Universidade de Brasília
Departamento de Estatística

AVALIAÇÃO DO PERFIL DOS BENEFICIÁRIOS DO PROGRAMA BOLSA FAMÍLIA EM 2018

Ana Theresa Figueiredo Camurça Martins

Relatório apresentado ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Ana Theresa Figueiredo Camurça Martins

AVALIAÇÃO DO PERFIL DOS BENEFICIÁRIOS DO PROGRAMA BOLSA FAMÍLIA EM 2018

Orientador: Leandro Tavares Correia

Relatório apresentado ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Agradecimentos

- Agradeço à minha família, aos meus pais e ao meu irmão, pelo apoio que sempre me deram durante toda a minha vida e principalmente durante essa graduação.
- Faço um agradecimento especial ao meu orientador, Leandro Tavares Correia, pelo incentivo e paciência pela dedicação do seu escasso tempo ao meu projeto de pesquisa.
- Agradeço aos membros da Banca, Maria Teresa Leão e Luís Gustavo Vinha, por terem aceitado ao convite, bem como pela sua disponibilidade e atenção.
- Agradeço, também, à Universidade de Brasília, ao Departamento de Estatística e ao corpo docente do meu curso pelo aprendizado e oportunidades ofertados.

Resumo

O presente estudo visa avaliar o perfil dos domicílios beneficiários do Programa Bolsa Família (PBF), utilizando a base de dados da Pesquisa Orçamentária Familiar - POF (2017-2018) do IBGE e, observar se há diferenças significativas na vulnerabilidade social entre os domicílios urbanos beneficiários e os dos não beneficiários que apresentam renda total de até R$ 2.000,00.

Primeiramente foi feita uma Análise Fatorial exploratória com as variáveis binárias do questionário sobre as condições de vida e, em seguida, foi construída a Regressão Logística e a Floresta Aleatória para a classificação dos domicílios beneficiários e dos não beneficiários. Ademais, realizou uma comparação entre essas técnicas estatísticas.

Os resultados obtidos na Regressão Logística e na Floresta Aleatória indicaram que o grupo dos beneficiários é o que possui maior vulnerabilidade social e as variáveis Insegurança Alimentar e Região Geográfica foram as mais significativas para a classificação dos grupos. Por fim, em 2018, os resultados sinalizaram que o PBF está alcançando os domicílios alvo.

Palavras-chaves: Bolsa Família; Análise Multivariada; Análise Fatorial; Regressão Logística; Floresta Aleatória.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

O Programa Bolsa Família (PBF) foi criado em 2003, no primeiro Governo do Luiz Inácio Lula da Silva. Este Programa surgiu a partir de estudos elaborados para viabilizar uma política de transferência de renda condicionada, com a finalidade do combate à fome no país (CAMPELLO; NERI, 2013).

Esta modalidade de política pública é reconhecida e prestigiada mundialmente, cujo objetivo fundamental é o de garantir os direitos sociais às famílias mais pobres e, por consequência, possibilitar um impacto positivo em prol da melhoria das condições de vida e da criação de oportunidades para milhares de brasileiros socialmente vulneráveis. Contudo, a intenção inicial do PBF é o de aliviar imediatamente as consequências da miséria (BRAUW et al., 2005).

Com a instituição do PBF houve a unificação dos programas oficiais não contributivos anteriormente existentes tais como: Bolsa Escola, Bolsa Alimentação, Cartão Alimentação e Auxílio Gás. Ressalta-se que esses dois programas sociais – Bolsa Escola e Bolsa Família – já estavam sendo implementados desde 2001 e abrangiam mais de 5 (cinco) mil municípios brasileiros (PASE; MELO, 2017; ORTIZ; CAMARGO, 2021; SOUZA et al., 2019).

Para participar do PBF são necessários cumprir requisitos específicos como: estar incluído no Cadastro Único; crianças matriculadas em escolas; gestantes realizando os exames de pré-natal, puérperas realizando os exames pós-natal, dentre outras exigências. Segundo Campello e Neri (2013), o PBF visa contribuir para interrupção do ciclo intergeracional da pobreza. Para tanto, uma série de estratégias de integração de ações públicas são requeridas para aprimorar o acesso aos serviços de saúde, educacionais e assistência social (CRISTÓVÃO, 2023; GERMANN; MEDEIROS, 2022).

No ano de 2003, época do surgimento do PBF, o índice de pobreza extrema no Brasil alcançou 28,2% da população e em 2019 este índice foi reduzido para 6,5%, segundo o IBGE (2018). Como efeito do êxito desta política pública, em 2014, registrou-se um importante marco para o país que consistiu na retirada do Brasil da lista de países que integram o Mapa da Fome da Organização das Nações Unidas (ONU), o que significa que a subalimentação (consumo de alimentos em quantidade insuficiente para atender às necessidades nutricionais do indivíduo) reduziu-se a menos que 5% da população. Paralelamente, se evidenciaram os efeitos positivos na esfera econômica com a queda do coeficiente de Gini, que passou de 0,594 em 2001 para 0,543 em 2019 (GERMANN; ME-

DEIROS, 2022).

Em meados da década de 70, o IBGE começou a realizar pesquisa qualitativa de âmbito nacional destinada a mensurar os orçamentos das famílias, os aspectos nutricionais e as condições de vida dos brasileiros, com o objetivo de subsidiar as políticas públicas nacionais. Assim, no período entre 10 de julho de 2017 e 10 de julho de 2018 foi realizada a quarta pesquisa nacional denominada de Pesquisa de Orçamento Familiar (POF) de 2017-2018, com fins de mensurar as estruturas de consumo das famílias e possibilitar o esboço do perfil das condições de vida da população, a partir da análise de seus orçamentos domésticos e, por conseguinte, subsidiar o estabelecimento de prioridades no campo das políticas públicas relativas à saúde, educação, nutrição, distribuição de alimento, dentre outras. Nesta POF houve, também, a análise dos dados relativos aos beneficiários e não beneficiários do PBF (IBGE, 2018).

Em 2018, o valor básico do auxílio do PBF destinado à cada família equivale a R$ 182,00, com a previsão de valores adicionais destinados a dois grupos distintos: grupo de crianças de até 15 anos de idade e de gestantes e o grupo de jovens entre 16 e 17 anos de idade que correspondiam, respectivamente, ao valor de R$ 41,00 e de R$ 48,00, sendo autorizado, no máximo, 5 benefícios para o primeiro grupo citado e, no máximo, 2 benefícios para o outro grupo (IBGE, 2018).

Destaca-se que esses valores são atualizados ao longo dos anos e, em 2023, com a aprovação da Medida Provisória nº 1116/23, foi definido o benefício básico de R$ 600,00, sendo necessário possuir uma renda familiar *per capita* de até R$ 218,00, havendo valor adicional de R$ 150,00 para cada criança com idade de até 6 anos (Benefício da Primeira Infância) e; para cada criança entre 7 e 12 anos, adolescente entre 13 e 18 anos e gestante foi adicionado o valor de R$ 50,00, possibilitando que o somatório dos valores ultrapasse a quantia de R$ 1.000,00 (CRISTÓVÃO, 2023).

Assim, face às considerações expostas e reconhecendo a relevância do PBF para o país, foi desenvolvido um estudo, utilizando-se a base de dados da POF de 2017-2018, para avaliar as características socioeconômicas do grupo dos domicílios beneficiários e do grupo dos não beneficiários do PBF. A partir dessa análise, observar se há diferença na vulnerabilidade social entre estes dois grupos.

Por meio do uso de diferentes técnicas estatísticas, a partir das respostas relacionadas às condições de vida, utilizando-se a análise descritiva para observar o perfil dos grupos dos beneficiários e dos não beneficiários, depois a análise fatorial exploratória com três fatores para analisar o comportamento das variáveis entre si relativas ao banco de

dados selecionado para o estudo. Em seguida, serão utilizadas as técnicas de Regressão Logística e de Floresta Aleatória para a classificação dos domicílios como beneficiários e não beneficiários e para identificar as variáveis mais determinantes para esta classificação.

Este estudo foi organizado em quatro outras seções seguidas da Introdução: Referencial Teórico (Seção 2), no qual se abordou os conceitos técnicos relativos à Análise Fatorial, a Regressão Logística e a Floresta Aleatória; Metodologia (Seção 3), que tratou do conjunto de dados selecionados e os procedimentos empregados para análise; Resultados (Seção 4), no qual foram apresentados as respostas obtidas a partir das diferentes técnicas estatísticas utilizadas e; por fim a Conclusão (Seção 5) deste estudo.

2 Referencial Teórico

Nesta seção são apresentados os conceitos técnicos específicos utilizados para o estudo.

2.1 Análise Fatorial

Nesta subseção são apresentados a Análise Fatorial exploratória (AFE) e a Análise Fatorial para dados binários utilizadas para o estudo.

2.1.1 Análise Fatorial Exploratória

A AFE é uma técnica multivariada segundo Joseph et al. (2009), e o seu principal propósito é o de identificar a estrutura subjacente entre as variáveis na análise. Isso se torna possível por meio da análise das covariâncias entre as variáveis presentes na base de dados, o que também permite, caso exista, a identificação de fatores que explicam a variabilidade dessa base de dados. Também, auxilia a compreender melhor os dados possibilitando encontrar determinados padrões (REVELLE, 2009; HASTIE; TIBSHIRANI; WAINWRIGHT, 2015).

Adicionalmente, a redução da base de dados é efetuada por meio dos cálculos de correlações e examinando se é possível agrupar ou não os dados, de acordo com os valores encontrados na matriz de correlação e com a proximidade entre os resultados (REVELLE, 2009).

Outrossim, segundo Kim e Mueller (1978), a AFE se baseia no pressuposto fundamental de que alguns fatores subjacentes, que são em menor número que as variáveis observadas, são responsáveis pela covariação entre as variáveis.

Ademais, a partir desses fatores podem ser inferidas as variáveis latentes, ou seja, variáveis que não são diretamente observáveis, e para tanto é necessário que uma série de variáveis estejam correlacionadas para identificar os grupos.

Para demonstrar a Análise Fatorial, considera-se a base de dados como a matriz ($\mathbf{W}$) de n escores de desvio para $\mathbf{N}$, onde cada elemento, w_{ij}, representa as respostas da i-ésima observação ao j-ésimo item. Então, a matriz de covariância, Cov, é (REVELLE, 2009):

$$Cov = \mathbf{N}^{-1}\mathbf{W}\mathbf{W}^{T} \tag{2.1.1}$$

$$sd = \sqrt{\text{diag}(Cov)},$$

assim, consegue-se identificar a matriz de correlação ($\mathbf{R}$), onde I_{sd} é a diagonal da matriz com elementos $= 1/sd_i$,

$$\mathbf{R} = I_{sd}CovI_{sd}. \tag{2.1.2}$$

É possível com produto de dois fatores calcular uma aproximação da matriz de correlação $\mathbf{R}$ pela equação a seguir

$$\mathbf{R} \approx \mathbf{C}\mathbf{C}', \tag{2.1.3}$$

considerando que n é o número de variáveis em $\mathbf{R}$, então o i-ésimo componente, C_i, é uma soma linear das variáveis,

$$C_i = \sum_{j=1}^{n} w_{ij}x_j. \tag{2.1.4}$$

O modelo de fatores aparenta ser semelhante, mas com a adição de uma matriz diagonal de singularidades $\mathbf{U}^2$, e considerando que a matriz de correlação ($\mathbf{R}$) é formada como o produto matricial de um vetor $\mathbf{F}\mathbf{F}'$ é possível ter a equação a seguir:

$$\mathbf{R} \approx \mathbf{F}\mathbf{F}' + \mathbf{U}^2, \tag{2.1.5}$$

e com as variáveis descritas como somas lineares ponderadas dos fatores desconhecidos

$$x_i \approx \sum_{j=1}^{n} w_{ij}F_j. \tag{2.1.6}$$

Em seguida, é fundamental calcular o autovalor, que representa a quantia de variância explicada por um fator e mede a variância em todas as variáveis que é devida a este fator. A seguir, é calculado o autovetor de cada autovalor, que seria a orientação do vetor (REVELLE, 2009).

As fórmulas, a seguir, são para obter os cálculos dos autovetores e autovalores respectivamente. O valor da matriz de correlação ($\mathbf{R}$) definida na equação 2.1.2, onde λ é a matriz diagonal dos autovetores e $\mathbf{x}_i$ é cada autovetor e λ_i é cada autovalor:

$$\mathbf{x}_i\mathbf{R} = \lambda_i\mathbf{x}_i, \tag{2.1.7}$$

e o conjunto de n autovetores são soluções para a equação,

$$\mathbf{RX} = \mathbf{X}\lambda. \tag{2.1.8}$$

Destaca-se essa relação entre os autovetores e a matriz de desvio padrão na equação abaixo:

$$\sum_{i=1}^{q} \lambda_i = s_1^2 + s_2^2 + \ldots + s_q^2 = \mathrm{t}r(sd), \tag{2.1.9}$$

o somatório dos autovetores (λ_i) é igual à soma das variâncias que é a matriz diagonal do desvio padrão (REVELLE, 2009).

