Universidade de Brasília
Departamento de Estatística

Modelo Econométrico de share de uso do solo no Mato Grosso, Brasil

Gustavo Maia Rodrigues Gomes
Igor Ribeiro Mendonça

Monografia apresentada para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2017

Gustavo Maia Rodrigues Gomes
Igor Ribeiro Mendonça

Modelo Econométrico de share de uso do solo no Mato Grosso, Brasil

Orientador:
Prof. Dr. **Antônio Eduardo Gomes**

Monografia apresentada para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2017

Dedicatória

Dedicamos este trabalho à todos que nos ajudaram durante nossa graduação, vida profissional e pessoal, especialmente pais, irmãos, familiares e amigos.

Gustavo Gomes e Igor Mendonça

Agradecimentos

Agradecimento especial para: José, Tatiana, Jarbas, Dirce, Celeste, Sheila, Paulo, Valentina, Márcio, Lucas, Luísa, Júlia, Marcelo, Juliana, Marli, Ademar, Sandro, Henrique, Bruno, Marilene, Marília, Marta, José, Isabelle, Raner, Rodrigo, Bruna, Camila, Luciana, Raquel, Caio, Bento, Janine, André, Pedro, Kaique, Carolina, Gustavo, Felipe, Antônio, Augusto, Augusto, Gabriel, Rafael, Isabella, Giovana, Markus, Fernando, Ramon, Vinícius, Lucas, Gabriel, Adriano, Michel, Felipe, Higor, Eduardo, Flávio, Carolina, Carolina, Thaís, Aghata, Mayara, Natália, Júlia, Thaisa, Julia, Kauana, Larissa, Rachael, Garrett, Hans, Galymzhan, Kuki, Béltran, Dani, Pali, Molly, Lukas, Maxi, Jacob, Dilma, Júlio, Gabriel, Victor, Rafael, João, Ornélio, raphael, Laura, Kiko, Fabiana, Janine, Flávio, Vera, Gustavo, Gustavo, Marcos, Denise, Rodrigo, Alice, Bárbara, Bárbara, Gustavo, Alex, Lourenzzo, Kendrick, Kanye, Obama, Anthony, William, Igor, Gabriel, Luan, Samuel, Marília, Maiara, Maraísa, Bruno, Marrone, Wesley, Danilo, Vitor, Nícolas, Rodrigo, Deni, Luísa, Beatriz, Lorena, Flávia, Quentin, Christopher, Bryan, Vincent, Carlos, Luiz, Luiz, Cristiano, Tait, Ticiano, Ellen, Thaís, Cecília, Márcio, Bruno, George, Kevin, Elias, Rodrigo, Robert, Stan, Antônio e outros que foram importantes em certo momento até aqui, mas não foram mencionados.

Igor Ribeiro Mendonça

Aos meus pais que nunca mediram esforços para proporcionar uma educação de qualidade para mim e por terem me ensinado as maiores virtudes que carrego hoje comigo.

Aos meus mentores, Antônio Eduardo Gomes pela paciência na orientação e incentivo; e Alexandre Xavier Ywata de Carvalho pela sugestão do tema e discussões explanatórias que tornaram possível a conclusão desta monografia.

Ao meu irmão Guilherme por sempre confiar no meu potencial e nunca deixado que eu tomasse decisões ruins na minha caminhada acadêmica e pessoal.

Á minha irmã Giovanna por ser um exemplo pra mim, que apesar de tudo que ela já viveu, no final ela sempre carrega um sorriso reconfortante.

Aos meus familiares, em especial, minha avó Antônia, por terem sempre incentivado e, algumas vezes proporcionado, oportunidades que desenvolveram o meu conhecimento cultural.

Por último, não menos importante, à Deus e meus amigos, pois, sem sombra de dúvidas, foram a base do suporte onde me apoiei nos momentos difíceis nesta graduação.

Gustavo Maia Rodrigues Gomes

Sumário

Resumo

Este trabalho propõe o uso de metódos lineares em modelos de shares de uso do solo nos municípios do estado do Mato Grosso. Uma necessidade em dias de tamanha degradação ambiental, um modelo que possa tentar quantificar políticas públicas e as decisões de uso do solo com base em variáveis pertinentes. A base de dados foi dividida em base de treino e base de teste, depois foi utilizado o recurso do StepWise Selection para selecionar dentre as inúmeras variáveis coletadas quais as que mais influenciam na porcentagem de uso do solo para cada uso do solo, após ter um modelo adequado segundo critérios de escolha de variável são feitas validações cruzadas do modelo com estatísticas de precisão da predição e por último testa-se como o modelo treinado consegue prever a base teste.

Palavras-chave: Modelos lineares, Modelos de uso do solo, Modelo econométrico.

Abstract

In this project we propose the use of linear models in land share models in the counties of Mato Grosso, Brazil. An important resource due to the enormous environment degradation we see these days, a model that will be able to quantify public policy and the decision on the use of the land based in pertinent variables. The data was divided: training and test, it was used stepwise selection to select among several colected variables which were more influential in the percentage of the use of land in each land use, after that cross validations of the model are performed with precision statistics of prediction. At last we test the trained model in the test data to see its capability of prediction.

Keywords: Linear models, Land share models, Econometric models.

Lista de Figuras

Lista de Tabelas

1 Introdução

Um exame histórico, ainda que superficial, da alocação de recursos no uso do solo no Brasil reflete de imediato que, cada vez mais, boa parte do país está sendo usado para agropecuária e para cidades, enquanto, cada vez menos, para florestas. As florestas do Brasil estão perdendo espaço para o crescimento econômico, para a construção de estradas, para a pecuária em larga escala, para expansão da fronteira agrícola e para o aumento da densidade populacional (Arraes et. al, 2012)

É fato que tanto as atividades industriais quanto as atividades agrícolas colaboram para o enriquecimento do Brasil medido pelo PIB principalmente devido à modernização da estrutura de trabalho e ao aumento da produtividade (Brugnaro e Bacha, 2009). Em contrapartida, elas causam problemas ambientais como degradação intensa das águas, já que muitas fontes naturais de água acabaram devido ao mau uso e manejo incorreto das mesmas; como a erosão do solo causada pelo uso intensivo do solo aliado ao manejo inadequado da água, entre outros danos causados ao meio ambiente, (De Deus e Bakonyi, 2012).

Surge então, como propósito deste trabalho, a necessidade de um estudo mais aprofundado que consiga quantificar de maneira coerente esses ganhos e perdas de ativos ambientais (Chakir et. al, 2014) e, assim, melhor alocar, do ponto de vista econômico ambiental, os recursos no solo do Brasil. Portanto, a criação de um modelo econométrico se torna evidente.

No tocante à disponibilidade de dados, devemos evidenciar dois grupos de dados mais utilizados: individuais e agregados. Os dados individuais envolvem modelos discretos de escolha para selecionar alguma categoria de uso do solo e são mais penosos de serem obtidos devido aos custos. Portanto, dados individuais são mais utilizados quando queremos analisar localmente o uso do solo.

Já os dados agregados são mais utilizados quando queremos analisar o uso do solo de regiões com dados heterogêneos e fazer predições do uso de solo para regiões, como escrito por Chakir et. al (2014). Os nossos dados são por município e nosso objetivo é fazer predições acerca do uso no Mato Grosso (temos o caso de dados agregados).

Os dados biofísicos são formados por 45 variáveis que trazem informações do solo e do clima de 141 municípios do estado do Mato Grosso mais 6 variáveis sociais, com alguns dados populacionais e índices sociais. Com respeito aos dados econômicos, temos basicamente 4 tipos de varíaveis que serão utilizadas no modelo: produção, valor de produção, área plantada (somente para agricultura) e área colhida (somente para agricultura) para a agricultura temporária, agricultura permanente, extração vegetal, silvicultura e pecuária para todos os municípios do estado nos 5 anos do estudo: 2004, 2008, 2010, 2012 e 2014,

resultando em um total de 16 variáveis. Para obter a variável dependente usa-se uma base com as porcentagens de uso do solo, pastagens, florestas, agrícolas e outros para todos os municípios e os 5 anos do estudo.

É comum na literatura de modelos de share de uso do solo (porcentagem em cada município que é destinada para cada uso do solo) a escolha do modelo da forma logística (Chakir et. al, 2014), principalmente pelo fato de que este permite uma transformação logarítmica deixando a parte das variáveis independentes lineares. Dessa forma é possível aplicar as técnicas lineares para estimação dos parâmetros.

A partir deste modelo, temos como objetivo quantificar a relação entre nossas variáveis independentes (produção, valor de produção, biofísicos e sociais) com o share de uso do solo, podendo ser usado para avaliar os efeitos das políticas públicas ambientais, tais como impostos e subsídios para cada setor, ou mesmo predizer as mudanças do uso do solo.

2 Revisão de Literatura

2.1 Modelos econométricos

Econometria é uma área da ciência econômica, onde são usados modelos matemáticos e aplicadas diversas técnicas estatísticas para medir e estimar relações econômicas. É possível estudar teorias macroeconômicas, políticas públicas governamentais, grandes questões de interesse da área em geral, e também para medir simples dados, como a relação de salário de trabalhadores e horas dedicadas ao estudo, tempo de experiência ou talvez notas de alunos em matemática e a frequência nas aulas destes, atendimento em cursos supletivos, presença em plantões de dúvidas.

Essa área de modelagem, por mais que utilize na sua maioria modelos estatísticos, como exemplo a regressão linear, múltipla, é imprescindível sua existência, porque a análise e a discussão de resultados podem variar muito do que um estatístico analisaria. Economistas aprofundaram e focaram em assuntos de seu interesse e até adequaram novas técnicas para desenvolverem melhores resultados.

Economistas podem ter conhecimento teórico suficiente para saber que certas variáveis têm efeitos sobre outras. Mesmo que não tenham, é interessante a aplicação de um modelo que consiga relacioná-las, basicamente temos:

$$Y_i = f(X_1, X_2, ..., X_k) \tag{1}$$

Y_i é um conjunto de observações que está em função das variáveis explicativas $X_1, ..., X_k$. Com essa estrutura, vários tipos de dados podem ser usados para mensurar a variável dependente, Y, por exemplo, Y_i sendo o salário de trabalhadores, X_{1i} anos de experiência de cada trabalhador e X_{2i}, anos de estudo, estimado por regressão a seguinte relação:

$$Y_i = 450 * X_{1i} + 1567 * X_{2i}, \tag{2}$$

$i = 1, 2, ..., t$

Pode ser feita uma análise bem básica, pela visão econométrica, para questões de ilustração que tendo anos de estudo fixo, cada ano de experiência do trabalhador aumenta em 450,00 Reais o seu salário e tendo anos de experiência fixos, cada ano de estudo aumenta em 1567,00 Reais o salário do trabalhador para os dados propostos.

2.2 Modelo multinível

Modelos multinível são modelos estatísticos paramétricos que podem variar em diferentes níveis, é conhecido na literatura por ser uma generalização de modelos lineares. Os dados para esta análise podem ser obtidos através de uma pesquisa amostral (o caso deste projeto), experimento, bastante usado na ciência, estudo pela observação de um fenômeno, por censo ou variações e junções destas formas.

Os dados são aninhados em mais de um nível. Por exemplo, podemos querer estimar a relação da exportação de soja em cada ano em vários países pela produção de cada ano para cada país. Neste caso os dados tem dois níveis: ano e país.

As suposições do modelo multinível são as mesmas da regressão linear, condicionadas com o tipo de dados para este caso. São elas: linearidade (também pode ser aplicado para relações não lineares), homocedasticidade, normalidade, e independência das observações.

Assim, temos no sistema, três componentes: a variável resposta (o que queremos estimar pelo método estatístico de regressão), as variáveis explicativas e a função que liga esses dois componentes (esta pode ser linear ou não linear).

2.3 Modelos Econométricos de Shares do Uso do Solo

A metodologia para análise do uso de shares do solo possui algumas variações com base na maneira de obtenção dos dados que são utilizados nos modelos. Logo, será apresentado a metodologia para dados individuais[1] como também para dados agregados.

2.3.1 Modelos para Dados de Indivíduos

Os modelos para dados de indivíduos se dividem em dois tipos com base em sua natureza temporal. O modelo estático é apropriado para os casos onde a decisão do proprietário da terra quanto ao uso é baseada nos lucros realizados ao final do período sendo o mesmo problema repetido no final dos períodos subsequentes.

O Modelo dinâmico se torna necessário na maioria das análises, uma vez que em muitos casos, os retornos acontecem em períodos posteriores ao da decisão de alocação (como o caso de floresta) e também quando a decisão de uso tem conexões intertemporais entre usos (como por exemplo, o caso de alternância entre agricultura temporária e pastagem).

[1] Fora do escopo deste trabalho, uma vez que não foram encontrados dados desse nível de especificação.

2.3.1.1 Modelo Estático

O caminho natural para a modelagem quando tratamos de dados individuais está em encontrar os valores ótimos de cada partição para cada uso do solo onde o Administrador da Terra $n_i(n_i = 1, ..., N_i)$ na região $i(i = 1, ..., I)$ é assumido de maximizar os lucros advindos do uso $k(k = 1, ..., K)$ em uma terra de qualidade $j(j = 1, ..., J)$. (Miller e Platinga, 1999)

Seja $\mathbf{x}(t, n_i)$ um vetor com as informações econômicas advindas de variáveis como preço, produção, custo ou outras variáveis de decisões econômicas da região i e $a_{jk}(t, n_i)$ a área de qualidade j da fazenda que é dedicada ao uso k. Portanto, $\pi_{jk}(\mathbf{x}(t, n_i), a_{jk}(t, n_i), n_i)$ é uma função restrita de lucro.

É fácil a visualização de que cada fazendeiro deve selecionar a área alocada para cada uso de maneira em que temos o valor máximo da função restrita de lucro. Então, devemos maximizar:

$$\sum_k \pi_{jk}(\mathbf{x}(t, n_i), a_{jk}(t, n_i), n_i) \tag{3}$$

sujeito a

$$\sum_k a_{jk}(t, n_i) = A_j(t, n_i) \tag{4}$$

para cada j onde $A_j(t, n_i)$ é a area total disponível para a área de qualidade j. Logo, $A(t, n_i) = \sum_j A_j(t, n_i)$.

Temos que $a^*_{jk}(\mathbf{x}(t, n_i), A_j(t, n_i), n_i)$ são as alocações ótimas advindas da solução de Kuhn-Tucker para o problema. Portanto, a partilha da terra ótima alocada para o uso k é:

$$f_k(\mathbf{X}(t, n_i), t, n_i) = \frac{1}{A(t, n_i)} \sum_k a^*_{jk}(\mathbf{x}(t, n_i), A_j(t, n_i), n_i) \tag{5}$$

onde temos que $\mathbf{X}(t, n_i)$ é um H-vetor que possui tanto os dados econômicos quanto variáveis do tipo biofísico.

2.3.1.2 Modelo Dinâmico

Quando analisamos sobre a ótica de um modelo dinâmico, temos que trabalhar com a esperança do valor presente de lucro e alocar a área para o uso que possuir a maior esperança dentre eles. Portanto o modelo é similar ao exposto em (5), mas com $\mathbf{X}(t, n_i)$ medindo ganhos futuros ajustados aos usos alternativos. Logo, a parcela da propriedade do agente n_i que será dedicada ao uso k será:

$$s_k(t, n_i) = f_k(\mathbf{X}(t, n_i), t, n_i) + u_k(t, n_i) \tag{6}$$

onde $u_k(t, n_i)$ representa o erro advindo de fatores externos àqueles que se encontram no modelo após a alocação (como por exemplo um período de chuva atípico na região). Então, assumimos que os erros não são correlacionados com as variáveis explicativas do modelo de tal forma que $E[X_h(t, n_i)u_k(t, n_i)] = 0$ para todo n_i, h, t e k.

Outro fator importante é que, como as partilhas dadas a cada uso devem somar 1, temos que $\sum_k u_k(t, n_i) = 0$.

2.3.2 Modelo para Dados Agregados

Quando mudamos o foco para dados agregados, trabalharemos com informações de propriedades de terra agrupadas já que esta é a natureza deste tipo de dado. Portanto, utilizaremos dados que podem ser encontrados sobre a forma de censo ou de amostra. Teremos, então, que a área observada que é alocada para o uso k no município i é:

$$y_k(t, i) = \sum_{n_i=1}^{N_i} w(t, n_i)[s_k(t, n_i) + v_k(t, n_i)] + \bar{v_k}(t) = p_k(t, i) + \epsilon_k(t, i) \tag{7}$$

A interpretação de $w(t, n_i)$ é de que esta quantidade representa a porção relativa da terra na região i que é administrada pelo indivíduo n_i caso os dados venham de um censo, mas caso venham de uma amostra, a quantidade $w(t, n_i)$ representa o peso dado ao indivíduo n_i.

Já $v_k(t, n_i)$, interpretamos como sendo o erro potencial de amostra associado com cada observação e $\bar{v}_k(t)$ como sendo o erro de amostra agregado. É assumido que os erros são variáveis aleatórias com média zero e variância finita, que não são correlacionados com as variáveis explicativas, que não são correlacionados ao longo do tempo, dos municípios e dos indivíduos, mas são correlacionados ao longo das categorias de uso (por construção).

Na terceira igualdade de (7), temos que $p_k(t, i)$ é a função das variáveis econômicas e/ou biofísicas para o município i no tempo t para o uso k como também representa a porção esperada para o uso k no município i no tempo t, enquanto $\epsilon_k(t, i)$ é o erro que possui a mesma estrutura de média-variância que $u_k(t, n_i)$. Por último, dado que as observações para município estão sujeitas a erros de coleta e de medidas, consideramos $\mathbf{X}(t, i)$ e $\epsilon(t, i)$ como processos estocásticos mas mantendo a igualdade:

$$E[X_h(t, i)\epsilon_k(t, i)] = 0 \tag{8}$$

que deve ser real para todo t, i, h e k para que as variáveis explicativas sejam não correlacionadas com os erros agregados.

Para encontrarmos $p_k(t, i)$, muitos autores primeiramente fazem uma alteração na

variável $y_k(t,i)$ a ser explicada, onde divide-se todos os $y_k(t,i)$ por $y_1(t,i)$ e depois toma-se os logarítimos naturais dos resultados, para assim, encontrarmos que:

$$ln(\frac{y_k(t,i)}{y_1(t,i)}) = \beta_k^{'}\mathbf{X}(t,i) - \beta_1^{'}\mathbf{X}(t,i) \tag{9}$$

É linear nos parâmetros do modelo. Portanto, se fizermos $\beta_1 = 0$, normalizaremos os parâmetros e assim o modelo é identificado. Então, para chegarmos em $p_k(t,i)$, basta substituirmos o resultado de (7) em:

$$p_k(t,i) = \frac{exp(\beta_k^{'}\mathbf{X}(t,i))}{\sum_k exp(\beta_k^{'}\mathbf{X}(t,i))} \tag{10}$$

Teremos, então, $T(K-1)$ equações em $H(K-1)$ parâmetros desconhecidos, que podem ser estimados por Mínimos Quadrados se $H < T$.

3 Metodologia

O modelo utilizado no nosso trabalho é baseado na revisão de literatura dos modelos de shares de uso do solo. Entretanto, por temos mais variáveis explicativas que observações para os modelos específicos de cada município, trabalharemos com um modelo de regressão para dados combinados, ou seja, trabalharemos com uma regressão para a base de dados que possui informações de todos os municípios tornando $H < T$. Nossa análise é baseada em 3 usos (Agricultura, Pecuária e Floresta). A primeira etapa da metodologia foi encontrar as estimativas dos parâmetros para cada um destes usos através da Regressão Linear por Mínimos Quadrados onde utilizamos apenas os dados econômicos. Em uma segunda etapa, o mesmo problema foi analisado utilizando os dados biofísicos em adição aos dados econômicos.

3.1 Material

Os dados usados neste relatório foram retiradas de diferentes lugares, especificamente são três fontes. A variável dependente foi obtida através do projeto TerraClass do Instituto Nacional de Pesquisas Espaciais (INPE). Este projeto utiliza imagens de satélite para obter o mapeamento do uso e cobertura do solo de alguns estados do país, com os dados do estado do Mato Grosso contendo a área do uso do solo para cada um dos 141 municípios do estado para os anos de 2004, 2008, 2010, 2012 e 2014, assim como a área total de cada município. É possível criar a porcentagem de cada diferente tipo de uso do solo em relação município: mata e floresta, agrícola e pastagem.

