UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

CAMILA FERREIRA DA SILVA

Estimação de indicador de agricultura produtiva e sustentável utilizando modelos de pequenas áreas para dados agropecuários no Brasil

Recife

2024

CAMILA FERREIRA DA SILVA

Estimação de indicador de agricultura produtiva e sustentável utilizando modelos de pequenas áreas para dados agropecuários no Brasil

Trabalho apresentado ao Programa de Pós-graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito para obtenção do grau de Mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientadora: Dra. Fernanda De Bastiani

Coorientador: Dr. Cristiano Ferraz

Recife

2024

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Silva, Camila Ferreira da.
Estimação de indicador de agricultura produtiva e sustentável utilizando modelos de pequenas áreas para dados agropecuários no Brasil / Camila Ferreira da Silva. - Recife, 2024.
121f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Programa de Pós-graduação em Estatística, 2024.
Orientação: Fernanda De Bastiani.
Coorientação: Cristiano Ferraz.

1. Amostragem; 2. Desagregação; 3. ODS ONU; 4. Pequena área; 5. PNAgro; 6. Regressão. I. Bastiani, Fernanda De. II. Ferraz, Cristiano. III. Título.

UFPE-Biblioteca Central

CAMILA FERREIRA DA SILVA

ESTIMAÇÃO DE INDICADOR DE AGRICULTURA PRODUTIVA E SUSTENTÁVEL UTILIZANDO MODELOS DE PEQUENAS ÁREAS PARA DADOS AGROPECUÁRIOS NO BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestra em Estatística.

Aprovada em: 22 de agosto de 2024.

BANCA EXAMINADORA

Prof. Dr. Cristiano Ferraz
Presidente/Co-Orientador, UFPE

Profª Drª Maria Cristina Falcão Raposo
Examinadora Interna à Instituição, UFPE

Profª Drª Denise Britz do Nascimento Silva
Examinadora Externa à Instituição, ENCE/IBGE

“A força dos meus sonhos.”

AGRADECIMENTOS

A Deus que realmente a cada dia me fortalece.

Ao alicerce familiar.

A todos os amigos com quem compartilhei essa experiência e colaboraram.

Aos meus orientadores, Fernanda De Bastiani e Cristiano Ferraz, pela oportunidade e compreensão, a confiança e os ensinamentos.

Aos membros da banca examinadora, as professoras Dra. Cristina Raposo e Dra. Denise Britz, pela disponibilidade e pelo empenho dedicado à avaliação da minha dissertação.

A todos os professores entre componentes curriculares e extracurriculares da Universidade Federal de Pernambuco, essenciais no meu crescimento, solícitos e empenhados na aprendizagem dos alunos.

A Alethea Gabriela Candia Calderón, Consultora da Organização das Nações Unidas para a Alimentação e a Agricultura - FAO Regional para a América Latina e o Caribe, e Maxwell Merçon Tezolin Barros de Almeida, Gerente Técnico do Censo Agropecuário, da Coordenação de Estatísticas Agropecuárias (COAGRO/IBGE), ambos pelo papel desempenhado e esforço na efetuação do cálculo de indicador de agricultura produtiva e sustentável utilizando dados do Censo Agropecuário 2017.

Ao Octavio Costa de Oliveira, Coordenador da COAGRO/IBGE - Coordenação de Estatísticas Agropecuárias, e ao Michael Rahija, Estatístico Regional da FAO Regional para a América Latina e o Caribe, pelo reconhecimento à importância desse trabalho e permissão de acesso aos dados dos indicadores 2.3.1 e 2.3.2, fundamentais para o desenvolvimento dessa dissertação.

A professora Dra. Andrea Diniz da Silva pela recepção e hospitalidade, e o ambiente acadêmico da Escola Nacional de Ciências Estatísticas (ENCE/IBGE), fundamentais para enriquecer minha experiência e expandir meu conhecimento.

A Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE) pela concessão da mobilidade de auxílio financeiro VIII.AMD - Auxílio de Mobilidade Discente.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, pelo apoio financeiro.

RESUMO

Os indicadores dos Objetivos de Desenvolvimento Sustentável (ODS) foram criados para monitorar avanços relativos à metas estabelecidas. O ODS 2 da Organização das Nações Unidas, cujo propósito é atingir um nível de Fome Zero e Agricultura Sustentável até 2030, tem dentre suas metas a 2.3, que preconiza dobrar a produtividade agrícola e a renda dos pequenos produtores de alimentos, particularmente das mulheres, dos povos indígenas, agricultores familiares, pastores e pescadores. Esta meta é monitorada por dois indicadores, um deles é o 2.3.1, definido pelo volume de produção por unidade de trabalho por dimensão do estabelecimento agrícola, pastoril e florestal. Essa dissertação se propõe a estudar a viabilidade de estimação do indicador 2.3.1 para domínios subnacionais com nível de desagregação municipal. Embora o Brasil ainda não possua uma Pesquisa Nacional Agropecuária (PNAgro), estudos têm sido realizados ao longo dos anos para tornar possível a realização de tal levantamento, suprindo uma necessidade de estatísticas agropecuárias nacionais entre censos agropecuários, amparada por método probabilístico de amostragem, que há muito tem sido sentida. Prover estimativas com nível de desagregação avançado, como o municipal, raramente está dentre os objetivos de levantamentos nacionais, como uma PNAgro, por razões orçamentárias. No entanto, a disponibilidade de estudos que possibilitem a identificação de modelos de estimação de pequenas áreas com potencial para gerar estatísticas agropecuárias municipais representa ganho metodológico que vem a acrescentar aos diversos motivos pelos quais o Brasil se beneficiaria com uma PNAgro. O estudo apresentado nessa dissertação é uma iniciativa de contribuição nesta direção, na medida em que demonstra uma situação de estimação do indicador 2.3.1 tendo como base dados de uma amostra aleatória nacional de município e uma análise específica para os municípios de Pernambuco. Os resultados obtidos permitiram a identificação de um conjunto de dados auxiliares promissor, bem como de um modelo de regressão factível para a estimação de pequenas áreas, com base em informações de uma amostra agropecuária nacional simulada.

Palavras-chaves: Amostragem. Desagregação. ODS ONU. Pequena área. PNAgro. Regressão.

ABSTRACT

Indicators of the Sustainable Development Goals (SDG) are created to monitor progress relative to established targets. The United Nations'SDG 2 whose purpose is to achieve a level of Hanger Zero and Sustainable Agriculture, by 2030 has the 2.3 among its targets, that precognises double the agricultural productivity and the incomes of small-scale food producers, particularly women, indigenous peoples, family farmers, pastoralists and fishers. Such Target is monitored by two indicators, one of them, the 2.3.1, defined by the volume of production per labor unit by classes of farming/pastoral/forestry enterprise size. This thesis' goal is to study the viability of estimating indicator 2.3.1 for subnational domains with municipality's disaggregation level. Although there is not yet a National Agricultural Survey (PNAgro) in Brazil, studies have been carried out along years to evaluate its feasibility in practice. Such a survey would supply a long-lasting need for national agricultural statistics supported by probabilistic sampling methods, between Brazilian agricultural censuses. Furnishing disaggregated estimates at advanced level, such as by municipality, seldom is among the goals of national surveys such as a PNAgro, due to budget limitations. However, the availability of studies that can identify the potentiality of using small area estimation models to generate municipal agricultural statistics, represents a methodological gain that adds to the many reasons Brazil can benefit from a PNAgro survey. The study introduced in this thesis hopes to represent a contribution in that direction, as it demonstrates a 2.3.1 SDG indicator estimation scenario based on a random national sample of municipalities and a specific analysis for the municipalities of Pernambuco. The results led to the assembling of a promising auxiliary database as well as the identification of a suitable regression model to furnish small area estimates based on a simulated national agricultural sample.

Keywords: Sampling. Disaggregation. UN SDG. Small area. PNAgro. Regression.

LISTA DE FIGURAS

LISTA DE CÓDIGOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A Organização das Nações Unidas (ONU) no mês de setembro de 2015 propôs os Objetivos de Desenvolvimento Sustentável (ODS) como um plano de ação global focado em 17 objetivos. A princípio, conforme apresentado no relatório global de índice e painéis de ODS (SACHS et al., 2016), a Comissão de Estatística da ONU definiu 169 metas com 231 indicadores para monitorá-las no intuito de erradicar a extrema fome e pobreza, proporcionar bem-estar e educação e proteger o planeta, com o prazo até o ano de 2030. O objetivo primordial *Leave No One Behind* (LNB) é a premissa fundamental na implementação e no progresso da Agenda 2030 para o Desenvolvimento Sustentável. Anualmente milhões de dólares são investidos na Agropecuária em países de baixo e médio rendimento e ainda existe uma escassez de dados precisos relacionados aos principais aspectos e fatores desse setor. Consequentemente, importantes decisões são efetuadas sobre esses recursos sem qualquer base de informação sólida. O ODS 2 (Fome Zero e Agricultura Sustentável) oferece uma oportunidade para analisar e mitigar o problema de dados na Agropecuária. Todavia, dados agrícolas e pecuaristas de boa qualidade são essenciais para alcançar e medir o progresso em direção a este objetivo.

É pragmático o intuito do 2º objetivo sendo acabar com a fome, promover agricultura sustentável e compreende a segurança alimentar e melhorias nas condições de nutrição. Entre as suas 8 metas com 14 indicadores, é a meta 2.3 que está relacionada em dobrar a produtividade agrícola e a renda dos pequenos produtores de alimentos, particularmente das mulheres, povos indígenas, agricultores familiares, pastores e pescadores, inclusive por meio de acesso seguro e igual à terra, outros recursos produtivos e insumos, conhecimento, serviços financeiros, mercados e oportunidades de agregação de valor e de emprego não agrícola. Ademais, garantir que os sistemas de produção de alimentos atinjam sustentabilidade a partir de práticas resilientes.

Estimar determinados indicadores de abrangência nacional, por exemplo um indicador da meta 2.3, o 2.3.1 que se refere ao volume de produção por unidade de trabalho por dimensão da empresa agrícola/pastoril/florestal, exige considerar o efeito de planos amostrais complexos utilizados em pesquisas por amostragem probabilística que originam tais dados agropecuários. Tal efetividade também depende da capacidade do país gerar estimativas dos indicadores para domínios subnacionais, como o municipal. No entanto, obter estimativas para determinados níveis de desagregação pode representar um desafio quando os respectivos tamanhos de amostra são muito pequenos ou até nulos. Nessas ocasiões, é possível a utilização de modelos de

estimação de pequenos domínios, também chamados de modelos de estimação de pequenas áreas (SAE - *small area estimation*), para auxiliar na elaboração de estimativas mais precisas.

Entre os métodos SAE estão as possíveis abordagens de estimação indireta ou sintética associadas ao nível de unidade ou de área, em que as técnicas usadas permitem combinar os dados de pesquisa com informações auxiliares provenientes de fontes de dados adicionais. Perante os procedimentos de estimação indireta em nível de área nesse estudo são apresentadas as abordagens baseadas em um modelo de regressão (*model-based*). Sobretudo, a explanação a partir de um modelo linear simples e pelo método de Fay-Herriot (FH), e a conjectura de modelos com base em uma componente espacial, como um modelo geoaditivo.

O Brasil ainda não possui uma Pesquisa Nacional Agropecuária (PNAgro), mas vários estudos têm sido realizados ao longo dos anos para tornar possível a concretização de tal levantamento, que viria a suprir uma necessidade de estatísticas agropecuárias nacionais entre censos agropecuários, tendo como base um método probabilístico de amostragem. Gerar estimativas com nível de desagregação avançado, como o municipal, raramente está dentre os objetivos de uma Pesquisa como uma PNAgro, por limitações orçamentárias. No entanto, a disponibilidade de estudos como o descrito nessa dissertação, que possibilitem identificar modelos de estimação de pequenas áreas com potencial para gerar estatísticas agropecuárias municipais, representa ganho metodológico que vem a acrescentar aos diversos motivos pelos quais o país se beneficiaria com uma PNAgro. O estudo apresentado nessa dissertação é uma iniciativa de contribuição nessa direção, na medida em que enseja estudar, na condição do que poderia ser viável em um contexto próximo a uma PNAgro, a estimação de um indicador de agricultura produtiva e sustentável, especificamente o 2.3.1, para domínios subnacionais com nível de desagregação municipal. Conforme relata o Instituto Brasileiro de Geografia e Estatística (IBGE), até o momento em que essa dissertação foi defendida, o indicador 2.3.1 ainda está em análise/construção neste país.

O desenvolvimento dessa dissertação é com base em uma circunstância experimental, criada para simular aspectos reais de estimação do indicador 2.3.1 na região Nordeste do Brasil, em particular, no estado de Pernambuco. Tal circunstância consiste na seleção de uma amostra aleatória simples de municípios do Brasil, tendo como cadastro a base de dados municipais fornecida pela Organização das Nações Unidas para Alimentação e Agricultura (FAO - *Food and Agriculture Organization*). Assim, se pretende gerar uma situação de estimação de pequenos domínios, a partir de um contexto similar ao de uma pesquisa amostral de âmbito nacional, como uma PNAgro, guardadas as devidas proporções de simplicidade do plano amostral rea-

lizado. Nesse estudo, no âmbito dos municípios do estado de Pernambuco, estimativas para os domínios são geradas a partir de um modelo de regressão linear simples e do modelo FH, baseados na presença de variávies auxiliares (explicativas, independentes ou preditivas) provenientes do último censo agropecuário e de fontes de dados confiáveis e consolidadas, como da FAO, da Fundação Oswaldo Cruz (Fiocruz) e do Instituto Militar de Engenharia (IME). Por conseguinte, os resultados obtidos permitiram o agrupamento de um conjunto de variáveis auxiliares promissor e um modelo de regressão factível para a estimação de pequenas áreas.

O perfil integrativo da dissertação mediante uma característica multidisciplinar remete também a difundir o estudo imparcial no crescimento de concepções acerca das Tecnologias da Geoinformação e a área de conhecimento da Estatística e contribuir na estimação ou na análise dos indicadores de sustentabilidade da ONU no cumprimento da Agenda 2030, útil à sociedade brasileira e eminente ao estado de Pernambuco.

1.1 JUSTIFICATIVA

Mundialmente é admitida a ausência de informações dos indicadores ODS e estritamente relativa ao setor Agropecuário afeta pelo menos 800 milhões de pessoas de forma indireta, uma vez que importantes decisões no âmbito são realizadas sem as devidas evidências empíricas. Tais evidências, quando disponíveis, têm maior alcance na medida em que é possível gerar estimativas desagregadas por diversas áreas (domínios).

Entretanto, a exigência de estatísticas desagregadas impõe a necessidade de diversos aspectos e recursos, pois requer mais informações e dados para representar adequadamente os subgrupos populacionais nos cálculos subjacentes. No contexto de estruturação dos indicadores, a desagregação pode demandar tamanhos de amostras maiores. Assim, se o procedimento probabilístico através do plano de amostragem não puder fornecer estimativas precisas, devido ao tamanho da amostra ser pequeno naquela área de interesse, com a aplicação das técnicas SAE existe a possibilidade de produzir resultados precisos.

Em uma conjuntura global retratando o ODS 2, esses seguimentos fornecem *insights* valiosos para políticas agrícolas eficazes, proporcionando uma compreensão detalhada das demandas e realidades típicas de comunidades rurais e agricultores familiares.

No Brasil, apesar dos avanços proporcionados pelo Sistema Nacional de Pesquisa Agropecuária (SNPA), a criação de uma PNAgro ainda é necessária para consolidar e ampliar os esforços de pesquisa e desenvolvimento. Uma PNAgro poderia centralizar dados, otimizar re-

cursos e permitir o alinhamento de estratégias com as necessidades do setor Agropecuário. Atualmente, o sistema de pesquisas agropecuárias contínuas do IBGE é delineado em função de levantamentos cadastrais e subjetivos (COAGRO, 2011). As pesquisas cadastrais investigam painéis das unidades especializadas apurando dados de produção pecuária e estoques agrícolas, enquanto as subjetivas abrangem produção agrícola, pecuária, silvicultura e extrativismo vegetal, coletados de forma indireta em consultas com especialistas ou registros administrativos em que a unidade de investigação é o município. Pelo fato de não serem estruturadas por um processo de amostragem probabiblística, essas pesquisas não possuem medida de precisão ou estimativas de erro. Esse aspecto se soma à heterogeneidade nos modos de aquisição dos dados estatísticos pelos agentes de coleta, potencialmente sujeitos a vieses.

O país possui uma grande diversidade socioeconômica que através dessas técnicas estatísticas pode garantir que intervenções e investimentos direcionados à este ramo atendam de maneira concludente e equitativa as diferentes realidades regionais, ao gerar dados precisos. Em nível estadual, como é o caso do estado de Pernambuco, a obtenção de estimativas precisas é mais vital, pois permite uma alocação eficiente de recursos e políticas direcionadas às premências dos agricultores locais, considerando as peculiaridades climáticas, geográficas e socioeconômicas da região Nordeste.

1.2 OBJETIVO GERAL

O objetivo geral dessa dissertação é estudar a viabilidade de estimação do indicador 2.3.1 para domínios subnacionais com nível de desagregação municipal, considerando a circunstância experimental que simula pequenos domínios a partir de uma amostra aleatória simples nacional de municípios.

1.2.1 Objetivos Específicos

Para alcançar o objetivo geral da pesquisa, viabilizando a estruturação e execução de maneira adequada dessa dissertação, foram definidos os seguintes objetivos específicos:

- Elaborar uma revisão de literatura sobre modelos de estimação de pequenas áreas;
- Produzir uma análise descritiva de dados provenientes das estimativas do indicador 2.3.1;

- Construir uma base de dados consolidada com informações auxiliares para o Brasil, provenientes de diversas fontes com potencial de uso em modelos de estimação de pequenas áreas para estatísticas agropecuárias;
- Investigar ao menos dois modelos de regressão com potencial para gerar estimação de pequenas áreas no contexto brasileiro e efetuar a modelagem para uma região específica;
- Analisar as estimativas do indicador 2.3.1 para o estado de Pernambuco no cenário de simulação de pequenas áreas obtidas da região Nordeste.

2 ESTRUTURA DA DISSERTAÇÃO

A estruturação da dissertação dispõe uma visão geral dos capítulos e seções que compõem o trabalho, desde a contextualização inicial até a análise final dos resultados.

Inicialmente, é contextualizada a importância da pesquisa no âmbito dos ODS propostos pela ONU, especialmente da meta 2.3 do ODS 2, que se relaciona com a produtividade agrícola e a renda dos pequenos produtores. Com enfoque no indicador 2.3.1 que mede a produtividade agrícola por unidade de trabalho.

A seção Abordagens de Estimação em Pequenas Áreas apresenta uma sinopse de métodos de estimação direta e indireta aplicáveis para obter estimativas para pequenos domínios. E explora o uso de modelos para combinar dados de pesquisa com informações auxiliares.

Em materiais e métodos são relatadas as ferramentas computacionais e abordagens usadas na obtenção dos resultados. Também são retratados: a análise exploratória de dados populacional e amostral, o conjunto de variáveis preditoras e técnicas automáticas de seleção, o ajuste de modelo estatístico pelo estimador sintético de regressão e modelo Fay-Herriot.

Especificamente, no levantamento das observações se destaca a relevância de entender o fenômeno em estudo e como algumas informações auxiliares podem contribuir para uma análise consolidada desse desempenho produtivo e auxiliar na estimação do indicador. Em seguida, a ênfase na caracterização do estado de Pernambuco, à levar em conta particularidades locais que podem ser ocultadas em uma análise mais abrangente.

Após essas elucidações, são expostos os resultados com discussão sobre as implicações desses apuramentos. Em seguida, as considerações finais fornecem um resumo dos principais efeitos da pesquisa, as conclusões do estudo e sugestões para pesquisas futuras. É incluído o referencial bibliográfico, assegurando a fundamentação teórica e metodológica do estudo. Por fim, nos apêndices encontram-se o código computacional desenvolvido na obtenção dos resultados e os gráficos de componentes residuais no modelo linear simples ajustado, e o anexo A contém a descrição das variáveis da base de dados auxiliares.

3 OS OBJETIVOS DE DESENVOLVIMENTO SUSTENTÁVEL

Os Objetivos de Desenvolvimento Sustentável representam uma abordagem holística e abrangente para enfrentar os desafios globais mais prementes, estabelecendo uma agenda compartilhada para o desenvolvimento sustentável até 2030. Ao reconhecer a complexidade e a interdependência dos desafios enfrentados pela humanidade, esses objetivos promovem uma abordagem integrada e colaborativa, envolvendo governos, sociedade civil, setor privado e outras partes interessadas em um esforço conjunto para alcançar metas ambiciosas e transformadoras. Além disso, representam não apenas um chamado à ação, mas também um compromisso moral e ético com as gerações presentes e futuras.

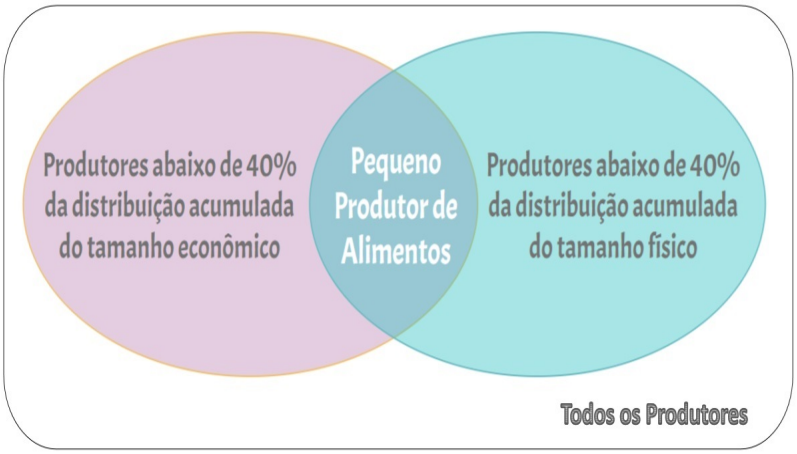
3.1 A META 2.3

Diante do ODS 2 (Fome Zero e Agricultura Sustentável), os indicadores 2.3.1 e 2.3.2 são específicos da meta 2.3 que consiste em, segundo a ONU até 2030, dobrar a produtividade agrícola e a renda dos pequenos produtores de alimentos, particularmente das mulheres, povos indígenas, agricultores familiares, pastores e pescadores, inclusive por meio de acesso seguro e igual à terra, outros recursos produtivos e insumos, conhecimento, serviços financeiros, mercados e oportunidades de agregação de valor e de emprego não agrícola. Entretanto, conforme destaca o Instituto de Pesquisa Econômica Aplicada (IPEA), no Brasil se tem uma adequação, a de aumentar a produtividade agrícola e a renda dos pequenos produtores de alimentos, particularmente de mulheres, agricultores familiares, povos e comunidades tradicionais, visando tanto à produção de autoconsumo e garantia da reprodução social dessas populações quanto ao seu desenvolvimento socioeconômico, por meio do acesso seguro e equitativo: i) à terra e aos territórios tradicionalmente ocupados; ii) à assistência técnica e extensão rural, respeitando-se as práticas e saberes culturalmente transmitidos; iii) a linhas de crédito específicas; iv) aos mercados locais e institucionais, inclusive políticas de compra pública; v) ao estímulo ao associativismo e cooperativismo; e vi) a oportunidades de agregação de valor e emprego não-agrícola (IPEA, 2019).

A definição do pequeno produtor de alimentos é uma combinação de dois critérios, a dimensão física expressa pela quantidade de terras exploradas e/ou pelo número de cabeças de gado, e a dimensão econômica da exploração expressa pelo valor total de produção agrícola

(KHALIL; CANDIA, 2023) (figura 1). Ambos os critérios são aplicados em termos relativos, a fim de melhorar a comparabilidade internacional. A FAO propôs um conceito global de pequenos produtores alimentares com o objetivo de calcular números internacionalmente comparáveis para todos os países, regiões e territórios.

Figura 1 – Identificação do pequeno produtor de alimentos



Fonte: Adaptada de KHALIL; CANDIA (2023)

Em suma, o indicador 2.3.1 visa relacionar o volume de produção ao esforço de trabalho empregado e à dimensão da empresa, em que a produtividade se define através da relação entre os produtos obtidos e a quantidade de insumos utilizados. Ou seja, a sua estrutura tem o formato de uma razão que é interpretada como o volume de produção por unidade de trabalho por dimensão da empresa agrícola/pastoril/florestal sendo a quantidade média que pode ser produzida em um determinado tempo:

$$\mathsf{I}_{2.3.1} = \frac{\sum_{j=1}^{N_a} \left(\frac{\sum_k V_{kj}^t \, p_{kj}^t}{Q_j^t} \right)}{N_a} \, ,$$

em que o cálculo é elaborado para um determinado ano t sendo da produção de um produto k do pequeno produtor de alimentos j. Desse modo, V_{kj}^t representa o volume físico do produto k vendido pelo pequeno produtor de alimentos j durante o ano t, p_{kj}^t é o preço constante de venda recebido pelo produtor j do produto k no ano t, Q_j^t é a quantidade de dias trabalhados por j em t. N_a é o número de pequenos produtores de alimentos (j é apenas um deles), se refere a entidades de produção (como exploração agrícola) e não a trabalhores individuais.

O segundo indicador da meta 2.3, o 2.3.2 calcula a renda média dos pequenos produtores de alimentos, por sexo e condição de indígena, sendo um aspecto fundamental para que esses produtores tenham acesso a um estilo de vida sustentável e decente. Com base nas

informações apresentadas do indicador 2.3.1 sobre a produtividade agrícola por unidade de trabalho, o indicador 2.3.2 pode ser descrito da seguinte maneira:

$$\mathsf{I}_{2.3.2} = \frac{\sum_{j=1}^{N_a} (\sum_k (V_{kj}^t \; p_{kj}^t - C_{kj}^t))}{N_a} \; ,$$

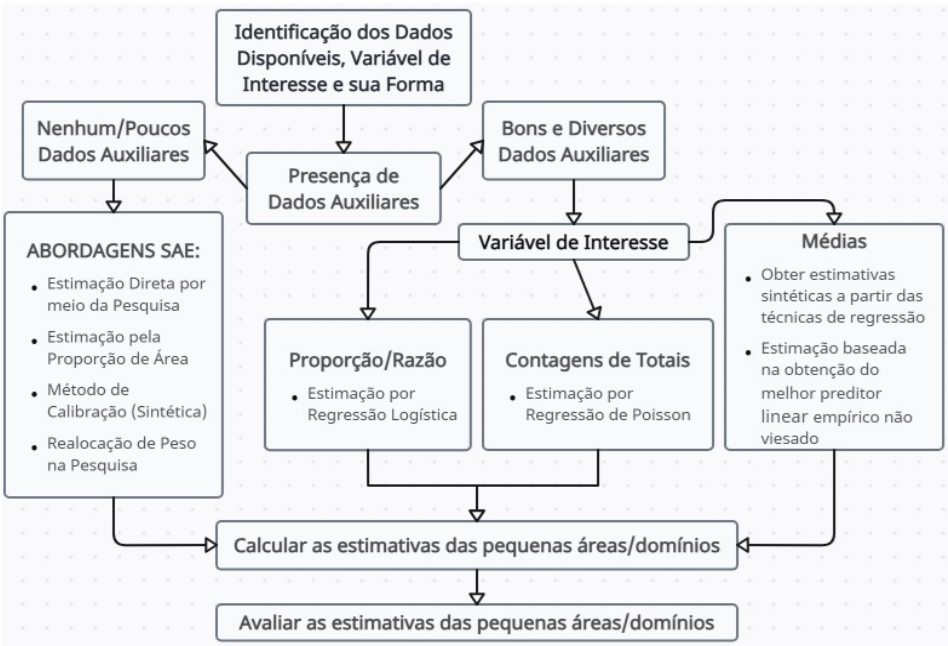
em que C_{kj}^t é o custo de produção do produto k especificado pelo pequeno produtor de alimentos j durante o ano t.

4 ABORDAGENS DE ESTIMAÇÃO EM PEQUENAS ÁREAS

A estimação de pequenas áreas está atrelada à produção de estimativas mais precisas, a exemplificar o resultado de médias, contagens e quantis para áreas em que se tem pequenos tamanhos de amostras ou nenhuma amostra está diponível. Ou seja, não é o tamanho da área que causa o possível problema de estimação, mas o tamanho amostral sendo insufiente para efetuar inferências satisfatórias apenas com os dados da pesquisa. Pequenos domínios incluem tanto uma região geográfica (distrito, setor censitário, município) sendo a componente ao nível de área, quanto um grupo demográfico (idade, gênero, raça, escolaridade) sendo a componente ao nível de unidade, entre outros.

Usualmente, as abordagens adotadas podem ser divididas em concordância com a proveniência dos dados. Uma abordagem direta usa apenas as observações que foram coletadas de uma pesquisa por amostragem em que não se têm dados auxiliares ou a presença deles é escassa, e está atribuída ao método baseado em um plano amostral (*design-based*). Enquanto uma abordagem indireta utiliza observações e dados auxiliares também em modelos matemáticos e estatísticos, a qual segue demais ramificações seja por um método baseado em modelo ou assistido por um modelo (*model based* ou *model assisted*). Sucintamente, para ilustrar com clareza a sequência das etapas envolvidas em um encadeamento SAE (*small area estimation*), a figura 2 representa abordagens que podem ser efetuadas .

Figura 2 – Etapas de um processo SAE



Fonte: Adaptada de ADB (2022)

4.1 ABORDAGEM DIRETA

Considere U como o conjunto de elementos da população-alvo, de tamanho N, para a qual existem D domínos de interesse. Defina U_{d} como o conjunto de elementos de U pertencentes ao domínio d, tal que U_{d} possui N_{d} elementos. O conjunto U é definido como:

$$U = \bigcup_{\mathrm{d}=1}^{\mathrm{D}} U_{\mathrm{d}}$$

e

$$\mathrm{N} = \sum_{\mathrm{d}=1}^{\mathrm{D}} \mathrm{N}_{\mathrm{d}} \ .$$

Pesquisas agropecuárias nacionais frequentemente tem como população-alvo o conjunto de todos os produtores agropecuários de um país. No caso do Brasil, uma Unidade da Federação (UF), como o estado de Pernambuco, pode ser um exemplo de domínio de interesse. Neste caso, U_{d} representaria o conjunto de todos os produtores agropecuários do Brasil, que estão localizados em Pernambuco.

Parâmetros de interesse relacionados a um domínio U_{d} são funções de variáveis de interesse observadas em cada elemento $i \in U_{\mathrm{d}}$. Médias, totais, proporções e índices como os abordados nessa dissertação, são alguns exemplos de parâmetros para os quais é comum se desejar gerar estimativas para a população-alvo U (nacionais) e para domínios U_{d} (UF).

Imagine que um plano de amostragem probabilística seja utilizado para selecionar uma amostra S, de tamanho n, dentre os N elementos de U, como por exemplo, uma amostra aleatória simples (AAS) sem reposição. Considere, para efeito de ilustração, que exista interesse em estimar um total populacional para uma variável y de interesse, em um domínio d, estabelecido por:

$$Y_{\mathrm{d}} = \sum_{i \in U_{\mathrm{d}}} y_i \ ,$$

em que y_i é o valor da variável de interesse y observado no elemento i.

