MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

INPE-14800-TDI/1243

DISTRIBUIÇÃO ESPACIAL DO RISCO ASSOCIADO A EVENTOS RAROS POR GEOESTATÍSTICA BINOMIAL E SIMULAÇÃO CONDICIONADA

Eduardo Celso Gerbi Camargo

Tese de Doutorado do Curso de Pós-Graduação em Sensoriamento Remoto, orientada pelos Drs. Antonio Miguel Vieira Monteiro e Suzana Druck Fucks, aprovada em 29 de março de 2007.

INPE
São José dos Campos
2007

Publicado por: **esta página é responsabilidade do SID**

Instituto Nacional de Pesquisas Espaciais (INPE)
Gabinete do Diretor – (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 – CEP 12.245-970
São José dos Campos – SP – Brasil
Tel.: (012) 3945-6911
Fax: (012) 3945-6919
E-mail: pubtc@sid.inpe.br

Solicita-se intercâmbio
We ask for exchange

MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

INPE-14800-TDI/1243

DISTRIBUIÇÃO ESPACIAL DO RISCO ASSOCIADO A EVENTOS RAROS POR GEOESTATÍSTICA BINOMIAL E SIMULAÇÃO CONDICIONADA

Eduardo Celso Gerbi Camargo

Tese de Doutorado do Curso de Pós-Graduação em Sensoriamento Remoto, orientada pelos Drs. Antonio Miguel Vieira Monteiro e Suzana Druck Fucks, aprovada em 29 de março de 2007.

INPE
São José dos Campos
2007

528.711.7

Camargo, E. C. G.
Distribuição espacial do risco associado a eventos raros por geoestatística binominal e simulação condicionada / Eduardo Celso Gerbi Camargo. - São José dos Campos: INPE, 2007.
148 p. ; (INPE-14800-TDI/1243)

1. Análise espacial. 2. Geoestatística.
3. Semivariograma do risco. 4. Cokrigeagem binomial.
5. Simulação seqüencial condicionada. I. Título.

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de Doutor(a) em Sensoriamento Remoto.

Dra. Corina da Costa Freitas

Presidente / INPE / SJCampos - SP

Dr. Antonio Miguel Vieira Monteiro

Orientador(a) / INPE / SJCampos - SP

Dra. Suzana Druck Fucks

Orientador(a) / EMBRAPA / Brasília - DF

Dr. Gilberto Câmara

Membro da Banca / INPE / SJCampos - SP

Dra. Marília Sá Carvalho

Convidado(a) / FIOCRUZ / Rio de Janeiro - RJ

Dr. Paulo Justiniano Ribeiro Jr.

Convidado(a) / UFPR / Curitiba - PR

Aluno (a): Eduardo Celso Gerbi Camargo

São José dos Campos, 29 de Março de 2007

"O segredo é colocar-se a caminho, num fazer-se e perfazer-se constantes, num empenho e aperfeiçoamento contínuos"

Frei Nilo Agostini

Dedico a meus pais,
(in memoriam)
a minha esposa Rita de Cássia, e ao meu
querido filho Eduardo Augusto, pelo amor,
compreensão e constante incentivo.

AGRADECIMENTOS

Esta tese pertence em parte a um conjunto de pessoas pelo apoio, paciência e amizade que sempre revelaram e a quem nunca conseguirei exprimir a minha gratidão.

A minha orientadora, antes amiga, Dra. Suzana Druck, pelo apoio constante e encorajamento. Obrigado pelo seu respeito, sua amizade, sua ética e pelos conhecimentos pródigos que me foram repassados sem restrições.

Ao meu orientador, antes amigo, Dr. Antônio Miguel Vieira Monteiro, gratidão pela grande oportunidade, incentivo e ensinamento. A sua marcante simplicidade, formação íntegra e segura explica a grande admiração que o amigo merece.

À amiga, Dra. Corina da Costa Freitas, à sua disponibilidade irrestrita, sua forma exigente, crítica e criativa de argüir as idéias apresentadas, creio que deram norte a este trabalho, facilitando o alcance de seus objetivos. Obrigado por seres tão especial.

Ao amigo, Dr. Gilberto Câmara meus sinceros agradecimentos por sua permanente solicitude em todas as fases deste trabalho, que muito me ajudaram a superar situações limites, estimulando-me a seguir em frente.

Ao amigo, Dr. Paulo Justiniano Ribeiro Junior meus sinceros agradecimentos pela disposição para discutir o trabalho em meados de setembro de 2005, bem como por seus ensinamentos, críticas e contribuições.

À amiga, Dra. Marilia Sá Carvalho, meus sinceros agradecimentos pela disponibilidade e disposição em analisar este trabalho.

Ao Instituto Nacional de Pesquisas Espaciais - INPE, pela oportunidade, o apoio e o incentivo ao meu crescimento pessoal e capacitação científica.

Ao amigo, Dr. Carlos Alberto Felgueiras, obrigado pelos ensinamentos e disposição em discutir parte deste trabalho.

Ao amigo, Dr. João Ricardo de Freitas Oliveira, minha gratidão pelo apoio e dedicação na revisão deste trabalho.

Ao meu amigo de sala, Lauro Tsutomu Hara, obrigado, não só pelo agradável convívio a mais de vinte anos, mas também pela paciência e compreensão ao aceitar o estresse natural decorrente do desenvolvimento de uma tese.

Aos meus amigos Júlio César Lima d'Alge, Cláudio Clemente Faria Barbosa, Fábio Furlan Gama, Ana Paula Dutra de Aguiar, Jussara de Oliveira Ortiz, Leila Maria G. Fonseca, Silvana Amaral Kampel, Lúbia Vinhas e Janete da Cunha, minha gratidão, pelos momentos de partilha na exaustiva caminhada de escrever uma tese.

Às secretárias do curso de pós-graduação em Sensoriamento Remoto, Maria Etelvina Rennó e Vera Gabriel da Silva, pela dedicação, a eficiência, a responsabilidade e o trabalho profissional.

Quando mencionamos nomes, incorremos no risco de esquecer alguém e, por isso, quero agradecer a todos que, embora não tenham sido mencionados, estiveram presentes comigo durante este percurso.

Finalmente agradeço à minha esposa, Rita de Cássia, e a meu filho, Eduardo Augusto, pela infinita paciência, carinho e compreensão.

RESUMO

Muitos eventos de interesse em políticas públicas como saúde e segurança são de baixa frequência de ocorrência ou eventos raros, como por exemplo, os vários tipos de câncer, os diversos tipos de violência e outros. Esses eventos se manifestam em pessoas, as quais não estão distribuídas aleatoriamente no espaço, e devido a isso, ao se trabalhar com registros de saúde e segurança para avaliar riscos, deve-se estimar a probabilidade do evento ocorrer. Neste contexto, ferramentas de análise que permitam produzir uma avaliação do risco e de sua distribuição espacial potencializam os meios de vigilância e, conseqüentemente, possibilitam fornecer informações importantes para o desenho de políticas de promoção da saúde e segurança, considerando novas estratégias de controle e prevenção. Esta tese oferece uma contribuição nesta direção. Uma metodologia geoestatística é empregada para a estimação e mapeamento do risco em eventos raros. Supõe-se que o risco é uma variável aleatória contínua e espacialmente correlacionada, cujos valores não são diretamente observados. A informação disponível são dados de taxa agregados por unidades de área (municípios, distritos, setores censitários e outros), definida como sendo a razão entre o número de eventos ocorridos numa determinada área e o número de pessoas expostas à ocorrência desse evento. Considera-se o número de casos ou eventos ocorridos numa determinada área como sendo uma variável aleatória com distribuição binomial. Um estimador para o semivariograma do risco é proposto para a análise e definição da estrutura de correlação espacial do risco. Este estimador considera na estimação de seus parâmetros a instabilidade que se observa nos dados e a sua tendência. Esta instabilidade é decorrente de áreas com pequenas populações. Seu comportamento é verificado através da co-krigeagem binomial e de um estudo de simulação que objetiva avaliar a estrutura de correlação espacial do risco estimada versus a simulada. A estrutura de correlação espacial do risco imposta pelo estimador proposto, conjuntamente com a informação disponível são empregadas no procedimento de co-krigeagem binomial para a obtenção de uma superfície da média da distribuição do risco. Para explorar outros momentos da distribuição do risco, emprega-se o procedimento de simulação seqüencial condicionada não-paramétrica. Um conjunto de realizações alternativas igualmente representativas do risco é gerado. Isto possibilita avaliar a probabilidade do campo aleatório em estudo exceder um dado valor de corte e, posteriormente, estabelecer cenários de risco mais adequados, por exemplo, para fins de planejamento de ações de vigilância e/ou intervenção. Como demonstração da técnica proposta, um estudo de caso é conduzido para o risco de homicídio na cidade de São Paulo, no triênio 2002 - 2004. Os resultados obtidos são condizentes com a estrutura urbana da cidade, através da observação de outros estudos, e apontam um caminho para o avanço nos sistemas de apoio a vigilância e a tomadas de decisões.

SPATIAL DISTRIBUTION OF RISK IN RARE EVENTS FOR BINOMIAL GEOSTATISTIC AND CONDITIONAL SIMULATION

ABSTRACT

Many events of interest in public policy, such as health care and security, are of low occurrence or rare events. The occurrence of various types of cancer and of diverse types of violence are some examples. These events are associated with people, who are not distributed randomly in space; therefore, when working with registers of health care and security to evaluate risks, the event probability must estimated. In this context, tools of analysis that allow to produce an evaluation of the risk, as well as its spatial distribution, enhance surveillance. Consequently, they make possible to provide important information towards the development of policies that promote better health care and security, considering new strategies of control and prevention. This thesis offers a contribution in this direction. A geostatistic methodology is used to estimate and map the risk in rare events. The risk is assumed to be a continuous and spatially correlated variable, whose values that are not observed directly. The available information are rates that are aggregated by units of area (towns, districts, census sectors and others), defined as being the ratio between the number of occurred events in one determined area and the number of people susceptible to the occurrence of this event. The number of cases or events occurred in one determined area is considered a variable with binomial distribution. An estimator for the risk semivariogram is presented for analysis and definition of the spatial correlation structure of the risk. This estimator considers the instability that is observed in the data and its trend in order to estimate its parameters. This instability is related with areas of the small populations. Its behavior is verified through the binomial co-kriging and a simulation study that evaluates the structure of spatial correlation of the estimated risk against the simulated one. The structure of spatial correlation of the risk imposed by the considered estimator and the available information are used in the binomial co-kriging procedure to obtain the average risk distribution surface. To explore other moments of the risk distribution, a non-parametric conditional sequential simulation procedure is used. A set of equally representative alternative realizations of the risk is generated. This makes possible to evaluate the probability of the random field in study exceed a given cut value, and, later, to establish more adequate risk scenarios, for example, action planning for monitoring and/or intervention. As a demonstration of the proposed technique, a case study is performed for the homicide risk in the city of São Paulo, for the period between 2002 and 2004. The results are suitable with its urban structure, through the observation of other studies, pointing a way to the advance of monitoring and decisions making support systems.

