non-use). Nutrition and sales price had significant but smaller effects, while most differences observed in soils were not significant. The coefficients estimated for CW using GWR are presented in Table 4.7. Similar to the estimation of CW, all coefficients presented very small variation between different locations. Both AIC (140561.10) and adjusted-R^2 (9.42%) presented in fact the same values for both models.

For the global model fitted for FD, results presented in Table 4.8 indicate that important variables are: animal category (estimates of 1.32 with P-value < 0.001, and 1.50 with P-value < 0.001, for the categories cow and steer, respectively, compared to bull), climate where animals were raised (differences of up to 1.40 with P-value < 0.001, when compared to the baseline class "yellow") and technical consulting (estimate 0.19 with P-value < 0.001 for use compare to non-use). Similar to results presented for AS, nutrition and sales price had significant but smaller effects, while most differences observed in soils were not significant. The coefficients estimated for CW using GWR are presented in Table 4.9. Similar to the estimation of CW and AS, all coefficients presented very small variation between different locations. Similar to AS, AIC (45596.68) and adjusted-R^2 presented the same values for both models (35.72%).

Table 4.6: Global model estimation of the outcome age at slaughter (AS) and respective coefficients for explanatory variables.

Variables	Coefficients	Standard Error	T value	Pr (>|t|)
Intercept	47.6046823	2.7529454	17.292	< 2e-16 ***
Animal category_Steer	3.0218006	0.2274306	13.287	< 2e-16 ***
Animal category_Cow	5.2098628	0.1777798	29.305	< 2e-16 ***
Technical consulting_Yes	-0.9955125	0.1570937	-6.337	2.39e-10 ***
PNFpc	0.0019936	0.0001722	11.579	< 2e-16 ***
FPpc	-0.0133705	0.0031185	-4.288	1.82e-05 ***
FApc	-0.0058702	0.0015378	-3.817	0.000135 ***
Cattle sales price	-0.0628332	0.0063216	-9.939	< 2e-16 ***
Corn 3mo before price	0.0241027	0.0088222	2.732	0.006300 **
Climate_Am[1]	-1.0916004	0.4333647	-2.519	0.011780 *
Climate_As[1]	-4.5873286	1.9465557	-2.357	0.018451 *
Climate_Aw[1]	-3.5075208	0.4377450	-8.013	1.18e-15 ***
Climate_BSh[1]	-6.7283077	5.0569426	-1.331	0.183366
Climate_Cfa[1]	-2.5943594	0.5422489	-4.784	1.73e-06 ***
Climate_Cfb[1]	0.8253458	3.5687911	0.231	0.817109
Climate_Cwa[1]	-5.2018743	0.9073481	-5.733	1.00e-08 ***
Climate_Cwb[1]	1.3898002	1.4402930	0.965	0.334585
Soil_Light blue[2]	-1.3660958	2.6874780	-0.508	0.611235
Soil_Gray[2]	-2.1954405	2.6177009	-0.839	0.401653
Soil_Cream[2]	-4.7292504	2.6423985	-1.790	0.073508 .
Soil_Orange[2]	-2.9809973	2.6154885	-1.140	0.254405
Soil_Pink[2]	3.6554306	2.6655386	1.371	0.170276
Soil_Purple[2]	-1.8153114	2.6790589	-0.678	0.498038
Soil_Light green[2]	-2.9482710	2.6244452	-1.123	0.261286
Soil_Dark green[2]	-6.3694289	2.6788676	-2.378	0.017433 *

Significance (P-value) codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1]In relation to the base level Af.;

[2]In relation to the base level yellow.

Table 4.7: Adaptive Gaussian Geographically Weighted model estimation of the outcome age at slaughter (AS) and respective coefficients for explanatory variables.

