UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE VETERINÁRIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS VETERINÁRIAS

Presente e futuro da análise de dados de fatores associados à soroprevalência da diarreia viral bovina

Pós-graduado: Gustavo Machado
Orientador: Luis Gustavo Corbellini
Co-orientadora: Luciana Neves Nunes

Porto Alegre, fevereiro de 2016

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE VETERINÁRIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS VETERINÁRIAS

Presente e futuro da análise de dados de fatores associados à soroprevalência da diarreia viral bovina

Autor: Gustavo Machado

Trabalho apresentado como requisito parcial para a obtenção do grau de Doutor em Ciências Veterinárias na área de Medicina Veterinária, subárea Medicina Veterinária Preventiva, especialidade Epidemiologia Animal

Orientador: Luis Gustavo Corbellini
Co-orientadora: Luciana Neves Nunes

Porto Alegre 2016

Gustavo Machado

PRESENTE E FUTURO DA ANÁLISE DE DADOS DE FATORES ASSOCIADOS À SOROPREVALÊNCIA DA DIARREIA VIRAL BOVINA

Aprovado em 24 de fevereiro 2016

APROVADO POR:

Prof. Dr. Luis Gustavo Corbellini

Orientador e Presidente da Comissão

Prof. Dr. André Felipe Streck

Membro da Comissão

Profa. Dr. Álvaro Vigo

Membro da Comissão

Prof. Dr. Fernando Spilki

Membro da Comissão

CIP - Catalogação na Publicação

Machado, Gustavo
Presente e futuro da análise de dados de fatores associados à soroprevalência da diarreia viral bovina / Gustavo Machado. -- 2016.
103 f.

Orientadora: Luis Gustavo Crobellini.
Coorientadora: Luciana Neves Nunes.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Veterinária, Programa de Pós-Graduação em Ciências Veterinárias, Porto Alegre, BR-RS, 2016.

1. BVDV. 2. epidemiologia. 3. amostras de tanque de leite. 4. regressão. 5. Random Forest. I. Gustavo Crobellini, Luis, orient. II. Neves Nunes, Luciana, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

DEDICATÓRIA

Dedico essa tese aos meus pais, pelo apoio incondicional

AGRADECIMENTOS

Aos meus pais: Ivone e Ademir, pela dedicação posta na minha educação, pela ajuda em construir sempre o que foi desejado, pelos incentivos a novas experiências e pelo amor dedicado em todos os momentos, ao meu irmão Tiago que mesmo que tenhamos morado longe nos últimos 10 anos sempre me inspirou e serve de exemplo de pessoa e profissional.

Agradecimento a todas as pessoas das quais já contribuíram para meu crescimento seja pessoal ou profissional, independente te terem tomado conhecimento disso, todas pessoas com quem já conversei de alguma forma contribuíram para que eu pudesse estar escrevendo essa tese.

Ao meu orientador e amigo, Prof. Luis Gustavo Corbellini, pela confiança e responsabilidade depositada em mim, pelo exemplo de competência e caráter oferecido.

À CAPES pelo auxílio financeiro

Presente e futuro da análise de dados de fatores associados à soroprevalência da diarreia viral bovina

RESUMO

O vírus da diarreia viral bovina (BVDV) causa uma das doenças mais importantes de bovinos em termos de custos econômicos e sociais, uma vez que é largamente disseminado na população de gado leiteiro. Os objetivos do trabalho foram estimar a prevalência em nível de rebanho e investigar fatores associados aos níveis de anticorpos em leite de tanque através de um estudo transversal, bem como discutir e comparar diferentes técnicas de modelagem, as tradicionais como regressão e as menos usuais para este fim, como as de *Machine learning* (ML) como *Random Forest*. O estudo transversal foi realizado no estado do Rio Grande do Sul para a estimação da prevalência de doenças reprodutivas baseados em amostras de tanque de leite, partindo de uma população total de 81.307 rebanhos. Foram coletadas 388 amostras de tanque de leite, e nas propriedades selecionadas foi aplicado um questionário epidemiológico. Como resultados se identificou uma prevalência de 23,9% ($IC_{95\%}$ = 19,8 - 28,1) de propriedades positivas. Através de análise de regressão de Poisson se identificou como fatores associados o BVDV: o exame retal como rotina para o diagnóstico de prenhes, Razão de Prevalência [PR] = 2,73 (IC $_{95\%}$: 1.87-3.98), contato direto entre animais (contato via cerca de propriedades lindeiras) (PR=1,63, IC $_{95\%}$: 1.13-2.95) e propriedades que não utilizavam inseminação artificial (PR=2.07, IC $_{95\%}$: 1.38-3.09) Na técnica de *Random Forest* pôde-se identificar uma dependência na ocorrência de BVDV devido a: inseminação artificial quando realizada pelo proprietário da propriedade ou capataz, o número de vizinhos que também possuem criação de bovinos, e em concordância com os resultados da regressão quanto a dependência da ocorrência de BVDV devido a palpação retal. Como resultado, pôde-se perceber que o BVDV está distribuído no estado do RS e caso seja de interesse do poder público, o desenvolvimento de um programa de controle da doença pode ser baseado nos resultados encontrados. Por outro lado, a contribuição deste estudo vai além das tradicionais análises realizadas em epidemiologia veterinária, principalmente devido os

bons resultados obtidos com a abordagem por ML neste estudo transversal. Por fim, a utilização de técnicas estatísticas mais avançadas contribuiu para elucidar melhor os fatores possivelmente envolvidos com a ocorrência de BVDV no rebanho leiteiro gaúcho.

Palavras-chave: BVDV; epidemiologia; amostras de tanque de leite; regressão; Random Forest.

PRESENT AND FUTURE OF DATA ANALYSIS OF ASSOCIATED FACTORS TO SEROPREVALENCE OF BOVINE VIRAL DIARRHEA

ABSTRACT

The bovine viral diarrhea virus (BVDV) causes one of the most important disease of cattle in terms of economic and social costs, since it is widely disseminated in dairy cattle population. The objectives were to estimate the herd level prevalence at and investigate factors associated with antibody levels in bulk tank milk through a cross sectional study, discuss and compare different modeling techniques such as the traditional regression with the ones less used for this approach machine learning (ML). The cross sectional study was conducted in Rio Grande do Sul state to estimate the prevalence of reproductive diseases based on bulk tank milk samples, from a total population of 81,307 herds. Milk samples from 388 bulk tank were sampled, and an epidemiological questionnaire was applied in each farm. The prevalence was 23.9% (95% CI 19.8 - 28.1). Through the Poisson regression analysis, the following factors associated with BVDV were found: routine use of rectal examination for pregnancy *(Prevalence Ratio [PR] = 2.73* (IC $_{95\%}$: 1.87-3.98)*, direct contact between/among animals (contact over the fence of neighboring farms) (PR = 1.63*, IC $_{95\%}$: 1.13-2.95*) and properties that did not use artificial insemination (PR = 2.07*, IC $_{95\%}$: 1.38-3.09*). On the other hand, using ML techniques, it was identify a dependency upon the occurrence of BVDV due to: artificial insemination when carried out by the owner of the property or foreman; the number of neighbors who also have cattle, and in accordance with the regression results as the dependence of the occurrence of BVDV due to routine use of rectal examination for pregnancy. BVDV is spread across the State and if the government's interest to launch a disease control program measures should be focusing mainly on better conditions and care in reproduction. On the other hand, the contribution of this study goes beyond traditional analyzes in veterinary epidemiology, mainly due to the good results obtained with the approach by ML in this cross-sectional study. Finally, the use of advances statistics techniques it has been*

made progress to better elucidate the factors possibly involved in the occurrence of BVDV in state dairy herds.

***Keywords**: BVDV; epidemiology; bulk tank milk; regression; Random Forest*

SUMÁRIO

1. INTRODUÇÃO

O Brasil tem o maior rebanho comercial do mundo com aproximadamente 211 milhões de bovinos no ano de 2012 (IBGE, 2011), sendo um dos maiores exportadores de carne bovina (MAPA, 2007; USDA, 2013). O Rio Grande do Sul (RS) possui cerca de 14 milhões de bovinos, o que lhe confere o 6º maior rebanho bovino do país (IBGE, 2011)

Segundo a Fundação Estadual de Economia e Estatística do RS (FEE), o produto interno bruto (PIB) agropecuário do estado cresceu cerca de 40% no ano 2013, influenciando positivamente o PIB estadual. Estudos demonstram ainda, que aproximadamente 1/3 do PIB do estado deve-se à participação do setor agropecuário (FEE, 2013; OLIVEIRA, 2010).

Uma das doenças que afeta principalmente o setor leiteiro é a diarreia viral bovina (BVD), enfermidade causada pelo BVDV, um vírus RNA de fita simples e envelopado, pertencente ao gênero *Pestivirus* da família *Flaviviridae* (GOYAL & RIDPATH, 2008). A infecção em bovinos de leite causa perdas econômicas significativas por diminuir o desempenho reprodutivo e a produção de leite (NISKANEN et al. 1995; GROOMS, 2006). A infecção pelo BVDV é uma enfermidade diretamente relacionada aos problemas reprodutivos que, consequentemente, implicam em perdas econômicas pelo fato da infecção ter a capacidade de ser transmitida para as próximas gerações (transmissão horizontal), resultando no nascimento de bezerros PI (persistentemente infectados), e afeta de forma marcante a produção leiteira (BAKER, 1995; HOUE, 1999). A soroprevalência em nível animal tem variado mundialmente entre 40 a 90% (HOUE, 1999; LINDBERG & HOUE, 2005), a prevalência de rebanhos com infecção ativa ou recentemente infectados varia de 47 a 100% (HOUE, 1994; SARRAZIN et al. 2012). Sabe-se que o rebanho bovino nacional está exposto ao BVDV, porém a exata prevalência e distribuição da doença através de estudos planejados, incluindo os principais fatores relacionados à presença destas enfermidades, ainda não estão totalmente esclarecidos. Para isso é importante definir o tipo de delineamento do estudo mais adequado para identificação de fatores relacionados à ocorrência de BVDV bem como as melhores abordagens analíticas no sentido de elucidar com maior precisão e segurança a influência desses fatores na ocorrência de BVDV na população leiteira do

Rio Grande do Sul.

A análise de dados em epidemiologia humana e animal tomaram dimensões relevantes nas últimas décadas com a maior disponibilidade computacional e o desenvolvimento de softwares livres. Inúmeros modelos estatísticos de regressão têm sido aplicados para a identificação de fatores de risco, porém a aplicação de algoritmos de *machine learning* (ML) é subutilizada na identificação de variáveis preditoras em estudos epidemiológicos na medicina veterinária. Os modelos ML são responsáveis pelo desenvolvimento de conhecimento em inúmeras áreas, sendo aplicados usualmente para reconhecimento de padrões supervisionado em conjuntos de dados. Tipicamente, algoritmos de ML são usados para treinar um modelo que permite separar amostras de diferentes classes (ex. saudável ou doente), baseado em um conjunto de preditores (ex. hábitos alimentares, tabagismo), para estimação de variáveis relevantes/importantes para o desfecho estudado.

Random Forest (RF) é um desses classificadores que tem se tornado muito popular devido a dois aspectos muito importantes para mineração de dados: alta acurácia nas predições e informações sobre a importância de cada preditora para a classificação. Segundo VERIKAS et al. (2011) outra vantagem da utilização de RF é o fato que seu desempenho pode ser comparado com outros classificadores como *support vector machine* (SVM) (CORTES & VAPNIK, 1995), *artificial neural networks* (RUMELHART, et al. 1986), *bayesian classifiers* (FRIEDMAN, et al. 1997), *logistic regression* (KLEINBAUM, et al. 1982), *k-nearest-neighbours* (FIXT & HODGES, 1989), análises discriminatórias lineares (FISCHER, 1936), *regularized discriminant analysis* (FRIEDMAN, 1997), *partial least squares* (PLS) (WOLD, 1975) e *decision trees* (CARTs) (BREIMAN, et al. 1984).

Tendo em vista a importância desse segmento no agronegócio, são necessárias pesquisas para a produção de insumos e formação de recursos humanos quanto aos métodos de diagnóstico e estudos epidemiológicos aplicados às doenças infecciosas. A presente tese de doutorado tem como objetivos contribuir para o desenvolvimento e aplicara técnica de RF para a obtenção de informações epidemiológicas veterinárias, visando um futuro estabelecimento desta abordagem na área, assim como, contribuir para a construção de estratégias de controle e erradicação de BVDV.

2. REVISÃO BIBLIOGRÁFICA

2.1. **Vírus da diarreia viral bovina**

O BVDV é o termo referido a um grupo diverso de vírus com genoma RNA de fita simples, membros do gênero *Pestivirus* da família *Flaviviridae* (RAUE et al. 2011; STÅHL & ALENIUS, 2012). Também fazem parte desta família os vírus da doença da fronteira (*border disease*) e o vírus da peste suína clássica. Os vírus da família *Flaviviridae* são vírus esféricos com diâmetro de 40 a 50 nm e são facilmente inativados pelo calor, detergentes, solventes orgânicos e radiação gama (RIDPATH, 2010a). As cepas de BVDV, independente do genótipo, estão limitadas a dois biotipos: BVDV não-citopático (NCP) e citopático (CP). Somente os NCP são capazes de atravessar a placenta, invadir o feto e estabelecer infecção persistente e são considerados "verdadeiros" BVDV (FLORES et al. 2005). Os NCP representam a grande maioria das amostras de campo que estão associadas às diversas manifestações clínicas da infecção por BVDV. O biotipo CP causa extensos danos nas células do cultivo, como vacuolização citoplasmática e destruição celular entre 48 a 72 horas, com menor ocorrência que o anterior são isolados quase que exclusivamente de animais afetados pela Doença das Mucosas (DM) (RIDPATH, 2010a). Esta diversidade antigênica entre as cepas isoladas de BVDV são importantes para epidemiologia, diagnóstico e seleção das estratégias de imunização e controle da doença (BOTTON et al. 1998). O BVDV fica caracterizado por sua diversidade e capacidade de estabelecer dois tipos principais de infecção: animais com infecção persistente (PI), considerados a principal fonte de infecção, e animais com infecção transitória (TI), considerada usualmente uma fonte de infecção menos importante (PETERHANS & SCHWEIZER, 2010).

2.2. **Infecção e doença clínica**

Basicamente existem duas formas de infecção por BVDV. A forma mais importante ocorre quando animais susceptíveis entram em contato com BVDV durante a gestação, quando ocorre a exposição do feto *in utero* com a cepa NCP do BVDV

anteriormente ao desenvolvimento completo do sistema imune fetal, o que ocorre em torno dos 125 dias de gestação (transmissão vertical) (CASARO et al. 1971). O vírus possui tropismo por células germinativas logo, as placas de *Peyer* da mãe e o feto são os principais locais de multiplicação e durante a viremia, o vírus pode atravessar a placenta e infectar o feto (GROOMS, 2004). Nestes casos, o vírus é reconhecido como próprio, e os animais nascidos vivos (fracos ou normais) podem eliminar o vírus durante toda vida, sendo reconhecidos como PI (MCCLURKIN et al. 1984; CASAUBON et al. 2012). Após o nascimento, esses animais não irão soroconverter e apresentarão viremia persistente (HANON et al. 2012). Animais PI geralmente são mais eficientes em transmitir o vírus do que animais denominados transitoriamente infectados (TI) pelo fato de secretarem grande quantidade de vírus por períodos prolongados (BROCK et al. 1998). Devido ao prejuízo ao sistema imune do animal PI, esses animais são particularmente susceptíveis a outras infecções, o que, em parte, explica a alta mortalidade dos animais quando jovens em comparação a animais sadios (HOUE, 1992; 1999). Alguns animais PI podem permanecer clinicamente normais e serem selecionados para reprodução (MCCLURKIN et al. 1979) e assim retransmitir a infecção às gerações subsequentes (STÅHL & ALENIUS, 2012).

A segunda forma, menos importante, denominada TI, ocorre quando os animais imunocompetentes ficam expostos ao BVDV (transmissão horizontal). Neste caso, o BVDV é adquirido primeiramente através de aerossóis, que infecta a mucosa nasal. O contato direto focinho-focinho entre um animal infectado e um animal livre de BVDV é considerado a via mais efetiva de transmissão de BVDV horizontalmente, apesar de haver relatos de transmissão indireta pelo uso de formigas para contenção e alojamento de animais em estábulos contaminados (NISKANEN & LINDBERG, 2003). Em curto período de tempo após a infeção, animais TI apresentam viremia e o vírus pode ser secretado através das secreções e excreções por alguns dias (4-15) (MCCLURKIN et al. 1984; BROCK et al. 1998). A transmissão horizontal já foi demonstrada e pode ocorrer em apenas uma hora de contato direto com o animal PI (TRÅVÉN et al. 1991). O contato direto entre animais susceptíveis e animais PI, principalmente através da cerca, é considerada a forma mais comum de introdução da infecção em rebanhos livres (SMITH et al. 2009; STOTT et al. 2010; VOAS, 2012). É importante ressaltar que a

soroconversão em rebanhos livres, ou seja, na ausência de animais PI, é um indicativo que a transmissão através de animais TI ocorreu de fato. No entanto, sua disseminação é mais lenta (MEYLING et al. 1990; MOERMAN et al. 1993).

A infecção por BVDV pode resultar em um amplo espectro de manifestações clínicas, partindo de curso subclínico a sinais flutuantes e possivelmente a morte (BAKER, 1995). A maioria dos isolados de ambos os genótipos (BVDV-1 e BVDV-2) apresenta baixa virulência, frequentemente com curso subclínico. Foi estimado que 70 a 90% das infecções causadas por BVDV são subclínicas (BOLIN & GROOMS, 2004). Nesta forma da doença, os animais desenvolvem apenas febre moderada e leucopenia.

As características do animal que influenciam no resultado das manifestações clínicas estão relacionadas com o estado imunológico, estágio de prenhes, idade do feto em gestação e condição de estresse imposto pelo ambiente (BAKER, 1995). A forma mais comum da doença afeta principalmente a reprodução, diminuindo o desempenho reprodutivo, aumentando taxas de retorno ao cio, malformações, aborto e pode estar acompanhada por febre (BAKER, 1995; NISKANEN et al. 1995).

O BVDV é um dos patógenos que fazem parte do complexo de doença respiratória bovina (BRDC). A forma respiratória da doença apresenta manifestações tanto do trato respiratório superior (tosse, descarga nasal e ocular) quanto trato respiratório inferior (frequência respiratória aumentada e ausculta de sons ásperos vindo do pulmão) (RAUE et al. 2011), além de sinais gerais menos determinados afetando o sistema respiratório (NETTLETON & ENTRICAN, 1995; RIDPATH, 2010b). O desenvolvimento de sinais clínicos respiratórios é dependente de inúmeros fatores como: virulência da cepa, tipo de infecção (TI ou PI), tempo de exposição (fetal ou pós-fetal) e a presença de infecções secundárias (RIDPATH, 2010b).

A produção de leite também fica comprometida pela infecção por BVDV, principalmente por queda da imunidade e por comprometimento das defesas da glândula mamária (LAUREYNS et al. 2012). Este fato foi investigado em condições de campo e ficou sugerido o efeito deletério negativo que o BVDV causa na contagem de células somáticas (LAUREYNS et al. 2012).

Outra forma da doença é a aguda ou superaguda, caracterizada por hemorragias e trombocitopenia. Esta forma pode estar presente tanto em bezerros como em animais

adultos e a espécie responsável por este tipo de infecção é a BVDV-2 (RIDPATH et al. 1994).

Finalmente, uma forma da doença mais agressiva, a doença das mucosas (DM), é severa e inevitavelmente fatal. Ocorre quando um animal PI se torna superinfectado por uma cepa CP derivada de uma NCP (BOLIN et al. 1985; BROWNLIE & CLARKE, 1993). A DM acomete principalmente animais de até dois anos de idade (PETERHANS et al. 2010). Na ausência de medidas de controle, a presença de animais PI foi estimada em 0,5% a 2% da população de um rebanho infectado, o que consequentemente leva a baixa incidência da DM, que é caracterizada por baixa taxa de ataque, porém com altas taxas de letalidade (BROWNLIE, 1990; HOUE, 1999). Outros autores sugerem que a DM é desenvolvida quando um animal PI é infectado simultaneamente por cepas CP e NCP (CHASE, 2012). Pode-se concluir que essa forma da doença é uma consequência tardia da infecção persistente por BVDV e o diagnóstico definitivo deve ser acompanhado por isolamento viral (BAKER, 1995).

2.3. **Diagnóstico**

Inúmeros métodos de identificação de animais infectados por BVDV foram desenvolvidos, incluindo isolamento viral de amostras de soro, sangue total e outros tecidos; imunofluorecência em tecidos; imuno-histoquímica em tecidos; ELISA (s) realizados em amostras de soro e leite, além de inúmeras técnicas de diagnóstico molecular (DUBOVI, 2012). O isolamento viral é considerado o teste padrão ouro e recomendado pela Organização Mundial de Saúde Animal (OIE). O melhor método a ser utilizado depende da situação em que o animal ou rebanhos se encontram, da idade dos animais, se estão vivos ou mortos e quais os objetivos para o teste, identificar animais PI ou TI.

Dentre os testes diagnósticos mais empregados estão os sorológicos, utilizados para determinar: 1) se o animal ou rebanho entrou em contato com o vírus; 2) se um bezerro possui anticorpos colostrais; 3) se o animal ou rebanho está adequadamente imunizado; 4) se o animal ou rebanho possui uma infecção ativa; 5) se o bezerro foi infectado *in utero* (DUBOVI, 2012). Há uma diferença regional quanto à escolha do teste

sorológico. Nos EUA, onde a infecção e vacinação estão presentes, a soroneutralização é mais frequentemente utilizada; já na Europa o teste de ELISA é comumente utilizado (DUBOVI, 2012). No geral, o ELISA é o teste frequentemente utilizado para amostras de soro e/ou leite (BEAUDEAU et al. 2001a).

Para o diagnóstico de um volume grande de amostras, o ELISA é considerado mais reprodutível do que a soroneutralização e, também, economicamente mais viável (GONDA et al. 2012). Existem basicamente duas configurações de ELISA, o indireto e o direto (competitivo), os quais podem ser utilizados para detectar anticorpos em leite, plasma e soro (KATZ & HANSON, 1987; NISKANEN et al. 1991). O uso do ELISA direto tem aumentado desde os anos 90, pelo fato de que os testes até então utilizados demandavam muito tempo e pessoal treinado (KAMPA et al. 2007). O ELISA de captura NS2/3 detecta BVDV em leucócitos e amostras de tecido utilizando anticorpos monoclonais específicos contra a proteína NS2/3, e este teste tem sido utilizado com sucesso na identificação de animais PI em programas de controle de BVDV na Noruega (SYNGE et al. 1999). Ainda, foi desenvolvido um ELISA com anticorpos monoclonais contra a glicoproteína E^{rns}, que é uma proteína estrutural secretada por células infectadas durante a replicação viral e pode ser detectada diretamente no soro (KUHNE et al. 2005). Outro ELISA direto foi desenvolvido para a identificação anticorpos contra a proteína p80/NS3, que permite a diferenciação entre anticorpos vacinais e anticorpos produzidos pela infecção natural, porém vale ressaltar que animais vacinados com vacinas vivas também desenvolvem anticorpos contra a proteína p80/NS3 (NIZA-RIBEIRO et al. 2005).

O ELISA indireto usa o BVDV por completo como antígeno para medir a resposta contra o espectro de proteínas imunogênicas presentes (PATON et al. 1991). Inúmeros kits comerciais de ELISA indiretos para detecção de anticorpos estão disponíveis. A adaptação da técnica de ELISA para detecção de anticorpos em amostras de leite de tanque de expansão constitui uma alternativa barata e factível na evolução dos programas de controle da BVDV em rebanhos leiteiros. O teste pode fornecer informações a respeito do *status* de um grande grupo de animais (vacas em lactação) com apenas uma amostra (EIRAS et al. 2012). Por este fato, inúmeros países da Europa vêm monitorando seus rebanhos por vários anos através de amostras de leite de tanque

(BEAUDEAU et al. 2001b; RIKULA et al. 2005). Estudos recentes demonstraram que a sensibilidade do teste utilizado em amostras de tanque de leite foi capaz de identificar animais PI quase em 100% dos casos, porém a especificidade do teste foi limitada (RIKULA et al. 2005; HOUE et al. 2006). Também foi identificada alta correlação entre os níveis de anticorpos em amostras de tanque de leite detectados por ELISA indireto e a prevalência de BVDV em amostras de soro de vacas positivas (NISKANEN et al. 1991).

Para o diagnóstico de BVDV em rebanhos, o *spot test* permitiu a identificação de rebanhos infectados dispensando a necessidade de coleta de 100% dos animais. A detecção de anticorpos contra BVDV no *spot test*, que é direcionado para animais jovens (animais de 8 a 18 meses), é um indicativo da presença de infecção corrente (HOUE, 1994), no entanto, é muito mais laborioso se comparado com o teste em amostras de tanque de leite.

Basicamente, quando se tratar de um rebanho livre de BVDV, a maioria do rebanho provavelmente se apresentará soronegativo, com isso o resultado do ELISA resultará em níveis baixos ou indetectáveis de anticorpos (JUNTTI et al. 1987; NISKANEN, 1993). A ausência de anticorpos indica que o rebanho está possivelmente livre da doença (PATON et al. 1998). A principal vantagem da amostragem de tanque de leite é a praticidade e o baixo custo associado, facilitando a logística dos estudos, assim como a diminuição dos riscos de acidentes associados à coleta de soro.