SUMÁRIO

Pág.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

CAPÍTULO 1 - INTRODUÇÃO 25
1.1 Motivação 25
1.2 Objetivo do Trabalho 29
1.3 Contribuições da Tese 30
1.4 Organização do Trabalho 31

CAPÍTULO 2 - REVISÃO BIBLIOGRÁFICA 35
2.1 Introdução 35
2.2 A influência da escala sobre dados agregados por unidade de área 36
2.3 Instabilidade de pequenas áreas 38
2.4 Abordagens da literatura 41

CAPÍTULO 3 - MODELAGEM DO RISCO POR GEOESTATÍSTICA BINOMIAL 47
3.1 Introdução 47
3.2 O modelo 48
3.3 Estimador para o semivariograma do risco 50
3.4 Estimador para o semivariograma do risco segundo Oliver 51
3.5 Estimador proposto para o semivariograma do risco 53
3.6 Componentes do estimador proposto para o semivariograma do risco 55
3.7 Co-krigeagem binomial 57
3.8 Simulação da distribuição empírica do semivariograma do risco 61
3.9 Construção de cenários do risco por simulação seqüencial condicionada 65

CAPÍTULO 4 - ESTUDO DE CASO: CENÁRIOS DO RISCO DE HOMICÍDIO NA CIDADE DE SÃO PAULO NO TRIÊNIO 2002 - 2004 71
4.1 Introdução 71
4.2 Área de estudo 74
4.3 Os dados de homicídios 75
4.4 Fluxograma de trabalho para geração de cenários do risco de homicídio 75
4.5 Análise preliminar dos dados 77
4.5.1 Estatísticas descritivas das taxas de homicídios 77
4.5.2 A instabilidade das taxas observadas 79
4.5.3 A tendência espacial dos dados de homicídios 81
4.6 Definição das zonas de risco 83
4.7 Análise da estrutura de correlação espacial do risco de homicídio 86
4.7.1 Impacto da estrutura de correlação espacial nas estimativas do risco de homicídio 93
4.7.2 Simulação da distribuição do semivariograma do risco 96
4.8 Estimação da superfície do risco de homicídio 100
4.9 Construção de cenários do risco de homicídio 106

CAPÍTULO 5 - CONSIDERAÇÕES FINAIS 119

REFERÊNCIAS BIBLIOGRÁFICAS 123

APÊNDICE A - DEDUÇÃO DO FORMALISMO PARA O SEMIVARIOGRAMA DO RISCO. 129

APÊNDICE B - TAXA DE HOMICÍDIO NA CIDADE DE SÃO PAULO EM 2002. 133

APÊNDICE C - TAXA DE HOMICÍDIO NA CIDADE DE SÃO PAULO EM 2003. 137

APÊNDICE D - TAXA DE HOMICÍDIO NA CIDADE DE SÃO PAULO EM 2004. 141

APÊNDICE E - DISTRITOS DA CIDADE DE SÃO PAULO. 145

APÊNDICE F - SÍNTESE DO FORMALISMO POR INDICAÇÃO PARA CONSTRUÇÃO DA FUNÇÃO DE DISTRIBUIÇÃO ACUMULADA CONDICIONADA - FDAC. 147

LISTA DE FIGURAS

Pág.

2.1 - Exemplo de unidades de área: (a) antes da redução de escala e (b) após a redução de escala..... 37

2.2 - Taxa de câncer de mama por distrito na cidade de São Paulo, em 2003. 39

2.3 - Taxa de câncer de mama na cidade de São Paulo em 2003 em função da população feminina em risco por distrito. 40

3.1 - Componentes do estimador proposto para o semivariograma do risco..... 55

3.2 - Diagrama de simulação para geração da distribuição empírica para o semivariograma do risco..... 63

3.3 - Exemplo ilustrativo de um valor de corte obtido da mediana da função de distribuição acumulada de R(**u**)..... 65

3.4 - Síntese do procedimento para construção de cenários por simulação seqüencial por indicação. 69

4.1 - Evolução da taxa média de homicídios na cidade de São Paulo de 1996 a 2005. . 72

4.2 - Mapa da cidade de São Paulo com destaque dos 96 distritos. 74

4.3 - Fluxograma de trabalho para geração de cenários do risco de homicídio. 76

4.4 - Esquema de cor *double-ended*..... 77

4.5 - *Box plot* das taxas de homicídios: (a) 2002, (b) 2003 e (c) 2004..... 79

4.6 - Instabilidade das taxas de homicídios versus a população em risco em 2002. 80

4.7 - Instabilidade das taxas de homicídios versus a população em risco em 2003. 80

4.8 - Instabilidade das taxas de homicídios versus a população em risco em 2004. 81

4.9 - Agrupamento estatístico por quintil em 2002: (a) taxas observadas de homicídios e (b) estimador de média móvel espacial. 82

4.10 - Agrupamento estatístico por quintil em 2003: (a) taxas observadas de homicídios e (b) estimador de média móvel espacial. 82

4.11 - Agrupamento estatístico por quintil em 2004: (a) taxas observadas de homicídios e (b) estimador de média móvel espacial. 83

4.12 - Estratificação e realce das zonas de risco: (a) 2002, (b) 2003 e (c) 2004. 85

4.13 - Dependência espacial do risco de homicídio em 2002. Estimador: (a) Oliver et al. (1998); (b) proposto com $W = 1$ e (c) proposto com $W = 2$. 88

4.14 - Dependência espacial do risco de homicídio em 2003. Estimador: (a) Oliver et al. (1998); (b) proposto com $W = 1$ e (c) proposto com $W = 2$. 89

4.15 - Dependência espacial do risco de homicídio em 2004. Estimador: (a) Oliver et al. (1998); (b) proposto com $W = 1$ e (c) proposto com $W = 2$. 90

4.16 - Estimativas do risco de homicídio e variâncias das estimativas nos centróides dos distritos, em 2002. 93

4.17 - Estimativas do risco de homicídio e variâncias das estimativas nos centróides dos distritos, em 2003. 94

4.18 - Estimativas do risco de homicídio e variâncias das estimativas nos centróides dos distritos, em 2004. 94

4.19 - Comportamento da estrutura de correlação espacial do risco de homicídio em relação à distribuição simulada do semivariograma do risco, para 2002: (a) segundo Oliver et al. (1998), (b) oriundo do estimador proposto com $W = 1$ e (c) conforme estimador proposto com $W = 2$. 97

4.20 - Comportamento da estrutura de correlação espacial do risco de homicídio em relação à distribuição simulada do semivariograma do risco, para 2003: (a) segundo Oliver et al. (1998), (b) oriundo do estimador proposto com $W = 1$ e (c) conforme estimador proposto com $W = 2$. 98

4.21 - Comportamento da estrutura de correlação espacial do risco de homicídio em relação à distribuição simulada do semivariograma do risco, para 2004: (a) segundo Oliver et al. (1998), (b) oriundo do estimador proposto com $W = 1$ e (c) conforme estimador proposto com $W = 2$. 99

4.22 - Evolução da média da distribuição do risco de homicídio, com detalhe da mancha urbana na imagem de fundo: (a) 2002, (b) 2003 e (c) 2004. 100

4.23 - Evolução das variâncias das estimativas do risco de homicídio: (a) 2002, (b) 2003 e (c) 2004..... 106

4.24 - *Box plot* das estimativas do risco de homicídio nos dos centróides: (a) 2002, (b) 2003 e (c) 2004..... 107

4.25 - Semivariogramas por indicação segundo valores de cortes em 2002: (a) 1º quintil, (b) 2º quintil, (c) 3º quintil e (d) 4º quintil. 108

4.26 - Semivariogramas por indicação segundo valores de cortes em 2003: (a) 1º quintil, (b) 2º quintil, (c) 3º quintil e (d) 4º quintil. 109

4.27 - Semivariogramas por indicação segundo valores de cortes em 2004: (a) 1º quintil, (b) 2º quintil, (c) 3º quintil e (d) 4º quintil. 110

4.28 - Em 2002: (a) Mapa do risco de homicídio por co-krigeagem binomial; (b), (c) e (d) realizações equiprováveis de R(**u**) oriundas da simulação. 111

4.29 - Evolução de cenários otimistas do risco de homicídio, com detalhe da mancha urbana na imagem de fundo: (a) 2002, (b) 2003 e (c) 2004. 112

4.30 - Evolução de cenários otimistas, com exposição somente das áreas em que o risco de homicídio excede 40 mortes por cem mil habitantes: (a) 2002, (b) 2003 e (c) 2004. 113

4.31 - Evolução de cenários pessimistas do risco de homicídio, com detalhe da mancha urbana na imagem de fundo: (a) 2002, (b) 2003 e (c) 2004. 114

4.32 - Evolução de cenários pessimistas, com exposição somente das áreas em que o risco de homicídio excede 40 mortes por cem mil habitantes: (a) 2002, (b) 2003 e (c) 2004. 114

4.33 - Evolução de cenários medianos do risco de homicídio, com detalhe da mancha urbana na imagem de fundo: (a) 2002, (b) 2003 e (c) 2004. 115

4.34 - Evolução de cenários medianos, com exposição somente das áreas em que o risco de homicídio excede 40 mortes por cem mil habitantes: (a) 2002, (b) 2003 e (c) 2004. 115

4.35 - Campos de incertezas gerados por simulação seqüencial por indicação: (a) 2002, (b) 2003 e (c) 2004. Cenários medianos do risco de homicídio: (d) 2002, (e) 2003 e (f) 2004. 116

E.1 - Distritos da cidade de São Paulo 145

F.1 - Exemplo ilustrativo do formalismo por indicação para 4 valores de corte..... 148

LISTA DE TABELAS

Pág.