Variables	Minimum	Q1	Median	Q3	Maximum
Intercept	47.604660	47.604673	47.604678	47.604683	47.604700
Animal category_Steer	3.021797	3.021803	3.021807	3.021811	3.021811
Animal category_Cow	5.209847	5.209862	5.209873	5.209881	5.209900
Technical consulting_Yes	-0.995518	-0.995514	-0.995512	-0.995511	-0.995510
PNFpc	0.001994	0.001994	0.001994	0.001994	0.002000
FPpc	-0.013371	-0.013371	-0.013371	-0.013370	-0.013400
FApc	-0.005870	-0.005870	-0.005870	-0.005870	-0.005900
Cattle sales price	-0.062834	-0.062833	-0.062833	-0.062833	-0.062833
Corn 3mo before price	0.024102	0.024103	0.0241029	0.024103	0.024103
Climate_Am[1]	-1.091611	-1.09161	-1.091604	-1.091603	-1.091600
Climate_As[1]	-4.587376	-4.587342	-4.587320	-4.587296	-4.587300
Climate_Aw[1]	-3.507534	-3.507526	-3.507519	-3.507512	-3.507513
Climate_BSh[1]	-6.728321	-6.728297	-6.728292	-6.728286	-6.728300
Climate_Cfa[1]	-2.594372	-2.594363	-2.594357	-2.594354	-2.594354
Climate_Cfb[1]	0.825299	0.825322	0.825339	0.825356	0.825400
Climate_Cwa[1]	-5.201876	-5.201871	-5.201866	-5.201862	-5.201900
Climate_Cwb[1]	1.389784	1.389793	1.389799	1.389804	1.389804
Soil_Light blue[2]	-1.366136	-1.366110	-1.366088	-1.366070	-1.366070
Soil_Gray[2]	-2.195452	-2.195440	-2.195434	-2.195431	-2.195431
Soil_Cream[2]	-4.729261	-4.729254	-4.729251	-4.729250	-4.729250
Soil_Orange[2]	-2.981004	-2.980998	-2.980996	-2.980994	-2.981000
Soil_Pink[2]	3.655395	3.655416	3.655436	3.655455	3.655500
Soil_Purple[2]	-1.815346	-1.815326	-1.815317	-1.815313	-1.815313
Soil_Light green[2]	-2.948277	-2.948271	-2.948265	-2.948262	-2.948300
Soil_Dark green[2]	-6.369442	-6.369429	-6.369420	-6.369415	-6.369415

[1]In relation to the base level Af.

[2]In relation to the base level yellow.

Table 4.8: Global model estimation of the outcome fat deposition (FD) and respective coefficients for explanatory variables.

Variables	Coefficients	Standard Error	T value	Pr (>|t|)
Intercept	0.937600	0.24710	3.794	0.000149 ***
Age	-0.004768	0.000636	-7.494	6.99e-14 ***
Animal category_Steer	1.321000	0.02035	64.902	< 2e-16 ***
Animal category_Cow	1.507000	0.01618	93.116	< 2e-16 ***
Technical consulting_Yes	0.196400	0.01401	14.021	< 2e-16 ***
PNFpc	0.000037	0.000015	2.424	0.015372 *
FPpc	0.000214	0.000278	0.769	0.441883
FApc	0.000936	0.000137	6.831	8.67e-12 ***
Cattle sales price	0.003775	0.000565	6.685	2.37e-11 ***
Corn 3mo before price	0.002568	0.000786	3.267	0.001089 **
Climate_Am[1]	-0.195500	0.03862	-5.062	4.19e-07 ***
Climate_As[1]	-0.942700	0.17350	-5.435	5.55e-08 ***
Climate_Aw[1]	-0.106900	0.03907	-2.736	0.006227 **
Climate_BSh[1]	1.404000	0.4506	3.117	0.001832 **
Climate_Cfa[1]	-0.026840	0.04834	-0.555	0.578778
Climate_Cfb[1]	0.106800	0.3180	0.336	0.736918
Climate_Cwa[1]	-0.240800	0.08091	-2.976	0.002926 **
Climate_Cwb[1]	-0.678200	0.1283	-5.285	1.27e-07 ***
Soil_Light blue[2]	0.394800	0.2394	1.649	0.099231 .
Soil_Gray[2]	0.588200	0.2332	2.522	0.011682 *
Soil_Cream[2]	0.320200	0.2355	1.360	0.173874
Soil_Orange[2]	0.514800	0.2330	2.209	0.027173 *
Soil_Pink[2]	0.339100	0.2375	1.428	0.153372
Soil_Purple[2]	0.245800	0.2387	1.030	0.303094
Soil_Light green[2]	0.215600	0.2338	0.922	0.356560
Soil_Dark green[2]	0.438400	0.2387	1.836	0.066326 .

Significance (P-value) codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1]In relation to the base level Af.

[2]In relation to the base level yellow.

Table 4.9: Adaptive Gaussian Geographically Weighted model estimation of the outcome fat deposition (FD) and respective coefficients for explanatory variables.