2.4. **Prevalência**

O BVDV é endêmico na maioria dos países e a prevalência pode ser estimada em níveis de anticorpos ou pela ocorrência de animais PI. Em todos os países onde dados de prevalência em nível de rebanho estão disponíveis, a média fica em torno de 55% de rebanhos positivos (HOUE, 1995). A soroprevalência em nível animal tem variado de 60 a 90% (HOUE, 1999; LINDBERG & HOUE, 2005) e a proporção de animais PI varia de 0,1 a 2% (BROWNLIE, 1990; FREY et al. 1996; HOUE, 1999). A prevalência de rebanhos com infecção ativa, ou recentemente infectados, varia de 70 a 100% (HOUE, 1994). Entretanto, há algumas diferenças entre regiões e países, o que pode estar relacionado com diferenças em densidade animal, instalações, vacinação, sistemas de

manejo e principalmente a presença de animais PI (HOUE, 1999). Um estudo em amostras de leite na Suíça identificou a presença de anticorpos em 83,7% dos rebanhos e 45,5% com infecção ativa ou recente causada por BVDV (NISKANEN, 1993). Em outros países, como no Irã, a prevalência de rebanhos foi de 94% e 52,5% de infecções ativas ou recentes (GAROUSSI et al. 2008); no Peru foram identificados níveis maiores (95%) de rebanhos positivos (STÅHL et al. 2008). Na Tailândia, níveis de infecção menores foram encontrados: 73% em nível de rebanho e uma proporção de infecção ativa ou recente de 13% (KAMPA et al. 2004). Mais recentemente, um estudo na Bélgica identificou 47,4% de rebanhos com anticorpos positivos e a presença de 4,4% de antígenos específicos para BVDV; já em nível animal foram 32,9% de positivos para presença de anticorpos e apenas 0,3% dos animais positivos para a presença de antígenos de BVDV, sendo que dos 44,4% dos rebanhos positivos, aproximadamente 60% dos animais amostrados eram jovens (SARRAZIN et al. 2012). Outro estudo realizado em amostras de tanque de leite na Escócia identificou frequência de níveis de anticorpos de acordo com os padrões Suecos de 12,7%, 22,3%, 44,5% e 20,5%, que são classificados em nível crescente pela presença de anticorpos de 0 a 3, respectivamente. Porém, um achado importe foi que 73% dos rebanhos possuíam níveis elevados de anticorpos sugerindo infecção ativa ou introdução recente da infecção (HUMPHRY et al. 2012).

No Brasil, a prevalência estimada através de estudos que apresentam descrição clara do processo amostral demonstra índices de BVDV variando de 43% a 90% (POLETTO et al. 2004; THOMPSON et al. 2006; QUINCOZES et al. 2007, ALMEIDA et al. 2013). Em países da América Latina, como Uruguai e Chile, a prevalência variou de 69 a 77,8% (REINHARDT et al. 1990; GUARINO et al. 2008). Em especial no Rio Grande do Sul estudos recentes baseados em amostras probabilísticas estimaram a prevalência de BVDV em 43% a nível de rebanhos (ALMEIDA, et al. 2013), seguido por um estudo que estimou em 24% (MACHADO, et al. 2014).

2.5. **Fatores associados à ocorrência de BVDV**

A identificação de fatores de risco[1] para a ocorrência de uma doença, particularmente BVDV, é desejável, pois auxilia na criação de futuras estratégias de controle da doença. Os principais fatores associados com infecção por BVDV já identificados foram: tamanho de rebanho; distância de propriedades vizinhas com criação de bovinos; número de vizinhos com rebanhos infectados; compra de animais sem teste negativo para BVDV; pastejo de várias categorias animais no mesmo piquete; contato direto entre animais de vizinhos (cerca-cerca); não possuir assistência técnica; fazendas estarem em área de alta prevalência de BVDV; vacinação para BVDV; alojamento de fêmeas prenhas com bezerros; proporção de vacas secas no rebanho. Outros fatores já foram especulados como potenciais fatores associados, porém não foram totalmente elucidados: ovinos e bovinos pastejando em mesmo piquete; queda de cerca; reutilização de agulhas pelo Médico Veterinário; presença de animais selvagens em pastejo com bovinos; presença de árvores nos piquetes dos bezerros; origem da água fornecida aos animais; monitoramento de abortos; entre outros (HOUE, 1999; VALLE et al. 1999; LUZZAGO et al. 2008; TALAFHA et al. 2009; HUMPHRY et al. 2012; SARRAZIN et al. 2012)

2.6. ***Machine learning*** **(ML)**

Machine learning, também conhecida como aprendizado de máquina, é uma área da inteligência artificial que visa o desenvolvimento de técnicas computacionais capazes de adquirir conhecimento de forma automática a partir de um conjunto de exemplos. Aprendizado de máquina é genericamente um conjunto de algoritmos que tomam decisões baseado em experiências acumuladas por meio da solução bem-sucedida de problemas anteriores (WEISS & KULIKOWSKI, 1991). Existe uma grande variedade de algoritmos de ML, dentre os mais utilizados na atualidade temos os métodos *ensemble* de classificação, *Random Forest* (RF), *Support Vector Machine* (SVM) e *Gradient Boosting*

[1] Fatores de risco é uma nomenclatura mais adequada para estudos epidemiológicos longitudinais. Nessa tese, será utilizada como análogo a fator de risco a terminologia “fator associado” a um determinado desfecho.

Machine (GBM). Nos próximos itens serão abordados brevemente os principais métodos utilizados nesta tese de doutorado.

2.7. **Classificador *Ensemble***

Trata-se de um método que utiliza ou combina múltiplos classificadores para melhorar a robustez, assim como alcançar um classificador com melhor desempenho em relação a qualquer classificador constituinte, ou são algoritmos de aprendizado que constroem um conjunto de classificadores e combinam seus votos para classificar um novo evento (DIETTERICH, 2000). Assim, este algoritmo é mais resiliente a perturbações quando comparado com um classificador simples (SYARIF, et al. 2012). Estes métodos utilizam "divide para conquistar", onde um problema complexo é decomposto em múltiplos sub-problemas, os quais são mais facilmente entendidos e resolvidos (SYARIF, et al. 2012).

Um classificador *ensemble* possui melhor acurácia do que técnicas de classificação simples (SYARIF, et al. 2012). O sucesso da abordagem *ensemble* depende da diversidade nos classificadores individuais, no que diz respeito a casos erroneamente classificados (LEE & CHO, 2010). De acordo com POLIKAR (2006), há quatro maneiras de atingir a diversidade do método primeiro é utilizar dados diferentes para treinar um classificador simples, o segundo é utilizar parâmetros distintos para o treinamento, e terceiro utilizar diferentes preditores para treinar o classificador e por fim combinar diferentes tipos de classificadores.

2.8. ***Random Forest***

Random forest é um classificador composto por uma coleção de árvores $\{h_k(x)\}$, k=1,2....,L, onde H_k são amostras aleatórias independentes e identicamente distribuídas e cada árvore vota na classe mais popular para a entrada x (BREIMAN, 2001). Os modelos de RF podem ser utilizados sem a afinação dos parâmetros do algoritmo, todavia um modelo melhor geralmente é gerado com a otimização de pouquíssimos parâmetros (BREIMAN, 2001). RF treina árvores de decisão (AD) individual baseado em amostras e

sua designação de classe e variáveis. Cada árvore na floresta é constituída por um subconjunto de amostras e variáveis (Figura 1-TOUW, et al. 2012). RF aplica o mesmo método de *bagging* para produzir amostras aleatórias de conjuntos de treinamento "boostraps" para cada AD gerada.

Suponha que uma floresta de AD (CARTs) seja construída baseada em um banco de dados. Para todas as árvores, um subconjunto de treino é criado através de amostragem aleatória (ex. amostra de pacientes) com reposição, resultando em um subconjunto de treino, ou um subconjunto *boostrap*, contendo mais ou menos 2/3 das observações presentes no banco de dados originais. O restante das observações no banco de dados é chamado de amostra `out-of-bag` (OOB). As árvores são construídas a partir do subconjunto de dados do *boostrap* através de partição recursiva (Figura 1- TOUW, et al. 2012). Para cada nodo, variáveis preditoras são aleatoriamente selecionadas a partir do conjunto de todas preditoras disponíveis e posteriormente avaliadas em termos de sua habilidade em dividir o banco de dados (Figura 1- TOUW, et al. 2012). A preditora resultante com a maior redução na impureza é escolhida para a separação do banco em dois nodos 'parentes', resultando em dois nodos distintos 'nodos filhas'. Em RF, uma das medidas de impureza é o índice *Gini.* A diminuição no índice *Gini* está relacionada no aumento na qualidade de ordem das classes da amostra introduzido pela divisão da AD.

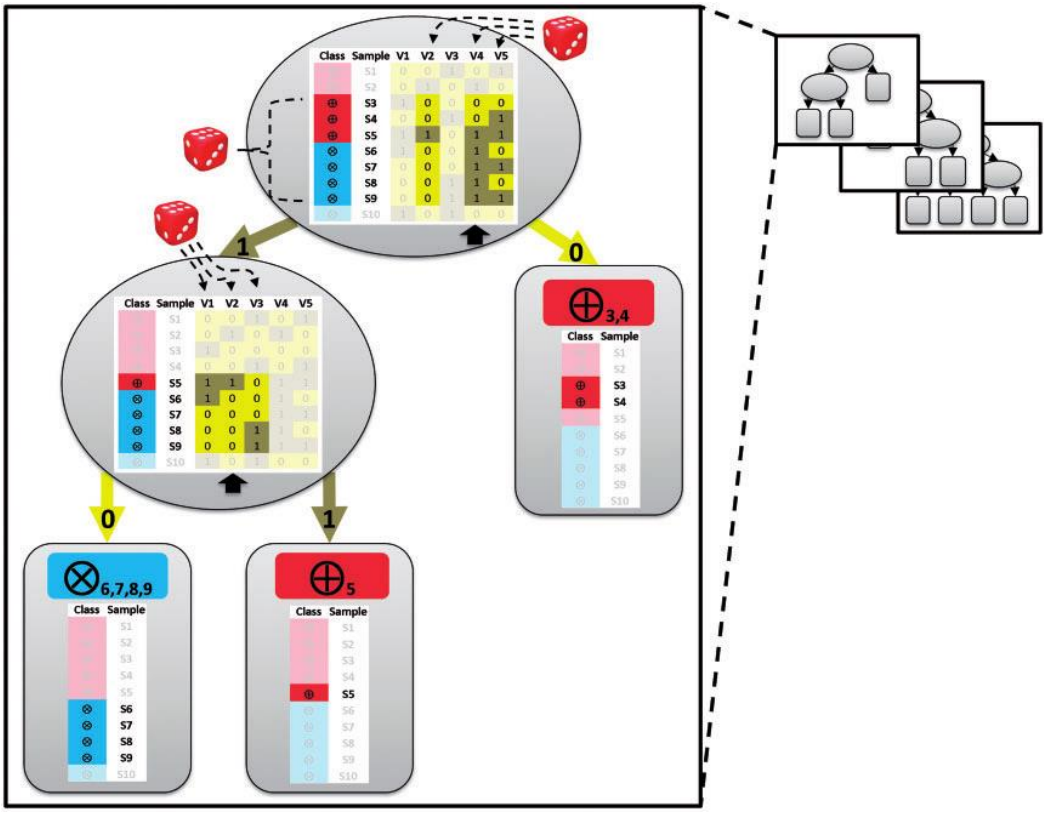


Figura 1- *Treinamento de uma árvore de uma RF. A árvore é construída baseada no banco de dados (mostrado dentro do círculo no topo). Esta matriz consiste em amostras (S1-S10-individuos) pertencentes a duas classes (indivíduos saudáveis-azul, doentes-vermelho) e as medidas para cada amostra para cada preditora (V1-V5). O dado representa o sorteio, linhas tracejadas: amostras e variáveis selecionadas aleatoriamente. Para cada árvore, um subconjunto é formado por boostrap, através da amostragem aleatória com reposição de observações do banco até que sejam amostradas tantas amostras quanto há no banco de dados. A amostra aleatória vai conter 63% das amostras do banco de dados original. No exemplo, o bootstrap contém sete amostras únicas (amostras S3-S9, as amostras não selecionadas foram S1, S2 e S10). Para cada nodo (indicado como círculos) variáveis são aleatoriamente selecionadas (neste caso três, as outras duas são mostradas em segundo plano; por padrão, o método RF escolhe a raiz quadrada do número de variáveis presentes no banco) e sofre uma avaliação quanto a sua habilidade de divisão do banco. As variáveis resultantes na maior redução de impureza são escolhidas para a definição da regra de divisão. No caso, o nodo do topo foi a variável V4 (seta preta) e para o segundo nodo no lado esquerdo V2. Este*

processo e repetido até que os nodos fiquem puros (assim chamadas folhas; indicado pelos retângulos), ou seja, eles contêm amostras da mesma classe (na folha do lado direito em vermelho-doentes).

Após a divisão do banco de dados a partir do subconjunto do *boostrap* ser concluída no topo do nodo, o processo de divisão é repetido. A compartimentação é finalizada quando os nodos terminais ou "folhas" estejam ou puras ou contenham apenas amostras pertencentes à mesma classe ou ainda contenham um número específico de amostras. Geralmente uma AD é construída até o nodo terminal esteja puro, mesmo que isso resulte em um nodo contendo apenas uma única amostra. A AD é assim construída até a sua maior dimensão, sem passar pelo processo de podagem ou "pruning". Após a *floresta* ter sido completamente construída, o processo de treinamento está completo (TOUW, et al. 2012).

É preciso reforçar que uma única AD é considerado um classificador limitado, pelo fato de ser treinado em um subconjunto dos dados, no entanto a combinação de todas AD em uma floresta é considerada um bom classificador (BREIMAN, 2001). A taxa de erro de classificação esperada para novas amostras são geralmente estimadas por método de validação cruzada como *leave-one-out* ou *K-fold cross-validation* (STONE, 1974). O erro de classificação da RF depende da força das árvores individuais da floresta e da correlação entre pares de árvores nas florestas (BREIMAN, 2001). O que BREIMAN justifica para que a correlação das árvores seja diminuída e, portanto menores os erros de classificação, se dá pelo uso dos processos de randomização (*bagging* e seleção aleatória das preditoras).

2.8.1. ***Importância da variável***

A estimativa da importância de cada variável preditora é fundamental para a interpretação e seleção de variáveis para a aplicação prática do modelo. O escore de importância pode ser útil para a identificação de biomarcadores (FUSARO, et al. 2009), seleção de preditoras em programas de vigilância de salmonelose em suínos (ABRAHANTES et al. 2009), identificar fatores de risco para H1N1, e para PRRS em

suínos (HOLTKAMP et al. 2012; LARISON et al. 2014) ou como um filtro para remoção ou não de preditoras que não agregam informações ao modelo (JIANG, et al. 2007). As medidas de importância para preditoras em RF são basicamente duas. Primeiramente, a redução média em classificações é baseada na permutação. Para cada AD, a acurácia da classificação da amostra OOB é determinada tanto com e sem a permutação aleatória dos valores das variáveis (TOUW, et al. 2012). A precisão da previsão após a permutação é subtraída da previsão da precisão antes da permutação e assim calculada sobre todas as AD na floresta, resultando no valor de importância das variáveis por permutação (TOUW, et al. 2012). A segunda medida de importância de preditoras é o índice de *Gini* a qual é calculada como a soma do decréscimo da impureza de *Gini* de cada nodo da floresta para a qual a variável foi utilizada para a divisão.

2.9. ***Support Vector Machine*** **(SVM) e** ***Gradient Boosting Machine*** **(GBM)**

O *Support Vector machine* é baseado em ideias simples que se originaram na estatística do teorema do aprendizado (VAPNIK, 1999). A simplicidade vem do fato de que os modelos SVM aplicam um método linear simples aos dados, mas em um espaço de alta dimensão não linearmente relacionados com o espaço de entrada (KARATZOGLOU & MEYER, 2006). No entanto, podemos pensar em SVMs como algoritmos lineares em um espaço de alta dimensão, na prática, isso não envolve nenhuma computação na alta dimensão. Esta simplicidade combinada com o estado da performática de outras maquinas de aprendizagem (classificadores, regressões, e *novelty detection*) contribuíram para a popularidade de SVM.

Em *gradient boosting machines*, ou simplesmente, GBMs, o processo de aprendizagem é consecutivo, onde novos modelos se encaixam sequencialmente para fornecer uma estimativa mais acurada da variável resposta (NATEKIN & KNOLL, 2013). A principal ideia por trás do algoritmo é a construção de uma nova base de modelos para que sejam maximamente correlacionados com o gradiente negativo da função de perda, associados com todo o conjunto. As funções de perda aplicadas no modelo podem ser arbitrárias, mas para que tenham maior intuição, se a função de erro é a clássica perda do erro ao quadrado (NATEKIN & KNOLL, 2013). A alta flexibilidade

dos modelos GBM permitem customizações complexas e sua aplicação a muitos problemas em diversas áreas do conhecimento (NATEKIN & KNOLL, 2013). No entanto, os algoritmos de GBM são relativamente simples de serem implementados, o que permite ao pesquisador experimentar diversos designs de modelo (NATEKIN & KNOLL, 2013). Ainda, os modelos GBM já demostraram enorme sucesso não somente na resolução de problemas práticos, mas também em desafios de modelos de aprendizagem (BISSACCO, et al. 2007; HUTCHINSON et al. 2011; PITTMAN & BROWN, 2011; JOHNSON & ZHANG, 2012).

3. OBJETIVOS

3.1. **Gerais**

Estimar a prevalência e identificar os principais fatores associados ao BVDV em rebanhos leiteiros do Rio Grande do Sul.

3.2. **Específicos**

- Estimar a prevalência de BVDV em amostras de tanque de leite do Rio Grande do Sul.
- Identificar fatores associados ao BVDV utilizando modelos clássicos de regressão.
- Utilizar classificadores, especialmente Random Forest para identificação de preditoras importantes para o controle de BVDV no Rio Grande do Sul.
- Comparar diferentes classificadores (em termos de desempenho) utilizando os dados do inquérito de BVDV em amostras de tanque de leite.

4. RESULTADOS E DISCUSSÃO

Os resultados e discussão dessa tese serão apresentados em forma de artigo científico. Cada subtítulo desse capítulo corresponde a um artigo científico.

4.1. **Capítulo 1: What variables are important in predicting BVDV (Bovine Viral Diarrhea Virus)? A Random Forest approach**

O presente estudo já foi concluído e um artigo científico de nome ***What variables are important in predicting BVDV (Bovine Viral Diarrhea Virus)? A Random Forest approach*** esta publicado no periódico *Veterinary Research* 2015, **46**:85 doi:10.1186/s13567-015-0219-7.

What variables are important in predicting bovine viral diarrhea virus? A random forest approach

Gustavo Machado[1*]
[*]Corresponding author
Email: gustavoetal@gmail.com

Mariana Recamonde Mendoza[2]
Email: mari.mendoza@gmail.com

Luis Gustavo Corbellini[1]
Email: lgcorbellini@hotmail.com

[1] Laboratory of Veterinary Epidemiology, Faculty of Veterinary, Federal University of Rio Grande do Sul (UFRGS), Av. Bento Gonçalves 9090, CEP 91540-000, Porto Alegre, RS, Brazil

[2] Experimental and Molecular Cardiovascular Laboratory, Experimental Research Center, Hospital de Clínicas de Porto Alegre (HCPA), Av. Ramiro Barcelos, 2350, CEP 99010-115, Porto Alegre, RS, Brazil

Abstract

Bovine viral diarrhea virus (BVDV) causes one of the most economically important diseases in cattle, and the virus is found worldwide. A better understanding of the disease associated factors is a crucial step towards the definition of strategies for control and eradication. In this study we trained a random forest (RF) prediction model and performed variable importance analysis to identify factors associated with BVDV occurrence. In addition, we assessed the influence of features selection on RF performance and evaluated its predictive power relative to other popular classifiers and to logistic regression. We found that RF classification model resulted in an average error rate of 32.03% for the negative class (negative for BVDV) and 36.78% for the positive class (positive for BVDV).The RF model presented area under the ROC curve equal to 0.702. Variable importance analysis revealed that important predictors of BVDV occurrence were: a) who inseminates the animals, b) number of neighboring farms that have cattle and c) rectal palpation performed routinely. Our results suggest that the use of machine learning algorithms, especially RF, is a promising methodology for the analysis of cross-sectional studies, presenting a satisfactory predictive power and the ability to

identify predictors that represent potential risk factors for BVDV investigation. We examined classical predictors and found some new and hard to control practices that may lead to the spread of this disease within and among farms, mainly regarding poor or neglected reproduction management, which should be considered for disease control and eradication.

Introduction

Bovine viral diarrhea virus (BVDV) has a single-stranded, positive-sense RNA genome and belongs to the genus *Pestivirus* of the family *Flaviviridae* [1], causing one of the most common and economically important viral diseases of cattle [2]. Several BVDV control strategies have been proposed and launched in many countries based on information about prevalence, incidence and associated risk factors, which is the baseline knowledge required for designing and implementing effective regional control actions [3].

A number of studies based on traditional risk factors identification approaches (logistic regression mainly) have been performed on BVDV [4-8], and the knowledge about major risk factors are related to the following: biosecurity [6], reproduction management [2,6,9,10], herd size [5,8], animal introduction [2,4,5,11], direct contact with other animals (from the same species or not) [4,11-13], communal grazing [4,5], age of animals [5,14], artificial insemination (AI) [15], and natural mating [13]. Nonetheless, usual epidemiologic analytic frameworks like logistic regression are often limited for the analysis of high-dimensional, imbalanced and nonlinear data, and may be poorly adapted to epidemiological datasets with a large number of predictor variables (parameters) in relation to the number of observations given the high susceptibility to overfitting [16,17].

Feature selection methods provided by machine learning (ML) approaches are an interesting, flexible and robust alternative for identifying predictors that contribute to disease occurrence. Among these, the random forest (RF) algorithm [18] has been regarded as one of the most precise prediction methods, having advantages such as ability to determine variable importance, ability to model complex interactions among independent variables, and flexibility to perform several types of statistical data analysis, including regression, classification and unsupervised learning [19]. Briefly, RF builds a collection of decision trees based on randomly and independently selected subsets of data, and a simple majority vote among all trees in the forest is taken for class prediction. A clear difference from traditional statistical frameworks is that RF performs a data-driven analysis without making a priori assumptions about the structure of data or the relationship between the response and independent variables, and is less sensitive to spatial autocorrelation and multicollinearity issues [17,20]. Its high predictive power has been supported by previous comparative studies with other ML methods [21-25].

The use of RF allows for a new way of modeling and extracting information from observational data, thus contributing to a better understanding of a target system and mechanism that are, in general, complex and nonlinear. However, according to the authors' knowledge, there are a limited number of studies in veterinary epidemiology that

adopt ML-based methods, and most of them still neglect the importance of proper and careful tuning of models parameters [26-28]. For example, RF was used in a cross-sectional study that aimed at assessing risk factors that may have led to spillover of pH1N1 from humans to swine in Cameroon, Central Africa [26]. In human epidemiology, RF has been already applied in Diabetic Retinopathy (DR) classification analysis for early detection of this illness based on clinical and fundus photography data [16]. Results suggested that RF was a valuable tool to diagnose DR, producing higher classification accuracy than logistic regression, and that the most relevant variables detected by this ML algorithm are meaningful and correlate well with known risk factors.

In this paper, we aim to investigate the use of RF in the analysis of cross-sectional data collected in a BVDV prevalence study. As previously discussed, the application of this ML algorithm is still uncommon for this type of task. Hence, this study has the following main objectives: (1) train a RF model that provides a good predictive power for the collected data, (2) perform a variable importance analysis using the RF model and the well-established Gini index method to identify potential BVDV predictors, (3) investigate the effect of feature selection on the overall performance of the RF model, carefully assessing the impact on the accuracy and the sensitivity-specificity balance, and, finally, (4) compare RF performance with that obtained by other popular ML algorithms and by logistic regression, examining their predictive power and robustness on the scenario of interest.

Materials and methods

Based on data collected from a prevalence study of reproductive disease in dairy cattle in the State of Rio Grande do Sul, Brazil, a RF model was trained and evaluated with respect to model accuracy, followed by variable importance analysis.

Study design-data collection

Study area and target population

Rio Grande do Sul is the southernmost state of Brazil, with a total area of 268 781.896 km² and 497 municipalities. The cattle population is approximately 13.5 million, 10% of which are dairy cattle [29]. Rio Grande do Sul is the second largest milk-producing state, in which milk production is clustered in six well-defined regions [30]. The study area is explained in more detail in [31].

The target population of data collection included all dairy herds in the state of Rio Grande do Sul. According to the official data from the Office of Agriculture, Livestock and Agribusiness of the State of Rio Grande do Sul 81 307 dairy herds were registered. Descriptive statistics of the studied population can be found in Additional file 1.

Survey design and sample collection

First, a cross-sectional survey was performed to estimate the BVDV, *Neospora caninum* and Infectious Bovine Rhinotracheitis (IBR) prevalence in dairy herds based on (bulk tank milk) BTM samples and to identify the associated risk factors, required by the Office of Agriculture, Livestock and Agribusiness of the State of Rio Grande do Sul. A one-stage stratified random sample design was used. Those farms from which one BTM sample was collected were considered a sampling unit. A stratified sample, which was proportional to the herd population present in each of the seven regions, was performed, and each herd was randomly sampled from all the individual strata. These regions are subdivisions of Brazilian states that are grouped according to proximity and share common agroecological characteristics. The sample size was calculated using R Foundation for Statistical Computing, Vienna, Austria (Package EpiCalc), considering the following parameters: total dairy herds registered at the moment (81 307), 50% expected prevalence, 95% confidence interval, and 5% of absolute precision. The minimum sample size required was 384 dairy herds; however, 388 herds were collected to have a safety margin of extra farm samples.

Bulk tank milk collection

For each herd, a total of 12 mL of milk was collected directly from the milk container immediately after the entire volume had been homogenized. During sampling and transportation, the raw milk was kept under refrigeration between 2 and 8 °C without preservatives. Following an overnight rest, a 1.2 mL sample of skim milk was collected and kept at −20 °C until analysis.

Serological assay and interpretation

The SVANOVIR BVDV p80-AB blocking BVDV ELISA (enzyme-linked immunosorbent assay) was used to detect the BTM samples positive for anti-BVDV antibodies. This blocking ELISA was developed to identify antibodies against the protein p80/NS3, which enables the differentiation between vaccination antibodies and antibodies produced by natural infection**.** All milk samples were centrifuged for 15 min at 2000 × *g*, according to the manufacturer's instructions. The absorbance at a single wavelength of 450 nm (A_{450}) was determined using a spectrophotometer (Asys Expert Plus, Asys Hitech GmbH, Austria). For the herd prevalence, the percent of inhibition (PI) values were calculated in the same manner as the positive control, as well as for each sample, using the following formula:

$$\boldsymbol{PI} = \frac{\boldsymbol{OD_{Negative\ control} - OD_{Sample\ or\ Positive\ control}}}{\boldsymbol{OD_{Negative\ control}}} \times \mathbf{100} \qquad (1)$$

Herds with PI ≥ 30% were considered to have a high probability to harboring an active infection and/or to have at least one positive cow contributing to the sample.

Random forest

In this study we built a RF classifier based on the epidemiological observational data collected from a set of BVDV positive (24%) and negative (76%) farms. The model training process is represented in the flowchart of the study (Figure 1). Since RF algorithm is not routinely used in veterinary epidemiology, we dedicate this section to explain its basis.