A abordagem de estimação direta consiste em estimar parâmetros como o total populacional Y_{d} usando apenas a informação dos valores de y observados nos elementos selecionados para a amostra S que pertencem ao domínio d. Defina

$$\mathrm{S}_{\mathrm{d}} = \mathrm{S} \cap U_{\mathrm{d}} \ ,$$

e uma variável indicadora de pertencimento ao domínio $\delta_{i\mathrm{d}} = 1$, se $i \in U_{\mathrm{d}}$, e $\delta_{i\mathrm{d}} = 0$, caso contrário. Defina ainda uma variável indicadora de inclusão na amostra $\mathrm{I}_i = 1$, se $i \in \mathrm{S}$, e

$\mathrm{I}_i = 0$, caso contrário. A variável aleatória I_i é tal que:

$$\mathrm{I}_i \sim Bernoulli(\pi_i) \ ,$$

com $\pi_i = P(\mathrm{I}_i = 1)$ sendo a probabilidade do elemento i ser incluído na amostra S. O estimador de Horvitz-Thompson (H-T) para Y_d é dado por:

$$\hat{Y}_\mathrm{d} = \sum_{i \in \mathrm{S}} \frac{y_i}{\pi_i} \delta_{i\mathrm{d}} = \sum_{i \in \mathrm{S_d}} \frac{y_i}{\pi_i} \ ;$$

É útil escrever $\hat{Y}$ em função das variáveis indicadoras de inclusão na amostra, quando se deseja demonstrar as propriedades deste estimador. Nesse caso, tem-se:

$$\hat{Y}_\mathrm{d} = \sum_{i \in U} \frac{y_i}{\pi_i} \mathrm{I}_i \delta_{i\mathrm{d}} = \sum_{i \in U_\mathrm{d}} \frac{y_i}{\pi_i} \mathrm{I}_i \ .$$

Ao submeter $\mathrm{E_p}(.)$ como o valor esperado com base no plano amostral p, verifica-se, por exemplo, que $\hat{Y}_\mathrm{d}$ é centrado para Y_d:

$$\mathrm{E_p}(\hat{Y}_\mathrm{d}) = \sum_{i \in U} \frac{y_i}{\pi_i} \delta_{i\mathrm{d}} \mathrm{E_p}(\mathrm{I}_i) = \sum_{i \in U} \frac{y_i}{\pi_i} \delta_{i\mathrm{d}} \pi_i = \sum_{i \in U} y_i \delta_{i\mathrm{d}} = \sum_{i \in U_\mathrm{d}} y_i = Y_\mathrm{d} \ .$$

É possível mostra ainda que a variância de $\hat{Y}_\mathrm{d}$ é dada por:

$$\mathrm{Var_p}(\hat{Y}_\mathrm{d}) = \sum_{i \in U} \sum_{l \in U} (\pi_{il} - \pi_i \pi_l) \frac{y_i}{\pi_i} \frac{y_l}{\pi_l} \mathrm{I}_i \delta_{i\mathrm{d}} \mathrm{I}_l \delta_{l\mathrm{d}} = \sum_{i \in U_\mathrm{d}} \sum_{l \in U_\mathrm{d}} (\pi_{il} - \pi_i \pi_l) \frac{y_i}{\pi_i} \frac{y_l}{\pi_l} \ ,$$

com $\pi_{il} = P(\mathrm{I}_i = 1, \mathrm{I}_l = 1)$. Além disso, o seguinte estimador é centrado para $\mathrm{Var_p}(\hat{Y}_\mathrm{d})$:

$$\hat{\mathrm{Var}}_\mathrm{p}(\hat{Y}_\mathrm{d}) = \sum_{i \in \mathrm{S}} \sum_{l \in \mathrm{S}} \frac{(\pi_{il} - \pi_i \pi_l)}{\pi_{il}} \frac{y_i}{\pi_i} \frac{y_l}{\pi_l} \delta_{i\mathrm{d}} \delta_{l\mathrm{d}} = \sum_{i \in \mathrm{S_d}} \sum_{l \in \mathrm{S_d}} \frac{(\pi_{il} - \pi_i \pi_l)}{\pi_{il}} \frac{y_i}{\pi_i} \frac{y_l}{\pi_l} \ .$$

Estimativas para um parâmetro geral Θ_d, provenientes de um estimador de Horvitz-Thompson, são chamadas de estimativas diretas. Nessa dissertação, considera-se a notação $\hat{\Theta}_\mathrm{d}^{dir}$ para enfatizar quando se está fazendo uso de tal estimador.

4.2 ABORDAGEM INDIRETA

Especificamente, toda a metodologia SAE baseada em um modelo de regressão é uma abordagem indireta. De antemão, as estimações sob pequenas áreas compartilham a mesma estrutura de notação de uma amostra de população, embora com distintas especificações (KHALIL; CANDIA, 2023). Logo, convém considerar U a população finita de N elementos, particionados em D domínios, $U_1, U_2, \ldots, U_\mathrm{D}$ de tamanho $\mathrm{N}_1, \mathrm{N}_2, \ldots, \mathrm{N}_\mathrm{D}$ respectivamente.

Em seguida, Ω representa o conjunto de todas as possíveis amostras probabilísticas s de tamanho n que advém de U $(\mathrm{s} \in \Omega)$ associadas à probabilidade de seleção da amostra $p(\mathrm{s})$, em d denotando o d-ésimo domínio desagregado.

Os métodos tradicionais de estimadores indiretos, como os estimadores sintéticos por regressão são baseados em modelos implícitos. A estimação sintética por regressão é uma técnica apropriada para variáveis dependentes quantitativas, onde os estimadores tomam emprestado força ao sintetizar dados de áreas diferentes. É a específica relação linear da variável dependente e as variáveis auxiliares, quanto mais estiverem correlacionadas, mais eficiente deve ser o estimador. Assim, os parâmetros de pequenas áreas são estimados por meio de uma regressão linear múltipla, assumindo que a variável de interesse depende linearmente das variáveis auxiliares ou covariáveis (ADB, 2022). A relação do estimador sintético por regressão $\hat{\Theta}_{\mathrm{d}}^{ST}$, do estimador direto $\hat{\Theta}_{\mathrm{d}}^{dir}$ e das variáveis independentes ao nível de área é baseada no seguinte modelo:

$$\Theta_{\mathrm{d}} = \beta_0 + \mathbf{x}_{\mathrm{d}}'\beta_1 + \upsilon + \epsilon \ , \quad \upsilon + \epsilon \overset{\text{i.i.d.}}{\sim} (0, \sigma_\upsilon^2 + \sigma_\epsilon^2) \ ,$$

em que Θ_{d} é o vetor da variável resposta ou do parâmetro de interesse, β os coeficientes de regressão, $\mathbf{x}'$ a matriz das variáveis explicativas, υ é o resíduo em nível de área e ϵ o termo residual agrupado dentro da área. Logo, de acordo com o ajuste de um modelo de regressão às covariáveis por área a partir dos vetores $\hat{\beta}$ dos estimadores de quadrados mínimos ponderados para calcular os coeficientes de regressão, um estimador sintético por regressão é expresso por:

$$\hat{\Theta}_{\mathrm{d}}^{ST} = \hat{\beta}_0 + \mathbf{x}_{\mathrm{d}}'\hat{\beta} \ .$$

É constatado que o estimador sintético por regressão é essencialmente um estimador não centrado (viciado, viesado ou tendencioso). Esses tipos de estimadores têm uma grande influência de informações de outras áreas, portanto podem ter pequena variância, mas um grande viés. Mesmo assim, quando a pequena área não apresenta fortes efeitos individuais em relação aos coeficientes de regressão, o estimador sintético será eficiente, com um pequeno erro quadrático médio (EQM) (RAHMAN, 2008). Uma estruturação para o EQM de $\hat{\Theta}_{\mathrm{d}}^{ST}$ é apresentada por Molina e Rao (2015) como:

$$\mathrm{EQM}_{\mathrm{p}}(\hat{\Theta}_{\mathrm{d}}^{ST}) = \mathrm{E}_{\mathrm{p}}(\hat{\Theta}_{\mathrm{d}}^{ST} - \hat{\Theta}_{\mathrm{d}}^{dir})^2 - \mathrm{Var}_{\mathrm{p}}(\hat{\Theta}_{\mathrm{d}}^{ST} - \hat{\Theta}_{\mathrm{d}}^{dir}) + \mathrm{Var}_{\mathrm{p}}(\hat{\Theta}_{\mathrm{d}}^{ST}) \ . \tag{4.1}$$

Diante da equação 4.1, Molina e Rao (2015) determinam uma aproximação amparada em um estimador centrado para o EQM de $\hat{\Theta}_{\mathrm{d}}^{ST}$ sendo $\hat{\mathrm{EQM}}_{\mathrm{p}}(\hat{\Theta}_{\mathrm{d}}^{ST}) \approx (\hat{\Theta}_{\mathrm{d}}^{ST} - \hat{\Theta}_{\mathrm{d}}^{dir})^2 - \hat{\mathrm{Var}}_{\mathrm{p}}(\hat{\Theta}_{\mathrm{d}}^{dir})$.

Contudo, esse estimador pode ser instável (pequenas alterações nos dados podem causar grandes alterações nos valores previstos) e também fornecer resultados negativos.

Em contrapartida, um estimador composto é uma soma ponderada entre o estimador direto e o sintético sendo uma opção à escolha de um em detrimento do outro para equilibrar o grau do viés. O autor Rahman (2008) destaca que a estimação composta é uma maneira natural de equilibrar o potencial viés de um estimador sintético contra a instabilidade de um estimador direto, escolhendo um peso apropriado ξ que esteja entre 0 e 1. Ou seja, a precisão do estimador pode ser melhorada com a seleção de um peso ξ. O estimador composto $\hat{\Theta}_{\mathrm{d}}^{COM}$ de um pequeno domínio pode ser delineado por:

$$\hat{\Theta}_{\mathrm{d}}^{COM} = \xi\hat{\Theta}_{\mathrm{d}}^{dir} + (1-\xi)\hat{\Theta}_{\mathrm{d}}^{ST}, \quad 0 < \xi < 1 \ .$$

O EQM de $\hat{\Theta}_{\mathrm{d}}^{COM}$ é apresentado conforme a autora Molina (2019) define:

$$\mathrm{EQM}(\hat{\Theta}_{\mathrm{d}}^{COMP}) \approx \xi^2 \mathrm{Var}(\hat{\Theta}_{\mathrm{d}}^{dir}) + (1-\xi)^2 \mathrm{EQM}(\hat{\Theta}_{\mathrm{d}}^{ST}) \ .$$

O peso ξ consegue ser uma compensação entre o viés do estimador sintético e a variância do estimador direto. Quando os tamanhos amostrais por área são relativamente pequenos, o estimador sintético supera o estimador direto e à medida que o tamanho da amostra aumenta, o estimador direto se torna mais adequado. Para encontrar o peso apropriado diversas propostas podem ser verificadas de acordo com o delineamento de uma determinada pesquisa, entretanto também pode ser atribuído arbitrariamente. E ainda Rahman (2008) pressupõe um ξ ótimo (ξ^*) aproximado através de uma formulação a partir do EQM dos estimadores direto e sintético, definido da seguinte forma:

$$\xi^* \approx \frac{\mathrm{EQM}(\hat{\Theta}_{\mathrm{d}}^{ST})}{[\mathrm{EQM}(\hat{\Theta}_{\mathrm{d}}^{dir}) + \mathrm{EQM}(\hat{\Theta}_{\mathrm{d}}^{ST})]} \ , \quad 0 < \xi^* < 1 \ .$$

Resumidamente, reafirma-se que apesar da possibilidade de ter melhor precisão, o estimador composto é viesado e depende intrinsecamente da seleção de ξ.

Com uma metodologia baseada na aplicação de modelos lineares de efeitos mistos (modelos mistos) as técnicas fundamentadas em modelos de ligação explícita podem fornecer melhorias significativas no procedimento de estimação indireta.

Quando os dados auxiliares são classificados em nível de área relacionam o estimador direto $\hat{\Theta}_{\mathrm{d}}^{dir}$ do pequeno domínio às informações auxiliares da área especificada, e são os efeitos aleatórios responsáveis pela variação entre áreas que não pode ser explicada pela inclusão de variáveis explicativas.

Nesse seguimento, o modelo proposto por Fay e Herroit (FH) em 1979 é o mais popular e comumente utilizado. O modelo FH é constituído por duas partes. A primeira consiste na relação entre o estimador direto centrado $\hat{\Theta}_{\mathrm{d}}^{dir}$ e o parâmetro desconhecido Θ_{d} para a pequena área, dada por (MOLINA; RAO, 2015):

$$\hat{\Theta}_{\mathrm{d}}^{dir} = \Theta_{\mathrm{d}} + e_{\mathrm{d}}, \quad e_{\mathrm{d}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2), \ \ \mathrm{d} = 1, 2, \ldots, \mathrm{D} \ ,$$

em que e_{d} é o erro amostral relacionado ao $\hat{\Theta}_d^{dir}$ e as D áreas selecionadas na amostra. E ainda, os erros associados e_{d}'s são variáveis aleatórias independentes e identicamentes distribuídas com média $\mathrm{E_p}(e_{\mathrm{d}}|\Theta_{\mathrm{d}}) = 0$ e variância amostral $\mathrm{Var}(e_{\mathrm{d}}|\Theta_{\mathrm{d}}) = \sigma_e^2$.

A segunda parte é destacada com atribuição do efeito aleatório ν_{d} na relação que o parâmetro Θ_{d} tem com as variáveis explicativas $\mathbf{x}_{\mathrm{d}}$. O erro aleatório representa a diferença entre as áreas não captadas pelas variáveis auxiliares, ele modela a variação não observada entre as áreas e a sua variância σ_ν^2 é estritamente maior que zero. Essa relação é retratada por YONG (2021) por meio de um modelo de regressão linear simples:

$$\Theta_{\mathrm{d}} = \mathbf{x}_{\mathrm{d}}'\beta + \nu_{\mathrm{d}}, \quad \nu_{\mathrm{d}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\nu^2), \ \ \mathrm{d} = 1, 2, \ldots, \mathrm{D} \ .$$

Na equação anterior, β é o vetor de dimensão $p \times 1$ dos coeficientes de regressão, $\mathbf{x}_{\mathrm{d}} = (\mathbf{x}_{\mathrm{d}1}, \mathbf{x}_{\mathrm{d}2}, \ldots, \mathbf{x}_{\mathrm{d}p})\prime$ o vetor-p de covariáveis em nível de área linearmente relacionado ao parâmetro Θ_{d} e o efeito aleatório ν_{d} estimado com base na suposição de normalidade.

Consequentemente, através das duas partes combinadas obtém-se o modelo linear misto:

$$\hat{\Theta}_{\mathrm{d}}^{dir} = \mathbf{x}_{\mathrm{d}}'\beta + \nu_{\mathrm{d}} + e_{\mathrm{d}} \ , \quad \nu_{\mathrm{d}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_\nu^2), \ e_{\mathrm{d}} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2) \ , \ \ \mathrm{d} = 1, 2, \ldots, \mathrm{D} \ . \tag{4.2}$$

Ao minimizar o EQM diante desse modelo misto, obtém-se o melhor preditor linear não viciado (BLUP) para o parâmetro de interesse. O BLUP sob o modelo será:

$$\hat{\Theta}_{\mathrm{d}}^{BLUP} = \mathbf{X}\tilde{\beta} + \gamma(\hat{\Theta}^{dir} - \mathbf{X}\tilde{\beta}) \ , \tag{4.3}$$

em que a componente sintética é dada por $\mathbf{X}\tilde{\beta}$, os coeficientes de regressão $\tilde{\beta}$ e $\mathbf{X}$ a matriz das variáveis independentes, e γ é o fator de encolhimento.

Entretanto, quando concernem aos parâmetros desconhecidos da equação 4.2 devem ser estimados os coeficientes de regressão β e a variância do efeito aleatório ν_{d} do pequeno domínio d, σ_ν^2. Desse modo, uma das abordagens de estimação utilizadas com mais frequência na prática aplicam o melhor preditor empírico linear não viciado (EBLUP) conforme a seguinte estrutura, sob uma perspectiva frequentista:

$$\hat{\Theta}_{\mathrm{d}}^{EBLUP} = \hat{\gamma}_{\mathrm{d}}\hat{\Theta}_{\mathrm{d}}^{dir} + (1 - \hat{\gamma}_{\mathrm{d}})\mathbf{x}_{\mathrm{d}}'\hat{\beta} \ , \tag{4.4}$$

em que $\hat{\beta}$ são os estimadores de quadrados mínimos ponderados dos parâmetros de regressão e $\hat{\gamma}_{\mathrm{d}}$ é o fator de encolhimento em d, $\hat{\gamma}_{\mathrm{d}} = \hat{\sigma}_{\nu}^2/(\hat{\sigma}_{\nu}^2 + \sigma_e^2)$. Essa abordagem é cabível aos modelos mistos para variáveis contínuas, sendo o EBLUP uma combinação poderada do estimador direto e do estimador sintético de regressão, ou seja, pode ser associado como um estimador composto. Todavia, reiterando que perante o delineamento de cada pesquisa, é possível qualificar outras abordagens como: a melhor predição empírica (EBP), hierárquica bayesiana (HB), empírica de Bayes (EB), entre outras.

A partir desses princípios existem diversas prerrogativas na aplicação de modelos em nível de área, como em circunstâncias da área de conhecimento da Estatística Espacial: a presença de correlação espacial, heterocedasticidade de efeitos aleatórios e assim por diante. Por essa razão, a seguir é contextualizado um direcionamento de investigação para trabalhos futuros na exploração de demais métodos de estimação de pequenas áreas com potencial para aprimorar a precisão de estatísticas agropecuárias em níveis subnacionais.

Sobretudo, na obtenção das estimativas de um domínio é possível explorar os fundamentos baseados a partir do par de coordenadas associado a sua referência espacial. O pressuposto é que os modelos geoaditivos (*geoadditive models*) consideram os possíveis efeitos lineares ou não lineares das covariávies e analisam a distribuição espacial da variável em estudo. Sob a suposição de aditividade, existe a possibilidade de lidar com tais efeitos de covariância ao mesclar um modelo aditivo, que considera a relação entre as variáveis e um modelo de krigagem que considera a correlação espacial, ao expressar ambos como um modelo linear misto (BOCCI; PETRUCCI, 2016). Um modelo geoaditivido pode ser formulado como:

$$\Theta_{\mathrm{d}} = \beta_0 + f(z_{\mathrm{d}}) + g(t_{\mathrm{d}}) + \beta_1'(\phi_{\mathrm{d}}) + \mathbf{G}(\phi_{\mathrm{d}}) + \epsilon_{\mathrm{d}} \ ,$$

em que, conforme descrevem os autores Pusponegoro et al. (2019), z_{d} e t_{d} representam as observações em dois preditores z e t, enquanto f e g são enunciadas como *smooth* (suaves) sendo funções de z e t respectivamente (*smooth function* - função que tem derivadas contínuas até alguma ordem desejada em algum domínio), Θ_{d} o dado da variável resposta Θ para a d-ésima área e os β's os coeficientes de regressão. A componente espacial $\mathbf{G}$ com $\phi_{\mathrm{d}} \in \mathbb{R}^2$, retrata a localização geográfica e $\{\mathbf{G}(\phi) : \phi_{\mathrm{d}} \in \mathbb{R}^2\}$ é um processo estocástico estacionário com média zero. O erro ϵ_{d} é assumido como uma variável aleatória, com média zero e variância σ_{ϵ}^2, independente e identicamente distribuído.

Ao supor a independência do efeito entre domínios no modelo FH, Pusponegoro e Rachmawati (2018) afirmam que regularmente é desconsiderada a primeira lei da geografia de

Tobler (todas as coisas são parecidas, mas as coisas mais próximas se parecem mais do que coisas mais distantes). Assim, com base em um modelo de regressão espacial simultâneo autoregressivo (SAR) cujo efeito aleatório é uma função da matriz de ponderação espacial e dos coeficientes autorregressivos, é desenvolvido um modelo SAE com a suposição de que a dependência espacial é incorporada ao efeito aleatório e a estrutura do modelo é dada por:

$$\hat{\Theta}_{\mathrm{d}} = \mathbf{X}\beta + \mathbf{W}(\mathbf{I} - \rho\mathbf{D})^{-1}\mathbf{v} + \mathbf{e}\ , \tag{4.5}$$

em que $\mathbf{X}\beta$ é o termo *large scale*, a superfície de tendência não espacial que depende de covariáveis, $\mathbf{W}$ é uma matriz $n \times n$ de constantes positivas conhecidas, $\mathbf{v}$ o vetor de efeitos aleatórios espacialmente correlacionados, $\mathbf{e}$ o vetor dos erros de amostragem relacionado ao estimador direto da área de interesse, a matriz de proximidade $\mathbf{D}$ indica se as áreas são vizinhas ou não, $\mathbf{I}$ é a matriz identidade e ρ é o parâmetro que caracteriza a dependência espacial (coeficiente autorregressivo espacial).

Então, um estimador que leva em conta os efeitos aleatórios da correlação espacial no BLUP é chamado de: o melhor preditor linear não viciado espacial (SBLUP). De acordo com Salvati (2004), a equação 4.5 baseada no modelo SAR e estabelecida a matriz variância-covariância de $\mathbf{v}$ e $\mathbf{e}$ sendo $\mathbf{C}_{sar} = \sigma^2_{\mathbf{v}}[(\mathbf{I} - \rho\mathbf{D})(\mathbf{I} - \rho\mathbf{D}')]^{-1}$, o SBLUP é definido por:

$$\hat{\Theta}^{sar} = \mathbf{x}_{\mathrm{d}}\hat{\beta} + \mathrm{b}'_{\mathrm{d}}\{\sigma^2_{\mathbf{v}}[(\mathbf{I}-\rho\mathbf{D})(\mathbf{I}-\rho\mathbf{D}')]^{-1}\}\mathbf{W}'\{\mathrm{diag}(\tau_{\mathrm{d}}) + \sigma^2_{\mathbf{v}}[(\mathbf{I}-\rho\mathbf{D})(\mathbf{I}-\rho\mathbf{D}')]^{-1}\}^{-1}(\hat{\Theta}_{\mathrm{d}} - \mathbf{X}\hat{\beta})\ ,$$

com $\mathrm{b}'_{\mathrm{d}} = (0, 0, \ldots 0, 1, 0, \ldots 0)$ em que 1 se refere a d-ésima e restabelecendo $\mathrm{diag}(\tau_{\mathrm{d}})$ como a variância amostral dos estimadores diretos dos parâmetros da área de interesse.

Posteriormente, Salvati (2004) delibera que a diferença entre o modelo SAR e o modelo autoregressivo condicional (CAR) está na estrutura da matriz de variância-covariância $\mathbf{C}$. Assim, no modelo CAR sendo a matriz $\mathbf{C}_{car} = \sigma^2_{\mathbf{v}}(\mathbf{I} - \rho\mathbf{D})^{-1}$, tem-se o SBLUP:

$$\hat{\Theta}^{car} = \mathbf{x}_{\mathrm{d}}\hat{\beta} + \mathrm{b}'_{\mathrm{d}}[\sigma^2_{\mathbf{v}}(\mathbf{I} - \rho\mathbf{D})^{-1}]\mathbf{W}'[\mathrm{diag}(\tau_{\mathrm{d}}) + \sigma^2_{\mathbf{v}}(\mathbf{I} - \rho\mathbf{D})^{-1}]^{-1}(\hat{\Theta}_{\mathrm{d}} - \mathbf{X}\hat{\beta})\ .$$

Portanto, as sugestões para novos estudos nessa área de pesquisa visam explorar a presença de dados espaciais e investigar modelos de regressão espacial.

5 MATERIAIS E MÉTODOS

A explanação para obter os dados da meta 2.3 corresponde a seguinte sequência estruturada baseada nos critérios de universalidade propostos pela FAO, uma vez que os pequenos produtores de alimentos precisam ser identificados e monitorados:

1. Coletar os dados sobre as unidades de produção por produtor;
2. Calcular o tamanho da terra e o tamanho econômico do respectivo produtor, além de estabelecer os limites correspondentes (absolutos e relativos);
3. Definir os produtores que estão na categoria de pequeno produtor de alimentos;
4. Calcular os indicadores 2.3.1 e 2.3.2.

Diante disso, na produção de estimavas diretas a partir de uma amostra probabílistica s, é possível assumir a relação da produtividade do trabalho do i-ésimo pequeno produtor de alimentos em um determinado domínio d, a variável de interesse y. A exemplificar uma forma de obter as estimativas diretas do indicador 2.3.1, o estimador direto de H-T no d-ésimo pequeno domínio é dado a seguir, sendo o peso amostral w_i o inverso da probabilidade de inclusão π_i do elemento i:

$$\hat{\bar{Y}}_{2.3.1,\mathrm{d}} = \frac{\sum\limits_{i \in S_\mathrm{d}} w_i \; y_{2.3.1,i}}{\sum\limits_{i \in S_\mathrm{d}} w_i} \; .$$

Em consequência do indicador ODS 2.3.1 definido como "Em análise/construção" no Brasil (figura 3), essa dissertação adota na modelagem do estudo de simulação dos métodos SAE os próprios valores desse indicador fornecido pela FAO como o resultado das estimativas diretas. A descrição do conteúdo das estimativas diretas para o experimento está na seção 5.2.2.

Os seguimentos para o experimento são adquiridos e organizados através do desenvolvimento em linguagem de programação R via ambiente de desenvolvimento integrado (IDE) e também manipulações em linguagem SAS®. Na aquisição das variáveis independentes é fundamental a manipulação via Sistemas de Informação Geográfica (SIG). Entretanto, é importante destacar a contrariedade que existe na acessibilidade, disponibilidade e formatação de dados, principalmente na aquisão de variáveis auxiliares, devido a diversos fatores como privacidade e questões políticas. Por consequência, para garantir uma possível interação e confiabilidade de dados a sua obtenção deve ser efetuada em órgãos das esferas governamentais federal, estadual e municipal, por exemplo o IBGE.

Figura 3 – Conjuntura dos dados da meta 2.3 no Brasil

Fonte: IBGE (2024)

5.1 SUPORTE COMPUTACIONAL

Uma série de rotinas computacionais é efetuada para avaliar a eficácia de diferentes abordagens, critérios e transformações. Cada teste foi projetado e executado levando em consideração fatores-chave e parâmetros cruciais. Após uma análise detalhada dos desfechos, constatou-se um seguimento que apresentava desempenho e precisão.

Portanto, para obtenção dos resultados foi elaborada uma rotina computacional em uma linguagem de programação orientada a objeto, em R. É semelhante a linguagem S e desenvolvida por Ross Ihaka e Robert Gentleman na década de 90 sendo um projeto sob os termos *General Public License* (GNU) gratuito e livre, bastante extensível com variedades de estatísticas e grande facilidade na produção de gráficos. Neste caso, a IDE utilizada é o RStudio versão 4.9.2 com a linguagem R versão 4.1.3.

Um código SAS (*The SAS System*) foi criado para visualizar os dados e gráficos através do SAS® *OnDemand for Academics* que contém todos os programas necessários para selecionar amostras, calcular estimativas e elaborar gráficos mediante dados para pesquisas. As manipulações para o desenvolvimento do *script* foram realizadas no ambiente de desenvolvimento web, o SAS® Studio.

As funcionalidades do SIG QGIS permitiram visualizar, editar e analisar dados espaciais. O software gratuito de código aberto tanto na versão 2.18.18 quanto 3.28.9 possui uma interface intuitiva e grande extensibilidade. Assim, possibilitou a integração de dados, realização de análises espaciais, criação de mapas ou imagens mais personalizadas e o manuseio de ferramentas de processamento de imagens.

A dissertação foi organizada e redigida em LaTeX , o qual é um sistema de alta qualidade para preparação de documentos desenvolvido por Leslie Lamporte em 1985. É baseada no formato de modelo de trabalho acadêmico em conformidade com a ABNT NBR 14724:2011, dissertação de mestrado.

Todo o procedimento foi executado em um notebook com processador Intel(R) Core(TM) i5, 8GB de memória RAM, SSD de 250GB e sistema operacional de 64bits Windows 11.

5.2 O CONJUNTO DE DADOS E INFORMAÇÕES

Os autores PINHEIRO et al. (2009) afirmam que quando um levantamento de dados é efetuado a respeito de um determinado assunto, eles costumam ser estruturados e organizados em formato tabular, usualmente armazenados em formato eletrônico chamado de planilhas. Na tabela ou matriz de dados cada linha corresponde a uma observação ou unidade de investigação e cada coluna a uma variável que corresponde à realização de uma característica, mas costumam também ser referidas como atributos, propriedades, entre outros.

A identificação de variáveis independentes é uma etapa crucial no processo de modelagem estatística, dessa forma é importante considerar alguns passos na escolha destes dados. É fundamental adotar critérios que assegurem a pertinência e a relevância ao escolher variávies auxiliares para análises e selação em um modelo, sendo imperativo que os dados escolhidos possuam um embasamento teórico sólido que justifique a sua inclusão. A princípio, é necessário ponderar a relação dos dados com o estudo, ou seja, as variáveis explicativas devem estar diretamente relacionadas ao fenômeno ou problema de interesse na pesquisa, a variável resposta ou dependente. Ou seja, a variável selecionada como preditora atua como um possível fator explicativo da variável dependente.

Posteriormente, realizar uma análise descritiva dos dados para entender a distribuição e a variabilidade dos termos, constatar possíveis padrões ou tendências e detectar *outliers*. De acordo com a complexidade do modelo, verificar a correlação entre as variáveis independentes e a variável dependente e também entre si para evitar incluir um número excessivo de dados. A base de dados deve apresentar uma gama suficiente de valores para possibilitar a realização de correlações significativas. Caso uma variável não demonstre variação entre os casos se torna inviável a sua aplicação. A presença de forte correlação entre as variáveis pode distorcer os resultados, comprometendo a validade das conclusões obtidas.

Após selecionar as variáveis para inclusão no modelo, validá-lo ajuda a garantir que o

modelo seja robusto e generalizável para novos dados, capaz de capturar de forma precisa e abrangente as relações existentes entre a variável resposta. Por fim, na interpretação dos resultados é preciso examinar os coeficientes adquiridos, a significância estatística, a magnitude dos efeitos e a relevância prática das variáveis incluídas.

5.2.1 Levantamento das Observações

O IBGE elaborou e conduziu o último Censo Agropecuário do Brasil em 2017, o qual tem como principal objetivo a coleta de dados e informações sobre a organização, dinâmica e estrutura dos setores florestal, aquícola e agropecuário do território brasileiro. Concisamente, os dados foram obtidos através de entrevistas com os responsáveis do local ou estabelecimento, baseadas em questionários previamente formulados, apoiados por um dispositivo móvel de coleta. Através de um rígido controle de qualidade para assegurar a precisão e a segurança dos dados, principalmente contra possíveis vazamentos das informações ou acessos não autorizados. Ao fornecer uma ampla variedade de dados proporciona entender a situação destes setores, a análise de padrões, o desenvolvimento de estratégias, a formulação de políticas públicas, entre outros. Além disso, evidenciou a incorporação de tecnologias e práticas sustentáveis propiciando uma visão holística do avanço tecnológico na Agropecuária.

O levantamento de dados desempenha um papel substancial em todo procedimento dessa dissertação. Com isso, é importante destacar que a pesquisa no Brasil é fundamental para embasar políticas públicas, orientar investimentos e promover o desenvolvimento socioeconômico do país. Aliás, a disponibilidade de dados e informações detalhadas, de qualidade, confiáveis e atualizadas facilita o trabalho de acadêmicos, pesquisadores e profissionais que buscam compreender os desafios e oportunidades enfrentados pelo Brasil em diferentes áreas de conhecimento. O acesso à informação é essencial para análises aprofundadas e embasadas, permitindo uma compreensão mais precisa da realidade em estudo.

Justamente, a partir da coleta de dados, esse censo é um reflexo preciso dos diversos setores tanto da agricultura e da pecuária quanto do agronegócio no país, ou seja, é a aquisição de numerosas variáveis viabiliza toda essa reprodução. Comumente, algumas variáveis são destacadas devido a sua clareza e simplicidade em:

- Características do Produtor: sexo, idade e escolaridade do produtor; condição de posse da terra.

- Características do Estabelecimento: tipo de estabelecimento (ex. familiar, não familiar, agroindustrial); localização geográfica (ex. município, estado, região); área total e área utilizada.
- Mão de Obra: quantidade de trabalhadores familiares e assalariados.
- Uso da Terra: áreas destinadas a lavouras, pastagens, matas, reservas legais, entre outros.
- Produção Agropecuária: culturas plantadas (ex. grãos, hortaliças, frutas); criação de animais (ex. aves, bovinos, suínos); quantidades produzidas e colhidas.
- Rendimento e Renda: rendimento monetário proveniente da produção; fontes de renda do estabelecimento.
- Tecnologia e Equipamentos: uso de máquinas agrícolas e equipamentos; práticas sustentáveis adotadas; uso de tecnologia da informação e comunicação no estabelecimento.
- Aspectos Ambientais: práticas ambientais utilizadas para a preservação.

Logo, algumas dessas variáveis disponibilizadas no Censo Agro pelo IBGE (2017) podem contribuir para uma análise consolidada desse desempenho produtivo do indicador 2.3.1 perante o produtor de alimentos e na elaboração de uma medida mais precisa. No contexto brasileiro, as variáveis relevantes podem incluir: dados sobre o número de pessoas envolvidas no trabalho agrícola em cada estabelecimento (mão de obra ou quantidade de trabalhadores), a quantidade total de produtos agrícolas produzidos considerando diferentes tipos de culturas e criação de animais (volume de produção ou produção agropecuária) e a área total utilizada, a extensão de terras dedicadas a lavouras, pastagens e outros usos (dimensão da empresa ou características do estabelecimento e uso da terra).

Note que essas variáveis vão além dos aspectos puramente de caráter quantitativo e abordam uma gama mais ampla de fatores que influenciam a viabilidade e a sustentabilidade da Agropecuária, bem como o bem-estar das comunidades rurais. Essa métrica é essencial para entender não apenas a eficiência e eficácia de produção, mas também sua sustentabilidade a longo prazo, especialmente no contexto dos pequenos produtores de alimentos.

