APLICAÇÃO DE MÉTODOS DE ANÁLISE ESPACIAL NA

CARACTERIZAÇÃO DE ÁREAS DE RISCO A SAÚDE

Marilia Sá Carvalho

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DE PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA BIOMÉDICA.

Aprovada por:

Prof. Flávio Fonseca Nobre, Ph.D.

Prof. Cláudio José Struchiner, Ph.D.

Prof. Jurandir Nadal, D.Sc.

Prof. Christovam de Castro Barcellos, D.Sc.

Prof. Renato Martins Assunção, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

MAIO DE 1997

CARVALHO, MARILIA SÁ

Aplicação de métodos de análise espacial na caracterização de áreas de risco à saúde [Rio de Janeiro] 1997

IX,179, 29,7cm (COPPE/UFRJ, D.Sc., Engenharia Biomédica, 1997)

Tese - Universidade Federal do Rio de Janeiro, COPPE

1. Análise de dados espaciais

2. Perfil de risco à saúde

I. COPPE/UFRJ II. Título (série)

Para

Clarice, Carolina e Marcos,

que comigo viveram

todas as etapas de mais este "parto".

Vocês são maravilhosos.

UM POUCO DA HISTÓRIA E O MUITO OBRIGADA

Este trabalho certamente começa com o Cláudio (Struchiner), e o prazer que senti destrinchando os autovalores e autovetores da "*analise de correspondance*" durante o mestrado. Gostei, quis mais, e fui buscar na COPPE. Super bem recebida pelo Flávio (Nobre), e depois muito bem cuidada pelos meus orientadores substitutos. Primeiro foi o Ronney (Panerai), preciso, meticuloso e corretíssimo, pena que foi embora. Depois o Fernando (Infantosi), totalmente taqui-psíquico e multi-processador, e o Jurandir (Nadal), solidário e atencioso, ouvia minhas reclamações e buscava apoiar minhas loucas idéias. Todos os professores e funcionários do Programa de Engenharia Biomédica foram ótimos: competentes, dedicados, sérios. Neste tempo a proposta de trabalho era a malária em Tocantins: bem que eu tentei, mas não foi para a frente. Apenas uma queixa: eu queria mais responsabilidade e trabalho.

Como a malária não aconteceu, fui capturada, ou melhor, ajudei a soprar o furacão do projeto SIG-FIOCRUZ, financiado pelo PAPES/FIOCRUZ. Bons amigos - Chico (Viacava), Christovam (Barcellos), Jorge (Machado) - e muita confusão. A confusão passou, os amigos ficaram e me aturaram. Sem esquecer mamãe Adália (Figueiredo), com sua paciência e boa vontade, e a Fátima (Pina), com a malha de setores censitários urgente, para ontem. Neste projeto retornei ao meu velho objeto: populações urbanas, contrastes sociais, metrópoles caóticas.

A volta do Flávio trouxe no início muita contradição: ele me queria mais presente na COPPE, e eu estava até o pescoço com o projeto na FIOCRUZ; ele não acreditava que as coisas se viabilizassem, e eu era muito otimista; ele tentava me direcionar e eu não aceitava. Mas reclamar do orientador é de praxe, e depois deste reinício conturbado as coisas se acertaram. Acho que ele se acostumou com meu jeito, e passou a confiar mais em mim. E eu só posso dizer que o admiro, muito. Neste meio tempo minhas três mestrandas muito queridas - Eleonora (d'Orsi), Enirtes (Melo) e Tatiana (Campos) - me ensinaram a diversidade e, de tanto repetir para elas, a realmente aceitar que "a tese é apenas **um** trabalho". Convidei o Cláudio para co-orientador, e, mesmo com os vários cursos que fiz com ele, usufrui menos do que gostaria. Espero ter outras oportunidades.

Depois de ultrapassar alguns graves problemas pessoais, e comemorando simplesmente estarmos vivos (viu, D.Léa?), a partir do final de 1995 comecei a realmente objetivar a tese. Infelizmente tive que desistir da análise da mortalidade, proposta inicial, por problemas na localização dos endereços. Como saldo uma contribuição decisiva para a construção de um cadastro nacional de endereços e para a criação de um sistema automático de localização, que, após o precioso auxílio da Juliana (Pereira), poderá ser adequado aos caóticos sistemas de endereços existentes na cidade. Investi, por sugestão do Flávio, no estudo de técnicas de classificação "*fuzzy*", que também vai ficar para depois, afinal "a tese é apenas **um**...".

E também por sugestão do Flávio, simplifiquei, reduzi e estou chegando ao fim. Se fosse recomeçar, faria muita coisa diferente, o que prova o quanto aprendi no processo. Em uma apresentação informal do trabalho para o Prof. Trevor Bailey recebi sugestões, entre as quais a análise de *cluster* sobre os valores interpolados apresentada. Quando perguntei sobre a validade disso, ouvi o seguinte comentário: "tenta, se der bons resultados aí você começa a se preocupar". Bem, estou de fato preocupada: acho que tenho resultados interessantes e todo um mundo para explorar. Nestas páginas tem menos do que eu gostaria, mas a tese...é só a tese.

No balanço final destes seis anos (ufa!), um saldo muito positivo. Primeiro, o projeto SIG-FIOCRUZ, e tudo que se desenvolveu a partir dele: cadastro de endereços, sistema de localização automático, malha digital de setores censitários e a relação privilegiada com o IBGE. Além disso, tudo que aprendi e as perspectivas de continuar trabalhando com microáreas - localizando, agregando, estabilizando - e com estatística espacial, ajudando a mostrar as diferenças no "*Rio uma cidade de cidades misturadas*". E, mais importante, as pessoas: novos amigos com quem quero continuar trabalhando e antigos amigos que quero voltar a encontrar depois deste período de quarentena.

Um capítulo especial para Oswaldo (G. Cruz). Sem ele esta tese estaria muito diferente. Foi o congresso de Atlanta, o artigo da Statistics in Medicine, séries temporais, análise espacial, setores censitários, o S-Plus, o SAS, a Internet, e sei lá o que mais. Sem ele, eu também estaria muito diferente. Mas esta é outra história...

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

APLICAÇÃO DE MÉTODOS DE ANÁLISE ESPACIAL NA CARACTERIZAÇÃO DE ÁREAS DE RISCO A SAÚDE

Marilia Sá Carvalho

Maio/1997

Orientadores: Flávio Fonseca Nobre
Cláudio José Struchiner

Programa: Engenharia Biomédica

O objetivo geral deste trabalho é estudar a aplicação de diversos métodos de análise espacial visando, a partir de indicadores socioeconômicos por microáreas, caracterizar regiões urbanas segundo seu perfil de risco à saúde. Foram utilizadas as variáveis do censo demográfico de 1991 e a malha digital dos setores censitários da 20ª Região Administrativa do Rio de Janeiro. Os seguintes métodos foram aplicados: componentes principais, análise de aglomerados por partição, algoritmo de agregação de áreas, teste para autocorrelação espacial, métodos de análise exploratória aplicados a dados espaciais, interpolação linear e por regressão local ponderada (*loess*), variograma e interpolação por Krigeagem universal. Foram utilizados os pacotes estatísticos SAS™ e S-Plus™ e o aplicativo de mapeamento MAP-INFO™. Os métodos empregados permitiram caracterizar as áreas segundo perfil socioeconômico, incorporando sua localização relativa. As técnicas são de difícil execução e exigem razoável familiaridade prévia com os modelos e com os dados, de forma a incorporar os aspectos realmente essenciais, cabendo sempre avaliar a necessidade real de precisão na interpolação, frente à qualidade dos dados e ao modelo de determinação de doenças subjacente.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

APPLICATION OF SPATIAL DATA ANALYSIS METHODS FOR THE CHARACTERIZATION OF HEALTH RISK AREAS

Marilia Sá Carvalho

May/1997

Advisors: Flávio Fonseca Nobre
Claudio José Struchiner

Department: Biomedical Engineering

The main purpose of this work is to study methods of spatial data analysis applied to socioeconomic variables, in small areas, in order to identify health risk profile of urban region. Data comes from the 1991 demographic census and digital map of census tracts of the 20th Administrative Region of Rio de Janeiro. The methods used were: principal components and multivariate cluster analysis, areal aggregation algorithm, test for spatial clustering, exploratory data analysis methods for spatial data, linear interpolation, weighted local regression (loess), variogram and universal kriging interpolation. The statistical packages SAS™ and S-Plus™, and the mapping software MAP-INFO™ were used. The methods applied allowed the socioeconomic characterization of the areas, treating the space as continuous. The techniques are difficult to apply and demand thorough knowledge of the models and the data, in order to grasp the essential aspects. The need of such a precision in interpolating should be assessed, face to data quality and the underlying model of disease determination.

Índice

1. INTRODUÇÃO 1

2. BASES TEÓRICAS 12

2.1. Análise exploratória espacial e formas de representação 12

2.2. Modelagem da tendência 14

2.3. Variograma 16

2.4. Aglomerado espacial: pontos e áreas 25

2.5. Krigeagem 29

3. MATERIAL E MÉTODOS 34

3.1. Área do estudo 34

3.2. Indicadores 35

3.2.1. Localização geográfica 35

3.2.2. Dados demográficos e socioeconômicos 36

3.3. Análise exploratória 39

3.4. Classificação multivariada 40

3.5. Análise exploratória espacial 41

3.6. Autocorrelação espacial 42

3.7. Modelagem da tendência 43

3.8. Modelagem por variograma 44

3.9. Krigeagem 46

3.10. Reclassificação: análise multivariada sobre modelagem espacial 46

4. RESULTADOS 48
4.1. Análise exploratória 48
4.2. Classificação multivariada 56
4.3. Autocorrelação espacial 61
4.4. Análise exploratória espacial: detectando tendência 62
4.5. Análise exploratória espacial: valores atípicos 74
4.6. Modelagem da tendência 80
4.7. Variogramas amostrais 90
4.8. Modelagem do variograma 98
4.9. Krigeagem universal 106
4.10. Recriando a classificação 109
4.11. Comparando os resultados 114
5. DISCUSSÃO 117
5.1. Classificação e risco 117
5.2. Geoestatística e indicadores sociais em áreas urbanas 121
5.2.1. Análise exploratória espacial 123
5.2.2. Tendência 126
5.2.3. Variograma 127
5.2.4. Interpolação e classificação multivariada 129
6. CONCLUSÃO 132
7. REFERÊNCIAS BIBLIOGRÁFICAS 135
ANEXO 1 141
ANEXO 2 162

1. INTRODUÇÃO

O escopo deste trabalho é a caracterização socioeconômica de regiões urbanas, a partir de informações referentes a microáreas, montando assim o pano de fundo sobre o qual se poderá analisar as condições de saúde da população. A idéia geral é, através de método multivariado de classificação e da incorporação de aspectos da contigüidade espacial existente, possibilitar, na medida em que se viabilize a localização dos eventos de saúde em microáreas, o estudo da ocorrência de doenças nas populações.

O processo de seleção das variáveis do censo demográfico, utilizando análise de componentes principais, juntamente com a classificação multivariada por partição dos setores censitários da Região Metropolitana do Rio de Janeiro foi desenvolvido em comunicação apresentada no III Congresso Brasileiro de Epidemiologia (Carvalho et al., 1995), sendo a versão integral, recém-submetida à publicação, apresentada no Anexo 1. No decorrer deste trabalho, visando o desenvolvimento de métodos de tratamento da instabilidade de indicadores de saúde em micro-regiões foi publicado artigo (Carvalho et al., 1996), apresentado no Anexo 2. O corpo principal da tese refere-se à aplicação de métodos de análise de dados espaciais a variáveis socioeconômicas, comparando classificação multivariada onde se modelou espacialmente os indicadores, com classificação onde se considera cada área independente da sua localização espacial.

A utilização de mapas e a preocupação com a distribuição geográfica de diversas doenças é bem antiga. O médico (cirurgião naval) escocês James Lind publicou em 1768 um livro chamado "*An Essay on Diseases Incidental to Europeans in Hot Climates*" no qual procura explicações para a distribuição de doenças, chegando

inclusive a atribuir riscos a determinadas áreas geográficas especificas (Barret, 1991). Desde então, diversos trabalhos foram escritos na geografia médica, descrevendo variações geográficas na distribuição das doenças. Destaca-se, entre outros, o estudo de John Snow sobre as origens do cólera, que utilizou técnicas de mapeamento para relacionar os casos de cólera e pontos de coleta de água.

Entretanto, apesar do tempo decorrido desde estes precursores, ainda são poucos os trabalhos, no Brasil, que incorporam métodos de análise de dados espaciais. Levantamento feito em duas publicações nacionais dedicadas à saúde pública - Cadernos de Saúde Pública e Revista de Saúde Pública - apontou a pequena utilização destas técnicas. Entre 127 artigos originais, de revisão e análise apresentados nos anos de 1993 e 1994 (distribuídos até março/95) nas duas publicações, apenas 25 consideravam, de alguma forma, a localização espacial. Entre estes, apenas 15 apresentavam mapas, utilizando as seguintes técnicas: mapa de localização simples, mostrando pontos de coleta de amostra ou a região de realização do estudo - 10 artigos; mapa de padrões, comparando a ocorrência do fenômeno em regiões diferentes - 5 artigos. Nenhum artigo apresentou de fato análise de dados espaciais.

Internacionalmente os métodos de análise espacial têm sido empregados na área da saúde nas seguintes situações:

- quando o evento em estudo é gerado por fatores ambientais de difícil detecção a nível do indivíduo;
- na delimitação de áreas homogêneas segundo intervenção pretendida;

- quando o evento em estudo e os fatores relacionados têm distribuição espacialmente condicionada;
- no estudo de trajetórias entre localidades.

No primeiro caso, o fenômeno mais estudado, em torno do qual se desenvolveu parcela ponderável de alguns métodos, foi a hipótese de aumento na incidência de leucemias ocasionado por contaminação ambiental por usinas nucleares (Hills & Alexanders,1989). Outros exemplos, abordando a síndrome de morte súbita infantil (Rodrigues et al.,1992) e a doença de Hodgkin's (Glaser,1990, Ross & Davis,1990), apresentam hipótese causal de contagiosidade.

A delimitação espacial de regiões é tradicional na geografia médica quando, por exemplo, são definidas as estratégias de controle da malária, separando curto, médio e longo prazo para a implantação de determinadas medidas. Mais recentes são as tentativas de elaborar diagnósticos de saúde de populações delimitando áreas de risco diferenciado. O objetivo central dos métodos utilizados é agrupar, distinguir ou calcular gradientes entre localidades segundo algum critério de similaridade (Cortinovis et al.,1993, Verhasselt & Mansourian, 1991).

O estudo da ocorrência de doenças a partir de sua localização espacial é bastante difundido, particularmente enquanto elemento de identificação de possíveis hipóteses causais, sejam estas relacionadas a ambiente, utilização de serviços de saúde ou análise comportamental dos usuários (Stimson,1980). Um contra-exemplo interessante é o estudo da relação entre dureza da água e doença coronariana, onde a associação

encontrada deve-se à existência de autocorrelação espacial, não considerada no modelo estatístico utilizado (Huel et al.,1978a, Huel et al.,1978b).

As análises de trajeto são úteis no planejamento da oferta de serviços de saúde (Francis & Schneider, 1984) e na análise dos deslocamentos populacionais (de vetores, hospedeiros ou parasitas), cujos fluxos são determinantes na compreensão dos mecanismos de propagação endemo/epidêmica (Smallman-Raynor & Cliff, 1991).

As técnicas de análise espacial podem ser sistematizadas, a partir do objeto e do tipo de dado disponível em (Bailey & Gatrell, 1995):

- distribuição de pontos ("*point pattern*") - quando o objeto da análise é a posição relativa de objetos ou eventos precisamente localizados, sejam estes casos de doenças ou espécies vegetais em estudos de ecologia ambiental;
- geoestatística - conjunto de técnicas aplicadas que pressupõem a continuidade espacial do objeto, utilizadas na estimativa e interpolação, por exemplo, de fatores ambientais cuja distribuição é contínua;
- dados de áreas - quando a ocorrência do fenômeno em estudo é mensurada a partir de dados agregados por área, como é o caso de taxas de morbi-mortalidade por município;
- deslocamento - quando o objeto de estudo é o acesso e o fluxo entre regiões, inclusive otimizando trajetórias e estudando a localização de equipamentos urbanos.

Um desdobramento interessante do emprego de métodos de análise espacial é sua incorporação aos estudos ecológicos. Recentes publicações vêm resgatando o papel

deste tipo clássico de investigação em epidemiologia, onde a ênfase está nas doenças da população e não do indivíduo, onde a pergunta que se deseja responder não é sobre as causas dos casos de doença, mas sobre as causas da incidência da doença (Rose,1985). O interesse focaliza-se não na doença **em** populações, mas na doença **de** populações, o objetivo é ver a *"floresta não as á rvores"* (Poole,1994).

A fim de compreender como um contexto afeta a saúde de grupos populacionais através de seleção, distribuição, interação, adaptação e outras respostas, torna-se necessário medir efeitos em nível de grupo, uma vez que medidas em nível individual não podem dar conta destes processos. Susser (1994) destaca que, sem medir estes contextos, não se pode explicar os padrões de mortalidade ou morbidade, a propagação de epidemias, a transmissão sexual de doenças nem a transferência de comportamentos ou valores. As pessoas vivem em grupos, e a análise a nível individual não capta os efeitos dessa dimensão, incluindo as interações entre uma pessoa e outras na transmissão de infecções, comportamentos ou valores.

Visando o estudo da doença **da** população, diagnósticos de saúde de áreas determinadas são freqüentemente realizados, a partir de dados relativos à situação socioeconômica, meio-ambiente, perfil de morbi-mortalidade, disponibilidade de equipamentos urbanos, utilização de serviços de saúde. Ou seja, para uma dada divisão geográfica, político-administrativa, são estudados os indicadores disponíveis na área, em geral como parte de um processo de identificação de um problema particular ou como uma etapa do planejamento (Castellanos,1990).

Um dos problemas usualmente encontrados nestes trabalhos é a extensão da base territorial de cada unidade geográfica, onde é freqüente que estejam agregados grupos sociais distintos (favelas e áreas nobres). Assim, o indicador calculado representa uma média entre populações diferentes. Evidentemente os agravos à saúde e o acesso aos equipamentos urbanos não estão distribuídos homogeneamente na população. Identificar os **diferentes** grupos populacionais, onde de fato exista relação entre os indivíduos componentes que permita caracterizá-los enquanto um grupo, é um problema relacionado à escala do estudo. As unidades de coleta e de análise da informação devem apresentar resolução - definida a partir da menor área para a qual estão disponíveis informações - adequada ao fenômeno que se deseja estudar.

Quanto menor a escala, maior a população e a área da unidade de estudo, menor a resolução e, portanto, menor a homogeneidade interna e a capacidade de distinguir diferenças. Aumentar a escala e a resolução traz outros problemas: à medida em que diminui a área e a população, diminui também a ocorrência do evento estudado. Assim, a contrapartida do aumento na homogeneidade é a instabilidade dos indicadores nos grupos.

A escolha da unidade de análise sofre, além disso, das limitações dos dados disponíveis: em geral a área de referência é definida a partir das divisões político-administrativas usuais, tais como bairros, distritos ou municípios, que, particularmente no Rio de Janeiro, apresentam composição muito heterogênea (ver Anexo 1). Por isso a busca de partição territorial adequada aos objetivos é parte essencial deste tipo de estudo. Assim, partir de uma unidade de coleta com a maior resolução possível e

agregá-la em função do propósito da pesquisa pode ser um caminho interessante (ver Anexo 2).

Por outro lado, um dos limites dos estudos ecológicos tradicionais é tratar cada grupo populacional como independente dos outros. Cabe lembrar que:

> "*Independência é um pressuposto muito conveniente que faz grande parte da teoria da estatística matemática tratável. Entretanto, modelos que envolvem dependência estatística são freqüentemente mais realísticos. Duas classes de modelos que têm sido comumente usados envolvem estruturas de correlação intraclasse e estruturas de correlação serial. Estes oferecem pouca aplicabilidade a dados espaciais, onde a dependência está presente em todas as direções e fica mais fraca à medida em que aumenta a dispersão na localização dos dados.*" (Cressie,1991)

Ou seja, não podem ser tratados como independentes grupos populacionais vizinhos. Por vizinhos entende-se a utilização de algum critério espacial de proximidade, seja esta definida através de fronteiras comuns, distância por estradas ou alguma forma de medir o volume de interações entre os grupos.

Um aspecto importante ao "olhar a floresta" é a caracterização de risco. Aplicações de técnicas e métodos de análise desenvolveram-se nos últimos 30 anos, especialmente em três metodologias: mensuração de risco (*Health Risk Appraisal*), enfoque de risco (*Risk Approach*) e manejo de risco (*Risk Analysis/Management*) (Hayes,1992).

A primeira baseia-se na análise dos atributos do indivíduo - genéticos e comportamentais - tendo por objetivo reduzir a mortalidade precoce através da modificação de comportamentos considerados de **risco**. Nesta metodologia pouca ou nenhuma atenção é dada aos aspectos coletivos da saúde, reduzidos a **fatores de risco**. Estes são características ou circunstâncias que são acompanhadas de um aumento da probabilidade de que um evento adverso ocorra. A principal intervenção decorrente desta abordagem está voltada para aspectos educacionais.

Na metodologia do enfoque de risco, desenvolvida pela Organização Mundial da Saúde (OMS), o objetivo é a detecção de grupos populacionais prioritários para a alocação de recursos de saúde, aumentando a eficiência da aplicação de recursos públicos em países não desenvolvidos economicamente. As fontes do **risco** neste caso são mais amplas, envolvendo atributos individuais e aspectos sócio-ecológicos (Hayes,1992, MS/OPAS,1983).

A esfera do manejo de risco é mais geral que as anteriores. Envolve avaliação de produtos potencialmente perigosos, saúde do consumidor, meio ambiente, sendo que as fontes de **risco** são assumidas como externas ao indivíduo. As medidas decorrentes desta abordagem incorporam aspectos de controle industrial, elaboração de leis de proteção do consumidor e do meio ambiente (Fiskel,1990).

A metodologia utilizada na identificação de grupos de risco - cuja delimitação não deve ser entendida estaticamente, com limites rígidos entre um grupo e outro - incorpora, dependendo do enfoque adotado, elementos que permitam inferir a existência de uma probabilidade semelhante da ocorrência de determinados agravos entre

os indivíduos do grupo. São utilizados: (i) indicadores da ocorrência passada do agravo, supondo a manutenção das mesmas condições que o geraram - por exemplo, altas taxas de mortalidade infantil; (ii) a ocorrência de agravo, quando este é raro e permite a diferenciação dos grupos - ocorrência de casos de intoxicação por mercúrio, caracterizando o risco de um grupo profissional ou de área contaminada; (iii) presença de fatores de risco conhecidos - fumantes; (iv) aspectos ambientais reconhecidamente importantes - ausência de saneamento básico.

Além destes, é grande o papel dos dados socioeconômicos, enquanto preditores importantes das condições de saúde, seja como causa direta, por exemplo a ausência de saneamento aumentando a incidência de cólera, ou indireta, onde a renda ou escolaridade relacionam-se ao modo de vida. A seleção de indicadores de pobreza, válidos e viáveis de serem coletados em condições reais, é uma problema a ser enfrentado (Carstairs,1995, Eames et al.,1993). A construção de índices - medidas compostas combinando vários indicadores que procuram descrever de forma global um problema complexo (Goldberg & Dab,1987) - para caracterização de um perfil socioeconômico é objeto de diversas investigações (Gordon,1995, Cruz,1996, d'Orsi,1996, Verhasselt & Mansourian, 1991).

Outro aspecto fundamental é a confluência de riscos em determinados grupos. É o caso, por exemplo, de alguns grupos populacionais específicos, como os "meninos de rua", onde é grande a ocorrência de homicídios, atropelamentos, intoxicações por drogas, AIDS, doenças transmissíveis em geral, gravidez precoce, etc. Análises realizadas em áreas geográficas distintas sugerem que os grupos populacionais de menor

renda além de apresentarem as mais altas taxas de mortalidade infantil, concentram também a maior ocorrência de doença coronariana, de óbitos por acidentes vasculares e de alguns tipos de câncer, caracterizando-se a “transição epidemiológica” - onde agravos à saúde tradicionalmente relacionados à miséria e os comuns às sociedades ricas coexistem temporal e espacialmente.

A população está submetida a diferentes pressões: das doenças transmissíveis, facilitadas pela precariedade do saneamento ambiental, às doenças cardiovasculares, cujos determinantes estão relacionados, entre outros, à dieta inadequada, consumo de cigarros e insuficiente assistência à saúde. Ou seja, o que caracteriza o perfil epidemiológico dito de transição é a ocorrência simultaneamente e nos mesmos grupos populacionais, de agravos gerados por fatores de risco substancialmente diversos, tais como dieta inadequada contribuindo para a mortalidade por infarto do miocárdio e inexistência de saneamento básico agravando o quadro das doenças transmissíveis. A caracterização de **mesmo** grupo populacional, entretanto, depende da resolução espacial utilizada. Ao diminuir a unidade de análise, pode ser possível diferenciar a ocorrência de agravos entre grupos anteriormente confundidos devido à escala da análise.

A perspectiva geral deste trabalho foi o aprimoramento de uma análise de perfil de risco de populações urbanas. Por um lado, pela primeira vez no Brasil estavam disponíveis dados demográficos e base cartográfica digital (a malha de setores censitários) para microáreas de regiões densamente povoadas. Além disso ainda são raros, aqui, trabalhos utilizando uma perspectiva espacial na análise de dados na área da saúde. São particularmente pouco freqüentes, fora da área estrita da geoestatística,

aplicações onde se modela a continuidade do território, sem considerar os limites administrativos usuais. Neste contexto, o objetivo deste trabalho é o estudo de diversos métodos de análise espacial aplicados a indicadores do censo demográfico, visando identificar regiões urbanas segundo seu perfil socioeconômico. Pretende-se modelar as condições de vida, enquanto "cenário" de determinação das condições de saúde nas grandes cidades, buscando estudar as doenças **das** populações.

2. BASES TEÓRICAS

Apesar dos textos tradicionais das disciplinas relacionadas à saúde coletiva citarem que a epidemiologia estuda o processo saúde-doença no espaço e no tempo, (Lillienfeld,1976) observa-se que apenas as aplicações das técnicas estatísticas relacionadas ao tempo se encontram bem desenvolvidas e sistematizadas. Procuraremos aqui apresentar as bases teóricas das técnicas utilizadas, especialmente aquelas aplicadas à análise espacial, incluindo: aspectos de análise exploratória de dados espaciais; detecção de aglomerados espaciais; variograma e interpolação por Krigeagem.

2.1. Análise exploratória espacial e formas de representação

As técnicas de análise exploratória aplicadas a dados espaciais são essenciais ao desenvolvimento das etapas da modelagem estatística espacial, em geral muito sensível ao tipo de distribuição, à presença de valores extremos e à ausência de estacionariedade. As técnicas empregadas são, em geral, adaptações das ferramentas usuais. Assim, na investigação do tipo de distribuição utiliza-se *boxplots*, histogramas, *qqplots*, entre outras. Quando a distribuição é muito afastada da normalidade, é comum a transformação das variáveis, utilizando-se, por exemplo, as funções logarítmica ou exponencial.

A investigação de *outliers* é feita considerando a localização do ponto no espaço. As técnicas mais comuns são:

- visualização de valores extremos nos mapas;

- gráfico de médias e medianas segundo linhas e colunas dos pontos amostrados - permite identificar a flutuação das medidas ao longo de duas direções, sugerindo a presença de valores discrepantes quando a diferença entre estas é grande, e a tendência ao longo de uma direção quando os valores variam suavemente (Cressie,1991);
- *boxplots* de indicadores de diferenças entre pares de valores mensurados - permite identificar diferenças importantes entre pares de pontos, que quando ocorrem em pequenas distâncias representam saltos no valor da variável, indicativos de não estacionariedade local (Cressie,1991).

As técnicas de mapeamento permitem descrever de forma gráfica diversos fenômenos em saúde, desde a distribuição de padrões de morbimortalidade até a alocação de serviços, passando pelos estudos de acessibilidade. Neste trabalho estas técnicas serão utilizadas como ferramentas de análise exploratória e de representação de resultados. Abaixo são destacados os principais tipos de mapas que serão utilizados neste trabalho.

Localização de pontos: É o tipo mais simples, onde, sobre uma base cartográfica, são assinalados os pontos onde se localizam unidades de saúde, fontes de poluentes, casos de doenças, etc. Este tipo de gráfico é comumente utilizado em vigilância epidemiológica, permitindo visualizar o espalhamento dos casos em uma área. Em geral, cada ponto representa um caso, ou um número definido de casos, podendo representar diferentes equipamentos urbanos, através da utilização de símbolos ou códigos de cores (Gesler,1986).

Padrão: Utilizado freqüentemente para comparação de áreas nas investigações epidemiológicas e nos serviços de saúde. A divisão territorial mais usual é a geopolítica: municípios, estados, distritos, regiões administrativas ou bairros. Os indicadores, sejam eles coeficientes de mortalidade, taxas de incidência ou índices de utilização de serviços de saúde, são calculados para cada subdivisão do mapa, permitindo a divisão da região em classes, segundo algum critério, estatístico ou não. A cada classe será associado um padrão (hachura) ou cor, que preencherá a subdivisão do mapa (Dunn,1987, Marshall,1991).

Contorno: É a representação de uma variável tridimensional sobre um plano. Este tipo de técnica é bastante usada em cartografia, na representação de altitude. Sobre a superfície do mapa desenham-se linhas de contorno, chamadas isolinhas, que delimitam áreas onde uma variável tem a mesma grandeza. Como resultado final, obtém-se uma figura geométrica, que mostra a distribuição da variável estudada, podendo ser usadas, por exemplo, para indicar o nível de prevalência de uma doença. As curvas de contorno são construídas por interpolação de valores medidos em diversos pontos, variando o método de interpolação utilizado (Isaaks & Shrivastava, 1989, Cressie, 1991).

2.2. Modelagem da tendência

A variação em larga escala observada na análise exploratória, seja em gráficos de contorno, tridimensionais, ou em gráficos de médias e medianas por linhas e colunas, pode ser modelada utilizando polinômios de diversos graus ou modelos “locais” de ajuste.

Considere-se um processo espacial onde o valor da variável é uma função polinomial de sua posição no espaço mais um erro aleatório. Neste modelo de regressão com duas variáveis - as coordenadas espaciais - os coeficientes podem ser ajustados através de mínimos quadrados. (Bailey & Gatrell,1995)

O modelo usual de regressão múltipla utilizando notação vetorial é:

$$\boldsymbol{V(s) = x^T(s)\beta + \varepsilon(s)} \qquad 1.$$

onde, $\boldsymbol{V(s)} \rightarrow$ Variável aleatória representando o processo no ponto $\boldsymbol{s}$,
$\boldsymbol{x^T(s)\beta} \rightarrow$ Tendência (ou seja, o valor médio $\boldsymbol{\mu(s)}$),
$\boldsymbol{\varepsilon(s)} \rightarrow$ Flutuação aleatória com média zero

O vetor $\boldsymbol{x(s)}$ consiste em p funções das coordenadas espaciais (s_1, s_2), do ponto amostrado s. Para uma superfície de tendência linear é apenas $(1, s_1, s_2)$, para quadrática é $(1, s_1, s_2, s_1^2, s_2^2, s_1.s_2)$, e assim sucessivamente. $\boldsymbol{\beta}$ é o vetor de dimensão $(p+1)$ dos parâmetros a serem ajustados. O pressuposto básico deste modelo assume que os erros têm variância constante e são independentes em cada local, sendo a covariância, conseqüentemente, igual a zero. Não há efeitos de segunda ordem presentes no processo. Neste contexto, é feito o ajuste do modelo por mínimos quadrados ordinários. Como no caso de dados espaciais se espera que exista correlação entre os resíduos, os resultados devem ser analisados cautelosamente, sendo os intervalos de confiança e testes de ajuste do modelo pouco confiáveis. Uma forma de incorporar a não estacionariedade na variância dos resíduos é a utilização do método de mínimos quadrados ponderados. Neste caso, cada observação na regressão recebe peso inversamente proporcional à sua variância, ou seja, as observações com maior variância

contribuem menos para o cálculo.

No caso de haver dependência espacial dos resíduos, um dos métodos utilizados é denominado mínimos quadrados generalizados, onde se inclui na estimativa dos parâmetros do modelo um erro, também com média zero, mas não necessariamente independente nos diferentes pontos. Neste caso, incorpora-se ao cálculo dos mínimos quadrados a matriz de covariâncias. O principal problema neste caso é que não se conhece a matriz de covariância, que tampouco pode ser estimada a partir das amostras, uma vez que em geral somente se dispõe de uma observação em cada local. Em geral, a modelagem de processos espaciais incorpora simultaneamente efeito de primeira ordem - tendência - e de segunda - covariância - através da Krigeagem universal.

Outra família de métodos potencialmente úteis para estimar tendência são as regressões não paramétricas. Os principais aspectos destes modelos são o uso de algum método de alisamento e a incorporação de um parâmetro relacionado ao tamanho da janela adotada. Dentre estes, o mais simples é a média móvel, onde o método de alisamento é a média, e o número de observações incorporada ao cálculo da média em cada ponto é a dimensão da janela.

Um modelo relativamente simples, e que apresenta comportamento melhor nos extremos da série, é a regressão linear local ponderada, denominada *loess*. Neste caso, ao invés de se calcular a média em cada janela, estima-se os parâmetros de um plano, por mínimos quadrados. O peso das observações diminui à medida em que se afasta do ponto estimado, sendo uma regressão local ponderada (Hastie & Tibshirani, 1990).

2.3. Variograma

Pouco comum na literatura sanitária, mas certamente a mais difundida nos livros texto de geoestatística, a estatística utilizada para análise da estrutura da dependência espacial onde se considera a continuidade do terreno é o variograma. Este é a medida da variância das diferenças dos valores medidos em todos os pontos separados a uma distância fixa. À medida em que aumenta a distância, o variograma tende a igualar a variância total. Se houver estacionariedade de primeira e segunda ordem, o variograma expressa o grau de dependência entre todos os pontos da superfície.

Considerando um processo estocástico ***Z(s)***, o variograma é a variância da diferença entre a realização do processo em *(s)* e em *(s + h)*, onde $\gamma(h)$ é o semi-variograma e *h* é o número de *lags* ou passos entre os pares (Cressie,1991).

$$\boldsymbol{Var(Z(s+h)-Z(s))} = 2\gamma(\boldsymbol{h}) \qquad 2.$$

O método clássico de estimar o variograma amostral é apresentado na equação (Eq.) 3. Ainda que a denominação mais precisa seja **semi-variograma**, uma vez que se considera a diferença dos valores da variável nos pontos de cada par somente uma vez, será adotada a denominação mais simples de **variograma**.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j)} (v_i - v_j)^2 \text{ , onde: } N \rightarrow \text{número de pares de pontos;} \qquad 3.$$
$$v \rightarrow \text{valor da variável nos pontos } i \text{ e } j$$

Cressie (1991) propõe um método robusto - menos sensível a valores extremos - de estimar o variograma amostral que consiste em:

$$\gamma(h) = \frac{1}{2} \cdot \frac{\left\{ \frac{1}{N(h)} \sum_{(i,j)} |v_i - v_j|^{1/2} \right\}^4}{\left(0{,}457 + \frac{0{,}494}{N(h)} \right)} \qquad 4.$$

A justificativa para este estimador é que para variáveis normais $\left(Z(s_i) - Z(s_j)\right)^2$ é uma variável aleatória com distribuição qui-quadrado com um grau de liberdade. A transformação que a torna mais próxima a uma distribuição normal é a raiz quarta, ou seja, $\left|Z(s_i) - Z(s_j)\right|^{1/2}$, a raiz quadrada da diferença absoluta. A inclusão do denominador visa corrigir o viés da distribuição (Cressie & Hawkins,1980).

Outras estatísticas também utilizadas na modelagem de processos espacialmente distribuídos são as funções de autocovariância e autocorrelação. A autocorrelação espacial, ou correlograma, é a função de covariância normalizada pela variância, sendo a autocovariância definida como:

$$Cov(h) = \frac{1}{N(h)} \sum_{i,j} \left[v_i v_j - \frac{1}{n} \sum_{k=1}^{n} v_k \right]^{\frac{1}{2}} \qquad 5.$$

É importante observar que a função de autocovariância espacial decresce com a distância, à medida em que diminui a covariância entre os pontos, enquanto que o variograma aumenta com a distância até estabilizar-se em um platô (*sill*) quando desaparece a dependência espacial. Quando, entretanto, existe tendência espacial em alguma direção, mesmo sem a dependência em pequena escala, o variograma não se

estabiliza, assim como a autocovariância cai muito lentamente a zero. Neste caso é necessário que a tendência seja retirada para possibilitar a modelagem da dependência espacial em distâncias curtas. A tendência pode ser modelada através de polinômios de diversos graus, globais ou locais (*loess*).

