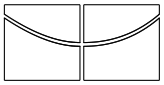


**Universidade de Brasília**

**Instituto de Ciências Exatas**

**Departamento de Estatística**


# Mapeamento ótimo de doenças através da minimização simultânea do viés e da variância

por


**Bárbara de Almeida e Silva Lima de Matos**


Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística


**Orientador: Prof. Dr. André Luiz Fernandes Cançado**


**Dezembro de 2012**

**Bárbara de Almeida e Silva Lima de Matos**

# Mapeamento ótimo de doenças através da minimização simultânea do viés e da variância

**Universidade de Brasília**

**Brasília, Dezembro de 2012**

Termo de Aprovação

Bárbara de Almeida e Silva Lima de Matos

# Mapeamento ótimo de doenças através da minimização simultânea do viés e da variância

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Data da defesa: 18 de dezembro de 2012

Orientador:

---

Prof. Dr. André Luiz Fernandes Cançado

Departamento de Estatística, Universidade de Brasília - UnB

Comissão Examinadora:

---

Prof. Dr. Carlos Henrique Ribeiro Lima

Departamento de Engenharia Civil e Ambiental, UnB

---

Prof. Dr. Luiz Henrique Duczmal

Departamento de Estatística, Universidade Federal de Minas Gerais - UFMG

Brasília, Dezembro de 2012

**MATOS**, BÁRBARA DE ALMEIDA E SILVA LIMA DE
Mapeamento ótimo de doenças através da minimização simultânea do viés e da variância, (UnB - IE, Mestre em Estatística, 2012).

Dissertação de Mestrado - Universidade de Brasília. Departamento de Estatística - Instituto de Ciências Exatas.

1. Mapeamento de doenças 2. Otimização Multiobjetivo 3. Estatística Espacial

*À minha mãe Cristina*

# Agradecimentos


- Ao meu orientador André Luiz Fernandes Cançado;
- Aos professores da banca examinadora Luiz Henrique Duczmal e Carlos Henrique Ribeiro Lima;
- Às professoras Maria Teresa Leão Costa e Claudete Ruas;
- À Dr.ª Iozenita Garcia da Silva Lima e ao Dr. Ricardo Granja Pontes;
- Aos técnicos-administrativos de rede do EST - Cláudia Diogenes Daniel e Sérgio Araújo;
- Aos amigos Daniel Guimarães Pena e Renata Assis de Matos;
- Aos meus chefes René Mallet Raupp e José Joaquim Vieira de Araújo.

# Sumário

# Lista de Figuras

# Lista de Tabelas

## Resumo

Propõe-se uma metodologia para gerar estimativas de taxa de doença nas regiões de um mapa. A construção do estimador baseia-se na minimização simultânea de funções definidas como viés e variância local. O novo procedimento de estimação adota duas abordagens que consideram a informação espacial. As estimativas do novo procedimento são alcançadas por meio do algoritmo *Multi-objective Particle Swarm Optimization* (MOPSO). Foram realizadas comparações entre os resultados da nova metodologia e os provenientes do estimador bayesiano empírico local de Marshall. Conforme os critérios adotados neste trabalho, conclusões satisfatórias foram obtidas com o novo método e superaram os resultados do estimador bayesiano empírico local de Marshall. Uma aplicação com dados de câncer de mama na Nova Inglaterra, Nordeste dos Estados Unidos, é apresentada.



**Palavras-chave:** *Mapeamento de doenças, Otimização Não-linear, Otimização Multiobjetivo, MOPSO, Estatística Espacial.*

**Abstract**

In this work it is proposed a methodology to produce estimates of disease rate in regions of a map. The construction of the estimator is based on the simultaneous minimization of functions defined as bias and local variance. The new estimation procedure adopts two approaches, both of which consider the spatial information. Estimates of the new procedure are achieved through the Multiobjective Particle Swarm Optimization (MOPSO) algorithm. The proposed methods were compared to Marshall's local empirical Bayes estimator. According to the criteria adopted in this work, satisfactory conclusions were obtained and the results indicates that the new method is comparable to, or better than, Marshall's local empirical Bayes estimator. An application to breast cancer data in New England, Northeast of the United States, is presented.

**Key words:** *Mapping Disease, Nonlinear Optimization, Multi-objective Optimization, MOPSO, Spatial Statistics.*

# Capítulo 1

# Introdução

Epidemiologia é o ramo da ciência que estuda, na população, a ocorrência, a distribuição e os fatores determinantes dos eventos relacionados com a saúde. Esta definição é apenas uma das dezenas encontradas na literatura especializada. Em sentido amplo, a epidemiologia é entendida como estudo do comportamento coletivo da saúde e da doença. Sua finalidade geral é concorrer para reduzir os problemas de saúde na população (Pereira, 1995). Sendo assim, o conhecimento da distribuição das doenças, dos fatores que a determinam e das possibilidades de êxito de intervenções converge para a busca do bem-estar da população e, consequentemente, para o alcance desse objetivo.

A descoberta de causas de uma doença assim como a intervenção em uma conjuntura existente julgada insatisfatória constituem, por sua vez, etapas seguintes de um primeiro passo: o conhecimento adequado da situação. A percepção do cenário de saúde de uma população é caracterizada pela determinação de frequências, pelo estudo da distribuição dos eventos e pelo consequente diagnóstico dos principais problemas de saúde ocorridos. O diagnóstico inclui a identificação dos segmentos da população afetados, em maior ou menor proporção, pelos problemas. Em meio a este cenário, os indicadores são medidas-síntese que contêm informação relevante

sobre determinados atributos e dimensões do estado de saúde, bem como sobre o desempenho do sistema de saúde. Por meio de comparações absolutas ou relativas, há uma quantificação da relação entre saúde e doença na população. Os indicadores de mortalidade, de morbidade, indicadores nutricionais, demográficos, sociais, ambientais e de serviços de saúde compõem, por sua vez, exemplos dos principais indicadores de saúde.

O sistema de estatísticas vitais compreende o estudo de nascimentos, casamentos e óbitos. Na área da saúde, seu campo de atuação restringe-se aos nascimentos e, principalmente, aos óbitos. No Brasil, algumas cidades e alguns estados dispõem de dados de natalidade e de mortalidade abrangentes e de relativa qualidade há bastante tempo. Contudo, devido à sua enorme extensão territorial, há dificuldades na obtenção de informações confiáveis para todo o país. As imperfeições existentes não invalidam, todavia, a utilização das estatísticas, as quais podem ser obtidas por meio de estimativas indiretas. Estas são baseadas em recenseamentos e em inquéritos. Já as informações oriundas de registros constituem as estimativas diretas. Seja por meio de estimativas diretas ou indiretas, as declarações de óbito alimentam dois sistemas paralelos de divulgação de estatísticas vitais: o do Instituto Brasileiro de Geografia e Estatística (IBGE) e o do Ministério da Saúde.

Historicamente, o primeiro indicador utilizado em avaliações de saúde coletiva, e ainda hoje o mais empregado, é o de mortalidade (Pereira, 1995). A definição objetiva de morte assim como a obrigatoriedade de seu registro apresentam facilidades operacionais para manutenção da informação. O registro obrigatório resulta na formação de bases de dados, mantidas pelo governo, e proporciona a preparação de estatísticas sob diversas formas. A interpretação destas estimativas, embora possa ser superficial, fornece um diagnóstico da situação.

Os indicadores que expressam a mortalidade da população são numerosos. Em geral, eles referem-se ao que acontece em uma população no período de um ano.

Um dos mais simples indicadores é o coeficiente geral de mortalidade ou taxa bruta de mortalidade. São obtidos de forma muito simples: número total de óbitos ocorridos em uma população e em um determinado período dividido pela quantidade de habitantes no mesmo período. A partir da idéia da taxa bruta de mortalidade, utiliza-se o conceito taxa bruta de doença no desenvolvimento deste trabalho. Além disso, pelo fato de as duas medidas serem uma média geral, as características da taxa bruta de mortalidade são aplicáveis à taxa bruta de doença. Já o numerador desta estatística adapta-se ao contexto da doença em questão.

Sob a suposição de um modelo probabilístico, o número de ocorrências do fenômeno em estudo em uma região pode ser considerado como a realização de uma variável aleatória. Por conseguinte, a taxa bruta de um evento é o estimador mais simples para o risco de ocorrência deste evento. O cálculo desta estimativa em regiões contíguas e a observação da variabilidade do fenômeno constituem o mapeamento de taxas de doenças na área de estatística espacial.

Um grande problema associado ao uso de taxas é a instabilidade que elas demonstram ao expressar o risco de um determinado evento quando ele é raro e a população da região de ocorrência é pequena. As variações bruscas que, porventura, são refletidas pelas taxas podem não estar relacionadas com o fenômeno, mas referem-se à variabilidade das observações. Flutuações aleatórias casuais, como a ocorrência de um ou dois casos do evento a mais ou a menos numa localidade, promovem variações substanciais nas taxas brutas se a sua população for pequena. Por outro lado, este efeito não é verificado em localidades de população grande.

Em um mapa, a mensuração da variabilidade de uma doença por meio da taxa bruta concentra-se na informação de cada região separadamente. Entretanto, muitas regiões são limítrofes entre si ou próximas. Logo, um evento em uma região afeta seus vizinhos à proporção de suas distâncias. Neste sentido, a inserção desta informação na produção de estimativas deve indicar contextos mais informativos

às diversas regiões de um mapa e à condição de saúde como um todo.

A inclusão de informações espaciais no procedimento de estimação promoveu a definição de duas medidas, *viés* e *variância local*. Estas foram construídas a fim de que esses dados fossem considerados. Seguidamente, propõe-se a minimização simultânea destas estatísticas para posterior determinação das taxas de doença. De outra forma, este novo procedimento proposto envolve a obtenção de taxas de doença por meio da minimização simultânea do *viés* e da *variância local*. Juntamente à proposição de um procedimento de estimação de taxas de doença, comparam-se estas com as estimativas bayesianas empíricas do artigo de Marshall (1991).

Esta dissertação está organizada em mais cinco capítulos. Neste capítulo, há um breve histórico da aplicação da taxa bruta de doença ou de mortalidade em alguns artigos. O estimador bayesiano empírico e seus variantes são expostos no capítulo 2. No capítulo 3, situam-se o procedimento proposto e o algoritmo *Multiobjective Particle Swarm Optimization* (MOPSO). Este foi empregado para encontrar as estimativas de doença nas regiões. O capítulo 4 indica duas medidas de performance de métodos e suas interpretações perante o resultado das simulações. Características observadas nas simulações também são constatadas em um problema com dados reais, no capítulo 5. Por fim, estão as seções Considerações Finais e Trabalhos Futuros, no capítulo 6.

## 1.1 Objetivos

### 1.1.1 Objetivo Geral

Desenvolver um procedimento de estimação do vetor de taxas de doença que minimize, simultaneamente, as funções *viés* e *variância local*.

### 1.1.2 Objetivos Específicos

Apresentação de um procedimento geral de estimação de taxas de doenças e de duas abordagens. Em sequência, foram realizadas as ações:

- Implementação das abordagens propostas;
- Obtenção das estimativas de cada método em termos de viés e de variância local;
- Comparação dessas medidas entre si em termos de viés e de variância local;
- Comparação das estimativas de cada método com o estimador bayesiano empírico local proposto em Marshall (1991) em termos de viés e de variância local;
- Aplicação das abordagens em dados de câncer de mama da Nova Inglaterra.

## 1.2 Revisão da Literatura

Inicialmente, a possibilidade de grande variação de taxas é apresentada em Tsutakawa et al. (1985). Neste artigo, as taxas de mortalidade para os tipos de câncer e grupos específicos de idade e sexo são calculadas para um grande número de cidades. A rara ocorrência de mortes por câncer específico na maioria das cidades de tamanhos pequeno e moderado promove a alta dispersão das referidas taxas. Sob a suposição de que o número de mortes segue um processo de Poisson, um método bayesiano empírico é usado para obter taxas ajustadas de mortalidade por câncer. Estas são, portanto, mais estáveis para uso em comparações entre as cidades e para prever as tendências futuras da mortalidade. O método é ilustrado

utilizando dados de câncer de estômago e de bexiga em cidades do Missouri (EUA). Manton et al. (1989), por sua vez, calculam taxas de mortalidade de câncer padronizadas por idade por meio de procedimento empírico bayesiano de dois estágios. Estas taxas também são ajustadas ao tamanho das populações.

A metodologia apresentada por Marshall (1991) estima taxas de mortalidade infantil, corrigidas com base nos valores observados, a partir de conceitos de inferência bayesiana. Primeiro, o estimador bayesiano empírico global calcula uma média ponderada entre a taxa bruta da localidade e a taxa global da região (razão entre o número total de casos e a população total). Já o estimador bayesiano empírico local inclui efeitos espaciais. Seu cálculo utiliza somente os vizinhos geográficos da área na qual se deseja estimar a taxa e, consequentemente, há convergência em direção a uma média local em vez de uma média global. As taxas corrigidas são menos instáveis, pois levam em conta no seu cálculo não só a informação da área, mas também a informação de sua vizinhança.

No artigo de Maiti (1998), o procedimento utilizado é o mapeamento da taxa de mortalidade padronizada em diferentes regiões geográficas (risco relativo). Desta forma, perante diversos tamanhos de população e padrões espaciais desiguais, uma abordagem de bayes hierárquico é apresentada para regularizar os riscos relativos e proporciona as medidas de incerteza associadas a essas estimativas.