Variables	Minimum	Q1	Median	Q3	Maximum
Intercept	0.93760	0.93761	0.93761	0.93761	0.93761
Age	-0.0047683	-4.7675e-03	-4.7675e-03	-4.7675e-03	-0.0048
Animal category_Steer	1.3211	1.3211	1.3211	1.3211	1.3211
Animal category_Cow	1.5069	1.5069	1.5069	1.5069	1.5069
Technical consulting_Yes	0.19644	0.19644	0.19644	0.19644	0.19644
PNFpc	0.000044	0.000044	0.000044	0.000044	0.000044
FPpc	0.000214	0.000214	0.000214	0.000214	0.000214
FApc	0.000936	0.000936	0.000936	0.000936	0.000936
Cattle sales price	0.003775	0.003775	0.003775	0.003775	0.003776
Corn 3mo before price	0.002568	0.002568	0.002568	0.002568	0.002568
Climate_Am[1]	-0.19547	-0.19547	-0.19547	-0.19547	-0.19547
Climate_As[1]	-0.94271	-0.94271	-0.94271	-0.94271	-0.94271
Climate_Aw[1]	-0.10688	-0.10688	-0.10688	-0.10688	-0.10688
Climate_BSh[1]	1.4043	1.4043	1.4043	1.4043	1.4043
Climate_Cfa[1]	-0.026838	-0.026838	-0.026838	-0.026838	-0.026838
Climate_Cfb[1]	0.10682	0.10682	0.10682	0.10682	0.10682
Climate_Cwa[1]	-0.24077	-0.24077	-0.24077	-0.24077	-0.24077
Climate_Cwb[1]	-0.67819	-0.67819	-0.67819	-0.67819	-0.67819
Soil_Light blue[2]	0.39477	0.39477	0.39477	0.39477	0.39477
Soil_Gray[2]	0.58819	0.58819	0.58819	0.58819	0.58819
Soil_Cream[2]	0.32019	0.32019	0.32019	0.32019	0.32019
Soil_Orange[2]	0.51483	0.51483	0.51483	0.51483	0.51483
Soil_Pink[2]	0.33911	0.33911	0.33911	0.33911	0.33911
Soil_Purple[2]	0.24583	0.24583	0.24583	0.24583	0.24583
Soil_Light green[2]	0.21559	0.21559	0.21559	0.21559	0.21559
Soil_Dark green[2]	0.43836	0.43836	0.43836	0.43836	0.43836

[1]In relation to the base level Af.

[2]In relation to the base level yellow.

For the global model fitted for CQ, results presented in Table 4.10 indicate that the most important variables were: animal category (coefficient estimates of 0.35, with P-value < 0.001, and 0.26, with P-value < 0.001, for the categories cow and steer, respectively, compared to bull), followed by climate (differences of up to -0.44, with P-value < 0.001, when compared to the baseline) and soil quality where animals were raised (differences of up to -0.14, with P-value < 0.001, when compared to the baseline). Technical consulting had significant but smaller effects, while nutrition sold to the farms had small effects and was only highly significant for one type of nutrition (FApc). The coefficients estimated for CW using GWR are presented in Table 4.11. Similar to the estimation of CW, AS and FD, all coefficients presented virtually no variation between different locations. The AIC for GWR (-12585.38) was slightly better than LR (-12558.36) and adjusted-R^2 presented the same value for both models (35.48%).

Table 4.10: Global model estimation of the outcome carcass quality (CQ) and respective coefficients for explanatory variables.

Variables	Coefficients	Standard Error	T value	Pr (>|t|)
Intercept	-0.009567	0.055790	-0.171	0.863854
Animal category_Steer	0.3525	0.004609	76.477	< 2e-16 ***
Animal category_Cow	0.2645	0.003603	73.418	< 2e-16 ***
Technical consulting_Yes	0.04187	0.003184	13.150	< 2e-16 ***
PNFpc	-0.000008	0.000003	-2.530	0.011412 *
FPpc	-0.000017	0.000063	-0.277	0.782084
FApc	0.000231	0.000031	7.410	1.31e-13 ***
Cattle sales price	0.002717	0.000128	21.207	< 2e-16 ***
Corn 3mo before price	0.003595	0.000179	20.104	< 2e-16 ***
Climate_Am[1]	-0.032380	0.008783	-3.687	0.000228 ***
Climate_As[1]	-0.443400	0.039450	-11.240	< 2e-16 ***
Climate_Aw[1]	0.011230	0.008872	1.265	0.205714
Climate_BSh[1]	0.145600	0.102500	1.421	0.155445
Climate_Cfa[1]	0.023310	0.010990	2.121	0.033950 *
Climate_Cfb[1]	-0.024360	0.072330	-0.337	0.736284
Climate_Cwa[1]	-0.043010	0.018390	-2.339	0.019347 *
Climate_Cwb[1]	-0.162300	0.029190	-5.560	2.74e-08 ***
Soil_Light blue[2]	0.108300	0.054470	1.988	0.046861 *
Soil_Gray[2]	0.142800	0.053050	2.692	0.007112 **
Soil_Cream[2]	0.1271e-01	0.053550	2.373	0.017635 *
Soil_Orange[2]	0.120900	0.053010	2.281	0.022583 *
Soil_Pink[2]	0.095610	0.054020	1.770	0.076787 .
Soil_Purple[2]	0.065640	0.054300	1.209	0.226722
Soil_Light green[2]	0.053170	0.053190	1.000	0.317530
Soil_Dark green[2]	0.096330	0.054290	1.774	0.076043 .