Random forest is an example of a machine learning method for classification and regression analysis that uses an ensemble of randomized decision trees to define its output. The algorithm constructs a collection of decision trees using the traditional classification and regression trees methodology (CART) [32] (Figure 2A) and combines the predictions from all trees as its final output when predicting the class of new instances (Figure 2B), making it accurate and robust in relation to other ML algorithms [18]. In classification tasks, as is the case in the current study, combination is performed by means of majority voting among the individual decision trees. Briefly, when classifying new instances from an input variables vector, the mode of the classes returned by the classification performed by individual trees is defined as the final output of the RF model. Hence, supposing we have 100 trees in the forest, among which 70 predict a particular instance as positive for BVDV and the other 30 predict it as negative, the final RF prediction would be positive for BVDV given the majority of votes for this class.

Each decision tree composing the forest has the standard flowchart-like structure, in which internal (split) nodes test variables and branch out according to their possible values, and leaf (terminal) nodes assign a classification for all instances that reach the leaf. The tree growing process in RF is also based in binary recursive splitting that aims at maximizing the decrease of impurity at each node, where impurity can be evaluated by heterogeneity for classification trees (if the response is of categorical type). Nonetheless, in constructing the ensemble of trees, RF incorporates two types of randomness. First, each tree is built using a random bootstrap sample of the original training data (~2/3 of samples), drawn by sampling with replacement (Figure 2A). Second, at each candidate split in the tree growing process, a subset of variables is randomly selected among all available variables to decide node splitting, and the best split among these variables is chosen based on the smallest node impurity [18,33]. Here, we adopt the well-known Gini index as a measure of node impurity. The tree growing procedure is performed recursively until a minimum node size is reached, which is parameterized by the user, or until no further improvement can be made [34]. The two main parameters of the RF algorithms are the number of random variables (predictors) to evaluate at each node split and the number of trees to grow in the ensemble.

The methodology underlying the RF algorithm has interesting properties that make them especially appealing for classification tasks. To begin with, the mechanism applied for tree growing allows the estimation of the most important variables for classification, and generates an internal unbiased estimate of the generalized error drawn from the data left out of the bootstrap sample used as a training set, called out-of-bag (OOB) data, which corresponds to about ~1/3 of the original data. In addition, the fact that the predicted class represents the mode of the outputs returned by individual trees gives robustness to this ensemble classifier in relation to a single tree. Finally, the bootstrapping procedure and

the out-of-bag estimates make RF more accurate and less sensitive to issues such as overfitting, outliers and confounding in comparison to other statistical and machine learning methods [18,33].

In this study, the learning process was carried out with the randomForest and caret packages for the R statistical environment [35,36].

Data preparation

Given the severe class imbalance observed in the data and the general difficulty of machine learning methods to handle this issue [37], we have incorporated a down-sampling procedure in the model learning functions provided by the caret R package, which samples the majority class to make its frequency closer to the rarest class. This procedure aims at avoiding the ML algorithm's tendency to be strongly biased towards the majority class, consequently misclassifying a lot of instances related to the minority class.

The original dataset was randomly and uniformly (i.e., maintaining the same proportion of classes as in the original dataset) split into a training set (80% of observations) and an independent testing set (20% of observations). This subdivision reflects an attempt to compose a minimum sample size that would be representative in future applications of the model and is a common strategy for evaluating ML models when external validation data is not available. The training set was applied for training our classifier using a cross-validation process and the testing set was further used to compare models performance based on independent test data.

Variables

The set of 40-predictor variables collected in the survey performed and used to train the BVDV classification model were: (1) who inseminates the animals, (2) number of neighboring farms that have cattle, (3) what proportion of the farm income is based on milk production, (4) for how many years has the farm produced milk, (5) frequency of technical assistance, (6) is rectal palpation performed routinely, (7) the number of different inseminators in the last year, (8) what is the origin of the bulls, (9) frequency of veterinary assistance, (10) are the animals placed in quarantine before introduction, (11) what is the origin of animals brought into the farm, (12) how often does the fence between/among farms that hold cattle collapse, (13) how the cows are milked, (14) was there an increase in abortions, (15) does calving occur in closed barns, (16) number of cows lactating at the sampling moment, (17) were animals vaccinated for BVDV, (18) was there a rise of mating failure, (19) do animals share the same feed and water containers, (20) number of cows not lactating at the sampling moment, (21) is colostrum stock available, (22) total farm area in hectares, (23) are paddocks available for sick animals, (24) who administers the medications, (25) is blood from a sick animal injected into the healthy ones ("Premunição"), (26) within the last year have animals been sent to fairs, (27) has the farmer seen weak born calves, (28) were pregnant cows introduced, (29) total area for cow farming, (30) has the farmer seen weak calves, (31) were new

animals introduced in the last year, (32) possibility of direct contact (over the fence) between animals from the neighboring farm, (33) animals are grouped based in age category, (34) is the inseminator always the same, (35) does the farm have technical assistance, (36) is natural mating used, (37) does the farm have bulk milk tank, (38) is artificial insemination used, (39) does calving occurs in the fields, (40) does the farm have veterinary assistance. See Additional file 2 for the frequency of important predictor variables.

Model training

The RF model was trained with the training set derived from the original data (i.e., 80% of data) and the complete set of variables using the randomForest package in R. The number of trees induced in the training process was configured to 500 trees following the suggestion of the authors, and the number of variables (*mtry*) randomly sampled as candidates for node splitting during the tree growing process was optimized using the caret package in the R environment. In training the model, we adopted a repeated 10-fold cross-validation technique to better estimate its performance and generalization power, and to prevent overfitting and artificial accuracy improvement due to use of the same data for training and testing the classifier.

Once the model was trained, we investigated the effect of multicollinearity over the performance of RF. For this purpose, we computed the correlation matrix for the set of 40 variables using Pearson correlation and identified highly correlated predictors among our independent variables. Next, we selected some of the highly correlated variables to discard from the analysis based on plausibility criteria and repeated the RF training process without these variables, comparing its performance with the original RF model.

An interesting property of RF is that it naturally provides estimates of variable importance, which are computed during model training by evaluating the average decrease in the nodes' impurity measured by Gini index. The importance of a variable is defined as the Gini index reduction for the variable summed over all nodes for each tree in the forest, normalized by the number of trees [38]. Hence, the higher the Gini importance, the more relevant that variable is for maintaining the predictive power of the RF model. Although RF are capable of modeling a large number of variables and achieving good prediction performance, finding a small number of variables with equivalent or better prediction ability is highly desired not only because it is helpful for interpretation, but also easy for practical use as strategies for disease control [38].

Thus, after running the first round of model training and obtaining the Gini importance for each of the 40-predictor variables of our data set, we performed a restricted forward feature selection and verified the impact of variables inclusion over the model's predictive accuracy in an incremental fashion. This step aims at identifying irrelevant variables that may mislead the algorithm and increase the generalization error [39]. Specifically, we trained several RF models, starting from a model trained upon a single variable, and subsequently adding new variable one at a time, from the most relevant to the least relevant. For each of the classifiers generated, we evaluated its performance by

computing the AUC score, specificity and sensitivity for the OOB data. Based on this analysis, we selected the top important predictor variables that optimized model's performance and ran the training process again, generating a simplified RF classifier that considers only the most impactful variables.

Finally, we explored the relevance of variables for classification results by partial dependence plots, which are useful for providing insights of the marginal effect of a given variable over the desired outcome. The partial dependence of a variable's effect is best understood by examining general patterns in relation to the values of the predictor variable rather than the specific values of partial dependence [40]. Because we are modeling binary classification (i.e., presence/absence of BVDV), partial dependence values are given in "logit" scale and are computed in relation to the probability for the positive class [19].

Model performance assessment

The model performance was assessed by computing the total prediction accuracy (ACC), specificity (SPE) and sensitivity (SEN) based on the confusion matrix. This matrix quantifies the number of instances in the test data classified as false positive (FP), true positive (TP), false negative (FN), and true negative (TN). We also plotted the area under the Receiver Operating Characteristic (ROC) curve. The area under the ROC curve gives us the AUC score, interpreted as the probability that a classifier will rank a random chosen positive instance higher than a random negative one.

$$\boldsymbol{ACC} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (2)$$

$$\boldsymbol{SPE} = \frac{\boldsymbol{TN}}{\boldsymbol{TN+FP}} \times \mathbf{100}\% \quad (3)$$

$$\boldsymbol{SEN} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \times \mathbf{100}\% \quad (4)$$

Comparing RF to other machine learning methods

In order to assess the predictive power of RF in comparison to other ML techniques, we performed a comparative evaluation of the RF classifier with two other popular methods, namely Support Vector Machine (SVM) and Gradient Boosting Machine (GBM), which have also not been extensively assessed in veterinary epidemiology. SVM was introduced by [41] and is based on a statistical-learning technique known as structural risk minimization [41,42], being first used in observational epidemiology studies in 2005 [43]. GBM, on the other hand, is an ensemble method that combines regression trees with weak individual predictive performances into a single model with high performance [34,40].

For such comparison, we adopted the same procedure used for RF training, i.e., 10 repetitions of 10-fold cross-validation, assuring that the exact same data points are used in each step of model training and testing. In other words, we maintained the same

subsampling of the training data used in the cross-validation process. In addition, we applied the caret R package to train SVM and GBM models, tuning some of the parameters involved in order to carry a fair comparison with RF. Based on the results from cross-validation, we performed a first round of comparison among models, contrasting their AUC score, sensitivity and specificity drawn from the average confusion matrix. Finally, the differences between models performance in terms of AUC scores were assessed with a pairwise Wilcoxon rank test in order to test for statistical significance.

Comparing RF to logistic regression

Since we are interested in suggesting the use of RF as an alternative method for traditional statistical approaches, we also assessed its performance relative to logistic regression, which is frequently used for the analysis of risk factors. Logistic regression was estimated with the glm() function in R environment and performance evaluation was carried out based on 10 repetitions of 10-fold cross-validation using the caret R package. To assure a fair comparison, we run the logistic regression analysis with the same distribution of data used for RF training among folds and across all repetitions of cross-validation.

Models evaluation on independent testing data

In addition to evaluating the methods performance using cross-validation, we also assessed their predictive accuracy with an independent test set derived from the original data. As aforementioned, during data preparation the original data set was subdivided in training data (80%) and testing data (20%), which is not used in the cross-validation procedure and thus can be regarded as an independent test set.

This approach is recommended when no external independent data are naturally available [44,45], which is the case in our study. Although cross-validation is well known for providing precise and unbiased estimative of the predictive accuracy and generalization power of ML classifiers, we opted to follow the common practice and conduct another comparison among models with explicitly independent data.

Results

Performance of the RF model

The confusion matrix for the tuned RF model trained with all available predictor variables (n = 40) and *mtry*=25 (optimized value computed by caret R package), averaged over the 10 repetitions of the 10-fold cross-validation, is shown in Table 1. We evaluated the confusion matrix for the final RF model, obtaining the following performance metrics: ACC: 67.42% (±3.69); SPE: 67.65% (±3.85) and SEN: 62.26% (±3.44). Despite optimizing parameters and adopting a down-sampling procedure, RF had an average error

rate of 32.03% for the negative class (negative for BVDV) and 36.78% for the positive class (positive for BVDV), with a standard deviation of 1.30% and 2.46%, respectively.

Analysis of the correlation matrix computed for the set of 40 variables (Additional file 3) suggested that a small set of independent variables is highly correlated. Based on plausibility criteria, we eliminated the highly correlated variables, namely (5) frequency of technical assistance, (9) frequency of veterinary assistance, (11) what is the origin of the animals brought into the farm and (30) has the farmer seen weak calves, and repeated the training process. We observed a minimal change in the RF model performance after the elimination of correlated variables, with the highest (but still modest) impact found for sensitivity, i.e., an increase from 62.26% to 65.10%.

Variable importance

We performed a variable importance analysis assessing the average decrease in the nodes' impurity measured by the Gini index during the construction of the random forest model. Figure 3 presents the result of this analysis, with the variables ranked by their Gini importance. As we may observe, the variables (1) who inseminates the animals, (2) the number of neighboring farms that have cattle, (3) what proportion of the farm income is based on milk production and (4) for how many years has the farm produced milk are the four most important variables for BVDV prediction found in this analysis, since they are associated to the highest Gini importance.

The result of the restricted forward feature selection carried after variable importance analysis can be seen in Figure 4. The best performance balance considering AUC score, specificity and sensitivity, as well as model complexity, seems to be associated with the model trained with the top 25-predictor variables. Hence, the RF training procedure was repeated for this subset of variables (Figure 3), optimizing model's parameters by means of the caret package in R. The best tune for *mtry* was 16, and the classification results for this model are shown in the confusion matrix depicted in Table 2. We noticed that the model trained with 25 variables, generated after feature selection, presented a slight increase in the average accuracy (ACC: 67.75%) and specificity (SPE: 67.98%) in relation to the model trained with the total set of variables, whilst no variation was observed for sensitivity. Nonetheless, this increase is not statistically significant, and hence in this scenario feature selection does not seem to introduce important benefits to the performance of the RF model.

To better understand the effects of the most important variables over classification results, we explored the partial dependence plots for the top 25-predictors (Figure 5), which give a graphical depiction of the marginal effect of a variable on the class probability. Greater y-values indicate that an observation for a specific variable is associated with higher probability for classifying new instances as BVDV positive.

As this analysis suggests, (B) the number of neighboring farms that have cattle and (G) the number of different inseminators in the last year had a strong linear correlation with BVDV. Moreover, we observed that disease occurrence was highly influenced by

observations related to some specific variables, mainly by (A) insemination performed by the owner or farmer, (C) milk production representing about 61-80% of far income, (E) technical assistance conducted annually, (F) rectal palpation performed routinely, (I) veterinary assistance held annually, (J) animals placed in quarantine before introduction, (M) milking process performed in an automatic fashion, (X) administration of medications performed by a technician and (Y) the regional habit of injecting blood from a sick animal into a healthy one ("Premunição"). In contrast, there was no significant relationship between BVDV occurrence and the variables (O) does calving occurs in closed barns, (P) number of cows lactating at the sampling moment, (S) do animals share the same feed and water containers, (T) number of cows not lactating at the sampling moment, (U) is colostrum stock available and (W) are paddocks available for sick animals.

Comparative evaluation of RF, SVM and GBM

The results of the comparative analysis based on the average AUC scores, computed as the mean of the area under the ROC curves over all repetitions of cross-validation, were 0.702 for RF, 0.690 for GBM and 0.687 for SVM. The highest specificity was achieved by SVM (69.45% ± 4.05), followed by RF (67.65% ± 3.85) and GBM (66.15% ± 2.58). On the other hand, RF achieved the highest sensitivity (62.26% ± 3.44), followed by GBM (61.73% ± 5.33) and SVM (57.60% ± 4.73).

In a visual analysis of density distributions of AUC scores obtained for each classifier (Figure 6A), RF presents a distribution slightly shifted to the right in relation to others, indicating a tendency in provide a better predictive accuracy than GBM and SVM. Nonetheless, differences among methods performance in terms of AUC scores are not statistically significant according to a pairwise Wilcoxon Ranked Sum test using the Benjamini-Hochberg procedure to correct for multiple comparisons. The lowest *p*-value was associated to the comparison between RF and SVM (*P*-value = 0.064), followed by the comparison between RF and GBM (*P*-value = 0.075).

We also compared the distribution of sensitivity and specificity metrics across all repetitions of cross-validation following the same methodology, and we found that SVM has better specificity performance than RF and GBM (*P*-value < 0.05), while both RF and GBM outperform SVM in terms of sensitivity (*P*-value < 0.05).

Comparison between RF and logistic regression

As expected according to our theoretical motivation, we observed a superior performance of RF relative to logistic regression. While RF had an average AUC score of 0.702, the model estimated by logistic regression achieved an AUC score of 0.610 across all repetitions of cross-validation. The density plots drawn from the cross-validation procedure makes evident the better predictive power of RF, which presents an AUC scores distribution shifted to the right of that related to logistic regression (Figure 6B).

Moreover, we observed that the classification provided by RF is much more balanced in terms of sensitivity and specificity than logistic regression. The average specificity was 67.65% (±3.85) for RF and 61.36% (±3.33) for logistic regression, while the average sensitivity achieved by these methods were 62.26% (±3.44) and 56.30% (±3.84) for RF and logistic regression, respectively.

Models evaluation with independent testing data

In addition to the comparative analysis carried out among classifiers using cross-validation, we evaluated the models' predictive accuracy with independent test data. Results in terms of the ROC curves are shown in Figure 7A for the ML algorithms. The corresponding AUC scores are 0.697 for RF, 0.703 for SVM and 0.785 for GBM.

Differently from the cross-validation technique that ensures every instance in the data set will be used exactly once for model validation, the initial partitioning of data is performed a single time in a random fashion, and may generate a testing data set for which GBM, fortunately, have a superior performance – an effect that is out of our control. To test for this possibility, we repeated the process of model training and testing 10 times, each of which with a random (and thus potentially different) partitioning of data into training and testing sets, keeping the proportions of 80% and 20%, respectively. We performed this procedure for the three classifiers, i.e., RF, SVM and GBM, and compared their average performance for the independent test data across all repetitions. We observed that RF outperforms the other classifiers in 6 out of the 10 repetitions, while in the remaining 4 the best performance is achieved by GBM (Additional file 4). Although the average AUC score of RF is only slightly better than GBM, 0.7466 vs. 0.7301, the worst and best performances achieved by RF show a performance gain of 12.09% and 7.13% in relation to the worst and best models trained by GBM, respectively.

Regarding the comparative evaluation between RF and logistic regression, similarly to what was observed from the cross-validation procedure, RF presented a more robust performance for independent testing data in relation to logistic regression. The ROC curves are shown in Figure 7B, corroborating the better predictive accuracy of RF in contrast to logistic regression.

Discussion

In this study, we trained a RF model based on cross-sectional data derived from an investigation for BVDV prevalence carried in Southern Brazil, aiming to identify important predictors for disease occurrence and to evaluate the predictive power of this machine learning model in this specific domain. To the best of our knowledge, this is one of the few studies in veterinary epidemiology that performs an investigation based on machine learning algorithms adopting a careful training process, which encompasses parameters optimization and a strategy to treat a severe class imbalance problem. In addition, it was also the first time that a comparative evaluation among RF, SVM and GBM models was held in this context, adopting appropriate methods for model tuning and a repeated 10-fold cross-validation technique.

Based on the classification results by RF, we noticed that our model's performance has shown an overall good predictive accuracy and quite balanced sensitivity and specificity across all repetitions of the cross-validation. The data-driven analysis carried by RF, without a priori assumptions about the relationship between the dependent and independent variables, has a great potential to outperform the traditional logistic regression, as experimentally verified for our data, suggesting that RF could be a valuable tool in cross-sectional studies. The reader should be aware that our results do not come from basic measures of total classification accuracy and error rates; instead, we have adopted robust evaluation approaches and made important interventions for training and optimizing the machine learning classifiers, providing a more appropriate application of these methods to our scenario. Specifically, we have optimized the number of predictor variables selected for splitting a new node during the production of the decision trees, and we decided to not optimize the number of trees in the forest based on the former discussion that RF is not very sensitive to this last parameter [35].

Despite its satisfactory performance, our classifier has missed on average more positive than negative cases of BVDV, even after the application of the down-sampling strategy (Table 1). Most standard algorithms assume or expect balanced class distribution or equal misclassification costs [46], so when a complex imbalanced data set is used, these algorithms fail to properly represent the distributive characteristics of the data and resultantly provide unfavorable accuracies across the classes of the data [46]. In our data we found an imbalance in a form commonly referred to as "intrinsic", which means the imbalance is a direct result of the nature of the data space [46]. We analyzed the effects of the down-sampling procedure over classifiers performance, comparing the results obtained from training with and without handling the data imbalance issue, and we observed that all three methods suffered impact from the severe data imbalance over their sensitivity. When training is carried without treating this issue, models' sensitivities were in the approximate range of 11.5% to 20%, which is clearly lower than the values of 57.60%, 61.73% and 62.26% achieved by SVM, GBM and RF, respectively. Hence, we observed that adopting this pre-processing strategy in data sets containing classes that are highly under-represented in comparison to others may introduce important benefits for data analysis, although in this case it did not completely solved this issue.

The final variables ranking in a descending order of importance as provided by RF's variable importance analysis (Figure 3) suggests that the main variables involved in BVDV prediction are related to reproduction-associated factors, movement of many people into and out of the farms, and direct contact among animals, as we discuss further. Feature selection has been previously shown to result in slight error reductions [47], and this step is normally performed in order to remove variables that do not contribute to the performance of the model, either because they do not play an important role on error reduction or because they have a minimal effect on the discriminant power of the RF classifier [48]. One can notice that although performance improvement was not so expressive after feature selection (Table 2), we still observed a slight gain in terms of accuracy and specificity. The top 25 variables model is therefore more efficient, as it provides a performance as good as the model trained with the complete set of 40 variables despite the reduction in model complexity.

Regarding the results of variable importance analysis, we discuss only the most relevant variables due to space limitation. The most impactful variable for BVDV prediction was related to farms that perform AI (Figure 5A), a factor that has been considered a predictor for BVDV globally, especially when semen is used from untested bulls or when farms use AI along with natural mating in order to "guarantee" the success of a pregnancy, a common and unsafe practice in Brazil [10]. AI is an important route of transmission of BVDV because semen remains infective, which is evident by the demonstration that susceptible cows can become infected following insemination [15,49,50,51]. A remarkable new association that we found was that when AI is performed by the owner or someone that is responsible for the farm, a common reproductive practice in Brazil and other countries, the influence on BVDV cases was evidently harmful, increasing the probability of disease occurrence. It was also reaffirmed that the number of neighboring cattle farms where there is chances of direct contact between cattle over the fence was a predictor for BVDV [13]. Others have identified the direct contact over fence lines one of the hardest to control [52]. In our analysis, we showed that the partial dependence of BVDV on this variable increases as the numbers of neighbors' increases, and that BVDV occurrence rises abruptly when there are three neighboring farms. The occurrence of BVDV was also influenced by factors related to milk production. When milk production was reported to represent 61 to 80% of farm income (Figure 5C), we observed a high association with BVDV, most likely due to milk production with intensive pressure on cow performance. It was found that farms that have produced milk for up to nine years had the highest influence on disease occurrence in contrast to farms that have been harvesting milk for longer periods (Figure 5D) this fact may be related to the inexperience of the farmer.

Partial dependence analysis also suggested that rectal palpation performed routinely (Figure 5F) causes significant influence on BVDV occurrence. It has been found that indirect transmission of BVD virus can be spread by veterinary equipment such as nose tongs, needles and protective rubber gloves worn during rectal examination [53,54]. Others [55] had also reported that rectal palpation performed consecutively on different animals without proper hygiene (e.g., without replacing glove between animals) might play an important role in the transmission of BVDV. Moreover, the number of different inseminators that had visited the farm in the past year showed a linear influence on BVDV (Figure 5G). We observed that as the number of inseminators increases, the chances of predicting positive cases of BVDV were also higher, probably due to intense people movement acting as fomites.

In order to compare the RF model against other classifiers that have similar literature, a repeated 10-fold cross-validation was performed, averaging model accuracy measures over all repetitions. We found a better overall performance of RF in relation to SVM and GBM, especially in terms of specificity and sensitivity balance, but results were very close among ensemble-based algorithms (i.e., RF and GBM). Although the difference between the AUC scores of these two classifiers are not statistically significant, we found based on visual analysis of kernel density estimates that the probability distribution of RF is shifted to the right of GBM and SVM distributions, which suggests that RF has a tendency to produce higher AUC scores (i.e., achieve best performance) in relation to the

latter. Others had previously found similar results when testing the performance of all tree classifiers, but in the previous study, GBM and SVM performed relatively better than RF [56]. The poor results related to SVM may be due to the fact that the performance and prediction results of this classifier are heavily dependent on the chosen values for the tuning parameters [57-59]. Although we adopted a parameter optimization procedure based on grid search methods that minimize total error rates, a more exhaustive study towards the evaluation of classifier's performance upon parameters optimization, combined with the application of other optimization techniques, could lead to an even better performance. However, this analysis is out of the scope of our work.

Surprisingly, for tests with independent data, GBM showed an improved performance, which is better and more balanced than the performance achieved by RF and SVM. This may indicate a better generalization power of this algorithm, but it may also be an artifact of data partitioning, which randomly generates a test set for which GBM has a more favorable chance of producing accurate classification. However, due to the random nature of the procedure, repeated partitioning of the original data into training and testing sets may produce results with large variability, both qualitative and quantitative, and consequently provide less consistent insights than the analysis performed with cross-validation. We verified this effect by repeating 10 times the complete training process, from data preparation (and consequently data partitioning) to models evaluation, based on which we observed significant variance in methods performance. Briefly, RF and GBM were always the top-performing classifiers, but in 6 out of the 10 repetitions, RF outperformed GBM, showing that the outcome of this comparison is highly dependent on initial data partitioning. Hence, we emphasize that the 10-fold cross-validation technique is more powerful in reducing overfitting and more precise for assessing the predictive power of machine learning methods, providing an unbiased estimative of how a classifier model will generalize to an independent data set.

It should be noted that GBM is functionally similar to RF because it creates an ensemble of trees and uses randomization during this process. This fact could support the similar results observed for these two methods. However, whereas RF builds the trees in parallel and these trees "vote" simultaneously on the preferred class during prediction, GBM creates a series of trees in which the prediction receives incremental improvement by each tree in the series [60].

In life sciences, random forests have been used to analyze genomic data [61,62], in ecology they have been successfully used as classifiers [19,63,64], and herein they are used for cross-sectional studies in veterinary epidemiology. Random forests proved to have good accuracy, sensitivity and specificity, showing a discriminant power that is highly competitive with other ML-based methods for detecting biologically plausible predictors of BVDV. Based on these results, we believe that RF is a promising computational approach for cross-sectional studies in veterinary epidemiology and should be more frequently considered as an alternative for traditional statistical methods.

Moreover, our model demonstrated a novel use of observational data that goes beyond the previously identified predictors. The application of machine learning extends the

usefulness of classical risk factors found on the basis of traditional statistical approaches. Based on the proposed RF model, we could take a closer look at some classical predictors and found important details regarding their relationship with disease occurrence, mainly regarding reproduction management, which should be considered for disease control and eradication. One should take this investigation further ahead in order to clarify how the important reproduction variables contribute to BVDV in other countries.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GM and MRM designed and conducted data analysis. GM collected the samples. LGC supervised the whole study.

Acknowledgements

We thank the field team who carried out all the blood sampling and questionnaire interviews. Part of this study was supported by FUNDESA and CNPq. MRM acknowledges the financial support from CAPES and HCPA.