Estudo da distribuição da taxa de suicídio na Inglaterra foi apresentado em Saunderson and Langford (1996). Em Ali et al. (2009), o mapeamento dos padrões espaciais de hospitalização por cólera durante os períodos de pré e pós-vacinação em Bangladesh foi realizado. Estes artigos constituem aplicações da estimação empírica bayesiana. No primeiro caso, há comparação dos métodos de mapeamento pela taxa de mortalidade padronizada (risco relativo) e pela estimativa empírica de Bayes. Observou-se superioridade da última medida, pois os riscos relativos mostraram-se altamente dependentes do tamanho da população das áreas

estudadas. Já no segundo artigo, o risco relativo de uma vila é fortemente influenciado pelas estimativas em vilas vizinhas e somente, indiretamente, influenciado pelas estimativas das vilas restantes no mapa.

Adicionalmente à estimação da taxa de mortalidade bruta, alguns procedimentos que envolvem tal medida são revisados ou propostos. A procura por possibilidades de estimar, de forma consistente, heterogeneidades espacial e espacial-temporal, via modelos de mistura não-paramétricos, é apresentada em Böhning (2003). Em Wakefield (2007), tem-se a regressão espacial para estimar a associação entre risco relativo e fatores de risco em potencial.

Neste trabalho, apresenta-se um novo procedimento para obtenção de taxas de doença. Esta metodologia baseia-se na minimização simultânea de duas medidas, as quais são definidas como *viés* e *variância local*. Estas estatísticas foram construídas com o intuito de incluir a informação espacial no cálculo das estimativas de taxa de doença em cada região de um mapa.

# Capítulo 2

# Estudo de estimadores de taxas de doenças

## 2.1 Estimação Bayesiana Empírica

Métodos empíricos de Bayes são utilizados há bastante tempo. Suas raízes podem ser rastreadas para trabalhos na década de 1940. A idéia de uma análise paramétrica bayesiana empírica não é nova, mas o primeiro grande trabalho nesta área apareceu no início da década de 1970 em uma série de trabalhos de Efron e Morris (1972,1973,1975), Casella (1985). Em seu trabalho, Efron and Morris (1975) mostraram que o estimador de James-Stein para $k \geq 3$ médias normais foi superior ao estimador de máxima verossimilhança em termos da função risco ou erro quadrático médio.

Numa configuração simples, o estimador de James-Stein é exposto a seguir. Primeiramente, supor a existência de $k$ parâmetros $\theta_1$, $\theta_2$, $\theta_3$, ..., $\theta_k$, $k \geq 3$. Para

cada $\theta_i$, observam-se variáveis aleatórias normais independentes:

$$X_i/\theta_i \sim N(\theta_i, \sigma^2).$$

Pretende-se estimar o vetor $\theta = \theta_1, \ldots, \theta_k$, com a menor avaliação da medida Erro Quadrático Médio (Perda Quadrática Média):

$$L(\theta, \hat{\theta}) = \sum (\hat{\theta}_i - \theta_i)^2,$$

onde $\hat{\theta} \equiv (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ é a estimativa de $\theta$.

Suponha também que a distribuição a priori de $\theta_i \sim N(\mu, \tau^2)$, $i = 1, \ldots, p$.

A distribuição a posteriori, distribuição de $\theta_i/X_i$, é

$$N\left(\delta^B(X_i), \frac{\sigma^2\tau^2}{(\sigma^2+\tau^2)}\right),$$

$i = 1, \ldots, p$.

Conforme Casella (1985), a estimativa bayesiana para $\theta_i$, $\delta^B(X_i)$, é dada por:

$$\delta^B(X_i) = \frac{\sigma^2}{(\sigma^2+\tau^2)}\mu + \frac{\tau^2}{(\sigma^2+\tau^2)}X_i.$$

Deve-se notar que $\delta^B(X_i)$ é uma média ponderada de $\mu$, a estimativa a priori, e de $X_i$, a estimativa amostral. Na situação empírica bayesiana, as estimativas de $\mu$ e de $\tau^2$ provém dos dados, a distribuição marginal de $X_i$. Sendo assim,

$$f(X_i) \sim N(\mu, \sigma^2 + \tau^2), i = 1, \ldots, p.$$

Obtidas as informações $\mu$ e $\tau^2$, o melhor estimador bayesiano linear de $\theta_i$, em termos da perda quadrática do erro (Efron and Morris,1973), é o estimador de contração (Marshall, 1991):

$$\widehat{\theta_i} = \mu_i + C_i * (X_i - \mu_i),$$

onde:

$$C_i = \frac{Var(\theta_i)}{Var(X)} = \frac{\tau^2}{\sigma^2 + \tau^2}$$

é a razão entre a variância a priori de $\theta_i$ e a variância incondicional de $X_i$.

Os estimadores James-Stein de $\theta_i$ são construídos por considerar $x_i$ como uma amostra da densidade marginal de $X_i$ para estimar $\mu_i$ e $C_i$ na equação anterior (Marshall, 1991).

No artigo de Marshall (1991), a média a posteriori de $\theta_i$ é obtida de forma análoga ao que foi apresentado:

Inicialmente, supor que $\theta_i \sim$ distribuição a priori com média $m_i$ e variância, $A_i$. Neste caso, $\theta_i$ é a taxa de doença da região $i$ e $X_i$ é a proporção de pessoas com certa característica na região $i$.

$$\theta_i \sim (m_i, A_i), i = 1, 2, \ldots, k.$$

e

$$X_i/\theta_i \sim (\theta_i, \theta_i/n).$$

Em termos da perda quadrática média, dados $m_i$ e $A_i$, o melhor estimador linear bayesiano de $\theta_i$ é o estimador de *encolhimento*.

$$\widehat{\delta_i} = m_i + C_i * (x_i - m_i),$$

onde

$$C_i = (A_i)/(A_i + (m_i/n_i)) = var_\theta(\theta)/var_X(X);$$

e

$x_i = p_i$ é a proporção de pessoas com certa característica.

Esta é a razão entre a variância a priori de $\theta_i$ e a variância incondicional de $X_i$. Os estimadores empíricos global e local são apresentados a seguir:

### 2.1.1 Estimador bayesiano empírico global

A taxa bayesiana empírica global é uma média ponderada entre a taxa bruta da localidade e a taxa global da região (razão entre o número total de casos e a população total). Se a localidade apresentar uma população considerável, sua taxa apresentará pequena variabilidade e ela permanecerá praticamente inalterada. Se, por outro lado, a localidade apresentar uma população pequena, a estimativa da taxa bruta terá grande variância e pouco peso será atribuído a essa taxa. A distribuição a priori para $\theta_i$ em todas as estimativas bayesianas são não-espaciais; isto é, a média e a variância a priori são estabelecidas como constantes para todas as áreas.

Seja uma região particionada em $N$ áreas, indexadas por $i$ (i=1, ..., N). O estimador global para a taxa na área $i$, $\widehat{\theta}_i$, é:

$$\widehat{\theta}_i = \widetilde{m} + \widehat{C}_i(x_i - \widetilde{m}),$$

onde

$$\widehat{C}_i = \frac{s^2 - (\widetilde{m}/\bar{n})}{s^2 - (\widetilde{m}/\bar{n}) + (\widetilde{m}/n_i)}$$

e

$$\widehat{\theta}_i = \widetilde{m}, \text{quando } s^2 < (\widetilde{m}/\bar{n}),$$

em que:

- $\widetilde{m}$ é a taxa média global;
- $s^2$ é a variância amostral;
- $N$ é o número de regiões;
- $n_i$ é a população da área $i$ ou número de pessoas sob risco na área $i$;

- $n = \sum n_i$ é a quantidade total de pessoas sob risco;
- $\bar{n} = n/N$ é a média de pessoas sob risco por região;
- $x_i$= $p_i$ é a proporção de pessoas com certa característica.

### 2.1.2 Estimador bayesiano empírico local

O estimador bayesiano empírico pode ser generalizado para incluir efeitos espaciais, ao se exigir que a estimativa ajustada para uma área se aproxime de uma média de sua vizinhança em vez de uma média global. Considera-se como vizinhança da área $i$ todas as demais áreas que compartilham fronteira com a $i$-ésima área além da própria área. Com isso, adiciona-se uma suavidade espacial ao modelo, pois as estimativas bayesianas globais não variam segundo a configuração espacial das áreas. Para o cálculo de estimativas bayesianas empíricas locais, modifica-se a distribuição a priori de $\theta_i$ para permitir que a média e a variância sejam relacionadas à vizinhança de $i$, em vez de permanecerem constantes para todas as áreas. Então, a taxa observada em pequenas populações irá convergir para uma média local em vez de uma média global. O estimador $\widetilde{\theta}_i$ fica da seguinte forma:

$$\widetilde{\theta}_i = \widetilde{m}_i + \widetilde{C}_i(x_i - \widetilde{m}_i),$$

onde

$$\widetilde{C}_i = \frac{s_i^2 - (\widetilde{m}_i/\bar{n}_i)}{s_i^2 - (\widetilde{m_i}/\bar{n}_i) + (\widetilde{m}_i/n_i)}$$

e

$$\widehat{\theta}_i = \widetilde{m}_i, \text{quando } s_i^2 < (\widetilde{m}_i/\bar{n_i}),$$

Neste caso:

- $\widetilde{m}_i$ é a taxa da doença sobre a vizinhança $i$;
- $s_i^2$ é a variância na vizinhança $i$;
- $n_i$ é a população da área $i$ ou número de pessoas sob risco na área $i$;
- $\bar{n_i}$ é a média da população dos vizinhos;
- $x_i = p_i$ é a proporção de pessoas com certa característica.

# Capítulo 3

# Mapeamento de doenças

O perfil da população de uma região ou de várias regiões pode ser traçado por meio de demandas de saúde. O panorama resultante possibilita o direcionamento de políticas públicas conforme as necessidades constatadas. Para que estas políticas sejam implementadas, as informações obtidas são apresentadas por meio de indicadores. Seus elementos principais são os quantitativos da população e de casos de doença; o aperfeiçoamento do indicador em determinado contexto depende do objetivo do pesquisador. Em relação às variáveis demográficas, parte-se da premissa de que elas já foram consideradas na distribuição da população das regiões, motivo pelo qual este tópico não é comentado neste trabalho.

Seja o mapa da figura 3.1. Ele está dividido em $n$ regiões e cada ponto representa o centróide. Suponha que o número de eventos na $i$-ésima região possa ser representado por uma variável aleatória $C_i$ cujo valor observado é $c_i$, $i = 1, \ldots, n$. Com a população $n_i$ da $i$-ésima região, o estimador mais simples é a taxa bruta. Esta medida é representada pela quantidade $\tilde{\theta}_i = c_i/n_i$.

Com as duas informações disponíveis por região, propõe-se que a estimativa da taxa de uma região receba influência das taxas das áreas vizinhas nas suas devidas

Figura 3.1: Mapa dividido em regiões e seus respectivos centróides

proporções.

Em problemas desta natureza, um bom conjunto de estimativas $\hat{\boldsymbol{\theta}}$ deveria incluir informação a respeito da estrutura espacial do mapa. Sob a hipótese de correlação espacial, deve haver forte dependência entre as taxas correspondentes a regiões próximas. Por outro lado, supõe-se fraca dependência, ou sua ausência, entre taxas de regiões distantes. Assim, regiões vizinhas devem ter taxas parecidas, enquanto o contrário é esperado entre locais distantes. Procura-se, então, um conjunto de estimativas $\hat{\boldsymbol{\theta}}$ que também minimize a variância entre as regiões próximas. Neste trabalho, esta medida será denominada variância local e seus principais componentes são a existência de vizinhança de cada região e a quantidade de vizinhos. Em seguida, são definidas as funções *viés* e *variância local*.

A função *viés* é definida da seguinte forma:

$$B(\hat{\theta}) = \sum_{i=1}^{n} |\hat{\theta}_i - \tilde{\theta}_i|,$$

onde $\tilde{\theta}_i$ é a taxa bruta observada na região $i$ e $\hat{\theta}_i$ é a estimativa da taxa da região $i$. Sabe-se que $B(\hat{\theta})$ assume seu mínimo quando $\hat{\theta}_i = \tilde{\theta}_i$.

Em relação à existência e à quantidade de vizinhos, a matriz de vizinhanças $W$

indica o posicionamento relativo das regiões entre si e é simétrica. As regiões $i$ e $j$ são vizinhas se $w_{i,j} = 1$. Caso contrário, $w_{ij} = 0$. A diagonal principal da matriz $W$ é igual a 1.

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{pmatrix}$$

Previamente à exposição do conceito de variância local, apresenta-se a definição de vizinhança da região $i$. Consideremos a vizinhança da região $i$ como sendo $S_i$, ou seja, o conjunto de regiões vizinhas à $i$-ésima região, incluindo ela própria, $i = 1, \dots, n$.

