Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Associação entre Variáveis Socioeconômicas para Estimação da Matriz OD Utilizando Análise de Componentes Principais

por

Lucas Silva de Castro 10/0049818

Brasília

2014

Lucas Silva de Castro 10/0049818

Associação entre Variáveis Socioeconômicas para Estimação da Matriz OD Utilizando Análise de Componentes Principais

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Dr. Alan Ricardo da Silva

Brasília

2014

Dedico esse trabalho aos meus pais e meus irmãos que sempre estiveram comigo nos momentos mais importantes de minha vida e foram importantes na minha formação pessoal e acadêmica.

Lucas Silva de Castro

Agradecimentos

Agradeço primeiramente a Deus, por ter me guiado na conclusão de mais uma etapa da minha vida.

Ao professor Alan Ricardo da Silva pela orientação, dedicação e grande auxílio no desenvolvimento desse trabalho.

Ao SAS *Institute* Brasil, por possibilitar a utilização do *software* em parceria com o Departamento de Estatística da Universidade de Brasília.

Aos meus amigos, cuja a amizade e companherismo me ajudaram em diversos momentos de decisões.

Aos meus pais e meus irmãos, pelo apoio, paciência e sabedoria transmitidos à mim nos momentos importantes dessa etapa.

Resumo

A matriz de Origem-Destino (OD) é uma matriz que apresenta o fluxo de viagens entre origens e destinos (usualmente utilizadas microrregiões e UF) e sua estimativa é vital para decisões governamentais.

Este trabalho busca, através das técnicas multivariadas de componentes principais e análise fatorial, ajustar um modelo de regressão com diversas variáveis sócio-econômicas para estimar a matriz OD do Brasil. Através das componentes/fatores, esse trabalho também busca verificar se é possível agrupar as microrregiões a partir das componentes geradas e se existe diferença entre os *clusters* gerados pelas variáveis originais e pelas componentes.

Os resultados mostraram que não foi possível obter uma estimativa satisfatória para a matriz OD através das técnicas de componentes principais e análise fatorial devido à explicação de mais de 90% da variabilidade dos dados na primeira componente, pois futuras previsões são prejudicadas em consequência da difícil interpretação da componente. No entanto, os agrupamentos criados pelas variáveis originais e pelas componentes foi o mesmo. Por fim, utilizou-se o modelo gravitacional clássico com população e renda para estimativa da matriz OD.

Lista de Figuras

Sumário

Capítulo 1

INTRODUÇÃO

Segundo o MTR (2012), Transporte Rodoviário é definido como: "Transporte realizado sobre rodas nas vias de rodagem pavimentadas ou não para transporte de mercadorias e pessoas, sendo na maioria das vezes realizados por veículos automotores (ônibus, caminhões, veículos de passeio, etc.)".

Os transportes realizados pelos veículos automotores tem um ponto de origem e um destino. A matriz Origem-Destino (OD) é uma matriz que apresenta o número de viagens realizadas entre os diferentes pontos de origem-destino, onde as origens/destinos apresentados pela matriz são os municípios para onde os veículos se deslocaram. Geralmente, devido à dificuldade de coleta de dados, a matriz OD é feita para as microrregiões ou mesorregiões ou até mesmo por estados.

Existem diferentes variáveis sócioeconomicas que justificam o deslocamento de um veículo a um determinado município. A locomoção de pessoas pode ser explicada por motivos relacionadas ao trabalho, saúde, educação, lazer etc. Para o transporte de mercadorias temos variáveis relacionadas ao nível de atividades na área industrial, comercial e também pecuária.

A justificativa atribuída à locomoção de pessoas em muitos casos não é precisa,

pois lidamos em muitas ocasiões com a vontade das pessoas, tornando mais complicada a estimação da matriz OD. Já ao analisarmos a locomoção de cargas, o destino traçado para essas cargas é geralmente justificado pela sua comercialização ou a necessidade da carga no munícipio definido como destino, facilitando a estimativa da matriz OD.

Existem diversas formas de se estimar a matriz OD, entre elas, o modelo gravitacional proposto por Romanatto (2011) que explica o fluxo de viagens realizadas entre UF's do Brasil através de variáveis sócioeconômicas e a distância entre elas. Como podem existir diversas variáveis sócioeconômicas explicativas no modelo de Romanatto (2011), é provável a existência do problema de multicolinearidade, e uma forma de sanar esse problema é utilizar a técnica de componentes principais que tem como utilidade de acordo com Johnson and Wichern (2007), explicar a estrutura de variância-covariância de um conjunto de variáveis através de combinações lineares destas variáveis.

1.1 OBJETIVOS

O objetivo geral do trabalho é buscar uma associação entre variáveis socioeconomicas para a estimação da matriz origem-destino utilizando Análise de Componentes Principais.

Os objetivos específicos são:

- Verificar se é possível agrupar municípios com características similares utilizado a análise de agrupamentos, a partir das componentes geradas.

- Realizar Análise Multivariada de variáveis socioeconomicas utilizando o *software* SAS 9.2;

Capítulo 2

MATRIZ ORIGEM-DESTINO

2.1 INTRODUÇÃO

Segundo Levine (2010), a matriz Origem-Destino (OD) é uma matriz que apresenta o número de viagens realizadas entre os diferentes pontos de origem-destino. A matriz OD utiliza o número previsto de viagens oriundos de um ponto de origem e o número previsto de viagens com o fim em um determinado ponto de destino, onde a partir da matriz determinada, são realizadas comparações com a matriz OD verdadeira, afim de verificar a razoabilidade da estimação da matriz.

No presente estudo temos como origens/destinos que estruturarão a matriz OD as microrregiões, mesorregiões e os estados presentes no território brasileiro. O nível de municípios não será utilizado visto que algumas cidades não possuem ligações diretas com outras cidades, sendo o caso mais comum a concentração de serviços em cidades pólos, como por exemplo a cidade de São Paulo.

A coleta de dados relacionados à locomoção de pessoas pode ser realizadas através das agências reguladoras relacionadas a cada meio de transporte utilizado, como ANTT e ANAC. Já a coleta de dados sobre cargas pode ser feita através dos postos

de fiscalização das Secretarias de Fazenda dos estados, onde é atestado a nota fiscal da mercadoria que contém, entre outras coisas, a origem e o destino da carga.

2.2 FORMA DE OBTENÇÃO DE DADOS

Uma parte importante para a construção da matriz OD é a obtenção dos dados, para o estudo será fundamental dados referentes aos transportes de uso pessoal ou para a locomoção de mercadorias.

Existem diversas formas de obtenção de dados, onde os propostos pelo estudo realizado serão explicados a seguir.

2.2.1 Passageiro

Diariamente são realizadas diversas viagens através de empresas de viagens ,onde as viagens realizadas devem atender diversas normas como o limite de passageiros no veículo transportador. O cumprimento dessas normas são fiscalizados pela Agência Nacional de Transporte Terrestre (ANTT, 2013) e a Agência Nacional de Aviação Civil - ANAC.

Para cada viagem realizada, a agência reguladora como a ANTT (2013) solicita o cumprimento das normas como a informação de quantos passageiros estão presentes em cada viagem e a origem e o destino de cada passageiro, sendo essas informações fundamentais para o estudo.

Portanto, a melhor forma considerada pelo estudo de obtermos os dados é através da agência reguladora responsável pela fiscalização das viagens realizadas.

Caso não seja possível a obtenção desses dados, pode ser coletada uma amostra

de viagens realizadas pelas empresas de viagens, atribuindo pesos à variáveis que influenciam na quantidade de viagens realizadas, como as empresas de viagens e cidades onde são feitas essas viagens.

2.2.2 Carga

No Brasil existe uma grande polarização na produção, tanto em âmbito industrial, quanto agrícola e também pecuária. Essa polarização gera a necessidade de comercialização entre empresas alocadas em diferentes partes do país.

A comercialização dessas mercadorias levam ao transporte delas, onde caminhões saem todos os dias de um determinado estado para outro, onde será entregue a mercadoria. Muitos municípios que participam da comercialização dessas mercadorias não possuem uma ligação direta com o município qual a mercadoria foi negociada, essa dificuldade obriga a realização de escalas em cidades pólos, como por exemplo a cidade de São Paulo.

No percurso feito pelo transporte dessa carga, os veículos passam por postos de fiscalização das Secretarias de Fazenda dos estados, onde deve ser apresentada a GIA/ST, que de acordo com SEFAZDF (2013) permite ao contribuinte substituto localizado em outra UF informar as operações e/ou prestações interestaduais realizadas com contribuintes localizados no DF, ou a ausência de movimento, se for o caso.

Portanto, a obtenção dos dados pode ser feita através das Secretarias de Fazendas dos Estados e seus postos de Fiscalização, onde podemos conseguir a partir da GIA/ST, informações importantes para análise.

Caso não seja possível a obtenção dos dados da maneira citada, pode ser coletada uma amostra dos transportadores de carga, atribuindo pesos às variáveis que tem influência nas viagens realizadas, dependendo do tipo de carga.

2.3 ESTIMAÇÃO DA MATRIZ ORIGEM-DESTINO

A estimação da matriz OD pode ser feita de diversas formas, desde a contagem de viagens realizadas entre as origens/destinos propostos para a matriz até métodos de estimação mais complexos, como os propostos por Calixto (2011) e Guerra (2011).

Para o estudo realizado, um modelo importante para a estimativa da matriz OD é o modelo proposto por Romanatto (2011).

2.3.1 MODELO GRAVITACIONAL

A utilização de modelos gravitacional se fez presente nos anos 60, com idéias de aplicações na teoria de comércio internacional. As variáveis utilizadas nessas aplicações normalmente eram o PIB, o fluxo do comércio bilateral e a distância entre os elementos da amostra, normalmente países.

Segundo Romanatto (2011), um modelo gravitacional que explica o fluxo de viagens entre UF's do país através de variáveis sócioeconômicas e a distância existente entre elas tem a seguinte forma:

$$F_{ij} = e^{\alpha} \prod_{k=1}^{p} X_k^{\beta_k} e^{u_{ij}} \tag{2.1}$$

onde

- F_{ij} é o fluxo entre as unidades i e j;

- X_k é o vetor de variáveis explicativas;

- α e β_k são os parâmetros a serem estimados;

- u_{ij} representa os erros aleatórios, independentes, normalmente distribuídos com média zero e variância σ^2

Linearizando o modelo temos:

$$ln(F_{ij}) = \alpha + \beta_k[ln(X_k)] + u_{ij} \tag{2.2}$$

Capítulo 3

ANÁLISE MULTIVARIADA

3.1 INTRODUÇÃO

Como pode ser visto anteriormente, diversas variáveis influenciam na estimação de uma matriz origem-destino, e portanto, a Análise Multivariada, que consiste de diversas técnicas que buscam explicar os dados a partir de uma grande quantidade de variáveis utilizando-as simultaneamente, pode ser utilizada para a estimativa da matriz OD.

As técnicas multivariadas que serão utilizadas para a estimativa da matriz OD serão a Análise de Componentes Principais, Análise Fatorial e a Análise de Cluster. A seguir será abordado mais sobre cada técnica utilizada no estudo.

3.2 ANÁLISE DE COMPONENTES PRINCIPAIS

A análise de componentes principais tem como interesse, segundo Johnson and Wichern (2007), explicar a estrutura de variância-covariância de um conjunto de variáveis através de algumas combinações lineares destas variáveis. Seus objetivos principais são redução de dados e melhor interpretação dos mesmos.

Em um universo em que possuimos p variáveis para representarmos a sua variabiliade, temos que k componentes podem representar uma grande parte dessa variabilidade. Logo, essas componentes contém quase o mesmo tanto de informação quanto às variáveis originais do universo, assim podemos substituir as variáveis originais pelas componentes para realizarmos a análise dos dados.

Com a análise de componentes principais são revelados relacionamentos entre as variáveis que anteriormente não foram identificados, e a partir desses relacionamentos novas interpretações podem ser feitas.

3.2.1 AS COMPONENTES PRINCIPAIS

Segundo Johnson and Wichern (2007) as componentes principais, em sua forma algébrica são combinações lineares das p variáveis aleatórias X_1, X_2,..., X_p. Geometricamente, as combinações lineares representam a seleção de um novo sistema de coordenadas, sendo este obtido através da rotação do sistema original com X_1, X_2,..., X_p como as coordenadas dos eixos. Esses novos eixos representam as direções com a máxima variância e nos fornece uma forma mais simples e parcimoniosa da matriz de variância-covariância.

Uma parte fundamental no cálculo das componentes principais, é o cálculo dos autovalores e autovetores da matriz de covariância. Considerando um vetor aleatório $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$ com matriz de covariância $\mathbf{\Sigma}$, a obtenção dos autovalores e autovetores se dá por:

$$\text{Autovalor: } \left| \mathbf{\Sigma} - \lambda \mathbf{I} \right| = 0 \tag{3.1}$$

$$\text{Autovetor: } \mathbf{\Sigma x} = \lambda \mathbf{x} \tag{3.2}$$

A partir dos autovalores e autovetores obtidos, onde os autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, as componentes principais para essas variáveis são:

$$\begin{aligned} Y_1 &= \mathbf{a}_1'\mathbf{x} = a_{11}x_1 + \ldots + a_{1p}x_p \\ Y_2 &= \mathbf{a}_2'\mathbf{x} = a_{21}x_1 + \ldots + a_{2p}x_p \\ \vdots &= \vdots \\ Y_p &= \mathbf{a}_p'\mathbf{x} = a_{p1}x_1 + \ldots + a_{pp}x_p \end{aligned}$$

ou

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_p \end{bmatrix} = \begin{bmatrix} a_{11} & \ldots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \ldots & a_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \tag{3.3}$$

$$Var(Y_i) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i, \quad i = 1, \ldots, p. \tag{3.4}$$

$$Cov(Y_i, Y_k) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_k, \quad i, k = 1, \ldots, p. \tag{3.5}$$

As componentes principais $Y_1, Y_2, \ldots, Y_p$ são não correlacionadas e as variâncias são as maiores possíveis. E a variância para cada componente apresenta um comportamento de decrescimento, onde temos $Var(Y_1) \geq Var(Y_2) \geq \ldots \geq Var(Y_p)$.

