UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

PARAMETRIC AND SEMI-PARAMETRIC MODELS FOR PREDICTING GENOMIC BREEDING VALUES OF COMPLEX TRAITS IN NELORE CATTLE

Rafael Espigolan

Zootecnista

2017

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

PARAMETRIC AND SEMI-PARAMETRIC MODELS FOR PREDICTING GENOMIC BREEDING VALUES OF COMPLEX TRAITS IN NELORE CATTLE

Rafael Espigolan

Orientadora: Profa. Dra. Lucia Galvão de Albuquerque

Coorientador: Dr. Daniel Gustavo Mansan Gordo

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Genética e Melhoramento Animal

2017

Espigolan, Rafael
E77p Parametric and semi-parametric models for predicting genomic breeding values of complex traits in Nelore cattle / Rafael Espigolan. – – Jaboticabal, 2017
vi, 63 p. : il. ; 29 cm

Tese (doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2017
Orientadora: Lucia Galvão de Albuquerque
Coorientador: Daniel Gustavo Mansan Gordo
Banca examinadora: Lenira El Faro Zadra, Ricardo Vieira Ventura, Danísio Prado Munari, Gerardo Alves Fernandes Júnior
Bibliografia

1. Acurácia. 2. Seleção Genômica. 3. Regressão RKHS. 4. SNP. 5. Características de carcaça. I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 636.2:636.082

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação – Serviço Técnico de Biblioteca e Documentação - UNESP, Câmpus de Jaboticabal.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO: PARAMETRIC AND SEMI-PARAMETRIC MODELS FOR PREDICTING GENOMIC BREEDING VALUES OF COMPLEX TRAITS IN NELORE CATTLE

AUTOR: RAFAEL ESPIGOLAN
ORIENTADORA: LUCIA GALVÃO DE ALBUQUERQUE
COORIENTADOR: DANIEL GUSTAVO MANSAN GORDO

Aprovado como parte das exigências para obtenção do Título de Doutor em GENÉTICA E MELHORAMENTO ANIMAL, pela Comissão Examinadora:

Pós-doutorando DANIEL GUSTAVO MANSAN GORDO
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Pesquisadora Dra. LENIRA EL FARO ZADRA
APTA / Ribeirão Preto, SP

Prof. Dr. RICARDO VIEIRA VENTURA
Universidade de São Paulo / Pirassununga/SP

Prof. Dr. DANISIO PRADO MUNARI
Departamento de Ciências Exatas / FCAV / UNESP - Jaboticabal

Pós-doutorando GERARDO ALVES FERNANDES JÚNIOR
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Jaboticabal, 23 de fevereiro de 2017.

DADOS CURRICULARES DO AUTOR

Rafael Espigolan, nascido em Orlândia, estado de São Paulo, no dia 29 de maio de 1988, filho de Alcides Espigolan e Maria Dolores Danielli Espigolan. Zootecnista formado pela Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, Câmpus de Jaboticabal, obtendo seu título em 2010. Durante a graduação, foi bolsista da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), sob orientação da Profa. Dra. Lucia Galvão de Albuquerque. Em março de 2011, ingressou no curso do Programa de Pós-Graduação em Genética e Melhoramento Animal da mesma faculdade, inicialmente como bolsista da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e, posteriormente, FAPESP, sob orientação da Profa. Dra. Lucia Galvão de Albuquerque, obtendo o título de mestre em 27 de fevereiro de 2013. Em março de 2013, ingressou no curso de doutorado no mesmo Programa de Pós-Graduação sob a orientação da Profa. Dra. Lucia Galvão de Albuquerque, inicialmente como bolsista CAPES e, posteriormente, FAPESP. Em setembro de 2015, iniciou o estágio sanduíche no exterior por meio da Bolsa Estágio de Pesquisa no Exterior (BEPE) fornecida pela FAPESP, permanecendo nove meses no *Instituto Nacional de Investigación Agropecuaria – INIA Las Brujas* localizado em Canelones, Uruguai, sob supervisão do PhD. Ignacio Aguilar. Obteve o grau de doutor em 23 de fevereiro de 2017, sob orientação da Profa. Dra. Lucia Galvão de Albuquerque e coorientação do Dr. Daniel Gustavo Mansan Gordo.

Love hides in the strangest places.

Love hides in familiar faces.

Love comes when you least expect it.

Love hides in narrow corners.

Love comes for those who seek it.

Love hides inside the rainbow.

Love hides in molecular structures.

Love is the answer.

The Doors

À minha alma gêmea, Bárbara Antoniassi

e nossa preciosa filha Raffaella

pelos incontáveis momentos de amor e alegria.

Dedico.

Aos meus queridos pais,

Alcides e Maria Dolores.

Ofereço.

AGRADECIMENTOS

À minha orientadora, Profa. Dra. Lucia Galvão de Albuquerque, pela oportunidade, apoio, confiança e ensinamentos que contribuíram muito para o desenvolvimento deste trabalho e para meu amadurecimento intelectual.

Ao meu coorientador, Dr. Daniel Gustavo Mansan Gordo pela paciência e amizade, além da constante disposição em ajudar e auxiliar no meu desenvolvimento acadêmico.

Aos professores Dr. Fernando Baldi e Dr. Henrique Nunes de Oliveira pelas valiosas sugestões no Exame Geral de Qualificação e pelo auxílio e amizade desde a Iniciação Científica.

Aos professores Dr. Danísio Prado Munari (meu estimado conterrâneo), Dr. Ricardo Vieira Ventura, Dra. Lenira El Faro Zadra e ao pós-doutorando Gerardo Alves Fernandes Júnior pelas sugestões que melhoraram muito a qualidade da Tese final.

Aos meus queridos amigos Ana Fabrícia, Ana Cristina, Anderson, Andrés, André, Bianca, Bresolin, Camila, Costa, Daiane, Dani Jovino, Dani Beraldo, Diego, Diércles, Diogo, Elisa, Fabielli, Inaê, Gabriela, Gerardo, Giovana, Gordo, Guilherme Leão, Henrique Mandí, Hermenegildo, Laiza, Larissa, Lucas, Luciana, Lúcio, Malane, Marcos Lemos, Mariana Berton, Medeiros, Tonussi, Samuel, Thaise, Venturini e William pelo carinho e amizade durante esses quatro anos do Doutorado, que foram inesquecíveis. A todos os outros colegas e companheiros da “salinha” e da faculdade por fazerem parte de minha formação e pelo apoio.

Ao pesquisador PhD. Ignacio Aguilar, do *Instituto Nacional de Investigación Agropecuaria – INIA Las Brujas,* Uruguai, por me receber no período do Doutorado Sanduíche, pelas conversas e ensinamentos transmitidos, que fizeram com que os anos de 2015 e 2016 fossem um marco na minha vida profissional e pessoal.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos no início do curso de Doutorado.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pela concessão das bolsas de estudo no país (Processo FAPESP Nº 2014/00779-0) e no exterior (Processo FAPESP Nº 2015/13084-3), que permitiram a realização desse Projeto, do Doutorado Sanduíche e a participação em cursos e congressos, o que contribuiu imensamente para minha formação profissional.

SUMÁRIO

Página

MODELOS ESTATÍSTICOS PARAMÉTRICOS E SEMIPARAMÉTRICOS PARA A PREDIÇÃO DE VALORES GENÉTICOS GENÔMICOS DE CARACTERÍSTICAS COMPLEXAS EM BOVINOS DA RAÇA NELORE

RESUMO - O melhoramento genético animal visa melhorar a produtividade econômica das futuras gerações de espécies domésticas por meio da seleção. A maioria das características de interesse econômico na pecuária é de expressão quantitativa e complexa, isto é, são influenciadas por vários genes e afetadas por fatores ambientais. As análises estatísticas de informações de fenótipo e pedigree permite estimar os valores genéticos dos candidatos à seleção com base no modelo infinitesimal. Uma grande quantidade de dados genômicos está atualmente disponível para a identificação e seleção de indivíduos geneticamente superiores com o potencial de aumentar a acurácia de predição dos valores genéticos e, portanto, a eficiência dos programas de melhoramento genético animal. Vários estudos têm sido conduzidos com o objetivo de identificar metodologias apropriadas para raças e características específicas, o que resultará em estimativas de valores genéticos genômicos (GEBVs) mais acurados. Portanto, o objetivo deste estudo foi verificar a possibilidade de aplicação de modelos semiparamétricos para a seleção genômica e comparar a habilidade de predição com os modelos paramétricos para dados reais (características de carcaça, qualidade da carne, crescimento e reprodutiva) e simulados. As informações fenotípicas e de pedigree utilizadas foram fornecidas por onze fazendas pertencentes a quatro programas de melhoramento genético animal. Para as características de carcaça e qualidade da carne, o banco de dados continha 3.643 registros para área de olho de lombo (REA), 3.619 registros para espessura de gordura (BFT), 3.670 registros para maciez da carne (TEN) e 3.378 observações para peso de carcaça quente (HCW). Um total de 825.364 registros para peso ao sobreano (YW) e 166.398 para idade ao primeiro parto (AFC) foi utilizado para as características de crescimento e reprodutiva. Genótipos de 2.710, 2.656, 2.749, 2.495, 4.455 e 1.760 animais para REA, BFT, TEN, HCW, YW e AFC foram disponibilizados, respectivamente. Após o controle de qualidade, restaram dados de, aproximadamente, 450.000 polimorfismos de base única (SNP). Os modelos de análise utilizados foram BLUP genômico (GBLUP), single-step GBLUP (ssGBLUP), Bayesian LASSO (BL) e as abordagens semiparamétricas Reproducing Kernel Hilbert Spaces (RKHS) e Kernel Averaging (KA). Para cada característica foi realizada uma validação cruzada composta por cinco "folds" e replicada aleatoriamente trinta vezes. Os modelos estatísticos foram comparados em termos do erro do quadrado médio (MSE) e acurácia de predição (ACC). Os valores de ACC variaram de 0,39 a 0,40 (REA), 0,38 a 0,41 (BFT), 0,23 a 0,28 (TEN), 0,33 a 0,35 (HCW), 0,36 a 0,51 (YW) e 0,49 a 0,56 (AFC). Para todas as características, os modelos GBLUP e BL apresentaram acurácias de predição similares. Para REA, BFT e HCW, todos os modelos apresentaram ACC similares, entretanto a regressão RKHS obteve o melhor ajuste comparado ao KA. Para características com maior quantidade de registros fenotípicos comparada ao número de animais genotipados

(YW e AFC) o modelo ssGBLUP é indicado. Considerando o desempenho geral, para todas as características estudadas, a regressão RKHS é, particularmente, uma alternativa interessante para a aplicação na seleção genômica, especialmente para características de baixa herdabilidade. No estudo de simulação, genótipos, pedigree e fenótipos para quatro características (A, B, C e D) foram simulados utilizando valores de herdabilidade baseados nos obtidos com os dados reais (0,09, 0,12, 0,36 e 0,39 para cada característica, respectivamente). O genoma simulado consistiu de 735.293 marcadores e 1.000 QTLs distribuídos aleatoriamente por 29 pares de autossomos, com comprimento variando de 40 a 146 centimorgans (cM), totalizando 2.333 cM. Assumiu-se que os QTLs explicavam 100% da variação genética. Considerando as frequências do alelo menor maiores ou iguais a 0,01, um total de 430.000 marcadores foram selecionados aleatoriamente. Os fenótipos foram obtidos pela soma dos resíduos (aleatoriamente amostrados de uma distribuição normal com média igual a zero) aos valores genéticos verdadeiros, e todo o processo de simulação foi replicado 10 vezes. A ACC foi calculada por meio da correlação entre o valor genético genômico estimado e o valor genético verdadeiro, simulados da 12ª a 15ª geração. A média do desequilíbrio de ligação, medido entre os pares de marcadores adjacentes para todas as características simuladas foi de 0,21 para as gerações recentes (12ª, 13ª e 14ª), e 0,22 para a 15ª geração. A ACC para as características simuladas A, B, C e D variou de 0,43 a 0,44, 0,47 a 0,48, 0,80 a 0,82 e 0,72 a 0,73, respectivamente. Diferentes metodologias de seleção genômica implementadas neste estudo mostraram valores similares de acurácia de predição, e o método mais adequado é dependente da característica explorada. Em geral, as regressões RKHS obtiveram melhor desempenho em termos de ACC com menor valor de MSE em comparação com os outros modelos.

Palavras-chave: acurácia, características de carcaça, regressão RKHS, seleção genômica, SNP

PARAMETRIC AND SEMI-PARAMETRIC MODELS FOR PREDICTING GENOMIC BREEDING VALUES OF COMPLEX TRAITS IN NELORE CATTLE

ABSTRACT - Animal breeding aims to improve economic productivity of future generations of domestic species through selection. Most of the traits of economic interest in livestock have a complex and quantitative expression i.e. are influenced by a large number of genes and affected by environmental factors. Statistical analysis of phenotypes and pedigree information allows estimating the breeding values of the selection candidates based on infinitesimal model. A large amount of genomic data is now available for the identification and selection of genetically superior individuals with the potential to increase the accuracy of prediction of genetic values and thus, the efficiency of animal breeding programs. Numerous studies have been conducted in order to identify appropriate methodologies to specific breeds and traits, which will result in more accurate genomic estimated breeding values (GEBVs). Therefore, the objective of this study was to verify the possibility of applying semi-parametric models for genomic selection and to compare their ability of prediction with those of parametric models for real (carcass, meat quality, growth and reproductive traits) and simulated data. The phenotypic and pedigree information used were provided by farms belonging to four animal breeding programs which represent eleven farms. For carcass and meat quality traits, the data set contained 3,643 records for rib eye area (REA), 3,619 records for backfat thickness (BFT), 3,670 records for meat tenderness (TEN) and 3,378 observations for hot carcass weight (HCW). A total of 825,364 records for yearling weight (YW) and 166,398 for age at first calving (AFC) were used as growth and reproductive traits of Nelore cattle. Genotypes of 2,710, 2,656, 2,749, 2,495, 4,455 and 1,760 animals were available for REA, BFT, TEN, HCW, YW and AFC, respectively. After quality control, approximately 450,000 single nucleotide polymorphisms (SNP) remained. Methods of analysis were genomic BLUP (GBLUP), single-step GBLUP (ssGBLUP), Bayesian LASSO (BL) and the semi-parametric approaches Reproducing Kernel Hilbert Spaces (RKHS) regression and Kernel Averaging (KA). A five-fold cross-validation with thirty random replicates was carried out and models were compared in terms of their prediction mean squared error (MSE) and accuracy of prediction (ACC). The ACC ranged from 0.39 to 0.40 (REA), 0.38 to 0.41 (BFT), 0.23 to 0.28 (TEN), 0.33 to 0.35 (HCW), 0.36 to 0.51 (YW) and 0.49 to 0.56 (AFC). For all traits, the GBLUP and BL models showed very similar prediction accuracies. For REA, BFT and HCW, models provided similar prediction accuracies, however RKHS regression had the best fit across traits considering multiple-step models and compared to KA. For traits which have a higher number of animals with phenotypes compared to the number of those with genotypes (YW and AFC), the ssGBLUP is indicated. Judged by overall performance, across all traits, the RKHS regression is particularly appealing for application in genomic selection, especially for low heritability traits. Simulated genotypes, pedigree, and phenotypes for four traits A, B, C and D were obtained using heritabilities based on real data (0.09, 0.12, 0.36 and 0.39 for each trait, respectively). The simulated genome

consisted of 735,293 markers and 1,000 QTLs randomly distributed over 29 pairs of autosomes, with length varying from 40 to 146 centimorgans (cM), totaling 2,333 cM. It was assumed that QTLs explained 100% of genetic variance. Considering Minor Allele Frequencies greater or equal to 0.01, a total of 430,000 markers were randomly selected. The phenotypes were generated by adding residuals, randomly drawn from a normal distribution with mean equal to zero, to the true breeding values and all simulation process was replicated 10 times. ACC was quantified using correlations between the predicted genomic breeding value and true breeding values simulated for the generations of 12 to 15. The average linkage disequilibrium, measured between pairs of adjacent markers for all simulated traits was 0.21 for recent generations (12, 13 and 14), and 0.22 for generation 15. The ACC for simulated traits A, B, C and D ranged from 0.43 to 0.44, 0.47 to 0.48, 0.80 to 0.82 and 0.72 to 0.73, respectively. Different genomic selection methodologies implemented in this study showed similar accuracies of prediction, and the optimal method was sometimes trait dependent. In general, RKHS regressions were preferable in terms of ACC and provided smallest MSE estimates compared to other models.