References

1. Simmonds P, Becher P, Collet MS, Gould EA, Heinz FX, Meyers G, Monath T, Pletnev A, Rice CM, Stiansny K, Thiel HJ, Weiner A, Bukhet J (2011) Family Flaviviridae. In: King AMQ, Adams EB, Carstens EJ, Lefkowitz (eds), Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses 2011, 9th ed. Elsevier Academic Press, United States: p. 1003-1020

2. Houe H (1999) Epidemiological features and economical importance of bovine virus diarrhoea virus (BVDV) infections. Vet Microbiol 64:89-107

3. Niza-Ribeiro J, Pereira A, Souza J, Madeira H, Barbosa A, Afonso C (2005) Estimated BVDV-prevalence, -contact and -vaccine use in dairy herds in Northern Portugal. Prev Vet Med 72:81-85

4. Valle PS, Martin SW, Tremblay R, Bateman K (1999) Factors associated with being a bovine-virus diarrhoea (BVD) seropositive dairy herd in the Mùre and Romsdal County of Norway. Prev Vet Med 40:165-177

5. Presi P, Struchen R, Knight-Jones T, Scholl S, Heim D (2011) Bovine viral diarrhea (BVD) eradication in Switzerland--experiences of the first two years. Prev Vet Med 99:112-121

6. Humphry RW, Brülisauer F, McKendrick IJ, Nettleton PF, Gunn GJ (2012) Prevalence of antibodies to bovine viral diarrhoea virus in bulk tank milk and associated risk factors in Scottish dairy herds. Vet Rec 171:445

7. Rodrigo SL, Perea A, García-Bocanegra I, Jos AA, Vinicio JD, Ramos R, Carbonero A (2012) Seroprevalence and risk factors associated with bovine viral diarrhea virus (BVDV) infection in non-vaccinated dairy and dual purpose cattle herds in Ecuador. Trop Anim Health Prod 44:645-649

8. Sarrazin S, Veldhuis A, Méroc E, Vangeel I, Laureyns J, Dewulf J, Caij AB, Piepers S, Hooyberghs J, Ribbens S, Van Der Stede Y (2012) Serological and virological BVDV prevalence and risk factor analysis for herds to be BVDV seropositive in Belgian cattle herds. Prev Vet Med 108:28-37

9. Gard JA, Givens MD, Stringfellow DA (2007) Bovine viral diarrhea virus (BVDV): epidemiologic concerns relative to semen and embryos. Theriogenology 68:434-442

10. Chaves NP, Bezerra DC, Sousa VE, Santos HP, Pereira HM (2010) Frequency of antibodies and risk factors of bovine viral diarrhea virus infection in non-vaccinated dairy cows in the Maranhense Amazon region, Brazil. Ciên Rur 40:1448-1451

11. Luzzago C, Frigerio M, Piccinini R, Dapra V, Zecconi A (2008) A scoring system for risk assessment of the introduction and spread of bovine viral diarrhoea virus in dairy herds in Northern Italy. Vet J 177:236-241

12. Lindberg AL, Alenius S (1999) Principles for eradication of bovine viral diarrhoea virus (BVDV) infections in cattle populations. Vet Microbiol 64:197-222

13. Machado G, Egocheaga RMF, Hein HE, Miranda ICS, Neto WS, Almeida LL, Canal CW, Stein M, Corbellini LG: Bovine Viral Diarrhoea Virus (BVDV) in dairy cattle: a matched case–control study. Transbound Emerg Dis in press

14. Mainar-Jaime RC, Berzal-Herranz B, Arias P, Rojo-Vazquez FA (2001) Epidemiological pattern and risk factors associated with bovine viral-diarrhoea virus (BVDV) infection in a non-vaccinated dairy-cattle population from the Asturias region of Spain. Prev Vet Med 52:63-73

15. Almeida LL, Miranda ICS, Hein HE, Santiago NW, Costa EF, Marks FS, Rodenbusch CR, Canal CW, Corbellini LG (2013) Herd-level risk factors for bovine viral diarrhea virus infection in dairy herds from Southern Brazil. Res Vet Scie 93:901-907

16. Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT (2014) Application of random forests methods to diabetic retinopathy classification analyses. PLoS One 9:e98587

17. Mansiaux Y, Carrat F (2014) Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with H1N1pdm influenza infections. BMC Med Res Methodol 14:99

18. Breiman L(2001) Random forests. Mach Learn 45:5-32

19. Cutler RD, Edwards TC, Beard KH, Cutler KT, Gibson HJ, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783-2792

20. Breiman L (2001) Statistical Modeling. The Two Cultures**.** Stat Scie 16:199-231

21. Benito GM, Blazek R, Neteler M, Sánchez DR, Sainz-Ollero H, Furlanello C (2006) Predicting habitat suitability with machine learning models: the potential area of Pinus sylvestris L. in the Iberian Peninsula. Ecol Model 197:383-393

22. Peters J, De Baets B, Verhoest NEC, Samson R, Degroeve S, Becker PD, Huybrechts W (2007) Random forests as a tool for ecohydrological distribution modelling. Ecol Model 207:304-318

23. Slabbinck B, De Baets B, Dawyndt P, De Vos P (2009) Towards large-scale FAMEbased bacterial species identification using machine learning techniques. Syst Appl Microbiol 32:163-176

24. Kampichler C, Wieland R, Calmé S, Weissenberger H, Arriaga-Weiss S: Classification in conservation biology (2010) a comparison of five machine-learning methods. Ecol Inform 5:441-450

25. Pino-Mejías R, Cubiles VMD, Anaya-Romero M, Pascual-Acosta A, Jordán-López A, Bellinfante-Crocci N (2010) Predicting the potential habitat of oaks with data mining models and the R system. Environ Model Softw 25:826-836

26. Larison B, Njabo KY, Chasar A, Fuller T, Harrigan RJ, Thomas TB (2014) Spillover of pH1N1 to swine in Cameroon: an investigation of risk factors. BMC Vet Res 10:55

27. Barco L, Macin M, Ruffa M, Saccardim C, Minorello C, Zavagin P, Lettini AA, Olsen JE, Ricci A (2012) Application of the Random Forest method to analyse epidemiological and phenotypic characteristics of Salmonella 4,[5],12:i:- and Salmonella Typhimurium strains. Zoonoses Public Health 59:505-512

28. Holtkamp DJ, Lin H, Wang C, O'Connor AM (2012) Identifying questions in the American Association of Swine Veterinarian's PRRS risk assessment survey that are important for retrospectively classifying swine herds according to whether they reported clinical PRRS outbreaks in the previous 3 years. Prev Vet Med 106:42-52

29. Instituto Brasileiro de Geografia e Estatística (IBGE): Pesquisa pecuária municipal, efetivo dos rebanhos por tipo de rebanho, Brasil. http://www.sidra.ibge.gov.br. Accessed 13 Jan 2014

30. Zoccal R, Assis AG, Evangelista SRM(2006) Distribuição geográfica da pecuária leiteira no Brasil. Embrapa Gado de Leite, Juiz de Fora, http://ainfo.cnptia.embrapa.br/digital/bitstream/item/65271/1/CT-88-Distribuicao-geografica-da-pecuaria.pdf. Accessed 10 Jul 2014

31. Silva GS, Costa E, Bernardo FA, Groff FHS, Todeschini B, Santos DV, Machado G (2014) Cattle rearing in Rio Grande do Sul, Brazil. Acta Scient Vet 42:1215.

32. Breiman L, Friedman JH, Olsen RA, Stone CJ (1984) Classification and Regression Trees: Chapman & Hall/CRC, Belmont

33. Mendoza RM, Fonseca GC, Loss-Morais G, Alves R, Margis R, Bazzan ALC: RFMirTarget (2013) Predicting Human MicroRNA Target Genes with a Random Forest Classifier. PloS One 8:e70153

34. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York

35. Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2:18-22

36. Kuhn M. caretClassification and Regression Training. 2014. http://CRAN.R-project.org/package=caret. R package version 5.15–052, Accessed 20 Dec 2014

37. Nguyen GH, Bouzerdoum A, Phung SL (2009) Learning pattern classification tasks with imbalanced data sets. In Pattern Recognition, Vukovar, Croatia,, pp 193-208

38. Xi C, Ishwaran H (2012) Random forest for genetic data analysis. Gonomics 99:323-329

39. Domingos P (2012) A few useful things to know about machine learning. Commu ACM 55:78-87

40. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189-1232

41. Vapnik V (1995) The nature of statistical learning theory. Springer, New York, p314

42. Noble WS (2006) What is a support vector machine? Nat. Biotechnol 24:1565-1567

43. Hermans PG, Fradkin D, Muchnik IB, Morgan KL (2006) Prevalence of wet litter and the associated risk factors in broiler flocks in the United Kingdom. Vet Rec 158:615-622

44. Efron B, Tibshirani R (1997) Improvements on cross-validation: the .632+ bootstrap method. J Am Statist Assoc 92:548-560

45. Gerds T, Schumacher M (2007) Efron-type measures of prediction error for survival analysis. Biomet 63:1283-128

46. He H, Garcia AE (2009) Learning from Imbalanced Data. IEEE Trans Knowl Data Eng 21:1263-1284

47. Svetnik V, Liaw A, Tong C, Wang T (2004) Application of Breiman's Random Forest to modeling structure–activity relationships of pharmaceutical molecules. In: Roli F, Kittler, J Windeatt T (Eds.), Multiple Classifier Systems, Fifth International Workshop, MCS 2004, Proceedings, 9-11 June 2004, Cagliari, Italy. Lecture Notes in Computer Science, v. 3077. Springer, Berlin, pp334-343

48. Xiong C, Johnson D, Xu R, Corso JJ (2012) Random forests for metric learning with implicit pairwise position dependence. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 958-966

49. Altamiranda EA, Kaiser GG, Weber N, Leunda MR, Pecora A, Malacari DA, Morán O, Campero CM, Odeón AC (2012) Clinical and reproduction consequencs of using BVDV-contaminated semen in artificial insemination in a beef herd in Argentina. An Repr Scie 133:146-152

50. Kirkland P, Mackintosh S, Moyle A (1994) The outcome of widespread use of semen from a bull persistently infected with pestivirus. Vet Rec 135:527-529

51. Paton DJ, Brockman S, Wood L (1999) Insemination of susceptible and preimmunized cattle with bovine viral diarrhoea virus infected semen. Bra Vet J 146:171-174

52. Niskanen R, Lindberg A (2003) Transmission of bovine viral diarrhoea virus by unhygienic vaccination procedures, ambient air, and from contaminated pens. Vet J 165:125-130

53. Gunn HM (1993) Role of fomites and flies in the transmission of bovine viral diarrhoea virus. Vet Rec 132:584-585

54. Lang-Ree JR, Vatn T, Kommisrud E, Loken T (1994) Transmission of bovine viral diarrhoea virus by rectal examination. Vet Rec 135:412-413

55. Goyal SM, Ridpath JF (2005) Bovine Viral Diarrhea Virus Diagnosis, Management and Control. Blackwell, Iowa

56. Ogutu JO, Piepho H, Schulz-Streeck T (2011) A comparison of random forest, boosting and support vector machines for genomic selection. BMC Proc 5(Suppl 3):S11

57. Duin RPW (1996) A note on comparing classifiers. Pat Recog Lett 17:529-536

58. Meyer D, Leischa F, Hornik K (2003) The support vector machine under test. Neurocom 55:169-186

59. Lim TS, Loh WY, Shih YS (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach Lear 40:203-228

60. Olinsky A, Kennedy K, Kennedy BB (2014) Assessing gradient boosting in the reduction of misclassification error in the prediction of success for actuarial majors. *Math Departt J Articl 5:12-16*

61. Jiang P, Wanf W, Ma W, Sun X, Lu Z (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res 35(Suppl 2):W339-W344

62. Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, Sun X: Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics 2009 25:30-35

63. Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181-199

64. Drew CA, Wiersma Y Huerrmann F (2011) Predictive modeling in landscape ecology. Springer, New York

Figure legends

Figure 1 Flowchart of the study design. Representation of each step of the study.

Figure 2 Random forest model. Example of training and classification processes using random forest. A) Each decision tree in the ensemble is built upon a random bootstrap sample of the original data, which contains positive (green labels) and negative (red labels) examples. B) Class prediction for new instances using a random forest model is based on a majority voting procedure among all individual trees. The procedure carried for each tree is as follows: for each new data point (i.e., X), the algorithm starts at the root node of a decision tree and traverse down the tree (highlighted branches) testing the variables values in each of the visited split nodes (pale pink nodes), according to each it selects the next branch to follow. This process is repeated until a leaf node is reached, which assigns a class to this instance: green nodes predict for the positive class, red nodes predict for the negative class. At the end of the process, each tree casts a vote for the preferred class label, and the mode of the outputs is chosen as the final prediction.

Figure 3 Variable importance analysis performed by RF. The set of 40 variables used for classification, ordered by their importance as estimated by the RF model.

Figure 4 Result of restricted forward feature selection. Performance of the RF model evaluated by means of a restricted forward feature selection. Several RF classifiers were trained adding each of the predictor variables at a time, following the rank obtained from the variable importance analysis, which is based on the mean decrease of Gini index.

Figure 5 Partial dependence plots for the top 25 variables. Partial dependence plots for the top 25 variables with the variable importance scores as calculated by random forests. Plots show the partial dependence of a Relative Occurrence Index value for BVDV on each predictor variable; the y-axis is given in log scale [the logit function gives the log-odds, or the logarithm of the odds $p/(1-p)$].

Figure 6 Comparative evaluation of RF against GBM, SVM and logistic regression based on repeated cross-validation. The performance of the models over several resamples are summarized by a kernel density estimator, which indicates a narrow distribution and slightly shifted to the right (higher values) for RF A) in relation to SVM and GBM and B) in relation to logistic regression.

Figure 7 Evaluation of models performance for independent test data. A) RF, SVM and GBM were also compared using an independent test set, which corresponds to the 20% portion of data that was not used in the training and cross-validation procedures. According to the ROC curves, the GBM classifier outperforms RF and SVM. B) Relative to the logistic regression, a traditional statistical approach used for the analysis of risk factors, RF achieved a more robust performance.

Additional files

Additional file 1 Descriptive statistics on the study population. A descriptive analysis has been performed in order to show an overview of the study population.

Additional file 2 Frequency of important predictor variables. The prevalence of important predictor variables obtained by serological assay results provides details of disease occurrence in the study population.

Additional file 3 Correlation matrix for predictor variables. Negative correlation is represented by red ellipses pending to the left; positive correlation is represented by blue ellipses pending to the right. The exact correlation values are given in the upper panel.

Additional file 4 Models performance for 10 randomly generated independent test data sets. The AUC scores are computed for 10 repetitions of model training and testing. In each repetition, a random portion of 80% of data is used for training, and the remaining 20% for testing (independent data).

Table 1 Classification performance of RF model for the 40 variables. Confusion matrix for the RF model trained with the complete set of predictor variables (n = 40) and a down-sampling procedure, estimated by averaging the results over ten repetitions of 10-fold cross-validation. Standard deviations are given in parenthesis[*].

		Real	
		BVDV-negative	BVDV-positive
Predicted	BVDV-negative	114.0 (6.5)	2.83 (0.25)
	BVDV-positive	54.5 (6.5)	4.67 (0.25)

[*] Performance metrics: ACC: 67.42 (Sd. 3.69); SPE: 67.65 (Sd. 3.85) and SEN: 62.26 (Sd. 3.44)

Table 2 Classification performance of RF model for the top 25 variables. Confusion matrix for the RF model trained with the top 25-predictor variables selected after variable importance analysis, estimated by averaging the results over ten repetitions of 10-fold cross-validation. Standard deviations are given in parenthesis[*].

		Real	
		BVDV-negative	BVDV-positive
Predicted	BVDV-negative	114.55 (6.8)	2.8 (0.20)
	BVDV-positive	53.95 (6.8)	4.7 (0.20)

[*] Performance metrics: ACC: 67.75 (Sd. 3.69); SPE: 67.98 (Sd. 3.85) and SEN: 62.26 (Sd. 3.33).

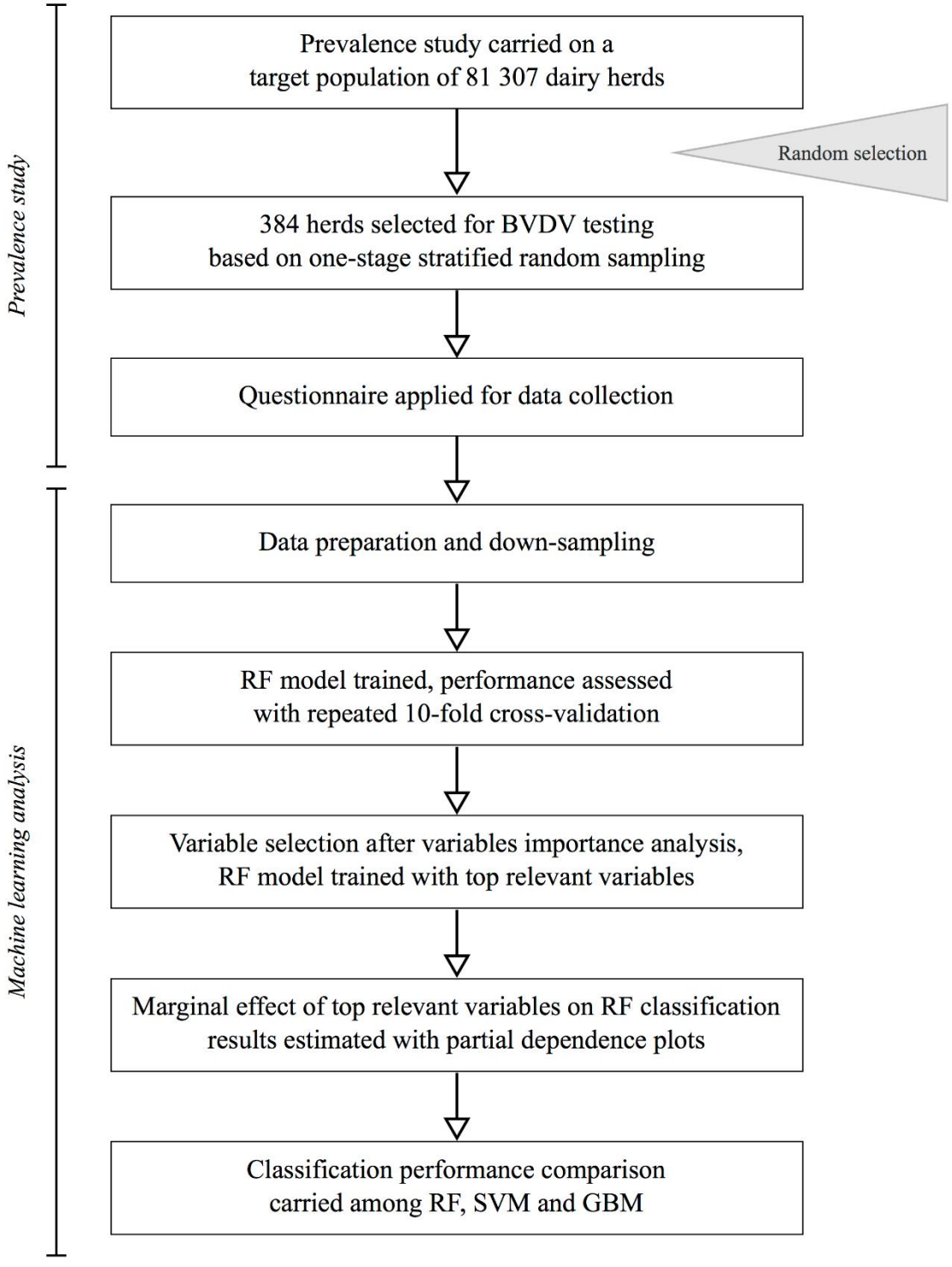

Prevalence study
Prevalence study carried on a target population of 81 307 dairy herds
Random selection
384 herds selected for BVDV testing based on one-stage stratified random sampling
Questionnaire applied for data collection
Machine learning analysis
Data preparation and down-sampling
RF model trained, performance assessed with repeated 10-fold cross-validation
Variable selection after variables importance analysis, RF model trained with top relevant variables
Marginal effect of top relevant variables on RF classification results estimated with partial dependence plots
Classification performance comparison carried among RF, SVM and GBM

Figure 1

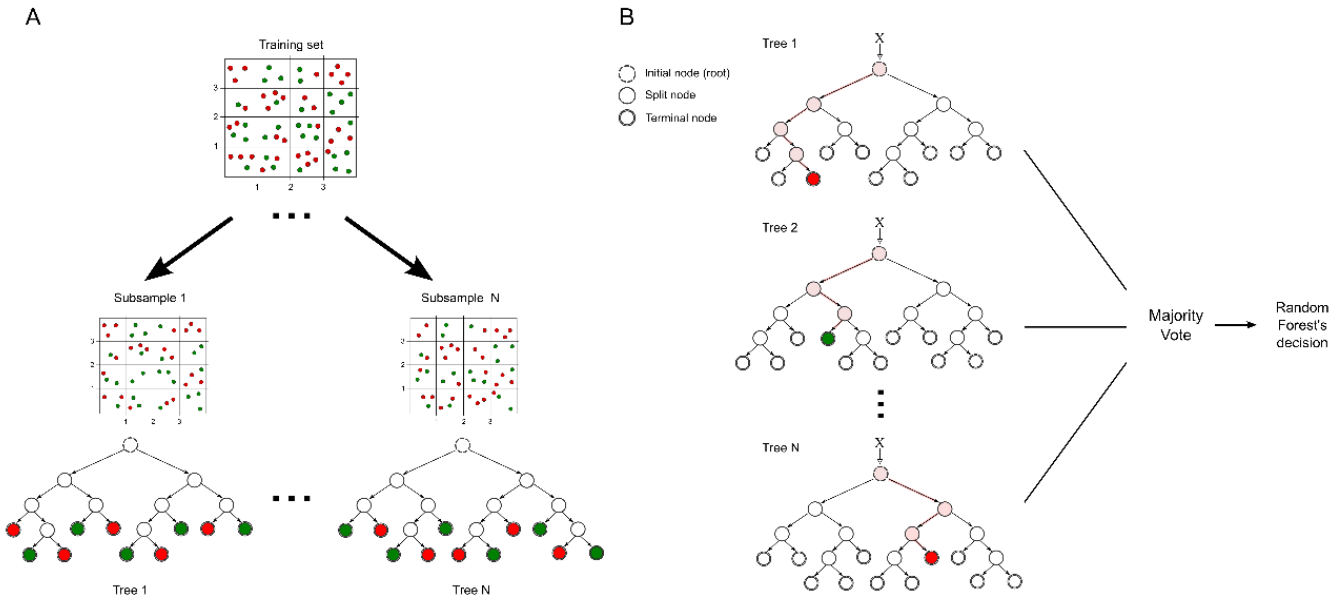

A
Training set
Subsample 1
Subsample N
Tree 1
Tree N
B
Initial node (root)
Split node
Terminal node
Tree 1
Tree 2
Tree N
X
Majority Vote
Random Forest's decision

Figure 2

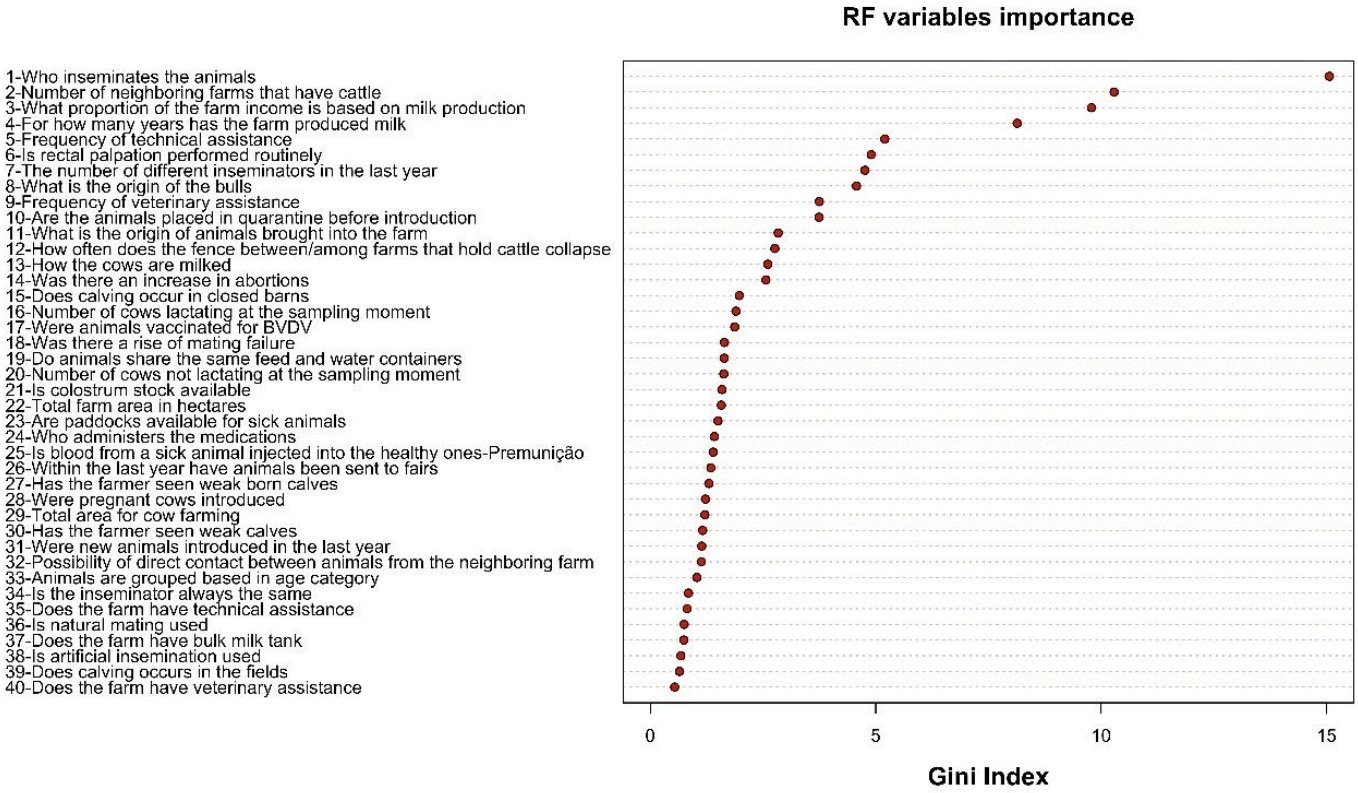

RF variables importance
1-Who inseminates the animals
2-Number of neighboring farms that have cattle
3-What proportion of the farm income is based on milk production
4-For how many years has the farm produced milk
5-Frequency of technical assistance
6-Is rectal palpation performed routinely
7-The number of different inseminators in the last year
8-What is the origin of the bulls
9-Frequency of veterinary assistance
10-Are the animals placed in quarantine before introduction
11-What is the origin of animals brought into the farm
12-How often does the fence between/among farms that hold cattle collapse
13-How the cows are milked
14-Was there an increase in abortions
15-Does calving occur in closed barns
16-Number of cows lactating at the sampling moment
17-Were animals vaccinated for BVDV
18-Was there a rise of mating failure
19-Do animals share the same feed and water containers
20-Number of cows not lactating at the sampling moment
21-Is colostrum stock available
22-Total farm area in hectares
23-Are paddocks available for sick animals
24-Who administers the medications
25-Is blood from a sick animal injected into the healthy ones-Premunição
26-Within the last year have animals been sent to fairs
27-Has the farmer seen weak born calves
28-Were pregnant cows introduced
29-Total area for cow farming
30-Has the farmer seen weak calves
31-Were new animals introduced in the last year
32-Possibility of direct contact between animals from the neighboring farm
33-Animals are grouped based in age category
34-Is the inseminator always the same
35-Does the farm have technical assistance
36-Is natural mating used
37-Does the farm have bulk milk tank
38-Is artificial insemination used
39-Does calving occurs in the fields
40-Does the farm have veterinary assistance
0
5
10
15
Gini Index