A equação dos autovalores e autovetores e a expressão $\mathbf{x}_i(\mathbf{R}-\lambda_i I) = 0$ mostra que $\mathbf{x}_i$ é um autovetor associado ao autovalor λ_i da matriz $\mathbf{R}$. A equação pode ser rearranjada para resolver os autovalores e autovetores,

$$\mathbf{x}_i\mathbf{R} - \lambda_i\mathbf{x}_i I = 0 \Leftrightarrow \mathbf{x}_i(\mathbf{R} - \lambda_i I) = 0. \tag{2.1.10}$$

O fato de que os vetores que compõem $\mathbf{X}$ são ortogonais significa que podem ser decompostos conforme a equação 2.1.11,

$$\mathbf{X}\mathbf{X}^\top = I \tag{2.1.11}$$

$$\mathbf{R} = \mathbf{X}\lambda\mathbf{X}^\top,$$

ou seja, é possível recriar a matriz de correlação $\mathbf{R}$ em termos de um conjunto ortogonal de vetores (os autovetores) escalonados por seus autovalores associados (REVELLE, 2009).

2.1.2 Análise Fatorial para dados binários

A Análise Fatorial para dados binários é possível utilizando-se o coeficiente de correlação policórica. O seu objetivo é estimar a correlação entre as variáveis e essa métrica é muito comum em estudos psicométricos (STARKWEATHER, 2014).

O coeficiente de correlação policórica é uma medida de associação bivariada e pode ser estimado pelo método de máxima verosimilhança Drasgow (2006). Neste método, a

probabilidade conjunta (P_{ij}) de se observar o valor x_i para a variável X e o valor y_i para a variável Y é estimada por

$$P_{ij} = \int_{\gamma_{i-1}}^{\gamma_i} \int_{\tau_{j-1}}^{\tau_j} \phi(x, y; \rho)\, dy\, dx, \tag{2.1.12}$$

onde $\phi(x, y; \rho)$ é a função de densidade normal bivariada com coeficiente de correlação ρ de X e Y que são as variáveis latentes. Os limites das integrais foram definidos a partir das variáveis latentes, que $\gamma_{i-1} \leq X \leq \gamma_i$ e $\tau_{i-1} \leq Y \leq \tau_i$.

A função de verossimilhança de uma amostra é:

$$L = k \prod_{i=1}^{r} \prod_{j=1}^{s} P_{ij}^{n_{ij}}, \tag{2.1.13}$$

sendo que:

- k é uma constante;
- n_{ij} é o número de observações;
- $n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$, r é a quantidade de variáveis e s o tamanho amostral.

Para obter o estimador de máxima verossimilhança para ρ, primeiramente é feita a transformação logarítmica da equação 2.1.7 e em seguida a equação é maximizada, denominada l. Respeitando os parâmetros, a equação 2.1.8 é o resultado esperado:

$$\frac{\partial l}{\partial \rho} = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} P_{ij} \left[\Phi(\lambda_i, \tau_j, \rho) - \Phi(\lambda_{i-1}, \tau_j, \rho) - \Phi(\lambda_i, \tau_{j-1}, \rho) + \Phi(\lambda_{i-1}, \tau_{j-1}, \rho) \right]. \tag{2.1.14}$$

Igualando a derivada parcial a 0 e resolvendo a equação obtém-se a estimativa de ρ. Contudo, a solução do estimador de máxima verosimilhança é obtida de forma iterativa. Para resolução da derivada parcial em ordem ρ é necessário os parâmetros da função de verosimilhança. A estimação destes parâmetros, por sua vez, exige a derivada parcial de l em ordem a cada um dos coeficientes e a resolução do sistema com todas as derivadas parciais iguais a 0 (MAROCO, 2010; DRASGOW, 2006; MARTINSON; HAMDAN, 1975).

Após o cálculo do coeficiente de correlação policórica, é possível aplicar a análise fatorial e observar a variável latente e quais variáveis possuem uma carga fatorial maior para a Análise Fatorial.

2.2 Regressão Logística

A Regressão Logística é utilizada quando a variável resposta é qualitativa, sendo possível estimar a probabilidade de ocorrência de determinado fenômeno, a partir da análise das variáveis explicativas e da variável resposta.

Segundo o Agresti (2007), a Regressão Logística utiliza o Método de Máxima Verossimilhança para estimar seus parâmetros, maximizando a probabilidade da amostra ter sido observada e o seu resultado é uma probabilidade de determinado evento ocorrer ou não. A variável resposta Y possui uma distribuição de Bernoulli, na qual representa a probabilidade de tal evento ter sucesso (π) ou fracasso (1 - π).

O objetivo da Regressão Logística é modelar a probabilidade de que a variável resposta pertença a uma determinada categoria com base em um conjunto de variáveis preditoras. A Regressão Logística modela a probabilidade de que a variável resposta seja um "sucesso"(ou seja, igual a 1) ou "fracasso"(ou seja, igual a 1) com base em um conjunto de variáveis preditoras. Logo, o resultado da variável resposta está entre 0 ou 1.

Ademais, o parâmetro β é a taxa de crescimento ($\beta > 0$) ou de decrescimento ($\beta < 0$) da curva para $\pi(x)$. O modelo de regressão com a transformação *logito* é o seguinte:

$$logito(\pi(X)) = \ln[\frac{\pi}{1-\pi}] = (\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p), \tag{2.2.1}$$

Ao retirar a transformação *logito*, isto é, a probabilidade de sucesso dividida pela probabilidade de fracasso, os parâmetros da função não ficam lineares,

$$\pi(X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}. \tag{2.2.2}$$

A estimação dos parâmetros depende da variável Y_i que representa o sucesso ou o fracasso em uma amostra de tamanho n com independência das observações. A equação 2.2.3 indica a função de máxima verossimilhança,

$$L(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \left(\pi_i^{Y_i} + (1-\pi_i)^{(1-Y_i)}\right), \tag{2.2.3}$$

aplicando o logaritmo que transforma o produto em uma função de soma e considerando $1 - \pi_i = [1 + \exp(\beta_0 + \beta_1 X_i)]^{-1}$ temos essa equação,

$$\ln L(\beta) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1 X_i+\ldots+\beta_p X_i) - \sum_{i=1}^{n} \ln\left[1+\exp(\beta_0+\beta_1 X_i+\ldots+\beta_p X_i)\right], \quad (2.2.4)$$

onde:

- ln é o logaritmo natural;
- $X_1, X_2, \ldots, X_p$ são as variáveis explicativas independentes;
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes da regressão, que representam o impacto das variáveis independentes na log-odds (logaritmo das chances) do evento ocorrer;

após a aplicação do *log*, é necessário derivar a equação 2.2.4 e é possível observar como é estimada a equação 2.2.5

$$\hat{\pi_i} = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 X_1+\ldots+\hat{\beta}_p X_p)}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 X_1+\ldots+\hat{\beta}_p X_p)}. \quad (2.2.5)$$

A partir do resultado da Regressão Logística, é possível montar a matriz de confusão, que é uma tabela na qual se permite a visualização do desempenho de um modelo classificatório e para obter esse desempenho é possível calcular a acurácia e a eficiência. Considerando que o $logito(\pi) = 1$ são os positivos e o $logito(\pi) = 0$ são os negativos. Segue a Tabela 1 com a representação da matriz de confusão:

Tabela 1: Tabela da Matriz de Confusão.

		Referência	
		0	1
Predição	0	VN	FN
	1	FP	VP

Abaixo segue a equação da acurácia do cálculo da Matriz de Confusão:

$$\text{Acurácia} = \frac{VP+VN}{VP+VN+FP+FN}, \quad (2.2.6)$$

onde:

- VP: Verdadeiro positivo - classificou os positivos corretamente;
- VN: Verdadeiro negativo - classificou os negativos corretamente;

- FP: Falso positivo - classificou incorretamente os negativos como positivos;
- FN: Falso negativo - classificou incorretamente os positivos como negativos.

2.3 Árvores e Florestas aleatórias

As Árvores e Florestas aleatórias são modelos utilizados para classificação e previsão. Para isto, podem ser utilizadas, como variável resposta, variáveis categorizadas ou contínuas. É comum se utilizar dados de treinamento e dados de validação, que consiste em selecionar uma amostra do conjunto de dados para construir o modelo e o restante do banco de dados é utilizado para validar o modelo construído (MORETTIN; SINGER, 2021).

Ademais, essa técnica é popular pela sua facilidade na interpretação e na implementação, porém é menos precisa que um modelo de regressão em geral. Segundo o Morettin e Singer (2021), quando são definidas as variáveis preditoras $(X_1, X_2, ..., X_n)$, o algoritmo usado na construção de árvores de decisão consiste essencialmente na determinação das regiões (retângulos mutuamente exclusivos) em que o espaço das variáveis preditoras é particionado.

Conforme ainda o Morettin e Singer (2021), para o vetor de variáveis preditoras $\mathbf{X} = (X_1, \ldots, X_p)$, o algoritmo obedece a seguinte sequência de passos:

1. Selecione uma variável preditora X_j e um limiar (ou ponto de corte) t, de modo que a divisão do espaço das variáveis preditoras nas regiões $\{\mathbf{X} : X_j < t\}$ e $\{\mathbf{X} : X_j \geq t\}$ corresponda ao menor erro de predição (ou de classificação).

2. Para todos os pares (j, t), considere as regiões $R_1(j, t) = \{\mathbf{X} : X_j < t\}$ e $R_2(j, t) = \{\mathbf{X} : X_j \geq t\}$ e encontre o par (j, t) que minimiza o erro de predição (ou de classificação) adotado.

3. Repita o procedimento, agora dividindo uma das duas regiões encontradas, obtendo três regiões; depois divida cada uma dessas três regiões minimizando o erro de predição (ou de classificação).

4. Continue o processo até que algum critério de parada (obtenção de um número mínimo fixado de elementos em cada região, por exemplo) seja satisfeito.

As Árvores de Decisão são construídas uma única vez e possuem alta variabilidade. Uma alternativa para esse problema é criar dados de treinamento e de validação

e utilizar uma Floresta Aleatória, que consiste em um "conjunto de árvores". No caso das Florestas Aleatórias, diversas árvores são geradas com os dados de treinamento, e as categorias que mais se repetem são selecionadas para a classificação final.

Ademais, as Florestas Aleatórias são indicadas para diminuir o viés e a variância das árvores. Segundo Morettin e Singer (2021), para a construção de cada nó de cada árvore, ao invés de escolher a melhor variável preditora dentre as p disponíveis no conjunto de treinamento, deve-se optar pela melhor delas dentre um conjunto de $m < p$ (variáveis preditoras) selecionadas ao acaso. Usualmente, escolhe-se $m \approx \sqrt{p}$, considerando que m é o tamanho do conjunto de dados e p a quantidade de variáveis preditoras.

Em relação ao problema da alta variabilidade, as Florestas Aleatórias forçam cada divisão considerar apenas um subconjunto dos preditores. Portanto, em média, $(p-m)/p$ as divisões não consideraram nem mesmo o preditor forte, dando assim mais chances aos outros preditores. Pode-se pensar nesse processo como uma forma de "descorrelacionar"as árvores, tornando a média das árvores resultantes menos variável e, portanto, mais confiável. Usar um valor pequeno de m na construção de uma Floresta Aleatória será tipicamente útil quando houver um grande número de preditores correlacionados (JAMES et al., 2013).

As Florestas Aleatórias são compostas por diversas árvores de decisão em amostras de treinamento *bootstrap*, que é uma técnica de reamostragem. Nesse caso, a técnica consiste em obter B réplicas *bootstrap* do conjunto de dados de treinamento. Assim, divide-se aleatoriamente as observações em um conjunto de treinamento e um conjunto de teste, e aplica-se as Florestas Aleatórias ao conjunto de treinamento para valores diferentes do número de variáveis de divisão m. As Florestas Aleatórias não causam *overfitting* se aumentarmos B, então, na prática, usa-se um valor de B suficientemente grande para que a taxa de erro se estabilize (JAMES et al., 2013).

Para o diagnóstico de um modelo da Floresta Aleatória é utilizada a figura do Índice de Impureza de Gini, que mede a probabilidade de uma instância qualquer ser escolhida aleatoriamente. Quanto menor o valor, maior o grau de pureza daquele nó. Atinge seu mínimo (zero) quando todos os casos no nó caem em uma única categoria de destino. Outro parâmetro essencial para o diagnóstico é o número de variáveis testadas por divisão (B), que controla quantas variáveis são consideradas em cada nó para determinar a melhor divisão. O ajuste desse parâmetro pode ajudar a reduzir o *overfitting*. O último critério do diagnóstico é a taxa de erro estimada OOB, que é uma medida importante da precisão do modelo, calculada usando as amostras de treinamento. Esta taxa de erro fornece uma estimativa do erro do modelo.

3 Metodologia

Trata-se de um estudo exploratório no qual se utiliza o banco de dados da POF de 2017-2018, realizada pelo IBGE. A coleta de dados dessa POF foi efetuada entre 10 de julho de 2017 e 10 de julho de 2018, por meio de uma amostragem por conglomerados em 2 estágios por estratificação, sendo que, o primeiro estágio, consistiu em escolher aleatoriamente as áreas de setores censitários e, o segundo estágio, selecionar os domicílios de cada setor, nos quais serão aplicados os questionários (IBGE, 2018).

3.1 Conjunto de dados

A metodologia utilizada pela POF é a amostragem probabilística em 2 estágios: primeiro são selecionadas aleatoriamente as áreas determinadas pelo próprio IBGE e depois selecionam quais casas vão realizar os questionários.

A pesquisa é feita com duração de 1 ano e são ao todo 7 questionários que abrangem os seguintes temas: domicílio e suas características, estrutura familiar e orçamento. Os questionários que foram mais utilizados neste estudo são os 5 e 6 referentes aos rendimentos familiares e as condições de vida respectivamente. Nas Tabelas 2, 3 e 4 seguem os dicionários das variáveis utilizadas nas respectivas etapas desse estudo.