Keywords: accuracy, carcass traits, RKHS regression, genomic selection, SNP

CHAPTER 1 – General Considerations

1. INTRODUCTION

The beef cattle industry stands out in the Brazilian agribusiness, since the country has the world's second largest cattle herd with over 213 million heads (MAPA, 2015). A portion of 80% of the Brazilian herd is composed of Zebu breeds (*Bos indicus*), which are animals of proven rusticity and adaptation to the predominant environment in Brazil and among these breeds, we can highlight the Nelore, with 90% of this portion (ABIEC, 2016).

In face of growing world demand for food and animal protein, the development and implementation of new technologies is necessary in beef cattle production to meet consumer market standards. In order to increase productivity in beef cattle, one of the alternatives is the use of selection, which changes allele frequencies by choosing, as parents, genetically superior animals for certain traits of economic interest.

One of the most efficient tools for identifying genetically superior animals is selection, which is based on genetic evaluations. The genetic evaluations and selection of animals are performed in all animal breeding programs aiming at greater production efficiency. Animals with higher growth rates or sexual precocity directly implicate in the shortening of the production cycle, allowing greater economic return to the breeder (BOLIGON et al., 2009).

Nowadays, a tool used to improve selection of animals is the use of single nucleotide polymorphisms (SNP) markers, widely distributed throughout the genome, assuming that genetic markers are in linkage disequilibrium with quantitative trait loci (QTL). SNPs are suitable for genotyping of animals and can be used in genetic evaluations, facilitating the identification of superior animals and selecting candidates for a particular trait, even if there is no pedigree information (CLARKE et al. 2014).

Genomic selection refers to the use of genome wide dense marker genotypes for breeding value estimation and subsequently for selection. The genomic

predictions can be obtained by estimating the effects of thousands of single nucleotide polymorphisms (SNP) markers spread throughout the genome. However, the main challenge is the estimation of many effects from a limited number of observations (BENNEWITZ; SOLBERG; MEUWISSEN, 2009).

To deal with this problem, Meuwissen, Hayes and Goddard (2001) proposed Bayesian methods that use informative priors and the authors showed, through simulations, that these parametric methods are able to estimate genomic breeding values with remarkably high accuracy, even for individuals without phenotypic records.

Under a parametric approach, several analytical methods have been proposed for genome-based prediction of genetic values, and these differ with respect to assumptions about the marker effects (MEUWISSEN; HAYES; GODDARD, 2001; HABIER et al. 2011). For example, Genomic Best Linear Unbiased Predictor (GBLUP), Single-Step Genomic Best Linear Unbiased Predictor (ssGBLUP) and Bayesian models.

According to VanRaden (2008), GBLUP estimates breeding values using a matrix of genomic relationships (**G** matrix) instead of pedigree information. In this model, only phenotypes from genotyped animals are included in the analysis. On the other hand, Mizstal, Legarra and Aguilar (2009) proposed a procedure known as ssGBLUP (Single-Step Genomic Best Linear Unbiased Predictor) in which the matrices A and G are combined in a matrix H. In this case all the phenotypic information, from both, genotyped and non-genotyped animals, is used to predict the genomic values.

Although many works are focused on the use of parametric models, the assumptions for these approaches, such as normality, linearity and independent explanatory variables, do not always hold. In this way, Gianola et al. (2006) proposed a semi-parametric method called Reproducing Kernel Hilbert Spaces (RKHS) regression to model the relationship between the phenotype and the markers, capable of accounting for complex epistatic models without explicitly modeling then.

The RKHS regression has been used in many areas of application, such as scatter-plot smoothing (WAHBA, 1990), spatial statistics (CRESSIE, 1993), and classification problems (VAPNIK, 1998). However, there are few articles dealing and evaluating the use of RKHS regressions for genomic-enable prediction of genetic values in animal breeding considering complex traits.

The main difference between a standard parametric mixed model and a semi-parametric regression model is the attempt to capture unknown forms of interaction among many loci that, arguably, parametric models are not able to explore properly (GIANOLA et al., 2006;. GIANOLA; VAN KAAM, 2008;. DE LOS CAMPOS; GIANOLA; ROSA, 2009).

According to Gianola et al. (2006) and Gianola and van Kaam (2008), the RKHS regression have been suggested as an alternative to multiple linear regression for capturing complex interaction patterns that may be difficult to account for in a linear model framework. In this approach, the information of the markers is included in a positive definite matrix, called kernel, which is used as incidence matrix in the regression model. Still, this semi-parametric approach has several attractive features, such: the methodology can be used with almost any type of information set, which is particularly important because techniques for characterizing genomes change rapidly; and the computations are performed in a n-dimensional space, which gives a great computational advantage to RKHS regression relative to some parametric models, especially when $p>n$ (DE LOS CAMPOS et al. 2010a)

Several studies have been developed to determine the predictive ability of different methods and statistical models with the inclusion of information provided by markers distributed throughout the genome (MUIR, 2007; GONZÁLEZ-RECIO, 2008; BENNEWITZ; SOLBERG; MEUWISSEN, 2009; SONESSON; MEUWISSEN, 2009; ; DE LOS CAMPOS et al. 2010a, LONG et al., 2010, WINKELMAN; JOHNSON; HARRIS, 2015). However, for Nelore beef cattle, there are only a few studies exploring models with semi-parametric approach with a set of real data and considering information provided by the SNP markers.

2. GENERAL OBJECTIVE

The objective of this study was to verify the possibility of applying semi-parametric models for genomic selection and to compare their ability of prediction with those of parametric models for carcass, meat quality, growth and reproductive traits.

2.1. Specific objectives

- To compare the accuracy of genomic predictions obtained through the parametric models GBLUP, ssGBLUP, Bayes LASSO and the semi-parametrics models RKHS regression and KA for a set of real data.
- To compare the accuracy of genomic predictions obtained using parametric models and the semi-parametrics models RKHS regression and KA for a set of simulated data.

3. LITERATURE REVIEW

3.1. Traits of economic interest

Selection of genetically superior animals for traits of economic interest has been one of the most important tools applied by breeding programs to increase meat production aiming to increase producers' profit. In beef cattle, the rib eye area is one of the most used regions to evaluate carcass quality in meat production.

According to Boggs and Merkel (1990), the rib eye area is an indicator of carcass composition and it is related to carcass muscularity. On the other hand, the thickness of fat cover is an important trait in the meat industry especially to protect the carcass after slaughter. An adequate quality carcass must have enough fat covering to guarantee its preservation and the quality for consuming (CUNDIFF et al., 1993). The backfat thickness becomes an important trait whereas the

conventional processing of cattle carcasses refrigeration after slaughter consists to down the temperature to around 7ºC, which may result in excessive contraction of the sarcomeres, resulting in tougher meat. According to Luchiari Filho (2000), the backfat thickness has been used as an efficient indicator of carcass finishing, while hot carcass weight is used as a classification and typification trait in the slaughterhouses, being directly related to the producers' payment. However, it is important consider that the excessive fat increases the cost of meat for the consumer and requires more cleaning of carcasses prior to weighing and paying the producer (RESENDE et al., 2014).

The meat tenderness is the most variable and important sensorial factor that affects consumer satisfaction with meat (SAVELL et al., 1987, 1989; HUFFMAN et al., 1996; MILLER et al., 2001; DEVITT; WILTON; MANDELL, 2002). A market research has shown that improving meat tenderness increases the probability of consumers buying meat and the price they are willing to pay (BOLEMAN et al., 1997; PLATTER et al., 2003).

The use of yearling weight in the selection indices allows the identification of animals with high potential of post-weaning growth. This trait presents heritability estimates ranging from 0.29 to 0.48 (PEREIRA; RIBEIRO; SILVA, 2005; YOKOO et al., 2007; PEDROSA et al., 2008; TONUSSI et al., 2015). The yearling weight is important because, in breeding programs, the productive traits, such as weights obtained at different ages, are widely used as a selection criterion.

The reproductive traits associated with sexual precocity are determinant for the economic efficiency of the production system (BOLIGON et al., 2010). Among these, age at first calving (AFC) is a reproductive trait that has been used as indicator of female's sexual precocity because it can be observed early, and it is easily obtained (DIAS; EL FARO; ALBUQUERQUE, 2004). However, selecting directly for decresing AFC is not simple, since some producers delay the entry of females into reproduction season, determining age or weight for the beginning of reproductive life, which makes difficult to identify sexually precocious females. In addition, this trait usually presents heritability estimates of low to moderate magnitudes (0.09 to 0.28)

(MERCADANTE; LOBO; OLIVEIRA, 2000; PEREIRA; ELER; FERRAZ, 2000; BOLIGON et al., 2010).

3.2. Genomic Selection using parametric models

According to Meuwissen, Hayes and Goddard (2001 and 2013), genomic selection is a form of marker-assisted selection on a genome-wide scale. Under a parametric approach, several methods to estimate the marker effects have been strongly studied and compared in order to find the most suitable for predicting genomic values for different traits. According to Solberg et al. (2009), the estimation of marker effects can be treated as a multiple regression problem, in which phenotypes for a trait of economic relevance are the response variable, while the genotypes for SNP markers are the explanatory variables. This situation, typically, constitutes a “large p, small n problem” because the number of phenotypes is generally much lower than the number of markers.

Accuracy of prediction is strongly dependent on many factors such as linkage disequilibrium (MEUWISSEN; HAYES; GODDARD, 2001), effective population size (GODDARD, 2009), marker density (MOSER et al., 2009), allele frequency distribution (LETTRE, 2011), number of genotyped animals (VANRADEN et al., 2009; DAETWYLER et al., 2010; CALUS, 2011), heritability of the traits and the method used to estimate marker effects (LOURENCO et al., 2014).

The first statistical methods using genomic data were proposed by Meuwissen, Hayes and Goddard (2001) in a simulation study. The authors compared the methods of least-squares, BLUP, and Bayesian analyses (BayesA and BayesB) in terms of their accuracy of predicting total breeding value, and reported that least-squares method showed limitations, such as low predictive ability and the impossibility of estimating all effects simultaneously. The GBLUP, BayesA and BayesB statistical methods were more accurate, thus, have become widely used.

The GBLUP method consists of assembling genomic kinship matrix (**G** matrix) using the information of the SNPs markers. Posteriorly, the **G** matrix is used in

replacement to the matrix of kinship based on the pedigree file (**A** matrix) for the resolution of the mixed models equations. According to VanRaden (2008), the **G** matrix is considered more accurate than the **A** matrix, because the first one informs the observed proportion of chromosomal segments shared by the individuals, making possible to differentiate the relationship between full siblings, identify previously unknown kinship relationships and correct possible genealogy errors. The GBLUP estimates all marker effects at the same time and assumes the same variance for all SNP (MEUWISSEN; HAYES; GODDARD, 2013). A disadvantage of this method is the overestimation of the variance of markers with no effect and underestimation of the variance of high effect markers that can harm accuracy of prediction (TIEZZI; MALTECCA, 2015). Bolormaa et al. (2013), working with cattle populations of *Bos Taurus*, *Bos Indicus*, synthetic breeds, using the GBLUP and BayesR (a modification of BayesCπ) methods, reported Genomic Estimated Breeding Values (GEBVs) accuracies for carcass and meat quality ranging from 0.17 to 0.33, and the BayesR showed the greatest average accuracy across traits.

As proposed by Misztal et al. (2009), if phenotypes, pedigrees, and genotypes are available, a simple way to incorporate genomic information into genetic evaluations is by the single-step genomic BLUP (ssGBLUP). This approach consists of incorporating phenotypes, pedigrees and genomic information into only one step of evaluation. With this procedure, the relationship matrix based on pedigree (**A** matrix) is combined with a genomic relationship matrix based on information from SNP markers (**G** matrix), into a single matrix (**H** matrix). Thus, all SNP markers are considered simultaneously with the phenotypic information of genotyped and non-genotyped animals. Another advantage of this method is that all phenotypic information is used for pedigree effects and known markers, and allows the estimation of genetic values by any model (WANG et al., 2012). The ssGBLUP is also suitable for multi-trait analysis (CHEN et al., 2011).

The Bayesian approach may or not assume different variances on all segments of chromosomes considering that very few SNPs have very high effect and the majority of SNPs have very small or null effect (VANRADEN, 2008). Frequently,

the number of animals is much smaller than the number of marker effects to estimate so, the final marker effect estimates are strongly influenced by the prior information (MEUWISSEN; HAYES; GODDARD, 2013). Using data belonging to 543 Angus and 400 Charolais steers, Li, Chen and Vinsky (2014) reported accuracies of predicting genomic breeding values for longissimus muscle area of 0.36 and 0.24, carcass average backfat thickness of 0.33 and 0.46, and hot carcass weight of 0.35 and 0.36, for Angus and Charolais, respectively. In this work, the authors showed the accuracies as an average of GBLUP and BayesB, and both models obtained similar results. Neves et al. (2014) published a study with genomic selection in Nelore cattle for several characteristics, such as growth and carcass traits measured by visual scores. The authors reported that Bayesian regression models (Bayes C and Bayesian LASSO) outperformed the implementation of GBLUP.

Recently, Fernandes Júnior et al. (2016) worked with data from, approximately, 1,500 Nelore males for rib eye area, backfat thickness and hot carcass weight, and they estimated empirical prediction accuracies considering three models: Bayesian ridge regression, Bayes C and Bayesian Lasso. The authors concluded that all models presented similar predictive performance, although Bayesian Lasso, Bayes C and Bayesian ridge regression showed the highest accuracies for rib eye area, backfat thickness and for hot carcass weight, respectively. All these accuracies were calculated using the phenotype adjusted for fixed effects.

3.3. Genomic Selection using semi-parametric models

According to Huang et al. (2012) there is increasing evidence that complex traits are the product of synergistic forces spanned by large numbers of genetic polymorphisms within the genome. To explore the complexity of traits using a linear approach, dominance and epistasis may be accommodated by adding appropriate interactions between marker covariates to the model. However, the number of predictor variables is extremely large and modelling interactions is only feasible to a

limited degree (DE LOS CAMPOS et al., 2010a). Thus, it seems reasonable to argue that genotypes and phenotypes may be connected in forms that are not well addressed by the linear additive models that are standard in quantitative genetic area (MOROTA el al., 2013).

Taking the aforementioned into consideration, Gianola, Fernando and Stella (2006) and Gianola and van Kaam (2008) proposed the use of Reproducing Kernel Hilbert Spaces (RKHS) regressions for estimating breeding values with genomic data and capturing epistatic interactions of complex traits, although in a non-explicit manner. According to the authors, the information of the markers is included in a positive (semi) definite matrix of order n x n (n is the number of phenotypes), called kernel (**K** matrix), which is used as incidence matrix in the regression model. The elements of **K** matrix can be estimated using a Gaussian kernel, where the distances between genotypes are obtained by means of Euclidean metric space calculations, creating genetic relatedness in terms of "spatial" distance (DE LOS CAMPOS et al., 2010a).

The spatial genetic distance between two individuals is given by the squared Euclidean norm, where a positive bandwidth parameter θ controls overall smoothness of the kernel function. Thus, a small Euclidean distance between two individuals reflects a strong similarity in state between their genotypes, in other words, as θ increases, the kernel approaches zero, producing a "sharp" or "local" kernel. On the other hand, as θ decreases (close to zero), the kernel approaches one, that is, a situation where the two individuals "match" perfectly, providing a "global" kernel (MOROTA el al., 2013).

The bandwidth parameter θ of the RKHS regression can be chosen either using a sequence (grid) of values greater than zero or by means of Bayesian methods. From a Bayesian perspective, one possibility is to treat θ as random; however, this is computationally demanding because the RKHS needs to be recomputed every time the θ is updated. To overcome this problem, De los Campos et al. (2010a) proposed using a multikernel approach (named kernel averaging, KA). The KA consists, for instance, in defining the two most extreme values for θ,

obtaining two RKHS which can be fitted using a multikernel model with many random effects as RKHS. According to De los Campos et al. (2010a), KA offers a computationally convenient method for kernel selection, since it is not necessary to evaluate θ over a grid of values.

There are few studies for Nelore beef cattle that explore the accuracies of prediction using models with semi-parametric approach, such as RKHS regressions. However, there are several researches with simulated data, chickens, maize and wheat lines and mice population. Using simulated data, Long et al. (2010) reported that the use of non-parametric Radial Basis Function regression, similar to RKHS regression, in the presence of complex relationships between phenotypes and genotypes (non-linearity and non-additivity), showed an accuracy of prediction 7% higher than Bayes A in the estimation of genomic values using SNP marker information. Sun et al. (2012) simulated data of maize lines, stipulating 12 scenarios for the application of genomic selection. The simulated and real data were analyzed using the Ridge Regression BLUP, BayesA, BayesB and RKHS regression models. The authors concluded that the last one delivered greater predictive ability, particularly when epistasis impacts traits expression.