4.1 - Estatísticas das taxas de homicídios no triênio 2002 - 2004. 78

4.2 - Sumário das zonas de risco em 2002, 2003 e 2004. 84

4.3 - Modelos teóricos de semivariogramas do risco, em 2002. 88

4.4 - Modelos teóricos de semivariogramas do risco, em 2003. 89

4.5 - Modelos teóricos de semivariogramas do risco, em 2004. 90

4.6 - Síntese das proporções obtidas sob cada alternativa. 92

4.7 - Síntese das médias das estimativas nos 96 centróides, valores expressos por 100 mil habitantes. 95

4.8 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos do centro da cidade de São Paulo. 101

4.9 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos em torno do centro da cidade de São Paulo. 102

4.10 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos da periferia Sul da cidade de São Paulo. 103

4.11 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos da periferia Leste da cidade de São Paulo. 104

4.12 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos da periferia Norte da cidade de São Paulo. 105

4.13 - Valores de cortes (por 100 mil habitantes) para os anos de 2002, 2003 e 2004. 107

4.14 - Parâmetros dos modelos esféricos de semivariogramas segundo os valores de cortes, para o ano de 2002. 108

4.15 - Parâmetros dos modelos esféricos de semivariogramas segundo os valores de cortes, para o ano de 2003. 109

4.16 - Parâmetros dos modelos esféricos de semivariogramas segundo os valores de cortes, para o ano de 2004. 110

B.1 - Taxa de homicídio na cidade de São Paulo em 2002. 133

C.1 - Taxa de homicídio na cidade de São Paulo em 2003. 137

D.1 - Taxa de homicídio na cidade de São Paulo em 2004. 141

LISTA DE SIGLAS E ABREVIATURAS

CID-10 – Classificação Internacional de Doenças -10ª revisão.

DATASUS – Departamento de Informática do SUS

EQM – Erro Quadrático Médio

FDAC – Função de Distribuição Acumulada Condicional

GSLIB – Geostatistical Software Library

IBGE – Instituto Brasileiro de Geografia e Estatística

LAT – Latitude

LONG – Longitude

MATLAB – MATrix LABoratory

PROAIM – Programa de Aprimoramento das Informações de Mortalidade

PRODAM – Processamento de Dados do Município de São Paulo

RNIS – Rede Nacional de Informações em Saúde

SMS – Secretaria Municipal da Saúde

SFMSP – Serviço Funerário do Município de São Paulo

SIM – Sistema de Informação sobre a Mortalidade

SINAN – Sistema de Informação de Agravos de Notificação

SINASC – Sistema de Informações de Nascidos Vivos

SPRING – Sistema de Processamento de Informações Georeferenciadas

SSG – Simulação Seqüencial Gaussiana

SSI – Simulação Seqüencial por Indicação

SUS – Sistema Único de Saúde

V.A. – Variável Aleatória

CAPÍTULO 1

INTRODUÇÃO

1.1 Motivação

Indicadores construídos a partir da coleta e sistematização de ocorrências pontuais agregadas por unidades de área político-administrativas ou operacionais, como estados e municípios ou setores censitários respectivamente, têm sido cada dia mais utilizados para a produção de mapas. Os mais comuns são os mapas gerados a partir de taxas brutas observadas, calculadas como a razão entre o número de ocorrências em uma determinada área e o número de pessoas expostas àquelas ocorrências. Os mapas, em geral, apresentam as taxas normalizadas, que se referem às ocorrências para cada 100, 1000, 10.000, 100.000 habitantes e consideram o conjunto das ocorrências dentro de um intervalo de tempo definido (mensal ou anual, por exemplo). A taxa bruta é o estimador mais simples e comumente utilizado para quantificar o risco de um desfecho particular como o óbito para uma certa doença ou a morte decorrente de algum tipo de violência, associado a uma população vivendo em uma área geográfica definida. Os mapas de taxas são então utilizados para representar a distribuição espacial da variável risco, não diretamente observada, e passam a orientar desde a formulação de políticas sociais, até o planejamento de intervenções, informando as ações de monitoração, vigilância e controle. Em particular, nas áreas da saúde coletiva e da segurança pública, um conjunto de decisões importantes para a qualidade e eficácia das políticas de contenção e proteção confia nesta quantificação para o risco estimado.

No Brasil, é no setor de saúde que encontramos um dos sistemas nacionais mais organizados direcionados para a coleta, sistematização, processamento e disseminação de dados e informações para apoiar a pesquisa, o planejamento, a gestão e os serviços em sua área fim. O DATASUS[1], Departamento de Informática do SUS – Sistema Único

[1] http://w3.datasus.gov.br/datasus/datasus.php

de Saúde, tem essa responsabilidade e organiza a RNIS – Rede Nacional de Informações em Saúde. Sistemas como o SINAN[2] – Sistema de Informação de Agravos de Notificação e o SIM[3] – Sistema de Informação sobre a Mortalidade são fundamentais para apoiar o desenho de políticas setoriais. A informação básica produzida e disponibilizada são taxas obtidas a partir dos agravos notificados e da captação de dados sobre mortalidade. Estas taxas, utilizadas como estimativa para o risco não observado, são apresentadas como mapas. Sobre estes mapas, estudos para a determinação e caracterização dos padrões espaciais das ocorrências, a detecção de agregados espaciais significativos (*clusters*), e a produção de mapas de "áreas quentes" (*hot-spots*) são usualmente conduzidos. Estes produtos derivados são freqüentemente utilizados pelos organismos de gestão para auxiliar nas decisões de alocação de recursos e na definição das estratégias de vigilância e controle, apoiando o desenho e implementação de políticas setoriais na área da saúde coletiva, da segurança pública e da análise de violência e criminalidade.

No entanto, a quantificação do risco a partir das taxas observadas pode ser muito pouco confiável quando estas taxas são calculadas para unidades de área com pequenas populações associadas a elas, e para acontecimentos com uma baixa freqüência de ocorrência [Beato et al. (1997) e Assunção et al. (1998)], denominado neste trabalho de *eventos raros*. Além disso, um outro problema que pode afetar a quantificação do risco é a tendência que se observa na informação disponibilizada, decorrente da heterogeneidade das áreas componentes da região de estudo (isto é, áreas com características demográficas e territoriais distintas) [Câmara et al. (2004)]. Os dados e indicadores que encontramos disponibilizados em sistemas como o SIM, o SINAN, o SINASC[4] - Sistema de Informações de Nascidos Vivos, o PROAIM[5] - Programa de Aprimoramento das Informações de Mortalidade no Município de São Paulo, o Sistema de Informações Municipais da Fundação SEADE[6], têm esta natureza. Os mapas produzidos a partir destes indicadores, subsidiam as diversas esferas de gestão na saúde

[2] http://dtr2004.saude.gov.br/sinanweb/index.php
[3] http://www.datasus.gov.br/catalogo/sim.htm
[4] http://www.datasus.gov.br/catalogo/sinasc.htm
[5] http://www6.prefeitura.sp.gov.br/secretarias/saude/publicacoes/0005
[6] http://www.seade.gov.br/

pública e na segurança auxiliando as análises de situação e o planejamento e avaliação de suas ações e dos seus programas.

A qualidade de decisões de gestão de políticas sociais importantes, como as da saúde coletiva e da segurança pública, pode ser afetada. Os sistemas de informação disponíveis, que já representam um imenso esforço e grande avanço no setor, fornecem as taxas observadas por unidade de área geográfica. Para unidades de área contendo pequenas populações as taxas tornam-se muito instáveis [Beato et al. (1997) e Assunção et al. (1998)]. Os mapas de risco são produzidos com base nestas taxas, na maior parte das vezes, sem observar os problemas da instabilidade e tendência presente nos dados. Desta maneira, informações importantes para o desenho de políticas de promoção da saúde e da segurança, considerando novas estratégias de controle e prevenção, são obtidas por um estimador "pobre" para o risco não observado.

Métodos de suavização (*smoothing*) têm sido utilizados em uma tentativa de tornar as estimativas para o risco não observado mais confiáveis. A idéia central passa pela filtragem das variações locais em pequenas escalas dos mapas de taxas agregadas, melhorando a análise das tendências regionais [Pickle (2002), Waller e Gotway (2004)]. Os métodos vão desde o uso de técnicas determinísticas [Kafadar (1994), Mungiole et al. (1999) e Talbot et al. (2000)] ao uso de técnicas estatísticas mais sofisticadas como os modelos na abordagem bayesiana completa (*full Bayesian Models*) [Christensen (2002) e Best (2005)], e os modelos geoestatísticos [Webster et al. (1994), Diggle et al. (1998), Kelsall e Wakefield (2002), Goovaerts et al. (2005) e Goovaerts (2005)]. É neste contexto que se insere o presente trabalho. Trata-se de uma metodologia probabilística baseada no paradigma geoestatístico, que objetiva a estimação e o mapeamento do risco associado a eventos raros, a partir de dados de taxa agregados por unidades de área.

A metodologia apresentada neste trabalho está baseada diretamente na proposta inicial de Lajaunie (1991), que considera o número de ocorrências do evento investigado uma variável aleatória com distribuição binomial. A partir desta consideração dá-se origem a função semivariograma do risco, que têm como parâmetros: i) o semivariograma

empírico das taxas observadas; ii) a média do risco; iii) as variâncias do risco e as populações sujeitas às ocorrências do evento investigado, ambas agregadas por unidades de área. Do ponto de vista prático, a grande dificuldade na utilização da formulação imposta por Lajaunie (1991) consiste na estimação das variâncias em cada uma das áreas componentes da região de estudo, porque o risco é desconhecido e, conseqüentemente, isto dificulta a estimação para o semivariograma do risco.

Oliver et al. (1998) apresentam uma alternativa para o semivariograma do risco sob a hipótese de homocedasticidade das taxas observadas e aplica a co-krigeagem binomial para produzir um mapa da distribuição do risco de câncer em crianças no centro-oeste da Inglaterra.

A alternativa apresentada por Oliver et al. (1998), para modelagem de eventos raros, pode ser em alguns casos limitada, porque propõe modelar um processo que supõe ser intrínseco, quando na realidade pode não ser. Em geral, eventos raros podem apresentar tendências, isto é, zonas de baixo e alto risco em localizações específicas dentro da região de estudo. Este fato pode ocorrer em regiões nas quais há uma configuração geométrica bastante heterogênea das áreas que compõem a região de estudo (agregação de grupos sociais distintos, diferenças em população e área), que é o caso das grandes cidades brasileiras. Além disso, uma outra questão importante a ser considerada é que a instabilidade que se observa nos dados, decorrente de áreas com pequenas populações, pode interferir diretamente na estimação da estrutura de correlação espacial do risco.