Figure 3

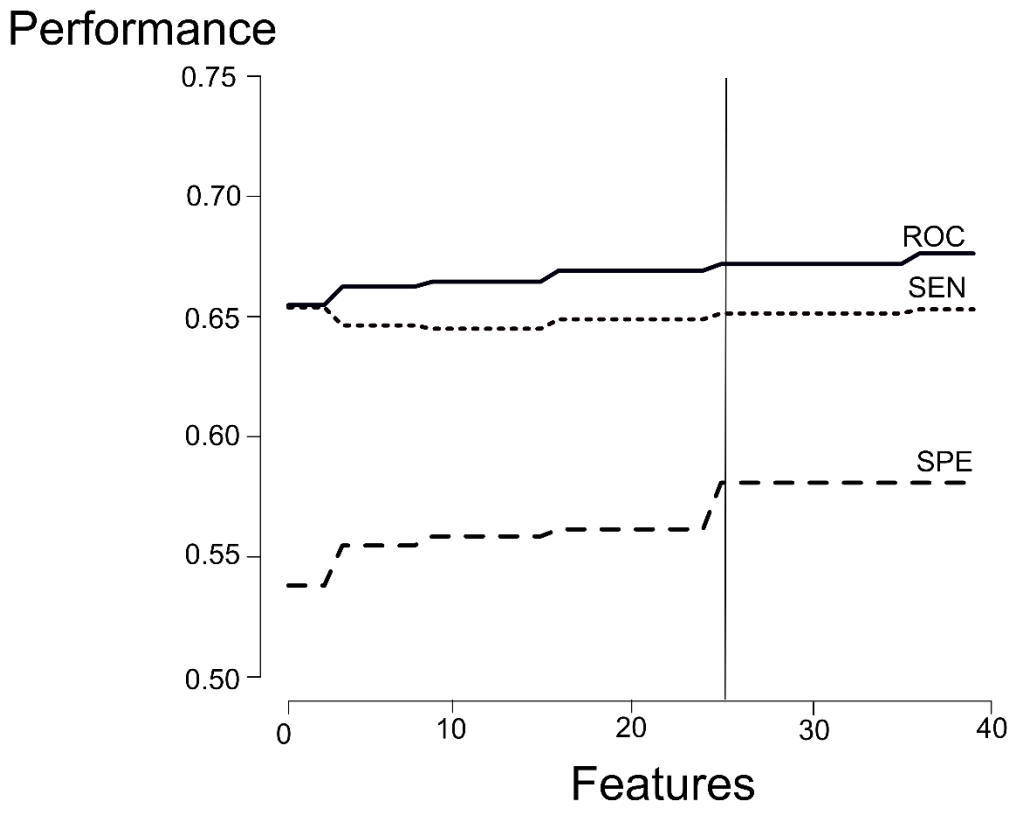

Performance
0.75
0.70
0.65
0.60
0.55
0.50
ROC
SEN
SPE
0
10
20
30
40
Features

Figure 4

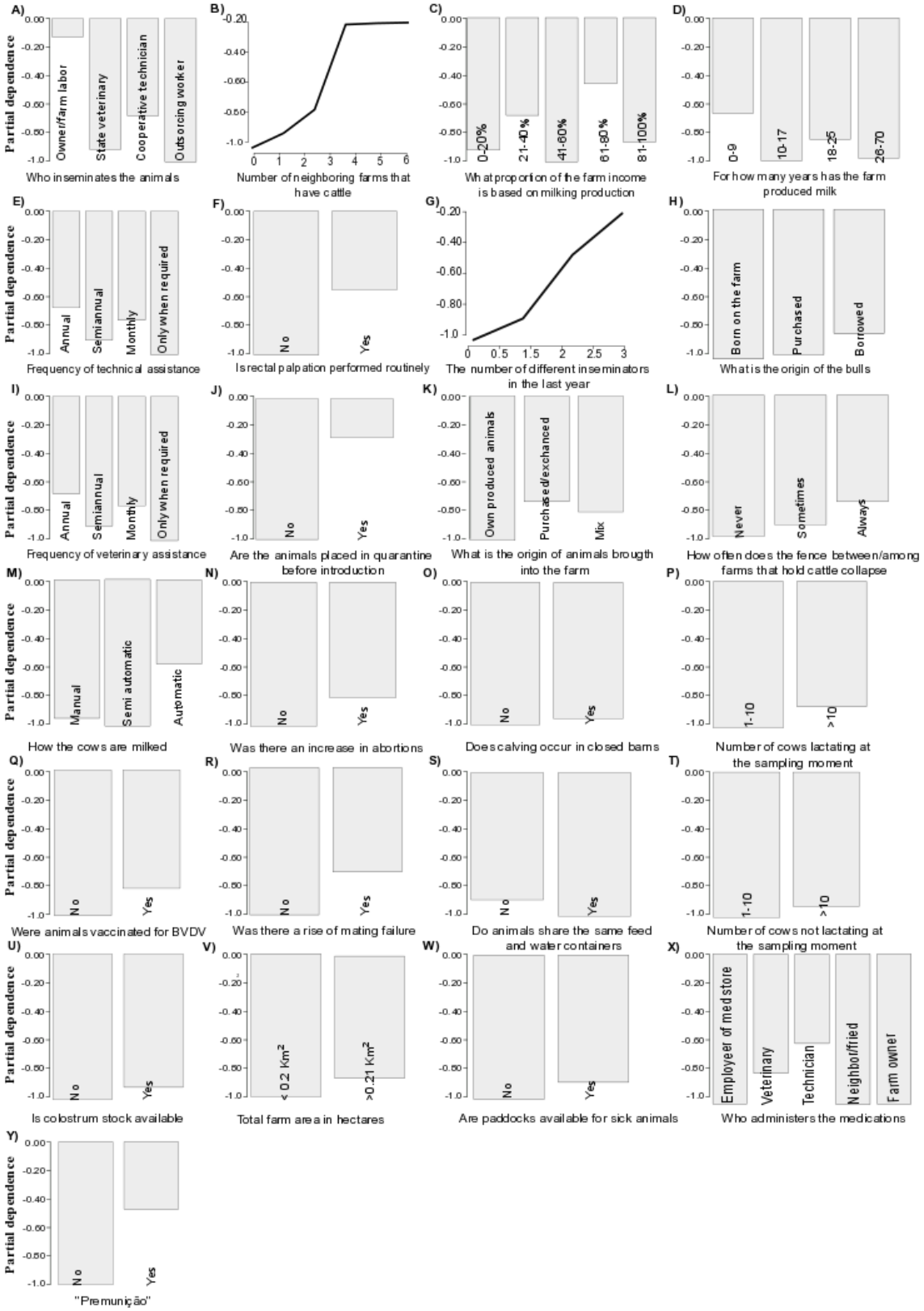

A)
Partial dependence
Owner/farm labor
State veterinary
Cooperative technician
Outsorcing worker
Who inseminates the animals
B)
Number of neighboring farms that have cattle
C)
0-20%
21-40%
41-60%
61-80%
81-100%
What proportion of the farm income is based on milking production
D)
0-9
10-17
18-25
26-70
For how many years has the farm produced milk
E)
Annual
Semiannual
Monthly
Only when required
Frequency of technical assistance
F)
No
Yes
Is rectal palpation performed routinely
G)
The number of different inseminators in the last year
H)
Born on the farm
Purchased
Borrowed
What is the origin of the bulls
I)
Annual
Semiannual
Monthly
Only when required
Frequency of veterinary assistance
J)
No
Yes
Are the animals placed in quarantine before introduction
K)
Own produced animals
Purchased/exchanced
Mix
What is the origin of animals brougth into the farm
L)
Never
Sometimes
Always
How often does the fence between/among farms that hold cattle collapse
M)
Manual
Semi automatic
Automatic
How the cows are milked
N)
No
Yes
Was there an increase in abortions
O)
No
Yes
Does calving occur in closed barns
P)
1-10
>10
Number of cows lactating at the sampling moment
Q)
No
Yes
Were animals vaccinated for BVDV
R)
No
Yes
Was there a rise of mating failure
S)
No
Yes
Do animals share the same feed and water containers
T)
1-10
>10
Number of cows not lactating at the sampling moment
U)
No
Yes
Is colostrum stock available
V)
< 0.2 Km²
>0.21 Km²
Total farm area in hectares
W)
No
Yes
Are paddocks available for sick animals
X)
Employeer of medstore
Veterinary
Technician
Neighbor/friend
Farm owner
Who administers the medications
Y)
No
Yes
"Premunição"

Figure 5

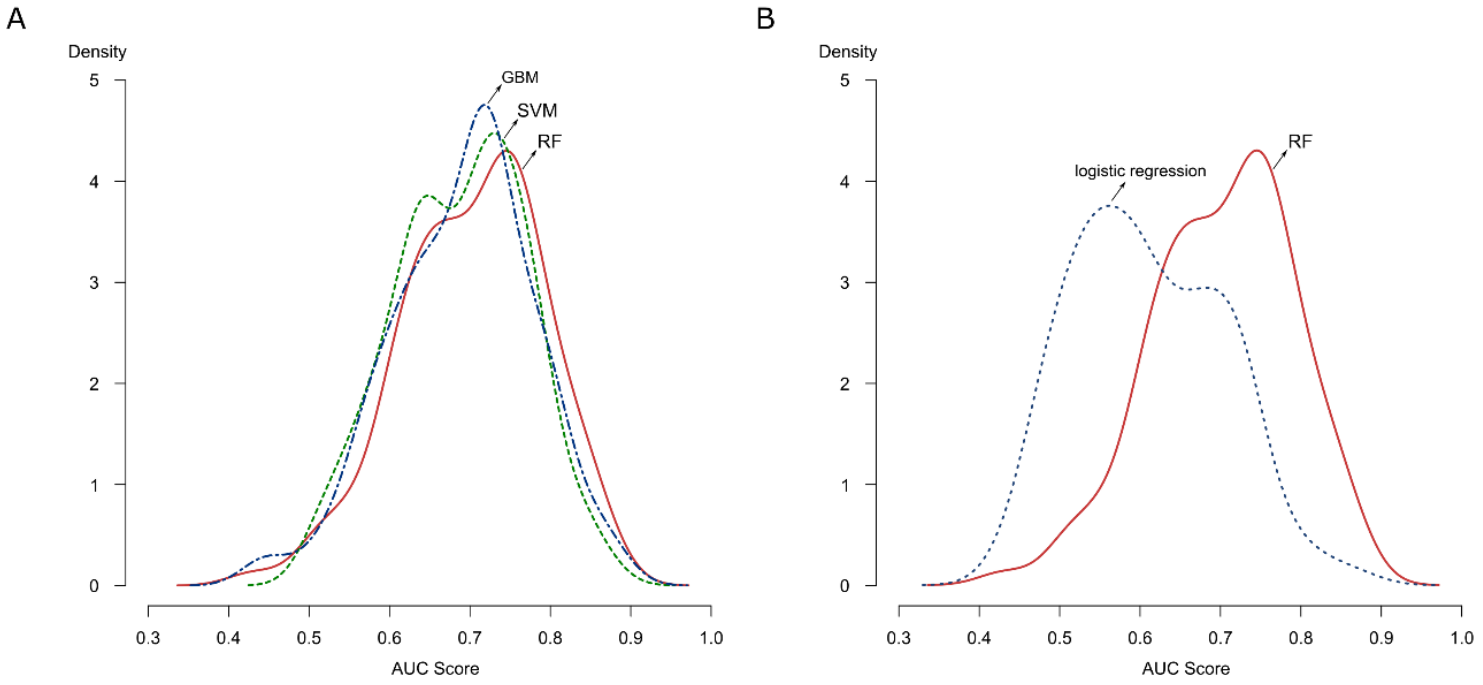

A
Density
5
4
3
2
1
0
GBM
SVM
RF
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
AUC Score
B
Density
5
4
3
2
1
0
logistic regression
RF
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
AUC Score

Figure 6

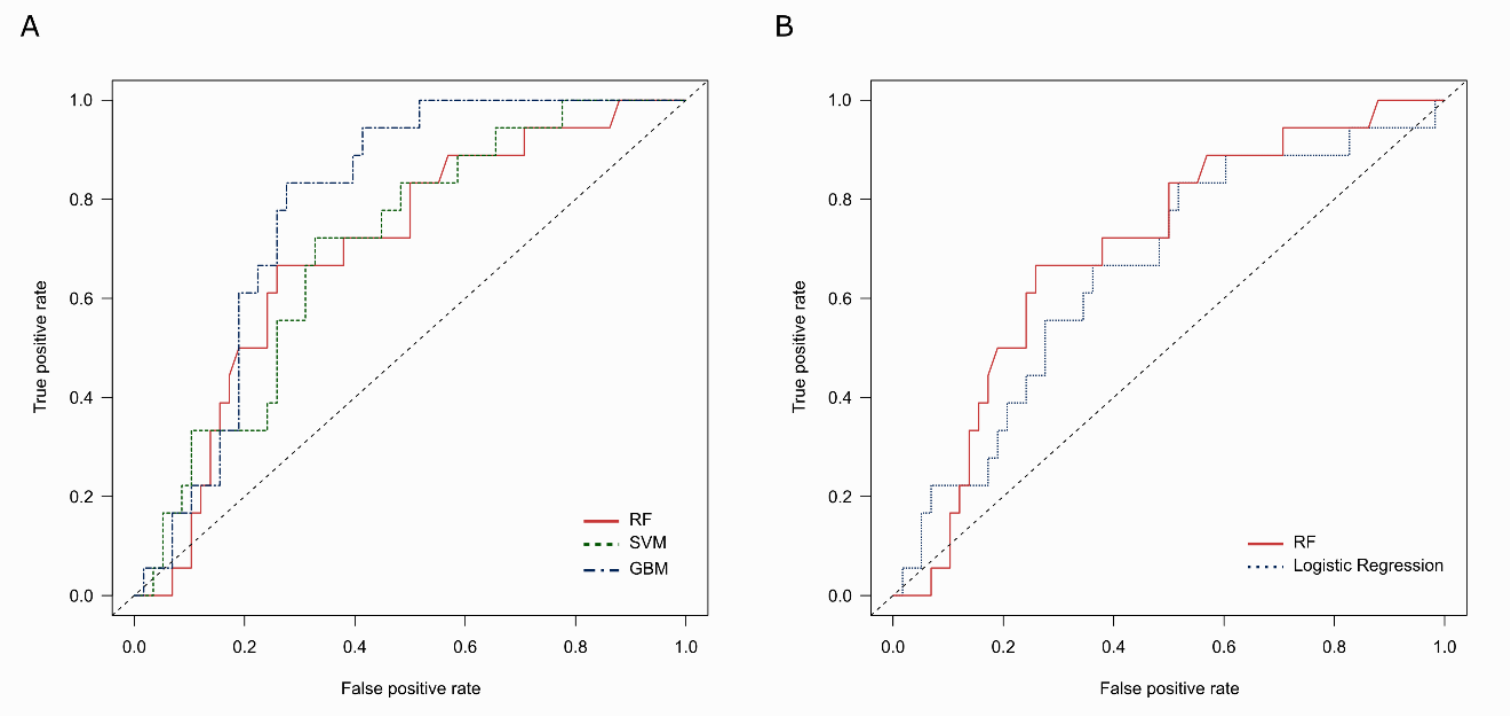

A
B
True positive rate
False positive rate
RF
SVM
GBM
Logistic Regression
0.0
0.2
0.4
0.6
0.8
1.0

Figure 7

Additional File 1. Descriptive statistics on the study population

A questionnaire with both close and open questions (n=40) was applied in 388 farms sampled in this study. We present a general description of the studied farms. The number of people working in the farms was small; with 67.20% reporting that up to two people were involved on dairy activities on daily bases. The average farm size (land) was 33.74 hectares (range: 1-1500). On top of that, 59% of the farm (land) was used for dairy related activity (range: 2-100%). About the experience on dairy production of each farm measured in years, an average of 19.5 years (range: 1-70) was found, and in 25.3% of the families/companies, the milk production represents 41-60% of total farm income. The median number of cows per farm was equal to 3 (range: 1-57), and on the day of the sampling the median of lactating cows was 11.5 (range: 0-130). Moreover, 75% of the farms had up to 21 lactating cows, with median of 3 not lactating cows (Figure 1).

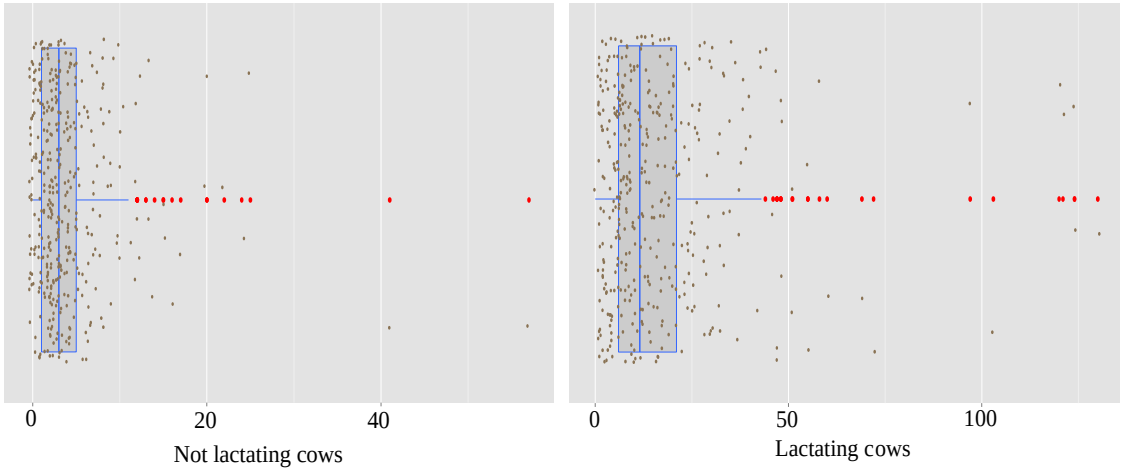


Figure 1. Distribution of count number of cows-red dots represents the outliers' farms.

About the sanitary management, veterinary assistance was reported to be present on 89.7% of the farms and the visits were done predominantly when required by the owner. Most of the owners have participated on agglomeration events, sending animals either for a show, fairs or rodeos. In 53.6% of the farms own only animals born and raised on the own farm, but when an outsourcing animal was introduced, 67.85% reported that the animals were pregnant. Quarantine of any introduced animal was done in 28.4% of the farms and most of them kept such animals apart from the herd for less

than 30 days. The availability of paddocks destined to keep sick animals away from healthy animals was found on 40.7% of the farms.

Related to reproductive practices, 54.5% answered that they adopted natural mating practices and 80.2% reported the use of artificial insemination (AI) and bulls simultaneously. When IA was used, 57% of the farms reported that it was always the same person that provided the service, 18% said to have more than one person inseminating the animals, and 42.4% of the farms, the service was provided by a veterinary company (Figure 2). Farms that routinely used rectal examination for pregnancy diagnostic represent 41.8% of the study population, and the ones that noticed that the responsible for doing the procedure did not changed rubber gloves between/among different animals were 40.6% of the total farms. In 61.2% of the farms, the origin of the bulls was related to an external source, i.e., they were not born in the property. Among the farmers that keep a bull as permanent member of the herd, 37% had bought the animal and 24.2% just borrowed the animal from a nearest neighbor.

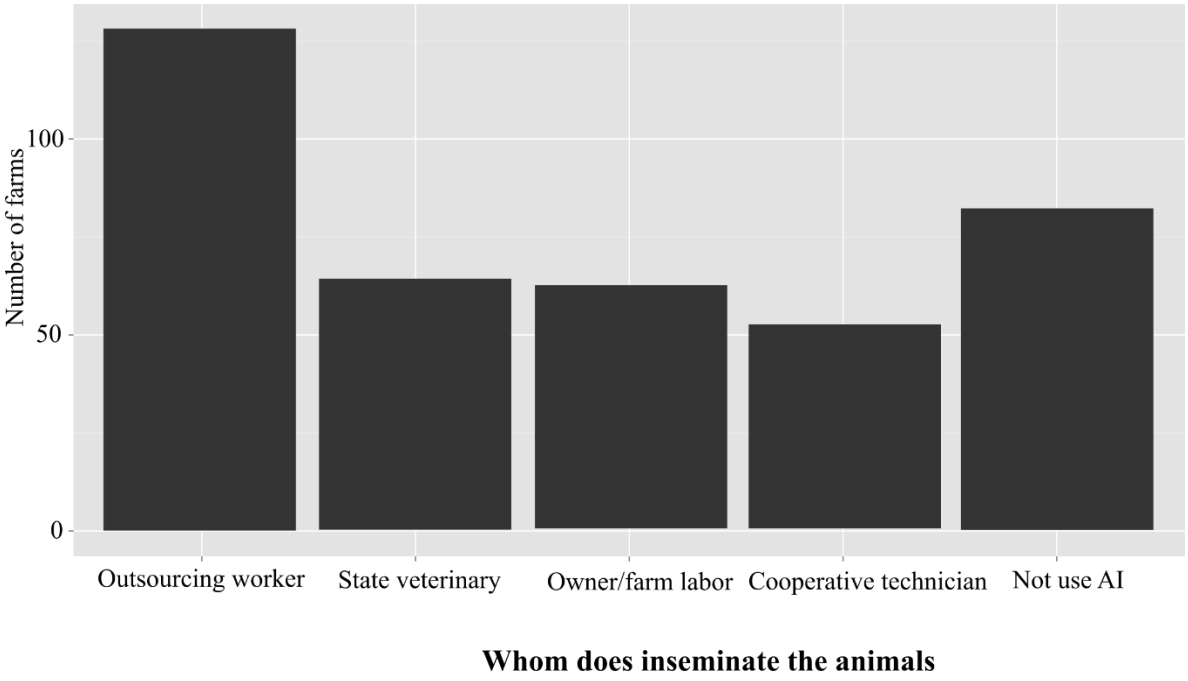


Figure 2. Artificial insemination been performed by different knowledge persons.

About calving, in 90.7% of the answers it occurred out on the fields. Answers related to the management after calving showed that 43% of the farms did not allowed

the calves to suck colostrum and were immediately separated from their mothers, while 39% of the farms were keeping colostrum stock (a pool of milk from many cows to feed on all new calves). The rise of weak newborns and abortion occurrence was observed in 31.4% e 25.5% of the farms, respectively.

The use of the same needle in diffent animals were praticated on 78.5% of the farms, which may be involved on the spread of many diseases including BVDV. The possibility of a direct contact over the fences lines between/among animals from neighboring farms was reported as positive in 55.4% of the answers. On the herd management, 59.4% of owners divided animals by age in different fields for feeding purpose. The collapse of fence and movement of animals' between/among farms was reported to happen sometimes on 42.3% of the farms.