Significance (P-value) codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1]In relation to the base level Af.

[2]In relation to the base level yellow.

Table 4.11: Adaptive Gaussian Geographically Weighted model estimation of the outcome carcass quality (CQ) and respective coefficients for explanatory variables.

Variables	Minimum	Q1	Median	Q3	Maximum
Intercept	-0.009569	-0.009568	-0.009568	-0.009567	-0.009600
Animal category_Steer	0.352510	0.352510	0.352510	0.352510	0.352510
Animal category_Cow	0.264530	0.264530	0.264530	0.264530	0.264530
Technical consulting_Yes	0.041867	0.041867	0.041867	0.041867	0.041867
PNFpc	-0.000008	-0.000008	-0.000008	-0.000008	-0.000008
FPpc	-0.000017	-0.000017	-0.000017	-0.000017	-0.000017
FApc	0.000231	0.000231	0.000231	0.000231	0.000231
Cattle sales price	0.002717	0.002717	0.002717	0.002717	0.002717
Corn 3mo before price	0.003595	0.003595	0.003595	0.003595	0.003595
Climate_Am[1]	-0.032383	-0.032383	-0.032383	-0.032383	-0.032383
Climate_As[1]	-0.443450	-0.443450	-0.443450	-0.443450	-0.443450
Climate_Aw[1]	0.011227	0.011227	0.011227	0.011227	0.011227
Climate_BSh[1]	0.145600	0.145600	0.145600	0.145600	0.145600
Climate_Cfa[1]	0.023307	0.023307	0.023307	0.023307	0.023307
Climate_Cfb[1]	-0.024359	-0.024359	-0.024359	-0.024359	-0.024359
Climate_Cwa[1]	-0.043014	-0.043013	-0.043013	-0.043013	-0.043013
Climate_Cwb[1]	-0.162300	-0.162300	-0.162300	-0.162300	-0.162300
Soil_Light blue[2]	0.108260	0.108260	0.108260	0.108260	0.108260
Soil_Gray[2]	0.142810	0.142810	0.142810	0.142810	0.142810
Soil_Cream[2]	0.127110	0.127110	0.127110	0.127111	0.127111
Soil_Orange[2]	0.120890	0.120890	0.120890	0.120890	0.120900
Soil_Pink[2]	0.095605	0.095606	0.095607	0.095608	0.095608
Soil_Purple[2]	0.065637	0.065639	0.065640	0.065641	0.065643
Soil_Light green[2]	0.053168	0.053168	0.053168	0.053168	0.053168
Soil_Dark green[2]	0.096327	0.096328	0.096328	0.096328	0.096329

[1]In relation to the base level Af.

[2]In relation to the base level yellow.

Table 4.12: Comparison between linear regression (LR) and geographically weighted regression (GWR) in terms of model fit.

Model	AIC	Adjusted-R^2
	CW	
LR	173204.20	76.99%
GWR	173176.20	76.98%
	AS	
LR	140561.10	09.31%
GWR	140561.10	09.31%
	FD	
LR	45596.68	35.72%
GWR	45568.68	35.72%
	CQ	
LR	-12558.36	35.48%
GWR	-12585.38	35.48%

4.5 Discussion

For all outcome variables studied in this analysis to evaluate meat production and quality in Brazil (CW, AS, FD and CQ) there was very little (and in some cases virtually non-existent) variation between model coefficients estimated using LR and GWR. This indicates that the influence of the explanatory variables tested in this study on meat production and quality outcomes does not vary according to the farm geographical location. In terms of model fit (AIC and adjusted-R^2), LR and GWR presented similar or identical results. More specifically, (as presented in Table 4.12), CW, AS and FD had exactly the same AIC result (173204.20, 140561.10, and 45568.68, respectively) and CQ had slightly better AIC results for GWR (-12585.38), when compared to LR (-12585.36). Regarding adjusted-R^2, AS, FD and CQ results were exactly the same between

models (9.31%, 35.72% and 35.48%, respectively) while for CW results for LR (76.99%) were slightly better than GWR (76.98%).