Avançando nesta direção, a metodologia apresentada neste trabalho propõe um novo estimador para o semivariograma do risco, o qual incorpora na estimação de seus parâmetros a instabilidade que se observa nos dados e sua tendência. Seu comportamento é verificado através da co-krigeagem binomial e de um estudo de simulação que objetiva avaliar a estrutura de correlação espacial do risco estimada versus a simulada. Mapas do risco de ocorrência do evento investigado são obtidos por co-krigeagem binomial. Esses mapas representam a média da distribuição do risco sobre a região de estudo. Para explorar outros momentos da distribuição do risco, emprega-se o procedimento de simulação seqüencial condicionada não-paramétrica (Goovaerts,

1997). Um conjunto de realizações equiprováveis ou igualmente representativas do risco é gerado. A partir deste conjunto (mapas simulados) é possível avaliar a probabilidade do campo aleatório em estudo exceder um dado valor de corte e, posteriormente, estabelecer cenários de risco, por exemplo, para fins de planejamento de ações de vigilância e/ou intervenção.

1.2 Objetivo do Trabalho

Premissas:

1) muitos eventos de interesse em políticas públicas como saúde e segurança são de baixa freqüência de ocorrência ou eventos raros, por exemplo, os vários tipos de câncer, os diversos tipos de violência, e outros. Esses eventos se manifestam em pessoas, e devido às características demográficas e territoriais vinculadas a uma determinada região de estudo a sua distribuição ocorre de modo desigual;

2) é importante mapear o risco de ocorrência destes eventos através de técnicas estatísticas mais robustas, que permitam produzir uma avaliação do risco e de seu padrão espacial. Isto potencializa os meios de vigilância e, conseqüentemente, possibilitam fornecer informações mais precisas para o desenho de políticas de promoção da saúde e segurança, considerando novas estratégias de controle e prevenção;

3) considera-se como informação disponível os dados de taxa agregados por unidades de área, como, por exemplo, municípios, distritos, setores censitários, e outros;

4) supõe-se que o risco de ocorrência do evento raro é uma variável aleatória contínua e espacialmente correlacionada, cujos valores podem ser estimados em todas as localizações da região de estudo. Esta hipótese decorre do fato que os levantamentos censitários muitas vezes impõem limites rígidos nas unidades de área a partir de critérios puramente operacionais que não têm relação direta com o evento modelado. Este fato leva a idéia de dissolver os limites das unidades de área em superfícies

contínuas, de forma a modelar melhor a real continuidade de, por exemplo, setores censitários em regiões urbanas densamente povoadas.

Dadas as premissas acima, o objetivo deste trabalho é desenvolver, implementar e aplicar uma metodologia para estimar a distribuição espacial do risco associado a eventos raros.

1.3 Contribuições da Tese

Configuram-se as principais contribuições deste trabalho:

1) tratar o problema por geoestatística binomial;

2) novo estimador para o semivariograma de risco, que é empregado para a análise da dependência espacial do risco. Seus parâmetros são estimados levando em conta a instabilidade que se observa nos dados e sua tendência;

3) uma metodologia baseada na co-krigeagem binomial e num esquema de simulação, para verificar o comportamento do estimador proposto para o semivariograma do risco;

4) uso do procedimento de simulação seqüencial condicionada não-paramétrica para: i) geração de cenários do risco para um nível de corte pré-estabelecido; ii) geração de mapas de incerteza;

5) determinação do risco de homicídio na cidade de São Paulo, no triênio 2002 – 2004, a partir de dados de taxa agregados para os 96 distritos político-administrativos.

Este trabalho se insere no contexto da linha de pesquisa e desenvolvimento da Divisão de Processamento de Imagens (DPI) do Instituto Nacional de Pesquisas Espaciais (INPE), que desde os anos 90, dentre outras atividades, vem investindo sistematicamente em novos procedimentos de tratamento, análise e modelagem de dados espaciais. Podemos citar os trabalhos de: a) Camargo (1997) – refere-se à introdução de procedimentos geoestatísticos num ambiente computacional de

geoprocessamento; b) evoluindo, Felgueiras (1999) – estabelece a modelagem ambiental com tratamento de incertezas sob o paradigma geoestatístico por indicação, c) Bönisch (2001) – apresenta uma metodologia em geoprocessamento ambiental com tratamento de incerteza, para o caso do zoneamento pedoclimático para a soja no estado de Santa Catarina e d) Ortiz et al. (2004) - modelagem de fertilidade do solo por simulação estocástica com tratamento de incertezas. Os trabalhos citados, assim como outros, são partes da evolução do Sistema para Processamento de Informações Georeferenciadas (SPRING) desenvolvido pela DPI [Câmara et al. (1996)]. Nos dias atuais, podemos citar o desenvolvimento da biblioteca de classes TerraLib [Vinhas e Ferreira (2005)] e produtos derivados como TerraView[7], TerraCrime uma parceria da DPI/INPE com o Laboratório de Estatística Espacial (LESTE) da Universidade Federal de Minas Gerais (UFMG) e a integração da TerraLib + R [Andrade Neto et al. (2005)] em parceria com Laboratório de Estatística e Geoinformação (LEG) da Universidade Federal do Paraná (UFPR).

1.4 Organização do Trabalho

Além deste capítulo, que descreve a motivação, o objetivo e a contribuição deste trabalho, este documento possui mais quatro Capítulos, os quais estão sintetizados nos parágrafos seguintes.

O Capítulo 2 apresenta uma revisão bibliográfica para estimação e mapeamento do risco em eventos raros, a partir de dados agregados por unidades de área. Inicialmente são discutidos os problemas decorrentes da definição de escala sobre um sistema de unidades de área. Depois, apresenta-se o problema da instabilidade que se observa nos dados, principalmente, nos casos de pequenas áreas em que se calcula taxas sobre um universo populacional reduzido. Finalmente, apresenta-se uma síntese de alguns métodos encontrados na literatura, sob diferentes abordagens, que permitem produzir uma avaliação do risco e de sua distribuição espacial.

[7] http://www.dpi.inpe.br/terraview

O Capítulo 3 introduz um modelo probabilístico para estimação e mapeamento do risco associado a eventos raros. Trata-se de uma abordagem geoestatística, que considera o risco um campo aleatório contínuo cujos valores podem ser estimados em todas as localizações da área de estudo. Considera-se como informação disponível os dados de taxa agregados por unidades de área, definida como sendo a razão entre o número de eventos ocorridos na unidade de área e o número de pessoas expostas à ocorrência desse evento. O número de casos ou eventos ocorridos numa determinada unidade de área é modelado como uma variável aleatória com distribuição binomial. A partir desta informação, desenvolve-se um novo estimador de semivariograma, que considera na estimação de seus parâmetros a instabilidade que se observa nos dados e a sua tendência. Esta instabilidade é decorrente de áreas com pequenas populações. O emprego deste estimador possibilita avaliar a dependência espacial do risco e, posteriormente, a definição de sua estrutura de correlação espacial. Para verificar o comportamento do estimador proposto emprega-se um esquema de simulação, que objetiva avaliar a estrutura de correlação espacial do risco estimada versus a simulada. Estimativas do risco são obtidas através do emprego do procedimento de co-krigeagem binomial. A formulação imposta no sistema de co-krigeagem binomial considera as covariâncias diretas entre as taxas observadas e as covariâncias cruzadas entre as taxas observadas e o risco. O resultado deste procedimento representa a média da distribuição do risco, e tem por objetivo fornecer uma primeira idéia das áreas de altos e baixos valores do risco decorrente do fenômeno em estudo. Finalmente, para explorar outros momentos da distribuição do risco, emprega-se o procedimento de simulação seqüencial condicionada não-paramétrica. Isto possibilita avaliar a probabilidade do campo aleatório em estudo exceder um dado valor de corte e, posteriormente, estabelecer diferentes cenários de risco.

O Capítulo 4 aplica o modelo de risco geoestatístico binomial, desenvolvido no Capítulo anterior, em um estudo de caso, no qual objetiva-se estimar e mapear o risco de homicídio na cidade de São Paulo, durante os anos de 2002 a 2004. A cidade de São Paulo está dividida em 96 distritos, e a informação disponível é a taxa de homicídio anual por 100 mil habitantes para cada distrito. Inicialmente, uma análise preliminar é

conduzida para detectar as principais características dos dados observados. Essa análise é realizada através de estatísticas descritivas, de gráficos que possibilitam verificar a instabilidade da informação decorrente de pequenas populações, e do emprego de técnicas de análise espacial que permitem detectar possíveis tendências presentes nos dados. Para verificar a dependência espacial do risco de homicídio na cidade de São Paulo e, posteriormente, definir sua estrutura de correlação espacial, aplica-se o estimador proposto de semivariograma. Estabelecida a estrutura de correlação espacial do risco de homicídio, para cada ano investigado, a mesma é avaliada através de um esquema de simulação, conforme descrito no Capítulo 3. Seguindo, uma superfície da média da distribuição do risco de homicídio é obtida por co-krigeagem binomial. Trata-se de um resultado suavizado que mostra as principais tendências do fenômeno investigado. Finalmente, para complementar o estudo de caso, cenários do risco de homicídio foram gerados por simulação seqüencial condicionada não-paramétrica. De um modo geral, os resultados obtidos são condizentes com a estrutura urbana da cidade, através de observações de outros estudos, e apontam um caminho para o avanço nos sistemas de apoio a vigilância e tomadas de decisões.

No Capítulo 5 apresentam-se as conclusões a respeito deste trabalho, com críticas e sugestões para futuros desenvolvimentos.

CAPÍTULO 2

REVISÃO BIBLIOGRÁFICA

2.1 Introdução

Muitos eventos de interesse em políticas públicas como saúde e segurança são de baixa frequência de ocorrência ou eventos raros. Como exemplo, podemos citar os vários tipos de câncer (mama, pulmão, útero, bexiga, etc), os diversos tipos de violência (homicídios, suicídios e acidentes de transportes), e outros. Esses eventos se manifestam em pessoas, as quais não estão distribuídas aleatoriamente no espaço, e devido a isso, ao se trabalhar com registros de saúde e segurança para avaliar riscos, deve-se estimar a probabilidade do evento ocorrer, ponderando-se pela população em risco.