Tabela 2: Os códigos e o significado das variáveis utilizadas na Análise Descritiva.

Nome das Variáveis	**Categorias**
Trabalho	Conta própria Setor privado Setor público Trabalhador doméstico Militar do exército Empregador Trabalhador não remunerado
Renda familiar*	
Renda *per capita**	
Preocupação dos moradores com os alimentos acabarem nos últimos 3 meses	Sim Não
Existência de carteira assinada	Sim Não
Contribuição na previdência social	Sim Não
Região Geográfica	Centro-Oeste Nordeste Norte Sudeste Sul

[0]* As variáveis renda familiar e renda *per capita* são quantitativas contínuas.

Tabela 3: Os códigos e o significado das variáveis utilizadas na Análise Fatorial.

Código	**Nome da variável**
V61061	No seu domicílio há problema de pouco espaço?
V61062	No seu domicílio há problema de casa escura, com pouca iluminação natural?
V61063	No seu domicílio há problema de telhado com goteira?
V61064	No seu domicílio há problema de fundação, paredes ou chão úmidos?
V61065	No seu domicílio há problema de madeira das janelas, portas ou assoalhos deteriorados?
V61066	No seu domicílio há problema de mosquitos ou outros insetos, ratos, etc.?
V61067	No seu domicílio há problema de fumaça, mau cheiro, barulho ou outros problemas ambientais causados pelo trânsito ou indústria?
V61068	No seu domicílio há problema de estar localizado próximo a rio, baía, lago, açude ou represa poluídos?
V61069	No seu domicílio há problema de estar localizado em área sujeita a inundação?
V610610	No seu domicílio há problema de estar localizado em encosta ou área sujeita a deslizamento?
V610611	No seu domicílio há problema de violência ou vandalismo na sua área de residência?
V6108	Nos últimos três meses, os moradores deste domicílio tiveram a preocupação de que os alimentos acabassem antes de poderem comprar ou receber mais comida?
V6109	Nos últimos três meses, os alimentos acabaram antes que os moradores deste domicílio tivessem dinheiro para comprar mais comida?
V6110	Nos últimos três meses, os moradores deste domicílio ficaram sem dinheiro para ter uma alimentação saudável e variada?
V6111	Nos últimos três meses, os moradores deste domicílio comeram apenas alguns poucos tipos de alimentos que ainda tinham porque o dinheiro acabou?

Tabela 4: Os códigos e o significado das variáveis utilizadas na Regressão Logística e Floresta Aleatória.

Nome das Variáveis	**Categorias**
Região Geográfica	Centro-Oeste Nordeste Norte Sudeste Norte
No seu domicílio há problema de madeira das janelas, portas ou assoalhos deteriorados?	Sim Não
No seu domicílio há problema de fumaça, mau cheiro, barulho ou outros problemas ambientais causados pelo trânsito ou indústria?	Sim Não
Insegurança Alimentar**	Segurança alimentar Insegurança alimentar leve Insegurança alimentar moderada Insegurança alimentar grave

[0]** A variável insegurança alimentar é medida através de uma inferência a partir de outras variáveis.

3.2 Procedimento de análise

Primeiramente, foi feita uma Análise descritiva com as variáveis que possuem caráter social, para observar as diferenças entre o grupo de beneficiários e o de não beneficiários. Em seguida, selecionou-se os domicílios urbanos e com renda total do domicílio de até R$ 2.000,00, pois o auxílio do Bolsa Família está vinculado à renda. A finalidade é analisar o perfil do domicílio beneficiário com renda parecida ao do domicílio não beneficiário e observar suas diferenças e semelhanças.

O passo seguinte foi a realização de uma Análise Fatorial Exploratória com as variáveis binárias, ou seja, as que só existem duas opções de resposta (sim ou não) do questionário sobre as condições de vida. A partir do resultado obtido, foi feita uma comparação entre os 3 fatores obtidos, seguida de uma comparação com as variáveis que não foram utilizadas, para verificar como os fatores se comportam.

Posteriormente, foi realizada uma comparação das médias em cada variável utilizada na Análise Fatorial e entre os grupos de domicílios que recebem o benefício e aqueles que não o recebem. Foi efetuada, também, uma análise com outras variáveis que não foram utilizadas no cálculo da Análise fatorial, como Insegurança Alimentar, Região Geográfica, Instrução e o Tipo de Trabalho do provedor, para analisar o desempenho do resultado obtido e observar se há diferenças.

Em complementação foi feita a Regressão Logística, na qual foi observada uma grande diferença nas proporções dos domicílios que são beneficiários e aqueles que não são beneficiários dentre o conjunto de domicílios selecionado. Na Tabela 5 pode-se observar as respectivas proporções em uma amostra de tamanho correspondente a 12153 domicílios.

Tabela 5: Tabela com a distribuição relativa dos domicílios beneficiários e não beneficiários.

Grupo	Proporção
Não beneficiário	0,80
Beneficiário	0,20

Ainda sobre a Tabela 5, tem-se a proporção de 1/5 de domicílios beneficiários. Adianta-se que o objetivo dessa pesquisa é avaliar o perfil do beneficiário do PBF. A Regressão Logística pode não ter uma consistência adequada em virtude da pequena quantidade relativa de domicílios que são beneficiários.

Assim, o banco de dados foi separado em dois grupos. Em relação ao grupo dos domicílios que recebem o benefício, estes foram repartidos de forma aleatória em 15 partes resultando em 14 amostras aleatórias simples de tamanho 160 e 1 de tamanho 159. Quanto ao outro grupo, foram selecionadas 15 amostras aleatórias de tamanho 160 e, em seguida, cada amostra foi agregada com a outra, formando 14 amostras de tamanho 320 e uma de tamanho 319; de forma que todos domicílios beneficiários sejam contemplados nessa seleção de amostras.

Segue a Equação 3.2.1 da Regressão Logística utilizada nas 15 amostras obtidas:

$$\begin{aligned} logito(\pi) = \beta_0 &+ \beta_1 \text{Região}_{\text{Nordeste}} + \beta_2 \text{Região}_{\text{Norte}} + \beta_3 \text{Região}_{\text{Sudeste}} + \beta_4 \text{Região}_{\text{Sul}} \\ &+ \beta_5 \text{Moradia}_{\text{Sim}} + \beta_6 \text{Bairro}_{\text{Sim}} \\ &+ \beta_8 \text{Inseg Alimen}_{\text{leve}} + \beta_9 \text{Inseg Alimen}_{\text{moderada}} + \beta_{10} \text{Inseg Alimen}_{\text{grave}}, \end{aligned} \tag{3.2.1}$$

as variáveis Região e Insegurança Alimentar não foram utilizadas na seção da Análise Fatorial, mas as demais variáveis foram incluídas no cálculo.

Ademais, na aplicação da Regressão Logística foi feito um balanceamento amostral entre o grupo de beneficiários e o de não beneficiários e, diante disso, foi necessário uma correção no cálculo do intercepto que é o Método de Correção a Priori. Este método envolve a modelagem da Regressão Logística de maneira padrão, calculando os estimadores através da máxima verossimilhança. No entanto, simultaneamente, essas estimativas são ajustadas com base em informações prévias ou conhecimentos anteriores.

De acordo com esse método, os estimadores de máxima verossimilhança $\hat{\beta}_i$ do modelo são consistentes e eficientes. Porém, para garantir que o estimador de β_0 também seja consistente, é necessário corrigi-lo usando a seguinte fórmula:

$$\hat{\beta}_0 - \ln\left(\frac{1-\tau}{\tau} \cdot \frac{y}{1-y}\right), \tag{3.2.2}$$

onde τ representa a proporção de sucessos na população e y a proporção de sucessos na amostra.

Após a criação das amostras, foram feitas as respectivas Regressões Logísticas para continuar a comparação dos grupos e também foram analisados os valores dos β, o impacto de cada variável e quais variáveis são relevantes para o modelo. Para a construção desse modelo foi utilizado o Método de Correção a Priori no cálculo do intercepto. Essa correção foi utilizada porque as amostras foram balanceadas.

A amostra, dentre as 15 que foram anteriormente criadas, que teve o melhor resultado no cálculo da acurácia foi a escolhida para fazer o diagnóstico na regressão e de seus respectivos resíduos.

Em seguida, para a construção da Floresta Aleatória foram agregadas as 15 amostras balanceadas em uma única amostra, e foi feita a análise e a comparação dos dois métodos de classificação - Regressão Logística e Floresta Aleatória.

, que mede a probabilidade de uma instância qualquer ser escolhida aleatoriamente. Quanto menor o valor, maior o grau de pureza daquele nó. Atinge seu mínimo (zero) quando todos os casos no nó caem em uma única categoria de destino.

Acrescenta-se ainda, que o tipo de Floresta Aleatória utilizada nesse estudo é a classificação e foram feitas 1000 Árvores Aleatórias. Para o diagnóstico do modelo foi utilizada a figura do Índice de Impureza de Gini, a taxa de erro estimada OOB e a Floresta Aleatória teve apenas 2 nós.

4 Resultados

Nesta seção são apresentados os resultados obtidos neste estudo.

4.1 Análise Descritiva

No primeiro momento foi efetuada a Análise descritiva para comparar o perfil dos grupos dos domicílios beneficiários e dos não beneficiários do PBF.

Tabela 6: Comparando os beneficiários e os não beneficiários em relação à categoria de trabalho.

	Beneficiários	Não Beneficiários
Conta própria	40,30%	31,20%
Setor privado	33,20%	45,20%
Setor público	3,20%	6,20%
Trabalhador doméstico	19,70%	14,30%
Militar do exército	0,05%	0,10%
Empregador	0,28%	0,50%
Trabalhador não remunerado	3,20%	2,50%

As modalidades de trabalho possuem algumas diferenças entre os dois grupos e as maiores diferenças estão entre "Conta própria"e "Setor privado". A Tabela 7 seguinte trata sobre a renda familiar e revela diferenças em todas as medidas, exceto no valor mínimo.

Tabela 7: Estatística resumo da renda familiar, em reais, entre beneficiários e os não beneficiários.

	Mín	1º Qu.	Mediana	Média	3º Qu.	Máx.
Beneficiário	84	818	1236	1209	1603	1999
Não Beneficiário	0	1144	1449	1385	1706	2000

Pode-se perceber que, devido ao benefício do PBF, nenhum domicílio beneficiário possui renda igual a zero, bem como pode-se constatar que, com exceção do valor mínimo, todas as medidas são inferiores no grupo dos beneficiários.

Tabela 8: Estatística resumo da renda *per capita* disponível, em reais, entre beneficiários e os não beneficiários.

	Mín	1º Qu.	Mediana	Média	3º Qu.	Máx.
Beneficiário	-1,50	12,05	17,38	19,43	24,32	85,94
Não Beneficiário	-15,27	4,22	6,44	8,03	10,71	93,25

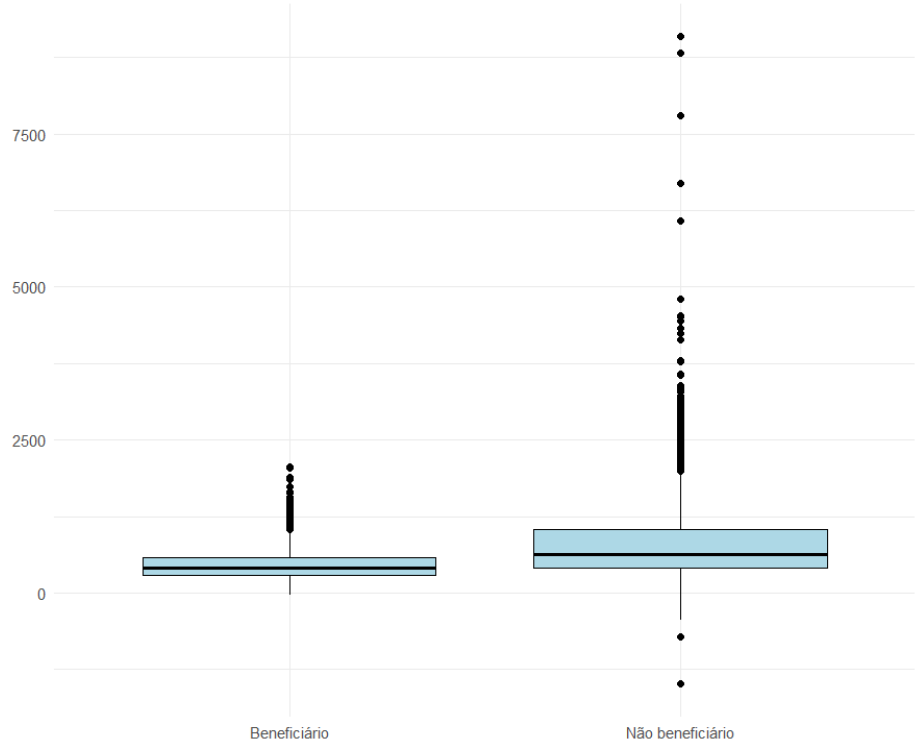

Figura 1: Boxplot comparando a renda *per capita* disponível entre os beneficiários e os não beneficiários.

A renda *per capita* disponível é, segundo o IBGE, a soma dos rendimentos monetários e não monetários, menos impostos diretos, contribuições sociais e outras deduções compulsórias ou quase compulsórias. Isso justifica o por quê desses valores negativos da Tabela 8. Os beneficiários possuem maiores valores de renda *per capita* em relação aos não beneficiários, com exceção do valor máximo. Porém, na Tabela 8, os valores correspondentes aos beneficiários são menores com exceção do valor mínimo.