Baseado no variograma amostral, modela-se o variograma, ajustando-se funções conhecidas aos valores calculados através da Eq. 3 ou da Eq. 4. No ajuste do modelo são considerados os parâmetros relacionados à escala, à extensão da continuidade e ao valor onde o variograma se estabiliza (*sill*), e à forma da dependência espacial. São utilizadas diferentes técnicas, sendo a mais usual o ajuste por mínimos quadrados.

A análise da variação em pequena escala procura dimensionar o grau de continuidade em distâncias muito pequenas. Teoricamente, sendo o processo contínuo, quando *h* tende a zero o variograma também tende a zero. Entretanto, devido a descontinuidades geradas por variações que ocorrem em escala abaixo da freqüência de amostragem (*1/h*), é necessário incorporar ao modelo o denominado "efeito pepita" ("*nugget effet*"). Como nada pode ser afirmado para intervalos menores do que o menor valor, em módulo, da distância entre duas amostras, a forma de incorporar estas descontinuidades é acrescentar uma constante (*Co*) ao modelo. Esta é calculada através da extrapolação do variograma estimado para intervalos muito pequenos. A importância de incorporar o efeito pepita na modelagem do variograma está relacionada à predição. Um aspecto prático relaciona-se à freqüência de amostragem, que é muitas vezes irregular. Por isso utiliza-se definir uma tolerância angular e linear na determinação dos intervalos *h*.

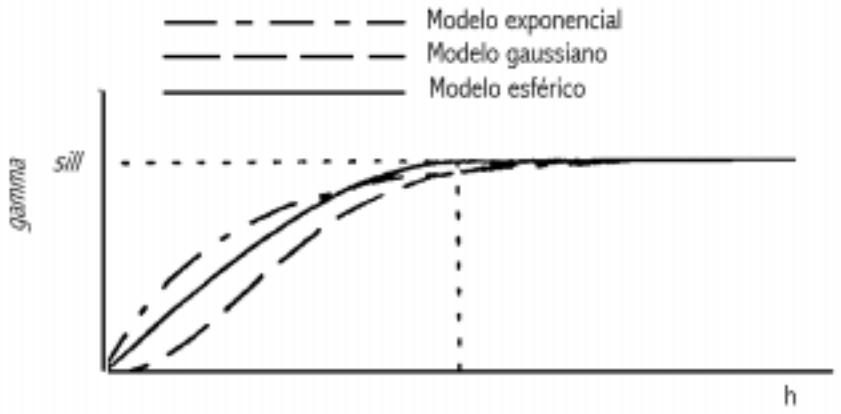
A inclinação da curva da função do variograma em relação ao número de intervalos indica o grau de dependência espacial da variável. O valor de *h* onde o variograma se estabiliza, o alcance (*range*), é a distância até onde existe dependência espacial.

As funções mais utilizadas são os modelos esférico, Gaussiano e exponencial, cujas equações são apresentadas no Quadro 1, onde *a* é o valor do alcance, distância onde a dependência espacial desaparece, e *h* é o intervalo, ou seja, o número de passos na grade de amostragem. (Isaaks & Shrivastava,1989). Conforme pode ser observado na Figura 2.1, os modelos exponencial e Gaussiano apresentam comportamento oposto nas menores distâncias entre pares de pontos, sendo que no segundo a diminuição da relação entre pontos próximos ocorre mais lentamente. O modelo esférico cresce linearmente nos menores intervalos.

Quadro 2.1 - Modelos de variograma mais usuais

MODELO	EQUAÇÃO	OBSERVAÇÕES
Esférico (6.)	$\gamma(h) = \begin{cases} 1{,}5\frac{h}{a} - 0{,}5\left(\frac{h}{a}\right)^3 & \text{, se } h \le a \\ 1 & \text{, caso contrário} \end{cases}$	comportamento linear próximo à origem
Exponencial (7.)	$\gamma(h) = 1 - \exp\left(\frac{-3h}{a}\right)$	atinge o platô assintoticamente; na prática, considera-se o valor de *a* onde o variograma atinge 95% do platô.
Gaussiano (8.)	$\gamma(h) = 1 - \exp\left(\frac{-3h^2}{a^2}\right)$	também assintótico, com crescimento parabólico próximo à origem

Figura 2.1 - Modelos de variogramas



Gamma ou γ é valor do variograma;
h é o número de intervalos entre dois pontos;
sill é o patamar máximo atingido de γ.

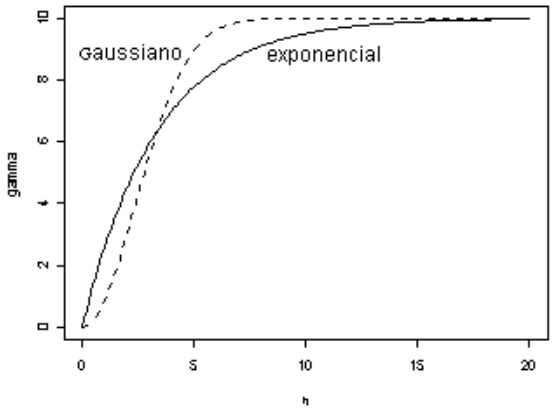
Fonte: ISAAKS,E.H. & SHRIVASTAVA,R.M., 1989, pag.374.

São apresentados a seguir alguns modelos de variogramas freqüentes na literatura, com variações nos parâmetros de forma a permitir visualizar a influência de cada um na modelagem.

Figura 2.2 - Efeitos de alterações nos modelos de variograma

Figura 2.2.A

Variação no modelo adotado: Exponencial e Gaussiano



9.A - exponencial:

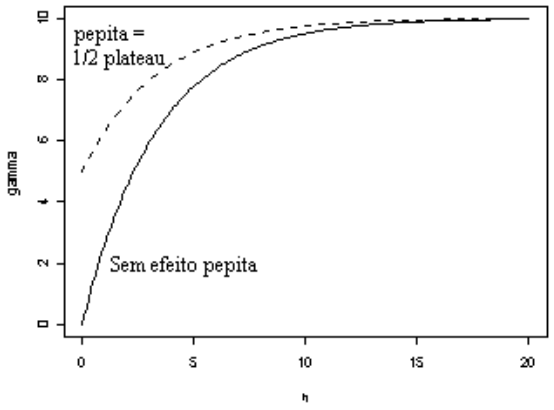
$$\gamma(h) = 10\left(1 - \exp\left(\frac{-3|h|}{10}\right)\right)$$

9.B - Gaussiano:

$$\gamma(h) = 10\left(1 - \exp\left(\frac{-3|h|}{10}\right)^2\right)$$

Figura 2.2.B

Variação no efeito pepita



10.A - sem efeito pepita:

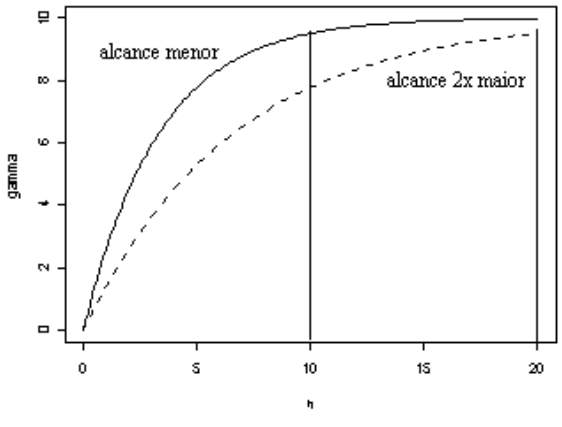
$$\gamma(h) = 10\left(1 - \exp\left(\frac{-3|h|}{10}\right)\right)$$

10.B - efeito pepita = 50% do platô

$$\gamma(h) = \begin{cases} 5 + 5\left(1 - \exp\left(\frac{-3|h|}{10}\right)\right) \\ \text{ou } 0 \text{ , se h = 0} \end{cases}$$

Figura 2.2.C

Variação no alcance (*range*)



11.A - amplitude menor:

$$\gamma(h) = 10\left(1 - \exp\left(\frac{-3|h|}{10}\right)\right)$$

11.B - com o dobro da amplitude:

$$\gamma(h) = 10\left(1 - \exp\left(\frac{-15|h|}{100}\right)\right)$$

Nos processos distribuídos espacialmente, a dependência espacial (intensidade, distância, ciclicidade) pode variar conforme a direção adotada. Assim, a análise deve ser feita em várias direções, detectando-se os eixos de maior e menor anisotropia. Esta pode ser definida como a existência de diferenças na covariância em função da direção e não apenas da distância entre os pontos. Embora todos os parâmetros do variograma possam mudar conforme a direção, são mais encontradas as denominadas anisotropia geométrica e zonal. A primeira ocorre quando a amplitude (*range*) difere conforme a direção e o platô (*sill*) permanece o mesmo. Já a anisotropia zonal apresenta diferentes valores do variograma amostral para uma mesma amplitude. A Figura 2.2.C, é um exemplo de anisotropia geométrica. O método mais simples de identificá-la é através de um diagrama onde são desenhados segmentos orientados conforme as direções analisadas, cujo tamanho é diretamente proporcional à distância entre os pontos para um dado valor do variograma na direção desejada, denominado *rose diagram* (Isaaks & Shrivastava,1989).

Para calcular o variograma omnidirecional corrigido para anisotropia, no somatório das diferenças $(v_i - v_j)^2$ da Eq.3, cada parcela é ponderada segundo um fator que considera a direção do par e a razão entre o maior e menor alcance. Ou seja, ao invés de se considerar um intervalo *h* fixo entre os pares, onde a cada valor de *h* são calculadas e somadas todas as diferenças, neste caso o valor de *h* não é fixo, mas varia conforme a direção entre os pontos de cada par. Ao invés de um círculo de raio *h* definindo cada intervalo de cálculo do variograma amostral (*lag*), considera-se uma figura elíptica, onde o eixo maior corresponde à direção de maior alcance (*range*).

Na prática, isto é feito através da alteração das coordenadas de cada ponto de um par, de forma semelhante à uma rotação de eixos com reescalonamento. Cada coordenada é corrigida por um fator que considera o ângulo entre os pontos do par considerado e a razão entre o maior e menor alcance (*range*) atingidos. Para tal, multiplica-se as coordenadas pela matriz simétrica *A*, abaixo:

$$A = \begin{bmatrix} \cos^2\theta + r \cdot \text{sen}^2\,\theta & (1-r)\cdot \text{sen}\,\theta \cdot \cos\theta \\ (1-r)\cdot \text{sen}\,\theta \cdot \cos\theta & \text{sen}^2\,\theta + r \cdot \cos^2\theta \end{bmatrix} \qquad 12.$$

onde:

$r \rightarrow$ razão entre maior e menor distâncias atingidas a um dado valor de γ;
$\theta \rightarrow$ direção do eixo da maior distância.

O efeito desta correção é que, ao estimar o variograma amostral na direção onde o alcance é maior, o valor absoluto de *h* também será maior, e o número de intervalos percorridos até atingir a máxima variância será igual em todas as direções.

Uma vez ajustado um modelo aos dados, pode-se analisar este ajuste através da distribuição dos resíduos. Considerando $\{Z(s(i)) \mid i = 1,\ldots,n\}$ um processo aleatório multivariado normal, e γ seu variograma, o indicador *F* (Eq.13) apresenta distribuição qui-quadrado com um grau de liberdade (Cressie,1991). Assim, é possível estabelecer pontos de corte baseados na distribuição esperada de *F*, identificando pares de pontos cuja diferença, considerando a distância *d* entre eles, apresente uma dada probabilidade *p* de ocorrer.

$$\boldsymbol{F} = \frac{\left(Z(s_i) - Z(s_j)\right)^2}{2\gamma(\boldsymbol{d})} \qquad 13.$$

O diagnóstico é feito baseado na identificação gráfica destes pares de pontos com similaridade ou diferença no valor da variável *F* acima do previsto pelo modelo para uma dada distância entre pares. Estes pares “muito” semelhantes ou “muito” diferentes são ligados por uma linha. Teoricamente, o total de linhas desenhadas no gráfico deve ser limitado a um percentual do total de pares existentes até a distância escolhida, ou seja, o *p* definido. O excesso de linhas no gráfico indica ajuste precário. Além disso, a dispersão destas pelo terreno também deve ser considerada na identificação de não estacionariedades locais (Barry,1996).

2.4. Aglomerado espacial: pontos e áreas

A detecção de aglomerado espacial pode ser feita a partir da análise da distribuição de pontos (*point pattern*), onde se avalia se a distância entre os pontos de ocorrência de eventos é ou não aleatória, ou entre áreas, onde se diagnostica se a freqüência na ocorrência de eventos em áreas apresenta distribuição condicionada pela posição espacial das regiões estudadas.

O aspecto central da análise de pontos, tal como desenvolvido por Knox (1964) e posteriormente generalizado por Mantel (1967), é a detecção de um número acima do esperado de pares de casos excessivamente próximos (segundo critério preestabelecido), ou de distâncias entre pares de casos muito pequenas. A distribuição base para o cálculo dos valores esperados é, em geral, uma distribuição de Poisson. Por distância entende-se, além da medida de espaço, usualmente a Euclidiana, uma medida de tempo, estabelecendo uma proximidade espaço-temporal na detecção do *cluster*.

A detecção do aglomerado é feita a partir do cálculo de uma estatística *Z*, definida pela Eq.14, onde *n* é o número de observações e *m* cada par de pontos.

$$Z = \sum_{m=1}^{n/2} [f \cdot g] \qquad 14.$$

Na equação acima, *f* é função da distância geográfica entre os pares e *g* da distância temporal. O somatório dos produtos destas distâncias entre todos os pares de pontos possíveis permite o cálculo da estatística *Z*, que quando comparada a uma distribuição de Poisson permitirá verificar se a probabilidade da distribuição destas distâncias está abaixo do esperado apenas por casualidade, caracterizando a aglomeração do fenômeno em estudo. As funções *f* e *g* devem ser escolhidas de forma a evitar artefatos relacionados à escala, seja normalizando, seja trabalhando com alguma transformação algébrica da função (Mantel,1967). Um aspecto central das diferentes técnicas estatísticas de detecção de aglomerados espaciais relaciona-se à especificidade e sensibilidade do método utilizado. Em geral, é esperado uma certa proporção de casos agregados por mero acaso, que serão falsos positivos, como também a existência de aglomerados não detectados, caracterizando falsos negativos (Chen et al.,1984, Openshaw et al.,1988, Alexander et al.,1988, Besag & Newell, 1991).

A detecção de similaridade entre áreas pode ser feita através do índice de Moran *I*, que é uma medida de correlação espacial usada para detectar afastamentos de uma distribuição espacial aleatória. Tais afastamentos indicam a existência de padrões espaciais, como por exemplo aglomerado ou tendência espacial. Utiliza-se como

ponderador uma matriz de conexão, que representa a estrutura e/ou a dependência espacial das áreas envolvidas. Assim, é testado se as áreas conectadas apresentam maior semelhança quanto ao indicador estudado do que o esperado num padrão aleatório. Os valores obtidos de Moran *I* encontram-se entre *-1* e *1*, quantificando o grau de autocorrelação existente, sendo positivo para correlação direta, negativo quando inversa. A hipótese nula (Ho) é a de completa aleatoriedade espacial, quando o indicador se distribui ao acaso entre as áreas sem relação com a posição destas. (Eq.15)

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{.j} Z_i Z_j}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{.j} \cdot \sum_{i=1}^{N} Z_i^2} \qquad 15.$$

onde: $N \rightarrow$ Número de áreas,
$X_i \rightarrow$ Indicador do evento na área i,
$Z_i \rightarrow$ Diferença entre a indicador X_i e a média de X
$w_{ij} \rightarrow$ Pesos atribuídos conforme a conexão entre as áreas i e j

Algumas variações deste modelo são o teste *C* de Geary apresentado na Eq. 18 (Cliff & Ord, 1981) e o teste *Ipop* de Moran, na Eq. 19 (Oden, 1995). O primeiro (*C* de Geary) difere do teste *I* de Moran por utilizar a diferença entre os pares, enquanto que Moran utiliza a diferença entre cada ponto e a média global. Assim, o indicador *C* de Geary assemelha-se ao variograma, e o *I* de Moran ao covariograma (Bailey & Gatrell, 1995).

$$C = \frac{(N-1)\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} \cdot \sum_{i=1}^{N} Z_i^2} \qquad 16.$$

O teste *Ipop* considera que a diferença entre as populações das áreas é sensível à ocorrência de aglomerado intra-área - ou seja, a ocorrência de elevado número de casos numa pequena população de um único município - além dos aglomerado entre áreas, onde municípios com muitos casos são adjacentes. A hipótese nula (Ho) assume que a variação geográfica do número de casos segue a variação geográfica do tamanho da população, sendo particularmente útil quando a população das áreas é muito variável.

$$Ipop = \frac{N^2\sum_{i=1}^{m}\sum_{j=1}^{m} w_{.j}(e_i - d_i)(e_j - d_j) - N(1-2\bar{b})\sum_{i=1}^{m} w_{.j}e_i - N\bar{b}\sum_{i=1}^{m} w_{ii}d_i}{\left(X^2\sum_{i=1}^{m}\sum_{j=1}^{m} d_i d_j w_{.j} - X\sum_{i=1}^{m} d_i w_{ii}\right)\bar{b}(1-\bar{b})} \qquad 17.$$

onde: m → Número de áreas
N → Número total de casos em todas as áreas.
n_i → Número de casos na área i
e_i → Proporção de casos na área i $(e_i=n_i/N)$
X → População total em todas as áreas
x_i → Tamanho da população na área i
d_i → Proporção de população na área i $(d_i=x_i/N)$
Z_i → Diferença entre a taxa X_i e a média de X
w_{ij} → Pesos atribuídos conforme a conexão entre as áreas i e j
b → Prevalência média (N/X)

Um aspecto fundamental na utilização destes testes é a escolha da matriz de vizinhança $\boldsymbol{W}$, onde cada elemento w_{ij} representa uma medida de proximidade espacial entre as áreas A_i e A_j. A escolha de w_{ij} depende do tipo de dado, de região e dos

mecanismos particulares da dependência espacial. Alguns critérios usuais são (Bailey & Gatrell,1995):

$w_{ij} = \begin{cases} 1 \text{ , se o centróide de } A_i \text{ é o mais próximo de } A_j \text{ e} \\ 0 \text{ , caso contrário} \end{cases}$

$w_{ij} = \begin{cases} 1 \text{ , se o centróide de } A_i \text{ está dentro de distância especificada de } A_j \text{ (}buffer\text{) e} \\ 0 \text{ , caso contrário} \end{cases}$

$w_{ij} = \begin{cases} 1, \text{ se } A_i \text{ tem fronteira comum com } A_j \text{ e} \\ 0, \text{ caso contrário} \end{cases}$

$w_{ij} = \frac{l_{ij}}{l_i}$, onde l_{ij} é o comprimento da fronteira comum entre A_i e A_j e l_i é o perímetro de A_i

$w_{ij} = \frac{1}{d}$, onde d é a distância entre os centróides de A_i e A_j

2.5. Krigeagem

A Krigeagem é um procedimento de interpolação de valores em uma superfície contínua. Difere dos outros métodos de interpolação por fazer uso explicitamente da variância entre os valores observados, o variograma. Baseia-se na combinação linear ponderada dos dados disponíveis, aplicada ao modelo de variograma adotado (Isaaks & Shrivastava,1989). O método de cálculo do vetor de pesos consiste na minimização da variância do erro, garantindo viés zero. Este método de interpolação permite o cálculo do erro padrão para cada valor estimado (Bailey & Gatrell, 1995).

Seja,

$$\hat{v} = \sum_{j=1}^{n} (w_j . v) \qquad 18.$$

Sendo $\hat{v}$ o estimador do valor da variável em estudo no ponto desejado. Duas condições devem ser satisfeitas para a predição através da Krigeagem: a variância do erro deve ser mínima e o viés nulo. Para garantir que o viés seja zero, o valor esperado dos resíduos deve ser zero. A equação abaixo (Eq. 19) assume valor nulo se a soma dos pesos w_i for igual a 1.

$$E\{R(s_0)\} = E\left\{\sum_{i=1}^{n} w_i . v(s_i) - v(s_0)\right\} = \sum_{i=1}^{n} w_i . E\{v(s_i)\} - E\{v(s_0)\} \qquad 19.$$

Para minimizar a variância faz-se a derivação parcial em relação à covariância de cada par $v_i v_j$ e iguala a zero. Constrói-se assim um sistema de equações lineares, onde, para garantir a ausência de viés se agrega a restrição:

$$\sum w_i = 1 \qquad 20.$$

O vetor de pesos é estimado a partir de

$$\begin{bmatrix} C_{11} & \dots & C_{1n} & 1 \\ \dots & \dots & \dots & \dots \\ C_{n1} & \dots & C_{nn} & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \bullet \begin{bmatrix} w_1 \\ \dots \\ w_n \\ \mu \end{bmatrix} = \begin{bmatrix} C_{10} \\ \dots \\ C_{n0} \\ 1 \end{bmatrix} \qquad 21.$$

Onde $\boldsymbol{C}$ é a matriz de covariância à qual se acrescenta mais uma linha e umacoluna preenchidas pelo valor 1, e na posição (n+1),(n+1) por zero. O vetor dos pesos é $\boldsymbol{w_i}$ e μ é o operador de Lagrange[1], utilizado na resolução do sistema de equações; $\boldsymbol{D}$ é o vetor da covariância entre o valor previsto no modelo para o ponto a estimar, e os pontos definidos como estimadores. Assim, o cálculo dos pesos para a estimação no ponto desejado é feito através de:

$$\boldsymbol{W = C^{-1} \cdot D} \qquad 22.$$

E a variância da Krigeagem, ou erro médio quadrático de predição (σ_e^2), é função da variância total observada (σ^2), do vetor de pesos e da matriz de covariância:

$$\sigma_e^2 = \sigma^2 - \boldsymbol{D^T} \cdot \boldsymbol{C}^{-1} \cdot \boldsymbol{D} \qquad 23.$$

Da mesma forma como se estima a variável em um ponto determinado, é possível estimar a média de uma área, através da Krigeagem em bloco, que obedece ao mesmo modelo, substituindo valores pontuais por médias encontradas em uma grade regular. Além disso, assim como é feita a análise univariada, é possível trabalhar com duas ou mais variáveis, utilizando as funções de variograma cruzado e a co-Krigeagem. O objetivo é estimar (interpolar) o valor de uma variável *V* em um ponto/região, explorando, além da autocorrelação espacial da variável *V*, a correlação cruzada com

[1] Na minimização da variância, ao se incluir o somatório de pesos igual a um (Eq.21), obtém-se *n+1* equações com *n* incógnitas. A solução é introduzir nova variável, denominada operador de Lagrange.

outra(s) variável(eis) *U*, buscando melhorar a estimativa. O modelo utilizado é a combinação linear dos *n* valores de *v* e dos *m* valores de *u*, nos pontos *i* e *j*.

$$\hat{v}_0 = \sum_{i=1}^{n} a_i \cdot v_i + \sum_{j=1}^{m} a_j \cdot u_j \qquad 24.$$

A modelagem é igual à Krigeagem, mas utiliza as variáveis aleatórias *U* e *V* (ou outras) modeladas a partir da autocovariância e da covariância cruzada. Busca-se encontrar vetor de pesos a partir dos valores das duas variáveis localizadas nos pontos amostrados, mantendo o viés igual a zero e a variância do erro mínima. Para tal, o sistema de equações gerado pela derivação parcial em relação à covariância de cada par deve incorporar a restrição de que o valor esperado dos erros será zero. Sendo o erro (*R*) definido como $R = U_i - U_0$, onde *U* é o valor da variável no ponto *0* que se deseja estimar e em *i*, então:

$$R = \sum_{i=1}^{n} a_i U_i + \sum_{j=1}^{m} b_j V_j - U_0 \qquad 25.$$

O viés será zero se:

$$E\{\hat{U}_0\} = E\left\{\sum_{i=1}^{n} a_i U_i + \sum_{j=1}^{m} b_j V_j\right\} \qquad 26.$$

E conseqüentemente,

$$\sum_{i=1}^{n} a_i = 1 \quad e \quad \sum_{j=1}^{m} b_j = 0 \qquad 27.$$

Outra alternativa que também garante a condição acima é apresentada na Eq. 28, onde $\hat{m}_v$ e $\hat{m}_u$ são as médias das variáveis *V* e *U*, sendo o restante do processo é igual à Krigeagem.

$$\hat{U}_0 = \sum_{i=1}^{n} a_i U_i + \sum_{j=1}^{m} b_j (V_j - \hat{m}_V + \hat{m}_U) \qquad 28.$$

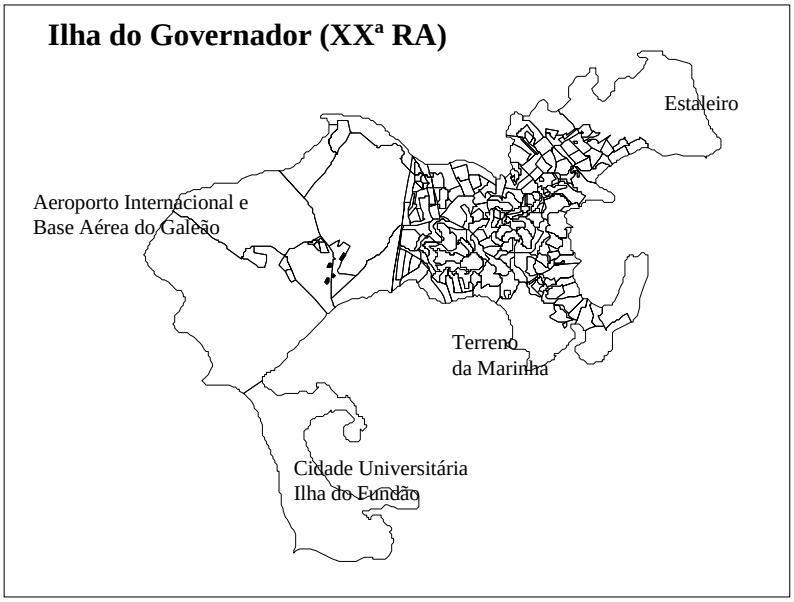
A Krigeagem ordinária assume estacionariedade de primeira ordem, ou seja, média constante. Quando existe tendência, duas alternativas são possíveis: sua retirada e modelagem do resíduo, ou a Krigeagem universal, que consiste na modelagem simultânea incorporando a tendência e a pequena escala. No primeiro caso, pode-se utilizar um conjunto de métodos de modelagem da tendência, inclusive alguns mais "locais", mas não é possível estimar o erro associado ao modelo. Na Krigeagem universal, como são estimados conjuntamente variação global e local, a estimativa dos parâmetros do variograma fica alterada pela não estacionariedade. Pode-se argumentar que como somente são utilizados pares de pontos a distâncias relativamente próximas, a tendência pouco influi na estimativa dos parâmetros do variograma. Ainda assim, sugere-se que o variograma seja modelado a partir de dados onde a tendência tenha sido retirada, e, utilizando os parâmetros deste, seja feito a Krigeagem universal com ajuste simultâneo da tendência (Bailey & Gatrell,1995).

3. MATERIAL E MÉTODOS

3.1. Área do estudo

A área estudada foi a 20ª Região Administrativa - Ilha do Governador - composta por 14 bairros e 225 setores censitários, onde localizam-se residências de classe média e favelas, indústrias, o Aeroporto Internacional do Galeão e o Campus da Universidade Federal do Rio de Janeiro. A Ilha foi, na década de 70, área de grande expansão imobiliária, especialmente de conjuntos habitacionais voltados para populações de médio poder aquisitivo. A base cartográfica foi digitalizada e posteriormente convertida para o aplicativo MAPINFO© a partir das plantas originais dos setores censitários na escala de 1:5.000, obtidas através de convênio com a FIBGE (Gráfico 3.1)

Gráfico 3.1
Base Cartográfica - Setores censitários da Ilha do Governador



A análise estatística espacial - mapas de contorno, modelagem de tendência, variograma e Krigeagem - foi realizada excluindo os setores censitários do aeroporto, cidade universitária e indústria naval, devido às grandes áreas e pequenas populações destes setores censitários especiais. Os diferentes métodos de interpolação utilizados pressupõem continuidade espacial, existente no terreno, exceto nas regiões excluídas, como pode ser observado na imagem de satélite (Gráfico 3.2).

Gráfico 3.2
Imagem da satélite da Ilha do Governador e limite da área estudada

3.2. Indicadores

3.2.1. Localização geográfica

Foram utilizadas basicamente duas variáveis espaciais: o contorno e as coordenadas dos centros geométricos dos setores censitários.

3.2.2. Dados demográficos e socioeconômicos

Todas as variáveis do questionário global do censo demográfico de 1991 foram consolidadas por setor censitário. O setor censitário é definido em função das necessidades operacionais do censo, uma vez que corresponde à área que pode ser percorrida por um recenseador durante o período de coleta de dados (FIBGE, 1993). O questionário do censo divide-se em duas partes: a primeira dedicada ao domicílio e ao chefe da família, e a segunda dedicada a cada morador (FIBGE,1990). Algumas definições das variáveis coletadas pela FIBGE serão apresentadas aqui visando esclarecer conceitos utilizados e melhorar a compreensão (FIBGE, 1993).

Os setores censitários são classificados - segundo sua situação - em urbanos e rurais. No município do Rio de Janeiro todos os setores foram classificados como urbanos. Quanto ao tipo, o setor pode ser:

Coletivos - Setores exclusivamente de habitações coletivas, como asilos, creches, quartéis, presídios, delegacias, etc. Quatro setores censitários especiais, onde

se localizam quartéis da aeronáutica, por suas características peculiares, foram excluídos da análise.

Normais - Setores onde há estrutura urbana tradicional, com ruas, endereços, ainda que a propriedade do terreno eventualmente não seja bem definida;

Sub-normais - A definição destes setores é: "(...) *conjunto constituído por unidades habitacionais ocupando ou tendo ocupado até período recente terrenos de propriedade alheia, dispostos, em geral, de forma desordenada e densa, e carentes, em sua maioria, de serviços públicos essenciais*" (FIBGE, 1994).

Assim, em setores ditos "normais", eventualmente podem ser encontrados domicílios precários, e em setores considerados pelo censo "sub-normais" o tipo de moradia e situação de urbanização pode ser normal. Para fins deste trabalho, consideramos os setores "sub-normais" como sendo favelas.

Quanto às características do domicílio, a primeira subdivisão separa-os em particulares e coletivos. Os primeiros dividem-se em permanentes e improvisados. A pesquisa das demais características dos domicílios limitou-se aos domicílios particulares permanentes, e pelo mesmo motivo os indicadores criados têm como denominador os domicílios particulares permanentes.

O processo de criação e seleção dos indicadores, apresentado mais detalhadamente no artigo "*Método multivariado de classificação socioeconômica de*

microáreas urbanas - os setores censitários da Região Metropolitana do Rio de Janeiro" (Anexo 1), constou de:

- criação dos indicadores a partir dos dados brutos de cada setor censitário, em geral proporções e médias de ocorrência por setor censitário;
- organização das variáveis nos seguintes blocos temáticos: saneamento (água, esgoto e coleta de lixo); ocupação do terreno (propriedade e tipo de construção); escolaridade (anos de estudo dos chefes de família, alfabetização da população); domicílio (média de cômodos e moradores); demográficos (mulheres chefes de família e mediana etária por sexo); renda (somente do chefe da família);
- seleção dos indicadores em cada bloco temático: o conjunto de variáveis de cada bloco foi submetida à análise de componentes principais, que orientou a escolha das duas ou três que apresentaram maior correlação com os dois pimeiros eixos, sendo selecionadas nesta etapa 15 variáveis com maior poder de explicação da variância total (a descrição mais detalhada do processo é apresentada no Anexo 1);
- nova análise de componentes principais sobre estas variáveis selecionadas, em conjunto com informações oriundas da literatura permitiu reduzí-las a seis,

Buscou-se selecionar variáveis menos colineares, ainda que a decisão final tenha considerado a importância descrita na literatura de alguns indicadores (Duchiade,1991, Barcellos & Machado, 1991, Verhasselt & Mansourian, 1991). Os seis indicadores selecionados foram:

- *Renda média do chefe do domicílio:* renda nominal média do chefe de família (este indicador foi reescalonado para média zero, desvio padrão 1, após a fase de análise exploratória inicial).
- *Proporção de chefes de família com segundo grau completo ou superior:* número de chefes de família com escolaridade igual ao segundo grau completo ou terceiro grau completo ou incompleto em relação ao total de chefes de família.
- *Proporção de população alfabetizada:* número de habitantes alfabetizados acima de 5 anos em relação à população total acima de 5 anos.
- *Proporção de domicílios alugados:* número de domicílios alugados em relação ao total de domicílios particulares permanentes.
- *Proporção de domicílios com abastecimento de água proveniente da rede pública geral:* número de domicílios com abastecimento de água proveniente da rede pública geral em relação ao total de domicílios particulares permanentes.
- *Proporção de domicílios ligados à rede pública de esgoto:* número de domicílios ligados à rede pública de esgoto em relação ao total de domicílios particulares permanentes.

3.3. Análise exploratória

Na análise exploratória de dados utilizou-se principalmente recursos gráficos, disponíveis no pacote estatístico S-Plus©. Os gráficos da análise univariada apresentam, pela ordem:

- histograma;
- diagrama de caixa ou *boxplot* (Tukey, 1990), onde a linha clara indica a mediana, o símbolo "o", quando presente, a média da distribuição, a altura da caixa o intervalo interquartílico, a linha pontilhada engloba 1,5 vezes o intervalo interquartílico a partir do centro, e os traços horizontais os valores extremos (*outliers*) da distribuição (StatSci,1993);
- gráfico de comparação da distribuição com a distribuição normal (*normal probability plot* ou *qqplot*) - através da comparação entre a distribuição da variável (eixo Y) e uma distribuição Gaussiana com mesma média e desvio padrão (eixo X), é possível identificar se a variável tem distribuição normal (quando os pontos do gráfico localizam-se todos sobre a linha reta traçada); se apresenta desvio (*skewness*), quando os pontos se dispõem em "U"; a forma em "S" indica que é mais espalhada nos extremos que a distribuição Gaussiana.
- sumário das medidas de tendência central e dispersão.

Na análise bivariada utilizou-se o diagrama de espalhamento (*scatter-plot*). Na análise exploratória espacial foram empregados mapas de padrão, feitos com o MAPINFO™, e mapas de contorno, construídos com o S-Plus™.

3.4. Classificação multivariada

Os setores foram agrupados em análise classificatória multivariada pelo algoritmo *k-means* (Hartigan,1975), que é um método não hierárquicos de partição em k grupos usando critérios que diminuem a variância intra-grupos e maximizam a

variância inter-grupos (mais detalhes no Anexo 1). Utilizou-se o pacote estatístico SAS™, selecionando-se inicialmente cinco setores como sementes iniciais. A cada iteração os centróides dos grupos eram atualizados. Grupos com menos de 2% do total de setores censitários foram descartados. O critério de convergência utilizado baseou-se na diferença da distância entre dois centróides em duas passagens: quando esta fosse menor do que 2% da menor distância entre as sementes originais se interrompia a realocação de setores censitários. O número máximo de iterações sem atingir convergência foi definido em 10 passagens. Observou-se empiricamente que, quando não se chegava à convergência até este número, em geral os grupos criados eram pouco consistentes, e ciclos posteriores não melhoravam a convergência.

3.5. Análise exploratória espacial

Em todas as etapas da análise espacial foi utilizando o módulo SpatialStat do S-Plus™. Os mapas de contorno para estudo preliminar de tendência e autocorrelação espacial utilizaram algoritmo de interpolação linear por triangulação. Neste a variável é apresentada em dez classes obedecendo a escala cromática que vai do vermelho ao violeta, segundo a ordem natural: vermelho, laranja, amarelo, verde, azul, violeta e cores intermediárias. Todas as técnicas de interpolação utilizadas estimaram valores para pontos em uma grade regular de 32 por 32.

Foram feitos ainda *boxplot* e gráficos de médias e medianas por linhas e colunas para investigar existência de tendência (Cressie,1991). Para construir estes gráficos a localização dos centróides foi organizada alocando-os em matriz regular, cujas

coordenadas foram divididas por 100. Para os indicadores renda média do chefe da família e proporção de chefes de família com escolaridade igual ou superior ao segundo grau foi feita a rotação dos eixos destes gráficos, em 45º no sentido horário e 30º anti-horário, visando investigar a direção da tendência.

A identificação de pontos atípicos e de vizinhos com grande contraste - para localizar regiões onde não há estacionariedade local - foi feita utilizando os primeiros intervalos entre pares de pontos da nuvem do variograma. Este é definido como o diagrama de espalhamento dos valores do variograma como uma função da distância entre os pares. Foram apresentados apenas os pares separados por distâncias de até 500 m. Também foi utilizado o indicador de contraste proposto por Cressie (1991): a raiz quadrada das diferenças entres os pares. Neste caso foram apresentados os *boxplots* deste indicador, agrupando os pares conforme a distância de separação, até uma distância de 2500 m.