Additional File 2. The frequency or central tendency measurement of important predictor variables

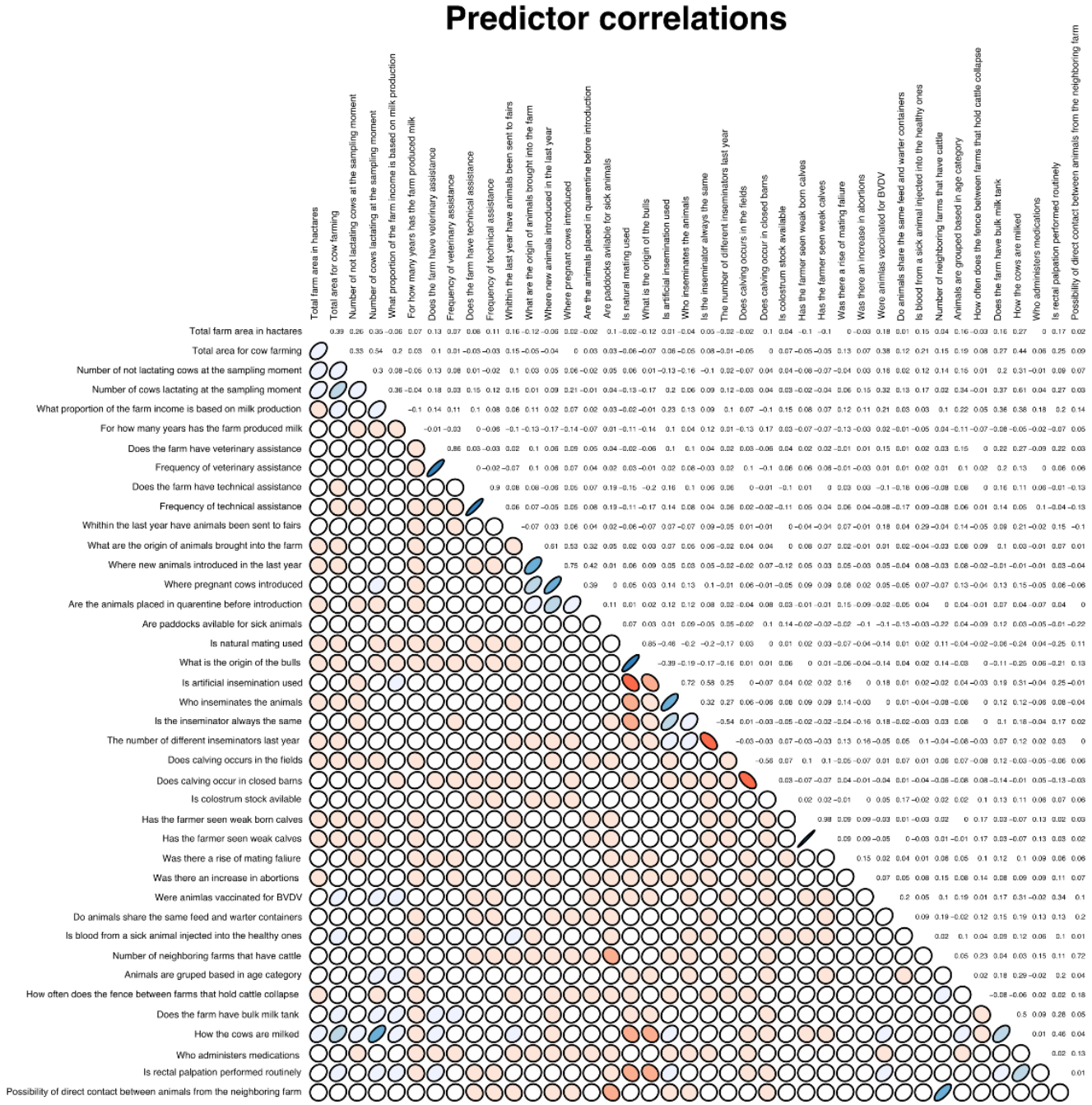
Variables	Frequency (%)or median	
	BVDV (+)	BVDV (-)
(1) Who inseminates the animals		
Owner/farm labor	30 (45.59)	32 (13.58)
State veterinary	8 (11.76)	56 (23.05)
Cooperative technician	10 (14.71)	42 (17.28)
Outsourcing worker	19 (27.94)	112 (46.09)
(2) Number of neighboring farms that have cattle	1	1
(3) What proportion of the farm income is based on milk production		
0-20%	11 (11.96)	64 (21.84)
21-40%	17 (18.48)	45 (15.36)
41-60%	20 (21.74)	78 (26.62)
61-80%	23 (25.00)	38 (12.97)
81-100%	21 (22.83)	68 (23.21)
(4) For how many years has the farm produced milk	17	19
(5) Frequency of technical assistance		
Annual	1 (2.38)	5 (3.09)
Semester	6 (14.29)	19 (11.73)
Monthly	17 (40.48)	54 (33.33)
Only when needed	18 (42.86)	84 (51.85)

(6) Is rectal palpation performed routinely		
No	34 (30.63)	192 (65.08)
Yes	59 (53.15)	103 (34.92)
(7) Number of different inseminators in the last year	1	1
(8) What is the origin of the bulls		
Born in the farm	41 (50.00)	64 (40.25)
Purchased	18 (21.95)	55 (34.59)
Borrowed	23 (28.05)	40 (25.16)
(9) Frequency of veterinary assistance		
Annual	3 (3.53)	1 (0.38)
Semester	2 (2.35)	2 (0.75
Monthly	25 (29.41)	38 (14.29)
Only when requested	55 (64.71)	225 (84.59)
(10) Are animals placed in a quarantine before introduction		
No	25 (25.77)	80 (76.19)
Yes	17 (17.53)	25 (23.81)
(11) What is the origin of animals brought into the farm		
Own produced animals	39 (41.94)	169 (57.29)
Only purchased or exchanged	3 (3.23)	11 (3.73)
Mix (own **and** purchased or exchanged)	51 (54.84)	115 (38.98)
(12) How often does the fence between/among farms that hold cattle collapse		
Never	40 (43.01)	156 (52.88)

Sometimes	44 (47.31)	120 (40.68)
Always	9 (9.68)	19 (6.44)
(13) How the cows are milked		
Manual	9 (9.68)	44 (14.92)
Semi-automatic	41 (44.09)	169 (57.29)
Automatic	43 (46.24)	82 (27.80)
(14) Was there an increase in abortions		
No	65 (69.89)	224 (75.93)
Yes	28 (30.11)	71 (24.07)
(15) Does calving occur in closed barns		
No	76 (81.72)	220 (74.58)
Yes	17 (18.28)	75 (25.42)
(16) Number of cows lactating at the sampling moment	17	10
(17) Were animals vaccinated for BVDV		
No	85 (91.40)	449 (84.98)
Yes	8 (8.90)	44 (15.02)
(18) Was there a rise of mating failure		
No	63 (74)	203 (68.81)
Yes	30 (32.26)	92 (31.19)
(19) Do animals share the same feed and water containers		
No	17 (25.37)	46 (22.66)
Yes	50 (74.63)	157 (77.34)
(20) Number of cows not lactating at the sampling moment	17	10

(21) Is colostrum stock available		
No	55 (59.14)	182 (61.69)
Yes	38 (40.86)	113 (38.31)
(22) Total farm area in hectares		
< 0.2 Km^2	40 (43.01)	162 (54.92)
> 0.21 Km^2	53 (56.99)	133 (45.08)
(23) Are paddocks available for sick animals		
No	56 (60.22)	174 (58.98)
Yes	37 (39.78)	121 (41.02)
(24) Who administers the medications		
Employer of the med store	6 (8.82)	5(2.06)
Veterinary	25 (36.76)	28 (11.52)
Technician	8 (11.76)	56 (23.05)
Neighboring/friend	10 (14.71)	42 (17.28
Farm Owner	19 (27.94)	112 (46.09)
(25) Is blood from a sick animal injected into the healthy ones		
No	70 (76.92)	279 (96.54)
Yes	21 (23.08)	10 (3.46)

Additional File 3. Correlation matrix for independent variables. Negative correlation is represented by red elipses pending to the left; positive correlation is represented by blue elipses pending to the right. The exact correlation values are given in the upper panel.



Additional File 4. Models performance for 10 randomly generated independent test data sets. The AUC scores are computed for 10 repetitions of model training and testing. In each repetition, a random portion of 80% of data is used for training, and the remaining 20% for testing (independent/external data). The best performance in each repetition is shown in boldface. The mean and standard deviations for each classifier after the 10 repetitions are also reported. We observe that RF achieves the best performance in six repetitions, while GBM provides the best model in the remaining four. The AUC scores for the worst and best model trained with RF are 0.6925 and 0.8764, while GBM's worst and best performance are 0.6178 and 0.8180. This corresponds to a performance gain of 12.09% and 7.13% in the worst and best cases, respectively.

	RF	**SVM**	**GBM**
Repetition 1	**0.6925**	0.6388	0.6178
Repetition 2	**0.8170**	0.7480	0.7873
Repetition 3	**0.6944**	0.6800	0.6388
Repetition 4	**0.7198**	0.6522	0.6762
Repetition 5	0.7447	0.6704	**0.7557**
Repetition 6	**0.7303**	0.6954	0.7279
Repetition 7	**0.8764**	0.7576	0.8180
Repetition 8	0.7112	0.7155	**0.7595**
Repetition 9	0.7749	0.6867	**0.7902**
Repetition 10	0.7049	0.6379	**0.7298**
Mean	0.7466	0.6883	0.7301
SD	0.0598	0.0419	0.0666

Additional File 5. Comparison between Random Forest (RF) and Logistic Regression

We assessed the RF performance relative to logistic regression, which is a very conventional statistical approach for the analysis of risk factors. Logistic regression was estimated with the glm() function in R environment, and performance assessment was carried based on 10 repetitions of 10-fold cross-validation using the caret R package. To assure a fair comparison, we run the logistic regression analysis with the same distribution of data used for RF training among folds and across all repetitions of cross-validation.

The average AUC scores for RF and logistic regression based on the repeated 10-fold cross validation were 0.702 (±0.08) and 0.610 (±0.09), respectively. The density plots drawn from the cross-validation procedure makes evident the better performance achieved by RF, given by the shift of RF AUC scores distribution to the right (Figure 1).

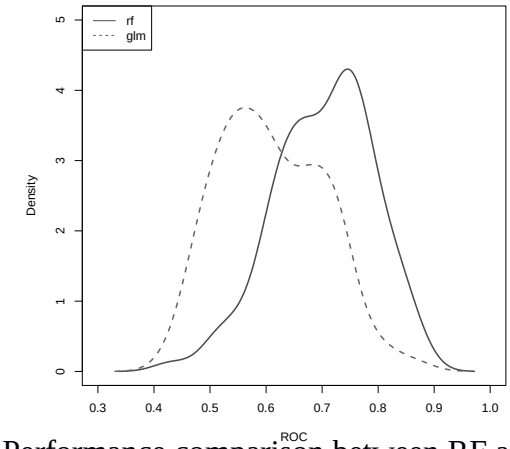


Figure 1. Performance comparison between RF and logistic regression in terms of AUC score distribution.

Moreover, we observed that the classification provided by RF is much more balanced in terms of sensitivity and specificity than logistic regression. The average sensitivity was 0.676 (±0.044) for RF in contrast to 0.613 (±0.047) for logistic regression, and the average specificity was 0.622 (±0.173) for RF and 0.563 (±0.175) for logistic regression, respectively. The density plots for specificity and sensitivity are shown in Figures 2-A and 2-B. In both plots, the curves associated to RF are dislocated to the right of logistic regression curves, meaning a tendency of RF tin producing higher values .

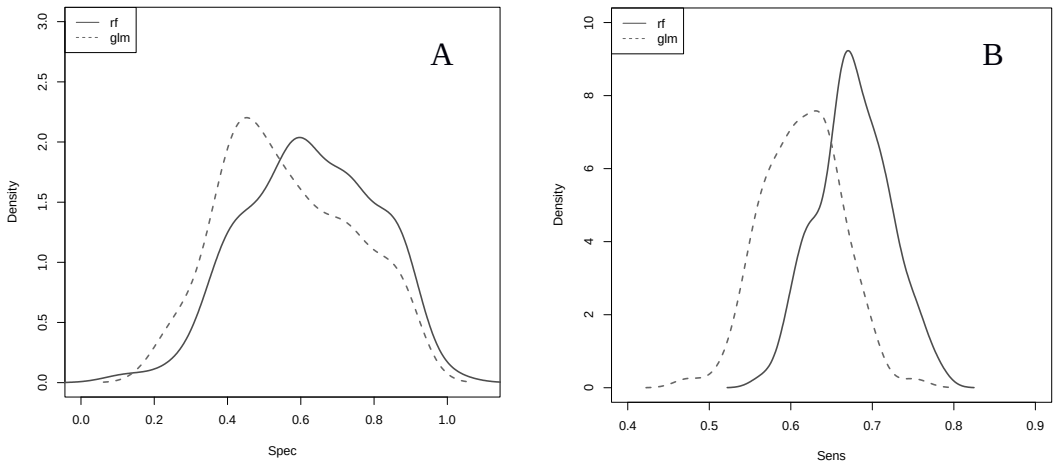


Figure 2. Comparison between RF and logistic regression on specificity and sensibility: a- Specificity of RF better than logistic regression; b- Sensitivity of RF performed better than logistic regression.

Further tests with independent test data, i.e., the 20% portion of data separated for model testing during data preparation, corroborates the previous finding. RF also outperforms the traditional statistical approach in classifying new instances, reaching an AUC score of 0.692 in contrast to 0.657 for logistic regression (Figure 3).

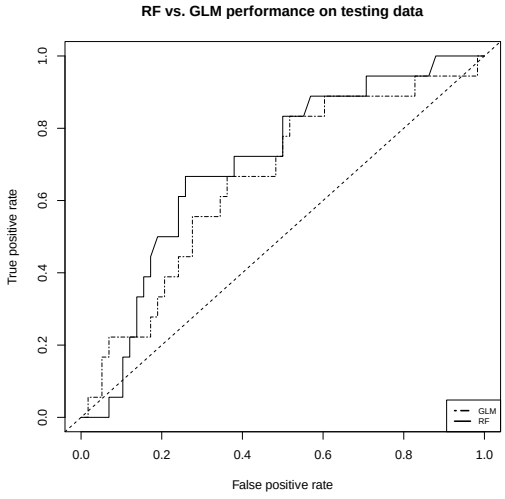


Figure 3. ROC curves for independent test set show that RF also performs better than traditional logistic regression when classifying external data.

4.2. **Capítulo 2: Herd prevalence and associated factors of bovine viral diarrhea virus antibodies in bulk tank milk in southern Brazil**

O presente trabalho já foi concluído e um artigo científico intitulado de *Herd prevalence and associated factors of bovine viral diarrhea virus antibodies in bulk tank milk in southern Brazil* foi redigido no formato do periódico ***Frontier in Veterinary Science,*** e revisado quanto ao idioma inglês pela ***American Journal Experts***.

Herd prevalence and associated factors of bovine viral diarrhea virus antibodies in bulk tank milk in southern Brazil

Gustavo Machado[1*], Fernanda Simone Marks[1], Waldemir Santiago Neto[1], Diego Viali dos Santos[1], Antonio Augusto Medeiros[1,2], André Mendes Ribeiro Corrêa[2], Heber Eduardo Hein[2], Ugo Souza[3], Cláudio Wageck Canal[4], Luciana Neves Nunes[5], Luis Gustavo Corbellini[1*]

[1] Laboratório de Epidemiologia Veterinária (EPILAB), Faculdade de Veterinária, Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves 9090, CEP 91540-000, Porto Alegre, RS, Brazil.

[2] Secretaria da Agricultura e Pecuária (SEAP-RS), Brasil, Av. Getúlio Vargas, 1384, CEP 95150-900, Porto Alegre, RS, Brazil.

[3] Instituto de Pesquisas Veterinárias Desidério Finamor (IPVDF), Fundação Estadual de Pesquisa Agropecuária (FEPAGRO), Eldorado do Sul, RS, Brazil.

[4] Laboratório de Virologia, Faculdade de Veterinária, Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves 9090, CEP 91540-000, Porto Alegre, RS, Brazil.

[5] Departamento de Estatística, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil.

***Correspondence**: Laboratório de Epidemiologia Veterinária (EPILAB), Faculdade de Veterinária, Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves 9090, CEP 91540-000, Porto Alegre, RS, Brazil.
gustavoetal@gmail.com: luis.corbellini@ufrgs.br;

Keywords: BVDV; epidemiology; milk; antibody; associated factors; reproduction.

Abstract
A cross-sectional study was performed to estimate the herd prevalence of antibodies against bovine viral diarrhea virus (BVDV) in bulk tank milk (BTM) and possible associated factors. Samples were randomly selected from a population of 81,307 dairy herds in the state of Rio Grande do Sul, Brazil. The estimated prevalence was 23.9% ($CI_{95\%}$ = 19.8 - 28.1). A Poisson regression analysis resulted in the following associated factors: routine use of rectal examination for pregnancy (Prevalence Ratio [PR] = 2.73); direct contact over fences among cattle of neighboring farms (PR = 1.63); and farms that did not use artificial insemination (PR = 2.07). The results of this study showed that BVDV is present statewide and that control strategies should be designed and applied, particularly those involving better reproductive practice.

Introduction

Bovine viral diarrhea virus (BVDV) is a member of the genus *Pestivirus* of the family *Flaviviridae* (1). BVDV is one of the most common and economically important viruses of cattle (2). The genus *Pestivirus* comprises other critical pathogens such as classical swine fever virus and border disease virus of sheep.

BVDV infections are globally endemic, impairing the dairy industry by reduced milk production and reproductive performance, delay of growth, increased susceptibility to other diseases and causing indirect market-related issues (3,4,5). The maintenance of BVDV within cattle herds and its transmission to susceptible hosts commonly occurs as a result of exposure to persistently infected (PI) cattle that harbor and shed the virus throughout their life (6). After infection occurs in an immunocompetent host, a neutralizing antibody response follows that can last for many years and can be detected in sera or milk (7).

Many BVDV control strategies have been proposed in different countries, and these were based on information about prevalence and/or incidence, which serves as baseline knowledge for designing and implementing effective regional or wider control programs (8). Among the benefits of estimating herd prevalence are monitoring the progress of the infection, gaining insights for better decision making, and use in developing health-control measures (9). Currently, some countries have been working on BVDV eradication programs (10,11), and others have successfully eradicated the disease (12,13). To estimate the prevalence of BVDV antibodies, enzyme-linked immunosorbent assays (ELISA) are the most frequently used diagnostic technique for serum and/or milk samples (14). Detection of antibodies from bulk tank milk (BTM) is considered an inexpensive and reliable alternative for monitoring the disease, and the results, in turn, can be applied to control strategies within dairy herds because the test can provide information about the status of a large group of animals or individual milk samples (14,15).

Many studies have estimated dairy and beef herd prevalence worldwide (16,15,17), and BVDV is known to spread within Brazilian dairy herds (18,19,20,21). There is a lack of epidemiological studies that include the entire dairy population of such a large area as the entire state of Rio Grande do Sul in southern Brazil, and the previous studies differ greatly in their herd prevalence results, which vary from 24.3% (21) to 82.35% (24).

We aimed to estimate the herd prevalence and associated factors for BVDV by focusing on the reproductive management and practices of an important dairy milk region to corroborate earlier findings obtained from studies made in a small dairy population in the same region (20,21).

Material and Methods

Study area and target population

Rio Grande do Sul is the southernmost state of Brazil (Fig. 1), has a total area of 268,781.896 km² and 497 municipalities, and shares borders with two countries:

Argentina and Uruguay (Fig. 1). The cattle population is approximately 13.5 million, 10% of which are dairy cattle (25,26). More information about the studied area and herd details are available (27).
The target population included all dairy herds in the state of Rio Grande do Sul. According to the official data from the Office of Agriculture, Livestock and Agribusiness of the State of Rio Grande do Sul (SEAP-RS), 81,307 dairy herds were registered in 2013.

Survey design and sample collection

A cross-sectional survey was performed to estimate the BVDV prevalence in dairy herds based on BTM samples and to identify the associated factors. Considering the following parameters: total dairy herds 81.307,00 registered at the moment, 50% expected prevalence, 95% confidence interval, and 5% of absolute precision, the minimum sample size required was 384 dairy herds; however, 388 herds were collected proportional to the stratum (n=7 regions, **Table 1**), in order to facilitate the logistic for sample collection. More details about the sample design and bulk tank milk collection can be found elsewhere (28).

Serological assay and interpretation

The SVANOVIR BVDV p80-AB blocking BVDV ELISA was used to detect the BTM samples positive for anti-BVDV antibodies. All milk samples were centrifuged for 15 minutes at 2000x*g*, according to the manufacturer's instructions. The absorbance at a single wavelength of 450 nm (A_{450}) was determined using a spectrophotometer (Asys Expert Plus, Asys Hitech GmbH, Austria). For the herd prevalence, the percent of inhibition (PI) values were calculated in the same manner as the positive control, as well as for each sample, using the following formula:

$$PI = \frac{OD_{Negative\ control} - OD_{Sample\ or\ Positive\ control}}{OD_{Negative\ control}} x100 \quad (1)$$

Results were interpreted according to the manufacturer´s scheme and were considered positive if ≥ 30% for the identification of herds with a high probability to harboring an active infection and/or at least one positive cow was contributing to the sample and were interpreted as negative if the herd PI was < 30%.

Questionnaire and interview

A questionnaire was designed to gather information about the potential factors associated with BVDV transmission and/or its maintenance within a herd. It was applied during visits to the 388 selected herds in November 2013. The questionnaire was developed in consultation with experts' knowledge of BVDV and based on previous studies. The structured questionnaire had 40 "close-ended" questions grouped into five main categories: general farm characteristics, biosecurity (reproductive management), farm sanitary conditions, and general management and farm facility structure. It was previously tested with five nonparticipating farmers to identify potential sources of

misinterpretation and to further refine the questions. Seven graduate students were trained to perform the interviews. Each personal interview lasted from 20 to 40 min. The interviews were performed face-to-face, and a copy of the questionnaire is available from the corresponding author upon request.

Data management

Epi Info 7 (CDC) was used to enter data and to track information quality. For every hundred questionnaires and laboratory results entered in the databases, 5% were randomly sampled and double-checked to verify the data quality. The spatial location (GIS) of each sampled farm in the survey was determined with a handheld global positioning system (GPS) and then plotted on a map using GIS software ArcView 10 (ESRI, Redlands, CA, USA).

Statistical analysis

For the estimation of herd prevalence, those herds with a percentage of inhibition value ≥ 30% were considered positive for BVDV. The overall prevalence and the 95% confidence interval (CI) were calculated taking sampling weight into consideration. Weights were computed as the inverse of the probability of a farm being sampled within each of the seven regions (29).

Univariable analysis

All variables collected by the questionnaire were tested for frequency distribution; continuous variables were tested by histogram, mean, standard deviation and range. The whole statistical process was carried out with R-language, v.3.1.1 (R Development Core Team, 2012). Variables with large amounts of missing data (> 10%) and limited variability (< 20%) were not included in the univariable analysis (n = 4). A univariable analysis was initially used to examine the association between positivity in BTM samples and the remaining 36 independent variables. A prevalence ratio (PR) with robust variance was applied to assess the impact of individual factors on outcomes (30). A multiple Poisson regression was used to estimate the prevalence ratio and 95% confidence interval (CI) of the estimates (31). A Poisson approach was choose, as recommended (32). Subsequently, all variables with $P \leq 0.20$ were select for inclusion in the analysis. Variance inflation factors (VIF) were estimate to verify the relations among all selected independent variables to check for potential collinearity (by a multivariable approach), in which a coefficient > 2.50 was considered high; when a high VIF was found, a variable with a lower *P*-value was considered for the model.

Multivariable analysis

Multivariable models were built in a manual forward method; each remaining variable was selected by the Akaike Information Criterion (AIC) and added to the best previous model. A backwards elimination step was use, resulting in a final model in which only variables with a $P \leq 0.05$ were retained. Confounding effects were investigated by

checking changes in the point estimates of the variables that remained in the model, based on conceptual criteria. Changes in parameter estimate > 25% were considered to be a confounder and were retained in the model until the final model; finally, two-way interaction terms between variables with biological plausibility were investigated. We used deviance performance as a goodness of fit test for the overall model.

The current study was approved by the Animal Welfare Committee of the Universidade Federal do Rio Grande do Sul, Brazil, number: 20711

Results

There were 93 BTM BVDV antibody-positive herds among the 388 herds sampled, and the prevalence estimated was 23.9% ($CI_{95\%}$ = 19.8 - 28.1) with a design effect of 0.94. The frequency of positive samples by each region can be observed in **Table 1**.

Two independent variables showed a VIF > 2.50: presence of technical assistance on the farm and the frequency of technical assistance. The presence of technical assistance was kept in the model because it had the lowest *P-value*. In the univariable analysis, 17 presented a *P-value* ≤ 0.20 (**Table 2**) for the presence of BVDV antibodies in BTM. The final model identified three variables as significantly associated with BVDV ($P < 0.05$) (**Table 3**): the routine use of rectal examination for pregnancy diagnosis (PR = 2.73; $IC_{95\%}$ = 1.74 – 4.29; $P < 0.01$); direct contact over fences among cattle from neighboring farms (PR = 1.63; $IC_{95\%}$ = 1.05 – 2.53; $P = 0.02$); and farms that did not use AI (PR = 2.07; $IC_{95\%}$ = 1.28 – 3.35; $P = 0.002$). None of the two-way interaction terms were significant at 5%, and the total farm area was the only confounding effect forced into the model because it changed the parameter estimates by more than 25%. The model goodness-of-fit was tested by a deviance chi-squared test and was found not to be significant ($P > 0.05$).

Discussion

The aim of this study was to estimate the herd seroprevalence of BVDV in the state of Rio Grande do Sul, which was found to be 23.9%, and to determine which variables would be classified as associated factors to BVDV seropositivity. We found an important association between the presence of antibodies and rectal examinations being performed on the cows. We also found one of the same associated factor that was previously identified (direct contact over fences) (21). The current study was performed because we believed that reproductive practices were the main factors involved in BVDV spread in the target population and that the veterinary office and the milk industry needed robust information about the situation of BVDV in dairy herds. This study is also complementary to other study (28) that our group has published, and what we found is that no matter the statistics method used, the findings are very similar, mainly about the most relevant associated factors that may be involved on BVDV transmission and spread within the studied area.

The estimated prevalence was lower than our recent study made in a smaller area, perhaps because the p80 protein ELISA used in this study that is focused on the detection of antibodies directed against a nonstructural viral protein, which is produced by actively reproducing virus (21). The prevalence based on the BTM samples showed that BVDV is present across the studied region and that the major milk production area (Northwest, **Figure 1**) has a percentage of positive herds similar to that of the whole state. However, the regions with no tradition in dairy cattle (27) showed the highest percentage of positive herds (Southwest and Western central) (Table 1, **Figure 1**). One study conducted in Brazil estimated the prevalence of positive milk in individual cows as varying from 5.26 to 70.83% among the different farms (22), and we found a prevalence of 48.8% and 24.3% (20,21) in two smaller areas of Rio Grande do Sul. In Finland, a very low prevalence of positive herds was found, which was most likely related to the low cattle and herd density in that country (33). Another study that used ELISA for BTM revealed a high level of exposure to BVDV (73%) and a lower proportion (13%) of herds with high BVDV antibody titers (34). Studies on herd prevalence were conducted in countries that share borders with Brazil, i.e., Argentina and Uruguay, and the studies found that 93.1% and 100% of the herds were positive, respectively (35,36). More recently, a publication showed that UK dairy herds have a prevalence of 58.6% (37).

The Poisson regression model reflects associated factors with BVDV infection detected by the presence of antibodies in BTM samples. The explanatory variables identified in the final model as being associated with BVDV seropositivity included the routine practice of rectal examination for pregnancy verification. Previous studies have suggested that gloves used during examination may be an important route of horizontal transmission of BVDV, and veterinarians performing rectal palpation on consecutive cows may do so without replacing gloves, occasionally simply wiping the gloves with some flannel. Nevertheless, we also asked the information about reusing rectal gloves, but what we found was that this variable was not significantly associated with BVDV (P=0.42). Perhaps, rectal palpation is a confounder for a variable not measured or the question performed to get information about replacing gloves is socially acceptable and more prone to false negative answers about replacing gloves between cows. We performed a different approach for the same data set, thought the Machine Learning and identify that rectal examination was also one of the predictor for BVDV (28).

The indirect transmission of BVDV can occur through the use of veterinary equipment such as nose tongs, needles and the protective rubber gloves worn during rectal examination (38, 39). Also (40) suggested that rectal palpation between different animals may play an important role in the transmission of BVDV; therefore, the veterinarians and technicians who are responsible for reducing disease spread by applying biosecurity measures are not doing so, and this practice will continue to make BVDV a dangerous risk.

Direct contact over fences among animals of neighboring farms was also considered an associated factor for BVDV. The most common route of BVDV transmission is known to be direct contact between animals, and this risk should be closely evaluated (41, 10, 15, 42). Direct contact has also been identified as the most important factor in a case control study (OR = 2.3; $CI_{95\%}$ = 1.27 – 4.24; $P < 0.05$) (43), which is in accordance with

previously findings (21); the direct contact was also identified as the second most important predictor for BVDV by Machine Learning approach using the same data set (28). Care should be taken with PI animals because they play a substantially larger role in BVDV transmission than do transiently infected cattle (44). There is a serious risk of spreading BVDV to a BVDV-free herd by over-the-fence pasture contact with an infected herd; thus, BVDV will continue to circulate, and the costs in terms of biosecurity and herd breakdown will continue to fall on those who have worked to control BVDV (45). Therefore, the prevention of contact between neighboring herds along fence lines decreases the risk of herd infection (46). Furthermore, if exposure to other herds sharing fence lines or communal pasture continues, removing the source of the infection inside the herd (culling PI animals) may not solve the risk of infections (47).