González-Recio et al. (2008) used four different statistical approaches (BLUP, linear regression on SNPs, Kernel regression on SNPs and RKHS regression) in broiler chickens. Considering information from SNPs, the authors concluded that the two non-linear statistical methods (kernel and RKHS) fitted the data better, with smaller mean squared error, greater ability of prediction and accuracy. Working with 394 records of food conversion rate in chickens with SNP information, González-Recio et al. (2009) found that Bayes A and RKHS regression were equally accurate when all 3481 SNPs were included in the models. However, the semi-parametric approach, RKHS regression, with 400 pre-selected informative SNPs was more accurate than Bayes A with all SNPs.

Pérez-Rodríguez et al. (2012) analyzed 306 elite wheat lines genotyped with 1717 diversity array technology. The authors examined the predictive ability of parametric models such as Bayesian Lasso, Bayesian Ridge Regression, Bayes A

and Bayes B, and RKHS regression, Bayesian regularized neural networks and Radial basis function neural networks as semi-parametric and non-parametric models. Results showed a consistent superiority of RKHS regression and Radial basis function neural networks over the parametric models.

Analyzing data from 1940 mice with 12,226 SNP markers located in autosomes, Neves, Carvalheiro and Queiroz (2012) worked with several traits, such as weight at 6 weeks, weight growth slope and body length and ten different statistical methods. The authors reported that Ridge Regression GBLUB, RKHS regression and Support Vector Regression are particularly appealing for application in genomic selection.

According to Gianola, Fernando and Stella (2006) and Long et al. (2010), semi-parametric models, such as RKHS regression, do not impose strong assumptions on the phenotypic-genotypic relationship. Therefore, these models have the potential of capturing interactions among loci, which can account for epistatic effects that are not captured by linear additive regression models, although it can be difficult to confirm, given the difficulty to model interactions explicitly (Neves et al. 2012).

4. REFERENCES

ABIEC - **Associação Brasileira das Indústrias Exportadoras de Carnes**. Available at: <http://www.abiec.com.br/3_rebanho.asp>. Accessed: November 23, 2016.

BENNEWITZ, J., SOLBERG, T. MEUWISSEN, T. Genomic breeding value estimation using nonparametric additive regression models. **Genetics Selection Evolution**, v.41, p.20, 2009.

BOGGS, D. L.; MERKEL, A. R. **Live animal carcass evaluation and selection manual**. 3 ed. Dubuque, Iowa, Kendall/Hunt Publishing Co., 1990. 211p.

BOLEMAN, S. J.; MILLER, R. K.; TAYLOR, J. F.; CROSS, H. R.; WHEELER, T. L.; KOOHMARAIE, M.; SHACKELFORD, S. D.; MILLER, M. F.; WEST, R. L.; JOHNSON, D. D.; SAVELL, J. W. Consumer evaluation of beef of known categories of tenderness. **Journal of Animal Science**, v.75, p.1521–1524, 1997.

BOLIGON, A. A.; ALBUQUERQUE, L. G.; MERCADANTE, M. E. Z.; LÔBO, R. B. Herdabilidades e correlações entre pesos do nascimento à idade adulta em rebanhos da raça Nelore. **Revista Brasileira de Zootecnia**, v. 38, p. 2320-2326, 2009.

BOLIGON, A. A.; ALBUQUERQUE, L. G.; MERCADANTE, M. E. Z.; LOBO, R. B. Study to relations between age at first calving, average weight gains and weights from weaning to maturity in Nellore cattle. **Revista Brasileira de Zootecnia**, v.39, p.746- 751, 2010.

BOLORMAA, S.; PRYCE, J.E.; KEMPER, K.; SAVIN, K.; HAYES, B.J.; BARENDSE, W.; ZHANG, Y.; REICH, C.M.; MASON, B.A.; BUNCH, R.J.; HARRISON, B.E.; REVERTER, A.; HERD, R.M.; TIER, B.; GRASER, H.U.; GODDARD, M.E. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality in Bos taurus, Bos indicus, and composite beef cattle. **Journal of Animal Science**, v.91, p.3088-3104, 2013.

CALUS, M. P. L.; VEERKAMP, R. F. Accuracy of multi-trait genomic selection using different methods. **Genetic Selection Evolution**, v. 43, n. 26, 2011.

CLARKE, S. M.; HENRY, H. M.; DODDS, K. G.; JOWETT, T. W. D.; MANLEY, T. R.; ANDERSON, R. M.; MCEWAN, J. C. A high throughput single nucleotide polymorphism multiplex assay for parentage assignment in New Zealand sheep, **PLOS ONE**, v. 9, p. 1-11, 2014.

CHEN, C. Y., MISZTAL, I., AGUILAR, I., TSURUTA, S., MEUWISSEN, T. H. E., AGGREY, S.E., WING, T., MUIR, W.M. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. **Journal of Animal Science**, v.89, p.23-28, 2011.

CRESSIE, N. **Statistics for Spatial Data**. New York Wiley, 1993.

CUNDIFF, L. V.; KOCH, R. M.; GREGORY, K. E.; CROUSE J. D.; DIKEMAN, M. E. Characteristics of diverse breeds in cycle IV of the cattle germoplasm evaluation program. **Beef Research-Progress Report**, v. 4, p. 63-71, 1993.

DAETWYLER, H. D.; PONG-WONG, R.; VILLANUEVA, B.; WOOLLIAMS, J. A. The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, v. 185, p. 1021-1031, 2010.

DE LOS CAMPOS, G., GIANOLA, D., ROSA, G. J. M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal of Animal Science**, v.87, p.1883-1887, 2009.

DE LOS CAMPOS, G., GIANOLA, D., ROSA, G. J. M., WEIGEL, K. A., CROSSA, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. **Genetic Research**, v.92, p.295-398, 2010a.

DEVITT, C. J. B.; WILTON, J. W.; MANDELL, I. B. In: **World Congress on Genetics Applied to Livestock Production**, v.31, p.455–458, 2002.

DIAS, L. T.; EL FARO, L.; ALBUQUERQUE, L. G. Efeito da idade de exposição de novilhas à reprodução sobre estimativas de herdabilidade da idade ao primeiro parto em bovinos Nelore. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v.56, p.370-373, 2004.

FERNANDES JÚNIOR, G. A.; ROSA, G. J. M.; VALENTE, B. D.; CARVALHEIRO, R.; BALDI, F.; GARCIA, D. A.; GORDO, D. G. M.; ESPIGOLAN, R.; TAKADA, L.; TONUSSI, R.L.; ANDRADE, W. B. F.; MAGALHÃES, A. F. B.; CHARDULO, L. A. L.; TONHATI, H.; ALBUQUERQUE, L. G. Genomic prediction of breeding values for carcass traits in Nellore cattle. **Genetics Selection Evolution**, v.48, n.7, 2016.

GIANOLA, D., FERNANDO, R. L., STELLA, A. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics**, v.173, p.1761-1776, 2006.

GIANOLA, D.; VAN KAAM, B. C. H. M. Reproducing kernel Hilbert spaces regression methods for genomic prediction of quantitative traits. **Genetics**, v.178, p.2289-2303, 2008.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of longterm response. **Genetics**, v. 136, p. 245-257, 2009.

GONZÁLEZ-RECIO, O.; GIANOLA, D.; LONG, N.; WEIGEL, K. A.; ROSA, G. J. M.; AVENDAÑO, S. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. **Genetics**, v.178, p.2305-2313, 2008.

GONZÁLEZ-RECIO, O.; GIANOLA, D.; ROSA, G. J. M.; WEIGEL, K. A.; KRANIS, A. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. **Genetics Selection Evolution**, v. 41, n. 3, 2009.

HABIER, D.; FERNANDO, R. L.; KIZILKAYA, K.; GARRICK, D. J. Extension of the Bayesian alphabet for genomic selection. **BMC Bioinformatics**, v.12, 2011.

HUANG, W. G.; RICHARDS, S.; CARBONE, M. A.; ZHU, D.; ANHOLT, R. R.; AYROLES, J. F.; DUNCAN, L.; JORDAN, K. W.; LAWRENCE, F.; MAGWIRE, M. M.; WARNER, C. B.; BLANKENBURG, K.; HAN, Y.; JAVAID, M.; JAYASEELAN, J.; JHANGIANI, S. N.; MUZNY, D.; ONGERI, F.; PERALES, L.; WU, Y. Q.; ZHANG, Y.; ZOU, X.; STONE, E. A.; GIBBS, R. A.; MACKAY, T. F. Epistasis dominates the genetic architecture of Drosophila quantitative traits. **Proceedings of the National Academy of Sciences of the United States of America**, v.109, p.15553-15559, 2012.

HUFFMAN, K. L.; MILLER, M. F.; HOOVER, L. C.; WU, C. K.; BRITTIN, H. C.; RAMSEY, C. B. Effect of beef tenderness on consumer satisfaction with steaks consumed in the home and restaurant. **Journal of Animal Science**, v.74, p.91–97, 1996.

LETTRE, G. Recent progress in the study of the genetics of height. **Human Genetics**, v. 129, p. 465-472, 2011.

LI, C.; CHEN, L.; VINSKY, M. Genomic prediction for feed efficiency and carcass traits in Angus and Charolais beef cattle. **Beef and Range Report** - University of Alberta, 2014.

LONG, N., GIANOLA, D., ROSA, G.J.M., WEIGEL, K.A., KRANIS, A., GONZALEZ-RECIO, O. Radial basis function regression methods for predicting quantitative traits using SNP markers. **Genetics Research**, v.92, p.209-225, 2010.

LOURENCO D. A.; MISZTAL, I.; TSURUTA , S.; AGUILAR, I.; EZRA, E.; RON, M.; SHIRAK, A.; WELLER, J. I. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. **Journal of Dairy Science**, v. 97, p. 1742-1752, 2014.

LUCHIARI FILHO, A. **Pecuária da carne bovina**. São Paulo, p. 134, 2000.

MAPA – **Ministério de Agricultura, Pecuária e Abastecimento**. Available at: <http://www.agricultura.gov.br/animal/especies/bovinos-e-bubalinos>. Accessed: November 15, 2016.

MERCADANTE, M. E. Z.; LOBO, R. B.; OLIVEIRA, H. N. Estimativas de (co)variância entre características de reprodução e de crescimento em fêmeas de um rebanho Nelore. **Revista Brasileira de Zootecnia**, v.29, p.997-1004, 2000.

MEUWISSEN, T. H., HAYES, B. J., GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker map. **Genetics**, v.157, p.1819-1829, 2001.

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Accelerating Improvement of Livestock with Genomic Selection. **Annual Review of Animal Biosciences**. v 1, p. 221–237, 2013.

MILLER, M. F.; CARR, M. A.; RAMSEY, C. B.; CROCKETT, K. L.; HOOVER, L. C. Consumer thresholds for establishing the value of beef tenderness. **Journal of Animal Science**, v.79, p.3062-3068, 2001.

MISZTAL, I.; LEGARRA, A.; AGUILAR, I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. **Journal of Dairy Science**, v.92, n.9, p.4648-4655, 2009.

MOROTA, G.; BODDHIREDDY, P.; VUKASINOVIC, N.; GIANOLA, D.; DENISE, S. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. **Frontiers in Genetics**, v.5, p.1-9, 2014.

MOROTA, G.; KOYAMA, M.; ROSA, G. J. M.; WEIGEL, K. A.; GIANOLA, D. Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. **Genetics Selection Evolution**, v. 45, n. 17, p.1-10, 2013.

MOSER, G.; TIER, B.; CRUMP, R. E.; KHATKAR, M. S.; RAADSMA, H. W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome wide SNP markers. **Genetics Selection Evolution**, v. 41, n. 56, 2009.

MUIR, W. M. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. **Journal of Animal Breeding and Genetics**, v.124, p.342-355, 2007.

NEVES, H. H. R.; CARVALHEIRO R.; O'BRIEN, A. M. P.; UTSUNOMIYA, Y. T.; CARMO, A. S.; SCHENKEL, F. S.; SÖLKNER, J.; MCEWAN, J. C.; VAN TASSELL, C. P.; COLE, J. B.; SILVA, MARCOS, V. G. B.; QUEIROZ, S. A.; SONSTEGARD, T. S.; GARCIA, J. F. Accuracy of genomic predictions in Bos indicus (Nellore) cattle. **Genetics Selection Evolution**, v. 46, n.17. 2014.

NEVES, H. H. R.; CARVALHEIRO, R.; QUEIROZ, S. A. A comparison of statistical methods for genomic selection in a mice population. **BMC Genetics**, v.13, 2012.

PÉREZ-RODRÍGUEZ, P.; GIANOLA, D.; GONZÁLEZ-CAMACHO, J. M.; CROSSA, J.; MANÈS, Y.; DREISIGACKER, S. Comparison between linear and non-parametric regression models for Genomic-Enabled prediction in wheat. **Genes Genomes Genetics**, v. 2, p.1595-1605, 2012.

PEREIRA, E.; ELER, J.P.; FERRAZ, J.B.S. Correlação genética entre perímetro escrotal e algumas características reprodutivas na raça Nelore. **Revista Brasileira de Zootecnia**, v.29, p.1676-1683, 2000.

PEREIRA, J. C. C.; RIBEIRO, S. H. A.; SILVA, M. A.; BERGMANN, J. A. G.; COSTA, M. D. Análise genética de características ponderais e reprodutivas de fêmeas bovinas Tabapuã. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v.57, p.231-236, 2005.

PEDROSA, V. B.; ELER, J. P.; SILVA, M. R.; FERRAZ, J. B. S; BALIEIRO, J. C. C. Estimativas de parâmetros genéticos do peso adulto e de características de escore visual em animais da raça Nelore. In: VII SIMPOSIO BRASILEIRO DE MELHORAMENTO ANIMAL, 2008, São Carlos - SP. **Anais...** São Carlos: 2008. CD-ROM.

PLATTER, W. J.; TATUM, J. D.; BELK, K. E.; SCANGA, J. A.; SMITH, J. C. Effects of repetitive use of hormonal implants on beef carcass quality, tenderness, and consumer ratings of beef palatability. **Journal of Animal Science**, v.81, p.984–996, 2003.

RESENDE, F. D.; GESUALDI JÚNIOR, A.; QUEIROZ, A. C.; FARIA, M. H.; VIANA, A. P. Carcass characteristics of feedlot-finished Zebu and Caracu cattle. **Revista Brasileira de Zootecnia**, v. 43, n. 2, p. 67-72, 2014.

SAVELL, J. W.; BRANSCON, R. E.; CROSS, H. R.; STFFLER, D. M.; WISE, J. W.; GRIFFIN, D. B.; SMITH, G. C. National consumer retail beef study: Palatability evaluations of beef loin steaks that differed in marbling. **Journal of Food Science,** v.52, p.517-519, 1987.

SAVELL, J. W.; CROSS, H. R.; FRANCIS, J. J.; WISE, J. W.; HALE, D. S.; WILKES, D. L.; SMITH, G. C. National consumer retail beef study: Interaction of trim level, price, and grade on consumer acceptance of beef steaks and roasts. **Food Quality**, v.12, p.251-274, 1989.

SOLBERG, T. R.; SONESSON, A. K.; WOOLLIAMS, J. A., MEUWISEEN, T. H. E. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selection Evolution**, 41:29, 2009.

SONESSON, A. K, MEUWISSEN T. H. E. Testing strategies for genomic selection in aquaculture breeding programs. **Genetic Selection Evolution**, v.41, p.37, 2009.

SUN, X.; MA, P.; MUMM, R. H. Nonparametric Method for Genomics-Based Prediction of Performance of Quantitative Traits Involving Epistasis in Plant Breeding, **PLoS ONE**, v.7, 2012.

TIEZZI, F.; MALTECCA, C. Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. **Genetics Selection Evolution**, v. 47(1), p.1-113, 2015.

TONUSSI, R. L.; ESPIGOLAN, R.; GORDO, D. G. M.; MAGALHÃES, A. F. B; VENTURINI, G. C.; BALDI, F.; OLIVEIRA, H. N.; CHARDULO, L. A. L.; TONHATI, H.; ALBUQUERQUE, L. G. Genetic association of growth traits with carcass and meat traits in Nellore cattle. **Genetics and Molecular Research**, v.14, p.18713 - 18719, 2015.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v.91, p. 4414-4423, 2008.

VANRADEN, P. M.; VAN TASSELL, C. P.; WIGGANS G. R.; SONSTEGARD T. S.; SCHNABEL, R. D.; TAYLOR, J. F.; SCHENKEL, F. S. Invited review: reliability of genomic predictions for North American Holstein bulls. **Journal of Dairy Science**, v. 92, p.16-24, 2009.

VAPNIK, V. **Statistical Learning Theory**. New York, NY: Wiley, 1998.

WAHBA, G. Spline Models for Observational Data. Philadelphia, PA: **Society for Industrial and applied Mathematics**, 1990.