As variáveis independentes são separadas em dois tipos: as de produção e as biofísicas. Para a primeira, foram colhidas informações de diferentes pesquisas do Instituto Brasileiro de Geografia e Estatística (IBGE): Produção Agrícola Municipal (PAM), Produção da Extração Vegetal e Silvicultura (PEVS) e Pesquisa Pecuária Municipal (PPM). Estas três pesquisas continham informações a nível municipal do estado do Mato Grosso sobre produção e valor de produção agrícola, pecuária, extração vegetal e silvicultura para os cinco anos: 2004, 2008, 2010, 2012 e 2014, obtendo as seguintes variáveis:

Tabela 1: Variáveis de produção e valor de produção

Variável	Unidade
área plantada agricultura temporária	Hectares
área colhida agricultura temporária	Hectares
Produção agrícola temporária	Toneladas
Valor da produção agrícola temporária	Milhares de reais
área plantada agricultura permanente	Hectares
área colhida agricultura permanente	Hectares
Produção agrícola permanente	Toneladas
Valor da produção agrícola permanente	Milhares de reais
Produção da extração vegetal	Toneladas
Valor da produção da extração vegetal em toneladas	Milhares de reais
Produção da extração vegetal	Metros cúbicos
Valor da produção da extração vegetal em metros cúbicos	Milhares de reais
Produção da silvicultura	Metros cúbicos
Valor da produção da silvicultura	Milhares de reais
Efetivo de animais da pecuária	Quantidade
Valor da produção de origem animal	Milhares de reais

Os dados biofísicos são os mesmos utilizados na publicação "Clusterização espacial e não espacial: um estudo aplicado à agropecuária brasileira", do Instituto de Pesquisa Econômica Aplicada (IPEA). Estes usam dados do IBGE para retratar diferentes características em nível municipal do Brasil. São elas: características físicas, aspectos do solo, tipos de clima, fertilidade, limitações do solo, seca, tipos de solo, temperatura, topologia. Com isso temos as seguintes variáveis:

Tabela 2: Variáveis de tipos de solo dos municípios

Variável	Unidade
Solo bom	Porcentagem
Solo bom a regular	Porcentagem
Solo regular a bom	Porcentagem
Solo regular	Porcentagem
Solo regular a bom	Porcentagem
Solo regular a restrito	Porcentagem
Solo restrito a desfavorável	Porcentagem
Solo restrito	Porcentagem

Tabela 3: Variáveis de fertilidade dos municípios

Variável	Unidade
Fertilidade alta	Porcentagem
Fertilidade baixa	Porcentagem
Fertilidade média	Porcentagem
Fertilidade média a alta	Porcentagem
Fertilidade muito baixa	Porcentagem

Tabela 4: Variáveis de temperatura dos municípios

Variável	Unidade
Quente	Porcentagem
Subquente	Porcentagem
Mesotérmico branco	Porcentagem
Mesotérmico mediano	Porcentagem

Tabela 5: Variáveis de limitações do solo dos municípios

Variável	Unidade
Alta salinidade	Porcentagem
Baixo nível de nutrientes	Porcentagem
Baixo nível de nutrientes e textura grosseira	Porcentagem
Declives acentuados, pouca profundidade e textura grosseira	Porcentagem
Declives acentuados, drenagem restrita e excesso de alumínio	Porcentagem
Declives acentuados	Porcentagem
Disponibilidade de nutrientes entre média e baixa	Porcentagem
Excesso de sódio, drenagem restrita e risco de inundação	Porcentagem
Praticamente sem limitações	Porcentagem
Impedimento de drenagem e risco de inundação	Porcentagem

Tabela 6: Variáveis de seca dos municípios

Variável	Unidade
Sem seca	Porcentagem
Sub seca	Porcentagem
seca entre 1 a 2 meses	Porcentagem
seca por 3 meses	Porcentagem
seca entre 4 a 5 meses	Porcentagem
seca por 6 meses	Porcentagem
seca entre 7 a 8 meses	Porcentagem
seca entre 9 a 10 meses	Porcentagem
seca por 11 meses	Porcentagem

Tabela 7: Variáveis de topografia dos municípios

Variável	Unidade
Plano e suavemente ondulado	Porcentagem
Plano a ondulado	Porcentagem
Forte ondulado	Porcentagem
Ondulado a montanhoso	Porcentagem
Montanhoso a escarpado	Porcentagem

Tabela 8: Variáveis de clima dos municípios

Variável	Unidade
Semiárido	Porcentagem
Semiúmido	Porcentagem
Superúmido	Porcentagem
úmido	Porcentagem

Em adição a estes dados biofísicos desta publicação do IPEA, temos alguns dados sociais e índices, com base em dados do IBGE e do IPEA. São estes:

Tabela 9: Variáveis sociais e índices sociais dos municípios

Variável	Unidade
População rural em 2010	Quantidade
População urbana em 2010	Quantidade
índice de desenvolvimento humano municipal em 2010	índice
índice de desenvolvimento humano municipal da educação em 2010	índice
índice de desenvolvimento humano municipal da longevidade em 2010	índice
índice de desenvolvimento humano municipal da renda em 2010	índice

3.2 Regressão Linear para cada uso k utilizando apenas dados econômicos e a variável de localidade dada pelo código de município do IBGE

Para darmos prosseguimento à etapa de regressão linear dos usos, primeiramente, calculamos o valor de um quarto uso $y_k(t,i)$ onde $k = 1$, chamado de *outros* que é $y_1(t,i) = 1 - \sum_{k=2}^{4} y_k(t,i)$. Então determinamos a igualdade:

$$ln(\frac{y_k(t,i)}{y_1(t,i)}) = \beta_k' \mathbf{X}(t,i) - \beta_1' \mathbf{X}(t,i),$$

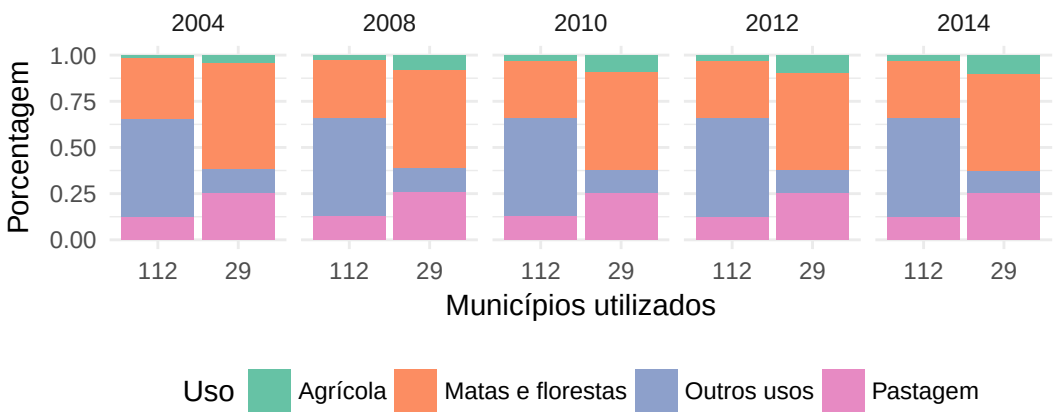
onde fixamos $\beta_1 = 0$, resultando em nossa função de regressão linear para o uso k:

$$ln(\frac{y_k(t,i)}{y_1(t,i)}) = \beta_k' \mathbf{X}(t,i).$$

Então, selecionamos os municípios que possuiam $ln(y_k(t,i)) < -15$ e que possuiam o uso *outros* inferior a 25% do total de uso para o município, uma vez que assumimos que, neste caso, houve erro de coleta de dados, pois os municípios que não se encontravam neste intervalo possuíam valores para algum share k aproximadamente igual a 0 e para *outros* um valor superior à um quarto do total da região, o que não é esperado quando trabalhamos com dados agregados se as categorias de uso forem apropriadamente definidas.

O gráfico a seguir exemplifica o exposto, uma vez que, para 112 dos 141 municípios do estado do Mato Grosso, tivemos um percentual de mais de 50% para o uso *outros* enquanto seus valores percentuais para o uso *agricultura* foram ínfimos. Já para os 29 municípios de estudo deste trabalho, tivemos uma quantidade bem menor do uso *outros*, enquanto os demais usos tiveram um aumento em seus valores percentuais se comparado com os municípios retirados da análise.

Figura 1: Distribuição do uso do solo dos 112 versus os 29 municípios



Os dados, assim definidos, foram divididos em **base de dados de treinamento** onde incluímos os dados para 2004, 2008, 2010 e 2012, e utilizamos esta base para treinar o modelo. A outra base foi chamada de **base de dados de teste** que possui somente os dados para 2014 e foi usada para testar a eficácia do modelo treinado para os anos anteriores, em 2014.

Na base de treinamento, as regressões para cada uso foram estimadas utilizando todas as variáveis econômicas explicativas, de onde se iniciou um processo de *StepWise* baseado no Critério de Informação de Akaike (AIC) para a escolha das variáveis relevantes ao problema.

Em seguida, uma análise de pontos discrepantes evidenciou que alguns municípios possuíam uma variabilidade grande na variável a ser explicada ao longo dos anos, enquanto tal variabilidade não foi seguida pelas variáveis econômicas do modelo. Então, assumimos que esta discrepância ocorreu em virtude de erros na coleta de dados para a variável explicada e/ou ausência de alguma variável que controlasse tal variabilidade. Logo, preferimos restringir nosso modelo aos 20 municípios que não foram identificados com este problema, o que resultou nos modelos parciais de uso aceitos pelas análises de resíduos. Por conseguinte, para estimarmos o valor predito para $y_k(t,i)$, bastou-se efetuar a substituição dos modelos parciais em:

$$p_k(t,i) = \frac{exp(\beta_k' \mathbf{X}(t,i))}{\sum_k exp(\beta_k' \mathbf{X}(t,i))}$$

3.2.1 Validação do Modelo

Após a criação do modelo, dois tipos de validação cruzadas foram analisadas e para cada modelo criado pelo processo, foram calculados as estatísticas MAPE (Média Percentual Absoluta do Erro), MAD (Desvio Padrão Absoluto da Média) e MSD (Desvio Padrão Quadrático da Média).

3.2.1.1 Validação Cruzada - *Leave One Observation Out* (VClooo):

Na validação cruzada especificada nesta seção, treinamos 80 modelos sendo que a única diferença entre eles é que em cada um retiramos uma observação diferente dos demais modelos e então calculamos a média das estatísticas MAPE, MAD e MSD para eles. O valor resultante foi usado para comparação com o modelo original e assim aferir sua eficácia de previsão.

3.2.1.2 Validação Cruzada - *Leave Five Observations Out* (VClfoo):

Este tipo de validação cruzada segue a mesma idéia da validação cruzada *Leave One Observation Out*, entretanto, ao invés de retirarmos apenas uma observação dos dados, neste caso, retiramos 5 observações aleatórias, ou seja, treinamos 16 modelos. Então, foi calculado a média das estatísticas MAPE, MAD e MSD para comparação com o modelo original.

3.2.2 Previsão do Modelo para a Base de Dados de Teste

Nesta última etapa do modelo com as variáveis econômicas, testamos como o modelo gerado pelos anos 2004, 2008, 2010 e 2012 conseguiu prever os resultados para 2014. Portanto, utilizando a base de dados de teste, substituímos os valores da matriz $\mathbf{X}(t, i)$ na função de estimação, calculamos os valores preditos, geramos os gráficos de valores preditos versus valores reais e estudamos as diferenças entre a previsão e o valor real.

3.3 Regressão Linear para cada uso k utilizando apenas dados econômicos e biofísicos

A regressão linear utilizando os dados biofísicos em adição aos econômicos segue a mesma estrutura da regressão linear apenas com dados econômicos e o código de localidade do IBGE. A diferença encontra-se apenas na substituição do código de município do IBGE por variáveis que descrevem o ambiente (tabelas 2 a 9).

Notamos durante o trabalho, que o código de município do IBGE carrega consigo o efeito de um bloco atribuído às variáveis de localidade no tempo (dados de painel). Portanto, a metodologia nesta seção busca ser possível, ao invés de estimarmos os valores dos shares de uso para 2014, estimar os valores dos shares de uso de um município dado que ele não se encontra no modelo, mas pertence a um dos 20 municípios que foram utilizados na sub-seção anterior.

Primeiramente, restringimos os nossos dados aos municípios que foram utilizados para previsões acerca de 2014. Numa segunda etapa, treinamos o modelo pelo Método dos Mínimos Quadrados, de onde efetuamos novamente o StepWise AIC para selecionarmos as variáveis econômicas e biofísicas que seriam importantes no modelo para, por último, efetuar a análise dos resíduos. Novamente, efetuamos a validação cruzada para aferirmos como os resultados de nossa análise estatística se generalizarão para um conjunto de dados independentes. No caso, este conjunto é representado pelas observações que são retiradas no processo de validação. Por último, avaliamos o grau de previsão dada pela substituição

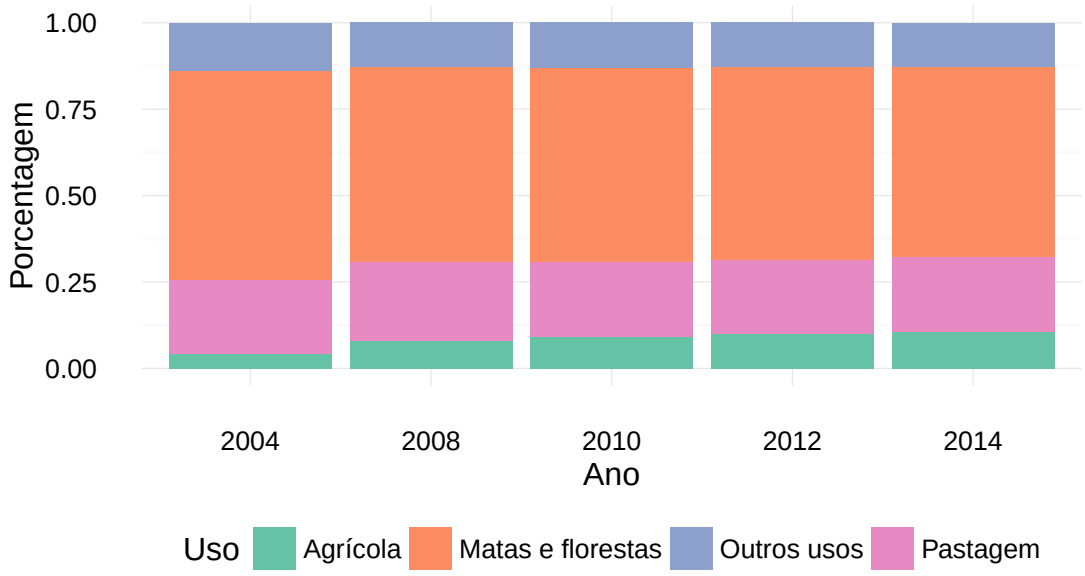
do código do IBGE por dados biofísicos.

4 Resultados e discussões

4.1 Análise Descritiva

A figura a seguir exemplifica a distribuição do uso do solo nos 20 municípios escolhidos para a modelagem da seção 4.2. É perceptível o aumento do uso do solo para agricultura, saindo de mais ou menos 4% em 2004 para aproximadamente 11% em 2014, enquanto pastagens e outros continuam em média no mesmo patamar, respectivamente 22% e 12%. As matas e florestas decaem de 60% para 55%, evidenciando um claro desmatamento e mudança na decisão do uso do solo em apenas 14% dos municípios do estado.

Figura 2: Distribuição do uso do solo nos municícipios usados no modelo



4.1.1 Dados econômicos

O estado é um grande produtor rural, com um número bastante expressivo de produção agrícola no cenário nacional. As principais culturas produzidas na agricultura temporária são: soja, cana-de-açúcar, milho e arroz. Agricultura permanente: banana, látex coagulado, coco-da-baía, café em grão. O principal produto extraído da extração vegetal é o carvão vegetal, correspondendo quase unicamente da produção em toneladas, enquanto na produção em metros cúbicos o cenário é dividido quase igualmente entre lenha e madeira em tora, seguindo este panorama, na silvicultura, o produto extraído é unicamente madeira em tora para outras finalidades. Na área agropecuária, o efetivo de animais é dividido entre galináceos e bovinos, representando juntos quase 100% para

todos os anos do efetivo de cabeças de animais, justificando o fato do valor de produção de origem animal ser quase exclusivamente advindo do leite bovino e ovos de galinha em todos os anos deste estudo.

Na próxima tabela[2][3], verifica-se o comportamento das variáveis econômicas, o mínimo para todas é igual a 0, destacando a produção agrícola temporária, com um máximo de 703 milhões de Reais, e média de 93 milhões de reais. Isso se deve ao fato de que dentro do modelo está um dos maiores municípios produtores de soja do país, Sorriso, sendo que o próprio estado concentra sua produção agrícola temporária em soja e cana-de-açúcar. Para os anos deste estudo, esses dois produtos juntos contabilizam em média 80% da produção total para este tipo de plantio.

Tabela 10: Estatísticas descritivas das variáveis usadas no modelo da seção 4.2

Variável	Máximo	Média	Desvio P.
área plantada agricultura temporária	231100	47370	52951,45
área colhida agricultura temporária	231100	47290	52968,17
Produção agrícola temporária	2182000	324200	508950,1
Valor da produção agrícola temporária	R$ 703400	R$ 93870	R$ 116888
área plantada agricultura permanente	4674	427,9	830,16
área colhida agricultura permanente	4143	384,9	750,86
Produção agrícola permanente	12000	1130	1962,01
Valor da produção agrícola permanente	R$ 15440	R$ 1583	R$ 2612,44
Produção da extração vegetal*	14760	1118	2700,01
Valor da produção da extração vegetal*	R$ 9009	R$ 610	R$ 1441,16
Produção da extração vegetal**	264700	42060	55565,12
Valor da produção da extração vegetal**	R$ 49190	R$ 4154	R$ 7545,14
Produção da silvicultura	313000	5690	32068,84
Valor da produção da silvicultura	R$ 22000	R$ 432	R$ 2287,65
Efetivo de animais da pecuária	3774000	481400	739319,80
Valor da produção de origem animal	R$ 64000	R$ 5456	R$ 9990,75

Fonte: Instituto Brasileiro de Geografia e Estatística. * produção em toneladas. **metros cúbicos

A produção de origem animal representa um valor bastante expressivo, com um máximo de 64 milhões de Reais, média de 5 milhões. Como foi dito, esta produção advém do forte número bovino e galináceo, seguido da produção da extração vegetal em metros cúbicos com um máximo de 49 milhões de reais, média de 4 milhões.

[2]Os dados desta tabela são de 20 municípios, os que foram usados no modelo da seção 4.2, para a seção 4.3 que usa dados biofísicos foram usados menos municípios, a análise descritiva dos dados econômicos para este conjunto encontra-se no anexo A.1 As seções referentes aos modelos explicam melhor a decisão de exclusão de municípios.

[3]As unidades de cada variável encontram-se na tabela 1.

4.1.2 Dados biofísicos e sociais

A parte biofísica[4] que descreve, de certa maneira, muito bem as características principais dos municípios em relação a diferentes fatores: topografia, solo, clima, temperatura, limitação do solo. É evidenciada uma média relativamente bastante alta de solo regular. Por volta de metade da área dos municípios usados no modelo tem este solo. Juntamente dos outros fatores naturais que propiciam uma produção agrícola e pecuária e o grande avanço tecnológico em corretivos do solo, adubos, defensivos agrícolas e até maquinários, é possível uma grande produção nestas áreas.

Tabela 11: Estatísticas descritivas das variáveis de tipos de solo dos municípios

Variável	Máximo	Média	Desvio Padrão
Solo bom	100%	8,54%	22,74%
Solo bom a regular	49,4%	0,74%	5,12%
Solo regular a bom	0%	0%	0%
Solo regular	100%	51,78%	36,52%
Solo regular a bom	0%	0%	0%
Solo regular a restrito	12,54%	0,15%	1,22%
Solo restrito a desfavorável	0%	0%	0%
Solo restrito	0,07%	8,47%	0,78%

Fonte: Instituto Brasileiro de Pesquisa Econômica Aplicada

É observado uma grande área dos municípios que tem em média uma fertilidade baixa ou muito baixa (por volta de 90% da área), mas seguindo a mesma lógica do parágrafo anterior, não exclui a possibilidade de uso de corretivos para o solo ou adubos químicos.

Tabela 12: Estatísticas descritivas das variáveis de tipos de fertilidade dos municípios

Variável	Máximo	Média	Desvio Padrão
Fertilidade alta	100%	8,55%	22,74%
Fertilidade baixa	100%	51,78%	36,52%
Fertilidade média	49,4%	0,74%	5,11%
Fertilidade média a alta	21%	0,22%	1,98%
Fertilidade muito baixa	100%	38,71%	35,21%

Fonte: Instituto Brasileiro de Pesquisa Econômica Aplicada

[4]Todas as variáveis biofísicas tem minímo igual a 0%: Tabelas 11 a 17.

O estado do Mato Grosso está situado na zona tropical, com uma média de quase 27ºC no ano, segundo o site *climate-data.org*. É possivel ver na próxima tabela este resultado. Em média aproximadamente quase toda a área dos municípios usados, possuem uma temperatura quente.

Tabela 13: Estatísticas descritivas das variáveis de tipos de temperatura dos municípios

Variável	Máximo	Média	Desvio Padrão
Quente	100%	96,21%	13,14%
Subquente	94,48%	3,79%	13,14%
Mesotérmico branco	0%	0%	0%
Mesotérmico mediano	0%	0%	0%

Fonte: Instituto Brasileiro de Pesquisa Econômica Aplicada

Os municípios usados no modelo apresentam em média 51% do seu espaço com um baixo nível de nutrientes e 38% de alta sanilidade, como mostra a tabela 14. Isto exemplifica novamente o mencionado nos parágrafos anteriores: esses tipos de características são passíveis de serem corrigidas com a aplicação de produtos químicos agrícolas.

Tabela 14: Estatísticas descritivas das variáveis de limitações dos solos dos municípios

Variável	Máximo	Média	Desvio Padrão
Alta salinidade	100%	38,71%	35,22%
Baixo nível de nutrientes	100%	51,78%	36,52%
Baixo nível de nutrientes e textura grosseira	0%	0%	0%
Declives acentuados, pouca profundidade e textura grosseira	12,54%	0,15%	1,22%
Declives acentuados, drenagem restrita e excesso de alumínio	0%	0%	0%
Declives acentuados	8,47%	0,07%	0,78%
Disponibilidade de nutrientes entre média e baixa	0,07%	8,47%	5,12%
Excesso de sódio, drenagem restrita e risco de inundação	0%	0%	0%
Praticamente sem limitações	100%	8,54%	22,74%
Impedimento de drenagem e risco de inundações	0%	0%	0%

Fonte: Instituto Brasileiro de Pesquisa Econômica Aplicada

A tabela a seguir mostra, juntamente com a tabela 16, uma forte característica de clima que é inerente a vários estados do país: o inverno seco devido à uma frente fria seca que chega por volta de maio no sul/sudeste e sobe passando pelo centro-oeste, chega por volta de maio e fica até o fim do inverno e meados do início da primavera.

Tabela 15: Estatísticas descritivas das variáveis de seca dos municípios

Variável	Máximo	Média	Desvio Padrão
Sem seca	0%	0%	0%
Sub seca	0%	0%	0%
Seca entre 1 a 2 meses	100%	3,65%	16,47%
Seca por 3 meses	100%	36,18%	43,98%
Seca entre 4 a 5 meses	100%	60,17%	45,85%
Seca por 6 meses	0%	0%	0%
Seca entre 7 a 8 meses	0%	0%	0%
Seca entre 9 a 10 meses	0%	0%	0%
Seca por 11 meses	0%	0%	0%

Fonte: Instituto Brasileiro de Pesquisa Econômica Aplicada

Visualiza-se que a área analisada é em suma úmida e superúmida, provavelmente pela localidade do estado do Mato Grosso, vizinho da Amazônia e a forte predominância do bioma pantanal.

Tabela 16: Estatísticas descritivas das variáveis de clima dos municípios

Variável	Máximo	Média	Desvio Padrão
Semiárido	0%	0%	0%
Semiúmido	100%	60,17%	45,85%
Superúmido	0%	0%	0%
úmido	100%	39,83%	45,85%

Fonte: Instituto Brasileiro de Pesquisa Econômica Aplicada

A predominância em média em torno de 61% da área dos municípios com a topografia plana e suavemente ondulada é um excelente indicativo que esta pode ser usada para fins agrícolas ou pecuários, caso a área tenha possibilidade devido a fatores de condições e limitações do solo.

Tabela 17: Estatísticas descritivas das variáveis de topografia dos municípios

Variável	Máximo	Média	Desvio Padrão
Plano e suavemente ondulado	100%	61,07%	1,22%
Plano a ondulado	12,54%	0,15%	1,22%
Forte ondulado	8,47%	0,07%	35,29%
Ondulado a montanhoso	0%	0%	0%
Montanhoso a escarpado	100%	38,71%	35,22%

Fonte: Instituto Brasileiro de Pesquisa Econômica Aplicada

Na tabela que segue, demonstra alguns dados sociais importantes dos municípios escolhidos, o IDHM da longevidade tem uma média muito alta, o que indica uma expectativa média de vida alta ao nascer, o IDHM da educação, com alcance entre 32% a 72%, indica que tem municípios com uma escolaridade da população adulta e fluxo escolar da população jovem entre muito baixo a alto e o IDHM da renda, indica uma variação da

renda per capita de baixo a muito alto, segundo o Atlas do Desenvolvimento Humano no Brasil.