Portanto, como resultado dessas combinações lineares, considerando $\mathbf{\Sigma}$ a matriz de covariância associada ao vetor aleatório $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$, os autovalores e autovetores de $\mathbf{\Sigma}$ dados por $(\lambda_1, e_1), \ldots, (\lambda_p, e_p)$. Temos que a i-ésima componente principal é dada por:

$$Y_i = \mathbf{e}_i'\mathbf{x} = e_{i1}x_1 + e_{i2}x_2 + \cdots + e_{ip}x_p \tag{3.6}$$

com

$$\begin{cases} Var(Y_i) = \mathbf{e_i'}\mathbf{\Sigma}\rho = \lambda_i, \quad i = 1, \ldots, p. \\ Cov(Y_i, Y_k) = \boldsymbol{\rho}_i\mathbf{\Sigma}\boldsymbol{\rho_k} = 0, \quad i \neq k \end{cases} \tag{3.7}$$

Observe que a Equação (3.6) é a Equação (3.3) calculada com os respectivos autovalores e autovetores da matriz de covariância $\mathbf{\Sigma}$.

Para cada componente, é calculada a proporção da variância total que é explicada pela variância da componente. A proporção explicada pela k-ésima componente principal é :

$$\text{Proporção} = \frac{Var(Y_k)}{\sum_{i=1}^{p} Var(Y_i)} = \frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i} \tag{3.8}$$

Segundo Johnson and Wichern (2007), para a realização de uma análise satisfatória é desejável que pelo menos 80% da variância total seja explicada pelas l componentes, onde $l \leq p$, podendo as p variáveis serem substituídas pelas l componentes com pouca perda de informação.

3.3 ANÁLISE FATORIAL

O objetivo principal da análise fatorial é descrever, se possível, o relacionamento da covariância entre muitas variáveis em termos de poucas, mas não observáveis, variáveis aleatórias chamadas *fatores*.

A Análise Fatorial pode ser considerada uma extensão da análise de componentes principais, pois ambas buscam aproximar a matriz de covariância $\mathbf{\Sigma}$. Entretanto, o modelo de análise fatorial apresenta uma aproximação mais elaborada que a aproximação feita a partir da análise de componentes principais.

3.3.1 MODELO FATORIAL

O modelo de análise fatorial é:

$$
\begin{aligned}
X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + \varepsilon_1 && (3.9)\\
X_2 - \mu_2 &= l_{21}F_2 + l_{22}F_2 + \ldots + l_{2m}F_m + \varepsilon_2 && (3.10)\\
\vdots &= \vdots && (3.11)\\
X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \ldots + l_{pm}F_m + \varepsilon_p && (3.12)
\end{aligned}
$$

ou matricialmente

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} \tag{3.13}$$

onde l_{ij} é a carga da i-ésima variável no j-ésimo fator, sendo assim $\mathbf{L}$ a matriz de carga de fatores. Para a execução do modelo, é necessário algumas suposições para os fatores $\mathbf{F}$ e para os erros $\boldsymbol{\varepsilon}$.

$$
\begin{aligned}
E(\mathbf{F}) &= 0 && (3.14)\\
Cov(\mathbf{F}) &= E(\mathbf{FF}') = \mathbf{I} && (3.15)\\
E(\varepsilon) &= 0 && (3.16)\\
Cov(\boldsymbol{\varepsilon\varepsilon}') &= \boldsymbol{\Psi} && (3.17)
\end{aligned}
$$

onde

$$\boldsymbol{\Psi} = \begin{bmatrix} \Psi_1 & 0 & \ldots & 0 \\ 0 & \Psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Psi_p \end{bmatrix} \tag{3.18}$$

A partir dessas suposições e a estrutura de modelo apresentada acima é constituído o Modelo Fatorial Ortogonal (MFO). Para o MFO, a estrutura de matriz de

covariância $\mathbf{\Sigma}$ é:

$$Var(X_i) = {l_{i1}}^2 + \ldots + {l_{in}}^2 + \Psi_i \tag{3.19}$$

onde Ψ_i é a variância específica atribuída a variável i e ${l_{i1}}^2 + \ldots + {l_{in}}^2$ são as comunalidades existentes entre as variáveis X_i.

3.3.2 MÉTODOS DE ESTIMAÇÃO

Na estimativa através da análise fatorial, o objetivo é obter um modelo para verificar os relacionamento da matriz de covariância $\mathbf{\Sigma}$. O problema inicial na análise fatorial é a estimativa das cargas fatoriais l_{ij} e a das variâncias específicas Ψ_i. Serão descritos os dois métodos utilizados para estimação dos parâmetros.

MÉTODO DE COMPONENTES PRINCIPAIS

A decomposição espectral nos fornece uma forma de fatorar a matriz de covariância $\mathbf{\Sigma}$. Seja o par (λ_i, e_i) de autovalor-autovetor de $\mathbf{\Sigma}$ com $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, então

$$\mathbf{\Sigma} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \ldots + \lambda_p \mathbf{e}_p \mathbf{e}_p' = \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1 & \ldots & \sqrt{\lambda_p}\mathbf{e}_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \vdots \\ \sqrt{\lambda_p}\mathbf{e}_p' \end{bmatrix} \tag{3.20}$$

A matriz de cargas tem a j-ésima coluna dada por $\sqrt{\lambda_i}\mathbf{e}_i$.

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{0} = \mathbf{L}\mathbf{L}' \text{ (Se } m = p) \tag{3.21}$$

Portanto, temos:

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi} = \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1 & \ldots & \sqrt{\lambda_p}\mathbf{e}_p \end{bmatrix} \begin{bmatrix} \sqrt{\lambda_1}\mathbf{e}_1' \\ \vdots \\ \sqrt{\lambda_p}\mathbf{e}_p' \end{bmatrix} + \begin{bmatrix} \Psi_1 & 0 & \ldots & 0 \\ 0 & \Psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Psi_p \end{bmatrix} \tag{3.22}$$

onde $\Psi_i = \sigma_{ii} - \sum_{i=1}^{n} lij^2$

Para aplicarmos esse método aos dados, é necessário primeiro centrar as observações. As observações centradas tem a mesma matriz de covariância $\mathbf{\Sigma}$ que as observações originais. Caso as unidades das variáveis não sejam comparáveis, utiliza-se variáveis padronizadas cuja a matriz de covariância é a matriz de correlação $\mathbf{R}$.

Se o número de fatores não é determinado por considerações a priori, a escolha pode ser baseado nos autovalores estimados da mesma maneira das componentes principais. Assim, o número de fatores retidos no modelo é aumentado até que uma proporção da variância explicada seja satisfatória para o estudo, onde para calcularmos essa proporção utilizamos a Equação (3.8) da análise de componentes principais.

MÉTODO DE MÁXIMA VEROSSIMILHANÇA

Se os fatores comuns $\mathbf{F}$ e os fatores específicos $\boldsymbol{\varepsilon}$ são normalmente distribuídos, então as estimativas de máxima verossimilhança (MV) para as cargas fatoriais e variâncias específicas podem ser obtidas.

Quando $\mathbf{F}$ e $\boldsymbol{\varepsilon}$ são conjuntamente normais, as observações $X_j - \boldsymbol{\mu} = \mathbf{L}\mathbf{F}_j + \boldsymbol{\varepsilon}_j$ são normais e a verossimilhança é dada por:

$$L(\boldsymbol{\mu}, \boldsymbol{\varepsilon}) = \frac{1}{(2\Pi)^{-\frac{np}{2}}|\mathbf{\Sigma}|^{-\frac{1}{2}}} e^{-\frac{1}{2}Tr[\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})\Sigma^{-1}(\mathbf{x}_j-\bar{\mathbf{x}})'+n(\bar{\mathbf{x}}-\boldsymbol{\mu})(\bar{\mathbf{x}}-\boldsymbol{\mu})']} \tag{3.23}$$

que depende de $\mathbf{L}$ e $\mathbf{\Psi}$ através de $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}' + \mathbf{\Psi}$. Para que $\mathbf{L}$ seja bem definida, pode-se impor a condição de singularidade $\mathbf{L}'\mathbf{\Psi}^{-1}\mathbf{L} = \mathbf{\Lambda}$.

As estimativas de MV $\hat{\mathbf{L}}$ e $\hat{\mathbf{\Psi}}$ devem ser obtidas pela maximização de $L(\boldsymbol{\mu}, \mathbf{\Sigma})$. As estimativas de MV das comunalidades são:

$$\hat{h}_i^2 = \hat{l}_{ij}^2 + \ldots + \hat{l}_{pj}^2 \text{ i=1,...,p.} \tag{3.24}$$

$$\text{Proporção da Variância Explicada Devido ao } j\text{-ésimo fator} = \frac{\hat{l}_{ij}^2 + \ldots + \hat{l}_{pj}^2}{S_{11} + \ldots + S_{pp}} \tag{3.25}$$

Se as variâncias forem padronizadas tal que $\mathbf{Z} = \mathbf{V}^{\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu})$, então a matriz de covariância de $\boldsymbol{\rho}$ de $\mathbf{Z}$ é dada por

$$\boldsymbol{\rho} = \mathbf{V}^{\frac{1}{2}}\boldsymbol{\Sigma}\mathbf{V}^{-\frac{1}{2}} = (\mathbf{V}^{\frac{1}{2}}\mathbf{L})(\mathbf{V}^{\frac{1}{2}}\mathbf{L})' + \mathbf{V}^{-\frac{1}{2}}\boldsymbol{\Psi}\mathbf{V}^{-\frac{1}{2}} \tag{3.26}$$

Assim, a matriz de cargas fatoriais é dada por $\mathbf{L}_Z = \mathbf{V}^{\frac{1}{2}}\mathbf{L}$ e a matriz de variância específicas é dada por $\boldsymbol{\Psi}_Z = \mathbf{V}^{-\frac{1}{2}}\boldsymbol{\Psi}\mathbf{V}^{\frac{1}{2}}$. Consequentemente, a estimativa de MV para $\boldsymbol{\rho}$ é

$$\hat{\boldsymbol{\rho}} = \mathbf{L}_Z\mathbf{L}_Z' + \boldsymbol{\Psi}_Z \tag{3.27}$$

ou

$$\hat{l_{ij}} = \hat{l_{Z,ij}}\sqrt{\sigma_{ii}^2} \tag{3.28}$$

e

$$\hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\Psi}_{Z,i}}\hat{\sigma_{ii}} \tag{3.29}$$

sendo na amostra utilizada $[\frac{(n-1)}{n}]$ $\mathbf{S}$ ao invés de $\mathbf{S}$, ou seja, fazendo a divisão de matriz de covariância por n.

ROTAÇÃO DE FATORES

Toda carga fatorial obtida de cargas iniciais por uma formação ortogonal tem a mesma habilidade de reproduzir a matriz de covariância ou correlação. Da álgebra matricial, temos que uma transformação ortogonal corresponde a uma rotação dos eixos de coordenadas. Por essa razão, a transformação ortogonal das cargas fatoriais,

como também a transformação ortogonal dos fatores é chamada de **Rotação de Fatores**, e o objetivo dessa rotação é facilitar a interpretação dos dados.

Se $\hat{\mathbf{L}}$ é uma matriz $p \times p$ das cargas fatoriais obtidas por qualquer método, então

$$\mathbf{L}^* = \mathbf{L}'\mathbf{T} \tag{3.30}$$

onde

$$\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I} \tag{3.31}$$

é uma matriz de cargas "rotacionadas". Ademais, a matriz de covariância (ou correlação) estimada permanece inalterada, dado que

$$\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\mathbf{\Psi}} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{L}' + \hat{\mathbf{\Psi}} = \hat{\mathbf{L}}^*\hat{\mathbf{L}}^{*'} + \hat{\mathbf{\Psi}} \tag{3.32}$$

A matriz de resíduos, as comunalidades h_i^2, as variâncias específicas Ψ_i também permanecem inalteradas.

3.4 ANÁLISE DE *CLUSTERS*

Após a obtenção das componentes principais, a análise de *cluster* será realizada visando o agrupamento dos municípios através das componentes geradas.

A análise de *cluster* ou de agrupamento, diferente das demais técnicas de análise multivariada, consiste em uma quantidade de *clusters* desconhecidos, que tem como objetivo principal atribuir observações para esses grupos. Para essa análise, não são feitas suposições sobre o número de grupos na análise, nem sobre a estrutura dos mesmos. O Agrupamento é feito com base em critérios pré-definidos.

O objetivo principal da análise de *cluster* é descobrir grupos naturais de itens ou

variáveis. No entanto, é necessário primeiro determinarmos uma escala quantitativa para medirmos as associações existentes entre as variáveis.

3.4.1 MEDIDAS DE SIMILARIDADE

A maioria das tentativas para produzir uma estrutura de grupo simples a partir de um conjunto de dados complexo necessitam de medidas de similaridade. Geralmente, a subjetividade está envolvida na escolha de uma medidia de similaridade. Pontos importantes para o agrupamento são: a natureza das variáveis (discreta, contínua, binária), a escala de medição (nominal, ordinal, intervalo, razão), além de conhecimento sobre o assunto. Quando unidades são agrupadas, a proximidade é geralmente indicada por algum tipo de distância.

Considere os vetores aleatórios $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$ e $\mathbf{y}' = [y_1, y_2, \ldots, y_p]$. A distância euclidiana entre esses dois vetores aleatórios de p dimensões é:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2} = \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})} \quad (3.33)$$

A distância estatística para os mesmos vetores aleatórios é da forma

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'\mathbf{S}^{-1}(\mathbf{x} - \mathbf{y})} \quad (3.34)$$

onde $\mathbf{S}$ é a matriz de variância e covariância amostral. Entretanto, sem um prévio conhecimento dos diferentes grupos, as amostra não poderão ser computadas. Por essa razão, a distância euclidiana é geralmente utilizada para agrupamento de unidades.

3.4.2 MÉTODOS DE AGRUPAMENTO HIERÁRQUICOS

As técnicas de agrupamento hierárquico procedem ou através de sucessivas fusões

ou de sucessivas divisões. Os métodos hierárquicos aglomerativos iniciam com as unidades separadas. Primeiramente, as unidades que possuem mais semelhanças são agrupadas, e após a formação dos grupos, são agrupados novamente, considerando as similaridades entre os *clusters* criados anteriormente, assim, a tendência desse método, é que em seu final, teremos somente um *cluster*.