WANG, H.; MISZTAL, I.; AGUILAR, I., LEGARRA, A.; MUIR, W. M. Genome-wide association mapping including phenotypes from relatives without genotypes. **Genetics Research**, v. 94, p. 73-83, 2012.

WINKELMAN, A. M., JOHNSON, D. L., HARRIS, B. L. Application of genomic evaluation to dairy cattle in New Zealand. **Journal of Dairy Science**, v.98, p.659-675, 2015.

YOKOO, M. J. I.; ALBUQUERQUE, L. G.; LOBO, R. B.; SAINZ, R. D.; CARNEIRO JÚNIOR, J. M.; BEZERRA, L. A. F.; ARAUJO, F. R. C. Estimativas de parâmetros genéticos para altura do posterior, peso e circunferência escrotal em bovinos da raça Nelore. **Revista Brasileira de Zootecnia**, v.36, n.6, p.1761-1768, 2007.

CHAPTER 2 – Accuracy of genomic predictions obtained through parametric and semi-parametric models using a set of real data

ABSTRACT - The aim of this study was to compare the accuracy of genomic predictions using parametric and semi-parametric statistical models for carcass, meat quality, growth and reproductive traits in Nelore cattle. The phenotypic and pedigree information used were provided by farms belonging to three animal breeding programs which represent eleven farms. For carcass and meat quality traits, the data set contained 3,643 records for rib eye area (REA), 3,619 records for backfat thickness (BFT), 3,670 records for meat tenderness (TEN) and 3,378 observations for hot carcass weight (HCW). A total of 825,364 records for yearling weight (YW) and 166,398 for age at first calving (AFC) were used as growth and reproductive traits. Genotypes of 2,710, 2,656, 2,749, 2,495, 4,455 and 1,760 were available for REA, BFT, TEN, HCW, YW and AFC, respectively. After quality control, approximately 450,000 single nucleotide polymorphisms (SNP) were used in analysis. Methods of analysis were genomic BLUP (GBLUP), single-step GBLUP (ssGBLUP), Bayesian LASSO (BL) and the semi-parametric approaches Reproducing Kernel Hilbert Spaces (RKHS) regression and Kernel Averaging (KA). A five-fold cross-validation with thirty random replicates was carried out and models were compared in terms of their prediction mean squared error and accuracy of prediction (ACC). The ACC ranged from 0.39 to 0.40 (REA), 0.38 to 0.41 (BFT), 0.23 to 0.28 (TEN), 0.33 to 0.35 (HCW), 0.36 to 0.51 (YW) and 0.49 to 0.56 (AFC). For all traits, the GBLUP and BL models showed very similar prediction accuracies. For REA, BFT and HCW, genomic models provided similar prediction accuracies, however RKHS regression had the best fit across traits considering multiple-step models and compared to KA. The application of the bandwidth parameter for RKHS regressions was clearly trait-specific and dependent. Judged by overall performance, across all traits, the RKHS regression is particularly appealing for application in genomic selection, especially for low heritability traits. For traits which have a higher number of animals with phenotypes compared to the number of those with genotypes (YW and AFC), the ssGBLUP is indicated.

Keywords: *Bos taurus indicus***,** genomic selection, Kernel Averaging, meat quality, RKHS regression

1. INTRODUCTION

Selection of animals for economical relevant traits is, traditionally, carried out through estimated genetic values obtained from individuals and/or their relatives phenotypes, considering the proportion of common alleles by descent and the heritability of the trait (VAN ARENDONK; TIER; KINGHORN, 1994). This has been a successful approach adopted by beef cattle breeding programs. However, the genetic progress has been relatively slow for traits measured in only one sex or after slaughter in addition to traits of low heritability and difficult or high cost of measurement (GODDARD; HAYES, 2009).

Nowadays, in order to improve response to selection, single nucleotide polymorphisms (SNP) markers, widely distributed throughout the genome, assuming that genetic markers are in linkage disequilibrium to quantitative trait loci (QTL), have been included in the models (CLARKE et al. 2014). Nevertheless, incorporating markers into models for predicting genetic values places important statistical and computational challenges since models including dense molecular markers should be able to cope with the curse of dimensionality; being flexible enough to capture the complexity of quantitative traits and amenable for computations (DE LOS CAMPOS et al. 2010a).

Taking into consideration the aforementioned, some methods have been proposed for dealing with the large amount of genomic information currently available. Those considering a parametric approach, can be highlighted: Genomic Best Linear Unbiased Predictor (GBLUP), Single-Step Genomic Best Linear Unbiased Predictor (ssGBLUP) and Bayesian models.

Instead of pedigree information, GBLUP estimates breeding values using a matrix of genomic relationships (**G** matrix) (VAN RADEN et al. 2008). On the other hand, Legarra et al. (2009) and Misztal et al. (2009) proposed a single-step procedure (ssGBLUP) that consists of integrating the pedigree (**A** matrix) and genomic information (**G** matrix) into a single matrix (**H**) to predict the genomic breeding value. Several studies reported that ssGBLUP is computationally efficient

and accurate for genomic evaluation purposes (AGUILAR et al., 2010; TSURUTA et al., 2011; CHEN et al., 2011; CHRISTENSEN et al., 2012).

Under a semi-parametric approach, Gianola et al. (2006) and Gianola and van Kaam (2008) proposed the Reproducing Kernel Hilbert Spaces (RKHS) regression for predicting genomic values. This model uses weaker assumptions than traditional fully parametric models and allows accounting for non-additive effects without explicit modeling. The RKHS regression can potentially pick up various forms of gene action without placing hightly parametrized structures that require making strong assumptions a priori about the distribution and genetic architecture. Another semi-parametric model, named kernel averaging (KA), was proposed by De los Campos et al. (2010a) and consists of the multikernel approach using RKHS regressions.

For Nelore beef cattle, there are few studies that explore models with semi-parametric approach with a set of real data and considering information provided by the SNP markers. Therefore, the objective of this study was to compare the accuracy of genomic predictions using parametric and semi-parametric statistical models for carcass, meat quality, growth and reproductive traits in Nelore cattle.

2. MATERIAL AND METHODS

2.1. Phenotypic and genotypic data

Phenotypic and pedigree information were provided by four animal breeding programs - DeltaGen, Nelore Qualitas, CRV Lagoa/PAINT and Cia. de Melhoramento – including eleven farms. Traits analyzed were: carcass and meat quality traits, like rib eye area (REA), backfat thickness (BFT), meat tenderness (TEN), hot carcass weight (HCW), yearling weight (YW) and age at first calving.

Animals were reared on pasture, finished in feedlot for about 90 days and slaughtered when they were, on average, 697 ± 98 days of age. Contemporary groups (CG) for carcass and meat quality traits were defined as farm and year of birth and management group at yearling.

The CG for YW and AFC were defined by the effects of farm and season of birth, management group at weaning and at yearling and, for YW, the sex of animals was added. For all traits in this study, CG with records outside the interval given by the mean of the group plus or minus three standard deviations were discarded. Additionally, CG with fewer than three observations were excluded from the analysis.

Measurements of HCW were obtained at slaughter, for each animal. After a 24 to 48 hours chill, samples of *Longissimus dorsi* muscle with bone, and approximately, 2.54 cm of thickness were collected between the 12^{th} and 13^{th} rib of the left half-carcass and immediately frozen at −20°C for later analyses. Point counting on a plastic grid (where each square corresponds to 1 cm^2) was used to measure REA, in which the grid was placed on the sample and the sum of all squares corresponds to the REA of the animal. For the determination of BFT, the layer of subcutaneous fat located at an angle of 45 degrees from the geometric center of the sample was measured in millimeters with a caliper. Records of TEN were obtained after cooking meat samples according to the methodology proposed by Wheeler, Koohmaraie and Shackelford (1995), and using a mechanical SALTER equipment with an Warner-Bratzler probe with a capacity of 25 kg and a speed of 20 cm/minute. The shearing was performed in ½-inch cylinders, taken from the central region of the sample. The mean age of the animals at slaughter was 697 ± 98 days.

The AFC is defined as the age of the cow/heifer at the moment of first calving. Mating season for cows begins around the second half of november, with a duration of approximately 70 days. For heifers of 14 to 16 months of age, there is an early mating season, between the months of February and April, with a duration of, approximately, 60 dias. All heifers are exposed to reproduction regardless of weight and body condition. The mating systems used are: artificial insemination, controlled mating and multiple sire, with sire:cow ratio of 1:50. Heifers are evaluated for pregnancy by rectal palpation nearly 60 days after the early mating season. A mean of 1046 ± 109 days of age to achieve the first calving was found. The structure of the data is shown in Table 1.

Table 1. Descriptive statistics for rib eye area (REA), backfat thickness (BFT), meat tenderness (TEN), hot carcass weight (HCW), yearling weight (YW) and age at first calving (AFC)

Trait	N_{total}	Mean	SD	Min	Max	CG	N_{sire}	N_{dam}	P_{unk}
REA (cm^2)	3643	67.80	8.23	43.00	95.00	136	412	3266	42
BFT (mm)	3619	4.04	1.86	1.00	11.00	136	412	3243	42
TEN (kgf)	3670	6.02	1.76	1.60	11.30	136	414	3290	42
HCW (kg)	3378	271.30	24.51	197.20	346.80	123	371	3116	42
YW (kg)	825364	286.60	37.66	182.00	414.00	24808	6513	446598	41
AFC (days)	166398	1046	108.91	400	1300	7865	3241	127951	43

N_{total} = number of animals with phenotypes; SD = standard deviation; Min = minimum; Max = maximum; CG = number of contemporary groups; N_{sire} = number of known sires; N_{dam} = number of known dams; P_{unk} = percentage of unknown sires

Genotypic data from 4,847 animals (2,812 males and 2,035 females) were used. Seventy percent of these genotypes were obtained using a panel of 777,962 single nucleotide polymorphisms (SNP) from the Illumina Bovine HD chip and the rest was genotyped using a GeneSeek® Genomic Profiler (GGP) HDi 80K (GeneSeek Inc., Lincoln, NE) from NEOGEN, which were built specifically for *Bos taurus indicus* breeds, and contains 74,085 markers. Genotypes from GGP HDi 80K chip were imputed to the HD chip using the FImpute software considering pedigree information (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014). For quality control of genotypes only autosomal SNPs were considered, and SNPs with minor allele frequency less than 0.03, a Hardy-Weinberg equilibrium *p* value less than 10^{-5} and a call rate less than 0.95 were excluded. For samples, a call rate of at least 0.90 was required.

The genotypes were defined as 0 (AA), 1 (AB), and 2 (BB) and the missings were imputed using allele frequency estimates from the data (PÉREZ; DE LOS CAMPOS, 2014; MOROTA et al. 2014). After phenotypic and genotypic editing the final dataset is shown in Table 2.

Table 2. Information about the genotypic data for rib eye area (REA), backfat thickness (BFT), meat tenderness (TEN), hot carcass weight (HCW), yearling weight (YW) and age at first calving (AFC)

Trait	N_{total}	N_{gen}	N_{SNP}	SNP Chip
REA (cm^2)	3643	2710	453431	Illumina Bovine HD 700K; GGP HDi 80K
BFT (mm)	3619	2656	453342	Illumina Bovine HD 700K; GGP HDi 80K
TEN (kgf)	3670	2749	453580	Illumina Bovine HD 700K; GGP HDi 80K
HCW (kg)	3378	2495	453920	Illumina Bovine HD 700K; GGP HDi 80K
YW (kg)	825364	4455	449681	Illumina Bovine HD 700K; GGP HDi 80K
AFC (days)	166398	1760	491341	Illumina Bovine HD 700K

N_{total} = number of animals with phenotypes; N_{gen} = number of animals with phenotypes and genotypes; N_{SNP} = number of SNP markers after quality control

Phenotypes adjusted for fixed effects (Y*) were used as a response variable in genomic analysis and were estimated using an animal model considering fixed effects of CG (for AFC) and linear effect of age at slaughter (for REA, BFT, TEN and HCW) or linear effect of age of animal at recording (for YW). For Y* estimation, the analyses for all traits were performed in a single-trait scheme using AIREMLF90 program, which are part of the BLUPF90 family (MISZTAL et al., 2002). Descriptive statistics for Y* across traits are summarized in Table 3.

Table 3. Descriptive statistics for phenotypes adjusted for fixed effects (Y*) on rib eye area (REA), backfat thickness (BFT), meat tenderness (TEN), hot carcass weight (HCW), yearling weight (YW) and age at first calving (AFC)

Trait	N	Mean	SD	Minimum	Maximum
REA (cm^2)	3643	0.33	6.84	-22.86	32.58
BFT (mm)	3619	0.05	1.50	-5.09	6.79
TEN (kgf)	3670	0.01	1.30	-4.71	5.12
HCW (kg)	3378	0.28	16.98	-64.64	78.90
YW (kg)	825364	5.66	27.71	-190.60	240.40
AFC (days)	166398	-1.98	78.15	-703.10	408.90

N = number of animals with phenotypes; SD = standard deviation

The variance components and heritability estimates were obtained using the same models described for Y*, without considering genotypic information (Table 4).

Table 4. Variance components and heritability estimates for rib eye area (REA), backfat thickness (BFT), meat tenderness (TEN), hot carcass weight (HCW), yearling weight (YW) and age at first calving (AFC)

Trait	σ_a^2	σ_e^2	h^2 (SD)
REA (cm^2)	18.84	30.04	0.39 (0.084)
BFT (mm)	0.43	1.92	0.18 (0.062)
TEN (kgf)	0.15	1.60	0.09 (0.044)
HCW (kg)	40.89	260.02	0.14 (0.057)
YW (kg)	282.26	495.50	0.36 (0.003)
AFC (days)	759.53	5538.40	0.12 (0.005)

σ_a^2 = additive genetic variance; σ_e^2 = residual variance; h^2 = heritability; SD = standard deviation of the heritability estimates

2.2. Statistical Models

The analyses were performed using the BLUPF90 family programs (MISZTAL et al., 2002) available at http://nce.ads.uga.edu/wiki/doku.php, the Bayesian Generalized Linear Regression (BGLR) package (PÉREZ; DE LOS CAMPOS, 2014) and R software (R Development Core Team, 2017).

2.2.1. Genomic Best Linear Unbiased Predictor (GBLUP)

The GBLUP model is similar to BLUP using a genomic relationship matrix (**G**) instead **A**, producing direct genomic value (DGV) based on SNP effects. Solutions from GBLUP can be obtained with the model showed below:

$$\mathbf{Y^* = 1\mu + Zg + e,}$$

where **Y*** is the vector of phenotypes adjusted for fixed effects, μ is the overall mean, **1** is a vector of ones, **Z** is an incidence matrix of markers effects, **g** is a vector of marker effects, and **e** is a vector of residual effects. It was assumed **g** ~ N(0, **G**σ^2_g), where σ^2_g is the variance of markers and **G** is the genomic relationship matrix. Random residuals were assumed **e** ~ N(0, **I**σ^2_e), where **I** is a identity matrix and σ^2_e is the residual variance.

According to VanRaden (2008), the **G** matrix can be obtained from at least three ways. For this study, we chose the following:

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{\mathbf{2} \sum_{\mathbf{j=i}}^{\mathbf{m}} \mathbf{p_j} \, (\mathbf{1} - \mathbf{p_j})},$$

where **M** is a matrix of marker alleles with n lines (n = number of genotyped animals) and m columns (m = number of markers), and **P** is a matrix containing: 2($\mathbf{p_j}$-0.5), with $\mathbf{p_j}$ being the frequency of the second allele. Elements of **M** are set to 0 and 2 for both homozygous and to 1 for the heterozygous genotype.

2.2.2. Single-step Genomic Best Linear Unbiased Predictor (ssGBLUP)

The model used in ssGBLUP consists in combining **A** and **G** into a single matrix (**H**). Thus, the inverse of the numerator relationship matrix ($\mathbf{A}^{-1}$) was replaced by $\mathbf{H}^{-1}$, which combines pedigree and genomic information.

$$\mathbf{Y^* = Zu + e},$$

where **Y*** is the vector of phenotypes adjusted for fixed effects, **u** is the vector of direct additive genetic effects, and **Z** is an incidence matrix. Considering an infinitesimal model, var(**u**) = $\mathbf{H}\boldsymbol{\sigma}^2_{\mathbf{u}}$, where **H** is a combined relationship matrix that

integrates the genomically derived relationships (**G** matrix) with population-based pedigree relationships (**A** matrix) and $\boldsymbol{\sigma}_{\mathbf{u}}^{\mathbf{2}}$ is the additive genetic variance, and var(**e**) = $\mathbf{I}\boldsymbol{\sigma}_{\mathbf{e}}^{\mathbf{2}}$ where $\boldsymbol{\sigma}_{\mathbf{e}}^{\mathbf{2}}$ is the residual variance.