Tabela 18: Estatísticas descritivas das variáveis sociais e índices sociais dos municípios

Variável	Mínimo	Máximo	Média	Desvio Padrão
População rural em 2010	152	21020	4377	3466,4
População urbana em 2010	944	540800	20010	57384,1
IDHM em 2010	53,8%	78,5%	68,54%	3,81%
IDHM educação em 2010	32,4%	72,6%	57,97%	6,28%
IDHM longevidade em 2010	76,1%	85%	81,78%	1,72%
IDHM renda em 2010	56,7%	80%	68,2%	4,47%

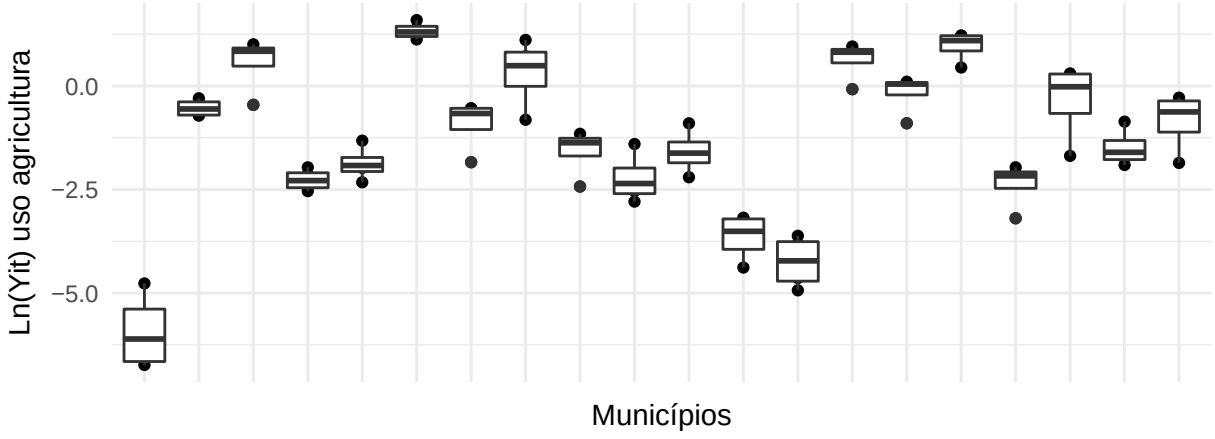
Fonte: Instituto Brasileiro de Pesquisa Econômica Aplicada

4.2 Modelo com dados econômicos e código de munícipio do IBGE

Os resultados para este tipo de modelo foram elaborados com base em análises parciais de três regressões, cada uma para um tipo de solo específico.

O *código de município* para cada regressão foi estabelecido como sendo uma variável de efeito fixo, uma vez que estamos trabalhando com dados em painel. Logo, as distribuições das quantidades à serem explicadas variam bastante entre municípios (como exemplo, para agricultura, figura 3). Desta maneira, cada município possui um intercepto único que controla o efeito do bloco *localidade*.

Figura 3: Distribuição da porcentagem do uso do solo para agricultura por municípios



Já para a variável *ano*, como podemos ver na figura 4 para agricultura, os boxplots ilustram que não houve uma variação tão grande entre anos se comparado com as distribuições entre municípios e também que o valor médio alterou-se de maneira discreta com o passar dos anos. Portanto, preferimos trabalhar com *ano* sendo uma variável númerica, a fim de mensurar este crescimento ou decaimento discreto dado pelo aumento do ano.

Figura 4: Distribuição da porcentagem do uso do solo para agricultura por ano

4.2.1 Uso do solo para agricultura

O nosso processo de análise selecionou como variáveis importantes no modelo final aquelas descritas na tabela 19. Cabe salientar que o modelo possui uma listagem com 19 dos 20 municípios pois a variável *Código de Município* se transforma em 19 variáveis binárias *dummies* para cada código (com excessão para o vigésimo município) com valores 0 ou 1, onde 0 representa ausência do município i na observação x e 1 representa sua presença. Já o vigésimo município não se encontra no modelo pois ele não possui nenhum grau de liberdade, sendo determinado pelos casos onde temos o valor 0 para todas as variáveis *dummies* criadas.

Tabela 19: Coeficientes e significância do modelo de regressão para o uso do solo para agricultura

Coeficiente	Estimativa	Significância
Intercepto	-227,8	***
Arenápolis	5,754	***
Cláudia	6,711	***
Confresa	3,887	***
Curvelândia	4,407	***
Denise	5,006	***
Feliz Natal	5,019	***
Itanhanga	6,486	***
Itauba	4,705	***
Matupá	4,115	***
Nova Santa Helena	4,747	***
Novo Mundo	2,375	***
Peixoto de Azevedo	2,165	***
Porto dos Gaúchos	6,65	***
Querência	6,159	***
Sinop	5,556	***
Terra Nova do Norte	3,802	***
União do Sul	5,283	***
Nova Guarita	4,892	***
Nova Maringá	5,457	***
Ano	0,1102	***
área plantada da agricultura permanente	0,001244	***
área colhida da agricultura permanente	-0,001221	0,000425***
Produção da agricultura temporária	0,000001311	0,0138*

Nota: Aqueles que possuem 3, 2, 1 asterisco ou (.) têm significância de 0, 0,001, 0,01 e 0,05, respectivamente.

Como resultado, vemos que cada município possui seu intercepto próprio e que a variável *ano* interfere no valor do percentual de uso para agricultura com uma tendência crescente pois temos um valor positivo para sua estimativa.

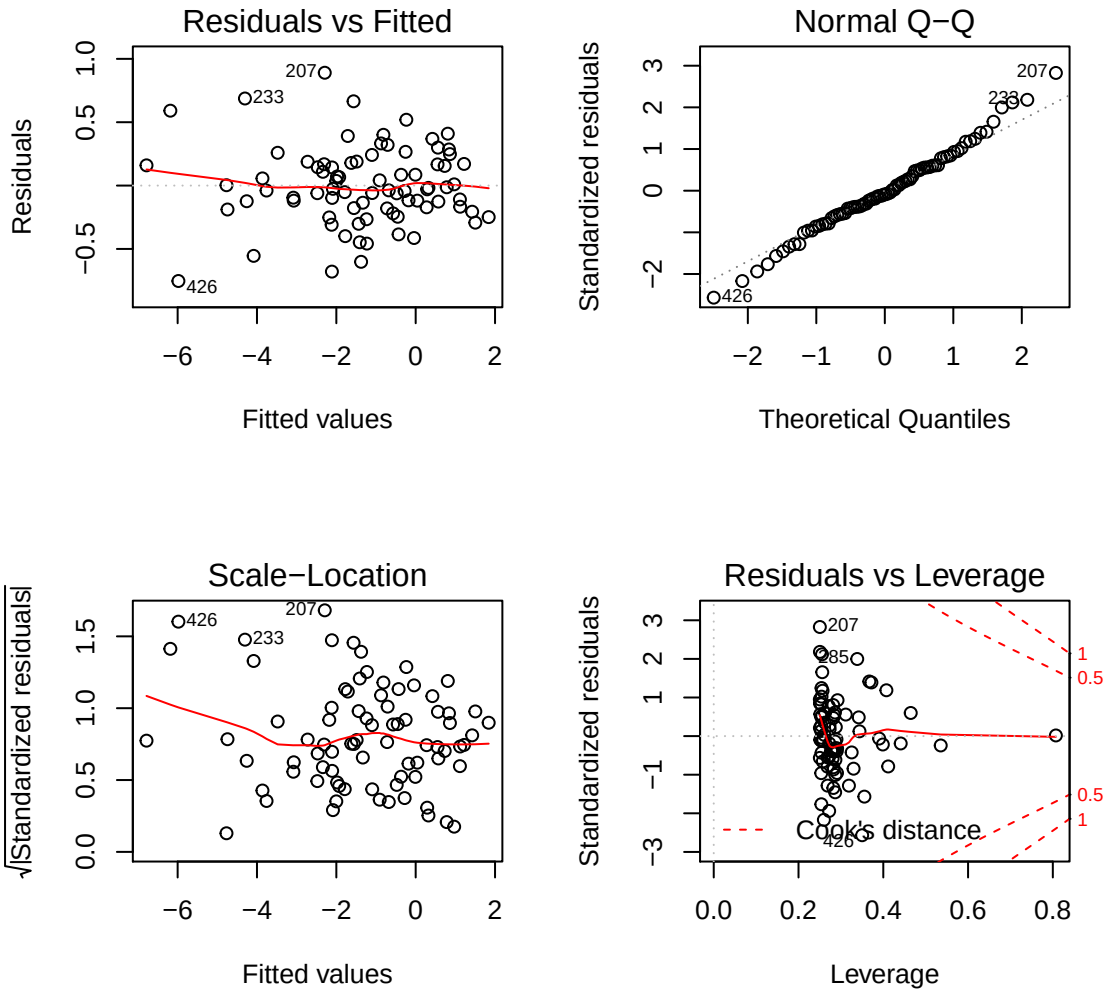
É interessante notar também que *área plantada* e *área colhida* para *agricultura permanente* interferem no percentual de uso para agricultura em praticamente mesma magnitude mas em sentidos opostos, uma vez que para o primeiro temos um valor positivo, crescente no aumento da variável a ser explicada, enquanto para o outro temos um valor negativo que representa um decréscimo no valor da variável explicada. Em uma análise específica do caso, vemos que isto se dá em virtude de que, para se plantar de maneira *permanente*, abre-se regiões para a plantação, ou seja, qualquer quantidade usada para plantação, por se tratar de uma cultura permanente, é esperada em resultar em um aumento no percentual do uso para agricultura. Porém, área colhida surge da necessidade de incorporar no resultado da variável explicada, praticamente a parcela da área plantada que ainda não foi colhida, uma vez que a estimativa dos parâmetros para essas variáveis

foram extremamente próximos (1,8% de diferença).

Por último, verificamos que a única variável relevante para mensurar os impactos da agricultura temporária na variável explicada foi a variável que representa produção para este tipo de agricultura, o que faz sentido pois, neste caso, estamos tratando de culturas que estão sujeitas ao replantio rápido após a colheita e que possuem uma quantidade limitada de produção por área determinada. Portanto, quanto maior a produção, maior será a quantidade de área utilizada.

O modelo elaborado foi aceito pela análise de resíduos, como podemos aferir na figura 5. No primeiro gráfico, que revela os valores preditos versus os resíduos, temos a suposição de que os residuos não possuem comportamento sistemático com relação à predição e, portanto, são aleatórios.

Figura 5: Gráficos do modelo de uso do solo para agricultura



O quadrante representado pelo gráfico QQ de normalidade revela que temos evi-

dências suficientes para acreditar na normalidade dos resíduos uma vez que quase todas as observações se encontram sobre o corte tranversal do gráfico. No terceiro quadrante, vemos que a suposição de homoscedasticidade do modelo está satisfeita, uma vez que a linha que corta o gráfico possui um comportamento linear, ou seja, um comportamento de igual variância para os erros. No último gráfico, de Resíduos versus Leverage, temos informações suficientes para acreditar que não há pontos discrepantes que sejam muito influentes para o modelo, já que nenhum ponto ultrapassou os limites de 0,5 ou 1 de *Leverage*.

A tabela 20 indica que os modelos utilizando a base treino, e as bases das validações cruzadas tiveram valores similares das estatísticas MAPE, MSD e MAD, o que é um bom sinal pois estabelece que o modelo teve pouca alternância com a retirada sistemática de observações no processo de validação. Como o desvio absoluto médio (MAD) é expressa na mesma unidade dos dados, indica que o desvio absoluto médio vai na base de treino até a base de teste de 1% a 5%, o que é um bom resultado.

Tabela 20: Estatísticas de precisão do modelo de uso do solo para agricultura

Estatística	Base de treino	VClooo	VClfoo	Base de teste
MAPE	21,68%	32,06%	34,83%	104,56%
MAD	0,01123	0,01574	0,01672	0,04963
MSD	0,000343	0,0007102	0,0007483	0,016805

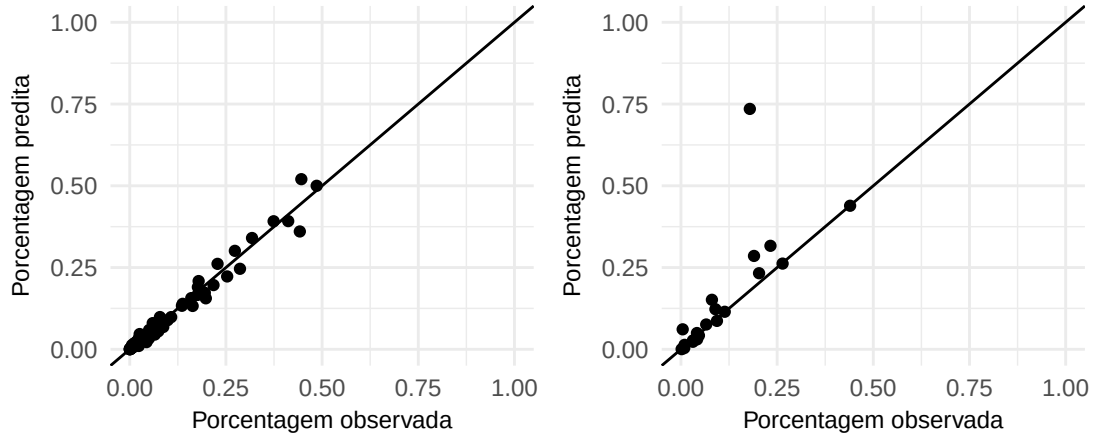
Já para a base de teste, é esperado que o valor destas estatísticas fossem um pouco maiores pois os dados utilizados são completamente diferentes daqueles usados nos outros modelos. No nosso problema, esses valores foram superiores ao esperado, como vemos que MAPE em 104% é um péssimo indicador da acurácia do modelo preditor.

Porém, com uma análise dos valores preditos, foi notado que duas observações somam 75% dos 104%. Em um dos casos o valor observado é 0,004494398 e o valor predito é 0,060813. Esse caso isolado equivale a 60% dentro dos 104% do indicador MAPE. A figura 7 sustenta nossa argumentação, uma vez que vemos que para a base de teste, as predições foram próximas aos valores reais tendo 3 pontos que não estão em concordância com suas estimativas.

A figura a seguir também revela que a maioria das observações se encontravam sobre a linha horizontal ou acima dela, portanto, temos a informação de que, no geral, os valores preditos estão super estimando o valor real. Acreditamos que estes resultados poderiam ser melhorados com a adição da variável *custo de fertilizantes* já que ela controlaria o efeito negativo de fertilizantes na variável explicada. O crescimento no preço de fertilizantes não segue a mesma velocidade que o crescimento de produção, pois devido à inércia da industria brasileira de fertilizantes, o Brasil se tornou refém de 3 industrias estrangeiras que conseguem controlar o mercado, (Saab e Paula, 2015). Portanto, temos a suposição

de que a ausência no modelo desta váriavel tenha sido o responsável por estes valores super estimados, o que motiva um estudo mais aprofundado e que apenas não fez parte do escopo deste trabalho em virtude de falta de fonte de dados confiáveis ou que possuíam estes dados de forma histórica.

Figura 6: Porcentagem observada versus porcentagem predita para base de treino e base de teste da agricultura, respectivamente



4.2.2 Uso do solo para pastagem

O modelo parcial de estimação de $\beta_k' \mathbf{X}(t,i)$ teve como resultado o exposto na tabela 21. Vemos que o *ano* para o caso de pastagem possui uma tendência negativa, ou seja, o valor esperado de y_{itPec} decresce em cerca de 1,6% a cada ano. Outra importância dos resultados refere-se ao fato de termos uma relação positiva entre o aumento da área esperada para pastagem e o aumento da área colhida para agricultura temporária aferida pelo valor positivo do parâmetro atribuído à última, o que é coerente com a realidade local uma vez que em algumas situações, parte da área colhida da agricultura temporária torna-se pastagem.

Por último, vemos que a variável relevante para quantificar os dados econômicos relativo à pecuária foi a variável *Efetivo da Pecuária*. Entretanto, foi aceito sob a ótica de 20% de significância. Acreditamos que poderíamos melhorar este resultado caso tivéssemos incluído no modelo dados de custos de produção agropecuária. Porém, seguindo a mesma linha de custos para agricultura, não encontramos dados confiáveis que pudéssemos utilizar.

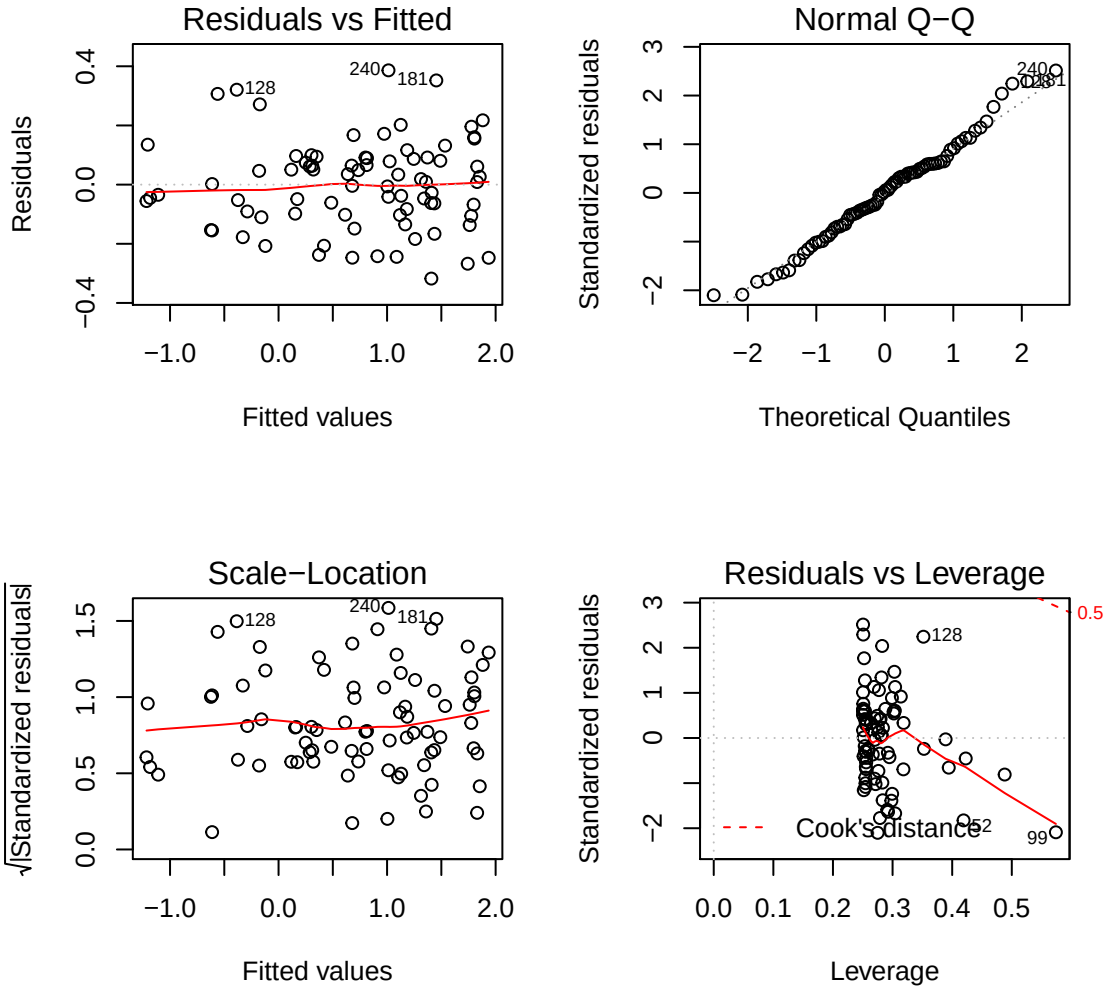
Tabela 21: Coeficientes e significância do modelo de regressão para o uso do solo para pastagem

Coeficiente	Estimativa	Significância
Intercepto	33,06	0,056201(.)
Arenápolis	75,13	***
Cláudia	-1,054	***
Confresa	0,02494	0,853672
Curvelândia	0,3144	0,050032(.)
Denise	0,3305	0,046572*
Feliz Natal	-2,392	***
Itanhanga	-0,4253	0,021048*
Itauba	-0,818	***
Matupá	-0,09557	0,518287
Nova Santa Helena	0,2898	0,040743*
Novo Mundo	-0,7369	0,000748***
Peixoto de Azevedo	-0,7123	***
Porto dos Gaúchos	-0,4496	0,014002*
Querência	-1,739	***
Sinop	-2,125	***
Terra Nova do Norte	0,6959	***
União do Sul	-0,9584	***
Nova Guarita	0,7705	***
Nova Maringá	-1,323	***
Ano	-0,01595	0,064533(.)
área colhida da agricultura temporária	0,00000278	0,070694(.)
Efetivo da pecuária	0,0000001521	0,186975

Nota: Aqueles que possuem 3, 2, 1 asterisco ou (.) têm significância de 0, 0,001, 0,01 e 0,05, respectivamente.

A figura de análise de resíduos cria informações suficientes para acreditarmos que os resíduos são aleatórios uma vez que não encontramos nenhum padrão sistemático no primeiro gráfico. No segundo quadrante, que mede se os resíduos são normais, verificamos que quase todas as observações se encontram sobre a linha diagonal. Portanto, não há evidências para contestar a normalidade dos desvios. O gráfico $Scale-Location$ mantêm a suposição de que os resíduos possuem mesma variância entre eles. Portanto, a necessidade de homoscedasticidade dos resíduos é satisfeita. O último quadrante esclarece que não há outlier na base de dados que seja muito influente para o modelo já que todas as observações encontram-se sob os limites de distância de Cook.

Figura 7: Gráficos do modelo de uso do solo para pastagem



Observa-se aqui, um ajuste muito melhor dos dados que no uso do solo anterior. O erro percentual absoluto médio (MAPE) para a base de teste e as validações cruzadas em torno de 9% é um ótimo indicador de que este modelo consegue explicar bem dados que não estavam incluídos dentro do modelo de treino. O MSD é uma estatística que é mais afetada por valores outliers. Quando este é maior que o MAD, pode ser um indicador de que há outliers. Pode-se ver o contrário neste caso.