A partir da análise exploratória espacial o indicador de alfabetização foi transformado, passando-se a trabalhar com o *logaritmo da proporção de população analfabeta*.

3.6. Autocorrelação espacial

Foi utilizada a estatística *I* de Moran para detectar a presença de autocorrelação espacial em todos os indicadores (Cliff & Ord, 1981). Além do cálculo da probabilidade (*p*) de ocorrência do valor calculado supondo uma distribuição normal, este foi estimado também através de simulação de Monte Carlo, com 1000 permutações. Utilizou-se

como critério de significância o valor de *p* obtido por simulação de Monte Carlo abaixo de 0,05.

A matriz de vizinhanças utilizada baseou-se exclusivamente no inverso da distância entre os centróides de cada setor censitário. Não foi utilizado qualquer outro ponderador, seja população ou existência de fronteira em comum, uma vez que se trata de área inteiramente urbana, já excluídos os setores com densidade demográfica diferenciada e sem acidentes geográficos ou qualquer restrição ao fluxo de pessoas.

3.7. Modelagem da tendência

Em primeiro lugar buscou-se estabelecer a direção principal da tendência observada nos gráficos de médias e medianas por linhas e colunas, através da rotação das coordenadas, procurando concentrar a maior parte da variação em grande escala, quando possível, em apenas um eixo.

A tendência espacial observada em alguns indicadores foi modelada através de regressão linear local ponderada pela distância (*loess*). A dimensão da janela local da regressão foi escolhida examinando-se graficamente o ajuste obtido contra os resíduos. Uma vez que o objetivo era modelar apenas a tendência em larga escala, quanto maior a janela, menor a captação da variação em pequena escala pelo modelo de regressão local. Por outro lado, quanto menor a janela, menor a capacidade de captar a tendência, que se expressa em larga escala. Assim, foi escolhida a maior janela onde ficasse preservada a ausência de estrutura observável através da inspeção gráfica dos resíduos. No caso de variáveis onde a análise exploratória espacial mostrou apenas a existência de tendência,

sem efeito de segunda ordem visível, utilizou-se para a interpolação o *loess* com janela de 20% do total de pontos. Os resíduos dos modelos *loess* foram analisados através de gráficos de médias e medianas por linhas e colunas e sua autocorrelação espacial foi testada.

3.8. Modelagem por variograma

A estrutura espacial foi analisada através do variograma calculado segundo o método clássico (Isaaks & Shrivastava, 1989) e pelo método robusto (Cressie,1991). O intervalo amostral considerado foi de 250 m, distância suficiente para garantir um número de pares necessário à estabilização do variograma amostral. Foram calculados 20 intervalos, com tolerância de meio intervalo, até a distância de 2.874 m. O primeiro intervalo foi de 95,67 m.

Para estudar possível anisotropia espacial, foram calculados os variogramas amostrais em quatro direções: Norte-Sul (0º), Sudoeste-Nordeste (45º), Leste-Oeste(90º) e Sudeste-Noroeste(135º). Para todos se estabeleceu tolerância angular de 22,5º, ou seja, o intervalo de cada eixo estudado vai até intervalo do próximo eixo. Os gráficos com os variogramas direcionais apresentam uma barra superior onde para cada direção é colocado o intervalo de tolerância, em coloração diferente. Nestes gráficos os quatro variogramas são representados na mesma escala visando comparabilidade. Nos variogramas direcionais utilizou-se também o método robusto de cálculo (Cressie,1991), nos quais se eliminou os pontos do primeiro intervalo, por apresentarem pequeno número de pares.

Para estudar as diferenças quanto ao alcance entre as direções, foi feito um gráfico dos variogramas direcionais, somente até 2000 m, e interpolou-se os valores obtidos através de regressão local ponderada (*loess*, com janela abrangendo 2/3 do total de pontos). A interpolação linear, como em curvas de nível, permitiu identificar a distância, em cada direção, onde o variograma atingiu determinado valor (γ) - aproximadamente 70% do valor máximo (*range*) para cada variável. Esta distância está registrada nos gráficos (**d0=**distância). Neste caso, o *loess* foi utilizado apenas para permitir uma interpolação dos pontos do variograma, diminuindo a flutuação aleatória ocasionada pelo pequeno número de pares em cada direção. As distâncias assim estimadas foram colocadas em diagrama direcional, *rose diagram*, cujo objetivo é identificar direção e razão de anisotropia. A escolha da correção da anisotropia geométrica foi feita a partir de gráfico onde se apresenta o variograma omnidirecional corrigido para anisotropia segundo duas direções e razões.

Utilizou-se dois métodos de ajuste para a seleção do modelo do variograma: mínimos quadrados não lineares e mínimos quadrados ponderados (Cressie,1991), sempre utilizando o variograma amostral somente até a distância de 3.000 m, ou 12 intervalos. Foram testados diferentes modelos: esférico, Gaussiano e exponencial para cada uma das variáveis. A seleção final entre dois modelos diferentes foi feita a partir do somatório do quadrado do resíduo entre cada modelo e variograma amostral, cujo valor é apresentado nos gráficos para cada modelo e método de ajuste. Quanto aos parâmetros, optou-se sempre pelo modelo ajustado a partir do método ponderado. O ajuste do modelo foi analisado por inspeção visual através do mapeamento dos pares de pontos excessivamente diferentes (acima do percentil 99) ou excessivamente

semelhantes (abaixo do percentil 1), dentro de um raio de 500 m. Ou seja, se espera que apenas 1% dos pares separados por distância de até 500 m sejam mais ou menos semelhantes do que o previsto pelo modelo.

3.9. Krigeagem

A interpolação foi feita utilizando a Krigeagem universal, onde são ajustados simultaneamente polinômio e variograma. Utilizou-se os parâmetros e modelo de variograma estimados para os resíduos do *loess*, como aproximação inicial. Dados estes valores, ajustou-se o polinômio, utilizando inicialmente 5 termos: coordenada X, quadrado da coordenada X, coordenada Y, quadrado da coordenada Y, interação entre coordenadas X e Y. Os termos cujos coeficientes fossem inferiores à metade do termo com maior coeficiente foram eliminados, e os resíduos deste polinômio foram novamente modelados através do variograma. Os parâmetros assim obtidos foram reutilizados na estimativa dos coeficientes do polinômio, e assim sucessivamente, até que a variação em qualquer dos parâmetros do variograma fosse inferior a 2%.

3.10. Reclassificação: análise multivariada sobre modelagem espacial

Os indicadores de saneamento - *proporção de casas ligadas à rede de água* e *proporção de casas ligadas à rede de esgotos* - foram modelados através de interpolação linear simples por triangulação de Delaunay (Bailey & Gatrell, 1995). As variáveis *renda média do chefe da família, proporção de chefes de família com escolaridade igual ou superior ao segundo grau* e *logaritmo da proporção de*

analfabetos foram modeladas por Krigeagem universal. A variável *proporção de casas alugadas* foi modelada através do *loess*. Sobre os valores estimados por estas técnicas para uma grade regular de 32 por 32 foi feita classificação multivariada aplicando exatamente os mesmos procedimentos utilizados anteriormente.

Visando comparar os resultados, a classificação baseada nas variáveis originais foi transposta para grade regular de 32 por 32, onde cada ponto recebe a classificação do setor censitário dentro do qual está contido.

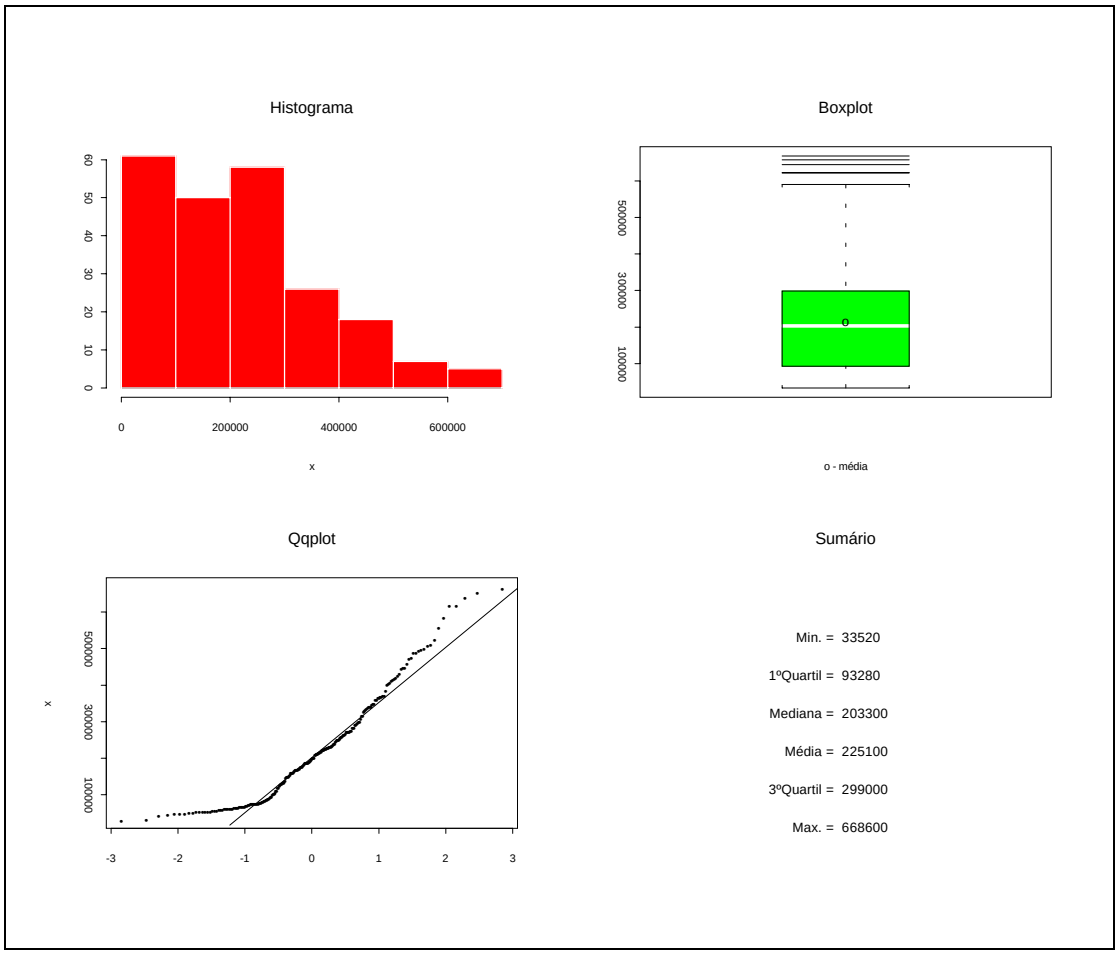
A comparação baseou-se na diferença simples entre as duas categorizações obtidas, atribuindo-se valor 1 ao grupo com melhor situação socioeconômica, 2 ao seguinte, e assim sucessivamente, hierarquizando a classificação. Quanto maior o valor absoluto, maior a diferença entre os dois modelos; diferença positiva indica que no modelo baseado na interpolação espacial das variáveis, o ponto estimado pertence a um grupamento com piores indicadores socioeconômicos e vice-versa.

4. RESULTADOS

4.1. Análise exploratória

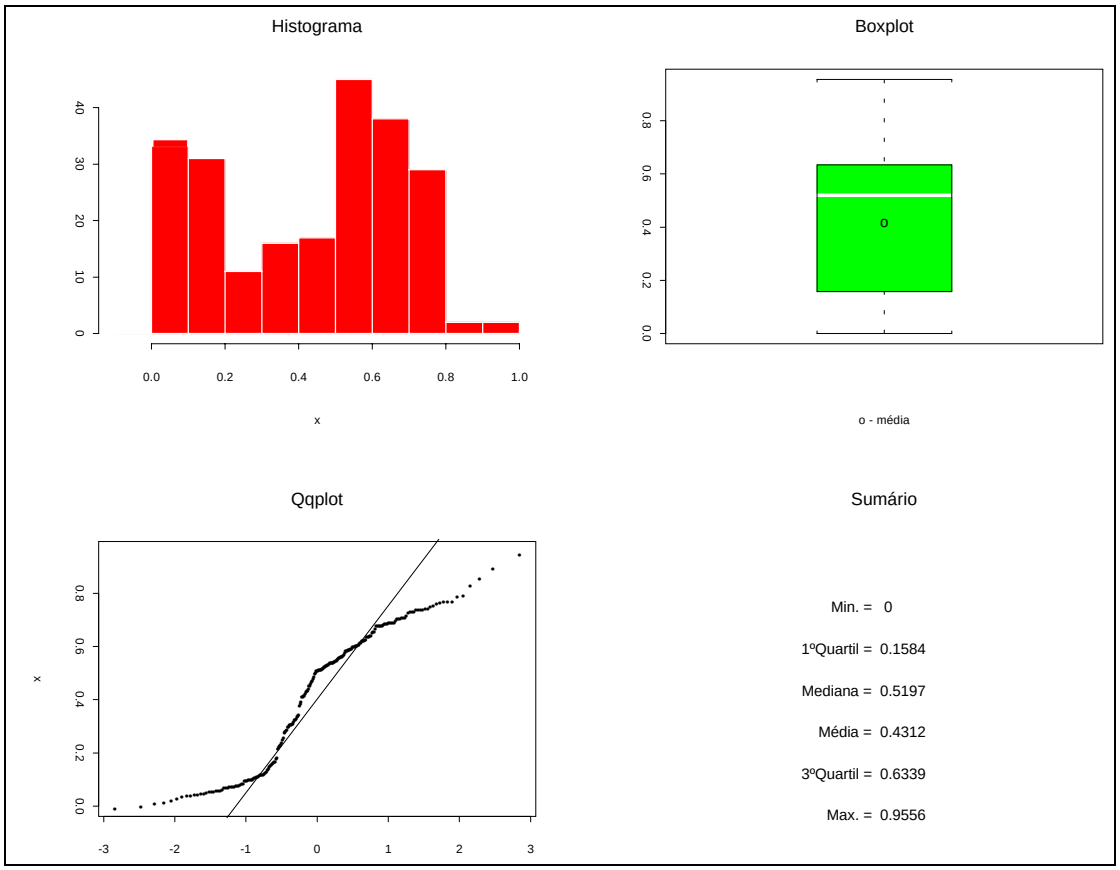
Nos Gráficos 4.1 a 4.6 estão apresentados os resultados da análise exploratória univariada dos indicadores. A renda média do chefe da família (Gráfico 4.1) apresenta distribuição com desvio para a esquerda, e média e mediana bastante próximas. Pode-se observar valores atípicos no extremo superior da distribuição, onde a renda média do chefe da família é muito superior. O *qqplot* apresenta boa aderência apenas no centro da distribuição.

Gráfico 4.1
Análise exploratória da Renda Média do Chefe da Família (bruta)
Ilha do Governador, Censo Demográfico de 1991



A proporção de chefes da família com segundo grau completo (Gráfico 4.2) apresenta distribuição bi-modal, caracterizando dois tipos de setores censitários: aqueles onde a escolaridade é baixa - menos de 20% dos chefes de família apresentam o segundo grau - e aqueles onde pelo menos 60% tem escolaridade mais elevada. A média é inferior à mediana da distribuição, e o *qqplot* sugere que a variável é oriunda de duas populações diferentes.

Gráfico 4.2
Análise exploratória da *Proporção de Chefes da Família com 2° Grau Completo*
Ilha do Governador, Censo Demográfico de 1991



Os indicadores relacionados a saneamento - proporção de casas ligadas à rede pública de água (Gráfico 4.3) e proporção de casa ligadas ao sistema de esgotos (Gráfico 4.4) - apresentam distribuição semelhante, muito concentrada nos valores mais altos indicando boa cobertura na região. É expressiva a presença de setores atípicos. No caso do esgoto, um grupo de setores apresenta grande parte das residências sem ligação à rede, segundo uma distribuição em dois estratos, conforme pode ser visualizado no

qqplot. Estas duas variáveis permitem identificar setores onde parte das residências carece de acesso a equipamentos urbanos básicos considerados como bem disseminados na área estudada.

Gráfico 4.3
Análise exploratória da *Proporção de Casas Ligadas na Rede Geral de Água*
Ilha do Governador, Censo Demográfico de 1991

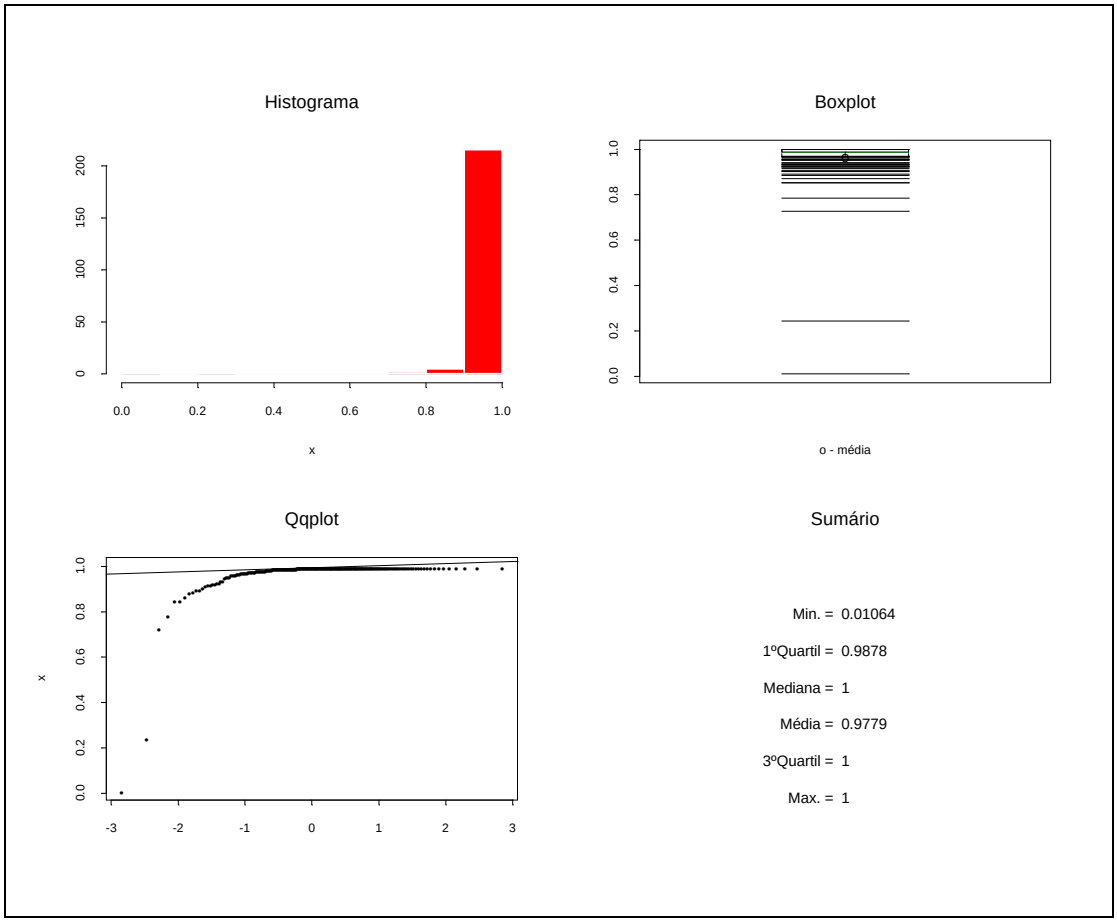
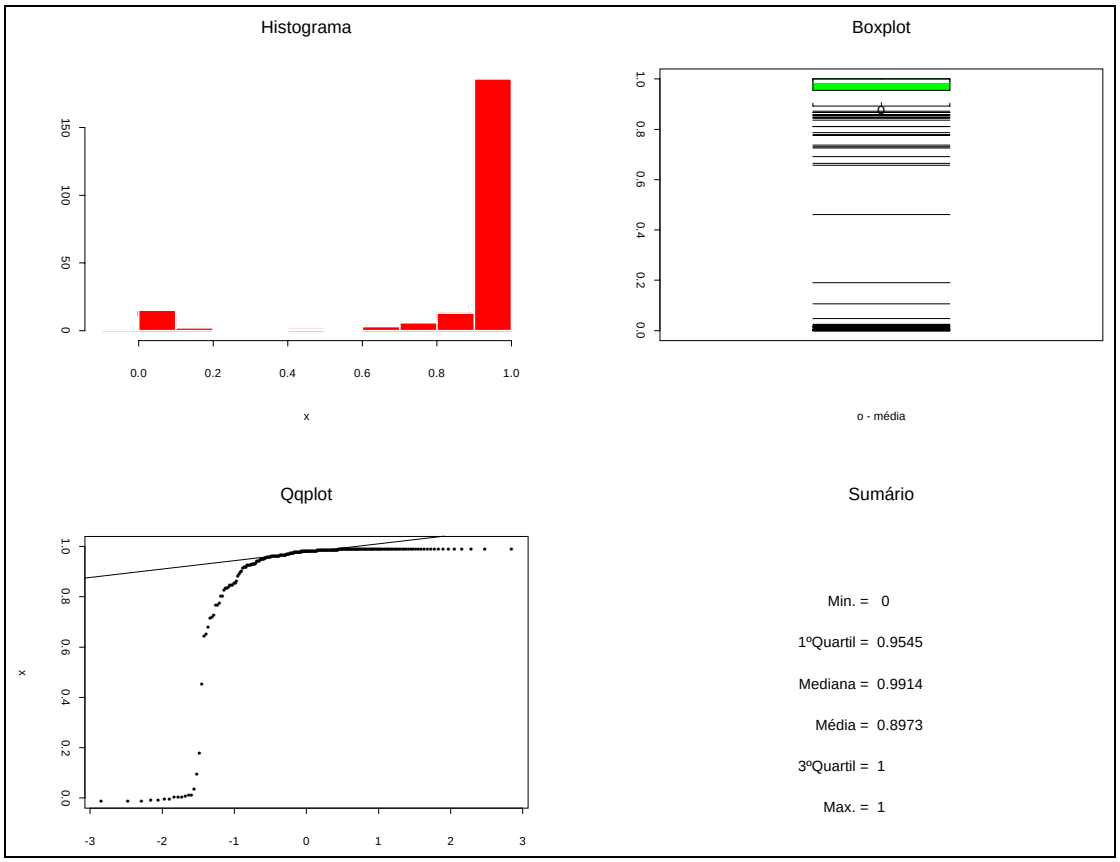
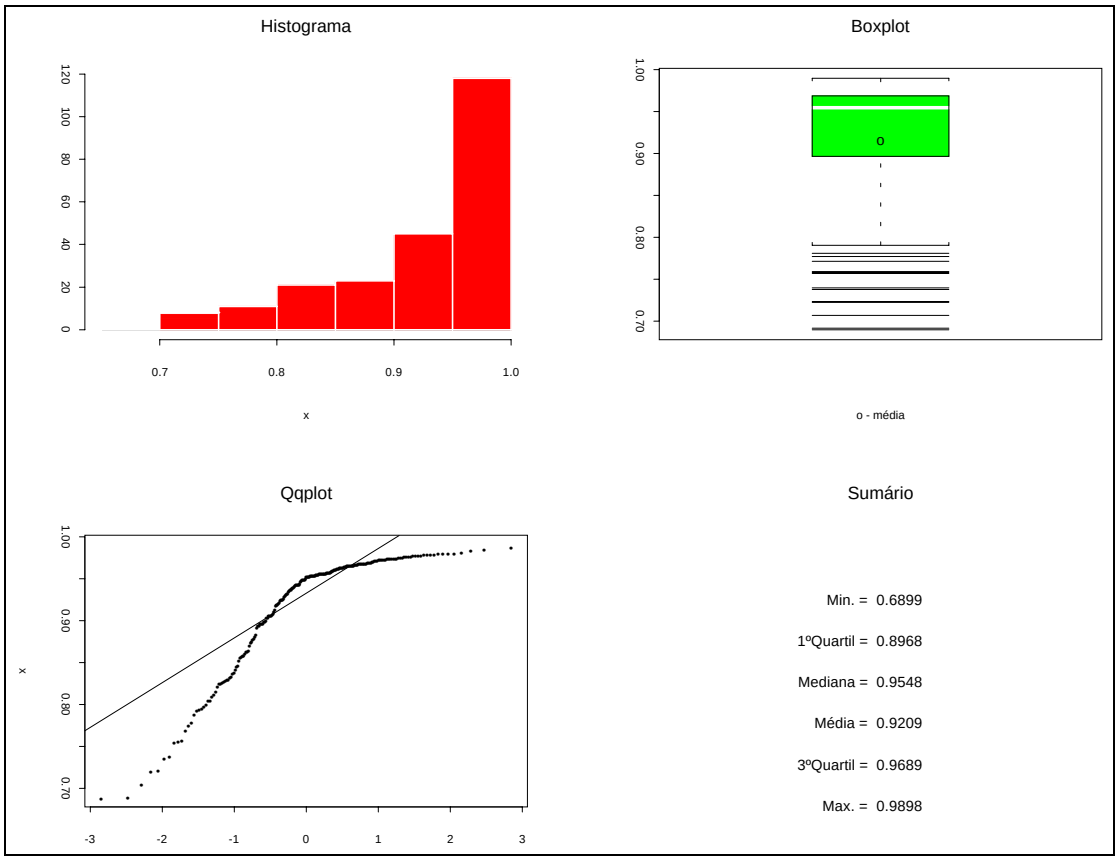


Gráfico 4.4
Análise exploratória da *Proporção de Casas Ligadas à Rede de Esgotos*
Ilha do Governador, Censo Demográfico de 1991



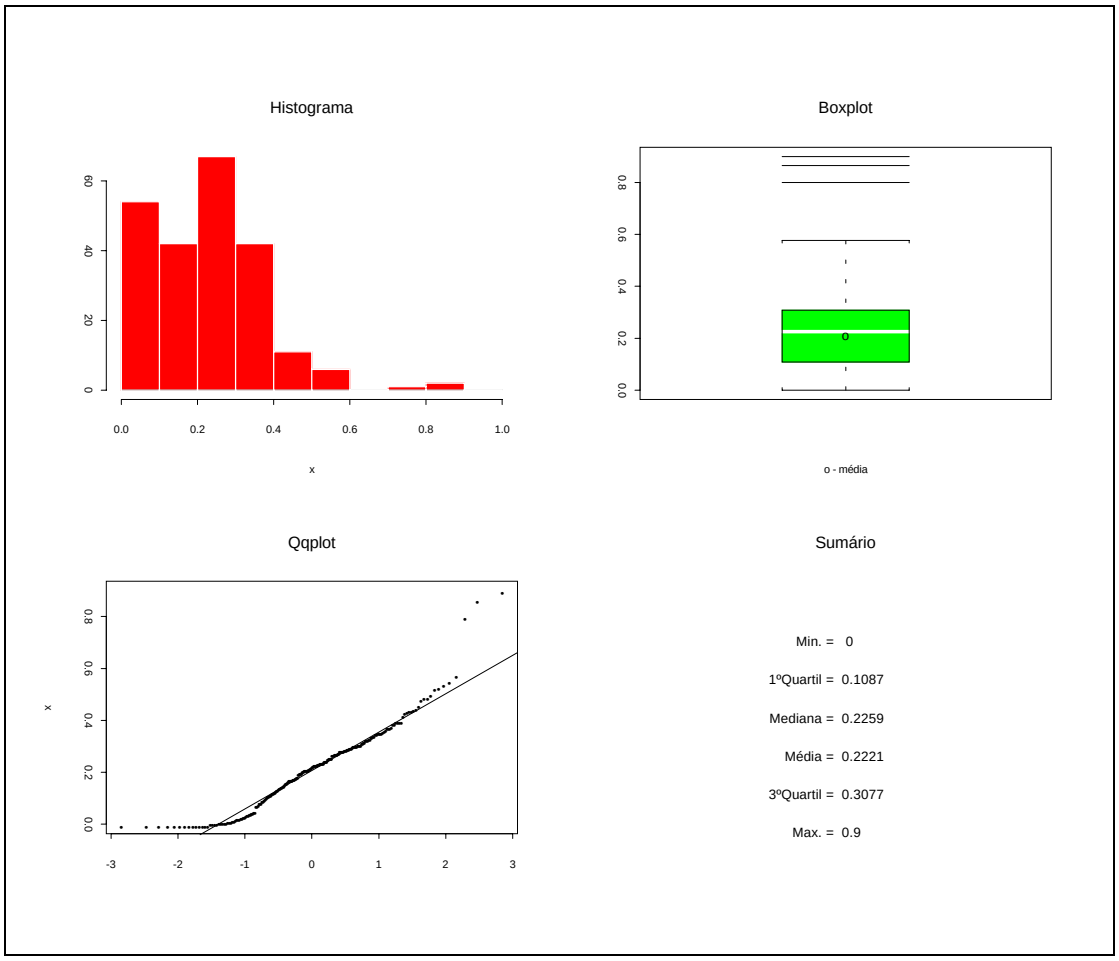
A proporção de população alfabetizada (Gráfico 4.5) apresenta curva de distribuição semelhante a uma exponencial, concentrando-se em valores próximos a 100%. A média (92,1%) e a mediana (95,5%) indicam que apenas pequena parcela da população acima de cinco anos não sabe ler e escrever. Entretanto, a existência de grande número de setores com valores atipicamente baixos sugere que os analfabetos não se distribuem homogeneamente por toda a área - acompanhando a distribuição de crianças ainda não alfabetizadas - mas concentram-se em algumas áreas específicas.

Gráfico 4.5
Análise exploratória da *Proporção de População Alfabetizada*
Ilha do Governador, Censo Demográfico de 1991



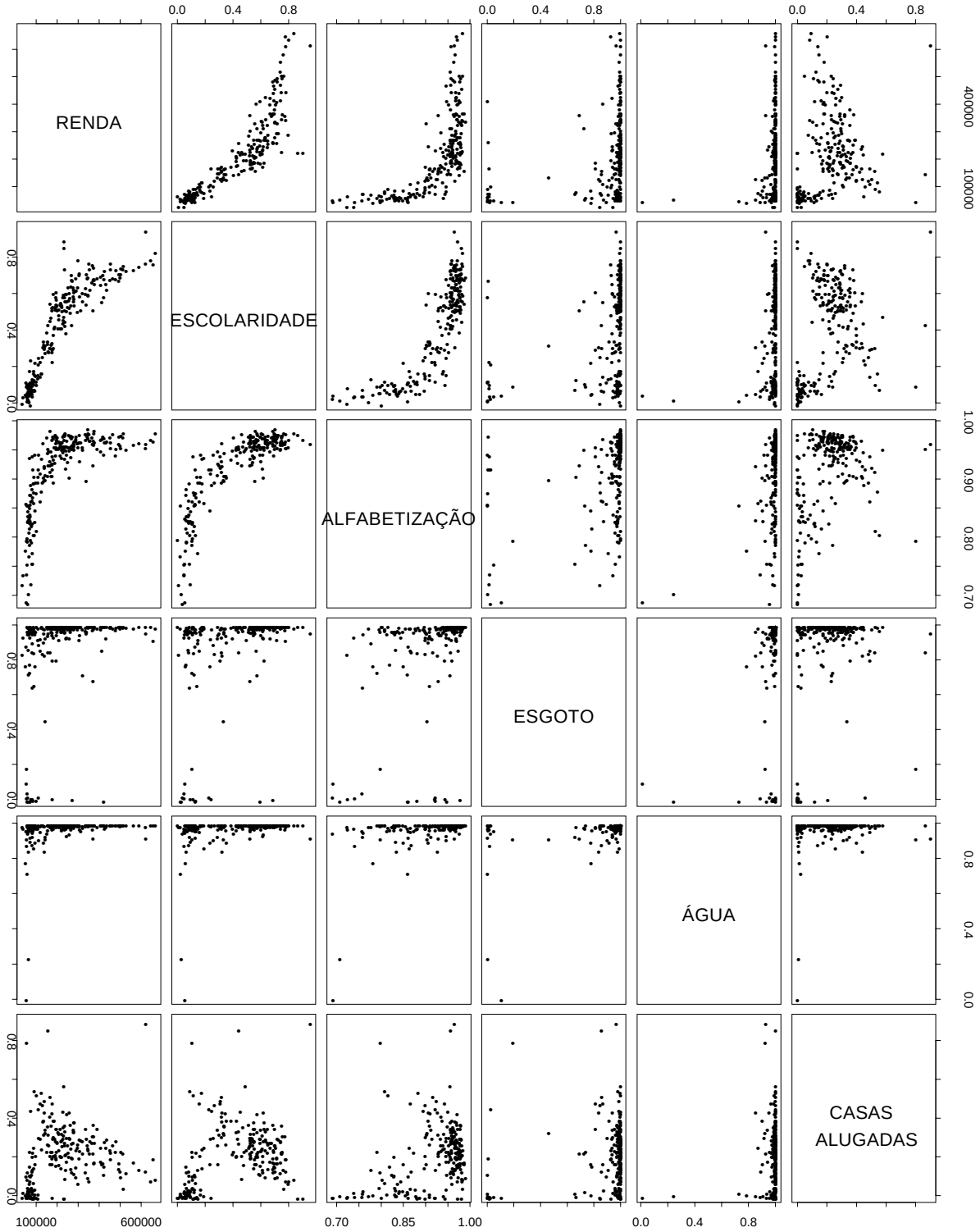
A proporção de casas alugadas (Gráfico 4.6) é, em geral baixa, (média=22,2%), e em 1/4 dos setores inferior a 11%.

Gráfico 4.6
Análise exploratória da *Proporção de Casas Alugadas*
Ilha do Governador, Censo Demográfico de 1991



O cruzamento dos indicadores dois a dois (Gráfico 4.7) mostra que as variáveis renda média, escolaridade do chefe da família e alfabetização apresentam correlação positiva, embora não linear. As demais variáveis, com seu perfil de distribuição muito concentrado, não apresentam correlação perceptível graficamente.

Gráfico 4.7
Diagrama de espalhamento das variáveis socioeconômicas duas a duas
Ilha do Governador, Censo Demográfico de 1991



4.2. Classificação multivariada

A partir das variáveis selecionadas foram obtidos cinco grupos de setores censitários. A Tabela 4.1 apresenta o valor médio de cada indicador nos grupos criados, o número de setores censitários e, para cada variável, o valor de $R^2/(1-R^2)$, onde R^2 mede a variância entre os grupos, e o $(1-R^2)$ mede a variância intra-grupo. Assim, quanto mais alto o valor desta razão, maior o peso do indicador na construção dos grupos. Renda e escolaridade do chefe da família, seguidos de proporção de casas ligadas à rede de esgotos e proporção de população alfabetizada, foram as variáveis que mais contribuíram para a classificação obtida. Ainda que os indicadores de acesso à rede de água e de casas alugadas apresentem $R^2/(1-R^2)$ próximo a zero, foram mantidos por permitirem melhor descrição dos grupos criados, sem alterar sua composição.

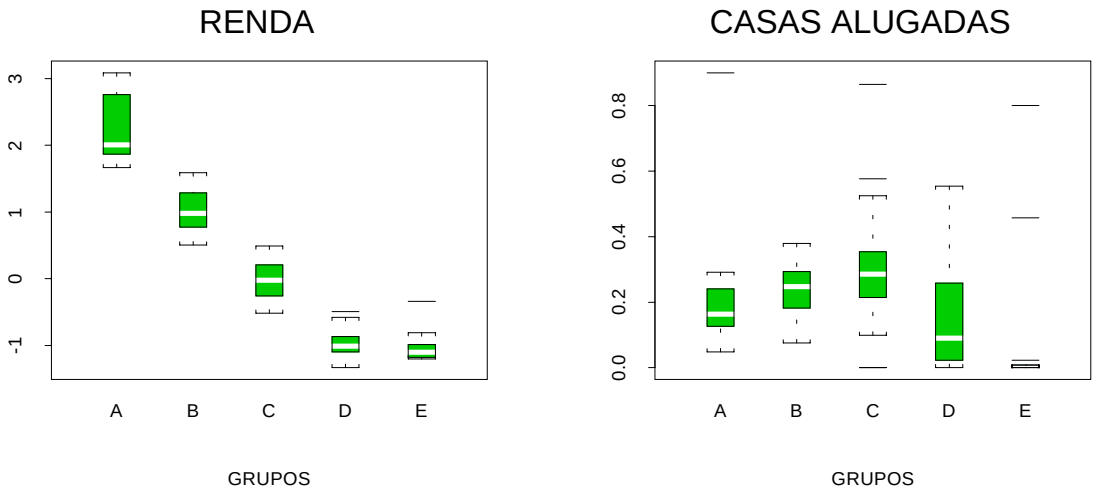
Tabela 4.1
Grupos segundo número de setores censitários componentes (N), média de cada indicador, e valor da razão variância entre (R^2)/variância intra grupos ($1-R^2$).