A variância local é assim definida:

$$V(\hat{\theta}) = \sum_{i=1}^{n} \frac{(\hat{\theta}_i - \overline{\theta}_i)^2}{n},$$

onde $\hat{\theta}_i$ é a estimativa da taxa $i$ e $\overline{\theta}_i$ é a média das taxas na vizinhança da região $i$, isto é

$$\overline{\theta}_i = \frac{\sum_{j=1}^{n} \tilde{\theta}_j I(j \in S_i)}{|S_i|},$$

em que $|S_i|$ denota o número de elementos em $S_i$:

$$|S_i| = \sum_{j=1}^{n} w_{ij},$$

onde $w_{i,j} = 1$ se as regiões $i$ e $j$ são vizinhas; caso contrário, $w_{ij} = 0$.

Além disso, $I(j \in S_i)$ é igual a 1 se a região $j$ está na vizinhança da região $i$ e zero, caso contrário.

## 3.1 Definição do problema

A questão apresentada envolve encontrar um vetor de estimativas $\hat{\boldsymbol{\theta}}$ que concilie dois critérios simultaneamente. Tem-se, portanto, um problema de otimização cuja função objetivo é $\mathbf{f} = (B, V)$. Este problema está condicionado também a restrições, representadas de forma genérica por $g(.)$ e $h(.)$. Estas serão especificadas ao longo do trabalho.

$$\boldsymbol{x}^* = arg\min_{\boldsymbol{x}} \quad \mathbf{f}(\boldsymbol{x}) = (f_1(x), \ldots, f_n(x))$$

$$\text{s.a.} \quad \begin{cases} \boldsymbol{g}(\boldsymbol{x}) \leq 0 \\ \boldsymbol{h}(\boldsymbol{x}) = 0 \end{cases}$$

onde $\boldsymbol{x}^*$ é o vetor que minimiza $\mathbf{f}(\boldsymbol{x})$ e $\boldsymbol{g}$ e $\boldsymbol{h}$ são funções vetoriais, isto é $g(x) = (g_1(x), \ldots, g_r(x))$ e $h(x) = (h_1(x), \ldots, h_s(x))$.

## 3.2 Metodologia Proposta

A estimativa $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \ldots, \hat{\theta}_n\}$ para o vetor de taxas $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ deve satisfazer a condição de minimização simultânea das funções *viés* e *variância local*.

Suponhamos que $\hat{\theta}_i$, a estimativa da taxa de doença na região $i$, seja uma combinação linear das taxas brutas observadas: $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$. Isto é:

$$\hat{\theta}_i = \alpha_{i1}\tilde{\theta}_1 + \cdots + \alpha_{in}\tilde{\theta}_n, \quad i = 1, \ldots, n$$

O procedimento de estimação consiste em obter um vetor de coeficientes $\boldsymbol{\alpha_i} = \{\alpha_{i1}, \ldots, \alpha_{in}\}$, onde $i$ refere-se à $i$-ésima região, $i = 1, \ldots, n$.

Cada vetor $\boldsymbol{\alpha_i}$ corresponde a uma região do mapa. Na fórmula da estimativa, o elemento $\alpha_{ij}$ representa a intensidade da influência das informações da região

$j$ na composição do estimador da $i$-ésima região. Em sequência, a junção dos $n$ vetores $\boldsymbol{\alpha_i}$ compõe a matriz de coeficientes $A = \{\alpha_{ij}\}$.

$$\mathbf{A} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \\ \alpha_{n1} & \alpha_{n2} & \dots & \alpha_{nn} \end{pmatrix}$$

A seguir, estão descritas as duas variações para a obtenção da matriz de coeficientes $A = \{\alpha_{ij}\}$.

### Abordagem 1

Seja $f_i(\cdot)$ a função de densidade de uma distribuição normal bivariada, com média $\mu_i = [x_i \quad y_i]'$ (ou seja, a média é dada pelo centróide da $i$-ésima região) e a matriz de variância-covariância

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

Define-se

$$\alpha_{ij} = \frac{f_i(x_j, y_j)}{\sum_{k=1}^{n} f_i(x_k, y_k)}$$

Esta modelagem é ilustrada na Figura 3.2. Com os $\alpha_{ij}$'s assim definidos, $\tilde{\theta}_i$ sempre é o termo de maior peso na combinação:

$$\hat{\theta}_i = \alpha_{i1}\tilde{\theta}_1 + \cdots + \alpha_{iN}\tilde{\theta}_n,$$

pois $f_i(\cdot)$ assume seu máximo em $(x_i, y_i)$.

De modo geral, a taxa de uma região $i$ será tão mais influenciada por dados de uma região $j$ quanto mais próximos estiverem seus centróides $(x_i, y_i)$ e $(x_j, y_j)$. Isto é, os pesos $\alpha_{ij}$ serão maiores à medida que as distâncias entre os centróides

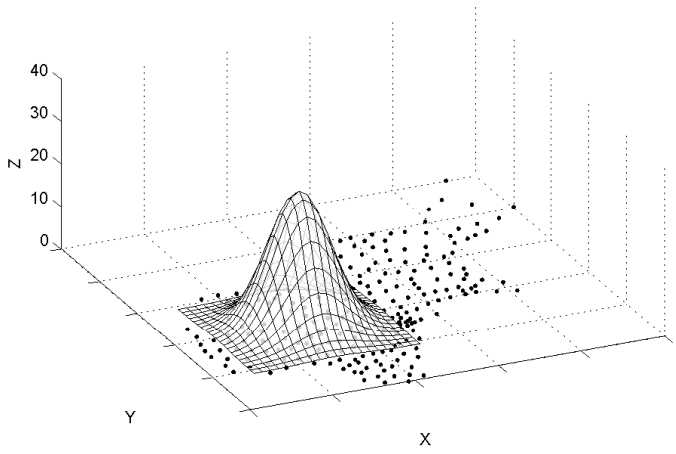


Figura 3.2: Modelagem da variabilidade pela densidade normal; o vetor de médias de cada região é seu respectivo centróide e a matriz de covariância é a identidade

são menores. Dessa forma, incorpora-se à estimativa $\hat{\theta}_i$ a estrutura espacial, uma vez que os $\alpha$'s dependem da distância entre as regiões $i$ e $j$. A relação inversa entre distância dos centróides e peso das taxas está ilustrada na Figura 3.3.

Como o centróide constitui uma das poucas informações do problema em questão, suas coordenadas também influenciam na composição da matriz $\Sigma$. Dependendo da escala do mapa, os valores de referência de um local (ou o vetor de médias da normal bivariada) são extremamente grandes. Logo, há reflexo na magnitude das variâncias assim como dificuldade de interpretação dos dados. As coordenadas dos centróides, as médias, foram então padronizadas em escala de zero a um.

Com a padronização das médias, a variância $\sigma^2$ assume valores entre zero e uma constante $v > 0$. Interessante atentar que, quando $\sigma \to 0$, então $\hat{\theta}_i \to \tilde{\theta}_i$. Isto é, essa estimativa tem baixo viés e alta variância local. Por outro lado, quando $\sigma \to \infty$, as taxas tendem todas ao mesmo valor (à média das taxas). Na prática,

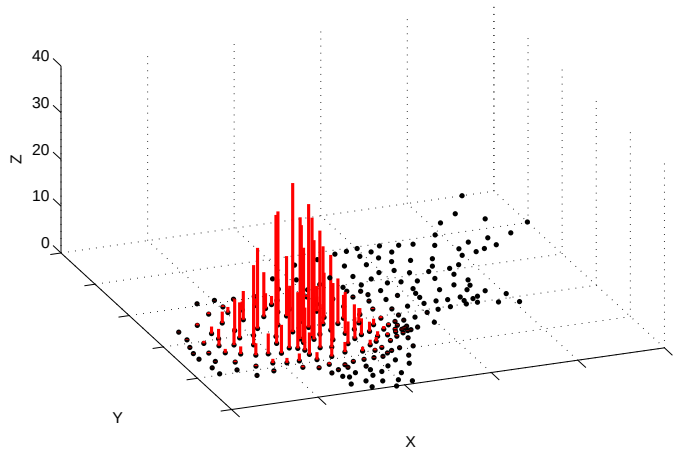


Figura 3.3: Exemplificação da relação entre pesos e distância entre os centróides

como as coordenadas dos centróides estão padronizadas entre 0 e 1, a adoção de um limite superior $v = 1$ é suficiente. Quando $\sigma$=1, a densidade será praticamente constante sobre o mapa.

**Abordagem 2**

Nesta abordagem, a estimativa $\hat{\theta}_i$ recebe influência unicamente das taxas das regiões vizinhas; há também ponderação pelas distâncias destas em relação à região $i$. Novamente, os pesos são caracterizados pela densidade da normal, centrada em $(x_i, y_i)$ e com variância $\sigma^2$. Nesta situação, a matriz $A$ de coeficientes teria elementos dados por:

$$\alpha_{ij} = \frac{f_i(x_j, y_j) I(j \in S_i)}{\sum_{k=1}^{N} [f_i(x_k, y_k) I(k \in S_i)]}$$

## 3.3 Obtenção das estatísticas *viés* e *variância*

As estatísticas $B$ e $V$ constituem o passo final de todo o processo de estimação do vetor $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$. A obtenção destas estimativas depende dos coeficientes $\alpha_{ij}$'s, os quais são função do vetor $\boldsymbol{\sigma}$ da densidade da distribuição normal. Em decorrência disto, $\boldsymbol{\theta}$ é caracterizado como uma função de $\boldsymbol{\sigma}$. Este assunto será tratado nas próximas seções.

### 3.3.1 Determinação do vetor $\boldsymbol{\sigma}$

Nas duas abordagens propostas, a estimativa da taxa de doença de uma região é a combinação linear das taxas observadas de outras regiões. A intensidade da influência de uma região em outra é caracterizada pelos $\alpha_{ij}$'s, os quais são delineados pela densidade da distribuição normal bivariada.

Em se tratando do parâmetro $\sigma$ da densidade das normais bivariadas, as seguintes definições foram convencionadas: a matriz $\Sigma$ é diagonal e os $\sigma$'s das duas variáveis aleatórias normais marginais são idênticos. Devido a esta igualdade, diz-se que cada região apresenta um $\sigma$. A determinação do vetor $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ constitui a incógnita para encontrar o melhor método de estimação. A configuração mais simples para o mapa ocorre quando este vetor é composto de elementos iguais. De outra forma, um vetor $\boldsymbol{\sigma}$ cuja composição é arbitrária também indica alternativas de solução para $\hat{\boldsymbol{\theta}}$. Ao longo do texto, deve-se atentar para duas notações: a letra em negrito $\boldsymbol{\sigma}$ refere-se a um vetor, ao mesmo tempo em que $\sigma$ corresponde a uma coordenada deste vetor.

### Combinação simples do vetor $\boldsymbol{\sigma}$

No modelo mais simples de composição do vetor $\boldsymbol{\sigma}$, há o seguinte problema de otimização.

$$\boldsymbol{\sigma}^* = arg\min_{\boldsymbol{\sigma}} \quad \mathbf{f}(\boldsymbol{\sigma}) = (B(\boldsymbol{\sigma}), V(\boldsymbol{\sigma}))$$

$$\text{s.a.} \quad \begin{cases} \sigma_i = \sigma \quad i = 1, \ldots, n \\ \sigma > 0 \end{cases}$$

Neste problema, a restrição $\sigma_i = \sigma$ implica que, em todas as regiões, o mesmo valor de $\sigma$ é usado na densidade normal bivariada para obtenção das estimativas.

Uma breve ilustração da versão simples do vetor $\boldsymbol{\sigma}$ é apresentada na figura 3.4. Neste gráfico, encontram-se os pares $(B, V)$ resultantes do emprego dos dois métodos propostos em apenas uma amostra. Especificamente, são 200 pares $(B, V)$ os quais foram originados de 200 valores de $\sigma$ da matriz $\Sigma$.

Pela Figura 3.4, o resultado proveniente dos dois métodos é semelhante quanto à distribuição das duas funções - $B$ e $V$. Contudo, parece não haver par ordenado que minimize ambas as funções simultaneamente. Quando o viés tende a diminuir, a variância aumenta e vice-versa. A resolução desta questão não engloba a simples comparação entre dois escalares - relação de superioridade, de inferioridade ou de igualdade. Torna-se necessário, então, observar as coordenadas $(B, V)$ consoante os valores da função-objetivo $f(\boldsymbol{\sigma}) = (B(\boldsymbol{\sigma}), V(\boldsymbol{\sigma}))$.

A avaliação conjunta entre as funções-objetivo e suas coordenadas remete à busca pelas melhores soluções, já que não é possível encontrar aquela melhor. Neste contexto, os conceitos de *dominância*, de *solução Pareto-ótima* e de *conjunto Pareto-ótimo* são pertinentes.