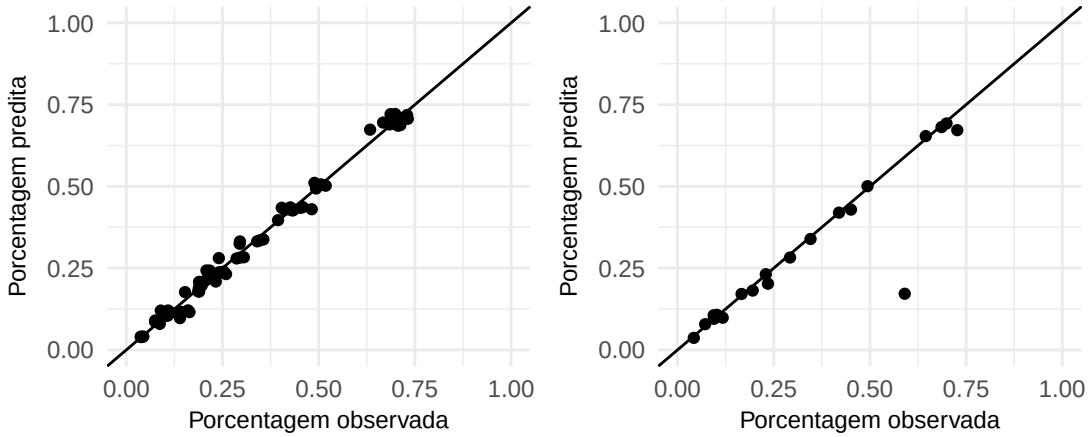
Tabela 22: Estatísticas de precisão do modelo de uso do solo para pastagem

Estatística	Base de treino	VClooo	VClfoo	Base de teste
MAPE	6,817%	9,74%	9,61%	8,698%
MAD	0,01517	0,02139	0,02194	0,032
MSD	0,0003873	0,0007755	0,0008199	0,009048

A figura 8 estabelece, no primeiro gráfico, a relação entre os valores preditos e os

reais para a base de treinamento, ou seja, mostra como o modelo treinado com os anos de treinamento (2004, 2008, 2010 e 2012) conseguiu estimar as observações dele mesmo, as observações da base de treinamento.

Figura 8: Porcentagem observada versus porcentagem predita para base de treino e base de teste da pastagem, respectivamente



O segundo gráfico proporciona a mesma análise alterando-se apenas a base de dados testada pelo modelo, que no caso, é a base de dados teste. Com exceção de apenas uma observação, todas se encontram próximas de seus valores reais.

4.2.3 Uso do solo para florestas

O modelo gerado para o uso *Floresta* evidenciou que a alteração da fronteira de florestas é mais diretamente relacionado com as variáveis que determinam o local da observação no tempo do que com as que determinam os fatores econômicos da realidade local.

É possível notarmos que a variável *ano* diz que com o aumento de uma unidade em seu valor, temos em média um decaimento de 2,36% na variável explicada, portanto, a sugestão de que com o passar do tempo, mantendo todas as outras variáveis constantes, temos um decréscimo no percentual de florestas nos municípios.

As duas últimas variáveis da tabela são as únicas variáveis econômicas que foram admitidas em influenciar, ainda que de maneira discreta, no valor esperado da variável explicada. Uma possível interpretação para a pouca quantidade de dados econômicos é em virtude da dificuldade encontrada em se contabilizar produtividade para este tipo de uso do solo, uma vez que o objetivo final da classe *floresta* não é gerar lucro como nos outros *shares*.

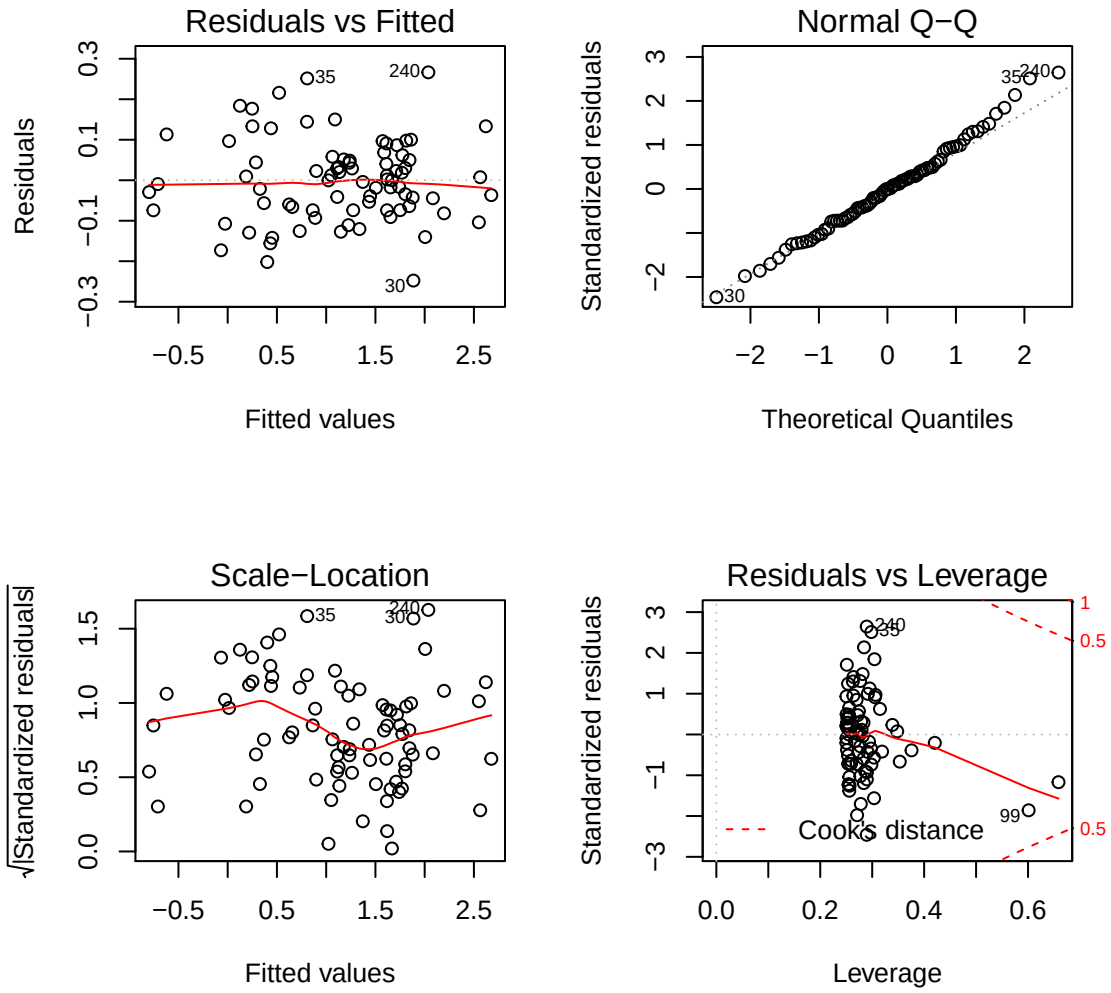
Tabela 23: Coeficientes e significância do modelo de regressão para o uso do solo para florestas

Coeficiente	Estimativa	Significância
Intercepto	48,43	***
Arenápolis	-1,115	***
Cláudia	0,5519	***
Confresa	-0,4781	***
Curvelândia	-1,841	***
Denise	-0,9542	***
Feliz Natal	0,4689	***
Itanhanga	0,2916	0,006585**
Itauba	-0,05735	0,528156
Matupá	0,6083	***
Nova Santa Helena	0,2905	0,001339**
Novo Mundo	-0,4033	0,004716**
Peixoto de Azevedo	0,5828	***
Porto dos Gaúchos	0,5349	***
Querência	0,08824	0,325884
Sinop	-0,6498	***
Terra Nova do Norte	-0,7185	***
União do Sul	0,9759	***
Nova Guarita	-0,8237	***
Nova Maringá	0,4065	***
Ano	-0,02356	***
área colhida da agricultura temporária	0,000003226	0,002418**
Valor de produção de origem animal	0,000005917	0,064598(.)

Nota: Aqueles que possuem 3, 2, 1 asterisco ou (.) têm significância de 0, 0,001, 0,01 e 0,05, respectivamente.

Assim como para *Pastagem* e *Agricultura*, e seguindo a mesma lógica de análise, vemos que os gráficos de resíduos criam informações suficientes para acreditarmos que os resíduos são aleatórios, normalmente distribuídos, com igual variância e sem observações que influenciam em grande magnitude os resultados do modelo.

Figura 9: Gráficos do modelo de uso do solo para florestas



O ajuste do uso *floresta* é o melhor dos três casos para esta seção, possui os menores valores de MAPE, não indicando existência de valores outliers, pelo fato do MSD ser bem menor que o MAD, o modelo consegue explicar muito bem o observado.

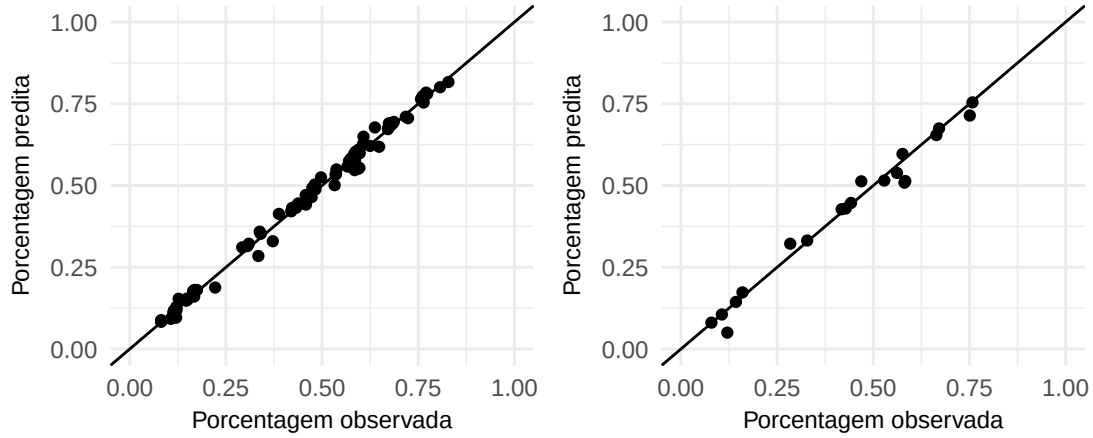
Tabela 24: Estatísticas de precisão do modelo de uso do solo para floresta

Estatística	Base de treino	VClooo	VClfoo	Base de teste
MAPE	3,794%	5,26%	5,24%	6,989%
MAD	0,01316	0,01834	0,01772	0,02208
MSD	0,0003152	0,0006308	0,00059	0,001067

A figura 10 estabelece, no primeiro gráfico, a relação entre os valores preditos para a base de treinamento e seus valores reais, enquanto que no segundo gráfico temos a mesma relação, alterando-se apenas para a base teste. Vemos que em ambos os gráficos a variabilidade dos pontos ao redor da reta horizontal é pequena e não temos nenhum

ponto que esteja muito longe das mesmas.

Figura 10: Porcentagem observada versus porcentagem predita para base de treino e base de teste da floresta, respectivamente



4.3 Modelo com dados econômicos e biofísicos

O modelo analisado nesta seção busca livrar-nos da necessidade do conhecimento do efeito fixo gerado pela variável *Código de Município do IBGE* alterando-a para um conjunto de variáveis que determinem a realidade sócio-espacial da observação.[5]

Atualmente, em virtude dos avanços em tecnologia dos fertilizantes e defensivos agrícolas é possível se plantar nos mais diversos tipos de solo, apesar de solos mais férteis serem preferidos. Portanto, temos a suposição de que as regiões escolhidas dependem de fatores como topografia e declividade (no caso de agricultura, procura-se regiões planas contínuas) e fatores que medem o clima local, como temperatura e presença de classes de clima de seca ao longo do ano.

4.3.1 Uso do solo para agricultura

Os resultados das estimativas para os coeficientes da regressão parcial para Agricultura podem ser vistos na tabela 25, é possível a visualização de que, com exceção da variável *Valor de produção da agricultura temporária*, todas as variáveis forem aceitas sob a ótica de 5% de significância.

A tabela 25 também deixa claro a magnitude de interferência de certas variáveis no modelo, como podemos ver que o aumento percentual de regiões com secas de 1 a 2 meses e temperatura muito elevadas atuam negativamente na variável explicada, enquanto

[5]Foram encontrados dados biofísicos para 117 dos 141 municípios, o que restringiu este modelo a ter menos municípios que o outro.

Tabela 25: Coeficientes e significância do modelo de regressão para o uso do solo para agricultura

Coeficiente	Estimativa	Significância
Intercepto	-163,7	***
Ano	0,07846	***
área plantada da agricultura permanente	0,001103	***
área colhida da agricultura permanente	-0,000933	0,016503*
Produção da agricultura permanente	-0,0001111	0,032169*
Produção da agricultura temporária	0,000003516	***
Valor de produção da agricultura temporária	0,00000154	0,073892(.)
Produção da extração vegetal	-0,0005052	0,001197**
Valor de produção da extração vegetal	0,0007993	0,001593**
Produção da extração vegetal em m^3	-0,000007374	0,003204**
Valor de produção da extração vegetal em m^3	0,0000533	0,007912**
Produção da silvicultura em m^3	-0,000005625	0,001379**
População urbana em 2010	-0,0000271	***
IDHM em 2010	48,63	***
IDHM da educação em 2010	-19,65	***
Seca entre 1 a 2 meses	-631,1	***
Seca por 3 meses	3,429	***
Temperatura quente	-21,02	***
Topologia montanhosa a escarpada	1,339	***
Fertilidade Alta	18,71	***

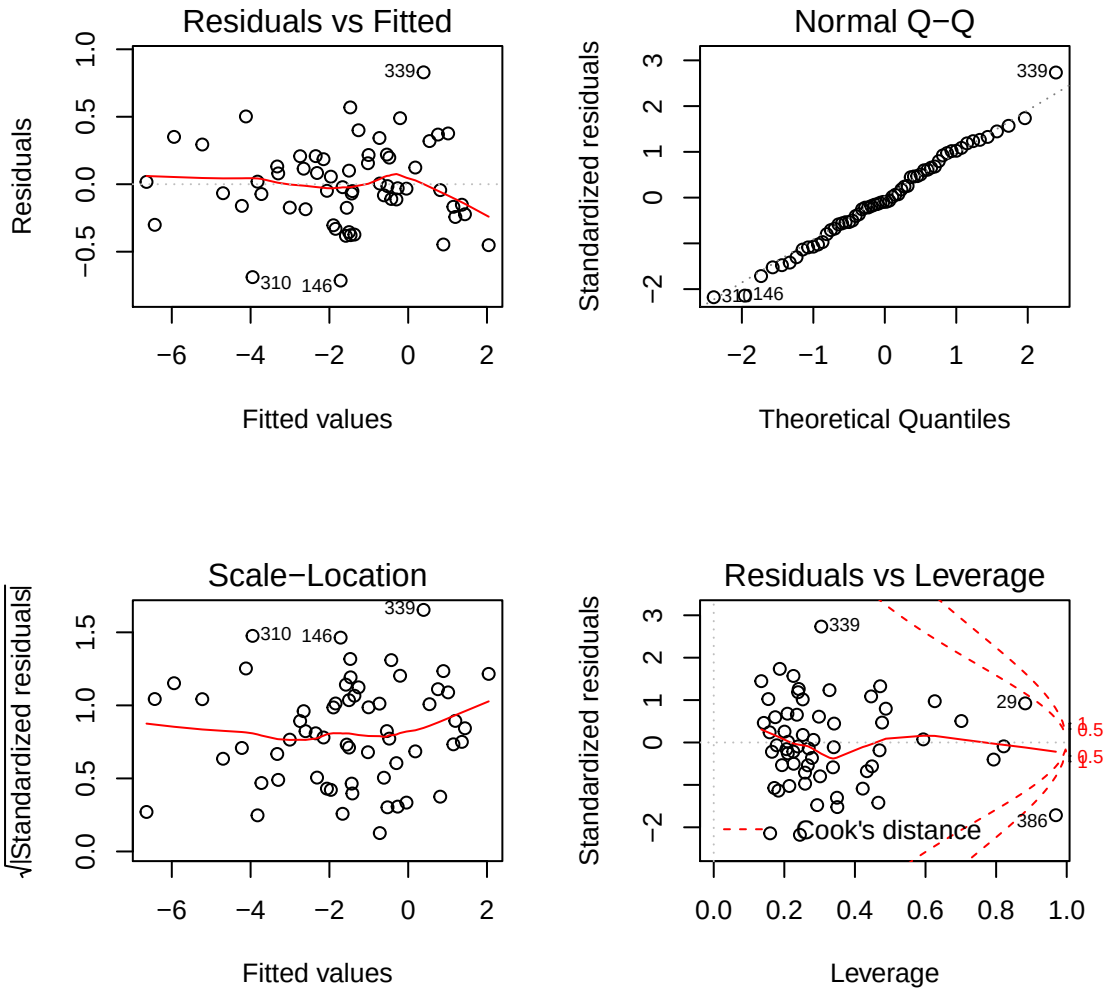
Nota: Aqueles que possuem 3, 2, 1 asterisco ou (.) têm significância de 0, 0,001, 0,01 e 0,05, respectivamente.

regiões com secas de 3 meses e regiões de fertilidade alta atuam de maneira positiva no resultado final.

A análise de resíduos para esta regressão, como podemos ver na figura a seguir, evidenciou que os resíduos não apresentam um comportamento sistêmico em relação aos valores preditos, possuem um comportamento normal e não quebram o pressuposto de homoscedasticidade. Porém, o último gráfico evidenciou que duas observações estão sendo muito influentes para o modelo. São elas, observações dos municípios Sinop e Feliz Natal, ambos da microrregião de Sinop.

Um estudo mais aprofundado do ambiente ecônomico desta microrregião ressalva que ela tem o seu crescimento atrelado à decadência da atividade mineradora na região, e por este motivo, em virtude da escassez de minérios e a sua localização estratégica no centro do estado de Mato Grosso, esta microrregião se tornou uma região de polarização urbana, referência na área de serviços, e no setor agroindustrial com empresas processadoras/esmagadoras de arroz e de soja. Portanto, temos a sugestão de que os dois municípios desta região não se adaptaram bem ao modelo, justamente por ter uma realidade histórico-cultural diferente dos demais.

Figura 11: Gráficos do modelo de uso do solo para agricultura



Neste ajuste conseguimos ver estatísticas piores que no modelo de dados econômicos e código do IBGE para município do uso do solo de agricultura. Como descrito na seção 4.2.1, explicações para esse possível ajuste ruim também se aplicam aqui. A média percentual absoluta do erro (MAPE) é elevado por duas observações muito discrepantes.

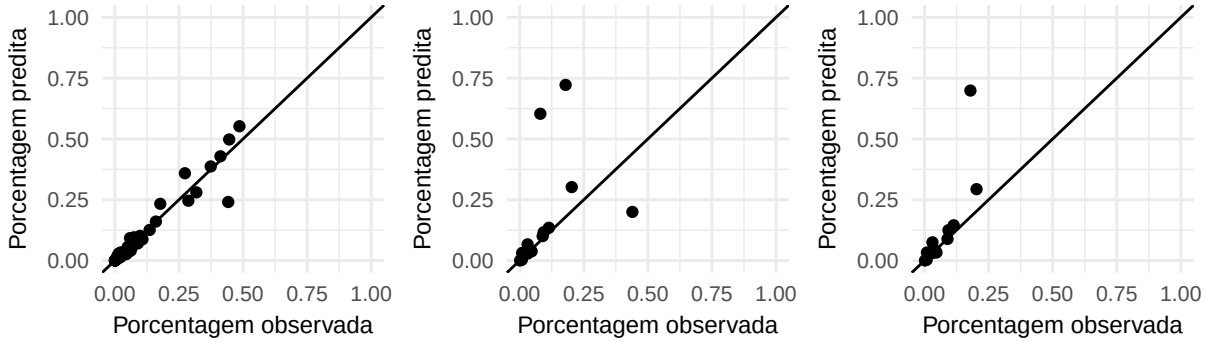
Tabela 26: Estatísticas de precisão do modelo de uso do solo para agricultura

Estatística	Base de treino	VClooo	VClfoo	Base de teste
MAPE	25,33%	51,42%	52,02%	108,41%
MAD	0,01513	0,03238	0,03359	0,1024
MSD	0,001112	0,01025	0,01069	0,04257

No tocante à figura de predição para os dados de 2014, torna-se evidente que os municípios de Sinop e Feliz Natal não se adequaram bem ao modelo, sendo os responsáveis pelos grandes desvios entre os valores preditos e os reais. Já com a ausência destes dois na

modelagem, vemos que apenas o município de Denise não se adequou mas não encontramos um motivo latente que levasse uma análise especial para este caso.

Figura 12: Porcentagem observada versus porcentagem predita para base de treino, base de teste e base de teste sem 2 municípios da agricultura, respectivamente



A média dos desvios para as predições do modelo foi 5,33% quando incluímos Denise na análise, mas sem ela, os desvios caem para 1,14% em média de diferença entre o real e o esperado para agricultura.

4.3.2 Uso do solo para pastagem

A modelagem para pecuária teve as variáveis da tabela 27 como relevantes. Com exceção de *Produção da extração vegetal* e de *Temperatura quente*, todas as variáveis foram aceitas sob a ótica de 5% de significância. As variáveis *IDHM 2010* e *topologia plana a ondulada* são as mais positivamente correlacionadas com a variável explicada, sendo que os $IDHMs$ parciais e *Fertilidade Alta* são as mais negativamente correlacionadas.

Tabela 27: Coeficientes e significância do modelo de regressão para o uso do solo para pastagem

Coeficiente	Estimativa	Significância
Intercepto	46,12	***
Produção da agricultura temporária	-0,0000005784	***
Produção da extração vegetal	0,00002794	0,130332
População rural em 2010	-0,00008397	***
População urbana em 2010	0,00003365	***
IDHM em 2010	359,5	***
IDHM da educação em 2010	-134,8	***
IDHM da longevidade em 2010	-135,0	***
IDHM da renda em 2010	-152,8	***
Seca por 3 meses	-0,4503	0,021226*
Temperatura quente	1,301	0,160004
Topologia montanhosa a escarpada	-0,6663	0,00229**
Topologia plana a ondulada	39,53	***
Fertilidade Alta	-3,679	***

Nota: Aqueles que possuem 3, 2, 1 asterisco ou (.) têm significância de 0, 0,001, 0,01 e 0,05, respectivamente.