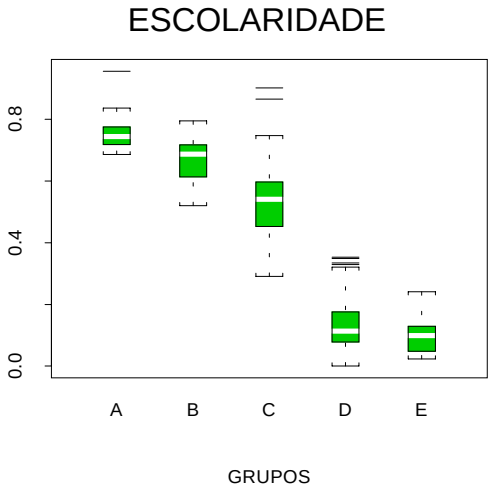
Grupo	N	Renda*	Escolar.	Alfabet.	Esgoto	Água	Aluguel
A	18	2,238	0,755	0,967	0,990	0,995	0,210
B	40	1,002	0,671	0,967	0,949	0,995	0,238
C	87	-0,027	0,528	0,957	0,971	0,995	0,286
D	66	-0,964	0,140	0,855	0,926	0,973	0,159
E	14	-1,025	0,100	0,815	0,033	0,827	0,093
Total	**225**	**0**	**0,431**	**0,920**	**0,897**	**0,978**	**0,222**
$\frac{R^2}{1-R^2}$	**-**	**11,73**	**5,94**	**1,91**	**3,91**	**0,26**	**0,18**

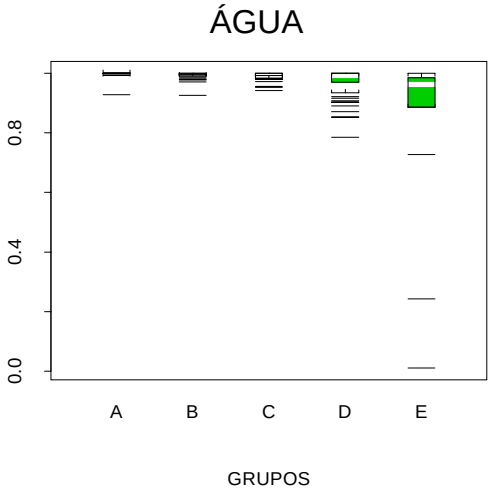
* Renda reescalonada para média zero, desvio padrão 1

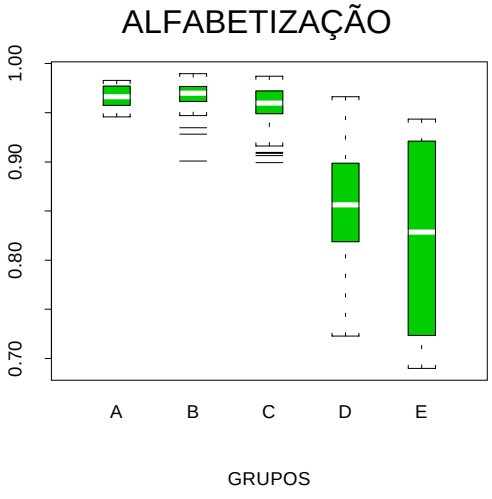
No Gráfico 4.8 são apresentados os *boxplots* de cada indicador segundo a classificação criada. A renda média do chefe da família, variável com maior peso na construção dos grupos, apresenta padrão decrescente a partir do grupo *A*, com pequena superposição dos valores, exceto nos dois últimos, com renda semelhante. A proporção de chefes de família com escolaridade igual ou superior ao segundo grau, ainda com padrão decrescente entre os grupos, os divide em dois tipos: de *A* a *C*, com valores altos, *D* e *E* em patamar bem inferior. O histograma deste indicador já sugeria um comportamento nitidamente dual. A proporção de casas ligadas à rede de esgotos apresenta apenas o grupo *E* com valores discrepantes dos demais. Neste grupo se concentrou a pequena parcela de setores censitários, observada no histograma, cuja proporção de casas ligadas à rede de esgotos era próxima a zero. A proporção de população alfabetizada apresenta o limite superior da distribuição muito semelhante entre os grupos, variando a amplitude da distribuição, mais uma vez discrepante nos grupos *D* e *E*. A proporção de casas ligadas à rede de água pouco discrimina os grupos. Entretanto, apenas no grupo *E* a distribuição atinge valores muito baixos. A proporção de casas alugadas pouco difere entre os grupos. Esta é a única variável que não decresce sempre entre os grupos *A* e *E*, onde a mediana aumenta gradativamente até o grupo *C* e depois volta a cair.

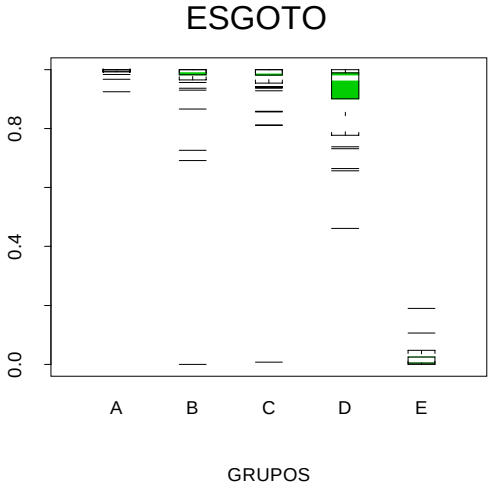
Gráfico 4.8
***Boxplots* dos indicadores socioeconômicos segundo classificação**
Ilha do Governador, Censo Demográfico de 1991











Os grupos criados podem ser descritos sumariamente conforme se segue:

A → composto por apenas 18 setores, apresenta a maior renda média do chefe da família, melhor situação em relação a todos os indicadores, proporção de casas alugadas mais baixa que a média geral;

B → com 40 setores, tem a renda média e a escolaridade do chefe da família um pouco menor, demais indicadores semelhantes, maior proporção de casas alugadas que em *A*;

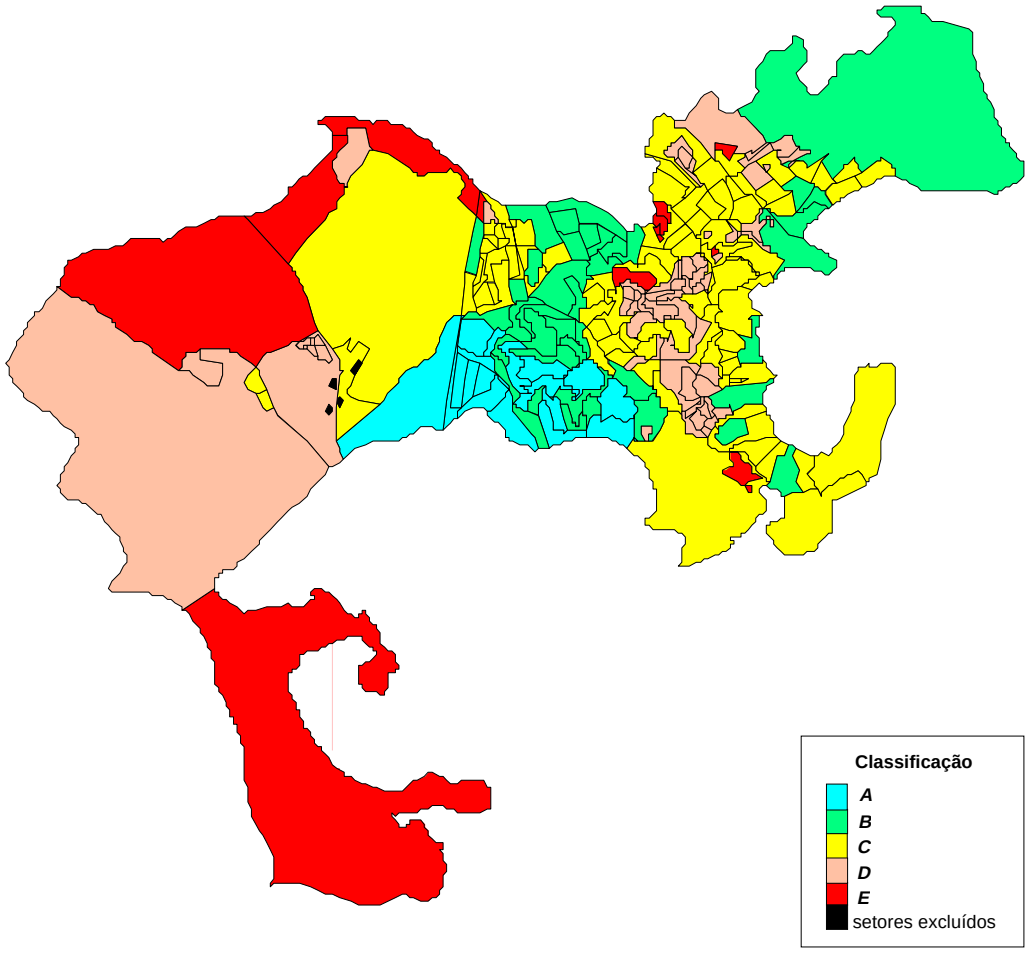
C → no maior grupo (87 setores), a renda média do chefe da família encontra-se na média geral da população (-0,027 desvios padrão) e a proporção de chefes de família com escolaridade igual ou superior ao segundo grau é 52,8% (superior à média geral de 43,1%); o acesso às redes de água e de esgotos e a proporção de população alfabetizada apresentam distribuição idêntica aos grupos anteriores, e a proporção de casas alugadas tem a maior média (28,6%);

D → com 66 setores apresenta, junto com o grupo *E*, as menores renda média do chefe da família e proporção de chefes de família com escolaridade igual ou superior ao segundo grau, diferindo substancialmente pela proporção de casas ligadas à rede de esgotos; a proporção de população alfabetizada apresenta média (85,5%) inferior à média geral (92,1%);

E → com 14 setores é o menor grupo, caracterizando-se principalmente pela proporção de casas ligadas à rede de esgotos no extremo inferior da distribuição; embora a proporção média de casas ligadas à rede pública de água não seja muito diferente do restante, a distribuição inclui os valores mais baixos; a proporção de população alfabetizada apresenta a menor média e a maior amplitude de distribuição e praticamente inexistem casas alugadas neste grupo.

O Gráfico 4.9 apresenta os setores censitários da 20ª Região Administrativa (Ilha do Governador) segundo o resultado da classificação socioeconômica multivariada. No artigo *Spatial partitioning using multivariate cluster analysis and a contiguity algorithm* (Anexo 2) é apresentada discussão mais detalhada da análise de aglomerados na região.

Gráfico 4.9
Classificação dos Setores Censitários
Ilha do Governador, 1991



A malha dos setores censitários foi apresentada no Gráfico 3.1. A distribuição espacial dos setores do aeroporto, cidade universitária e indústria naval, com grande extensão territorial e pequena população, orientou a retirada destas áreas na etapa da análise espacial, tendo sido incluídos apenas 203 setores censitários, com visível continuidade na ocupação do território (Gráfico 3.2).

4.3. Autocorrelação espacial

Os resultados da estatística *I* de Moran para o teste de autocorrelação espacial são apresentados na Tabela 4.2. Utilizando como critério de significância o valor de $p < 0,05$ (por simulação de Monte Carlo), foi detectada presença de autocorrelação espacial em todos os indicadores, exceto acesso à rede pública de água.

Tabela 4.2
Indicadores segundo estatísticas de autocorrelação espacial

Indicador	Autocorrelação	*"I"*	p (bicaudal)	p (permut.)
Renda	0,7421	13,59	4,795e-42	0
Escolaridade	0,8399	15,37	2,746e-53	0
Alfabetização	0,7957	14,56	4,933e-48	0
Rede de Esgotos	0,3642	6,715	1,88e-11	0
Água	-9,066e-4	0,07355	0,9414	0,35
Casas Alugadas	0,3421	6,313	2,736e-10	0

Renda, escolaridade e alfabetização apresentaram forte autocorrelação espacial. A proporção de casas alugadas e de casas ligadas a rede de esgotos, embora tenham *p* abaixo do limiar estipulado (0,05), apresentam correlação fraca. A proporção

de residências ligadas à rede pública de água não apresenta qualquer correlação espacial.

4.4. Análise exploratória espacial: detectando tendência

Nos Gráficos 4.10 a 4.15 são apresentados os mapas de contornos dos indicadores. Embora o diagrama de espalhamento (Gráfico 4.7) indicasse correlação entre os indicadores de renda e escolaridade, nestes mapas a sobreposição entre as áreas de alta renda e alta escolaridade não é tão evidente. A renda apresenta superfície com transição suave, e aparente tendência na direção de Sudoeste-Nordeste (Gráfico 4.10). As áreas de renda mais baixa (de vermelho a amarelo) ocupam a maior parte de território, com pequenas ilhas de renda mais alta - desvio padrão acima da média global - concentradas mais a Sudoeste. A escolaridade (Gráfico 4.11) apresenta região de altas proporções mais extensa, mantendo entretanto aproximadamente a mesma distribuição territorial, sugerindo que as áreas em verde e azul no mapa de renda se espalharam sobre as áreas amarelas e laranjas neste. Estas diferenças entre os mapas refletem as diferenças também encontradas nos histogramas das variáveis, onde a escolaridade apresenta uma parcela importante dos setores censitários distribuídos entre os de alta proporção de chefes com escolaridade acima do segundo grau, que ficam, conseqüentemente, mais visíveis no mapa.

Gráfico 4.10
Mapa de contorno da *Renda Média do Chefe da Família*
Ilha do Governador, excluída área sem continuidade espacial

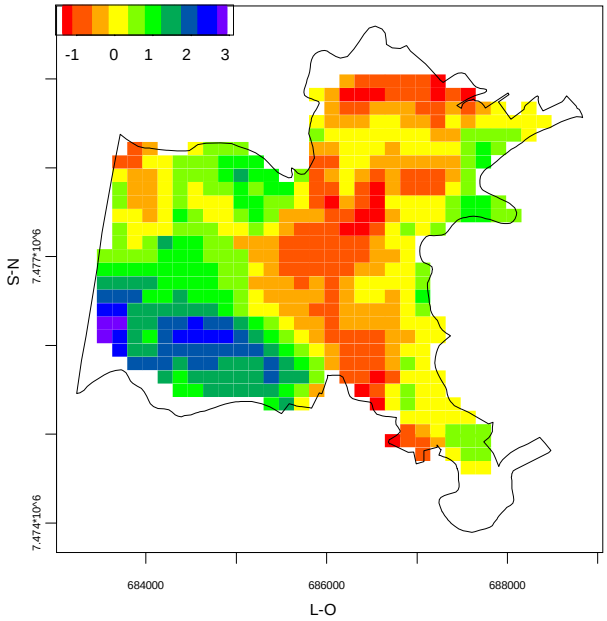
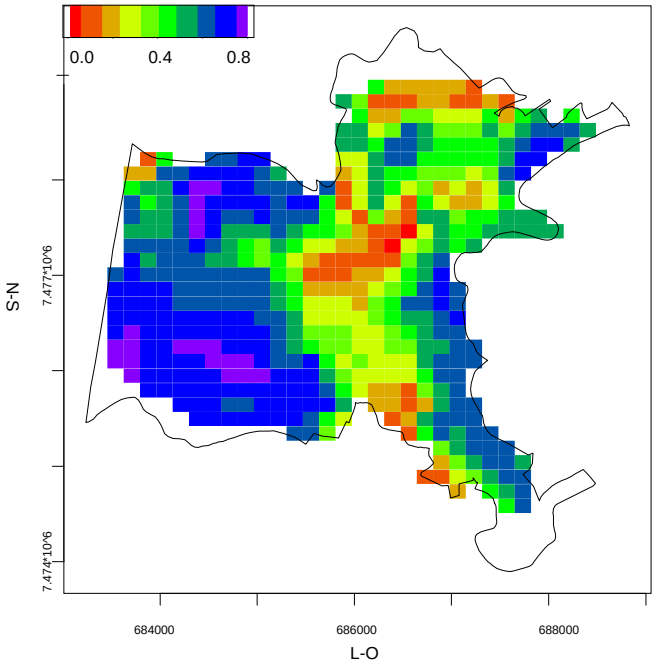


Gráfico 4.11
Mapa de contorno da *Proporção de Chefes da Família com 2° Grau Completo*
Ilha do Governador, excluída área sem continuidade espacial



A proporção de casas ligadas à rede pública de água é praticamente uniforme, com pequenas “ilhas” sem acesso a este recurso (Gráfico 4.12). O acesso à rede de esgoto apresenta áreas maiores com valores baixos (Gráfico 4.13). Ambos apresentam transições bruscas entre áreas de altos e baixos valores. É interessante observar que, embora exista superposição de áreas carentes de rede de água e de esgoto, em diversos pontos cada fator ocorre independentemente.

Gráfico 4.12
Mapa de contorno da *Proporção de Casas Ligadas na Rede Geral de Água*
Ilha do Governador, excluída área sem continuidade espacial

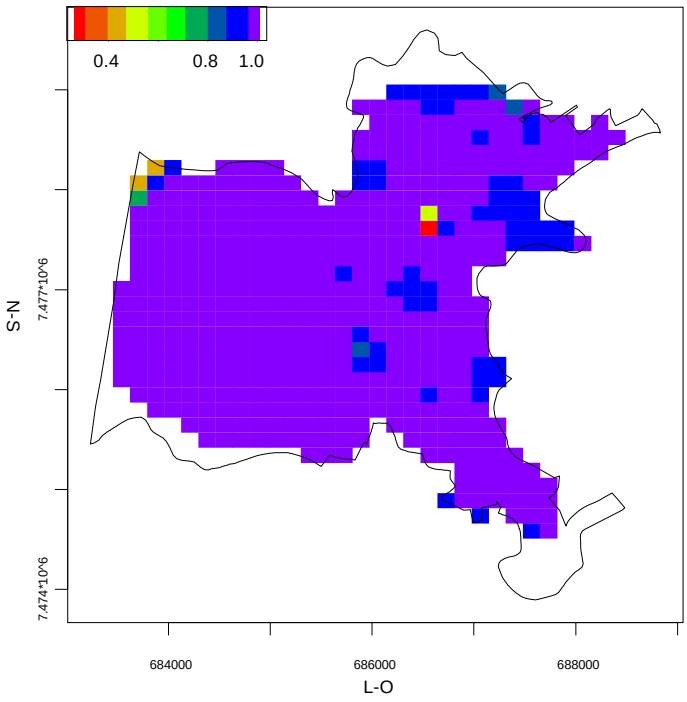
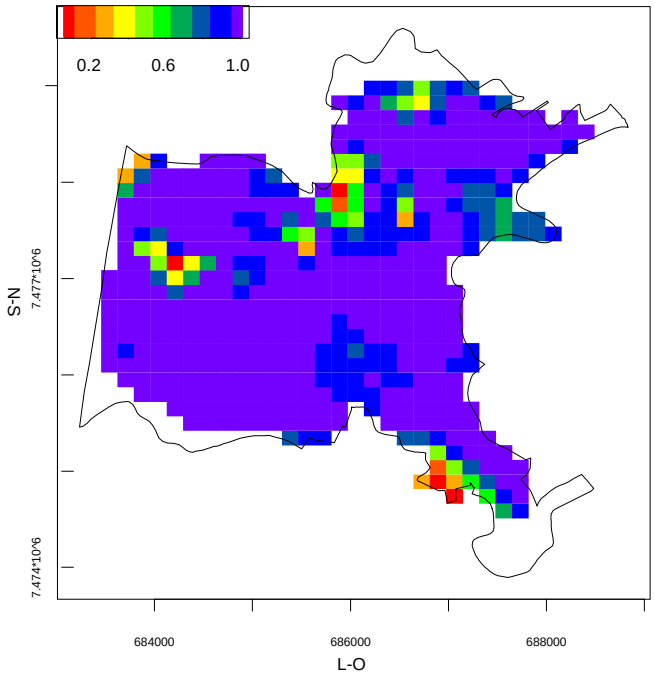


Gráfico 4.13
Mapa de contorno da *Proporção de Casas com Ligação à Rede de Esgotos*
Ilha do Governador, excluída área sem continuidade espacial



A proporção de população alfabetizada (Gráfico 4.14) ainda apresenta extensas áreas de altos valores, mas as transições são menos abruptas que as anteriores. A proporção de casas alugadas (Gráfico 4.15) tem padrão bem diferente de todos os demais: pequenas áreas no extremo superior e no centro do mapa, em região de baixos valores para renda e escolaridade, concentram a maior parte dos setores censitários com residências alugadas.

Gráfico 4.14
Mapa de contorno da *Proporção de População Alfabetizada*
Ilha do Governador, excluída área sem continuidade espacial

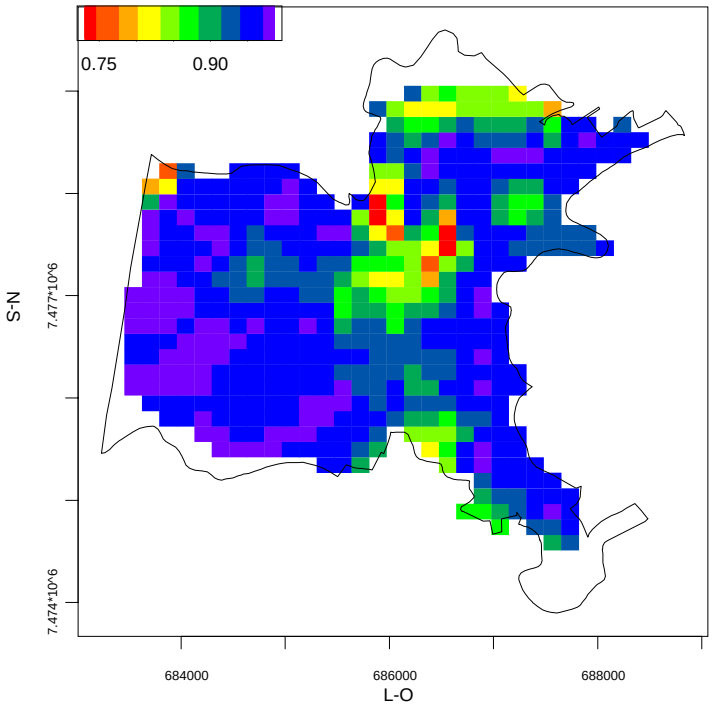
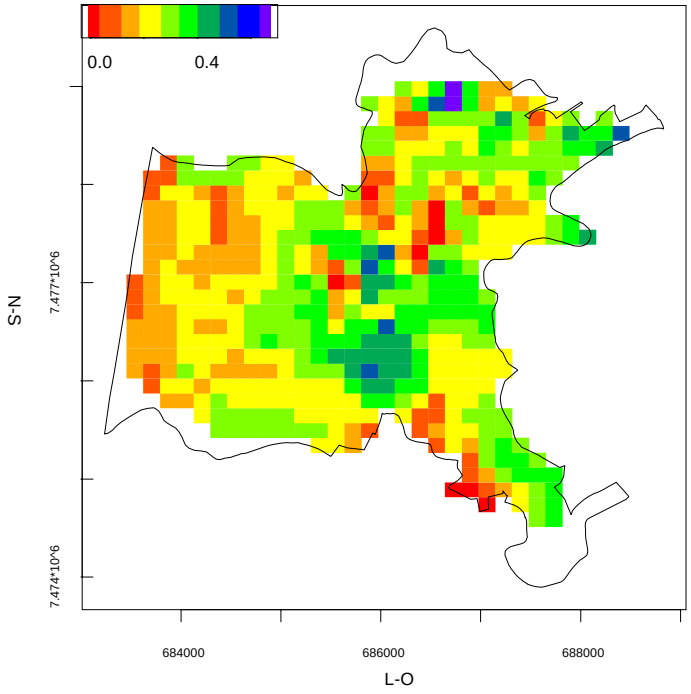


Gráfico 4.15
Mapa de contorno da *Proporção de Casas Alugadas*
Ilha do Governador, excluída área sem continuidade espacial



Visando identificar com mais precisão a tendência observada nos mapas de contorno foram criados gráficos comparando médias e medianas dos indicadores ao longo de duas direções - Leste-Oeste e Norte-Sul. Estes gráficos permitem visualizar possíveis eixos de tendência espacial e valores extremos dos indicadores selecionados.

A proporção de residências com acesso à rede pública de água ou com acesso à rede de esgotos não apresenta direção preferencial: somente algumas linhas e colunas onde se localizam valores atípicos (Gráfico 4.16).

No Gráfico 4.17 pode-se observar que a renda média do chefe da família apresenta entre as colunas 10 e 30 do gráfico por colunas tendência decrescente visível, para médias e medianas. Nas linhas, também parece haver algum deslocamento para valores mais altos no sentido leste. A proporção de chefes de família com escolaridade igual ou superior ao segundo grau é bastante semelhante ao anterior, embora o trecho de declive nas colunas seja mais compacto, ocupando uma região menor. Esta possível tendência nos dois indicadores estaria dividida entre linhas e colunas, sugerindo uma direção de 45º, que será explorada através da rotação dos eixos de referência. A proporção de casas alugadas não apresenta nas linhas qualquer indicação de tendência, mas nas colunas há um aumento suave, porém regular, das médias e medianas ao longo de toda a região. Neste caso, a orientação da tendência é ao longo do eixo Leste-Oeste.

Gráfico 4.16
Médias e medianas por linhas e colunas:
Ligação às redes de Água e Esgotos
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial

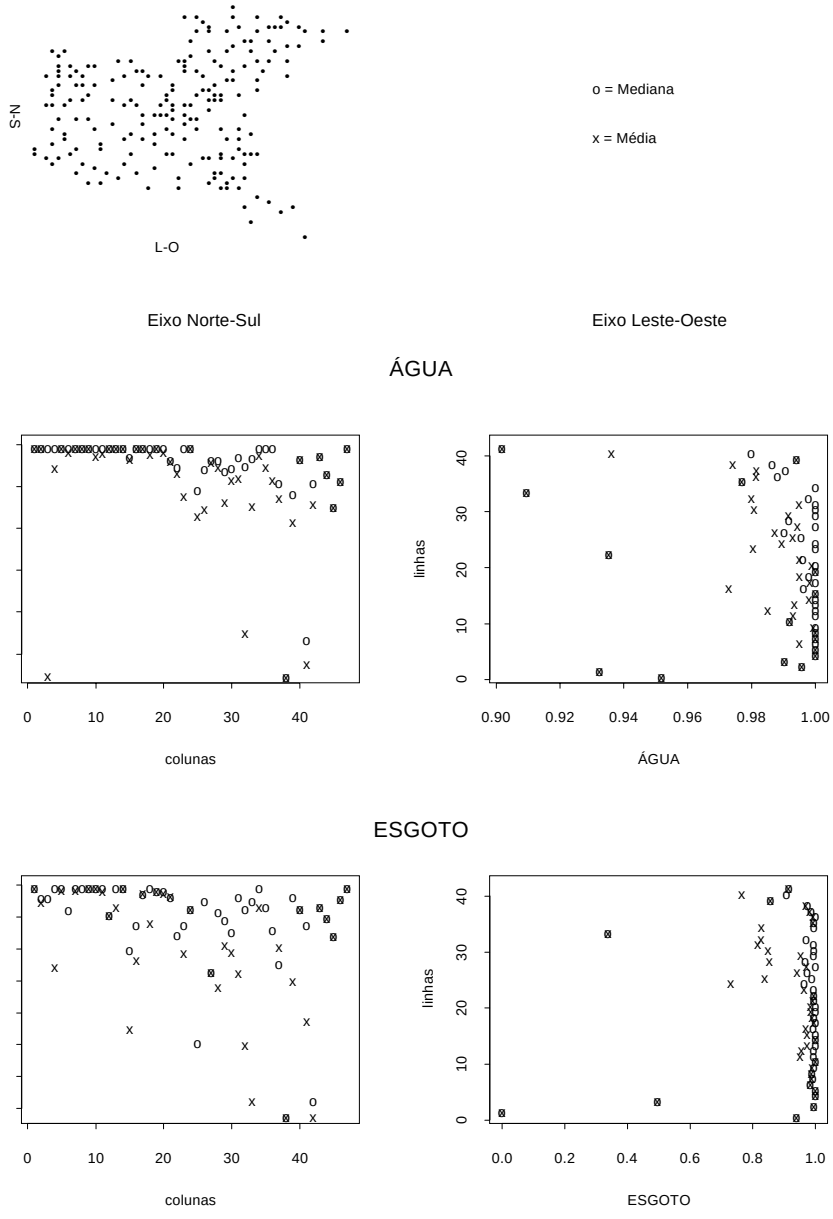
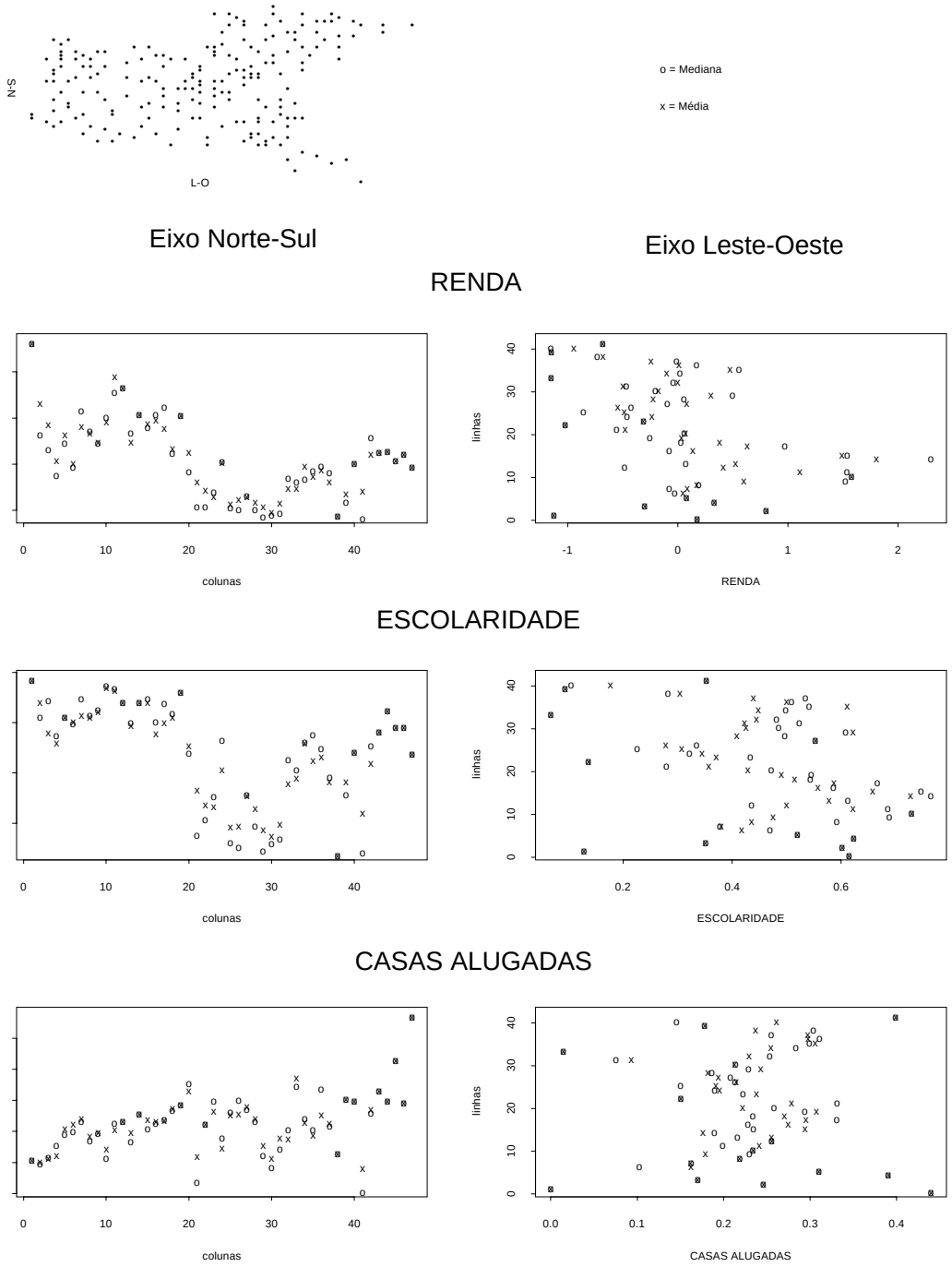


Gráfico 4.17
Médias e medianas por linhas e colunas:
Renda e Escolaridade do Chefe da Família;
Proporção de Casas Alugadas
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial

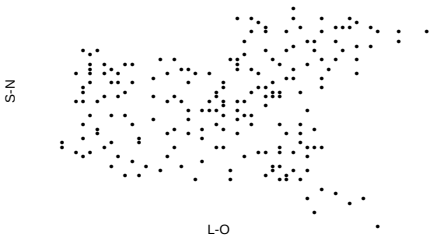


Em relação à proporção de população alfabetizada (Gráfico 4.18) o gráfico por linhas somente apresenta variações que podem ser caracterizadas enquanto valores extremos, e no gráfico por colunas, no trecho entre a coluna 20 e 35, encontram-se em valores mais baixos. Não foi possível, entretanto, observar uma linha claramente definida, ficando inconclusivo o diagnóstico da tendência. Entretanto, frente ao tipo de distribuição do indicador, optou-se por realizar transformação logarítmica no complemento da variável, que pode então ser caracterizada como percentual de analfabetos, visando aproximar sua distribuição de uma normal, e refazendo em seguida a análise exploratória.

O gráfico de médias e medianas da variável transformada (Gráfico 4.18) permite visualizar mais claramente um trecho ascendente nas colunas, até a coluna 25, sendo interessante verificar o comportamento após rotação dos eixos de coordenadas.

O histograma e *boxplot* da variável transformada apresentam perfil de distribuição menos concentrado em valores altos, embora ainda com acentuado desvio à esquerda (Gráfico 4.19). O mapa de contorno neste gráfico permite verificar que a nova variável não apresenta transições tão bruscas como antes, indicando haver tendência discreta na direção Sudoeste-Nordeste, no trecho entre o limite inferior esquerdo e o meio da imagem. Foram estudados o comportamento dos indicadores de renda e escolaridade do chefe da família em duas rotações: a 45º (sentido horário) e 30º (anti-horário).