Matrix $\mathbf{H}^{-1}$ can be obtained as follows (AGUILAR et al., 2010):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix},$$

where $\mathbf{G}^{-1}$ is the inverse of genomic relationship matrix and $\mathbf{A}_{\mathbf{22}}^{-\mathbf{1}}$ is the inverse of pedigree-based numerator relationship matrix for genotyped animals.

2.2.3. Bayesian LASSO (BL)

The general model for genomic prediction using BL model, expressed in matrix notation is:

$$\mathbf{Y^* = 1}\mu \mathbf{+ Wg + e,}$$

where **Y*** is the vector of phenotypes adjusted for fixed effects, μ is the overall mean, **1** is a vector of ones, **g** is the vector of marker effects, **W** contains the genotype (coded 0 = AA, 1 = AB and 2 = BB) for each individual and each marker, and **e** is the vector of residual effects. According to De Los Campos et al. (2013) and Pérez and De Los Campos (2014), the marginal prior distribution assigned to **g** in the BGLR package, is a double exponential function, which includes a parameter λ^2 that was treated as unknown, with a prior distribution λ^2 ~ gamma(*r, s*). Still, the BGLR

package considers by default that *r* = 1.1 and calculates the scale parameter *s* based on the "prior" R^2 of the model.

For BL, samples from posterior distributions were obtained by the Gibbs sampler based on 80,000 Monte Carlo Markov Chain (MCMC) samples with the first 40,000 discarded as burn-in. After burn-in, samples were thinned at a rate of 10. Convergence diagnostics, statistical and graphical analysis of Gibbs sampling were checked by visual inspection of trace plots of variance components using the Coda (RAFTERY; LEWIS, 1992) package.

2.2.4. Reproducing Kernel Hilbert Spaces (RKHS) regression

The RKHS regressions with a semi-parametric approach can be formulated, as described in Gianola and van Kaam (2008), as:

$$\mathbf{Y^* = 1}\mu + \mathbf{K}_\theta\boldsymbol{\alpha} + \mathbf{e},$$

where **Y*** is the vector of phenotypes adjusted for fixed effects, μ is the overall mean, **1** is a vector of ones. The non-parametric term is given by $\mathbf{K}_\theta\boldsymbol{\alpha}$, where $\mathbf{K}_\theta$ is a positive definite matrix of kernels, dependent on a bandwidth parameter (θ), **α** is a vector of non-parametric coefficients and **e** is the vector of residuals effects, with **α** and **e** assumed to be independently distributed.

Thus, let **M** denote the *n* x *p* matrix of genotypes (coded 0 = AA, 1 = AB and 2 = BB) for *n* genotyped animals and *p* SNP markers, and θ is a bandwidth parameter (GIANOLA; VAN KAAM, 2008). The matrix **K** is based on Euclidean distance matrix ($\mathbf{EDM}$) in a Gaussian kernel (WINKELMAN; JOHNSON; HARRIS, 2015):

$$\mathbf{K} = \exp(-\theta * \mathbf{EDM}),$$

where

$$\mathbf{EDM}_{ij} = \sum_{k} \left(\mathbf{m}_{ik} - \mathbf{m}_{jk}\right)^2,$$

and **m** are the elements of **M.** Therefore, the $\mathbf{EDM}_{ij}$ matrix is the squared Euclidean distance between individuals i and j calculated based on their genotypes for SNP markers. The matrix $\mathbf{EDM}$ can be conveniently derived using:

$$\mathbf{EDM}_{ij} = \sqrt{\mathbf{S}_i + \mathbf{S}_j - 2\mathbf{MM}'_{ij}},$$

where

$$\mathbf{S} = \mathrm{Diag}[\mathbf{MM}'].$$

The parameter θ in this study was set for a grid of values equal to 0.1, 0.2, 0.5, and ranging from 1 to 10. For RKHS regression, a MCMC with 300,000 samples was run and the first 50,000 were discarded. Subsequently, 250,000 samples were obtained and thinned at a rate of 50, resulting in 5000 mildly correlated samples for posterior inference. Convergence diagnostics, statistical and graphical analysis of Gibbs sampling were checked by visual inspection.

2.2.5. Kernel Averaging (KA)

The KA was fitted using two RKHS: the first one with the lowest value in the grid sequence for bandwidth parameter ($\theta_L = 0.1$) and the second with the highest value ($\theta_H = 10$). The "multikernel model" was described with as many random effects as kernels (DE LOS CAMPOS et al., 2010a):

$$\mathbf{Y^* = 1}\mu + \mathbf{K_{\theta_L}\alpha_L} + \mathbf{K_{\theta_H}\alpha_H} + \mathbf{e},$$

where **Y*** is the vector of phenotypes adjusted for fixed effects, μ is the overall mean, **1** is a vector of ones. The non-parametric terms is given by $\mathbf{K_{\theta_L}\alpha_L}$ and $\mathbf{K_{\theta_H}\alpha_H}$, where $\mathbf{K_{\theta_L}}$ and $\mathbf{K_{\theta_H}}$ are positive definite matrices of kernels, dependent of two bandwidth parameters ($\boldsymbol{\theta_L = 0.1}$ and $\boldsymbol{\theta_H = 10}$), $\boldsymbol{\alpha_L}$ and $\boldsymbol{\alpha_H}$ are vectors of non-parametric coefficients and **e** is the vector of residual effects, with $\boldsymbol{\alpha_L}$, $\boldsymbol{\alpha_H}$ and **e** assumed to be independently distributed. The **K** matrix was calculated as described in topic 2.2.4. For KA, a MCMC with 300,000 samples was run and the first 50,000 were discarded. Subsequently, 250,000 samples were obtained and thinned at a rate of 50, resulting in 5000 mildly correlated samples for posterior inference.

2.2.6. Prediction of genomic values

As proposed by Meuwissen, Hayes and Goddard (2001), the GEBVs of the animals, in this case for parametric models, are a function of SNP markers' genotypes and effects, and were obtained using the following equation:

$$\mathrm{GEBV_i} = \sum_{j=1}^{p} w_{ij}\,\hat{g}_j,$$

where *p* is the number of SNP; w_{ij} is the genotype of animal *i* for SNP *j* (coded as 0, 1 or 2), and $\hat{g}_j$ is the estimated SNP substitution effect for SNP *j* that was estimated from the training population.

The prediction of GEBVs of animals using RKHS regression and KA was made using the following equation (MOROTA et al. 2014):

$$\mathbf{GEBV} = \mathbf{K}^*(\theta)\hat{\boldsymbol{\alpha}},$$

where $\mathbf{K}^*(\theta)$ is a matrix with the distance kernel between genotypes with a bandwidth parameter (θ), and $\hat{\boldsymbol{\alpha}}$ is the estimated vector of non-parametric regression coefficients in the training set.

2.2.7. Cross-validation and criteria for the comparison of models

The animals were divided into 5 subsets of, approximately, equal sizes, and each subset was sequentially taken as a testing-set while the remaining ones were used to train the predictive model using different statistical methods. This five-fold cross-validation process was repeated and replicate thirty times and for each run of cross-validation, the same training and test set were used for all the models to guarantee a fair comparison.

Models were compared in terms of their prediction mean squared error (MSE) and accuracy of prediction (ACC). The ACC was defined as the correlation between the Y* observed and the Y predicted divided by the square root of heritability of each trait. This division was made to account for the fact that phenotypes adjusted for fixed effects were used instead of true breeding value (PRYCE et al., 2012).

The MSE was used as a measure of the prediction ability of the models, which combines quality assessment in terms of variance and bias of predictions. MSE can be calculated as follows:

$$\mathbf{MSE} = \sum_{\mathbf{i} \in \mathbf{TST}} (\mathbf{y_i^*} - \hat{\mathbf{y}}_\mathbf{i})^2 \,/\, \mathbf{N_{TST}}$$

where $\mathbf{y_i^*}$ is the adjusted phenotype inside the testing set, $\hat{\mathbf{y}}_\mathbf{i}$ is the predicted phenotype for animal **i** that belongs to the testing set and $\mathbf{N_{TST}}$ is the number of animals in the testing set.

3. RESULTS AND DISCUSSION

In RKHS regression, the choice of bandwidth is a central element of model specification. Therefore, the accuracy of prediction was evaluated over a grid of values of θ for all traits studied, as shown in Figure 1.

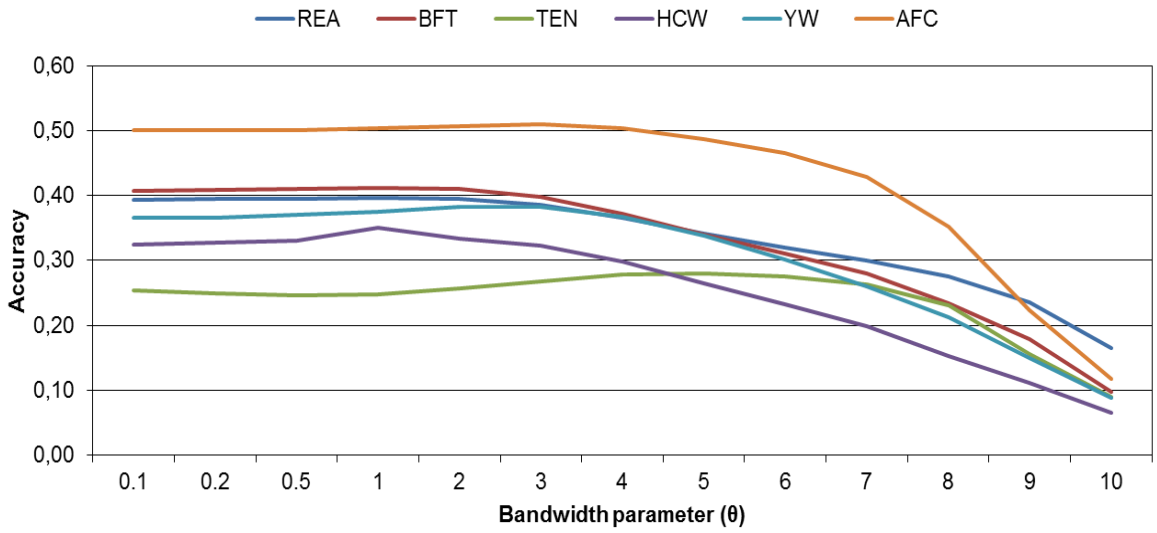

Figure 1. Accuracy of prediction across bandwidth parameters (θ) for rib eye area (REA), backfat thickness (BFT), meat tenderness (TEN), hot carcass weight (HCW), yearling weight (YW) and age at first calving (AFC)

Except for TEN, the accuracy of prediction for all traits quickly decreases with θ equal to or greater than 5. Indeed, with more extreme values of the bandwidth parameter, marker information is virtually lost, in other words, choosing θ = 0.1 gives a kernel matrix full of values very close to one and θ = 10 gives a kernel matrix with very low correlations in the off-diagonal, similar to an identity matrix (Figure 2). This general pattern was also reported by De los Campos et al. (2010a) who worked with RKHS regression using marker and phenotypic information of 599 wheat lines in four environments.

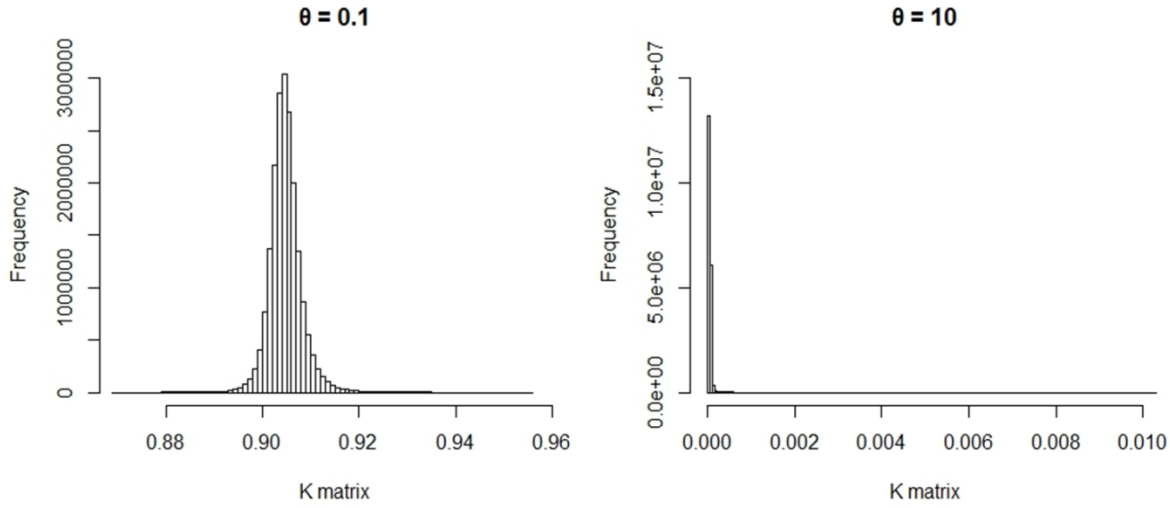

Figure 2. Histogram of the evaluations of off-diagonal elements from K matrix by value of the bandwidth parameter (θ=0.1 left and θ=10, right)

Considering TEN, higher accuracies were achieved between θ = 4 and θ = 6, which differs from that found for the other traits, illustrating that the optimal choice of RKHS regression dependent of a bandwidth parameter, may be trait dependent. These results are in agreement with the research conducted by De los Campos et al. (2010b) that, using data from 1446 US Jersey sires genotyped with the BovineSNP50BeadChip, compared the predictive ability of RKHS regression over a grid of values of the bandwidth parameter for predicted transmiting abilities of milk

production, protein content and daughter pregnancy rate. The authors reported that higher predictive abilities were found with different bandwidth parameters for each trait.

In general, the accuracy of prediction ranged from low to moderate (Table 5). Although it is expected that traits with higher heritability estimates lead to higher accuracy of prediction, this was not fully verified in the present study. For instance, the heritability for AFC was low (0.12). However the accuracy of prediction for this trait was higher (0.56) compared to REA (0.40) which has high heritability estimate (0.39). The aforementioned can be explained by the fact that the accuracy of prediction is strongly dependent on many factors, in addition to heritability, such as linkage disequilibrium (MEUWISSEN; HAYES; GODDARD, 2001), effective population size (GODDARD, 2009), marker density (MOSER et al., 2009), number of genotyped animals (VANRADEN et al., 2009; DAETWYLER et al., 2010; CALUS, 2011;), allele frequency distribution (LETTRE, 2011), and the method used to estimate marker effects (LOURENCO et al., 2014).

Table 5. Accuracy of prediction (ACC), standard deviation (SD) and mean squared error (MSE) for rib eye area (REA), backfat thickness (BFT), meat tenderness (TEN), hot carcass weight (HCW), yearling weight (YW) and age at first calving (AFC) obtained with different models and the average of five fold cross-validation with thirty random replicates. The best prediction model for each trait is in bold

Trait	Models	ACC ± SD	MSE
REA (cm^2)	GBLUP	0.39 ± 0.032	45.14
	ssGBLUP	0.39 ± 0.026	49.35
	BL	0.39 ± 0.029	45.12
	RKHS $_{(\theta=1.0)}$	**0.40 ± 0.032**	**45.10**
	KA	0.39 ± 0.032	45.14
BFT (mm)	GBLUP	0.41 ± 0.035	2.13
	ssGBLUP	0.38 ± 0.031	2.40
	BL	0.40 ± 0.033	2.14
	RKHS $_{(\theta=1.0)}$	**0.41 ± 0.035**	**2.12**
	KA	0.41 ± 0.034	2.13
TEN (kgf)	GBLUP	0.23 ± 0.037	1.72
	ssGBLUP	0.27 ± 0.032	1.83
	BL	0.23 ± 0.039	1.73
	RKHS $_{(\theta=5.0)}$	**0.28 ± 0.038**	**1.70**
	KA	0.26 ± 0.038	1.71
HCW (kg)	GBLUP	0.33 ± 0.035	261.23
	ssGBLUP	0.34 ± 0.039	280.73
	BL	0.33 ± 0.036	261.19
	RKHS $_{(\theta=1.0)}$	**0.35 ± 0.035**	**260.65**
	KA	0.33 ± 0.036	261.23
YW (kg)	GBLUP	0.36 ± 0.028	478.40
	ssGBLUP	**0.51 ± 0.002**	**447.24**
	BL	0.36 ± 0.028	477.26
	RKHS $_{(\theta=2.0)}$	0.38 ± 0.028	475.78
	KA	0.37 ± 0.028	476.23
AFC (days)	GBLUP	0.49 ± 0.040	15200.80
	ssGBLUP	**0.56 ± 0.006**	**14858.74**
	BL	0.49 ± 0.036	15178.62
	RKHS $_{(\theta=2.0)}$	0.51 ± 0.040	15154.97
	KA	0.50 ± 0.042	15162.34

θ = bandwidth parameter for kernel regression method

For REA, BFT and HCW, genomic models provided similar prediction accuracies, however, the RKHS regression with θ = 1.0 showed the lowest MSE estimate. Still, for these traits, we have not found major differences between the GBLUP, BL and ssGBLUP models, because the number of animals with phenotypes is not much larger than the number of those with genotypes.