A forma mais usual de se considerar a população na avaliação de riscos é através da agregação dos dados em unidades discretas de área. Essa agregação dos dados pode ocorrer por conveniência ou pode simplesmente refletir a forma como os dados foram disponibilizados, como por exemplo, por municípios, por distritos, por setores censitários, e outros.

Os eventos observados num sistema de unidades de área são, geralmente, expressos sob a forma de indicadores. A taxa bruta é o indicador (*estimador*) mais simples para o risco de ocorrência de um evento, definida como a razão entre o número de eventos ocorridos numa determinada área e o número de pessoas expostas à ocorrência desse evento. A medida obtida traduz o risco de "doença" ou óbito numa determinada população, para um intervalo de tempo definido, habitualmente um ano, sendo o seu valor multiplicado por uma constante, que pondera para um determinado número de pessoas (exs: 1.000 habitantes; 10.000 habitantes; 100.000 habitantes).

A estratégia de dados agregados por unidades de área, no entanto, apresenta alguns problemas [Beato et al. (1997), Assunção et al. (1998), Anselin et al. (2000) e Câmara et al. (2004)]:

1) as estimativas obtidas dentro de um sistema de unidades de área são funções das diversas maneiras que estas unidades podem ser agrupadas, ou da escala[8] de trabalho imposta. Este problema é descrito mais adiante na Seção 2.2;

2) associado às áreas componentes de uma região de estudo, existe também o problema da instabilidade que se observa nos dados, principalmente nos casos de pequenas áreas em que se calcula taxas sobre um universo populacional reduzido. Este problema é descrito com mais detalhes na Seção 2.3.

O presente Capítulo está organizado da seguinte forma: a Seção 2.2 limita-se a apresentar de forma direta e objetiva a influência da escala sobre as estimativas de indicadores agregados por unidades de área. Uma descrição mais ampla para o problema de dados agregados por área pode ser visto em [Openshaw (1984), Wakefield (2004) e Câmara et al. (2004)]. Na Seção 2.3 trata-se do problema da instabilidade de taxas em pequenas áreas. Para elucidar melhor este problema, considerou-se a distribuição da taxa de mortalidade de câncer de mama na cidade de São Paulo, em 2003. Na Seção 2.4 são apresentadas algumas abordagens da literatura para a análise, estimação e mapeamento do risco em eventos raros, a partir de dados agregados por unidades de área.

2.2 A influência da escala sobre dados agregados por unidade de área

Esta Seção aborda o problema da definição de escala inerente a dados agregados por unidades de área. Conforme já mencionado, a forma mais usual de se considerar a população na avaliação de riscos é através da agregação dos dados em unidades discretas de áreas, para posteriormente calcular indicadores como proporções, taxas, médias, e outros, ajustados dentro do sistema de unidades de área. Essa estratégia, no entanto, pode apresentar problemas; isto é, para uma mesma população estudada, a definição espacial das fronteiras das áreas afeta os resultados obtidos. As estimativas obtidas dentro de um sistema de unidades de área são funções das diversas maneiras que

[8] Neste trabalho, o conceito de escala se refere aos diferentes níveis de resolução espacial das unidades de área. Uma redução na escala significa áreas geográficas mais agregadas (em outras palavras, uma redução na escala significa ter áreas geográficas com áreas maiores e vice-versa, conforme Figura 2.1).

estas unidades podem ser agrupadas; pode-se obter resultados diferentes simplesmente alterando as fronteiras destas áreas [(Openshaw, 1984) e Câmara et al. (2004)]. Para enfatizar melhor este problema, considere o exemplo ilustrado na Figura 2.1.

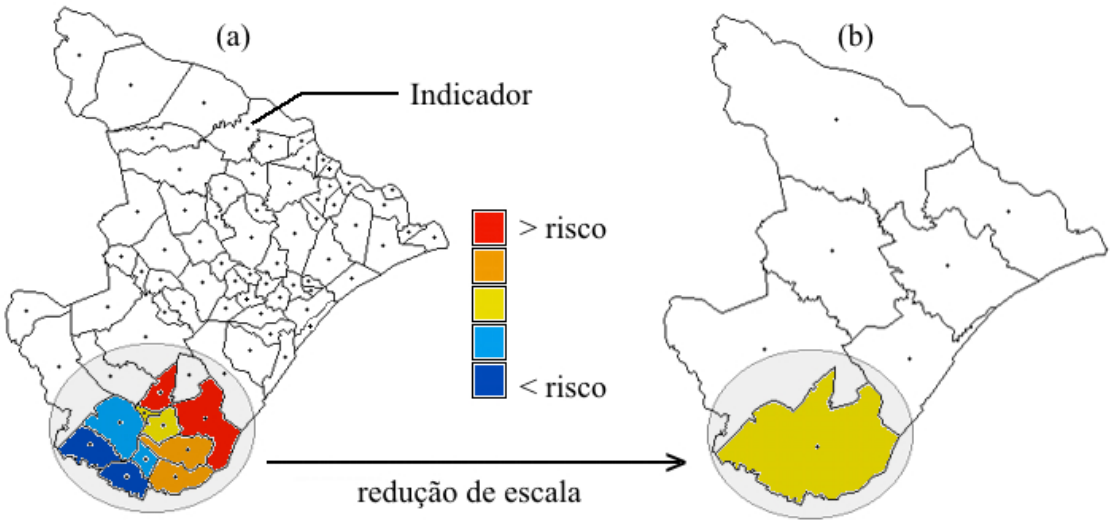


FIGURA 2.1 - Exemplo de unidades de área: (a) antes da redução de escala e (b) após a redução de escala.

O mapa mostrado na Figura 2.1 apresenta duas situações distintas. Inicialmente a região de estudo é composta de 68 unidades de área, Figura 2.1(a), e após a redução de escala, tem-se apenas 7 unidades de área, Figura 2.1(b). Tomando como referência a zona Sul em destaque das Figuras 2.1(a) e (b), ao se trabalhar com escalas reduzidas corre-se o risco de misturar numa mesma zona áreas completamente discrepantes. A diversidade das áreas componentes de uma zona é diluída numa média geral buscando-se representar as características de todas as áreas de forma simplificada através de um único indicador. Então, quanto maior a zona, maior a heterogeneidade das áreas componentes e, portanto, menos representativo o valor zonal para expressar a situação das áreas componentes (Câmara et al., 2004). Em outras palavras, uma redução de escala tende a homogeneizar os dados, reduzir as flutuações aleatórias e possivelmente reforçar as correlações que, assim, aparentam ser mais fortes que em áreas menores.

Deve ser observado também, que dependendo do nível de agregação estabelecido para as unidades de área, pode-se ter uma geometria de amostragem resultante, normalmente imposta pelos centróides das áreas componentes, pouco representativa da região de

estudo - conforme ilustrado anteriormente na Figura 2.1 (b). Neste caso, o emprego de modelos inferenciais que visam quantificar a dependência espacial entre as amostras e posteriormente construir superfícies contínuas do fenômeno investigado, tendem a produzir resultados mais suavizados, principalmente, nas regiões da área de estudo onde a geometria de amostragem é limitada.

De um modo geral, os métodos de análise espacial para dados agregados por áreas disponíveis na literatura são aplicados considerando-se um nível de agregação de áreas pré-estabelecido ou fixo; então, sempre que possível, uma análise detalhada que respeite e investigue a heterogeneidade existente nos dados exige uma definição de escala compatível com a realidade a ser investigada. A próxima Seção aborda a questão da instabilidade de pequenas áreas.

2.3 Instabilidade de pequenas áreas

Na Seção 2.2 apresentou-se o problema do efeito da escala na unidade de área, com a recomendação final de utilizar, sempre que possível, a melhor resolução espacial disponível. Associado às áreas componentes da região de estudo, existe também o problema da instabilidade que se observa nos dados, principalmente nos casos de pequenas áreas em que se calcula taxas sobre um universo populacional reduzido [Beato et al. (1997), Assunção et al. (1998) e Câmara et al. (2004)].

Para elucidar melhor o problema, considere a Figura 2.2, na qual se apresenta um mapa temático com mortalidade de câncer de mama[9] na cidade de São Paulo, em 2003. Nesse mapa, a cidade de São Paulo está dividida em 96 distritos, e a taxa de câncer de mama anual para cada distrito expressa o número de óbitos por 100 mil habitantes.

[9] Dados obtidos do Programa de Aprimoramento das Informações de Mortalidade – PROAIM, que é coordenado pela Secretaria de Saúde do Município de São Paulo – SSMSP.

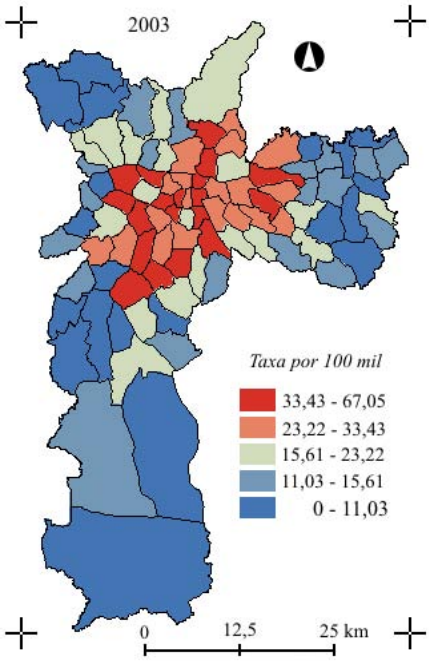


FIGURA 2.2 - Taxa de câncer de mama por distrito na cidade de São Paulo, em 2003.

Numa primeira análise, o mapa ilustrado na Figura 2.2 impressiona pelas altas taxas de mortalidade de vários distritos, com 19 distritos apresentando taxa maior que 33 mortes por 100 mil habitantes, e dois casos com taxas acima de 60 mortes por cem mil habitantes. Um observador desatento poderia concluir que todos esses distritos apresentam um grave problema decorrente do fenômeno investigado. Na realidade, muitos desses valores extremos ocorrem nos distritos com pequenas populações, pois a divisão imposta na cidade esconde enormes diferenças na população feminina em risco, variando de 4400 a 189232 habitantes por distrito. Por exemplo, suponha um distrito em que a população feminina é constituída no máximo por 5 mil habitantes, a ocorrência de um único caso de câncer de mama nesta população levaria à uma taxa de 20 por 100 mil, enquanto que a adição de apenas mais um caso faria a taxa pular para 40 por 100 mil. Esta instabilidade é uma característica de taxas de pequenas populações. Já para um distrito com 150 mil habitantes, a taxa de 20 por 100 mil ocorre quando 30 casos forem registrados. Para a taxa dobrar para 40 por 100 mil, como antes, é necessário a ocorrência de 30 casos adicionais, o que é mais improvável de acontecer.