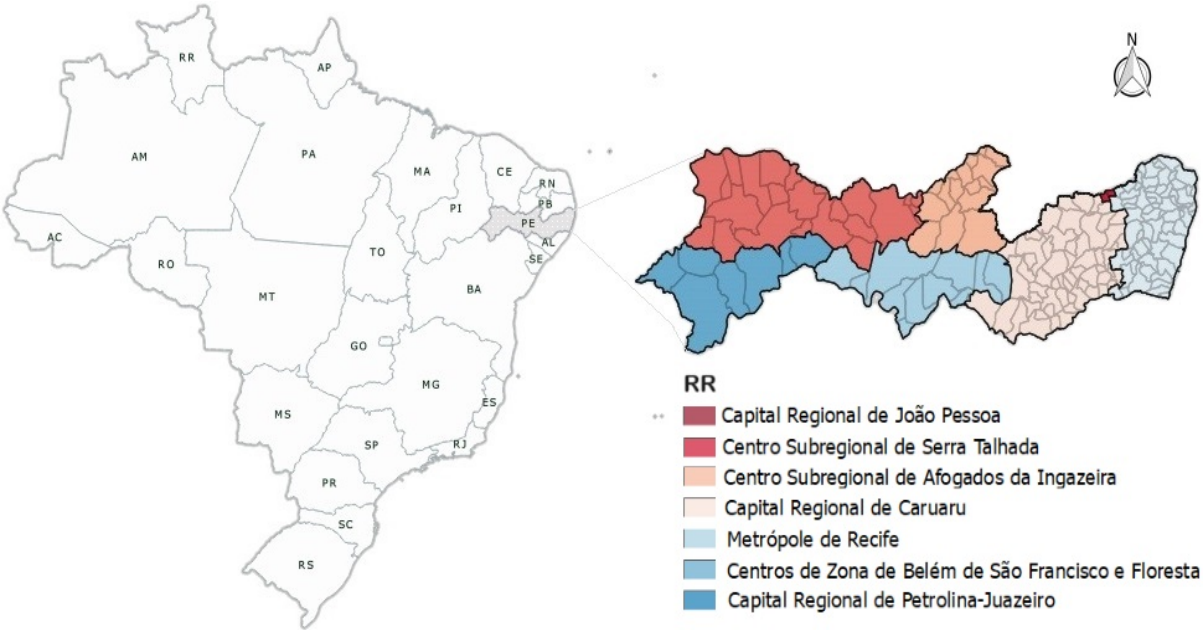
Entretanto, a eficiência no uso de recursos é fundamental e a resiliência às mudanças climáticas é essencial para garantir a segurança alimentar e o sustento dessas comunidades. Desse modo, variáveis preditoras que indicam a capacidade dos pequenos produtores de se adaptarem a eventos climáticos extremos, como secas e enchentes, são vitais para indicar a

continuidade da produção agrícola e a proteção dos meios de vida enquanto minimiza o impacto ambiental negativo. Outros dados auxiliares relevantes estão relacionados à práticas agrícolas sustentáveis, como o uso de tecnologias de precisão, gestão eficiente da água e conservação do solo. Demais aspectos importantes são: a segurança alimentar e nutricional, o acesso as vias abertas à circulação e renda adequada, a sustentabilidade ambiental, a conservação dos recursos naturais e assim por diante.

5.2.1.1 Caracterização do estado de Pernambuco

O estado de Pernambuco possui 185 municípios, com área territorial de aproximadamente 98.068 KM^2 sendo 1.415 KM^2 de área urbanizada. A sua economia agrícola é caracterizada por uma riqueza de setores produtivos que desempenham papéis distintos, mas complementares no panorama Agropecuário local. Entre os segmentos se destacam a fruticultura, a produção de cana-de-açúcar e a agricultura familiar, cada qual conferindo singularidade e vitalidade à paisagem agrícola da região. A figura 4 mostra as divisas do Brasil e destaca as Regiões Rurais (RR) presentes em Pernambuco. Para a confecção da seguinte imagem o produto digital é fornecido gratuitamente pelo IBGE na portal de Geociências.

Figura 4 – Os estados brasileiros e as Regiões Rurais de Pernambuco



Fonte dos dados básicos: IBGE(2022). Elaborada pela autora (2024)

As RR delineadas buscam facilitar a disseminação de dados estatísticos em consonância com as crescentes exigências sociais quanto à elaboração de informações baseadas em segmentações

territoriais mais refinadas, as quais reflitam os diversos padrões de assentamento desenvolvidos ao longo do tempo pela sociedade brasileira em sua vasta extensão territorial. Assim, representa uma contribuição aos estudos rurais conduzidos pelo do IBGE a partir do ano de 2015.

A fruticultura emerge sustentada por um contexto geoclimático favorável e uma diversidade edafoclimática propícia ao cultivo de uma série de espécies frutíferas alavancando uma significativa produção destinada tanto ao mercado interno quanto à exportação. Paralelamente, a produção de cana-de-açúcar figura como uma atividade de longa tradição e relevância econômica e constitui a base para uma próspera indústria sucroalcooleira, cuja produção de açúcar, etanol e seus derivados contribuem também para a geração de empregos e a dinamização das economias locais. Já a agricultura familiar se consolida como uma matriz multifacetada de culturas de subsistência e comerciais, abarca desde a produção de milho, feijão e mandioca até a hortifruticultura, contribuindo substancialmente para a segurança alimentar, a preservação de práticas agrícolas tradicionais e a promoção da sustentabilidade ambiental.

Ademais, setores complementares como a avicultura, a ovinocaprinocultura e a produção leiteira conferem ainda mais diversidade e robustez à produtividade agropecuária de Pernambuco, refletindo a adaptabilidade dos agricultores locais diante das demandas do mercado e das condições ambientais.

Portanto, na modelagem estatística do estudo dessa dissertação, diversas variáveis auxiliares são determinantes para o sucesso dessas atividades na ramo da Agropecuária, principalmente na Agricultura. Com isso, dados e informações relacionadas ao tipo de clima, do solo e a sua fertilidade, precipitação pluviométrica média, as tecnologias agrícolas utilizadas (como sistemas de irrigação), podem influenciar no padrão de cultivo, na produtividade e na comercialização, consequentemente na obtenção dos resultados.

5.2.2 O Conjunto da Variável Resposta

A base da dados fornecida pela FAO, viabilizada uma planilha em armazenamento eletrônico enviada no formato .xlsx, consta os resultados dos indicadores ODS (IODS) 2.3.1 (que mede a produtividade agrícola por unidade de trabalho) e 2.3.2 (que se refere a renda média dos pequenos produtores de alimentos, por sexo e condição de indígena) correspondente a escala do pequeno produtor de alimento para 5560 municípios brasileiros. As demais especificações das colunas dessa base estão dispostas com referência ao código do município definido pelo IBGE para idenficiar o local (p. ex. Recife $=$ 2611606), a quantidade de estabelecimentos

agropecuários utilizados no cálculo desses IODS e algumas estatísticas descritivas dos valores dos IODS, como os valores de mínimo e máximo. Esse conjunto de dados elaborado pela FAO não está disponível ao público, sendo baseado nos dados do Censo Agro de 2017 do Brasil.

A tabela 1 apresenta os 10 municípios não contemplados (não são fornecidos no arquivo), ou seja, quaisquer dados não estão disponíveis na planilha. Portanto, para o experimento dessa dissertação, no procedimento de modelagem, o conjunto da variável resposta consiste nos resultados obtidos pela FAO do indicador ODS 2.3.1, que são as estimativas diretas.

Tabela 1 – Municípios do Brasil sem dados da meta 2.3

Código	**Município**	**UF**	**Região**
3525003	Jandira	SP	Sudeste
3510609	Carapicuíba	SP	Sudeste
3541000	Praia Grande	SP	Sudeste
3500600	Águas de São Pedro	SP	Sudeste
3505708	Barueri	SP	Sudeste
3506359	Bertioga	SP	Sudeste
3303203	Nilópolis	RJ	Sudeste
3305109	São João de Meriti	RJ	Sudeste
4307708	Esteio	RS	Sul
2919926	Madre de Deus	BA	Nordeste

Fonte dos dados básicos: IBGE(2022). Elaborada pela autora (2024)

5.2.3 O Conjunto das Variáveis Explicativas

Na seleção de variáveis independentes de um conjunto de dados de alta dimensionalidade, é fundamental utilizar métodos de seleção que sejam eficientes e robustos, principalmente que não sejam custosos computacionalmente. Por consequência, a base da dados auxiliar utilizada nessa dissertação é composta por 87 variáveis sendo unificadas em uma estrutura tabular. A tabela com a nomenclatura, unidade de medida, fonte de dados e o ano de referência das variáveis explicativas está disponível no Anexo A.

O conteúdo do conjunto das variáveis explicativas abrange os indicadores temáticos municipais fornecidos pelo IBGE no Censo Agro 2017, que contém o quociente entre a área total e a quantidade de estabelecimentos agropecuários localizados no município, informações sobre atividades econômicas desses estabelecimentos, entre outros. E também alguns dados do IBGE identificados na fase de levantamento das observações, por exemplo o índice de Gini, e o Índice de Desenvolvimento Humano Municipal (IDHM) proporcionado pelo IPEA (2013)

baseado no Censo Demográfico de 2010. Já para retratar a cobetura vegetal do local são adotados os índices de vegetação (em nível estatal) disponilizados pela FAO (2022), como o Índice de Vegetação por Diferença Normalizada (NDVI) e o Índice da Saúde da Vegetação (VHI). E incluindo a contribuição do estudo realizado pela Fiocruz e o IME, no qual foi elaborado um conjunto de dados auxiliar compreendendo um total de 11.556 medidas estatísticas para cada município brasileiro (ABDALLA et al., 2022). Na pesquisa citada, com publicação na revista *Nature*, foram utilizadas 19 camadas temáticas obtidas de diferentes agências governamentais brasileiras e internacionais, dependendo da natureza do tema cada camada poderá possuir múltiplas classes ou variáveis, assim totalizando 642 variáveis temáticas. Aliás, o trabalho também é notável devido ao pré-processamento de dados, em um contexto levou cerca de 600 horas para obter os dados de uso e cobertura do solo, enquanto outros dados com relação a temperatura, precipitação, altitude, dentre outros, durou aproximadamente 1 mês de processamento ininterrupto em um computador, somando 195 GB de dados brutos. Diante da base explicativa atribuída ao experimento de modelagem dessa dissertação, a tabela 2 informa o quantitativo de variáveis independentes por fonte dos dados.

Tabela 2 – Quantidade de variáveis explicativas por fonte de dados da base de dados auxiliar

BASE AUXILIAR	**FONTE DE DADOS**				**TOTAL**
Variáveis Disponíveis	IPEA	FAO	FIOCRUZ	IBGE	
	1	6	35	45	87

Fonte: Elaborada pela autora (2024)

Segundo o agrupamento construído, as informações coletadas estão relacionadas consoante o clima, as tecnologias agrícolas, a atividade econômica e assim por diante. Portanto, em síntese, a priporidade na distinção das informações é baseada na natureza dos dados, nos objetivos da análise, nas características relevantes ao indicador de interesse (2.3.1) e na disponibilidade de recursos computacionais.

5.3 COMPONENTES DE UM MODELO ESTATÍSTICO

Em qualquer uma das técnicas ou aplicações para estimação de pequenos domínios, o êxito está atrelado a um conjunto parcimonioso de preditores na obtenção de resultados com um poder preditivo ótimo, atingindo uma menor variância obtida no ajuste do modelo. Inclusive, os procedimentos automáticos de seleção de variáveis auxiliares, por exemplo o procedimento *stepwise*, são úteis para encontrar o melhor modelo de regressão.

Na implementação da abordagem SAE baseada em modelos de regressão as suposições subjacentes precisam ser cuidadosamente validadas. Ademais, o viés da estimativa na pequena área precisa ser calculado para verificar a precisão do resultado. Geralmente é feito por meio do EQM, o erro quadrático médio que fornece um critério combinando precisão (variância) e acurácia (viés) das estimativas. No entanto, por si só não é uma medida irrestrita. Logo, ao interpretar o EQM é indispensável considerar outros aspectos do modelo.

É fundamental efetuar uma avaliação e validação do modelo, verificar se as suposições necessárias para a análise de regressão e a adequação do modelo ao ajuste são atendidas, como a independência e suposições de normalidade dos erros e a homocedasticidade dos termos de erro.

5.3.1 Modelo de Efeitos Fixos

O modelo de regressão linear simples precisa ser ajustado com as estimativas diretas sendo definidas como a variável dependente. E atráves das variáveis auxiliares é implementado o ajuste do modelo por quadrados mínimos ordinários (QMO), cuja finalidade é alcançar o melhor ajuste linear ao minimizar a soma dos quadrados das diferenças entre os valores observados e os preditos pelo modelo. Em seguida, deve ser averiguada as premissas: análise de resíduos, verificar a independência dos erros, dentre outras. Se essas proposições não forem satisfeitas, ainda é possível reajustar o mesmo modelo por meio de quadrados mínimos ponderados.

Seja o modelo $y_{\mathrm{d}} = \beta_0 + \beta_1 \mathbf{x}'_{\mathrm{d}1} + \ldots + \beta_{13} \mathbf{x}'_{\mathrm{d}13} + \upsilon_{\mathrm{d}}$ com 13 variáveis independentes selecionadas, em que y_{d} é o parâmetro de interesse, $\mathbf{x}'$ as variáveis auxiliares, β os coeficientes de regressão e υ_{d} são os erros residuais independentes e identicamente distribuídos (i.i.d.). Assim, o preditor via quadrados mínimos da variável de interesse é simplificado como:

$$\hat{y}_{\mathrm{d}}^{ST} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{X}_{\mathrm{d}1} + \ldots + \hat{\beta}_{13} \mathbf{X}_{\mathrm{d}13} \ ,$$

$\hat{\beta}_{\mathrm{p}}$ os estimadores para β e $\mathbf{X}_{\mathrm{dp}}$ são os conjuntos conhecidos das variáveis explicativas no domínio d, $\mathrm{z} = 1, \ldots, 13$.

Para um estimador sintético por regressão o principal problema está atrelado ao viés no procedimento de estimação, enquanto a variância pode ser estimada por diversas maneiras, inclusive de forma analítica. Logo, uma aproximação amparada em um estimador centrado para o EQM foi exposta na seção 4.2.

Como uma maneira de equilibrar o potencial viés de um estimador sintético contra a

instabilidade de um estimador direto, é escolher um peso ξ que esteja entre 0 e 1. Então, de acordo com o delineamento da pesquisa dessa dissertação é aplicada a formulação do quociente elaborado por Drew, Singh e Choudhry (1982). O peso ξ proposto por esses três autores corresponde ao caso em que o estimador é dependente do tamanho amostral. Nessa condição, é defindo M sendo a quantidade total de estabelecimentos agropecuários do município d utilizados no cálculo do indicador ODS 2.3.1, como funções do quociente $\hat{\mathrm{M}}_{\mathrm{d}}/\mathrm{M}_{\mathrm{d}}$. Ao atribuir um valor $\kappa > 0$ predeterminado que é escolhido subjetivamente para controlar a contribuição da componente sintética, ξ assume a forma de (COELHO, 1996):

$$\xi = \begin{cases} 1 & se \quad \hat{\mathrm{M}}_{\mathrm{d}} \geq \kappa \mathrm{M}_{\mathrm{d}} \\ \hat{\mathrm{M}}_{\mathrm{d}}/(\kappa \mathrm{M}_{\mathrm{d}}) & se \quad \hat{\mathrm{M}}_{\mathrm{d}} < \kappa \mathrm{M}_{\mathrm{d}} \end{cases} .$$

É esperado que sejam obtidos resultados mais precisos da variável dependente. Entretanto, reitera-se que apesar da possibilidade de ter melhor precisão, o estimador composto é viesado e depende da seleção de um ξ adequado.

Na modelagem efetuada nessa dissertação, o ajuste do modelo de efeitos fixos foi elaborado em linguagem R, conforme as funções do pacote (biblioteca) 'car' diponívies em R Core Team (2022).

5.3.2 Modelo de Efeitos Mistos

Na formulação do modelo de regressão pelo método de FH, utilizam-se as estimativas diretas como a variável resposta ligada a um vetor de variáveis explicativas. O vetor dos estimadores diretos é incorporado com suas informações no modelo à gerar estimativas mais precisas, por exemplo para municípios, e assim é obtido o estimador de pequenas áreas (COELHO, 1996).

No cálculo do BLUP é considerado que as matrizes de covariâncias são conhecidas. Habitualmente, na prática essas são substituídas por estimativas e o estimador resultante é o conhecido EBLUP. As variâncias desconhecidas podem ser estimadas atráves do método dos momentos (proposto por Fay e Herriot), método de máxima verossimilhança (MV), máxima verossimilhança restrita (MVR), dentre outros.

No experimento de modelagem dessa dissertação, da mesma forma que são necessários dados anteriores dos valores do IODS 2.3.1, é exigida a componente da variância amostral para que ela seja estimada. Por consequência, é construída uma variância fictícia. Sob essa situação, a partir da base concedida pela FAO tem-se um único valor para cada município do

IODS 2.3.1, representando a média municipal. Logo, a variância dos dados é zero, pois não existe variação entre os valores dentro de cada município.

Entretanto, fornecida a quantidade total de estabelecimentos (M), em que foi efetuado o cálculo para retirar cada média. E esse fato implica que cada valor é uma média de uma amostra (y_f) com um determinado tamanho. Porém, se existe um único estabelecimento (M=1) a variação é 0. Assim, é criada uma variância ponderada, onde pondera-se a variância pelo inverso do tamanho da amostra (número total de estabelecimentos) para cada município:

$$\mathrm{LVar_p} = \frac{1}{\mathrm{M}} \sum_{f=1}^{\mathrm{M}} y_f^2 - (\frac{1}{\mathrm{M}^2}(\sum_{f=1}^{\mathrm{M}} y_f)^2) \ . \tag{5.1}$$

Na fórmula 5.1 a subtração dessas duas componentes segue o princípio de que é a diferença entre a média dos quadrados dos valores e o quadrado da média.

No ajuste do modelo FH, quando essa variância é estimada, por exemplo via MVR, o EQM do estimador EBLUP é (MOLINA; RAO, 2015):

$$\mathrm{EQM}(\hat{\Theta}_{\mathrm{d}}^{EBLUP}) = \hat{\gamma}_{\mathrm{d}}\sigma_e^2 + (1-\hat{\gamma}_{\mathrm{d}})^2 \mathbf{x}_{\mathrm{d}}'(\sum_{\mathrm{d}=1}^{D} \hat{\gamma}_{\mathrm{d}} \mathbf{x}_{\mathrm{d}}'\mathbf{x}_{\mathrm{d}})^{-1}\mathbf{x}_{\mathrm{d}}) + 2(1-\hat{\gamma}_{\mathrm{d}})^2 \hat{\gamma}_{\mathrm{d}} \hat{\sigma}_{\nu}^{-2} \bar{\mathrm{v}}(\hat{\sigma}_{\nu}^2) \ .$$

Pontualmente, é a variância assintótica de $\hat{\sigma}_{\nu}^2$ que depende do método de estimação utilizado para σ_{ν}^2, o elemento $\bar{\mathrm{v}}(\hat{\sigma}_{\nu}^2)$.

Outra medida crucial nesse ajuste é o coeficiente de variação (CV), sendo uma medida de dispersão. O CV é uma medida de variabilidade da estimativa que, conforme relatam os autores Kreutzamann et al. (2019), normalmente os instituitos de estatística usam para quantitificar a incerteza associada às estimativas. E o define como $\mathrm{CV} = \sqrt{\mathrm{E\hat{Q}M}(\widetilde{\Theta}_{\mathrm{d}})}/\widetilde{\Theta}_{\mathrm{d}}$, em que $\widetilde{\Theta}_{\mathrm{d}}$ representa a estimativa obtida no domínio d.

Na modelagem realizada nessa dissertação, o ajuste do modelo de efeitos mistos foi desenvolvido em linguagem R, conforme as funções do pacote (biblioteca) 'emdi' sendo elaborado por esses mesmo autores, KREUTZMANN et al. (2019).

Resumidamente, na aplicação do modelo FH deve-se constar que os totais da variável de interesse para cada um dos D domínios da população estão ligados a um vetor de variáveis explicativas através de um modelo de regressão linear. E assumirá que a relação entre a variável de interesse e as covariáveis é a mesma na população total e na amostra. Com essa relação, se estima os parâmetros do modelo para depois aplicar esses parâmetros aos valores das covariáveis na população obtendo estimativas mais precisas para o total populacional.

6 RESULTADOS E DISCUSSÕES

Em algumas análises estatísticas, os métodos aplicados juntamente com os resultados são componentes que podem evoluir de forma iterativa e simultânea. Onde os métodos podem ser estabelecidos ou adaptados conforme surgem os resultados, essa apresentação contínua reflete melhor a natureza iterativa desse processo de investigação. Assim, a seção 6 descreve o encadeamento do experimento de simulação dessa dissertação refletindo a dinâmica do estudo.

6.1 ANÁLISE EXPLORATÓRIA DE DADOS

A análise descritiva dos dados verifica diversas medidas resumo. Por meio dos dados concedidos pela FAO do IODS 2.3.1 (que se refere ao volume de produção por unidade de trabalho por dimensão da empresa agrícola/pastoril/florestal) definido na seção 3.1, são destacadas algumas destas estatísticas, as quais estão apresentadas na tabela 3 e concerne à 5560 municípios do Brasil (conjunto populacional).

Tabela 3 – Medidas descritivas dos dados do indicador 2.3.1 no conjunto populacional

Estatísticas Descritivas	**Variável de Interesse** Indicador ODS 2.3.1
Valor Mínimo	0,42
1º Quartil	7,59
Mediana	13,91
3º Quartil	20,19
Valor Máximo	111,57
Amplitude	111,15
IQR	12,60
Média	15,01
Variância	91,87
Desvio Padrão	9,58
Coef. de Variação	63,87
Assimetria	1,72
Curtose	7,51

Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

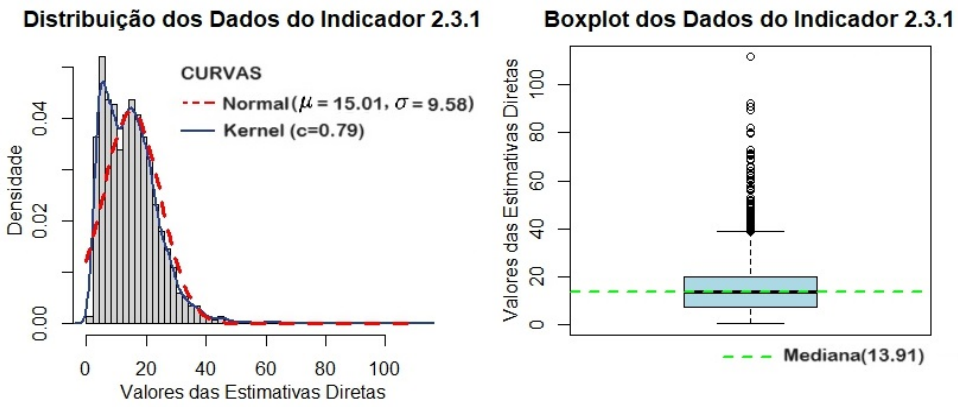
Com respeito as medidas de posição, a média é de 15 unidades, que segundo a UNDS (2024) a unidade de medida corresponde a *Constant PPP USD* 2017 (CPU). A mediana ou segundo quartil (Q_2) é 13,9 CPU, ademais a média amparada com um limiar de 10% é 14 CPU. Em suma, CPU é o produto interno bruto convertido em dólares usando taxas da paridade do poder de compra (WBG, 2021). Perante as medidas de dispersão mais comuns,

a variância é aproximadamente 91,8 CPU^2, assim o desvio padrão é de 9,58 CPU, o desvio médio absoluto 9,34 CPU e 12,6 a distância interquartil (IQR). No caso das medidas de forma tem-se 1,72 de assimetria, sendo 1,724 o coeficiente de Fisher-Pearson, de 1,725 o coeficiente de Fisher-Pearson Ajustado e 0,344 o coeficiente de Pearson 2, já 7,51 é o valor da curtose.

O resultado da média e do desvio padrão é um indicativo de uma dispersão considerável em relação à média sendo 63,87 % o coeficiente de variação (CV). Os valores extremos variam de 0,4 a 111,6 refletindo uma ampla gama de observações. A assimetria dos dados indica uma inclinação à direita (assimetria positiva), e com os resultados dos coeficientes nota-se que a distribuição dos dados tem uma forte assimetria positiva, onde a cauda à direita terá maior impacto sobre a média, levando-a se mover em sua direção. A curtose revela uma distribuição altamente leptocúrtica, em que os dados têm mais valores extremos (os *outliers*), com caudas mais pesadas e um pico mais alto do que o esperado em uma distribuição normal.

A figura 5 reafirma e evidencia a interpretação geral postulada ilustrando a análise exploratória efetuada. A representação do histograma com a curva de densidade ocasiona a avaliação da não normalidade e a presença de assimetria positiva na distribuição dos dados. O boxplot concede os quartis, a mediana (linha verde) e os valores extremos mostrando um alongamento maior na parte superior da caixa, evidência de assimetria à direita. A notável presença de *outliers* pode distinguir uma variedade de situações, como características verdadeiramente incomuns no fenômeno em estudo. Que é o caso de um município situado na região Centro-Oeste, com um valor acima de 100 CPU (devio ao critério de confidencialidade outros dados e informações não podem ser divulgadas).

Figura 5 – Histrograma e Boxplot das estimativas do IODS 2.3.1 do conjunto populacional

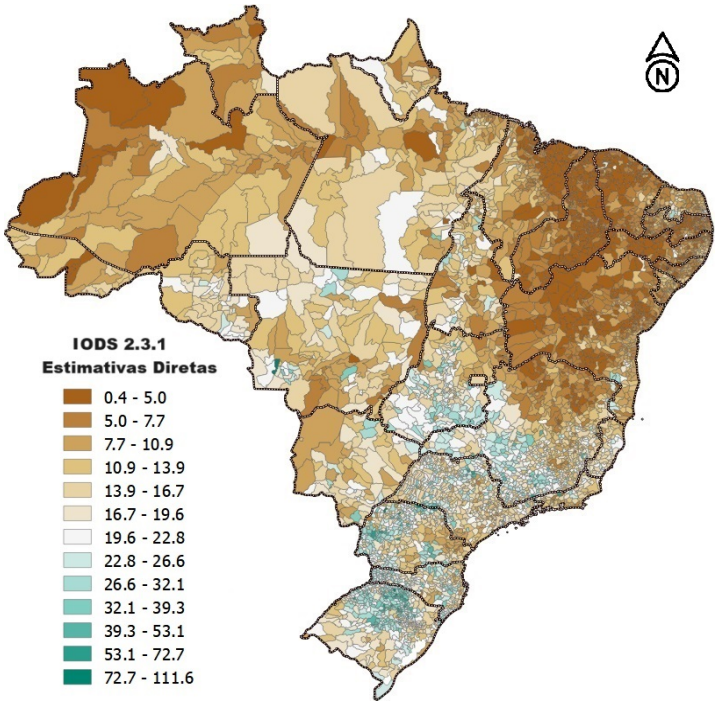


Fonte dos dados básicos: IBGE(2022). Elaborada pela autora (2024)

A imagem 6 é elaborada em ambiente de desenvolvimento SIG (QGIS 2.18.28), ilustra os resultados da classificação das estimativas diretas do IODS 2.3.1 no Brasil (2017) em modo

graduado por 13 classes através da regra de Sturges, a partir da técnica de classificação de Quebras Naturais (Jenks). Nessa ilustração, é evidente que existem concentrações específicas da disposição dos dados, em que os valores médios mais baixos estão concentrados nas regiões Norte e Nordeste do Brasil, enquanto os valores médios mais altos nas regiões Sul e Sudeste.

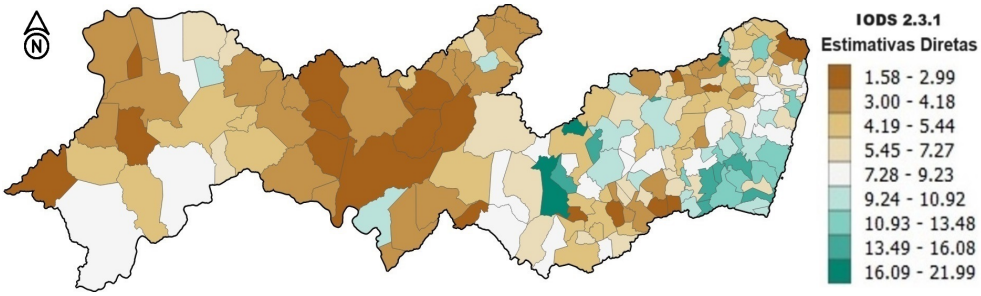
Figura 6 – Representação das estimativas diretas do IODS 2.3.1 no Brasil em 2017 do conjunto populacional



Fonte dos dados básicos: IBGE(2022). Elaborada pela autora (2024)

Diante dos 185 municípios, na extensão territorial pernambucana os resultados com os valores médios mais altos estão no litoral do estado (classificação a partir da técnica de Quebras Naturais em modo graduado por 9 classes obtidas pela regra de Sturges), e mais concentrados na Região de Desenvolvimento (RD) Mata Sul (figura 7). Localidade onde a atividade principal agropecuária e agroindustrial é a cana-de-açúcar e seus derivados, porém o turismo também é uma atividade bastante significativa com belezas naturais de suas cachoeiras e praias.

Figura 7 – Representação das estimativas diretas do IODS 2.3.1 em Pernambuco do conjunto populacional



Fonte dos dados básicos: IBGE(2022). Elaborada pela autora (2024)

6.1.1 A Amostra Probabilística: Fragmento da População

As Pesquisas por amostragem probabilística frequentemente usam frações amostrais da ordem de 1%, 5% ou até mesmo 10%, justificadas por um estudo de planejamento de tamanho de amostra que prevê uma margem de erro máxima admissível para um determinado nível de confiança de interesse. Perante a circunstância experimental desse estudo, optou-se por considerar, ad hoc, uma amostra aleatória simples com tamanho correspondente a uma fração amostral de aproximadamente 10%, o que resultou em 556 municípios (o conjunto amostral).

Ao refletir as característica da população condizente com as estimativas do IODS 2.3.1, a tabela 4 expõe determinadas estatísticas descritivas e a figura 8 demonstra a representação gráfica da amostra.

Tabela 4 – Medidas descritivas de dados do indicador 2.3.1 no conjunto amostral

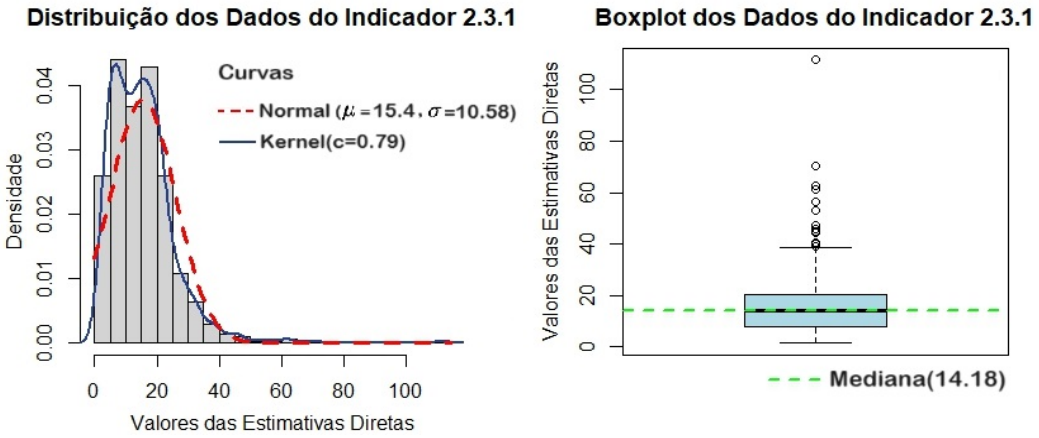
Estatísticas Descritivas	**Variável Dependente** Indicador ODS 2.3.1
Valor Mínimo	1,48
1º Quartil	7,71
Mediana	14,18
3º Quartil	20,18
Valor Máximo	111,57
Amplitude	110,09
IQR	12,47
Média	15,40
Variância	111,95
Desvio Padrão	10,58
Coef. de Variação	68,69
Assimetria	2,47
Curtose	14,29

Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

A média de 15,4 CPU e o desvio padrão de 10,6 CPU apontam uma considerável variabilidade, a variância é de 112 CPU^2. A mediana de 14,2 CPU sugere uma distribuição centralizada, porém a assimetria de 2,47 revela uma distribuição fortemente inclinada à direita, enquanto a alta curtose de 14,3 denota a presença de caudas longas e picos acentuados. As estatísticas descritivas também mostram que o primeiro quartil é 7,7 e 20,2 é o terceiro quartil, com um valor mínimo de 1,5 e máximo de 111,6.