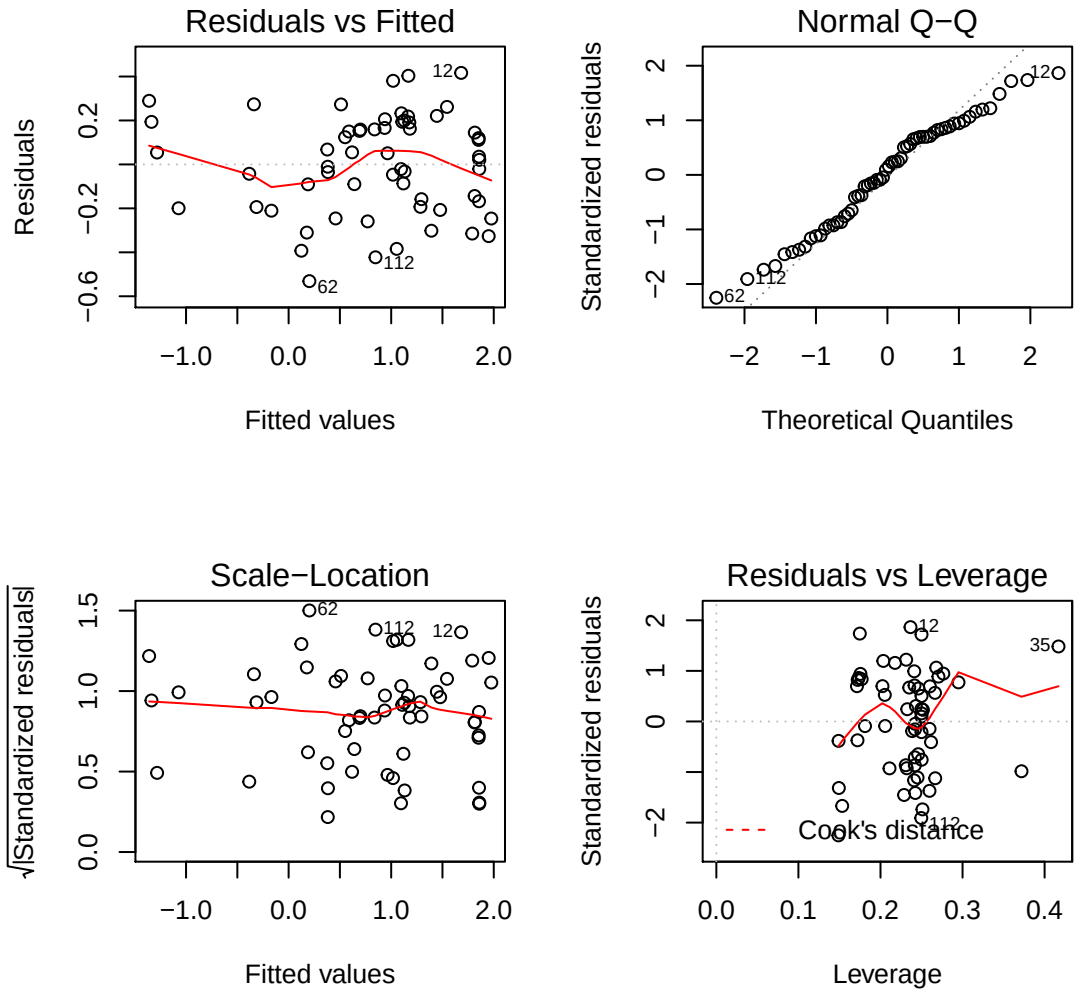
As médias percentuais absolutas neste caso, mostram um bom ajuste, não tanto quanto no caso anterior da seção 4.3.2 e média dos desvios em torno de 2,5% a 5%, indicam ser um bom indicador. As estatísticas não defendem valores discrepantes.

Tabela 28: Estatísticas de precisão do modelo de uso do solo para pastagem

Estatística	Base de treino	VClooo	VClfoo	Base de teste
MAPE	12,02%	16,39%	16,75%	23,11%
MAD	0,02512	0,03214	0,03350	0,05357
MSD	0,001156	0,001789	0,001971	0,01236

O diagnóstico dos resíduos está na figura abaixo, onde vemos que os resíduos possuem comportamento aleatório e normal, admitindo o pressuposto de homoscedasticidade e sem valores muito influentes para o modelo.

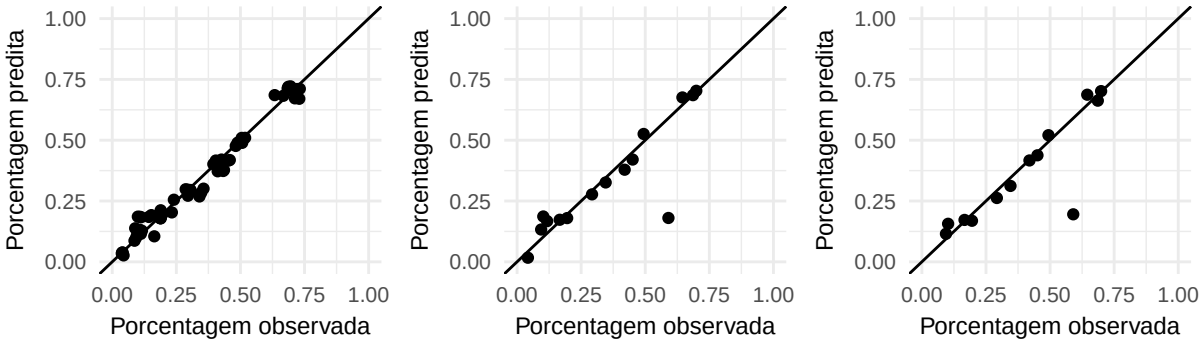
Figura 13: Gráficos do modelo de uso do solo para pastagem



Análise das estimações dos modelos indica novamente o erro de estimação para o município Denise, uma vez que mesmo após a remoção de Sinop e de Feliz Natal no modelo, ainda assim, o erro de estimação para Denise se faz presente em grande magnitude. Entretanto, para os demais municípios, o erro não foi grande pois a maioria das observações se encontram ao redor das retas diagonais dos gráficos.

Para o caso de pastagem, a média dos desvios foi de 6,37% para o modelo que inclui Sinop e Feliz Natal onde diminuiu para 3,90% se tirarmos a estimativa para Denise. Já no modelo que não inclui Sinop e Feliz Natal, a média dos desvios foi de 5,84% e com a remoção da estimativa para Denise, tivemos um valor igual a 2,92%.

Figura 14: Porcentagem observada versus porcentagem predita para base de treino, base de teste e base de teste sem 2 municípios da pastagem, respectivamente



4.3.3 Uso do solo para florestas

As váriáveis explicativas do modelo de uso para florestas na tabela 29, refletem que a porcentagem de floresta no município está inversamente relacionada com variáveis de produção de agricultura, representando este avanço da fronteira agrícola em relação as florestas, ainda que de uma maneira discreta, medida pelas estimativas delas.

Tabela 29: Coeficientes e significância do modelo de regressão para o uso do solo para floresta

Coeficiente	Estimativa	Significância
Intercepto	-110,2	***
área plantada da agricultura temporária	-0,000005065	***
Produção da agricultura temporária	-0,0000003556	***
População rural em 2010	-0,0001081	***
IDHM da longevidade em 2010	81,1	***
Seca entre 1 a 2 meses	-84,3	***
Temperatura quente	1,023	0,058601(.)
Topologia montanhosa a escarpada	44,86	***
Fertilidade alta	43,74	***
Fertilidade baixa	45,44	***

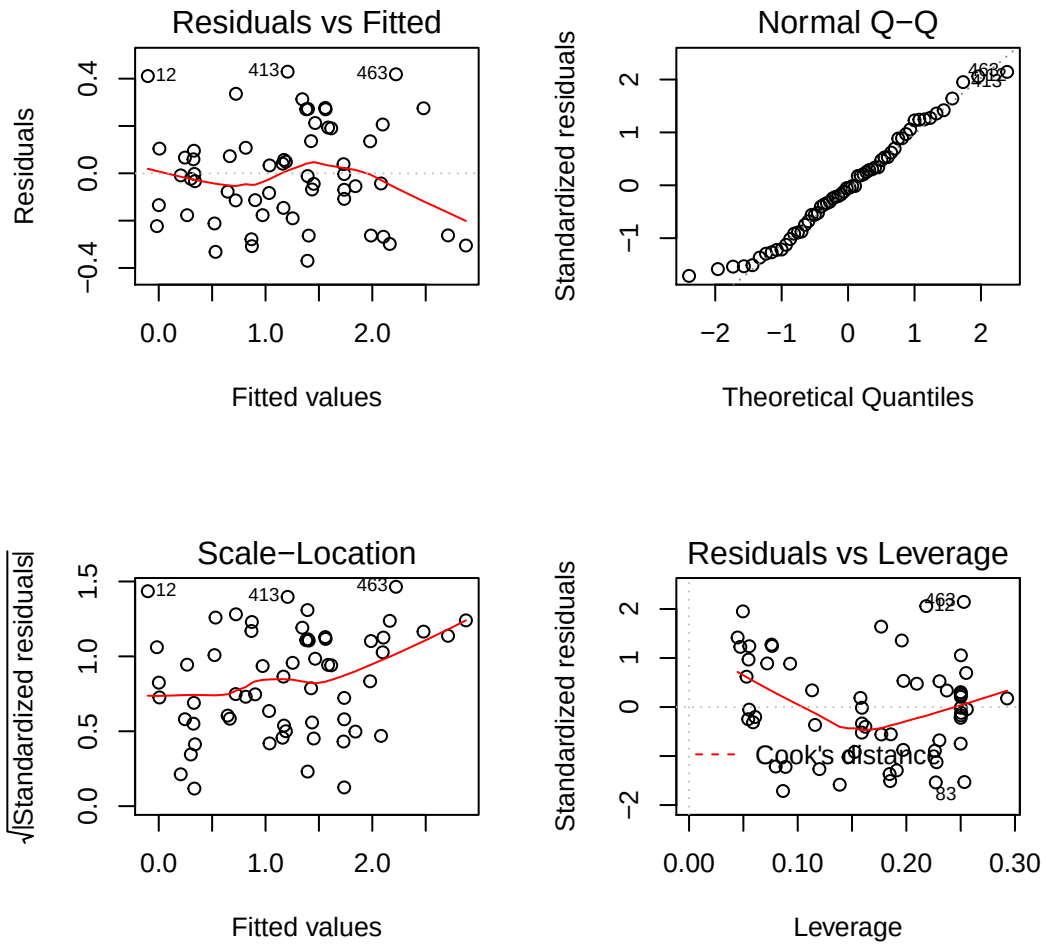
Nota: Aqueles que possuem 3, 2, 1 asterisco ou (.) têm significância de 0, 0,001, 0,01 e 0,05, respectivamente.

As variáveis que descrevem o ambiente, ao contrário das variáveis econômicas, refletiram, no geral, uma associação positiva com a variável explicada. Vemos que as regiões com maiores esperanças de vida ao nascer refletido pelo *índice de Desenvolvimento Humano Municípal para Longevidade* aumentam o valor esperado da porcentagem de florestas no município, bem como regiões com topologia montanhosa a escarpada, regiões com temperatura considerada quente, regiões com fertilidade alta e regiões com fertilidade baixa. A única variável de classe biofísica que interfere negativamente no resultado esperado, é

a variável *Seca de 1 a 2 meses*, ou seja, é esperado que os municípios diminua o número de florestas à medida que aumenta o número de regiões que possuem seca de 1 a 2 meses no ano.

O diagnóstico dos resíduos pode ser encontrado através da figura 13, onde vemos que os resíduos aparentam ter um comportamento aleatório, normal e não temos no modelo nenhuma observação muito influente, pois todas encontram-se dentro dos limites de distância de Cook.

Figura 15: Gráficos do modelo de uso do solo para floresta



O único gráfico que merece uma atenção especial é o gráfico *scale − location* que averigua o pressuposto de homoscedasticidade do modelo, uma vez que a linha horizontal está levantada em sua cauda direita. Entretanto, nota-se que apenas 3 observações compõem o final da cauda enquanto o resto encontra-se espalhado no começo ou meio da linha, logo, acreditamos que tenha sido em virtude da variabilidade dos dados e não temos informações suficientes para rejeitarmos a hipótese de igual variância.

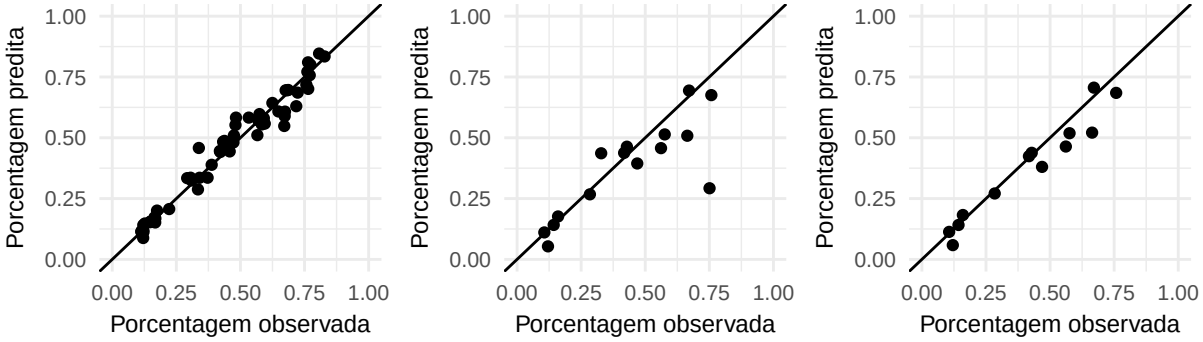
Vemos boas estimativas nesta tabela, assim como nos casos anteriores, do uso para *pastagem* deste modelo e do anterior e *floresta* do modelo anterior, há um bom ajuste.

Tabela 30: Estatísticas de precisão do modelo de uso do solo para floresta

Estatística	Base de treino	VClooo	VClfoo	Base de teste
MAPE	7,45%	10,29%	10,88%	17,75%
MAD	0,03088	0,04546	0,04678	0,08195
MSD	0,001764	0,00795	0,008218	0,01867

Na figura 16, vemos que as estimativas estão próximas dos valores reais quando o modelo estima suas próprias observações, quando ele estima os valores para 2014 incluindo Sinop e Feliz Natal, e quando ele estima os valores para 2014 excluíndo estes municípios da modelagem. É possível notarmos que não há ponto discrepante, apesar da variabilidade das estimativas ao redor da reta diagonal do uso *floresta* terem sido maiores que para os outros usos.

Figura 16: Porcentagem observada versus porcentagem predita para base de treino, base de teste e base de teste sem 2 municípios da floresta, respectivamente



5 Considerações Finais

O estudo deste trabalho teve origem na necessidade de conhecer a realidade espaço-temporal de determinados municípios, e assim, compreender melhor os fatores que interferem nas escolhas dos usos da terra para, posteriormente, utilizar os resultados para auxiliar na criação de políticas públicas de meio ambiente.

Primeiramente, a análise feita evidenciou que é possível trabalhar com os municípios em uma base de dados agregada e também tornou claro que os efeitos fixos de municípios podem ser controlados por variáveis que indiquem a estrutura biofísica dos mesmos, ou seja, podemos quantificar a classe dada pelo município i e, assim, tornar o modelo independente do código do IBGE.

Segundo, os resultados para agricultura demonstraram que este uso é dependente de variáveis econômicas mas, principalmente, de alterações no ambiente biofísico e social. Portanto, uma mudança neste uso deve ser pensada através de uma mudança na realidade espacial, seja alterando o espaço (como por exemplo, planificar alguma região), seja influenciando em alterações sociais (como por exemplo, aumentando o IDHM de alguma região). Para pecuária, concluimos que este uso está relacionado com a variação de agricultura temporária, não possui uma variação grande de ano à ano e está ligado com o espaço, uma vez que são preferidas para pastagem regiões planas a montanhosas, em detrimento de regiões montanhosas a escarpadas. Também prefere-se regiões quentes e com fertilidade baixa. Já em floresta, constatamos que este uso está relacionado com variáveis econômicas dos demais usos de maneira discreta, mas principalmente relacionado com a realidade espacial, uma vez que os fatores que mais influenciaram os valores esperados dos percentuais de floresta foram fatores como temperatura, o tipo de seca, o tipo de topografia e o tipo de fertilidade.

Por último, para atingir o objetivo final do trabalho que é o de auxiliar na criação de políticas públicas para o Meio Ambiente, entendemos, do ponto de vista econômico, que estas políticas não devem ser pautadas em uma tentativa de mudança nos valores de produção para a pecuária, mas em mudanças nas variáveis de produção para agricultura temporária. Uma possível solução seria criar incentivos para a diminuição de habitantes no meio rural, causando uma baixa da mão-de-obra no campo e polarizando regiões antes rurais. Sob a ótica da natureza biofísica do local, deduzimos que políticas públicas de meio ambiente devem ser criadas no sentido de se preservar as regiões que possuem características biofísicas propícias para florestas.

Referências

ADAMI, M., ALMEIDA, C., COUTINHO, A., DESSAY, N., DINIZ, C., DURIEUX, L., ESQUERDO, J. C. D. M., GOMES, A., and VENTURIERI, A. (2016). **High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data.** *Acta Amazonica*, 46(3):291–302.

ALBUQUERQUE, P. H. M., BASSO, G. G., CARVALHO, A. X. Y. d., GUIMARÃES, L. F. D., LAURETO, C. R., MOREIRA, G. C. C., and PENA, M. G. (2017). **Clusterização espacial e não espacial: um estudo aplicado à agropecuária brasileira**. *Texto para discussão*, 2279:18–19.

ARRAES, R. d. A., MARIANO, F. Z., and SIMONASSI, A. G. (2012). **Causas do desmatamento no Brasil e seu ordenamento no contexto mundial.** *Revista de Economia e Sociologia Rural*, 50(1):119–140.

AY, J. S., CHAKIR, R., and LE GALLO, J. (2014). **Individual vs. aggregate models of land use changes: Using spatial econometrics to improve predictive accuracy.** Statistics seminar, Toulouse: TSE, april 1.

BACHA, C. J. C. and BRUGNARO, R. (2009). **Análise da participação da agropecuária no PIB do Brasil de 1986 a 2004.** *Estudos Econômicos*, 39(1):127–159.

BAKONYI, S. M. C. and DE DEUS, R. M. (2012). **O impacto da Agricultura sobre o Meio Ambiente.** *Reget*, 7(7):1306–1315.

CHAKIR, R. (2009). **Land use: econometric methods and modelings issues.** *UMR Économie Publique*.

HARDIE, I. W. and PARKS, P. J. (1997). **Land use with heterogenous land quality: an application of an area base model**. *American Journal of Agricultural Economics*, 79(2):299–310.

LE GALLO, J. and PLANTINGA, A. J. (2011). **Predicting land use allocation in france: a spatial panel data analysis**. *UMR Économie Publique*.

MILLER, D. J. and PLANTINGA, A. J. (1999). **Modeling land use decisions with aggregate data**. *American Journal of Agricultural Economics*, 81(1):180–194.

PAULA, R. and SAAB, A. (2015). **O mercado de fertilizantes no Brasil diagnósticos e propostas de políticas.** *Revista de Política Agrícola*, 17.

PLANTINGA, A. J. (1996). **The effect of agricultural policies on land use and environmental quality**. *American Journal of Agricultural Economics*, 78(4):1082–1091.

Anexos

A.1 Análise dos dados para os municípios usados no modelo da seção 4.3

Tabela 31: Análise descritiva dos dados econômicos usados no modelo da seção 4.3

Variável	Mínimo	Máximo	Média	Desvio Padrão
área plantada agricultura temporária	170	231088	49501,12	57214,56
área colhida agricultura temporária	170	231088	49399,28	57241,39
Produção agrícola temporária	2200	2181534	378228,71	574806,97
Valor da produção agrícola temporária	R$ 902,00	R$ 703387,00	R$ 100527,99	R$ 127656,66
área plantada agricultura permanente	0	4674	531,91	935,59
área colhida agricultura permanente	0	4143	475,77	847,58
Produção agrícola permanente	0	12000	1330,55	2223,62
Valor da produção agrícola permanente	R$ 0	R$ 15435,00	R$ 1882,65	R$ 2931,44
Produção da extração vegetal*	0	11666	944,85	2348,50
Valor da produção da extração vegetal*	R$ 0	R$ 5833,00	R$ 524,68	R$ 1266,58
Produção da extração vegetal**	465	264726	40677,36	55725,30
Valor da produção da extração vegetal**	R$ 26,00	R$ 49188,00	R$ 4223,96	R$ 8248,07
Produção da silvicultura	0	313000	7519,27	36906,05
Valor da produção da silvicultura	R$ 0	R$ 22200,00	R$ 569,27	R$ 2631,13
Efetivo de animais da pecuária	17420	3774433	598144,29	820334,24
Valor da produção de origem animal	R$ 169,00	R$ 64005,00	R$ 6895,15	R$ 11181,98

Fonte: Instituto Brasileiro de Geografia e Estatística
*Produção em toneladas, ** metros cúbicos

A.2 Código para criação das bases de dados

```
##-----------------------------------------------------
## Subindo bases
##-----------------------------------------------------

uso_da_terra_mt <- read.delim("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/planilha
agricola_temporaria <- read.delim("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/plan
agricola_permanente <- read.delim("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/plan
floresta_exve <- read.delim("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/planilhas
floresta_silv <- read.delim("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/planilhas
pecuaria <- read.delim("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/planilhas modif

##-----------------------------------------------------
## Juntando bases, transformando unidade de km2 para hectare em algumas variáveis
## e tirando variáveis repetidas
##-----------------------------------------------------

dados <- cbind(uso_da_terra_mt, agricola_permanente, agricola_temporaria,
               floresta_exve, floresta_silv, pecuaria)
dados[,2:5] <- 100*dados[,2:5]
dados <- dados[,-c(11, 16, 17, 22, 23, 28, 29, 32, 33, 36)]

##-----------------------------------------------------
## Criando arquivo csv
##-----------------------------------------------------

write.csv(dados, 'C:/Users/igor/Google Drive/TCC/Dados/mato grosso/planilhas modificadas

##-----------------------------------------------------
## Subindo dados biofísicos
##-----------------------------------------------------

load("C://Users//igor//Google Drive//TCC//Programação//solos.Rda")
solos[, 8:56] <- apply(solos[, 8:56], 2, function(x) as.numeric(as.character(x)))
solos <- solos[, -c(2, 3, 4)]
amc_mun <- read.delim("C:/Users/igor/Google Drive/TCC/Dados/dados gerais/amc_mun.txt")
base <- merge(amc_mun, solos, by.x = 'codamc97', by.y = 'codamc97')
```

```
##-----------------------------------------------------
## Base
##-----------------------------------------------------

dados2 <- dados

##-----------------------------------------------------
## Indexando base solos com a base de dados de produção
##-----------------------------------------------------

index <- match(agricola_permanente$codigo, base$UFMUNDV)
index2 <- na.omit(index)
index3 <- which(index != is.na(index))
dados2[, 27:78] <- NA
dados2[index3, 27:78] <- base[index2, 4:55]

##-----------------------------------------------------
## Renomeando colunas dos dados biofísicos dentro da base dados2
##-----------------------------------------------------

nomes <- colnames(base[, 4:55])
nomes2 <- colnames(dados2)
nomes2[27:78] <- nomes
colnames(dados2) <- nomes2

##-----------------------------------------------------
## Criando CSV
##-----------------------------------------------------

dados2 <- na.omit(dados2)
write.csv(dados2, 'C:/Users/igor/Google Drive/TCC/Dados/mato grosso/planilhas modif
```