Os métodos de divisão hierárquica age de maneira oposta aos métodos aglomerativos. Em um primeiro momento, um *cluster* de unidades é divido em dois *clusters*, onde os elementos de um *clusters* estão "distantes" dos elementos do outro. Esse processo de divisão continua até que se possua um elemento em cada grupo.

Para o nosso estudo, utilizaremos os procedimentos aglomerativos, em particular, o método de agrupamento hierárquico proposto por (Joe H. Ward, 1963).

MÉTODO DE AGRUPAMENTO HIERÁRQUICO DE WARD

O procedimento hierárquico proposto por Joe H. Ward (1963) é baseado em minimizar a "perda de informação" ao unir dois grupos.

A perda de informação é tomada como o crescimento de um critério de soma de quadrados dos erros (ESS).

$$ESS = \sum_{j=1}^{k} (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}}) \tag{3.35}$$

onde $\mathbf{x}_j$ é a medida multivariada associada com o j-ésimo elemento e $\bar{\mathbf{x}}$ é a média de todos os elementos.

A partir desses cálculos, constrói-se um dendograma (Figura 3.1), que é construído baseado na distância desses elementos aos *cluster* formados a partir do banco de dados.

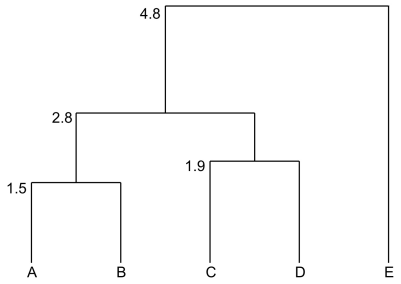

Figura 3.1: Dendograma

Tanto no método de Joe H. Ward (1963), quanto para os demais métodos, a característica que diferencia os métodos hierárquicos dos não-hierárquicos é a impossibilidade de se realocar um elemento para outro *cluster*. Os métodos não hierárquicos, em que é possível a realocação dos elementos entre os *clusters*, será melhor explorada a seguir.

3.4.3 MÉTODOS DE AGRUPAMENTO NÃO-HIERÁRQUICOS

As técnicas de agrupamento não-hierárquicas são direcionadas para o agrupamento de elementos, ao invés das variáveis, em K *clusters*. O número de *clusters* K, pode ser especificado antecipadamente ou determinado durante o processo de agrupamento. Devido à não determinação da matriz de distâncias (similaridades) e à não computação dos dados, o método não-hierárquico pode ser aplicado para bancos de dados maiores do que os bancos em que as técnicas hierárquicas podem ser aplicadas.

Nos métodos não-hierárquicos é necessário que se predetermine a quantidade de *clusters*. Para o estudo, é determinado a fim de comparação com o *rank* estabelecido

pelo IBGE (2007) que o número de *clusters* será 4.

O método não-hierárquico começa de qualquer partição de elementos de grupos ou de sementes aleatórias. Uma maneira de se iniciar o procedimento é selecionar aleatoriamente pontos centrais e particionar aleatoriamente os elementos em grupos. O método utilizado no estudo será um dos mais conhecidos métodos não-hierárquicos, o método de k-médias.

K-MÉDIAS

O método de k-médias consiste em seguir os seguintes passos:

- Separe os elementos em K $clusters$ iniciais.
- A partir de uma lista dos elementos, atribua cada elemento ao *cluster* em que o ponto central (média) é o mais próximo. Recalcule o ponto central do *cluster* recebendo o novo elemento e perdendo um elemento.
- Repita o passo anterior até que não sejam mais realizadas alocações de elementos.

Capítulo 4

MATERIAL E MÉTODOS

4.1 INTRODUÇÃO

Para a realização desse estudo, é necessário a obtenção de dados sobre o deslocamento exercido pelos transportes de carga e passageiro, como o município de origem e o município de destino ou as microrregiões, e afim de explicar o deslocamento desses transportes é necessária a obtenção de variáveis diversas sócioeconomicas, assim serão aplicadas diversas técnicas multivariadas com o objetivo de obtermos uma proposta de estimativa da matriz OD.

Portanto, nesse capítulo serão explicados as formas de obtenção desses dados, e como serão utilizadas as técnicas multivariadas para a análise desses dados.

4.2 DADOS DE TRANSPORTES E VARIÁVEIS SOCIOECONOMICAS

Os dados de transporte de passageiros serão obtidos através da Agência Nacional de Transportes Terrestres - ANTT (2013), pois a mesma é responsável pela fiscalização sobre esses transportes, onde a mesma solicita o cumprimento de normas como a capacidade máxima de passageiros, assim é possível contabilizar a quanti-

dade de viagens realizadas para o transporte de passageiros.

Os dados de transporte de cargas serão obtidos através da Secretaria de Fazenda do Distrito Federal - SEFAZDF (2013), dado que a mesma é responsável pela tributação sobre as cargas transportadas, assim são obtidas informações importantes para a análise, como mercadoria, origem/destino da mercadoria, peso da mercadoria e valor da mercadoria.

As variáveis socioeconômicas serão obtidas através de informações do IBGE (2013), onde teremos mais de 60 variáveis para a realização da análise multivariada para a construção da matriz origem-destino.

4.3 ANÁLISE MULTIVARIADA DOS DADOS

Com os dados obtidos, onde temos a quantidade de viagens entre uma origem e um destino como a variável resposta Y_i e diversas variáveis socioeconômicas como variáveis explicativas X_i's, é possível construir um modelo de regressão da seguinte forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon_{ij} \tag{4.1}$$

Porém, como possuimos um grande número de variáveis explicativas, a análise de regressão pode apresentar o problema de colinearidade entre as variáveis, enviesando os parâmetros do modelo de regressão.

Uma forma de solucionar o problema de multicolinearidade é por meio da análise de componentes principais e/ou análise fatorial, onde é possível obter componentes ou fatores que serão independentes entre si, resolvendo assim o problema da multicolinearidade.

Assim, será possível obter um modelo não mais em função das covariáveis originais, e sim das componentes ou fatores obtidos a partir delas. O novo modelo de regressão terá a seguinte forma:

$$Y = \beta_0^* + \beta_1^* Z_1 + \beta_2^* Z_2 + \ldots + \beta_n^* Z_k + \varepsilon_{ij} \tag{4.2}$$

onde os $Z_i's$ são as componentes geradas a partir das variáveis utilizadas.

Uma parte importante de uma análise de dados é a análise descritiva, onde nela é possível obter informações importantes sobre as variáveis que serão analisadas. Após feita a análise descritiva, para que as técnicas de componentes principais e fatorial possam ser aplicadas aos dados, é necessário que haja correlação entre os dados estudados, sendo assim serão calculadas as correlações entre as variáveis selecionadas para o estudo.

A partir da verificação de existência de correlação entre as variáveis presentes no estudo, é possível utilizar as técnicas multivariadas de componentes principais e/ou análise fatorial, a fim de que se possa gerar componentes ou fatores, independentes entre si, que explique a variável resposta, assim podendo ser construído o modelo proposto em 4.2.

A Análise de *Clusters* poderá ser utilizada caso as microrregiões sejam muito heterogêneas e um modelo único de regressão não tenha um ajuste razoável. Então pode-se agrupar as microrregiões a partir das componentes geradas e/ou a partir das variáveis originais. Na técnica não-hierárquica, o número de *clusters* construídos será 4, com o fim comparativo em relação ao *rank* proposto pelo IBGE (2007).

Capítulo 5

ANÁLISE DOS RESULTADOS

5.1 INTRODUÇÃO

Neste capítulo serão apresentados os resultados obtidos em relação aos transportes de passageiros. Devido à indisponibilidade dos dados para o transporte de carga, não foi possível realizar a análise para este tipo. Após a análise descritiva e de associação entre as variáveis, serão aplicadas as técnicas multivariadas propostas no estudo, utilizando o *software* SAS 9.2.

5.2 MODELO GERAL

Para o início do estudo, é necessário que se identifique as variáveis utilizadas no estudo e faça uma breve análise descritiva dessas variáveis. Portanto as variáveis selecionadas para o estudo são:

- Quantidade de Matrículas no Ensino Fundamental (MEF);
- Quantidade de Matrículas no Ensino Médio (MEM);
- Quantidade de Matrículas no Ensino Superior (MES);

- População Residente Alfabetizada a partir dos 5 anos de idade (PRA);
- População que Frequentava ou Frequenta Creche (PRFC);
- Quantidade de Pessoal Ocupado Assalariado(POA);
- Quantidade de Pessoal Ocupado Total (POT);
- Salários e Outras Remunerações (SOR);
- Arrecadação no Setor Industrial (IND);
- Arrecadação no Setor de Serviço(SERV);
- Arrecadação na Administração Pública (APU);
- Impostos Arrecadadados (IMP);
- População Residente (POPR);
- População de Homens Residentes (POPHR);
- População de Mulheres Residentes (POPMR);
- População Católica Residente (POPCR);
- População Espírita Residente (POPESPR);
- População Evangélica Residente (POPEVGR);
- Quantidade de Automóveis (AUTOMOVEL);
- Quantidade de Caminhões (CAMINHÃO);

- Quantidade de Tratores (TRATOR);
- Quantidade de Caminhonete (CAMINHONETE);
- Quantidade de Camionetas (CAMIONETA);
- Quantidade de Micro-Ônibus (MICROBUS);
- Quantidade de Motocicletas (MOTO);
- Quantidade de Motonetas (MOTONETA);
- Quantidade de Ônibus (ÔNIBUS);

Após a identificação das variáveis presentes no estudo, obteve-se as seguintes estatísticas descritivas:

Obs	Variável	N	Média	Desvio Padrão	Coeficiente de Variação	Mínimo	Máximo
1	MEF	558	53.233	121.506	2,283	291	1.882.022
2	MEM	558	15.012	37.779	2,517	49	628.289
3	MES	355	22.637	88.402	3,905	27	1.353.766
4	PRA	558	282.494	780.689	2,764	2.263	12.311.557
5	PRFC	558	106.755	274.003	2,567	888	4.438.357
6	POA	558	80.974	325.689	4,022	720	5.916.341
7	POT	558	93.500	372.548	3,984	764	6.838.329
8	SOR	558	1865608	9714878	5,207	8.870	185.299.247
9	IND	558	1623383	5743280	3,538	5.220	103.352.154
10	SERV	558	3853279	17784037	4,615	23267	330.303.121
11	APU	558	936.849	3956820	4,224	4.117	72.493.372
12	IMP	558	972.944	5027675	5,167	2.546	94.746.076
13	POPR	558	341.871	878.401	2,569	2.630	13.804.831
14	POPHR	558	167.404	417.600	2,495	1.292	6.559.587
15	POPMR	558	174.468	460.858	2,642	1.338	7.245.244
16	POPCR	558	220.950	478.635	2,166	1.201	8.014.121
17	POPESPR	553	6.960	37.088	5,329	4	620.518
18	POPEVGR	558	75.761	225.286	2,974	994	3.284.551
19	AUTOMOVEL	558	76.483	307.177	4,016	5	5.902.668
20	CAMINHÃO	558	4.266	9.531	2,234	8	156.807
21	TRATOR	558	883	2.115	2,396	0	32.302
22	CAMINHONETE	558	9.387	25.311	2,696	18	465.465
23	CAMIONETA	558	4.101	19.487	4,752	4	389.397
24	MICROBUS	558	571	2.157	3,777	0	39.677
25	MOTO	558	30.303	57.175	1,887	255	972.984
26	MOTONETA	558	5.419	8.962	1,654	15	133.064
27	ONIBUS	558	923	2.894	3,137	1	49.362

Figura 5.1: Estatísticas Descritivas

Observa-se que para MES, tem-se apenas 355 observações, uma quantidade muito menor que a apresentada pelas demais variáveis presentes no estudo, isso porque as

instituições de ensino superior não estão presentes em todas as microrregiões, com uma média de $22.636,66$ alunos matriculados no ensino superior e um coeficiente de variação de $3,905$, onde a menor quantidade de alunos matriculados no ensino superior observada é de 27 na microrregião Itaberaba-BA. Já a maior quantidade de alunos matriculados no ensino superior é de aproximadamente 1 milhão na microrregião São Paulo-SP.

Em relação à população residente alfabetizada (PRA), observa-se uma média de $282.494,206$ pessoas alfabetizadas por microrregião com um coeficiente de variação de $2,764$, onde a microrregião que apresenta a menor quantidade de cidadãos alfabetizados é Fernando de Noronha-PE, com 2.263 pessoas alfabetizadas. Já a microrregião que apresenta a maior quantidade de pessoas alfabetizadas é São Paulo-SP, com aproximadamente 12 milhões de alfabetizados.

Pode-se observar que a quantidade média por microrregião de pessoas ocupadas assalariadas (POA) é de $80.974,11$ com um coeficiente de variação de $4,022$, onde a microrregião que apresenta a menor quantidade de pessoas ocupadas assalariadas é Japura-AM com 720 indivíduos nessa situação. Já São Paulo-SP é a microrregião que apresenta a maior quantidade de pessoas que trabalham recebendo salário, com total aproximado de 6 milhões.

Nota-se que a média de população residente (POPR) é de $341.871,38$ de habitantes por microrregião com um coeficiente de variação de $2,569$, onde a microrregião que possui a menor população residente é Fernando de Noronha-PE com 2.630 habitantes. Já São Paulo-SP é a cidade com o maior número de residentes, com

aproximadamente 14 milhões.

Nota-se também que a quantidade de média de automóveis é de $76.482,64$ automóveis por microrregião com um coeficiente de variação de $4,016$, onde Japura-AM possui a menor quantidade de automóveis no território brasileiro com somente 5 automóveis. Já a cidade de São Paulo-SP possui a maior quantidade, com aproximadamente 6 milhões de automóveis.