Acerca da substancialidade da utilização dos métodos SAE atenta-se a quantidade de municípios por estado, por exemplo somente 16 municípios foram selecionados de Pernambuco, com 52 municípios o estado de Rondônia tem apenas 4 locais observados, enquanto nenhum

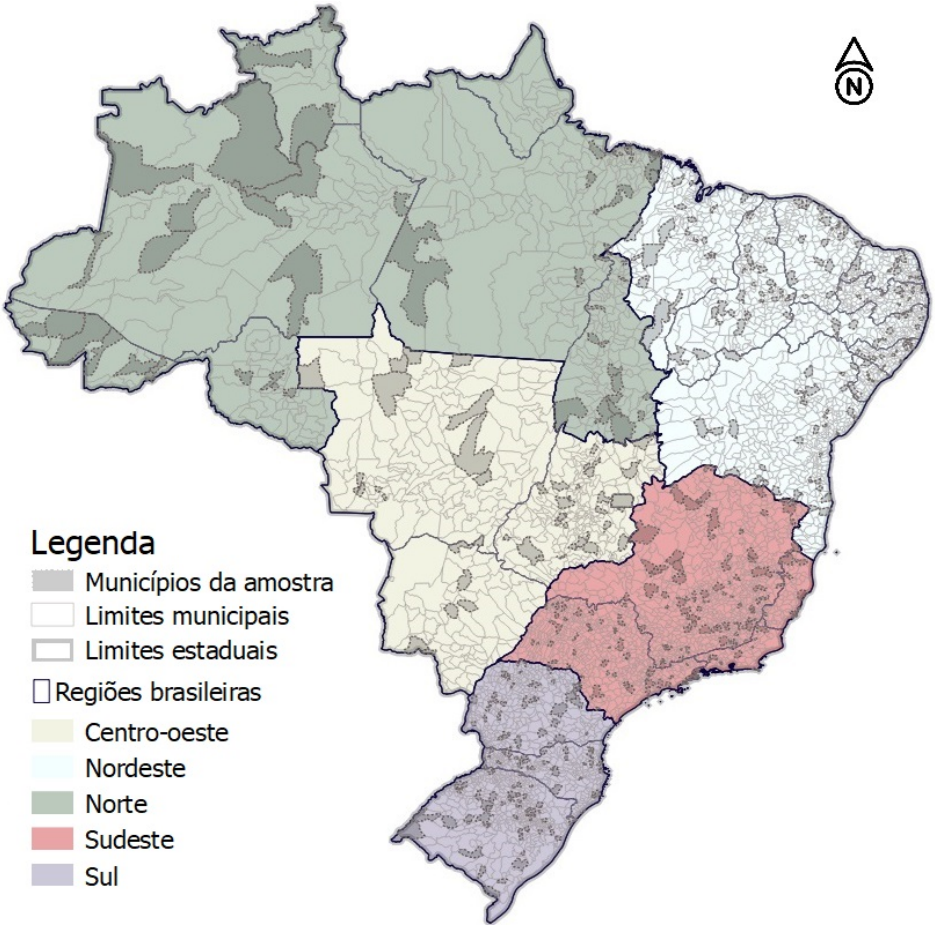
Figura 8 – Histograma e Boxplot das estimativas do IODS 2.3.1 de amostra



Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

município do Amapá foi elegido na amostra (figura 9). O quantitativo dos municípios pertencentes ao conjunto populacional e amostral, por Unidade da Federação (UF) e região, é evidenciado na tabela 5.

Figura 9 – Distinção dos municípios da amostra por região do Brasil



Fonte dos dados básicos: IBGE(2022). Elaborada pela autora (2024)

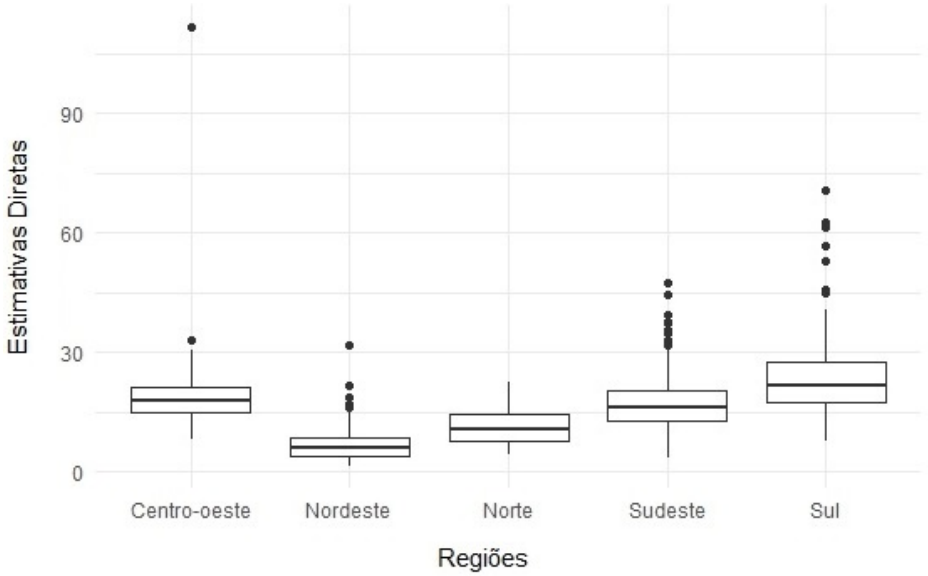
Tabela 5 – Representação do quantitativo municipal no Brasil dos conjuntos de dados

Território Brasileiro		Total do Conjunto			
Unidade Federativa	Região	Populacional		Amostral	
		Municipal	Regional	Municipal	Regional
Rondônia	Norte	52	450	4	55
Acre		22		7	
Amazonas		62		9	
Roraima		15		4	
Pará		144		15	
Amapá		16		0	
Tocantins		139		16	
Maranhão	Nordeste	217	1793	19	171
Piauí		224		29	
Ceará		184		19	
Rio Grande do Norte		167		16	
Paraíba		223		17	
Pernambuco		185		16	
Alagoas		102		16	
Sergipe		75		7	
Bahia		416		32	
Minas Gerais	Sudeste	853	1660	83	152
Espírito Santo		78		8	
Rio de Janeiro		90		10	
São Paulo		639		51	
Paraná	Sul	399	1190	42	130
Santa Catarina		295		31	
Rio Grande do Sul		496		57	
Mato Grosso do Sul	Centro-oeste	79	467	6	48
Mato Grosso		141		12	
Goiás		246		29	
Distrito Federal		1		1	

Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

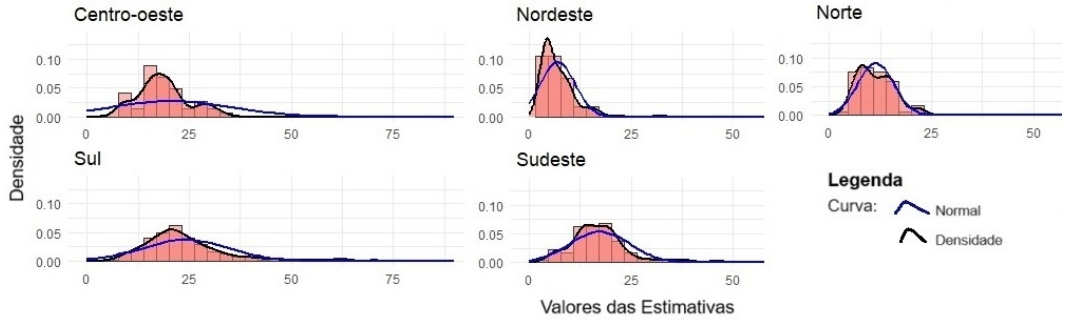
Toda a sistematização deve acomodar a variabilidade entre áreas, garantindo que as estimativas sejam precisas mesmo com tamanhos amostrais pequenos ou inexistente em certos domínios. E assim, se tornar essencial para a tomada de decisões, alocação de recursos e formulação de políticas públicas. Dessa forma, pondera-se um tema complexo e essencial para entender a geografia, a economia, a cultura e as políticas públicas do Brasil, a diversificação das regiões brasileiras. As cinco macrorregiões brasileiras apresentam características distintas que influenciam diretamente nos estudos e na formulação de política estatal. Na próxima visualização gráfica, as figuras 10 e 11 comprovam essas particularidades.

Figura 10 – Boxplot dos resultados do IODS 2.3.1 por região do Brasil da amostra



Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

Figura 11 – Histograma dos valores do IODS 2.3.1 por região do Brasil da amostra



Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

6.1.1.1 Abordagens para Modelar a Distribuição dos Dados

Para a análise subsequente na atribuição dos modelos de regressão é inevitável aplicar algumas estratégias para atender às suposições do modelo. Sucintamente, os resultados das estatísticas descritivas do conjunto populacional e amostral apontam que os dados apresentam uma assimetria à direita, existe uma maior concentração de valores nas porções inferiores da distribuição, com uma cauda estendendo-se para a direita em direção aos valores mais altos. Além de um coeficiente de assimetria positivo, a média é maior que a mediana e a presença de valores extremamente altos contribui para esse padrão. A princípio, identificar os *outliers* que poderiam estar distorcendo a distribuição foi importante para compreender sua influência sobre a distribuição dos dados e no entendimento do fenômeno em estudo. O experimento dessa dissertação explorou regras com base em critérios numéricos na identificação de *outliers* como:

no cálculo do intervalo interquartil (IQR), na determinação dos limites para valor abaixo de Q1-(1,5*IQR) ou acima de Q3+(1,5*IQR) e no número de desvios padrão (valores que estão a mais de 3 desvios padrão da média).

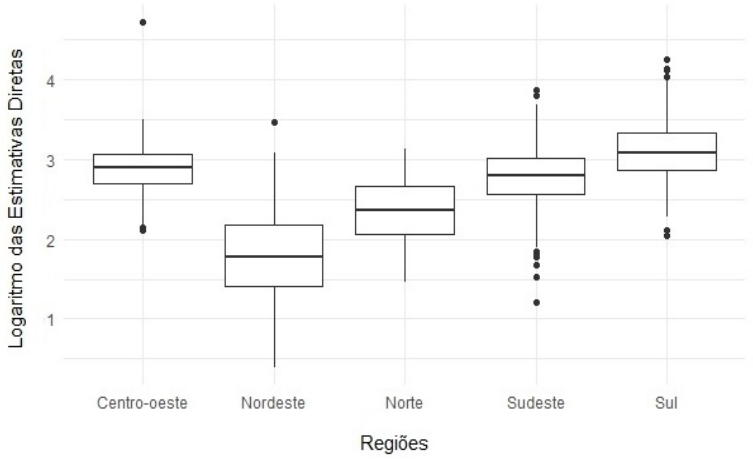
A escolha da técnica apropriada depende do problema específico em questão, das suposições sobre os dados e dos propósitos da análise. Se possível, para selecionar a distribuição mais adequada, é preferível usar o conjunto de dados da população. No entanto, muitas vezes, apenas os dados da amostra estão disponíveis, e é a partir deles que a seleção deverá ser feita. Desse modo, como a denotação de simulação nessa dissertação tem como escopo recriar uma situação real de estimação, as abordagens e análises são aplicadas diante da amostra. Todavia, a distribuição encontrada pode não representar completamente a distribuição da população.

Em seguida, é efetuada a aplicação de transformações de dados com o intuito de reduzir a assimetria e tornar a distribuição mais simétrica. Primeiramente, diante de transformações monotônicas, como a logarítmica ou a raiz quadrada. Também a verificação da transformação de Box-Cox nos dados e de ajuste de modelos estatísticos, capazes de capturar a conformação da distribuição dos dados. E técnicas de normalização, como a padronização (subtração da média e divisão pelo desvio padrão), são indagadas e averiguadas para transformar os dados em uma escala que porporcine a obtenção de estimativas mais precisas dos dados.

As diferenças macrorregionais podem influenciar a eficácia de modelos de regressão. Perante a análise exploratória já descrita, para a análise subsequente da modelagem estatística é inevitável que sejam consideradas. De fato, essa diversificação influenciou na simulação e algumas estratégias foram adotadas, mas foi consolidada a transformação logarítmica.

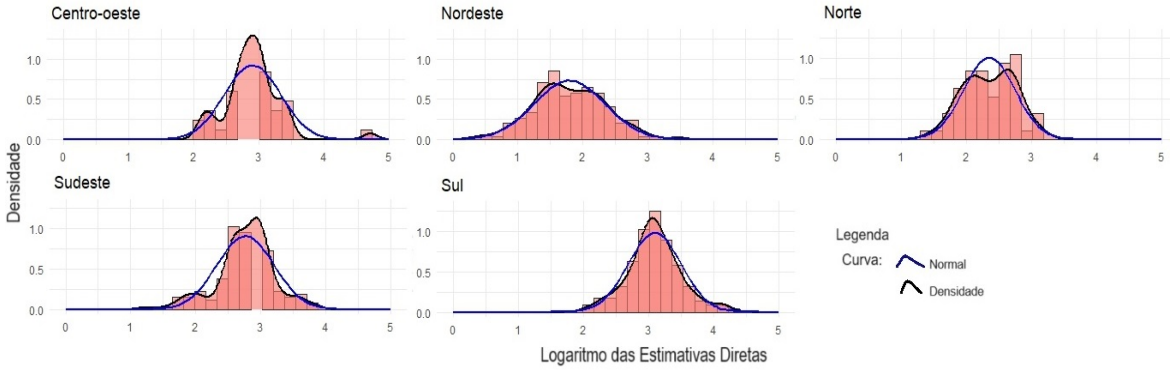
As imagens 12 e 13 representam os gráficos da transformação logarítmica (logaritmo neperiano) aplicada aos dados do indicador 2.3.1 (LODS) por região brasileira. Essa transformação resultou em uma distribuição significativamente menos assimétrica e com menor curtose das estimativas do IODS 2.3.1. A média e a mediana estão próximas, indicando que a distribuição central dos dados é mais simétrica do que antes. A dispersão reduzida, evidenciada pelo desvio padrão, sugere que os valores estão mais concentrados em torno do centro da distribuição. A assimetria e a curtose indicam uma distribuição mais normalizada e menos propensa a *outliers*. Porém, revelam que existem variações na distribuição dos dados, justamente a diversidade no Brasil. As regiões Centro-Oeste e Sudeste apresentam assimetrias mais acentuadas e curtoses positivas, com a indicação da presença de valores extremos. A região Sul tem uma distribuição levemente assimétrica e com caudas um pouco mais pesadas que a normal. E as regiões Norte e Nordeste possuem distribuições mais simétricas e achatadas, com caudas mais leves.

Figura 12 – Boxplot do logaritmo dos resultados do IODS 2.3.1 da amostra por região do Brasil



Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

Figura 13 – Histograma do logaritmo dos valores do IODS 231 da amostra por região do Brasil



Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

A tabela 6 exibe as estatísticas descritivas em um comparativo entre as estimativas do IDOS 2.3.1 com relação aos dados transformados pelo logaritmo neperiano dos valores desse indicador por região brasileira, o LODS.

Tabela 6 – Medidas descritivas dos valores do indicador 2.3.1 da amostra por região do Brasil

Estatísticas Descritivas	Macrorregiões Brasileiras									
	Norte		Nordeste		Centro-Oeste		Sudeste		Sul	
	ODS	LODS	ODS	LODS	ODS	LODS	ODS	LODS	ODS	LODS
Valor Mínimo	4,30	1,46	1,48	0,39	8,25	2,11	3,37	1,21	7,75	2,05
Mediana	10,5	2,36	5,9	1,78	18,0	2,89	16,3	2,79	21,7	3,08
Valor Máximo	22,7	3,12	32,0	3,47	111,6	4,71	47,4	3,86	70,5	4,26
IQR	6,62	0,614	4,68	0,762	6,49	0,362	7,36	0,451	10,27	0,462
Média	11,33	2,35	6,99	1,79	20,51	2,90	17,33	2,76	24,06	3,10
Variância	19,2	0,156	17,7	0,299	216,7	0,187	55,0	0,193	116,5	0,166
Desvio Padrão	4,38	0,394	4,20	0,547	14,72	0,433	7,42	0,439	10,79	0,408
Assimetria	0,57	-0,1097	2,01	0,0874	4,90	1,1758	1,24	-0,5406	1,70	0,1678
Curtose	-0,38	-0,872	7,03	-0,111	27,59	4,543	2,70	1,087	3,90	0,448

Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

6.1.2 Seleção de Variáveis Explicativas

Inicialmente, antes da composição da base auxiliar por 87 variáveis, uma seleção manual é organizada de acordo com a compatibilização das bases de dados explicativas apuradas para evitar a duplicidade das variáveis e reduzir o próprio conjunto. Posteriormente, é determinante constatar a possibilidade do tipo de ausência no preenchimento dos dados, assim as correspondências de NA (*Not Available*), traço (-) e célula vazia são cuidadosamente avaliadas. A presença desses termos pode ocasionar diferentes implicações e são capazes de ocorrer por diversas razões, como a indicação que um valor específico está ausente ou não foi coletado. Então, como uma das formas de tratamento foi efetuada a análise com exclusão, ou seja, a remoção das variáveis que contêm essas informações.

Em exemplificação, para impedir a duplicação de variáveis, as características de culturas como arroz e café permaneceram apenas os elementos referentes aos indicadores temáticos do Censo Agro de 2017 do Brasil. Sem pormenorizar, no agrupamento das 87 variáveis, são adotados todos os indicadores difundidos desse Censo (2017), do estudo de Abdalla et al. (2022) os indicadores de temperatura, os dados de precipitação e altitude, biomas, a extensão total de acesso as vias abertas à circulação e de uso e ocupação do solo, da FAO (2022) os índices de vegetação e a quantidade de estabelecimentos Agropecuários.

Em seguida, dentre as abordagens que são possíveis considerar tem-se a análise de correlação. Primeiramente, é priorizada a relação entre duas variáveis. Para entender os valores dos percentuais de correlação, é necessário notar que eles variam entre -1 e 1. Um valor próximo de 1 indica uma correlação positiva perfeita, um valor próximo de -1 uma correlação negativa perfeita, já um valor próximo de 0 uma correlação fraca ou inexistente. É observada a colinearidade e a multicolinearidade na base de dados auxiliar na investigação de que alguns desses elementos podem estar altamente correlacionados entre si, pois isso pode causar instabilidade numérica e problemas de interpretação nos modelos de regressão. A próxima imagem (figura 14) mostra o resultado da filtragem das correlações entre as 87 variáveis independentes acima de 0,75 (módulo do número) a partir do método de Spearman em lingaguem R, em que foram selecionadas 13 variáveis explicativas.

O coeficiente de correlação de Spearman ($\rho_{\mathbf{r}}$) irá medir a força e a direção da associação monotônica entre variáveis, mas não necessariamente lineares. É uma técnica não paramétrica baseada nos postos (*rankings*) dos dados sendo útil em casos onde existe a presença de valores extremos, não exigindo que os dados provenham de duas populações com distribuição

normal. O valor de 0,75 é definido de forma arbitrária, porém dentro de um intervalo para uma correlação forte entre 0,7 e 0,9.

Figura 14 – Relação das variáveis preditivas com correlação alta

```
> #coeficientes de correlação são fortes (absolutos acima de 0.75)
> forte_corr <- which((abs(corr_aux) > 0.75) & (abs(corr_aux) < 1),
+                     arr.ind = TRUE)
```

	rn	cn		rn	cn
1	V1	ESTABEL_ODS	11	bio7_mean	bio4_mean
2	Total_Area_URB_ha	POP_2022	12	bio4_mean	bio7_mean
3	POP_2022	Total_Area_URB_ha	13	Salt_flat_mean	Mangrove_mean
4	ESTABEL_ODS	V1	14	Mangrove_mean	Salt_flat_mean
5	bio3_mean	bio1_mean	15	ASI_AS_S1_FAO	Mean_VHI_AS_S1_FAO
6	bio4_mean	bio1_mean	16	ndvi_adm1_FAO	Mean_VHI_AS_S1_FAO
7	bio1_mean	bio3_mean	17	Mean_VHI_AS_S1_FAO	ASI_AS_S1_FAO
8	bio4_mean	bio3_mean	18	ndvi_adm1_FAO	ASI_AS_S1_FAO
9	bio1_mean	bio4_mean	19	Mean_VHI_AS_S1_FAO	ndvi_adm1_FAO
10	bio3_mean	bio4_mean	20	ASI_AS_S1_FAO	ndvi_adm1_FAO

Fonte: Elaborada pela autora (2024)

No entanto, é importante destacar que correlação não implica causalidade, e outros fatores podem influenciar a relação observada, exigindo uma interpretação cuidadosa dos resultados. Uma correlação alta entre a variável resposta e as independentes sugere que a auxiliar é uma boa preditora da variável dependente, o que pode fortalecer a validade da teoria ou hipótese subjacente (as previsões teóricas são suportadas pelos dados empíricos). Desse modo, é estabelecido o preceito de selecionar dentre os termos retidos no quesito colinearidade apenas os que apresentam uma correlação alta com a variável dependente. Ou seja, pelo método de Spearman será medida a intensidade da associação das variáveis auxiliares com a transformação logarítmica das estimativas do indicador do indicador ODS 2.3.1 (variável resposta), e entre as 13 variáveis explicativas selecionadas da figura 14 permanecerá aquela em que o resultado do coeficiente $|\rho_{\mathbf{r}}|$ (em módulo) da associação com relação a variável resposta é o mais alto.

A tabela 7 exibe os resultados da correlação de Spearman entre algumas variáveis explicativas e a resposta, em que a V37 (% de Assistência Técnica) representa a correlação mais alta da base de dados auxiliar com a variável resposta sendo $\rho_{\mathbf{r}} = 0{,}691$. Em resumo, no conjunto amostral o $|\rho_{\mathbf{r}}|$ varia de 0,008 a 0,691, indicando diferentes níveis de associação entre a variável **LODS** e as variáveis explicativas. Todo o agrupamento compreende uma correlação fraca à moderada com relação a variável dependente.

Por conseguinte, quantificar a multicolinearidade antes de aplicar técnicas de seleção automática de variáveis ou ajustar o modelo referido é essencial para garantir a estabilidade, interpretabilidade e eficácia do modelo. Justamente, para prevenir problemas decorrentes de variáveis altamente correlacionadas, assegurando que as inferências e decisões tomadas com

Tabela 7 – Valores do coeficiente correlação de Spearman entre algumas variáveis independentes

Associações com a variável LODS 2.3.1	
Variável Auxiliar	$\lvert\rho_{\mathbf{r}}\rvert$
V37	0,691
V25	0,687
ASI_AS_S1_FAO	0,658
Mean_VHI_AS_S1_FAO	0,655
bio7_mean	0,646
IDHM	0,644
bio4_mean	0,589
Caatinga_mean	0,587
bio3_mean	0,585
Savanna_formation_mean	0,580
bio1_mean	0,577
V26	0,554
ndvi_adm1_FAO	0,493
ESTABEL_ODS	0,394
V1	0,247

Fonte: Elaborada pela autora (2024)

base no modelo de regressão sejam precisas e confiáveis. Dessa maneira, tratato por Montgomery e Runger (2003), o fator de inflação da variância (VIF) é uma medida usual para efetuar essa avaliação e as diretrizes de interpretação dos resultados quantificados em: até 1, as variáveis auxiliares não são correlacionadas; entre 1,1 e 5, são relativamente correlacionadas, mas não significativas; entre 5,1 e 9,9, moderada multicolinearidade; apenas a partir de 10, a multicolinearidade afeta substancialmente as previsões do modelo, podendo resultar em sérios danos. Se o VIF ultrapassar 10, é crucial mitigar a multicolinearidade eliminando variáveis preditoras. Contudo, a remoção das variáveis deve ser retirada um de cada vez (REIS; COUTO; FERNANDES, 2015). O Dr. Iain Pardoe (2018) complementa no seu material que um VIF quantifica o quanto a variância é inflada, ou seja, as variâncias dos coeficientes de regressão estimados são inflados quando existe multicolinearidade. E existe um VIF para cada variável independente (preditora ou preditor) em um modelo de regressão. O Pardoe (2018) exemplifica que, $\mathrm{VIF_j}$ é apenas o fator pelo qual a variância do coeficiente de regressão estimado $\mathrm{b_j}$ é "inflada" pela existência de correlação entre as variáveis preditoras no modelo. O VIF para o j-ésimo preditor é dado por:

$$\mathrm{VIF_j} = \frac{1}{(1 - \mathrm{R_j^2})} \, ,$$

em que o $\mathrm{R_j^2}$ é o valor do coeficiente de determinação (R-quadrado) obtido ao regredir o j-ésimo preditor sobre as variáveis independentes restantes. Quando o resultado ultrapassa a

diretriz estabelecida, a variável explicativa está mais relacionada aos outros preditores do que à variável resposta.

Contudo, no estudo dessa dissertação é utilizada a medida fornecida por Fox e Weisberg (2019). Esses autores afirmam que $\mathrm{GVIF}^{1/(2\times \mathrm{df})}$ é a medida destinada à ser comparável entre termos (coeficientes de regressão) de diferentes dimensões, em que df são os graus de liberdade associados ao termo e GVIF denominado de VIF generalizado. Nessa condição, as primeiras variáveis detectadas acima de 10 são AREA, V8, V9, V10, V11, V12, V13, V14, V15, Amazon_Forest_mean e Cerrado_mean, apresentando forte multicolinearidade. A V10 (% de Atividade-Pecuária) registrou o maior valor do grupo sendo de 679301,68. Em seguida, após excluir a V10 e repetir o procedimento, é detecado o maior valor para AREA de 50,87.

Por fim, os componentes bio4_mean, Mean_VHI_AS_S1, ndvi_adm1, V1, V10, POP e AREA foram retirados do conjunto da base de dados auxiliar. A variável bio3_mean a cada repetição estava elevando o VIF dos demais, então o procedimento só é estabilizado após a substituição por bio1_mean. Exclusivamente, permanecem Mangrove_mean e Salt_flat_mean embasados na etapa de levantamento dos dados. Consequentemente, de acordo com a amostra obtida, a base de dados auxiliar é instituída com 80 variáveis explicativas.

6.1.2.1 Técnicas Automáticas para Seleção de Variáveis Explicativas

A seleção de variáveis explicativas envolve a escolha de um subconjunto de variáveis que contribuem significativamente para o modelo, eliminando aquelas que são redundantes ou irrelevantes. Isso auxilia a melhorar a interpretabilidade do modelo e reduzir sua complexidade, a eficiência computacional e a prevenção do fenômeno de sobreajuste (*overfitting*).

Após uma investigação detalhada de algumas dessas técnicas, no estudo dessa dissertação é priorizado o método *Stepwise*. Todavia, notabilizam-se a técnica de regularização Lasso (*Least Absolute Shrinkage and Selection Operator*), REF (*Recursive Feature Elimination*) e algoritmos de florestas aleatórias (*Random Forest*).

Sinteticamente, o componente de regularização Lasso (L_1) utiliza uma soma de valores absolutos de parâmetros e um coeficiente de penalização que os encolhe para zero. E se adequa para extração de variáveis no modelo, pois vincula pesos nulos a parâmetros que têm contribuição limitada para efeito de previsão (MORETTIN; SINGER, 2022). RFE é uma técnica que seleciona recursivamente um subconjunto de variáveis, esse método funciona removendo iterativamente as variáveis menos importantes do modelo, com base em uma métrica de

desempenho e avaliando o aproveitamento do modelo resultante em um conjunto de validação cruzada. Já o uso de algoritmos baseados em "floresta aleatória" têm a capacidade embutida de calcular a importância dos elementos, o IncNodePurity (Pureza Incremental do Nó) é uma métrica de importância que designa o quanto cada variável contribui para a redução da impureza nos nós das árvores. E o seu principal o objetivo é dividir os dados em subgrupos cada vez mais homogêneos com relação a variável resposta.

O método *Stepwise* é uma abordagem iterativa que adiciona ou remove variáveis do modelo com base em critérios predefinidos, como o AIC (Critério de Informação de Akaike), BIC (Critério de Informação Bayesiano) ou a especificação em nível de significância do valor F, a "estatística F". É essencial eleger as variáveis mais relevantes considerando coeficientes significativos (valor-$\mathrm{p} < 0{,}05$) e um bom ajuste do modelo (por exemplo, um R^2 alto).

Com pressuposto da diversidade brasileira, ao dividir a amostra por regiões, o estudo considera a variabilidade geográfica e socioeconômica garantindo que em um nível mais desagredado as áreas (estados, municípios) sejam devidamente representadas na amostra, respeitando suas particularidades. Portanto, nos demais procedimentos é priorizada a região Nordeste, com 171 municípios observados na amostra.

O código fonte elaborado via SAS® por sua usabilidade e praticidade de definir os argumentos, aplica o procedimento *Stepwise* na variável depende LODS na região Nordeste do Brasil. O AIC é a métrica atribuída baseada na função de log-verossimilhança penalizada pela quantidade de parâmetros do modelo, ou seja, o propósito é identificar o melhor ajuste aos resultados a partir do menor número de parâmetros e não requer testes estatísticos. Com a condição de entrada em nível de significância de 0,1 e permanência 0,05.

Código Fonte 1 – *Script* do procedimento *stepwise* em SAS.

```
/* Selecao de Variaveis Independentes */
        /* Metodo Stepwise */
proc glmselect data=amostra_ne plot=CriterionPanel;
model LODS_231 = *Inserir todas as variaveis*
/selection = stepwise (select=SL SLE=0.1 SLS=0.05 choose=AIC) stats=all;
run;
```

Fonte: Elaborada pela autora (2024)

Os resultados dessa implementação em SAS com as 80 variáveis auxiliares são expostos a seguir na figura 15. As variáveis mais relevantes captadas para o Nordeste envolve um total de 13 variáveis. O R^2 (R-quadrado) é 0,6612, significa que aproximadamente 66$\%$ da variabilidade na variável dependente é explicada pelas variáveis independente incluídas no

modelo. Essencialmente, quanto mais próximo o R^2 estiver de 1 melhor o modelo se ajusta aos dados. O *Adjusted R-squared* (R-quadrado ajustado) de 0,6331 foi o maior quantitativo em comparação com as outras aplicações mencionadas. As métricas de ajuste, com valores extremamente baixos de AIC e BIC sugerem um modelo melhor ajustado.

Figura 15 – Seleção de variáveis independentes pelo método *stepwise*

Stepwise Selection Summary

Step	Effect Entered	Number Effects In	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	Intercept	1	0.0000	0.0000	-32.6002	-32.5288	-205.4226	235.8061	-202.4586	51.3856	0.2970	0.00	1.0000
1	V8	2	0.1675	0.1626	-61.9488	-61.8051	-235.4044	170.0000	-228.6655	43.2598	0.2472	34.00	<.0001
2	V19	3	0.3038	0.2955	-90.5291	-90.2882	-264.1977	116.8150	-254.1042	36.6543	0.2068	32.90	<.0001
3	V21	4	0.4371	0.4270	-124.8648	-124.5012	-297.9859	64.8669	-285.2982	29.8751	0.1672	39.54	<.0001
4	V33	5	0.4864	0.4740	-138.5328	-138.0206	-311.3714	46.9163	-295.8245	27.6845	0.1525	15.93	<.0001
5	V20	6	0.5270	0.5127	-150.6291	-149.9420	-322.9346	32.4642	-304.7792	25.9548	0.1405	14.18	0.0002
6	V16	7	0.5744	0.5588	-166.6670	-165.7781	-337.7725	15.2964	-317.6754	23.9208	0.1264	18.24	<.0001
7	ID_GINI	8	0.5986	0.5814	-174.6843	-173.5663	-344.8521	7.4932	-322.5510	22.7816	0.1192	9.83	0.0020
8	Mean_VHI_AS_S2_FAO	9	0.6119	0.5927	-178.4295	-177.0545	-347.8556	4.1244	-323.1546*	22.3391	0.1153	5.54	0.0198
9	V6	10	0.6223	0.6012	-181.1176	-179.4572	-349.7745	1.8753	-322.7010	21.8791	0.1122	4.47	0.0359
10	V34	11	0.6330	0.6101	-184.0185	-182.0438	-351.7404	-0.4440	-322.4602	21.5185*	0.1090	4.65	0.0325
11	V29	12	0.6429	0.6182	-186.6712	-184.3527	-353.3501	-2.4315	-321.9712	21.9683	0.1061	4.39	0.0378
12	POP_Area_nao_densa	13	0.6525	0.6261	-189.3438	-186.6515	-354.8312	-4.3284	-321.5021	22.8821	0.1032	4.38	0.0380
13	V5	14	0.6612	0.6331*	-191.6653*	-188.5686*	-355.8792*	-5.8389*	-320.6821	22.9100	0.1006	4.02	0.0467

* Optimal Value of Criterion

Selection stopped because the candidate for entry has SLE > 0.1 and the candidate for removal has SLS < 0.05.

Fonte: Elaborada pela autora (2024)

A figura 16 apresentada os resultados da análise de variância (ANOVA):

Figura 16 – A análise de variância com as informações sobre a significância do modelo

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	13	33.57831	2.58295	23.57
Error	157	17.20801	0.10961	
Corrected Total	170	50.78632		

Fonte: Elaborada pela autora (2024)

Na análise anterior identifica-se um valor de 2,58 do quadrado médio e 0,11 é o erro quadrático médio, então a sua raiz é 0,33. Logo, 23,57 é o valor-F, sendo muito maior que o valor-F tabelado da distribuição F de Fisher-Snedecor (1,80) ao 5% nível de significância, 13 graus de liberade do modelo e 157 graus de liberdade. Assim, o modelo é significativamente melhor do que um modelo sem variáveis (modelo nulo).

A ANOVA mostrou evidências de que os grupos não possuem médias iguais, e existe pelo menos um grupo que se diferencia significativamente dos outros. Em geral, o modelo se ajusta

bem aos dados observados, possui capacidade preditiva razoável e é parsimônico, não inclui variáveis desnecessárias. Em geral, o modelo se ajusta bem aos dados observados, possui capacidade preditiva razoável e é parsimônico, não inclui variáveis desnecessárias.