A.3 Código para modelagem

```
dados <- read.csv2(file='/Users/Gustavo/Downloads/dados.csv', sep=',')
```

```
# dados.trainingR1 <- dados.training[ dados$codigo %in% listamuniciNorte, ]
# dados.trainingR2 <- dados.training[ dados$codigo %in% listamuniciNordeste, ]
# dados.trainingR3 <- dados.training[ dados$codigo %in% listamuniciSul, ]
# dados.trainingR4 <- dados.training[ dados$codigo %in% listamuniciMeiomeio, ]
# dados.trainingR5 <- dados.training[ dados$codigo %in% listamuniciMeionorte, ]
#
# dados.trainingR1$subreg <- 'norte'
# dados.trainingR2$subreg <- 'nordeste'
# dados.trainingR3$subreg <- 'sul'
# dados.trainingR4$subreg <- 'meiomeio'
# dados.trainingR5$subreg <- 'meionorte'

# dados.training <- rbind(dados.trainingR1, dados.trainingR2, dados.trainingR3, dados.tr



for(i in 3:10){
  dados[,i] <- as.numeric(as.character(dados[,i]))
}

dados[dados == 0] <- 0.0000000001

#normalizando tendo outros como referencia:

dados$lnagr <- log(dados$agr.porcentagem/dados$out.porcentagem)
dados$lnpec <- log(dados$pas.porcentagem/dados$out.porcentagem)
dados$lnflo <- log(dados$flo.porcentagem/dados$out.porcentagem)


#setando base de treinamento

dados.training <- dados[dados$ano != 2014,]

dados.test <- dados[dados$ano == 2014,]

###### Retirando as observacoes discrepantes em virtude
###### de dados mal coletados para agricultura ou pouca utilizacao
###### de agricultra no municipio e observacoes onde outros
```

```
###### eh maior que 20% do total de uso da terra para o municipio

training <- dados.training[ dados.training$lnagr > -15,]

training <- training[order(training$codigo),]

for(i in 1:length(training$out.porcentagem)){
if(training$out.porcentagem[i] < 0.2){
  training$d2[i] <- 0
  } else training$d2[i] <- 1
}

v3 <- NULL
var <- NULL
for(i in 1:length(training$out.porcentagem)){
  if(training$d2[i] == 1){
   var <- training$codigo[i]
   v3 <- c(var,v3)
  }
}

v3 <- unique(v3)
v3 <- c(v3,5106299)

training <- training[! training$codigo %in% v3,]

hist(training$lnagr)

# Ipiranga do norte - 5104526
# Nova Ubirata - 5106240
# Querencia - 5107065
# Sinop - 5107909
# Sorriso - 5107925

# trainIpi <- dados.training[dados.training$codigo == 5104526,]
# trainNov <- dados.training[dados.training$codigo == 5106240,]
# trainQue <- dados.training[dados.training$codigo == 5107065,]
# trainSin <- dados.training[dados.training$codigo == 5107909,]
# trainSor <- dados.training[dados.training$codigo == 5107925,]
```

```
#
# training <- rbind(trainIpi, trainNov, trainQue, trainSin, trainSor)
#
# hist(dados.training$lnagr)
# trainAgr <- dados.training[,c(28,2,11,14,15,18,19)]

###################################
####### Minimos Quadrados #########
###################################

#funcoes de regressao utilizando Quadrados Minimos para os tres casos:

### selecionando apenas os casos para o teste set onde temos
### informacao completa para todos os 4 anos

teste <- as.data.frame(table(training$codigo))
teste <- teste[ teste$Freq == 4,]

training <- training[training$codigo %in% teste$Var1,]

############## M o d e l o s ####################

training$codigo <- as.character(training$codigo)

# training$ano <- as.character(training$ano)

# test.set <- training[training$ano == 2014,]
# training <- training[training$ano != 2014,]

# listm <- read.csv('/Users/Gustavo/Downloads/munnorte.csv')
#
# training <- training[training$codigo %in% listm$V2,]



RegAgr <- lm(lnagr ~ . , data = training[c(2,11:28)])

pec <- as.formula(paste('lnpec ~', paste(colnames(training)[c(2,11:27)], collapse='+')))
flo <- as.formula(paste('lnflo ~', paste(colnames(training)[c(2,11:27)], collapse='+')))
```

```
##n <- names(training)[c(2,11:27)]
##pec <- as.formula(paste('lnpec ~ ', paste(n, collapse = ' + ')))
##RegPec <- lm(pec, data = subset(training, select = c("lnpec", n)))

RegPec <- lm(pec, data = training[c(2, 11:27, 29)])
RegFlo <- lm(flo, data = training[c(2,11:27,30)])

##### StepWise ######

library(MASS)
stepAgr <- stepAIC(RegAgr, direction="both")

summary(stepAgr)

stepPec <- stepAIC(RegPec, direction="both")
stepFlo <- stepAIC(RegFlo, direction="both")

summary(stepFlo)


######## Modelos selecionados no stepWise #######

RegAgr <- lm(lnagr ~ . , data = training[c(2,11,12,13,18,28)])

pec <- as.formula(paste('lnpec ~', paste(colnames(training)[c(2,11,17,26)], collaps
flo <- as.formula(paste('lnflo ~', paste(colnames(training)[c(2,11,17,27)], collaps

###COLOCAR ESSE TRAINING[C(2,11,17,26,29)]
RegPec <- lm(pec, data = training[c(2,11,17,26,29)])
RegFlo <- lm(flo, data = training[c(2,11,17,27,30)])

plot(RegAgr)
plot(RegPec)
plot(RegFlo)

###### Outliers ############

outAgr <- training[ rownames(training) %in% c(258,32,117),]
```

```
outPec <- training[ rownames(training) %in% c(139),]
outFlo <- training[ rownames(training) %in% c(141,123),]

#113 (houve uma grande variabilidade entre os anos para os valores da variÃ¡vel a ser ex

outGeral <- training[ rownames(training) %in% c(258,32,117,113,124,67,65, 141, 123, 139)

ta <- training[training$codigo %in% outGeral$codigo,]


####### Modelos sem os outliers ########

training <- training[!training$codigo %in% outGeral$codigo,]

RegAgr <- lm(lnagr ~ . , data = training[c(2,11,12,13,18,28)])

pec <- as.formula(paste('lnpec ~', paste(colnames(training)[c(2,11,17,26)], collapse='+'
flo <- as.formula(paste('lnflo ~', paste(colnames(training)[c(2,11,17,27)], collapse='+'

RegPec <- lm(pec, data = training[c(2,11,17,26,29)])
RegFlo <- lm(flo, data = training[c(2,11,17,27,30)])

plot(RegAgr)
plot(RegPec)
plot(RegFlo)

#Calculando as estimativas no logito:

exponenciais <- data.frame(exp(RegAgr$fitted.values),
                           exp(RegPec$fitted.values),
                           exp(RegFlo$fitted.values),
                           exp(RegAgr$fitted.values) +
                           exp(RegPec$fitted.values) +
                           exp(RegFlo$fitted.values) + 1)

exponenciais2 <- NULL
exponenciais2$agr <- exponenciais$exp.RegAgr.fitted.values. /   exponenciais$exp.RegAgr.

exponenciais2$pec <- exponenciais$exp.RegPec.fitted.values. /   exponenciais$exp.RegAgr.
```

```
exponenciais2$flo <- exponenciais$exp.RegFlo.fitted.values. /   exponenciais$exp.Re

exponenciais2 <- as.data.frame(exponenciais2)

#### Avaliando o scatterplot de y e sua estimativa para a base de
#### treinamento

plot(training$agr.porcentagem, exponenciais2$agr)
plot(training$pas.porcentagem, exponenciais2$pec)
plot(training$flo.porcentagem, exponenciais2$flo)

##### Estatisticas MAPE, MAD, MSD para medir acuracia da previsao do modelo na base

### Agricultura ###

exponenciais2$DesvioAgr <- abs(training$agr.porcentagem - exponenciais2$agr)/ train
exponenciais2$DesvioAgr2 <- abs(training$agr.porcentagem - exponenciais2$agr)
exponenciais2$DesvioAgr3 <- abs(training$agr.porcentagem - exponenciais2$agr)^2

MAPEagr1 <- sum(exponenciais2$DesvioAgr)/80
MADagr1 <- sum(exponenciais2$DesvioAgr2)/80
MSDagr1 <- sum(exponenciais2$DesvioAgr3)/80


### Pecuária ###

exponenciais2$DesvioPec <- abs(training$pas.porcentagem - exponenciais2$pec)/ train
exponenciais2$DesvioPec2 <- abs(training$pas.porcentagem - exponenciais2$pec)
exponenciais2$DesvioPec3 <- abs(training$pas.porcentagem - exponenciais2$pec)^2

MAPEPec1 <- sum(exponenciais2$DesvioPec)/80
MADPec1 <- sum(exponenciais2$DesvioPec2)/80
MSDPec1 <- sum(exponenciais2$DesvioPec3)/80

### Floresta ###

exponenciais2$DesvioFlo <- abs(training$flo.porcentagem - exponenciais2$flo)/ train
exponenciais2$DesvioFlo2 <- abs(training$flo.porcentagem - exponenciais2$flo)
```

```
exponenciais2$DesvioFlo3 <- abs(training$flo.porcentagem - exponenciais2$flo)^2

MAPEFlo1 <- sum(exponenciais2$DesvioFlo)/80
MADFlo1 <- sum(exponenciais2$DesvioFlo2)/80
MSDFlo1 <- sum(exponenciais2$DesvioFlo3)/80

##### Cross-Validation
########################

########### Leave one observation out

model1 <- NULL
model2 <- NULL
model3 <- NULL
exponenciais6 <- NULL
exponenciais7 <- NULL
MAPEagr <- NULL
MAPEagr2 <- NULL
MADagr <- NULL
MADagr2 <- NULL
MSDagr <- NULL
MSDagr2 <- NULL
MAPEpas <- NULL
MAPEpas2 <- NULL
MADpas <- NULL
MADpas2 <- NULL
MSDpas <- NULL
MSDpas2 <- NULL
MAPEflo <- NULL
MAPEflo2 <- NULL
MADflo <- NULL
MADflo2 <- NULL
MSDflo <- NULL
MSDflo2 <- NULL

for(i in 1:80){
  model1 <- lm(lnagr ~ . , data = training[-i,c(2,11,12,13,18,28)])
  model2 <- lm(pec, data = training[-i, c(2,11,17,26,29)])
  model3 <- lm(flo, data = training[-i, c(2,11,17,27,30)])
```

```
base1 <- training[i, c(2,11,12,13,18,28)]
base2 <- training[i, c(2,11,17,26,29)]
base3 <- training[i, c(2,11,17,27,30)]
predagr <- predict(model1, newdata = base1)
predpec <- predict(model2, newdata = base2)
predflo <- predict(model3, newdata = base3)

exponenciais6 <- data.frame(x = exp(predagr),
                            y = exp(predpec),
                            z = exp(predflo),
                            w = exp(predagr) +
                              exp(predpec) +
                              exp(predflo) + 1)


exponenciais7 <- NULL
exponenciais7$agr <- exponenciais6$x/exponenciais6$w

exponenciais7$pas <- exponenciais6$y/exponenciais6$w

exponenciais7$flo <- exponenciais6$z/exponenciais6$w

exponenciais7 <- as.data.frame(exponenciais7)

exponenciais7$DesvioAgr <- abs(training$agr.porcentagem[i] - exponenciais7$agr)/
exponenciais7$DesvioAgr2 <- abs(training$agr.porcentagem[i] - exponenciais7$agr)
exponenciais7$DesvioAgr3 <- abs(training$agr.porcentagem[i] - exponenciais7$agr)^

MAPEagr <- mean(exponenciais7$DesvioAgr)
MAPEagr2 <- c(MAPEagr, MAPEagr2)
MADagr <- mean(exponenciais7$DesvioAgr2)
MADagr2 <- c(MADagr, MADagr2)
MSDagr <- mean(exponenciais7$DesvioAgr3)
MSDagr2 <- c(MSDagr, MSDagr2)

exponenciais7$DesvioPas <- abs(training$pas.porcentagem[i] - exponenciais7$pas)/
exponenciais7$DesvioPas2 <- abs(training$pas.porcentagem[i] - exponenciais7$pas)
exponenciais7$DesvioPas3 <- abs(training$pas.porcentagem[i] - exponenciais7$pas)^
```

```
  MAPEpas <- mean(exponenciais7$DesvioPas)
  MAPEpas2 <- c(MAPEpas, MAPEpas2)
  MADpas <- mean(exponenciais7$DesvioPas2)
  MADpas2 <- c(MADpas, MADpas2)
  MSDpas <- mean(exponenciais7$DesvioPas3)
  MSDpas2 <- c(MSDpas, MSDpas2)

  exponenciais7$DesvioFlo <- abs(training$flo.porcentagem[i] - exponenciais7$flo)/ train
  exponenciais7$DesvioFlo2 <- abs(training$flo.porcentagem[i] - exponenciais7$flo)
  exponenciais7$DesvioFlo3 <- abs(training$flo.porcentagem[i] - exponenciais7$flo)^2

  MAPEflo <- mean(exponenciais7$DesvioFlo)
  MAPEflo2 <- c(MAPEflo, MAPEflo2)
  MADflo <- mean(exponenciais7$DesvioFlo2)
  MADflo2 <- c(MADflo, MADflo2)
  MSDflo <- mean(exponenciais7$DesvioFlo3)
  MSDflo2 <- c(MSDflo, MSDflo2)

}


mean(MAPEagr2)
mean(MADagr2)
mean(MSDagr2)
mean(MAPEpas2)
mean(MADpas2)
mean(MSDpas2)
mean(MAPEflo2)
mean(MADflo2)
mean(MSDflo2)

########### Leave FIVE observations out

###vai mudar toda vez que samplear
thevalues <- sample(x = 1:80,size = 1000,replace = TRUE)
thevalues.unique <- unique(thevalues)
lista <- list()
for(i in 1:16) {
  lista[[i]] <- thevalues.unique[(5*i-4):(5*i)]
```

```
}

model1 <- NULL
model2 <- NULL
model3 <- NULL
exponenciais6 <- NULL
exponenciais7 <- NULL
MAPEagr <- NULL
MAPEagr3 <- NULL
MADagr <- NULL
MADagr3 <- NULL
MSDagr <- NULL
MSDagr3 <- NULL
MAPEpas <- NULL
MAPEpas3 <- NULL
MADpas <- NULL
MADpas3 <- NULL
MSDpas <- NULL
MSDpas3 <- NULL
MAPEflo <- NULL
MAPEflo3 <- NULL
MADflo <- NULL
MADflo3 <- NULL
MSDflo <- NULL
MSDflo3 <- NULL


for(i in 1:16){
  model1 <- lm(lnagr ~ . , data = training[-lista[[i]],c(2,11,12,13,18,28)])
  model2 <- lm(pec, data = training[-lista[[i]], c(2,11,17,26,29)])
  model3 <- lm(flo, data = training[-lista[[i]], c(2,11,17,27,30)])
  base1 <- training[lista[[i]], c(2,11,12,13,18,28)]
  base2 <- training[lista[[i]], c(2,11,17,26,29)]
  base3 <- training[lista[[i]], c(2,11,17,27,30)]
  predagr <- predict(model1, newdata = base1)
  predpec <- predict(model2, newdata = base2)
  predflo <- predict(model3, newdata = base3)

  exponenciais6 <- data.frame(x = exp(predagr),
```

```
                              y = exp(predpec),
                              z = exp(predflo),
                              w = exp(predagr) +
                                exp(predpec) +
                                exp(predflo) + 1)


exponenciais7 <- NULL
exponenciais7$agr <- exponenciais6$x/exponenciais6$w

exponenciais7$pas <- exponenciais6$y/exponenciais6$w

exponenciais7$flo <- exponenciais6$z/exponenciais6$w

exponenciais7 <- as.data.frame(exponenciais7)

exponenciais7$DesvioAgr <- abs(training$agr.porcentagem[lista[[i]]] - exponenciais7$ag
exponenciais7$DesvioAgr2 <- abs(training$agr.porcentagem[lista[[i]]] - exponenciais7$a
exponenciais7$DesvioAgr3 <- abs(training$agr.porcentagem[lista[[i]]] - exponenciais7$a

MAPEagr <- mean(exponenciais7$DesvioAgr)
MAPEagr3 <- c(MAPEagr, MAPEagr3)
MADagr <- mean(exponenciais7$DesvioAgr2)
MADagr3 <- c(MADagr, MADagr3)
MSDagr <- mean(exponenciais7$DesvioAgr3)
MSDagr3 <- c(MSDagr, MSDagr3)

exponenciais7$DesvioPas <- abs(training$pas.porcentagem[lista[[i]]] - exponenciais7$pa
exponenciais7$DesvioPas2 <- abs(training$pas.porcentagem[lista[[i]]] - exponenciais7$p
exponenciais7$DesvioPas3 <- abs(training$pas.porcentagem[lista[[i]]] - exponenciais7$p

MAPEpas <- mean(exponenciais7$DesvioPas)
MAPEpas3 <- c(MAPEpas, MAPEpas3)
MADpas <- mean(exponenciais7$DesvioPas2)
MADpas3 <- c(MADpas, MADpas3)
MSDpas <- mean(exponenciais7$DesvioPas3)
MSDpas3 <- c(MSDpas, MSDpas3)

exponenciais7$DesvioFlo <- abs(training$flo.porcentagem[lista[[i]]] - exponenciais7$fl
```

```
  exponenciais7$DesvioFlo2 <- abs(training$flo.porcentagem[lista[[i]]] - exponencia
  exponenciais7$DesvioFlo3 <- abs(training$flo.porcentagem[lista[[i]]] - exponencia

  MAPEflo <- mean(exponenciais7$DesvioFlo)
  MAPEflo3 <- c(MAPEflo, MAPEflo3)
  MADflo <- mean(exponenciais7$DesvioFlo2)
  MADflo3 <- c(MADflo, MADflo3)
  MSDflo <- mean(exponenciais7$DesvioFlo3)
  MSDflo3 <- c(MSDflo, MSDflo3)

}


mean(MAPEagr3)
mean(MADagr3)
mean(MSDagr3)
mean(MAPEpas3)
mean(MADpas3)
mean(MSDpas3)
mean(MAPEflo3)
mean(MADflo3)
mean(MSDflo3)


############################################################
##### Testando utilizando o test set para o ano 2014 ########

dados.test$codigo <- as.character(dados.test$codigo)

# dados.test2 <- dado.test[ dados.test$codigo %in% tabelacodv3$V2,]
dados.test2 <- dados.test[dados.test$codigo %in% training$codigo, ]

PredAgr <- predict(RegAgr, newdata=dados.test2)
PredPec <- predict(RegPec, newdata=dados.test2)
PredFlo <- predict(RegFlo, newdata=dados.test2)

exponenciais <- data.frame(exp(PredAgr),
                           exp(PredPec),
                           exp(PredFlo),
```

```
                          exp(PredAgr) +
                            exp(PredPec) +
                            exp(PredFlo) + 1)


exponenciais$agr <- exponenciais$exp.PredAgr. /   exponenciais$exp.PredAgr....exp.PredPe

exponenciais$pec <- exponenciais$exp.PredPec. /   exponenciais$exp.PredAgr....exp.PredPe

exponenciais$flo <- exponenciais$exp.PredFlo. /   exponenciais$exp.PredAgr....exp.PredPe


plot(dados.test2$agr.porcentagem, exponenciais$agr)
plot(dados.test2$pas.porcentagem, exponenciais$pec)
plot(dados.test2$flo.porcentagem, exponenciais$flo)

final <- merge(dados.test2, tabelacod, by.x="codigo", by.y="V2")
##### Estatisticas MAPE, MAD, MSD para medir acuracia da previsao do modelo com a base t

### Agricultura ###

exponenciais$DesvioAgr <- abs(dados.test2$agr.porcentagem - exponenciais$agr)/ dados.tes
exponenciais$DesvioAgr2 <- abs(dados.test2$agr.porcentagem - exponenciais$agr)
exponenciais$DesvioAgr3 <- abs(dados.test2$agr.porcentagem - exponenciais$agr)^2

MAPEagr4 <- sum(exponenciais$DesvioAgr)/20
MADagr4 <- sum(exponenciais$DesvioAgr2)/20
MSDagr4 <- sum(exponenciais$DesvioAgr3)/20


### PecuÃ¡ria ###

exponenciais$DesvioPec <- abs(dados.test2$pas.porcentagem - exponenciais$pec)/ dados.tes
exponenciais$DesvioPec2 <- abs(dados.test2$pas.porcentagem - exponenciais$pec)
exponenciais$DesvioPec3 <- abs(dados.test2$pas.porcentagem - exponenciais$pec)^2

MAPEPec4 <- sum(exponenciais$DesvioPec)/20
MADPec4 <- sum(exponenciais$DesvioPec2)/20
MSDPec4 <- sum(exponenciais$DesvioPec3)/20
```

```
### Floresta ###

exponenciais$DesvioFlo <- abs(dados.test2$flo.porcentagem - exponenciais$flo)/ dado
exponenciais$DesvioFlo2 <- abs(dados.test2$flo.porcentagem - exponenciais$flo)