Tabela 9: Distribuição relativa dos beneficiários e dos não beneficiários em relação à preocupação dos moradores com alimentos acabarem nos últimos 3 meses.

	Sim	Não
Beneficiário	66,4%	33,6%
Não Beneficiário	48,5%	51,5%

Com base na análise da Tabela 9 e no teste de associação para comprovar se a diferença entre os dois grupos é realmente significativa obteve-se o resultado de um p-valor ≈ 0, ou seja, pode-se afirmar que há diferença quanto à preocupação da falta de alimentos até o fim do mês entre os grupos de beneficiários e dos não beneficiários.

Tabela 10: Distribuição relativa dos beneficiários e dos não beneficiários em relação à existência de carteira assinada.

	Sim	Não
Beneficiário	16,6%	83,4%
Não Beneficiário	45,0%	55,0%

Tabela 11: Distribuição relativa dos beneficiários e dos não beneficiários em relação à contribuição na previdência social.

	Sim	Não
Beneficiário	4,3%	95,7%
Não Beneficiário	7,5%	92,5%

Ao observar as Tabelas 10 e 11 e ao realizar o teste de associação, pode-se constatar que, em ambos os resultados, obteve-se o p-valor ≈ 0, ou seja, existe uma associação entre os grupos. Na Tabela 10 há uma associação entre ser beneficiário e ter carteira assinada, e na Tabela 11 confirma-se uma associação entre ser beneficiário e não contribuir para a previdência social.

Tabela 12: Distribuição relativa dos domicílios entrevistados por Região Geográfica brasileira.

Região Geográfica	
Centro-Oeste	7,4%
Nordeste	53,2%
Norte	15,4%
Sudeste	17,5%
Sul	6,5%

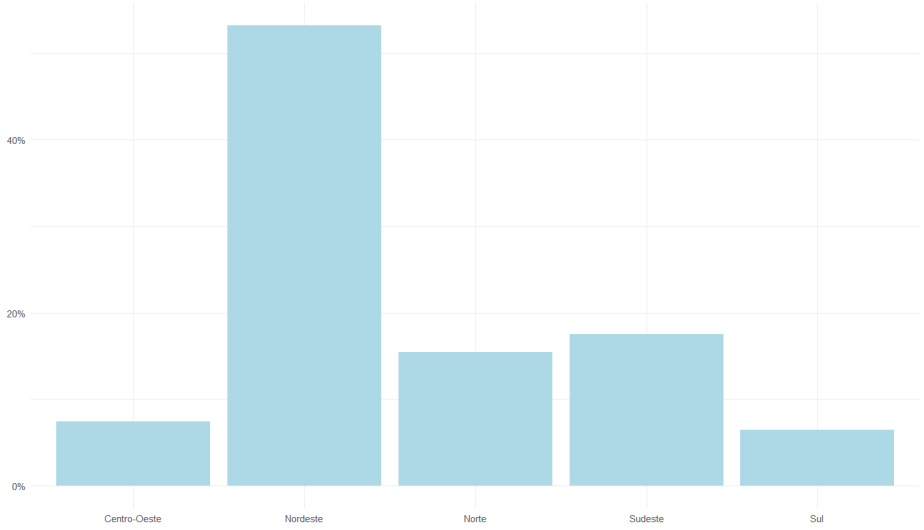

Figura 2: Histograma comparando as proporções dos domicílios por Região Geográfica.

Ao observar a Tabela 12 e a Figura 2 verifica-se que mais de 50% dos domicílios estão na região Nordeste e, nas demais regiões, os domicílios se distribuem em proporções semelhantes. Pode-se perceber que as regiões Centro-Oeste e Sul foram as que tiveram as menores proporções de domicílios.

4.2 Análise Fatorial

A primeira etapa, após a Análise exploratória, foi a de calcular as correlações policóricas entre as variáveis e na Figura 3 seguinte pode-se observar o resultado:

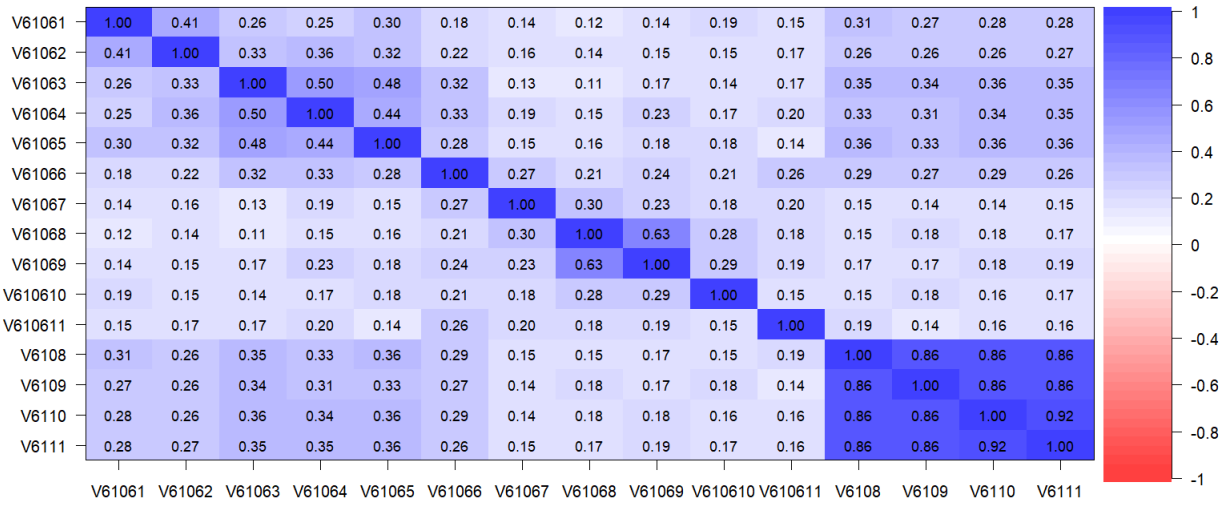

Figura 3: Gráfico de correlação tetracórica das variáveis.

É possível observar que não há correlação negativa entre as variáveis e que àquelas mais correlacionadas entre si são as quatro últimas, pelo fato de as perguntas do questionário serem bem semelhantes, o que implica no aumento das chances de terem tido respostas semelhantes para as quatro perguntas. Em seguida foi calculada os valores de cada variável e suas médias para os respectivos grupos.

Em seguida, foi feita uma Análise Fatorial com três fatores e o resultado obtido foi comparado com os grupos dos beneficiários e dos não beneficiários do PBF. E também observou-se as variáveis que mais impactam nos respectivos fatores.

Tabela 13: Resultado da Análise Fatorial e suas respectivas cargas fatoriais.

Pergunta	Código	Fator 1	Fator 2	Fator 3
No seu domicílio há problema de pouco espaço?	V61061	0,18	0,40	0,12
No seu domicílio há problema de casa escura, com pouca iluminação natural?	V61062	0,14	0,49	0,13
No seu domicílio há problema de telhado com goteira?	V61063	0,19	0,66	
No seu domicílio há problema de fundação, paredes ou chão úmidos?	V61064	0,16	0,65	0,15
No seu domicílio há problema de madeira das janelas, portas ou assoalhos deteriorados?	V61065	0,21	0,59	0,13
No seu domicílio há problema de mosquitos ou outros insetos, ratos, etc.?	V61066	0,16	0,40	0,24
No seu domicílio há problema de fumaça, mau cheiro, barulho ou outros problemas ambientais causados pelo trânsito ou indústria?	V61067		0,20	0,33
No seu domicílio há problema de estar localizado próximo a rio, baía, lago, açude ou represa poluídos?	V61068			0,84
No seu domicílio há problema de estar localizado em área sujeita a inundação?	V61069		0,14	0,74
No seu domicílio há problema de estar localizado em encosta ou área sujeita a deslizamento?	V610610		0,19	0,34
No seu domicílio há problema de violência ou vandalismo na sua área de residência?	V610611		0,24	0,22
Nos últimos três meses, os moradores deste domicílio tiveram a preocupação de que os alimentos acabassem antes de poderem comprar ou receber mais comida?	V6108	0,86	0,28	0,10
Nos últimos três meses, os alimentos acabaram antes que os moradores deste domicílio tivessem dinheiro para comprar mais comida?	V6109	0,87	0,24	0,12
Nos últimos três meses, os moradores deste domicílio ficaram sem dinheiro para ter uma alimentação saudável e variada?	V6110	0,91	0,27	0,12
Nos últimos três meses, os moradores deste domicílio comeram apenas alguns poucos tipos de alimentos que ainda tinham porque o dinheiro acabou?	V6111	0,91	0,27	0,12

Pode-se perceber, a partir da Tabela 13, que as variáveis V6108, V6109, V6110 e V6111 estão presentes nos três fatores. No primeiro fator as maiores cargas fatoriais foram compostas por variáveis sobre insegurança alimentar. Para o segundo fator quase todas as variáveis foram utilizadas com exceção da V61068 e as cargas fatoriais mais altas estão relacionadas às condições do domicílio. Em seguida, na construção do fator 3 também foram utilizadas quase todas as variáveis com exceção da V61063 e as maiores cargas fatoriais foram as variáveis sobre a localidade do domicílio e sua respectiva questão sanitária.

Segue a Tabela 14 com o somatório ao quadrado das cargas fatoriais dos respectivos fatores e o quanto essa Análise Fatorial explica sobre a variância do banco de dados que corresponde aproximadamente a 50%. O fator 1 é o que possui as maiores cargas fatoriais e a maior taxa de explicação da variância e esse valor indica como as variáveis sobre insegurança alimentar possuem forte impacto nesse banco de dados. Pode-se afirmar que o resultado da explicação da variância foi razoável, mas o banco de dados tem muita variabilidade considerando que a amostra é composta por domicílios de todo o país.

Tabela 14: Resultados da Análise Fatorial.

Fator	**Fator 1**	**Fator 2**	**Fator 3**
SS loadings	3,36	2,22	1,71
Proporção da variância	0,22	0,15	0,11
Variância acumulada	0,22	0,37	0,49

Em seguida, foi feita uma uma comparação da média dos três fatores obtidos na Análise Fatorial comparando os grupos dos beneficiários e dos não beneficiários do PBF.

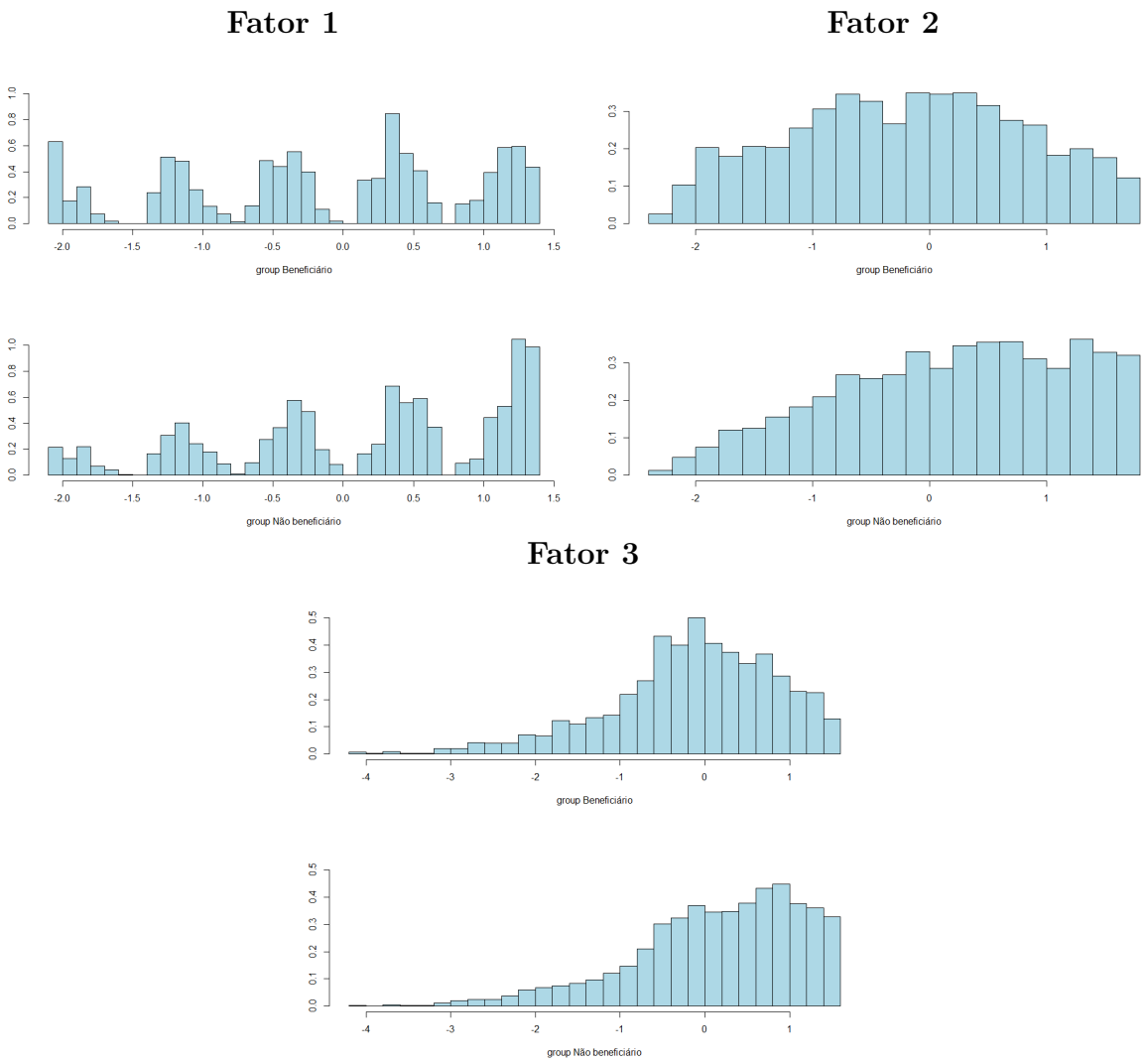

Figura 4: Gráfico comparando o resultado da média dos três fatores dos grupos de beneficiários e não beneficiários.