Consequentemente, o cojunto de variáveis independentes para a região Nordeste deve ser formado por:

- V5 - Área lavoura / Colheitadeira (ha): Média da área de lavouras por colheitadeira, por município, calculada pelo quociente entre a área total de lavouras e a quantidade de colheitadeiras;
- V6 - Área lavoura / Semeadeira (ha): Média da área de lavouras por semeadeira ou plantadeira, por município, calculada pelo quociente entre a área total de lavouras e a quantidade de semeadeiras;
- V8 - Atividade-Lavoura Temporária (%): Percentual de estabelecimentos pertencentes ao Grupo de Atividade Econômica Produção de lavouras temporárias, em relação ao total de estabelecimentos agropecuários do município;
- V16 - Uso das terras-Lavoura (%): Percentual de área classificada como lavoura (temporária + permanente) no tema Utilização das Terras, em relação à área total dos estabelecimentos agropecuários do município;
- V19 - Aves-Ovos (%): Percentual de estabelecimentos cuja finalidade principal da criação de aves é Produção de ovos, em relação ao total de estabelecimentos agropecuários do município;
- V20 - Bovinos-Corte (%): Percentual de estabelecimentos cuja finalidade principal da criação de bovinos é Corte, em relação ao total de estabelecimentos agropecuários do município;
- V21 - Bovinos-Leite (%): Percentual de estabelecimentos cuja finalidade principal da criação de bovinos é Leite, em relação ao total de estabelecimentos agropecuários do município;
- V29 - Rendimento-Café (kg/ha): Rendimento médio de café, definido pelo quociente entre a produção e a área colhida, por município;

- V33 - Cisterna (%): Percentual de estabelecimentos agropecuários com cisterna, em relação ao total de estabelecimentos no município;
- V34 – Utilização de Agrotóxicos (%): Percentual de estabelecimentos agropecuários com declaração de uso de agrotóxicos em relação ao total de estabelecimentos agropecuários no município. A variável totaliza os que usaram agrotóxicos e os que não precisaram usar em 2016;
- ID_GINI: O Índice de Gini é usado para medir o grau de concentração de renda em determinado grupo. Ele aponta a diferença entre os rendimentos dos mais pobres e dos mais ricos em que o resultado consiste em um número entre 0 e 1, onde 0 corresponde à completa igualdade;
- Mean_VHI_AS_S2: O Índice de Saúde da Vegetação (VHI) é calculado a partir de dados do Índice de Vegetação por Diferença Normalizada (NDVI), resultando da composição de dois subíndices, Índice de Condição da Vegetação (VCI) e o Índice de Condição de Temperatura (TCI). O VHI é um índice composto utilizado para computar o Índice de Estresse Agrícola (ASI) e permite identificar o início/fim da condição de estresse vegetativo, área afetada, intensidade e duração da seca e a sua relação com os eventuais impactos;
- POP_Area_nao_densa: A população total em área não densa que auxilia na identificação e classe do grau de urbanização municipal.

6.2 AJUSTE DE MODELO ESTATÍSTICO

A modelagem por região deve capturar a variabilidade geral e as características compartilhadas pelos municípios e permitir que a amostra também proporcione uma quantidade adequada de observações. Neste caso, o modelo pode utilizar as covariáveis e combinar dados de diferentes níveis para melhorar a precisão das estimativas, mesmo em áreas com poucas observações. Quando se tenta ajustar o modelo apenas por estado, a baixa quantidade de dados impede a utilização direta de métodos de regressão, já que o número de observações é insuficiente para garantir o ajuste. Uma vez que o modelo esteja ajustado para a região como um todo, ele pode ser usado para derivar estimativas em um nível mais desagregado dentro dessa região (e.g., estado), mesmo que a quantidade de dados em cada local seja pequena.

Em linguagem de programação R e apoiado em SAS, são ajustados dois modelos de regressão para fomentar e alcançar estimativas mais precisas do indicador 2.3.1. O estimador sintético por regressão é obtido através do método de QMO e os resultados ponderados, assim com os novos resultados designado o estimador composto. Em seguida, é aplicado o modelo Fay-Herriot e adquirido o eblup. Na tabela 8 são comparadas as métricas calculadas nesses ajustes, considerando a amostra da região Nordeste que contém 171 municípios, as 13 variáveis auxiliares selecionados pelo método *Stepwise* e o LODS 2.3.1 sendo a variável resposta:

Tabela 8 – Métricas dos modelos ajustados

Métricas	**Modelos por Regressão**	
	Sintético	Fay-Herriot
R-quadrado	0,661	0,67
R-quadrado ajustado	0,633	0,633
EQM	0,101	0,00413
REQM	0,317	0,0643
EAM	0,249	0,0253
EPAM	0,175	0,0114
Variância	0,11	0,0983

Fonte: Elaborada pela autora (2024)

Ambos os modelos constam o mesmo R-quadrado ajustado de 0,633. O R-quadrado com aproximadamente 67% indica a proporção razoável da variabilidade na variável dependente que pode ser explicada pelas variáveis independentes incluídas em cada modelo. O modelo FH apresenta os menores valores de EQM, REQM (Raiz do EQM), EAM (Erro Absoluto Médio) EPAM (Erro Percentual Absoluto Médio) e Variância, sendo um indício que ele é mais preciso e eficaz na previsão da variável resposta.

As tabelas 9 e 10 mostram os parâmetros estimados nos procedimentos de ajuste. Os coeficientes estimados para os dois modelos são bastante parecidos. O intercepto e a maioria das variáveis são estatisticamente significativos com valores-p menores que 0,05. Apenas a variável ID_GINI no ajuste do modelo FH com 0,68 está acima desse limiar. De qualquer forma, assegura que a desigualdade de renda exerce um impacto na variável dependente.

Em resumo, todas essas variáveis mostram significância estatística e têm efeitos variados (positivos e negativos) na variável dependente, indicando sua importância no modelo e desempenham papéis importantes na explicação da variabilidade da LODS 2.3.1.

A análise de resíduos é substancial para validar as suposições feitas durante o ajuste dos modelos de regressão. Na tabela 11 tem-se os resultados dos testes Shapiro-Wilk e Durbin-Watson para avaliar a normalidade dos resíduos e a presença de autocorrelação.

Tabela 9 – Coeficientes estimados por quadrados mínimos ordinários para o Nordeste

Parâmetros Estimados na Estimação Sintética por Regressão				
Variáveis	Estimativa	Erro Padrão	Valor-t	Valor-p
(Intercepto)	1,7149	0,464	3,70	0,0003
V8	-0,00673	0,00165	-4,09	<0,0001
V19	-0,00965	0,00153	-6,30	
V21	0,0242	0,00250	9,70	
V33	-0,00327	0,000974	-3,36	0,0010
V20	0,0111	0,00183	6,05	<0,0001
V16	0,00686	0,00174	3,93	0,0001
ID_GINI	0,626	0,306	2,04	0,04260
Mean_VHI_AS_S2_FAO	-1,15	0,578	-1,99	0,04794
V6	0,0000448	0,0000146	3,07	0,00255
V34	0,00296	0,00121	2,45	0,01542
V29	0,0000311	0,0000128	2,44	0,01585
POP_Area_nao_densa	0,00000397	0,00000164	2,42	0,01678
V5	-0,0000249	0,0000124	-2,00	0,04673

Fonte: Elaborada pela autora (2024)

Tabela 10 – Coeficientes estimados atráves do modelo FH o Nordeste

Parâmetros Estimados no Modelo Fay-Herriot				
Variáveis	Estimativa	Erro Padrão	Valor-t	Valor-p
(Intercepto)	1,77	0,467	3,79	0,0001
V8	-0,00673	0,00165	-4,07	<0,0001
V19	-0,00912	0,00153	-5,94	
V21	0,0236	0,00257	9,18	
V33	-0,0036	0,000965	-3,73	0,0002
V20	0,0102	0,00184	5,54	<0,0001
V16	0,00647	0,00173	3,73	0,0002
ID_GINI	0,557	0,305	1,83	0,06775
Mean_VHI_AS_S2_FAO	-1,15	0,575	-2,0	0,04523
V6	0,0000461	0,0000146	3,16	0,00156
V34	0,00339	0,00119	2,85	0,00441
V29	0,0000311	0,0000123	2,53	0,01139
POP_Area_nao_densa	0,00000455	0,00000166	2,74	0,00618
V5	-0,0000252	0,000012	-2,09	0,03660

Fonte: Elaborada pela autora (2024)

Tabela 11 – Aplicação de testes sob os resíduos dos modelos

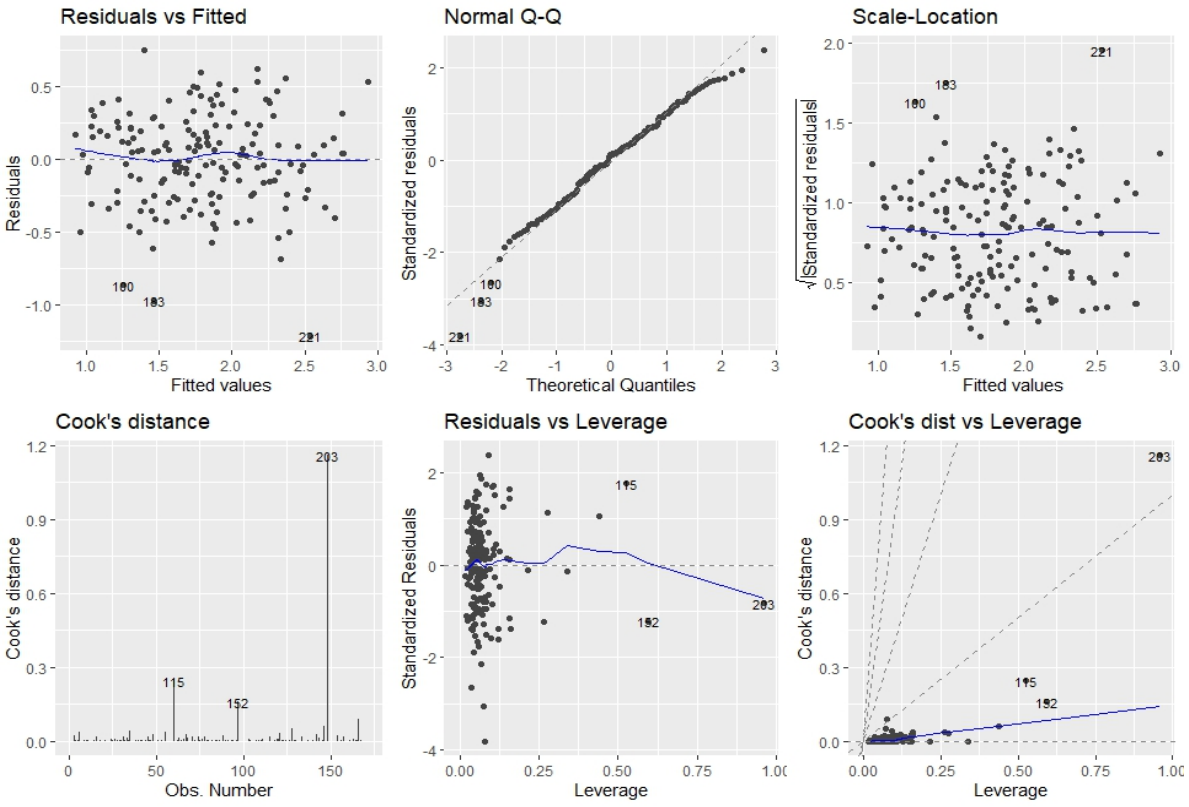
TESTES:		**Shapiro-Wilk**	**Durbin-Watson**
Termos	Modelo	Valor-p	Estatística
Resíduos	Linear	0.04	1.97
Resíduos Padronizados	simples	0.03	
Resíduos	Fay-Herriot	<0.0001	1.91
Resíduos Padronizados			1.84

Fonte: Elaborada pela autora (2024)

Ao nível de significância de 5%, para qualquer valor-p dos termos residuais dispostos na tabela anterior denota a rejeição da hipótese nula de normalidade. E os valores da estatística do teste de Durbin-Watson estão próximos de 2, apontando que não existe evidência significativa de autocorrelação nos resíduos do modelo, que os erros das observações adjacentes não estão correlacionados. Dessa maneira, os modelos apresentam resíduos que não seguem uma distribuição normal, porém a ausência de autocorrelação nos erros assinala que as estimativas dos coeficientes são válidas e não subestimam o erro padrão, tornando as inferências precisas.

O diagnóstico do ajuste dos modelos perante a variável LODS 2.3.1 é ilustrado nas sucessivas imagens de 17 até 20. O primeiro gráfico da figura 17 (*Residuals vs Fitted*) apresenta a linha de regressão na cor azul suavizada e os resíduos estão dispersos ao redor de zero sem um padrão claro. Entretanto, no *Scale-Location* a dispersão crescente para os valores ajustados maiores pode ser um indício de heterocedasticidade. No segundo gráfico, o *Normal Q-Q*, a maior parte dos resíduos segue a linha reta de normalidade, e se dispersando somente nas extremidades sem formar uma linha curva. Nos últimos três gráficos, notam-se observações que se destacam como influentes pelo possível alto impacto que podem exercer no ajuste do modelo. Especialmente, o domínio identificado pela numeração 203.

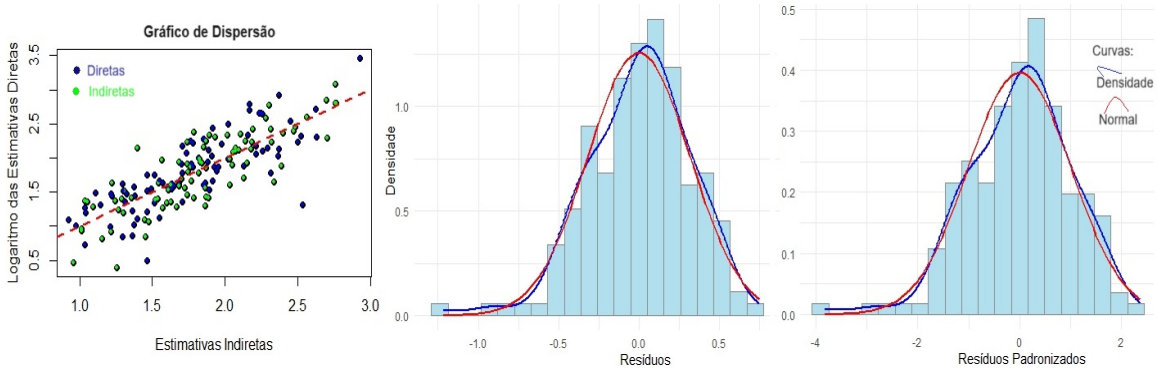
Figura 17 – Gráficos de resíduos do ajuste do modelo linear simples na região Nordeste



Fonte: Elaborada pela autora (2024)

O gráfico de dispersão da imagem 18 sugere que o modelo pode estar ajustando os valores de forma à "compensar" um possível viés, mas com variabilidade que reflete uma incerteza maior em áreas específicas. Em suma, os histogramas dos resíduos e dos resíduos padronizados exibem uma forma simétrica e as suas curvas de densidade ajustadas próximas a curva normal.

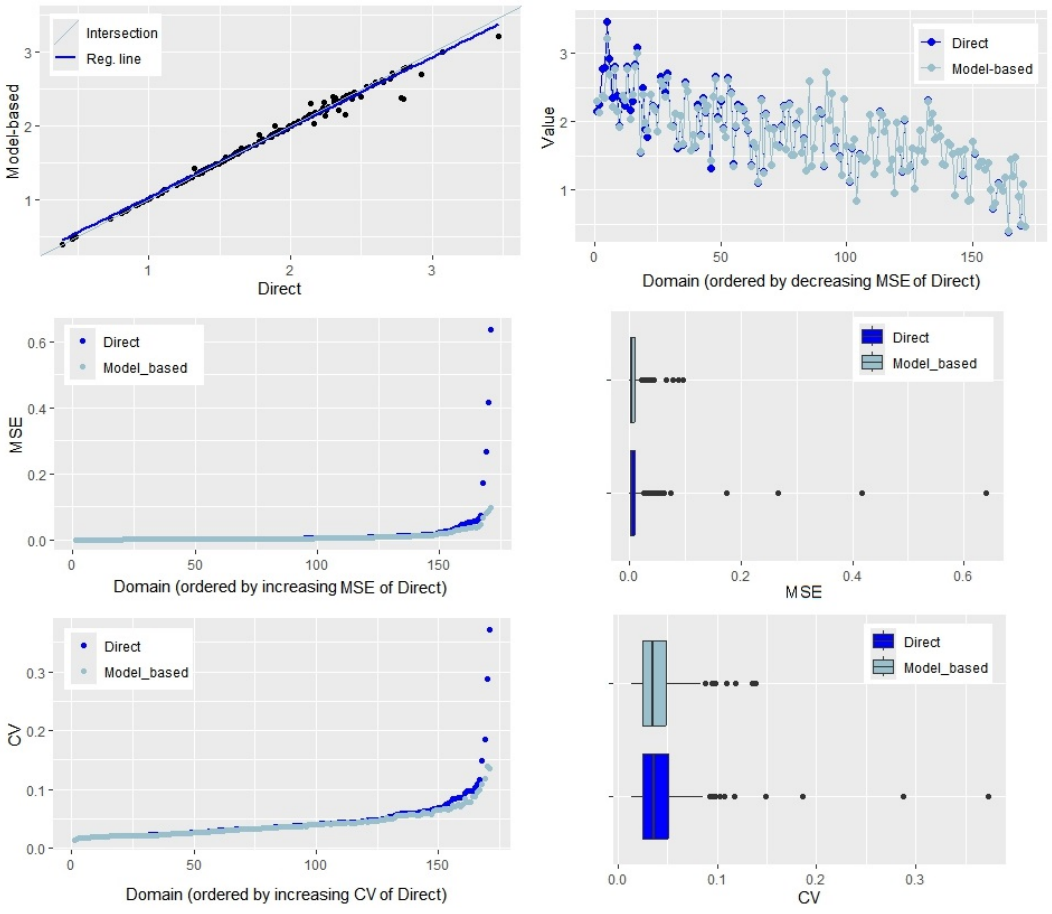
Figura 18 – Gráficos de resíduos do ajuste do modelo linear simples no Nordeste



Fonte: Elaborada pela autora (2024)

A figura 19 exibe os gráficos do ajuste pelo modelo FH na região Nordeste.

Figura 19 – Gráficos do ajuste pelo modelo FH na região Nordeste

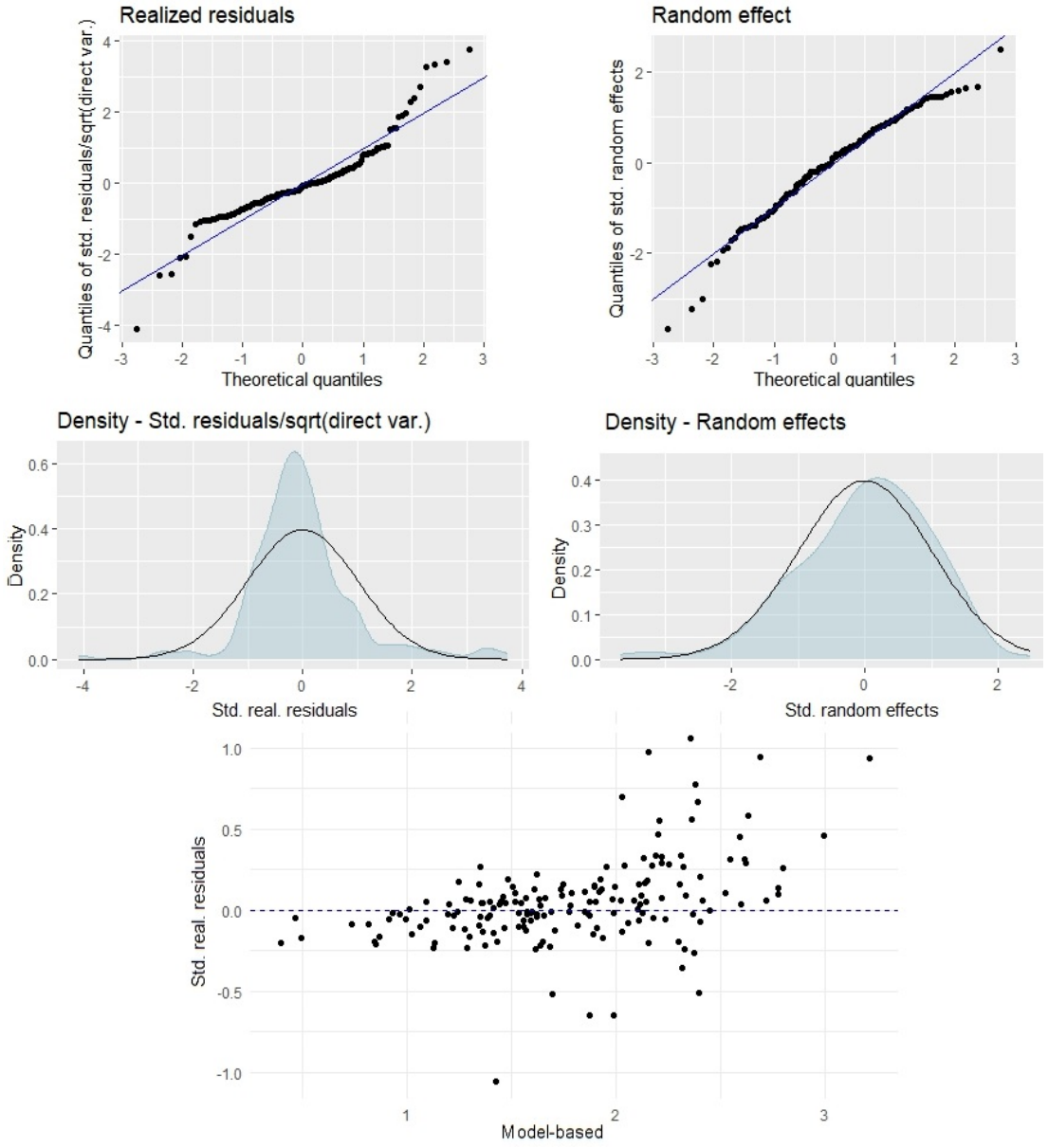


Fonte: Elaborada pela autora (2024)

O primeiro gráfico é o de dispersão onde as estimativas diretas e as indiretas estão fortemente alinhadas, como indicado pela linha de regressão azul e a linha diagonal de interseção. Com relação ao erro quadrático médio (*MSE*), as estimativas diretas e as obtidas no ajuste do modelo (*model-based* ou indiretas) seguem um padrão similar, com discrepâncias nos domínios com valores maiores. O boxplot *MSE* com a presença de *outliers* demonstra que em alguns domínios o ajuste não conseguiu reduzir substancialmmente o EQM. Acerca do CV, o CV das estimativas indiretas é consistentemente menor do que o das estimativas diretas, logo o modelo FH consegue diminuir os efeitos da alta variabilidade e do erro associados ao ajuste.

A figura 20 mostra os gráficos de erros pelo ajuste no modelo FH.

Figura 20 – Gráficos dos erros associados ao ajuste pelo modelo FH na região Nordeste



Fonte: Elaborada pela autora (2024)

É perceptível que os resíduos demonstram desvios substanciais da normalidade. Todavia, os gráficos de efeitos aleatórios (*Random effect*) indicam que eles se aproximam da normalidade. No gráfico Q-Q nota-se que o efeito se dispersa apenas nas extremidades. No seu histograma a curva de densidade apresenta uma forma simétrica e ajustada próxima a curva da normal. A variação não observada entre as áreas é essencialmente o conjunto de fatores que influenciam os resultados, mas que não foram explicitamente incluídos no modelo por não terem sido medidos. Os efeitos aleatórios no modelo FH servem para captar essas influências, e se esses efeitos estão bem ajustados, o modelo pode fornecer estimativas mais precisas, ajustando-se melhor às nuances e peculiaridades de cada área.

Sinteticamente, o modelo FH aplicado à região Nordeste mostra um desempenho adequado. Com algumas discrepâncias e leves assimetrias nas densidades, que devem ser investigadas, pois talvez afete a validade dos intervalos de confiança e testes de hipóteses. Entretanto, não comprometem significativamente a adequação do modelo para a maioria dos domínios.

Os gráficos de diagnóstico e as métricas de desempenho indicam que o modelo Fay-Herriot pode fornecer estimativas mais precisas que o ajuste do modelo linear simples por quadrados mínimos ordinários. O modelo FH demonstra maior precisão e menor variabilidade do estimador. As métricas apresentadas, como o EQM de 0,004 e 0,025 do EAM, são bem menores que o ajuste no modelo linear simples expondo estimativas mais acuradas. Contudo, a técnica pelo estimador sintético por regressão possui uma interpretação mais direta e pode ser preferível em ocasiões onde a simplicidade e a facilidade de implementação são cruciais.

Ao modelar por região a estimativa para pequenos domínios é refinada usando dados de áreas maiores e de covariáveis. Sob essas condições, o ajuste do modelo é feito em um nível mais agregado (região), onde existem dados suficientes, como no caso dos 171 municípios da região Nordeste. Ou seja, as características dos estados são modeladas dentro do contexto regional. E isso é possível devido ao compartilhamento de informações entre os municípios da mesma região. A partir desse ajuste, obtém-se as estimativas para os estados (e.g., Pernambuco). Os resultados para os estados (nível desagregado) são extraídos por essa agregação regional, permitindo que se utilizem os dados da região como um todo para "informar" as estimativas nos estados com menos observações.

No caso do estado de Pernambuco, com apenas 16 municípios na amostra, a quantidade de observações é insuficiente para suportar a complexidade de um modelo estatístico. Por essa razão, as estimativas para Pernambuco são derivadas do modelo ajustado no nível regional. A tabela 12 apresentada as estatísticas descritivas desses resultados.

Tabela 12 – Valores de medidas descrivitas referentres ao estado de Pernambuco

Estatísticas Descritivas	Variável Resposta	Estimativas			Diferenças Residuais		
	LODS_231	ST	COMP	FH	ST	COMP	FH
Valor Mínimo	0.46	0.78	0.82	0.46	-1.09	-1.10	-1.07
1º Quartil	1.35	1.44	1.52	1.41	-0.204	-0.27	-0.178
Mediana	1.73	1.70	1.70	1.67	-0.027	-0.004	-0.001
3º Quartil	2.18	1.94	1.93	1.94	0.323	0.289	0.305
Valor Máximo	3.09	3.11	2.91	3.17	1.65	1.46	1.69
Média	1.76	1.72	1.74	1.70	0.045	0.02	0.063
Desvio Padrão	0.54	0.39	0.33	0.41	0.42	0.44	0.42

Fonte dos dados básicos: IBGE(2017). Elaborada pela autora (2024)

O modelo FH possui a mediana do resíduo mais próxima de zero (-0,001), em média, não existe uma tendência sistemática de erro significativo. E a sua média (0,063) também é relativamente pequena. O 1º Quartil (-0,178) e o 3º Quartil (0,305) também estão bem distribuídos ao redor de zero, representando uma menor dispersão dos erros. No entanto, apresenta a maior amplitude entre os modelos (-1,07 a 1,69). Embora o resultado do estimador composto (COMP), calculado 0,2 o peso ξ, tenha uma média de resíduos mais próxima de zero (0,02), a mediana dos resíduos (-0,004) não é tão próxima de zero quanto a do modelo FH. Já os dados residuais do estimador sintético (ST) apresentam uma mediana mais distante de zero (-0,027), indicando uma maior presença de erros sistemáticos.

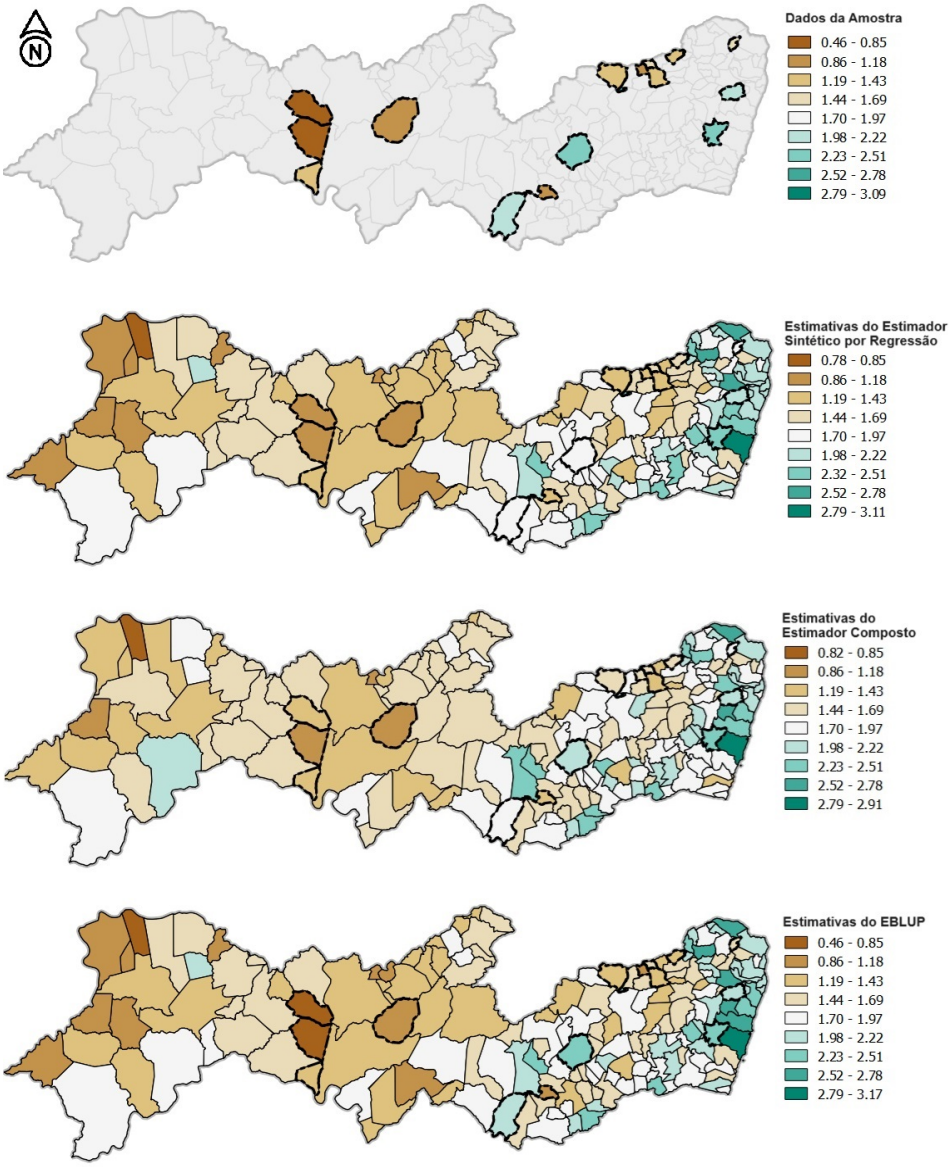
Na figura 21 tem-se a disposição espacial dos dados obtidos no processo de estimação, que também demonstra o comportamento do procedimento SAE no estado de Pernambuco.

A imagem 22 exibe os boxplot na comparação entre os métodos de estimação diante das estimativas obtidas em Pernambuco. O *outlier* mais extremo acima de cada boxplot assumindo o valor máximo, corresponde ao mesmo município. E o valor mais extremo abaixo do boxplot Composto e FH assumindo o valor mínimo, não corresponde ao mesmo município.

Nos gráficos de dispersão da figura 23, nota-se a mesma conclusão evidenciada no ajuste do modelo linear simples. Nesse caso os dois modelos podem estar ajustando os valores de forma a "compensar" um possível viés, mas com a variabilidade que reflete uma incerteza maior em determinados domínios.

A comparação ilustrada na figura 24 verifica a adequação e o desempenho dos diferentes modelos ajustado no Nordeste com os resultados das estimativas para Pernambuco. A variação observada em torno de zero para os três métodos indica que, no geral, não existe um viés sistemático claro, mas sim alterações no desempenho em diferentes faixas de valores das estimativas. Os pontos na cor azul representam o modelo FH que apesar de demonstrar variações

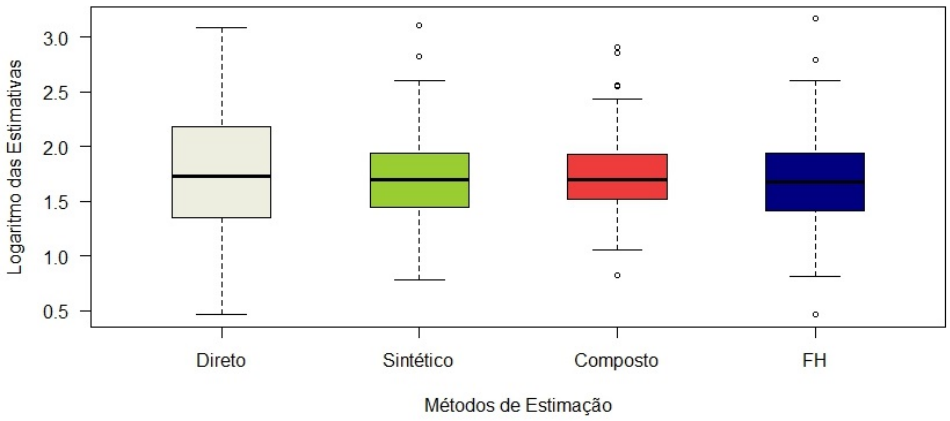
Figura 21 – Representação de estimação de pequenas áreas no estado de Pernambuco



Fonte dos dados básicos: IBGE(2022). Elaborada pela autora (2024)

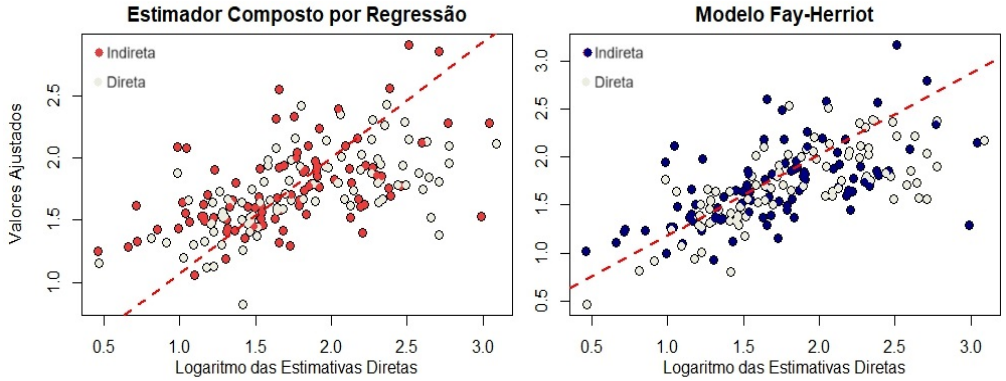
é o mais concentrado em torno da linha de referência no valor 0, e assim transmitindo uma maior precisão para os valores intermediários das estimativas. No modelo composto (pontos marrons) percebe-se uma dispersão considerável ao longo de todo o eixo x, o que indica uma variação maior nas diferenças entre as estimativas compostas e as diretas. Esse modelo é o mais espalhado ao longo de toda a gama do gráfico, sugerindo possíveis inconsistências no ajuste. E a variação dos pontos verdes (Sintético) é menor em torno da linha de referência, porém aumenta à medida que dos valores das estimativas diretas também aumentam. Logo, é um indício que nesse método existe uma maior precisão em uma faixa de valores menores das estimativas. Entretanto, a diferença nos resíduos entre os procedimentos é pequena, colocando que o estimador composto ainda pode fornecer estimativas razoavelmente boas (tabela 12).