**Definição 3.3.1** (Dominância)**.** *: Seja* $\mathbf{f}(\boldsymbol{\sigma}) = (B(\boldsymbol{\sigma}), V(\boldsymbol{\sigma}))$ *uma função definida*

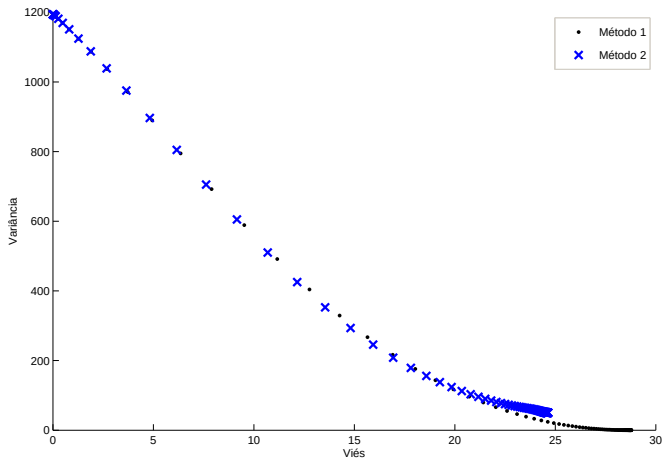


Figura 3.4: Distribuição dos pares *(B,V)* conforme método - simulação com 1 amostra

*em um espaço* $\Re^n$. *Um ponto* $\boldsymbol{\sigma_1} \in \Re^n$ ***domina*** *outro ponto* $\boldsymbol{\sigma_2} \in \Re^n$, *(denota-se* $\boldsymbol{\sigma_1} \prec \boldsymbol{\sigma_2}$*), se*

$$B(\boldsymbol{\sigma}_1) \leq B(\boldsymbol{\sigma}_2) \text{ e } V(\boldsymbol{\sigma}_1) \leq V(\boldsymbol{\sigma}_2)$$

*e se, pelo menos, uma das desigualdades é estrita, isto é:*

$$B(\boldsymbol{\sigma}_1) < B(\boldsymbol{\sigma}_2) \text{ ou } V(\boldsymbol{\sigma}_1) < V(\boldsymbol{\sigma}_2)$$

Quando não existe relação de dominância entre duas soluções, estas são denominadas soluções incomparáveis ou indiferentes.

A figura 3.5 exemplifica o conceito de dominância. Sejam 5 soluções para o vetor $\boldsymbol{\sigma}$: $\boldsymbol{\sigma_1}$, $\boldsymbol{\sigma_2}$, $\boldsymbol{\sigma_3}$, $\boldsymbol{\sigma_4}$ e $\boldsymbol{\sigma_5}$. Os pontos $P_i = (B(\sigma_i), V(\sigma_i))$, $i = 1, \ldots, 5$ representam os valores de viés e de variância obtidos por estas soluções.

Constata-se que a solução $\boldsymbol{\sigma_2}$ domina as soluções $\boldsymbol{\sigma_1}$, $\boldsymbol{\sigma_3}$, $\boldsymbol{\sigma_4}$ e $\boldsymbol{\sigma_5}$. Por outro lado, as soluções $\boldsymbol{\sigma_3}$ e $\boldsymbol{\sigma_5}$ são consideradas soluções incomparáveis ou indiferentes

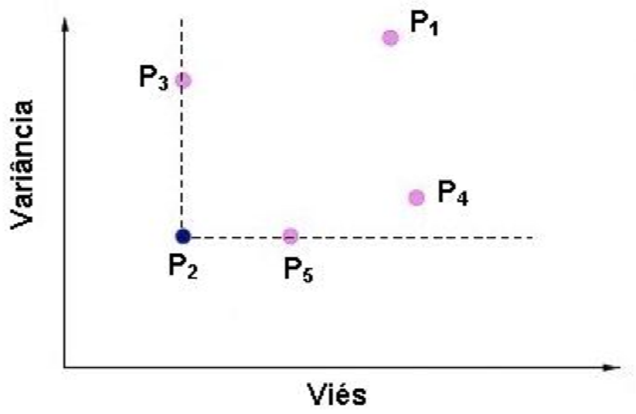


Figura 3.5: Relação de dominância em um espaço bi-objetivo

entre si. A solução $\boldsymbol{\sigma_5}$ domina as soluções $\boldsymbol{\sigma_1}$ e $\boldsymbol{\sigma_4}$, enquanto a solução $\boldsymbol{\sigma_3}$ domina a solução $\boldsymbol{\sigma_1}$.

**Definição 3.3.2** (Solução Pareto-ótima)**.** *: Uma solução* $\boldsymbol{\sigma} \in \Re^n$ *é Pareto-ótima, ou não-dominada, se não existe* $\boldsymbol{\sigma'} \in \Re^n$ *tal que* $\boldsymbol{\sigma'}$ *domina* $\boldsymbol{\sigma}$*.*

**Definição 3.3.3** (Conjunto Pareto-ótimo ou Fronteira de Pareto)**.** *: É aquele conjunto* $\boldsymbol{\sigma}$ *do espaço dos parâmetros formado por todas as soluções Pareto-ótimas.*

O conjunto de soluções $\widehat{\sigma^*}$ não-dominadas compõem o conjunto Pareto-Ótimo. Este, por sua vez, constitui a solução ótima de um problema de otimização e sua obtenção é o objetivo deste trabalho.

### Vetor $\boldsymbol{\sigma}$ livre

Na suposição de um modelo mais razoável para a composição do vetor $\boldsymbol{\sigma}$, o problema de otimização é apresentado da seguinte forma.

$$\boldsymbol{\sigma}^* = arg\min_{\boldsymbol{\sigma}} \quad \mathbf{f}(\boldsymbol{\sigma}) = (B(\boldsymbol{\sigma}), V(\boldsymbol{\sigma}))$$

$$\text{s.a.} \quad \{\sigma_i > 0, \quad i = 1, \ldots, n$$

Em comparação à primeira alternativa de composição do vetor $\boldsymbol{\sigma}$, removeram-se as restrições de igualdade para que fosse possível o uso de $\sigma$'s diferentes em regiões diferentes. Isto, por sua vez, inviabiliza uma busca sistemática no espaço de $\boldsymbol{\sigma}$'s como feito anteriormente. Como exemplo, suponhamos que haja 300 regiões em um mapa e que 200 valores de $\sigma$ fossem observados em cada região. Haveria, portanto, $200^{300}$ opções para a composição do vetor $\boldsymbol{\sigma}$. A avaliação de cada um desses vetores é uma tarefa inviável. Mediante esta dificuldade, aplica-se uma heurística para resolver o problema de otimização multiobjetivo.

## 3.4 Otimização

Segundo o Novo Dicionário Aurélio, otimização é o *processo pelo qual se determina o valor ótimo de uma grandeza.* Este processo é caracterizado como otimização linear ou não-linear e a grandeza refere-se ao delineamento do problema: função objetivo de interesse sujeita ou não a restrições. A programação linear trata do problema de minimizar ou maximizar uma função linear na presença de restrições lineares de igualdade e/ou desigualdade. Contudo, muitos problemas reais não podem ser modelados adequadamente como um problema linear devido à não-linearidade da função objetivo e/ou à não-linearidade de alguma de suas restrições (Bazaraa and Shetty, 1979).

Neste trabalho, será adotada a convenção de que um problema de otimização é representado por um problema de minimização. Caso se deseje maximizar uma

função, a conversão para um problema de minimização ocorre por meio da multiplicação dos coeficientes da função objetivo por (-1).

Dado um certo problema de otimização, procura-se o vetor $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]'$ que forneça a melhor solução possível no sentido de minimizar uma função. Contudo, o domínio da função nem sempre compreende todos os valores da variável $\boldsymbol{x}$. Situações deste tipo são exemplos de otimização restrita; de outra forma, a otimização é irrestrita.

Sob a premissa de que a otimização é restrita, um modelo de problema de otimização está apresentado a seguir:

$$\boldsymbol{x}^* = arg\min_{\boldsymbol{x}} \quad \mathbf{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$$

$$\text{s.a.} \quad \begin{cases} \boldsymbol{g}(\boldsymbol{x}) \leq 0 \\ \boldsymbol{h}(\boldsymbol{x}) = 0, \end{cases}$$

onde $\boldsymbol{x}$ representa o vetor de argumentos de entrada e $\boldsymbol{x}^*$, o vetor de argumentos cuja imagem $f(x)$ é mínima.

Neste modelo, há uma função $\mathbf{f}$ de $n$ variáveis cuja resposta é um vetor $(m \times 1)$; $\boldsymbol{g}(.)$: $\Re^n \rightarrow \Re^r$ e $\boldsymbol{h}(.)$: $\Re^n \rightarrow \Re^s$.

No contexto da otimização restrita, os conceitos de região factível e de ponto factível são pertinentes.

**Definição 3.4.1** (Região factível)**.** *: Conjunto dos pontos do espaço $\Re^n$ que satisfaz, simultaneamente, a todas as restrições (de desigualdade e de igualdade). Também chamada de conjunto factível ou de conjunto viável.*

**Definição 3.4.2** (Ponto factível)**.** *: Ponto pertencente à região factível.*

No caso de problemas de otimização restrita, a função objetivo $\mathbf{f}$ tem sentido ou a solução encontrada é viável (ponto viável) apenas na região do domínio da

função, ou seja, na região factível. Caso a solução esteja fora da região factível (ponto infactível), ela é considerada inviável.

Após a definição do problema e de suas restrições caso existam, a solução ótima é encontrada por meio do emprego/da utilização de alguma técnica de otimização. Todavia, o conhecimento das características da função-objetivo ou do vetor de funções-objetivo faz toda a diferença tanto no tempo despendido para encontrar a solução quanto na sua existência. Modalidade, diferenciabilidade, convexidade, linearidade constituem os atributos da função-objetivo para escolha de uma técnica de otimização e, consequentemente, obtenção mais rápida da solução. Um exemplo simples é aquele em que a função objetivo tem apenas um mínimo, é diferenciável, linear e convexa. Os pontos em que a derivada desta função se anula são facilmente encontrados.

A ausência de algum desses atributos na função-objetivo origina as dificuldades na sua minimização em problemas de otimização reais. Neste grupo, encontra-se a função-objetivo deste trabalho $\mathbf{f} = (B, V)$.

Os métodos de resolução de um problema de otimização não-linear são brevemente apresentados.

### 3.4.1 Otimização Não-Linear

A otimização não-linear trata de problemas em que a função-objetivo ou alguma das restrições do problema são funções não-lineares das variáveis envolvidas. Existem diversos métodos para resolver um problema de otimização não-linear, que podem ser divididos em métodos de direção de busca, métodos de exclusão de semi-espaços, métodos de otimização por populações.

Os primeiros métodos de otimização (minimização) de funcionais não-lineares foram desenvolvidos a partir da idéia básica de fazer o algoritmo evoluir encon-

trando novos pontos situados em direções para as quais o funcional decresça em relação ao ponto corrente (Takahashi, 2007). Estes são denominados métodos de direção de busca. Utilizam como informação primordial o gradiente e/ou a Hessiana da função-objetivo.

Os métodos de exclusão de semi-espaços são mais adequados quando a função-objetivo é unimodal, mas não é diferenciável. Para *compensar* a ausência deste atributo, acrescenta-se a premissa de existência de convexidade para a função-objetivo.

Por fim, os métodos de otimização por populações são adequados para problemas mais complexos, em que não há garantias sobre convexidade, unimodalidade, continuidade ou diferenciabilidade. Sua denominação provém do trabalho em conjunto de um grupo de soluções iniciais ou de partículas, a população. Desde a população inicial até aquela denominada como conjunto-solução, diversas populações são geradas a partir da troca de informações entre os componentes de uma população. A cada atualização de dados, uma nova solução-candidata ou população é formada e sua avaliação, consoante alguns critérios, definirá o prosseguimento do algoritmo (nova população) ou seu término.

### 3.4.2 Otimização por Enxame de Partículas

O algoritmo de Otimização por Enxame de Partículas (*Particle Swarm Optimization (PSO)*) é um método de otimização por populações, adequado para problemas não-lineares em variáveis contínuas. Foi desenvolvido por meio da simulação de um modelo social simplificado (Kennedy and Eberhart, 1995). A intenção original era simular a coreografia de um bando de pássaros ou seu comportamento social perante as adversidades do momento, tais como predadores, temperatura ambiental, existência de colisão entre eles, procura por comida etc.

De forma análoga à atuação dos pássaros na busca por um certo alvo, as partículas ou agentes *caminharão* em direção à solução ótima de uma função mono-objetivo. O espaço de busca é o $\Re^n$. O progresso das partículas nesta jornada ocorrerá mediante tantas avaliações de sua velocidade quanto o número de componentes da variável de entrada do problema; neste caso, o vetor $\boldsymbol{\sigma}$ contém $n$ incógnitas. A celeridade deste movimento está condicionada, por sua vez, à posição dos pontos consoante os aspectos: local onde a partícula apresentou a sua melhor avaliação da função-objetivo até o momento e lugar em que houve a melhor avaliação da função-objetivo por qualquer partícula em qualquer período. Este caso configura a organização mais simples de partículas: todas formam um único grupo. Neste trabalho, será considerada a configuração mais simples entre os agentes, embora existam diversos agrupamentos possíveis entre as partículas. E esta variedade influenciará o andamento do processo como um todo.

Na primeira iteração, a melhor posição em que cada partícula já esteve é a sua posição atual. O valor da função-objetivo de cada agente e o melhor valor desta função são mantidos. As partículas possuem velocidades iniciais. Nas iterações que se seguem, as avaliações da função-objetivo serão feitas na nova posição em que as particulas estiverem. Consequentemente, as novas velocidades resultarão da combinação de dados entre cada partícula em seus diversos períodos e a informação conjunta de todos os elementos até o momento. A velocidade e a posição das partículas são atualizadas para o prosseguimento do processo.

