LUIZ MAX FAGUNDES DE CARVALHO

APLICAÇÃO DE UM MÉTODO DE REGIONALIZAÇÃO AO ESTUDO DA DENGUE NO MUNICÍPIO DO RIO DE JANEIRO

Trabalho de Monografia a ser apresentado ao Instituto de Microbiologia Paulo de Góes, da Universidade Federal do Rio de Janeiro, como pré-requisito para a obtenção do grau de Bacharel em Ciências Biológicas: Microbiologia e Imunologia.

INSTITUTO DE MICROBIOLOGIA PAULO DE GÓES
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
RIO DE JANEIRO
MARÇO/ 2013

Sumário

Trabalho realizado no Programa de Computação Científica (PROCC/FIOCRUZ), da Fundação Oswaldo Cruz, sob a orientação do Dr. Oswaldo Gonçalves Cruz.

FICHA CATALOGRÁFICA

Carvalho, Luiz Max Fagundes.

Aplicação de um Método de Regionalização ao Estudo da Dengue no Município do Rio de Janeiro 72p. Monografia [Bacharelado em Ciências Biológicas: Microbiologia e Imunologia] Universidade Federal do Rio de Janeiro/ Instituto de Microbiologia Paulo de Góes, 2013.
Orientador: Oswaldo Gonçalves Cruz
Referências bibliográficas: 53-60.
1. Dengue 2. Epidemiologia 3. Estatística Espacial
I. Oswaldo Gonçalves Cruz (Orientador). II. UFRJ. Instituto de Microbiologia Paulo de Góes, Bacharelado em Ciências Biológicas: Microbiologia e Imunologia. III. Aplicação de um Método de Regionalização ao Estudo da Dengue no Município do Rio de Janeiro.

FOLHA DE APROVAÇÃO

LUIZ MAX FAGUNDES DE CARVALHO

APLICAÇÃO DE UM MÉTODO DE REGIONALIZAÇÃO AO ESTUDO DA DENGUE NO MUNICÍPIO DO RIO DE JANEIRO.

Rio de Janeiro, 04 de Março de 2013

Prof. Fernando Portela Câmara, Doutor, IMPG/UFRJ

Prof. Davis Fernandes Ferreira, Doutor, IMPG/UFRJ

Luis Paulo Vieira Braga, DG/UFRJ

Agradecimentos

Muitas são as pessoas que merecem ser agradecidas por este trabalho: as faxineiras, os técnicos, os cozinheiros, os professores, e, em última análise, cada brasileira ou brasileiro que trabalha de sol a sol para mover este país continental para frente, sustentando a existência de um ensino superior público de qualidade e gratuito com seus impostos. Há, no entanto, um grupo mais próximo, a quem agradeço em ordem cronológica:

Aos meus pais, por me manterem saudável, seco, quentinho e alimentado para que eu pudesse crescer e me dedicar ao que mais gosto: fazer e ensinar ciência;

Ao Mestre, Professor Doutor Fernando Portela Câmara, por ter me iniciado nos caminhos na epidemiologia quantitativa e da biomatemática e ter, mais que me ensinado ciência, me ensinado como *fazer* ciência.

A Adriana Fagundes Gomes, por ter sido sempre um ouvido atento para as minhas divagações e ter "quebrado o meu galho" por diversas vezes;

Ao Professor Doutor Luis Paulo Vieira Braga, por ter pacientemente me ensinado geoestatística, campo que aprendi a amar, e me introduzido ao ambiente R, que hoje é minha principal ferramenta de trabalho;

Ao Professor Doutor Davis Fernandes Ferreira, pela paciência dispensada sempre que eu furava um prazo ou esquecia um documento, bem como pelas sugestões muito pertinentes de alguém que entende da biologia da dengue para alguém como eu, que enxerga esta doença por um prisma matemático.

Ao Doutor Oswaldo Gonçalves Cruz, por ter me orientado na elaboração deste trabalho, sendo orientador paciente e compreensivo, e me ensinando mais sobre R e Linux do que eu pensei ser possível.

I- INTRODUÇÃO

1.1. A Doença

A dengue é uma doença febre aguda, cujo agente etiológico é o vírus da dengue (DENV), um arbovirus (**Ar**tropod **b**orn **virus**) da família *Flaviviridae* e do gênero Flavivirus (WILDER-SMITH *et al.*, 2010), transmitida pela picada da fêmea dos culicídios *Aedes albopictus e Ae. aegypti*. É hoje a arbovirose mais importante no mundo, causando entre 50 e 100 milhões de infecções todos os anos (GUZMAN *et al.*, 2010).

O contágio acontece quando um indivíduo susceptível é picado por um mosquito infectado. Cerca de 4 a 7 dias após a picada (chamado período de incubação intrínseco) inicia-se o quadro febril e surgem os primeiros sintomas. Os sintomas mais frequentes incluem cefaleia, artralgia, mialgia, leucopenia e exantema. Seu espectro clínico é amplo (CHRISPAL, A *et al.*, 2010) e inclui febre indiferenciada e oligossintomática; a febre dengue (DF), que é a forma clássica da doença e apresenta evolução geralmente benigna; e a febre hemorrágica da dengue (FHD), caracterizada por extravasamento plasmático e queda de contagem de plaquetas e que pode se complicar levando à síndrome do choque da dengue (SCD), quadro frequentemente letal (WHO, 2002; WHO, 2009).

A classificação proposta pela Organização Mundial de Saúde (OMS) (WHO, 2009) divide os casos de dengue em dois tipos: dengue sem gravidade (FD), e dengue com gravidade (FHD e SCD). A FHD corresponde à cerca de 1% do número de casos e, caso não tratada, pode evoluir para o óbito em 10 a 20% dos casos. O quadro ocorre devido ao aumento da permeabilidade vascular e do extravasamento plasmático e pode evoluir para a SCD, quadro ainda mais grave caracterizado pela queda da pressão arterial e da irrigação dos órgãos, complicação que frequentemente leva ao óbito.

1.2. Epidemiologia

Os primeiros relatos de uma doença febril semelhante à dengue datam do final do século XVIII, sendo de 1779 na ilha de Java, 1780 no estado americano da Filadélfia e de 1784 nas cidades de Cádiz e Sevilha na Espanha, denominada de “quenturas benignas de Sevilha” (PONS, 1960).

Acredita-se que o vírus da dengue evoluiu inicialmente na Ásia, espalhando-se posteriormente ao redor do mundo. O começo do século XIX foi marcado por grandes epidemias de uma doença febril compatível com a dengue, na Índia, Egito, e Peru. Na metade do século, em 1848, ocorreu uma segunda pandemia que se estendeu para Cuba, no Caribe, e nos estados da Luisiana, Florida, Carolina do Sul e Texas, nos EUA (GUBLER, 1997), e nos anos seguintes em diversos outros locais do mundo, como Panamá (KAY, 1984), e Grécia (BRÉS, 1979).

Cem anos mais tarde, logo após a Segunda Guerra Mundial, uma forma hemorrágica de dengue emergiu em diversos países do Sudeste Asiático, como Filipinas, Tailândia, Vietnã, Cingapura e Indonésia e persiste nesses países de forma endêmica. Até o começo dos anos 2000, a dengue afetava principalmente crianças (ENDY *et al*, 2002; TORRES, 2005), porém já se detecta um aumento da idade média dos pacientes (CUMMINGS *et al*, 2009).

A dengue ocorre em países da faixa tropical e subtropical do planeta, principalmente em áreas urbanas (REY, 2008). Estima-se que cerca de 2,5 bilhões de pessoas (2/5 da população mundial) (Figura 1) vivem em áreas de risco para a dengue. Segundo o World Health Assembly, a dengue constitui uma enfermidade capaz de representar emergência de saúde pública de caráter internacional devido a seu rápido alastramento que ultrapassa as fronteiras nacionais (WHO, 2002; WHO, 2009).

Figura 1. Mapa ilustrando a parcela mundial em áreas de risco para transmissão de dengue.

1.3. Dengue no Brasil e na cidade do Rio de Janeiro

No começo da década de 1980, entre os anos de 1981 e 1982 ocorreram os primeiros casos da doença no Brasil com a presença dos sorotipos DEN-1 e DEN-4 na cidade de Boa Vista-RR. Em 1986 o sorotipo DEN-1 emergiu em diversas cidades do Brasil (PONTES *et al.*, 1994), possivelmente oriundo do município do Rio de Janeiro (SCHATZMAYR *et al.*, 1986). Este mesmo município experimentou, em 1990, com a introdução do DENV-2 – também ocorrida no Nordeste-, os primeiros casos da forma hemorrágica da doença (NOGUEIRA *et al.*, 1990). A introdução do DEN-3 no município de Rio de Janeiro ocorreu no ano 2000, sendo este o este o sorotipo dominante na maior epidemia já experimentada pelo município, ocorrida no período 2001-2002 (NOGUEIRA *et al.*, 2005; CÂMARA *et*

al., 2007, CÂMARA *et al.*, 2009). Desde então a doença se alastrou pelo país e a maior parte dos estados da federação está sob risco alto ou muito alto de ocorrência da doença (Figura 2). Em julho de 2010 o sorotipo DENV-4 foi isolado no Brasil no estado de Roraima, provavelmente vindo da Venezuela, onde o DENV-4 circula de forma endêmica há muitos anos (Nota técnica SVS/MS, disponível em http://portal.saude.gov.br/portal/.cfm?id_area=1498). O Estado do Rio de Janeiro foi o primeiro a reportar casos de infecção pelo DENV-4 (NOGUEIRA & EPPINGHAUS, 2011) configurando-se o quadro de hiperendemicidade observado nos países do Sudeste Asiático, criando novos desafios para o controle da doença.

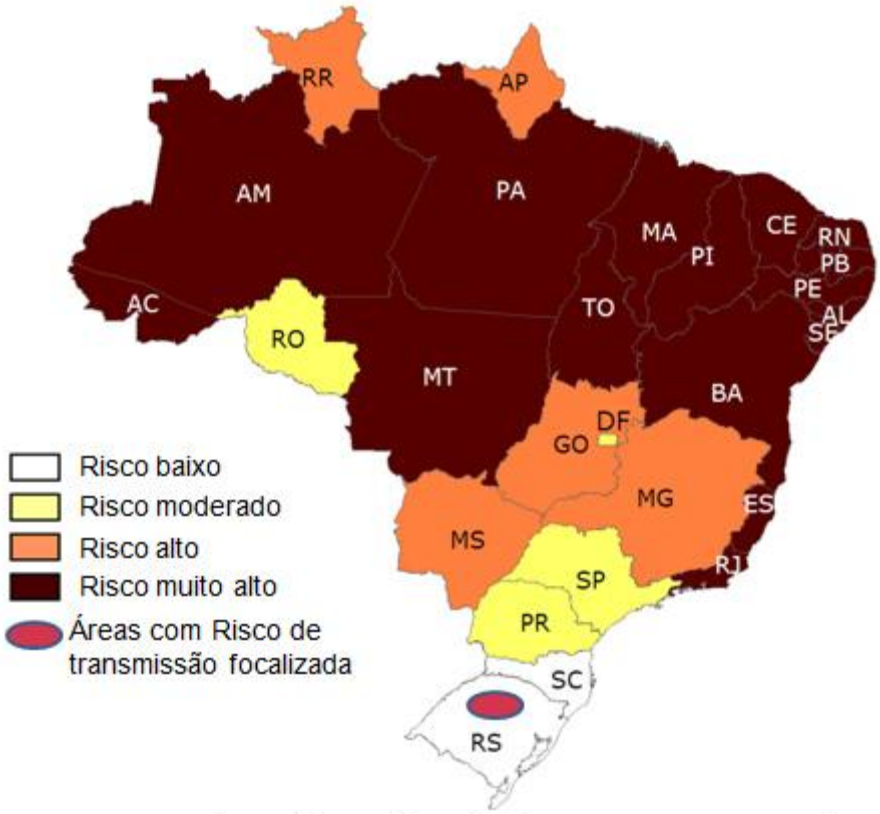


Figura 2. Mapa de risco para dengue no Brasil (ano de referência: 2011; fonte: SVS/MS)

1.4. Análise Espacial de Doenças

Análise espacial é o estudo quantitativo de fenômenos que são localizados no espaço. Compreender a distribuição espacial de doenças é um dos objetivos importantes da epidemiologia moderna e tem sido campo de intensa pesquisa nos últimos anos. O conhecimento do padrão espacial dos agravos pode ter influência decisiva na descoberta de fontes de exposição e na identificação de áreas de alto risco, sendo um elemento de suporte à decisão quando da elaboração de políticas públicas de saúde.

No Brasil, as primeiras aplicações tem suas raízes na década de 1950, com o uso de grandes computadores para estudos de planejamento urbano e análise ambiental (SANTOS & BARCELLOS, 2006). O progresso na aplicação de técnicas de geoprocessamento e análise espacial em saúde é devedor principalmente do desenvolvimento e difusão de softwares de fácil manipulação e do acesso a computadores de alto rendimento a baixo custo. Tais condições permitiram ampliar o número de usuários dos chamados Sistemas de Informações Geográficas (SIGs), apontados como os principais instrumentos de integração de informações epidemiológicas e variáveis climáticas e socioambientais e muito utilizados nos chamados estudos ecológicos[1]. A disponibilidade de dados, softwares e computadores criou também uma demanda por novos métodos de análise, capazes de se aliar às SIGs para propiciar as ferramentas necessárias aos diversos problemas que surgem em Epidemiologia Geográfica.

Os estudos em Epidemiologia Geográfica podem ser classificados em três abordagens básicas, de acordo com o tipo e escala dos dados analisados (SANTOS & SOUZA, 2007): Análise de Pontos, Geoestatística e Análise de Áreas. Para as duas primeiras referimos o leitor a DRUCK *et al* (2004) e SANTOS & SOUZA (2007). A última abordagem, sendo tema deste trabalho, será apresentada em mais detalhe na próxima seção.

1.5. Análise Espacial de Áreas

A análise de áreas trata basicamente do estudo quantidades (variáveis e covariáveis) associadas a entidades discretas (as áreas). Frequentemente os dados disponíveis em

[1] Chamamos estudos ecológicos àqueles em que a unidade de análise é uma população em vez do indivíduo e se procura relacionar fontes de exposição ambiental a medidas de risco. Estes estudos, embora limitados em vários aspectos, são importantes em epidemiologia por serem em geral mais baratos e rápidos.

grandes bases de dados secundários encontram-se agregados por unidades político-administrativas e setores censitários – as áreas. Neste contexto é que se inserem as ferramentas de análise espacial de áreas. É importante, no entanto considerar a escala de agregação dos dados de modo a entender a informação neles contida. Por exemplo, se agregarmos o número de casos de uma doença por país, teremos uma figura global desta doença, mas não seremos capazes de observar sua dinâmica local.

Um elemento importante a ser considerado na análise de áreas é a **instabilidade em pequenas áreas**. Este fenômeno ocorre principalmente quando utilizamos taxas brutas como incidência e mortalidade para eventos raros em populações pequenas. Numa população pequena, pequenas flutuações no número de casos podem produzir variações abruptas nas estimativas das taxas brutas. Por este motivo é comum aplicarmos técnicas de suavização espacial, de modo a corrigir as estimativas para uma área utilizando a informação dos seus **vizinhos**. Métodos comuns incluem as médias móveis espaciais e os estimadores empíricos bayesianos local e global. Para mais detalhes o leitor é referido a (GOSH & RAO, 1994 e SANTOS & SOUZA, 2007).

Neste contexto é necessário levar em conta o relacionamento entre as áreas, comumente expresso sob a forma de uma **matriz de vizinhança**. A matriz de vizinhança é uma matriz de conectividade que indica a relação espacial de cada área com as demais. Comumente usamos como matriz de vizinhança a **matriz de contiguidade** [2] (Figura 3) mas pode-se usar matrizes de distância geográfica, fluxo populacional ou que simbolizem qualquer relacionamento de interesse entre as áreas (SANTOS & SOUZA, 2007).

A análise espacial de áreas está entre as técnicas de epidemiologia geográfica mais comuns, muito devido à larga utilização de estudos ecológicos. A pesquisa em dengue é um exemplo onde essas técnicas tem sido aplicadas na tentativa de associar fatores socioambientais à incidência da doença.

[2] A matriz de contiguidade ou adjacência é uma matriz binária que assinala 1 (um) para áreas que fazem fronteira entre si e 0 (zero) para aquelas que não. Para uma definição formal, ver a definição de c_{ij} na seção 3.6.2.

MUNICÍPIO	Água Santa	Bebedouro	Cacimba	Nascente	Poço
Água Santa	0	1	0	1	1
Bebedouro	1	0	0	1	0
Cacimba	0	0	0	0	1
Nascente	1	1	0	0	0
Poço	1	0	1	0	0

Figura 3. Exemplo de matriz de vizinhança. Neste caso temos uma matriz de contiguidade ou adjacência. (Adaptado de SANTOS & SOUZA, 2007).

1.6. Análise Espacial da Dengue

A análise espacial da dengue vem crescendo nos últimos anos no Brasil e no mundo. Sendo transmitida por um mosquito vetor, acredita-se que a incidência da doença dependa de variáveis ambientais como temperatura e umidade bem como das condições socioeconômicas das populações sob risco (FLAUZINO, SOUZA-SANTOS, & OLIVEIRA, 2009; GOMES, 2011). Os estudos sobre a dengue em geral procuram associações entre as diferentes covariáveis ambientais e a doença no espaço, em geral representado por unidades distritais (FLAUZINO, SOUZA-SANTOS, & OLIVEIRA, 2009).