Após a análise descritiva dos dados, realiza-se o cálculo da correlação, e a partir dessa, observa-se que todas as variáveis são altamente correlacionadas positivamente, sendo a menor correlação observada de $0,5$. Quando da construção de um modelo de regressão para a variável dependente **PASSRODO** (quantida de passageiros transportados por ônibus na origem), espera-se que todos os parâmetros estimados tenham o mesmo sinal apresentado na matriz de correlação, entretanto, observa-se que algumas variáveis tem sinais negativos, sendo um indício de que a utilização de um modelo de regressão não é uma opção viável, pois o modelo apresenta o problema da multicolinearidade dos dados, como pode-se observar na Figura 5.2.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	26	2.70064E14	1.038708E13	130.66	<.0001
Error	326	2.591691E13	79499711065		
Corrected Total	352	2.959809E14			

Root MSE	281957	R-Square	0.9124
Dependent Mean	358764	Adj R-Sq	0.9055
Coeff Var	78.59113		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	39718	30709	1.29	0.1968
MEF	MEF	1	-14.38978	3.27877	-4.39	<.0001
MEM	MEM	1	-17.11077	8.71020	-1.96	0.0503
MES	MES	1	2.36897	1.80363	1.31	0.1900
PRA	PRA	1	-4.72756	1.45165	-3.26	0.0012
PRFC	PRFC	1	7.96097	3.26136	2.44	0.0152
POA	POA	1	-19.58992	6.82157	-2.87	0.0043
POT	POT	1	16.33288	6.24535	2.62	0.0093
SOR	SOR	1	0.04780	0.04206	1.14	0.2566
IND	IND	1	-0.02613	0.00857	-3.05	0.0025
SERV	SERV	1	-0.00076570	0.01227	-0.06	0.9503
APU	APU	1	0.02838	0.01673	1.70	0.0907
IMP	IMP	1	-0.01784	0.02356	-0.76	0.4494
POPR	POPR	B	-0.22633	3.47421	-0.07	0.9481
POPHR	POPHR	B	4.84637	6.53847	0.74	0.4591
POPMR	POPMR	0	0	.	.	.
POPCR	POPCR	1	2.01968	0.73562	2.75	0.0064
POPESPR	POPESPR	1	7.80344	3.50111	2.23	0.0265
POPEVGR	POPEVGR	1	5.77949	1.09103	5.30	<.0001
AUTOMOVEL	AUTOMOVEL	1	1.85493	1.36523	1.36	0.1752
CAMINHAO	CAMINHÃO	1	-10.48488	15.84187	-0.66	0.5085
TRATOR	TRATOR	1	28.97116	27.83949	1.04	0.2988
CAMINHONETE	CAMINHONETE	1	8.11144	7.48722	1.08	0.2794
CAMIONETA	CAMIONETA	1	-22.99370	20.77292	-1.11	0.2691
MICROBUS	MICROBUS	1	-62.15028	59.89939	-1.04	0.3002
MOTO	MOTO	1	-2.72787	1.93039	-1.41	0.1586
MOTONETA	MOTONETA	1	7.33760	4.62584	1.59	0.1137
ONIBUS	ONIBUS	1	10.61962	50.59621	0.21	0.8339

Figura 5.2: Modelo de Regressão Geral

Portanto, devido à multicolinearidade, é proposto pelo estudo que se utilize de técnicas multivariadas, nesse caso componentes principais, a fim de que se evite tal problema. As componentes geradas estão na Figura 5.3:

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	24.8018308	23.9201683	0.9186	0.9186
2	0.8816625	0.4386788	0.0327	0.9512
3	0.4429837	0.1305231	0.0164	0.9676
4	0.3124606	0.1092662	0.0116	0.9792
5	0.2031944	0.0982574	0.0075	0.9867
6	0.1049370	0.0169615	0.0039	0.9906

Eigenvectors

	Prin1	Prin2	Prin3
MEF	0.196800	0.122527	-.141132
MEM	0.198591	0.065012	-.135449
MES	0.197243	0.018005	-.062415
PRA	0.198177	0.107529	-.139889
PRFC	0.198077	0.108876	-.142734
POA	0.199953	0.004795	-.012819
POT	0.199922	-.008524	-.008436
SOR	0.197668	0.035129	0.107872
IND	0.188095	-.172774	-.030007
SERV	0.195330	0.102623	0.204095
APU	0.136917	0.566242	0.738635
IMP	0.196256	-.019960	0.045186
POPR	0.197839	0.112772	-.151072
POPHR	0.197952	0.107387	-.146761
POPMR	0.197716	0.117634	-.154958
POPCR	0.198271	0.033742	-.114268
POPESPR	0.192563	0.139230	-.107988
POPEVGR	0.193218	0.161219	-.154102
AUTOMOVEL	0.197804	-.087211	0.075742
CAMINHÃO	0.192557	-.227718	0.067941
TRATOR	0.168452	-.472490	0.251314
CAMINHONETE	0.193926	-.198107	0.148707
CAMIONETA	0.197010	-.078301	0.027821
MICROBUS	0.197481	0.027602	-.104730
MOTO	0.191948	-.182461	0.027388
MOTONETA	0.163921	-.383392	0.294293
ONIBUS	0.199355	0.023414	-.065438

Figura 5.3: Componentes Principais Geral

A partir das componentes geradas, nota-se que com uma componente já é explicado aproximadamente 91% da variabilidade dos dados, entretanto não é possível fazer nenhuma interpretação, pois a componente é uma média das variáveis selecionadas para o estudo, impossibilitando assim uma interpretação para essa abordagem.

Portanto, uma forma proposta para sanar esse problema é a realização das análises por grupos de variáveis, que são:

- Variáveis Escolares;
- Variáveis Econômicas;
- Variáveis Sociais;
- Variaveis de Frota;

Na seção a seguir serão apresentadas as análises realizadas para os grupos de variáveis apresentados acima.

5.3 VARIÁVEIS ESCOLARES

Observa-se na matriz de correlações da Figura 5.4, que todas as variáveis são altamente correlacionadas, com correlações de pelo menos $0,88$. Portanto a utilização de um modelo de regressão linear não se mostra viável, pois o uso resultaria em multicolinearidade dos dados, como pode ser verificado no modelo (Figura 5.5) que possui como variável resposta a quantidade de viagens realizadas (PASSRODO).

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	PASSRODO	MEF	MEM	MES	PRA	PRFC
PASSRODO	1.00000	0.88453	0.89426	0.91169	0.89691	0.89476
		<.0001	<.0001	<.0001	<.0001	<.0001
	531	531	531	353	531	531
MEF	0.88453	1.00000	0.99398	0.95751	0.99278	0.99735
	<.0001		<.0001	<.0001	<.0001	<.0001
	531	558	558	355	558	558
MEM	0.89426	0.99398	1.00000	0.97213	0.99306	0.99721
	<.0001	<.0001		<.0001	<.0001	<.0001
	531	558	558	355	558	558
MES	0.91169	0.95751	0.97213	1.00000	0.96646	0.97040
	<.0001	<.0001	<.0001		<.0001	<.0001
	353	355	355	355	355	355
PRA	0.89691	0.99278	0.99306	0.96646	1.00000	0.99640
	<.0001	<.0001	<.0001	<.0001		<.0001
	531	558	558	355	558	558
PRFC	0.89476	0.99735	0.99721	0.97040	0.99640	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	
	531	558	558	355	558	558

Figura 5.4: Matriz de Correlação de Variáveis Escolares

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.47523E14	4.950459E13	354.49	<.0001
Error	347	4.845795E13	1.396483E11		
Corrected Total	352	2.959809E14			

Root MSE	373695	R-Square	0.8363
Dependent Mean	358764	Adj R-Sq	0.8339
Coeff Var	104.16181		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	132729	27157	4.89	<.0001
MEF	1	1.98001	2.47096	0.80	0.4235
MEM	1	-7.35416	6.21735	-1.18	0.2377
MES	1	8.77034	1.23033	7.13	<.0001
PRA	1	0.58381	0.26831	2.18	0.0302
PRFC	1	-1.32806	1.79290	-0.74	0.4594

Figura 5.5: Modelo de Regressão para PASSRODO das Variáveis Escolares

Sendo verificada a colinearidade dos dados no modelo acima, é proposto pelo estudo a utilização das técnicas multivariadas a fim de sanar esse problema. As compontes estão na Figura 5.6:

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.93623238	4.88365940	0.9872	0.9872
2	0.05257299	0.04568939	0.0105	0.9978
3	0.00688360	0.00338937	0.0014	0.9991
4	0.00349422	0.00267741	0.0007	0.9998
5	0.00081681		0.0002	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
MEF	0.447921	-.380940	-.221906	0.659979	0.411627
MEM	0.449124	-.103167	-.562154	-.663616	0.176749
MES	0.440844	0.878480	0.019114	0.166151	0.077183
PRA	0.448532	-.206706	0.795907	-.283939	0.204945
PRFC	0.449588	-.172586	-.030122	0.125753	-.866812

Figura 5.6: Componentes Principais das Variáveis Escolares

Pode-se observar que para os autovalores obtidos para as variáveis, a representação das componentes se mostra satisfatória com duas componentes, com um percentual de variância explicada de $99,78\%$ pelas componentes. Porém, para uma

melhor visualização dos fatores, utiliza-se a técnica fatorial, pois os fatores gerados podem ser rotacionados. Os resultados da Análise Fatorial estão na Figura 5.7:

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	4.93623238	4.88365940	0.9872	0.9872
2	0.05257299	0.04568939	0.0105	0.9978
3	0.00688360	0.00338937	0.0014	0.9991
4	0.00349422	0.00267741	0.0007	0.9998
5	0.00081681		0.0002	1.0000

Rotated Factor Pattern

	Factor1	Factor2
MEF	0.80855	0.58673
MEM	0.76880	0.63656
MES	0.60729	0.79441
PRA	0.78337	0.61778
PRFC	0.78001	0.62522

Figura 5.7: Fatores Rotacionados Escolares

Verifica-se no primeiro fator que a variável MEF apresenta um maior coeficiente que as demais. Já no segundo fator a variável MES apresenta o maior coeficiente.

Com a obtenção dos fatores, que são independentes entre si, é proposto pelo estudo que se faça uma regressão com os fatores, sendo essa regressão sanada do problema de colinearidade. As Figuras 5.8, 5.9 e 5.10 apresentam a matriz de correlação dos fatores, a regressão construída e o gráfico de Resíduos por Valores Preditos respectivamente.

Pearson Correlation Coefficients, N = 355
Prob > |r| under H0: Rho=0

	Factor1	Factor2
Factor1	1.00000	0.00000 1.0000
Factor2	0.00000 1.0000	1.00000

Figura 5.8: Correlação dos Fatores Escolares

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.463988E14	1.231994E14	869.67	<.0001
Error	350	4.958207E13	1.416631E11		
Corrected Total	352	2.959809E14			

Root MSE	376382	R-Square	0.8325
Dependent Mean	358764	Adj R-Sq	0.8315
Coeff Var	104.91053		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	358395	20033	17.89	<.0001
Factor1	1	537338	20020	26.84	<.0001
Factor2	1	638353	20006	31.91	<.0001

Figura 5.9: Regressão dos Fatores Escolares para PASSRODO

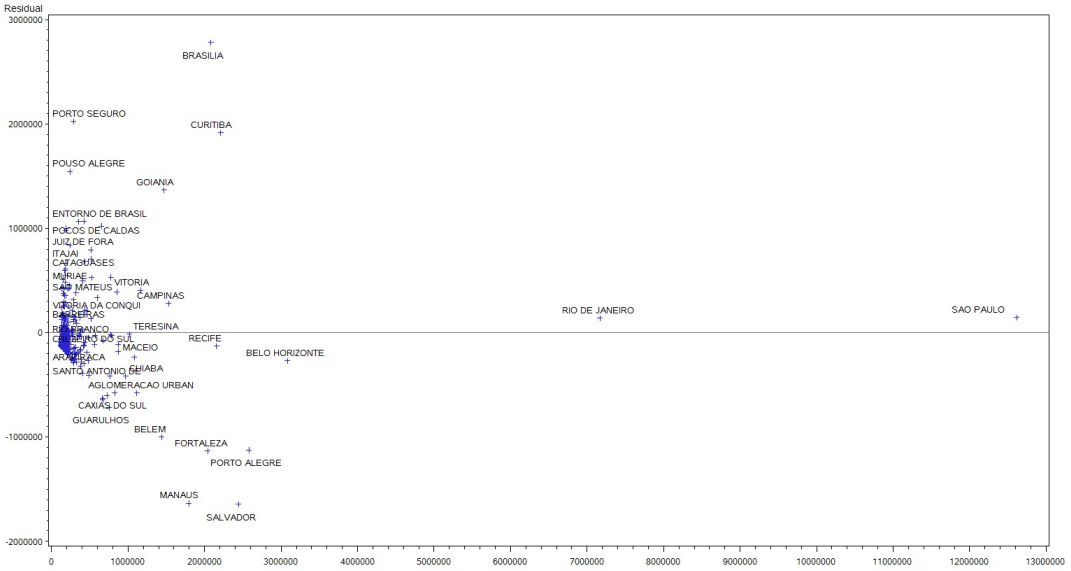

Figura 5.10: Resíduo x Valor Predito - Variáveis Escolares

A partir da regressão construída com os fatores, verifica-se que o coeficiente R^2 é de aproximadamente 83% o que aparentemente indica que o modelo é bem ajustado aos dados, porém ao observarmos o gráfico de resíduos, verificamos que os dados apresentam heterocedasticidade.

5.4 VARIÁVEIS ECONÔMICAS

Assim como nas variáveis escolares, observa-se na matriz de correlações (Figura 5.11) que todas as variáveis são altamente correlacionadas, com correlações de pelo menos $0,67$. Novamente, a utilização de um modelo de regressão linear não se mostra viável, pois o uso resultaria em um problema de multicolinearidade, como pode ser visto no modelo (Figura 5.12) que possui como variável resposta a quantidade de viagens realizadas (PASSRODO).