No tocante às informações individuais das partículas (posição e o respectivo valor da função-objetivo), elas compõem o domínio da variável aleatória *Melhor posição pessoal*. Esta variável se assemelha a uma memória autobiográfica à medida em que cada indivíduo lembra sua própria experiência. O ajuste da velocidade associado à variável *Melhor posição pessoal* é chamado de *nostalgia simples*. Portanto, cada indivíduo/agente tende a retornar ao lugar (ou ficar no local) onde a

função-objetivo alcançou a melhor avaliação.

A pesquisa do alvo ou da solução ótima inclui também cooperação das partículas entre si. Em cada iteração, elas disponibilizam informação sobre sua posição e o respectivo valor da função-objetivo. Uma vez que o conhecimento é difundido, os agentes convergem para a posição em que o resultado da função-objetivo é o melhor dentre todos os valores alcançados em todas as iterações. Este local é denominado *Melhor posição global* e é similar ao conhecimento público ou norma padrão que os agentes pretendem atingir/alcançar.

O encadeamento das informações individuais e conjuntas proporciona uma maior ou menor velocidade aos agentes do processo em direção à solução ótima. A expressão que dá origem, na $t$-ésima iteração, à velocidade das partículas é:

$$\boldsymbol{v}_i^t = \boldsymbol{v}_i^{t-1} + \phi_1 u_1(\boldsymbol{p_i} - \boldsymbol{x_i^t}) + \phi_2 u_2(\boldsymbol{g} - \boldsymbol{x_i^t}),$$

onde:

- $\boldsymbol{v}_i^{t-1}$ é a velocidade da $i$-ésima partícula no tempo $(t-1)$;
- $\phi_1$ e $\phi_2$ são taxas de cognição;
- $u_1$ e $u_2$ são números aleatórios entre 0 e 1;
- $\boldsymbol{p}_i$ é a melhor posição da $i$-ésima partícula;
- $\boldsymbol{g}$ é a melhor posição global;
- $\boldsymbol{x}_i^t$ é a posição da partícula $i$ no tempo $t$.

A atualização da posição das partículas é feita por meio da seguinte composição:

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \boldsymbol{v}_i^t$$

Os quatro vetores $\boldsymbol{v}_i^t$, $\boldsymbol{x}_i^t$ e $\boldsymbol{p_i}$ e $\boldsymbol{g}$ são $n$-dimensionais, onde $n$ é a dimensão do espaço de busca.

Valores elevados do incremento da *Melhor posição pessoal* em relação ao incremento da *Melhor posição global* resultam em dispersão de indivíduos isolados através do espaço do problema, enquanto o contrário resulta em aglomeração prematura dos agentes em direção aos mínimos locais. Assim, valores aproximadamente iguais dos dois incrementos tendem a resultados mais eficazes no domínio do problema.

No que tange às taxas de cognição $\phi_1$ e $\phi_2$, elas representam a energia que empurra cada partícula em direção à *Melhor posição pessoal* e à *Melhor posição global*, respectivamente. Além disso, o uso da velocidade da iteração anterior impõe um *momentum* à partícula. Isto evita que o algoritmo convirja de forma precoce a ótimos locais.

A seguir, apresenta-se o pseudocódigo do PSO:

1. Gerar partículas em posições aleatórias e com velocidades aleatórias;
2. Avaliar a função, que será minimizada, nas posições das partículas;
3. Atualizar a *Melhor posição pessoal* de cada partícula até o momento, quando a posição atual for melhor;
4. Atualizar a *Melhor posição global* até a iteração realizada;
5. Atualizar as velocidades e as posições das partículas;
6. Voltar ao passo 2 e repetir o procedimento até o passo 5 enquanto um critério pré-estabelecido não for alcançado.

## 3.5 Otimização Multiobjetivo

A otimização multiobjetivo abrange a minimização conjunta de mais de uma função e seu intuito é achar uma solução única que englobe o melhor compromisso entre os múltiplos objetivos. Mas, usualmente, existe mais de uma solução com essas características. A seleção da solução com o melhor compromisso requer a opinião do pesquisador ou do tomador de decisões (Grosan, 2003).

Em termos gerais, um problema de otimização multiobjetivo é descrito da seguinte forma:

$$\boldsymbol{x}^* = arg\min_{\boldsymbol{x}} \quad \mathbf{f}(\boldsymbol{x}) = (\boldsymbol{f}_1(x), \boldsymbol{f}_2(x), \ldots, \boldsymbol{f}_m(x))$$

$$\text{s.a.} \quad \begin{cases} \boldsymbol{g}(\boldsymbol{x}) \leq 0 \\ \boldsymbol{h}(\boldsymbol{x}) = 0 \end{cases}$$

onde $\boldsymbol{g}(\boldsymbol{x})$ e $\boldsymbol{h}(\boldsymbol{x})$ são funções vetoriais. No contexto deste trabalho, $\mathbf{f}(\boldsymbol{x}) = (\boldsymbol{f}_1(x), \boldsymbol{f}_2(x)) = (\boldsymbol{B}(\sigma), \boldsymbol{V}(\sigma))$.

Sob estas condições, pretende-se encontrar o vetor $\boldsymbol{\sigma}$ ou o conjunto de vetores $\boldsymbol{\sigma}$ que produza as menores respostas para $\boldsymbol{B}$ e $\boldsymbol{V}$. Neste sentido, foi empregado o algoritmo PSO adaptado para problemas de otimização multiobjetivo.

### 3.5.1 Otimização Multiobjetivo por Enxame de Partículas

No artigo de Coello and Lechuga (2002), há o propósito de estender a heurística PSO para tratar de problemas de otimização multiobjetivo. Esta aproximação usa o conceito de dominância de Pareto para determinar a direção de vôo de uma partícula e mantém os vetores não-dominados encontrados anteriormente em um

arquivo externo, que é usado depois por outras partículas para guiá-las em seus próprios vôos. Como no PSO, o espaço de busca é o $\Re^n$, onde $n$ é o número de regiões do mapa.

Semelhantemente à descrição feita para o PSO, as partículas ou agentes *caminharão* em direção à solução ótima. A função-objetivo a que se refere esta solução é uma função vetorial, cujo resultado não possibilita a comparação simples entre dois escalares. Logo, não há como determinar *a melhor solução até o momento*, o que seria a melhor posição global no PSO. Observam-se, portanto, melhores posições as quais são encontradas pela aplicação do conceito de dominância.

A adaptação do PSO para um MOPSO (*Multiobjective Particle Swarm Optimization*) concentra-se, em essência, na natureza das comparações entre soluções e no arquivo das melhores soluções encontradas. As soluções não-dominadas são guardadas temporariamente em um repositório. A cada iteração, as novas soluções geradas são comparadas com aquelas que se situam no repositório. São eliminadas dele as soluções que se tornam dominadas por alguma nova solução, enquanto outras novas soluções caracterizadas como não-dominadas são incluídas nesse arquivo.

Ao incluir uma nova solução no repositório, as seguintes situações podem ocorrer:

- (Caso 1): O repositório encontra-se vazio; a solução atual $N_s$ é aceita;
- (Caso 2): O repositório possui a solução $S_1$; A solução atual, $N_s$, é dominada por $S_1$. Logo, $N_s$ será descartada;
- (Caso 3): A solução $S_1$ não domina a nova solução $N_s$ e nem $N_s$ domina a solução $S_1$. Então, ambas comporão o repositório;
- (Caso 4): Suponha a existência de algumas soluções no repositório. Caso a nova solução $N_s$ domine alguma dessas soluções, aquela dominada sairá do

repositório e $N_s$ entrará.

Com o passar do tempo, o repositório tende a abrigar cada vez mais soluções não-dominadas. Várias propostas para limitação do tamanho do repositório/arquivo são apresentadas na literatura (Coello et al., 2004). Na implementação do problema deste trabalho, escolheu-se a limitação da quantidade de soluções arquivadas no repositório para que o algoritmo não demandasse muito tempo.

Uma vez que o repositório possui a capacidade limitada de soluções não-dominadas, é desejável guardar a solução não-dominada que apresenta, até o momento, o mínimo valor para cada uma das funções-objetivo. Quando uma solução com o menor viés e outra com a menor variância até o momento são encontradas, elas são retidas no repositório até que alguma solução melhor neste sentido possa substituí-las. Estas duas soluções são consideradas boas porque estão próximas à origem do plano cartesiano segundo um dos critérios estabelecidos. O preenchimento das demais vagas no repositório é feito a partir da escolha aleatória e sem reposição de soluções não-dominadas.

Para uma breve ilustração, seja $n$ a capacidade do repositório. Suponhamos que, a certa altura, o repositório tenha sido atualizado, de modo que novas soluções não-dominadas tenham sido encontradas e adicionadas ao repositório. De forma similar, soluções do repositório que tenham se tornado dominadas pelas novas soluções são eliminadas. Se, após esta atualização, o repositório passou a abrigar $N > n$ soluções, deve-se escolher aquelas $n$ que permanecerão. As soluções $\boldsymbol{\sigma_1}$ e $\boldsymbol{\sigma_2}$ que apresentam, respectivamente, os menores valores de viés e de variância são mantidas. As outras $(n-2)$ vagas são preenchidas aleatoriamente dentre as $(N-2)$ soluções remanescentes.

No que se refere às partículas, a atualização da melhor posição é feita sob regras mais simples. Se a atual posição domina a melhor solução pessoal, então esta última é substituída pela primeira. Caso contrário, nada ocorre.

Ainda sobre as partículas, poderia ter sido adotado o modelo de repositório geral. Assim, cada partícula teria seu próprio repositório. No entanto, isto demandaria maior uso de memória e maior esforço computacional. O modelo adotado, por sua vez, apresentou bons resultados, pois manteve o esforço computacional próximo ao do PSO.

A expressão que dá origem à velocidade das partículas no MOPSO é idêntica à formula indicada no PSO.

$$\boldsymbol{v}_i^t = \boldsymbol{v}_i^{t-1} + \phi_1 u_1(\boldsymbol{p_i} - \boldsymbol{x_i^t}) + \phi_2 u_2(\boldsymbol{g} - \boldsymbol{x_i^t}),$$

onde:

- $\boldsymbol{v}_i^{t-1}$ é a velocidade da $i$-ésima partícula no tempo $(t-1)$;
- $\phi_1$ e $\phi_2$ são taxas de cognição;
- $u_1$ e $u_2$ são números aleatórios entre 0 e 1;
- $\boldsymbol{p}_i$ é a melhor posição da $i$-ésima partícula;
- $\boldsymbol{g}$ é uma solução não-dominada escolhida aleatoriamente dentre as presentes no repositório;
- $\boldsymbol{x}_i^t$ é a posição da partícula $i$ no tempo $t$.

Os quatro vetores $\boldsymbol{v}_i^t$, $\boldsymbol{x}_i^t$, $\boldsymbol{p_i}$ e $\boldsymbol{g}$ são $n$-dimensionais, onde $n$ é a dimensão do espaço de busca.

A atualização da posição das partículas é feita por meio da seguinte composição:

$$\boldsymbol{x}_i^{t+1} = \boldsymbol{x}_i^t + \boldsymbol{v}_i^t$$

O pseudocódigo do MOPSO é apresentado a seguir:

1. Gerar partículas em posições aleatórias e com velocidades aleatórias;

2. Avaliar a função a ser minimizada;

3. Comparar as soluções encontradas quanto ao valor da função-objetivo e as soluções não-dominadas vão a um repositório;

4. Atualizar a memória de cada partícula;

5. A *Melhor posição global* é escolhida aleatoriamente no repositório;

6. Atualizar as velocidades e as posições das partículas;

7. Voltar ao passo 2 e repetir o procedimento até o passo 6 enquanto um critério pré-estabelecido não for alcançado.

O exemplo a seguir ilustra dois passos do MOPSO em um problema com duas funções-objetivo, $f_1$ e $f_2$, e duas variáveis de entrada, $x_1$ e $x_2$. Estão apresentadas a terceira e a última ($50^{\text{a}}$) iterações. A figura 3.6 mostra as soluções no espaço de busca após a terceira iteração e suas imagens no espaço de objetivos. O final do processo iterativo situa-se na figura 3.7. Nos quatro gráficos, os $\times$'s representam as soluções; os $\circ$'s são as soluções não-dominadas que estão no repositório até aquela iteração e as linhas contínuas representam as soluções Pareto-ótimas (desconhecidas) as quais o algoritmo procura.

Após a convergência do algoritmo (figura 3.7), o conjunto de soluções não-dominadas que compõe o repositório é uma aproximação do conjunto Pareto-ótimo do problema e a aproximação da Fronteira de Pareto é bastante satisfatória.

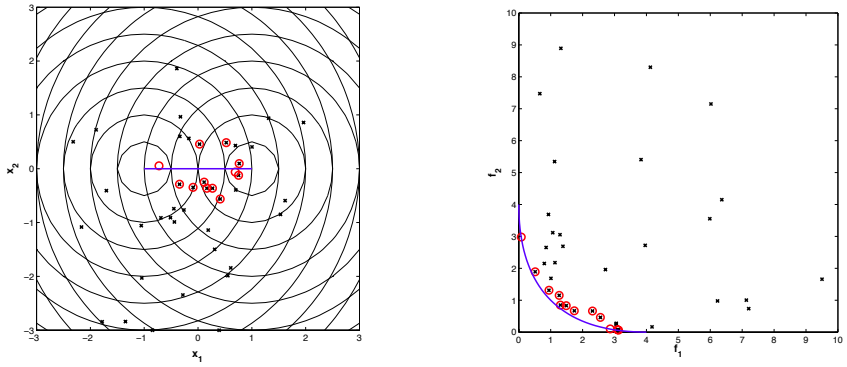

Figura 3.6: Espaço de busca (esquerda) e espaço de objetivos (direita) após três iterações.