Vários aspectos da dengue no espaço foram investigados, entre eles a associação da incidência da doença com variáveis ambientais (GOMES, 2011; TEIXEIRA & CRUZ, 2011) e socioeconômicas (TEIXEIRA & MEDRONHO, 2008) e a soroprevalência (HONÓRIO *et al*, 2009). A maior parte dos estudos é do tipo ecológico, investigando associações estatísticas entre a dengue e as diversas covariáveis ao nível de setores censitários, bairros. Os resultados são contraditórios quanto à influência de fatores socioeconômicos específicos, mas o consenso é de que as condições de vida da

população tem alguma influência sobre a incidência da dengue, muito embora nem sempre possamos detectar sinal estatístico significativo. Um estudo recente mostrou que a desigualdade social, expressa pelo índice de Gini[3], estava associada à incidência de dengue. No âmbito dos fatores ambientais, a temperatura, em especial a temperatura mínima, tem sido apontada como fator importante para a proliferação do vetor e estabelecimento de epidemia (CÂMARA *et al.*, 2009; GOMES, 2011). Surpreendentemente, os índices de infestação vetorial (Índice de Breteau, Índice de Infestação Predial e LIRAa) não tem sido associados com a incidência da doença, presumivelmente por não refletirem a real densidade populacional do mosquito (TEIXEIRA & CRUZ, 2011). A discordância entre os estudos sobre a associação da dengue e variáveis que sabemos importantes para o ciclo biológico da doença pode se dever a fatores confundidores como a instabilidade em pequenas áreas e a variabilidade intrínseca de certos processos, como a dinâmica da população do vetor.

1.7. Regionalização

Muitas vezes queremos ter uma visão global do padrão espacial de um fenômeno, de forma a sumarizar a informação espacial e eliminar o ruído que porventura possa estar presente. A **regionalização** (em inglês *regionalization* ou *region-building*) é a criação de agrupamentos de áreas contíguas no espaço, as *regiões*, a partir de algum critério de agrupamento. Esta abordagem é útil, por exemplo, quando queremos criar mesorregiões de saúde ou novas divisões administrativas com base em critérios objetivos (PAIVA, ALONSO & TARTARUGA, 2010). Nesse sentido a escolha informada de um nível de agregação adequado pode revelar padrões espaciais não presentes no conjunto de dados original (OPENSHAW & ALVANIDES, 1999).

1.7.1. Definição

Seja **R** uma região subdividida em N áreas A_i tal que $\mathbf{R} = \{A_1, A_2, A_3, ..., A_N\}$, em que a cada A_i está associado um vetor de atributos $\mathbf{v}_i$. Suponha que se deseja agrupar

[3] O indice de Gini é usado para medir a disparidade de renda (desigualdade) numa população. A medida, proposta pelo sociólogo e estatístico Conrado Gini (1884-1965) em 1912, varia de zero (igualdade perfeita) a um (total desigualdade), sendo muitas vezes expressa como um percentual.

essas áreas de tal forma a gerar K regiões em que as áreas sejam contíguas no espaço, isto é, que compartilhem fronteiras. A regionalização consiste em encontrar, dentro do universo de soluções dado por $N!/(N-K)!\,K!$, qual (i) atende à restrição de contiguidade e (ii) minimiza algum critério de dissimilaridade[4] operando em $\mathbf{v}_i$, de forma a criar áreas o mais homogêneas possível.

1.7.2. Abordagens Estatísticas para a regionalização

A regionalização é um problema combinatório do tipo *NP-hard* (ASSUNÇÃO *et al.*, 2006), tornando a exploração exaustiva do espaço de soluções impraticável mesmo para valores modestos de N e/ou K. Desta maneira, ao longo dos anos foram propostas diversas abordagens heurísticas para o problema, examinadas em mais detalhe a seguir. As abordagens estatísticas para o problema de regionalização podem ser enquadradas em quatro tipos básicos:

(i) Agrupamento não-espacial com posterior processamento espacial;

(ii) Agrupamento não-espacial incorporando as coordenadas geográficas como variáveis;

(iii) Agrupamento espacial a partir de um particionamento inicial aleatório (ad hoc);

(iv) Agrupamento espacial a partir de grafos de contiguidade.

No primeiro tipo temos o uso de técnicas clássicas de agrupamento não-hierárquico, que são posteriormente refinadas pelo "prunning" dos conglomerados obtidos para que estes atendam à restrição espacial de contiguidade (FOVELL & FOVELL, 1993; CARVALHO, CRUZ & NOBRE, 1996). Essa abordagem em dois passos, no entanto, não captura a estrutura de adjacência espacial corretamente, apresentando poder limitado para capturar padrões espaciais (ASSUNÇÃO *et al*, 2006).

Já para o segundo tipo, o componente espacial é incorporado de forma não restritiva, tratando-se as coordenadas geográficas das áreas como atributos adicionais (HAINING, WISE & MA, 2000). Muito embora seja possível obter bons resultados

[4] Podem, ainda, ser adicionadas restrições adicionais ao problema, como por exemplo, um número mínimo e/ou máximo de habitantes por agrupamento.

com essa metodologia, é preciso um grande número análises e ajustes até que se encontre a combinação correta de pesos para produzir resultados coerentes, isto é, agregados espacialmente contíguos. Além disso, essa metodologia se baseia em critérios de optimalidade conceitualmente irreconciliáveis, em que diferentes aspectos conflitantes devem ser maximizados. Para uma discussão detalhada o leitor é referido a OPENSHAW, ALVANIDES & WHALLEY, 1998.

No terceiro tipo de regionalização o componente geográfico é considerado explicitamente como condição restritiva para a formação dos aglomerados. Os grupos são dinamicamente formados e refinados de forma que apenas áreas contíguas sejam incorporadas, através de tentativa e erro. Um exemplo desta técnica é o Automated Zoning Procedure (AZP), proposto pelo Professor Stan Openshaw em 1977 (OPENSHAW, 1977; OPENSHAW & RAO, 1995). No AZP, entre os métodos de otimização do processo de tentativa e erro que podem ser utilizados incluem otimização de Monte Carlo (OPENSHAW, 1977) e "simulated annealing" (OPENSHAW e RAO, 1995; ALVANIDES, OPENSHAW & REES, 2002). Por se basear em técnicas complexas de otimização, o AZP se torna computacionalmente intensivo, e requer grandes recursos computacionais.

A quarta abordagem trata de transformar a regionalização num problema de particionamento de grafos. Cada área A_i é transformada no vértice de um grafo de contiguidade, onde as arestas conectam regiões contíguas no espaço, isto é, que compartilham fronteiras. Este grafo é então utilizado como base para a regionalização, através do seu particionamento para obtenção de subgrafos conexos. As diferenças entre os métodos desta última abordagem residem na atualização ou não da estrutura de vizinhança durante o processo de particionamento. Desta maneira, dividimos esta última classe de métodos em métodos de regionalização com estrutura de contiguidade dinâmica (GUO, 2008) e não dinâmica (ASSUNÇÃO *et al*, 2006). Neste trabalho utilizaremos uma técnica que se enquadra nesta última classe, a SKATER.

1.7.3. SKATER

A SKATER (do inglês, **S**patial '**K**'luster Analysis by **T**ree **E**dge **R**emoval) é um algoritmo que transforma a regionalização num problema de partição de grafos. Cada área

A_i é transformada num vértice do grafo G. As arestas, inicialmente, correspondem a todas as conexões de primeira ordem (contiguidade) entre os vértices. O primeiro problema combinatório a ser resolvido é encontrar a árvore geradora mínima (MST, na sigla em inglês) do grafo **G** de contiguidade de R. Tal tarefa é levada a cabo através do algoritmo de Prim (JUNGNICKEL, 1999). Um exemplo do algoritmo de Prim é mostrado no Apêndice 1 (seção 8.1), Figura A1.

Após a obtenção da MST, passa-se à tarefa de encontrar o melhor particionamento da mesma em K subgrafos, de acordo com uma função objetivo arbitrária. Em geral, essa função objetivo visa a minimizar a heterogeneidade dos grupos formados. Para o presente estudo, a função objetivo adotada será a soma de quadrados entre grupos de acordo com os atributos em $\mathbf{v}_i$. O particionamento de **G** é realizado através da exploração heurística do espaço de soluções, selecionando-se os particionamentos que minimizem a soma de quadrados entre os vértices dos subgrafos formados. Como em qualquer método de otimização, há o risco de se atingir máximos locais e desta forma obter soluções sub-ótimas. Na implementação original do método, Assunção *et al* (2006) atacam o problema utilizando duas funções-objetivo, f_1 e f_2, em conjunto. Enquanto f_1 avalia o critério de homogeneidade de uma solução candidata S_l[5] - nesse caso a soma de quadrados -, f_2 examina cinco soluções candidatas na vizinhança de S_l, e seleciona aquela com maior valor de f_1. Desta maneira, previne-se a aceitação de particionamentos muito desbalanceados e heterogêneos e diminui-se o risco de se obter uma solução que corresponde a um máximo local ao invés do máximo global desejado. Um esquema do algoritmo de particionamento da MST é apresentado no Apêndice, Figura A2.

A escolha do vértice de G a ser utilizado como ponto de partida tem grande influência no número de iterações do algoritmo necessárias para que se obtenha uma solução. Para ilustrar as propriedades heurísticas da SKATER, a Tabela 1 mostra o esforço computacional demandado para realizar a SKATER, medido em tempo computacional, para vários números de áreas e variáveis. Além disso, apresentamos o coeficiente de variação do tempo computacional como uma medida da influência do ponto de partida sobre o número de iterações necessário.

[5] O índice "l" de S_l refere-se à solução candidata obtida removendo-se o vértice l de **G**.

Um resumo das etapas da SKATER é apresentado na Figura 4, onde mostramos a regionalização do município de Belo Horizonte baseada em 5 variáveis socioeconômicas. Os dados e o código necessários à confecção da figura podem ser encontrados na documentação da função `skater()` do pacote **spdep** (PEBESMA & BIVAND, 2005) do ambiente R (R CORE TEAM, 2012).

			Tempo de Execução		Coeficiente de Variação (%)	
Mapa	No. Áreas	Variáveis	Média	I.C. 95 %	Média	I.C. 95 %
Belo Horizonte	98	4	$5{,}48 \times 10^{-3}$	$4{,}38\text{-}6.63 \times 10^{-3}$	147	129-170
Niterói	48	15	$2{,}64 \times 10^{-3}$	$8{,}33\text{-}40{,}90 \times 10^{-3}$	254	175-409
Eire	26	9	$1{,}86 \times 10^{-3}$	$0{,}07\text{-}3{,}07 \times 10^{-3}$	302	220-373
Equador	20	4	$1{,}26 \times 10^{-3}$	$1{,}00\text{-}3{,}50 \times 10^{-3}$	292	212-326
Oregon	36	45	$2{,}22 \times 10^{-3}$	$1{,}11\text{-}3{,}61 \times 10^{-3}$	274	200-358
EUA	48	3	$2{,}79 \times 10^{-3}$	$1{,}45\text{-}4{,}37 \times 10^{-3}$	235	178-311
Brasil (UF)	27	2	$1{,}80 \times 10^{-3}$	$0{,}74\text{-}2{,}96 \times 10^{-3}$	302	223-381
Carolina do Norte (EUA)	100	6	$6{,}82 \times 10^{-3}$	$5{,}30\text{-}8{,}20 \times 10^{-3}$	128	105-150
Auckland (Nova Zelândia)	167	2	$10{,}38 \times 10^{-3}$	$8{,}98\text{-}12{,}39 \times 10^{-3}$	82	66-95
Rio de Janeiro (Estado)	92	18	$5{,}87 \times 10^{-3}$	$4{,}23\text{-}7{,}71 \times 10^{-3}$	141	110-171
Rio de Janeiro (Municipio)	157	12	$14{,}17 \times 10^{-3}$	$12{,}86\text{-}15{,}47 \times 10^{-3}$	57	49-66
Minas Gerais (Estado)	853	12	$6{,}27 \times 10^{-3}$	$5{,}41\text{-}6{,}96 \times 10^{-3}$	18	16-19
São Paulo (Estado)	645	31	$4{,}89 \times 10^{-3}$	$4{,}65\text{-}5{,}07 \times 10^{-3}$	23	21-24

Tabela 1. Tempo computacional da SKATER em diversas condições de número de áreas e variáveis. Intervalos de Confiança obtidos a partir de 500 réplicas de cada experimento. As análises foram feitas utilizando o ambiente R versão 2.13 numa máquina Core i5 4 GB DDR3.

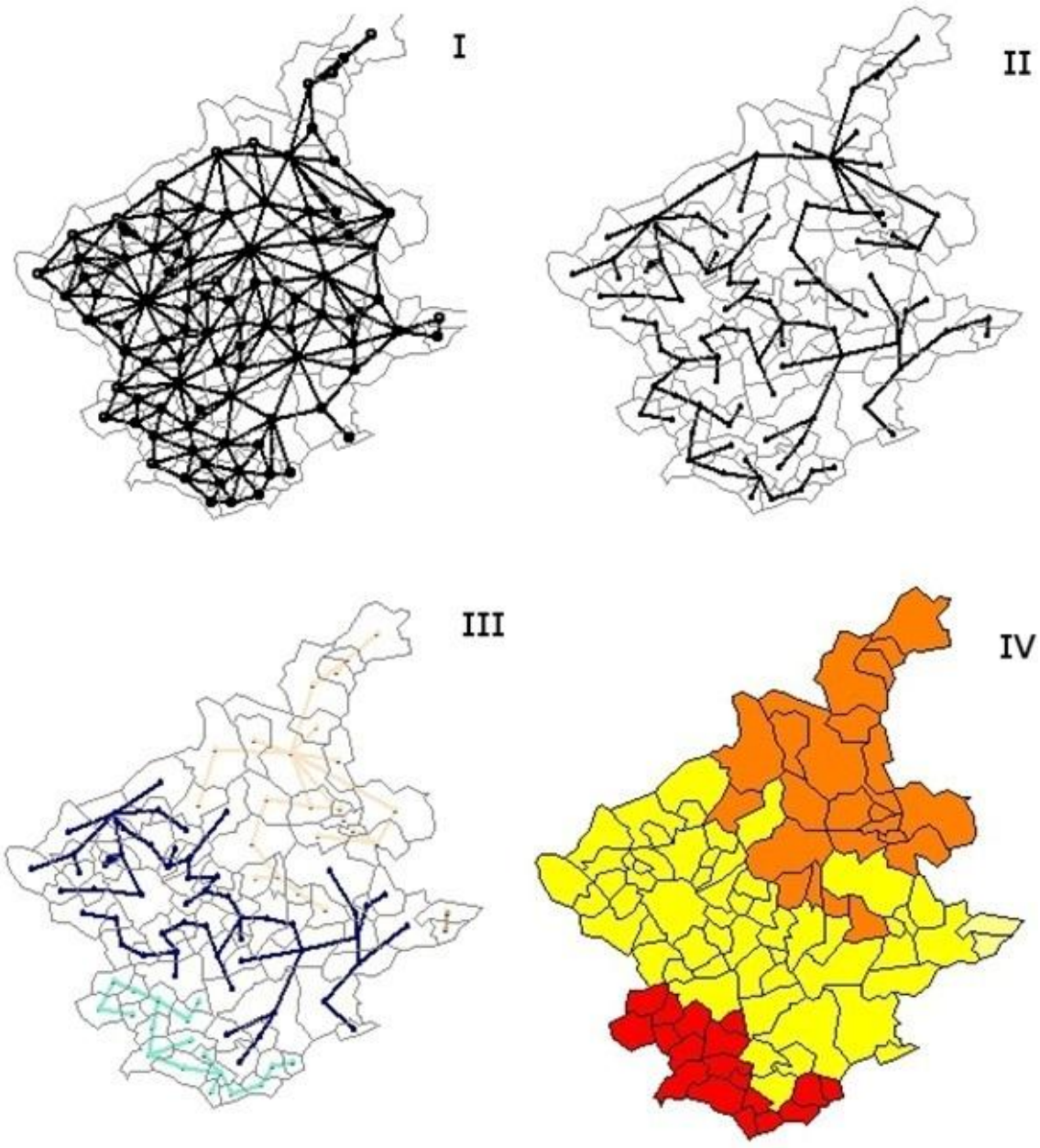


Figura 4. Representação das etapas da SKATER: (I) obtenção do grafo de contiguidade; (II) obtenção da MST; (III) particionamento da MST; (IV) resultado final da regionalização. Dados e mapas derivados do exemplo da função `skater()` do pacote **spdep** do ambiente R.

II – OBJETIVOS

2.1. Motivação

O entendimento do padrão espacial da dengue pode ser uma ferramenta útil à elaboração de políticas públicas de saúde mais efetivas por permitir que as autoridades identifiquem áreas de maior risco com precisão e exerçam ações localizadas, reduzindo-se a carga da doença e os custos financeiros. Neste sentido, a regionalização, ao apontar áreas homogêneas para a incidência da doença, pode contribuir para a elucidação do padrão global da dengue e para a elaboração de novas mesorregiões de saúde. Por incorporar automaticamente restrições como número mínimo de habitantes por região, a SKATER se mostra uma metodologia promissora para a obtenção de áreas balanceadas, que sejam úteis para este propósito.

2.2. Objetivo Geral:

Obter áreas homogêneas para a incidência da dengue no município do Rio de Janeiro, utilizando dados de incidência, variáveis socioeconômicas e ambientais. Estudar regionalizações com diversos critérios de população mínima por região e diversos números de regiões (K).

2.3. Objetivos Específicos

- Aplicar a metodologia SKATER aos dados de dengue, índices de infestação vetorial e índice de desenvolvimento social (IDS) no Município do Rio de Janeiro;
- Estudar as regiões obtidas utilizando diferentes variáveis-base para a SKATER;
- Estudar a regionalização para diferentes números de regiões em anos pré-epidêmicos (2001 e 2007) e epidêmicos (2002 e 2008).
- Realizar a regionalização para diversos números de regiões (K), definidos de acordo com vários critérios de população mínima por região.

III- METODOLOGIA

3.1. Área de estudo

O município do Rio de Janeiro está localizado na Região Sudeste do Brasil, situado a 23º 04' 10" de Latitude Sul e 43º 47' 40" de Longitude Oeste, com uma população estimada de 6.093.472 habitantes, possui área total de 1.224,56 Km² (IBGE, http://www.ibge.gov.br/; Armazém de dados, http:// www.armazemdedados.rio.rj.gov.br/). O clima é tropical atlântico com baixa amplitude térmica anual e temperatura média de 23 ºC. Climaticamente, o ano pode ser dividido em uma estação quente e chuvosa e outra com temperaturas mais amenas. O volume pluviométrico acumulado anual em torno de 1000 mm, com as chuvas concentradas nos meses de dezembro a março.