Figura 22 – Boxplot das estimativas em PE dos métodos de estimação de pequenas áreas



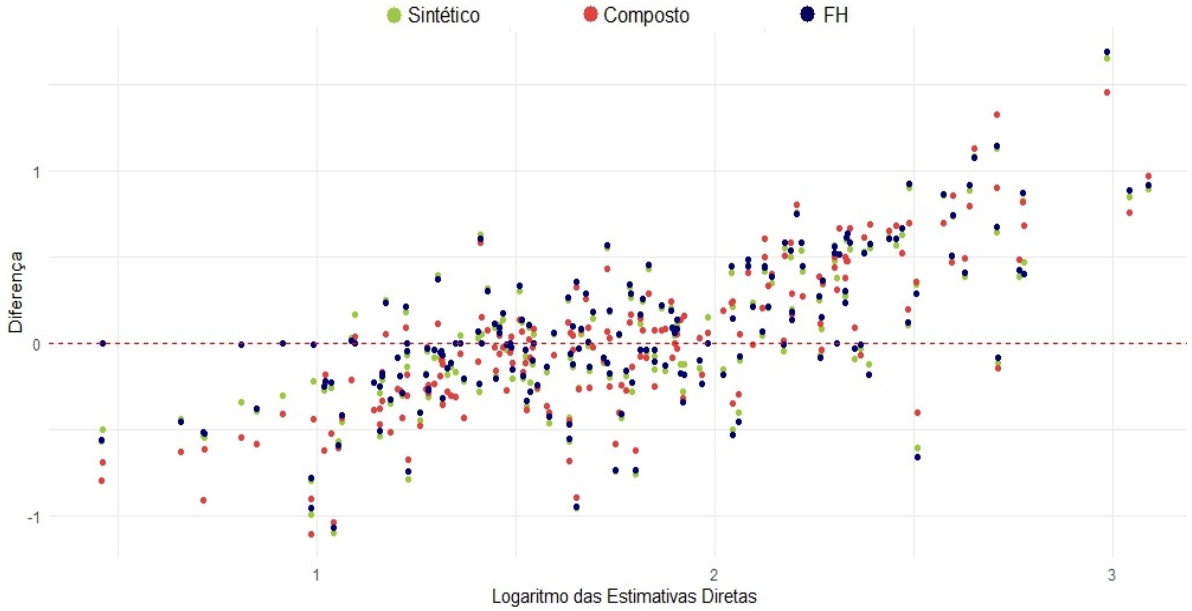
Fonte: Elaborada pela autora (2024)

Figura 23 – Gráficos de dispersão das estimativas em PE dos modelos ajustados no Nordeste



Fonte: Elaborada pela autora (2024)

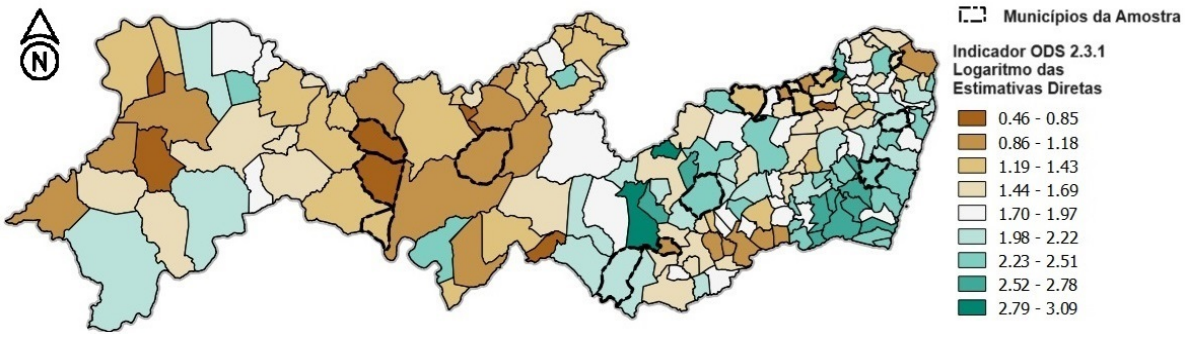
Figura 24 – Diferença das estimativas obtidas em PE entre os métodos de estimação modelados no Nordeste



Fonte: Elaborada pela autora (2024)

De acordo com a análise exploratória do conjunto populacional e a comparação dos resultados obtidos e evidenciados em Pernambuco, é constatado o comportamento dos métodos de estimação. O conjunto populacional, com os dados dispostos espacialmente na figura 25, demonstra que no território pernambucano os valores médios mais altos do LODS 2.3.1 estão concentrados no litoral do estado. A modelagem do experimento dessa dissertação também comprova essa concentração (figura 21). Entretanto, não é predominante a centralização dos resultados das estimativas indiretas na RD Mata Sul.

Figura 25 – Representação do logaritmo nos dados do IODS 2.3.1 em PE do conjunto populacional



Fonte dos dados básicos: IBGE(2022). Elaborada pela autora (2024)

Por fim, em nível de desagregação contextualizado por estimação de pequenas áreas e também exemplificado com o caso do estado de Pernambuco. É possível concluir que o modelo proposto por Fay e Herriot é o mais adequado para estimar os dados da variável dependente, principalmente devido à sua maior precisão e menor variabilidade do estimador. Embora o modelo FH geralmente forneça resultados com maior precisão, os demais procedimentos podem se mostrar adequados em cenários específicos, especialmente diante da disponibilidade de dados e facilidade de implementação. A escolha entre eles deve ser avaliada com base nas condições do estudo, garantindo que o método selecionado seja o mais apropriado para as necessidades e limitações da pesquisa.

7 CONSIDERAÇÕES FINAIS

O estudo apresenta um experimento de simulação utilizando modelos de pequenas áreas em análise do indicador ODS 2.3.1, que mede o volume de produção por unidade de trabalho por dimensão da empresa agrícola/pastoril/florestal, para domínios subnacionais, com foco nos municípios do estado de Pernambuco. No território brasileiro, em um cenário onde os censos agropecuários são realizados a cada dez anos e ainda não possui uma Pesquisa Nacional Agropecuária, onde as pesquisas disponíveis são subjetivas e cadastrais, a aplicação de técnicas de estimação de pequenas áreas é crucial para gerar estimativas precisas e desagregadas. A implementação de uma PNAgro, amparada por métodos probabilísticos de amostragem, permitiria a produção de estatísticas mais detalhadas e frequentes, beneficiando a formulação de políticas públicas e a tomada de decisões baseadas em evidências concretas.

A adoção de metodologias de estimação de pequenas áreas no Brasil pode representar um avanço significativo na produção de dados agropecuários, promovendo o desenvolvimento sustentável e a segurança alimentar. A combinação de diferentes modelos estatísticos, como o proposto por Fay e Herriot e um modelo linear simples, pode otimizar a obtenção de dados e garantir a precisão das estimativas, aproveitando as vantagens de cada abordagem. A viabilidade de tais modelos foi demonstrada nesta dissertação, a modelagem evidenciou a necessidade de uma base de dados explicativa eficaz e a importância de técnicas SAE para melhorar a precisão das estimativas em domínios subnacionais brasileiros.

A disponibilidade de um conjunto de dados auxiliar consolidado, contendo informações de fontes de dados confiáveis, como a FAO e o IBGE, é fundamental. A seleção sistemática e automática dessa base de dados diante de um conjunto de alta dimensionalidade, envolveu a análise de correlação, a verificação do VIF e aplicação do método *stepwise*, em que o intuito é remover características irrelevantes ou redundantes no modelo. A presença de dados auxiliares consistentes são vitais para ajudar à obter precisão nas estimativas. A qualidade de consistência permite que esses dados reflitam adequadamente as realidades locais dos municípios analisados, e ajuda a melhorar a acurácia dos modelos de estimação, reduzindo erros e aumentando a confiabilidade dos resultados.

Na modelagem estatística, entre os métodos SAE verificados, o modelo FH forneceu o melhor ajuste para os dados. O modelo FH identificou menores erros sistemáticos, evidenciados pelas métricas de avaliação, como o EQM e o EAM, que apresentaram valores significati-

vamente menores em comparação a abordagem indireta atráves do estimador sintético por regressão. O ajuste do modelo linear simples por quadrados mínimos ordinários também possui suas vantagens, principalmente pela sua simplicidade de implementação e utilização de recursos computacionais. A combinação de ambos os modelos pode, portanto, ser uma abordagem promissora, aproveitando as vantagens de cada um para produzir estimativas precisas diante de métodos robustos.

Para trabalhos futuros, é recomendável planejar a propriedade *Benchmarking*, assegurando o ajuste diante das estimativas produzidas e garantir a consistência em níveis superiores. Inclusive, também verificar melhorias a partir dos modelos propostos nessa dissertação. Além disso, explorar a presença de dados espaciais ou investigar modelos de regressão espacial pode aprimorar ainda mais a precisão das estimativas.

A elaboração de estatísticas mais desagregadas e granulares direciona intervenções eficientes, reconhecendo subgrupos específicos da população que não estão sendo beneficiados perante a Agenda 2030. A efetividade dos ODS depende de específicas e consolidadas regras metodológicas e também da acessibilidade, viabilidade e disponibilidade de dados. Consequentemente, os objetivos são consolidados e se transformam em uma ferramenta para combater conflitos dentro e entre pessoas, situações e nações, à caminho da prosperidade e do desenvolvimento sustentável. Portanto, a identificação e validação de métodos estatísticos como uma forma de gerar estimativas mais precisas para os indicadores de sustentabilidade da ONU é importante para todas as camadas da sociedade.

REFERÊNCIAS

ABDALLA, L. dos S.; AUGUSTO, D. A.; CHAME, M.; DUFEK, A. S.; OLIVEIRA, L.; KREMPSER, E. Statistically enriched geospatial datasets of brazilian municipalities for data-driven modeling. *Scientific Data 9*, v. 489, 2022. Disponível em: <https://doi.org/10.1038/s41597-022-01581-2>. Acesso em: 03/11/2023.

BOCCI, C.; PETRUCCI, A. *Geoadditive small area model for the estimation of consumption expenditure in Albania. IN: Pratesi, Monica(ed.). Analysis of Poverty Data by Small Area Estimation*. 1. ed. The Atrium, Southern Gate, Terminus Rd, Chichester PO19 8SQ, Reino Unido: Wiley, 2016. (Wiley Series in Survey Methodology). ISBN 9781118815014.

COAGRO, C. de Agropecuária do I. *Proposta de Sistema Nacional de Pesquisas por Amostragem de Estabelecimentos Agropecuários – SNPA*. 2011. Disponível em: <https://www.ibge.gov.br/arquivo/projetos/prpa/SNPA_concepcao_e_conteudo2av.pdf>. Acesso em: 29/08/2024.

COELHO, P. S. *Estimação em Pequenos Domínios*. Lisboa.: ISEGI, 1996. ISSN 0872-895X. Disponível em: <https://run.unl.pt/bitstream/10362/7644/1/WP0048.pdf>. Acesso em: 01/06/2023.

FOOD; FAO, A. O. of the U. N. *Earth Observation - Brazil*. 2022. Disponível em: <https://www.fao.org/giews/earthobservation/country/index.jsp?lang=en&code=BRA>. Acesso em: 13/09/2023.

FOX, J.; WEISBERG, S. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage, 2019. Disponível em: <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.

IBGE. Censo agropecuário. *Rio de Janeiro: IBGE*, 2017. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/21814-2017-censo-agropecuario.html?=&t=downloads>. Acesso em: 15/05/2023.

IBGE. Geociências. download. *Rio de Janeiro: IBGE*, 2022. Disponível em: <https://www.ibge.gov.br/geociencias/downloads-geociencias.html>. Acesso em: 01/09/2023.

IBGE, I. B. de Geografia e E. *Indicadores Brasileiros para os Objetivos de Desenvolvimento Sustentável*. 2024. Disponível em: <https://odsbrasil.gov.br/objetivo/objetivo?n=2>. Acesso em: 13/05/2024.

IPEA, I. de P. E. A. *AtlasBR - Bases de Dados*. 2013. Disponível em: <http://www.atlasbrasil.org.br/acervo/biblioteca>. Acesso em: 20/04/2023.

IPEA, I. de P. E. A. *2. Fome Zero e Agricultura Sustentável*. 2019. Disponível em: <https://www.ipea.gov.br/ods/ods2.html>. Acesso em: 21/04/2024.

KHALIL, C. A.; CANDIA, S. di. *Integrating surveys with geospatial data through small area estimation to disaggregate SDG indicators at subnational level – Case study on SDG Indicators 2.3.1 and 2.3.2*. Roma: FAO, 2023. 46 p. ISBN 978-92-5-137545-7. Disponível em: <https://doi.org/10.4060/cc3944en>. Acesso em: 06/04/2024.

KREUTZMANN, A.-K.; PANNIER, S.; ROJAS-PERILLA, N.; SCHMID, T.; TEMPL, M.; TZAVIDIS, N. The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, v. 91, n. 7, p. 1–33, 2019.

MOLINA, I. Spatial empirical best linear unbiased prediction in small area estimation of poverty. *Series Estudios Estadísticos*, Comisión Económica para América Latina y el Caribe (CEPAL)„ p. 97, 2019. ISSN 1680-8789 (versión electrónica). (LC/TS.2018/82/Rev.1), Santiago. Disponível em: <https://www.cepal.org/es/publicaciones/44214-desagregacion-datos-encuestas-hogares-metodologias-estimacion-areas-pequenas>. Acesso em: 02/11/2023.

MOLINA, I.; RAO, J. *Small Area Estimation*. [S.l.]: John Wiley & Sons, Ltd, 2015. ISBN 9781118735855.

MORETTIN, P. A.; SINGER, J. da M. *Estatística e Ciência de Dados*. 1. ed. Rio de Janeiro: LTC, 2022. ISBN 9788521638162.

PARDOE, I. 10.7 - detecting multicollinearity using variance inflation factors. In: *STAT 462 - Applied Regression Analysis*. 'Eberly College of Science' The Pennsylvania State Universita. 517 Thomas Building, University Park, PA 16802, Estados Unidos.: [s.n.], 2018. Disponível em: <https://online.stat.psu.edu/stat462/node/180/>. Acesso em: 10/09/2024.

PINHEIRO, J. I.; CUNHA, S. B. da; CARVAJAL, S. R.; GOMES, G. C. *Estatística Básica: a arte de trabalhar com dados*. Rio de Janeiro: Elsevier, Ltda, 2009. 312 p. ISBN 9788535230307.

PUSPONEGORO, N. H.; DJURAIDAH, A.; FITRIANTO, A.; SUMERTAJAYA, I. M. Geo-additive models in small area estimation of poverty. *Journal of Data Science and Its Applications*, v. 2, p. 59–67, 04 2019.

PUSPONEGORO, N. H.; RACHMAWATI, R. N. Spatial empirical best linear unbiased prediction in small area estimation of poverty. *Procedia Computer Science*, Elsevier, v. 135, p. 712–718, 2018. ISSN 1877-0509. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050918315060>. Acesso em: 01/10/2023.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022. Disponível em: <https://www.R-project.org/>.

RAHMAN, A. A review of small area estimation problems and methodological developments. In: *A review of small area estimation problems and methodological developments*. [s.n.], 2008. ISBN 978-174-088-3030. ISSN 1443-5101. Disponível em: <https://api.semanticscholar.org/CorpusID:215941259>. Acesso em: 30/10/2024.

REIS, L. P.; COUTO, A. C. S.; FERNANDES, J. M. Modelagem Matemática para a Predição do Limite de Resistência de Aços Produzidos por uma Siderúrgica. In: *XXXV Encontro Nacional de Engenharia de Produção: Perspectivas Globais para a Engenharia de Produção. Anais...* Fortaleza: ENEGEP, 2015. p. 1–16. Disponível em: <https://abepro.org.br/biblioteca/TN_STO_207_230_27043.pdf>. Acesso em: 23/02/2024.

SACHS, J.; SCHMIDT-TRAUB, G.; KROLL, C.; DURAND-DELACRE, D.; TEKSOZ, K. *An SDG Index and Dashboards – Global Report. New York: Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN).* 2016. Disponível em: <https://www.unsdsn.org/resources/sdg-index-and-dashboards-2016/>. Acesso em: 06/07/2023.

SALVATI, N. Small area estimation by spatial models: the spatial empirical best linear unbiased prediction (spatial eblup). In: *Working Paper n 2004/04*. 'G. Peranti' Departamento de Estatística, Universidade de Florença. Viale Morgagni, 59 - Florença.: [s.n.], 2004. Disponível em: <https://api.semanticscholar.org/CorpusID:211037908>. Acesso em: 10/10/2023.

THE ASIAN DEVELOPMENT BANK. *Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices*: A practical guide for national statistics offices. Filipinas, 2022. 112 p. Disponível em: <https://www.adb.org/publications/small-area-estimation-guide-national-statistics-offices>. Acesso em: 18/02/2023.

UNDS, U. N. S. D. *SDG Indicators - Metadata repository: Indicator 2.3.1*. 2024. Disponível em: <https://unstats.un.org/sdgs/metadata/>. Acesso em: 09/01/2024.

WBG, W. B. G. *DataBank - Metadata Glossary*. 2021. Disponível em: <https://databank.worldbank.org/metadataglossary/world-development-indicators/series/NY.GDP.MKTP.PP.KD>. Acesso em: 09/01/2024.

YONG, Y. Small area estimation using fay-herriot area level model with sampling variance smoothing and modeling. *Survey Methodology, Statistics Canada, Catalogue No. 12-001-X*, v. 47, n. 2, p. 361–370, dec 2021. ISSN 1492-0921. Disponível em: <https://www150.statcan.gc.ca/n1/pub/12-001-x/2021002/article/00007-eng.pdf>. Acesso em: 18/05/2023.