Gráfico 4.18
Médias e medianas por linhas e colunas:
Proporção de População Alfabetizada e
Logaritmo da Proporção de Analfabetos
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial



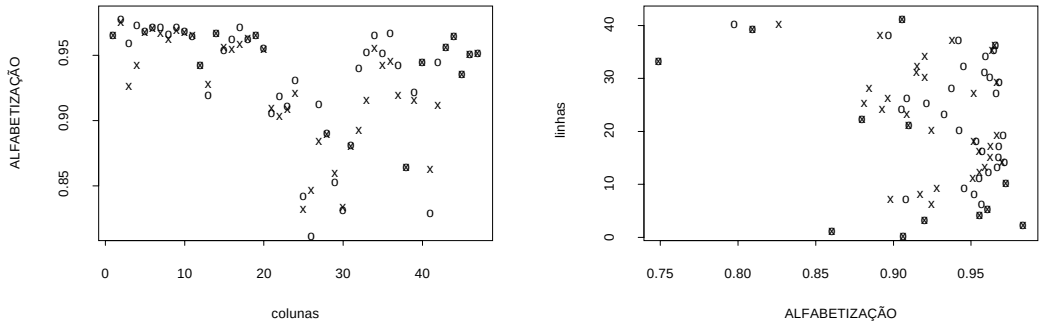
o = Mediana

x = Média

Eixo Norte-Sul

Eixo Leste-Oeste

ALFABETIZAÇÃO



ANALFABETISMO(log)

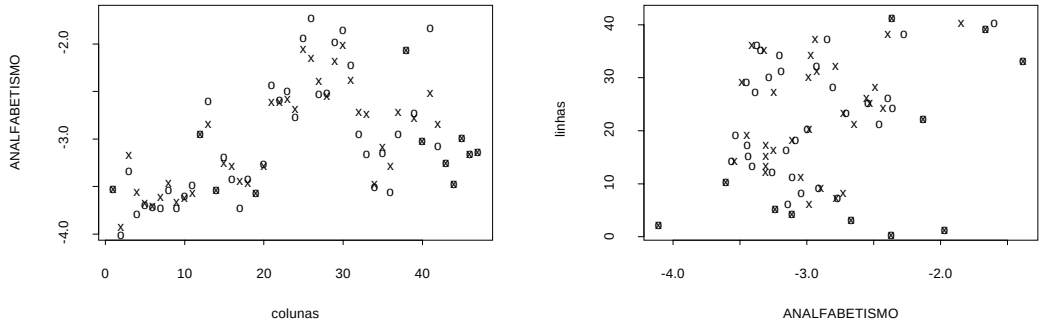
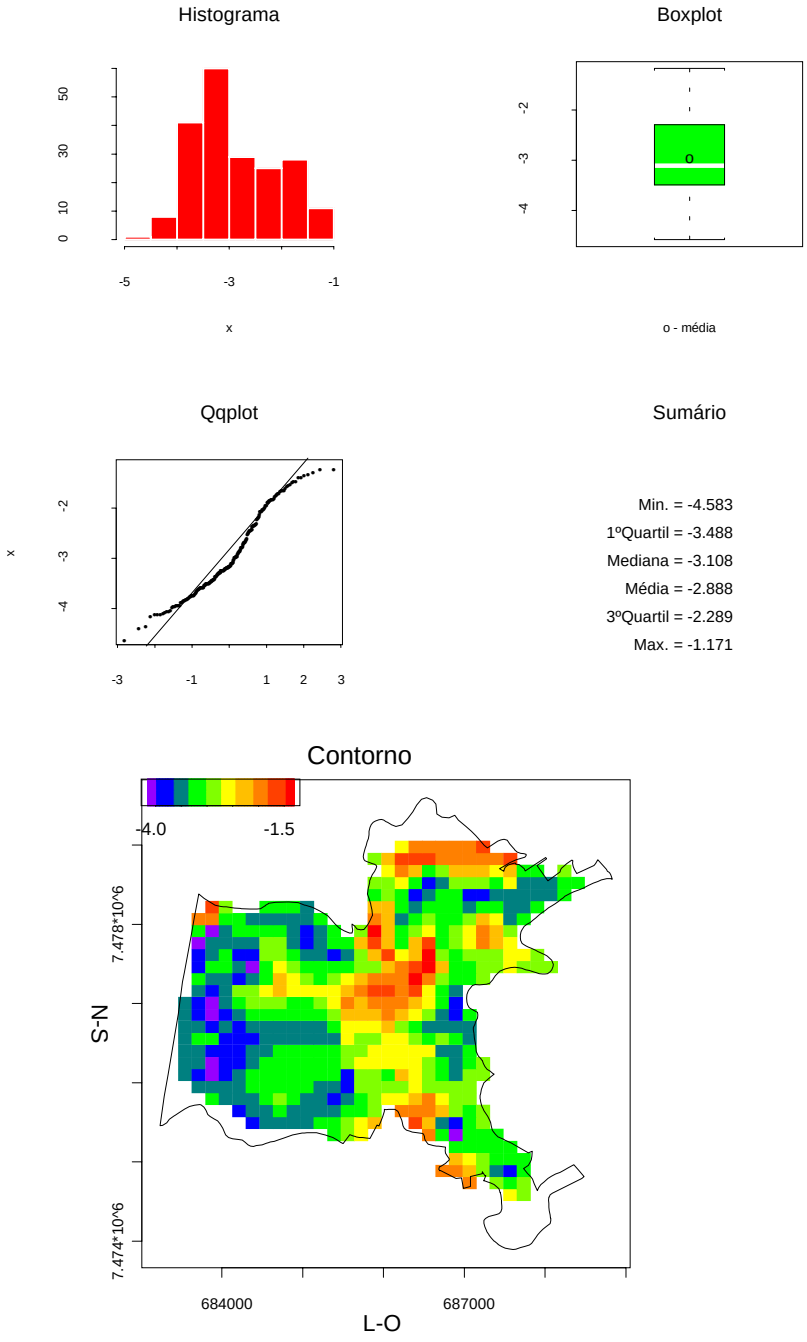


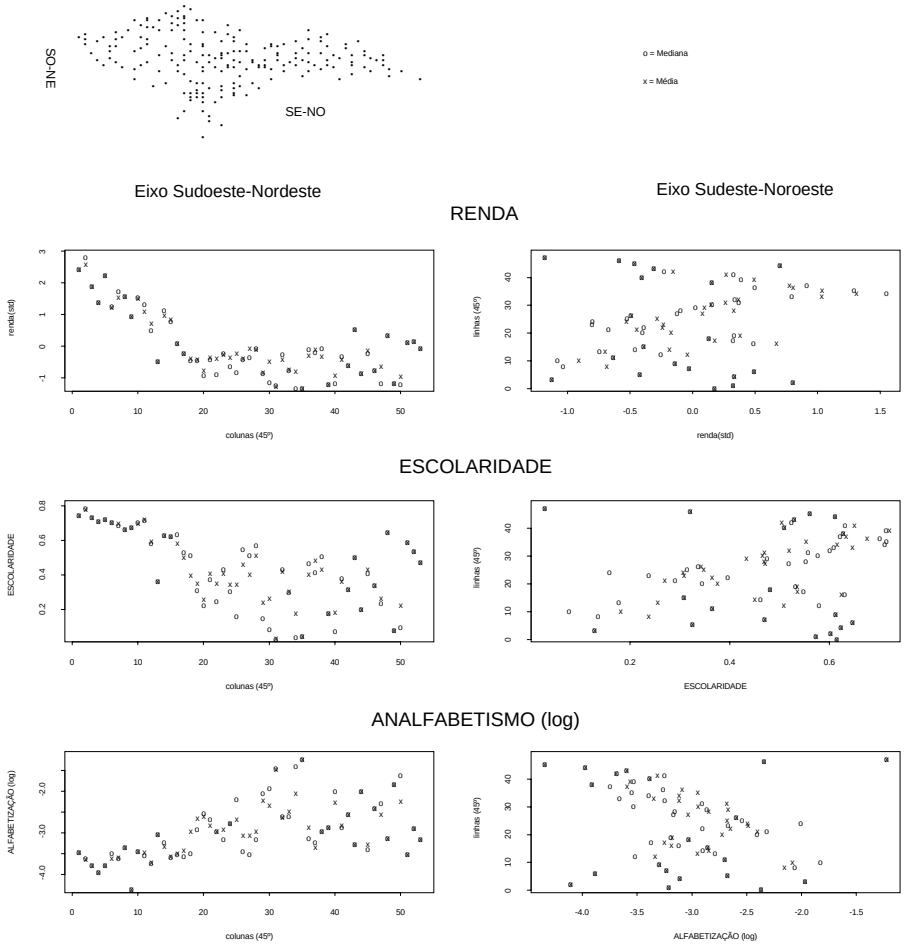
Gráfico 4.19
Análise exploratória do Logaritmo da Proporção de População Analfabeta
Ilha do Governador, Censo Demográfico de 1991



O gráfico de médias e medianas após rotação de 45º (sentido horário), permite visualizar mais claramente do que no gráfico original a existência de tendência da renda média do chefe da família, mais intensa nesta direção (Gráfico 4.20). Entretanto, no gráfico de linhas ainda há uma sugestão de variação em larga escala, com um ponto de inflexão por volta da linha 40. Aparentemente, a tendência existente não é linear ao longo de toda a região, mas apresenta um comportamento como um pico de alta renda com declive em todas as direções. Ou seja, existe umas tendência na distribuição espacial da renda média do chefe da família, principalmente a 45º, direção Noroeste-Sudeste. Esta variação em larga escala não é linear nem uniforme em todo o território.

No mesmo gráfico, o indicador escolaridade, entretanto, não se organizou de forma semelhante: a 45º aumentou a tendência visualizada nas linhas, ainda que de forma pouco estruturada. Ainda assim, a rotação a 45º parece concentrar a tendência na direção das colunas. A rotação de 45º no sentido horário da transformação logarítmica da variável proporção de população analfabeta permitiu verificar a existência de tendência, visível nas linhas, em forma de um trecho ascendente seguido de um descendente, caracterizando uma variação em larga escala. Examinando o comportamento dos indicadores após rotação de 30º (sentido anti-horário), não foi possível visualizar claramente a tendência, confirmando a opção por modelar a tendência observada com rotação de 45º do eixo de coordenadas.

Gráfico 4.20
Médias e medianas por linhas e colunas após rotação de 45°:
Renda, Escolaridade e Analfabetismo(log)
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial



4.5. Análise exploratória espacial: valores atípicos

Um aspecto importante de todos estes gráficos é o grande números de linhas e colunas onde média e mediana estão afastadas, sugerindo a presença de valores atípicos. A localização destes pontos foi feita através dos primeiros intervalos da nuvem do variograma e dos *boxplots* da raiz quadrada das diferenças entres os pares, indicador de contraste proposto por Cressie (1991).

A renda média do chefe da família apresentou alguns pares muito contrastantes, que no gráfico da nuvem do variograma apresentam altos valores de γ em pequenas distâncias inter pares (Gráfico 4.21). Este foram indetificados através do número de registro do par de setores censitários no banco de dados (por exemplo, setores censitários 101 e 102 apresentam alto valor de γ. Os *boxplots*, entretanto, não apresentaram valores extremos, ainda que média e mediana estejam bastante deslocadas no primeiro intervalo. Para a escolaridade (Gráfico 4.22) o número de valores atípicos aumenta, o que se confirma nos *boxplots*, com o aparecimento de 4 pares de valores aberrantes.

Gráfico 4.21
Identificação de pares contrastantes: *Renda*
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial

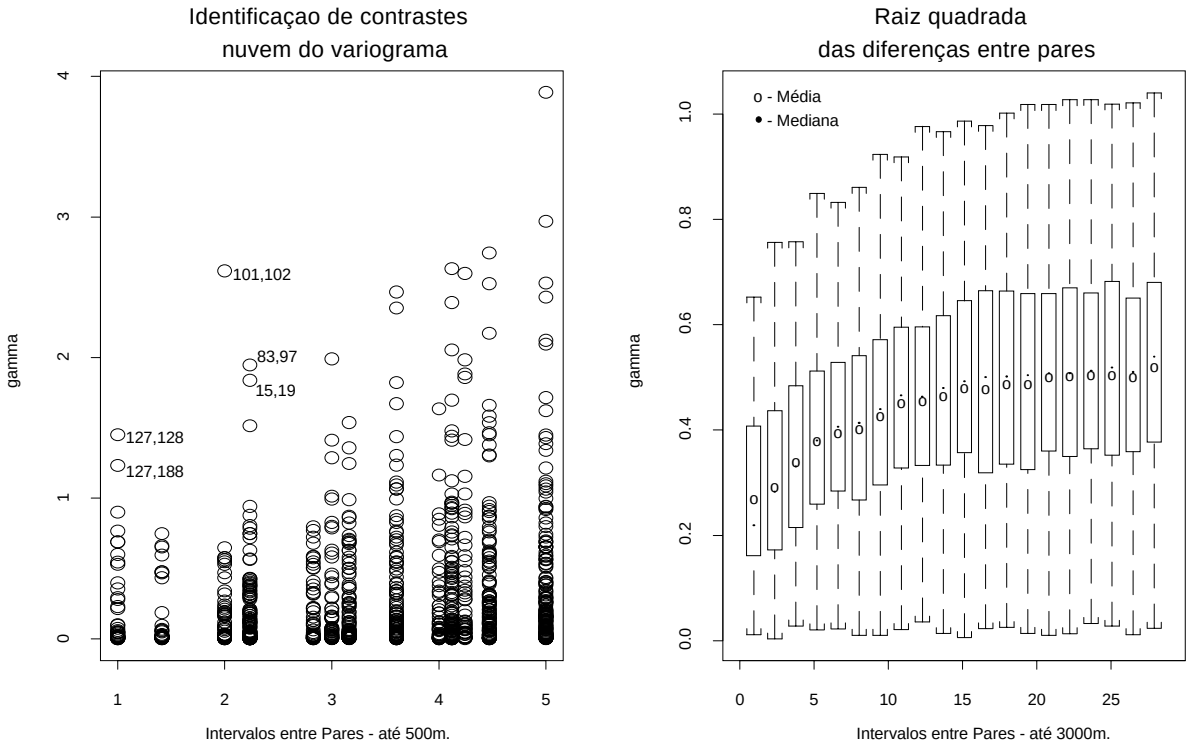
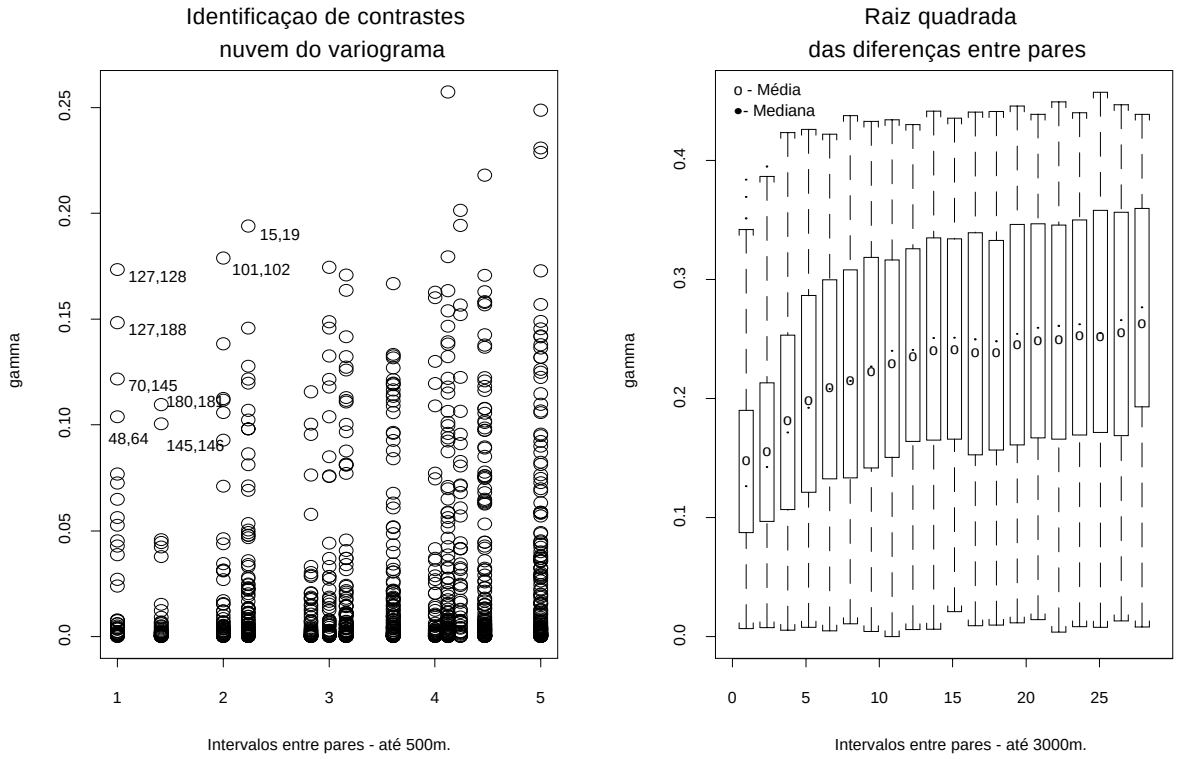


Gráfico 4.22
Identificação de pares contrastantes: *Escolaridade*
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial



O indicador de analfabetismo apresenta número maior de pares contrastantes, particularmente no segundo e terceiro intervalos (Gráfico 4.23). A proporção de casas alugadas, curiosamente, apresenta maior quantidade de pares atípicos, a partir do quarto intervalo, sugerindo um processo muito semelhante em pequenas distâncias e com saltos em distâncias médias (Gráfico 4.24).

Gráfico 4.23
Identificação de pares contrastantes: *Analfabetismo(log)*
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial

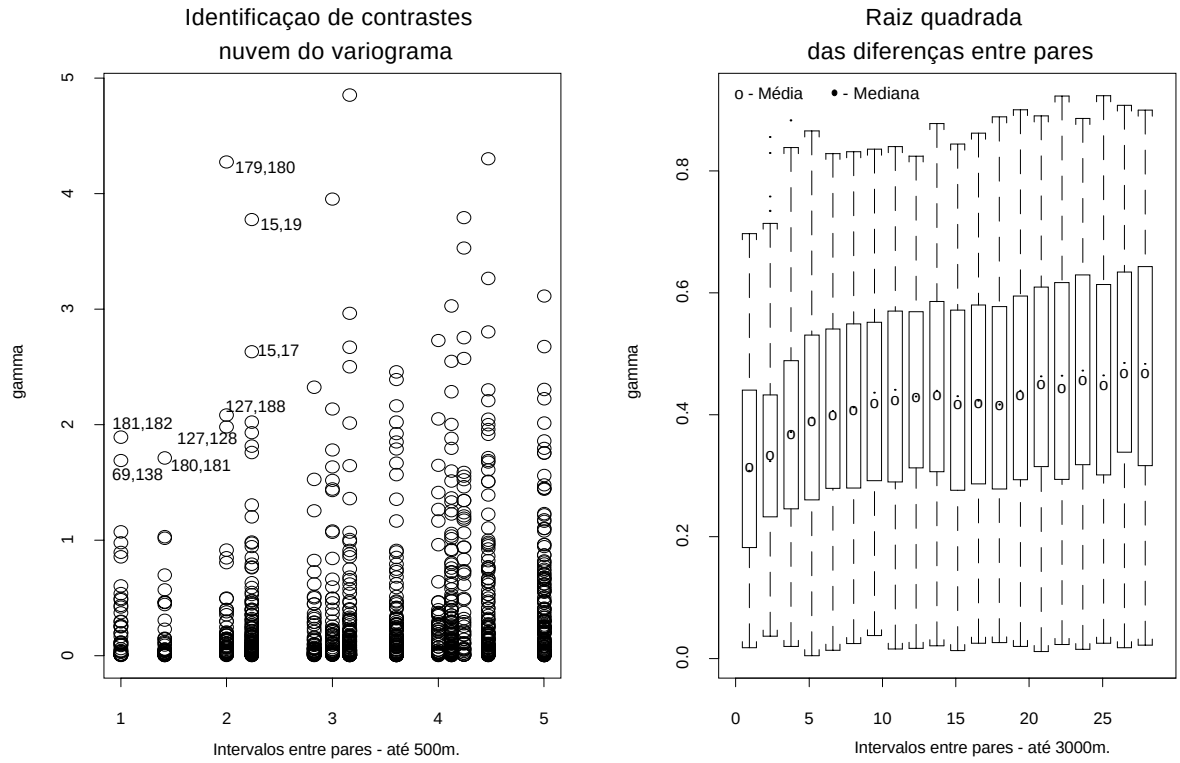
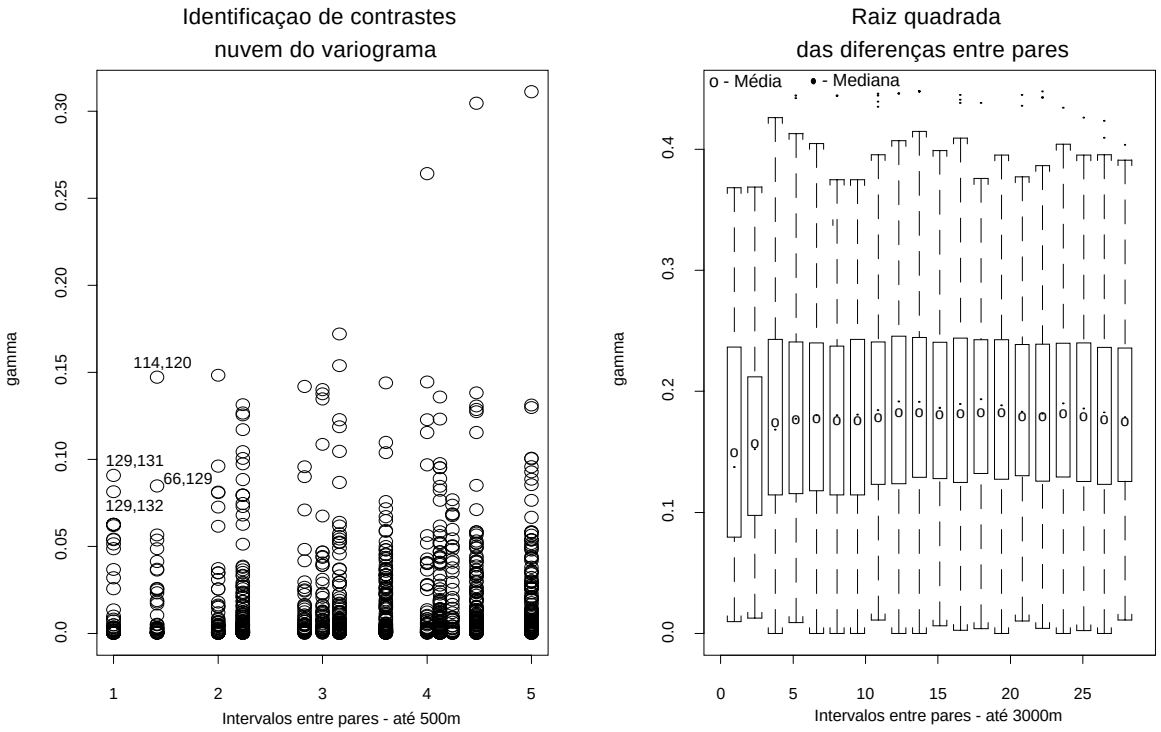


Gráfico 4.24
Identificação de pares contrastantes: *Casas Alugadas*
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial



A proporção de casas ligadas à rede de água e proporção de casas ligadas à rede de esgotos (Gráficos 4.25 e 4.26) apresentam número excessivo de valores atípicos e *boxplots* inteiramente deslocados para os valores mais baixos, com diversos pontos aberrantes. Estes indicadores apresentam grande contraste em áreas vizinhas.

Gráfico 4.25
Identificação de pares contrastantes: *Água*
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial

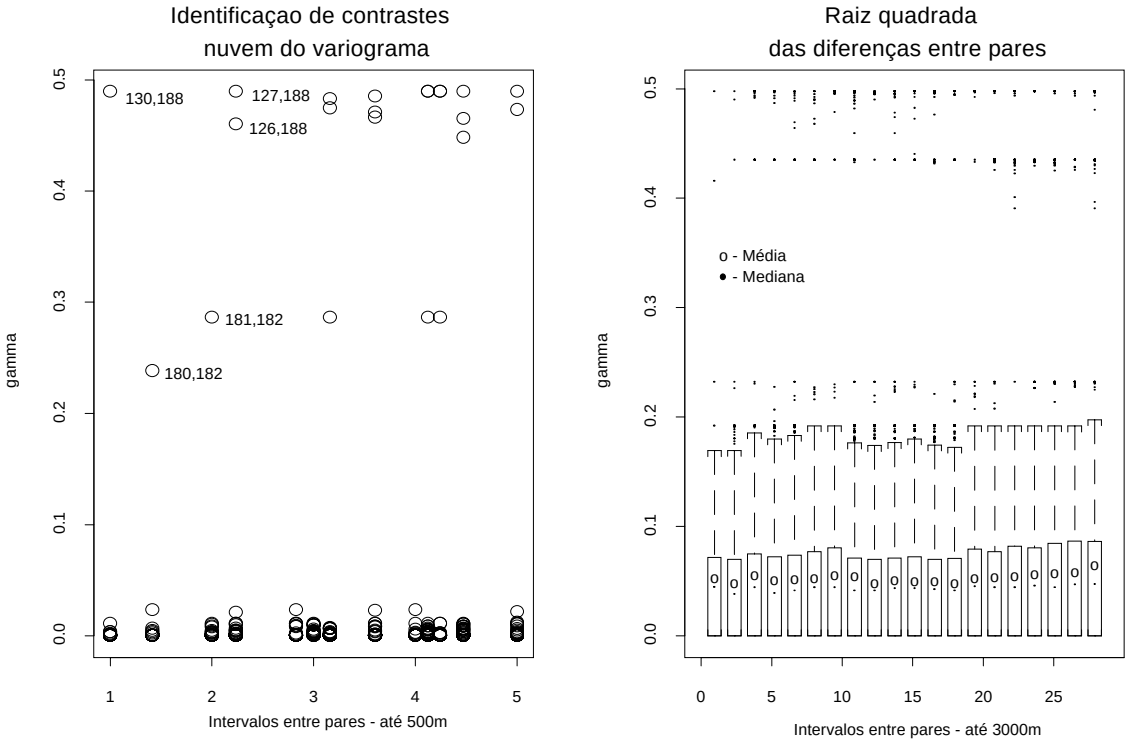
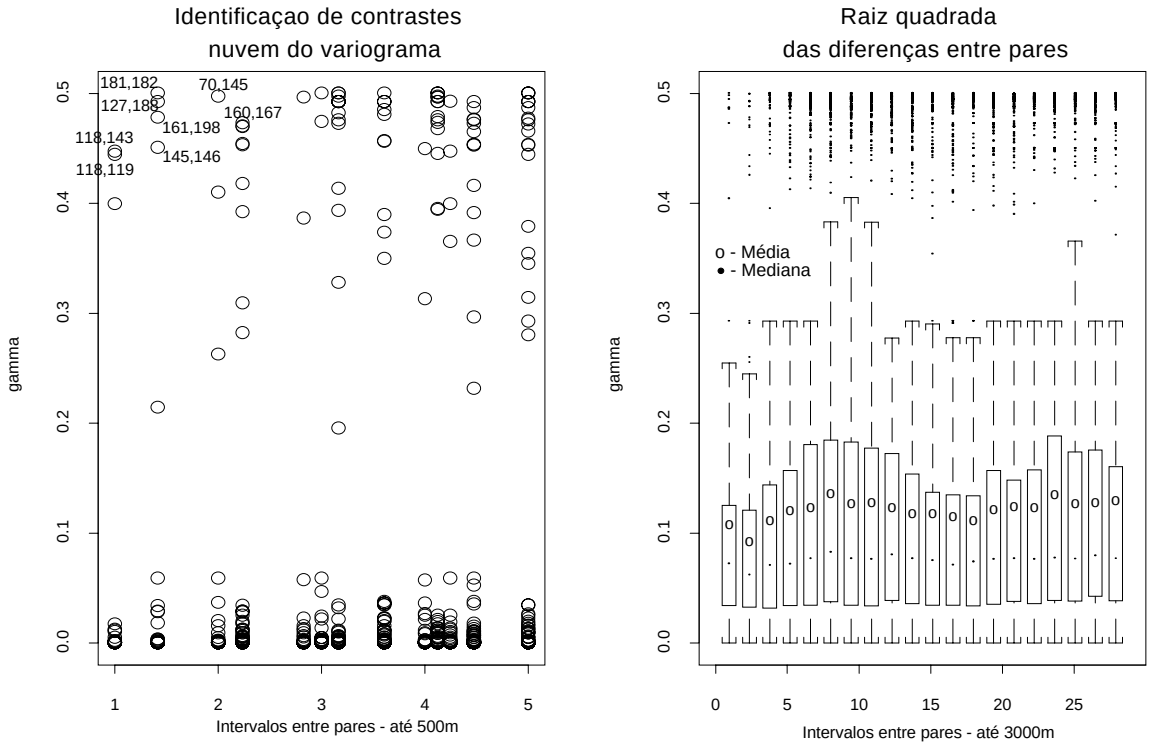
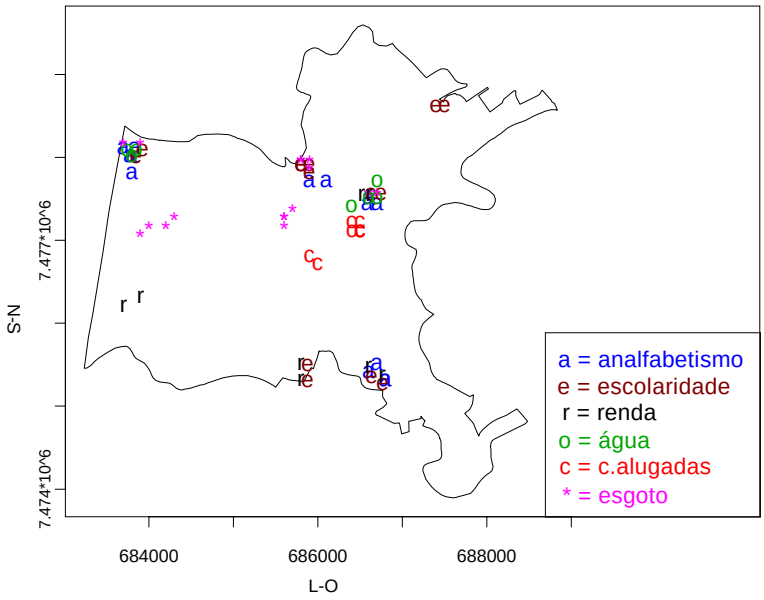


Gráfico 4.26
Identificação de pares contrastantes: *Esgoto*
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial



A localização espacial destes pontos foi feita no Gráfico 4.27, onde áreas vizinhas, com grandes contrastes nos valores das variáveis, são identificadas. Algumas pequenas áreas concentram diversas variáveis com “saltos” no valores, especialmente na região central do mapa. Os contrastes relacionados a renda, escolaridade e analfabetismo se superpõem em diversos locais. A ligação à rede de esgotos destaca-se dos demais indicadores quanto à região onde locais vizinhos apresentam diferenças na variável. Os contrastes identificados não podem ser considerados exceções, mas, ao contrário, são peculiaridades da região estudada, e caracterizam a inexistência de estacionariedade em alguns locais.

Gráfico 4.27
Identificação dos vizinhos com altos contrastes para cada indicador
Ilha do Governador, coordenadas regularizadas, excluída área sem continuidade espacial



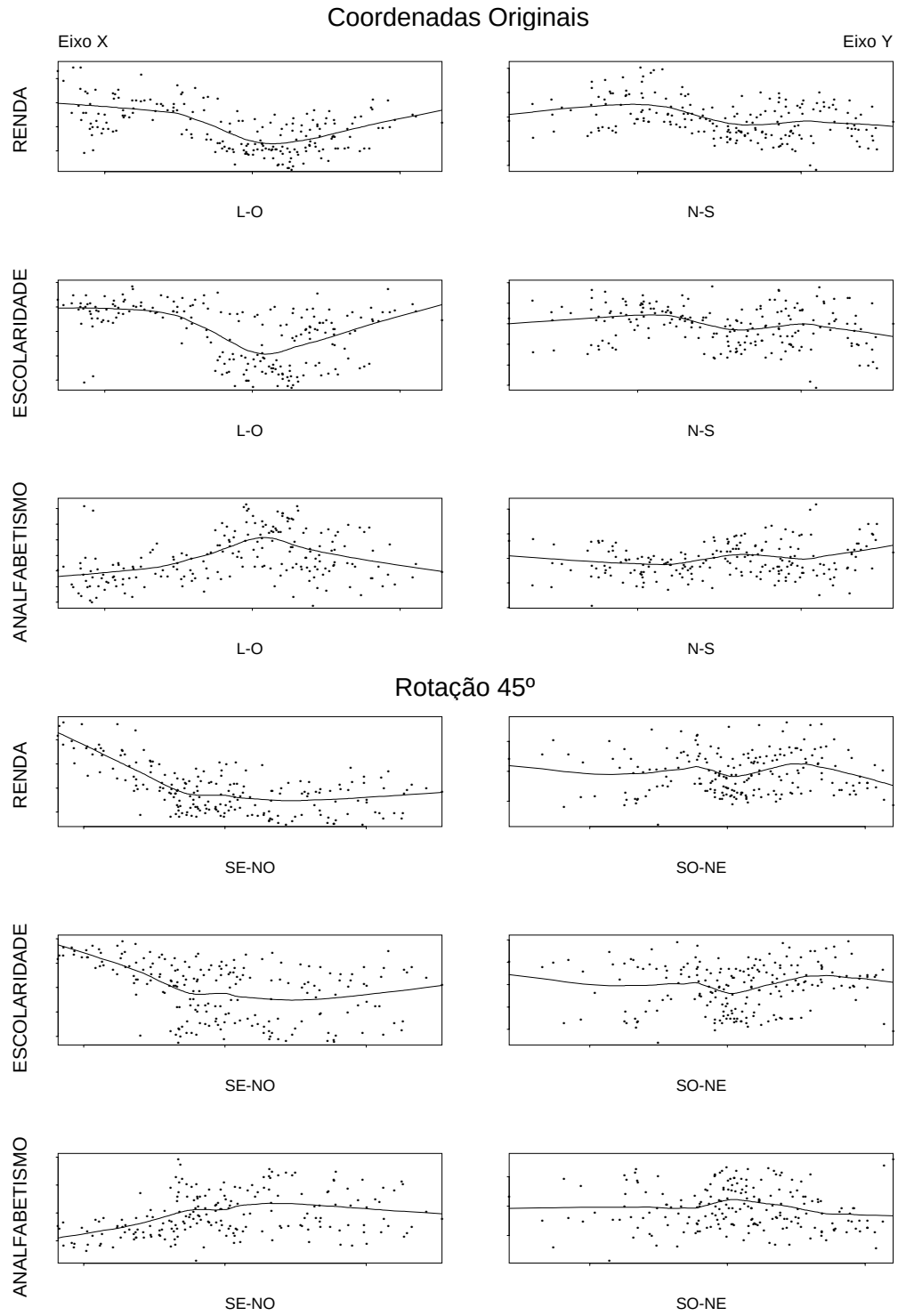
4.6. Modelagem da tendência

Baseado nas observações da análise exploratória de dados utilizando gráficos de médias e medianas de linhas e colunas, buscou-se modelar a tendência dos indicadores de renda, escolaridade, analfabetismo e propriedade da residência, ajustando-a através de *loess* - regressão linear local ponderada pela distância - de forma a permitir sua retirada e o estudo do variograma. Inicialmente se explorou o modelo obtido, variando a dimensão da janela local, buscando captar a tendência - variação em larga escala - sem eliminar a variação em pequena escala, cuja modelagem será feita através da

Krigeagem. As duas dimensões - latitude e longitude - foram analisadas isoladamente, sem considerar a interação entre elas, da mesma forma que na análise exploratória através dos gráficos de médias e medianas.

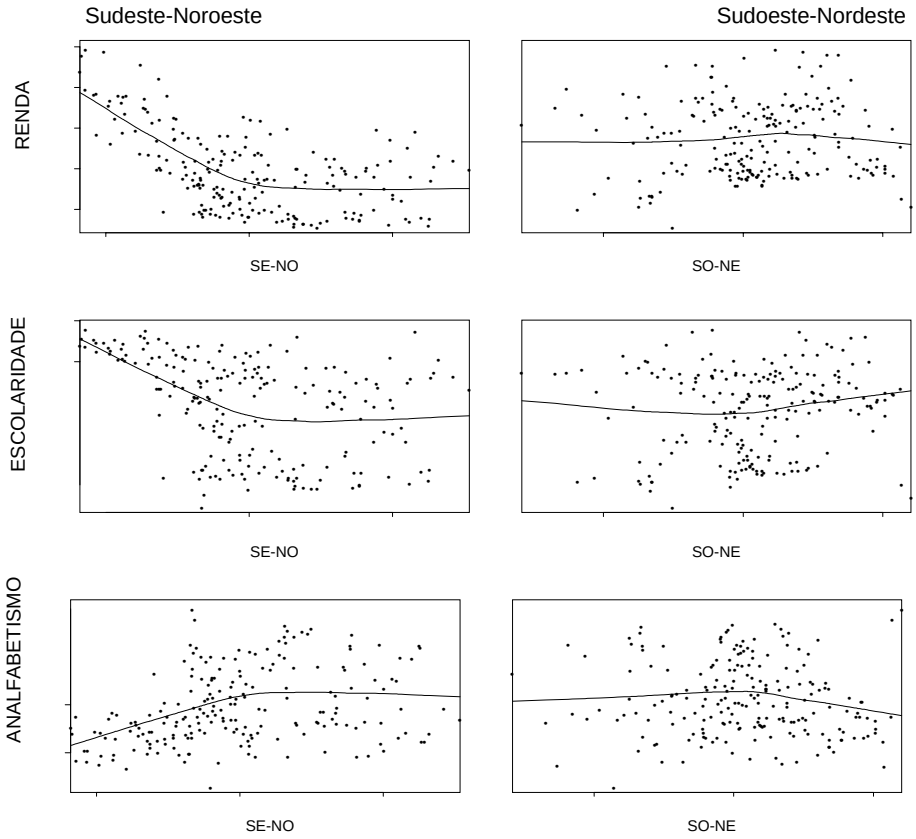
No Gráfico 4.28 pode-se analisar as curvas ajustadas aos indicadores de renda, escolaridade e analfabetismo, através do *loess*, com janela de 2/3 dos pontos. O ajuste foi feito para cada coordenada isoladamente, ficando o eixo X sempre à esquerda e o Y à direita do gráfico. Os pontos representam os resíduos, e o ajuste foi feito considerando os eixos originais (os três primeiros quadros) e após rotação de 45º (os três quadros seguintes). O comportamento das variáveis é muito parecido, ficando a tendência mais evidente ao longo do eixo das ordenadas (X) pós-rotação, no equivalente à direção Sudoeste-Nordeste. O analfabetismo também apresenta tendência neste eixo, porém em sentido inverso.