A cidade do Rio de Janeiro atualmente é divida em 5 áreas de planejamento, 19 subprefeituras, 38 regiões administrativas, e 160 bairros.

3.2. Fonte dos dados

Como o estudo cobre um período temporal em que aconteceram mudanças no número de bairros da cidade, decidimos manter a divisão em 157 bairros para tornar os resultados comparáveis em todo o período de estudo. Informações sobre incidência de dengue, índices de infestação vetorial como Indice de Breteau (IB) e Indice de Infestação Predial (IIP) nos bairros do município do Rio de Janeiro para os anos estudados (2001/2002 e 2006/2008) foram conseguidas junto à Secretaria Municipal de Saúde.

3.3. Imputação dos índices de infestação vetorial [Pode inverter com A. exploratória?]

Em virtude da presença de dados faltantes (“NAs”) nos atributos de infestação vetorial para alguns bairros, procedemos à imputação, isto é, à atribuição de valores a esses campos faltante através de um modelo estatístico. A imputação tem papel central em ciência estatística, sendo uma das maneiras de se contornar a falta de informação para certos atributos, sem que se perca todos os dados para um determinado indivíduo.

No caso da análise espacial de áreas, podemos utilizar a informação contida na estrutura de vizinhança para fazer "previsões" sobre os valores a serem imputados.

No presente estudo adotamos uma abordagem não paramétrica, em que a média atribuída a uma área A_i é a média das observações de seus vizinhos[6] (ver MARSHALL,1991 para uma abordagem paramétrica).

Para verificar a qualidade da imputação, calculamos o erro quadrático médio (MSE) de imputação, da seguinte maneira: para cada área que possuía um valor para o atributo imputado, realizamos o processo acima e depois calculamos o quadrado da diferença entre o valor obtido e o valor "real" do atributo. A média desses valores para cada variável foi tomada para comparação. O código utilizado para esta etapa está apresentado no Anexo 2 (seção 8.2.2).

3.4. Análise exploratória

A exploração dos dados é etapa importante em qualquer tratamento de dados. Ela permite que se conheça o conjunto de dados analisado, possibilitando a detecção de campos faltantes ou preenchidos de forma incorreta, os tipos de variáveis envolvidos e a variação das observações.

Em particular, o estudo da variabilidade dos dados e a obtenção de medidas de resumo são etapas importantes na tarefa de se analisar um conjunto de dados. Para o presente trabalho, utilizamos, principalmente, duas ferramentas de visualização e resumo dos dados: os diagramas de caixa (*box-plots*) e os mapas coropléticos. Adicionalmente, estudamos a presença de estrutura espacial nos dados utilizando os índices de autocorrelação espacial de Moran e Geary.

3.4.1. Diagramas de caixa (box-plots)

O diagrama de caixa, ou *box-plot*, é um gráfico utilizado para avaliar a distribuição empírica dos dados, mais especificamente os momentos empíricos. O *box-plot* é formado pelo primeiro e terceiro quartil e pela mediana. As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior ao limite inferior e do quartil superior até o maior valor não superior ao limite

[6] Neste trabalho utilizamos o grafo de contiguidade como estrutura de vizinhança. Talvez seja interessante notar que outras estruturas de vizinhança podem ser utilizadas, de forma a melhor descrever a estrutura de autocorreção espacial dos dados.

superior. O gráfico é especialmente útil na exploração visual dos dados em busca de valores aberrantes (*outliers*) (BRAGA, 2010).

Neste trabalho geramos diagramas de caixa para as variáveis de estudo de forma a ganhar informação sobre sua distribuição e variabilidade. Em conjunto com os mapas coropléticos (ver seção 3.4.3), esses gráficos nos permitem estudar a variabilidade espacial do fenômeno sob estudo.

3.4.2. Índices de autocorrelação espacial

Em análise espacial de áreas é comum estarmos interessados em estudar a estrutura de autocorrelação espacial das variáveis analisadas, isto é, a correlação entre uma variável num local com a mesma variável nas áreas vizinhas (SANTOS & SOUZA, 2007). Para tanto, podemos utilizar a estatística I de Moran[7], cuja expressão é da forma:

$$\mathrm{I} = \frac{n}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (1)$$

onde n é o número de áreas, x_i é o valor da variável na i-ésima área, $\overline{x}$ é a média amostral e $S_0 = \sum_i^n \sum_j^n w_{ij}$. Por conveniência, geralmente especificamos w_{ij} de modo que $S_0 = n$ (MORAN, 1950). Sendo uma medida de correlação, a estatística I de Moran assume valores no intervalo [-1,1], sendo valores próximos a -1 indicativos de extrema dissimilaridade espacial e valores próximos de 1 indicativos de alta similaridade espacial.

Outra medida de autocorrelação utilizada neste trabalho foi a estatística de autocorrelação C de Geary[8] (GEARY, 1954), definida como:

$$\mathrm{C} = \frac{(n-1)}{2S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad (2)$$

em que n, x_i, $\overline{x}$ e S_0 são definidos como em (1). Valores de C menores que 1 indicam autocorrelação espacial positiva, enquanto valores acima de 1 apontam autocorrelação

[7] Assim nomeado em homenagem ao seu proponente, o estatístico australiano Patrick Alfred Pierce Moran (1917-1988).
[8] Proposta pelo estatístico irlandês Robert Charles Geary (1896-1983) em 1954.

negativa. De (1) e (2) notamos que os índices de Moran e Geary são inversamente relacionados, sendo este último mais sensível à autocorrelação local (ANSELIN, 1995).

É também possível testar a significância dos índices obtidos, isto é, se existe suporte estatístico para a hipótese de existência de estrutura (autocorrelação) espacial. Para tanto, sob certas condições de regularidade, pode-se utilizar aproximações normais ou métodos de re-amostragem para obter uma distribuição amostral para os índices. Neste trabalho utilizamos esta segunda abordagem, lançando mão das rotinas de Monte Carlo `moran.mc()` e `geary.mc()`, implementadas no pacote **spdep,** para aproximar a distribuição das estatísticas I e C e assim obter p-valores para as estatísticas encontradas.

3.4.3. Mapas coropléticos

No contexto da análise espacial de áreas, normalmente existem variáveis associadas a cada área e as cores, ou tonalidades de cores, são usadas para representar as diferenças entre as áreas. A partir do conhecimento da variação dos dados é possível construir uma escala de cores com número suficiente de intervalos para que se tenha uma perspectiva visual da distribuição espacial da(s) variável(s) em estudo.

No presente estudo, geramos séries de mapas coropléticos para a incidência de dengue no município do Rio de Janeiro, de modo a visualizar o padrão espacial da doença nos anos escolhidos para este estudo. Geramos, da mesma maneira, mapas para as outras variáveis utilizadas no presente estudo.

3.5. Experimentos

Neste trabalho estudamos a regionalização das seguintes variáveis:

- Índice de Desenvolvimento Social (IDS);
- Os índices de infestação vetorial: índice de Breteau (IB);
- A incidência de dengue no município do Rio de Janeiro nos anos de 2001, 2002, 2006 e 2008.

Com vistas a obter regiões mais balanceadas oram realizados experimentos de regionalização para K igual a 6 (seis), 12 (doze) e 30 (trinta) regiões. Para cada um desses conjuntos de experimentos, utilizamos restrições de população mínima por região formada, a saber:

Tabela 2. Valores de restrição de população mínima por região.

Número de Regiões (K)	População Mínima (x1000 habitantes)
6	1000
12	500
30	200

3.6. Medidas de qualidade

No contexto da regionalização, muitas vezes é desejável avaliar a qualidade das divisões obtidas, principalmente quanto ao balanceamento e à homogeneidade das regiões. Neste trabalho apresentamos algumas propostas de medidas de qualidade para avaliar estes dois aspectos. A seguir detalhamos as medidas utilizadas.

3.6.1. Balanceamento

Para avaliar o balanceamento das regiões obtidas, computamos o número de áreas agrupadas em cada região. Para termos uma medida comparável entre regionalizações de mesmo número de regiões (K), calculamos a razão entre a entropia de Shannon da distribuição de áreas por região obtida e aquela de um vetor de mesmo tamanho com distribuição uniforme do número de áreas. A entropia de Shannon de um vetor aleatório discreto X é definida como:

$$H(\boldsymbol{X}) = -\sum_{i=1}^{n} P(x_i)\log\big(P(x_i)\big) \qquad (3)$$

Sendo máxima quando a distribuição de $\boldsymbol{X}$ é uniforme. Dessa maneira, para avaliar o balanceamento criamos um vetor K-dimensional $\boldsymbol{Y}$ tal que:

$$y_i = \left\| \frac{N}{K} \right\| \ \forall i \ \in \{1,2,\dots,K\} \qquad (4)$$

isto é, cada elemento de $\boldsymbol{Y}$ é o menor inteiro da razão entre o número de áreas N e o número de regiões K. Como $\boldsymbol{Y}$ é a divisão mais balanceada possível das N áreas em K regiões, quando calculamos

$$R(X) = \frac{H(\boldsymbol{X})}{H(\boldsymbol{Y})} \qquad (5)$$

temos uma medida do balanceamento da regionalização obtida normalizada pelo máximo teórico. A implementação do cálculo de $R(\boldsymbol{X})$ foi feita por meio da função `entr()`, descrita no anexo 3 (seção 8.3).

Adicionalmente, como um indicador de compactação espacial, calculamos a média do número de vizinhos de uma determinada área que foram agrupados numa mesma região. Esta medida pode ser encarada como um grau médio ($< k >$, na notação da teoria dos Grafos) intra-região, representando o número de vértices conectados a um dado vértice que estão na mesma região que este.

3.6.2. Homogeneidade

Outra medida importante quando se obtêm regiões é avaliar a homogeneidade das regiões obtidas, isto é a similaridade entre áreas que foram agrupadas dentro de uma mesma região. Durante a construção da árvore geradora mínima (MST), são calculadas as diferenças quadráticas entre os valores da(s) variável(s) nas diferentes áreas, e estes valores são utilizados como critério para a divisão da mesma em K regiões, sendo os cortes realizados sobre vértices de maior dissimilaridade.

Nesse ímpeto, utilizamos a soma de quadrados total (SSE) como medida indicadora da homogeneidade das regionalizações. Com vistas a obter medidas de homogeneidade que incorporassem o componente espacial, calculamos também a soma de quadrados ponderada pela distância entre as áreas (dSSE):

$$\mathrm{dSSE} = \sum_{k=1}^{K} \sum_{i \neq j}^{n} d_{ij} \left(x_{ki} - x_{kj}\right)^2 \qquad (6)$$

em que d_{ij} é a distância *great-circle* entre os centroides das áreas i e j.

Nesse contexto é interessante também ponderar a distância quadrática pela contiguidade, de forma que calculamos a soma de quadrados apenas entre áreas que estão conectadas no grafo de contiguidade (cSSE)

$$\text{cSSE} = \sum_{k=1}^{K} \sum_{i \neq j}^{n} c_{ij} \left(x_{ki} - x_{kj}\right)^2 \qquad (7)$$

onde c_{ij} é uma função indicadora que assume 1 se i e j são contíguos e 0 caso contrário.

A implementação computacional dessas medidas retorna valores tanto para cada área em separado como para as regionalizações em si. Desta maneira podemos detectar áreas extremamente dissimilares dentro de suas regiões.

3.7. Comparação entre regionalizações (concordância)

Quando utilizamos diversas variáveis ou combinações de variáveis como base para a regionalização, uma questão natural é se perguntar sobre a similaridade entre regionalizações, isto é, a sobreposição espacial entre duas regionalizações com mesmo K e variáveis-base diferentes.

Neste estudo utilizamos medidas de concordância comumente encontradas na literatura de análise de conglomerados (*cluster*), a saber: índice de Rand, índice de Rand ajustado para chance, índice de Fowlkes-Mallows e índice de Jaccard. Estas medidas estão implementadas na função `adjustedRand()` do pacote **clues** (WANG, QIU & ZAMAR, 2007). O leitor é direcionado a MILIGAN & COOPER, 1986 e WANG, QIU & ZAMAR, 2007 para uma descrição detalhada da estrutura destas métricas.

No presente trabalho comparamos, para cada K, as regionalizações obtidas para anos pré-epidemicos (2001 e 2006) e epidêmicos (2002 e2008), bem como as regionalizações utilizando a incidência de dengue como variável-base com aquelas baseadas nos índices de Breteau e de desenvolvimento social. Essas comparações visam a fornecer uma ideia sobre a possível diferença de padrão espacial da doença em diferentes estágios epidêmicos e também investigar a relação entre as regiões obtidas para dengue e aquelas obtidas para seus preditores.

3.8. Aspectos computacionais

3.8.1. Pacotes [Dependências?]

As análises realizadas neste estudo foram feitas no ambiente de computação estatística R (R CORE TEAM, 2012). A implementação da SKATER se dará através da utilização da função `skater()` do pacote **spdep** (PEBESMA & BIVAND, 2005). Para a geração de mapas e análises espaciais em geral foram utilizados os pacotes **maptools** (LEWIN-KOH *et al*., 2009) e **RColorBrewer** (NEUWIRTH, 2007). A seguir, detalhamos os procedimentos adotados, bem como a escrita dos *scripts* utilizados nas análises apresentadas. Para as comparações entre as regionalizações utilizamos as funções implementadas no pacote **clues** (WANG, QIU & ZAMAR, 2007). Para adicionar a funcionalidade de colapsar e exportar arquivos shape (.shp), utilizamos os pacotes **gpclib** e **shapefiles.**

3.8.2. Automatização da Regionalização

Para facilitar a realização de diversas regionalizações com diversas combinações de variáveis-base, número de regiões desejado e critérios de restrição, foram criados *wrappers* personalizados para juntar funções de interesse e facilitar a exportação de dados e a visualização de resultados.

A ideia por trás da elaboração desses scripts é facilitar a realização e análise simultânea de várias regionalizações. Para tanto, a abordagem básica é criar uma tabela com as comparações a serem feitas, gerar uma árvore de diretórios que conterá os resultados de cada experimento e exportar arquivos de texto contendo os dados gerados. Todos estes procedimentos podem ser controlados pelo usuário, de forma a filtrar aquilo que é ou não exportado e que análises são desejadas. Esses códigos são apresentados no apêndice 3, seção 8.3.

Todos os *scripts* utilizados neste trabalho foram escritos na linguagem R e tem por objetivo produzir gráficos em padrão de publicação e tabelas informativas de forma automática, para um grande número de combinações de números de regiões (K), variáveis-base e critérios de população mínima. O código foi escrito tendo em mente a generalidade,

podendo ser utilizado para outros estudos de regionalização com pouca ou nenhuma modificação.

Entre as funcionalidades do scripts escritos podemos destacar:

1. Geração e exportação automáticas de *box-plots* coloridos da(s) variável(s) de interesse, por região;
2. Geração e exportação automáticas de mapas coloridos com os resultados da regionalização;
3. Geração e exportação automáticas de tabelas de medidas de homogeneidade e balanceamento, por área e por região, conforme aplicável;
4. Exportação de arquivos .shp contendo as regiões formadas, bem como arquivos .dbf a eles associados, contendo as variáveis de interesse devidamente recalculadas, por região;
5. *Output* separado automaticamente por diretórios, facilitando a análise e organização dos resultados.

IV- RESULTADOS

4.1. Analise exploratória

4.1.1. Box-plots

A seguir apresentamos os diagramas de caixa obtidos para as variáveis estudadas.

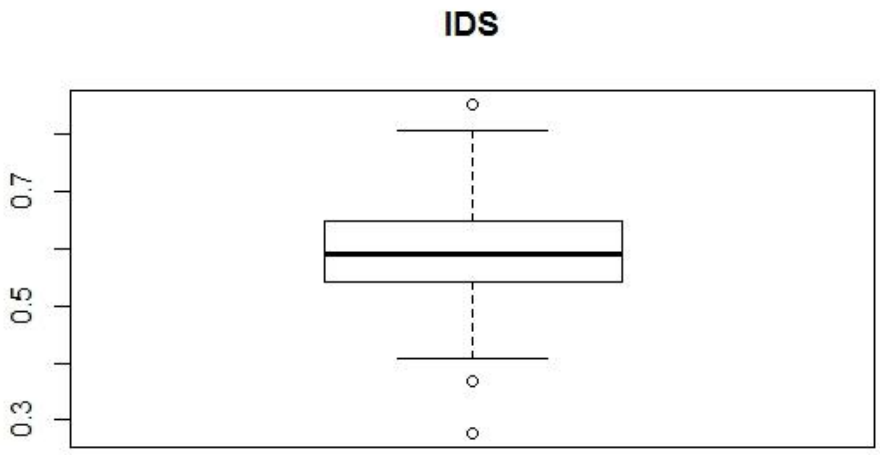


Figura 5. Diagrama de caixa para o índice de desenvolvimento social (IDS).