Tais problemas são típicos de recobrimentos espaciais sobre divisões político-administrativas em que se analisam áreas com valores muito distintos de população em risco. Em vários estudos tem-se mostrado que divisões políticas como distritos e municípios apresentam relações inversas de área e população, isto é, os maiores distritos em população tendem a ter menores áreas, e vice-versa [Câmara et al. (2004)]. Por isso mesmo, freqüentemente, o que mais chama a atenção num mapa temático de taxas são os valores extremos, muitas vezes ocasionados pelo número reduzido de observações, resultando em valores menos confiáveis (apenas flutuação aleatória).

Na linguagem estatística dizemos que as taxas possuem variâncias muito diferentes. A dependência da variância das taxas em função do tamanho da população pode ser vista com clareza na Figura 2.3, que mostra um gráfico das taxas de câncer de mama versus a população feminina em risco para os distritos da cidade de São Paulo, em 2003.

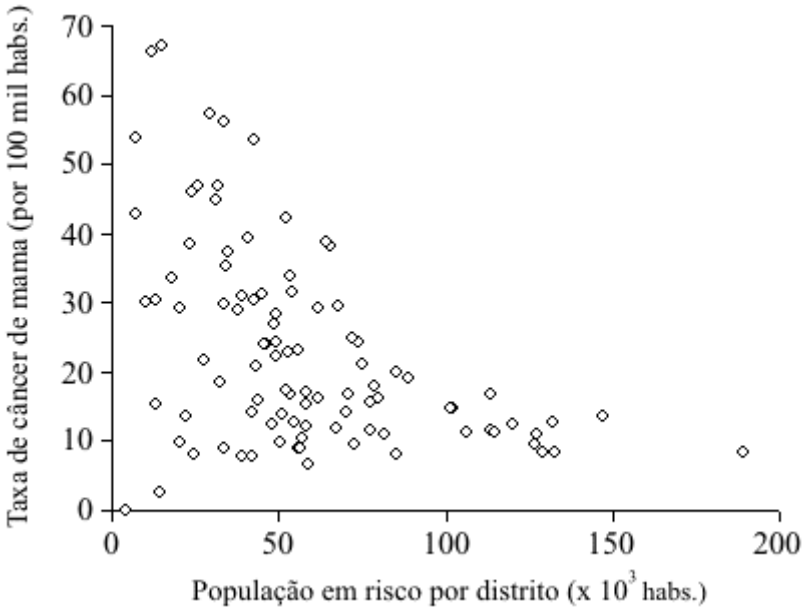


FIGURA 2.3 - Taxa de câncer de mama na cidade de São Paulo em 2003 em função da população feminina em risco por distrito.

No caso do município de São Paulo, a taxa média de câncer de mama em 2003 foi da ordem de 23 óbitos por cem mil habitantes. Como mostrado na Figura 2.3, os distritos com maior população apresentam taxas próximas da média da cidade. Conforme

diminui a população em risco, maior a variabilidade da taxa. Isto tende a produzir um efeito, denominado de “efeito funil”. Nos distritos com menor população, essa variação oscilou de zero a quase 70 por cem mil habitantes.

Esta Seção apresentou, através de um exemplo, o problema da instabilidade de taxas em pequenas populações. Mapas de eventos baseados diretamente nas estimativas de taxas são, normalmente, de difícil interpretação, e freqüentemente geram falsas conclusões. Para minimizar este problema é necessário empregar técnicas que permitam tratar a instabilidade intrínseca dos dados, de modo a obter estimativas mais próximas do risco real ao qual a população está exposta. A literatura aponta diversas soluções sob diferentes abordagens, conforme descrito na Seção seguinte.

2.4 Abordagens da literatura

Esta Seção apresenta uma síntese de alguns métodos encontrados na literatura, que podem ser empregados para a análise e estimação do risco, a partir de dados agregados por área. A análise aprofundada do formalismo dos métodos apontados, os conceitos teóricos, e implementações práticas, podem ser vistos nos trabalhos referenciados.

Muitos algoritmos de suavização estatísticos têm sido desenvolvidos para filtrar as variações locais em pequenas escalas dos mapas de mortalidade, melhorando as tendências regionais em grandes escalas [Kafadar (1994) e Lawson (2001)]. Esses métodos diferem extremamente quanto aos requerimentos computacionais, como também nas hipóteses subjacentes a respeito do padrão espacial e da distribuição dos valores do risco. Os algoritmos mais diretos são determinísticos e envolvem uma simples média ponderada de taxas vizinhas. Os pesos podem ser calculados, por exemplo, com o inverso do quadrado da distância entre o local a ser estimado e os centróides das áreas componentes (Talbot et al., 2000). Outros métodos mais elaborados consideram as observações vizinhas e a geometria espacial; versões ponderadas consideram a instabilidade intrínseca da informação oriunda de pequenas populações (Mungiole et al., 1999). Uma limitação dos métodos de suavização simples é que eles não são facilmente adaptáveis para considerar o padrão de variabilidade incorporado no

dado. Por exemplo, características importantes tais como a anisotropia (a variabilidade do fenômeno dependente da direção) e a distância de correlação espacial não são consideradas pelo método do inverso do quadrado da distância. Uma outra limitação importante é que, na ausência de alguma modelagem probabilística, a incerteza associada às estimativas não pode ser quantificada.

Métodos mais complexos têm sido desenvolvidos por estatísticos, combinando efeitos fixos com os efeitos aleatórios espacialmente estruturados e os efeitos não-correlacionados, conduzindo a modelos hierárquicos [Pickle (2000), Christensen e Waagepertesen (2002), Best et al. (2005)]. Muitos desses métodos foram desenvolvidos dentro de um paradigma bayesiano onde os parâmetros são tratados como quantidades aleatórias com distribuições a priori, os quais por sua vez, têm hiperparâmetros. A modelagem bayesiana completa atribui distribuições a priori para esses hiperparâmetros, o que permite que toda fonte de incerteza no modelo seja considerada. O custo para a flexibilidade de uma proposta bayesiana completa é a complexidade da estimação dos parâmetros do modelo. Este passo é executado usando-se procedimentos iterativos, tal como o método de Monte Carlo via cadeias de Markov (*Markov Chain Monte Carlo - MCMC*), que requer ajustes precisos de parâmetros, o que torna sua aplicação e interpretação desafiadora para os não estatísticos [Johnson (2004), Leyland e Davies (2005)]. Métodos bayesianos empíricos simplificam o procedimento de estimação [Clayton e Kaldor (1987), Marshall (1991), Martuzzi e Elliott (1996)]. A idéia central desses métodos é supor que as taxas observadas das diferentes áreas estão auto-correlacionadas, levando-se em conta o comportamento dos vizinhos para estimar uma nova taxa (filtrada) para as áreas de menor população. Embora os métodos empíricos menosprezem a variabilidade associada com a estimação do parâmetro, conduzindo a cálculos aproximados do risco, eles são facilmente implementados e empregados.

A modelagem probabilística na abordagem geoestatística está baseado na associação do conceito de variável regionalizada com o modelo de probabilidades, desenvolvida por Matheron (1963; 1970). A geoestatística fornece um conjunto de ferramentas estatísticas para a análise de dados distribuídos no espaço e no tempo, permite a descrição de padrões espaciais no dado, a incorporação de múltiplas fontes de

informação no mapeamento de atributos, a modelagem da incerteza espacial e sua propagação [Goovaerts (1997), Chilès e Delfiner (1999)]. Desde o seu desenvolvimento na industria de mineração, a geoestatística tem emergido como ferramenta primária para a análise de dados espaciais em vários campos, desde as ciências da Terra e atmosfera, na agricultura, nas ciências dos solos e hidrologia, estudos ambientais e mais recentemente na epidemiologia ambiental [Waller e Gotway (2004)].

A implementação tradicional dos métodos geoestatísticos, no entanto, não acomodam a heterocedasticidade das taxas de doenças e contagens; isto é, a variância em cada lugar assume valores em função do tamanho da população [Pickle (2002)]. Então, alternativas para o estimador de semivariograma e algoritmos de krigeagem apresentados por Matheron (1963; 1970) necessitam ser desenvolvidos para considerar a natureza específica dos dados de saúde.

Na literatura geoestatística encontram-se algumas abordagens que consideram o problema da não estacionariedade da variância causada pela variação do tamanho da população em risco. Uma solução que é mais direta de implementar é aplicar uma transformação nas taxas antes de executar a análise geoestatística clássica. Cressie (1993), (p. 385-402), analisando as taxas de incidência da síndrome de morte súbita em crianças em 100 condados na Carolina do Norte propõe dois tipos de transformações dos dados: uma para remover a dependência média-variância e outra para a correção da heterocedasticidade. O semivariograma tradicional foi então aplicado para os resíduos transformados. Ainda nessa linha, Berke (2004), para explorar a síndrome de morte súbita infantil na Carolina do Norte, propõe um procedimento de dois passos. Primeiro, as taxas são filtradas (suavizadas) através do estimador bayesiano empírico global [Marshall (1991), Martuzzi e Elliott (1996)], e depois são interpoladas por krigeagem para geração da superfície do risco. Apesar de sua simplicidade, a abordagem proposta por Berke (2004) apresenta algumas desvantagens, tais como, não consideram a incerteza associada com as taxas transformadas e a super suavização das estimativas do risco causada pela combinação dos métodos bayesiano empírico global e de krigeagem.

Kelsall e Wakefield (2002) propõem um procedimento geoestatístico mais complexo, baseado num modelo linear generalizado do dado regional, o qual é similar para os trabalhos de Diggle et al. (1998). Neste trabalho os autores investigaram a variabilidade espacial do risco de câncer de cólon do reto, associada com índices sócio-econômicos, nos distritos da cidade de Birmingham, na Inglaterra.