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	PASSRODO	POA	POT	SOR	IND	SERV	APU	IMP
PASSRODO	1.00000	0.91900	0.92017	0.92538	0.83083	0.92496	0.72218	0.89831
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	531	531	531	531	531	531	531	531
POA	0.91900	1.00000	0.99979	0.99106	0.93669	0.97482	0.67889	0.97582
	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558
POT	0.92017	0.99979	1.00000	0.99155	0.93852	0.97533	0.67280	0.97718
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558
SOR	0.92538	0.99106	0.99155	1.00000	0.92692	0.99074	0.71574	0.97992
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558
IND	0.83083	0.93669	0.93852	0.92692	1.00000	0.90604	0.55405	0.93456
	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558
SERV	0.92496	0.97482	0.97533	0.99074	0.90604	1.00000	0.77086	0.97737
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
	531	558	558	558	558	558	558	558
APU	0.72218	0.67889	0.67280	0.71574	0.55405	0.77086	1.00000	0.66245
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
	531	558	558	558	558	558	558	558
IMP	0.89831	0.97582	0.97718	0.97992	0.93456	0.97737	0.66245	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
	531	558	558	558	558	558	558	558

Figura 5.11: Matriz de Correlação de Variáveis Econômicas

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	2.732818E14	3.904026E13	574.23	<.0001
Error	523	3.555756E13	67987680472		
Corrected Total	530	3.088394E14			

Root MSE	260744	R-Square	0.8849
Dependent Mean	251433	Adj R-Sq	0.8833
Coeff Var	103.70323		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	49278	14641	3.37	0.0008
POA	1	-16.24000	2.16553	-7.50	<.0001
POT	1	16.31368	1.98968	8.20	<.0001
SOR	1	0.00709	0.01625	0.44	0.6627
IND	1	-0.02478	0.00610	-4.06	<.0001
SERV	1	-0.00477	0.00857	-0.56	0.5782
APU	1	0.04886	0.00823	5.94	<.0001
IMP	1	-0.01740	0.01532	-1.14	0.2567

Figura 5.12: Modelo de Regressão para PASSRODO das Variáveis Econômicas

Portanto é proposto pelo estudo a utilização das técnicas multivariadas a fim de sanar esse problema. As componentes estão na Figura 5.13:

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	6.34100457	5.81150394	0.9059	0.9059
2	0.52950063	0.44981373	0.0756	0.9815
3	0.07968691	0.04506022	0.0114	0.9929
4	0.03462669	0.02248964	0.0049	0.9978
5	0.01213705	0.00921030	0.0017	0.9996
6	0.00292674	0.00280933	0.0004	1.0000
7	0.00011741		0.0000	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
POA	0.393248	-.119936	-.198244	0.467141	0.285483	-.134997
POT	0.393217	-.131900	-.202523	0.428100	0.216622	-.185446
SOR	0.395132	-.046025	-.221985	0.085456	-.508659	0.725557
IND	0.372592	-.337981	0.861706	-.001923	-.065904	-.006674
SERV	0.394310	0.072419	-.190543	-.318746	-.557929	-.621846
APU	0.296453	0.909284	0.240216	0.044581	0.144666	0.065832
IMP	0.390407	-.140804	-.184007	-.698295	0.525627	0.172864

Figura 5.13: Componentes Principais Econômicas

Com as componentes formadas, verifica-se que o uso de duas componentes para o estudo já se mostra satisfatório, com $98,15\%$ da variância explicada, entretanto, a

interpretação da mesma é complicada. Portanto através da análise fatorial, é possível que se faça a rotação de fatores, melhorando assim a interpretação dos fatores. Os resultados estão na Figura 5.14:

Eigenvalues of the Correlation Matrix: Total = 7 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	6.34100457	5.81150394	0.9059	0.9059
2	0.52950063	0.44981373	0.0756	0.9815
3	0.07968691	0.04506022	0.0114	0.9929
4	0.03462669	0.02248964	0.0049	0.9978
5	0.01213705	0.00921030	0.0017	0.9996
6	0.00292674	0.00280933	0.0004	1.0000
7	0.00011741		0.0000	1.0000

2 factors will be retained by the NFACTOR criterion.

Rotated Factor Pattern

	Factor1	Factor2
POA	0.91153	0.39665
POT	0.91562	0.38896
SOR	0.88999	0.44616
IND	0.94168	0.23242
SERV	0.84698	0.52087
APU	0.33947	0.93799
IMP	0.91250	0.37989

Figura 5.14: Fatores Rotacionados Econômicos

Verifica-se que no primeiro fator o maior coeficiente é apresentado pela variável IND e no segundo fator o maior coeficiente é o da variável APU. As Figuras 5.15, 5.16 e 5.17 apresentam a matriz de correlação dos fatores, a regressão construída e o gráfico de Resíduos por Valores Preditos respectivamente.

Pearson Correlation Coefficients, N = 558
Prob > |r| under H0: Rho=0

	Factor1	Factor2
Factor1	1.00000	0.00000 1.0000
Factor2	0.00000 1.0000	1.00000

Figura 5.15: Correlação dos Fatores Econômicos

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.65025E14	1.325125E14	1596.89	<.0001
Error	528	4.381438E13	82981777023		
Corrected Total	530	3.088394E14			

Root MSE	288066	R-Square	0.8581
Dependent Mean	251433	Adj R-Sq	0.8576
Coeff Var	114.56938		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	244201	12502	19.53	<.0001
Factor1	1	584672	12218	47.85	<.0001
Factor2	1	367562	12214	30.09	<.0001

Figura 5.16: Regressão dos Fatores Econômicos

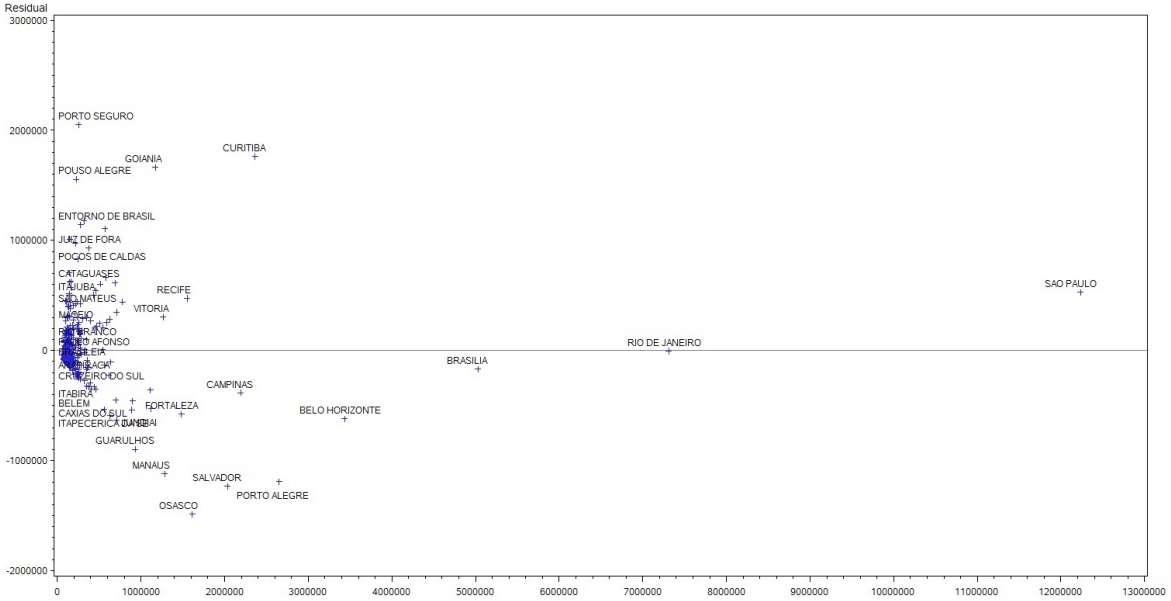

Figura 5.17: Resíduo x Valor Predito - Variáveis Econômicas

A partir da regressão construída com as componentes, verifica-se que o coeficiente R^2 é de aproximadamente 85% o que aparentemente indica que o modelo é bem ajustado aos dados, porém ao observarmos o gráfico de resíduos, verifica-se a presença de heterocedasticidade.

5.5 VARIÁVEIS SOCIAIS

Assim como nas variáveis anteriores, observa-se na matriz de correlações (Figura 5.18), que todas as variáveis são altamente correlacionadas, com correlações acima $0,90$. O modelo está na Figura 5.19.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	PASSRODO	POPR	POPHR	POPMR	POPCR	POPESPR	POPEVGR
PASSRODO	1.00000	0.89403	0.89432	0.89365	0.89573	0.88865	0.87418
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	531	531	531	531	531	528	531
POPR	0.89403	1.00000	0.99993	0.99994	0.98953	0.96023	0.98614
	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	553	558
POPHR	0.89432	0.99993	1.00000	0.99974	0.99001	0.95867	0.98623
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	553	558
POPMR	0.89365	0.99994	0.99974	1.00000	0.98897	0.96153	0.98594
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
	531	558	558	558	558	553	558
POPCR	0.89573	0.98953	0.99001	0.98897	1.00000	0.93580	0.95780
	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
	531	558	558	558	558	553	558
POPESPR	0.88865	0.96023	0.95867	0.96153	0.93580	1.00000	0.94178
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
	528	553	553	553	553	553	553
POPEVGR	0.87418	0.98614	0.98623	0.98594	0.95780	0.94178	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
	531	558	558	558	558	553	558

Figura 5.18: Matriz de Correlação - Variáveis Sociais

Os resultados da Análise de Componentes Principais está na Figura 5.20.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.607568E14	5.215137E13	567.72	<.0001
Error	522	4.795159E13	91861287125		
Corrected Total	527	3.087084E14			

Root MSE	303086	R-Square	0.8447
Dependent Mean	252549	Adj R-Sq	0.8432
Coeff Var	120.01073		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19891	20688	0.96	0.3368
POPR	B	-4.50939	1.41305	-3.19	0.0015
POPHR	B	2.35715	3.18096	0.74	0.4590
POPMR	0	0	.	.	.
POPCR	1	3.86268	0.42075	9.18	<.0001
POPESPR	1	19.53920	1.70384	11.47	<.0001
POPEVGR	1	5.04735	0.71853	7.02	<.0001

Figura 5.19: Modelo de Regressão para PASSRODO das Variáveis Sociais

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.88145961	5.80662096	0.9802	0.9802
2	0.07483865	0.03304171	0.0125	0.9927
3	0.04179694	0.04006226	0.0070	0.9997
4	0.00173468	0.00156456	0.0003	1.0000
5	0.00017012	0.00017012	0.0000	1.0000
6	0.00000000		0.0000	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
POPR	0.412057	-.121987	0.014794	-.384954	-.028650
POPHR	0.411976	-.142872	0.015738	-.279003	-.762384
POPMR	0.412079	-.103047	0.013937	-.480909	0.636197
POPCR	0.406966	-.355595	0.616464	0.565112	0.092472
POPESPR	0.399607	0.896847	0.118103	0.148379	-.003994
POPEVGR	0.406654	-.152665	-.778051	0.448694	0.068091

Figura 5.20: Componentes Principais Sociais

Verifica-se que 2 componentes explicam $99,27\%$ da variância dos dados, mas não é possível interpretá-las de maneira clara, assim, utiliza-se a técnica de análise fatorial para gerar os mesmos fatores e utilizar da rotação para obter uma melhor interpretação.

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	5.88145961	5.80662096	0.9802	0.9802
2	0.07483865	0.03304171	0.0125	0.9927
3	0.04179694	0.04006226	0.0070	0.9997
4	0.00173468	0.00156456	0.0003	1.0000
5	0.00017012	0.00017012	0.0000	1.0000
6	0.00000000		0.0000	1.0000

Rotated Factor Pattern

	Factor1	Factor2
POPR	0.78992	0.61299
POPHR	0.79342	0.60847
POPMR	0.78665	0.61701
POPCR	0.82127	0.55595
POPESPR	0.58857	0.80807
POPEVGR	0.78520	0.59817

Figura 5.21: Fatores Rotacionados Sociais

A partir dos fatores rotacionados (Figura 5.21), observa-se no primeiro fator a variável POPCR apresenta o maior coeficiente e no segundo fator apresenta a variável POPESPR apresenta o maior coeficiente.

Como nos grupos anteriores, a matriz de correlação dos fatores, a regressão construída e o gráfico de Resíduos por Valores Preditos estão nas Figuras 5.22, 5.23 e 5.24.

Pearson Correlation Coefficients, N = 553
Prob > |r| under H0: Rho=0

	Factor1	Factor2
Factor1	1.00000	0.00000 1.0000
Factor2	0.00000 1.0000	1.00000

Figura 5.22: Correlação dos Fatores Sociais

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.505149E14	1.252574E14	1130.02	<.0001
Error	525	5.819356E13	1.108449E11		
Corrected Total	527	3.087084E14			

Root MSE	332934	R-Square	0.8115
Dependent Mean	252549	Adj R-Sq	0.8108
Coeff Var	131.82920		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	245159	14491	16.92	<.0001
Factor1	1	489725	14254	34.36	<.0001
Factor2	1	464015	14190	32.70	<.0001

Figura 5.23: Regressão dos Fatores Sociais

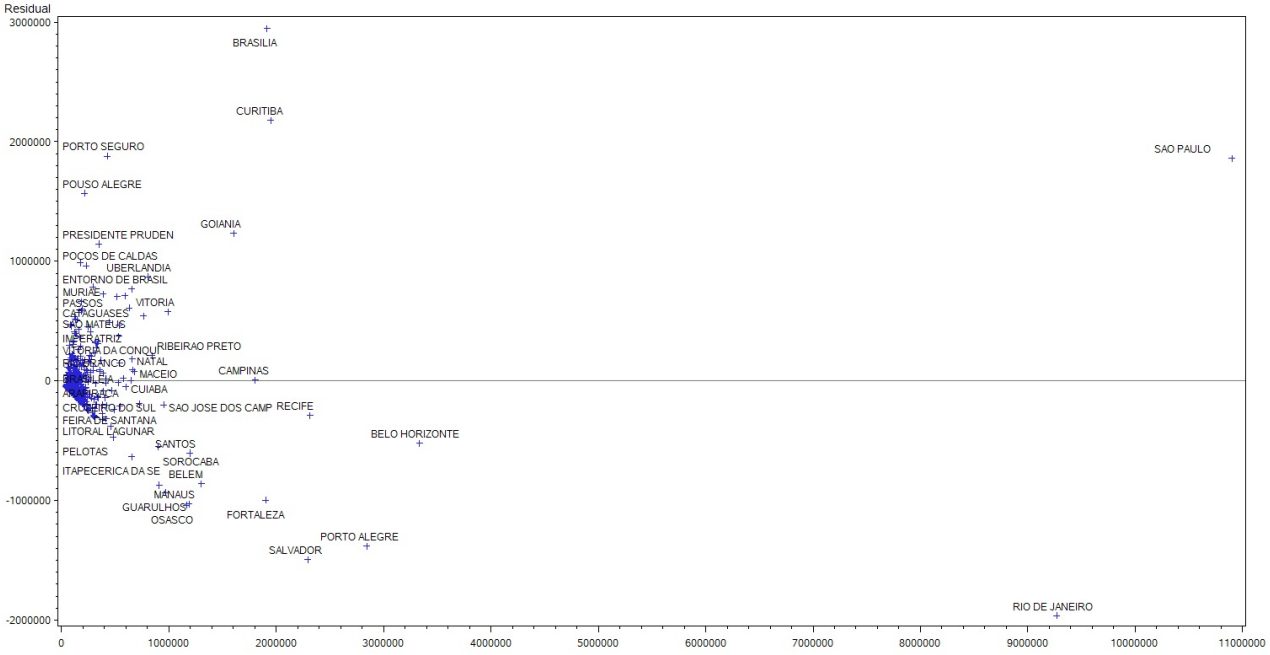

Figura 5.24: Resíduo x Valor Predito - Variáveis Sociais

A partir da regressão construída com as componentes, verifica-se que o coeficiente R^2 é de aproximadamente 82% o que aparentemente indica que o modelo é bem ajustado aos dados, porém ao observarmos o gráfico de resíduos, verificamos a

presença de heterocedasticidade.