A partir do resultado da Figura 4, percebe-se no primeiro gráfico que representa o Fator 1 que a maior frequência no grupo dos beneficiários está entre 0,0 e 0,5, enquanto no grupo dos não beneficiários está entre 1 e 1,5. O que mostra que dentre as variáveis utilizadas neste fator, o grupo dos beneficiários respondeu "Não"mais vezes do que o outro grupo, isso justifica uma maior frequência nos valores mais altos.

No Fator 2, a distribuição no grupo dos não beneficiários aparenta ser uma Normal e seus valores majoritariamente então em torno de zero, enquanto no grupo dos não beneficiários estão em torno de 0 e 1. No último gráfico, nos dois grupos existem poucos valores baixos, porém no grupo dos beneficiários, também existe uma concentração em torno de zero.

Constata-se também uma sobreposição na distribuição nos três gráficos, mostrando que não tem muita diferença na média do cálculo das médias entre os grupos e

que existe uma grande interseção entre os resultados.

Para analisar a consistência dos fatores, foram feitas algumas comparações com variáveis que não foram utilizadas no cálculo da Análise Fatorial.

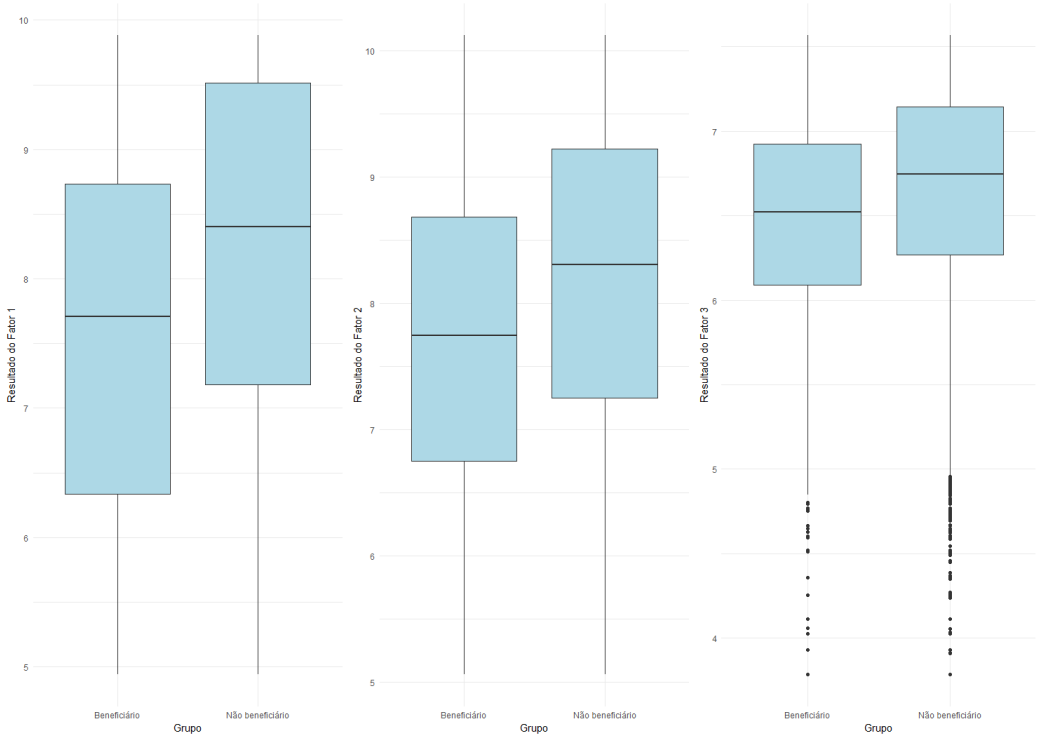

Figura 5: Boxplot comparando o resultado da Análise Fatorial nos grupos dos beneficiários e dos não beneficiários.

Na Figura 5, é o cálculo dos respectivos fatores entre o grupo dos beneficiários e dos não beneficiários e verifica-se que a mediana nos três gráficos é levemente superior no grupo dos não beneficiários. O primeiro gráfico é o que possui uma maior diferença, entre os grupos, indicando que a maior diferença seja em relação a questões sobre insegurança alimentar do que aos problemas de domicílio (Fator 2) e da localidade do domicílio e sua respectiva questão sanitária (Fator 3). O terceiro gráfico e o segundo gráfico possuem diferenças menores entre os grupos o que pode indicar uma diferença na qualidade de vida, porém esta diferença não é muito expressiva.

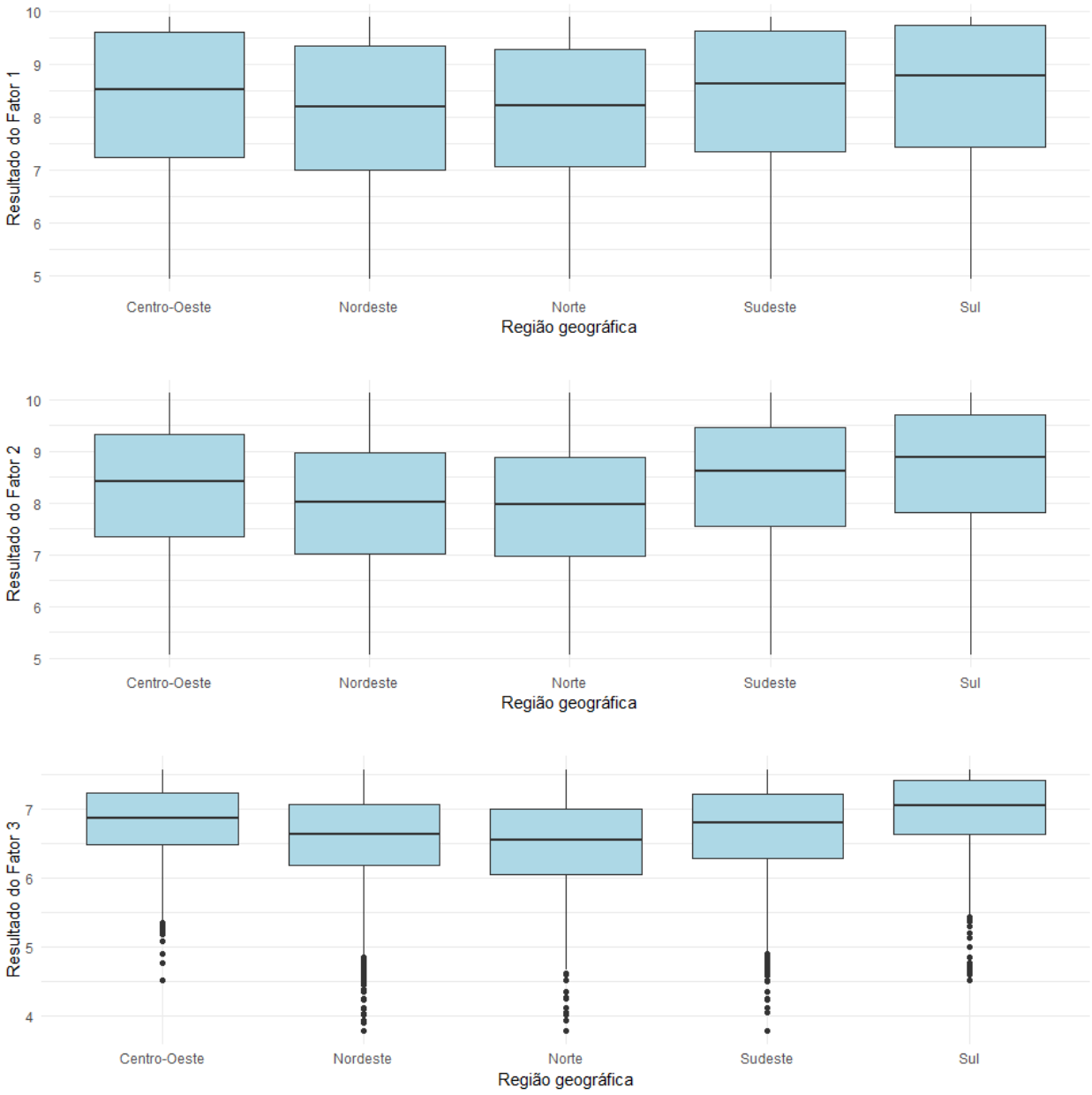

Figura 6: Gráficos comparando o desempenho dos fatores na variável Região Geográfica.

A Figura 6 apresenta o resultado nas respectivas Regiões Geográficas e observa-se que a região Norte e a Nordeste tiveram os menores valores, enquanto a região Sul teve o maior nos três gráficos. Ademais, segundo os dados do SENARC (2018), a região brasileira que mais recebe benefício em quantidades absolutas é a região Nordeste e as regiões menos beneficiadas são a Centro-Oeste e a Sul.

Tabela 15: Nome das categorias da variável Trabalho.

Código	Categoria
1	Conta própria
2	Setor privado
3	Setor público
4	Trabalhador doméstico
5	Militar do exército
6	Empregador
7	Trabalhador não remunerado

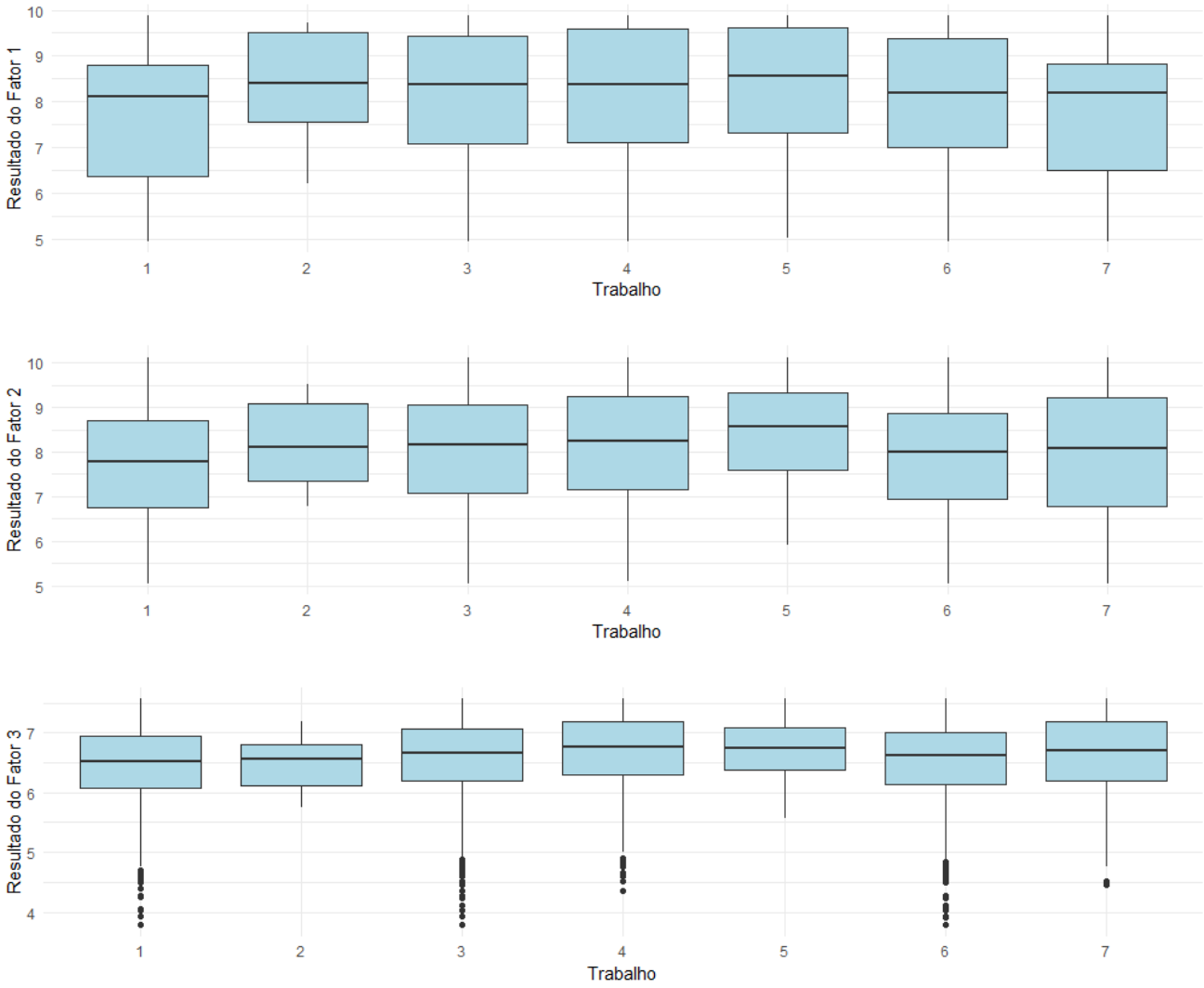

Figura 7: Gráficos comparando o desempenho dos fatores na variável Trabalho.