It should be noted that some explanatory variables included in all models were already intrinsically accounting for factors that are spatially heterogeneous. The climate and soil type of the farms where animals were raised were determined based on the geographical location of the farm. These two explanatory variables, at least to a certain extent, account for the environmental conditions in which animals were raised. Regarding climate, it is long known that temperature and humidity of the environment exerts influence in the body temperature regulation of beef cattle, with environmental temperature and humidity within the optimum range being responsible for increasing productivity (Finch 1986). Finch (1986) highlight that even small increases in body temperature (heat stress) can be harmful to metabolic process that affect production and quality. Extremely hot climates in Brazil likely exert effects that exceed the ability of the animal to regulate body temperature, being prejudicial to their performance. Furthermore, soil fertility is tightly related to the agricultural potential and the quality and nutrients of the pasture, available to beef cattle. As mentioned by Porto et al. (2012), pasture is the main ingredient of diets of Brazilian beef cattle at least for part of their lives (before being finished on feedlots) and the success of the production system depends on the adequacy of the diet to the nutritional needs of the animal (Millen et al., 2011).

To confirm this hypothesis that soil and climate naturally account for spatial heterogeneity, a reduced "global" and GWR models, excluding the variables soil and climate class were run (results presented in Table 4.A.1 and 4.A.2 of Appendix section). In the absence of these explanatory variables, the variability across coefficients estimated for different regions increased

considerably and the GWR reduced model outperformed the reduced "global model", in terms of model fit for all variables studied. Specifically, GWR results for CW were AIC = 173643.2 and adjusted-R^2 = 76.43%; AS were AIC = 141151.9 and adjusted-R^2 = 06.46%; for FD were AIC = 46011.76 and adjusted-R^2 = 34.25%; and lastly for CQ were AIC = -11849.13 and adjusted-R^2 = 33.01%, vs LR results for CW that were AIC = 173661.9 and adjusted-R^2 = 76.42%; for AS were AIC = 141154.7 and adjusted-R^2 = 06.45%; for FD were AIC = 46026.04 and adjusted-R^2 = 34.28%; and lastly for CQ were AIC = -11833.34 and adjusted-R^2 = 33.00%. It is evident, however, that when comparing the full global and reduced GWR model, the former model provides better fit (lower AIC and higher adjusted-R^2) for all production and quality variables studied. This indicates that the inclusion of variables that naturally account for spatial heterogeneity (such as climate and soil), when they are available, is a more powerful approach when compared to solely accounting for farm location.

As noted previously, after accounting for spatial heterogeneity, little variation was observed in the coefficients of the model. This indicates that, after accounting for these factors, physiological mechanisms that determine meat production and quality remain are consistent across different locations of Brazil. Therefore, the large variation in outcomes for different states of Brazil (presented in Table 4.2) might not stand from external factors exerting effects with different magnitude, but from the availability of the explanatory variables being different across geographical location. For example, the variable AS, which was both an outcome and explanatory variable, showed expressive variation across states. When used as an explanatory variable for CW and FD, that variation could explain the differences observed in the outcome across different regions.

Lastly, it was noticed that while the proportion of the variance in the outcome explained by the model (adjusted-R^2) was fairly high for CW (76.9%), it was much lower for FD (35.7%) and CQ (35.5%), and fairly low for AS (9.3%). This could be partially due to the "loss of information" in the data compression step implemented. Alternatively, it might indicate that there are other variables in the system that are important for explaining the variance of these outcomes and which were not accounted for in this model (i.e. variables that were not available). Regarding the data compression, due to computer memory constrains, it was the only alternative to make this analysis viable as most programming environments (such as R) do not allow allocating a matrix of 816K by 816K (R limits on each dimension of an array = 2^{31}), that would be necessary to run the analysis with the whole data. Variables not accounted for in the current analysis that are potentially important predictors to be evaluated are animal breed composition, genetic merit and health indicators. Such variables could potentially improve the quality of the analysis in future studies as they may significantly affect beef production outcomes. We recommend that keeping track of such information would be a good practice for the beef industry.

In this analysis, large scale integrated data from different sources was utilized to test the hypothesis that effects that affect meat production and quality vary across different regions of Brazil by contrasting results of LR and GWR models. Results indicate that when inferring meat quality and production in Brazil, including climate and soil of the farms where animals were raised naturally captures spatial heterogeneity. After accounting for those variables, animal physiology seems to be consistent across regions and very little spatial heterogeneity is captured by GWR models, presenting no advantage when compared to traditional LR.