5.6 VARIÁVEIS DE FROTA

Assim com nos demais grupos de variáveis, observa-se na matriz de correlações (Figura 5.25) que todas as variáveis são altamente correlacionadas, com correlações de pelo menos $0,78$. O resultado do modelo está na Figura 5.26.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	PASSRODO	AUTOMOVEL	CAMINHAO	TRATOR	CAMINHONETE	CAMIONETA	MICROBUS	MOTO	MOTONETA	ONIBUS
PASSRODO	1.00000	0.92265	0.87939	0.78156	0.90576	0.91556	0.89953	0.87707	0.77951	0.91095
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	531	531	531	531	531	531	531	531	531	531
AUTOMOVEL	0.92265	1.00000	0.95782	0.86072	0.97753	0.99628	0.97680	0.93952	0.80660	0.97414
	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558	558	558
CAMINHAO	0.87939	0.95782	1.00000	0.93643	0.97961	0.94009	0.92357	0.95712	0.84628	0.95119
	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558	558	558
TRATOR	0.78156	0.86072	0.93643	1.00000	0.90753	0.83887	0.79550	0.86674	0.83940	0.82758
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558	558	558
CAMINHONETE	0.90576	0.97753	0.97961	0.90753	1.00000	0.96582	0.93685	0.96487	0.84504	0.95698
	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558	558	558
CAMIONETA	0.91556	0.99628	0.94009	0.83887	0.96582	1.00000	0.97857	0.92745	0.79157	0.96923
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558	558	558
MICROBUS	0.89953	0.97680	0.92357	0.79550	0.93685	0.97857	1.00000	0.90991	0.76254	0.97884
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
	531	558	558	558	558	558	558	558	558	558
MOTO	0.87707	0.93952	0.95712	0.86674	0.96487	0.92745	0.90991	1.00000	0.87958	0.93625
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
	531	558	558	558	558	558	558	558	558	558
MOTONETA	0.77951	0.80660	0.84628	0.83940	0.84504	0.79157	0.76254	0.87958	1.00000	0.78448
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
	531	558	558	558	558	558	558	558	558	558
ONIBUS	0.91095	0.97414	0.95119	0.82758	0.95698	0.96923	0.97884	0.93625	0.78448	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
	531	558	558	558	558	558	558	558	558	558

Figura 5.25: Matriz de Correlação de Variáveis de Frota

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	2.689519E14	2.988355E13	390.33	<.0001
Error	521	3.98875E13	76559496313		
Corrected Total	530	3.088394E14			

Root MSE	276694	R-Square	0.8708
Dependent Mean	251433	Adj R-Sq	0.8686
Coeff Var	110.04662		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	54725	17411	3.14	0.0018
AUTOMOVEL	1	4.46489	0.68636	6.51	<.0001
CAMINHAO	1	-38.75564	10.25598	-3.78	0.0002
TRATOR	1	-16.45091	22.07152	-0.75	0.4564
CAMINHONETE	1	4.67674	3.83672	1.22	0.2234
CAMIONETA	1	-33.69717	8.87717	-3.80	0.0002
MICROBUS	1	-104.31796	38.26140	-2.73	0.0066
MOTO	1	-1.17600	1.01735	-1.16	0.2482
MOTONETA	1	14.89199	3.11155	4.79	<.0001
ONIBUS	1	146.89129	26.63939	5.51	<.0001

Figura 5.26: Modelo de Regressão para PASSRODO de Variáveis de Frota

O resultado das componentes obtidas está na Figura 5.27 e o da Análise Fatorial na Figura 5.28.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	8.27604995	7.88638878	0.9196	0.9196
2	0.38966117	0.21038434	0.0433	0.9629
3	0.17927684	0.10776384	0.0199	0.9828
4	0.07151300	0.02977752	0.0079	0.9907
5	0.04173547	0.02297328	0.0046	0.9954
6	0.01876219	0.00655771	0.0021	0.9974
7	0.01220448	0.00332154	0.0014	0.9988
8	0.00888294	0.00696899	0.0010	0.9998
9	0.00191395		0.0002	1.0000

Eigenvectors

	Prin1	Prin2	Prin3
AUTOMOVEL	0.342503	-.223357	-.004715
CAMINHÃO	0.342272	0.077299	-.267881
TRATOR	0.316759	0.451610	-.680913
CAMINHONETE	0.344084	-.007899	-.115968
CAMIONETA	0.339271	-.278716	0.047187
MICROBUS	0.333516	-.386279	0.151985
MOTO	0.337754	0.098733	0.215278
MOTONETA	0.303431	0.648826	0.611969
ONIBUS	0.338120	-.287238	0.070364

Figura 5.27: Componentes Principais das Variáveis de Frota

As componentes geradas explicam $96,29\%$ da variabilidade dos dados com 2 componentes, porém não se obtém a partir dessas componentes uma interpretação

satisfatória, portanto são gerados os fatores rotacionados, a fim de obter uma visualização melhor dos fatores.

Eigenvalues of the Correlation Matrix: Total = 9 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	8.27604995	7.88638878	0.9196	0.9196
2	0.38966117	0.21038434	0.0433	0.9629
3	0.17927684	0.10776384	0.0199	0.9828
4	0.07151300	0.02977752	0.0079	0.9907
5	0.04173547	0.02297328	0.0046	0.9954
6	0.01876219	0.00655771	0.0021	0.9974
7	0.01220448	0.00332154	0.0014	0.9988
8	0.00888294	0.00696899	0.0010	0.9998
9	0.00191395		0.0002	1.0000

Rotated Factor Pattern

	Factor1	Factor2
AUTOMOVEL	0.84464	0.52619
CAMINHÃO	0.72355	0.66958
TRATOR	0.51719	0.80148
CAMINHONETE	0.76172	0.63218
CAMIONETA	0.85972	0.49373
MICROBUS	0.89016	0.43164
MOTO	0.70500	0.67148
MOTONETA	0.40872	0.87119
ONIBUS	0.86060	0.48753

Figura 5.28: Fatores Rotacionados das Variáveis de Frota

A partir dos fatores rotacionados, verifica-se no primeiro fator um domínio das variáveis MICROBUS, AUTOMÓVEL, CAMIONETA e ÔNIBUS. Já o segundo fator apresenta um domínio das variáveis MOTONETA e TRATOR.

Como nos grupos anteriores, a matriz de correlação dos fatores, a regressão construída e o gráfico de Resíduos por Valores Preditos estão nas Figuras 5.29, 5.30 e 5.31.

Pearson Correlation Coefficients, N = 558
Prob > |r| under H0: Rho=0

	Factor1	Factor2
Factor1	1.00000	0.00000 1.0000
Factor2	0.00000 1.0000	1.00000

Figura 5.29: Correlação dos Fatores das Variáveis de Frota

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.608849E14	1.304424E14	1436.23	<.0001
Error	528	4.795451E13	90822930365		
Corrected Total	530	3.088394E14			

Root MSE	301368	R-Square	0.8447
Dependent Mean	251433	Adj R-Sq	0.8441
Coeff Var	119.86018		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	241021	13083	18.42	<.0001
Factor1	1	571585	12773	44.75	<.0001
Factor2	1	379310	12880	29.45	<.0001

Figura 5.30: Modelo de Regressão dos Fatores de Frota para PASSRODO

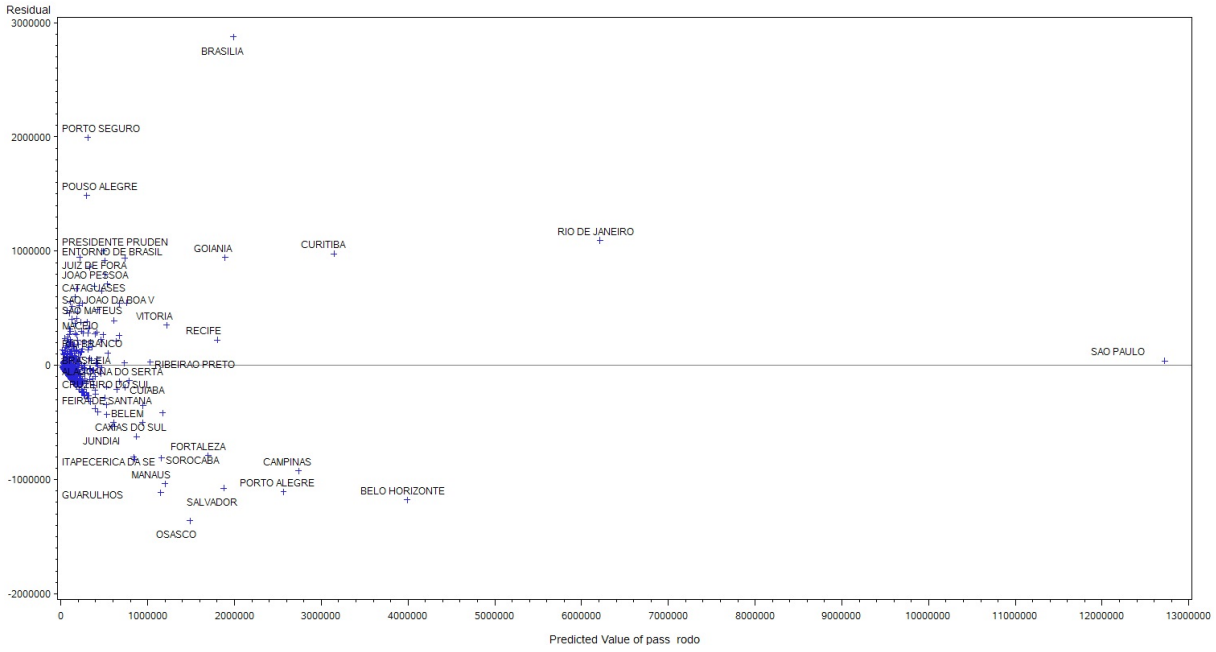

Figura 5.31: Resíduo x Valor Predito - Variáveis de Frota

A partir da regressão construída com as componentes, verifica-se que o coeficiente R^2 é de aproximadamente 78% o que aparentemente indica que o modelo é bem ajustado aos dados, porém ao observarmos o gráfico de resíduos, verificamos

novamente que os dados apresentam heterocedasticidade.

Como visto nos gráficos dos resíduos, microrregiões como São Paulo, Rio de Janeiro e Brasília ficam muito distantes do restante das microrregiões. Devido a essa heterogeneidade das regiões, sugere-se a construção de *clusters* a fim de sanar esse problema, onde a partir dos *clusters* gerados, podem ser construídos modelos de regressão para cada *cluster*. Portanto, a próxima seção, através da análise de *cluster*, verificará a existência de conglomerados para as microrregiões do Brasil e a possibilidade de se construir modelos de regressão a partir deles.

5.7 ANÁLISE DE CLUSTER

A partir do método hierárquico de ward é construído um dendograma, a fim de obter uma visualização dos possíveis *clusters* gerados a partir das variáveis originais. O dendograma construído está na Figura 5.32.

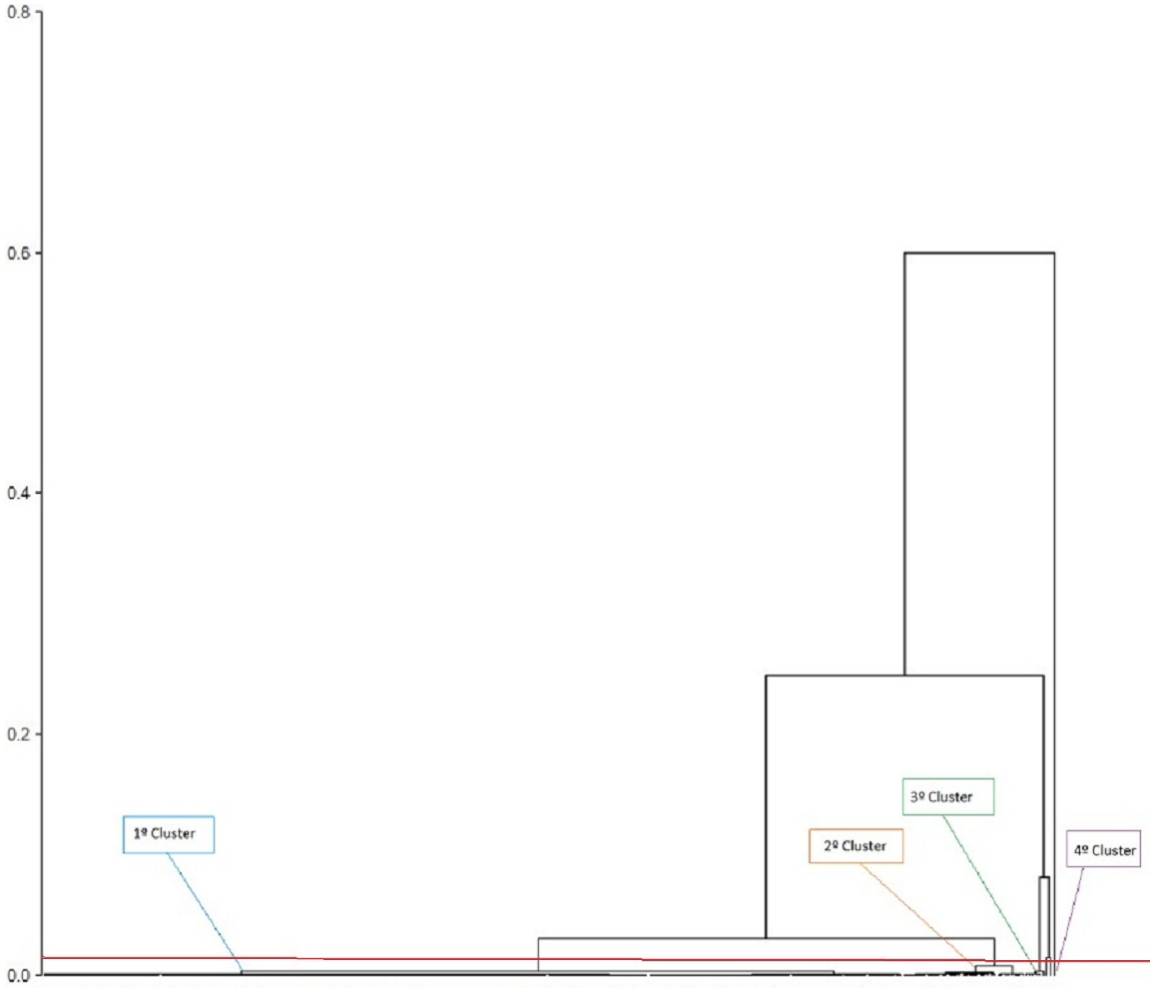

Figura 5.32: Dendograma - Variáveis Originais

Observa-se no dendograma construído que existem 4 *clusters*. Para se ter uma melhor visualização de quais microrregiões pertencem aos *clusters* obtidos, os *clusters* são gerados novamente a partir do método hierárquico e do método de *k-médias*

utilizando as variáveis originais e os fatores obtidos para fim comparativo.