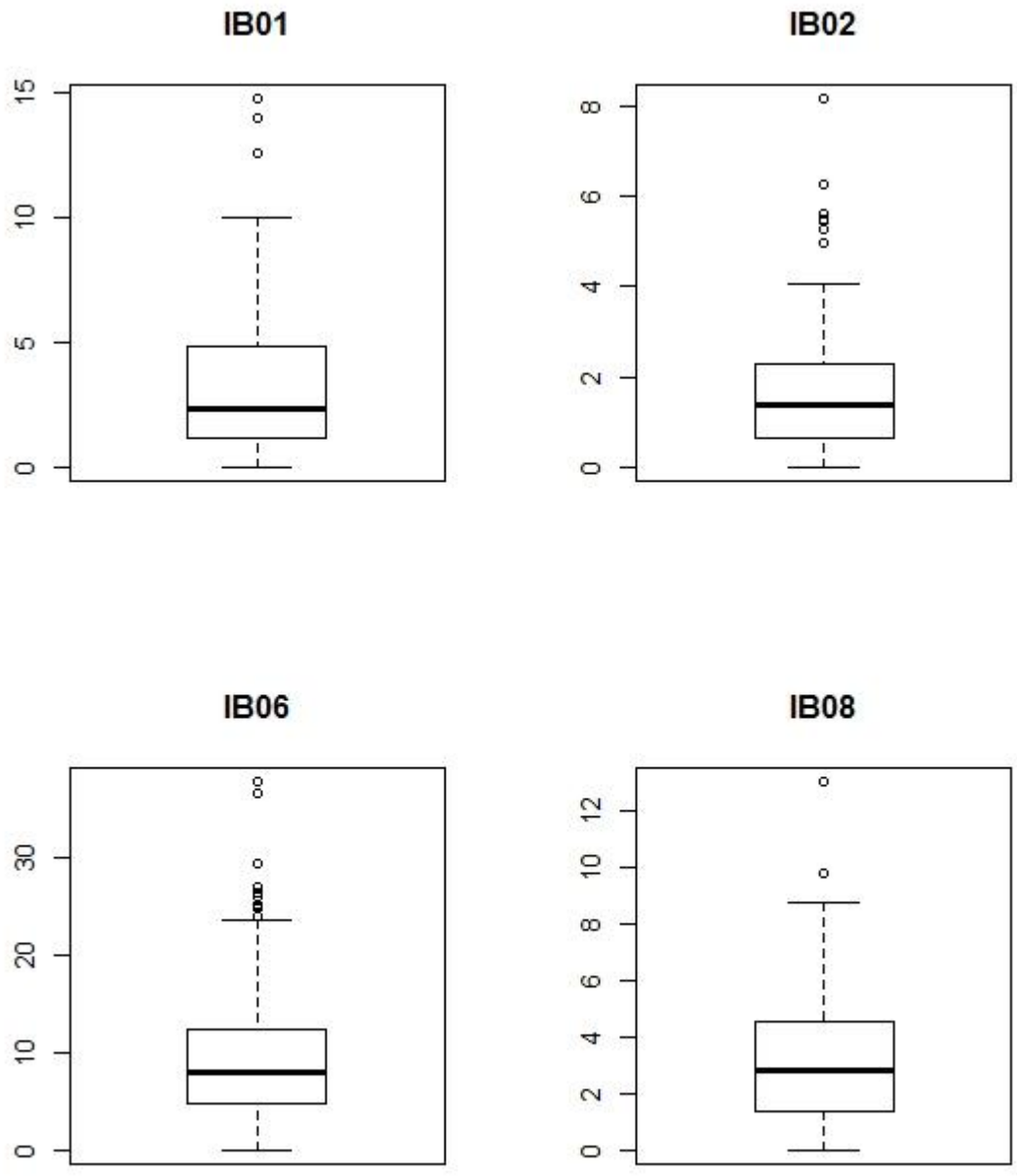


Figura 6. Diagramas de caixa para o índice de Breteau nos anos de 2001, 2002, 2006 e 2008.

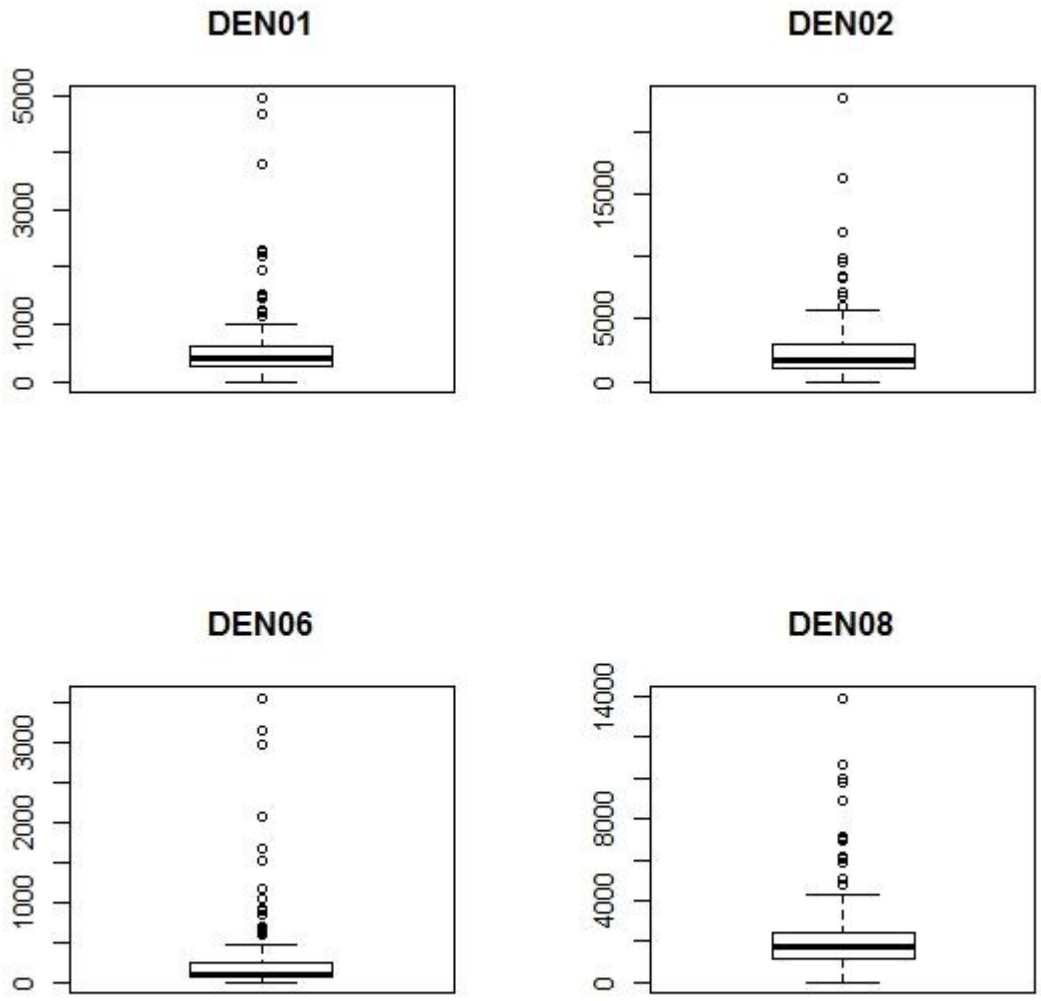


Figura 7. Diagramas de caixa para a incidência de dengue nos anos de 2001, 2002, 2006 e 2008.

4.1.2. Índices de Moran e Geary

Na tabela abaixo apresentamos os resultados da aplicação de rotinas de re-amostragem (Monte Carlo, neste caso) à variáveis de interesse com vistas à realização de testes de significância para os índices de autocorrelação espacial Moran e Geary. Para

cada variável, apresentamos a estatística obtida juntamente com um p-valor obtido a partir de 1000 realizações do algoritmo de re-amostragem.

Tabela 3. Índices de Moran e Geary para as variáveis analisadas. P-valores obtidos através de 1000 iterações do algoritmo de Monte Carlo. Em negrito os p-valores significativos ao nível de confiança de 0.05.

Variável	I de Moran	p-valor	C de Geary	p-valor
pop2000	-0.062	0.901	0.899	0.255
DEN01	-0.004	0.414	0.753	**0.046**
DEN02	-0.014	0.516	0.777	0.066
DEN06	-0.037	0.713	0.926	0.369
DEN08	0.065	0.067	0.960	0.366
IB01	0.270	0.001	0.711	**0.001**
IB02	0.336	0.001	0.682	**0.001**
IB06	0.214	0.001	0.735	**0.001**
IB08	0.316	0.001	0.732	**0.001**
IDS	0.043	0.151	0.922	0.129

4.1.3. Mapas coropléticos

Na Figura 8 mostramos os mapas coropléticos para a incidência de dengue nos anos estudados.

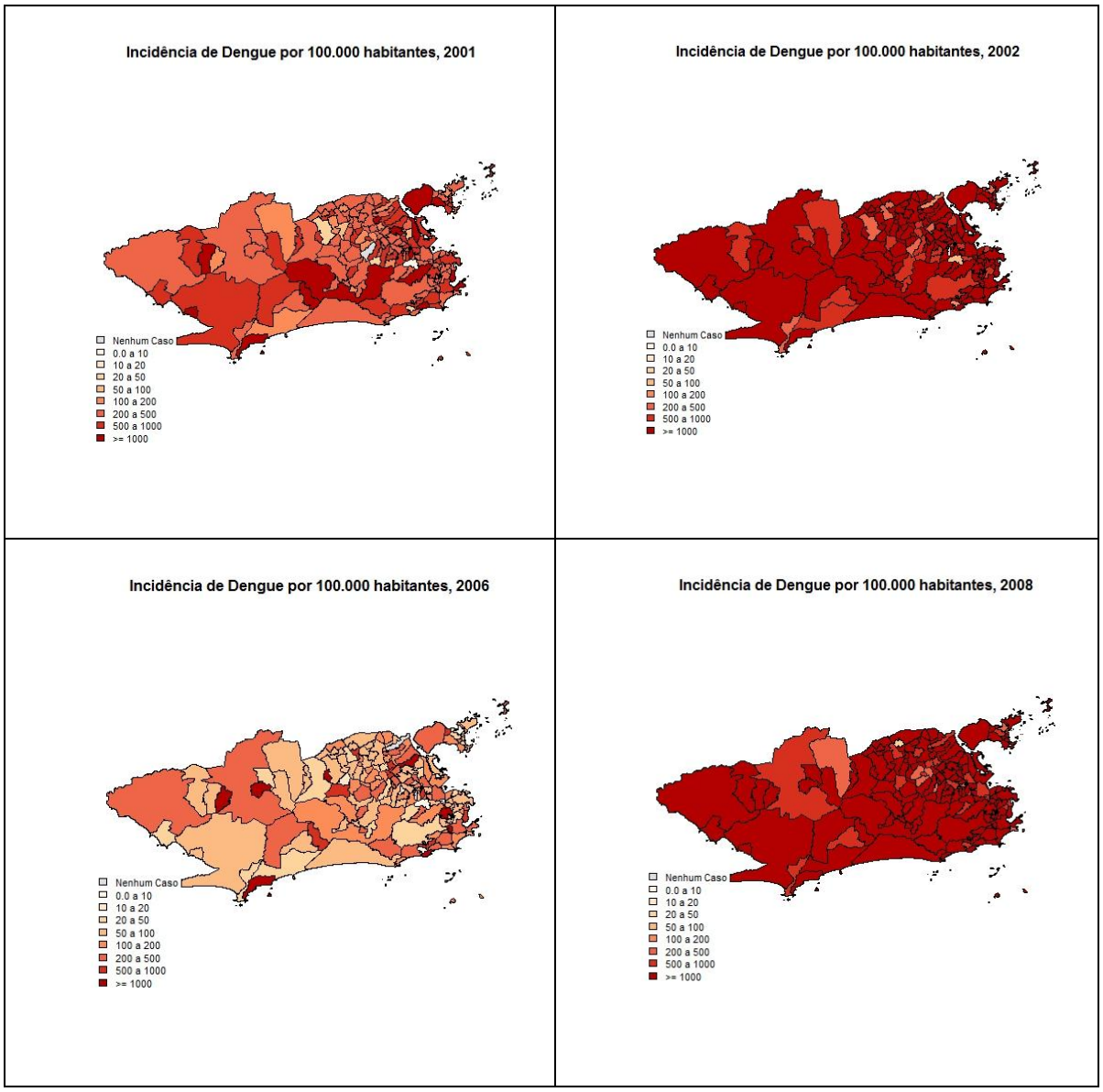


Figura 8. Mapas de cores para a incidência de dengue nos anos de 2001, 2002, 2006 e 2008

4.2. Regionalizações

Nesta seção apresentamos os mapas ilustrativos da regionalizações obtidas. Por conveniência, reportamos os resultados para cada número de regiões (K), facilitando a visualização de eventuais diferenças entre as regionalizações obtidas com diferentes variáveis-base.

4.2.1. Regionalizações com K=6

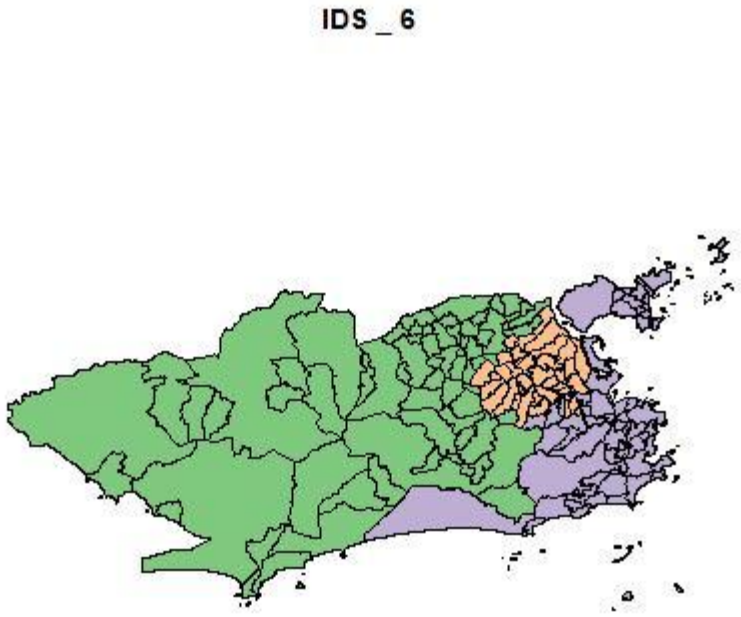


Figura 9. Regionalização para IDS, K=6, mínimo de 1 milhão de habitantes por região.

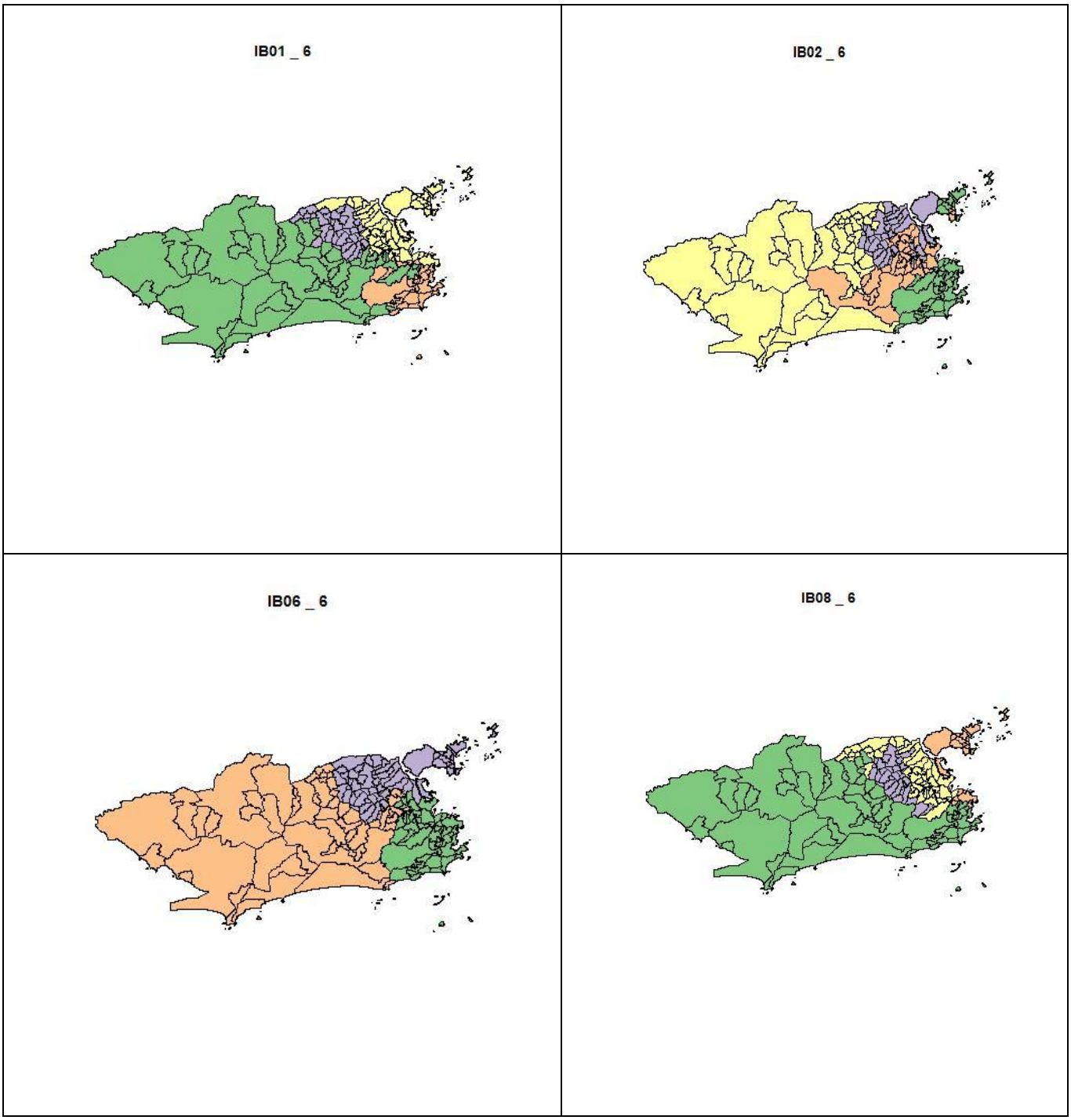


Figura 10. Regionalizações obtidas para o índice de Breteau em 2001, 2002, 2006 e 2008, K=6, mínimo de 1 milhão habitantes por região.

Figura 11. Regionalizações obtidas para a incidência de dengue em 2001, 2002, 2006 e 2008, K=6, mínimo de 1 milhão habitantes por região.

4.2.2. Regionalizações com K=12

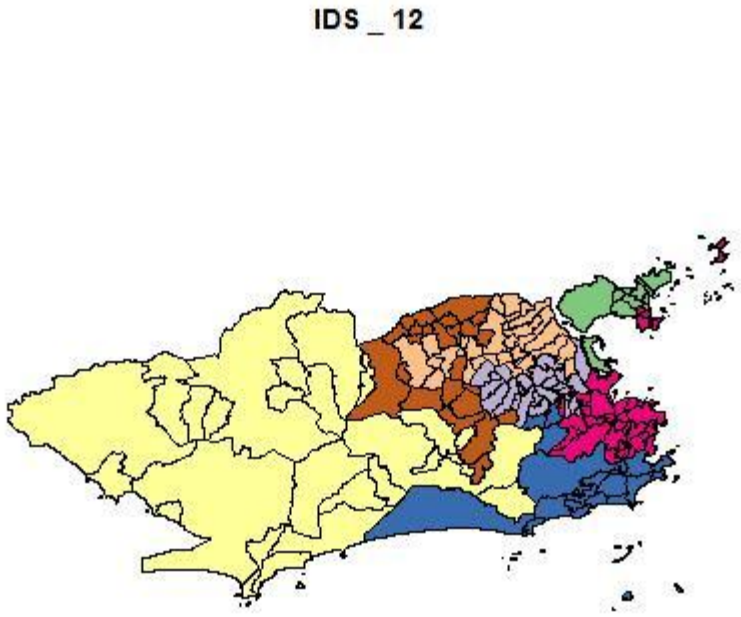


Figura 12. Regionalização para IDS, K=12, mínimo de 500 mil habitantes por região.