Uma outra abordagem é incorporar o impacto do tamanho da população diretamente no estimador do semivariograma empírico, conforme Rivoirard et al. (2000). Esta solução foi empregada por Goovaerts et al. (2005) conjuntamente com uma variante da krigeagem fatorial, onde os pesos do sistema de krigeagem são reescalonados a posteriori considerando-se o tamanho da população de cada observação. Esta abordagem foi aplicada em um estudo de caso para explorar relações de dependência de escala entre taxas de diferentes tipos de câncer. Conforme aponta Goovaerts et al. (2005) esta abordagem é relativamente direta e, como ilustrado em seu trabalho através de um extensivo estudo de simulação, proporciona estimativas mais próximas do risco real ao qual a população está exposta.

Algoritmos de krigeagem considerando a natureza binomial ou poisson do dado de contagem, foram inicialmente formulados por Lajaunie (1991), que apresenta a função de semivariograma do risco. A formulação imposta para o semivariograma do risco tem como parâmetros: i) o semivariograma empírico das taxas observadas; ii) a média do risco; iii) as variâncias do risco e as populações sujeitas às ocorrências do evento investigado, ambas agregadas por unidades de área. Do ponto de vista prático, a grande dificuldade na utilização da formulação imposta por Lajaunie (1991), consiste na estimação das variâncias risco em cada uma das áreas componentes da região de estudo, porque o risco é desconhecido e, conseqüentemente, isto dificulta a estimação para o semivariograma do risco. Posteriormente, Oliver et al. (1998) apresentam uma alternativa para o semivariograma do risco sob a hipótese de homocedasticidade das taxas observadas e aplica a co-krigeagem binomial para produzir um mapa da distribuição do risco de câncer em crianças no centro-oeste da Inglaterra. Neste caso, a variância do risco é estimada de forma iterativa como sendo o patamar de um modelo de ajuste teórico aplicado ao semivariograma do risco [Oliver et al. (1998)].

A hipótese imposta por Oliver et al. (1998) pode ser em alguns casos limitada, porque propõe modelar um processo que supõe ser estacionário, quando na realidade pode não ser, pois, em geral, muitos dos eventos raros de interesse de políticas públicas como saúde e segurança apresentam tendências, isto é, zonas de baixo e alto risco em localizações específicas dentro da região de estudo.

Avançando na direção proposta por Lajaunie (1991), este trabalho apresenta uma metodologia que emprega um novo estimador para o semivariograma do risco, que considera na estimação de seus parâmetros o problema da instabilidade que se observa nos dados e a sua tendência. Os detalhes de construção e o formalismo para o estimador proposto são apresentados no Capítulo 3, conforme segue.

CAPÍTULO 3

MODELAGEM DO RISCO POR GEOESTATÍSTICA BINOMIAL

3.1 Introdução

O presente capítulo introduz o modelo geoestatístico binomial para estimação e mapeamento do risco em eventos raros. Considera-se a informação disponível, os dados de taxa agregados por área (número de ocorrência do evento raro / população em risco). Para estabelecer a estrutura de correlação espacial do risco, um estimador de semivariograma é proposto, que considera na estimação de seus parâmetros a instabilidade dos dados, relacionadas à informação oriunda de pequenas populações, e sua tendência. Seu comportamento é verificado através da co-krigeagem binomial e de um estudo de simulação que objetiva avaliar a estrutura de correlação espacial do risco estimada versus a simulada. Uma superfície da média da distribuição do risco é estimada empregando-se o procedimento de co-krigeagem binomial. Trata-se de um resultado suavizado, que tem por objetivo fornecer uma primeira idéia das áreas de altos e baixos valores do risco. Para complementar, um conjunto de realizações alternativas igualmente representativas do risco é construído por simulação seqüencial condicionada não-paramétrica. Isto possibilita avaliar a probabilidade do campo aleatório em estudo exceder um dado valor de corte e, posteriormente, estabelecer cenários mais adequados, por exemplo, para fins de planejamento de ações de vigilância e/ou intervenção.

Este capítulo está organizado da seguinte forma: na Seção 3.2 conceitua-se o modelo binomial para eventos raros e seus respectivos momentos condicionais; na Seção 3.3 introduz-se o estimador para o semivariograma do risco, segundo Lajaunie (1991), e aborda-se algumas questões de ordem prática para o emprego desse estimador; na Seção 3.4 apresenta-se a alternativa proposta por Oliver et al. (1998) para o estimador do semivariograma do risco e suas limitações para modelagem de fenômenos raros; na Seção 3.5 mostra-se o estimador proposto para o semivariograma do risco e as correções impostas em seus respectivos parâmetros; a Seção 3.6 apresenta os componentes do estimador proposto para o semivariograma do risco e os efeitos esperados decorrentes

das correções impostas em seus parâmetros; na Seção 3.7 apresenta-se o estimador de co-krigeagem binomial, para estimação e mapeamento do risco; na Seção 3.8 apresenta-se um esquema de simulação para verificar o comportamento do estimador proposto e na Seção 3.9 trata-se de procedimentos de simulações seqüenciais condicionais para construção de cenários de risco do fenômeno em estudo.

3.2 O modelo

Considere uma região de estudo, *A*, composta de *N* áreas. Associado a cada área tem-se o número de casos de ocorrência do evento raro e a população residente, denotados por $l(\mathbf{u}_i)$ e $n(\mathbf{u}_i)$ respectivamente, em que o vetor de coordenadas espaciais $\mathbf{u}_i = (x_i, y_i)$, i = 1, ..., *N*, refere-se geograficamente ao centróide da i-ésima área. A variável aleatória (V.A.) $Z(\mathbf{u}_i)$ é denominada de taxa de ocorrência do evento raro e $z(\mathbf{u}_i)$, estabelecida pela razão $l(\mathbf{u}_i)/n(\mathbf{u}_i)$, é uma das possíveis realizações da V.A. $Z(\mathbf{u}_i)$.

Supõe-se que existe a ocorrência de um risco subjacente oriundo do evento raro, denotado por $R(\mathbf{u})$, decorrente de um processo estocástico $\{R(\mathbf{u}), \mathbf{u} \in A, A \subset \rfloor^2\}$ e que todo indivíduo em *A* está a ele submetido. Portanto, $R(\mathbf{u})$ é uma V.A. contínua e espacialmente correlacionada, cujos valores não são diretamente observados.

Neste ponto, é importante notar que há dois suportes geográficos distintos. Um refere-se às áreas geográficas que compõem a região de estudo *A*, para as quais existem valores observados $z(\mathbf{u}_i)$ da V.A. $Z(\mathbf{u}_i)$. O outro se refere à natureza do processo investigado, isto é, uma superfície contínua de $R(\mathbf{u})$. Uma discussão mais ampla da combinação de informações espaciais residindo em suportes geográficos distintos pode ser vista em Gotway e Young (2002) e Kyriakidis (2004).

Para estabelecer os momentos da distribuição de $Z(\mathbf{u}_i)$ em função de $R(\mathbf{u})$ é necessário conceituar o risco sob o mesmo suporte geográfico. Seja, $R(\mathbf{u}_i)$, i = 1, ..., *N*, o risco médio associado ao centróide do i-ésimo suporte geográfico de área v_i, é definido como (Lajaunie, 1991):

$$R(\mathbf{u}_i) = \frac{1}{v_i} \int_{v_i} R(\mathbf{u})\, d\mathbf{u} \tag{3.1}$$

Considera-se que o conceito de risco está associado à probabilidade, P($\mathbf{u}_i$), de ocorrer o evento raro em um indivíduo escolhido ao acaso, residindo no i-ésimo suporte geográfico de área v_i, durante um intervalo de tempo δt (Waller e Gotway, 2004). Assim:

$$R(\mathbf{u}_i) = \frac{P(\mathbf{u}_i)}{\delta t} \tag{3.2}$$

Entretanto, pode-se considerar sem perda de generalidade qualquer período de tempo como unidade. Assim, para $\delta t = 1 \Rightarrow R(\mathbf{u}_i) = P(\mathbf{u}_i)$.

A V.A. $L(\mathbf{u}_i)$ que mede o número de sucessos (o número de ocorrências do evento raro) nas $n(\mathbf{u}_i)$ repetições do i-ésimo suporte geográfico, tem distribuição binomial denotada por:

$$L(\mathbf{u}_i) \sim \text{Bi}\,[r(\mathbf{u}_i), n(\mathbf{u}_i)] \tag{3.3}$$

em que:

- Bi denota uma distribuição binomial;
- o parâmetro $r(\mathbf{u}_i) = p(\mathbf{u}_i)$;
- $n(\mathbf{u}_i)$ a população em risco no i-ésimo suporte geográfico.

Supõe-se também que os diferentes casos, relativos ao evento raro, ocorrem independentemente quando o risco é fixado (em outras palavras, o risco é a única fonte de correlação entre os casos). Então, as taxas observadas são variáveis com distribuição (Lajaunie, 1991):

$$Z(\mathbf{u}_i) \mid R \sim \frac{1}{n(\mathbf{u}_i)}\, \text{Bi}\left[r(\mathbf{u}_i),\, n(\mathbf{u}_i)\right] \tag{3.4}$$

A partir do modelo expresso na Equação (3.4) os seguintes momentos condicionais são estabelecidos (Lajaunie, 1991):

$$E[Z(\mathbf{u}_i) \mid R] = R(\mathbf{u}_i) \tag{3.5}$$

$$E\left[Z^2(\mathbf{u}_i) \mid R\right] = \frac{n(\mathbf{u}_i)\text{-}1}{n(\mathbf{u}_i)} R^2(\mathbf{u}_i) + \frac{R(\mathbf{u}_i)}{n(\mathbf{u}_i)} \tag{3.6}$$

$$E\left[Z(\mathbf{u}_i) Z(\mathbf{u}_j) \mid R\right] = R(\mathbf{u}_i) R(\mathbf{u}_j) \tag{3.7}$$

Os cálculos relativos aos momentos condicionais estabelecidos nas Equações (3.5), (3.6) e (3.7), estão disponibilizados no Apêndice A. Seguindo, a próxima seção apresenta o estimador para o semivariograma do risco.