A Figura 7 refere-se à comparação do resultado dos fatores da Análise Fatorial em relação à variável Trabalho e verifica-se que em relação à mediana não teve uma diferença

expressiva entre as categorias nos três casos, mas o resultado na categoria Empregador (6) e Conta Própria (1) aparecem um pouco abaixo das demais nos três casos.

Tabela 16: Nome das categorias da variável Insegurança Alimentar.

Código	Categoria
1	Segurança
2	Insegurança leve
3	Insegurança moderada
4	Insegurança grave

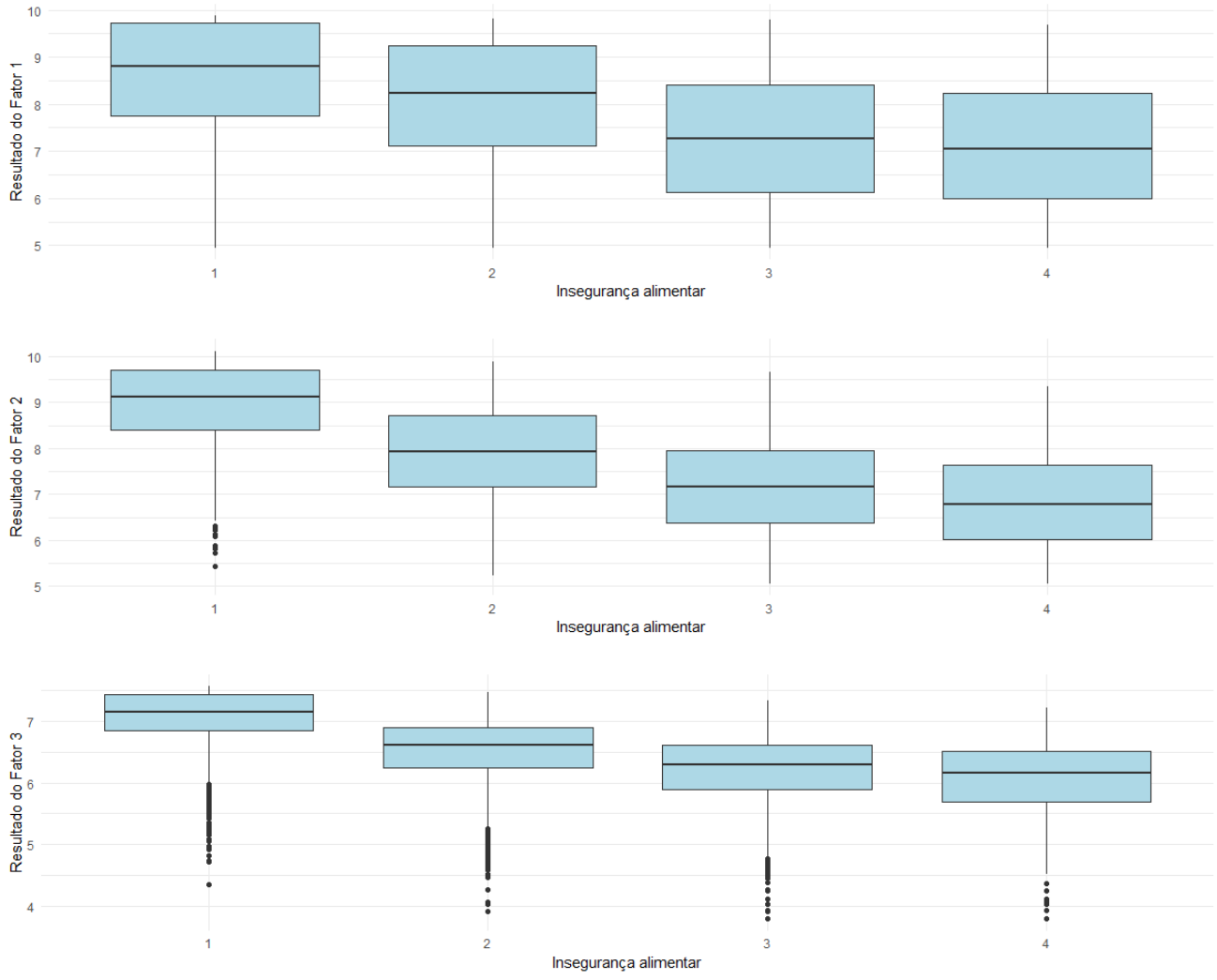

Figura 8: Gráficos comparando o desempenho dos fatores na variável Insegurança Alimentar.

Em seguida, tem-se a Figura 8 que trata da variável Insegurança Alimentar. O resultado é bem acurado, nos três gráficos, e quanto maior o grau de insegurança alimentar menor foi o resultado nos três fatores. Porém, o Fator 1 da Tabela 13 demonstra que as

maiores cargas fatoriais foram em perguntas relacionadas à essa variável e nas categorias insegurança moderada (3) e grave (4) obteve-se resultados aproximados.

Tabela 17: Nome das categorias da variável Instrução.

Código	Categoria
1	Sem instrução
2	Ensino Fundamental Incompleto
3	Ensino Fundamental Completo
4	Ensino Médio Incompleto
5	Ensino Médio Completo
6	Ensino Superior Incompleto
7	Ensino Superior Completo

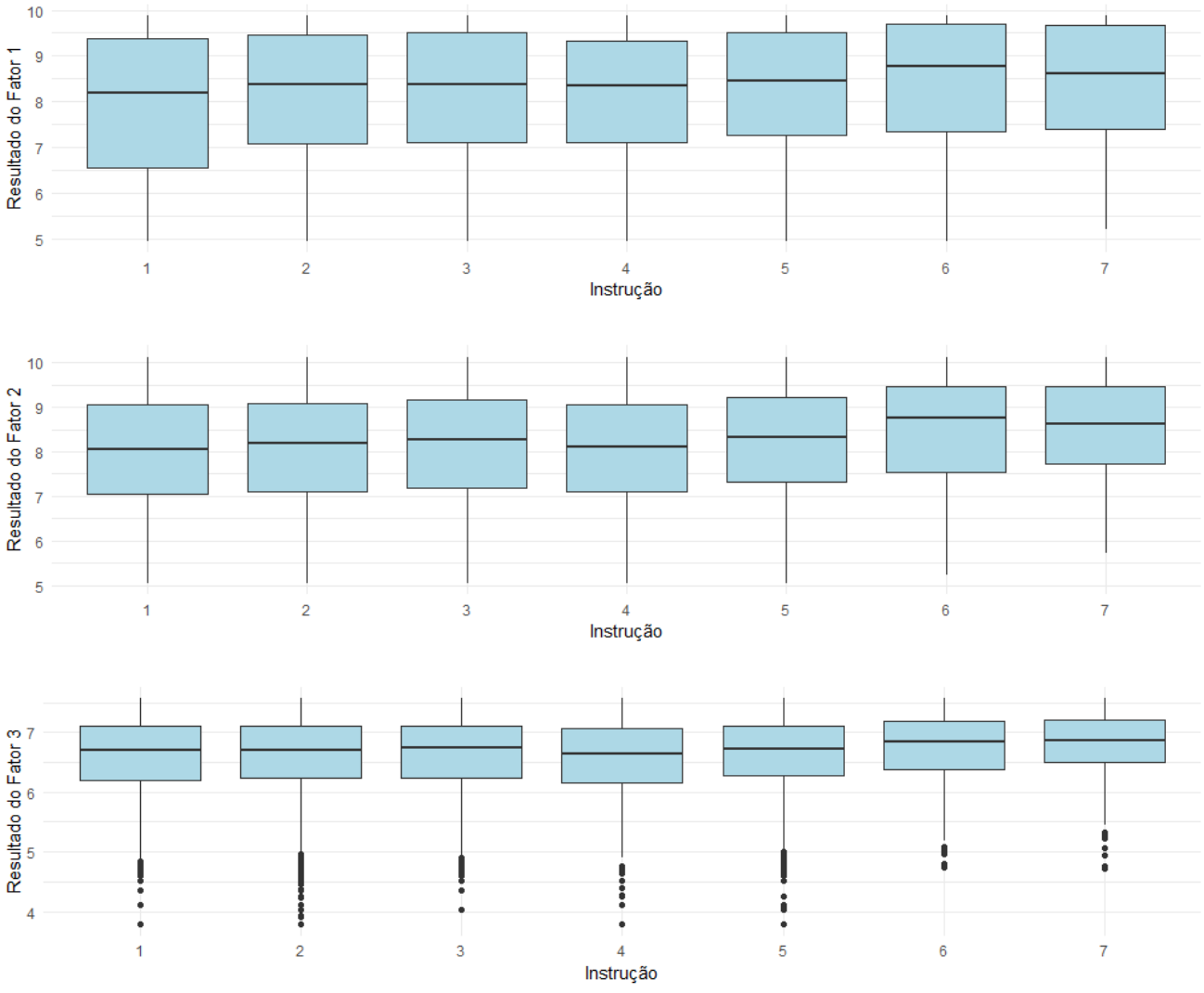

Figura 9: Gráficos comparando o desempenho dos fatores na variável Instrução.

Por último, a Figura 9 mostra o comportamento da variável Instrução do provedor do domicílio e observa-se que a diferença entre os boxplots é sutil. No entanto, a categoria Ensino Médio Incompleto (4) possui uma mediana abaixo das demais, enquanto Ensino Superior Incompleto (6) está levemente acima das demais nos três gráficos.

Após essa análise dos gráficos, foram realizados diversos testes de Análise de Variância para verificar se existe alguma média diferente nas respectivas categorias.

Tabela 18: Resultados da Análise de Variância para as variáveis de Trabalho, Região Geográfica, Insegurança Alimentar e Instrução.

Variável	**Fator 1**	**Fator 2**	**Fator 3**
Trabalho	F = 11,0 $P = 2,7 \times 10^{-10}$	F = 16,0 $P = 3,0 \times 10^{-12}$	F = 12,6 $P = 3,0 \times 10^{-14}$
Região	F = 54,5 $P = 2,0 \times 10^{-16}$	F = 101,5 $P = 2,0 \times 10^{-16}$	F = 79,7 $P = 2,0 \times 10^{-16}$
Insegurança Alimentar	F = 783,9 $P = 2,0 \times 10^{-16}$	F = 952,3 $P = 2,0 \times 10^{-16}$	F = 1587 $P = 2,0 \times 10^{-16}$
Instrução	F = 12,0 $P = 2,0 \times 10^{-13}$	F = 10,4 $P = 1,4 \times 10^{-11}$	F = 4,3 $P = 2,5 \times 10^{-4}$

É possível perceber que nas quatro variável dos três fatores o p-valor foi pequeno, aproximadamente zero, o que nos indica que há evidências de rejeição da hipótese nula, ou seja, existe pelo menos uma categoria que a média é diferente. Assim, pode-se afirmar que nem todas as médias são iguais e que, de fato, existe uma diferença no resultado da Análise Fatorial nas variáveis analisadas.

Na variável Trabalho, pode-se observar na Figura 7 que no Fator 1 as categorias Conta Própria (1) e Trabalho Não Remunerado (7) são as que mais diferem entre as demais. Enquanto no Fator 2 é a categoria é a Militar do Exército (5) e no Fator 3 é a Trabalhador doméstico (4).

Na Figura 6 sobre a variável Região, é possível observar que nos três Fatores as regiões Norte e Nordeste tiveram um resultado abaixo das demais regiões. A Figura 8 sobre Insegurança Alimentar, quanto mais grave for a insegurança menor é o resultado nos três fatores.

Na variável Instrução é possível perceber na Figura 9 que nos Fatores 1 e 2 a categoria Ensino Superior Incompleto (6) é levemente diferente das demais. Enquanto no Fator 3 as categorias estão mais consistentes, mas a categoria Ensino Médio Incompleto (4) teve um resultado um pouco menor que as demais.

4.3 Regressão Logística

Foram criadas 15 amostras aleatórias sem reposição com proporções desbalanceadas sendo 80% extraídas do grupo dos não beneficiários e 20% do grupo dos beneficiários. Também foram feitas as 15 amostras balanceadas com a mesma proporção, ou seja, 50% de cada grupo. Em seguida, foram analisados os resultados da Regressão Logística para cada grupo e abaixo seguem os gráficos dos resultados de cada β obtido nas respectivas amostras.