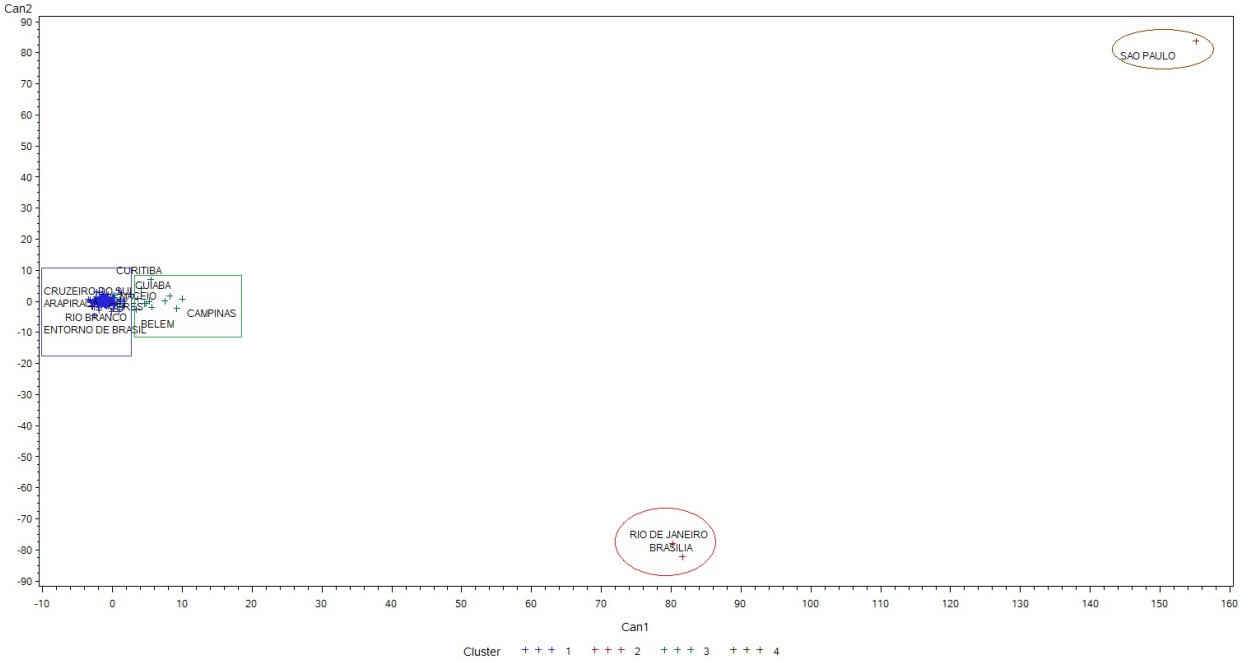

Figura 5.33: *Clusters* Gerados com Variáveis Originais - Método de *K-Médias*

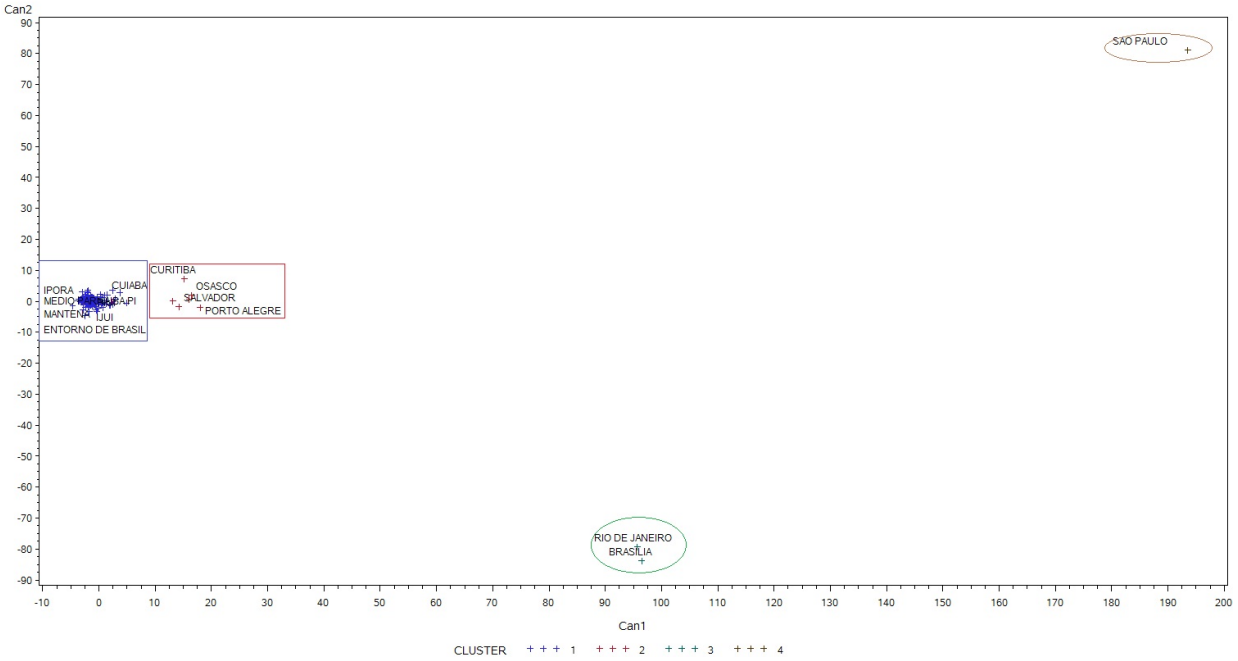

Figura 5.34: *Clusters* Gerados com Variáveis Originais - Método de Ward

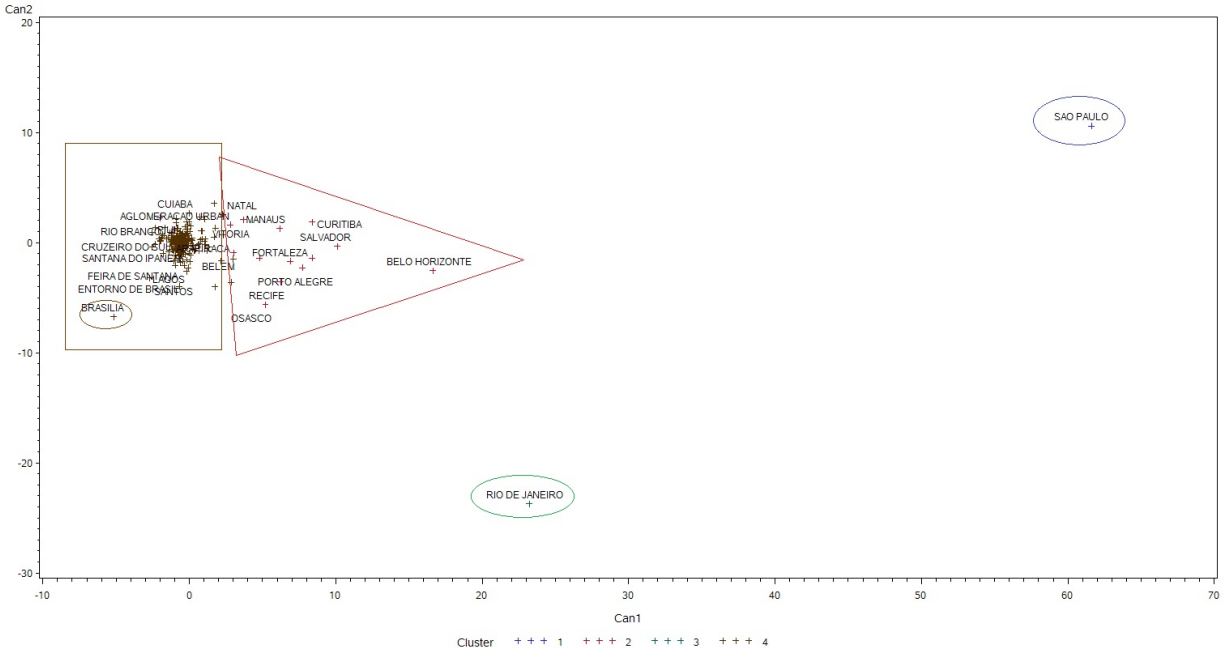

Figura 5.35: *Clusters* Gerados com Fatores - Método de *K-Médias*

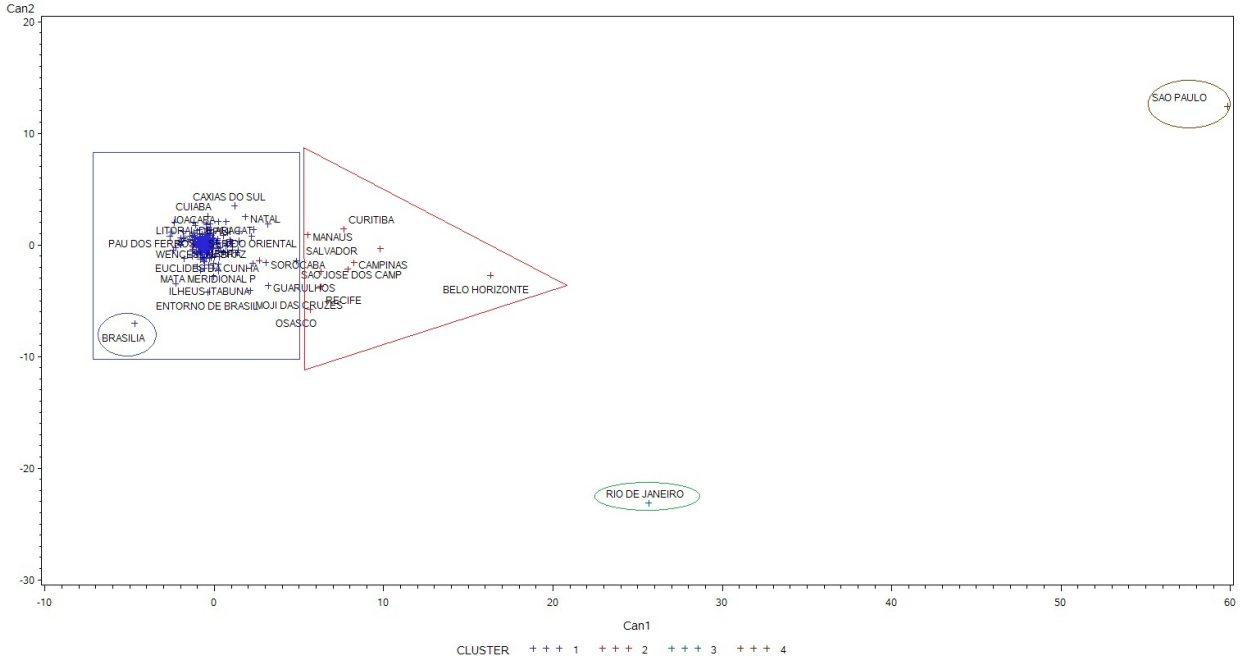

Figura 5.36: *Clusters* Gerados com Fatores - Método de Ward

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	ln_PASSRODO	Fator1_ escol	Fator1_ econ	Fator1_ soc	Fator1_ frota
ln_PASSRODO	1.00000	0.46900	-0.08230	0.49146	0.08345
		<.0001	0.1363	<.0001	0.1309
	329	329	329	329	329
Fator1_escol	0.46900	1.00000	0.45290	0.91183	0.27884
	<.0001		<.0001	<.0001	<.0001
	329	352	352	352	352
Fator1_econ	-0.08230	0.45290	1.00000	0.37427	-0.14628
	0.1363	<.0001		<.0001	0.0060
	329	352	352	352	352
Fator1_soc	0.49146	0.91183	0.37427	1.00000	0.10028
	<.0001	<.0001	<.0001		0.0602
	329	352	352	352	352
Fator1_frota	0.08345	0.27884	-0.14628	0.10028	1.00000
	0.1309	<.0001	0.0060	0.0602	
	329	352	352	352	352

Figura 5.37: Matriz de Correlação dos Fatores por Grupos

Com a construção dos *clusters*, pode ser observado que tanto para as variáveis originais, quanto para os fatores, não houve diferença significativa na construção dos *clusters* entre o método hierárquico ou o método de *k-médias*, mesmo o método hierárquico não utilizando a mesma quantidade de observações para a construção do *cluster*.

Já comparando os *clusters* de variáveis originais e os fatores, verifica-se uma diferença significativa entre os *clusters* construídos, onde pode-se notar visivelmente, que a cidade de Brasília mudou de *cluster*.

Ressalta-se que a diferença entre variáveis originais e os fatores ocorre pela forma que foram gerados os fatores, sendo obtidos fatores para cada tipo de variável presente no estudo. Logo os fatores obtidos apresentam correlações significativas, verificadas na Figura 5.37, o que pode vir a influenciar na construção dos *clusters*.