Taking into consideration the aforementioned, for YW and AFC, which have a higher number of phenotypes, single step model performed better than GBLUP, BL and RKHS regression. These results are in agreement with Onogi et al. (2014) who concluded that the implementation of genomic selection by ssGBLUP provided more accurate predictions than GBLUP for carcass weight and ribeye area using phenotypic records of 17,347 animals and only 616 genotyped sires of Japanese Black cattle breed.

Except for BFT, GBLUP and BL showed equal values for accuracy of prediction. Nevertheless, BL fitted better than GBLUP, with lowest MSE estimates, except for TEN. Similar results were found in the work conducted by Costa (2014) using genotypes from 1853 Nelore heifers for reproductive traits, including AFC, and evaluating GBLUP, BayesCπ and Improved Bayesian LASSO statistical models. The author found a slight superiority of the Bayesian models over GBLUP. Still, Fernandes Júnior et al. (2016), using approximately half of the same data set as the present work and applying Bayesian models, reported accuracy for REA (0.47) and HCW (0.37) greater than those found in this study. However, for BFT, the authors found an accuracy of 0.22, which was lower than the result we have found using the BL model (0.40). These differences can be explained by the variation in the number of animals in testing and training population (VANRADEN et al., 2009; CALUS, 2010; DAETWYLER et al., 2010; CALUS; VEERKAMP, 2011).

RKHS regression fitted best across all traits considering multiple-step models. On the other hand, KA showed accuracies of prediction very close to those obtained with parametric models, except ssGBLUP. This difference between RKHS and KA is, to some extent, expected because the analysis with RKHS was tested using thirteen values of bandwidth parameter, individually and for each trait, that is, in an optimized

way compared to KA. These results are in agreement with those reported by De los Campos et al. (2010a and 2010b) who describe that using KA gave accuracy of prediction similar to that achieved with best performing RKHS with θ from a grid of values. However, the authors concluded that KA outperformed RKHS regression for the majority of traits.

Although the analysis with RKHS over a sequence of values is more complex and time-consuming compared to KA, in the case of TEN, a low heritable trait (0.09), the RKHS regression with θ = 5.0 showed accuracies of prediction 18% and 7% better than those obtained with GBLUP and KA, respectively. This may suggest that RKHS regression has advantages when dealing with low heritability traits, which is similar to the results found by the research of González-Recio et al. (2008) who analyzed a low heritable trait (chicken mortality) using different parametric and non-parametric approaches. These authors found a higher predictive ability using RKHS regression than other methods, including the Bayesian regression.

4. CONCLUSIONS

Application of the bandwidth parameter for RKHS regressions was clearly trait-specific and dependent. Judged by overall performance, across all traits, the RKHS regression is particularly appealing for application in genomic selection, especially for low heritability traits. For traits, which have a higher number of animals with phenotypes compared to the number of those with genotypes (YW and AFC), the ssGBLUP is indicated.

5. REFERENCES

AGUILAR, I.; MISZTAL, I.; JOHNSON, D. L.; LEGARRA, A.; TSURUTA, S.; LAWLOR, T. J. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, v. 93, p. 743-752, 2010.

CALUS, M. P. L. Genomic breeding value prediction: methods and procedures. **Animal**, v. 42, p. 157-164, 2010.

CALUS, M. P. L.; VEERKAMP, R. F. Accuracy of multi-trait genomic selection using different methods. **Genetic Selection Evolution**, v. 43, n. 26, 2011.

CHEN, C. Y.; MISZTAL, I.; AGUILAR, I., TSURUTA, S.; MEUWISSEN, T. H. E.; AGGREY, S. E.; WING, T.; MUIR, W. M. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. **Journal of Animal Science**, v. 89, p. 23–28, 2011.

COSTA, R. B. **Associação e Seleção Genômica para características relacionadas à eficiência reprodutiva de fêmeas da raça Nelore**. 2014. Tese (Doutorado em Genética e Melhoramento Animal) - Faculdade de Ciências Agrárias e Veterinária, Universidade Estadual Paulista “Júlio de Mesquita Filho”, Jaboticabal, SP, 2014.

CHRISTENSEN, O. F.; MADSEN, P.; NIELSEN, B.; OSTERSEN, T; SU, G. Single-step methods for genomic evaluation in pigs. **Animal**, v. 6, p. 1565–1571, 2012.

CLARKE, S. M.; HENRY, H. M.; DODDS, K. G.; JOWETT, T. W. D.; MANLEY, T. R.; ANDERSON, R. M.; MCEWAN, J. C. A high throughput single nucleotide polymorphism multiplex assay for parentage assignment in New Zealand sheep, **PLOS ONE**, v. 9, p. 1-11, 2014.

DAETWYLER, H. D.; PONG-WONG, R.; VILLANUEVA, B.; WOOLLIAMS, J. A. The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, v. 185, p. 1021-1031, 2010.

DE LOS CAMPOS, G., GIANOLA, D., ROSA, G. J. M., WEIGEL, K. A., CROSSA, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. **Genetic Research**, v.92, p.295-398, 2010a.

DE LOS CAMPOS, G., GIANOLA, D., ROSA, G. J. M., WEIGEL, K. A., VAZQUEZ, A. I., ALLISON, D. B. Semi- Parametric Marker-enabled Prediction of Genetic Values using Reproducing Kernel Hilbert Spaces methods. **In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production**. Leipzig, Germany, 2010b.

DE LOS CAMPOS, G.; HICKEY, J. M.; PONG-WONG, R.; DAETWYLER, H. D.; CALUS, M. P. L. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, v.193, p.327-345, 2013.

FERNANDES JÚNIOR, G. A.; ROSA, G. J. M.; VALENTE, B. D.; CARVALHEIRO, R.; BALDI, F.; GARCIA, D.A.; GORDO, D.G.M.; ESPIGOLAN, R.; TAKADA, L.; TONUSSI, R. L.; ANDRADE, W. B. F.; MAGALHÃES, A. F. B.; CHARDULO, L. A. L.; TONHATI, H.; ALBUQUERQUE, L. G. Genomic prediction of breeding values for carcass traits in Nellore cattle. **Genetics Selection Evolution**, v.48, n.7, 2016.

GIANOLA, D., FERNANDO, R. L., STELLA, A. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics**, v.173, p.1761-1776, 2006.

GIANOLA, D.; VAN KAAM, B. C. H. M. Reproducing kernel Hilbert spaces regression methods for genomic prediction of quantitative traits. **Genetics**, v.178, p.2289-2303, 2008.

GODDARD, M. E.; HAYES, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Reviews**: Genetics, v. 10, p.381-391, 2009.

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of longterm response. **Genetica**, v. 136, p. 245-257, 2009.

GONZÁLEZ-RECIO, O.; GIANOLA, D.; LONG, N.; WEIGEL, K. A.; ROSA, G. J. M.; AVENDAÑO, S. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. **Genetics**, v.178, p.2305-2313, 2008.

LEGARRA, A.; AGUILAR, I.; MISZTAL, I. A relationship matrix including full pedigree and genomic information. **Journal of Dairy Science**, v. 92, p. 4656-4663, 2009.

LETTRE, G. Recent progress in the study of the genetics of height**. Human Genetics**, v. 129, p. 465-472, 2011.

LOURENCO D. A.; MISZTAL, I.; TSURUTA , S.; AGUILAR, I.; EZRA, E.; RON, M.; SHIRAK, A.; WELLER, J. I. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. **Journal of Dairy Science**, v. 97, p. 1742-1752, 2014.

MEUWISSEN, T.H. E.; HAYES, B.J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker map. **Genetics**, v.157, p.1819-1829, 2001.

MISZTAL, I.; LEGARRA, A.; AGUILAR, I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. **Journal of Dairy Science**, v.92, n.9, p.4648-4655, 2009.

MISZTAL, I., TSURUTA, S., STRABEL, T., AUVRAY, B., DRUET, T., LEE, D. H. BLUPF90 and related programs (BGF90). **In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production**, Montpellier, 2002.

MOROTA, G.; BODDHIREDDY, P.; VUKASINOVIC, N.; GIANOLA, D.; DENISE, S. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. **Frontiers in Genetics**, v.5, p.1-9, 2014.

MOSER, G.; TIER, B.; CRUMP, R. E.; KHATKAR, M. S.; RAADSMA, H. W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome wide SNP markers. **Genetics Selection Evolution**, v. 41, n. 56, 2009.

ONOGI, A.; KOMATSU, T.; SHOJI, N.; SIMIZU, K.; KUROGI, K.; YASUMORI, T.; TOGASHI, K.; IWATA, H. Genomic prediction in Japanese Black cattle: application of a single-step approach to beef cattle. **Journal of Animal Science**, v. 95, p. 1931-1938, 2014.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-Wide Regression and Prediction with the BGLR Statistical Package. **Genetics**, v.198, p.483-495, 2014.

PRYCE, J. E.; ARIAS, J.; BOWMAN, P. J.; DAVIS, S. R.; MACDONALD, K. A.; WAGHORN, G. C.; WALES, W. J. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. **Journal of Dairy Science**, v. 95, p. 2108-2119, 2012.

RAFTERY, A. E.; LEWIS, S. M. Comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. **Statistical Science**, v.7, p.493-497, 1992.

R Development Core Team (2016). **R: A language and environment for statistical computing**. Vienna, Austria. Available at: <http://www.R-project.org>. Accessed: January 12, 2017.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F. S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, v.15: 478-10.1186/1471-2164-15-478, 2014.

TSURUTA, S.; AGUILAR, I.; MISZTAL, I.; LAWLOR, T. J. Multiple trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. **Journal of Dairy Science**, v. 94, p. 4198–4204, 2011.

VAN ARENDONK, J. A. M.; TIER, B.; KINGHORN, B. P. Use of multiple genetic markers in prediction of breeding values. **Genetics**, v. 137, p.319-329, 1994.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v.91, p. 4414-4423, 2008.

VANRADEN, P. M.; VAN TASSELL, C. P.; WIGGANS G. R.; SONSTEGARD T. S.; SCHNABEL, R. D.; TAYLOR, J. F.; SCHENKEL, F. S. Invited review: reliability of genomic predictions for North American Holstein bulls. **Journal of Dairy Science**, v. 92, p.16-24, 2009.

WINKELMAN, A. M., JOHNSON, D. L., HARRIS, B. L. Application of genomic evaluation to dairy cattle in New Zealand. **Journal of Dairy Science**, v.98, p.659-675, 2015.

WHEELER, T. L.; KOOHMARAIE, M.; SHACKELFORD, S. D. **Standardized Warner-Bratzler shear force procedures for meat tenderness measurement**. Clay Center: Roman L. Hruska U. S. MARC. USDA, 7p., 1995.

CHAPTER 3 – Application of parametric and semi-parametric models to evaluate the accuracy of prediction using cattle simulated data

ABSTRACT – The objective of this study was to evaluate the accuracy of genomic predictions applying parametric and semi-parametric statistical models in a simulated cattle population. Genotypes, pedigree, and phenotypes for four traits A, B, C and D were simulated using heritabilities based on real data (0.09, 0.12, 0.36 and 0.39 for each trait, respectively). The simulated genome consisted of 735,293 markers and 1,000 QTLs randomly distributed over 29 pairs of autosomes, with length varying from 40 cM to 146 cM, totaling 2,333 cM. It was assumed that QTLs explained 100% of genetic variance. All markers were bi-allelic and for QTLs the amount of alleles per loci randomly ranged from 2 to 4. Considering Minor Allele Frequencies greater or equal to 0.01, a total of 430,000 markers were randomly selected from the last generation of the historical population to generate genotypic data for the selection population. The phenotypes were generated by adding residuals, randomly drawn from a normal distribution with mean equal to zero, to the true breeding values and all simulation process was replicated 10 times. Simulated phenotypes and genotypes of 2,600, 1,800, 4,500 and 2,600 animals were randomly selected from the last four generations for traits A, B, C and D, respectively. Methods of analysis were genomic BLUP (GBLUP), single-step GBLUP (ssGBLUP), Bayesian LASSO (BL) and the semi-parametric approaches Reproducing Kernel Hilbert Spaces (RKHS) regression and Kernel Averaging (KA). Models were compared in terms of their accuracy of prediction and mean squared error. Accuracy was quantified using correlations between the predicted genomic breeding value and true breeding values simulated for the generations of 12 to 15. Results were the mean of the 10 replicates generated in the simulation process. The average linkage disequilibrium, measured between pairs of adjacent markers for all simulated traits was 0.21 for recent generations (12, 13 and 14), and 0.22 for generation 15. The accuracy for simulated traits A, B, C and D ranged from 0.43 to 0.44, 0.47 to 0.48, 0.80 to 0.82 and 0.72 to 0.73, respectively. Different genomic selection methodologies implemented in this study showed similar accuracies of prediction, and the optimal method was sometimes trait dependent. In general, RKHS regressions were preferable in terms of accuracy of prediction and provided smallest mean squared error estimates compared to other models.

Keywords: beef cattle, genomic evaluation, RKHS regression, simulated genome

1. INTRODUCTION

The availability and knowledge of the genome of livestock populations bring a new and complementary source of information to that previously available for selection. In this case, information is obtained for a large number of single nucleotide polymorphisms (SNP) markers. However, only few thousands of individuals are genotyped leading to the so called curse of dimensionality problem also known as the “large p, small n” problem (DE LOS CAMPOS et al. 2010). This scenario generates an over-parameterization in traditional methods (MONTERO, 2013). Therefore it has been necessary to develop, test and implement new methods in the genome-enhanced evaluations.

Different approaches are currently used for estimating genomic values, and it is essential to assess the performance of diverse methodologies and to identify methods that provide the greatest accuracy of prediction in a given population. Therefore, genomic prediction methods can be categorized considering parametric (Genomic Best Linear Unbiased Predictor - GBLUP, Single-Step Genomic Best Linear Unbiased Predictor - ssGBLUP, Bayesian LASSO - BL) and semi-parametric (Reproducing Kernel Hilbert Spaces regression - RKHS) approaches.

The GBLUP method (VANRADEN, 2008) is similar to the traditional BLUP evaluations described by Henderson (1975). However, it uses a genomic relationship matrix built from molecular information instead of traditional pedigree relationship matrix. Individuals sharing identical by state genotype for a larger number of markers are expected to be genetically more similar and will have higher values in the corresponding cells of the matrix. The ssGBLUP method was proposed by Misztal et al. (2009) and consists of an evaluation where pedigree relationship is reinforced with contributions from the genomic relationship matrix, generating a matrix called **H**. From a Bayesian point of view, the BL method (PARK; CASELLA, 2008) considers a Laplace double exponential prior distribution on the markers effects and performs higher shrinkage on the marker coefficients estimates towards zero, producing an effect similar to the pre-selection of covariates (DE LOS CAMPOS et al. 2010).

As an alternative to parametric methods, Gianola, Fernando and Stella (2006) and Gianola and van Kaam (2008) proposed a semi-parametric method for the genomic evaluations. These models are more attractive than parametric methods because can, potentially, capture multiple and complex interactions that may exist in the biological and metabolic systems without posing richly parametrized structures that require making strong distribution and genetic architecture assumptions a priori. (MOROTA et al. 2013)

In literature, simulation studies mimicking a cattle population and exploring models with semi-parametric approach are scarce. Thus, the objective of this study was to evaluate the accuracy of genomic predictions applying parametric and semi-parametric statistical models in a simulated cattle population.

2. MATERIAL AND METHODS

The QMSim software version 1.10 (SARGOLZAEI; SCHENKEL, 2009) was employed to simulate pedigree, phenotypes and genotypes in a way that the simulated population had an extent and pattern of linkage disequilibrium (LD) consistent with that verified in *Bos indicus* beef cattle populations. The simulations were carried out using parameters related to the historical generations, defined in a similar way as in Brito et al. (2011).

2.1. Simulation of Population Structure

A historical population was created simulating 1,000 generations with effective population size constant of 1,000 animals. Hereafter, 1,020 historical generations were simulated in which the number of animals was gradually reduced from 1,000 to 200, ensuring that mutation-drift equilibrium was established and initial linkage disequilibrium was generated. For these procedures, the number of individuals of each sex was the same and the mating system was based on random union of gametes.

The population was expanded selecting randomly 100 founder males and 100 founder females from the last generation of the historical population. Posteriorly, eight generations were simulated with five offspring per dam and an exponential growth of the number of dams, also under random union of gametes and without selection.