exponenciais$DesvioFlo3 <- abs(dados.test2$flo.porcentagem - exponenciais$flo)^2

MAPEFlo4 <- sum(exponenciais$DesvioFlo)/20
MADFlo4 <- sum(exponenciais$DesvioFlo2)/20
MSDFlo4 <- sum(exponenciais$DesvioFlo3)/20

MAPEagr4
MADagr4
MSDagr4
MAPEPec4
MADPec4
MSDPec4
MAPEFlo4
MADFlo4
MSDFlo4

#### Analise de denise ####


#####################################################################
###### Modelo alterando codigo do ibge por dados biofisicos ######

microrregioes <- read.csv2(file='/Users/Gustavo/Documents/unb/10_semestre/TCC2/micr



#############################################################
### VariÃ¡veis de Qualidade ###################################

dados2 <- read.csv2(file='/Users/Gustavo/Downloads/dados2.csv', sep=',')


for(i in c(3:79)){
  dados2[,i] <- as.numeric(as.character(dados2[,i]))
```

```
}

dados2[dados2 == 0] <- 0.0000000001
dados2[is.na(dados2)] <- 0.0000000001

#normalizando tendo outros como referencia:

dados2$lnagr <- log(dados2$agr.porcentagem/dados2$out.porcentagem)
dados2$lnpec <- log(dados2$pas.porcentagem/dados2$out.porcentagem)
dados2$lnflo <- log(dados2$flo.porcentagem/dados2$out.porcentagem)

#setando base de treinamento

dados.Qtraining <- dados2[dados2$ano != 2014,]

dados2.test <- dados2[dados2$ano == 2014,]


####################################
####### Minimos Quadrados #########
####################################

#funcoes de regressao utilizando Quadrados Minimos para os tres casos:

### selecionando apenas os casos para o teste set onde temos
### informacao completa para todos os 4 anos

Qtraining <- dados.Qtraining[dados.Qtraining$codigo %in% training$codigo,]

############## M o d e l o s ####################

Qtraining$codigo <- as.character(Qtraining$codigo)


QRegAgr <- lm(lnagr ~ . , data = Qtraining[c(11:80)])

qpec <- as.formula(paste('lnpec ~', paste(colnames(Qtraining)[c(11:79)], collapse='+')))
qflo <- as.formula(paste('lnflo ~', paste(colnames(Qtraining)[c(11:79)], collapse='+')))
```

```
QRegPec <- lm(qpec, data = Qtraining[c(11:79,81)])
QRegFlo <- lm(qflo, data = Qtraining[c(11:79,82)])


##### StepWise ######

QstepAgr <- stepAIC(QRegAgr, direction="both")

summary(QstepAgr)

QstepPec <- stepAIC(QRegPec, direction="both")
QstepFlo <- stepAIC(QRegFlo, direction="both")

plot(QstepFlo)


######## Modelos selecionados no stepWise #######
#
# QRegAgr <- lm(lnagr ~ . , data = Qtraining[c(11,12,13,14,15,16,18,19,20,21,22,23,

QRegAgr <- lm(lnagr ~ . , data = Qtraining[c(11,12,13,14,18,19,20,21,22,23,24,29,31

QRegPec <- lm(lnpec ~ . , data = Qtraining[c(18,20,28,29,31,32,33,34,37,46,53,55,57

QRegFlo <- lm(lnflo ~ . , data = Qtraining[c(16,18,28,33,36,46,53,57,58,82)])

#11,13,16,18,20,21,22,25,26,28,29,31,32,34,36,46,53,57,58,82
#Calculando as estimativas no logito:

exponenciais3 <- data.frame(exp(QRegAgr$fitted.values),
                            exp(QRegPec$fitted.values),
                            exp(QRegFlo$fitted.values),
                            exp(QRegAgr$fitted.values) +
                              exp(QRegPec$fitted.values) +
                              exp(QRegFlo$fitted.values) + 1)

exponenciais4 <- NULL
exponenciais4$agr <- exponenciais3$exp.QRegAgr.fitted.values. /   exponenciais3$exp
```

```
exponenciais4$pec <- exponenciais3$exp.QRegPec.fitted.values. /   exponenciais3$exp.QReg

exponenciais4$flo <- exponenciais3$exp.QRegFlo.fitted.values. /   exponenciais3$exp.QReg

exponenciais4 <- as.data.frame(exponenciais4)


#### Avaliando o scatterplot de y e sua estimativa para a base de
#### treinamento

plot(Qtraining$agr.porcentagem, exponenciais4$agr)
plot(Qtraining$pas.porcentagem, exponenciais4$pec)
plot(Qtraining$flo.porcentagem, exponenciais4$flo)

##### Estatisticas MAPE, MAD, MSD para medir acuracia da previsao

### Agricultura ###

exponenciais4$DesvioAgr <- abs(Qtraining$agr.porcentagem - exponenciais4$agr)/ Qtraining
exponenciais4$DesvioAgr2 <- abs(Qtraining$agr.porcentagem - exponenciais4$agr)
exponenciais4$DesvioAgr3 <- abs(Qtraining$agr.porcentagem - exponenciais4$agr)^2

QMAPEagr <- sum(exponenciais4$DesvioAgr)/60
QMADagr <- sum(exponenciais4$DesvioAgr2)/60
QMSDagr <- sum(exponenciais4$DesvioAgr3)/60


### Pecuária ###

exponenciais4$DesvioPec <- abs(Qtraining$pas.porcentagem - exponenciais4$pec)/ Qtraining
exponenciais4$DesvioPec2 <- abs(Qtraining$pas.porcentagem - exponenciais4$pec)
exponenciais4$DesvioPec3 <- abs(Qtraining$pas.porcentagem - exponenciais4$pec)^2

QMAPEPec <- sum(exponenciais4$DesvioPec)/60
QMADPec <- sum(exponenciais4$DesvioPec2)/60
QMSDPec <- sum(exponenciais4$DesvioPec3)/60

### Floresta ###
```

```
exponenciais4$DesvioFlo <- abs(Qtraining$flo.porcentagem - exponenciais4$flo)/ Qtra
exponenciais4$DesvioFlo2 <- abs(Qtraining$flo.porcentagem - exponenciais4$flo)
exponenciais4$DesvioFlo3 <- abs(Qtraining$flo.porcentagem - exponenciais4$flo)^2

QMAPEFlo <- sum(exponenciais4$DesvioFlo)/60
QMADFlo <- sum(exponenciais4$DesvioFlo2)/60
QMSDFlo <- sum(exponenciais4$DesvioFlo3)/60

QMAPEagr
QMADagr
QMSDagr
QMAPEPec
QMADPec
QMSDPec
QMAPEFlo
QMADFlo
QMSDFlo




##### Cross-Validation
########################

########### Leave one observation out

model1 <- NULL
model2 <- NULL
model3 <- NULL
exponenciais6 <- NULL
exponenciais7 <- NULL
QMAPEagr <- NULL
QMAPEagr2 <- NULL
QMADagr <- NULL
QMADagr2 <- NULL
QMSDagr <- NULL
QMSDagr2 <- NULL
QMAPEpas <- NULL
```

```
QMAPEpas2 <- NULL
QMADpas <- NULL
QMADpas2 <- NULL
QMSDpas <- NULL
QMSDpas2 <- NULL
QMAPEflo <- NULL
QMAPEflo2 <- NULL
QMADflo <- NULL
QMADflo2 <- NULL
QMSDflo <- NULL
QMSDflo2 <- NULL

for(i in 1:60){
  model1 <- lm(lnagr ~ . , data = Qtraining[-i,c(11,12,13,14,18,19,20,21,22,23,24,29,31,
  model2 <- lm(lnpec ~ ., data = Qtraining[-i,c(18,20,28,29,31,32,33,34,37,46,53,55,57,8
  model3 <- lm(lnflo ~ ., data = Qtraining[-i,c(16,18,28,33,36,46,53,57,58,82)])
  base1 <- Qtraining[i, c(11,12,13,14,18,19,20,21,22,23,24,29,31,32,36,37,46,53,57,80)]
  base2 <- Qtraining[i, c(18,20,28,29,31,32,33,34,37,46,53,55,57,81)]
  base3 <- Qtraining[i,c(16,18,28,33,36,46,53,57,58,82)]
  qpredagr <- predict(model1, newdata = base1)
  qpredpec <- predict(model2, newdata = base2)
  qpredflo <- predict(model3, newdata = base3)


  exponenciais6 <- data.frame(x = exp(qpredagr),
                              y = exp(qpredpec),
                              z = exp(qpredflo),
                              w = exp(qpredagr) +
                                exp(qpredpec) +
                                exp(qpredflo) + 1)


  exponenciais7 <- NULL
  exponenciais7$agr <- exponenciais6$x/exponenciais6$w

  exponenciais7$pas <- exponenciais6$y/exponenciais6$w

  exponenciais7$flo <- exponenciais6$z/exponenciais6$w
```

```
  exponenciais7 <- as.data.frame(exponenciais7)

  exponenciais7$DesvioAgr <- abs(Qtraining$agr.porcentagem[i] - exponenciais7$agr)/
  exponenciais7$DesvioAgr2 <- abs(Qtraining$agr.porcentagem[i] - exponenciais7$agr)
  exponenciais7$DesvioAgr3 <- abs(Qtraining$agr.porcentagem[i] - exponenciais7$agr)

  QMAPEagr <- mean(exponenciais7$DesvioAgr)
  QMAPEagr2 <- c(QMAPEagr, QMAPEagr2)
  QMADagr <- mean(exponenciais7$DesvioAgr2)
  QMADagr2 <- c(QMADagr, QMADagr2)
  QMSDagr <- mean(exponenciais7$DesvioAgr3)
  QMSDagr2 <- c(QMSDagr, QMSDagr2)

  exponenciais7$DesvioPas <- abs(Qtraining$pas.porcentagem[i] - exponenciais7$pas)/
  exponenciais7$DesvioPas2 <- abs(Qtraining$pas.porcentagem[i] - exponenciais7$pas)
  exponenciais7$DesvioPas3 <- abs(Qtraining$pas.porcentagem[i] - exponenciais7$pas)

  QMAPEpas <- mean(exponenciais7$DesvioPas)
  QMAPEpas2 <- c(QMAPEpas, QMAPEpas2)
  QMADpas <- mean(exponenciais7$DesvioPas2)
  QMADpas2 <- c(QMADpas, QMADpas2)
  QMSDpas <- mean(exponenciais7$DesvioPas3)
  QMSDpas2 <- c(QMSDpas, QMSDpas2)

  exponenciais7$DesvioFlo <- abs(Qtraining$flo.porcentagem[i] - exponenciais7$flo)/
  exponenciais7$DesvioFlo2 <- abs(Qtraining$flo.porcentagem[i] - exponenciais7$flo)
  exponenciais7$DesvioFlo3 <- abs(Qtraining$flo.porcentagem[i] - exponenciais7$flo)

  QMAPEflo <- mean(exponenciais7$DesvioFlo)
  QMAPEflo2 <- c(QMAPEflo, QMAPEflo2)
  QMADflo <- mean(exponenciais7$DesvioFlo2)
  QMADflo2 <- c(QMADflo, QMADflo2)
  QMSDflo <- mean(exponenciais7$DesvioFlo3)
  QMSDflo2 <- c(QMSDflo, QMSDflo2)

}


mean(QMAPEagr2)
```

```
mean(QMADagr2)
mean(QMSDagr2)
mean(QMAPEpas2)
mean(QMADpas2)
mean(QMSDpas2)
mean(QMAPEflo2)
mean(QMADflo2)
mean(QMSDflo2)


########### Leave FIVE observations out


thevalues <- sample(x = 1:60,size = 1000,replace = TRUE)
thevalues.unique <- unique(thevalues)
lista <- list()
for(i in 1:12) {
  lista[[i]] <- thevalues.unique[(5*i-4):(5*i)]
}

model1 <- NULL
model2 <- NULL
model3 <- NULL
exponenciais6 <- NULL
exponenciais7 <- NULL
QMAPEagr <- NULL
QMAPEagr3 <- NULL
QMADagr <- NULL
QMADagr3 <- NULL
QMSDagr <- NULL
QMSDagr3 <- NULL
QMAPEpas <- NULL
QMAPEpas3 <- NULL
QMADpas <- NULL
QMADpas3 <- NULL
QMSDpas <- NULL
QMSDpas3 <- NULL
QMAPEflo <- NULL
QMAPEflo3 <- NULL
```

```
QMADflo <- NULL
QMADflo3 <- NULL
QMSDflo <- NULL
QMSDflo3 <- NULL

for(i in 1:12){
  model1 <- lm(lnagr ~ . , data = Qtraining[-lista[[i]],c(11,12,13,14,18,19,20,21,2
  model2 <- lm(lnpec ~ ., data = Qtraining[-lista[[i]],c(18,20,28,29,31,32,33,34,37
  model3 <- lm(lnflo ~ ., data = Qtraining[-lista[[i]],c(16,18,28,33,36,46,53,57,58
  base1 <- Qtraining[lista[[i]], c(11,12,13,14,18,19,20,21,22,23,24,29,31,32,36,37,
  base2 <- Qtraining[lista[[i]], c(18,20,28,29,31,32,33,34,37,46,53,55,57,81)]
  base3 <- Qtraining[lista[[i]],c(16,18,28,33,36,46,53,57,58,82)]
  qpredagr <- predict(model1, newdata = base1)
  qpredpec <- predict(model2, newdata = base2)
  qpredflo <- predict(model3, newdata = base3)


  exponenciais6 <- data.frame(x = exp(qpredagr),
                              y = exp(qpredpec),
                              z = exp(qpredflo),
                              w = exp(qpredagr) +
                                exp(qpredpec) +
                                exp(qpredflo) + 1)


  exponenciais7 <- NULL
  exponenciais7$agr <- exponenciais6$x/exponenciais6$w

  exponenciais7$pas <- exponenciais6$y/exponenciais6$w

  exponenciais7$flo <- exponenciais6$z/exponenciais6$w

  exponenciais7 <- as.data.frame(exponenciais7)

  exponenciais7$DesvioAgr <- abs(Qtraining$agr.porcentagem[lista[[i]]] - exponencia
  exponenciais7$DesvioAgr2 <- abs(Qtraining$agr.porcentagem[lista[[i]]] - exponenci
  exponenciais7$DesvioAgr3 <- abs(Qtraining$agr.porcentagem[lista[[i]]] - exponenci

  QMAPEagr <- mean(exponenciais7$DesvioAgr)
```

```
  QMAPEagr3 <- c(QMAPEagr, QMAPEagr3)
  QMADagr <- mean(exponenciais7$DesvioAgr2)
  QMADagr3 <- c(QMADagr, QMADagr3)
  QMSDagr <- mean(exponenciais7$DesvioAgr3)
  QMSDagr3 <- c(QMSDagr, QMSDagr3)

  exponenciais7$DesvioPas <- abs(Qtraining$pas.porcentagem[lista[[i]]] - exponenciais7$p
  exponenciais7$DesvioPas2 <- abs(Qtraining$pas.porcentagem[lista[[i]]] - exponenciais7$
  exponenciais7$DesvioPas3 <- abs(Qtraining$pas.porcentagem[lista[[i]]] - exponenciais7$

  QMAPEpas <- mean(exponenciais7$DesvioPas)
  QMAPEpas3 <- c(QMAPEpas, QMAPEpas3)
  QMADpas <- mean(exponenciais7$DesvioPas2)
  QMADpas3 <- c(QMADpas, QMADpas3)
  QMSDpas <- mean(exponenciais7$DesvioPas3)
  QMSDpas3 <- c(QMSDpas, QMSDpas3)

  exponenciais7$DesvioFlo <- abs(Qtraining$flo.porcentagem[lista[[i]]] - exponenciais7$f
  exponenciais7$DesvioFlo2 <- abs(Qtraining$flo.porcentagem[lista[[i]]] - exponenciais7$
  exponenciais7$DesvioFlo3 <- abs(Qtraining$flo.porcentagem[lista[[i]]] - exponenciais7$

  QMAPEflo <- mean(exponenciais7$DesvioFlo)
  QMAPEflo3 <- c(QMAPEflo, QMAPEflo3)
  QMADflo <- mean(exponenciais7$DesvioFlo2)
  QMADflo3 <- c(QMADflo, QMADflo3)
  QMSDflo <- mean(exponenciais7$DesvioFlo3)
  QMSDflo3 <- c(QMSDflo, QMSDflo3)

}


mean(QMAPEagr3)
mean(QMADagr3)
mean(QMSDagr3)
mean(QMAPEpas3)
mean(QMADpas3)
mean(QMSDpas3)
mean(QMAPEflo3)
mean(QMADflo3)
```

```
mean(QMSDflo3)



############################################################
##### Testando utilizando o test set para o ano 2014 ########

dados2.test$codigo <- as.character(dados2.test$codigo)

# dados.test2 <- dado.test[ dados.test$codigo %in% tabelacodv3$V2,]
dados2.test2 <- dados2.test[ dados2.test$codigo %in% Qtraining$codigo,]

QPredAgr <- predict(QRegAgr, newdata=dados2.test2)
QPredPec <- predict(QRegPec, newdata=dados2.test2)
QPredFlo <- predict(QRegFlo, newdata=dados2.test2)

exponenciais5 <- data.frame(exp(QPredAgr),
                            exp(QPredPec),
                            exp(QPredFlo),
                            exp(QPredAgr) +
                              exp(QPredPec) +
                              exp(QPredFlo) + 1)


exponenciais5$agr <- exponenciais5$exp.QPredAgr. /   exponenciais5$exp.QPredAgr....

exponenciais5$pec <- exponenciais5$exp.QPredPec. /   exponenciais5$exp.QPredAgr....

exponenciais5$flo <- exponenciais5$exp.QPredFlo. /   exponenciais5$exp.QPredAgr....

plot(dados2.test2$agr.porcentagem, exponenciais5$agr)
plot(dados2.test2$pas.porcentagem, exponenciais5$pec)
plot(dados2.test2$flo.porcentagem, exponenciais5$flo)

exponenciais5$agr - dados2.test2$agr.porcentagem
exponenciais5$pec - dados2.test2$pas.porcentagem
exponenciais5$flo - dados2.test2$flo.porcentagem
```

```
#######################

exponenciais5$DesvioAgr <- abs(dados2.test2$agr.porcentagem - exponenciais5$agr)/ dados2
exponenciais5$DesvioAgr2 <- abs(dados2.test2$agr.porcentagem - exponenciais5$agr)
exponenciais5$DesvioAgr3 <- abs(dados2.test2$agr.porcentagem - exponenciais5$agr)^2

QMAPEagr4 <- sum(exponenciais5$DesvioAgr)/15
QMADagr4 <- sum(exponenciais5$DesvioAgr2)/15
QMSDagr4 <- sum(exponenciais5$DesvioAgr3)/15


### Pecuária ###

exponenciais5$DesvioPec <- abs(dados2.test2$pas.porcentagem - exponenciais5$pec)/ dados2
exponenciais5$DesvioPec2 <- abs(dados2.test2$pas.porcentagem - exponenciais5$pec)
exponenciais5$DesvioPec3 <- abs(dados2.test2$pas.porcentagem - exponenciais5$pec)^2

QMAPEPec4 <- sum(exponenciais5$DesvioPec)/15
QMADPec4 <- sum(exponenciais5$DesvioPec2)/15
QMSDPec4 <- sum(exponenciais5$DesvioPec3)/15

### Floresta ###

exponenciais5$DesvioFlo <- abs(dados2.test2$flo.porcentagem - exponenciais5$flo)/ dados2
exponenciais5$DesvioFlo2 <- abs(dados2.test2$flo.porcentagem - exponenciais5$flo)
exponenciais5$DesvioFlo3 <- abs(dados2.test2$flo.porcentagem - exponenciais5$flo)^2

QMAPEFlo4 <- sum(exponenciais5$DesvioFlo)/15
QMADFlo4 <- sum(exponenciais5$DesvioFlo2)/15
QMSDFlo4 <- sum(exponenciais5$DesvioFlo3)/15

QMAPEagr4
QMADagr4
QMSDagr4
QMAPEPec4
QMADPec4
QMSDPec4
QMAPEFlo4
QMADFlo4
```

```
QMSDFlo4

################## tirando 2 municípios outliers
#4 (1,2,3), 5 (1,3), 8 (2,3), 13 (1,3)
# 5103452,5103700,5108907,5107909 respectivamente
# Denise, Feliz Natal, Nova MaringÃ¡, Sinop respectivamente

ff <- c(5103700,5107909)

Qtraining2 <- Qtraining[!Qtraining$codigo %in% ff,]

QRegAgr <- lm(lnagr ~ . , data = Qtraining2[c(11,12,13,14,18,19,20,21,22,23,24,29,3

QRegPec <- lm(lnpec ~ . , data = Qtraining2[c(18,20,28,29,31,32,33,34,37,46,53,55,5

QRegFlo <- lm(lnflo ~ . , data = Qtraining2[c(16,18,28,33,36,46,53,57,58,82)])


#################################################################
##### Testando utilizando o test set para o ano 2014 ########

dados2.test$codigo <- as.character(dados2.test$codigo)

# dados.test2 <- dado.test[ dados.test$codigo %in% tabelacodv3$V2,]
dados2.test3 <- dados2.test[ dados2.test$codigo %in% Qtraining2$codigo,]

QPredAgr <- predict(QRegAgr, newdata=dados2.test3)
QPredPec <- predict(QRegPec, newdata=dados2.test3)
QPredFlo <- predict(QRegFlo, newdata=dados2.test3)

exponenciais6 <- data.frame(exp(QPredAgr),
                            exp(QPredPec),
                            exp(QPredFlo),
                            exp(QPredAgr) +
                              exp(QPredPec) +
                              exp(QPredFlo) + 1)


exponenciais6$agr <- exponenciais6$exp.QPredAgr. /   exponenciais6$exp.QPredAgr....
```

```
exponenciais6$pec <- exponenciais6$exp.QPredPec. /   exponenciais6$exp.QPredAgr....exp.Q

exponenciais6$flo <- exponenciais6$exp.QPredFlo. /   exponenciais6$exp.QPredAgr....exp.Q

plot(dados2.test3$agr.porcentagem, exponenciais6$agr)
plot(dados2.test3$pas.porcentagem, exponenciais6$pec)
plot(dados2.test3$flo.porcentagem, exponenciais6$flo)

exponenciais5$agr - dados2.test2$agr.porcentagem
exponenciais5$pec - dados2.test2$pas.porcentagem
exponenciais5$flo - dados2.test2$flo.porcentagem
```