Para verificar essa diferença, a Figura 5.38 apresenta a análise de cluster com as componentes geradas para todas as variáveis, assim verificando que é possível

utilizar análise de clusters com as variáveis originais ou com os fatores gerados.

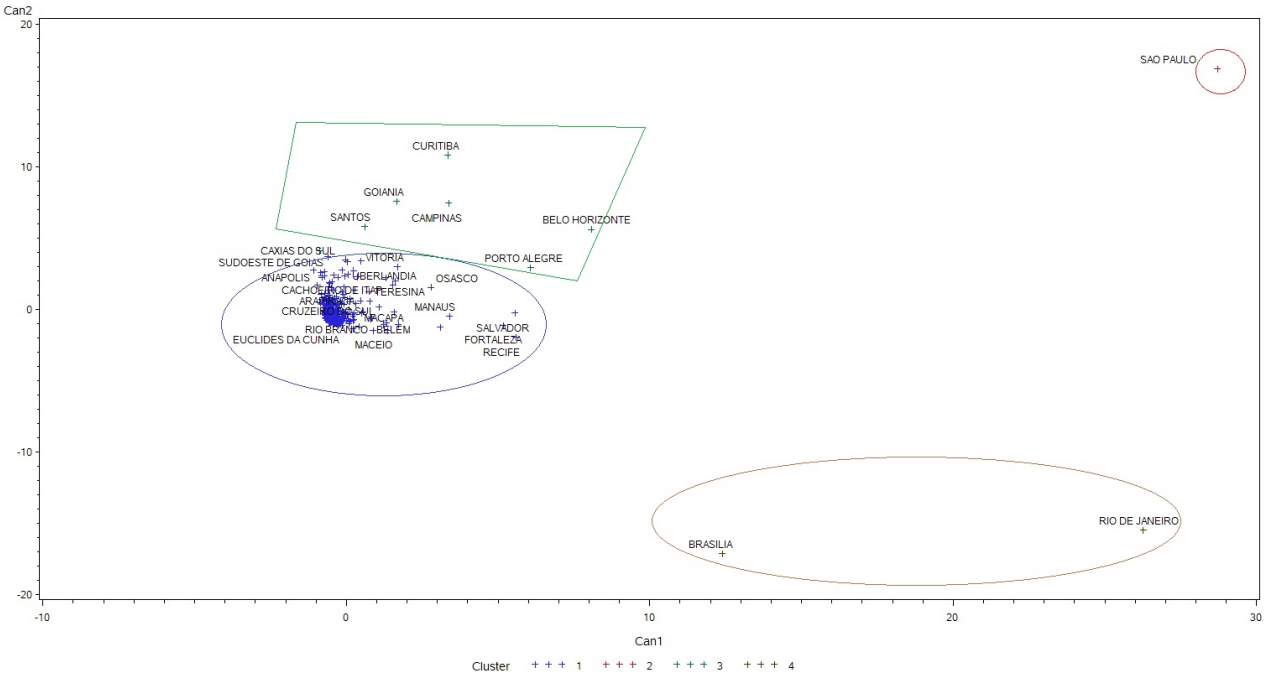

Figura 5.38: Cluster Gerados com Fatores Gerais - Método de *K-Médias*

Pode-se observar que os *clusters* obtidos apresentam diferentes tamanhos, onde foram construídos *clusters* com somente uma microrregião, e outros com um número maior de microrregiões. Portanto, não é possível a construção de modelos de regressão para os *clusters* gerados. Logo, o estudo sugere, a fim de sanar o problema de tamanho dos *clusters* a utilização de UF's ao invés dos *clusters* na estimativa do modelo geral.

5.8 MODELO PARA MATRIZ OD UF

Após ser feita a análise para os diferentes tipos de variáveis presentes no modelo, percebe-se que a construção de um modelo de regressão por *cluster* não é viável, pois como visto na seção anterior, existem *clusters* com muito poucas microrregiões, assim não sendo viável a construção de modelos de regressão por *clusters*. Portanto, é proposto pelo estudo para evitar esse problema, que o modelo final construído seja por UF.

Para a construção do modelo, é necessário gerar novas componentes ou fatores por UF, portanto, como verificado anteriormente a correlação entre as variáveis e por consequência o modelo de regressão apresentar multicolinearidade, serão obtidos os fatores rotacionados das variáveis originais, para melhor interpretação delas. As componentes geradas para os grupos apresentados anteriormente estão nas Figuras 5.39 a 5.42.

Eigenvalues of the Correlation Matrix: Total = 5 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	4.92333082	4.85437050	0.9847	0.9847
2	0.06896032	0.06446292	0.0138	0.9985
3	0.00449740	0.00176834	0.0009	0.9994
4	0.00272906	0.00224666	0.0005	0.9999
5	0.00048240		0.0001	1.0000

Rotated Factor Pattern

	Factor1	Factor2
MEF	0.81994	0.57180
MEM	0.76694	0.63925
MES	0.58455	0.81112
PRA	0.72528	0.68641
PRFC	0.79564	0.60527

Figura 5.39: Fatores Rotacionados Escolares UF

Eigenvalues of the Correlation Matrix: Total = 7 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	6.68406888	6.40183911	0.9549	0.9549
2	0.28222977	0.25611024	0.0403	0.9952
3	0.02611952	0.01938330	0.0037	0.9989
4	0.00673622	0.00614691	0.0010	0.9999
5	0.00058931	0.00041085	0.0001	1.0000
6	0.00017845	0.00010060	0.0000	1.0000
7	0.00007785		0.0000	1.0000

Rotated Factor Pattern

	Factor1	Factor2
POA	0.88006	0.46964
POT	0.88296	0.46377
SOR	0.85387	0.51792
IND	0.89821	0.43072
SERV	0.83493	0.54685
APU	0.46352	0.88587
IMP	0.88172	0.46230

Figura 5.40: Fatores Rotacionados Econômicos UF

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	5.85036519	5.72289602	0.9751	0.9751
2	0.12746917	0.10713168	0.0212	0.9963
3	0.02033749	0.01862732	0.0034	0.9997
4	0.00171017	0.00159220	0.0003	1.0000
5	0.00011797	0.00011797	0.0000	1.0000
6	0.00000000		0.0000	1.0000

Rotated Factor Pattern

	Factor1	Factor2
POPR	0.78175	0.62339
POPHR	0.78664	0.61725
POPMR	0.77701	0.62910
POPCR	0.85442	0.51623
POPESPR	0.53485	0.84171
POPEVGR	0.66948	0.73450

Figura 5.41: Fatores Rotacionados Sociais UF

Eigenvalues of the Correlation Matrix: Total = 9 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	8.64534284	8.44904934	0.9606	0.9606
2	0.19629350	0.11833900	0.0218	0.9824
3	0.07795450	0.02070100	0.0087	0.9911
4	0.05725350	0.04375374	0.0064	0.9974
5	0.01349977	0.00753559	0.0015	0.9989
6	0.00596418	0.00323781	0.0007	0.9996
7	0.00272637	0.00212768	0.0003	0.9999
8	0.00059869	0.00023204	0.0001	1.0000
9	0.00036665		0.0000	1.0000

Rotated Factor Pattern

	Factor1	Factor2
AUTOMOVEL	0.77164	0.62646
CAMINHÃO	0.67958	0.72754
TRATOR	0.51360	0.84940
CAMINHONETE	0.70461	0.70632
CAMIONETA	0.79947	0.58499
MICROBUS	0.86074	0.50506
MOTO	0.73518	0.64031
MOTONETA	0.57553	0.79486
ONIBUS	0.80449	0.58647

Figura 5.42: Fatores Rotacionados de Frota UF

Pode-se observar que em relação aos fatores por microrregião, houve um pequeno aumento na variabilidade explicada pelas componentes, sendo mantido na maioria dos casos o cenário proposto nas análises por microrregião, exceto pelas variáveis de frota, que agora verifica-se no primeiro fator um maior coeficiente para a variável MICRO-ÔNIBUS. Porém no segundo fator é verificado um maior coeficiente para a variável TRATOR.

Após a obtenção dos fatores aplicou-se a transformação apresentadada na seção 2.3.1. O modelo resultante está na Figura 5.43.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	175.93750	35.18750	46.60	<.0001
Error	322	243.14193	0.75510		
Corrected Total	327	419.07943			

Root MSE	0.86896	R-Square	0.4198
Dependent Mean	14.67780	Adj R-Sq	0.4108
Coeff Var	5.92026		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	16.86930	0.56418	29.90	<.0001
Fator1_escol	1	1.21821	0.23884	5.10	<.0001
Fator1_econ	1	-0.50374	0.06177	-8.16	<.0001
Fator1_soc	1	0.13629	0.14405	0.95	0.3448
Fator1_frota	1	-0.39050	0.11330	-3.45	0.0006
log_dist	1	-0.31017	0.07689	-4.03	<.0001

Figura 5.43: Modelo de Regressão dos Fatores UF

Verifica-se no modelo proposto um coeficiente de determinação baixo, com aproximadamente 42% da variabilidade dos dados explicadas, além também do problema da multicolinearidade dos dados, assim, retornando para o problema inicial do estudo.

Portanto, um possível modelo para a estimação da matriz OD tem a seguinte forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \tag{5.1}$$

onde:

- Y são as viagens realizadas pelo modo ônibus entre uma origem e um destino;
- X_1 é o log da população estimada na UF de origem;
- X_2 é o log da população estimada na UF de destino;
- X_3 é o log dos salários;

- X_4 é o log da distância entre as origens e destinos.

Finalmente o modelo para estimação da Matriz OD é dado na Figura 5.44.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1377.54669	344.38667	106.21	<.0001
Error	252	817.12971	3.24258		
Corrected Total	256	2194.67640			

Root MSE	1.80072	R-Square	0.6277
Dependent Mean	9.18246	Adj R-Sq	0.6218
Coeff Var	19.61040		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.74309	3.18317	2.75	0.0065
log_populacaoestimada_o	1	0.95503	0.17444	5.47	<.0001
log_populacaoestimada_d	1	0.49607	0.18111	2.74	0.0066
log_salarios2011_cap_o	1	0.54683	0.24902	2.20	0.0290
log_dist	1	-3.21446	0.17767	-18.09	<.0001

Figura 5.44: Modelo de Matriz OD

Nesse caso verifica-se um coeficiente de determinação de aproximadamente 62%. Nota-se também que os β estão com os sinais de acordo com o esperado, com destaque para o sinal negativo da variável *log_dist*, mostrando que quanto maior a distância entre as cidades, menor a incidência de viagens nesse trajeto.

Capítulo 6

CONCLUSÃO

Como pode ser observado nas análises, a tentativa de construir um modelo geral utilizando a técnica de componentes principais não foi viável, pois as componentes obtidas a partir do mesmo são muito difíceis de interpretar. No entanto, a ideia de separar as variáveis selecionadas pelo estudo em grupos foi de grande valia para o desenvolvimento das análises.

Como era esperado, a utilização de componentes ou fatores na construção de um modelo de regressão linear é útil quando é constatado o problema de multicolinearidade dos dados, onde o modelo de regressão com os fatores apresenta um coeficiente de determinação similar ao coeficiente do modelo com as variáveis originais.

A tentativa de construir *clusters* a partir das componentes/fatores gerados a partir de grupos de variáveis não teve o resultado esperado da igualdade entre os *clusters* gerados através das variáveis originais. Essa diferença se deve à obtenção dos fatores por grupos de variáveis, o que tornou os fatores não ortogonais.

A construção de um modelo de regressão para os fatores por UF não apresentou um resultado satisfatório, pois o mesmo não explicava muito sobre a variabilidade dos dados, além de apresentar o problema da multicolinearidade dos dados.

Uma alternativa para futuras análises sobre o assunto é utilizar variáveis proporcionais à estrutura da microrregião, como por exemplo população matriculada no ensino superior pela população total da microrregião alvo.

Referências Bibliográficas

ANTT (2013). Transporte de passageiros. Technical report, Agência Nacional de Transportes Terrestres. URL `https://appweb.antt.gov.br/AV/AvPublico/index.asp`. Acesso em 24 set. 2013.

Calixto, I. C. A. C. (2011). Proposta de um método de estimação de matrizes origem-destino baseado em programação linear fuzzy para redes viárias brasileiras congestionadas. Master's thesis, UFG. URL `http://www.inf.ufg.br/mestrado/sites/www.inf.ufg.br.mestrado/files/uploads/Dissertacoes/IacerCoimbra.pdf`. Acesso em 02 set. 2013.

Chatterjee, S. & Hadi, A. S. (2006). *Regression Analysis by Example*, (4th ed.). Wiley.

Guerra, A. L. (2011). Determinação de matriz origem/destino utilizando dados do sistema de bilhetagem eletrônica. Master's thesis, UFMG. URL `http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/BUOS-8NWF3Z/disserta__o_bilhetagem_r12.pdf?sequence=1`. Acesso em 02 set. 2013.

IBGE (2007). Região de influência das cidades. Technical report, Instituto Brasileiro de Geografia e Pesquisa. http://www.ibge.gov.br/home/geociencias/geografia/regic.shtm.

IBGE (2013). Downloads ibge. Technical report. URL `http://downloads.ibge.gov.br/.Acessoem14nov.2013`, institution = Instituto Brasileiro de Geografia e Estatística,.

Joe H. Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of The American Statistical Association*, 58(301):236–244.

Johnson, R. A. & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, (6th ed.). Pearson.

Levine, N. (2010). *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations*, (3.3 ed.). Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC. July.

MTR (2012). Transporte rodoviário do brasil. Technical report, Ministério dos Transportes. URL `http://www2.transportes.gov.br/bit/02-rodo/rodo.html`. Acesso em 26 ago. 2013.

Romanatto, E. (2011). Análise de clusters e aplicação do modelo gravitacional aos fluxos de comércio do estado de goiás. *Indic. Econ. FEE, Porto Alegre*, 39(2):87–96.

SEFAZDF (2013). Gia-st/sicopi - guia nacional de informação e apuração do icms substituição tributária. Technical report, Secretaria de Estado de Fazenda do Distrito Federal. URL `http://www.fazenda.df.gov.br//area.cfm?id_area=393`. Acesso em 19 set. 2013.