At the end of expansion process, 400 males and 10,000 females from the last generation of expanded population were randomly selected, including the founder animals. The recent population was spanned over 15 generations and the selected males and females from each generation were randomly mated. This step reproduced a selected cattle population, with one offspring per dam and about 50% of male progeny. Replacement rate of sires and dams was kept constant over generations at 60% and 20%, respectively. The rate of unknown sires was approximately 40% over the fifteen recent generations. At the end of this process, the recent population comprised 160,400 animals, including sires, dams and their offspring.

2.2. Simulation of Genome

The simulated genome consisted of 735,293 markers and 1,000 QTLs randomly distributed over 29 pairs of autosomes, with length varying from 40 cM to 146 cM, thus identical to the real bovine genome based on Btau_4.6.1 assembling (SNELLING et al., 2007) totaling 2,333 cM. It was assumed that QTLs explain 100% of genetic variance. The number of markers and QTLs per chromosome ranged from 12,931 to 46,495 and from 121 to 438, respectively. All markers were bi-allelic and for QTLs the amount of alleles per loci randomly ranged from 2 to 4 considering that allele effects were sampled from a gamma distribution with a shape parameter equal to 0.4 (HAYES; GODDARD, 2001).

In order to establish mutation-drift equilibrium in historical generations, the rates of marker genotyping error and recurrent mutation (for markers and QTL) were 0.005 and 10^{-4}, respectively. Considering Minor Allele Frequencies (MAF) greater or

equal to 0.01, a total of 430,000 markers were randomly selected from the last generation of the historical population to generate genotypic data for the selection population. The true breeding value (TBV) of each individual was the sum of the QTL allele substitution effects. The phenotypes were generated by adding residuals, randomly drawn from a normal distribution with mean equal to zero, to the TBVs. All the simulation process was replicated 10 times. The parameters of the simulation processes are showed in the Table 1.

Table 1. Options and parameters used in the data simulation processes

Options for simulation process	Parameters
Historical Population (HP)	
Phase 1	1000 generations / 1000 animals
Phase 2	1020 generations / 1000 to 200 animals
Expanded Population (EP)	
Number of founder males from HP (Phase 2)	100
Number of founder females from HP (Phase 2)	100
Number of generations	8
Number of offspring per dam	5
Recent Population (RP)	
Number of founder males from EP	400
Number of founder females from EP	10000
Number of generations	15
Number of offspring per dam	1
Proportion of male progeny	0.5
Mating design	random
Replacement ratio for sires	0.6
Replacement ratio for dams	0.2
Selection/culling design	EBV
Breeding value estimating method	BLUP animal method
Rate of missing sire	0.4
Heritability of the trait	0.09, 0.12, 0.36 and 0.39
Phenotypic variance	1.0
Genome	
Number of chromosomes	29
Total length	2333cM
Number of markers	735293
Marker distribution	evenly spaced
Number of QTL	1000
QTL distribution	random
MAF for markers	0.01
MAF for QTL	0.01
Additive allelic effects for markers	neutral
Additive allelic effects for QTL	Gamma distribution (shape = 0.4)
Rate of missing marker genotypes	0
Rate of marker genotyping error	0.005
Rate of recurrent mutation	0.0001

2.3. Simulated traits

Four traits were simulated, each one characterized by a level of heritability, sex in which phenotypes were measured and number of animals with phenotypic and genotypic information (Table 2).

Table 2. Definition of the simulated traits, phenotypes and genotypes

Trait	Heritability	N_{total}	N_{gen}	Sex
A	0.09	3500	2600	Male
B	0.12	10000	1800	Female
C	0.36	40000	4500	Male/Female
D	0.39	3500	2600	Male

N_{total} = number of simulated animals with phenotypes; N_{gen} = number of simulated animals with phenotypes and genotypes; Sex = sex in which simulated phenotypes and genotypes were available

The traits heritabilities, number of animals and genotypes were chosen in order to simulate a set of data similar to the real data structure described in Chapter 2 of the present work. Taking this into account, the traits A and B mimicked a low heritable selection criteria, with B only expressed in females, representing criteria similar to reproductive traits. Traits C and D simulated a moderately heritable selection criteria where C is expressed in both sexes, indicating selection criteria similar to growth-related traits. All information of animals used in the analyzes were randomly selected from the last four generations of simulated recent population (generations 12 to 15).

2.4. Statistical Models

The analyses were performed using the BLUPF90 family programs (MISZTAL et al., 2002) available at http://nce.ads.uga.edu/wiki/doku.php, the Bayesian

Generalized Linear Regression (BGLR) package (PÉREZ; DE LOS CAMPOS, 2014) and R software (R Development Core Team, 2017).

2.4.1. Genomic Best Linear Unbiased Predictor (GBLUP)

The GBLUP model is similar to BLUP using a genomic relationship matrix (**G**) instead **A**, producing direct genomic value (DGV) based on SNP effects. Solutions from GBLUP can be obtained with the model showed below:

$$\mathbf{y = 1}\mu \mathbf{+ Zg + e,}$$

where **y** is the vector of simulated phenotypes, μ is the overall mean, **1** is a vector of ones, **Z** is an incidence matrix of markers effects, **g** is a vector of marker effects, and **e** is a vector of residual effects. It was assumed $\mathbf{g} \sim N(0, \mathbf{G}\sigma^2_g)$, where σ^2_g is the variance of markers and **G** is the genomic relationship matrix. Random residuals were assumed $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$, where **I** is an identity matrix and σ^2_e is the residual variance.

According to VanRaden (2008), the **G** matrix can be obtained from at least three ways. For this study, we chose the following:

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{\mathbf{2}\sum_{\mathbf{j=i}}^{\mathbf{m}} \mathbf{p_j}\,(\mathbf{1} - \mathbf{p_j})},$$

where **M** is a matrix of marker alleles with n lines (n = number of genotyped animals) and m columns (m = number of markers), and **P** is a matrix containing: $2(\mathbf{p_j}\text{-}0.5)$, with

$\mathbf{p_j}$ being the frequency of the second allele. Elements of **M** are set to 0 and 2 for both homozygous and to 1 for the heterozygous genotype.

2.4.2. Single-step Genomic Best Linear Unbiased Predictor (ssGBLUP)

The model used in ssGBLUP consists in combining **A** and **G** into a single matrix (**H**). Thus, the inverse of the numerator relationship matrix ($\mathbf{A^{-1}}$) was replaced by $\mathbf{H^{-1}}$, which combines pedigree and genomic information.

$$\mathbf{y = Zu + e},$$

where **y** is the vector of simulated phenotypes, *β* is the vector of fixed effects, **u** is the vector of direct additive genetic effects, and **Z** is an incidence matrix. Considering an infinitesimal model, var(**u**) = $\mathbf{H\sigma_u^2}$, where **H** is a combined relationship matrix that integrates the genomically derived relationships (**G** matrix) with population-based pedigree relationships (**A** matrix) and $\boldsymbol{\sigma_u^2}$ is the additive genetic variance, and var(**e**) = $\mathbf{I\sigma_e^2}$ where $\boldsymbol{\sigma_e^2}$ is the residual variance.

Matrix $\mathbf{H^{-1}}$ can be obtained as follows (AGUILAR et al., 2010):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix},$$

where $\mathbf{G^{-1}}$ is the inverse of genomic relationship matrix and $\mathbf{A_{22}^{-1}}$ is the inverse of pedigree-based numerator relationship matrix for genotyped animals.

2.4.3. Bayesian LASSO (BL)

The Bayesian counterpart of the LASSO model (PARK; CASELLA, 2008; DE LOS CAMPOS et al. 2009) was used to estimate SNP coefficients in the training population. The general model for genomic prediction using BL model was expressed in matrix notation:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Wg} + \mathbf{e},$$

where **y** is the vector of simulated phenotypes, μ is the overall mean, **1** is a vector of ones, **g** is the vector of marker effects, **W** contains the genotype (coded 0 = AA, 1 = AB and 2 = BB) for each individual and each marker, and **e** is the vector of residual effects. According to De Los Campos et al. (2013) and Pérez and De Los Campos (2014), the marginal prior distribution assigned to **g** in the BGLR package, is a double exponential function, which includes a parameter λ^2 that was treated as unknown, with a prior distribution $\lambda^2 \sim$ gamma(*r, s*). Moreover, the BGLR package considers by default that $r = 1.1$ and calculates the scale parameter *s* based on the "prior" R^2 of the model.

Samples from posterior distributions were obtained by the Gibbs sampler based on 70,000 Monte Carlo Markov Chain (MCMC) samples with the first 35,000 discarded as burn-in. After burn-in, samples were thinned at a rate of 10. Convergence of the chain was checked by visual inspection using the Coda (RAFTERY; LEWIS, 1992) package and inferences on the parameters were made on the mean posterior estimates after burn-in.

2.4.4. Reproducing Kernel Hilbert Spaces (RKHS) regression

As described in Gianola and van Kaam, the RKHS regressions with a semi-parametric approach can be formulated as follows:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{K}_\theta\boldsymbol{\alpha} + \mathbf{e},$$

where **y** is the vector of simulated phenotypes, μ is the overall mean, **1** is a vector of ones. The non-parametric term is given by $\mathbf{K}_\theta\boldsymbol{\alpha}$, where $\mathbf{K}_\theta$ is a positive definite matrix of kernels, dependent on a bandwidth parameter (θ), **α** is a vector of non-parametric coefficients and **e** is the vector of residuals effects, with **α** and **e** assumed to be independently distributed.

Thus, let **M** denote the *n* x *p* matrix of genotypes (coded 0 = AA, 1 = AB and 2 = BB) for *n* genotyped animals and *p* SNP markers, **m** are the elements of **M** and θ is a bandwidth parameter (GIANOLA; VAN KAAM, 2008). The matrix **K** is based on Euclidean distance matrix ($\mathbf{EDM}$) in a Gaussian kernel (WINKELMAN; JOHNSON; HARRIS, 2015):

$$\mathbf{K} = \exp(-\theta * \mathbf{EDM}),$$

where

$$\mathbf{EDM}_{ij} = \sum_k \left(\mathbf{m}_{ik} - \mathbf{m}_{jk}\right)^2,$$

Therefore, the $\mathbf{EDM}_{ij}$ matrix is the squared Euclidean distance between individuals i and j calculated based on their genotypes for SNP markers. The matrix $\mathbf{EDM}$ can be conveniently derived using:

$$\mathbf{EDM}_{ij} = \sqrt{\mathbf{S}_i + \mathbf{S}_j - 2\mathbf{MM}'_{ij}},$$

where

$$\mathbf{S} = \mathrm{Diag}[\mathbf{MM}'].$$

The parameter θ in this simulation study was set for a grid of values equal to 0.1, 0.2, 0.5, and ranging from 1 to 10. For RKHS regression, a MCMC with 150,000 samples was run and the first 50,000 were discarded. Subsequently, 100,000 samples were obtained and thinned at a rate of 50, resulting in 2000 mildly correlated samples for posterior inference.

2.4.5. Kernel Averaging (KA)

In this simulation work, the KA was fitted using two RKHS: the first one with the lowest value for bandwidth parameter in the grid sequence ($\theta_L = 0.1$) and the second with the highest value ($\theta_H = 10$). The "multikernel model" was described with as many random effects as kernels (DE LOS CAMPOS et al., 2010):

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{K}_{\boldsymbol{\theta}_L}\boldsymbol{\alpha}_L + \mathbf{K}_{\boldsymbol{\theta}_H}\boldsymbol{\alpha}_H + \mathbf{e},$$

where **y** is the vector of simulated phenotypes, μ is the overall mean, **1** is a vector of ones. The non-parametric terms is given by $\mathbf{K}_{\boldsymbol{\theta}_\mathbf{L}}\boldsymbol{\alpha}_\mathbf{L}$ and $\mathbf{K}_{\boldsymbol{\theta}_\mathbf{H}}\boldsymbol{\alpha}_\mathbf{H}$, where $\mathbf{K}_{\boldsymbol{\theta}_\mathbf{L}}$ and $\mathbf{K}_{\boldsymbol{\theta}_\mathbf{H}}$ are positive definite matrices of kernels, dependent of two bandwidth parameters ($\boldsymbol{\theta}_\mathbf{L} = \mathbf{0.1}$ and $\boldsymbol{\theta}_\mathbf{H} = \mathbf{10}$), $\boldsymbol{\alpha}_\mathbf{L}$ and $\boldsymbol{\alpha}_\mathbf{H}$ are vectors of non-parametric coefficients and **e** is the vector of residuals effects, with $\boldsymbol{\alpha}_\mathbf{L}$ and $\boldsymbol{\alpha}_\mathbf{H}$ and **e** assumed to be independently distributed. The **K** matrix is calculed as described in topic 3.2.5. For KA, a MCMC with 300,000 samples was run and the first 50,000 were discarded. Subsequently, 250,000 samples were obtained and thinned at a rate of 50, resulting in 5000 mildly correlated samples for posterior inference.

2.4.6. Prediction of genomic values

As proposed by Meuwissen, Hayes and Goddard (2001), the GEBVs of the animals, in this case for parametric models, will be a function of SNP markers' genotypes and effects, using the following equation:

$$\mathrm{GEBV_i} = \sum_{\mathrm{j=1}}^{\mathrm{p}} \mathrm{w_{ij}}\,\hat{\mathrm{g}}_\mathrm{j},$$

where *p* is the number of SNP; w_{ij} is the genotype of animal *i* for SNP *j* (coded as 0, 1 or 2), and $\hat{g}_j$ is the estimated SNP substitution effect for SNP *j* that was estimated from the training population.

The prediction of GEBVs of animals using RKHS regression and KA was made using the following equation (MOROTA et al. 2014):

$$\mathbf{GEBV} = \mathbf{K}^*(\theta)\hat{\boldsymbol{\alpha}},$$

where $\mathbf{K}^{*}(\theta)$ is a matrix with the distance kernel between genotypes with a bandwidth parameter (θ), and $\hat{\boldsymbol{\alpha}}$ is the estimated vector of non-parametric regression coefficients in the training set.

2.4.7. Criteria for the comparison of models

Models were compared in terms of their accuracy of prediction (ACC) and mean squared error (MSE). The ACC is a common measurement of predictive ability in genetic prediction studies (GODDARD; HAYES, 2007; LUAN et al., 2009) and was quantified using correlations between the predicted genomic breeding value (GEBV) and true breeding values (TBV) simulated for the generations of 12 to 15. The results were the mean of the 10 replicates generated in the simulation process.

The MSE was used as a measure of the prediction ability of the models, which combines quality assessment in terms of variance and bias of predictions and can be calculated as follows:

$$\mathbf{MSE} = \sum(\mathbf{TBV} - \mathbf{GEBV})^{\mathbf{2}} \,/\, \mathbf{N}$$

where $\boldsymbol{N}$ is the number of animals for each trait.

3. RESULTS AND DISCUSSION

The average linkage disequilibrium (r^2), measured between pairs of adjacent markers for all simulated traits was 0.21 for recent generations 12, 13 and 14, and 0.22 for generation 15. Indeed, all chromosomes were simulated using the same parameters for all traits, and therefore, differences of LD between them were not expected. For Nelore cattle, values of r^2 of 0.17 (ESPIGOLAN et al. 2013), 0.29 (NEVES et al. 2014) and 0.31 (FERNANDES JÚNIOR et al. 2016) have been

reported. Taking that into account, LD values estimated in the present study were similar and are within the range described in the literature for real Nelore populations.

The accuracies of prediction of the medium-heritability simulated traits (Traits C and D) were greater than the accuracy of the low-heritability traits (Traits A and B) as shown in Figure 1 for the bandwidth parameters used in RKHS regression and for GBLUP, ssGBLUP and Bayesian Lasso in Table 3.