Finally, farms that did not use AI (i.e., only bulls) were positively associated with BVDV seropositivity. It has been proven that BVDV can be horizontally transmitted by natural mating and AI (48). Some bulls may not exhibit obvious symptoms when carrying the virus; however, they may be transiently or persistently infected or may have prolonged testicular infection or persistent testicular infection (PTI) (49). A single seropositive, nonviremic bull was retrospectively identified in an earlier study as shedding virus in the semen (50); even more problematic are those PI bulls that continuously shed large amounts of virus in multiple body secretions, including semen (49). We have previously identified natural mating as associated factors for BVDV in a restricted area of the state (21), and exposure to this risk shows a nine times greater chance of being BVDV-infected.

Other studies compared farms that used only AI for reproduction with farms that used only natural mating and found increased odds for BVDV infections (1.90), which were clearly due to the use of infected bulls (24). More recently, the presence of bulls on the farm have been found to be a significant influence on the number of positive samples for BVDV in the UK (37).

Conclusions

In summary, we show that BVDV on dairy herds in Rio Grande do Sul is present across the state. This study has provided crucial information about BVDV, which is particularly important for the veterinary office because the information can be used to develop a state control program. This information is also important to the milk industry, which can take this opportunity to work with farmers to apply sanitary measures that will improve the health condition of the herds. Important associated factors, such as rectal examinations, have been identified and highlighted the role of veterinarians in applying adequate reproductive practices. Farms that did not use AI should be advised by veterinarians and technicians about the risks of using infected bulls. Finally, the results provide important epidemiological contributions to the factors that have historically been believed to be associated with BVDV-infected herds, and future research studies should address other reproductive practices in more detail.

Acknowledgements

Financial support was provided by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq/Brazil). Part of this study was supported by FUNDESA.

Conflict of interest

The authors certify that there is no conflict of interest with any researcher or organization regarding the material used in this manuscript.

References

1. Simmonds P, Becher P, Collet MS, Gould EA, Heinz FX, Meyers G, Monath T, Pletnev A, Rice CM, Stiansny K, Thiel HJ, Weiner A, Bukhet J. Family *Flaviviridae*. In: King, A.M.Q., M.J. Adams, E.B. Carstens, E.J. Lefkowitz (eds), Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses, 9th edn. pp. 1003–1020. (2011) Elsevier Academic Press, United States.
2. Houe H. Epidemiological features and economical importance of bovine virus diarrhoea virus (BVDV) infections. *Vet Microbiol* (1999) 64:89-107. doi:10.1016/S0378-1135(98)00262-4
3. Baker JC. The clinical manifestations of bovine viral diarrhea infection. *Vet Clin N Am-food A* (1995) 11:425-445.
4. Houe H. Economic impact of BVDV infection in dairies. *Biologicals* (2003) 31:137-143. doi:10.1016/S1045-1056(03)00030-7
5. Ståhl K, Alenius S. BVDV control and eradication in Europe—an update. *Jpn J Vet Res* (2012) S31-39.
6. Brownlie J, Clarke MC, Howard CJ, Pocock DH. Pathogenesis and epidemiology of bovine virus diarrhoea virus infection of cattle. *An Reche Vet* (1987) 18:157-166.
7. Houe H. Epidemiology of bovine viral diarrhea virus. *Vet Clin N Am-food A* (1995) 11:521-547.
8. Niza-Ribeiro J, Pereira A, Souza J, Madeira H, Barbosa A, Afonso C. Estimated BVDV-prevalence, -contact and -vaccine use in dairy herds in Northern Portugal. *Prev Vet Med* (2005) 72:81-85.
9. Humphry RW, Brülisauer F, McKendrick IJ, Nettleton PF, Gunn GJ. Prevalence of antibodies to bovine viral diarrhoea virus in bulk tank milk and associated risk factors in Scottish dairy herds. *Vet Rec* (2012) 171:445. doi:10.1136/vr.100542
10. Sandvik T. Progress of control and prevention programs for bovine viral diarrhea virus in Europe. *Vet Clin N Am-food A* (2004) 20:151-169. doi:10.1016/j.cvfa.2003.12.004
11. Ridpath JF. Bovine viral diarrhea virus: global status. *Vet Clin N Am-food A* (2010) 26:105-121. doi:10.1016/j.cvfa.2009.10.007
12. Houe H, Lindberg A, Moennig V. Test strategies in bovine viral diarrhea virus control and eradication campaigns in Europe. *J Vet Diagn Invest* (2006) 18:427-436. doi: 10.1177/104063870601800501

13. Presi P, Struchen R, Knight-Jones T, Scholl S, Heim D. Bovine viral diarrhea (BVD) eradication in Switzerland--experiences of the first two years. *Prev Vet Med* (2011) 99:112-121. doi:10.1016/j.prevetmed.2011.01.012
14. Beaudeau F, Belloc C, Seegers H, Assié S, Sellal E, Joly A. Evaluation of a blocking ELISA for the detection of bovine viral diarrhoea virus (BVDV) antibodies in serum and milk. *Vet Microbiol* (2001) 80:329-337. doi:10.1016/S0378-1135(01)00322-4
15. Ståhl K, Rivera H, Vågsholm I, Moreno-López J. Bulk milk testing for antibody seroprevalences to BVDV and BHV-1 in a rural region of Peru. *Prev Vet Med* (2002) 56:193-202. doi:10.1016/S0167-5877(02)00161-7
16. Paton DJ, Christiansen KH, Alenius S, Cranwell MP, Pritchard GC, Drew TW. Prevalence of antibodies to bovine virus diarrhoea virus and other viruses in bulk tank milk in England and Wales. *Vet Rec* (1998) 142:385-391.
17. Brülisauer F, Lewis FI, Ganser AG, McKendrick IJ, Gunn GJ. The prevalence of bovine viral diarrhoea virus infection in beef suckler herds in Scotland. *Vet J* (2010) 186:226-231. doi:10.1016/j.tvjl.2009.08.011
18. Canal CW, Strasser M, Hertig C, Masuda A, Peterhans E.. Detection of antibodies to bovine viral diarrhoea virus (BVDV) and characterization of genomes of BVDV from Brazil. *Vet Microbiol* (1998) 63:85-97. doi:10.1016/S0378-1135(98)00232-6
19. Chaves NP, Bezerra DC, Sousa VE, Santos HP, Pereira HM. Frequency of antibodies and risk factors of bovine viral diarrhea virus infection in non-vaccinated dairy cows in the Maranhense Amazon region, Brazil. *Ciên Rur* (2010) 40:1448-1451. doi.org/10.1590/S0103-84782010005000089
20. Almeida LL, Miranda ICS, Hein HE, Santiago Neto W, Costa EF, Marks FS, Rodenbusch CR, Canal CW, Corbellini LG. Herd-level risk factors for bovine viral diarrhea virus infection in dairy herds from Southern Brazil. *Res Vet Sci* (2013) 95 (3): 901-907. doi: 10.1016/j.rvsc.2013.08.009
21. Machado G, Egocheaga RM, Hein HE, Miranda IC, Neto WS, Almeinda LL, Stein MC, Corbellini LG. Bovine Viral Diarrhoea Virus (BVDV) in Dairy Cattle: A Matched Case-Control Study. Transbound Emerg Dis (2014) Ahead of print. doi: 10.1111/tbed.12219
22. Dias FC, Samara SI. Detection of antibodies to the bovine viral diarrhoea virus in serum, in individual milk and in bulk tank milk from unvaccinated herds. *Braz J Vet Res A Scie* (2003) 40:161-168. doi.org/10.1590/S1413-95962003000300001
23. Flores EF, Weiblen R, Vogel FSF, Roehe PM, Alfieri AA, Pituco EM. A infecção pelo vírus da Diarréia Viral Bovina (BVDV) - histórico, situação atual e perspectivas. *Pesq Vet Bras* (2005) 25:125-134. doi.org/10.1590/S0100-736X2005000300002
24. Quincozes CG, Fischer G, de Oliveira Hübner S. Prevalence and factors associated with bovine viral diarrhea virus infection in South of Rio Grande do Sul. *Semina-Ciên Agrá* (2007) 28:269-275.
25. Instituto Brasileiro de Geografia e Estatística (IBGE), 2010. Pesquisa pecuária municipal, efetivo dos rebanhos por tipo de rebanho, Brazil. Available at http://www.sidra.ibge.gov.br (accessed January 2014).

26. Zoccal R, Assis AG, Evangelista SRM. Distribuição geográfica da pecuária leiteira no Brasil. Embrapa Gado de Leite, Juiz de Fora. (2006) http://cnpgl.embrapa.br/nova/publicacoes/circular/CT88.pdf (accessed January 2015).
27. Silva GS, Costa E, Bernardo FA, Groff FHS, Todeschini B, Santos DV, Machado G. Cattle rearing in Rio Grande do Sul, Brazil. *Acta Scien Vet* (2014) 42:2015
28. Machado G, Mendoza MR, Corbellini, LG. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. *Vet Research* (2015) 46:85.
29. Dohoo I, Martin W, Stryhn H. Veterinary Epidemiology Research. 2nd edition. Atlantic Veterinary College inc. Prince Edward Island (2010) 865p.
30. Medronho RA, Bloch KV, Raggio R; Werneck GL. Epidemiologia. 2nd ed. São Paulo: Atheneu. (2009) 790p.
31. Deddens JA; Petersen MR. Approaches for estimating prevalence ratios. *Occup Envir Med* (2008) 65:501–506. doi:10.1136/oem.2007.034777
32. Barros AJD, Hirakata V. Alternatives for regression in cross-seccional studies: an empirical comparision of models that directly estimate the prevalence ratio. *BMC Med Res Meth* (2003) 3:21. doi:10.1186/1471-2288-3-21
33. Niskanen R. Relationship between the levels of antibodies to bovine viral diarrhoea virus in bulk tank milk and the prevalence of cows exposed to the virus. *Vet Rec* (1993) 133:341-344.
34. Kampa J, Ståhl K, Moreno-López J, Chanlun A, Aiumlamai S, Alenius S. BVDV and BHV-1 infections in dairy herds in northern and northeastern Thailand. *Acta Vet Scand* (2004) 45:181-192. doi: 10.1186/1751-0147-45-181
35. Carbonero A, Maldonado A, Perea A, García-Bocanegra I, Borge C, Torralbo A, Arenas-Montes A, Arenas-Casas A. Factores de riesgo del síndrome respiratorio bovino en terneros lactantes de Argentina. *Arch Zoot* (2011) 60:41-51. DOI: 10.4321/S0004-05922011000100005
36. Guarino H, Núñez A, Repiso MV, Gil A, Dargatz DA. Prevalence of serum antibodies to bovine herpesvirus-1 and bovine viral diarrhea virus in beef cattle in Uruguay. *Prev Vet Med* (2008) 85:4-40. doi:10.1016/j.prevetmed.2007.12.012
37. Williams D, Winden SV. Risk factors associated with high bulk milk antibody levels to common pathogens in UK dairies. Vet Rec (2014) 174(23). doi:10.1136/vr.102049
38. Gunn HM. Role of fomites and flies in the transmission of bovine viral diarrhoea virus. *Vet Rec* (1993) 132:584-585.
39. Lang-Ree JR, Vatn T, Kommisrud E, Løken T. Transmission of bovine virus diarrhoea virus by rectal examination. *Vet Rec* (1994) 135:412-413.
40. Goyal SM, Ridpath JF. Bovine Viral Diarrhea Virus Diagnosis, Management and Control. (2005) Iowa: Blackwell.
41. Barrett DJ, More SJ, Graham DA, O'Flaherty J, Doherty ML, Gunn HM. Failure to spread bovine virus diarrhoea virus infection from primarily infected calves despite concurrent infection with bovine coronavirus. *Vet J* (2002) 163:251-259. doi:10.1053/tvjl.2001.0657
42. Tinsley M, Lewis FI, Brülisauer F. Network modeling of BVD transmission. *Vet Res* (2012) 43:1-11. doi:10.1186/1297-9716-43-11

43. Valle PS, Martin SW, Tremblay R, Bateman K. Factors associated with being a bovine-virus diarrhoea (BVD) seropositive dairy herd in the Mùre and Romsdal County of Norway. *Prev Vet Med* (1999) 40:165-177. doi: 10.1016/S0167-5877(99)00030-6
44. Lindberg A, Houe H. Characteristics in the epidemiology of bovine viral diarrhea virus (BVDV) of relevance to control. *Prev Vet Med* (2005) 72:55-73. doi:10.1016/j.prevetmed.2005.07.018
45. Voas S. Working together to eradicate BVD in Scotland. *Vet Rec* (2012) 170:278-279.
46. Smith RL, Sanderson MW, Renter DG, Larson RL, White BJ. A stochastic model to assess the risk of introduction of bovine viral diarrhea virus to beef cow-calf herds. *Prev Vet Med* (2009) 88:101-108.
47. Smith RL, Sanderson MW, Renter DG, Larson R, White B. A stochastic risk-analysis model for the spread of bovine viral diarrhea virus after introduction to naive cow-calf herds. *Prev Vet Med* (2010) 95:86-98. doi:10.1016/j.prevetmed.2010.02.009
48. Perry GH. Risk assessment of transmission of bovine viral diarrhea virus (BVDV) in abattoir-derived in vitro produced embryos. *Theriog* (2007)68:38-55. doi:10.1016/j.theriogenology.2007.03.022
49. Newcomer BW, Toohey-Kurth K, Zhang Y, Brodersen BW, Marley MS, Joiner KS, Zhang Y, Galik PK, Riddell KP, Givens MD. Laboratory diagnosis and transmissibility of bovine viral diarrhea virus from a bull with a persistent testicular infection. *Vet Microbiol* (2014) 170:246-257. doi:10.1016/j.vetmic.2014.02.028
50. Voges H, Horner GW, Rowe S, Wellenberg GJ. Persistent bovine pestivirus infection localized in the testes of an immuno-competent, non-viraemic bull. *Vet Microbiol* (1998) 61:165e75. doi:10.1016/S0378-1135(98)00177-1

Table 1 Distribution of the sampled herds and the absolute and relative frequencies of positive herds in the state of Rio Grande do Sul, Brazil.

Region	Total herds (N)	Positive herds (n)
1-Northwest	58,456 (72%)	20% (57/277)
2-Northeast	8,475 (10%)	19% (8/41)
3-Southwest	1,910 (2%)	66% (6/9)
4-Western central	1,178 (1%)	83% (5/6)
5-Eastern central	5,401 (7%)	18% (5/27)
6-Metropolitan	2,563 (3%)	33% (4/12)
7-Southeast	3,324 (4%)	50% (8/16)
Total	81,307 (100%)	24% (93/388)

Table 2 Definition and distribution of explanatory variables retained at the unadjusted univariable analysis (399 herds)*.

Variables	Frequency (%) or median	*P*–value	PR (IC 95%)
Inseminator was always the same		0.13	
No	29		-
Yes	71		1.56 (0.95-2.55)
Who does AI the animals		0.19	
Owner/farm labor	2		-
State veterinary	2		0.96 (0.18-5.12)
Cooperative technician	1		0.25 ((0.04-1.42)
Outsourcing worker	95		0.38 (0.06-2.14
Total farm area[1]		0.08	
< 0.2 Km^2	52		-
≥ 0.2 Km^2	48		1.43 (1.02-2.02)
Observed weak calves		0.16	
No	76		-
Yes	24		0.68 (0.44-1.06)
Does rectal examination as routine		<0.001	
No	56		-
Yes	44		2.42 (1.70-3.44)
How much cows represents for farm income		0.01	
0-20%	20		-
21-40%	16		1.86 (0.99-3.52)
41-60%	25		1.39 (0.75-2.57)
61-80%	16		2.57 (1.40-4.68)
81-100%	23		1.60 (0.87-2.96)
Areas used to farm cow		0.009	
<0.2 Km^2	59		-
≥ 0.2 Km^2	41		1.71 (1.21-2.40)
Dry cow's –not lactating		0.01	
<3 animals	47		-
≥3 animals	53		1.68 (1.77-2.40)
How many years does produce milk	2	0.14	0.98 (0.97-1.00)
Does perform "*Premunição*"		0.03	
No	80		-
Yes	10		2.08 (1.17-3.70)
Whether farm had technical assistance		0.17	
No	48		-
Yes	52		0.72 (0.53-1.06)
Frequency of veterinary		0.04	

assistance			
Annual	2		-
Semester	2		3.80 (1.29-11.18)
Monthly	16		1.50 (0-81-2.78)
Only when requested	80		7.46 (0.42-1.31)
Have send animal to fairs last year		0.14	
No	95		-
Yes	5		1.77 (0.93-3.38)
Direct contact over the fence		0.01	
No	45		-
Yes	55		1.68 (1.17-2.43)
The use of AI		0.08	
No	20		1.50 (1.02-2.20)
Yes	80		-
Does keep mortality records		0.10	
No	43		-
Yes	57		1.42 (0.99-2.03)
Whether farm separated animals by age		0.01	
No	41		-
Yes	59		1.72 (1.18-2.52)

[1] Variable used to control for potential confounding, * Poisson regression.

Table 3 Final Poisson regression model of those variables associated with BVDV for n = 388 BTM samples

Variables	Estimate (β) [a]	S.E. [b]	*P*-value	PR (CI: 95%)
Confounder (Total farm area)			0.45	
< 0.2 Km^2				-
≥ 0.2 Km^2	0.15	0.21		1.17 (0.82-1.66)
Associated factors				
Routine use of rectal examination			<0.001	
Yes	1.00	0.23		2.73 (1.87-3.98)
No	-	-		-
Direct contact over the fences			0.02	
Yes	0.49	0.22		1.63 (1.13-2.35)
No	-	-		-
The use of AI			0.002	
Yes	-	-		-
No	0.72	0.24		2.07 (1.38-3.09)

[a,b] Results given with Estimate (β), standard errors (S.E.), *P*-values and PR with 95% CI.

Figure legends

Figure 1

The location of 388 samples and 93 positive herds tested for antibodies against BVDV in the state of Rio Grande do Sul, Brazil. The names of the regions are indicated: 1) Northwest Region, 2) Northeast Region, 3) Southwest Region, 4) Western central Region, 5) Eastern central Region, 6) Metropolitan Region, and 7) Southeast Region.

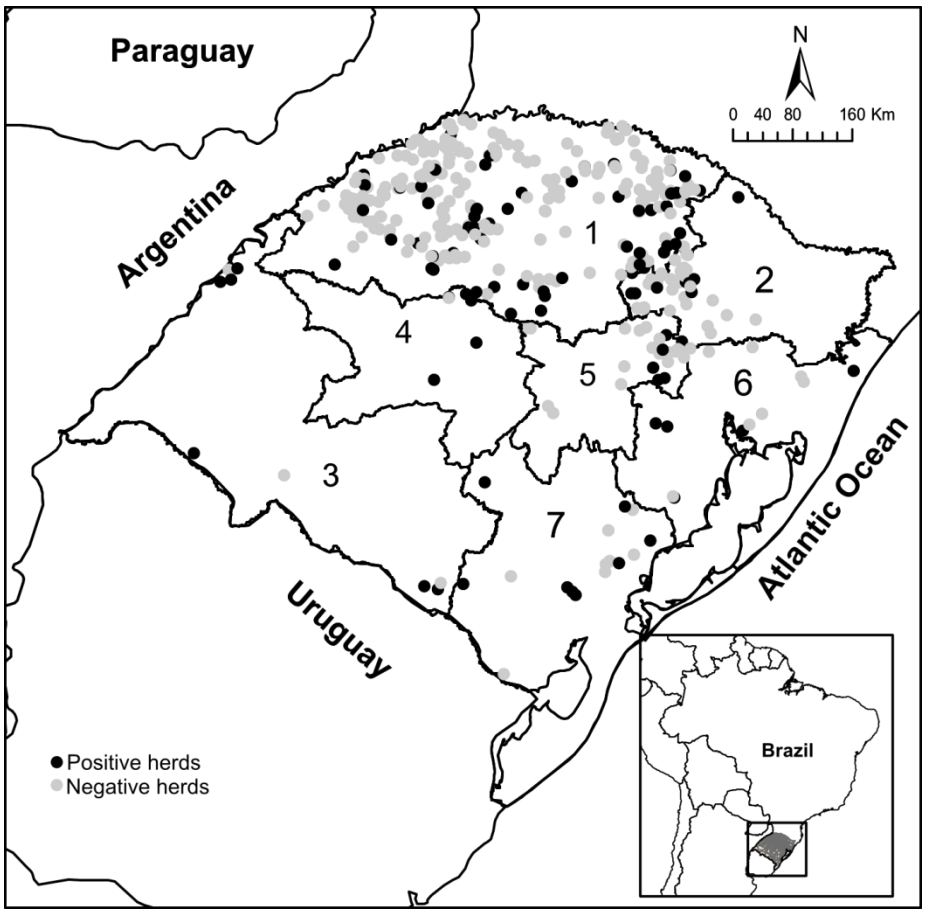

Paraguay
Argentina
Uruguay
Atlantic Ocean
Brazil
N
0 40 80 160 Km
1
2
3
4
5
6
7
Positive herds
Negative herds

Figure 1

5. CONSIDERAÇÕES FINAIS

O vírus da diarreia viral bovina (BVDV) tem sido estudado em muitos países e a abordagem tradicional (regressão) tem sido aplicada classicamente para a identificação de variáveis associadas a ocorrência de BVDV, principalmente em países onde a doença já foi ou está em fase de erradicação.

O desafio principal desta tese foi o de utilizar uma abordagem analítica por RF frente às clássicas análises de regressão na identificação de fatores associados a BVDV (KALMAR, 2014). Com relação ao desempenho preditivo das técnicas utilizadas nesse estudo, quando se utilizou a média dos escores de AUC para comparação da validação cruzada entre RF e a regressão logística, ficou demostrada a superioridade do desempenho da RF, em concordância com outros trabalhos científicos que vem mostrando melhor desempenho do RF quando comparado com a regressão logística (CASANOVA et al. 2014). Esses resultados reforçam o potencial que a RF tem para ser utilizado na identificação e classificação de preditores em estudos transversais de prevalência.

Quando analisado os resultados dos dois capítulos desta tese pode-se perceber uma aderência quanto as variáveis identificadas como associadas a BVDV.

A utilização da palpação retal nos rebanhos foi identificada como variável significativa tanto através da análise de regressão como na análise por RF. Quando analisamos a concordância entre este resultado deve-se pensar que existe um grande indicativo de que esta prática de manejo esteja realmente interferindo na distribuição de BVDV, principalmente dentro dos rebanhos e que medidas educativas e de treinamento podem contribuir para a interrupção desta rota de transmissão.

Já o contato direto entre animais de propriedades vizinhas foi identificado via regressão, e através de RF, a variável numérica representada pelo “número de propriedades vizinhas com criação de bovinos”, foi identificada como preditora da ocorrência de BVDV. Neste caso existe um desafio muito grande para o combate ou atenuação desta variável, uma vez que a construção de cerca dupla já foi utilizada em alguns países, porém é uma medida economicamente inviável para muitos países, especialmente o Brasil. Talvez os achados mais discordantes entre a análise de regressão

e RF tenham ficado a cargo da não utilização da inseminação artificial como fator associado com o aumento da ocorrência de BVDV através da regressão tradicional e, por outro lado os resultados obtidos por RF, identificaram que quando a inseminação era realizada ou pelo próprio proprietário ou pelo responsável da propriedade se tinham maiores probabilidades de ocorrência de BVDV. Apesar da discordância, ambos estão diretamente relacionados com o manejo reprodutivo, seja através da utilização de touro ou a utilização de sêmen não testado ou por falhar na realização da inseminação por pessoal com pouco ou nenhum treinamento técnico e sanitário.

Pode-se concluir que os achados encontrados reforçam o que vem sendo encontrado como fatores associados à BVD no estado do Rio Grande do Sul. Muitos dos fatores encontrados estão ligados à reprodução, o que solidifica as necessidades de se considerar como ponto central da atuação do serviço veterinário oficial e da iniciativa privada, o manejo reprodutivo e uma melhor educação continuada, para que caso venha-se a introduzir um programa de controle e erradicação da BVDV em propriedades leiteiras no RS, sejam obtidos os melhores níveis de redução da ocorrência em menor espaço de tempo possível.

Por fim, deve-se apontar como uma vantagem da utilização dos métodos tradicionais de regressão em comparação a RF, a facilidade na realização das análises e interpretação dos resultados, porém quando comparados os desempenhos dos modelos de regressão e RF, este último mostrou-se melhor. Por outro lado, a complexidade para a realização das análises via RF pode ser destacada como um limitante, assim como a não existência de uma medida de efeito simples para a interpretação dos resultados como é o caso das razões de chances, risco relativo ou razão de prevalências, por exemplo, que são produzidas pelas regressões. Não obstante, a utilização de RF tende a se popularizar nas áreas de medicina humana, e com isso no futuro próximo pode vir a se tornar uma aliada de peso nas análises de fatores associados em estudos rotineiros de prevalência em medicina veterinária.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ABRAHANTES, J.C., BOLLAERTS, K., AERTS, M., OGUNSANYA, V., STEDE, Y. Salmonella sorosurveliance: Different statistical methods to categorize pig herds based on sorological data. **Preventive Veterinary Medicine**, v. 89, p. 59-66, 2009.

ARENHART, S., BAUERMANN, F. V., OLIVEIRA, S. A. M., WEIBLEN, R., FLORES, E. F. Shedding and transmission of bovine viral diarrhea virus by persistently infected calves. **Pesquisa Veterinária Brasileira**, v. 29, p. 736–742, 2009.

BAKER, J. C. The clinical manifestations of bovine viral diarrhea infection. **Veterinary Clinics of North America: Food Animal Practice**, v. 11, p. 425–445, 1995.

BEAUDEAU, F., BELLOC, C., SEEGERS, H., ASSIÉ, S., SELLAL, E., JOLY, A. Evaluation of a blocking ELISA for the detection of bovine viral diarrhoea virus (BVDV) antibodies in serum and milk. **Veterinary Microbiology**, v. 80, p. 329–337. 2001a.

BEAUDEAU, F., ASSIÉ, S., SEEGERS, H., BELLOC, C., SELLAL, E., JOLY, A. Assessing the within-herd prevalence of cows antibody-positive to bovine viral diarrhoea virus with a blocking ELISA on bulk tank milk. **Veterinary Record**, v. 149, p. 236–240, 2001b.

BISSACCO, A., YANG, M.H., SOATTO, S. Fast human pose estimation using appearance and motion via multi-dimensional boosting regression, in IEEE **Conference on Computer Vision and Pattern Recognition**, 2007.