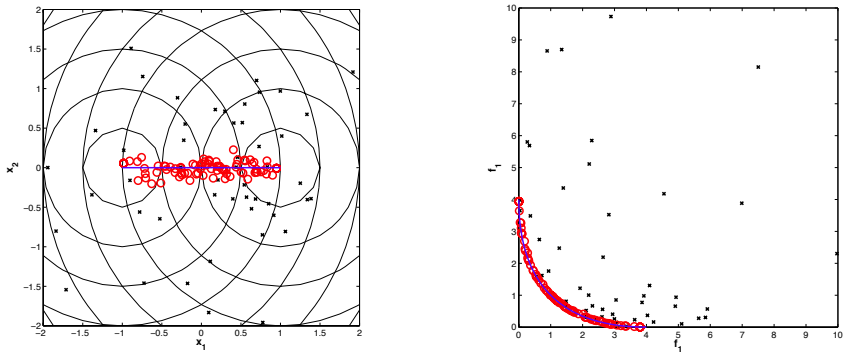

Figura 3.7: Espaço de busca (esquerda) e espaço de objetivos (direita) após a última iteração (50ª).

# Capítulo 4

# Simulações e Resultados Numéricos

## 4.1 Cenários e Amostras

A produção das estimativas de $B$ e de $V$ provém da modelagem dos pesos das estimativas $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \ldots, \hat{\theta}_n\}$. A composição destes pesos depende dos valores do vetor $\boldsymbol{\sigma}$, das coordenadas geográficas das regiões e da matriz de vizinhanças. Como as coordenadas geográficas e a matriz de vizinhança são conhecidas, a determinação do vetor $\boldsymbol{\sigma}$ faz-se necessária.

A metodologia proposta foi aplicada em dois blocos: no primeiro, as regiões do mapa possuem a mesma população e, no segundo, as populações são diferentes. Independentemente da composição do vetor $\boldsymbol{\sigma}$, quatro bancos de dados foram disponibilizados em cada cenário: população das regiões, número de casos de doença de cada região, coordenadas geográficas de cada região e a matriz de vizinhança

ou de adjacências (matriz $W$).

Para reproduzir diferentes distribuições e concentrações de população e de casos, alguns cenários foram criados. Eles baseiam-se em um mapa formado por uma malha regular de 203 regiões hexagonais. No bloco de populações iguais, cada região possui população de 1.000 indivíduos, totalizando 203.000 indivíduos em todo o mapa. O número total de casos foi de 20.300. Já no bloco de população diferente, a população total é de 26.292.572 pessoas e o total de casos, 52.563.

No bloco de população diferente, a composição da população e do número de casos de cada região foi baseada em dados de morte por câncer de mama na Nova Inglaterra no período de 1988 a 1992. A Nova Inglaterra é constituída por 245 condados e, neste trabalho, cada condado foi considerado como uma região do mapa. Destas 245 regiões, as informações das 203 primeiras regiões da Nova Inglaterra compuseram os dados do bloco de população diferente. Sendo assim, os totais de população e de casos são resultantes da soma das respectivas informações de cada região.

Os cenários foram criados também para verificar se a distribuição geográfica da população e dos casos influencia na qualidade das estimativas. No cenário homogêneo, os casos são gerados sob a suposição de não-aglomeração. Isto é, a probabilidade de um indivíduo ser um caso é a mesma em qualquer região do mapa. Nos outros cenários, essa probabilidade é maior em certas regiões e menor no restante do mapa. O conjunto de regiões, em que a probabilidade de que um indivíduo seja um caso é maior, é chamado de *conglomerado espacial*.

Os cenários com conglomerados podem variar quanto à forma e quanto à intensidade. A forma pode ser simples, dupla ou irregular - figuras 4.1, 4.2 e 4.3 nesta ordem. Os adjetivos fraco, moderado e forte caracterizam a intensidade do conglomerado.

Em cada um dos blocos, aplicaram-se duas metodologias:

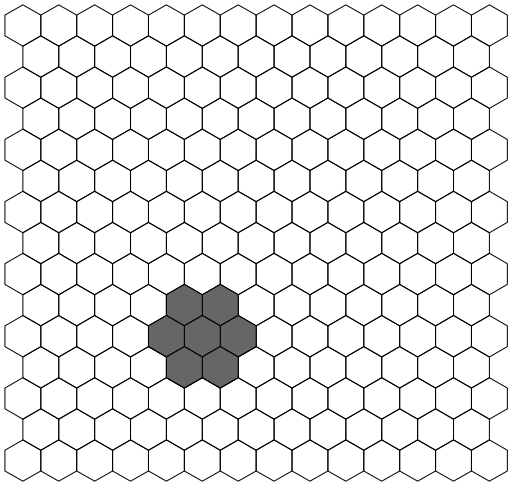

Figura 4.1: Conglomerado com forma simples

1. aquela em que há varredura dos possíveis valores de sigma no intervalo [0,1], com a restrição de sigmas iguais em todas as regiões;

2. aquela em que há utilização do MOPSO, relaxando a restrição de igualdade de sigmas.

No primeiro bloco, todas as regiões do mapa possuem a mesma população. Os cenários criados foram:

- Cenário 1: Ausência de conglomerado;
- Cenário 2: Conglomerado simples com fraca intensidade;
- Cenário 3: Conglomerado simples com sinal moderado;
- Cenário 4: Conglomerado simples com sinal forte;
- Cenário 5: Conglomerado duplo com intensidade moderada;

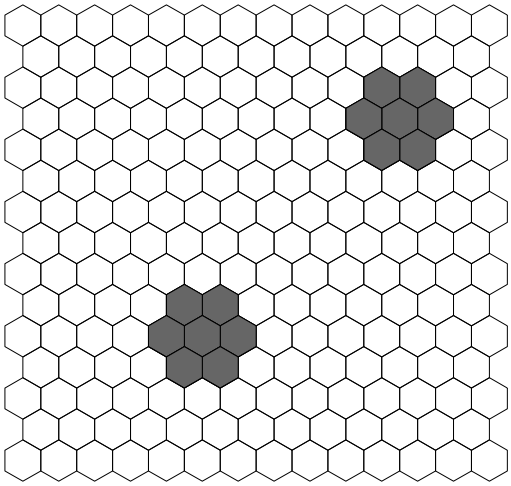

Figura 4.2: Conglomerado com forma dupla

- Cenário 6: Conglomerado duplo com sinal forte;
- Cenário 7: Conglomerado irregular com forte intensidade.

No segundo bloco, a forma, a localização e a força de sinal dos conglomerados foram idênticos àqueles do primeiro bloco. Porém, o mapa apresenta populações variadas para as regiões.

Nas ocasiões em que o vetor $\boldsymbol{\sigma}$ é composto por elementos iguais, o padrão da ocorrência de doença de um determinado cenário foi propagado em 1000 amostras por meio da distribuição multinomial. Seguidamente, 200 valores foram escolhidos para este parâmetro em cada amostra: de 0,005 a 1 com passos de 0,005. A partir de cada um dos 200 vetores, foram originadas as taxas de doença estimadas em cada região, o vetor $\hat{\boldsymbol{\theta}}$. Por fim, foram obtidas as estimativas $(B, V)$. Paralelamente, calculou-se a estimativa do vetor $\hat{\boldsymbol{\theta}}$ pelo método empírico local. Como exemplo, supor que haja duas amostras. Para cada uma, 200 valores de $\sigma$ originam os 200

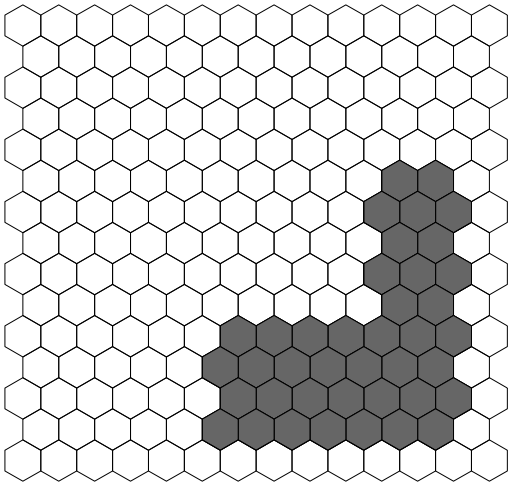

Figura 4.3: Conglomerado com forma irregular

pares $(B, V)$. Pelo método de Marshall, obtém-se um par $(B, V)$ por amostra. Quando estas estimativas são plotadas em um gráfico, forma-se uma curva para cada amostra (figura 4.4).

Cabe destacar que os pares $(B, V)$ do método empírico local estão bem afastados da origem. Por este motivo, este método não foi considerado no caso em que o vetor $\boldsymbol{\sigma}$ é livre.

Resumidamente, os itens que se seguem mostram os passos para a obtenção das estimativas $(B, V)$ pelos 3 métodos a partir do vetor $\boldsymbol{\sigma}$ composto por valores iguais.

- Geração do vetor de estimativas $\hat{\boldsymbol{\theta}}$;
- Para cada estimativa $\hat{\boldsymbol{\theta}}$, calculam-se seus respectivos valores de viés e de variância.

Quando o vetor $\boldsymbol{\sigma}$ é livre, os cenários das simulações são idênticos àqueles

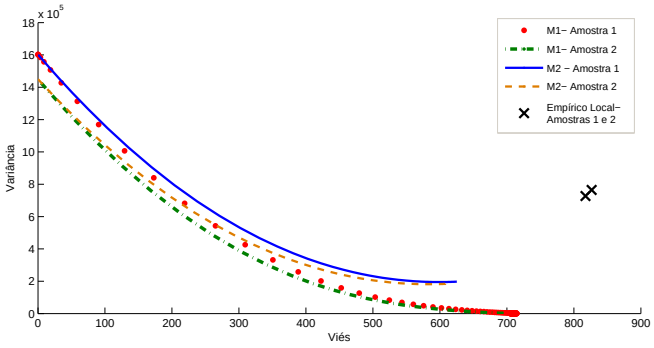


Figura 4.4: Distribuição de pares *(B,V)* conforme método empregado para duas amostras

do caso anterior. Todavia, a quantidade de amostras ou simulações foi reduzida para 100, pois o tempo despendido para a execução do algoritmo foi maior do que na primeira parte das simulações. Para esta segunda composição do vetor $\boldsymbol{\sigma}$, compararam-se apenas os resultados dos métodos 1 e 2.

O processo do MOPSO foi iniciado com as especificações do espaço de procura, das posições iniciais das partículas e de suas velocidades iniciais. O espaço de busca compreende todos os possíveis vetores $\boldsymbol{\sigma} \in [0,1]^n$, isto é $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_n)$, em que $n$ é o número de regiões e $0 < \sigma_i \leq 1$. Cada partícula é uma solução-tentativa e, portanto, a população de partículas é o conjunto de vetores $\boldsymbol{\sigma_1}$, $\boldsymbol{\sigma_2}$,..., $\boldsymbol{\sigma_N}$, onde $N$ é o tamanho da população. Os parâmetros utilizados no algoritmo foram:

- Número de partículas ($N$): 50;
- Tamanho do repositório: 100;
- Número máximo de iterações: 50;
- Velocidades iniciais: zero;
- Taxas de cognição $\phi_1 = \phi_2 = 2$.

A continuação do procedimento se dá pela aplicação do conceito de dominância entre as partículas (pares $(B, V)$) e as soluções não-dominadas que estão no repositório naquela iteração. Com as diversas estimativas de $(B, V)$ calculadas, outras medidas são apresentadas para avaliar os métodos. Estas estatísticas encontram-se na seção seguinte.

## 4.2 Medidas de Performance

Muitas métricas para mensurar/medir a convergência de um conjunto de soluções não-dominadas em direção à *fronteira de Pareto* têm sido propostas. Quase todas elas foram construídas com o intuito de comparar diretamente dois conjuntos de soluções não-dominadas. Existem também aproximações que comparam um conjunto de soluções não-dominadas com um conjunto de soluções ótimas de Pareto se a verdadeira fronteira de Pareto é conhecida (Grosan, 2003).

Neste trabalho, a verdadeira fronteira não é conhecida. Será utilizada, portanto, uma medida que não considere esta informação, embora possa comparar a distribuição dos pares $(B, V)$ oriundos dos três métodos apresentados: 1, 2 e empírico local. Cabe ressaltar que haverá apenas um ponto $(B, V)$ resultante da aplicação do método empírico local.

As duas abordagens citadas fornecem estimativas $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_1, \ldots, \hat{\theta}_N\}$, as quais serão avaliadas conforme os critérios de menor viés e de menor variância. Assim, os pares $(B, V)$ mais próximos da origem são as soluções esperadas. Duas formas de avaliar o comportamento destes pontos são a distância euclidiana e a medida C, proveniente do Método da Cobertura.

### 4.2.1 Distância à origem

A distância euclidiana constitui uma das medidas mais simples na descrição de um fenômeno representado em um gráfico bidimensional. Neste caso, a origem do plano cartesiano é o valor utópico para as medidas oriundas das estimativas de taxa de doença.