Gráfico 4.28
Detecção de tendência: ajuste por *loess* e resíduos (janela =2/3)
Renda, Escolaridade e Analfabetismo(log)
Ilha do Governador, excluída área sem continuidade espacial



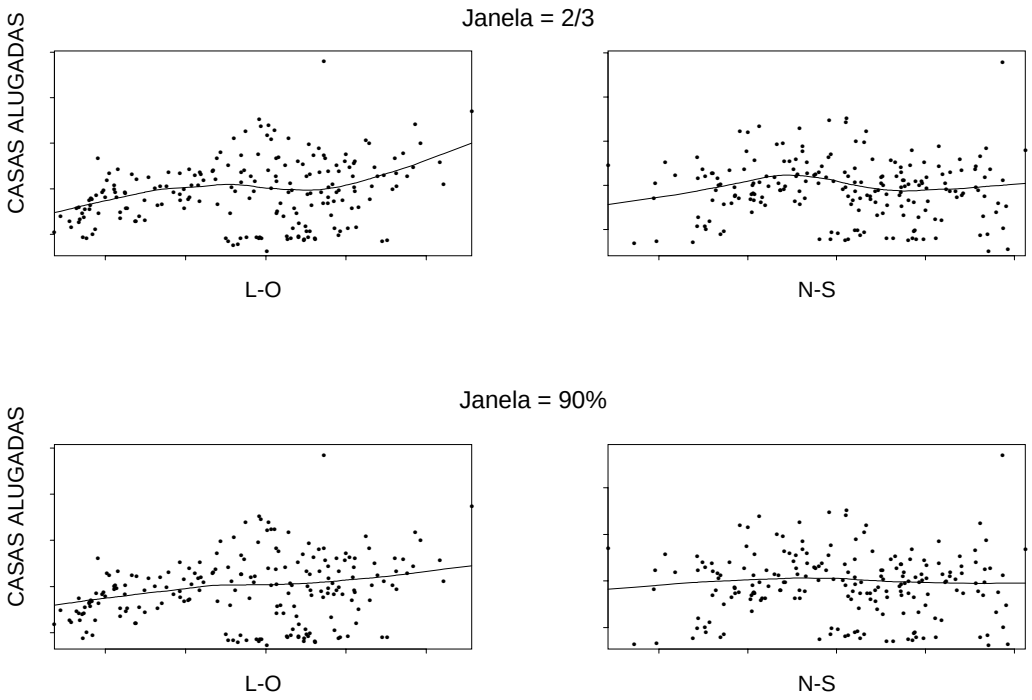
Ampliando a janela do *loess* para 90% dos pontos (Gráfico 4.29), pode-se observar que nas abcissas diminui bastante a tendência, que fica portanto mais concentrada no eixo vertical - direção Sudoeste-Nordeste. A distribuição dos resíduos no extremo esquerdo dos gráficos do eixo Sudoeste-Nordeste concentra-se um pouco mais em um dos lados, sugerindo que o modelo neste caso não captou inteiramente a tendência existente. Ainda assim, optou-se por modelar a tendência através de regressão local ponderada - o *loess*, utilizando janela de 90% da área - deixando a estrutura restante para modelar através do variograma.

Gráfico 4.29
Ajuste de tendência por *loess* com janela = 90% e resíduos
Renda, Escolaridade e Analfabetismo(log)
Ilha do Governador, excluída área sem continuidade espacial



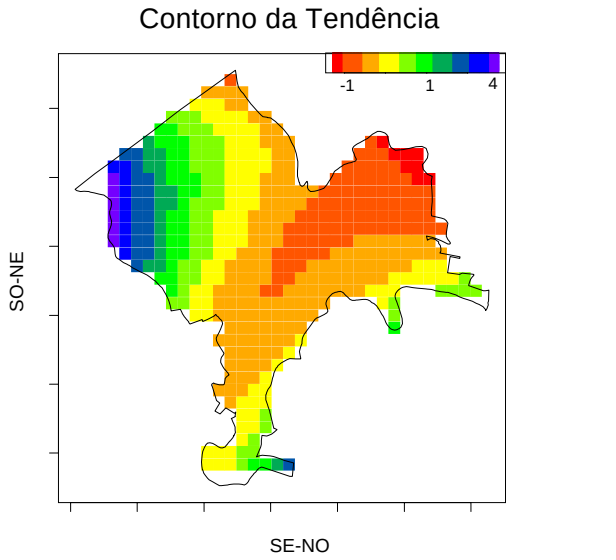
A análise do ajuste por regressão linear local ponderada da proporção de casas alugadas manteve a impressão inicial do gráfico de médias e medianas, onde a direção principal da tendência era Leste-Oeste (Gráfico 4.30). Além disso, os resíduos do *loess* ficaram distribuídos igualmente em ambos os gráficos, com janela de 2/3 e de 90%, optando-se, então, por estudar a dependência espacial sobre os resíduos do modelo com a maior janela.

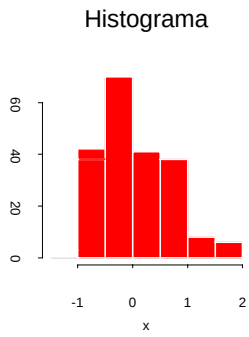
Gráfico 4.30
Ajuste de tendência por *loess* com janela = 2/3 e a 90% e resíduos
Proporção de Casas Alugadas
Ilha do Governador, excluída área sem continuidade espacial

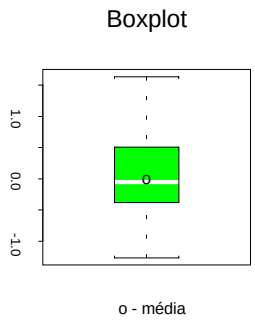


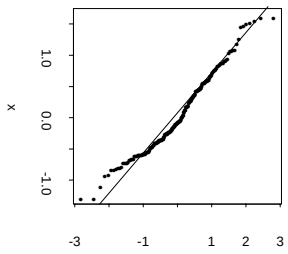
Os Gráficos 4.31 a 4.34 apresentam os mapas de contorno de tendência e a análise exploratória dos resíduos do *loess* de cada variável.

Gráfico 4.31
Mapa de contorno da tendência e análise exploratória dos resíduos do *loess*
Renda Média do Chefe da Família
Ilha do Governador, excluída área sem continuidade espacial





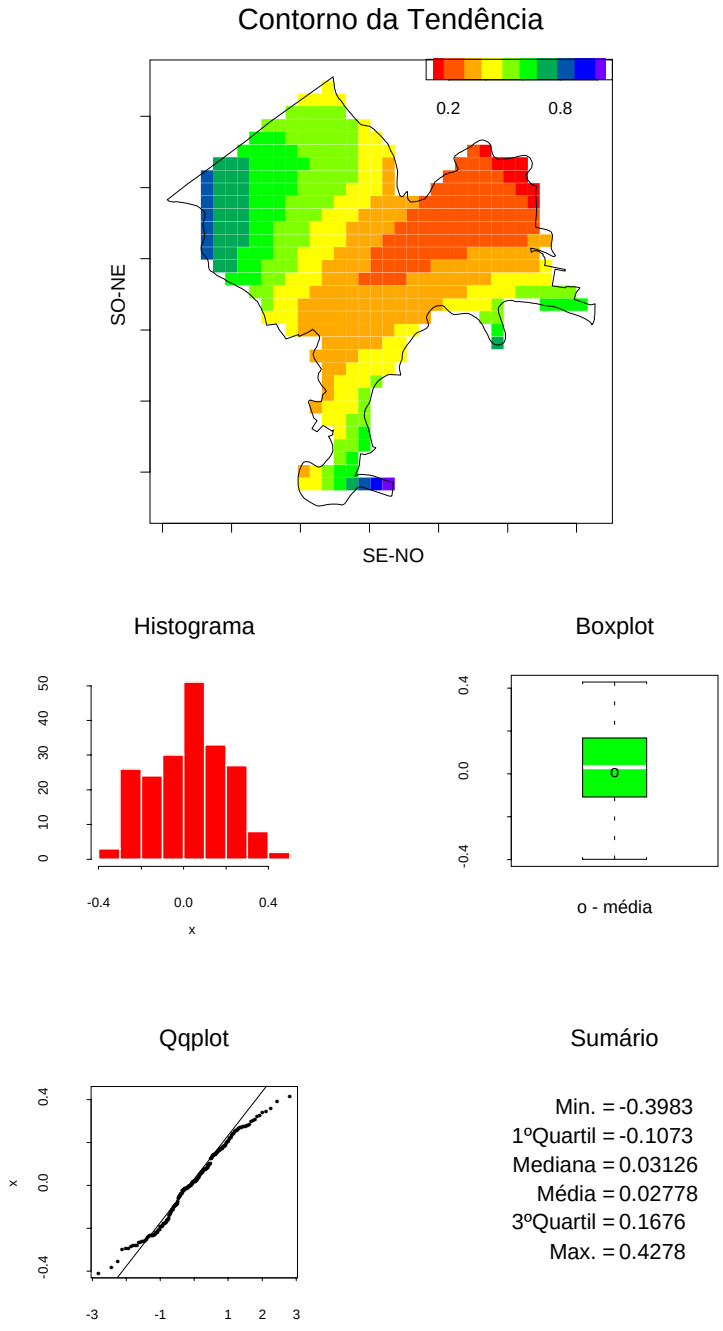




Sumário

Min. =	-1.265
1ºQuartil =	-0.3749
Mediana =	-0.0497
Média =	0.07483
3ºQuartil =	0.5063
Max. =	1.634

Gráfico 4.32
Mapa de contorno da tendência e análise exploratória dos resíduos do *loess*
Proporção de Chefes de Família com 2° Grau
Ilha do Governador, excluída área sem continuidade espacial



A tendência dos indicadores de renda, escolaridade e analfabetismo tem aparência similar nos mapas de contorno, embora não seja coincidente o tamanho da área onde os indicadores exprimem melhores condições de vida. A área da população com renda mais alta é menor do que a área ocupada por setores censitários com maior proporção de escolaridade acima do segundo grau, que por sua vez é maior do que a área onde praticamente inexiste analfabetismo (Gráficos 4.31 a 4.33).

O resíduo da renda média do chefe da família, apresenta distribuição um pouco assimétrica, com desvio à esquerda, poucos valores atípicos, média e medianas próximas (Gráfico 4.31). A proporção de chefes de família com escolaridade igual ou superior ao segundo grau tem distribuição mais próxima da normal (Gráfico 4.32). Os resíduos do logaritmo da proporção de analfabetos apresentam perfil bastante próximo de uma distribuição normal, com valores atípicos no extremo superior da distribuição (Gráfico 4.33).

A superfície da proporção de casas alugadas é diferente das outras em relação à direção da tendência e intensidade da variação, que é menor. Os resíduos apresentam alguns valores atípicos no extremo superior, e no trecho em torno de -0,3 se afastam da normalidade (Gráfico 4.34).

Gráfico 4.33
Mapa de contorno da tendência e análise exploratória dos resíduos do *loess*
Proporção de População Analfabeta (log)
Ilha do Governador, excluída área sem continuidade espacial

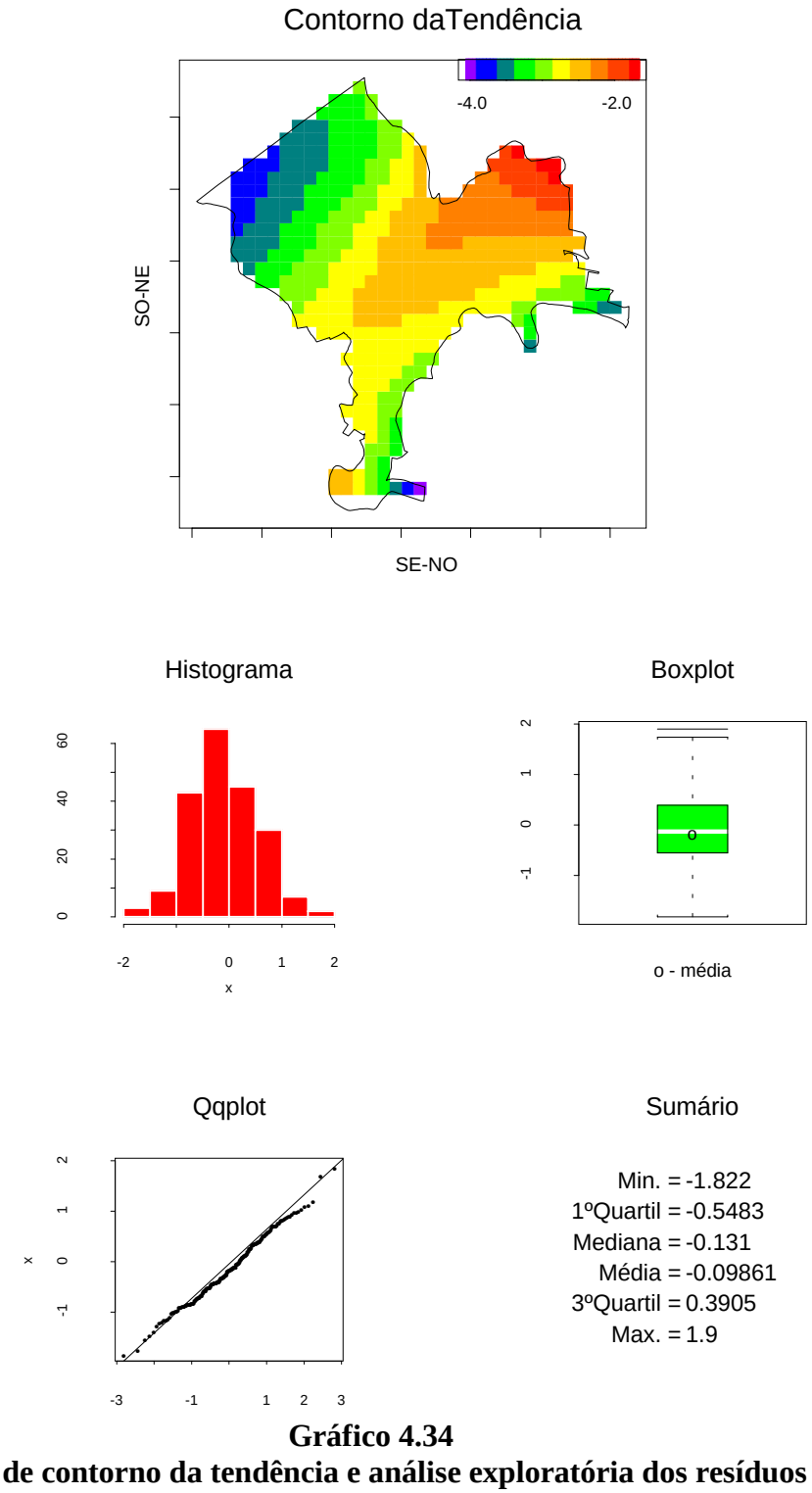
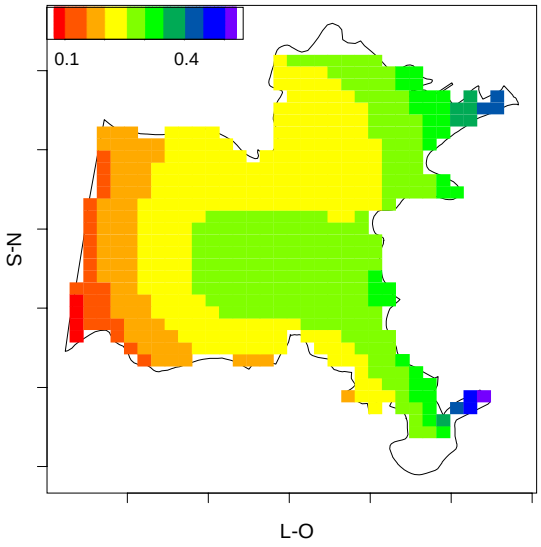
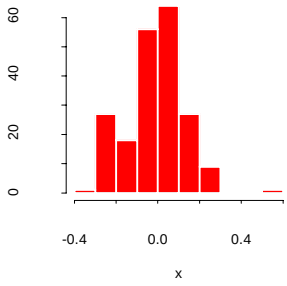


Gráfico 4.34
Mapa de contorno da tendência e análise exploratória dos resíduos do *loess*
Proporção de Casas Alugadas
Ilha do Governador, excluída área sem continuidade espacial

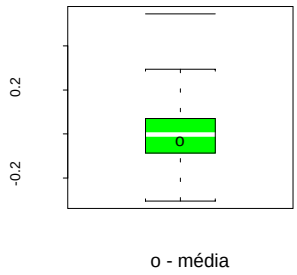
Contorno daTendência



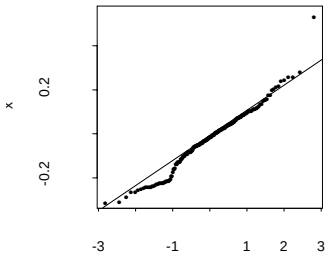
Histograma



Boxplot



Qqplot



Sumário

Min. = -0.3044
1ºQuartil = -0.08454
Mediana = -0.002202
Média = -0.01071
3ºQuartil = 0.06943
Max. = 0.5459

A autocorrelação dos resíduos do *loess* foi testada através do teste *I* de Moran: houve diminuição para os indicadores de renda, escolaridade e analfabetismo (Tabela 4.3). A situação do indicador referente à propriedade do imóvel manteve-se semelhante.

Tabela 4.3
Indicadores após eliminação da tendência
segundo estatísticas de autocorrelação espacial

Indicador	Correlação	*"I"*	p (bicaudal)	p (permut.)
Renda	0,3844	7,081	1,431e-12	0
Escolaridade	0,5104	9,372	7,084e-21	0
Log Analfabetos	0,4443	8,172	3,035e-16	0
Casas Alugadas	0,3213	5,934	2,958e-9	0

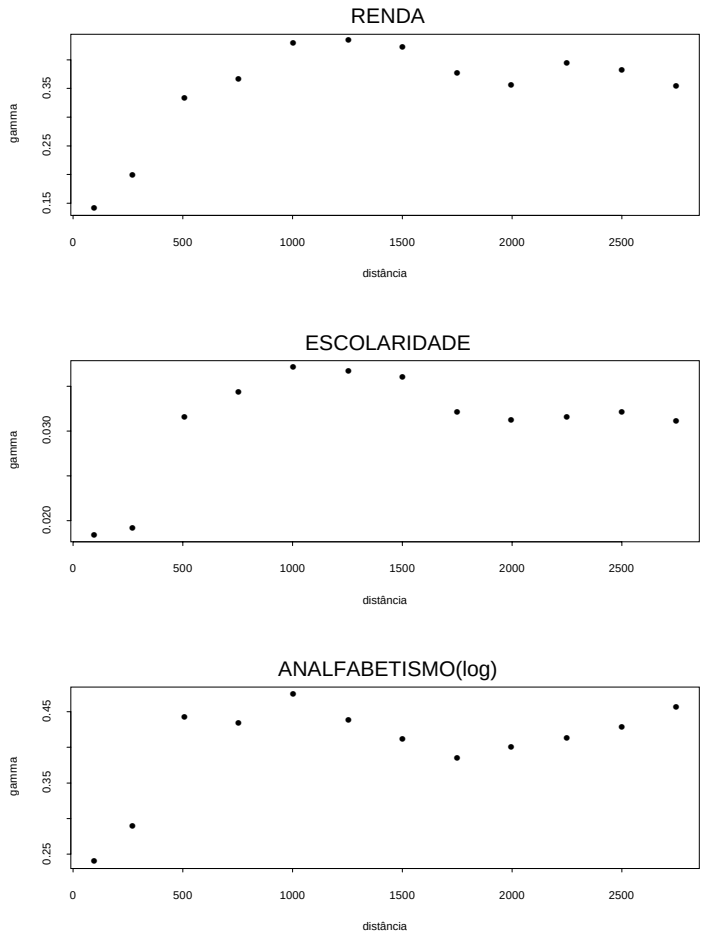
Refeitos os gráficos de médias e medianas por linhas e colunas observou-se que para as quatro variáveis a tendência havia sido eliminada. Modelos *loess* com janelas menores foram experimentados, e quando testados a autocorrelação havia desaparecido inteiramente, uma vez que a regressão linear local ponderada, nestes casos, tinha incorporado a variação em pequena escala.

4.7. Variogramas amostrais

Os variogramas amostrais omnidirecionais calculados para cada uma das variáveis apresentaram 38 pares no primeiro intervalo, o que permite razoável estabilidade. Os variogramas omnidirecionais utilizando o método robusto (Cressie,1991), também calculados não serão apresentados por apresentarem forma muito semelhante ao anterior.

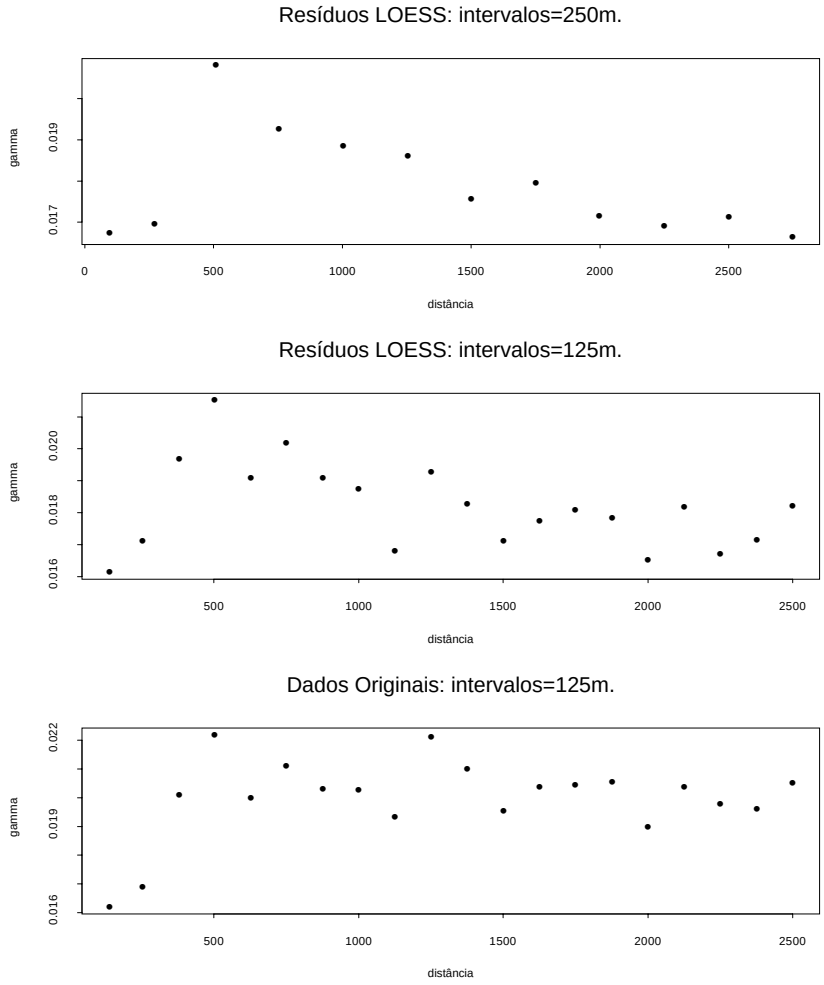
O variograma amostral omnidirecional dos resíduos do *loess* dos indicadores de renda e escolaridade do chefe da família tem forma semelhante e patamar máximo atingido a uma distância de aproximadamente 1.000 m. A escolaridade apresenta, no início (distâncias menores do que 250 m), comportamento diferenciado, com ascensão não linear. O logaritmo da proporção de analfabetos tem comportamento oscilante entre distâncias de 500 e 1.250 m. As oscilações apresentadas após 1.500 m nestas três variáveis sugerem ciclicidade (Gráfico 4.35).

Gráfico 4.35
Variograma amostral dos resíduos do *loess*
Renda, Escolaridade e Analfabetismo (log)
Ilha do Governador, excluída área sem continuidade espacial



Os resíduos da modelagem da tendência para o indicador de aluguel residencial apresentam variograma plano (Gráfico 4.36). É interessante observar os variogramas com intervalo de 125 m, antes e após retirar a tendência. Em ambos, o trecho ascendente (*range*), que corresponde à presença de dependência espacial, situa-se abaixo de 250 m, e o trecho de estabilização (*sill*), relacionado à variância total do indicador, apresenta declive onde foi retirada a tendência. Neste caso, pode-se supor que não há variação espacial modelável através de variograma. O resultado significativo do teste de Moran deve-se possivelmente apenas à tendência espacial observada. Assim, este indicador não será modelado através deste método, mas apenas através do *loess*, onde, diminuindo-se a janela para 20% do total de pontos, é possível modelar a tendência observada, ao mesmo tempo em que variações em escalas pequenas também são consideradas.

Gráfico 4.36
Variograma amostral em diferentes intervalos, dados originais e resíduos,
Proporção de Casas Alugadas
Ilha do Governador, excluída área sem continuidade espacial



Os variogramas direcionais foram analisados visando estudar a anisotropia. Um aspecto importante é o número de pontos em cada intervalo de distância entre pares, que podem gerar grande oscilação nos valores encontrados (Tabela 4.4). Assim, o primeiro intervalo, que apresenta no máximo 12 pares (a 45º), não será considerado na análise direcional, sendo excluído do variograma.

Tabela 4.4
Número de pares em cada trecho do variograma, segundo quatro direções (0º, 45º, 90º, 135º)

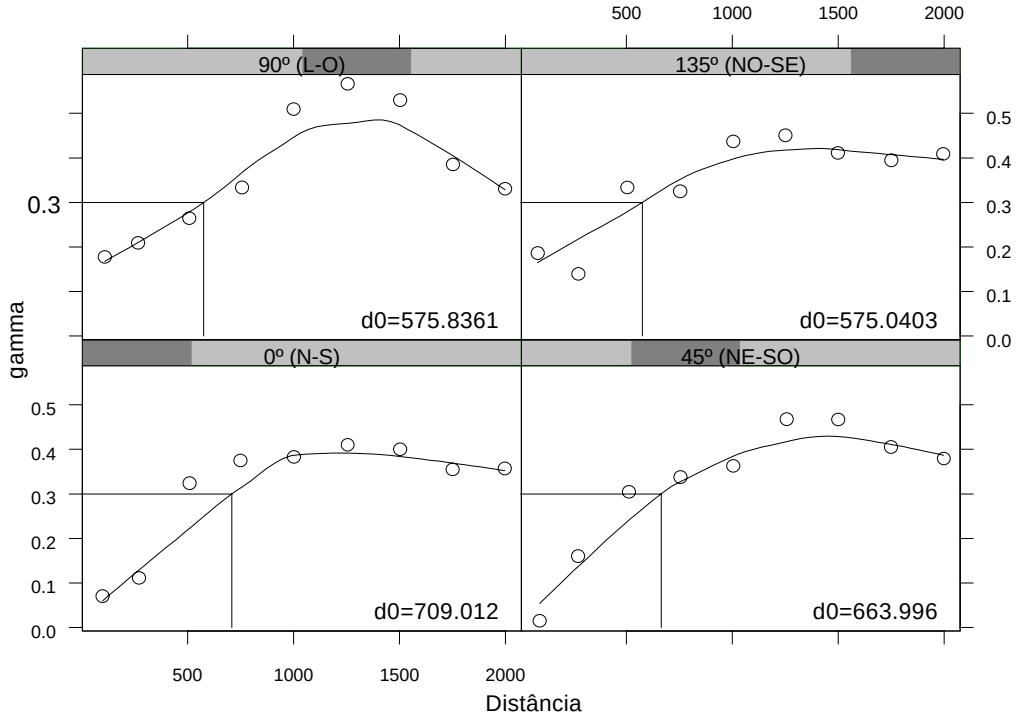
Intervalo	0º	45º	90º	135º
1	9	12	10	7
2	152	134	150	144
3	253	255	258	246
4	328	339	317	328
5	381	375	372	363
6	422	442	410	405
7	429	469	457	399
8	431	477	446	399
9	391	504	502	346
10	364	491	495	321
11	294	485	483	287
12	249	418	437	252
Média	**308,58**	**366,75**	**361,42**	**291,42**

Nos Gráficos 4.37 a 4.39 são apresentados os variogramas direcionais dos resíduos do *loess*, estimados pelo método robusto, onde indica-se na barra superior do gráfico a direção e, através da diferença de cores, o intervalo de tolerância

considerado. Na investigação da anisotropia foi estimada a distância onde um dado valor do variograma direcional, cerca de 70% do patamar, foi atingido, registrado no gráfico através do indicador **d0**.

O indicador de renda do chefe da família (Gráfico 4.37) apresenta variogramas semelhantes a 45º e 135º, quanto ao patamar atingido e ao alcance, embora o variograma a 135º apresente maior oscilação no trecho inicial. O variograma a 0º parece ter o patamar máximo um pouco mais baixo e a curva a 90º apresenta três pontos - entre 1.000 e 2.000 m aproximadamente - com valores muito acima dos demais, e queda bem visível imediatamente depois. Estes pontos com valores muito altos no variograma podem estar destacados graças ao padrão espacial descontínuo da variável, que apresenta alguns “saltos” em distâncias relativamente próximas. Os intervalos menores, estudados através da nuvem do variograma, apresentam este tipo de situação, característica de não estacionariedade local. Entretanto, este perfil de contrastes é típico da região, não sendo possível remover estes pontos. O modelo de variograma será ajustado considerando o patamar médio obtido no variograma omnidirecional. A distância estimada entre os pares onde valor calculado do variograma (γ) atingiu 0,3 variou entre 693 m, a 135º, e 578 m, a 90 - diferença relativa menor do que 20% - indicando pequena anisotropia geométrica.

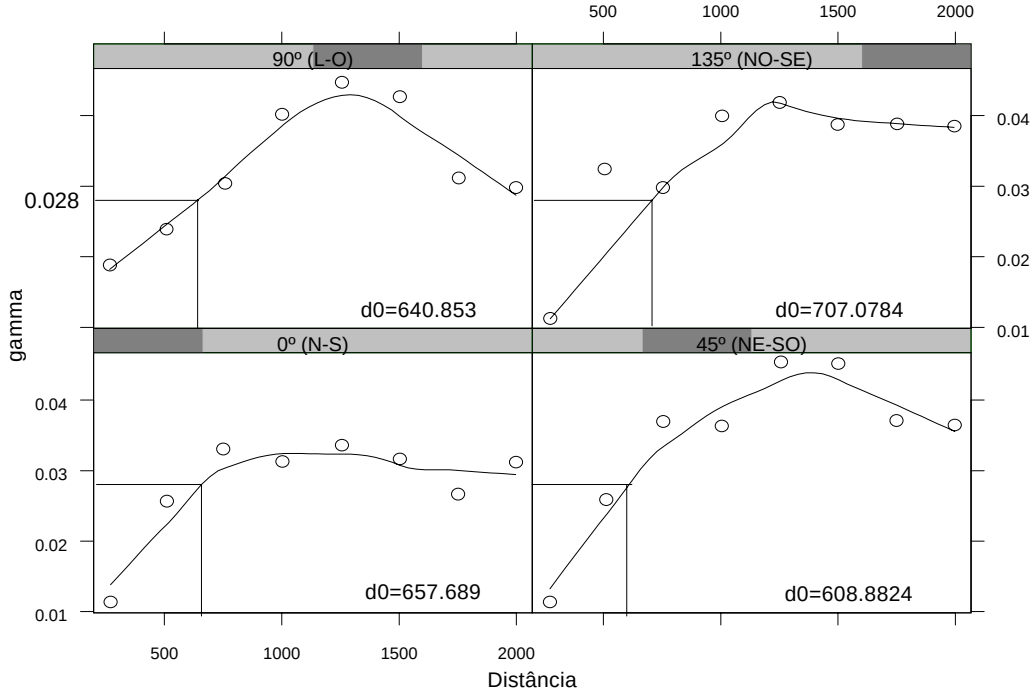
Gráfico 4.37
Variograma direcional e medida de anisotropia
Renda Média do Chefe da Família
Ilha do Governador, excluída área sem continuidade espacial



A proporção de chefes de família com escolaridade igual ou superior ao segundo grau (Gráfico 4.38) apresentou variogramas direcionais dos resíduos do *loess* semelhantes a 45º e 135º. A 90º, a curva se assemelha ao variograma da renda média do chefe da família na mesma direção, com valores altos entre 1.000 e 1.500 m, e queda posterior. O variograma robusto a 0º apresenta patamar entre 750 e 1.500 m, voltando a subir posteriormente. Este tipo de curva, sugestivo de permanência da tendência por se localizar a distâncias maiores e após um patamar de estabilização, pode ser excluído simplesmente manipulando-se a distância máxima a ser considerada, definida a priori para 2.000 m. O patamar de estabilização a 90º parece encontrar-se em valores inferiores ao máximo atingido (a uma distância de 1.200 m aproximadamente). A

distância estimada entre os pares onde o valor calculado do variograma atingiu 0,028 variou entre 713 m, a 135º, e 601 m a 45º, novamente menor do que 20% de variação relativa. O formato das curvas é muito semelhante ao da variável anterior.

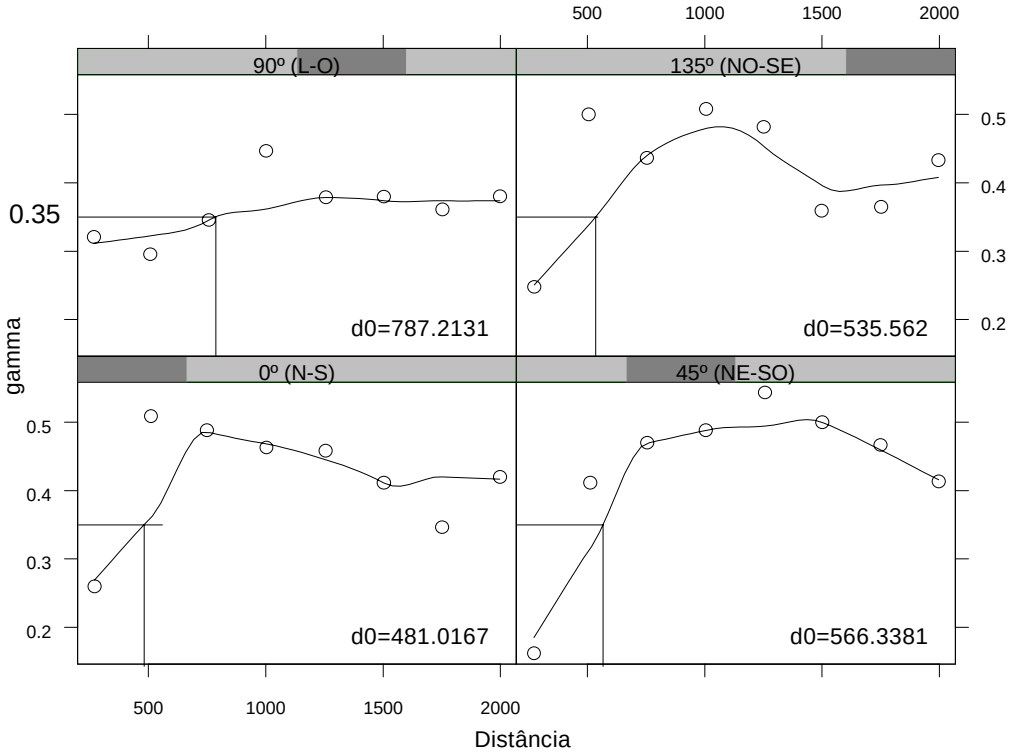
Gráfico 4.38
Variograma direcional e medida de anisotropia
Proporção de Chefes da Família com 2° Grau
Ilha do Governador, excluída área sem continuidade espacial



Os variogramas direcionais dos resíduos do logaritmo da proporção de analfabetos são bem diferentes conforme a direção: a 45º apresenta-se mais regular, com pequena queda após 1.500 m; a 0º, também apresenta queda, e torna a aumentar após os 1.800 m; a 135º sobe rapidamente e flutua em torno do patamar atingido, e a 90º não apresenta autocorrelação muito visível, exceto pelo trecho inicial (Gráfico 4.39). As distâncias para γ de 0,35 são diferentes conforme a direção variando entre 456 m a 0º e

61 m a 90º (variação relativa de 70%), sendo indicado neste caso a correção da anisotropia geométrica.