BOLIN, S. R., GROOMS, D. L. Origination and consequences of bovine viral diarrhea virus diversity. **Veterinary Clinics of North America: Food Animal Practice**, v. 20, p. 51–68, 2004.

BOLIN, S. R., MCCLURKIN, A. W., CUTLIP, R. C., CORIA, M. F. Severe clinical disease induced in cattle persistently infected with noncytopathic bovine viral diarrhea virus by superinfection with cytopathic bovine viral diarrhea virus. **American Journal of Veterinary Research**, v. 46, p. 573–576, 1985.

BOTTON, S. A., DA-SILVA, A. M., BRUM, M. C. S., WEIBLEN, R., FLORES, E. F. Antigenic characterization of Brazilian bovine viral diarrhea virus isolates by monoclonal antibodies and cross-neutralization. **Brazilian Journal of Medical Biological Research**, v. 31, p. 1429–1438, 1998.

BRITO, M.A.V.P.; BRITO, J.R.F.; RIBEIRO, M.T.; VEIGA, V.M.O. Padrão de infecção intramamária em rebanhos leiteiros: exame de todos os quartos mamários das vacas em lactação. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 51, p. 129-135, 1999.

BREIMAN, L., FRIDMAN, J.H., OLSHEN, R.A., STONE, C.L. Classification and regression trees. **Chapman & Hall**, 1984.

BREIMAN, L. Bagging predictions. **Machine Learning**, v. 24, p. 123-140, 1996.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 2001.

BROCK, K. V., GROOMS, D. L., RIDPATH, J., BOLIN, S. R. Changes in levels of viremia in cattle persistently infected with bovine viral diarrhea virus. **Journal of Veterinary Diagnostic Investigation**, v. 10, p. 22–26, 1998.

BROWNLIE, J. The pathogenesis of bovine virus diarrhoea virus infections. **Revue Scientifique et Technique**, v. 9, p. 43–59, 1990.

BROWNLIE, J., CLARKE, M. C. Experimental and spontaneous mucosal disease of cattle: a validation of Koch's postulates in the definition of pathogenesis. **Intervirology**, v. 35, p. 51–59, 1993.

CASANOVA, R., SALDANA, S., CHEW, E.Y., DANIS, R.P., GREVEN, C.M., AMBROSIUS, W.T. Application of random forests methods to diabetic retinopathy classification analyses. PLoS One 9:e98587, 2014

CASARO, A. P., KENDRICK, J. W., KENNEDY, P. C. Response of the bovine fetus to bovine viral diarrhea-mucosal disease virus. **American Journal of Veterinary Research**, v. 32, p. 1543–1562, 1971.

CASAUBON, J., VOGT, H. R., STALDER, H., HUG, C., RYSER-DEGIORGIS, M. P. Bovine viral diarrhea virus in free-ranging wild ruminants in Switzerland: low prevalence of infection despite regular interactions with domestic livestock. **BMC Veterinary Research**, v. 8, p. 1–14, 2012.

CHASE, C. C. L. The impact of BVDV infection on adaptive immunity. **Biologicals**, v. 41, p. 52–60, 2012.

CLSI. **Performance Standards for Antimicrobial Disk and Dilution Susceptibility Tests for Bacteria Isolated from Animals:** Approved standard. 2. ed. Wayne, PA: CLSI / NCCLS, 2002. 86 p. NCCLS document M31-A2.

CORTES, C., VAPNIK V. Support-vector networks. **Machine Learning**, v. 20, p. 273–97, 1995.

COSTA, E.O., BENITES, N.R., MELVILLE, P.A., PARDO, R.B., RIBEIRO, A.R., WATANABE, E.T. Estudo etiológico da mastite clínica bovina. **Revista Brasileira de Medicina Veterinária**, v. 17, p. 156-158, 1995.

DIETTERICH, T.G., 2000. **Machine Learning.** In David Hemmendinger, Anthony

Ralston and Edwin Reilly (Eds.), The Encyclopedia of Computer Science, Fourth Edition, Thomson Computer Press. 1056-1059.

DUBOVI, E. J. Laboratory diagnosis of bovine viral diarrhea virus. **Biologicals**, v. 41, p. 8–13, 2012.

EIRAS, C., ARNAIZ, I., SANJUÁN, M. L., YUS, E., DIÉGUEZ, F. J. Bovine viral diarrhea virus: Correlation between herd seroprevalence and bulk tank milk antibody levels using 4 commercial immunoassays. **Journal of Veterinary Diagnostic Investigation**, v. 24, p. 549–553, 2012.

FEE. Fundação Estadual de Economia e Estatística (FEE). 2014. Em 2013, **PIB gaúcho cresce 5,8% e alcança o valor de R$ 310,5 bilhões**. Disponível em: <http://www.fee.rs.gov.br/indicadores/pib-rs/pib-trimestral/destaques//>. Acessado em: 03/2014.

FISCHER, R.A. The use of multiple measurements in taxonomic problems. **Annals of Human Genetics**, v. 7, p. 179-188, 1936.

FIXT, E, HODGES, J.L. Discriminatory analysis-nonparametric discrimination: consistency properties. **International Statistic Review**, v. 57, p. 238-247, 1989.

FLORES, E. F., WEIBLEN, R., VOGEL, F. S. F., ROEHE, P. M., ALFIERI, A. A., PITUCO, E. M., A. infecção pelo vírus da Diarréia Viral Bovina (BVDV) - histórico, situação atual e perspectivas. **Pesquisa Veteterinária Brasileira**, v. 25, p. 125–134, 2005.

FREY, H. R., FLEBBE, U., LIESS, B. Prevalence and clinical symptoms of persistent BVD-virus infection in cattle herds of Lower Saxony. **Praktische Tierarzt**, v. 77, p. 49–52, 1996.

FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifiers. **Machine Learning**, v. 29, p. 131-163, 1997.

FUSARO, V.A., MANI, D.R., MESIROV, J.P., CARR, S.A. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. **Nature Biotechnology**, v. 27, p. 190-198, 2009.

GAROUSSI, M. T., HAGHPARAST, A., ESTAJEE, H. Prevalence of bovine viral diarrhoea virus antibodies in bulk tank milk of industrial dairy cattle herds in suburb of Mashhad-Iran. **Preventive Veterinary Medicine**, v. 84, p. 171–176, 2008.

GONDA, M. G., FANG, X., PERRY, G. A., MALTECCA, C. Measuring bovine viral diarrhea virus vaccine response: Using a commercially available ELISA as a surrogate for serum neutralization assays. **Vaccine**, v. 30, p. 6559–6553, 2012.

GOYAL, S. M., RIDPATH, J. F. **Bovine Viral Diarrhea Virus**. 1. ed. Victoria: Wiley-Blackwell. 2008. 272 p.

GRAAT, E.A.M.; FRANKENA, K.; BOS, H. Principles and methods of sampling in animal diseases surveys. In: NOORDHUIZEN, J.P.T.M.; FRANKENA, K.; VAN DER HOOFD, C.M. **Application of quantitative methods in veterinary epidemiology**. Wageningen: Wageningen Pers, 1997. p.31-62.

GROOMS, D. L. Reproductive consequences of infection with bovine viral diarrhea virus. **Veterinary Clinics of North America: Food Animal Practice**, p. 20, v. 5–19, 2004.

GROOMS, D. L. Reproductive losses caused by bovine viral diarrhea virus and leptospirosis. **Theriogenology**, v. 66, p. 624–628, 2006.

GUARINO, H., NÚÑEZ, A., REPISO, M. V., GIL, A., DARGATZ, D. A. Prevalence of serum antibodies to bovine herpesvirus-1 and bovine viral diarrhea virus in beef cattle in Uruguay. **Preventive Veterinary Medicine**, v. 85, p. 34–40, 2008.

HANON, J. B., VAN DER STEDE, Y., ANTONISSEN, A., MULLENDER, C., TIGNON, M., VAN DEN BERG, T., CAIJ, B. Distinction Between Persistent and Transient Infection in a Bovine Viral Diarrhoea (BVD) Control Programme: Appropriate Interpretation of Real-Time RT-PCR and Antigen-ELISA Test Results. **Transboundary and Emerging Diseases**, v. 61, 156-162, 2012.

HOLTKAMP, D.J. LIN, H. WANG, C., OÇANNOR, A.M. Identifying questons in the American Association of Swine Veterinary's PRRS risk assessment survey that are important for retrospectively classifying swine herds according to wheter they reported clinical PRRS outbreaks in the previuos 3 years. **Preventive Veterinary Medicine**, v. 106, p. 42-52, 2012.

HOUE, H. Age distribution of animals persistently infected with bovine virus diarrhea virus in twenty-two Danish dairy herds. **Canadian Journal of Veterinary Research**, v. 56, p. 194–198, 1992.

HOUE, H. Bovine virus diarrhoea virus: detection of Danish dairy herds with persistently infected animals by means of a screening test of ten young stock. **Preventive Veterinary Medicine**, v. 19, p. 241–248, 1994.

HOUE, H. Epidemiology of bovine viral diarrhea virus. **Veterinary Clinics of North America: Food Animal Practice**, v. 11, p. 521–547, 1995.

HOUE, H. Epidemiological features and economical importance of bovine virus diarrhoea virus (BVDV) infections. **Veterinary Microbiology**, v. 64, p. 89–107, 1999.

HOUE, H., LINDBERG, A., MOENNIG, V. Test strategies in bovine viral diarrhea virus control and eradication campaigns in Europe. **Journal of Veterinary Diagnostic Investigation**, v. 18, p. 427–436, 2006.

HUMPHRY, R. W., BRÜLISAUER, F., MCKENDRICK, I. J., NETTLETON, P. F., GUNN, G.J. Prevalence of antibodies to bovine viral diarrhoea virus in bulk tank milk and associated risk factors in Scottish dairy herds. **Veterinary Record**, v. 171(18), p. 445, 2012.

HUTCHINSON, R. A., LIU, L.P., AND DIETTERICH, T. G. **Incorporating boosted regression trees into ecological latent variable models**, in *AAAI'11,* (San Francisco, CA), p. 1343–1348. 2011.

IBGE. Instituto Brasileiro de Estatística (IBGE), 2011. **Produção da Pecuária Municipal**. v. 39. IBGE, 60 p.

JIANG, P., WU, H., WANG, W., MA, W., SUN, X., LU, Z. MiPred:classificationofreal and pseudo microRNA precursors using random forest prediction model with combined features. **Nucleic Acids Research**, v. 35, p. 339–344, 2007.

JOHNSON, R., AND ZHANG, T. Learning Nonlinear Functions Using Regularized Greedy Forest. **Technical Report**. 2012.

JUNTTI, N., LARSSON, B., FOSSUM, C. The Use of Monoclonal Antibodies in Enzyme Linked Immunosorbent Assays for Detection of Antibodies to Bovine Viral Diarrhoea Virus. **Journal of Veterinary Medicine Series B**, v. 34, p. 356–363, 1987.

WÅLINDER, A. Evaluation of logistic regression and random forest classification based on prediction accuracy and metadata analysis. Institutionen för matematik, 2014.

KAMPA, J., STAHL, K., RENSTRÖM, L. H. M., ALENIUS, S. Evaluation of a commercial Erns-capture ELISA for detection of BVDV in routine diagnostic cattle serum samples. **Acta Veteterinaria Scandinavica**, v. 49, p. 1–7, 2007.

KAMPA, J., STÅHL, K., MORENO-LÓPEZ, J., CHANLUN, A., AIUMLAMAI, S., ALENIUS, S. BVDV and BHV-1 infections in dairy herds in northern and northeastern Thailand. **Acta Veteterinaria Scandinavica**, v. 45, p. 181–192. 2004.

KARATZOGLOU, A., MEYER, D. Support Vector Machines in R. **Journal of Statistical Software**, v. 15(9), p. 1-28, 2006.

KATZ, J. B., HANSON, S. K. Competitive and blocking enzyme-linked immunoassay for detection of fetal bovine serum antibodies to bovine viral diarrhea virus. **Journal of Virological Methods**, v. 15, p. 167–175, 1987

KLEINBAUM, D.G., KUPPER, L.L., CHAMBLESS, L.E. Logistic regression analysis of epidemiologic data: theory and practice. **Communications in Statistics**, v. 11, p. 485–

547, 1982.

KUHNE, S., SCHROEDER, C., HOLMQUIST, G. Detection of Bovine Viral Diarrhoea Virus Infected Cattle – Testing Tissue Samples Derived from Ear Tagging Using an Erns Capture ELISA. **Journal of Veterinary Medicine Series B**, v. 52, p. 272–277, 2005.

LANGONI, H., PINTO, M.P., DOMINGUES, P.F., LISTONI, F.J.P. Etiologia e susceptibilidade da mastite bovina subclínica. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 43, p. 507-515, 1991.

LAUREYNS, J., PIEPERS, S., RIBBENS, S., SARRAZIN, S., DE VLIEGHER, S., VAN CROMBRUGGE, J. M., DEWULF, J. Association between herd exposure to BVDV-infection and bulk milk somatic cell count of Flemish dairy farms. **Preventive Veterinary Medicine**, v. 109, p. 148-151, 2012.

LARISON, B., NJABO, K.Y., CHASAR, A., FULLER, T., HARRIGAN, R.J., SMITH, T.B. Spillover of pH1N1 in Camaroon: an investigation of risk factors. **BMC Veterinary Medicine**, v. 10, p. 1-8, 2014.

LEE, K.C., CHO, H. Performance of Ensemble Classifier for Location Prediction Task: Emphasis on Markov Blanket Perspective. **International Journal of u-and e-Service, Science and Technology**, v. 3, p. 1-12, 2010.

LINDBERG, A., HOUE, H. Characteristics in the epidemiology of bovine viral diarrhea virus (BVDV) of relevance to control. **Preventive Veterinary Medicine**, v. 72, p. 55–73, 2005.

LUZZAGO, C., FRIGERIO, M., PICCININI, R., DAPRA, V., ZECCONI, A. A scoring system for risk assessment of the introduction and spread of bovine viral diarrhea virus in dairy herds in Northern Italy. **Veterinary Journal**, v. 177, p. 236–241, 2008.

MACHADO G, EGOCHEAGA RM, HEIN HE, MIRANDA IC, NETO WS, ALMEINDA LL, STEIN MC, CORBELLINI LG. Bovine Viral Diarrhoea Virus (BVDV) in Dairy Cattle: A Matched Case-Control Study. Transbound Emerg Dis (2014) Ahead of print. doi: 10.1111/tbed.12219

MAPA. Ministério da Agricultura Pecuária e Abastecimento. 2013. **Exportação.** Disponível em: <http://www.agricultura. gov.br/animal/exportacao>. Acessado em: 10/ 2014.

MCCLURKIN, A. W., CORIA, M. F., CUTLIP, R. C. Reproductive performance of apparently healthy cattle persistently infected with bovine viral diarrhea virus. **Journal of the American Veterinary Medical Association**, v. 174, p. 1116–1119, 1979.

MCCLURKIN, A. W., LITTLEDIKE, E. T., CUTLIP, R. C., FRANK, G. H., CORIA, M. F., BOLIN, S. R. Production of cattle immunotolerant to bovine viral diarrhea virus.

Canadian Journal of Comparative Medicine, v. 48, p. 156–161, 1984.

MEYLING, A., HOUE, H., JENSEN, A. M. Epidemiology of bovine virus diarrhoea virus. **Revue Scientifique et Technique (International Office of Epizootics)**, v. 9, p. 75–93, 1990.

MOERMAN, A. A., STRAVER, P. J. P., DE JONG, M. C. M., QUAK, J. J., BAANVINGER, T. T., VAN OIRSCHOT, J. T. J. A long term epidemiological study of bovine viral diarrhoea infections in a large herd of dairy cattle. **Veterinary Record**, v. 132, p. 622–626, 1993.

NATEKIN, A. KNOLL, A. Gradient boosting machines, a tutorial. **Frontiers in Neurorobotics**, v. 7, p. 1-21, 2013.

NETTLETON, P. F., ENTRICAN, G. Ruminant pestiviruses. **British Veterinary Journal**, v. 151, p. 615–642, 1995.

NISKANEN, R. Relationship between the levels of antibodies to bovine viral diarrhoea virus in bulk tank milk and the prevalence of cows exposed to the virus. **Veterinary Record**, v. 133, p. 341–344, 1993.

NISKANEN, R., ALENIUS, S., LARSSON, B., JACOBSSON, S. O. Determination of level of antibodies to bovine virus diarrhoea virus (BVDV) in bulk tank milk as a tool in the diagnosis and prophylaxis of BVDV infections in dairy herds. **Archives of Virology**, **Supplement**, v. 3, p. 245–251, 1991.

NISKANEN, R., EMANUELSON, U., SUNDBERG, J., LARSSON, B., ALENIUS, S. Effects of infection with bovine virus diarrhoea virus on health and reproductive performance in 213 dairy herds in one county in Sweden. **Preventive Veterinary Medicine**, v. 23, p. 229–237, 1995.

NISKANEN, R., LINDBERG, A. Transmission of bovine viral diarrhoea virus by unhygienic vaccination procedures, ambient air, and from contaminated pens. **Veterinary Journal**, v. 165, p. 125–130, 2003.

NIZA-RIBEIRO, J., PEREIRA, A., SOUZA, J., MADEIRA, H., BARBOSA, A., AFONSO, C. Estimated BVDV-prevalence, -contact and -vaccine use in dairy herds in Northern Portugal. **Preventive Veterinary Medicine**, v. 72, p. 81–85, 2005.

OIE. World Organization for Animal Health. Harmonisation of national antimicrobial resistance surveillance and monitoring programs. In: OIE. **Terrestrial animal health code**. Paris: OIE, 2008b. (Capítulo 6.5. p. 1-7). Disponível em http://www.oie.int/eng/normes/Mcode/en_sommaire.htm. Acesso em 20/11/2008.

OLIVEIRA, L.S. **Contas Regionais: O desempenho da economia do RS em 2009. Indicadores Econômicos FEE**. v. 37(4), p. 7-28, 2010.

PATON, D. J., CHRISTIANSEN, K. H., ALENIUS, S., CRANWELL, M. P., PRITCHARD, G. C., DREW, T. W. Prevalence of antibodies to bovine virus diarrhoea virus and other viruses in bulk tank milk in England and Wales. **Veterinary Record**, v. 142, p. 385–391, 1998.

PATON, D. J., IBATA, G., EDWARDS, S., WENSVOORT, G. An ELISA detecting antibody to conserved pestivirus epitopes. **Journal of Virological Methods**, v. 31, p. 315–324,1991.

PETERHANS, E., BACHOFEN, C., STALDER, H., SCHWEIZER, M. Cytopathic bovine viral diarrhea viruses (BVDV): emerging pestiviruses doomed to extinction. **Veterinary Research**, v. 41, p. 41–44, 2010.

PETERHANS, E., SCHWEIZER, M. Pestiviruses: How to outmaneuver your hosts. **Veterinary Microbiology**, v. 142, p. 18–25, 2010.

PITTMAN, S. J., AND BROWN, K. A. Multi-scale approach for predicting fish species distributions across coral reef seascapes. Plos one, v. 6, p. 1-12, 2011.

POLETTO, R., KREUTZ, L. C., GONZALES, J. C., BARCELLOS, L. J. G. Prevalence of tuberculosis, brucelosis and viral infections in dairy cattle from the county of Passo Fundo, RS, Brazil. **Ciência Rural**, v. 34, p. 595–598, 2004.

POLIKAR, R. Ensemble Based Systems in Decision Making. **IEEE Circuits and Systems Magazine**, v. 6, p. 21-45, 2006.

QUINCOZES, C. G., FISCHER, G., DE OLIVEIRA HÜBNER, S. Prevalence and factors associated with bovine viral diarrhea virus infection in South of Rio Grande do Sul. **Semina-Ciências Agrárias**, v. 28, p. 269–276, 2007.

RAUE, R., HARMEYER, S. S., NANJIANI, I. A. Antibody responses to inactivated vaccines and natural infection in cattle using bovine viral diarrhoea virus ELISA kits: assessment of potential to differentiate infected and vaccinated animals. **Veterinary Journal**, v. 187, p. 330–334, 2011.

REINHARDT, G., RIEDEMANN, S., ERNST, S., AGUILAR, M., ENRIQUEZ, R., GALLARDO, J. Seroprevalence of bovine viral diarrhea/mucosal disease in southern Chile. **Preventive Veterinary Medicine**, v. 10, p. 73–78, 1990.

RIDPATH, J. F. Bovine viral diarrhea virus: global status. **Veterinary Clinics of North America: Food Animal Practice**, v. 26, p. 105–121, 2010a.

RIDPATH, J. F. The contribution of infections with bovine viral diarrhea viruses to **bovine respiratory disease. Veterinary Clinics of North America: Food Animal**

Practice, v. 26, p. 335–348, 2010b.

RIDPATH, J. F., BOLIN, S. R., DUBOVI, E. J. Segregation of bovine viral diarrhea virus into genotypes. **Virology**, v. 205, p. 66–74, 1994.

RIKULA, U., NUOTIO, L., AALTONEN, T., RUOHO, O. Bovine viral diarrhoea virus control in Finland 1998-2004. **Preventive Veterinary Medicine**, v. 72, p. 139–142, 2005.

RUMELHART DE, HINTON GE, WILLIAMS RJ. Learning representations by back-propagating errors. **Nature**, v. 323, p. 533–536, 1986.

SAINANI, K. L., POPAT, R. A. Understanding Study Design. **PM & R**, v. 3, p. 573–577, 2011.

SARRAZIN, S., VELDHUIS, A., MÉROC, E., VANGEEL, I., LAUREYNS, J., DEWULF, J., CAIJ, A. B., PIEPERS, S., HOOYBERGHS, J., RIBBENS, S., VAN DER STEDE, Y. Serological and virological BVDV prevalence and risk factor analysis for herds to be BVDV seropositive in Belgian cattle herds. **Preventive Veterinary Medicine**, v. 108, p. 28–37, 2012.

SMITH, R. L., SANDERSON, M. W., RENTER, D. G., LARSON, R. L., WHITE, B. J. A stochastic model to assess the risk of introduction of bovine viral diarrhea virus to beef cow-calf herds. **Preventive Veterinary Medicine**, v. 88, p. 101–108, 2009.

STÅHL, K., ALENIUS, S. BVDV control and eradication in Europe—an update. Japanese **Journal of Veterinary Research**, v. 60, p. S21–S39, 2012.

STÅHL, K., LINDBERG, A., RIVERA, H., ORTIZ, C., MORENO-LÓPEZ, J. Self-clearance from BVDV infections--a frequent finding in dairy herds in an endemically infected region in Peru. **Preventive Veterinary Medicine**, v. 83, p. 285–296, 2008.

STONE M. Cross-validatory choice and assessment of statistical predictions. **Journal of the Royal Statistical Society B (Methodological)**, v. 36, p. 111-147, 1974.

STOTT, A. W., HUMPHRY, R. W., GUNN, G. J. Modelling the effects of previous infection and re-infection on the costs of bovine viral diarrhoea outbreaks in beef herds. **Veterinary Journal**, v. 185, p. 138–143, 2010.

SYARIF, I., PRUGEL-BENNETT, A. WILLS, G. Application of bagging, boosting and stacking to intrusion detection. Machine Learning and Data Mining in Pattern Recognition. v. 7376, p. 594-602, 2012.

SYNGE, B. A., CLARK, A. M., MOAR, J. A., NICOLSON, J. T., NETTLETON, P. F., HERRING, J. A. The control of bovine virus diarrhoea virus in Shetland. **Veterinary Microbiology**, v. 64, p. 223–229, 1999.

TALAFHA, A. Q., HIRCHE, S. M., ABABNEH, M. M., AL-MAJALI, A. M., ABABNEH, M. M. Prevalence and risk factors associated with bovine viral diarrhea virus infection in dairy herds in Jordan. **Tropical Animal Health and Production**, v. 41, p. 499–506, 2009.

THOMPSON, J. A., DE MIRANDA HENRIQUES LEITE, R., GONÇALVES, V. S. P., LEITE, R. C., BANDEIRA, D. A., HERRMANN, G. P., MOREIRA, E. C., PRADO, P. E. F., LOBATO, Z. I. P., DE BRITO, C. P. T., LAGE, A. P. Spatial hierarchical variances and age covariances for seroprevalence to Leptospira interrogans serovar hardjo, BoHV-1 and BVDV for cattle in the State of Paraíba, Brazil. **Preventive Veterinary Medicine**, v. 76, p. 290–301, 2006.

TRÅVÉN, M. M., ALENIUS, S. S., FOSSUM, C. C., LARSSON, B. B. Primary bovine viral diarrhoea virus infection in calves following direct contact with a persistently viraemic calf. **Zentralbl Veterinarmed B**, v. 38, p. 453–462, 1991.

USDA. United States Department of Agriculture. 2013. **Foreign Agricultural Service. Livestock and Poultry: World Markets and Trade**, 25p. Disponível em: <http://www.fas.usda.gov/psdonline/circulars/livestock_poultry.pdf>. Acessado em: 10/2014.

VALLE, P., MARTIN, S., TREMBLAY, R., BATEMAN, K. Factors associated with being a bovine-virus diarrhoea (BVD) seropositive dairy herd in the Mùre and Romsdal County of Norway. **Preventive Veterinary Medicine**, v. 40, p. 165–177, 1999.

VALLE, P. S., SKJERVE, E., MARTIN, S. W., LARSSEN, R. B., OSTERAS, O., NYBERG, O. Ten years of bovine virus diarrhoea virus (BVDV) control in Norway: a cost-benefit analysis. **Preventive Veterinary Medicine**, v. 72, p. 189–207, 2005.

VAPNIK, V.N., 1999. An Overview of Statistical Learning Theory, IEEE, **Transactions on Neural Networks**, v. 10 n° 5.

VERIKAS, A., GELZINIS A, BACAUSKIENE M. Mining data with random forests: a survey and results of new tests. **Pattern Recognit**, v. 44, p. 330–49, 2011.

VOAS, S. Working together to eradicate BVD in Scotland. **Veterinary Record**, v. 170, p. 278–279, 2012.

WEISS, S.M., KULIKOWSKI, C.A. **Computer systems that learn: classification and prediction methods from statistics neural nets, machine learning, and expert systems**. Morgan Kanfmann Publishers inc. San Francisco, CA, USA. 1991.

TOUW, W.G., BAYJANOV, J.R., OVERMARS, L., BACKUS, L., BOEKHORST, J., WELS, M., VAN HIJUM, S.A.F.T. Data mining in the Life Sciences with Random

Forest: a walk in the park or lost in the jungle? **Briefings in bioinformatics**, v. 14, p. 315-326, 2012.

WOLD, H., 1975. Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Ap- proach. In J. Gani (Ed.), Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett, 520–540. **Academic Press**, London.