A distância bidimensional, $D$, entre dois pontos quaisquer $P_1 = (B_1, V_1)$ e $P_2 = (B_2, V_2)$ é assim expressa:

$$D = \sqrt{(B_1 - B_2)^2 + (V_1 - V_2)^2}$$

Seja o par $(B_1, V_1)$ gerado por um dos métodos apresentados. A sua distância em relação à origem é calculada por meio da fórmula anterior simplificada:

$$D = \sqrt{(B_1)^2 + (V_1)^2}$$

As medidas de distância foram calculadas somente nas simulações em que o vetor $\boldsymbol{\sigma}$ possui elementos iguais. Ainda nesta situação mais simples, a quantidade de estimativas $(B, V)$ é proporcional ao número de $\sigma$'s ou de vetores $\boldsymbol{\sigma}$ com elementos iguais. Perante as várias soluções $(B, V)$ encontradas, o interesse é descobrir se existe uma melhor ou algumas melhores. Trata-se, portanto, das soluções não-dominadas.

### 4.2.2 Método da Cobertura - Métrica $C$

A métrica $C$ foi introduzida e aperfeiçoada por Zitzler (1999) e possibilita a comparação entre dois conjuntos de soluções não-dominadas.

**Definição 4.2.1** (Cobertura de dois conjuntos)**.** *: Sejam X o conjunto de vetores-decisão para o problema considerado e A,B* $\subseteq X$ *dois conjuntos de vetor-decisão. A função C mapeia o par ordenado (A,B) no intervalo* $[0, 1]$*:*

$$C(A, B) = \frac{|a \in A / \nexists b \in B : b \preceq a|}{|A|},$$

*onde* $|.|$ *denota o número de elementos do conjunto-argumento.*

A medida $C(A, B)$ denota, então, a proporção de elementos do conjunto $A$ não-dominada por elementos do conjunto $B$. Assim:

- *C(A,B)* $= 1$ significa que não existe solução de $A$ dominada por qualquer solução de $B$;
- *C(A,B)* $= 0$ significa que todas as soluções de $A$ são dominadas por soluções de $B$;
- *C(A,B)* não necessariamente é igual a 1-*C(B,A)*.

## 4.3 Resultados

Inicialmente, são dispostos os resultados das simulações em que o vetor $\boldsymbol{\sigma}$ possui elementos iguais.

A primeira medida trata da média das 1000 menores distâncias euclidianas entre o par $(B, V)$ e a origem em cada cenário. As médias de distância dos pares (viés, variância) do Método 1 (M1) mostraram-se sempre menores em relação aos resultados obtidos para os métodos 2 (M2) e de Marshall. Para que os resultados do M1 fossem considerados como referência, as 3 medidas de distância de cada cenário foram divididas pelo valor da distância média de M1 do respectivo cenário.

Juntamente à informação de que, em média, os pares de $(B, V)$ do M1 são mais próximos da origem, a mensuração desta proporção está representada na tabela 4.1.

As médias das menores distâncias de M2 e de Marshall são superiores àquelas do M1 numa faixa entre 75 a 682 vezes. Além disso, as distâncias médias do método Marshall são também maiores do que as medidas encontradas pelo M2.

Tabela 4.1: Média das menores distâncias nas simulações e em cada cenário - População homogênea

| Cenário | M1 | M2 | Marshall |
|---|---|---|---|
| 1 | 1.0 | 191.4 | 449.9 |
| 2 | 1.0 | 212.7 | 536.4 |
| 3 | 1.0 | 230.9 | 584.3 |
| 4 | 1.0 | 261.8 | 681.8 |
| 5 | 1.0 | 220.9 | 537.6 |
| 6 | 1.0 | 257.3 | 660.4 |
| 7 | 1.0 | 75.3 | 202.3 |

Analogamente à descrição feita para os cenários com população homogênea, os resultados do M1 são mantidos como referência ao comportamento dos outros dois métodos para a população não-homogênea. As distâncias médias do M2 e de Marshall são também maiores do que as medidas do M1(tabela 4.2), mas ocorre em uma proporção muito menor do que se constata na população homogênea.

A performance dos três métodos também foi avaliada consoante a estatística Medida de Cobertura. A partir da aplicação do conceito de dominância entre os pontos a cada dois métodos, calculou-se a proporção de pontos não-dominada de um método em relação a outro. Deseja-se saber, portanto, a proporção de vezes que um método *não é pior que o outro*.

Tabela 4.2: Média das menores distâncias nas simulações e em cada cenário - População não-homogênea

| Cenário | M1 | M2 | Marshall |
|---|---|---|---|
| 1 | 1.0 | 2.3 | 7.1 |
| 2 | 1.0 | 5.5 | 6.8 |
| 3 | 1.0 | 6.5 | 8.5 |
| 4 | 1.0 | 5.9 | 9.3 |
| 5 | 1.0 | 5.8 | 8.3 |
| 6 | 1.0 | 6.4 | 8.2 |
| 7 | 1.0 | 5.5 | 7.3 |

Foram utilizadas quatro proporções na tabela 4.3 e são assim definidas:

- $C(1,2)$= proporção de pontos do M1 não-dominada por pontos do M2;
- $C(2,1)$= proporção de pontos do M2 não-dominada por pontos do M1;
- $C(M,1)$= proporção de pontos de Marshall não-dominada por pontos do M1;
- $C(M,2)$= proporção de pontos do Marshall não-dominada por pontos do M2.

Nota-se que, como o método de Marshall produz um único ponto, $C(M,1)$ e $C(M,2)$ só podem assumir os valores 0 e 1. O valor 0 indica que o ponto é dominado por M1 ou por M2. O valor 1 indica que o ponto não é dominado.

Quanto aos resultados da população homogênea (tabela 4.3), nota-se que quase todos os pontos/pares$(B,V)$ do M1 não foram dominados por pontos do M2 $(C(1,2))$. De outra forma, aproximadamente um quinto dos pontos do M2 não foram dominados por pares do M1 $(C(2,1))$. Os pares $(B,V)$ oriundos do método

de Marshall encontraram-se dominados por pontos dos métodos 1 e 2 em praticamente todas as simulações.

Tabela 4.3: Proporção de pontos de um método que são não-dominados por pontos de outro método

| População | Cenário | $C(1,2)$ | $C(2,1)$ | $C(M,1)$ | $C(M,2)$ |
|---|---|---|---|---|---|
| Homogênea | 1 | 0.9996 | 0.1896 | 0 | 0 |
| | 2 | 0.9992 | 0.1883 | 0 | 0 |
| | 3 | 0.9991 | 0.1886 | 0 | 0 |
| | 4 | 0.9986 | 0.1885 | 0 | 0 |
| | 5 | 0.9998 | 0.1894 | 0 | 0 |
| | 6 | 0.9997 | 0.1871 | 0 | 0 |
| | 7 | 0.9995 | 0.1865 | 0 | 0 |
| Não-homogênea | 1 | 0.9835 | 0.2054 | 0.9250 | 0.9070 |
| | 2 | 0.9999 | 0.1884 | 0.2280 | 0.7980 |
| | 3 | 0.9999 | 0.1870 | 0.2270 | 0.7070 |
| | 4 | 0.9999 | 0.1889 | 0.1080 | 0.5140 |
| | 5 | 1.0000 | 0.1881 | 0.1370 | 0.5440 |
| | 6 | 0.9999 | 0.1852 | 0.1650 | 0.7050 |
| | 7 | 1.0000 | 0.1882 | 0.1350 | 0.6450 |

Na tabela 4.3 para população não-homogênea, as proporções das duas primeiras colunas refletem análise similar àquela disposta para a população homogênea: os pares $(B, V)$ do M1, em sua maioria, não foram dominados por pontos do M2. Ademais, os pares $(B, V)$ do método de Marshall demonstraram alguma melhora na proporção de pares não-dominados em relação aos dados de população homogênea.

A tabela 4.4 mostra os resultados das simulações em que o vetor $\boldsymbol{\sigma}$ é livre.

Para os dois tipos de população, praticamente todos os pontos $(B, V)$ do M1 não foram dominados por pontos do M2 ($C(1, 2)$).

Tabela 4.4: Proporção de pontos de um método que são não-dominados por pontos de outro método

| População | Cenário | $C(1, 2)$ | $C(2, 1)$ |
|---|---|---|---|
| Homogênea | 1 | 1.0000 | 0.2280 |
| | 2 | 1.0000 | 0.2275 |
| | 3 | 1.0000 | 0.2671 |
| | 4 | 1.0000 | 0.2797 |
| | 5 | 1.0000 | 0.2808 |
| | 6 | 1.0000 | 0.2186 |
| | 7 | 1.0000 | 0.2386 |
| Não-homogênea | 1 | 0.9996 | 0.1751 |
| | 2 | 1.0000 | 0.2976 |
| | 3 | 1.0000 | 0.2848 |
| | 4 | 1.0000 | 0.3058 |
| | 5 | 1.0000 | 0.3003 |
| | 6 | 1.0000 | 0.3423 |
| | 7 | 1.0000 | 0.2806 |

Conforme as duas métricas e independentemente da composição do vetor $\boldsymbol{\sigma}$, as combinações de $(B, V)$ do método 1 proporcionaram os melhores resultados. As estimativas de taxas de doenças pelo método 1 responderam de forma mais satisfatória aos critérios propostos no trabalho.

# Capítulo 5

# Aplicação

A metodologia proposta é aplicada em dados de morte por câncer de mama no período de 1988 a 1992 na Nova Inglaterra. Esta região situa-se no nordeste dos Estados Unidos e é composta de 245 condados. Dentre a população de risco de 29.535.210 mulheres, houve 58.943 mortes.

O mapa da figura 5.1 mostra a distribuição das taxas brutas de doença nos diversos condados da Nova Inglaterra.

Em se tratando da proporção de pontos de um método não-dominados por outro método (tabela 4.3), os resultados são:

- $C(1,2) = 1$, isto é, nenhum ponto do M1 é dominado por qualquer ponto do M2;
- Apenas 12,5% dos pontos do M2 são não-dominados, quando comparados aos pontos do M1.

Diante destes resultados, o método 1 mostrou-se melhor do que o método 2. Desta forma, os dados da Nova Inglaterra serão observados somente conforme o método 1. A figura 5.2 mostra os pontos de $(B, V)$ que formam a Fronteira de

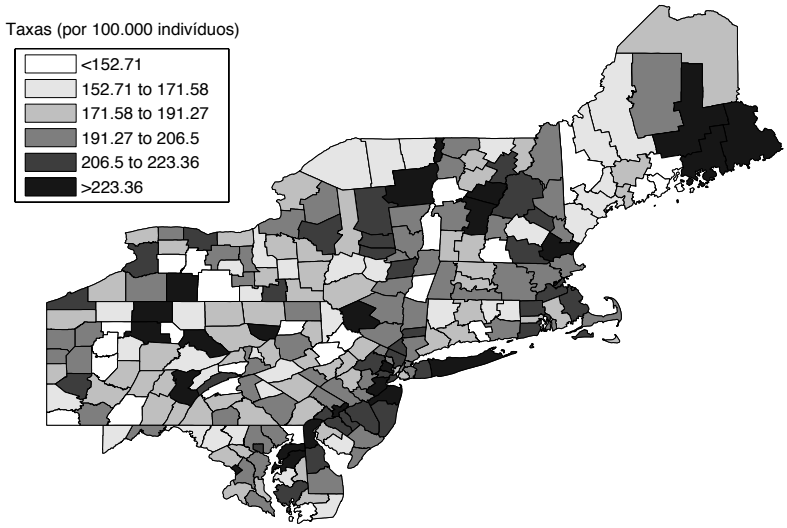


Figura 5.1: Taxa bruta

Pareto. As soluções estão bem próximas da origem dos eixos e o $\times$ parece ser uma solução que apresenta um bom compromisso entre ambos os critérios.

A solução com $B = 0$, isto é, aquela situada mais à esquerda no gráfico da figura 5.2, coincide com a taxa bruta. Já a solução com $V = 0$ resulta num mapa com todas as regiões coloridas com o mesmo tom. Em outras palavras, a estimativa de cada região seria igualada à média das taxas (soluções situadas ao longo do eixo $x$).

A distribuição das taxas que geram o par $(B, V)$, representado pelo $\times$, está apresentada na figura 5.3.

Esta solução (taxa ótima) produz um mapa com tonalidades mais parecidas entre regiões mais próximas. O mapa mais suave, obtido por essa solução, provavelmente se aproximaria do resultado advindo de uma regressão.