4.6 References

Albuquerque, P. H. M., F. A. S. Medina, and A. R. da Silva. 2016. “Geographically Weighted Logistic Regression applied to credit scoring models.” R. Cont. Fin. 28(73): 93–112.

Bello, N. M., and D. G. Renter. 2018. “Reproducible research from noisy data : revisiting key statistical principles for the animal sciences.” J. Dairy Sci. 101(7): 5679–5701.

Binbin, L., P. Harris, M. Charlton, C. Brunsdon, T. Nakaya, D. Murakami and I. Gollini. 2019. “GWmodel: an R package for exploring spatial heterogeneity.” R package version 2.1-3.

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton. 1996. “Geographically Weighted Regression: a method for exploring spatial nonstationarity.” Geogr. Anal. 28(4): 281–98.

“FAO.” 2014. “Meat consumption” http://www.fao.org/ag/againfo/themes/en/meat/background.html (November, 2, 2019).

Finch, V. A. 1986. “Body temperature in beef cattle: its control and relevance to production in the tropics.” J. Anim. Sci. 62(2): 531–42.

IBGE. 2018. “Indicadores IBGE estatística da produção pecuária.”: 37–40. ftp://ftp.ibge.gov.br/Producao_Pecuaria/Fasciculo_Indicadores_IBGE/abate-leite-couro-ovos_201802caderno.pdf.

Mena, C., C. Sepúlveda, E. Fuentes, Y. Ormazábal and I. Palomo. 2018. “Spatial analysis for the epidemiological study of cardiovascular diseases: a systematic literature search.” Geospat. Health 13(1): 11-19. doi: 10.1081/gh.2018.587.

Millen, D. D., R. D. L. Pacheco, M. D. B. Arrigoni, M. L. Galyean, and J. T. Vasconcelos. 2009.

"A snapshot of management practices and nutritional recommendations used by feedlot nutritionists in Brazil." J. Anim. Sci. 87(10): 3427–39. doi: 10.2527/jas.2009-1880.

Millen, D. D., R. D. L Pacheco, P. M. Meyer, P. H. M Rodrigues, and M. D. B. Arrigoni. 2011. "Current outlook and future perspectives of beef production in Brazil." Anim. Front. 1(2): 46–52. doi:10.2527/af.2011-0017.

Oliveira, M. 2018. "Contributions of Brazilian cattle." *Pesquisa FAPESP* (264). https://revistapesquisa.fapesp.br/en/2018/06/25/contributions-of-brazilian-cattle/.

Oliveira, M. 2019. "Produção da pecuária municipal 2018." *Catalog of the Instituto Brasileiro de Geografia e Estatística* 84(01014234): 1–8. https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=784.

Porto, M. O., et al. 2012. "Nutritional requirements of energy, protein and macrominerals for maintenance and weight gain of young crossbred Nellore × Holstein bulls on pasture." R. Bras. Zootec. 41(3): 734–45.

Stekhoven, D. J. 2013. "MissForest: nonparametric missing value imputation using random forest." R package version 1.4.

Zia, M., J. Hansen, K. Hjort, and C. Valdes. 2019. "Brazil once again becomes the world's largest beef exporter." *United States department of Agriculture - Economic Research Service*. https://www.ers.usda.gov/amber-waves/2019/july/brazil-once-again-becomes-the-world-s-largest-beef-exporter/ (October 21, 2019).

4.6 Appendix

Table 4.A1: Global model (without climate and soil) estimation of the outcomes carcass weight (CW), age when finished (AS), carcass fat deposition (FD), and carcass quality (CQ) with respective coefficients for explanatory variables.