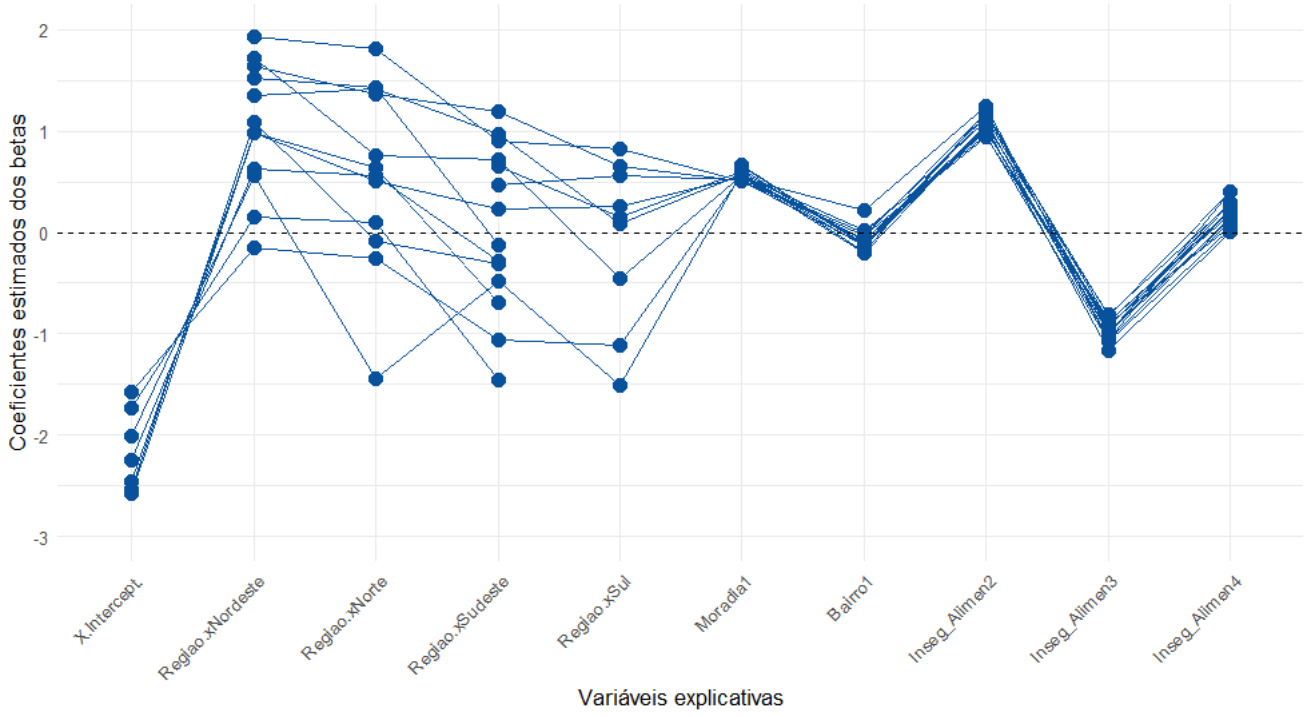

Figura 10: Gráfico em linha dos valores dos β da Regressão Logística em cada amostra não balanceada.

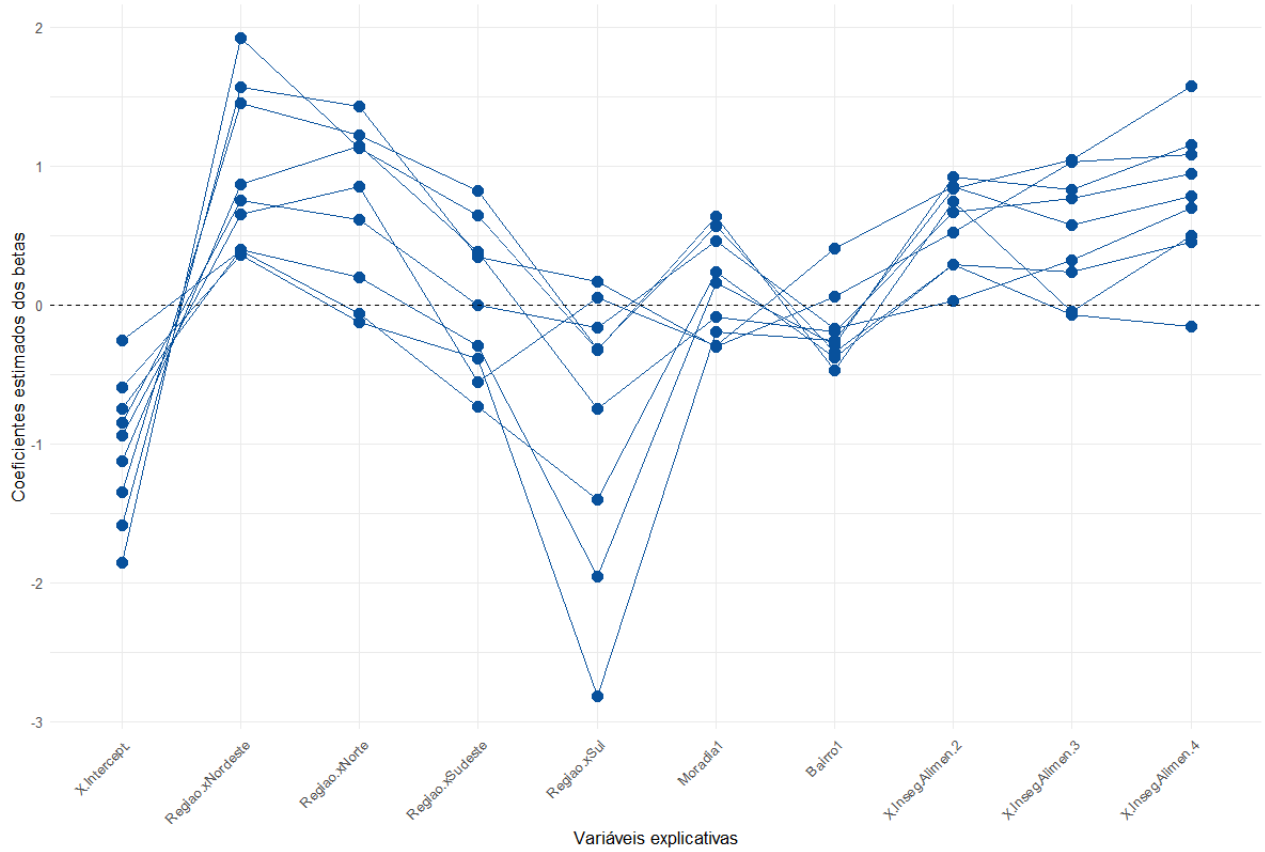

Figura 11: Gráfico em linha dos valores dos β da Regressão Logística em cada amostra balanceada.

Ao analisar a Figura 10 com a mesma escala da Figura 11, constata-se que algumas estimativas foram perdidas para manter a escala. Pode-se evidenciar que a Figura 10 possui estimativas menos consistentes, principalmente na variável Região. As demais variáveis parecem mais consistentes e com estimativas próximas entre as amostras, enquanto que na Figura 11 evidencia-se uma diferença maior nas estimativas. Apresenta-se, a seguir, a Tabela 19 com as proporções dos β maiores que zero referentes à amostra balanceada da Figura 11.

Tabela 19: A proporção de β que são não nulos e positivos dentre as 15 amostras.

Variável	**Nome da variável**	**Proporção de $\beta > 0$**
Intercept	Intercepto	0,00
Regiao.xNordeste	Nordeste	1,00
Regiao.xNorte	Norte	0,78
Regiao.xSudeste	Sudeste	0,44
Regiao.xSul	Sul	0,22
Moradia1	No seu domicílio há problema de madeira das janelas, portas ou assoalhos deteriorados	0,56
Bairro1	No seu domicílio há problema de fumaça, mau cheiro, barulho ou outros problemas ambientais causados pelo trânsito ou indústria	0,22
Inseg Alimen.2	Insegurança alimentar leve	1,00
Inseg Alimen.3	Insegurança alimentar moderada	0,78
Inseg Alimen.4	Insegurança alimentar grave	0,89

Pode-se evidenciar uma grande variação na variável Região, principalmente na região Sul, indicando uma baixa consistência em relação às demais variáveis. A região de referência é a Centro-Oeste e observando o gráfico da Figura 11 e a Tabela 19 pode-se constatar que a região Sul, em apenas 3 amostras tiveram β positivos, o que mostra que na maioria das amostras houve a diminuição do resultado da regressão, indicando que os domicílios da região Sul tem menor chance de receber o benefício em relação aos domicílios da região Centro-Oeste.

Em contraposição, nas regiões Norte e Nordeste, em mais de 75% dos β são positivos, o que indica que nessas regiões há maior chance dos domicílios serem beneficiários em relação aos domicílios situados no Centro-Oeste. Vale ressaltar que a região Nordeste teve todos os β positivos e não nulos, ou seja, em relação à região Centro-Oeste, todas as amostras da região Nordeste apresentam maiores chances de terem domicílios beneficiários.

Em relação às demais variáveis da regressão, Moradia1 ("No seu domicílio há problema de madeira das janelas, portas ou assoalhos deteriorados") mais de 50% dos β tiveram resultados maiores que zero, indicando que a presença desses problemas aumenta a chance do domicílio receber o benefício.

Quanto a variável Bairro1 ("No seu domicílio há problema de fumaça, mau cheiro, barulho ou outros problemas ambientais causados pelo trânsito ou indústria"), esta indica

que a presença de problemas no bairro diminui a probabilidade do domicílio ser beneficiário do PBF. Sendo a única variável que possui majoritariamente os β nulos ou negativos.

A variável Insegurança Alimentar possui uma proporção alta de β positivos em todas as categorias. Considerando que o valor de referência da Insegurança Alimentar é igual a 1, que significa Segurança Alimentar. Ou seja, qualquer grau de insegurança alimentar aumenta a chance do domicílio ser beneficiário.

Ademais, as amostras tem um comportamento consistente entre as regressões obtidas. Para o diagnóstico da Regressão Logística foi escolhida a amostra com a maior acurácia na matriz de confusão, ou seja, a matriz que tem a maior proporção de acerto ao tentar classificar os domicílios em beneficiários ou não beneficiários.

O gráfico da Figura 12 refere-se à Regressão Logística da amostra 10 e está ajustado dentro dos limites do intervalo de confiança e não tem nenhuma observação fora dos limites.

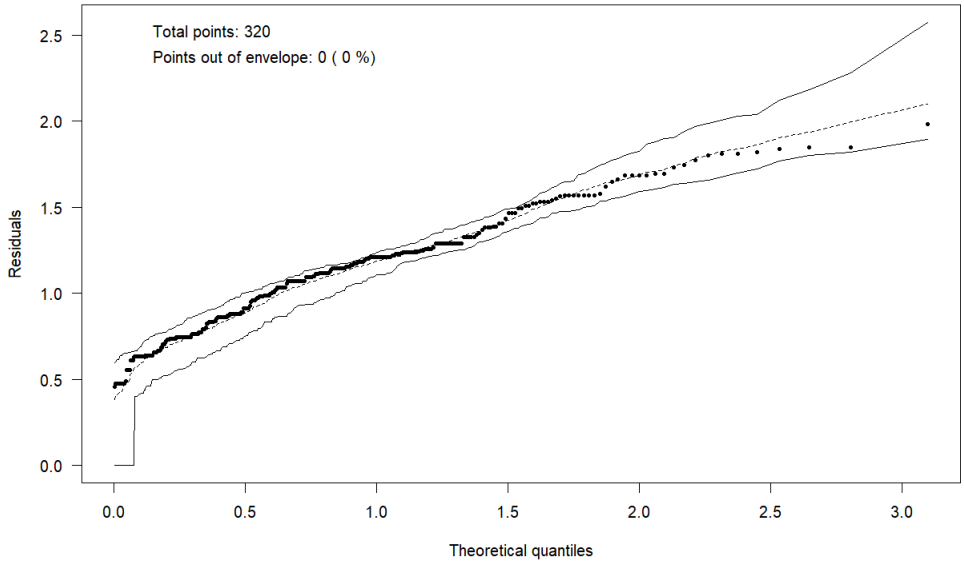

Figura 12: Gráfico envelope da Regressão Logística.

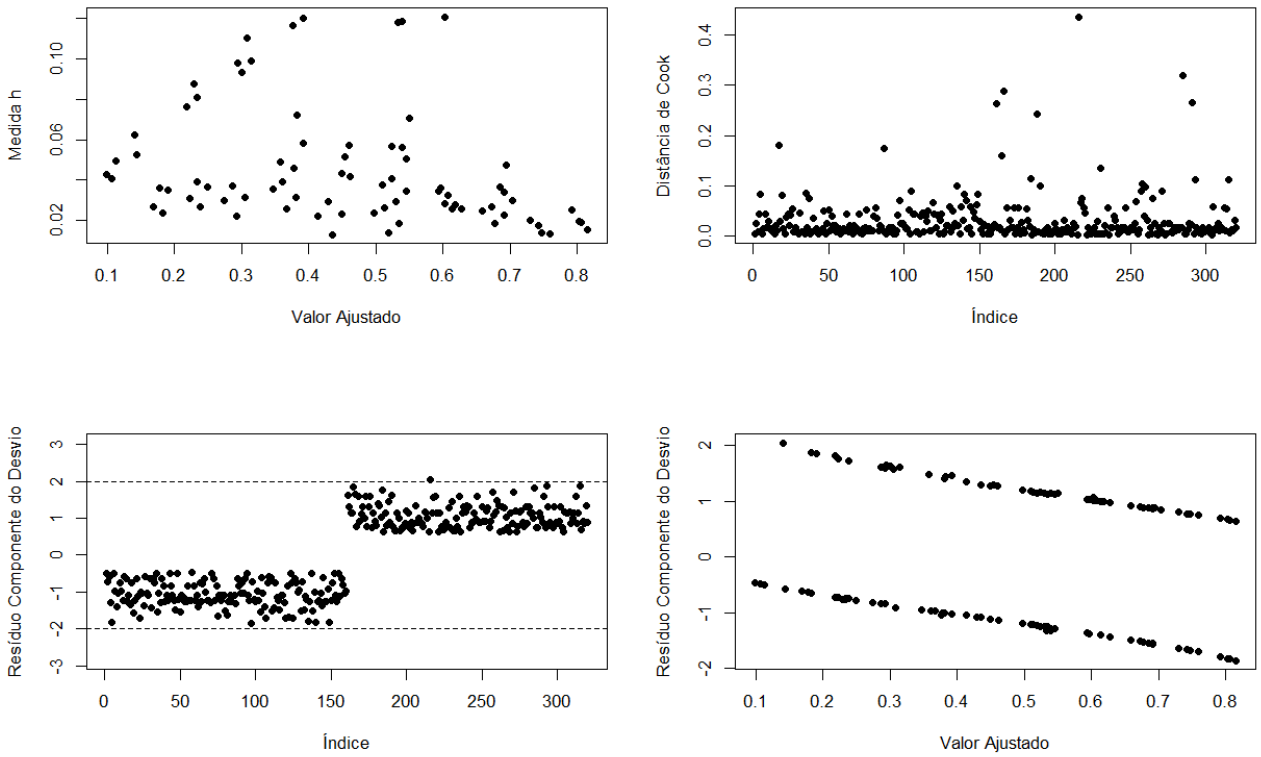

Figura 13: Gráfico de resíduos da Regressão Logística.