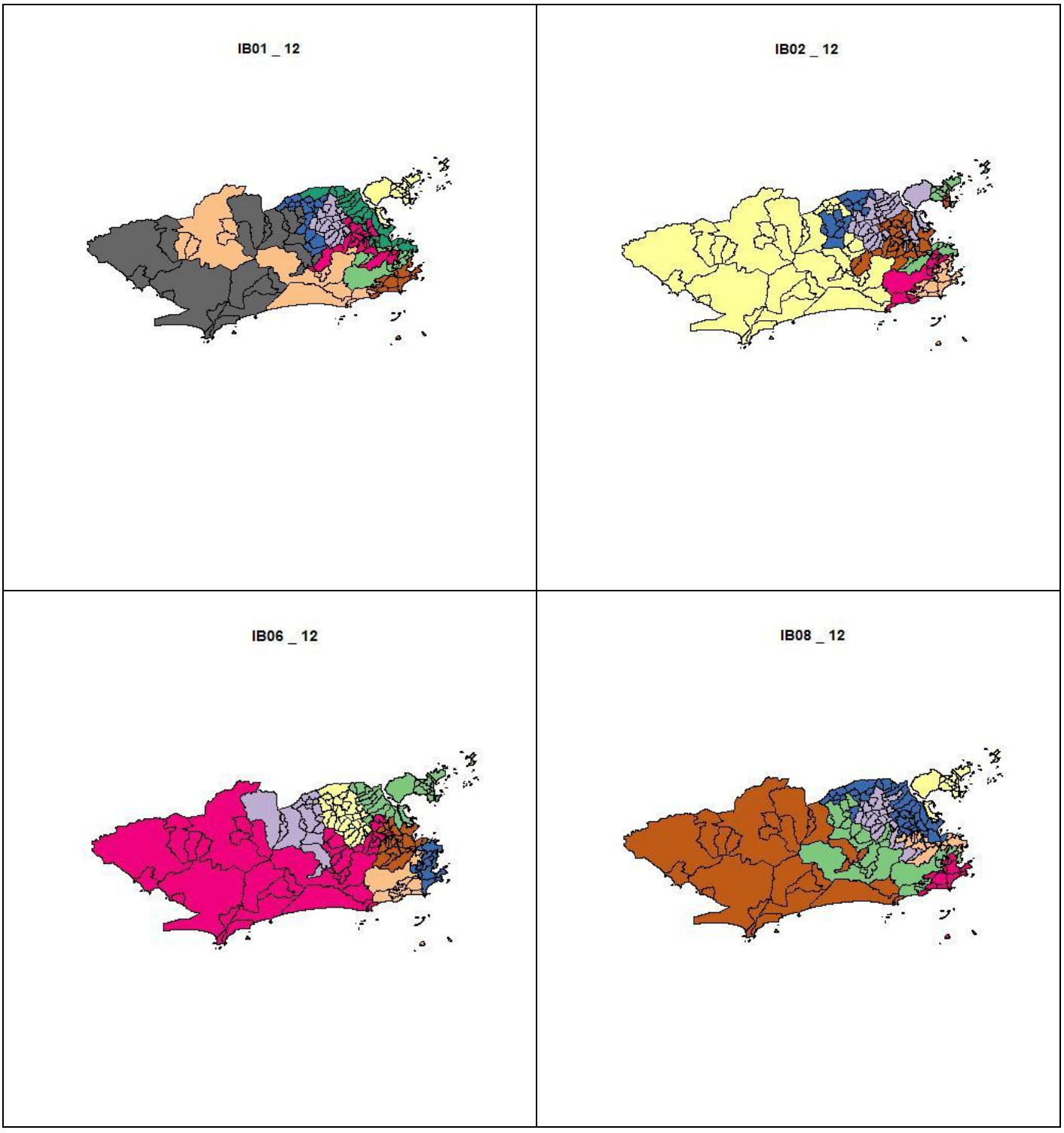


Figura 13. Regionalizações obtidas para o índice de Breteau em 2001, 2002, 2006 e 2008, K=12, mínimo de 500 mil habitantes por região.

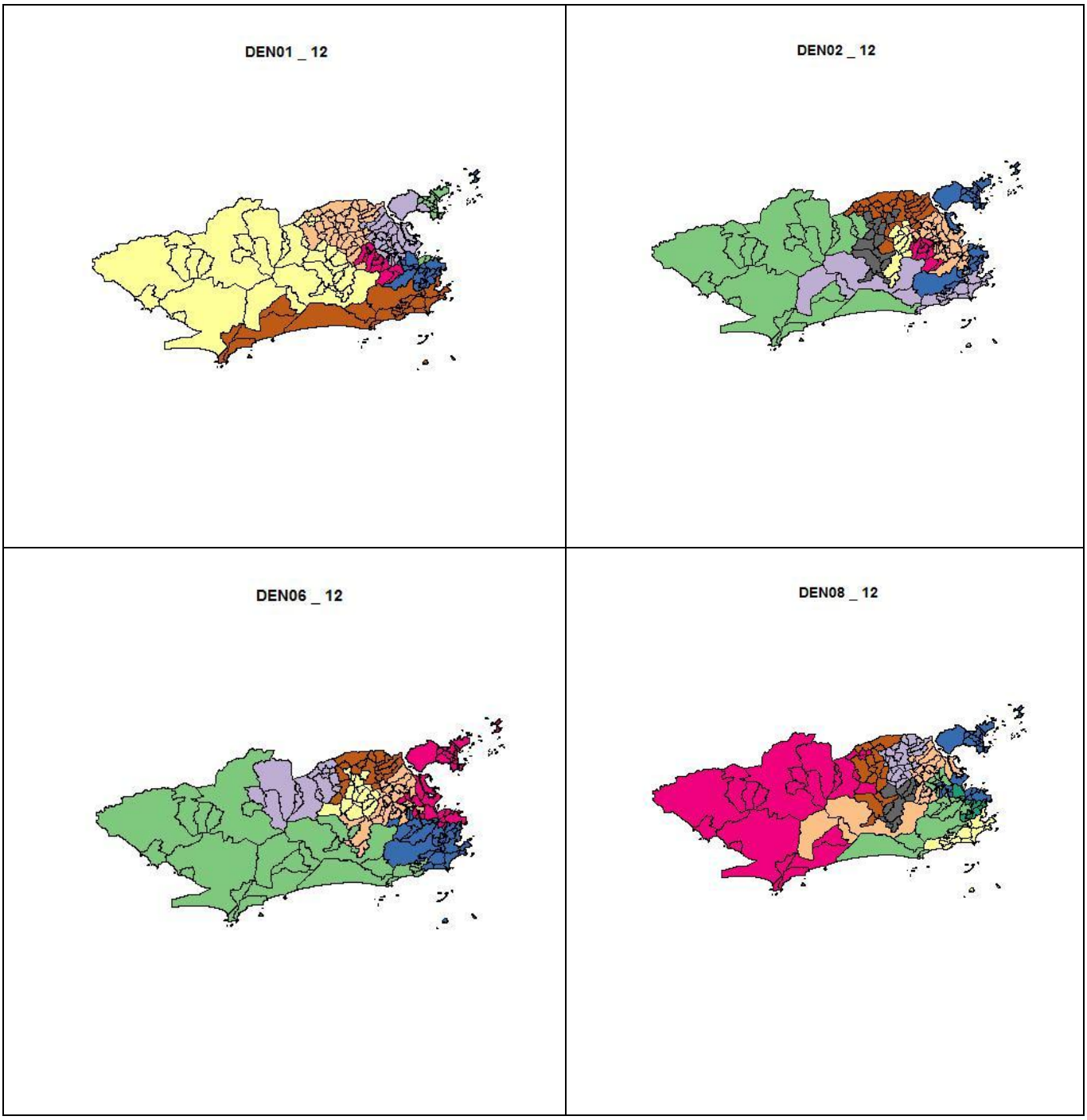


Figura 14. Regionalizações obtidas para a incidência de dengue em 2001, 2002, 2006 e 2008, K=12, mínimo de 500 mil habitantes por região.

4.2.3. Regionalizações com K=30

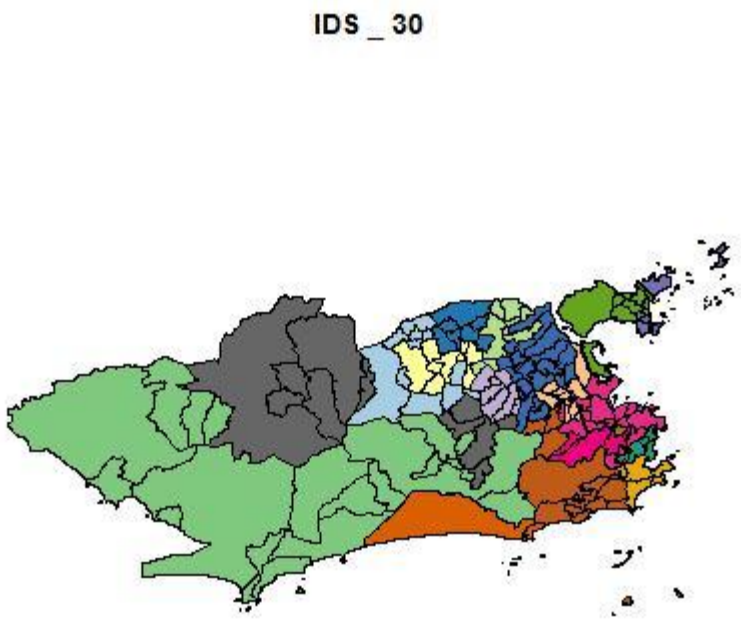


Figura 15. Regionalização para IDS, K=30, mínimo de 200 mil habitantes por região.

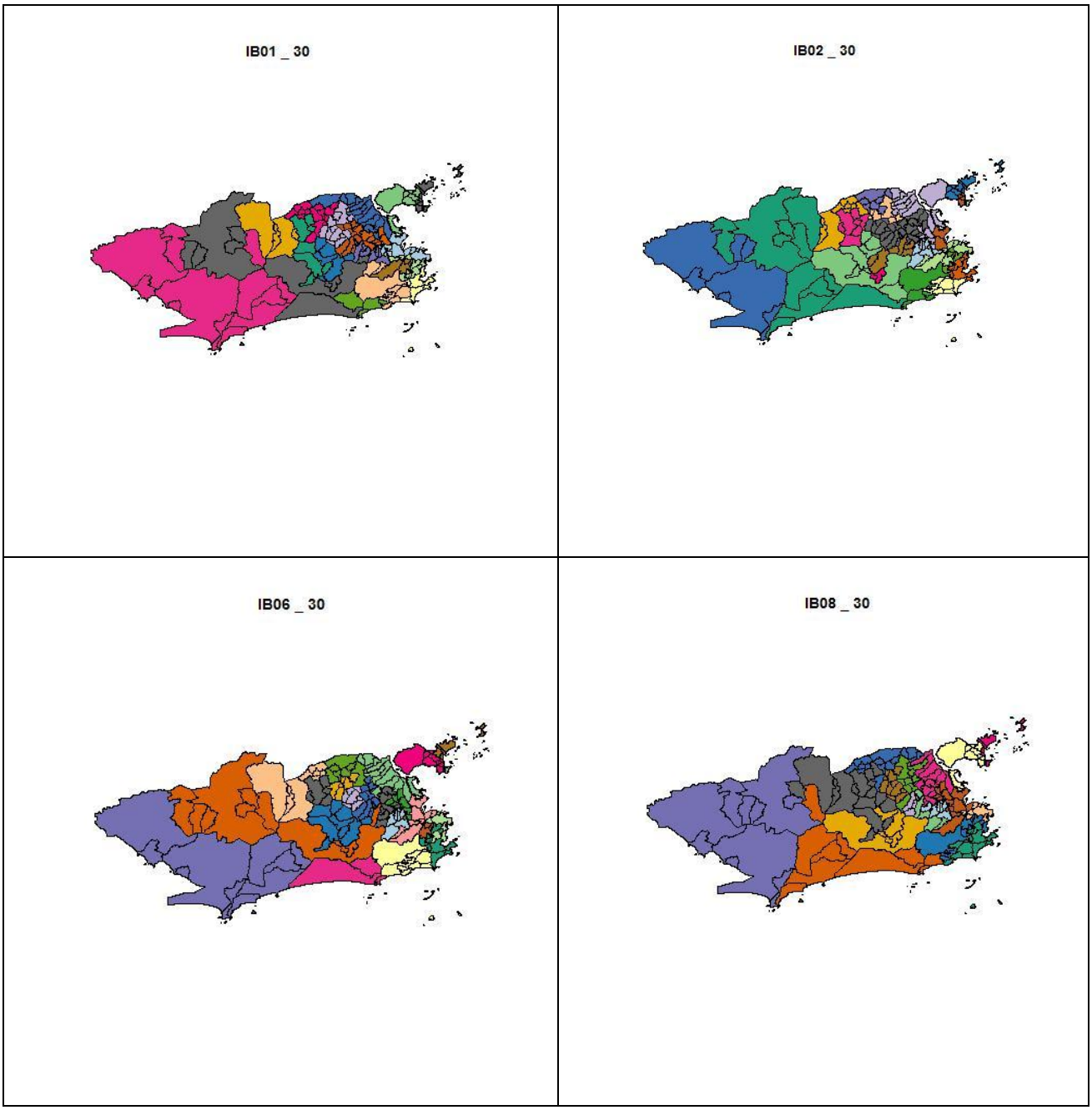


Figura 16. Regionalizações obtidas para o índice de Breteau em 2001, 2002, 2006 e 2008, K=30, mínimo de 200 mil habitantes por região.

Figura 17. Regionalizações obtidas para a incidência de dengue em 2001, 2002, 2006 e 2008, K=30, mínimo de 200 mil habitantes por região.

4.3. Qualidade das Regionalizações

Nesta seção apresentamos uma tabela mostrando as medidas propostas nas seções 3.6.1 e 3.6.2 para todos os experimentos realizados.

Tabela 4. Medidas de homogeneidade e balanceamento para as regionalizações obtidas neste trabalho. SSE- soma de quadrados; dSSE- soma de quadrados ponderados pela distância; cSSE soma de quadrados entre vizinhos. GMiR--grau médio intra-região. R(X) equivale à medida proposta na equação 5 (seção 3.6.1).

Experimento	Variável	K	SSE	dSSE	cSSE	GMiR	R(X)
1	DEN01	6	1.41E+08	5937.43	8186793.28	4.27	0.989
2	DEN02	6	2.29E+09	101212.50	186548707.85	4.23	0.951
3	DEN06	6	79185926	3263.33	9371942.99	4.27	0.972
4	DEN08	6	1.34E+09	54162.08	139710414.40	4.36	0.971
5	IB01	6	1427.959	0.06	138.48	4.15	0.966
6	IB02	6	412.1986	0.02	31.33	3.93	0.995
7	IB06	6	12272.23	0.48	836.96	4.59	0.997
8	IB08	6	1154.299	0.05	126.73	4.32	0.947
9	IDS	6	2.731519	0.00	0.20	4.68	0.973
10	DEN01	12	1.29E+08	6070.42	12154531.96	4.04	0.969
11	DEN02	12	2.29E+09	84660.82	275188364.08	3.63	0.968
12	DEN06	12	78062246	3344.69	12261254.52	3.72	0.974
13	DEN08	12	1.12E+09	43377.00	213004857.45	3.38	0.975
14	IB01	12	1043.912	0.05	172.89	3.24	0.983
15	IB02	12	400.4463	0.02	43.81	3.48	0.946
16	IB06	12	9927.151	0.46	1772.04	3.99	0.955
17	IB08	12	873.8895	0.04	136.94	3.55	0.974
18	IDS	12	2.295853	0.00	0.30	3.76	0.982
19	DEN01	30	1.18E+08	5802.02	18220585.14	2.92	0.953
20	DEN02	30	1.91E+09	66860.54	328823355.63	2.74	0.970
21	DEN06	30	66404124	3363.59	17729116.50	2.80	0.962
22	DEN08	30	1.06E+09	40744.30	282962141.76	2.96	0.959
23	IB01	30	706.3143	0.03	236.75	2.70	0.980
24	IB02	30	258.6538	0.01	73.67	2.64	0.969
25	IB06	30	7551.944	0.40	2820.04	2.86	0.978
26	IB08	30	622.0987	0.03	226.03	2.93	0.984
27	IDS	30	1.538392	0.00	0.50	2.99	0.943

4.4. Concordância

Neste trabalho buscamos comparar, principalmente, dois aspectos das regionalizações obtidas: (i) as diferenças entre anos pré- epidêmicos e epidêmicos; e (ii) as diferenças entre regionalizações obtidas para dengue e para covariáveis ambientais e sócio-economicas.

A seguir apresentamos duas tabelas que reúnem os resultados dessas comparações utilizando o índice de Jaccard, tanto para a relação entre a dengue nos diferentes anos quanto para a relação entre esta e as variáveis IDS e IB.

Tabela 5. Comparação entre as regionalizações para dengue em anos pré-epidêmicos e epidêmicos, índice de Jaccard.