APÊNDICE A – CÓDIGO COMPUTACIONAL

```
#################################################################
# --                                                          -- #
#      (UFPE)    Universidade Federal de Pernambuco            #
#      (PPGE) Programa de Pós-Graduação em Estatística         #
# --                                                          -- #
#-------------------------------------------------------------#
#               Estimação de Pequenas Áreas                   #
#                         (SAE)                               #
#-------------------------------------------------------------#
#                                                             #
# PROGRAMA: SAE_CAMILAF.R                                     #
# Versão: Jan/2024                                            #
#   Att.: Jul/2024                                            #
# Autora: Camila Ferreira da Silva                            #
#                                                             #
# Descrição: Com base nos dados fornecidos e a partir das     #
#   estimativas diretas obter estimativas mais precisas do    #
#   indicador ODS 2.3.1.                                      #
#                                                             #
# Informações: Dados do indicador disponilizados pela FAO.    #
#                                                             #
# Etapas:                                                     #
# 1) Análise de Dados                                         #
# 2) Técnicas de Seleção e Extração de Variáveis Auxiliares   #
# 3) Métodos de Estimação em Pequenas Áreas                   #
#                                                             #
#################################################################

#_____________________________________________________________#
#-------------------------------------------------------------#
```

```
# - Configurações -
#Limpar a memória
# Remoção dos objetos especificados entre ()
rm(list=ls())
#Fixar a quantidade de casas decimais
options(digits = 3)
#Reprodutibilidade de dados
set.seed(1687)
#Espeficicação do Diretório
# Verificação do diretório atual de trabalho
getwd()
# Alterar/Especificar o diretório do arquivo
setwd("C:/Users/camis/Desktop/SAE/Tab")
#Especificação das Bibliotecas
# Estimação/SAE
library(sae)
library(emdi)
#Manipulações
library(tidyverse)
library(readxl)
library(openxlsx)
#Análises
library(caret)
library(MASS)
library(stats)
library(car)
library(e1071)
library(nortest)
library(psych)
library(tseries)
library(fitdistrplus)
library(goftest)
library(Metrics)
```

```
library(summarytools)
#Visualização
library(ggplot2)
library(gridExtra)
library(ggfortify)

#-------------------- As Bases de Dados ---------------------#
# >> Verificação e Análise dos Dados e das Informações

# - Conjunto Populacional
# A base de dados consolidada com as estimativas diretas e
#as variáveis auxiliares dos 5560 municípios
estd_bra<-read_excel("C:/Users/camis/Desktop/SAE/Conjunto_Populacional.xlsx")
dim(estd_bra)
#Observar os 6 primeiros elementos da tabelas
head(estd_bra)
#Tipos de variáveis e contagem de observações não nulas
str(estd_bra)
glimpse(estd_bra)
#Verificar se existem valores ausentes
sum(is.na(estd_bra))
sapply(estd_bra, function(x) sum(is.na(x)))
#Medidas Resumo
summary(estd_bra)
estd_bra<-as.data.frame(estd_bra)
describe(estd_bra)
#Analisar a variável resposta: ODS_231
#Estatísticas descritivas
summary(estd_bra$ODS_231)
describe(estd_bra$ODS_231)
var(estd_bra$ODS_231) #variância
#IQR: Interquartile Range
# Distância interquartil (Q3-Q1: 20.2 - 7.6)
```

```
IQR(estd_bra$ODS_231)
#Coeficiente de Assimetria de Fisher-Pearson
cf_assfp <- skewness(estd_bra$ODS_231, type = 1)
print(paste("Coeficiente de Assimetria de Fisher-Pearson:", cf_assfp))
#Coeficiente de Assimetria de Fisher-Pearson Ajustado
n <- length(estd_bra$ODS_231)
aj_cf_assfp <- (sqrt(n * (n - 1)) / (n - 2)) * cf_assfp
print(paste("Coeficiente de Assimetria de Fisher-Pearson Ajustado:",
            aj_cf_assfp))
#Coeficiente de Assimetria de Pearson 2
med <- mean(estd_bra$ODS_231)
med_an <- median(estd_bra$ODS_231)
desv <- sd(estd_bra$ODS_231)
cf_assp2 <- 3 * (med - med_an) / desv
print(paste("Coeficiente de Assimetria de Pearson 2:", cf_assp2))

## - Análise Gráfica
plotdist(estd_bra$ODS_231, histo = TRUE, demp = TRUE)
par(mfrow=c(1,2))  # Dividir a área de plotagem em 1 linhas e 2 colunas
#Histograma
hist(estd_bra$ODS_231, breaks = 40, freq = FALSE,
     main = "Distribuição dos Dados do Indicador 2.3.1",
     xlab = "Valores das Estimativas Diretas", ylab = "Densidade")
# Curva de densidade
dens <- density(estd_bra$ODS_231)
lines(dens, col = "royalblue4", lwd = 2.8)
# Curva normal
curve(dnorm(x, mean = mean(estd_bra$ODS_231), sd = sd(estd_bra$ODS_231)),
      col = "red1", lwd = 3, lty=2, add = TRUE)
# Legenda
legend("topright", legend = c("Curva de Densidade", "Curva Normal"),
       col = c("royalblue4", "red1"), lwd = 2.5, bty = "n")
rug(estd_bra$ODS_231)
```

```
#BOXPLOT
boxplot(estd_bra$ODS_231,
        main = "Boxplot dos Dados do Indicador 2.3.1",
        ylab = "Valores das Estimativas Diretas", col = "lightblue")
abline(h = median(estd_bra$ODS_231),
       col = "green", lwd = 2, lty = 2.5) #linha horizontal na mediana
# Legenda
legend("topright", legend = c("Mediana"),
       col = c("green"), lwd = 2, lty = 2, bty = "n")
#Gráfico de Densidade (Density Plot)
plot(density(estd_bra$ODS_231),
     main = "Representação de Densidade das Estimativas IODS 231",
     xlab = "Valores das Estimativas Diretas", ylab = "Densidade")
#Gráfico de quantis (Q-Q Plot: Normal Q-Q plot)
qqnorm(estd_bra$ODS_231,
       main = "Q-Q Plot das Estimativas IODS 231",
       xlab = "Quantis Teóricos", ylab = "Quantis Observados")
qqline(estd_bra$ODS_231, col = "red", lwd = 2) #linha de referência

# - Conjunto Amostral
# A base de dados da amostra com 556 cidades brasileiras
am_bra<-read_excel("C:/Users/camis/Desktop/SAE/Conjunto_Amostral.xlsx")
dim(am_bra)
head(am_bra)
#Tipos de variáveis
str(am_bra)
#Medidas Resumo
summary(am_bra)
am_bra <- as.data.frame(am_bra)
describe(am_bra)

#Analisar a variável resposta: ODS_231
#Estatísticas descritivas
```

```
summary(am_bra$ODS_231)
describe(am_bra$ODS_231)
var(am_bra$ODS_231)
IQR(am_bra$ODS_231)
#Coeficiente de Assimetria de Fisher-Pearson
am_cf_assfp <- skewness(am_bra$ODS_231, type = 1)
print(paste("Coeficiente de Assimetria de Fisher-Pearson:", am_cf_assfp))

## - Análise Gráfica
plotdist(am_bra$ODS_231, histo = TRUE, demp = TRUE)
par(mfrow=c(1,2))  # Dividir a área de plotagem em 1 linhas e 2 colunas
#Histograma
hist(am_bra$ODS_231, breaks = 35, freq = FALSE, #am_bra$LODS_231
     main = "Distribuição dos Dados do Indicador 2.3.1",
     xlab = "Valores das Estimativas Diretas", ylab = "Densidade")
# Curva de densidade
dens <- density(am_bra$ODS_231)
lines(dens, col = "royalblue4", lwd = 2.8)
# Curva normal
curve(dnorm(x, mean = mean(am_bra$ODS_231), sd = sd(am_bra$ODS_231)),
      col = "red1", lwd = 3, lty=2, add = TRUE)
# Legenda
legend("topright", legend = c("Curva de Densidade", "Curva Normal"),
       col = c("royalblue4", "red1"), lwd = 2.5, bty = "n")
rug(am_bra$ODS_231)
#BOXPLOT
boxplot(am_bra$ODS_231,  #am_bra$LODS_231
        main = "Boxplot dos Dados do Indicador 2.3.1",
        ylab = "Valores das Estimativas Diretas", col = "lightblue")
abline(h = median(am_bra$ODS_231),
       col = "green", lwd = 2, lty = 2.5) #linha horizontal na mediana
legend("topright", legend = c("Mediana"),
       col = c("green"), lwd = 2, lty = 2, bty = "n")
```

```
#Gráfico de Densidade (Density Plot)
plot(density(am_bra$ODS_231),
     main = "Representação de Densidade das Estimativas IODS 231",
     xlab = "Valores das Estimativas Diretas", ylab = "Densidade")
#Gráfico de quantis (Q-Q Plot)
qqnorm(am_bra$ODS_231,
       main = "Q-Q Plot das Estimativas IODS 231",
       xlab = "Quantis Teóricos", ylab = "Quantis Observados")
qqline(am_bra$ODS_231, col = "red", lwd = 2) #linha de referência
qqPlot(am_bra$ODS_231, distribution = "norm")

##Avaliar a adequação da distribuição dos dados
#Teste de Shapiro-Wilk
shapiro.test(am_bra$ODS_231)
#Testes de Bondade de Ajuste (Cramér-von Mises)
cvm.test(am_bra$ODS_231, "pnorm",
         mean = mean(am_bra$ODS_231), sd = sd(am_bra$ODS_231))
# - - -
# / > Abordagens para Modelar a Distribuição dos Dados
# 1. Análise e Tratamento de Outliers
#Visualizar novamente o boxplot
boxplot(am_bra$ODS_231, main = "Boxplot dos Dados IODS 231",
        ylab = "Estimativas Direta")
#Identificar os outliers
out_am <- boxplot.stats(am_bra$ODS_231)$out
#Encontrar os índices dos outliers
id_out <- which(am_bra$ODS_231 %in% out_am)
#Obter os municípios associados aos outliers
name_out <- am_bra$name[id_out]
#Combinar os valores outliers com os municípios correspondentes
res_out <- data.frame(Domínio=name_out, ODS_231=out_am)
res_out <- res_out[order(res_out$ODS_231, decreasing = TRUE), ]
#Ordenação decrescente dos valores extremos
```

```
res_out  # "510835MT" - Município do Mato Grosso

# 2. Transformação de Dados
#Transformação Logarítmica Natural
LODS <- log(am_bra$ODS_231)
#Exibir e verificar os dados transformados
summary(LODS)
describe(LODS)

#Boxplot
boxplot(LODS, main = "Boxplot do Logaritmo Dados IODS 2.3.1",
        ylab = "Logaritmo das Estimativas Direta")
par(mfrow=c(1,2))
#Q-Q Plot
qqnorm(LODS)
qqline(LODS, col = "red", lwd = 2)
qqPlot(LODS, distribution = "norm")
#Criar o histograma com densidade e curva normal
ggplot(am_bra, aes(x =LODS)) +
  geom_histogram(aes(y = after_stat(density)), bins = 20,
                 color = "gray43", fill = "lightblue") +
  geom_density(color = "royalblue4", size = 1) +
  stat_function(fun = dnorm, args = list(mean = mean(LODS),
                    sd = sd(LODS)), color = "red1", size = 1) +
  labs(title = "      Histograma com Curva Normal e de Densidade
       da Transformação Logarítmica do IODS 2.3.1",
#labs(title = "Distribuição da variável dependente em transformação logarítmica")
    x = "Valores da Transformação", y = "Densidade") +  theme_minimal()

#Avaliar a adequação da distribuição dos dados
shapiro.test(LODS)
cvm.test(LODS, "pnorm", mean = mean(LODS), sd = sd(LODS))
```

```
# X. Alternativas Generalizadas
#Ajuste de distribuições
#Útil para modelar valores de dados positivos que são assimétricos à direita
distgam <- fitdist(am_bra$ODS_231, "gamma", optim.method="BFGS")
distgam
summary(distgam)
# Comparação visual dos dados observados e da distribuição ajustada
plot(distgam)
#Histograma
ggplot(am_bra, aes(x = ODS_231)) +
  geom_histogram(aes(y = ..density..), binwidth = 2,
                 fill = "blue", color = "black", alpha = 1) +
  geom_density(color = "green", size = 1) +
  stat_function(fun = function(x) dgamma(x, shape = 2.423, rate = 0.157),
                color = "red", size = 1.2) +
  labs(title = "Histograma com o Ajuste da Distribuição Gama",
       x = "Estimativas Diretas", y = "Densidade")
#Q-Q plot
qqPlot(am_bra$ODS_231, distribution = "gamma", shape = 2.423, rate = 0.157)

#Avaliar a adequação da distribuição
cvm.test(am_bra$ODS_231, "pgamma", shape = distgam$estimate["shape"],
         rate = distgam$estimate["rate"])
gofstat(distgam)

# - Base de Dados Auxiliares
#Conjunto das informações adicionais por município do Brasil
base_aux <- am_bra[,15:100]
head(base_aux)
dim(base_aux)
# 87 variáveis contando com a região
# / - Seleção de dados
base_aux <- as.data.frame(base_aux)
```

```
#Fatores pós corr
#base_exp <- base_aux[, !(names(base_aux) %in% c("POP_2022","V1","bio4_mean",
#                        "Mean_VHI_AS_S1_FAO", "ndvi_adm1_FAO", "bio3_mean"))]
# - -
# - Análise de Correlação (Preliminar)
#Calcular a matriz de correlação
corr_aux <- cor(base_aux, method = 'spearman')
print(corr_aux)
#Filtragem das correlações
# Identificar as posições na matriz onde os
# coeficientes de correlação são fortes (absolutos acima de 0.75)
forte_corr <- which((abs(corr_aux) > 0.75) & (abs(corr_aux) < 1),
                    arr.ind = TRUE)
dim(forte_corr)
data.frame(rn = row.names(corr_aux)[forte_corr[,1]],
           cn=colnames(corr_aux)[forte_corr[,2]])
# _ _
# Correlação entre as variáveis independentes e a resposta
base_sub<-cbind(ODS_231=am_bra$ODS_231, base_aux) #LODS_231
corr <- cor(base_sub, method = "spearman")
dim(base_sub)
names(base_sub)
#Selecionar aquelas com maior correlação absoluta
corr_ods <- abs(corr[,"ODS_231"]) #LODS_231
abs_corr <- names(sort(corr_ods, decreasing = TRUE))
print(abs_corr)
corr_ods
# Escrever as correlações em um arquivo CSV
dirt <- "C:/Users/camis/Desktop/SAE/corrSP.csv" #corrSPlog
write.csv(corr_ods, file = dirt, row.names = TRUE)

## Obs.: method = "kendall" também é testado
#com a transformação log >> O resultado das variáveis é o mesmo
```

```
#_____________________________________________________________#
#---------------- Redução de dimensionalidade ---------------#

# ~ Extração e Seleção de Características

## / > Abordagens para resolver singularidade
# / - Verificar a colinearidade/multicolinearidade
vars_sing <- glm(LODS_231 ~ as.factor(cod_reg)+ESTABEL_ODS+AREA_ha+
        POP_Area_nao_densa+Total_Area_URB_ha+IDHM+ID_GINI+V2+V3+V4+V5+
        V6+V7+V8+V9+V10+V11+V12+V13+V14+V15+V16+V17+V18+V19+V20+V21+V22+
        V23+V24+V25+V26+V27+V28+V29+V30+V31+V32+V33+V34+V35+V36+V37+V38+
        V39+Altitude_mean+bio1_mean+bio2_mean+bio7_mean+CHG_bio12_mean+
        CHG_bio15_mean+Roads_mean+Waterways_mean+Watercourse_mean+
        Dams_mean+Amazon_Forest_mean+Atlantic_Forest_mean+
        Caatinga_mean+Cerrado_mean+Pampa_mean+Pantanal_mean+
        Beach_dune_sand_spot_mean+Citrus_mean+Forest_formation_mean+
        Forest_plantation_mean+Grassland_mean+Mangrove_mean+
        Mining_mean+Other_non_forest_formations_mean+
        Other_non_vegetaded_areas_mean+Other_perennial_crop_mean+
        Salt_flat_mean+Other_temporary_crops_mean+River_lake_ocean_mean+
        Rocky_outcrop_mean+Savanna_formation_mean+Wetlands_mean+
        Wooded_restinga_mean+Mean_VHI_AS_S2_FAO+ASI_AS_S1_FAO+ASI_AS_S2_FAO,
        data = am_bra)

# Utilizar o VIF para quantificar a multicolinearidade
#Calcular o VIF para todas as variáveis independentes
vars_vif <- vif(vars_sing)
print(vars_vif)

#uma transformação do GVIF que uma comparação direta com o VIF
vars_vif_id <- format(vars_vif[,"GVIF^(1/(2*Df))"], scientific = FALSE)
vars_vif_id <- sort(vars_vif_id, decreasing = FALSE)
#Identificar as variáveis acima do limiar:
```

```
# até 1: não são correlacionadas
# 1 ~ 5: moderadamente correlacionadas
# < 10: pode causar sérios prejuízos para o modelo
print(vars_vif_id)

### Alternativa.: Em glm() usar o ODS e diversificar "family"
# Resultado: RETIRAR V10 (1ª mais alta)
# Repetir o processo e retirar AREA_ha (1ª mais alta)

# -- :: Variabilidade entre áreas :: --
# Verificação por macrorregiões do Brasil

# Estatísticas Descritivas
est_cod_reg <- am_bra %>%
  group_by(regiao) %>%
  summarise( tamanho = n(),
    soma = sum(ODS_231, na.rm = TRUE),
    media = mean(ODS_231, na.rm = TRUE),
    desvp = sd(ODS_231, na.rm = TRUE),
    minimo = min(ODS_231, na.rm = TRUE),
    maximo = max(ODS_231, na.rm = TRUE),
    curtose = kurtosis(ODS_231, na.rm = TRUE),
    assimetria = skewness(ODS_231, na.rm = TRUE),
    q25 = quantile(ODS_231, 0.25, na.rm = TRUE),
    mediana = median(ODS_231, na.rm = TRUE),
    q75 = quantile(ODS_231, 0.75, na.rm = TRUE),
    variância = var(ODS_231, na.rm = TRUE),
    IQR = IQR(ODS_231, na.rm = TRUE) )
View(est_cod_reg)
# Relatório descritivo
dfSummary(am_bra$ODS_231, headings = TRUE, method = 'render')
# Gerar relatórios descritivos por região
rb <- am_bra %>%
```

```
  group_by(regiao) %>%
  do(relatorio = dfSummary(.$ODS_231, headings = TRUE, method = 'render'))
# Estatísticas Descritivas
estlog_cod_reg <- am_bra %>%
  group_by(regiao) %>%
  summarise( tamanho = n(),
    soma = sum(LODS_231, na.rm = TRUE),
    media = mean(LODS_231, na.rm = TRUE),
    desvp = sd(LODS_231, na.rm = TRUE),
    minimo = min(LODS_231, na.rm = TRUE),
    maximo = max(LODS_231, na.rm = TRUE),
    curtose = kurtosis(LODS_231, na.rm = TRUE),
    assimetria = skewness(LODS_231, na.rm = TRUE),
    q25 = quantile(LODS_231, 0.25, na.rm = TRUE),
    mediana = median(LODS_231, na.rm = TRUE),
    q75 = quantile(LODS_231, 0.75, na.rm = TRUE),
    variância = var(LODS_231, na.rm = TRUE),
    IQR = IQR(LODS_231, na.rm = TRUE) )
View(estlog_cod_reg)

# Boxplot para a distribuição de IODS_231 por região
ggplot(am_bra, aes(x = regiao, y = ODS_231)) +
  geom_boxplot() + theme_minimal() +
  labs(title = "Boxplot dos Valores do IODS 2.3.1 por Região",
       x = "Regiões", y = "Estimativas Diretas")
# Boxplot para a distribuição de IODS_231 por região
ggplot(am_bra, aes(x = regiao, y = LODS_231)) +
  geom_boxplot() + theme_minimal() +
  labs(title = "Boxplot do Logaritmo dos Valores do IODS 2.3.1 por Região",
       x = "Regiões", y = "Logaritmo das Estimativas Diretas")

# Gráfico com histogramas, curvas de densidade e curva normal
# Calcular a média e o desvio padrão de ODS_231 por região
```

```
est_reg <- am_bra %>%
  group_by(regiao) %>%
  summarise(med_ODS_231 = mean(ODS_231, na.rm = TRUE),
    ds_ODS_231 = sd(ODS_231, na.rm = TRUE))
# Combinar os dados originais com as estatísticas calculadas
am_bra_reg <- am_bra %>%
  left_join(est_reg, by = "regiao")
# Função para adicionar a curva normal
normalc <- function(data) {
  ggplot(data, aes(x = ODS_231, fill = regiao)) +
    geom_histogram(aes(y = after_stat(density)), color = "gray20",
                   bins = 30, alpha = 0.5, position = "identity") +
    geom_density(alpha = 0.6, color ="black",size = 0.8) +
    stat_function(fun = dnorm, args = list(mean = data$med_ODS_231[1],
                sd = data$ds_ODS_231[1]), color = "blue", size = 1)+
    theme_minimal() +
   labs(title = paste("",data$regiao[1]),  #"Histograma do IODS 2.3.1 na Região"
         x = "Valores das Estimativas", y = "Densidade") +
    xlim(0,90) + ylim(0,0.14) +
    theme(legend.position = "none")}  # Remover a legenda
plots <- lapply(split(am_bra_reg, am_bra_reg$regiao), normalc)
do.call(grid.arrange, c(plots, ncol = 2))

# Gráfico com histogramas, curvas de densidade e curva normal
# Calcular a média e o desvio padrão de ODS_231 por região
estlog_reg <- am_bra %>%
  group_by(regiao) %>%
  summarise(
    med_LODS_231 = mean(LODS_231, na.rm = TRUE),
    ds_LODS_231 = sd(LODS_231, na.rm = TRUE))
# Combinar os dados originais com as estatísticas calculadas
am_bra_reg <- am_bra %>%
  left_join(estlog_reg, by = "regiao")
```

```
# Função para adicionar a curva normal
normalc <- function(datalog) {
  ggplot(datalog, aes(x = LODS_231, fill = regiao)) +
    geom_histogram(aes(y = after_stat(density)), color = "gray20",
                   bins = 30, alpha = 0.5, position = "identity") +
    geom_density(alpha = 0.6, color ="black",size = 0.8) +
    stat_function(fun = dnorm, args = list(mean = datalog$med_LODS_231[1],
                sd = datalog$ds_LODS_231[1]), color = "blue", size = 1)+
    theme_minimal() +
   labs(title = paste("",datalog$regiao[1]),  #"Histograma do IODS 2.3.1 na Região"
         x = "Logaritmo dos Valores das Estimativas", y = "Densidade") +
    xlim(0,05) + ylim(0,1.3) +
    theme(legend.position = "none")}  # Remover a legenda
plots <- lapply(split(am_bra_reg, am_bra_reg$regiao), normalc)
do.call(grid.arrange, c(plots, ncol = 2))

#-------------------------------------------------------------#
#                     REGIÃO NORDESTE                         #
#-------------------------------------------------------------#
am_ne <- am_bra[am_bra$regiao == "Nordeste", ]

#Analisar a variável resposta: ODS_231
summary(am_ne$ODS_231)
describe(am_ne$ODS_231)
var(am_ne$ODS_231)
IQR(am_ne$ODS_231)

# Transformação Logarítmica : LODS_231
summary(am_ne$LODS_231)
describe(am_ne$LODS_231)
var(am_ne$LODS_231)
IQR(am_ne$LODS_231)
```

```
## - Análise Gráfica
plotdist(am_ne$ODS_231, histo = TRUE, demp = TRUE)
plotdist(am_ne$LODS_231, histo = TRUE, demp = TRUE)
par(mfrow=c(2,2))  # Dividir a área de plotagem em 2 linhas e 2 colunas
#Histograma
hist(am_ne$ODS_231, breaks = 20, freq = FALSE,
     main = "Distribuição dos Dados do Indicador 2.3.1",
     xlab = "Estimativas Diretas", ylab = "Densidade")
# Curva de densidade
dens <- density(am_ne$ODS_231)
lines(dens, col = "royalblue4", lwd = 2.8)
# Curva normal
curve(dnorm(x, mean = mean(am_ne$ODS_231), sd = sd(am_ne$ODS_231)),
      col = "red1", lwd = 2.8, lty=2, add = TRUE)
# Legenda
legend("topright", legend = c("Curva de Densidade", "Curva Normal"),
       col = c("royalblue4", "red1"), lwd = 2.5, bty = "n")
rug(am_ne$ODS_231)
#BOXPLOT
boxplot(am_ne$ODS_231, horizontal = TRUE,
        main = "Boxplot dos Dados do Indicador 2.3.1",
        xlab = "Estimativas Diretas", col = "lightblue")
abline(v = median(am_ne$ODS_231),
       col = "green", lwd = 2, lty = 2.5) #linha horizontal na mediana
# Legenda
legend("topright", legend = c("Mediana"),
       col = c("green"), lwd = 2, lty = 2, bty = "n")
#Q-Q Plot
qqPlot(am_ne$ODS_231, distribution = "norm")
qqnorm(am_ne$ODS_231,
       main = "Q-Q Plot das Estimativas IODS 231",
       xlab = "Quantis Teóricos", ylab = "Quantis Observados")
qqline(am_ne$ODS_231, col = "red", lwd = 2) #linha de referência
```

```
#Histograma
hist(am_ne$LODS_231, breaks = 20, freq = FALSE, main="",
     xlab = "Logaritmo das Estimativas Diretas", ylab = "Densidade")
# Curva de densidade
dens <- density(am_ne$LODS_231)
lines(dens, col = "royalblue4", lwd = 2.8)
# Curva normal
curve(dnorm(x, mean = mean(am_ne$LODS_231), sd = sd(am_ne$LODS_231)),
      col = "red1", lwd = 2.8, lty=2, add = TRUE)
rug(am_ne$LODS_231)
#BOXPLOT
boxplot(am_ne$LODS_231,
        xlab = "Logaritmo Estimativas das Diretas", col = "lightblue",
        horizontal = TRUE)
abline(v = median(am_ne$LODS_231),
       col = "green", lwd = 2, lty = 2.5) #linha horizontal na mediana
# Legenda
legend("bottomright", legend = c("Mediana"),
       col = c("green"), lwd = 2, lty = 2, bty = "n")
#Q-Q Plot
qqPlot(am_ne$LODS_231, distribution = "norm")
qqnorm(am_ne$LODS_231,
       main = "Q-Q Plot do Logaritmo dos Dados IODS 231",
       xlab = "Quantis Teóricos", ylab = "Quantis Observados")
qqline(am_ne$LODS_231, col = "red", lwd = 2) #linha de referência

#Avaliar a adequação da distribuição dos dados
shapiro.test(am_ne$LODS_231)
cvm.test(am_ne$LODS_231, "pnorm",
mean = mean(am_ne$LODS_231), sd = sd(am_ne$LODS_231))

#Identificar os outliers
out_am_ne <- boxplot.stats(am_ne$LODS_231)$out
```

```
#Encontrar os índices dos outliers
id_out_ne <- which(am_ne$LODS_231 %in% out_am_ne)
#Obter os municípios associados aos outliers
name_out_ne <- am_ne$name[id_out_ne]
#Combinar os valores outliers com os municípios correspondentes
res_out_ne <- data.frame(Domínio=name_out_ne, LODS_231=out_am_ne)
res_out_ne <- res_out_ne[order(res_out_ne$LODS_231, decreasing = TRUE), ]
res_out_ne

#-------------------------------------------------------------#
#_____________________Estimador Direto_____________________#

#DEMONSTRAÇÃO.: Uma forma de obtenção de estimativas diretas
# Atráves do estimador de Horvitz-Thompson(H-T)
# Perante as Regiões
HTREG<- direct( y = "LODS_231", am_bra, smp_domains = "regiao",
                weights = "ESTABEL_ODS")
summary(HTREG)
predict(HTREG)
# Perante os Estados
HTET <- direct(y = "LODS_231", am_bra, smp_domains = "UF",
               weights = "ESTABEL_ODS")
summary(HTET)
predict(HTET)

#_____________________________________________________________#
#---------------- AJUSTE DE MODELO ESTATÍSTICO --------------#

# Sob o modelo em nível de área
#diante de uma transformação logarítmica
#::::::::::::::::::::::::::::::::::::::#
#:: Estimador Sintético por Regressão ::#
#::::::::::::::::::::::::::::::::::::::#
```

```
##Implementação de Quadrados Mínimos Ordinários
nereg_lin_bt <- lm(LODS_231 ~ V8+V19+V21+V33+V20+V16+
                     ID_GINI+Mean_VHI_AS_S2_FAO+V6+V34+
                     V29+POP_Area_nao_densa+V5,
                     data = am_ne)
#Exibir o modelo final (as métricas)
summary(nereg_lin_bt)
head(nereg_lin_bt)
#Variância
summary(nereg_lin_bt)$sigma^2
# Predições
nepred_lin_bt <- predict(nereg_lin_bt)
nereg_lin_bt$fitted.values
# Resíduos
neres_reg <- residuals(nereg_lin_bt)
# Resíduos padronizados
pa_resid_ne <- rstudent(nereg_lin_bt)
#Extração das métricas de desempenho
print(nereg_lin_bt)
mse(am_ne$LODS_231, nepred_lin_bt)
rmse(am_ne$LODS_231, nepred_lin_bt)
mae(am_ne$LODS_231, nepred_lin_bt)
mape(am_ne$LODS_231, nepred_lin_bt)
rmsle(am_ne$LODS_231, nepred_lin_bt)
rae(am_ne$LODS_231, nepred_lin_bt)
bias(am_ne$LODS_231, nepred_lin_bt)
# Realizar a ANOVA
var_negreg <- anova(nereg_lin_bt)
print(var_negreg)
#Variância do erro do modelo
sig12 <- anova(nereg_lin_bt)[2,3]
# Distribuição
family(nereg_lin_bt)
```

```
logLik(nereg_lin_bt) # Ligação
# Coeficientes
coef(nereg_lin_bt)
# Intervalo de Conf. dos coeficientes
confint(nereg_lin_bt)
# Variância-Covariância
vcov(nereg_lin_bt)

#A partir do modelo “melhor” ajustado obtido
# as proporções previstas
nesyn_reg_est <- predict(nereg_lin_bt, am_ne, se.fit=TRUE, type="response")
nesyn_reg_data<-cbind(am_ne$LODS_231, nesyn_reg_est$fit, nesyn_reg_est$se.fit)
nesyn_reg_data
## Intervalos de confiança para previsões
#das respostas médias
intc_pred <- predict(nereg_lin_bt, am_ne, interval="confidence")
#das respotas individuais
intc_pred <- predict(nereg_lin_bt, am_ne, interval="prediction")
print(intc_pred)

## Testes > no modelo
durbinWatsonTest(nereg_lin_bt)
ncvTest(nereg_lin_bt)
vif(nereg_lin_bt)
# Testes > nos resíduos
shapiro.test(neres_reg)
shapiro.test(pa_resid_ne)
durbinWatsonTest(neres_reg)
durbinWatsonTest(pa_resid_ne)

# Análise Gráfica
plot(nereg_lin_bt)
par(mfrow = c(1, 2))
```

```
autoplot(nereg_lin_bt, which = 1:6, ncol = 3, label.size = 3)

# Gráfico de componentes residuais
residualPlots(nereg_lin_bt, terms = ~ .)

par(mfrow = c(1, 3))
# Diagrama de resíduos
res_nelm <- fortify(nereg_lin_bt)
ggplot(res_nelm, aes(.fitted, .resid)) +
  geom_point() +
  geom_smooth(method = "loess", col = "red") +
  labs(title = "Resíduos vs Valores Ajustados",
       x = "Valores Ajustados", y = "Resíduos") + theme_minimal()
# Diagrama de resíduos padronizados
ggplot(res_nelm, aes(.fitted, .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  geom_smooth(method = "loess", col = "blue") +
  labs(title = "Resíduos Padronizados vs Valores Ajustados",
  x="Valores Ajustados", y="Resíduos Padronizados") + theme_minimal()

# Gráfico de alavancagem vs.resíduos padronizados
ggplot(res_nelm, aes(.hat, .stdresid)) +
  geom_point(aes(size = .cooksd), alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Alavancagem vs. Resíduos Padronizados",
    x="Leverage", y="Resíduos Padronizados")+theme_minimal()

# Q-Q Plot dos resíduos padronizados
ggplot(res_nelm, aes(sample = .stdresid)) +
  stat_qq() + stat_qq_line() +
  labs(title = "Q-Q Plot dos Resíduos Padronizados",
    x="Quantis Teóricos", y="Quantis Amostrais") + theme_minimal()
```

```
#Gráfico de Dispersão
plot(nepred_lin_bt, am_ne$LODS_231,
     xlab = "Estimativas Indiretas",
     ylab = "Logaritmo das Estimativas Diretas",
     col = "black",  # Cor para os valores ajustados
     pch = 16)  # Tipo de ponto
points(nepred_lin_bt, am_ne$LODS_231,
       col = c("green", "blue"),  # Cor para as estimativas diretas
       pch = 16, cex=0.5)  # Tipo de ponto
#a linha de referência
abline(a = 0, b = 1, col = "red", lwd = 2,lty = 2)
#Adiciona uma legenda
legend("topleft", legend = c("Estimativas Indiretas",
                             "Logaritmo das Estimativas Diretas"),
       col = c("green", "blue"), pch = c(16, 16))

#Histograma
m_resid <- mean(res_nelm$.resid)
s_resid <- sd(res_nelm$.resid)
ggplot(res_nelm, aes(x = .resid)) +
  geom_histogram(aes(y = ..density..), bins = 20, fill = "lightblue2",
  color = "gray55") + geom_density(color = "blue", linewidth = 1) +
  stat_function(fun = dnorm, args = list(mean = m_resid, sd = s_resid),
                color = "red", linewidth = 1) +
  labs(title = "Histograma e Curva de Densidade dos Resíduos",
    x = "Resíduos", y = "Densidade") + theme_minimal()

# Histograma de densidade com curva de densidade dos res padr.
# Calcular média e desvio padrão dos resíduos padronizados
mp_resid <- mean(res_nelm$.stdresid)
sp_resid <- sd(res_nelm$.stdresid)
ggplot(res_nelm, aes(x = .stdresid)) +
  geom_histogram(aes(y = ..density..), bins = 20, fill = "lightblue2",
```

```
  color = "gray55") + geom_density(color = "blue", linewidth = 1) +
  stat_function(fun = dnorm, args = list(mean = mp_resid, sd = sp_resid),
                color = "red", linewidth = 1) +
  labs(title = "Histograma e Curva de Densidade dos Resíduos Padronizados",
       x="Resíduos Padronizados", y="Densidade") + theme_minimal()

# Boxplot dos resíduos padronizados
ggplot(res_nelm, aes(y = .stdresid)) +
  geom_boxplot(fill = "white", color = "black") +
  labs(title = "Boxplot dos Resíduos Padronizados",
       y = "Resíduos Padronizados") + theme_minimal()

par(mfrow = c(1,1))
#Identificar pontos influentes
infIndexPlot(nereg_lin_bt, main="Pontos de Influência")
outlierTest(nereg_lin_bt)

#QQ-plot resíduos
par(mfrow = c(2,2))
plot(density(pa_resid_ne))
plot(density(neres_reg))
qqPlot(pa_resid_ne, main = "QQ-Plot dos Resíduos Padronizados")
qqPlot(neres_reg, main = "QQ-Plot dos Resíduos",
       xlab = "Quantil", ylab ="Resíduos")

#_____________________________________________________________#
#---------------- Qualidade das Estimativas -----------------#

#Extrair as Estimativas e o Erro Quadrático Médio (EQM ou MSE)
esti_pred <- nesyn_reg_est$fit
#As diferenças entre os valores observados e os valores preditos
# Calcular os resíduos
res_ne <- resid(nereg_lin_bt)
```

```
# Calcular os resíduos ao quadrado para cada observação
res_ne2 <- res_ne^2
esti_eqm <- res_ne2

## Possíveis conclusões acerca dos resultados
#Medidas de qualidade para comparar as estimativas diretas

# - A Raiz do Erro Quadrático Médio
reqm_mselin <- sqrt(esti_eqm)

# - É calculado o Erro Quadrático Médio Relativo (EQMR)
##o Coeficiente de Variação (CV)
eqmr_mselin <- 100*sqrt(esti_eqm)/esti_pred
eqmr_mselin

# - O Intervalo de Confiança
inf_mselin <- esti_pred - (1.96*sqrt(esti_eqm))
sup_mselin <- esti_pred + (1.96*sqrt(esti_eqm))

# - Construção de uma base de dados com todos os resultados
estim_pred_lin <- data.frame(Dominio = am_ne$name,
                             UF = am_ne$UF,
                             LODS_231 = am_ne$LODS_231,
                             Preditos = esti_pred,
                             EQM = esti_eqm,
                             REQM = reqm_mselin,
                             CV = eqmr_mselin,
                             EQMR = eqmr_mselin,
                             Limite_Inferior = inf_mselin,
                             Limite_Superior = sup_mselin)
dim(estim_pred_lin)
head(estim_pred_lin)
write.xlsx(estim_pred_lin, "ESTIMA_PRED_LM_NE.xlsx", rowNames = FALSE)
```

```
#_____________________________________________________#
#------------------ Visualização dos Dados ---------------#
#Os resultados armazenados em uma tabela
rest_sae <- read_excel("C:/Users/camis/Desktop/SAE/Tab/ESTIMA_PRED_LM_NE.xlsx")
summary(rest_sae)
attach(rest_sae)
# - Distribuição do conjunto de dados
#das estimativas obtidas no modelo de regressão linear simples
boxplot(list(Direto = LODS_231, Sintético = Preditos),
        xlab = "Métodos de Estimação", ylab = "Estimativas",
        main = "Comparação entre o Ajuste do modelo linear e o Direto",
        col = c("lightblue", "yellow2"), las=1, cex=0.75, boxwex=0.5)
plot(rest_sae$LODS_231, rest_sae$Preditos,
    xlab = "Logaritmo das Estimativas Diretas", ylab = "Estimativas Indiretas",
     col = c("lightblue", "yellow2"),pch = c(16, 16), lwd = 2,
     main = "Relação entre Estimativas Diretas e Indiretas da Amostra",
     cex=0.9)
legend("topleft", legend = c("Estimativas Diretas", "Estimativas Indiretas"),
       col = c("lightblue", "yellow2"), pch = c(16, 16))
# Adicionar a linha de referência
abline((lm(LODS_231 ~ Preditos)), col = "blue", lwd = 2)
detach(rest_sae)

# :::::::::::::::: # CASO AMOSTRAL # :::::::::::::::: #
# --- Predições em áreas não pertencentes a amostra --- #

# Parte 1:
# - É necessário extrair os betas(coeficientes) do modelo
# Visualizar coeficientes
ambetas <- as.matrix(coefficients(nereg_lin_bt), ncol=1)
ambetas
class(ambetas)
str(ambetas)
```

```
# Parte 2:
# - Multiplicar os betas com as variáveis explicativas
estd_ne<- estd_bra[estd_bra$regiao == "Nordeste", ]
aux_pop <- estd_ne %>%
  select(V8,V19,V21,V33,V20,V16,
    ID_GINI,Mean_VHI_AS_S2_FAO,V6,V34,
    V29,POP_Area_nao_densa,V5)
str(aux_pop)
head(aux_pop)
amx <- as.matrix(cbind(1, aux_pop)) #base auxiliar
class(amx)
str(amx)

## O vetor preditivo
# As previsões de ambos os domínios
#os observados na amostra e os não observados
# (converter para matriz se necessário)
am_pred <-  amx%*%ambetas
dim(am_pred)
am_pred

#Base de dados com a nomenclatura dos domínios
expl_pop <- estd_ne %>%
  select(name,
         V8,V19,V21,V33,V20,V16,
         ID_GINI,Mean_VHI_AS_S2_FAO,V6,V34,
         V29,POP_Area_nao_densa,V5)
am_pred <- data.frame(name = expl_pop$name, Preditos = am_pred)
dim(am_pred)

#Uma base de dados apenas com os domínios não observados
#Domínios que não foram observado
n_amostra <- setdiff(am_pred$name, unique(am_ne$name))
```

```
#n_amostra <- as.data.frame(n_amostra)
#dim(n_amostra) #1662 #NAME
#Predições em domínios não observados
pred_Nobs <- subset(am_pred, am_pred$name %in% n_amostra)
dim(pred_Nobs)
pred_Nobs

# - Parte 2.1: O cálculo do EQM
#Estimar a variância dos coeficientes Betas
## a variância apenas depende das observações amostrais
#Um subconjunto do conjunto de dados original auxiliar
info_aux_obs <- subset(expl_pop, am_pred$name %in%  #auxiliar
                         unique(am_ne$name)) %>% mutate(inter = 1)
str(info_aux_obs) #171x15
head(info_aux_obs)
class(info_aux_obs)

#Uma matriz com duas colunas das amostras dos dados originais
X_obs <- cbind(1,info_aux_obs[,2:15])
X_obs <- as.matrix(X_obs)
class(X_obs)

#O cálculo da matriz de covariância dos coeficientes estimados
lambda <- 1e-5
Vbeta_est <- solve(t(X_obs) %*% X_obs + lambda * diag(ncol(X_obs)))
Vbeta_esta <- as.matrix(Vbeta_est)
dim(Vbeta_esta) # Deve retornar (k, k) e k é o número de preditores
Vbeta_esta

# Obter a matriz de covariância dos coeficientes
Vbeta_cov <- vcov(nereg_lin_bt)
Vbeta_cov
dim(Vbeta_cov)
```

```
#Organizar as infos auxiliares nos domínios não observados
## Calcular o EQM
info_aux_Nobs <- subset(expl_pop, am_pred$name %in% n_amostra) %>%
  mutate(inter = 1)
dim(info_aux_Nobs)
str(info_aux_Nobs)
X_Nobs <- cbind(1, info_aux_Nobs[,2:15])
X_Nobs_m <- as.matrix(X_Nobs)
dim(X_Nobs_m) # Deve retornar (n, k)
# n é o número de observações
str(X_obs)

# Estimar o EQM de forma matricial
EQM_Nobs <- diag((X_Nobs_m %*% Vbeta_esta %*% t(X_Nobs_m)))
EQM_Nobs
#Proceder com o cálculo do Erro Quadrático Médio Relativo
EQMR_Nobs <- 100*sqrt(EQM_Nobs)/(pred_Nobs$Preditos)
EQMR_Nobs

# Parte 2.2 : Organização das Bases de Dados
pred_Nobs$EQM <- EQM_Nobs
pred_Nobs$EQMR <- EQMR_Nobs
head(pred_Nobs)
dim(pred_Nobs)
pred_Nobs <- as.data.frame(pred_Nobs)
# Domínios da Amostra
head(estim_pred_lin)
estim2_pred <- data.frame(name = am_ne$name,
                          Preditos = esti_pred,
                          EQM = esti_eqm,
                          EQMR = eqmr_mselin)
# Agregar os dados
AG_PRED <- rbind(estim2_pred, pred_Nobs)
```

```
#A base de dados completa
#os dados das estimativas diretas e indiretas do modelo ajustado
BC_PRED <- full_join(AG_PRED, estd_ne)
head(BC_PRED)
write.xlsx(BC_PRED, "ESTIMA_PRED_SYN_NE.xlsx", rowNames = FALSE)

# - Parte 3:
#::::::::::::::::::::::::::::::::::::::#
#::  Compondo o Estimador Sintético     #
#::::::::::::::::::::::::::::::::::::::#

# Soma dos estabelecimentos para cada município
# em que calcula-se o indicador 2.3.1
Nd_dir <- tapply(am_ne$ESTABEL_ODS, list(am_ne$name), sum)
delta <- 5
# Se delta for 1 dará mais peso ao estimador direto
#então quando o estimador sintético é aplicado
#nas áreas não observadas, então sempre será superestimado
Nd <- subset(am_ne$ESTABEL_ODS, am_ne$name %in% unique(estd_ne$name))
Nd.delta <- delta * Nd
phi1 <- ifelse(Nd_dir > Nd.delta, 1, Nd_dir / Nd.delta)
phi1 <- data.frame(name = am_ne$name, phi)
phi1
dim(phi1)
rownames(phi1)<- NULL
# A união da base
Estima <- full_join(BC_PRED, phi1)
dim(Estima)
# Preencher os valores ausentes na coluna phi com 0.2
Estima <- Estima %>% mutate(phi = ifelse(is.na(phi), 0.2, phi))
#Dados do Indicador
LODS <- as.numeric(log(estd_ne$ODS_231))
class(LODS)
```

```
RES_PREDC <- as.numeric(BC_PRED$Preditos)
class(RES_PREDC)
## :: Estimador Composto
Estima<- Estima %>% mutate(Preditos.comp = phi * LODS +
                              (1 - phi) * RES_PREDC , phi= NULL)
dim(Estima)
head(Estima)
write.xlsx(Estima, "ESTIMA_COMP_NE.xlsx", rowNames = FALSE)

# - Parte 4:
#Verificação dos Resultados
BC_PREDCOMP <-read_excel("C:/Users/camis/Desktop/SAE/tab/ESTIMA_COMP_PRED_NE.xlsx")
BC_PREDCOMP <- as.data.frame(BC_PREDCOMP)
dim(BC_PREDCOMP)
# Análise Gráfica
par(mfrow = c(1, 2))
plot(BC_PREDCOMP$LODS_231,
     BC_PREDCOMP$Preditos.comp,
     xlab = "Valores do Logaritmo das Estimativas Diretas",
     ylab = "Valores Ajustados",
     col = c("lightblue", "yellow2"),pch = c(16, 16), lwd = 2,
     main = "Relação entre Estimativas Diretas e Indiretas", cex=1)
legend("topleft", legend = c("Estimativas Diretas", "Estimativas Indiretas"),
       col = c("lightblue", "yellow2"), pch = c(16, 16))
# Adicionar uma linha de regressão
abline((lm(LODS_231 ~ Preditos.comp)), col = "blue", lwd = 2)

#Boxplot
boxplot(list(Direto = BC_PREDCOMP$LODS_231, Sintético = BC_PREDCOMP$Preditos,
             Composto = BC_PREDCOMP$Preditos.comp),
xlab = "Métodos de Estimação", ylab = "Logratimo das Estimativas",
main = "Boxplot entre o Ajuste do modelo linear e o Direto",
col = c("lightblue", "yellow2", "orange2"),las=1, cex=0.75, boxwex=0.5)
```

```
#:::::::::::::::::::::::::#
#::  Modelo Fay-Herriot ::#
#:::::::::::::::::::::::::#

# -- Variância -- #
#ELABORAÇÃO DA VARIÂNCIA FICTÍCIA
# Definir os dados de cada município
dtm <- subset(estd_bra, select = c(name, LODS_231, ESTABEL_ODS))
# Calcular a variância ponderada da média para cada domínio
LVAR_POND <- with(dtm,
(1 / ESTABEL_ODS) * LODS_231^2 - (1 / ESTABEL_ODS^2) * LODS_231^2)
# Ajustar para zero se houver apenas um valor para cada município
LVAR_POND <- ifelse(dtm$ESTABEL_ODS == 1, 0, LVAR_POND)
print(LVAR_POND)
# Criar um dataframe para armazenar os resultados
res_varp <- data.frame( name = dtm$name, LVAR_POND = LVAR_POND)
head(res_varp)
write.csv(res_varp, file = "res_varpond.csv", row.names = FALSE)

# Ajuste do modelo FH
am_ne <- as.data.frame(am_ne)
estihf_ne <- fh(LODS_231 ~ V8+V19+V21+V33+V20+V16+
                ID_GINI+Mean_VHI_AS_S2_FAO+V6+V34+
                V29+POP_Area_nao_densa+V5,
                vardir = "LVAR_POND",
                combined_data = am_ne,
                domains ="name", method = "reml",
                eff_smpsize = "ESTABEL_ODS",
                MSE = TRUE, mse_type = "analytical",
                B = c(100, 0) )
#Exibir o modelo final (as métricas)
print(estihf_ne)
summary(estihf_ne)
```

```
head(estihf_ne)
head(estimators(estihf_ne, MSE=TRUE, CV=TRUE))
# Predições
nepred_hf <- predict(estihf_ne)
print(nepred_hf)
estihf_ne$ind$FH
# Resíduos
neres_hf <- residuals(estihf_ne)
neres_hf <- as.vector(estihf_ne$model$real_residuals)
# Resíduos Padrozinados
pahf_res <- as.vector(estihf_ne$model$std_real_residuals)
#Extração das métricas de desempenho
mse(am_ne$LODS_231, nepred_hf$FH)
#mse(nepred_hf$Direct, nepred_hf$FH)
rmse(am_ne$LODS_231, nepred_hf$FH)
mae(am_ne$LODS_231, nepred_hf$FH)
mape(am_ne$LODS_231, nepred_hf$FH)
rmsle(am_ne$LODS_231, nepred_hf$FH)
rae(am_ne$LODS_231, nepred_hf$FH)
bias(am_ne$LODS_231, nepred_hf$FH)
print(var_negreg)
# Coeficientes
coef(estihf_ne)
fixed.effects(estihf_ne)
family(estihf_ne) # Distribuição
confint(estihf_ne) # Intervalo de Conf. dos coeficientes
vcov(estihf_ne) # Variância-Covariância

#A partir do modelo “melhor” ajustado obtido
# as proporções previstas
nesyn_reg_datahf <- cbind(am_ne$LODS_231, estihf_ne$ind$FH, estihf_ne$MSE)
nesyn_reg_datahf
## Intervalos de confiança para média de previsões
```

```
intc_predhf <- predict(estihf_ne, am_ne, interval="confidence")
print(intc_pred)

## Testes > nos resíduos
shapiro.test(neres_hf)
durbinWatsonTest(neres_hf)
## Testes > res. padronizados
shapiro.test(pahf_res)
durbinWatsonTest(pahf_res)

# Análise Gráfica
plot(neres_hf)
compare_plot(estihf_ne, MSE = TRUE, CV = TRUE)
plot(estihf_ne)

#Resíduos vs Valores Ajustados
ggplot(res_fh, aes(x = estihf_ne$ind$FH, y = neres_hf)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Resíduos vs Valores Ajustados",
       x="Valores Ajustados", y="Resíduos") + theme_minimal()
#Resíduos Padronizados vs Valores Ajustados
ggplot(res_fh, aes(x = estihf_ne$ind$FH, y = pahf_res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
  labs(title = "Resíduos Padronizados vs Valores Ajustados",
       x="Valores Ajustados", y="Resíduos Padronizados") + theme_minimal()

# QQ plot Resíduos Padronizados
ggplot(data.frame(sample = pahf_res), aes(sample = sample)) +
  stat_qq() + stat_qq_line() +
  labs(title = "QQ Plot dos Resíduos Padronizados",
    x="Quantis Teóricos", y="Quantis Amostrais") + theme_minimal()
```

```
#Gráfico de Dispersão
plot(estihf_ne$ind$FH, am_ne$LODS_231,
     xlab = "Estimativas Indiretas",
     ylab = "Logaritmo das Estimativas Diretas",
     col = "black",  # Cor para os valores ajustados
     pch = 16)  # Tipo de ponto
points(estihf_ne$ind$FH, am_ne$LODS_231,
       col = c("green", "blue"),  # Cor para as estimativas diretas
       pch = 16, cex=0.5)  # Tipo de ponto
#a linha de referência
abline(a = 0, b = 1, col = "red", lwd = 2,lty = 2)
#Adiciona uma legenda
legend("topleft", legend = c("Estimativas Indiretas",
                             "Logaritmo das Estimativas Diretas"),
       col = c("green", "blue"), pch = c(16, 16))

#Histograma e Curva de Densidade dos Resíduos
ggplot(data.frame(Residuos = neres_hf), aes(x = Residuos)) +
  geom_histogram(aes(y = ..density..), bins = 30,
      fill = "lightblue2", color = "gray55") +
  geom_density(color = "blue", size = 1) +
 stat_function(fun = dnorm, args = list(mean = mean(neres_hf), sd = sd(neres_hf)),
                color = "red", linewidth = 1) +
  labs(title = "Histograma e Curva de Densidade dos Resíduos",
    x = "Resíduos", y = "Densidade") + theme_minimal()

#Histograma res. padronizados
ggplot(data.frame(Residuos = pahf_res), aes(x = Residuos)) +
  geom_histogram(aes(y = ..density..), bins = 30,
      fill = "lightblue2", color = "gray55") +
  geom_density(color = "blue", size = 1) +
 stat_function(fun = dnorm, args = list(mean = mean(pahf_res), sd = sd(pahf_res)),
                color = "red", linewidth = 1) +
```

```
  labs(title = "Histograma e Curva de Densidade dos Resíduos Padronizados",
       x = "Resíduos Padronizados",
       y = "Densidade") + theme_minimal()

#QQ-plots dos resíduos
par(mfrow = c(1,1))
plot(density(pahf_res))
plot(density(neres_hf))
qqPlot(pahf_res, main = "QQ-Plot dos Resíduos Padronizados")
qqPlot(neres_hf, main = "QQ-Plot dos Resíduos",
       xlab = "Quantil",  ylab ="Resíduos")

#_____________________________________________________________#
#---------------- Qualidade das Estimativas -----------------#

#Extrair as Estimativas e o Erro Quadrático Médio (EQM ou MSE)
esti_predhf <- estihf_ne$ind$FH
esti_eqmhf <- estihf_ne$MSE$FH

## Possíveis conclusões acerca dos resultados
#Medidas de qualidade para comparar as estimativas diretas

# - A Raiz do Erro Quadrático Médio
reqm_msehf <- sqrt(esti_eqmhf)

# - É calculado o Erro Quadrático Médio Relativo (EQMR)
##o Coeficiente de Variação (CV)
eqmr_msehf <- 100*sqrt(esti_eqmhf)/esti_predhf

# - O Intervalo de Confiança
inf_msehf <- esti_predhf - (1.96*sqrt(esti_eqmhf))
sup_msehf <- esti_predhf + (1.96*sqrt(esti_eqmhf))
```

```
# - Construção de uma base de dados com todos os resultados
estim_predhf <- data.frame(Dominio = am_ne$name,
                           LODS_231 = am_ne$LODS_231,
                           Preditos = esti_predhf,
                           EQM = esti_eqmhf,
                           REQM = reqm_msehf,
                           CV = eqmr_msehf,
                           EQMR = eqmr_msehf,
                           Limite_Inferior = inf_msehf,
                           Limite_Superior = sup_msehf)
dim(estim_predhf)
head(estim_predhf)
write.xlsx(estim_predhf, "ESTIMA_PRED_FH_NE.xlsx", rowNames = FALSE)

#_____________________________________________________________#
#------------------ Visualização dos Dados ----------------#
#Os resultados armazenados em uma tabela
rest_sae <- read_excel("C:/Users/camis/Desktop/SAE/Tab/ESTIMA_PRED_FH_NE.xlsx")
summary(rest_sae)
attach(rest_sae)

# - Distribuição do conjunto de dados
# das estimativas obtidas por FH
boxplot(list(Direta = rest_sae$LODS_231, Indireta = rest_sae$Preditos),
        xlab = "Métodos de Estimação", ylab = "Estimativas",
        main = "Comparação entre o Ajuste do modelo misto e o Direto",
        col = c("lightblue", "yellow2"),
        las=1, cex=0.75, boxwex=0.5)
plot(rest_sae$LODS_231, rest_sae$Preditos,
    xlab = "Logaritmo das Estimativas Diretas", ylab = "Estimativas Indiretas",
     col = c("lightblue", "yellow2"),pch = c(16, 16), lwd = 2,
     main = "Relação entre Estimativas Diretas e Indiretas da Amostra",
     cex=0.9)
```

```
legend("topleft", legend = c("Estimativas Diretas", "Estimativas Indiretas"),
       col = c("lightblue", "yellow2"), pch = c(16, 16))
# Adicionar uma linha de regressão
abline((lm(LODS_231 ~ Preditos)), col = "blue", lwd = 2)
detach(rest_sae)

# ::::::::::::::::: # CASO AMOSTRAL # ::::::::::::::::: #
# --- Predições em áreas não pertencentes a amostra --- #

# Parte 1:
# - É necessário extrair os betas(coeficientes) do modelo
# Visualizar coeficientes
ambetas <- as.matrix(coefficients(estihf_ne), ncol=1)
class(ambetas)
str(ambetas)

# Parte 2:
# - Multiplicar os betas com as variáveis explicativas
estd_ne<- estd_bra[estd_bra$regiao == "Nordeste", ]
aux_pop <- estd_ne %>%
  select(V8,V19,V21,V33,V20,V16,
    ID_GINI,Mean_VHI_AS_S2_FAO,V6,V34,
    V29,POP_Area_nao_densa,V5)
str(aux_pop)
head(aux_pop)
amx <- as.matrix(cbind(1, aux_pop)) #base auxiliar
class(amx)
str(amx)

## O vetor preditivo
#As previsões de ambos os domínios os observados na amostra e os não observados
# (converter para matriz se necessário)
am_pred <-  amx%*%ambetas
```

```
#Base de dados com a nomenclatura dos domínios
expl_pop <- estd_ne %>%
  select(name,
         V8,V19,V21,V33,V20,V16,
         ID_GINI,Mean_VHI_AS_S2_FAO,V6,V34,
         V29,POP_Area_nao_densa,V5)
am_pred <- data.frame(name = expl_pop$name, Preditos = am_pred)
dim(am_pred)

#Uma base de dados apenas com os domínios não observados
#Domínios que não foram observado
n_amostra <- setdiff(am_pred$name, unique(am_ne$name))
#n_amostra <- as.data.frame(n_amostra)
#dim(n_amostra) #1662 #NAME
#Predições em domínios não observados
pred_Nobs <- subset(am_pred, am_pred$name %in% n_amostra)
dim(pred_Nobs)
pred_Nobs

# - Parte 2.1: O cálculo do EQM
#Estimar a variância dos coeficientes Betas
## a variância apenas depende das observações amostrais
#Um subconjunto do conjunto de dados original auxiliar
info_aux_obs <- subset(expl_pop, am_pred$name %in%  #auxiliar
                         unique(am_ne$name)) %>% mutate(inter = 1)
str(info_aux_obs) #171x15
head(info_aux_obs)
class(info_aux_obs)
#Uma matriz com duas colunas das amostras dos dados originais
X_obs <- cbind(1,info_aux_obs[,2:15])
X_obs <- as.matrix(X_obs)
class(X_obs)
```

```
# 1 - se estima a variância do efeito aleatório
sigma2_u <- salida$est$fit$refvar
##sigma2_u é uma constante representando a variância incondicional do erro

Vbeta_est <- solve(t(Vi*X_obs) %*% X_obs)
estihf_ne
Vi <- 1/(0.0983+(am_ne$LVAR_POND))
Vi <- 1/(am_ne$LVAR_POND)

#O cálculo da matriz de covariância dos coeficientes estimados
lambda <- 1e-5
Vbeta_est <- solve(t(Vi*X_obs) %*% X_obs + lambda * diag(ncol(X_obs)))
Vbeta_esta <- as.matrix(Vbeta_est)
dim(Vbeta_esta) # Deve retornar (k, k) e k é o número de preditores
Vbeta_esta

# Obter a matriz de covariância dos coeficientes
Vbeta_cov <- vcov(estihf_ne)
Vbeta_cov
dim(Vbeta_cov)
#Organizar as infos auxiliares nos domínios não observados
## Calcular o EQM
info_aux_Nobs <- subset(expl_pop, am_pred$name %in% n_amostra) %>%
  mutate(inter = 1)
dim(info_aux_Nobs)
str(info_aux_Nobs)
X_Nobs <- cbind(1, info_aux_Nobs[,2:15])
X_Nobs_m <- as.matrix(X_Nobs)
dim(X_Nobs_m) # Deve retornar (n, k) e n é o número de observações
str(X_obs)

# A estimativa da variância do efeito aleatório
s2u <- as.numeric(estihf_ne$framework$vardir)
```