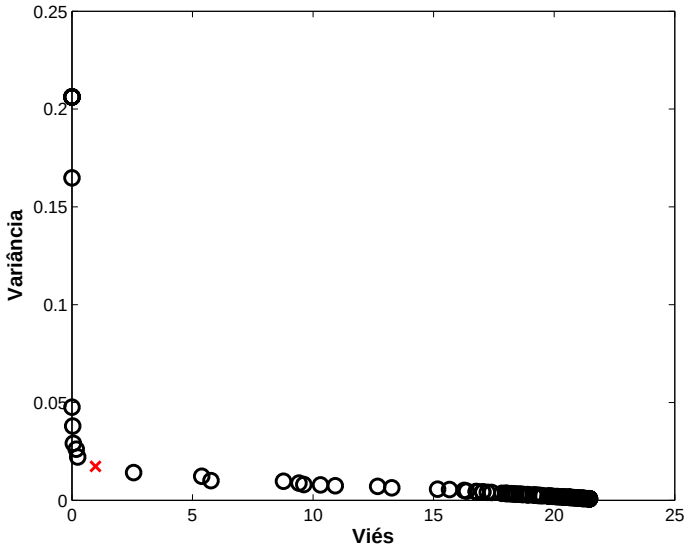


Figura 5.2: Fronteira de Pareto. A solução destacada (×) apresenta um bom compromisso entre B e V

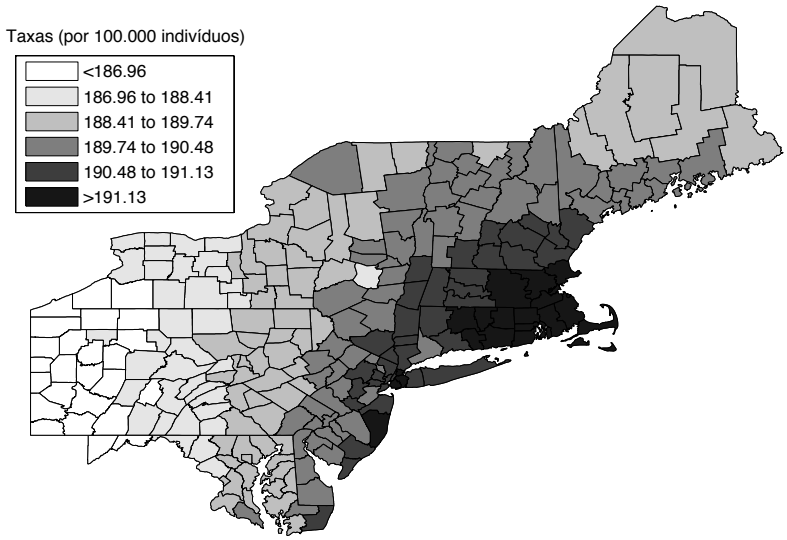


Figura 5.3: Taxa ótima

# Capítulo 6

# Considerações Finais e Trabalhos Futuros

## 6.1 Considerações Finais

Este trabalho apresentou um novo método de composição das taxas de doença para regiões em um mapa. Este procedimento caracteriza-se pela inclusão da informação da vizinhança no cômputo da taxa de doença de uma região. Sabe-se que os padrões de doença são mais parecidos entre regiões próximas do que entre regiões mais distantes. Portanto, a propagação da informação das vizinhanças na taxa de doença de uma região não é realizada uniformemente. Neste sentido, consideraram-se ponderações diferenciadas segundo a distância entre as regiões. Estas ponderações, por sua vez, correspondem à informação transmitida, cuja proporção depende do peso atribuído pela densidade da distribuição normal bivariada.

A modelagem do problema a partir de uma distribuição normal bivariada trouxe duas incógnitas, os seus parâmetros. Convencionou-se que o vetor de médias

de cada normal bivariada fosse igual ao centróide da sua respectiva região. Em relação à matriz $\Sigma$, definiu-se apenas a questão da independência das variáveis normais marginais.

Adicionalmente à modelagem dos pesos, as estimativas propostas seriam o resultado da minimização conjunta das medidas viés e variância local. Assim, estas medidas formam a função-objetivo vetorial de um problema de otimização multiobjetivo, cujas variáveis de interesse são os $\sigma$'s de cada região do mapa. Logo, a definição dos valores de $\sigma$ das normais marginais em todas as regiões constituiu o foco deste trabalho.

Diante de inúmeras possibilidades de composição do vetor $\boldsymbol{\sigma}$, estabeleceu-se o intervalo de valores possíveis. A constituição deste vetor seguiu dois caminhos: combinação simples e vetor livre. Na primeira alternativa, todos os elementos do vetor são iguais, enquanto combinações arbitrárias para o vetor $\boldsymbol{\sigma}$ são permitidas na segunda versão. A partir dos valores iniciais do vetor $\boldsymbol{\sigma}$ livre, as melhores combinações deste vetor para as regiões foram encontradas por meio do algoritmo MOPSO.

Com a determinação dos componentes do vetor $\boldsymbol{\sigma}$, as simulações indicaram que o método 1 promoveu respostas mais favoráveis do que os métodos 2 e empírico de Marshall. E os métodos 1 e 2 mostraram-se melhores do que o método de Marshall. Constatou-se também que a proporção de soluções do método 1 não-dominadas por soluções dos outros dois métodos foi relevante. Análise similar foi feita perante os resultados da aplicação dos métodos em dados reais de câncer de mama. Em particular, observaram-se diferenças entre as distribuições da taxa bruta e da taxa ótima no mapa da Nova Inglaterra. O padrão observado no mapa da taxa ótima é mais suave, pois existe uma transição suave de cores ao longo das regiões.

Este padrão provavelmente seria observado caso fosse realizada uma regressão. Neste caso, a taxa de cada região dependeria da taxa de outras regiões.

Foi empregado o algoritmo MOPSO. Houve tentativa de utilização dos métodos de direção de busca: Gradiente e Método de Newton, mas os resultados não foram satisfatórios. Ao mesmo tempo, o PSO é um algoritmo de implementação simples e eficiente.

## 6.2 Trabalhos Futuros

A minimização simultânea do viés e da variância poderia ser abordada de três formas: uma nova configuração de vizinhança para o MOPSO, aplicação de outros algoritmos e utilização de regressão.

Em relação à estrutura de vizinhança, a configuração utilizada neste trabalho foi a mais simples: todas as partículas são vizinhas entre si (conexão completa). Quando existem estruturas menores de vizinhança, as partículas se reportam aos seus líderes imediatos. Estes seriam equivalentes à melhor posição global no contexto de conexão completa *Melhor posição global*. Assim, a *Melhor posição global* de cada iteração seria substituída pelas *Melhores posições locais* das vizinhanças. As estruturas de rede de árvores ou hierárquica e de $k$-vizinhos imediatos são exemplos.

No que se refere à segunda alternativa citada, podem-se utilizar outros métodos de otimização. Dentre eles, outros algoritmos de população como os algoritmos genéticos. Por fim, a relação de dependência entre as taxas das regiões poderia ser observada em uma regressão simples e em uma regressão Poisson geograficamente ponderada.

# Referências Bibliográficas

M. Ali, M. Emch, M. Yunus, and J. Clemens. Modeling spatial heterogeneity of disease risk and evaluation of the impact of vaccination. *Elsevier*, 27:3724–3729, 2009.

M. S. Bazaraa and C. M. Shetty. *Nonlinear Programming - Theory and Algorithms*. John Wiley & Sons, 1979.

D. Böhning. Empirical bayes estimators and non-parametric mixture models for space and time-space disease mapping and surveillance. *Environmetrics*, 14: 431–451, 2003.

G. Casella. An introduction to empirical bayes data analysis. *J. Am. Statist. Ass.*, 39(2):83–87, 1985.

C. A. C. Coello and M. S. Lechuga. Mopso: A proposal for multiple objective particle swarm optimization. *Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2002)- Honolulu, Hawaii, USA*, 1:1051–1056, 2002.

C. A. C. Coello, G. T. Pulido, and M. S. Lechuga. Handling multiple objectives with particle swarm optimization. *IEEE Transactions on Evolutionary Computation*, 8(3):256–279, 2004.

B. Efron and C. Morris. Data analysis using stein's estimator and its generalizations. *J. Am. Statist. Ass.*, 70(350):311–319, 1975.

C. Grosan. Performance metrics for multiobjective optimization evolutionary algorithms. *Conference on Applied and Industrial Mathematics (CAIM)*, pages 1–13, 2003.

J. Kennedy and R. C. Eberhart. Particle swarm optimization. *In: Proceedings of the IEEE International Conference on Neural Networks IV. Piscataway: IEEE*, pages 1942–1948, 1995.

T. Maiti. Hierarchical bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69:339–348, 1998.

K. G. Manton, M. A. Woodburry, E. Stallard, W. B. Riggan, J. P. Creason, and A. C. Pellom. Empirical bayes procedures for stabilizing maps of us cancer mortality. *J. Am. Statist. Ass.*, 84:637–650, 1989.

R. J. Marshall. Mapping disease and mortality rates using empirical bayes estimators. *Journal of the Royal Statistical Society*, 40(2):283–294, 1991.

M. G. Pereira. *Epidemiologia: teoria e prática*. Guanabara Koogan, 1995.

T. R. Saunderson and I. H. Langford. A study of the geographical distribution of suicide rates in england and wales 1989-92 using empirical bayes estimates. *Soc. Sci. Med.*, 43(4):489–502, 1996.

R. H. C. Takahashi. *Otimização Escalar e Vetorial*. Notas de aula - OTEV - Universidade Federal de Minas Gerais - Departamento de Matemática - http://www.mat.ufmg.br/ taka/, 2007.

R. K. Tsutakawa, G. L. Shoop, and C. J. Marienfeld. Empirical bayes estimation of cancer mortality rates. *Statis. Med.*, 4:201–212, 1985.

J. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183, 2007.

E. Zitzler. Evolutionary algorithms for multiobjective optimization: Methods and applications. *PhD thesis, Swiss Federal Institute of Technology Zurich*, 1999.

# Apêndice A

# Descrição do problema de otimização

## A.1 Função *viés*, $B$

$$B(\hat{\theta}) = \sum_{i=1}^{n} |\hat{\theta}_i - \tilde{\theta}_i|,$$

onde $\tilde{\theta}_i$ é a taxa bruta observada na região $i$ e $\hat{\theta}_i$ é a estimativa da taxa da região $i$.

## A.2 Variância local, $V$

$$V(\hat{\theta}) = \sum_{i=1}^{n} \frac{(\hat{\theta}_i - \overline{\theta}_i)^2}{n},$$

onde $\hat{\theta}_i$ é a estimativa da taxa $i$ e $\overline{\theta}_i$ é a média das taxas na vizinhança da região $i$.

## A.3 Cálculo de $\hat{\theta}_i$, a estimativa da taxa de doença na região $i$

$\hat{\theta}_i$ é uma combinação linear das taxas brutas observadas: $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$. Assim,

$$\hat{\theta}_i = \alpha_{i1}\tilde{\theta}_1 + \cdots + \alpha_{in}\tilde{\theta}_n, \quad i = 1, \ldots, n$$

## A.4 Definição de $\alpha_{ij}$

O procedimento de estimação consiste em obter um vetor de coeficientes $\boldsymbol{\alpha_i} = \{\alpha_{i1}, \ldots, \alpha_{in}\}$, onde $i$ refere-se à $i$-ésima região, $i = 1, \ldots, n$.

### A.4.1 Abordagem 1

Seja $f_i(\cdot)$ a função de densidade de uma distribuição normal bivariada, com média $\mu_i = [x_i \quad y_i]'$ (ou seja, a média é dada pelo centróide da $i$-ésima região) e a matriz de variância-covariância

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

Define-se

$$\alpha_{ij} = \frac{f_i(x_j, y_j)}{\sum_{k=1}^{n} f_i(x_k, y_k)}$$

### A.4.2 Abordagem 2

Definição:

$$\alpha_{ij} = \frac{f_i(x_j, y_j) I(j \in S_i)}{\sum_{k=1}^{N} [f_i(x_k, y_k) I(k \in S_i)]},$$

onde:

- $S_i$ é a vizinhança da região $i$;

- $I(j \in S_i)$ é igual a 1 se a região $j$ está na vizinhança da região $i$ e zero, caso contrário;
- $f_i(\cdot)$: densidade da normal, centrada em $(x_i, y_i)$ e com variância $\sigma^2$.

## A.5 Obtenção das estatísticas *viés* e *variância*

As estatísticas $B$ e $V$ dependem dos coeficientes $\alpha_{ij}$'s, os quais são função do vetor $\boldsymbol{\sigma}$ da densidade da distribuição normal bivariada. Em decorrência disto, $\boldsymbol{\theta}$ é caracterizado como uma função de $\boldsymbol{\sigma}$.

### A.5.1 Determinação do vetor $\boldsymbol{\sigma}$

**Vetor $\boldsymbol{\sigma}$ composto por elementos iguais**

Problema de otimização neste caso:

$$\boldsymbol{\sigma}^* = arg\min_{\boldsymbol{\sigma}} \quad \mathbf{f}(\boldsymbol{\sigma}) = (B(\boldsymbol{\sigma}), V(\boldsymbol{\sigma}))$$

$$\text{s.a.} \quad \begin{cases} \sigma_i = \sigma \quad i = 1, \ldots, n \\ \sigma > 0 \end{cases}$$

Neste problema, a restrição $\sigma_i = \sigma$ implica que, em todas as regiões, o mesmo valor de $\sigma$ é usado na densidade normal bivariada para obtenção das estimativas.

**Vetor $\boldsymbol{\sigma}$ livre**

O problema de otimização é apresentado da seguinte forma:

$$\boldsymbol{\sigma}^* = arg\min_{\boldsymbol{\sigma}} \quad \mathbf{f}(\boldsymbol{\sigma}) = (B(\boldsymbol{\sigma}), V(\boldsymbol{\sigma}))$$

$$\text{s.a.} \quad \{\sigma_i > 0, \quad i = 1, \ldots, n$$