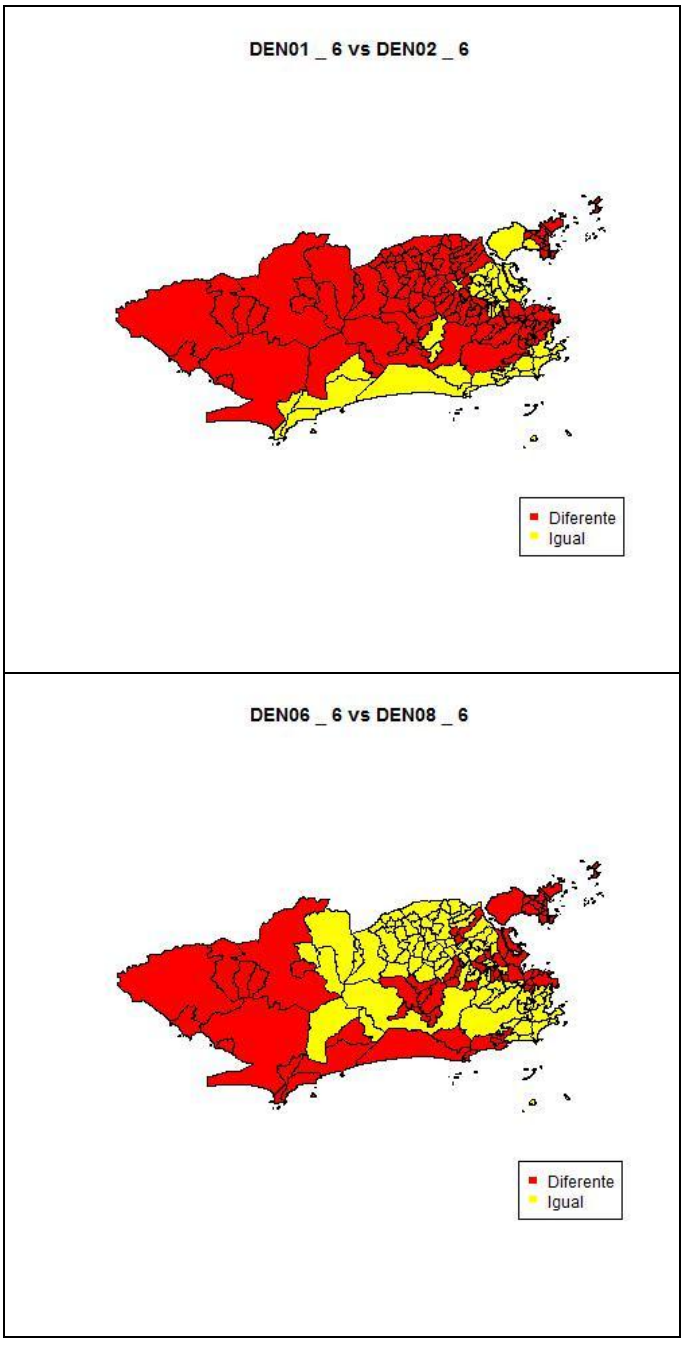
	ANO		
K	2001-2002	2006-2008	2002-2008
6	0.330	0.386	0.404
12	0.292	0.266	0.306
30	0.212	0.222	0.284

A Tabela 5, acima, mostra as comparações tanto para anos pré-epidêmicos e seus subsequentes anos epidêmicos quanto uma comparação entre anos epidêmicos, para os valores de K utilizados neste trabalho.

Tabela 6. Comparações entre as regionalizações para dengue e para indicadores de infestação vetorial e socioeconômicos, índice de Jaccard.

Ano	K	6	12	30
	DEN-IB	0.293558	0.24886	0.1843
2001	DEN-IDS	0.38573	0.318198	0.195804
	DEN-IB	0.284206	0.226746	0.194097
2002	DEN-IDS	0.407013	0.277778	0.215524
	DEN-IB	0.317804	0.263487	0.188091
2006	DEN-IDS	0.333748	0.273872	0.223893
	DEN-IB	0.268676	0.231315	0.240628
2008	DEN-IDS	0.499298	0.235835	0.227848

Apresentamos, também alguns exemplos de mapas de sobreposição, confeccionados com o intuito de facilitar a análise das diferenças e semelhanças entre regionalizações por prover uma abordagem gráfica. A Figura 18 mostra as comparações entre anos epidêmicos e não epidêmicos, e a Figura 19 mostra as comparações entre dengue e IDS. Para estes exemplos, apresentamos os resultados para K=6.


DEN01 _ 6 vs DEN02 _ 6
Diferente
Igual
DEN06 _ 6 vs DEN08 _ 6
Diferente
Igual

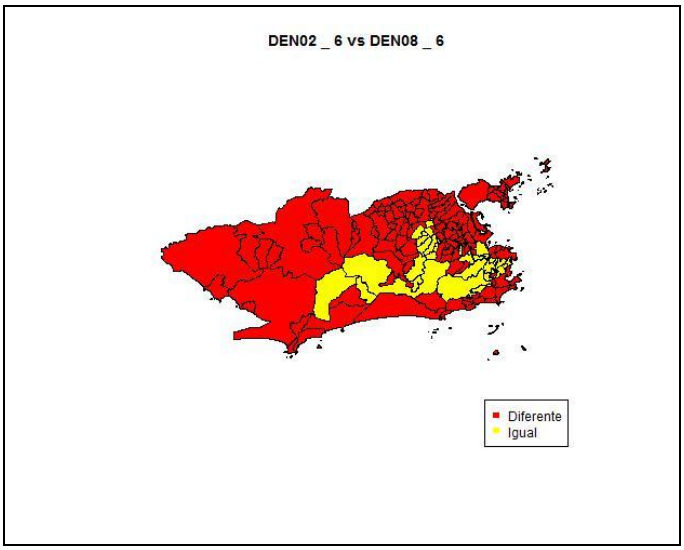


Figura 18. Comparações entre as regionalizações para dengue nos diferentes anos. Áreas marcadas em vermelho são aquelas que não foram agrupadas da mesma maneira nas duas regionalizações e as marcadas em amarelo aquelas que permaneceram agrupadas nas mesmas regiões.

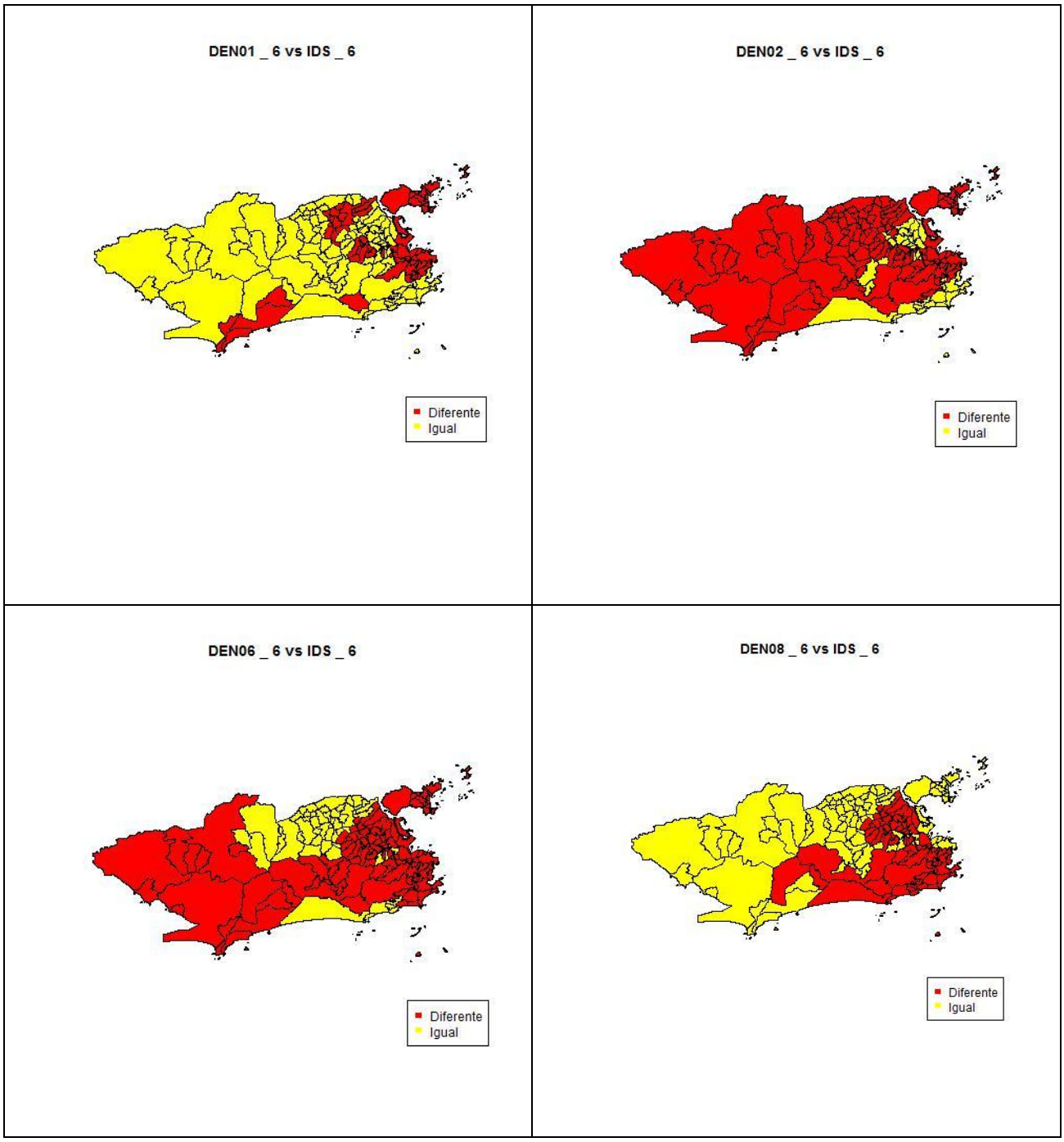


Figura 19. Comparações entre dengue e IDS para K=6 nos diversos anos. As cores são como na Figura 18.

V- DISCUSSÃO

Este estudo busca investigar a aplicação de um método de regionalização (SKATER) à dengue nos bairros do município do Rio de Janeiro, utilizando vários números de regiões desejadas (K) e critérios de população mínima. Mesmo com um desenho experimental

modesto (três níveis para K, três variáveis-base [dengue, IDS e IB]), a quantidade de informações geradas pela abordagem computacional proposta neste trabalho é muito grande. De forma a facilitar a compreensão do leitor, bem como a extração de informação utilizável das análises realizadas, discutimos a seguir os resultados mais importantes/relevantes do trabalho, sem nos preocuparmos em cobrir de forma extensiva todos os experimentos realizados.

5.1. Observações gerais

A partir da análise exploratória dos dados percebemos que tanto a dengue quando o índice de Breteau apresentam distribuições assimétricas e heterogêneas no espaço (Figuras 5, 6, 7 e 8), em acordo com estudos anteriores (CÂMARA et al, 2007; CÂMARA et al, 2009; TEIXEIRA & CRUZ,2011).

A análise da autocorrelação espacial, apresentada na Tabela 3, mostra que o índice de Breteau apresenta alta autocorrelação espacial global ($p \leq 0.001$), enquanto resultados significativos ao nível de 0.05 só foram obtidos para a dengue em 2001. Estes resultados, devem ser interpretados com cautela, já que os índices de Moran e Geary indicam autocorrelação espacial global, deixando escapara as nuances locais.

A heterogeneidade das variáveis estudadas no espaço se reflete nas regionalizações obtidas, na medida em que as regiões, embora atendam às respectivas restrições de população mínima, são, em geral, pouco compactas no espaço. Nas Figuras 9, 10 e 11, para K=6, percebemos, por exemplo que alguns dos bairros da zona Oeste da cidade são agrupados com bairros da zona norte e da região mais central da cidade, criando regiões que se alongam longitudinalmente.

Este é um resultado interessante, pois pode sugerir que alguns bairros da zona Oeste tem dinâmica epidêmica semelhante àquela dos bairros das regiões mais centrais. Semelhanças no sistema de abastecimento de água e/ou coleta de lixo, densidade populacional, IDS, entre outros, podem ser a explicação para tal. A regionalização para IDS (Figura 9) sugere este que este pode ser um fator semelhante entre essas regiões, já que a zona Oeste, bem como uma parte da zona Norte e alguns bairros da região central são agrupados juntos. É interessante notar que este padrão (zona Oeste e bairros centrais agrupados juntos) se repete para K=12 e K=30 (Figuras 12 e 15, respectivamente).

Observando as Figuras 10, 11, 13, 14, 16 e 17, não conseguimos discenir nehuma das variáveis como aquela que dá regiões mais compactas, para os três valores de K. Este resultado se reflete também na Tabela 4, discutida em mais detalhes na seção 5.3.

5.2. Anos epidêmicos e não-epidêmicos

Uma pergunta frequente no contexto da análise espacial da dengue é se há diferença de padrão espacial da doença em anos epidêmicos e não-epidêmicos, ou ainda mais interessantemente, pré-epidêmicos. Para tanto, buscamos comparar as regionalizações obtidas nestes primeiros com aquelas obtidas nos últimos.

Os resultados, apresentados na Figura 18 e na tabela 5, sugerem que há uma maior semelhança entre o padrão de anos epidêmicos, já que os anos de 2002 e 2008 apresentaram maior concordância (índice de Jaccard) entre si do que com seus respectivos anos pré-epidêmicos. É importante, no entanto, notar que essas diferenças não devem ser encaradas como definitivas, já que a abordagem aqui utilizada falha em fornecer os meios para um teste formal da hipótese de diferença nos índices de concordância.

Os mapas apresentados na Figura 18 mostram que os anos de 2006 e 2008 são espacialmente mais semelhantes entre si quando comparados com 2002 e 2008. Uma explicação para esta diferença está fora do escopo deste trabalho e pode envolver, entre outros fatores, diferenças na força de infeção das duas epidemias (2001-2002 e 2006-2008).

O conhecimento das semelhanças e diferenças entre anos epidêmicos, não epidêmicos e pré-epidêmicos poderia nos fornecer informações úteis para elaboração de programas de vigilância, na medida em que permitiria às autoridades perceber quando uma mudança de padrão espacial acontecesse.

5.3. Análise da qualidade da regiões obtidas

Os resultados apresentados na Tabela 5 sugerem que não existe um padrão consistente no que toca a quais variáveis-base dão origem a regionalizações mais compactas e/ou homogêneas. No caso da homogeneidade, uma deficiência da abordagem aqui apresentada é o fato das somas de quadrados não serem, necessariamente, comparáveis entre si. Uma correção é possível e deverá implementada.

De forma geral tanto o grau médio intra-região (GMiR) quanto a razão de entropia diminuíram com o valor de K. Este é um resultados esperado, já que - sob uma ótica probabilística - o aumento do número de regiões diminui a probabilidade de que dois vizinhos estejam na mesma região, ao mesmo tempo que aumenta a probabilidade de os grupos (regiões) apresentarem números diferentes de áreas, por mero acaso. Pensando nisso é que as comparações entre regionalizações foram feitas apenas entre experimentos com igual K.

De uma maneira geral, as medidas aqui propostas, especialmente aquelas de homogeneidade, podem ser estendidas para incorporar o componente espacial. Ponderar as distâncias quadráticas pela distância ou pela contiguidade entre as áreas pode oferecer novas perspectivas sobre as semelhanças e diferenças entre regionalizações. Estas medidas se mostram importante, sobretudo, no contexto em que o pesquisador dispõe de várias variáveis e gostaria, para um certo número K de regiões desejadas, escolher qual retorna a melhor divisão em termos de homogeneidade e compactação.

Uma dimensão adicional da discussão é que a SKATER pode ser encarada meramente como técnica de sumarização e redução do ruído dos dados, não sendo importante selecionar regionalizações, mas sim observar os padrões gerados, apenas.

5.4. Comparações

Um aspecto abordado no presente trabalho é a análise de concordância entre regionalizações. No contexto da análise de conglomerados, é de praxe analisar os diferentes particionamento obtidos através de medidas de concordância (MILIGAN & COOPER, 1986; WANG, QIU & ZAMAR, 2007). Por simplicidade, apresentamos, nas Tabelas 5 e 6, apenas os resultados obtidos para o índice de Jaccard.

Na Tabela 5 temos as comparações entre as regionalizações para dengue em anos epidêmicos e não epidêmicos, para vários valores de K. Os resultados foram discutidos na seção 5.2.

Os resultados apresentados na Tabela 6 mostram que a incidência de dengue no ano de 2008 teve alta concordância com o IDS para K=6, mas esta concordância não se manteve para outros valores de K. Esse resultado, tomado em conjunto com os resultados de qualidade e balanceamento da Tabela 4, sugerem que a natureza heurística da SKATER

e estocasticidade podem ter impacto relevante na nossa habilidade de comparar regionalizações obtidas por este método. Em outras palavras, torna-se difícil distinguir entre o que é variação espúria, aleatória, e o que são diferenças espaciais sensíveis.

Surpreendentemente, embora na análise de concordância os anos de 2002 e 2008 tenham obtidos os maiores índices, a análise no espaço (Figura 18) mostra diferenças substanciais entre as duas regionalizações. Uma explicação para isso é que o índice de Jaccard – assim como os outros índices calculados - não leva o componente espacial em consideração. Embora existam propostas na literatura, o desenvolvimento de medidas de concordância espacial se mostra uma possível direção para pesquisas futuras.

VI- CONCLUSÕES

A regionalização se mostra uma técnica poderosa para análise exploratória de dados de áreas, possibilitando a detecção de padrões espaciais a partir da criação de áreas contíguas e homogêneas no espaço. Por outro lado, é eficiente na redução do ruído e da variação espúria dos dados, e, desta forma, pode ser utilizada como preparação dos dados para técnicas mais refinadas de análise, como os modelos de regressão.

Neste trabalho foi possível estudar o padrão espacial da dengue no Município do Rio de Janeiro, identificar tendências gerais de variação espacial da doença e também de indicadores de infestação vetorial e sócio-econômicos, bem como comparar os padrões espaciais observados nos diferentes anos.

Concluímos, portanto, que a regionalização pode ser uma ferramenta útil na compreensão da variação espacial da dengue e na elaboração de políticas públicas de saúde mais efetivas, apresentando potencialidade de uso no desenho de mesorregiões de saúde que sejam mais significativas no combate à dengue no Município do Rio de Janeiro.

Para além disso, este trabalho permitiu gerar ferramentas automáticas de regionalização, de grande generalidade e riqueza de análises. Esperamos que estes códigos possam ser úteis aos pesquisadores que desejarem estudar um grande número de regionalizações, com vários números de regiões, variáveis-base e critérios de restrição.