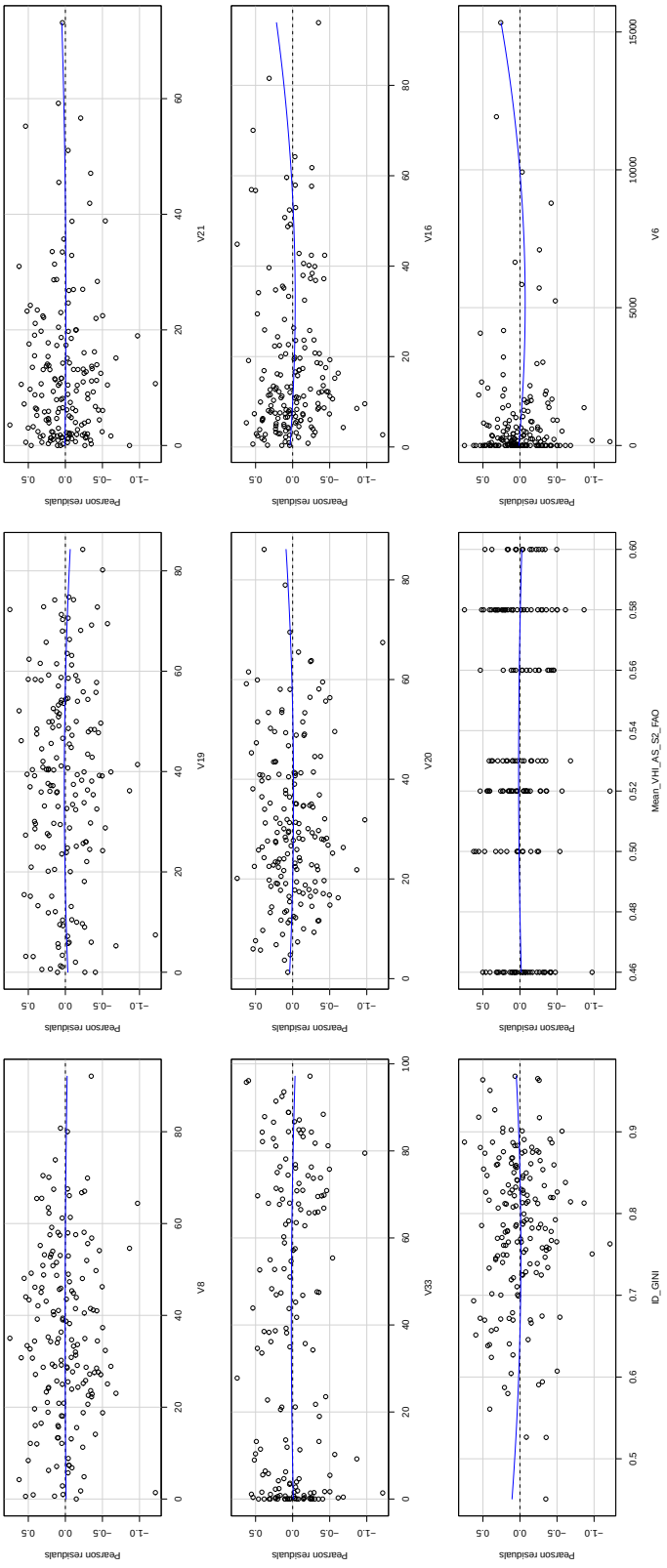
```
# Estimar o EQM de forma matricial
EQM_Nobs <- 0.0983 + diag((X_Nobs_m %*% Vbeta_esta %*% t(X_Nobs_m)))

#Proceder com o cálculo do Erro Quadrático Médio Relativo
EQMR_Nobs <- 100*sqrt(EQM_Nobs)/(pred_Nobs$Preditos)

# Parte 2.2 : Organização das Bases de Dados
pred_Nobs$EQM <- EQM_Nobs
pred_Nobs$EQMR <- EQMR_Nobs
head(pred_Nobs)
dim(pred_Nobs)
pred_Nobs <- as.data.frame(pred_Nobs)
# Domínios da Amostra
head(estim_predhf)
estim2_predhf <- data.frame(name = am_ne$name,
                            Preditos = esti_predhf,
                            EQM = esti_eqmhf,
                            EQMR = eqmr_msehf)
# Agregar os dados
AG_PRED <- rbind(estim2_predhf, pred_Nobs)
#A base de dados completa
#os dados das estimativas diretas e indiretas do modelo ajustado
BC_PREDHF <- full_join(AG_PRED, estd_ne)
head(BC_PREDHF)
write.xlsx(BC_PREDHF, "ESTIMA_PRED_NE_FH_POP.xlsx", rowNames = FALSE)

# - Parte 4:
#Verificação dos Resultados
BC_PREDHF<-read_excel("C:/Users/camis/Desktop/SAE/tab/ESTIMA_PRED_NE_FH_POP.xlsx")
BC_PREDHF <- as.data.frame(BC_PREDHF)

attach(BC_PREDHF)
# Análise Gráfica
```

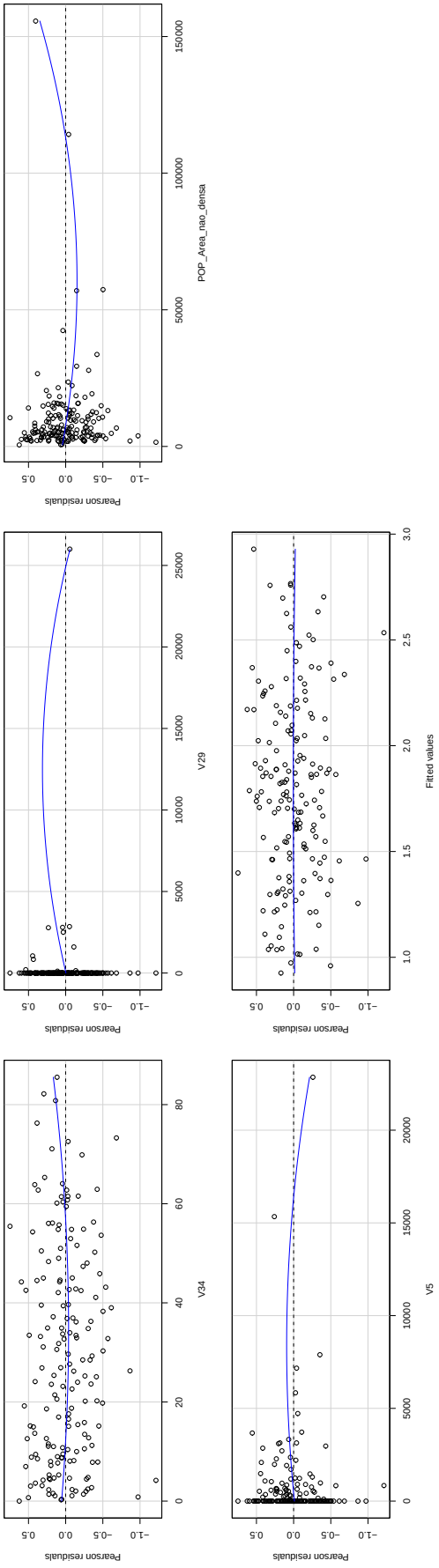
```
plot(BC_PREDHF$LODS_231, BC_PREDHF$Preditos,
     xlab = "Valores do Logaritmo das Estimativas Diretas",
     ylab = "Valores Ajustados",
     col = c("lightblue", "yellow2"),pch = c(16, 16), lwd = 2,
     main = "Relação entre Estimativas Diretas e Indiretas", cex=1)
legend("topleft", legend = c("Estimativas Diretas", "Estimativas Indiretas"),
       col = c("lightblue", "yellow2"), pch = c(16, 16))
# Adicionar uma linha de regressão
abline((lm(LODS_231 ~ Preditos)), col = "blue", lwd = 2)

#BOXPLOT
boxplot(list(Direta = BC_PREDHF$LODS_231, Indireta = BC_PREDHF$Preditos),
        xlab = "Métodos de Estimação", ylab = "Estimativas",
        main = "Boxplot entre o Ajuste do modelo FH e o Direto",
        col = c("lightblue", "yellow2"), las=1, cex=0.75, boxwex=0.5)
detach(BC_PREDHF)

#_____________________________________________________________#
#----------------- Análise Exploratória PE ----------------#

# Análise de Dados do Estado de Pernambuco
PREDPE<-read_excel("C:/Users/camis/Desktop/SAE/RES/ESTIMAS_POP_NE.xlsx")
PREDPE <- PREDPE[PREDPE$UF == "PE", ]
str(PREDPE)
summary(PREDPE)
describe(PREDPE)

# Análise Gráfica
par(mfrow = c(1, 2))
# Gráfico de Dispersão com Linha de Regressão
plot(PREDPE$LODS_231, PREDPE$Preditos.comp,
     xlab = "Logaritmo das Estimativas Diretas",
     ylab = "Valores Ajustados",
```

```
     col = "black", pch = 16, lwd = 2,
     main = "Estimador Composto por Regressão", cex = 1.2)
points(PREDPE$LODS_231, PREDPE$Preditos.comp,
       col = c("ivory2","brown2"), pch = 16, lwd = 2)
legend("topleft", legend = c("Estimativas Diretas", "Estimativas Indiretas"),
       col = c("ivory2","brown2"), pch = 16)
abline(lm(LODS_231 ~ Preditos.comp, data = PREDPE),
       col = "red", lwd = 2, lty=2)

# Gráfico de Dispersão com Linha de Regressão
plot(PREDPE$LODS_231, PREDPE$Preditos.FH,
     xlab = "Logaritmo das Estimativas Diretas",
     ylab = "Valores Ajustados",
     col = "black", pch = 16, lwd = 2,
     main = "Modelo Fay-Herriot", cex = 1.2)
points(PREDPE$LODS_231, PREDPE$Preditos.FH, col = c("ivory2","navy"),
       pch = 16, lwd = 2)
legend("topleft", legend = c("Estimativas Diretas",
                             "Estimativas Indiretas"),
       col = c("ivory2","navy"), pch = 16)
abline(lm(LODS_231 ~ Preditos.FH, data = PREDPE),
       col = "red", lwd = 2, lty = 2)

par(mfrow = c(1, 1))
# Boxplot entre métodos de estimação
bp <- boxplot(list(Direto = PREDPE$LODS_231,
             Sintético = PREDPE$Preditos.syn,
             Composto = PREDPE$Preditos.comp,
             FH = PREDPE$Preditos.FH),
        xlab = "Métodos de Estimação", ylab = "Logaritmo das Estimativas",
        main = "Boxplot entre o Ajuste de Modelos e o Direto",
        col = c("ivory2", "yellowgreen", "brown2", "navy"),
        las = 1, cex = 0.75, boxwex = 0.5)
```

```
# Função para retornar um data frame com geocode e outliers
ot_geo <- function(values, outliers, geocode_col) {
  data.frame(Geocode = geocode_col[values %in% outliers],
             Estimativa = outliers[outliers %in% values])}
# Gerar o frame para os outliers por geocode para cada método
ot_dir <- ot_geo(PREDPE$LODS_231, bp$out[bp$group == 1], PREDPE$geocode)
ot_syn <- ot_geo(PREDPE$Preditos.syn, bp$out[bp$group == 2], PREDPE$geocode)
ot_comp <- ot_geo(PREDPE$Preditos.comp, bp$out[bp$group == 3], PREDPE$geocode)
ot_FH <- ot_geo(PREDPE$Preditos.FH, bp$out[bp$group == 4], PREDPE$geocode)
# Exibição dos resultados por método
ot_dir
ot_syn
ot_comp
ot_FH

#Gráfico de Dispersão das Diferenças
PREDPE <- PREDPE %>%
  mutate(Dif_Sin_Dir = Preditos.syn - LODS_231,
         Dif_Comp_Dir = Preditos.comp - LODS_231,
         Dif_FH_Dir = Preditos.FH - LODS_231)
ggplot(PREDPE, aes(x = LODS_231)) +
  geom_point(aes(y = dif.syn, color = "Sintético")) +
  geom_point(aes(y = dif.comp, color = "Composto")) +
  geom_point(aes(y = dif.fh, color = "FH")) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Logaritmo das Estimativas Diretas", y = "Diferença",
       title = "Diferenças entre Estimativas Diretas e Outros Métodos") +
  scale_color_manual(values = c("Sintético" = "yellowgreen",
  "Composto" = "brown2", "FH" = "navy")) + theme_minimal() +
  theme(legend.position = "top", legend.title = element_blank())
```

APÊNDICE B – GRÁFICOS DE COMPONENTES RESIDUAIS PARA O MODELO LINEAR SIMPLES AJUSTADO




Pearson residuals
POP_Area_nao_densa
V29
Fitted values
V34
V5

ANEXO A – DESCRIÇÃO DAS VARIÁVEIS EXPLICATIVAS

Variável	Descrição	Ano	Fonte
ESTABEL_ODS	Quantidade de estabelecimentos agropecuários	2017	FAO
cod_reg	Codificação para as regiões brasileiras	2022	IBGE
POP	População total	2022	IBGE
AREA	Área total em hectare (ha)	2022	IBGE
POP_AREA_NAO_DENSA	População total em área não densa	2017	IBGE
TOTAL_AREA_URB	Área total urbanizada (ha)	2019	IBGE
IDHM	Índice de Desenvolvimento Humano Municipal	2010	IPEA
ID_GINI	Índice de Gini	2017	IBGE
V1	Número de estabelecimentos agropecuários	2017	IBGE
V2	Área média (ha) perante V1	2017	IBGE
V3	Pessoal Ocupado / Estabelecimento (Pessoa)	2017	IBGE
V4	Área lavoura / Adubadeira (ha)	2017	IBGE
V5	Área lavoura / Colheitadeira (ha)	2017	IBGE
V6	Área lavoura / Semeadeira (ha)	2017	IBGE
V7	Área lavoura / Trator (ha)	2017	IBGE
V8	Atividade-Lavoura Temporária (%)	2017	IBGE
V9	Atividade-Lavoura Permanente (%)	2017	IBGE
V10	Atividade-Pecuária (%)	2017	IBGE
V11	Atividade-Horticultura&Floricultura (%)	2017	IBGE
V12	Atividade-Sementes&Mudas (%)	2017	IBGE
V13	Atividade-Produção Florestal (%)	2017	IBGE
V14	Atividade-Pesca (%)	2017	IBGE
V15	Atividade-Aquicultura (%)	2017	IBGE
V16	Uso das terras-Lavoura (%)	2017	IBGE
V17	Uso das terras-Pastagem (%)	2017	IBGE
V18	Aves-Corte (%)	2017	IBGE
V19	Aves-Ovos (%)	2017	IBGE
V20	Bovinos-Corte (%)	2017	IBGE
V21	Bovinos-Leite (%)	2017	IBGE
V22	Rendimento-Arroz (kg/ha)	2017	IBGE
V23	Rendimento-Cana (kg/ha)	2017	IBGE
V24	Rendimento-Mandioca (kg/ha)	2017	IBGE
V25	Rendimento-Milho (kg/ha)	2017	IBGE
V26	Rendimento-Soja (kg/ha)	2017	IBGE
V27	Rendimento-Trigo (kg/ha)	2017	IBGE
V28	Rendimento-Cacau (kg/ha)	2017	IBGE
V29	Rendimento-Café (kg/ha)	2017	IBGE
V30	Rendimento-Laranja (kg/ha)	2017	IBGE
V31	Rendimento-Uva (kg/ha)	2017	IBGE
V32	Carga de Bovinos (n/ha)	2017	IBGE
V33	Cisterna (%)	2017	IBGE
V34	Utilização de Agrotóxicos (%)	2017	IBGE
V35	Despesa com Agrotóxicos (%)	2017	IBGE
V36	Uso de irrigação (%)	2017	IBGE
V37	Assistência Técnica (%)	2017	IBGE
V38	Agricultura familiar (%)	2017	IBGE
V39	Produtor com escolaridade até Ensino Fundamental (%)	2017	IBGE
Altitude_mean	Altitude Média (m)	2000	FIOCRUZ
bio01_mean	Média da temperatura média anual (ºC)	2000	FIOCRUZ
bio02_mean	Média da variação diurna de temperatura (ºC)		FIOCRUZ
bio03_mean	Isotermalidade média (ºC)		FIOCRUZ

Variável	Descrição	Ano	Fonte
bio04_mean	Média da sazonalidade da temperatura (ºC)	2000	FIOCRUZ
bio07_mean	Média da amplitude térmica anual (ºC)		FIOCRUZ
CHG_bio12_mean	Média da precipitação anual (mm)	2020	FIOCRUZ
CHG_bio15_mean	Média da sazonalidade de precipitação - CV (mm)		
Roads_mean	Média de trechos rodoviários e estradas (m)	2021	FIOCRUZ
Waterways_mean	Média de hidrovias (m)		
Watercourse_mean	Média de curso d'água (m)		
Dams_mean	Média de barragens (m)		
Amazon_Forest_mean	Delimitação dos biomas, que, com poucas exceções, seguem a delimitação original dos tipos de vegetação que compô-los (ha)	2006	FIOCRUZ
Atlantic_Forest_mean			
Caatinga_mean			
Cerrado_mean			
Pampa_mean			
Pantanal_mean			
Beach_dune_sand_spot_mean	Fios arenosos, de cor branco brilhante, onde não existe predominância de qualquer tipo de vegetação (ha)	2020	FIOCRUZ
Citrus_mean	Áreas predominantemente ocupadas por culturas cítricas (ha)	2020	FIOCRUZ
Forest_formation_mean	Agregado de formações florestais naturais, atravésda diversidade de fitofisionomias (ha)	2020	FIOCRUZ
Forest_plantation_mean	Espécies de árvores plantadas para fins comerciais (ha)	2020	FIOCRUZ
Grassland_mean	Tipos de vegetação com predominância de espécies herbáceas (ha)	2020	FIOCRUZ
Mangrove_mean	Formações florestais densas e perenes, muitas vezes inundadas pela maré e associadas ao ecossistema costeiro de Manguezal (ha)	2020	FIOCRUZ
Mining_mean	Áreas referentes à extração mineral em grande escala, com clara exposição do solo pela ação de máquinas pesadas (ha)	2020	FIOCRUZ
Other_non_forest_formations_mean	Vegetação herbácea com influência fluviomarina (ha)	2020	FIOCRUZ
Other_non_vegetaded_areas_mean	Classe mista que inclui maioritariamente áreas agrícolas em preparação, com solo exposto no início da primavera (ha)	2020	FIOCRUZ
Other_perennial_crop_mean	Áreas predominantemente ocupadas com presença de outras culturas perenes (ha)	2020	FIOCRUZ
Other_temporary_crops_mean	Áreas predominantemente ocupadas com presença de outras culturas temporárias (ha)	2020	FIOCRUZ
River_lake_ocean_mean	Rios, lagos, barragens, reservatórios e outros corpos d'água (ha)	2020	FIOCRUZ
Rocky_outcrop_mean	Áreas que não possuem nenhum tipo de vegetação, apenas a rocha (ha)	2020	FIOCRUZ
Salt_flat_mean	Apicuns ou Salgados são formações quase sempre desprovidas de vegetação arbórea associadas a uma área mais alta, hipersalina e menos alagada do mangue (ha)	2020	FIOCRUZ
Savanna_formation_mean	Tipos de vegetação com predominância de espécies de dossel semi-contínuo (ha)	2020	FIOCRUZ
Wetlands_mean	Zonas inundáveis com rede de lagoas interligadas, localizadas junto a cursos de água e em zonas de depressões que acumulam água (ha)	2020	FIOCRUZ
Wooded_restinga_mean	Áreas de vegetação com predominância de espécies arbóreas com influência marinha (ha)	2020	FIOCRUZ
Mean_VHI_AS_S1	Média do Índice de Saúde da Vegetação estação 1	2022	FAO
Mean_VHI_AS_S2	Média do Índice de Saúde da Vegetação estação 2		
ASI_AS_S1	Índice de Estresse Agrícola estação 1		
ASI_AS_S2	Índice de Estresse Agrícola estação 2		
ndvi_adm1	Índice de Vegetação por Diferença Normalizada		