A.4 Código para análise descritiva

```
##-----------------------------------------------------
## Subindo base 'dados', com todos os municípios
##-----------------------------------------------------

dados <- read_csv("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/planilhas modificada
dados <- dados[, -1]
codigo_mun <- read_csv("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/planilhas modif
codigo_mun <- codigo_mun[, -1]

##-----------------------------------------------------
## base com os municípios usados no modelo
##-----------------------------------------------------

municipios <- subset(dados, codigo %in% codigo_mun$V2)
municipios2 <- subset(dados, codigo %in% c("5100250", "5101308", "5103353", "5103452", "
                                           "5105606", "5108808", "5108907", "5106190", "
                                           "5106422", "5106802", "5107909", "5108055", "
municipios3 <- subset(dados, codigo %in% c(codigo_mun$V2, c("5103205", "5105580", "51056
                                                    "5107263", "5107263", "51073

municipios4 <- subset(dados, !(codigo %in% c(codigo_mun$V2, c("5103205", "5105580", "510
```

```
                                                        "5107263", "5107263",

df <- aggregate(cbind(agr_perm.producao, agr_temp.producao, flo.exve_producao, flo_
                flo_silv.producao_m3, pec.efetivo, pec.valor) ~ ano, data = municip
                sum, na.rm = TRUE)

df2 <- aggregate(cbind(agr_perm.producao, agr_temp.producao, flo.exve_producao, flo
                       flo_silv.producao_m3, pec.efetivo, pec.valor) ~ ano, data =
                 sum, na.rm = TRUE)

for(i in 1:7) {
  df[, i + 8] <- df[ , i + 1] / df2[, i + 1]
}

##-----------------------------------------------------
## Análise da produção da agricultura temporária
##-----------------------------------------------------

#--- Porcentagem por ano de cada cultura ---#

### precisa fazer essa base de novo com os municípios usados no modelo no IBGE

agr_temp <- read.delim("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/Análise de

a <- ggplot(agr_temp, aes(x = factor(Ano), y = Porcentagem, fill = Cultura)) +
 geom_bar(stat = "identity") + labs(x = "ano") +
  scale_fill_brewer(palette = "Set2") + theme_minimal()

#--- Porcentagem da produção dos municípios usados no modelo em relação ao estado

b <- ggplot(df, aes(x = factor(ano), y = V10)) + geom_bar(stat = "identity") +
  labs(x = "Ano", y = "Porcentagem") + scale_fill_brewer(palette = "Set2") + theme_

#--- distribuição da produção dos municípios usados no modelo

c <- ggplot(municipios, aes(x = factor(ano), y = agr_temp.producao)) + geom_boxplot
  labs(x = "Ano", y = "Produção") + scale_fill_brewer(palette = "Set2") + theme_min

summary(municipios$agr_temp.producao)
```

```
summary(municipios$agr_temp.valor)
summary(municipios$agr_temp.areaplantada)
summary(municipios$agr_temp.areacolhida)

sd(municipios$agr_temp.producao)
sd(municipios$agr_temp.valor)
sd(municipios$agr_temp.areaplantada)
sd(municipios$agr_temp.areacolhida)

##-----------------------------------------------------
## Análise da produção da agricultura permanente
##-----------------------------------------------------
agr_perm <- read.delim("C:/Users/igor/Google Drive/TCC/Dados/mato grosso/Análise descrit

#--- Porcentagem da produção dos municípios usados no modelo em relação ao estado

b <- ggplot(df, aes(x = factor(ano), y = V9)) + geom_bar(stat = "identity") +
  labs(x = "Ano", y = "Porcentagem") + scale_fill_brewer(palette = "Set2") + theme_minim

c <- ggplot(municipios, aes(x = factor(ano), y = agr_perm.producao)) + geom_boxplot() +
  labs(x = "Ano", y = "Produção") + scale_fill_brewer(palette = "Set2") + theme_minimal(

summary(municipios$agr_perm.producao)
summary(municipios$agr_perm.valor)
summary(municipios$agr_perm.areaplantada)
summary(municipios$agr_perm.areacolhida)

sd(municipios$agr_perm.producao)
sd(municipios$agr_perm.valor)
sd(municipios$agr_perm.areaplantada)
sd(municipios$agr_perm.areacolhida)
##-----------------------------------------------------
## Análise da produção da extração vegetal
##-----------------------------------------------------

#--- Porcentagem da produção dos municípios usados no modelo em relação ao estado

b <- ggplot(df, aes(x = factor(ano), y = V11)) + geom_bar(stat = "identity") +
  labs(x = "Ano", y = "Porcentagem") + scale_fill_brewer(palette = "Set2") + theme_minim
```

```
#--- distribuição da produção dos municípios usados no modelo

c <- ggplot(municipios, aes(x = factor(ano), y = flo.exve_producao)) + geom_boxplot
  labs(x = "Ano", y = "Produção") + scale_fill_brewer(palette = "Set2") + theme_min

summary(municipios$flo.exve_producao)
summary(municipios$flo.exve_valor)

sd(municipios$flo.exve_producao)
sd(municipios$flo.exve_valor)

##-----------------------------------------------------
## Análise da produção da extração vegetal de metros cúbicos
##-----------------------------------------------------

#--- Porcentagem da produção dos municípios usados no modelo em relação ao estado

b <- ggplot(df, aes(x = factor(ano), y = V12)) + geom_bar(stat = "identity") +
  labs(x = "Ano", y = "Porcentagem") + scale_fill_brewer(palette = "Set2") + theme_

#--- distribuição da produção dos municípios usados no modelo

c <- ggplot(municipios, aes(x = factor(ano), y = flo_exve.producao_m3)) + geom_boxp
  labs(x = "Ano", y = "Produção") + scale_fill_brewer(palette = "Set2") + theme_min

summary(municipios$flo_exve.producao_m3)
summary(municipios$flo_exve.valor_m3)

sd(municipios$flo_exve.producao_m3)
sd(municipios$flo_exve.valor_m3)

##-----------------------------------------------------
## Análise da produção da silvicultura de metros cúbicos
##-----------------------------------------------------

#--- Porcentagem da produção dos municípios usados no modelo em relação ao estado

b <- ggplot(df, aes(x = factor(ano), y = V13)) + geom_bar(stat = "identity") +
```

```
  labs(x = "Ano", y = "Porcentagem") + scale_fill_brewer(palette = "Set2") + theme_minim

#--- distribuição da produção dos municípios usados no modelo

c <- ggplot(municipios, aes(x = factor(ano), y = flo_silv.producao_m3)) + geom_boxplot()
  labs(x = "Ano", y = "Produção") + scale_fill_brewer(palette = "Set2") + theme_minimal(

summary(municipios$flo_silv.producao_m3)
summary(municipios$flo_silv.valor_m3)

sd(municipios$flo_silv.producao_m3)
sd(municipios$flo_silv.valor_m3)

##-----------------------------------------------------
## Análise do efetivo da pecuária
##-----------------------------------------------------

#--- Porcentagem da produção dos municípios usados no modelo em relação ao estado

b <- ggplot(df, aes(x = factor(ano), y = V14)) + geom_bar(stat = "identity") +
  labs(x = "Ano", y = "Porcentagem") + scale_fill_brewer(palette = "Set2") + theme_minim

#--- distribuição da produção dos municípios usados no modelo

c <- ggplot(municipios, aes(x = factor(ano), y = pec.efetivo)) + geom_boxplot() +
  labs(x = "Ano", y = "Produção") + scale_fill_brewer(palette = "Set2") + theme_minimal(

summary(municipios$pec.efetivo)

sd(municipios$pec.efetivo)

##-----------------------------------------------------
## Análise do valor da produção de origem animal
##-----------------------------------------------------

#--- Porcentagem da produção dos municípios usados no modelo em relação ao estado

b <- ggplot(df, aes(x = factor(ano), y = V15)) + geom_bar(stat = "identity") +
  labs(x = "Ano", y = "Porcentagem") + scale_fill_brewer(palette = "Set2") + theme_minim
```

```
#--- distribuição da produção dos municípios usados no modelo

c <- ggplot(municipios, aes(x = factor(ano), y = pec.valor)) + geom_boxplot() +
  labs(x = "Ano", y = "Produção") + scale_fill_brewer(palette = "Set2") + theme_min

summary(municipios$pec.valor)

sd(municipios$pec.valor)

##-----------------------------------------------------
## Análise dos municípios escolhidos no modelo 1
##-----------------------------------------------------

df3 <- aggregate(cbind(agr.area, pas.area, flo.area, out.area) ~ ano, data = munici
                 sum, na.rm = TRUE)

for (i in 1:5) {
df3[i, 6] <- sum(df3[i, 2:5])
}

df4 <- data.frame(Ano = rep(c(2004, 2008, 2010, 2012, 2014), 4),
                  Porcentagem = c(df3$agr.area,
                                  df3$pas.area,
                                  df3$flo.area,
                                  df3$out.area)/11594304,
                  Uso = c(rep("Agrícola", 5),
                          rep("Pastagem", 5),
                          rep("Matas e florestas", 5),
                          rep("Outros usos", 5)))


a <- ggplot(df4, aes(x = factor(Ano), y = Porcentagem, fill = Uso)) +
  geom_bar(stat = "identity") + labs(x = "Ano") +
  scale_fill_brewer(palette = "Set2") + theme_minimal() +
  theme(legend.position="bottom")

ggplot(df4, aes(x = factor(Ano), y = Porcentagem, group = Uso)) +
  geom_line(aes(color = Uso), size = .8) + labs(x = "Ano") +
```

```
  scale_color_brewer(palette = "Set2") + theme_minimal()

################### municípios descartados

descartados <- subset(dados, !(codigo %in% codigo_mun$V2))

df5 <- aggregate(cbind(agr.area, pas.area, flo.area, out.area) ~ ano,
                 data = descartados, sum, na.rm = TRUE)

for (i in 1:5) {
  df5[i, 6] <- sum(df5[i, 2:5])
}

df6 <- data.frame(Ano = rep(c(2004, 2008, 2010, 2012, 2014), 4),
                  Porcentagem = c(df5$agr.area, df5$pas.area, df5$flo.area, df5$out.area
                  Uso = c(rep("Agrícola", 5),
                          rep("Pastagem", 5),
                          rep("Matas e florestas", 5),
                          rep("Outros usos", 5)))

a <- ggplot(df6, aes(x = factor(Ano), y = Porcentagem, fill = Uso)) +
  geom_bar(stat = "identity") + labs(x = "Ano") +
  scale_fill_brewer(palette = "Set2") + theme_minimal()

ggplot(df6, aes(x = factor(Ano), y = Porcentagem, group = Uso)) +
  geom_line(aes(color = Uso), size = .8) + labs(x = "Ano") +
  scale_color_brewer(palette = "Set2") + theme_minimal()

################### todos municípios


df7 <- aggregate(cbind(agr.area, pas.area, flo.area, out.area) ~ ano,
                 data = dados, sum, na.rm = TRUE)

for (i in 1:5) {
  df7[i, 6] <- sum(df7[i, 2:5])
}

df8 <- data.frame(Ano = rep(c(2004, 2008, 2010, 2012, 2014), 4),
```

```
                  Porcentagem = c(df7$agr.area, df7$pas.area, df7$flo.area, df7$out
                  Uso = c(rep("Agrícola", 5),
                          rep("Pastagem", 5),
                          rep("Matas e florestas", 5),
                          rep("Outros usos", 5)))

b <- ggplot(df8, aes(x = factor(Ano), y = Porcentagem, fill = Uso)) +
  geom_bar(stat = "identity") + labs(x = "Ano") +
  scale_fill_brewer(palette = "Set2") + theme_minimal() +
  theme(legend.position="bottom")

ggplot(df8, aes(x = factor(Ano), y = Porcentagem, group = Uso)) +
  geom_line(aes(color = Uso), size = .8) + labs(x = "Ano") +
  scale_color_brewer(palette = "Set2") + theme_minimal()


##-----------------------------------------------------
## Análise dos municípios escolhidos no modelo 2 (com dados biofísicos)
##-----------------------------------------------------

df9 <- aggregate(cbind(agr.area, pas.area, flo.area, out.area) ~ ano, data = munici
                 sum, na.rm = TRUE)

for (i in 1:5) {
  df9[i, 6] <- sum(df9[i, 2:5])
}

df10 <- data.frame(Ano = rep(c(2004, 2008, 2010, 2012, 2014), 4),
                  Porcentagem = c(df9$agr.area,
                                  df9$pas.area,
                                  df9$flo.area,
                                  df9$out.area)/8633146,
                  Uso = c(rep("Agrícola", 5),
                          rep("Pastagem", 5),
                          rep("Matas e florestas", 5),
                          rep("Outros usos", 5)))


c <- ggplot(df10, aes(x = factor(Ano), y = Porcentagem, fill = Uso)) +
```

```
  geom_bar(stat = "identity") + labs(x = "Ano") +
  scale_fill_brewer(palette = "Set2") + theme_minimal() +
  theme(legend.position="bottom")

ggplot(df10, aes(x = factor(Ano), y = Porcentagem, group = Uso)) +
  geom_line(aes(color = Uso), size = .8) + labs(x = "Ano") +
  scale_color_brewer(palette = "Set2") + theme_minimal()

##-----------------------------------------------------
## Análise dos 29 municípios
##-----------------------------------------------------

df11 <- aggregate(cbind(agr.area, pas.area, flo.area, out.area) ~ ano, data = municipios
                  sum, na.rm = TRUE)

for (i in 1:5) {
  df11[i, 6] <- sum(df11[i, 2:5])
}

df12 <- data.frame(Ano = rep(c(2004, 2008, 2010, 2012, 2014), 4),
                   Porcentagem = c(df11$agr.area,
                                   df11$pas.area,
                                   df11$flo.area,
                                   df11$out.area)/15648289,
                   Uso = c(rep("Agrícola", 5),
                           rep("Pastagem", 5),
                           rep("Matas e florestas", 5),
                           rep("Outros usos", 5)))


d <- ggplot(df12, aes(x = factor(Ano), y = Porcentagem, fill = Uso)) +
  geom_bar(stat = "identity") + labs(x = "Ano") +
  scale_fill_brewer(palette = "Set2") + theme_minimal() +
  theme(legend.position="bottom")

ggplot(df12, aes(x = factor(Ano), y = Porcentagem, group = Uso)) +
  geom_line(aes(color = Uso), size = .8) + labs(x = "Ano") +
  scale_color_brewer(palette = "Set2") + theme_minimal()
```

```
##-------------------------------------------------
## Análise dos 141-29 municipios x 29 municipios
##-------------------------------------------------
b1 <- aggregate(cbind(agr.area, pas.area, flo.area, out.area) ~ ano, data = municip
                  sum, na.rm = TRUE)

for (i in 1:5) {
  b1[i, 6] <- sum(b1[i, 2:5])
}

b2 <- data.frame(Ano = rep(c(2004, 2008, 2010, 2012, 2014), 4),
                   Porcentagem = c(b1$agr.area,
                                   b1$pas.area,
                                   b1$flo.area,
                                   b1$out.area)/15648289,
                   Uso = c(rep("Agrícola", 5),
                           rep("Pastagem", 5),
                           rep("Matas e florestas", 5),
                           rep("Outros usos", 5)))

b3 <- aggregate(cbind(agr.area, pas.area, flo.area, out.area) ~ ano, data = municip
                 sum, na.rm = TRUE)

for (i in 1:5) {
  b3[i, 6] <- sum(b3[i, 2:5])
}

b4 <- data.frame(Ano = rep(c(2004, 2008, 2010, 2012, 2014), 4),
                  Porcentagem = c(b3$agr.area,
                                  b3$pas.area,
                                  b3$flo.area,
                                  b3$out.area)/74676381,
                  Uso = c(rep("Agrícola", 5),
                          rep("Pastagem", 5),
                          rep("Matas e florestas", 5),
                          rep("Outros usos", 5)))
```

```
teste <- rbind(b2, b4)
teste[1:20, 4] <- "29"
teste[21:40, 4] <- "112"

p <- ggplot() + geom_bar(data=teste, aes(x = V4, y = Porcentagem, fill=Uso), stat="ident
  theme_bw() + facet_grid( ~ Ano) + scale_fill_brewer(palette = "Set2") + theme_minimal(
  theme(legend.position="bottom")



##-----------------------------------------------------
## 9 municipios
##-----------------------------------------------------

mun <- c("5103205", "5105580", "5105622", "5107107", "5107248",
         "5107263", "5107263", "5107354", "5108501", "5108600")

grafs <- list()

for (i in 1:9) {
  base <- subset(dados, codigo %in% mun[i])
  base2 <- data.frame(Ano = rep(c(2004, 2008, 2010, 2012, 2014), 4),
                      Porcentagem = c(base$agr.porcentagem,
                                      base$pas.porcentagem,
                                      base$flo.porcentagem,
                                      base$out.porcentagem),
                      Uso = c(rep("Agrícola", 5),
                              rep("Pastagem", 5),
                              rep("Matas e florestas", 5),
                              rep("Outros usos", 5)))
  grafs[[i]] <- ggplot(base2, aes(x = factor(Ano), y = Porcentagem, fill = Uso)) +
    geom_bar(stat = "identity") + labs(x = "Ano") +
    scale_fill_brewer(palette = "Set2") + theme_minimal() +
    theme(legend.position="bottom", axis.text.x = element_text(angle=90, hjust=1),
          axis.title = element_text(size = 10))
}


grid_arrange_shared_legend(grafs[[1]], grafs[[2]], grafs[[3]],
```

```
                          grafs[[4]], grafs[[5]], grafs[[6]],
                          grafs[[7]], grafs[[8]], grafs[[9]],ncol = 3, nrow = 3)

abc <- subset(dados, codigo %in% "5103205")
qplot(factor(ano), agr_perm.areaplantada, data = abc) + geom_bar(stat = "identity")
  scale_fill_brewer(palette = "Set2") + theme_minimal()
qplot(factor(ano), agr_perm.areacolhida, data = abc) + geom_bar(stat = "identity")
  scale_fill_brewer(palette = "Set2") + theme_minimal()
qplot(factor(ano), agr_perm.producao, data = abc) + geom_bar(stat = "identity") + l
  scale_fill_brewer(palette = "Set2") + theme_minimal()


##-----------------------------------------------------
## 20 municipios
##-----------------------------------------------------

codigo_mun$V2

graf <- list()

for (i in 1:20) {
  base <- subset(dados, codigo %in% codigo_mun$V2[i])
  base2 <- data.frame(Ano = rep(c(2004, 2008, 2010, 2012, 2014), 4),
                      Porcentagem = c(base$agr.porcentagem,
                                      base$pas.porcentagem,
                                      base$flo.porcentagem,
                                      base$out.porcentagem),
                      Uso = c(rep("Agrícola", 5),
                              rep("Pastagem", 5),
                              rep("Matas e florestas", 5),
                              rep("Outros usos", 5)))
  graf[[i]] <- ggplot(base2, aes(x = factor(Ano), y = Porcentagem, fill = Uso)) +
    geom_bar(stat = "identity") + labs(x = "Ano") +
    scale_fill_brewer(palette = "Set2") + theme_minimal() +
    theme(legend.position="bottom", axis.text.x = element_text(angle=90, hjust=1, s
          axis.title = element_text(size = 9))
}

grid_arrange_shared_legend(graf[[1]], graf[[2]], graf[[3]], graf[[4]], graf[[5]],
```

```
                          graf[[6]], graf[[7]], graf[[8]], graf[[9]],graf[[10]],
                          graf[[11]],graf[[12]],graf[[13]],graf[[14]], graf[[15]],
                          graf[[16]], graf[[17]], graf[[18]], graf[[19]], graf[[20]],nc


##-----------------------------------------------------
## gráficos de porcentagem observada versus porcentagem predita
##-----------------------------------------------------

##### AGRICULTURA MODELO 1
a <- qplot(training$agr.porcentagem, exponenciais2$agr) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + theme(axis.title = element_text(size = 9))
  xlim(0, 1) + ylim(0, 1)



b <- qplot(dados.test2$agr.porcentagem, exponenciais$agr) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))

multiplot(a, b, cols = 2)

ggsave("x1.pdf", multiplot(a, b, cols = 2))

##### PASTAGEM MODELO 1
a <- qplot(training$pas.porcentagem, exponenciais2$pec) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))



b <- qplot(dados.test2$pas.porcentagem, exponenciais$pec) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))

multiplot(a, b, cols = 2)
```

```
ggsave("x2.pdf", multiplot(a, b, cols = 2))

##### FLORESTAS MODELO 1

a <- qplot(training$flo.porcentagem, exponenciais2$flo) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))


b <- qplot(dados.test2$flo.porcentagem, exponenciais$flo) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))

multiplot(a, b, cols = 2)

ggsave("x3.pdf", multiplot(a, b, cols = 2))

##### AGRICULTURA MODELO 2
a <- qplot(Qtraining$agr.porcentagem, exponenciais4$agr) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))

b <- qplot(dados2.test2$agr.porcentagem, exponenciais5$agr) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))

multiplot(a, b, cols = 2)

ggsave("x4.pdf", multiplot(a, b, cols = 2))

##### PASTAGEM MODELO 2
a <- qplot(Qtraining$pas.porcentagem, exponenciais4$pec) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
```

```
  theme(axis.title = element_text(size = 9))


b <- qplot(dados2.test2$pas.porcentagem, exponenciais5$pec) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))

multiplot(a, b, cols = 2)

ggsave("x5.pdf", multiplot(a, b, cols = 2))

##### FLORESTAS MODELO 2

a <- qplot(Qtraining$flo.porcentagem, exponenciais4$flo) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))


b <- qplot(dados2.test2$flo.porcentagem, exponenciais5$flo) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9))

multiplot(a, b, cols = 2)

ggsave("x6.pdf", multiplot(a, b, cols = 2))

##-----------------------------------------------------
## boxplots training
##-----------------------------------------------------

a <- qplot(factor(codigo), lnagr, data = training) + geom_boxplot() +
  theme_minimal() + labs(x = "Municípios", y = "Ln(Yit) uso agricultura") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank())

ggsave("agrcod.pdf", a)
```

```
b <- qplot(factor(ano), lnagr, data = training) + geom_boxplot() +
  theme_minimal() + labs(x = "Ano", y = "Ln(Yit) uso agricultura")

ggsave("agrano.pdf", b)


##-------------------------------------------------
##Três gráficos
##-------------------------------------------------
##### AGRICULTURA MODELO 2
a <- qplot(Qtraining$agr.porcentagem, exponenciais4$agr) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9), axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8))


b <- qplot(dados2.test2$agr.porcentagem, exponenciais5$agr) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9), axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8))

c <-  qplot(dados2.test3$agr.porcentagem, exponenciais6$agr) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9), axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8))

multiplot(a, b, c, cols = 3)

ggsave("x4.pdf", multiplot(a, b, c, cols = 3))

##### PASTAGEM MODELO 2
a <- qplot(Qtraining$pas.porcentagem, exponenciais4$pec) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9), axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8))
```

```
b <- qplot(dados2.test2$pas.porcentagem, exponenciais5$pec) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9), axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8))

c <-  qplot(dados2.test3$pas.porcentagem, exponenciais6$pec) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9), axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8))

multiplot(a, b, c, cols = 3)

ggsave("x5.pdf", multiplot(a, b, c, cols = 3))

##### FLORESTAS MODELO 2

a <- qplot(Qtraining$flo.porcentagem, exponenciais4$flo) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9), axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8))


b <- qplot(dados2.test2$flo.porcentagem, exponenciais5$flo) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9), axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8))

c <-  qplot(dados2.test3$flo.porcentagem, exponenciais6$flo) +
  labs(x = "Porcentagem observada", y = "Porcentagem predita") +
  theme_minimal() + geom_abline(slope = 1) + xlim(0, 1) + ylim(0, 1) +
  theme(axis.title = element_text(size = 9), axis.text.x = element_text(size=8),
        axis.text.y = element_text(size=8))
```

```
multiplot(a, b, c, cols = 3)

ggsave("x6.pdf", multiplot(a, b, c, cols = 3))


##-----------------------------------------------
## Tabela com análise descritiva dos dados do modelo 2
##-----------------------------------------------

tab <- rbind(Qtraining, dados2.test2)

apply(tab[, 12:27], 2, min)
apply(tab[, 12:27], 2, max)
apply(tab[, 12:27], 2, mean)
apply(tab[, 12:27], 2, sd)
```