VII- BIBLIOGRAFIA

ALVANIDES, S., OPENSHAW, S. & REES, P. **Designing your own geographies**. In: The Census Data System, P. Rees, D. Martin and P. Williamson (Eds), pp. 47–65, Chichester, Wiley, Reino Unido, 2002.

ANSELIN, L. **Local indicators of spatial association**, Geographical Analysis, 27:93–115, 1995.

ASSUNÇÃO, R.M., NEVES, M.C., CÂMARA, G., FREITAS, C.C. **Efficient Regionalization Techniques for Socio-Economic Geographical Units Using Minimum Spanning Trees**. International Journal Of Geographical Information Science, 20(7):797-811, 2006.

BATISTA, R. S.; IGREJA, R. P.; GOMES, A.P.; HUGGINS, D.W.; Medicina Tropical: **Abordagem atual das doenças infeciosas e parasitarias**. Editora Cultural Medica, Rio de Janeiro, 2001.

BRAGA, L. P. **Compreendendo Probabilidade e Estatística**. 1a. ed. Rio de Janeiro, Brasil, Editora E-papers servicos editoriais. 230p, 2010.

BRÉS, P. **Historical review of dengue 1: implications of its introduction in the western hemisphere in 1977**. In *Dengue in the Caribbean*, 1997. Washington, D.C.: PAHO, 1979.

CÂMARA, F.P. *et al.* **Estudo retrospectivo (histórico) da dengue no Brasil: características regionais e dinâmicas**. Rev da Soc Bras Med Trop, 40(2): 192-196 2007.

CÂMARA, F.P. *et al.* **Climate and dengue epidemics in State of Rio de Janeiro**. Rev. Soc. Bras. Med. Trop., 42(2): 137-140, 2009.

CÂMARA, G. *et al.* **Análise espacial de áreas**. In: Druck S, Carvalho MS, Câmara G, Monteiro AMV, editores. Análise espacial de dados geográficos. São Paulo: Instituto Nacional de Pesquisas Espaciais, 2002.

CARVALHO, M.S., CRUZ, O. G. & NOBRE, F.F. **Spatial partitioning using multivariate cluster analysis and a contiguity algorithm: application to Rio de Janeiro, Brazil**. Statistics In Medicine, 15:1885-1894, 1996.

CHRISPAL, A. et al. **Acute undifferentiated febrile illness in adult hospitalized patients: the disease spectrum and diagnostic predictors—an experience from a tertiary care hospital in South India**. Trop Doct; 40:230–234, 2010.

CUMMINGS, D. A. T.; IAMSIRITHAWORN, S.; LESSLER J. T.; McDERMOTT A. *et al.*; **The impact of demographic transition on dengue in Thailand: insights from a statistical analysis and mathematical modeling.** PloS Med., 9(6), 2009.

DRUCK, S. *et al* (org). **Análise Espacial de Dados Geográficos**. Empresa Brasileira de Pesquisa Agropecuária, Planaltina, Brasil, 2004.

ENDY, T.P. *et al.* **Epidemiology of inapparent and symptomatic acute dengue virus infection: a prospective study of primary school children in Kamphaeng Phet**, Thailand. Am. J. Epidemiol., 156(1):40-51, 2002.

FLAUZINO, R.F., SOUZA-SANTOS, R. & OLIVEIRA, R.M. **Dengue, geoprocessamento e indicadores socioeconômicos e ambientais: um estudo de revisão**. Rev Panam Salud Publica, 25(5):456–61 , 2009.

FOVELL, R.G. & FOVELL, M.Y.C., **Climate zones of the conterminous United States defined using cluster analysis.** Journal of Climate, 6:2103–2135, 1993.

GEARY, R.C. **The Contiguity Ratio and Statistical Mapping**. The Incorporated Statistician, 5(3):115-145, 1954.

GOSH, M. & RAO, J.N.K., **Small Area Estimation: An Appraisal**. Statistical Science 9:55-93, 1994.

GOMES, A.F. **Análise Espacial e Temporal da Relação entre Dengue e Variáaveis Metereológicas na Cidade do Rio de Janeiro no período de 2001 a 2009**. Dissertação de Mestrado. Escola Nacional de Saúde Pública, Fundação Oswaldo Cruz, 2011.

GUBLER, D. J.. **Dengue and dengue hemorrhagic fever: its history and resurgence as a global public health problem**. In: GUBLER, D. J. & KUMO, G. (Ed.) **Dengue and dengue hemorrhagic fever**. New York: CAB International, 1997.

GUO, D. **Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP)**. International Journal of Geographical Information Science. 22(7):801-823, 2008.

GUZMAN, MG *et al*. **Dengue: a continuing global threat**. Nat. Rev. Microbiol., 8(12,Suppl):S7-16, 2010.

HAINING, R., WISE, S. & MA, J., **Designing and implementing software for spatial statistical analysis in a GIS environment**. Journal of Geographical Systems, 2: 257–286, 2000.

HONÓRIO, N.A. *et al.* **Spatial evaluation and modeling of dengue seroprevalence and vector density in Rio de Janeiro, Brazil**. PLoS Negl Trop Dis.3:e545, 2009.

IBGE. **Rio de Janeiro.** Disponível em: (http://www.ibge.gov.br/cidadesat/xtras/perfil.php?codmun=330455&r=2). Acessado em: 30 nov. 2010.

JUNGNICKEL, D., **Graphs, Networks and Algorithms,** Berlin, Springer, Alemanha, 1999.

KAY, B. **Dengue fever: reappearance in northern Queensland after 26 years**. Medical Journal of Australia, 140:264-268, 1984.

MARSHALL, R. M. **Mapping disease and mortality rates using Empirical Bayes Estimators**, Applied Statistics, 40:283–294, 1991.

MILIGAN, G.W. & COOPER, M.C. **A study of the comparability of external criteria for hierarchical cluster analysis**. Multivariate Behavioral Research 21:441–458, 1986.

MORAN, P.A.P. **Notes on Continuous Stochastic Phenomena**. Biometrika 37(1):17-23, 1950.

NEUWIRTH, E. **RColorBrewer: ColorBrewer palettes. R package version 1.0-2.** http://CRAN.R-project.org/package=RColorBrewer , 2007.

NOGUEIRA R. M. R., MIAGOSTOVICH M. P., LAMPE E., SCHATZMAYR H. G. **Isolation of dengue virus type 2 in Rio de Janeiro**. Memórias do Instituto Oswaldo Cruz, Brasil, 85: 253, 1990.

NOGUEIRA, R. M. R. *et al.* **Dengue Virus Type 3, Brazil, 2002**. Emerging Infectious Diseases**,** Brasil, 1376-1381, 2005.

NOGUEIRA, R.M.R, EPPINGHAUS, A.L. **Dengue virus type 4 arrives in the state of Rio de Janeiro: a challenge for epidemiological surveillance and control**. Mem Inst Oswaldo Cruz.106:255-6, 2011.

OPENSHAW, S. **A geographical solution to scale and aggregation problems in regionbuilding, partitioning, and spatial modelling**. Transactions of the Institute of British Geographers, 2:459–472, 1977.

OPENSHAW, S. & RAO, L., **Algorithms for reengineering 1991 census geography**. Environment & Planning A, 27: 425–446, 1995.

OPENSHAW, S., ALVANIDES, S. and WHALLEY, S**. Some Further Experiments with Designing Output Areas for the 2001 UK Census**. Report (Leeds, UK: University of Leeds), 1998.

OPENSHAW, S. & ALVANIDES, S. **Zone design for planning and policy analysis**. In Geographical Information and Planning, J. Stillwell, S. Geertman and S. Openshaw (Eds), pp 299-315, Berlim, Springer, Alemanha, 1999.

PAIVA, C.A., ALONSO, J.A., TARTARUGA, I.P. **Em busca de uma divisão regional mais compatível com as múltiplas necessidades da pesquisa e do planejamento**. In: CONCEIÇÃO, Octávio A. C. *et al.* (Org.). O ambiente regional. Porto Alegre: FEE, 2010.

PEBESMA, E.J., BIVAND, R.S**. sp:classes and methods for spatial data. R package version 0.9-44**. http://CRAN.R-project.org/package=sp, 2005.

PONS, P. **Tratado de Patología y Clínica Médica**. 2.ed Barcelona: Salvat, 1960. p.647-650. v.4.

PONTES R. J. S., RUFFINO-NETTO A. **Dengue in an urban locality situated in the southeast region of Brazil: epidemiological aspects**. Rev. Saúde Pública, Brasil, v. *28* p.218–227, 1994.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2012.

REY, J.R.. **What is Dengue?** Disponível em: <http://edis.ifas.ufl.edu/pdffiles/IN/IN69900.pdf>. Acessado em: 07 mar. 2008.

SANTOS, S.M. & BARCELLOS, C. (org) **Abordagens espaciais na saúde pública.** Brasília : Ministério da Saúde, 2006.

SANTOS, S.M. & SOUZA, W.V. (org) **Introdução à Estatística Espacial para a Saúde Pública**. Brasília: Ministério da Saúde, 2007.

SCHATZMAYR, H. G.; NOGUEIRA, R. M. Ribeiro; TRAVASSOS DA ROSA, A. P. **An outbreak of dengue virus at Rio de Janeiro - 1986**. Memórias do Instituto Oswaldo Cruz, Brasil, 2(81): 245-246, 1986.

SOPER, F. L. The **elimination of urban yellow fever in the americas through the eradication of *aedes aegypti***. Am J Public Health Nations Health, 53(1), 7-16, 1963.

TEIXEIRA, T.R.A & MEDRONHO, R.A**. Indicadores sócio-demográficos e a epidemia de dengue em 2002 no Estado do Rio de Janerio, Brasil.** Cad.Saud.Pub., 24,2160-70, 2008.

TEIXEIRA, T.R.A & CRUZ, O.G. **Spatial modelling of dengue and socio-economic environmental indicators in the city of Rio de Janeiro, Brazil**. Cad. Saud. Pub.,27,591-602, 2011.

TORRES, E. M. **Dengue**. Rio de Janeiro, Editora Fiocruz, Brasil, 2005.

WANG, S., QIU, W. & ZAMAR, R. H. **CLUES: A non-parametric clustering method based on local shrinking.** Computational Statistics & Data Analysis, 52(1): 286-298. 2007.

WHO. **Dengue fever and dengue haemorrhagic fever prevention and control.** World Health Assembly Resolution WHA55.17, adopted by the 55th World Healthy Assembly

2002. Disponível em (http://www.who.int/gb/ebwha/pdf_files/WHA55/ewha5517.pdf). Acessado em: 21 fev. 2012

WHO. **Dengue: guidelines for diagnosis, treatment, prevention, and control.** World Health Organization, 2009. Disponível em (http://whqlibdoc.who.int/publications/2009/9789241547871_eng.pdf). Acessado em: 21 fev. 2012

WILDER-SMITH, A *et al.* **Update on dengue: epidemiology, virus evolution, antiviral drugs, and vaccine development.** Curr Infect Dis Rep, 12, 3:157-64, 2010.

VIII- ANEXOS

8.1. Anexo 1: esquemas dos algoritmos da metodologia SKATER (ASSUNÇÃO et al, 2006).

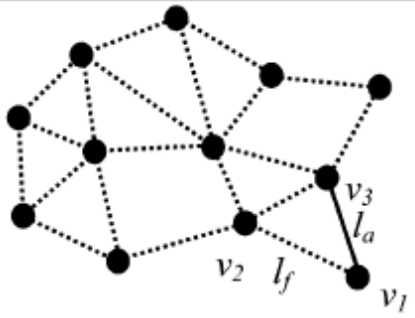
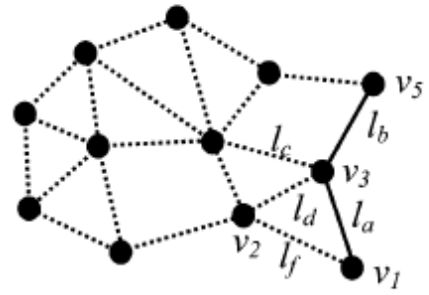
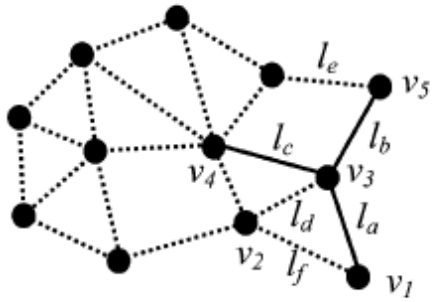
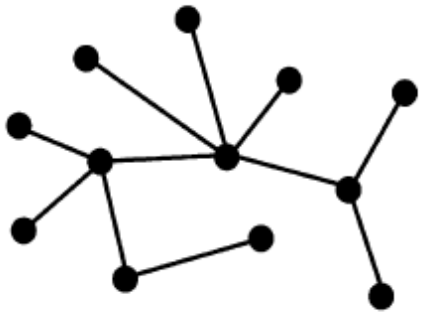
	**First iteration:** Set $T_1 = (V_1, L_1)$, where $V_1 = \{v_1\}$ and $L_1 = \phi$. Find the edge of lowest cost ($l_a \langle l_f$). Step 3: $T_2 \Rightarrow V_2 = \{v_1, v_3\}$ e $L_2 = \{l_a\}$. Step 4: Repeat Step 2.
	**Second iteration:** Find the edge of lowest cost ($l_b \langle l_c \langle l_d \langle l_f$). Set $T_3 \Rightarrow V_3 = \{v_1, v_3, v_5\}$ and $L3 = \{l_a, l_b\}$.
	**Third iteration:** Find the edge of lowest cost ($l_c \langle l_d \langle l_e \langle l_f$). Set $T_4 \Rightarrow V_4 = \{v_1, v_3, v_4, v_5\}$ and $L3 = \{l_a, l_b, l_c\}$.
	**Final Iteration:** $V_n = V$.

Figura A1. Representação do algoritmo de Prim para obtenção da árvore geradora mínima (MST). Retirado de ASSUNÇÃO *et al*, 2006.

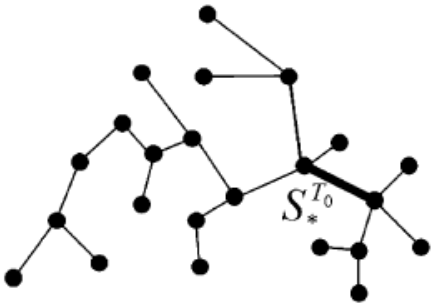
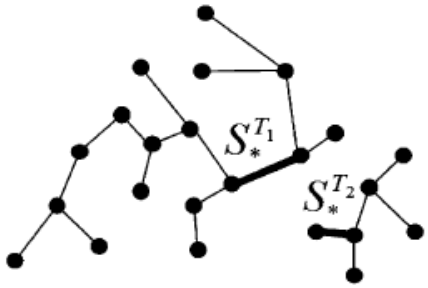
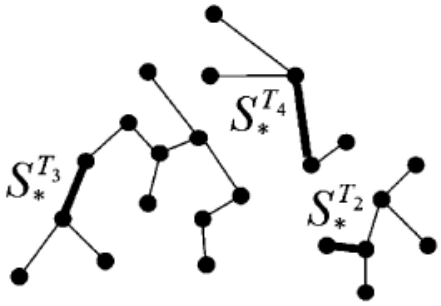
	Iteration 0: G* = MST. We select the edge which has the largest objective function. Cut out this edge leaving two trees (T_1 and T_2).
	Iteration 1: G* = (T_1, T_2). We compare the highest objective functions for T_1 and T_2. We split the tree T_1 since $f_1(S_*^{T_2}) \leq f_1(S_*^{T_1})$
	Iteration 2: G* = (T_2, T_3, T_4). We compare the highest objective functions for T_2, T_3 and T_4. We split the tree T_3 since $f_1(S_*^{T_2}) \leq f_1(S_*^{T_4}) \leq f_1(S_*^{T_3})$

Figura A2. Esquema do algoritmo de particionamento da MST. Retirado de ASSUNÇÃO *et al*, 2006.

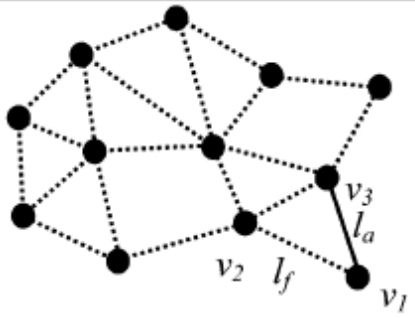
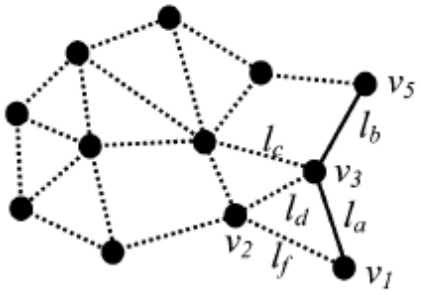
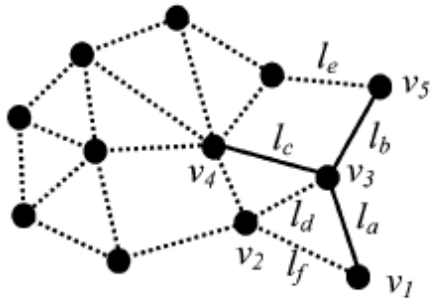
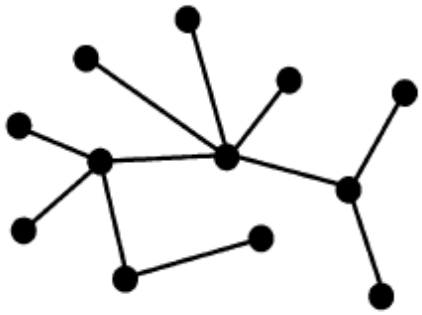
	**First iteration:** Set $T_1 = (V_1, L_1)$, where $V_1 = \{v_1\}$ and $L_1 = \phi$. Find the edge of lowest cost ($l_a \langle l_f$). Step 3: $T_2 \Rightarrow V_2 = \{v_1, v_3\}$ e $L_2 = \{l_a\}$. Step 4: Repeat Step 2.
	**Second iteration:** Find the edge of lowest cost ($l_b \langle l_c \langle l_d \langle l_f$). Set $T_3 \Rightarrow V_3 = \{v_1, v_3, v_5\}$ and $L3 = \{l_a, l_b\}$.
	**Third iteration:** Find the edge of lowest cost ($l_c \langle l_d \langle l_e \langle l_f$). Set $T_4 \Rightarrow V_4 = \{v_1, v_3, v_4, v_5\}$ and $L3 = \{l_a, l_b, l_c\}$.
	**Final Iteration:** $V_n = V$.