Em relação aos resíduos demonstrados nos gráficos da Figura 13, pode-se verificar que, no primeiro gráfico os resíduos estão bem dispersos, porém nas extremidades existem menos observações e entre 0,3 e 0,6 estão as observações com os maiores valores. Ademais, não tem nenhum ponto de alavanca específico. Pela distância de Cook (segundo gráfico) existe uma observação bem maior que as demais, e há outras bem próximas desta e a amplitude do eixo y é de apenas 0,4. No terceiro gráfico tem-se os resíduos em cada grupo que aparentam estar dentro dos limites, com poucas observações na linha. O último gráfico é o valor ajustado da regressão e os respectivos resíduos, o que aparenta estar uniforme.

Na Tabela 20 encontra-se a matriz de confusão, que é calculada a partir da regressão com uma acurácia de aproximadamente 0,75 e pode-se constatar que existem algumas observações que foram preditas e classificadas no grupo errado.

Tabela 20: Tabela com a Matriz de Confusão.

		Referência	
		0	1
Predição	0	125	45
	1	35	115

Acrescenta-se ainda, que existe uma quantidade considerável de observações que foram classificadas de forma equivocada. No entanto, aproximadamente 75% das observações foram classificadas corretamente. Porém, ao observar o banco de dados, foi possível constatar que existe uma grande interseção entre os grupos, principalmente na Figura 5 que mostra o resultado da Análise Fatorial nos respectivos grupos.

4.4 Floresta Aleatória

A Figura 14 e a Tabela 20 apresentam o modelo da Floresta Aleatória para a classificação dos domicílios em beneficiários ou não beneficiários do PBF.

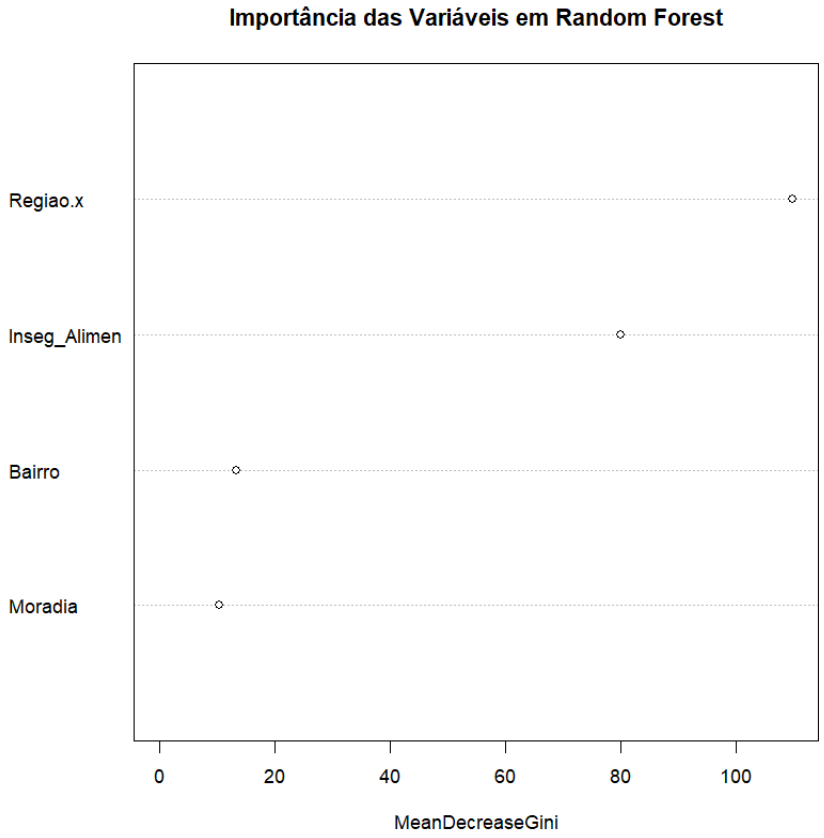

Figura 14: Gráfico indicando a importância das variáveis na Floresta Aleatória.

Tabela 21: Código das variáveis, seus respectivos significados e a importância medida pelo MeanDecreaseGini na Floresta Aleatória.

Código	**Nome da variável**	**Gini**
Região.x	Região Geográfica	110,0
Inseg_Alimen	Insegurança Alimentar	80,0
Bairro	No seu domicílio há problema de fumaça, mau cheiro, barulho ou outros problemas ambientais causados pelo trânsito ou indústria?	13,4
Moradia	No seu domicílio há problema de madeira das janelas, portas ou assoalhos deteriorados?	10,6

Ao observar a Figura 14, as variáveis mais significativas para o modelo da Floresta Aleatória foram "Região Geográfica"e "Insegurança Alimentar"e, assim como na Regressão Logística, estas foram as variáveis mais determinantes para a classificação dos domicílios em beneficiários ou em não beneficiários. Seguem as Tabelas 22 e 23 com os detalhes da Floresta Aleatória.

Tabela 22: Detalhes do Modelo Floresta Aleatória.

Detalhes do Modelo Floresta Aleatória	
Tipo de Random Forest	Classificação
Número de árvores	1000
Número de variáveis testadas por divisão	2
Taxa de erro estimada OOB	39,1%

Tabela 23: Matriz de confusão.

	0	**1**	**Taxa de erro de classe**
0	1398	1002	0,42
1	876	1522	0,37

Nesta Floresta Aleatória foram feitas 1000 árvores aleatórias. Cada nó foi dividido em apenas dois subconjuntos e a acurácia da matriz de confusão (Tabela 23) foi de aproximadamente 61%, um resultado consideravelmente menor que o da Regressão Logística da amostra 10.

Em adição, a Floresta Aleatória teve uma acurácia relativamente baixa. O valor obtido da taxa de erro estimada OOB foi alto, aproximadamente 40%. O grupo dos beneficiários teve uma taxa de erro menor que a taxa do grupo dos não beneficiários, como foi mostrado na Tabela 23. O resultado da matriz confusão indica novamente a interseção que existe entre estes dois grupos.

Ao se comparar os resultados anteriores com os resultados da Regressão Logística, pode-se afirmar que as variáveis Região Geográfica e Insegurança Alimentar tiveram os β em mais de 50% positivos e também estas foram as variáveis com maior impacto nos parâmetros da regressão. Adiciona-se, ainda, que ambos os modelos tiveram problemas de classificação, embora os dois modelos complementam. Salienta-se que a Floresta Aleatória teve um desempenho inferior na classificação dos grupos. Ademais, ao observar o resultado da Regressão Logística e da Floresta Aleatória, foi possível constatar o grande impacto que a variável Insegurança Alimentar teve para os respectivos modelos. É certo que, em ambos os modelos, a classificação teve erros na acurácia, comprovando que existe uma grande interseção entre os grupos como foi mostrado na Figura 4.

5 Conclusão

A variável com maior impacto nos modelos estatísticos utilizados neste estudo - Regressão Logística e Floresta Aleatória - foi a Insegurança Alimentar, a qual foi significativa para identificar o grupo dos domicílios em beneficiários do PBF. A Análise Fatorial, construída com três fatores - insegurança alimentar, condições do domicílio e localidade do domicílio e sua respectiva questão sanitária - complementa os modelos estatísticos e reforça a presença da insegurança alimentar no grupo dos beneficiários.

Os grupos de domicílios urbanos analisados - beneficiários e não beneficiários - com renda total de até R$2.000,00 apresentaram vulnerabilidades sociais semelhantes, havendo uma interseção nos resultados obtidos. Porém, os resultados obtidos em todas as técnicas estatísticas indicou que o grupo dos domicílios beneficiários apresentou condições de vida ainda mais vulneráveis do que o grupo dos não beneficiários. Portanto, em 2018, os resultados sinalizaram que o PBF está alcançando corretamente os domicílios alvo.

A Análise Fatorial teve uma taxa de explicação razoável e mostrou que o grupo dos domicílios beneficiários teve um resultado um pouco abaixo do grupo dos não beneficiários. Indicou, também, que os grupos não possuem tantas diferenças sociais. Em complementação, construiu-se a Regressão Logística, cujos resultados revelaram que as variáveis Região Geográfica e Insegurança Alimentar foram as que tiveram uma maior proporção de β positivos e que é algo significativo na classificação dos grupos.

Na Floresta Aleatória as duas variáveis mais significativas foram as mesmas identificadas na Regressão Logística. A partir da construção da matriz de confusão observou-se equívocos na classificação dos domicílios, porque os dois grupos são semelhantes e possuem interseções entre si.

Para o objetivo deste estudo, o banco de dados utilizado apresentou limitações, porque um dos seus objetivos incluía a percepção subjetiva da qualidade de vida. Na atual conjuntura, um número maior de domicílios são beneficiários de um valor de R$600,00, com a previsão dos adicionais para crianças, jovens e gestantes, evidenciando-se um perfil de domicílios beneficiários diferente ao que foi estudado.

Referências

AGRESTI, A. *An Introduction to Categorical Data Analysis*. 2nd. ed. Gainesville, Florida: Wiley, 2007. Department of Statistics, University of Florida.

BRAUW, A. D. et al. The impact of bolsa família on schooling. *International Food Policy Research Institute*, Washington, 2005.

CAMPELLO, T.; NERI, M. C. *Programa Bolsa Família: uma década de inclusão e cidadania*. 1a. ed.. ed. Brasília: IPEA, 2013. 494 p.

CRISTÓVÃO, D. Bolsa família começa a ser pago hoje com valor médio recorde; veja calendário. *Valor Investe*, Maio 2023. Disponível em: ⟨[https://valorinveste.globo.com/mercados/brasil-e-politica/programas-sociais/noticia/2023/05/18/calendario-bolsa-familia-mes-de-maio.ghtml]⟩.

DRASGOW, F. *Item Response Theory*. New York: Taylor & Francis, 2006.

GERMANN, C. B.; MEDEIROS, M. R. A. de. Programa bolsa familia, auxílio emergencial e auxílio brasil: a pobreza como foco programa bolsa familia, auxílio emergencial e auxílio brasil: poverty as a focus. *Brazilian Journal of Development*, v. 8, n. 6, p. 47473–47481, 2022.

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. *Statistical learning with sparsity: the lasso and generalizations*. [S.l.]: CRC press, 2015.

IBGE. *Pesquisa de Orçamentos Familiares 2017-2018*. [S.l.], 2018. Disponível em: ⟨https://www.ibge.gov.br/estatisticas/sociais/habitacao/9207-pesquisa-de-orcamentos-familiares.html⟩.

JAMES, G. et al. *An Introduction to Statistical Learning: with Applications in R*. [S.l.]: Springer, 2013.

JOSEPH, F. H. et al. *Análise multivariada de dados*. [S.l.: s.n.], 2009. E-book. ISBN 9788577805341.

KIM, J.-O.; MUELLER, C. W. *Factor analysis: Statistical methods and practical issues*. [S.l.]: sage, 1978. v. 14.

MAROCO, J. Integração do r nos menus do pasw statistics: Um exemplo de aplicação com o package'polycor'do r. *Boletim da Sociedade Portuguesa de Estatística*, Sociedade Portuguesa de Estatística, p. 71–80, 2010.

MARTINSON, M.; HAMDAN, A. An examination of performance standards. *Journal of Applied Psychology*, v. 60, n. 4, p. 345–350, 1975.

MORETTIN, P. A.; SINGER, J. M. *Estatística e Ciência de Dados*. Versão preliminar. São Paulo, SP: Departamento de Estatística, Universidade de São Paulo, 2021. Rua do Matão, 1010, São Paulo, SP 05508-090, Brasil.

ORTIZ, L. R. A.; CAMARGO, R. A. L. Breve histórico e dados para análise do programa bolsa família. *II Seminério Internacional de Pesquisa em Políticas Públicas e Desenvolvimento Social*, Volume do artigo, n. Número do artigo, p. 10, 2021.

PASE, H. L.; MELO, C. C. Políticas públicas de transferência de renda na américa latina. *Revista de Administração Pública*, v. 51, n. 2, p. 212–232, 2017.

REVELLE, W. *An introduction to psychometric theory with applications in R*. [S.l.]: Springer Evanston, IL, 2009.

SENARC, S. N. de Renda de C. *Bolsa Família Informa - Informe Nº 639*. 2018. Ministério do Desenvolvimento Social. Disponível em: ⟨https://www.gov.br/citacao-exemplo⟩.

SOUZA, P. H. G. F. d. et al. Os efeitos do programa bolsa família sobre a pobreza e a desigualdade: Um balanço dos primeiros quinze anos. *IPEA*, v. 2499, n. Número do artigo, p. 46, 2019.

STARKWEATHER, J. Factor analysis with binary items: A quick review with examples. *Benchmarks RSS Matters*, September 2014. Disponível em: ⟨http://web3.unt.edu/benchmarks/issues/2014/09/rss-matters⟩.