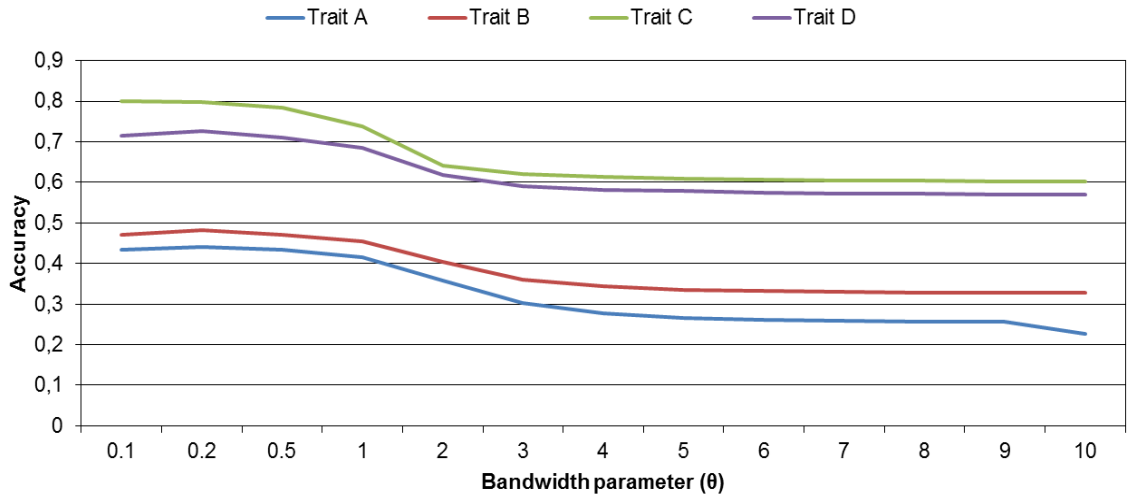

Figure 1. Accuracy of prediction across bandwidths parameters (θ) for simulated traits

Considering values for θ ranging from 0.1 to 0.5, the accuracy remains constant for all traits and quickly decreases with θ greater than 1. Similar behavior was observed in Chapter 2 with real data set and in a research conducted by Morota et al. (2013). The authors worked with Holstein and wheat data sets and evaluated the predictive performance of Gaussian Kernels (based on Euclidean Distance Matrix) for several bandwidth parameters, and they found that values for θ less than 1 had the best predictive correlation and lowest mean squared error. According to the authors, as the θ increases, the kernel matrix approaches zero, producing a “sharp” or “local” kernel. On the other hand, as θ tends to zero, the kernel matrix contains values very close to one, that is, a situation where the two animals “match” perfectly,

providing a “global” kernel. Figure 2 illustrates the effect of the bandwidth parameter on the Kernel matrix.

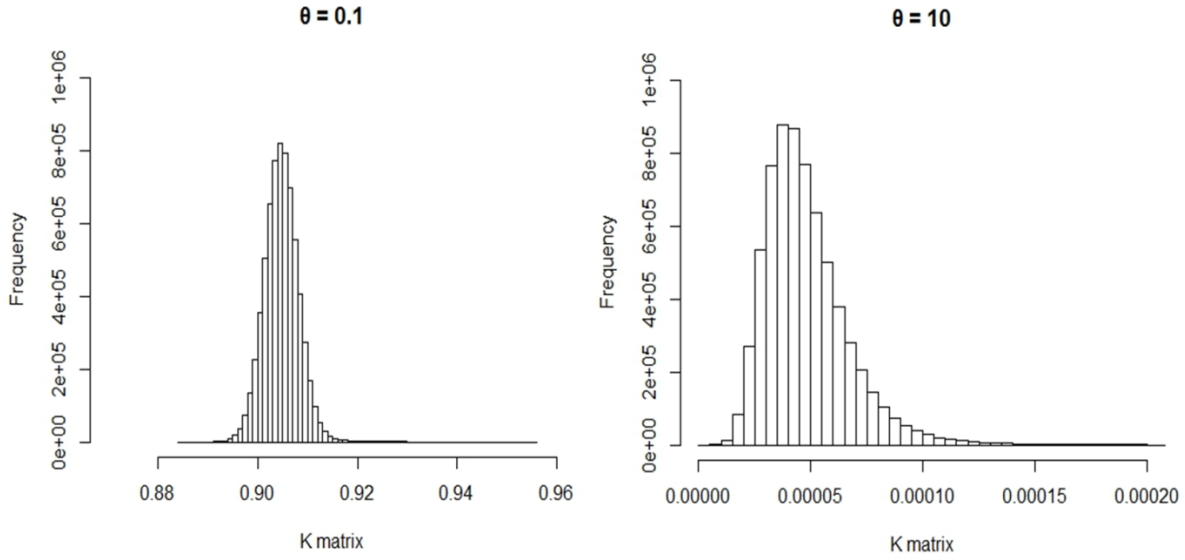

Figure 2. The influence of the bandwidth parameter on the off-diagonal elements of K matrix. On the left, θ=0.1 and right θ=10

As previously mentioned, with increasing of the heritability, the correlation between TBV and GEBV was increased and the difference between trait A (h^2 = 0.09) and C (h^2 = 0.36) was higher compared to trait A to D (h^2 = 0.39), perhaps because C was simulated for both sexes and has a greater number of genotyped animals, resulting in more reliable accuracies, even with heritability slightly lower than that for D. These results are in agreement with the works carried out by Daetwyler, Villanueva and Woolliams (2008), Howard, Carriquiry and Beavis (2014) and Atefi, Shadparvar and Hossein-Zadeh (2016) who have found that higher accuracies are achieved for high heritability traits, since there is a high contribution of gene effects for phenotypic variation.

Table 3. Accuracy of prediction (ACC), standard deviation (SD) and predicted mean squared error (MSE) for simulated traits obtained with different models and the average of ten simulated population. The best prediction model for each trait is in bold

Trait	Heritability	Models	ACC ± SD	MSE
A	0.09	GBLUP	0.43 ± 0.027	0.068
		ssGBLUP	0.43 ± 0.024	0.075
		BL	0.43 ± 0.029	0.080
		RKHS $_{(\theta=0.2)}$	**0.44 ± 0.026**	**0.066**
		KA	0.44 ± 0.022	0.071
B	0.12	GBLUP	0.47 ± 0.023	0.087
		ssGBLUP	0.47 ± 0.021	0.108
		BL	0.48 ± 0.024	0.102
		RKHS $_{(\theta=0.2)}$	**0.48 ± 0.023**	**0.085**
		KA	0.48 ± 0.023	0.089
C	0.36	GBLUP	0.80 ± 0.012	0.141
		ssGBLUP	**0.82 ± 0.010**	**0.118**
		BL	0.80 ± 0.012	0.138
		RKHS $_{(\theta=0.1)}$	0.80 ± 0.012	0.131
		KA	0.80 ± 0.015	0.140
D	0.39	GBLUP	0.72 ± 0.022	0.149
		ssGBLUP	0.72 ± 0.027	0.152
		BL	0.72 ± 0.026	0.145
		RKHS $_{(\theta=0.2)}$	**0.73 ± 0.022**	**0.142**
		KA	0.73 ± 0.028	0.148

θ = bandwidth parameter for kernel regression method

The accuracy for simulated traits A, B, C and D ranged from 0.43 to 0.44, 0.47 to 0.48, 0.80 to 0.82 and 0.72 to 0.73, respectively (Table 3). Among the investigated methods, ssGBLUP provided the highest accuracy of prediction for trait C (0.82). This result is expected, mainly for C, because the simulated trait has the highest amount of phenotypes (40,000 animals) and genotypes (4500), which makes the inverse of **H** matrix more informative and robust. Our results agree with those from Zhang et al. (2016) who worked with simulated datasets consisted of phenotypes for 13,000 animals, including 1540 animals genotyped for 45,000 SNP and compared the accuracy using GBLUP, Bayesian methods and a weighted ssGBLUP. The authors

reported higher accuracies obtained with ssGBLUP compared to those from GBLUP and Bayesian methods.

For traits A and D, RKHS regression and KA showed greater accuracies (0.44 for A and 0.73 for D) and outperformed GBLUP, ssGBLUP and BL. Similar results were described by Howard, Carriquiry and Beavis (2014), who investigated several parametric (ridge regression, Bayesian Lasso, Bayes A, Bayes C, GBLUP) and nonparametric methods (Nadaraya-Watson estimator, RKHS regression, support vector machine regression, neural networks) in a simulation study and they described the superior ability of nonparametric methods to accurately predict phenotypes and hypothesized that these methods also will enable more accurate predictions of individual genotypic value.

The BL, RKHS ($\theta = 1.0$) regression and KA showed slightly greater accuracy (0.39) for trait B compared to GBLUP (0.38) and ssGBLUP (0.38). These results differ from those recently reported by Atefi, Shadparvar and Hossein-Zadeh (2016). The authors showed in a simulation study that for models with only additive gene effects, RKHS method did not perform better than parametric methods such as Bayes A and BL. However, in their simulation only 500, 750 or 1000 bi-allelic SNP markers were considered and, in the implementation of RKHS regression, the bandwidth parameter, which clearly affect the accuracies and MSE, was chosen using arbitrary values.

The parameters of the genomic structure of the population simulated in this study showed that, although there are no relevant differences in terms of accuracy of prediction between all investigated methods, except for trait C, the RKHS regression provided smallest MSE estimates. Moreover, there was consistency between the accuracies and the MSE, in the sense that with the highest accuracies had the smallest MSE, which can be considered as a measure of the predictive variance and bias of predictions.

4. CONCLUSIONS

Different genomic selection methodologies implemented in this study showed similar accuracies of prediction, and the optimal method was sometimes trait dependent. In general, RKHS regressions were preferable in terms of accuracy of prediction and provided smallest MSE estimates compared to other models.

5. REFERENCES

AGUILAR, I.; MISZTAL, I.; JOHNSON, D. L.; LEGARRA, A.; TSURUTA, S.; LAWLOR, T. J. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, v. 93, p. 743-752, 2010.

ATEFI, A.; SHADPARVAR, A. A.; HOSSEIN-ZADEH, N. G. Comparison of whole genome prediction accuracy across generations using parametric and semi parametric methods. **Acta Scientiarum**, v.38, p.447-453, 2016.

BRITO, F. V.; NETO, J. B.; SARGOLZAEI, M.; COBUCI, J. A.; SCHENKEL, F. S. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. **BMC Genetics**, v. 12, p. 80-89, 2011.

DAETWYLER, H. D.; VILLANUEVA, B.; WOOLLIAMS, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. **PLOS ONE**, v.3(10), p.3395, 2008.

DE LOS CAMPOS, G.; GIANOLA, D.; ROSA, G. J. M.; WEIGEL, K. A.; CROSSA, J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. **Genetic Research**, v.92, p.295-398, 2010.

DE LOS CAMPOS, G.; HICKEY, J. M.; PONG-WONG, R.; DAETWYLER, H. D.; CALUS, M. P. L. Whole-genome regression and prediction methods applied to plant and animal breeding. **Genetics**, v.193, p.327-345, 2013.

DE LOS CAMPOS, G., NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J. M. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. **Genetics**, V.182, P.375–385, 2009.

ESPIGOLAN, R.; BALDI, F.; BOLIGON, A. A.; SOUZA, F. P. R.; GORDO, D. G. M.; TONUSSI, R. L.; CARDOSO, D. F.; OLIVEIRA, H. N.; TONHATI, H.; SARGOLZAEI, M.; SCHENKEL, F. S.; CARVALHEIRO, R.; FERRO, J. A.; ALBUQUERQUE, L. G. Study of whole genome linkage disequilibrium in Nellore cattle. **BMC Genomics**, v.14, p.305, 2013.

FERNANDES JÚNIOR, G. A.; ROSA, G. J. M.; VALENTE, B. D.; CARVALHEIRO, R.; BALDI, F.; GARCIA, D. A.; GORDO, D. G. M.; ESPIGOLAN, R.; TAKADA, L.; TONUSSI, R. L.; ANDRADE, W. B. F.; MAGALHÃES, A. F. B.; CHARDULO, L. A. L.; TONHATI, H.; ALBUQUERQUE, L. G. Genomic prediction of breeding values for carcass traits in Nellore cattle. **Genetics Selection Evolution**, v.48, n.7, 2016.

GIANOLA, D., FERNANDO, R. L., STELLA, A. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics**, v.173, p.1761-1776, 2006.

GIANOLA, D.; VAN KAAM, B. C. H. M. Reproducing kernel Hilbert spaces regression methods for genomic prediction of quantitative traits. **Genetics**, v.178, p.2289-2303, 2008.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, v. 124, p.323–330, 2007.

HAYES, B.; GODDARD, M. E. The distribution of the effects of genes affecting quantitative traits in livestock. **Genetics Selection Evolution**, v. 33, n.3, p.209-229, 2001.

HENDERSON, C. R. Best linear unbiased estimation and prediction under a selection model. **Biometrics**, v. 31, p. 423–447, 1975.

HOWARD, R.; CARRIQUIRY, A. L.; BEAVIS, W. D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **Genes, Genomes, Genetics,** v.4(6), p.1027-1046, 2014.

LUAN, T.; WOOLLIAMS, J. A.; LIEN, S.; KENT, M.; SVENDSEN, M.; MEUWISSEN, T. H. E. The accuracy of genomic selection in Norwegian red cattle assessed by cross- validation. **Genetics**, v. 183, p.1119–1126, 2009.

MEUWISSEN, T.H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker map. **Genetics**, v.157, p.1819-1829, 2001.

MISZTAL, I.; LEGARRA, A.; AGUILAR, I. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. **Journal of Dairy Science**, v.92, n.9, p.4648-4655, 2009.

MISZTAL, I., TSURUTA, S., STRABEL, T., AUVRAY, B., DRUET, T., LEE, D. H. BLUPF90 and related programs (BGF90). **In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production**, Montpellier, 2002.

MOROTA, G.; BODDHIREDDY, P.; VUKASINOVIC, N.; GIANOLA, D.; DENISE, S. Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits. **Frontiers in Genetics**, v.5, p.1-9, 2014.

MOROTA, G.; KOYAMA, M.; ROSA, G. J. M.; WEIGEL, K. A.; GIANOLA, D. Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. **Genetics Selection Evolution**, v. 45, n. 17, p.1-10, 2013.

MONTERO, J. A. J. **Genomic selection in small dairy cattle populations**. Ph.D Thesis - Polytechnic University of Valencia, Valencia, Spain, 2013.

NEVES, H. H. R.; CARVALHEIRO R.; O'BRIEN, A. M. P.; UTSUNOMIYA, Y. T.; CARMO, A. S.; SCHENKEL, F. S.; SÖLKNER, J.; MCEWAN, J. C.; VAN TASSELL, C. P.; COLE, J. B.; SILVA, MARCOS, V.G.B.; QUEIROZ, S. A.; SONSTEGARD, T. S.; GARCIA, J. F. Accuracy of genomic predictions in Bos indicus (Nellore) cattle. **Genetics Selection Evolution**, v. 46, n.17. 2014.

PARK, T.; CASELLA, G. The Bayesian Lasso. **Journal of the American Statistical Association,** v. 103, p. 681–686, 2008.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-Wide Regression and Prediction with the BGLR Statistical Package. **Genetics**, v.198, p.483-495, 2014.

R Development Core Team (2016). R: **A language and environment for statistical computing**. Vienna, Austria. Available at [http://www.R-project.org]. Accessed January 29, 2017.

RAFTERY, A. E.; LEWIS, S. M. Comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. **Statistical Science**, v.7, p.493-497, 1992.

SARGOLZAEI, M.; SCHENKEL, F. S. QMSim: a large-scale genome simulator for livestock. **Bioinformatics**, v.25, p.680-681, 2009.

SNELLING, W. M.; CHIU, R.; SCHEIN, J. E.; HOBBS, M.; ABBEY, C. A.; ADELSON, D. L.; AERTS, J,; BENNETT, G. L ; BOSDET, I. E.; BOUSSAHA, M.; BRAUNING, R.; CAETANO, A. R.; COSTA, M. M.; CRAWFORD, A. M.; DALRYMPLE, B. P.; EGGEN A.; VAN DER WIND, A. E.; FLORIOT, S.; GAUTIER, M.; GILL, C. A.; GREEN, R. D.; HOLT, R.; JANN, O.; JONES, S. J. M.; KAPPES, S. M., KEELE, J. W.; PONG, P. J.; LARKIN, M.; LEWIN, H. A.; MCEWAN, J. C.; MCKAY, S.; MARRA, M. A.; MATHEWSON, C. A.; MATUKUMALLI, L. K., MOORE, S. S.; MURDOCH, B.; NICHOLAS, F. W.; OSOEGAWA, R.; ROY, A.; SALIH, H.; SCHIBLER, L.; SCHNABEL, R. D.; L.; SILVERI, L.; SKOW, L. C.; SMITH, T. P. L.; SONSTEGARD, T. S.; TAYLOR, J. F.; TELLAM, R.; VAN TASSEL, C. P.; WILLIAMS, J. L., WOMACK, J. E.; WYE, N. H.; YANG, G.; ZHAO, S. for the International Bovine BAC Mapping Consortium: A physical map of the bovine genome. **Genome Biology**, 8:R165, 2007.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v.91, p. 4414-4423, 2008.

WINKELMAN, A. M., JOHNSON, D. L., HARRIS, B. L. Application of genomic evaluation to dairy cattle in New Zealand. **Journal of Dairy Science**, v.98, p.659-675, 2015.

ZHANG, X.; LOURENCO, D.; AGUILAR, I.; LEGARRA, A.; MISZTAL, I. Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS. **Frontiers in Genetics**, v.7, 2016.