Figura A1. Representação do algoritmo de Prim para obtenção da árvore geradora mínima (MST). Retirado de ASSUNÇÃO *et al*, 2006.

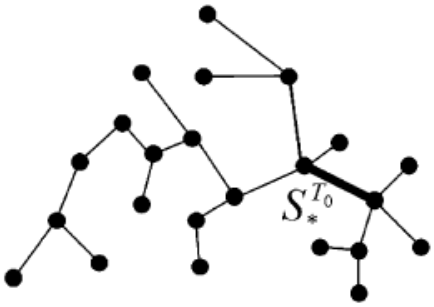
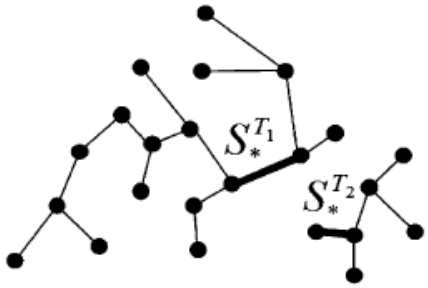
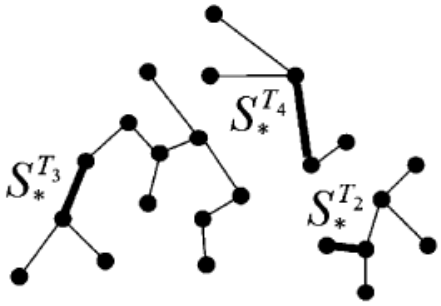
	Iteration 0: G* = MST. We select the edge which has the largest objective function. Cut out this edge leaving two trees (T_1 and T_2).
	Iteration 1: G* = (T_1, T_2). We compare the highest objective functions for T_1 and T_2. We split the tree T_1 since $f_1(S_*^{T_2}) \leq f_1(S_*^{T_1})$
	Iteration 2: G* = (T_2, T_3, T_4). We compare the highest objective functions for T_2, T_3 and T_4. We split the tree T_3 since $f_1(S_*^{T_2}) \leq f_1(S_*^{T_4}) \leq f_1(S_*^{T_3})$

Figura A2. Esquema do algoritmo de particionamento da MST. Retirado de ASSUNÇÃO *et al*, 2006.

8.2. Anexo 2: código R para imputação dos dados de infestação vetorial

C:\Users\paraftosa\Dropbox\Compartilhada_Luiz-Oswaldo\CODE\imputa_a_bagaca.R quinta-feira, 21 de fevereiro de 2013 17:08

```
####Pequeno e singelo script para imputar os dados de infestação vetorial#########
################################################################################
# carregando os dados
setwd("M:\\")
#setwd("~/")
set.seed(278326)
library(spdep)
load("rio_map.RData")
source("skater.mm.R")
banco<-data.frame(read.table("DADOS_LUIZ.txt",TRUE,sep="\t"))
dengue<-banco[,12:28]
iip<-banco[,29:38]
iib<-banco[,39:48]
rio<-bairros;rio.nb<-poly2nb(rio)
rio@data<banco
#################################################
# consertando a vizinhança
rio.nb[13][[1]]<-as.integer(c(14))
rio.nb[14][[1]]<-as.integer(c(rio.nb[14][[1]],13))
#
rio.nb[13][[1]]<-as.integer(c(98))
rio.nb[98][[1]]<-as.integer(c(rio.nb[98][[1]],13))
#
rio.nb[105][[1]]<-as.integer(c(104))
rio.nb[104][[1]]<-as.integer(c(rio.nb[104][[1]],105))
#########################################################
bin.mat<-nb2mat(rio.nb,style="B")# matriz de incidencia
data.2.fix<-data.frame(iib,iip);dt.2.fix<-data.2.fix
##
# A tal função. Note que tem dois for() loops, já que ela conserta todas as variáveis do
data.frame de uma vez
fix.NA<-function(data,mat){
  fixed.data<-data
  for (v in 1:ncol(data)){
   where<-which(is.na(data[,v]))
    for(i in 1:length(where)){
    neigh<-bin.mat[,where[i]]
    fix<-neigh*data[,v]
    fixed.data[where[i],v]<- mean(as.vector(fix[as.vector(which(fix>0))]))
  }
 }
return(fixed.data)
}
fixed<-fix.NA(data.2.fix,bin.mat)#data.frame consertado
onde<-which(is.na(iip$IIP01));result<-fixed$IIP01[onde];result # os valores imputados, IIP01
como exemplo
rio2<-rio;rio2@data<-fixed;rio2@data$IND<-ifelse(is.na(iip$IIP01)=="TRUE",1,0)
plot(rio2,col=heat.colors(2)[rio2$IND+1])
title("Onde tinha NA, IIP01");windows()
library(RColorBrewer)
cores <-brewer.pal(8,"Reds")
#cores <- c("#E0E0E0",cores)
quebra <- c(0,0.5,1,1.5,2,4,6,10)
indicor <- findInterval(fixed$IIP01,quebra)
plot(rio,col=cores[indicor])
title("Imputado, IIP01")
#####
```

-1-

8.3. Anexo 3: código R para automatização das regionalizações

O código abaixo apresentado contém funções escritas tanto para automatizar a regionalização quanto para implementar as medidas de homogeneidade e balanceamento propostas no
texto.

```
### Companion to main script 'regionalizator.R'###
# Copyleft (or the one to blame): Carvalho, LMF (2012)
# Last update: 15/01/2012
### WARNING: This code is a MESS, and I don't care a bit about it :0) It was written for my
own purposes
##_____________________________________________________________________________
#Loading packages
get.packs<- function(packs,repos){
  for (i in 1:length(packs)){
    if (!packs[i] %in% installed.packages()) {
      install.packages(packs[i],repos=repos)
    }
    library(packs[i],character.only = TRUE)
  }
}

packs<-c("spdep","maptools","RColorBrewer","spatstat","shapefiles","fields","gpclib","clues")
get.packs(packs,repos="http://cran.us.r-project.org")
spatstat.options(gpclib=TRUE)
gpclibPermit()
# Big collor pallete
bigpal<-c(brewer.pal(8,"Accent"),brewer.pal(8,"Dark2"),brewer.pal(12,"Paired"),brewer.pal(9,
"Pastel1"),brewer.pal(8,"Pastel2"),brewer.pal(9,"Set1"),brewer.pal(8,"Set2"),brewer.pal(12,
"Set3"))
#_____________________________________________________#
#                 Auxiliary functions
#_____________________________________________________#
regionalize<-function(shp,nb,data,k,crit,vec.crit,ID,export=FALSE,plot=FALSE){# wrapper para
juntar as funções relativas à regionalização
  if(missing(nb)){nb<-poly2nb(shp)}
  if(missing(data)){data<-shp@data}
  cost<-nbcosts(nb, data)
  mst <-mstree(nb2listw(nb, cost, style="B"),sample(1:length(nb),1))
  if(missing(vec.crit)){
    skater_reg<-skater(mst[,1:2], data=data,ncuts=k-1)
  }else skater_reg<-skater(mst[,1:2], data=data,ncuts=k-1,crit=crit,vec.crit=vec.crit)
  reg<-skater_reg$groups
  if(plot){plot(shp, col=bigpal[reg]);title(ID)}
  if(export){
    write.table(reg,file=paste("reg_vec",ID,".txt"),row.names=FALSE,col.names=FALSE)
  }
  k<-length(unique(reg))
  return(list(regionalization=reg,regions=k))
}

#__________________
areal.measures<-function(shp,nb,data,reg,export){# Medidas de dissimilaridade por área
  if(missing(nb)){nb<-poly2nb(shp)}
  if(missing(data)){data<-shp@data}
  if(!is.vector(data)){data<-rowMeans(data)} else data<-data
   n.mat<-nb2mat(nb,style="B")
   d.mat<-rdist.earth(coordinates(shp));d.mat<-(1/d.mat);diag(d.mat)<-0;d.mat<-d.mat/sum(
   d.mat)
   N<-length(reg)
  d_i<-colSums(n.mat) # degree
  His<-data.frame(ID=1:N,d_i,matrix(NA,nrow=N,ncol=4))
  for (i in 1:N){
```

```
### Companion to main script 'regionalizator.R'###
# Copyleft (or the one to blame): Carvalho, LMF (2012)
# Last update: 15/01/2012
### WARNING: This code is a MESS, and I don't care a bit about it :0) It was written for my
own purposes
##_____________________________________________________________________________
#Loading packages
get.packs<- function(packs,repos){
  for (i in 1:length(packs)){
    if (!packs[i] %in% installed.packages()) {
      install.packages(packs[i],repos=repos)
    }
    library(packs[i],character.only = TRUE)
  }
}

packs<-c("spdep","maptools","RColorBrewer","spatstat","shapefiles","fields","gpclib","clues")
get.packs(packs,repos="http://cran.us.r-project.org")
spatstat.options(gpclib=TRUE)
gpclibPermit()
# Big collor pallete
bigpal<-c(brewer.pal(8,"Accent"),brewer.pal(8,"Dark2"),brewer.pal(12,"Paired"),brewer.pal(9,
"Pastel1"),brewer.pal(8,"Pastel2"),brewer.pal(9,"Set1"),brewer.pal(8,"Set2"),brewer.pal(12,
"Set3"))
#_____________________________________________________#
#                 Auxiliary functions
#_____________________________________________________#
regionalize<-function(shp,nb,data,k,crit,vec.crit,ID,export=FALSE,plot=FALSE){# wrapper para
juntar as funções relativas à regionalização
  if(missing(nb)){nb<-poly2nb(shp)}
  if(missing(data)){data<-shp@data}
  cost<-nbcosts(nb, data)
  mst <-mstree(nb2listw(nb, cost, style="B"),sample(1:length(nb),1))
  if(missing(vec.crit)){
    skater_reg<-skater(mst[,1:2], data=data,ncuts=k-1)
  }else skater_reg<-skater(mst[,1:2], data=data,ncuts=k-1,crit=crit,vec.crit=vec.crit)
  reg<-skater_reg$groups
  if(plot){plot(shp, col=bigpal[reg]);title(ID)}
  if(export){
    write.table(reg,file=paste("reg_vec",ID,".txt"),row.names=FALSE,col.names=FALSE)
  }
  k<-length(unique(reg))
  return(list(regionalization=reg,regions=k))
}

#___________________
areal.measures<-function(shp,nb,data,reg,export){# Medidas de dissimilaridade por área
  if(missing(nb)){nb<-poly2nb(shp)}
  if(missing(data)){data<-shp@data}
  if(!is.vector(data)){data<-rowMeans(data)} else data<-data
   n.mat<-nb2mat(nb,style="B")
   d.mat<-rdist.earth(coordinates(shp));d.mat<-(1/d.mat);diag(d.mat)<-0;d.mat<-d.mat/sum(
   d.mat)
   N<-length(reg)
  d_i<-colSums(n.mat) # degree
  His<-data.frame(ID=1:N,d_i,matrix(NA,nrow=N,ncol=4))
  for (i in 1:N){
```

```
    same_reg<-which(reg==reg[i])
    His[i,3]<-sum(n.mat[i,same_reg])
    His[i,4]<-mean((data[i]-data[same_reg])^2) # Soma de quadrados (SSE)
    His[i,5]<-mean((data[i]-data[same_reg])^2*d.mat[i,same_reg]) # Soma de quadrados
    ponderada pela distância (dSSE)
    His[i,6]<-mean((data[i]-data[same_reg])^2*n.mat[i,same_reg]) # Soma de quadrados
    ponderada entre vizinhos (cSSE)
  }
  names(His)<-c("ID","Degree","cDegree","SSE","dSSE","cSSE")
  if(export){write.table(His,file=paste("Areal_measures",".txt",sep=""),row.names=F,sep="\t")}
  return(His)}

#_____________
entr<-function(P){ # R(X), a razão entre a entropia da regionalização obtida e um vetor
uniforme
  shannon.entropy <- function(p)
  {
    if (min(p) < 0 || sum(p) <= 0)
      return(NA)
    p.norm <- p[p>0]/sum(p)
    -sum(log2(p.norm)*p.norm)
  }
  K<-length(unique(P))
  N<-length(P)
  shannon.entropy(table(P))/shannon.entropy(rep(round(N/K),K))
}

#_____________
collapse.data.reg<-function(shp,nb,data,ID,export=FALSE,
                            reg,int_variables,map=FALSE){#função para colapsar as áreas
                            obtidas em regiões exportar um arquivo .shp
  if(missing(nb)){nb<-poly2nb(shp)}
  if(missing(data)){data<-shp@data}
  k<-length(unique(reg))
  which_var<-match(int_variables,names(data))
  VAR<-data.frame(matrix(NA,nrow=k,ncol=length(which_var)));names(VAR)<-names(data)[which_var
  ]
  for (p in 1:length(which_var)){
    VAR[,p]<-xtabs(data[which_var][,p]~reg)
  }
  TAB<-data.frame(
    ID = sort(unique(reg)),
    N_areas =as.vector(table(reg)),
    VAR)
  if(export){
    write.table(TAB,file=paste("SKATER_REGIONS_",ID,".txt",sep=""),sep="\t",row.names=FALSE)
  }
  if(map){
    new_shp<-unionSpatialPolygons(shp, reg)
    jpeg(file=paste("REG_AGG_",ID,"_.jpeg",sep=""))
    plot(new_shp,col=bigpal[sort(unique(reg),decreasing=T)])
    title(ID);if(export){dev.off()}
    data_shp<-TAB
    for (p in 1:ncol(data_shp)){
      dim(data_shp[,p])<-NULL
    }
    new_shp<-SpatialPolygonsDataFrame(new_shp, data_shp, match.ID = TRUE)
```

```
    if(export){
      writePolyShape(new_shp,fn=paste("SHP_",ID,sep=""))
    }
  }
  return(TAB)}

#_____________
automa.reg<-function(design,data,shp,nb,crit,
                     vec.crit,export=FALSE,plot=FALSE,int_variables,map,parent){# wrapper
                     para regionalização automática a partir de uma tabela de desenho
  if(missing(nb)){nb<-poly2nb(shp)}
  result<-data.frame(matrix(NA,ncol=5,nrow=nrow(design)));names(result)<-c("SSE","dSSE",
  "cSSE","CD","ER")
  Regionalizations<-list()
  reg.names<-rep(NA,nrow(design))
  for (i in 1:nrow(design)){
    setwd(paste(parent))
    var.id<-design$Var[i]
    exp.ID <- paste(var.id,"_",design$K[i])
    if(export){
      dir.create(paste("reg_",i,"_",exp.ID))
      setwd(paste("reg_",i,"_",exp.ID))}
    variable<-data[,match(var.id,names(data))]
    if(export=="TRUE" && plot=="TRUE"){
      jpeg(file=paste("SKATER_AREAS_",exp.ID,"_.jpeg",sep=""))
    }
    reg.names[i]<-paste(var.id)
    k<-design$K[i]
    R<-regionalize(shp=shp,nb=nb,data=variable,k,crit,vec.crit,ID=exp.ID,export=export,plot=
    plot)[[1]];if(export){dev.off()}
    Regionalizations[[i]]<-R
    His<- areal.measures(shp=shp,nb=nb,data=variable,reg=R,export=export)
    if(export=="TRUE" && plot=="TRUE"){
      jpeg(file=paste("Boxplot_",exp.ID,"_.jpeg",sep=""))
    }
    boxplot(variable~R,col=bigpal[unique(R)],main=exp.ID)
    if(export){dev.off()}
    #
    result$SSE[i]<-sum(His$SSE)
    result$dSSE[i]<-sum(His$dSSE)
    result$cSSE[i]<-sum(His$cSSE)
    result$CD[i]<-mean(His$cDegree)
    result$ER[i]<-entr(R)
    #
    cdata<-collapse.data.reg(shp=shp,data=data,ID=exp.ID,export=export,
                             reg=R,int_variables=int_variables,map=map)
    ##
  }
  names(Regionalizations)<-as.character(reg.names)
  save(Regionalizations,file=paste(parent,"\\Regionalizations.RData",sep=""))
  #print(Regionalizations)
  graphics.off()
  return(data.frame(design,result))}
#_____________
comp.reg<-function(reg1,reg2,shp,ID,export=FALSE){# função para comparar duas regionalizações
  pal<-rep("yellow",length(reg1))
  pal[which(reg1!=reg2)]<-"red"
```

```
  if(export){jpeg(file=paste("REG_DIFF_", ID, "_.jpeg", sep=""))}
  plot(shp, col=pal)
  legend(x="bottomright", legend=c("Diferente", "Igual"), col=c("red","yellow"), pch=15)
  title(paste(ID))
  return(adjustedRand(reg1, reg2))}

#__________
automa.comp.reg<-function(REG, grid, design, shp, export=FALSE){#Automatizando as comparações
  Comp<-data.frame(matrix(NA, ncol=5, nrow=nrow(grid)))
  names(Comp)<-c("Rand", "HA", "NA", "FM", "Jaccard")
  comp.name<-vector()
  for (i in 1:nrow(grid)){
    comp.name[i]<-paste(names(REG)[[grid[i,1]]], "_", names(REG)[[grid[i,2]]], sep="")
    ID<-paste(paste(design[grid[i,1],]$Var, "_", design[grid[i,1],]$K),
              "vs",
              paste(design[grid[i,2],]$Var, "_", design[grid[i,2],]$K),
              sep=" ")
    Comp[i,]<-comp.reg(reg1=REG[[grid[i,1]]],
                       reg2=REG[[grid[i,2]]],
                       shp=shp,
                       ID=ID,
                       export=export)
  }
  graphics.off()
  row.names(Comp)<-comp.name
  return(Comp)}
```