Gráfico 4.39
Variograma direcional e medida de anisotropia
Proporção de População Analfabeta (log)
Ilha do Governador, excluída área sem continuidade espacial

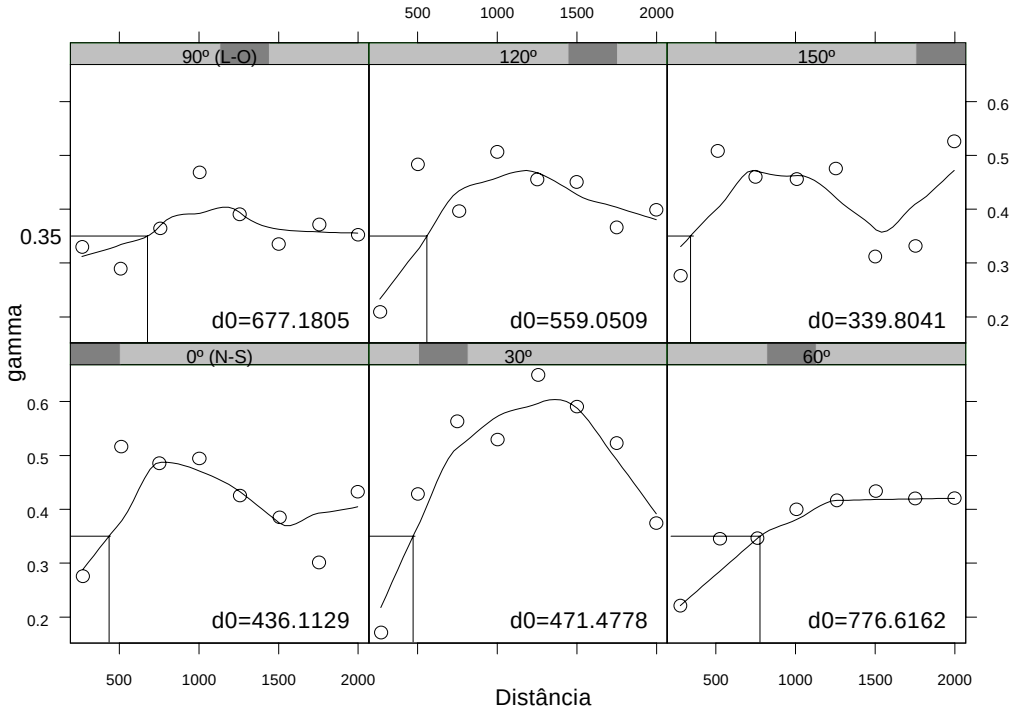


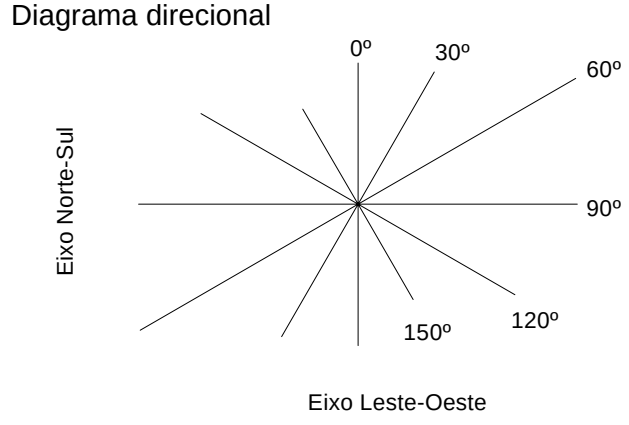
4.8. Modelagem do variograma

Visando definir mais precisamente o eixos de anisotropia para o indicador logaritmo da proporção de analfabetos, foram calculados variogramas em 6 direções: 0º, 30º, 60º, 90º, 120º e 150º, com o mesmo intervalo, tolerância de 15º, eliminando os pontos com menos de 20 pares (Gráfico 4.40). Baseado nas distâncias onde o valor do variograma atinge 0,35 (aproximadamente 70% do valor máximo) foi feito diagrama

direcional (*rose diagram*), que permite identificar a direção onde mais rapidamente sobe o valor de γ (60º) e a razão entre os dois eixos de maior e menor (150º) variação (Gráfico 4.40), que será usada na correção da anisotropia geométrica.

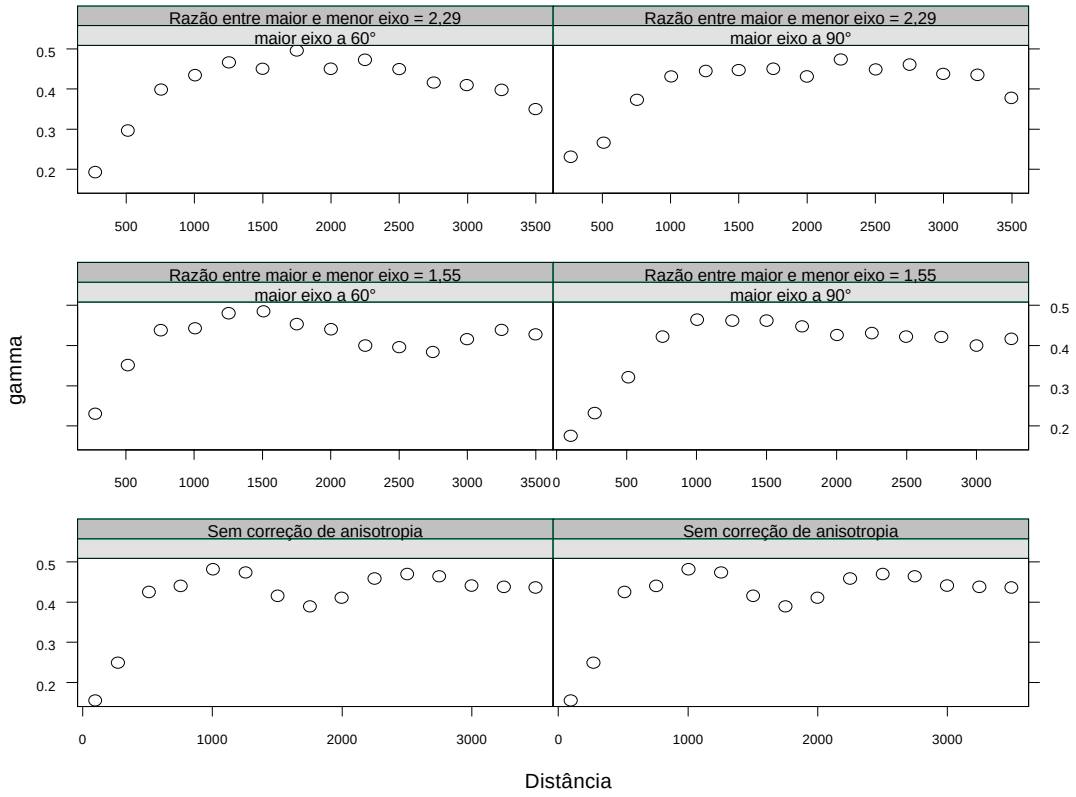
Gráfico 4.40
Variograma direcional e medida de anisotropia
Proporção de População Analfabeta (log)
Ilha do Governador, excluída área sem continuidade espacial





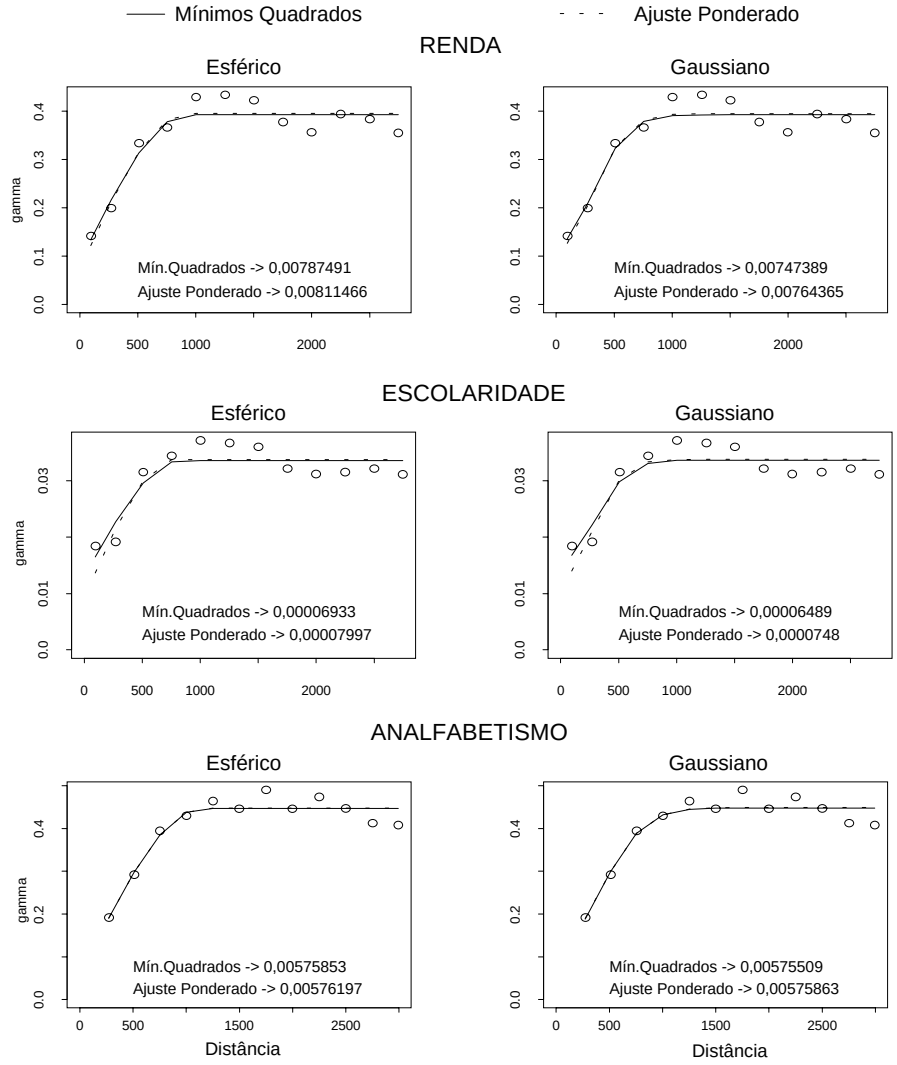
Como o diagrama direcional não é uma elipse perfeita, havendo dois pares de eixos potencialmente utilizáveis na correção da anisotropia geométrica, foram analisadas duas opções: pares de eixos a 60º e 150º ou a 0º e 90º, e razões entre distâncias de 2,3 e 1,6, sugeridas, respectivamente, pelos eixos examinados. O Gráfico 4.41 apresenta os variogramas com correção para anisotropia segundo os dois pares de eixos e razões entre as distâncias. Pode-se observar que, corrigindo a anisotropia segundo um eixo de 60º e razão de 2,3 (entre as distâncias observadas a 150º e 60º - quadro superior esquerdo), os pontos apresentam ascensão mais suave, tendo sido eliminada a flutuação observada no valor de γ entre 1.500 m e 2.200 m. Esta será a correção adotada na modelagem do variograma e na Krigeagem desta variável.

Gráfico 4.41
Variograma omnidirecional segundo diferentes correções de anisotropia
Proporção de População Analfabeta (log)
Ilha do Governador, excluída área sem continuidade espacial



Aos três indicadores selecionados para modelagem - os resíduos do *loess* da proporção de chefes de família com escolaridade igual ou superior ao segundo grau, da renda média do chefe da família normalizada e do logaritmo da proporção de analfabetos - foram ajustados modelos exponencial, esférico e Gaussiano. Os dois últimos, que tiveram o melhor ajuste, são apresentados graficamente para cada variável (Gráfico 4.42).

Gráfico 4.42
Ajustes de diferentes modelos de variograma
Renda, Escolaridade e Analfabetismo (log)
Ilha do Governador, excluída área sem continuidade espacial



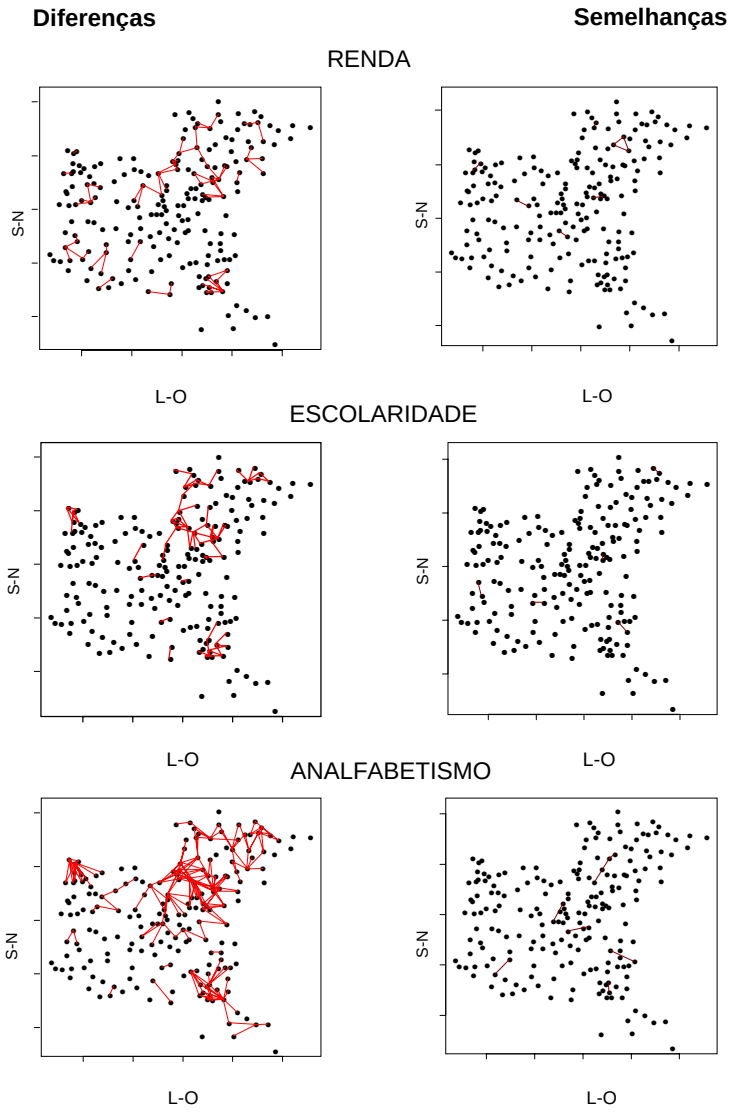
Optou-se por utilizar o modelo Gaussiano para os três indicadores, sugerido pelo formato da curva na fase inicial, com inflexão na fase ascendente, sugestiva deste tipo de modelo. Além disso, este modelo apresentou valores ligeiramente menores no indicador de ajuste adotado, ainda que muito próximos ao modelo esférico. A Tabela 4.5 apresenta os parâmetros dos modelos adotados para cada variável.

Tabela 4.5 - Parâmetros do variograma ajustado a cada indicador

Parâmetro	Renda	Escolaridade	Alfabetização
Modelo	Gaussiano	Gaussiano	Gaussiano
Pepita	0,112701	0,01272651	0,1255453
Patamar	0,2819259	0,02106183	0,3229618
Alcance	432	384	583
Valor Teste	0,00764365	0,0000748	0,00575863

O Gráfico 4.43 apresenta o diagnóstico dos modelos do variograma ajustado para cada indicador, através da identificação dos pares de pontos, dentro de um raio igual ao alcance, com semelhança maior ou dessemelhança maior do que o esperado considerando o percentil 1 de uma distribuição qui-quadrado. Em geral pode-se observar que a quantidade de pares dessemelhantes para cada distância estudada é muito superior aos semelhantes, corroborando o perfil observado na análise exploratória, onde é grande a presença de setores censitários atípicos e vizinhos com grandes diferenças.

Gráfico 4.43
Diagnóstico de ajuste dos modelos de variograma (percentil=99)
Renda, Escolaridade e Analfabetismo
Ilha do Governador, excluída área sem continuidade espacial



No modelo ajustado para a variável renda média do chefe da família (Gráfico 4.43 - primeiro quadro) nota-se alguns focos de grandes contrastes bem aparentes ao Norte, em torno de diversos pontos, e a Sudeste, mais concentrados. Pares excessivamente semelhantes neste raio são raros. O indicador de escolaridade apresenta poucos pares excessivamente semelhantes, da mesma forma que a renda. A distribuição

de pares excessivamente diferentes em relação ao previsto pelo modelo ocorre em uma região principal: Norte-Nordeste, indicando importante não estacionariedade local. A Sudeste, novamente, um pequeno aglomerado de setores com diferenças em relação aos vizinhos próximos indica, possivelmente, a existência neste local de ponto com valores muito discrepantes dos demais. O indicador de analfabetismo apresentou o pior ajuste de todos, com a maior quantidade de pares dessemelhantes, mais uma vez, localizados nas mesmas regiões.

O principal aspecto que caracteriza a discrepância entre o modelo e os dados relaciona-se aos valores atípicos, e aos grandes “saltos” no valor das variáveis em setores vizinhos, caracterizando-se a ausência de estacionariedade de segunda ordem nos indicadores. A alternativa visando melhorar o modelo seria ajustar separadamente cada sub-área identificada. Entretanto, como pode ser observado nestes mesmos gráficos de diagnóstico do ajuste, as diferenças excessivas ocorrem em distâncias muito pequenas, não permitindo a separação clara de sub-grupos, mas caracterizando uma distribuição espacial dos indicadores sociais em mosaico. Além disso, estas regiões fronteiriças apresentam padrão peculiar, sofrendo influência de ambos os extremos da distribuição. Assim, guardando ressalvas em relação aos modelos espaciais adotados, estes serão utilizados na interpolação e classificação da região em estudo.

4.9. Krigeagem universal

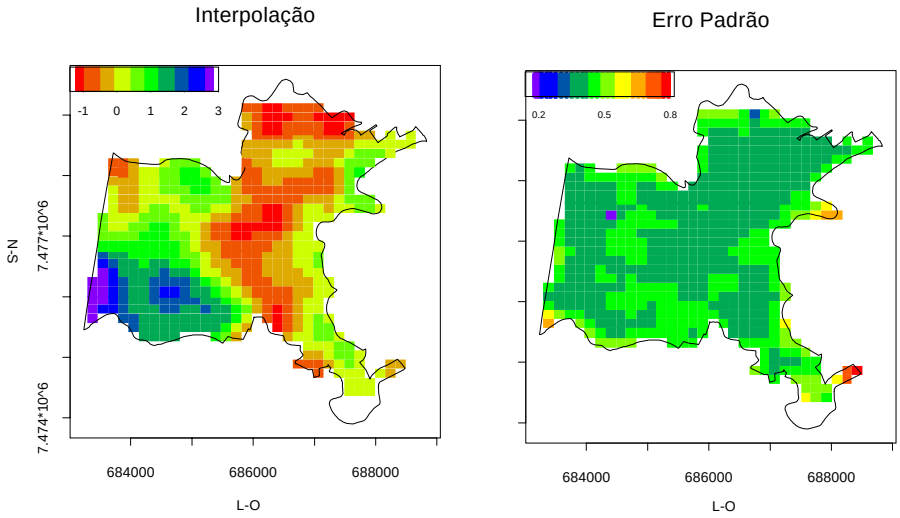
Visando recompor a estrutura da região incluindo tendência e interpolação, necessário à análise conjunta dos indicadores, a simples soma entre valores estimados pelo *loess* e pela Krigeagem, permitiria visualizar a área modelada. Optou-se, entretanto por utilizar Krigeagem universal, incorporando à estimativa em pequena escala um modelo paramétrico de tendência. A Tabela 4.6 apresenta os parâmetros ajustados a cada componente dos modelos.

Tabela 4.6 - Parâmetros dos modelos ajustados

VARIÁVEL		Renda	Escolaridade	Analfabetismo
Nº Iterações		**3**	**4**	**5**
Coeficientes do Polinômio (2º com duas variáveis)	**constante**	0	0,49	-2,97
	Coord.X	-0,72	-	-
	Coord.Y	-0,66	-	-
	X^2	1,03	0,27	-0,94
	Y^2	-	-0,28	0,93
	X*Y	1,02	-	-
Parâmetros do Variograma	**Modelo**	**Gaussiano**	**Gaussiano**	**Gaussiano**
	Alcance	571	626	460
	Patamar	0,437	0,039	0,338
	Pepita	0,124	0,015	0,206

Previsivelmente, uma vez que a variável foi normalizada, a constante ajustada pelo modelo polinomial para o indicador de renda é zero. Os parâmetros do variograma ajustados após a modelagem do polinômio foram alterados em relação aos parâmetros ajustados sobre os resíduos do *loess*, aumentando o alcance (*range*) e o valor do patamar atingido (*sill*). Pode-se observar no Gráfico 4.44 que a superfície interpolada da renda é bem semelhante ao gráfico de contorno por interpolação linear (4.10). Somente as pequenas inclusões de mais alta renda a Noroeste e Nordeste que na interpolação passam a ocupar menor área. A estrutura de erros apresenta-se homogênea, aumentando apenas nos limites da área, conforme esperado.

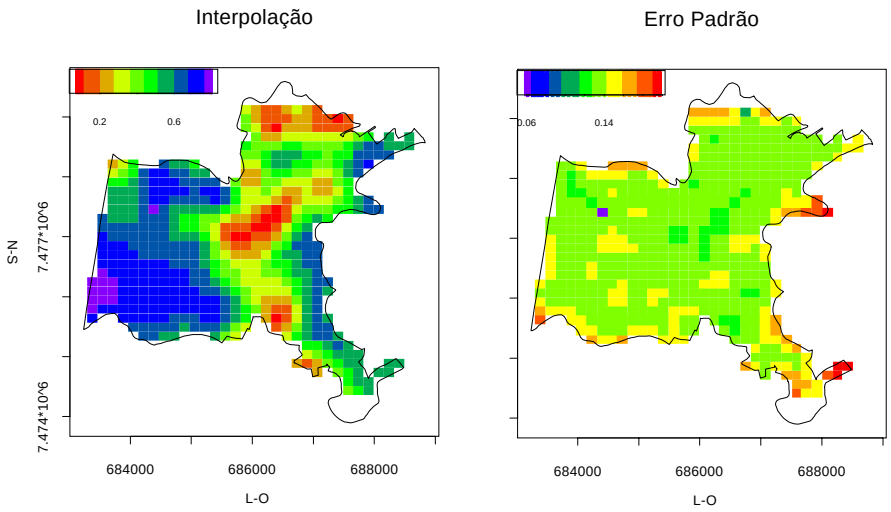
Gráfico 4.44
Krigeagem universal: superfície de interpolação e de erro padrão
Renda Média do Chefe da Família
Ilha do Governador, excluída área sem continuidade espacial



O polinômio ajustado à proporção de chefes de família com segundo grau ficou restrito a apenas dois termos - X^2 e Y^2 - além da média (0,49). Os parâmetros do

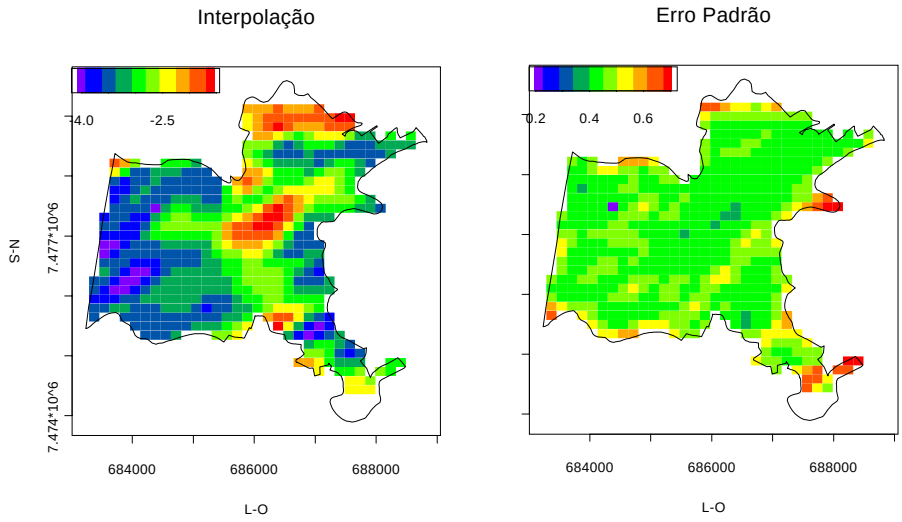
variograma sofreram alteração semelhante ao modelo da renda: o alcance aumentou para 626 m, e o patamar de estabilização (*sill*) ficou mais alto. O efeito pepita neste modelo é grande, como é característico de modelos Gaussianos, e responde por cerca de 1/3 do valor máximo do variograma. O Gráfico 4.45, com a superfície interpolada pela Krigeagem universal apresenta visível "alisamento" em relação ao mapa de contorno por interpolação linear (Gráfico 4.11), com diminuição das áreas com valores extremos (azuis e vermelhas). O erro padrão encontra-se distribuído homogeneamente, embora com valores relativamente mais altos do que para o indicador de renda.

Gráfico 4.45
Krigeagem universal: superfície de interpolação e de erro padrão
Proporção de Chefes da Família com 2º Grau
Ilha do Governador, excluída área sem continuidade espacial



Em relação ao indicador de analfabetismo houve diminuição no alcance do modelo, manteve-se o patamar e aumentou substancialmente o efeito pepita, que atinge cerca de 60% do valor do patamar. A superfície interpolada é muito semelhante à anterior, conforme pode ser observado no Gráfico 4.46.

Gráfico 4.46
Krigeagem universal: superfície de interpolação e de erro padrão
Proporção de População analfabeta (log)
Ilha do Governador, excluída área sem continuidade espacial



Comparando as Tabelas 4.5 e 4.6, pode-se observar que para as variáveis renda e escolaridade aumentou o alcance do modelo do variograma quando da incorporação da tendência através da Krigeagem universal, sugerindo que parte da estrutura espacial de pequena escala tinha sido retirada pelo *loess*, ainda que a janela fosse de 90%. No caso do modelo para analfabetismo, que sofreu o maior número de iterações, o aumento do efeito pepita pode significar que o modelo espacial que incorpora larga e pequena escala, deixa parte substancial da variância sem qualquer estrutura espacial modelável.

4.10. Recriando a classificação

A análise de *cluster* multivariado foi feita sobre os valores dos indicadores no pontos de uma grade de 32 por 32, obtidos pelos seguintes métodos:

- renda média do chefe da família (normalizada): Krigeagem universal;
- proporção de chefes de família com escolaridade igual ou superior ao

segundo grau: Krigeagem universal;

- logaritmo da proporção de analfabetos: Krigeagem universal;
- proporção de casas alugadas: ajuste por *loess*, com janela de 0,2;
- proporção de casas ligadas à rede de água: interpolação linear por triangulação;
- proporção de casas ligadas à rede de esgotos: interpolação linear por triangulação.

O *cluster* multivariado apresentou como variável mais importante a renda, seguida da escolaridade e do indicador de analfabetismo. As demais variáveis apresentaram pequeno peso na construção dos agrupamentos. (Tabela 4.7)

Tabela 4.7
Agrupamentos segundo número de pontos e média de cada indicador, e valor da razão variância entre/variância intra grupos

Grupo	N	Renda (std)	Escolar.	Analfabt. (log)	Esgoto	Água	Aluguel
A	**71**	1,736	0,705	-3,484	0,994	0,998	0,215
B	**80**	0,840	0,640	-3,425	0,926	0,996	0,234
C	**170**	0,135	0,547	-3,275	0,946	0,987	0,247
D	**113**	-0,427	0,356	-2,723	0,888	0,978	0,252
E	**74**	-0,884	0,192	-2,012	0,796	0,937	0,215
Total	**508**	**0,196**	**0,489**	**-3,021**	**0,915**	**0,981**	**0,237**
$\frac{R^2}{1-R^2}$	**-**	**11,80**	**7,02**	**3,87**	**0,14**	**0,13**	**0,04**

A caracterização dos grupos, resumida na Tabela 4.7, é muito semelhante à classificação anterior, sendo importantes as seguintes características:

A → composto por 71 pontos, apresenta maior renda média do chefe da família, melhor situação em relação a todos os indicadores, proporção de casas alugadas mais baixa que a média geral e igual ao grupo *E*, descrição esta muito semelhante a da classificação sobre os dados originais;

B → com 80 pontos, tem a renda média do chefe da família e proporção de chefes de família com escolaridade igual ou superior ao segundo grau um pouco menor, demais indicadores semelhantes;

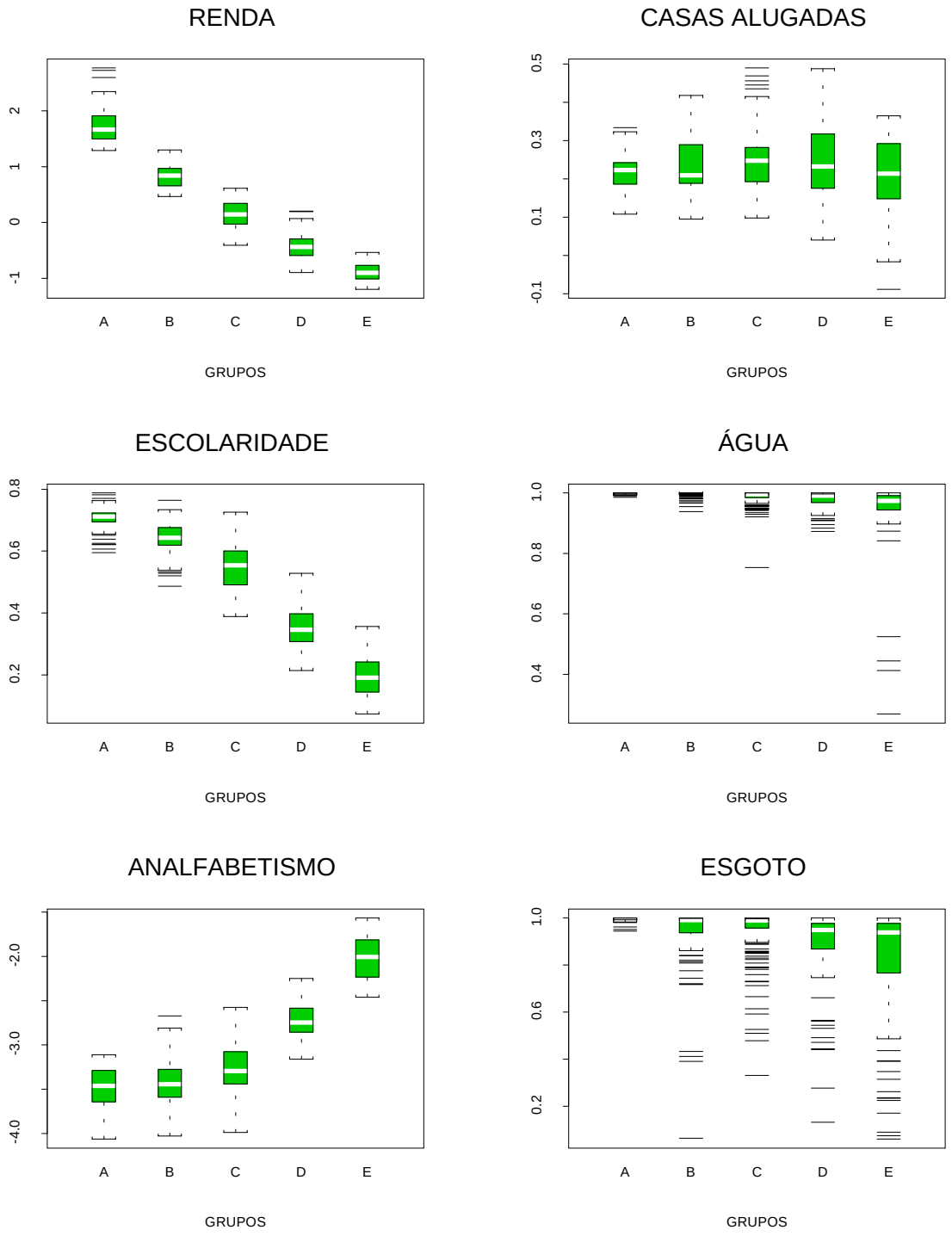
C → no maior grupo (170 pontos), a renda média do chefe da família encontra-se na média geral da população (0,135 desvios padrão) e a proporção de chefes de família com escolaridade igual ou superior ao segundo grau é superior à média geral; o acesso às redes de água e de esgotos, a proporção de população analfabeta e a proporção de casas alugadas apresentam distribuição semelhante ao grupo anterior;

D → com 113 pontos apresenta, junto com o grupo *E* as menores renda e proporção de chefes de família com escolaridade igual ou superior ao segundo grau, maior proporção de casas alugadas, diminuindo gradativamente as proporções de casas ligadas às redes de água e esgotos.

E → com 74 pontos, tamanho muito próximo ao grupo *A*, caracteriza-se pela proporção de casas ligadas à rede de água e esgotos no extremo inferior da distribuição; a proporção de população analfabeta apresenta a maior amplitude de distribuição e a proporção de casas alugadas neste grupo é semelhante a do grupo *A*.

As principais diferenças entre as duas classificações podem ser vistas no Gráfico 4.47, onde os indicadores de alfabetização e acesso à rede de esgotos sofreram as maiores alterações. O primeiro, graças essencialmente à transformação da variável. O segundo, provavelmente efeito do alisamento e interpolação, onde surgem valores intermediários anteriormente inexistentes, descaracterizando o grupo *E*. Os valores extremos anteriormente encontrados no indicador proporção de casas alugadas foram eliminado através do *loess*.

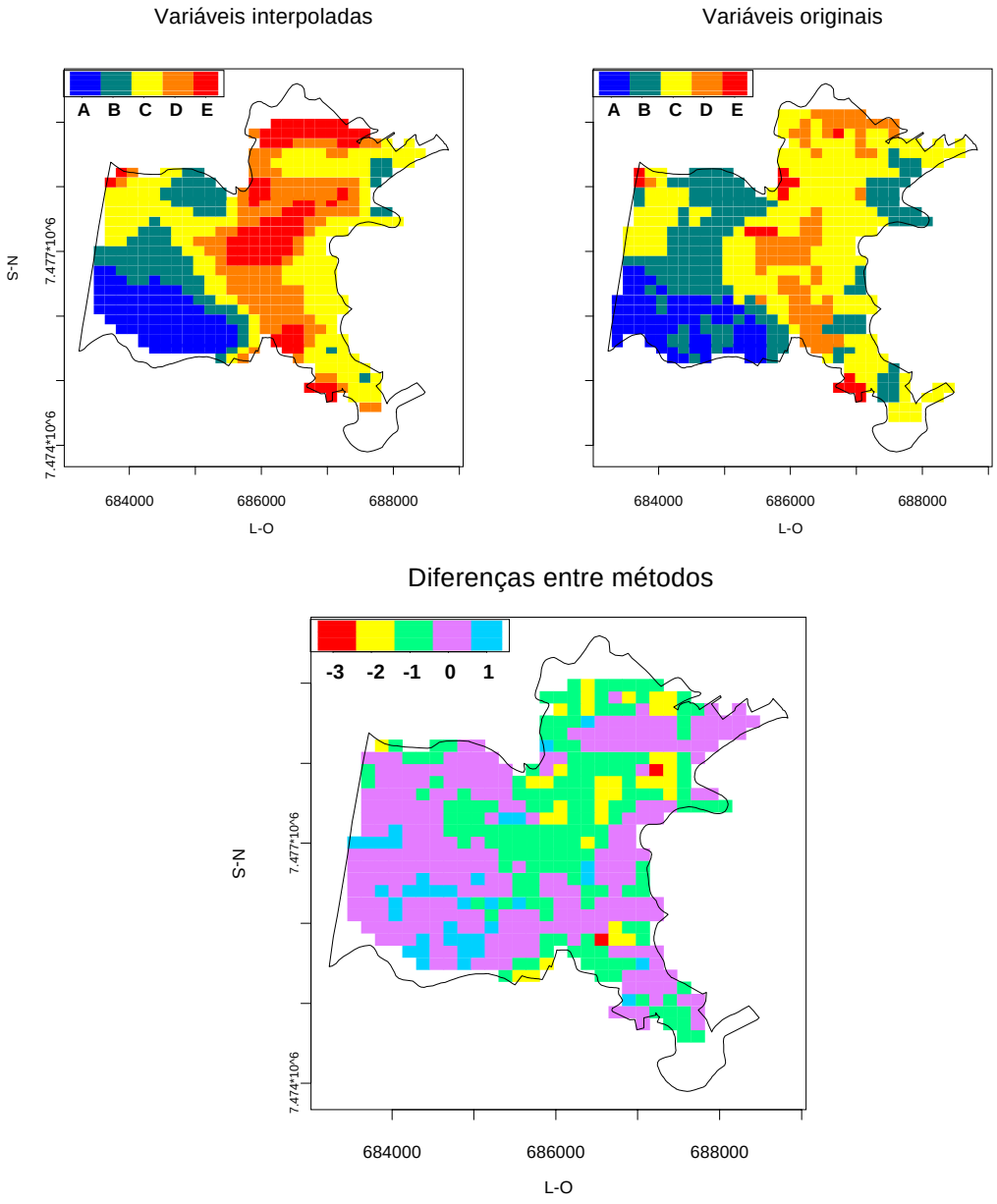
Gráfico 4.47
***boxplots* das variáveis após interpolação**
segundo grupos socioeconômicos resultantes de classificação multivariada
Ilha do Governador, excluída área sem continuidade espacial



4.11. Comparando os resultados

No Gráfico 4.48 pode-se visualizar a distribuição espacial da área segundo classificação socioeconômica, com as variáveis originais, após a interpolação espacial e a diferença entre os dois modelos. No mapa com a análise após interpolação espacial, observa-se que o centro do terreno pertence ao grupamento com piores condições, melhorando o perfil na direção Sudoeste. Embora a mesma tendência exista no outro mapa, com a classificação aplicada sobre as áreas originais, duas grandes diferenças aparecem: a maior extensão da área pertencente ao grupo *E* no primeiro quadro do gráfico, e a maior oscilação entre diferentes classificações no segundo. Este efeito decorre do alisamento gerado pela interpolação espacial.