Variables	Coefficients	SE	T value	Pr (>|t|)
	CW (AIC = 173661.9, Adjusted-R^2 = 76.42%)			
Intercept	246.258285	1.832269	134.401	<2e-16 ***
Age	0.424924	0.016331	26.019	<2e-16 ***
Animal category_Steer	-10.394702	0.508090	-20.458	<2e-16 ***
Animal category_Cow	-95.076178	0.416707	-228.161	<2e-16 ***
Technical consulting_Yes	6.035158	0.360814	16.727	<2e-16 ***
PNFpc	0.005288	0.000401	13.187	<2e-16 ***
FPpc	0.014593	0.007202	2.026	0.0428 *
FApc	0.006497	0.003567	1.822	0.0685 .
Cattle sales price	0.233807	0.014514	16.109	<2e-16 ***
Corn 3mo before price	-0.167700	0.020189	-8.306	<2e-16 ***
	AS (AIC = 141154.7, Adjusted-R^2 = 06.45%)			
Intercept	41.7829840	0.7432101	56.220	< 2e-16 ***
Animal category_Steer	3.5463950	0.2206198	16.075	< 2e-16 ***
Animal category_Cow	5.4426048	0.1779355	30.588	< 2e-16 ***
Technical consulting_Yes	-1.3273351	0.1574133	-8.432	< 2e-16 ***
PNFpc	0.0020095	0.0001747	11.505	< 2e-16 ***
FPpc	-0.0131758	0.0031465	-4.187	2.83e-05 ***
FApc	-0.0064553	0.0015583	-4.143	3.45e-05 ***
Cattle sales price	-0.0583653	0.0063298	-9.221	< 2e-16 ***
Corn 3mo before price	0.0288426	0.0088216	3.270	0.00108 **
	FD (AIC = 46026.04, Adjusted-R^2 = 34.28%)			
Intercept	1.246000	0.071050	17.541	< 2e-16 ***
Age	-0.005371	0.000633	-8.480	< 2e-16 ***
Animal category_Steer	1.365000	0.01970	69.274	< 2e-16 ***
Animal category_Cow	1.53400	0.016160	94.900	< 2e-16 ***
Technical consulting_Yes	0.22160	0.01399	15.838	< 2e-16 ***
PNFpc	0.000046	0.000015	2.967	0.00301 **
FPpc	0.000092	0.000279	0.330	0.74169
FApc	0.000907	0.000138	6.557	5.64e-11 ***
Cattle sales price	0.004294	7.629	0.000563	2.48e-14 ***
Corn 3mo before price	0.001233	0.000783	1.575	0.11533
	CQ (AIC = -11833.34, Adjusted-R^2 = 33.00%)			
Intercept	0.102400	0.015110	6.779	1.25e-11 ***
Animal category_Steer	0.355600	0.004486	79.257	< 2e-16 ***
Animal category_Cow	0.266600	0.003618	73.690	< 2e-16 ***
Technical consulting_Yes	0.051090	0.003201	15.960	< 2e-16 ***
PNFpc	-0.000063	0.000035	-1.838	0.0661 .
FPpc	-0.000086	0.000064	-1.342	0.1795
FApc	0.000223	0.000032	7.021	2.27e-12 ***
Cattle sales price	0.002703	0.000129	21.002	< 2e-16 ***
Corn 3mo before price	0.003363	0.000179	18.745	< 2e-16 ***

Significance (P-value) codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4.A2: Adaptive Gaussian Geographically Weighted model (without climate and soil) estimation of the outcomes carcass weight (CW), age when finished (AS), carcass fat deposition (FD), and carcass quality (CQ) with respective coefficient summary for explanatory variables.

Variables	Minimum	Q1	Median	Q3	Maximum
	CW (AIC = 173643.2, Adjusted-R^2 = 76.43%)				
Intercept	246.081283	246.183099	246.222959	246.292923	246.399999
Age	0.423288	0.424552	0.425486	0.426105	0.427400
Animal category_Steer	-10.413557	-10.408341	-10.404850	-10.400694	-10.396000
Animal category_Cow	-95.093816	-95.089981	-95.086088	-95.082362	-95.076200
Technical consulting_Yes	6.019106	6.031608	6.033749	6.038532	6.049300
PNFpc	0.005275	0.005286	0.005288	0.005292	0.005300
FPpc	0.014330	0.014610	0.014684	0.014753	0.0150010
FApc	0.006457	0.006482	0.006494	0.006498	0.006500
Cattle sales price	0.233007	0.233647	0.234011	0.234197	0.234800
Corn 3mo before price	-0.167870	-0.167651	-0.167596	-0.167547	-0.167400
	AS (AIC = 141151.9, Adjusted-R^2 = 06.46%)				
Intercept	41.765134	41.775979	41.779331	41.782929	41.796700
Animal category_Steer	3.543163	3.547589	3.551079	3.553639	3.558200
Animal category_Cow	5.434279	5.441878	5.448313	5.453380	5.461500
Technical consulting_Yes	-1.331617	-1.328151	-1.326271	-1.325389	-1.321900
PNFpc	0.002009	0.002010	0.002011	0.002011	0.002012
FPpc	-0.013201	-0.013192	-0.013189	-0.013185	-0.013200
FApc	-0.006490	-0.006470	-0.006460	-0.006450	-0.006440
Cattle sales price	-0.058690	-0.058440	-0.058410	-0.058310	-0.058100
Corn 3mo before price	0.028453	0.028762	0.028882	0.028959	0.029300
	FD (AIC = 46011.76, Adjusted-R^2 = 34.25%)				
Intercept	1.241400	1.244000	1.244800	1.246400	1.249100
Age	-0.005383	-0.005369	-0.005365	-0.005361	-0.005400
Animal category_Steer	1.364300	1.364500	1.364600	1.364700	1.364900
Animal category_Cow	1.533300	1.533400	1.533500	1.533600	1.533700
Technical consulting_Yes	0.221220	0.221560	0.221620	0.221750	0.222000
PNFpc	0.000088	0.000093	0.000095	0.000096	0.000100
FPpc	0.000046	0.000046	0.000046	0.000046	0.000046
FApc	0.000906	0.000907	0.000907	0.000907	0.000908
Cattle sales price	0.004260	0.004291	0.004307	0.004313	0.004315
Corn 3mo before price	0.001187	0.001228	0.001238	0.001248	0.001300
	CQ (AIC = -11849.13, Adjusted-R^2 = 33.01%)				
Intercept	0.101530	0.102080	0.102220	0.102520	0.103100
Animal category_Steer	0.355320	0.355420	0.355450	0.355500	0.355600
Animal category_Cow	0.266540	0.266560	0.266580	0.266600	0.266640
Technical consulting_Yes	0.050931	0.051049	0.051070	0.051124	0.051200
PNFpc	-0.000086	-0.000085	-0.000085	-0.000085	-0.000090
FPpc	-0.000007	-0.000007	-0.000007	-0.000007	-0.000007
FApc	0.000222	0.000222	0.000222	0.000223	0.000224
Cattle sales price	0.002693	0.002702	0.002706	0.002708	0.002710
Corn 3mo before price	0.003346	0.003359	0.003363	0.003370	0.003400