Gráfico 4.48
Classificação multivariada: variáveis interpoladas, originais e diferenças entre modelos
Ilha do Governador, excluída área sem continuidade espacial



A comparação entre os dois métodos de classificação mostra que apenas em pequenos trechos localizados a Nordeste e Sul a diferença entre eles é superior a mais de um nível. Como pode ser visto no último quadro do Gráfico 4.48, o tipo de diferença mais freqüente entre classificação tradicional e classificação com interpolação espacial ocorre no sentido de atribuir categoria inferior à anterior (em verde claro). Em pequenas áreas localizadas no mesmo trecho onde havia sido detectado grandes saltos entre as variáveis (Gráfico 4.27), e onde o diagnóstico dos modelos de variograma adotados indicavam ajuste pobre (Gráfico 4.43) esta diferença nas classificações chega a dois níveis (em amarelo). Quando a diferença ocorre no sentido de melhorar a classificação de um determinado ponto (em azul), esta é no máximo de um nível, e se concentra na região Sudoeste, sendo que este tipo de troca na classificação ocorre quase exclusivamente entre os grupos *B* e *A*.

5. DISCUSSÃO

5.1. Classificação e risco

A idéia básica por trás deste esforço de modelagem de indicadores socioeconômicos é discutir a ocorrência de doenças nos grupos populacionais, à luz de sua localização espacial. Embora usualmente somente questões relacionadas à dispersão de poluentes fabris sejam consideradas fatores de risco ambiental, ambiente de fato é algo mais amplo. Ainda que variáveis socioeconômicas sejam eventualmente consideradas fatores de risco associadas à ocorrência de doenças, são usualmente tratadas enquanto características do indivíduo. Neste trabalho, renda, escolaridade ou acesso à água, não são variáveis relacionadas ao indivíduo, mas caracterizam o perfil de um grupo populacional. Segundo Susser (1994) os padrões de mortalidade ou morbidade, a propagação de epidemias, a transmissão sexual de doenças e a transferência de comportamentos ou valores não podem ser compreendidos somente a partir do indivíduo. As pessoas vivem em grupos e somente a análise dos grupos permite captar comportamentos, valores, ou mesmo a transmissão de infecções que ocorre nessa escala.

Na caracterização de grupos populacionais utilizando fatores de risco socioeconômicos, são grandes as dificuldades. Trabalho recente analisando variáveis relativas à saúde do recém-nato e classificação socioeconômica de risco levanta o seguinte problema: "*Torna-se difícil avaliar o que é 'pior' do ponto de vista do recém-nascido: nascer em subúrbio tradicional, onde residem trabalhadores de*

baixa renda, em favela da região nobre da cidade, ou em um sítio com características ainda rurais." (d'Orsi, 1996). Ainda assim, encontrou-se associação entre perfil socioeconômico e Índice de Apgar[2], apesar dos limites da escala de análise utilizada (bairro). Entre renda, escolaridade ou acesso à água e o adoecer existem dezenas de passos intermediários, que vão do comportamento alimentar, determinado por complexa interação entre renda e hábitos culturais, até a disponibilidade e qualidade da assistência à saúde. Assim, mais do que estabelecer toda a cadeia de eventos que levou à doença, onde o peso de variáveis socioeconômicas é relativizado, buscou-se olhar os grupos populacionais visando estabelecer padrões que possam ser utilizados para orientar propostas de intervenção. Ou seja, além de identificar fatores causais de doenças, é necessário localizar, o mais precisamente possível, os grupos populacionais onde se deve atuar, direcionando as medidas preventivas, sejam elas de cunho ambiental, educacional ou assistencial.

Todo o trabalho de construção e análise dos indicadores socioeconômicos utilizados no artigo "*Método multivariado de classificação socioeconômica de microáreas urbanas - os setores censitários da Região Metropolitana do Rio de Janeiro*" (Anexo 1), permitiu amadurecer a seleção das variáveis, não somente a partir de suas características estatísticas, mas também segundo o que se desejava evidenciar na região estudada. As principais características da 20ª Região Administrativa, decisivas

[2] O Índice de Apgar consiste em uma nota de 0 a 10 resultante da avaliação das funções vitais do recém-nato no primeiro e no quinto minuto de vida, indicando as condições de saúde ao nascer.

para que fosse escolhida enquanto a área deste estudo, são a homogeneidade na ocupação do território e a heterogeneidade socioeconômica dos grupos populacionais residentes. Os grandes espaços vazios, setores censitários do aeroporto, cidade universitária e área industrial, localizam-se externamente à área, e simples exclusão facilitou a utilização das técnicas geoestatísticas. Por outro lado, conforme pode ser visto no Anexo 1, na Ilha do Governador encontra-se de favelas a mansões, em um microcosmo bastante representativo da cidade.

É característico do modelo de urbanização do Rio de Janeiro encontrar-se em uma dada região grupos sociais distintos, ou seja, favelas coexistindo com áreas nobres. Neste caso, os indicadores para estas regiões representam uma média entre grupos populacionais diferentes. Evidentemente os agravos à saúde e o acesso aos equipamentos urbanos não estão distribuídos homogeneamente na população. Diferenciar os grupos populacionais segundo os riscos potenciais e o acesso aos recursos terapêuticos é uma das questões-chave da prevenção em saúde. A utilização dos setores censitários da FIBGE como unidade mínima de análise permite discriminar o que, na menor divisão político-administrativa (bairro), ficava obscurecido. Ainda assim, no caso de alguns indicadores, particularmente o acesso à rede de água, ficou a impressão de que se deveria localizar pontualmente as residências sem acesso à este recurso, que nos pouco setores onde a cobertura não é de 100% não devem somar mais do que 6 domicílios.

Por um lado, se quanto menor a área para a qual se dispõe dos dados, maior é a capacidade de diferenciar objetos próximos, e maior a homogeneidade interna da

unidade adotada, por outro, a utilização de áreas muito pequenas, com população também muito pequena, introduz uma variabilidade espúria nos valores dos indicadores em estudo. A linha de base pode se aproximar de zero, impossibilitando o tratamento estatístico do fenômeno. Assim, é necessário estabelecer um compromisso entre estabilidade e homogeneidade, aspecto este tratado principalmente no artigo "*Spatial partition using multivariate...*" (Anexo 2).

Um aspecto importante do trabalho com microáreas urbanas relaciona-se à definição de setor censitário: "*região a ser coberta por um recenseador no período de um censo*" (FIBGE,1994), onde não há, evidentemente, qualquer perspectiva de análise de risco. É preocupação da FIBGE evitar a inclusão no mesmo setor censitário de favelas e áreas urbanas, ou de extrapolar os limites naturais de conjuntos habitacionais ou condomínios. Ainda assim, o critério que delimita o setor censitário em função do trabalho do recenseador implica, por exemplo, na divisão de um mesmo prédio em dois e até três setores, ou na estratégia de cobertura de áreas faveladas em espiral a partir de um dado ponto, até que dois recenseadores se encontrem. As fronteiras dos setores censitários não delimitam sequer unidades administrativas ou políticas, mas apenas operacionais. Assim, podem ser considerados artefatos de coleta do dado, que na essência é contínuo em todo o espaço analisado. Esta foi a perspectiva que norteou a utilização dos métodos de análise de dados espacialmente contínuos.

O outro aspecto considerado foi que, ainda que se possa pensar, por exemplo, na falta de acesso à água encanada como um fator de risco independente do local onde ocorre, suas conseqüências são mediadas por uma realidade local com influência

essencial no desenlace sobre a saúde do indivíduo. Por exemplo, entre dois grupos populacionais moradores em favelas com condições sanitárias semelhantes e de renda igualmente baixa, apresenta melhores condições de saúde aquele mais próximo a bairros mais ricos, e possivelmente melhor servidos quanto à assistência hospitalar de emergência, ainda que seja por tradição histórica. Também na direção inversa ocorre a "contaminação" socioeconômica: residir em locais nobres porém próximos a áreas com alto índice de violência aumenta, por exemplo, o risco de ser atingido por "balas perdidas", ainda que a vítima principal da violência urbana se encontre nos grupos economicamente mais marginalizados da sociedade.

Em síntese, quatro questões relacionadas à caracterização do risco de grupos populacionais foram desenvolvidas durante o período do doutoramento:

- a construção, análise e seleção de indicadores a partir dos dados secundários disponíveis - Anexo 1;
- o tratamento conjunto dos indicadores selecionados visando estabelecimento de um perfil único de risco - Anexos 2 e 3;
- métodos de análise para microáreas - Anexo 2;
- a incorporação da contiguidade espacial - corpo principal aqui apresentado.

5.2. Geoestatística e indicadores sociais em áreas urbanas

Neste item serão discutidas as técnicas apresentadas, quanto à capacidade de contribuir para a descrição, visualização, levantamento de hipóteses, modelagem e

diagnóstico dos ajustes, das seis variáveis selecionadas. Diversas outras técnicas também empregadas não foram incluídas nos resultados, por não acrescentarem qualquer informação útil à compreensão do problema. Ainda assim, eventualmente estas tentativas serão referidas, apenas como elemento para comparação.

O problema discutido neste trabalho - uma classificação socioeconômica baseada em variáveis associadas a áreas - deveria, a princípio, ser modelado a partir das ferramentas de análise de áreas. De fato algumas técnicas deste tipo foram empregadas, particularmente o teste de autocorrelação espacial de Moran. Entretanto, o próprio tipo de terreno, sem descontinuidades relevantes (Gráfico 4.2), onde as áreas são muito pequenas e foram criadas arbitrariamente visando apenas a operacionalização do censo, fez com que se levantasse a possibilidade de utilizar modelos onde o território fosse considerado uma superfície única. Assim, considerando o centróide de cada setor censitário como um ponto amostral onde foram medidas cada uma das variáveis utilizadas, desaparecem os limites de área, procedendo-se então como em qualquer modelagem geoestatística. Outro aspecto considerado foi a disponibilidade de um conjunto de métodos - da análise exploratória à modelagem de tendência - utilizados para terrenos contínuos que permitiu importantes percepções quanto aos indicadores e região estudada.

Por outro lado, as seis variáveis analisadas apresentam diferentes distribuições - estatística e espacial - permitindo avaliar a aplicabilidade dos métodos à ampla gama de problemas relacionados a indicadores sociais, e, por extensão, de saúde. Um aspecto importante a ser ressaltado na distribuição espacial das variáveis é o padrão em mosaico,

onde regiões vizinhas apresentam grandes contrastes, característico do modelo de urbanização do Rio de Janeiro (Mapa 2 no Anexo 2). Este aspecto é particularmente importante na modelagem espacial, uma vez que estas mudanças abruptas entre áreas vizinhas apresentam problemas especiais na modelagem.

Quanto à distribuição estatística, as variáveis relacionadas ao saneamento apresentam comportamento peculiar, onde o acesso à água e à rede de esgotos é de 100% em, respectivamente, mais de 50% e 35% dos setores censitários. O indicador relacionado ao acesso à rede de esgotos apresenta um grupo de 12 setores censitários onde praticamente inexiste ligação domiciliar à rede de esgotos. É importante observar que a falta de ligação à rede de água e à rede de esgotos não ocorrem no mesmos setores, como pode ser constatado nos Gráficos 4.12 e 4.13. Através destas variáveis é possível identificar microáreas que se destacam pela precariedade das residências e, provavelmente, de ocupação muito recente.

As demais variáveis - renda média do chefe da família, proporção de chefes de família com escolaridade igual ou superior ao segundo grau, logaritmo da proporção de analfabetos e proporção de casas alugadas - apresentam curva de distribuição mais suave e mais próxima da normalidade, havendo colinearidade entre as três primeiras (Gráfico 4.7).

Serão discutidos as seguir os principais aspectos da aplicabilidade dos métodos geoestatísticos aos dados analisados.

5.2.1. Análise exploratória espacial

Os métodos de análise exploratória de dados têm por objetivo, basicamente, descrever, levantar hipóteses e modelos apropriados, bem como identificar observações atípicas em relação ao modelo subjacente. No caso da análise estatística espacial, os mesmos objetivos se aplicam à localização. Por exemplo, são atípicos não só os valores extremos em relação ao conjunto, mas também aqueles que apresentam grandes diferenças em relação aos vizinhos. Outro aspecto a considerar nesta fase da análise é a detecção de tendência, da mesma forma que nos estudos de séries temporais. Entretanto, dois aspectos diferenciam os objetos: enquanto ao longo do tempo existe uma ordenação natural e periodicidade regular na coleta do dado, no espaço não existe direcionalidade a priori e as amostras não são regularmente espaçadas. Assim, grande atenção foi dada neste trabalho para as técnicas de análise exploratória espacial, que permitem estudar a estacionariedade de primeira e segunda ordens, global e local.

Entre as técnicas utilizadas, os gráficos de médias e medianas por linhas e colunas (Gráficos 4.16 a 4.18) mostraram-se bastante eficazes na análise das variáveis cuja distribuição é menos concentrada: renda, escolaridade e proporção de casas alugadas. Para as variáveis de saneamento, água e esgoto, este tipo de gráfico pouco acrescentou ao que já se sabia. Ou seja, onde não há padrão espacial detectável através dos mapas de contorno, esta técnica apenas confirmou o que já se havia observado. Em relação à proporção de população alfabetizada, foi examinando este gráfico, em conjunto com os tradicionais histograma e *boxplot*, que se decidiu pela transformação

logarítmica da variável, possibilitando então a identificação de direção da tendência.

Os gráficos da nuvem de pares do variograma amostral (Gráficos 4.21 a 4.26) permitiram identificar os pares de pontos cuja diferença é atípica, possibilitando mapear as áreas onde ocorre não estacionariedade local para cada variável. Este mapa (Gráfico 4.27), muito simples na concepção e execução, permite identificar áreas contrastantes para cada variável e para diferentes combinação entre elas. Os *boxplots* do indicador de diferença entre pares, mais uma vez, confirmam que este tipo de análise não é adequada aos indicadores de saneamento (Gráficos 4.25 e 4.26). Este recurso, entretanto, permite visualizar em relação à proporção de casas alugadas (Gráfico 4.24) o comportamento errático da variável em pares de pontos cuja distância está entre 800 e 3.000 m, comportamento este que ficou melhor explicado ao se modelar tendência e variação em pequena escala.

Após a análise exploratória dos indicadores de acesso à água e esgoto, diversas transformações das variáveis foram testadas, com o objetivo de verificar possível adequação aos métodos de modelagem propostos. Entretanto, a importância destas variáveis é muito maior na identificação de áreas com grande contraste em relação aos vizinhos, não cabendo, portanto, qualquer alisamento. Assim, frente aos dados disponíveis optou-se por simplesmente fazer a interpolação linear por triangulação. Este é o tipo de variável que nas condições da Ilha do Governador, área com boa distribuição da rede de água e esgoto, a localização pontual das residências, ou grupos de residências, não servidas pelos sistemas é, por si só, indicativo de risco diferenciado. Outra possibilidade de tratamento para estes indicadores é utilizá-los como variáveis

dicotômicas, a partir da definição de pontos de corte que caracterizam ter ou não acesso a este equipamento urbano. Neste caso os indicadores poderiam identificar diferentes extratos espaciais para análise.

Ainda na fase exploratória da análise, utilizando métodos mais comumente aplicados a áreas foi testada a autocorrelação espacial através da estatística *I* de Moran. Uma vez que era intenção partir para a análise como um terreno contínuo, utilizou-se como ponderador o inverso da distância entre os centróides de cada setor censitário.

5.2.2. Tendência

Antes da utilização do *loess* na modelagem da tendência espacial, foram ajustados aos indicadores polinômios de segundo e terceiro graus, com resultados visualmente pouco satisfatórios, onde as superfícies ajustadas pouca semelhança guardavam com a distribuição da variável. Também a remoção da tendência através da técnica de "*median polish*" (Cressie, 1991) foi experimentada. Neste caso, ao se analisar os resíduos do procedimento para modelagem através do variograma toda a estrutura espacial tinha sido eliminada junto com a tendência.

O *loess* apresentou excelente resultado na modelagem da tendência. Através da variação na amplitude da janela foi possível escolher aquela que, modelando a variação em larga escala, não eliminasse dos resíduos a estrutura espacial em pequenas distâncias. Os gráficos onde a curva *loess* ajustada são comparados aos resíduos, feitos para as coordenadas originais e após a rotação de 45º, foram decisivos na seleção do modelo adotado. Os resíduos, após a retirada da tendência por este método,

apresentam distribuição mais próxima da normalidade. O teste de Moran, após a retirada da tendência, permitiu verificar que nesta escala - distância entre setores censitários - ainda estava presente a estrutura espacial.

5.2.3. Variograma

O principal problema desta etapa da análise estatística espacial é a seleção do modelo a ser ajustado, seja quanto ao tipo, parâmetros e eventual anisotropia, freqüentemente feita por ajuste visual. Segundo Bailey & Gatrell (1995), "*Freqüentemente modelos de variograma são ajustados ao variograma amostral observado 'no olho',.... Particularmente, deve-se evitar a tendência de super ajustar o modelo de variograma tentando capturar cada e todas as dobras. O objetivo da modelagem é capturar a estrutura básica da dependência espacial;...*" A principal dificuldade no processo foi descobrir o que seria satisfatório. Diversos modelos foram ajustados, de forma interativa, até que se começasse a perceber o efeito de cada mudança de parâmetro nas curvas, e, adquirindo alguma experiência, se pudesse selecionar o tipo de modelo. Em seguida, selecionada a curva adequada, de modelos bastante simples, os parâmetros foram ajustados por mínimos quadrados ponderados, onde o peso de cada ponto é inversamente proporcional à variância (Cressie, 1991). Optou-se por este ajuste ao invés do método usual de mínimos quadrados porque desta forma é menor a importância de pontos com número reduzido de pares e com grande dispersão para uma dada distância.

O processo de investigação de anisotropia utilizando a técnica de diagramas

direcionais foi simples tanto na execução quanto na interpretação quando comparado a outras tentativas, como o mapa de contorno dos valores do variograma. É importante observar que, exceto para o primeiro intervalo, havia número suficiente de pares em cada direção: apesar da pequena extensão da região estudada, a malha de setores censitários permitiu uma cobertura detalhada da região. Novamente houve certa dose de arbitrariedade no exame da anisotropia, onde a comparação com outros exemplos foi fundamental na decisão. A observação inicial sugeria haver diferenças no patamar atingido, a 90º e 0º para as variáveis renda e escolaridade, o que significaria a presença de anisotropia zonal. Entretanto, comparando estes gráficos com os encontrados em outros trabalhos (Isaaks & Shrivastava, 1989 - pag.152), concluiu-se que as diferenças observadas eram de pouca monta, exceto para o indicador de analfabetismo, sendo a variação direcional essencialmente geométrica e não zonal. Os gráficos comparando variogramas direcionais, curvas *loess* ajustadas e intervalo para dado valor do variograma amostral, foram fundamentais para compreensão do processo anisotrópico. É importante observar que este tipo de modelo não paramétrico como o *loess*, utilizado na remoção da tendência, na modelagem espacial da variável logaritmo da proporção de analfabetos e na seleção de modelos para avaliação da anisotropia são menos freqüentes na literatura geoestatística.

O diagnóstico do ajuste dos modelos adotados corroborou o mapa de pontos contrastantes da análise exploratória, particularmente para renda e escolaridade. No caso do analfabetismo, a seleção de pares cuja diferença é superior ao esperado

aumentou a quantidade de pares identificados como tendo ajuste pouco consistente. Ou seja, foi possível comparar indicadores e modelos, quanto à intensidade e localização de problemas no ajuste do variograma.

5.2.4. Interpolação e classificação multivariada

Inicialmente optou-se por interpolar espacialmente os indicadores de renda, escolaridade e analfabetismo somando os valores ajustados pelo *loess* aos interpolados através da Krigeagem. Entretanto, ainda que visualmente esta solução fosse adequada, a Krigeagem universal possibilitou resolver simultaneamente as duas questões - tendência e variação em pequena escala - incluindo estimativa de erro padrão. Ainda assim, foi fundamental a utilização dos parâmetros obtidos na análise dos resíduos do *loess*. Uma vez que o processo de estimar alternadamente parâmetros do polinômio e do variograma, até que a diferença entre eles fosse menor do que o valor arbitrado, foi bastante trabalhoso. Por isso a seleção dos valores iniciais é importante, tendo sido todo o processo facilitado pela utilização dos valores estimados do variograma do resíduos do *loess*, que apresentavam valores muito próximos do ajuste final, como parâmetros iniciais. Além disso, "*...Krigeagem com modelo inadequado pode produzir piores estimativas do que outros modelos mais simples*" (Isaaks & Shrivastava, 1989).

As variáveis interpoladas através da Krigeagem foram utilizadas conjuntamente com variáveis interpoladas por outros métodos, cabendo discutir a utilização de cada um destes na composição da classificação multivariada. Em primeiro lugar, o *loess* apresenta uma solução interessante para o problema da variável proporção de casas

alugadas, onde a estrutura de dependência espacial existe ou em escala maior do que a disponível, intra setor censitário, ou em escala menor, como a tendência detectada. Quanto à interpolação linear utilizada para os indicadores de água e esgoto, o problema é mais delicado. A ausência de ligação à rede de água ou esgoto na Ilha do Governador é um evento raro. Assim, técnicas com grande potencial de alisamento não interessavam, uma vez que se pretendia preservar o poder discriminatório dessas variáveis quanto à detecção de áreas “ruins”. Por outro lado, simplesmente atribuir a cada ponto de cada setor censitário o valor da variável, excluía a possibilidade de trabalhar a transição: áreas próximas a residências sem água ou esgoto certamente sofrem o efeito das más condições de saneamento. Assim, como uma primeira aproximação, utilizou-se a interpolação linear simples, preservando o valor do indicador no centróide do setor censitário. Ainda assim, o efeito da interpolação sobre o indicador proporção de casas ligadas à rede de esgotos alterou o perfil do grupo *E*, antes melhor caracterizado. No caso destas variáveis, a localização pontual das residências sem acesso a saneamento básico e a criação de áreas tampão, onde os efeitos das condições locais se propagam seria uma solução interessante.

Entendendo que a classificação das áreas é apenas um método que pode ser considerado como análise exploratória de dados, cabe discutir a adequação das técnicas de interpolação apresentadas. É interessante relembrar uma característica básica da Krigeagem, não casualmente intitulada BLUE (*best linear unbiased estimator*). Ou seja, considerando que foi escolhido e ajustado corretamente o modelo de dependência espacial (variograma), entre os métodos de interpolação disponíveis a Krigeagem é o

melhor estimador possível. Entretanto, fica a questão: será que é necessário toda esta precisão neste tipo de estudos? Claro que depende dos objetivos: para testar hipóteses relacionando causa e efeito espacialmente distribuídos, o método mais consistente de interpolação é este. Entretanto, considerando o aspecto algo arbitrário da classificação socioeconômica, certamente um filtro não paramétrico como o *loess* teria possivelmente apresentado resultados próximos, dependendo do tamanho da janela utilizada.

Um aspecto importante a discutir é o modelo de classificação adotado. Considerando não só a continuidade do terreno, mas também a existência de ampla gama de possibilidades de classificação dos setores censitários, seja alterando a composição dos indicadores, ou o número de grupos criados, seria interessante construir um modelo onde cada setor fosse descrito a partir de seu grau de pertinência aos grupos, estabelecendo um critério de risco contínuo e possivelmente mais representativo do risco real. Um modelo de classificação por lógica difusa (Braga & Fucks, 1995) seria pertinente ao que se propõe neste trabalho.

6. CONCLUSÃO

As variáveis coletadas no censo demográfico de 1991 permitiram a construção de grande número de indicadores, abordando diversos aspectos: saneamento (água, esgoto e coleta de lixo); ocupação do terreno (propriedade e tipo de construção); escolaridade (anos de estudo dos chefes de família, alfabetização da população); domicílio (média de cômodos e moradores); demografia (mulheres chefes de família e mediana etária por sexo); renda (somente do chefe da família).

O método de classificação multivariado por partição, a partir das seis variáveis selecionadas através da análise de componentes principais - *renda média do chefe da família, proporção de chefes de família com escolaridade igual ou superior ao segundo grau, proporção de população alfabetizada, proporção de casas alugadas, proporção de casas ligadas à rede de esgotos* e *proporção de casas ligadas à rede de água,* permitiu a criação de grupos bem caracterizados, que podem ser considerados diferentes quanto ao perfil de risco à saúde.

As técnicas de análise exploratória de dados - mapa de contorno por interpolação linear, gráfico de médias e medianas por linhas e colunas, nuvem de pares do variograma amostral foram muito úteis na identificação de possíveis modelos para a distribuição espacial do fenômeno.

A regressão linear local (*loess*) permitiu modelar e visualizar a tendência dos indicadores de renda, escolaridade e analfabetismo, estudar os parâmetros do

variograma amostral e fazer a interpolação final do indicador proporção de casas alugadas. Além disso, o método é de fácil execução, sendo possível estudar o efeito de diferentes janelas.

A interpolação linear dos indicadores de saneamento, ainda que seja uma solução excessivamente simplificada, foi a única possível face os limites do dado, que neste caso deveria ser pontual.

A modelagem através do variograma e interpolação por Krigeagem universal mostrou-se adequada, permitindo tratar o espaço enquanto contínuo. Nos locais onde havia excesso de pares de pontos com diferenças acima do previsto, a interpenetração das características de setores censitários distintos, com "contaminação" entre uma área e outra, parece ser um forma adequada à modelagem dos eventos em microáreas urbanas. Um aspecto fundamental na discussão do método é a relação custo-benefício. As técnicas são de difícil execução e exigem razoável familiaridade prévia com os modelos e com os dados, de forma a incorporar os aspectos realmente essenciais, cabendo sempre avaliar a necessidade real de precisão na interpolação, frente à qualidade dos dados e ao modelo de determinação de doenças subjacente.

Na classificação obtida após a modelagem espacial, onde houve mudanças de classe, esta ocorreu graças à incorporação da localização relativa. Assim, áreas anteriormente classificadas em grupos socioeconomicamente afastados, quando próximas, foram reclassificadas, fazendo-se sentir o efeito da "vizinhança" - no sentido comum da palavra - no perfil social.

Surgiram diversas perspectivas interessantes de desdobramento deste trabalho no desenvolvimento da sub-área de análise espacial aplicada à saúde. A utilização de métodos difusos de classificação apresenta potencial interessante, que permitiria abordar também a classificação de forma contínua, além de possibilitar alterações no algoritmo de agregação de áreas (Anexo 2) que viabilizassem o tratamento de áreas pertencentes a um determinado grupo socioeconômico, inclusas em região classificada em outro grupo.

Entre os possíveis desdobramentos, o mais importante é a aplicação deste tipo de modelo a dados de morbi-mortalidade, que somente será possível se viabilizada a localização de eventos nos setores censitários, atividade esta já em andamento. A micro-localização deste tipo de dado permitiria a abordagem de problemas de saúde **de** grupos populacionais incorporando a localização espacial, utilizando técnicas que permitissem modelar diversos fatores de risco e problemas de saúde de um ponto de vista coletivo.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ALEXANDER,F.E., MCKINNEY,P.A., WILLIAMS,J., RICKETTS,T.J., CARTWRIGHT,R.A., 1991, "Epidemiological evidence for the 'two-disease hypothesis' in Hodkin's disease", **International Journal of Epidemiology**, v.20, n.2, pp.354-361.

BAILEY,T.C., GATRELL,A.C., 1995, **Interactive spatial data analysis**, 1 ed. Essex, Longman Scientific & Technical.

BARCELLOS, C., MACHADO, J.H., 1991, "Seleção de indicadores epidemiológicos para o saneamento", **BIO**, out/dez , pp. 37-41.

BARRETT F.A., 1991, "'SCURVY' Lind's medical geography", **Social Science and Medicine**, v.33, pp.347-353.

BARRY,R.P., 1996, "A diagnostic to assess the fit of a variogram model to spatial data", **Journal of Statistical Software**, v.1, n.1, pp.1-10, URL: http://www.stat.ucla.edu/journals/jsf/v01/i01/.

BESAG,J., NEWELL,J., 1991, "The detection of clusters in rare diseases", **J.R.Stat.Soc.A**, v.154, n.1, pp.143-155.

BRAGA,L.P.V., FUCKS.S.D., 1995, **Analysis and classification of soil properties by geoestatistical and fuzzy methods**, Monografias técnicas do Instituto de Matemática/UFRJ, nº 85.

CARSTAIRS,V., 1995, "Deprivation indices: their interpretation and the use in relation to health", **Journal of Epidemiology and Community Health**, v.49, suppl.2, pp.S3-S8.

CARVALHO, M.S., CRUZ, O.G., 1995, **Mapeamento de áreas de risco no Rio de Janeiro**. In: Relatório final do projeto Saúde e Qualidade Ambiental, convênio IPLAN-RIO/FIOCRUZ, pp.49-81.

CARVALHO, M.S., CRUZ, O.G., NOBRE, F.F., 1995, "Análise multivariada do censo 1991 por setores censitários - Região Metropolitana do Rio de Janeiro/Brasil". In: **Resumos do III Congresso Brasileiro de Epidemiologia**, pp.18, Salvador, Jun.

CARVALHO, M.S., CRUZ, O.G., NOBRE, F.F., 1996, "Spatial partition using multivariate cluster analysis and contiguity algorithm: application to Rio de Janeiro, Brazil", **Statistics in Medicine**, v.15, pp.1885-1894.

CASTELLANOS,P.L., 1990, "Sobre el concepto de salud-enfermedad. Descripción y explicación de la situación de salud", **Bol. Epidemiológico OPAS**, v.10, n.4, pp.1-7.

CHEN,R., MANTEL,N., KLINGBERG,M.A., 1984, “A study of three techniques for time-space clustering in Hodgkin's disease”, **Statistics in Medicine**, v.3, n.1, pp.173-184.

CLIFF,A.D., ORD,J.K., 1981, **Spatial processes: model and applications**, 1 ed., London, Pion.

CORTINOVIS,I., VELLA,V., NDIKU,J., 1993, “Construction of a socio-economic index to facilitate analysis of health data in developing countries”, **Social Science and Medicine**, v.36, n.8, pp.1087-1097.

CRESSIE,N., 1991, **Statistics for spatial data**. 1 ed., New York, Wiley Interscience.

CRESSIE,N., HAWKINS,D.M., 1991, “Robust estimation of the variogram”, **Mathematical Geology**, v.12, n.1, pp.115-125.

CRUZ,O.G., 1996, **Homicídios no Estado do Rio de Janeiro: análise da distribuição espacial e sua evolução.** Dissertação de mestrado, Faculdade de Saúde Pública/USP, São Paulo, SP, Brasil.

D’ÓRSI,E., 1996, **Perfil de nascimentos e condições socioeconômicas no Município do Rio de Janeiro: uma análise espacial**. Dissertação de mestrado, Escola Nacional de Saúde Pública/FIOCRUZ, Rio de Janeiro, RJ, Brasil.

DUCHIADE,M.P., 1991, **Mortalidade infantil por pneumonia na Região Metropolitana do Rio de janeiro, 1976-1986**. Dissertação de mestrado, Escola Nacional de Saúde Pública/FIOCRUZ, Rio de Janeiro, RJ, Brasil.

DUNN,R., 1987, “Variable-width framed rectangle charts for statistical mapping”, **The American Statistician**, v.41, n.2, pp.153-156.

EAMES,M., BEM-SCHLOMO,T., MARMOT,M.G., 1993, “Social deprivation and premature mortality: regional comparison across England”, **British Medical Journal**, v.307, n.8, pp.1097-1102.

FISKEL,J., 1990, “Risk analysis in the 1990's”, **Risk Analysis**, v.10, n.2, pp.195-196.

FRANCIS,A.M., SCHNEIDER,J.B., 1984, “Using computer graphics to map origin-destination data describing health care delivery system”, **Social Science and Medicine**, v.18, pp.405-420.

FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, DIRETORIA DE GEOCIÊNCIAS, 1990, **X Recenceamento geral do Brasil. Questionário Básico**, Rio de Janeiro, FIBGE.

FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA., 1993, **Censo Demográfico 1991 - Resultado do universo relativo às características da população e dos domicílios**, Rio de Janeiro, FIBGE.

GESLER,W., 1986, “The uses of spatial analysis in medical geography: a review”, **Social Science and Medicine**, v.23, pp.936-773.

GLASER,S.L., 1990, “Spatial clustering of Hodkin's diseases in San Francisco Bay area”, **Am. J. Epidem.**, v.132, suppl., pp.167-177.

GOLDBERG,M., DAB,W.,1987, “Complex indexes for measuring a complex phenomenon”. In: **Measurement in health promotion and protection**, nº 22, Regional Publications, WHO/IEA,

GORDON,D., 1995, “Census bases deprivation indices: their weighting and validation”, **Journal of Epidemiology and Community Health**, v.49, suppl.2, pp.S39-S44.

HARTINGAN, JA, 1975, **Clustering algorithms**, 1 ed., New York, Willey.

HASTIE,T.J., TIBSHIRANI,R.J., 1990, **Generalized additive model**, 1 ed., London, Chapman and Hill.

HAYES,M.V., 1992, “On the epistemology of risk: language, logic and social science”, **Social Science and Medicine**, v.35, pp.401-407.

HILLS,M., ALEXANDER,F., 1989, “Statistical methods used in assessing the risk of disease near a source of possible environmental pollution: a review”, **J.R.Stat.Soc.A**, v.152, part.3, pp.353-363.

HUEL, G., DERRIENNIC, F., DUCIMETIERE, P., LAZAR, P., 1978a, “Dureté de l'eau et mortalité cardiovasculaire. Analyse critique des arguments de pathologie géographique”, **Rev. Epidém. et Santé Publ.**, v.26, pp.349-359.

HUEL, G., THOMAZEAU, R., DERRIENNIC, F., LAZAR, P., 1978b, “Dureté de l'eau et mortalité cardiovasculaire. analyse portant sur 947 communes alsaciennes”, **Rev. Epidém. et Santé Publ.**, v.26, pp.381-390.

ISAAKS,E.H., SHRIVASTAVA,R.M., 1989, **Applied Geostatistics**. 1 ed., New York, Oxford University Press.

KNOX,E.G., 1964, “The detection of space-time interactions”, **Appl. Statist.**, v.13, n.1, pp.25-29.

LEBART,L., MORINEAU,A., FÉNELON,J.-P., 1986, **Traitement de données statistiques - méthodes et programmes**. 2 ed, Paris, Dunod.

LILIENFELD,A., 1976, **Foundations of epidemiology**, New York, Oxford University Press.

MANTEL,K.G., 1967, “The detection of disease clustering and a generalized regression approach”, **Cancer Res.**, v.27, pp.209-220.

MARSHALL,R.J., 1991, “A review of methods for the statistical analysis of spatial patterns of disease”, **J.R.Stat.Soc.A**, v.154, part.3, pp.421-241.

MINISTÉRIO DA SAÚDE/OPAS, 1983, **Manual sobre enfoque de risco na saúde materno-infantil**. 1 ed., vols 1,2,3, Brasília, Ministério da Saúde.

ODEN, N., 1995, “Adjusting Moran’s I for population density”, **Statistics in Medicine**, v.4, n.1, pp.17-26.

OPENSHAW,S., CRAFT,A.W., CHARLTON,M., BIRCH,J.M., 1988, “Investigation of leukaemia clusters by use of a geographical analysis machine”, **Lancet**, 1988:272-273.

POOLE,C., 1994, “Editorial: Ecologic analysis as outlook and method”, **American Journal of Public Health**, v.84, n.5, pp.715-716.

RODRIGUES,L.C., MARSHALL,T., MURPHY,M., OSMOND,C., 1992, “Space time clustering of births in SIDS: do perinatal infections play a role?”, **International Journal of Epidemiology**, v.4, pp.714-719.

ROSE,G, 1985, “Indivíduos enfermos y poblaciones enfermas”, **Bol. Epidemiológico OPAS**, v.6, n.3, pp.1-8.

ROSS A., SCOTT D, 1990, “Point pattern analysis of the spatial proximity of residences prior to diagnosis of persons with Hodgkin's disease”, **Am. J. Epidem**., v.132, suppl., pp.53-61.

SMALLMAN-RAYNOR,M., CLIFF,A., 1991, “The spread of human immunodeficiency virus type 2 into Europe: a geographical analysis”, **International Journal of Epidemiology**, v.20, n.2, pp.480-489.

STATISTICAL SCIENCES, INC., 1993, **S-Plus for windows user’s manual. version 3.1**, Seattle, Statistical Sciences, Inc.

STIMSON,R.J., 1980, “Spatial aspects of epidemiological phenomena and of the provision and utilization of health care services in Australia: a review of methodological problems and empirical analysis”, **Environment and Planning A**, v.12, pp.881-907.

SUSSER,M., 1994, “The logic in ecological: I. The logic of analysis”,. **American Journal of Public Health**, v.84, n.5, pp.825-829.

TUKEY,J.W., 1990, “Data-based graphics: visual display in the decades to come”, **Statistical Science**, v.5, pp.327-339.

VERHASSELT,Y., MANSOURIAN,B., 1991, “Método para la classificación de los países de acuerdo con sus indicadores de salud”, **Boletín de la Oficina Sanitária Panamericana**, v.110, pp.319-323.

ANEXO 1

Método multivariado de classificação socioeconômica de microáreas urbanas - os setores censitários da Região Metropolitana do Rio de Janeiro

Artigo submetido

ANEXO 2

Spatial partitioning using multivariate cluster analysis and a contiguity algorithm

Artigo publicado