Concluding remarks

This dissertation focuses on data analytics of large scale integrated beef cattle data from Brazil. Specifically, it (1) provided a framework for data integration, developing the concept of "farm-matching" and comparing efficiency of different deterministic, stochastic and machine learning approaches for "farm-matching"; (2) evaluated forecasting models for beef production and quality at the national level using a unique integrated large data collection, including farm, market and environmental factors; and (3) investigated how factors affecting meat production and quality vary across different regions of Brazil.

Effectively integrating data from different sources is one of the first and most crucial steps for subsequent data analytics. In Chapter 2, we extended concepts of record-linkage to agriculture, creating a framework for "farm-matching". When comparing different methods to perform farm-matching, the best results (in terms of accuracy, precision) were obtained with the machine learning methods Support Vector Machines and Bagged Clustering. Both methods provided very high accuracy and precision, being capable of learning the behavior of human experts on which information is relevant to consider a specific farm in two data bases as a match (i.e. the same farm). High performance was achieved even in the presence of typos, misspelling, abbreviations and missing information on farms attributes (farm name, farmer name, etc.). Results of this study may be relevant to future applications in agriculture, especially because data integration is often a bottleneck that can hamper the utilization of data analytics in agriculture.

Forecasting of trends is central to prepare for the future, allocating resources, and optimizing production systems. In Chapter 3, we compared statistical and machine learning methods for the task of forecasting meat production and quality at the national level, using large

scale integrated data from Brazil (with the method developed in Chapter 2). We showed that despite being a complex system, it is possible to forecast meat production and quality at the national level with moderate-high accuracy. Knowledge on which methods and variables are relevant for improved performance of such task is essential for future applications. In the current scenario, where agriculture is faced with the challenge of scaling up to fulfil demand in a sustainable manner and without increasing costs associated with production, forecasting will be central for optimizing the production chain, addressing the current challenge of agriculture.

Lastly, in Chapter 4, we explore the variability of meat production and quality across different regions of Brazil by contrasting linear regression (LR) and geographically weighted regression (GWR). Results indicate that after including climate and soil of the farms where beef animals were raised into the models, the GWR non-stationary method presented virtually no advantage when compared to LR. This indicates that climate and soil are effective on naturally capturing the spatial variability of different Brazilian beef production systems and after accounting for these factors, animal physiology seems consistent across different locations.

We argue that data analytics of beef production and quality could be enhanced if more variables (and even the ones already collected) are recorded in a systematic manner by different players in the production chain. We advise that this practice is implemented in future applications. Additionally, it should be noted that the data analytics tools presented in this dissertation are not restricted to beef cattle production and could easily be extended to other livestock species (such as dairy cattle, pigs, or poultry). Lastly, this dissertation accomplished the goal to show that there is value in extracting information from complex, messy, large data collections in animal sciences.

We suggest further implementations of data analytics in the field, as they can prove invaluable for addressing challenges lying ahead in animal agriculture.