3.3 Estimador para o semivariograma do risco

A partir dos momentos condicionais definidos nas Equações (3.5), (3.6) e (3.7), a seguinte relação (vide Apêndice A) é estabelecida (Lajaunie, 1991):

$$\gamma^R_{(\mathbf{u}_i, \mathbf{u}_j)} = \gamma^Z_{(\mathbf{u}_i, \mathbf{u}_j)} - \frac{1}{2}\left\{\frac{1}{n(\mathbf{u}_i)} + \frac{1}{n(\mathbf{u}_j)}\right\}\mu(1-\mu) + \frac{1}{2}\left\{\frac{\sigma^2_{R(\mathbf{u}_i)}}{n(\mathbf{u}_i)} + \frac{\sigma^2_{R(\mathbf{u}_j)}}{n(\mathbf{u}_j)}\right\} \tag{3.8}$$

em que:

- $\gamma^R_{(\mathbf{u}_i, \mathbf{u}_j)} = \frac{1}{2} E\left[R(\mathbf{u}_i) - R(\mathbf{u}_j)\right]^2$: é a função semivariograma do risco;
- $\gamma^Z_{(\mathbf{u}_i, \mathbf{u}_j)} = \frac{1}{2} E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\right]^2$: é a função semivariograma das taxas [Isaaks e Srivastava (1989), Goovaerts (1997), Deutsch e Journel (1998)];
- $n(\mathbf{u}_i)$ e $n(\mathbf{u}_j)$: são as populações no i-ésimo e j-ésimo suportes geográficos, com centróides em $\mathbf{u}_i$ e $\mathbf{u}_j$, respectivamente;
- $\mu = E[R(\mathbf{u})]$: refere-se ao risco médio;

- $\sigma^2_{R(\mathbf{u}_i)}$ e $\sigma^2_{R(\mathbf{u}_j)}$: são as variâncias do risco no i-ésimo e j-ésimo suportes geográficos, respectivamente.

A Equação (3.8) sugere como estimador a seguinte função:

$$\hat{\gamma}^R_{(\mathbf{u}_i,\mathbf{u}_j)} = \hat{\gamma}^Z_{(\mathbf{u}_i,\mathbf{u}_j)} - \frac{1}{2}\left[\frac{1}{n(\mathbf{u}_i)} + \frac{1}{n(\mathbf{u}_j)}\right]\hat{\mu}(1-\hat{\mu}) + \frac{1}{2}\left[\frac{\hat{\sigma}^2_{R(\mathbf{u}_i)}}{n(\mathbf{u}_i)} + \frac{\hat{\sigma}^2_{R(\mathbf{u}_j)}}{n(\mathbf{u}_j)}\right] \quad (3.9)$$

O estimador conforme a Equação (3.9) apresenta características diferentes do estimador tradicional de semivariograma, uma vez que o mesmo é função de um outro semivariograma e de parâmetros que devem ser estimados. Deve ser observado também, que o estimador $\hat{\gamma}^R_{(\mathbf{u}_i,\mathbf{u}_j)}$ depende das distâncias entre pontos e da localização espacial, uma vez que as populações, n($\mathbf{u}_i$) e n($\mathbf{u}_j$), e as variâncias $\sigma^2_{R(\mathbf{u}_i)}$ e $\sigma^2_{R(\mathbf{u}_j)}$, variam de lugar para lugar (Lajaunie, 1991).

De forma geral, o estimador do semivariograma quando modela um campo aleatório intrínseco depende somente do vetor distância **h**, em que $|\mathbf{h}| = \|\mathbf{u}_i - \mathbf{u}_j\|$, e neste caso é facilmente estimado.

Do ponto de vista prático, este fato não ocorre no estimador em consideração, Equação (3.9), o que traz grandes dificuldades no cálculo de suas estimativas (Lajaunie, 1991). Entretanto, Oliver et al. (1998) apresentam uma alternativa para o estimador do semivariograma do risco, conforme descrito a seguir.

3.4 Estimador para o semivariograma do risco segundo Oliver

A alternativa apresentada por Oliver et al. (1998) objetiva calcular o estimador do semivariograma do risco descrito na Equação (3.9) de forma aproximada, através de um estimador do semivariograma do risco que depende somente de **h**.

Inicialmente, Oliver et al. (1998) supõem que as variâncias nas localizações $\mathbf{u}_i$ e $\mathbf{u}_j$ são iguais, $\sigma^2_{R(\mathbf{u}_i)} = \sigma^2_{R(\mathbf{u}_j)} = \sigma^2_R$. Assim, a Equação (3.9) se reduz a:

$$\hat{\gamma}^R_{(\mathbf{u}_i,\mathbf{u}_j)} = \hat{\gamma}^Z_{(\mathbf{u}_i,\mathbf{u}_j)} - \frac{1}{2}\left[\hat{\mu}(1-\hat{\mu}) - \hat{\sigma}^2_R\right]\left[\frac{n(\mathbf{u}_i) + n(\mathbf{u}_j)}{n(\mathbf{u}_i)\,n(\mathbf{u}_j)}\right] \tag{3.10}$$

Em seguida, o fator $\left[\frac{n(\mathbf{u}_i) + n(\mathbf{u}_j)}{n(\mathbf{u}_i)\,n(\mathbf{u}_j)}\right]$ é modificado, colocando-se em seu lugar a sua média, a qual é calculada em função do vetor distância **h.** Desta forma, o estimador proposto por Oliver et al. (1998) é denotado por:

$$\hat{\gamma}^R_{(\mathbf{u}_i,\mathbf{u}_j)} = \hat{\gamma}^Z_{(\mathbf{u}_i,\mathbf{u}_j)} - \frac{1}{2}\left[\hat{\mu}(1-\hat{\mu}) - \hat{\sigma}^2_R\right]\overline{\left[\frac{n(\mathbf{u}_i) + n(\mathbf{u}_j)}{n(\mathbf{u}_i)\,n(\mathbf{u}_j)}\right]} \tag{3.11}$$

em que:

- $\hat{\gamma}^Z_{(\mathbf{h})} = \frac{1}{2\,M(\mathbf{h})}\sum_{i=1}^{M(\mathbf{h})}\left[z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})\right]^2$: é o estimador para o semivariograma empírico das taxas observadas, onde M(**h**) refere-se ao número total de pares de pontos disponíveis para uma certa distância de análise (*lag*) [Isaaks e Srivastava (1989), Goovaerts (1997), Deutsch e Journel (1998)];

- $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} z(\mathbf{u}_i)$

- $\hat{\sigma}^2_R$: é o estimador para a variância do risco. Segundo Oliver et al. (1998) é calculado de forma iterativa como sendo o patamar de um modelo de ajuste teórico aplicado ao semivariograma do risco;

- $\left[\overline{\frac{n(\mathbf{u}_i)+n(\mathbf{u}_i+\mathbf{h})}{n(\mathbf{u}_i)\,n(\mathbf{u}_i+\mathbf{h})}}\right]$: refere-se a média de todos os pares de localizações envolvidos no cálculo do vetor distância **h**. É calculado da seguinte forma (Oliver et al., 1998): $\frac{1}{M(\mathbf{h})}\sum_{i=1}^{M(\mathbf{h})}\frac{n(\mathbf{u}_i)+n(\mathbf{u}_i+\mathbf{h})}{n(\mathbf{u}_i)\,n(\mathbf{u}_i+\mathbf{h})}$.

3.5 Estimador proposto para o semivariograma do risco

A alternativa apresentada por Oliver et al. (1998), Equação (3.11), para modelagem de eventos raros, como, por exemplo, os vários tipos de câncer, os diversos tipos de violência e outros, pode ser em alguns casos limitada, porque propõe modelar um processo que supõe ser intrínseco, quando na realidade pode não ser. Em geral, eventos raros podem apresentar tendências, isto é, zonas de baixo e alto risco em localizações específicas dentro da região de estudo. Este fato pode ocorrer em regiões nas quais há uma configuração geométrica bastante heterogênea das áreas que compõem a região de estudo.

Uma outra questão importante a ser considerada é que a instabilidade que se observa nos dados, decorrente de áreas com pequenas populações (conforme discutido na Seção 2.3), pode interferir diretamente na estimação da estrutura de correlação espacial do risco; isto é, a V.A. $Z(\mathbf{u}_i)$ incorpora um erro que varia de lugar para lugar, dependendo do tamanho da população, $n(\mathbf{u}_i)$, associada. Esse erro se propaga nas estimativas das semivariâncias de $Z(\mathbf{u}_i)$ e conseqüentemente nas estimativas das semivariâncias do risco, podendo resultar em estruturas de correlações espaciais do risco que não condizem com a variabilidade do fenômeno investigado.

Para contornar os problemas mencionados acima, um novo estimador para o semivariograma do risco é proposto, denotado por $\hat{\gamma}^{*R}_{(\mathbf{h})}$. Este estimador incorpora duas modificações na estimação de seus parâmetros:

1) A primeira é incorporar o impacto do tamanho da população diretamente na estimação do semivariograma empírico das taxas, $\hat{\gamma}^{*Z}_{(\mathbf{h})}$, para remover a instabilidade das taxas;

2) A segunda é incorporar estimativas de médias e variâncias zonais para tratar da tendência dos dados. Para tal, a área em estudo é dividida inicialmente em *W* zonas de risco, supostamente homogêneas.

A incorporação dessas duas modificações em $\hat{\gamma}^{*R}_{(\mathbf{h})}$, conduz à seguinte notação:

$$\hat{\gamma}^{*R}_{(\mathbf{u}_i,\mathbf{u}_j)} = \hat{\gamma}^{*Z}_{(\mathbf{u}_i,\mathbf{u}_j)} - \frac{1}{2}\left[\hat{\mu}_w(1-\hat{\mu}_w) - \hat{\sigma}^2_{R_w}\right]\left[\frac{n(\mathbf{u}_i)+n(\mathbf{u}_j)}{n(\mathbf{u}_i)\ n(\mathbf{u}_j)}\right] \qquad (3.12)$$

em que:

- $\hat{\gamma}^{*Z}_{(\mathbf{u}_i,\mathbf{u}_j)} = \dfrac{1}{2\sum_{i,j=1}^{M(\mathbf{u}_i,\mathbf{u}_j)} n(\mathbf{u}_i)\,n(\mathbf{u}_j)} \sum_{i,j=1}^{M(\mathbf{u}_i,\mathbf{u}_j)} \{n(\mathbf{u}_i)\,n(\mathbf{u}_j)\left[z(\mathbf{u}_i)-z(\mathbf{u}_j)\right]^2\}$, é denominado estimador de semivariograma empírico ponderado pela população;

- $\hat{\mu}^*_w = \dfrac{1}{n_w}\sum_{i=1}^{n_w} z(\mathbf{u}_i)$ e $\hat{\sigma}^{2*}_{R_w} = \sum_{i=1}^{n_w} \dfrac{\left[z(\mathbf{u}_i)-\hat{\mu}^*\right]^2}{n_w-1}$, são escolhidos adequadamente dentre as *W* médias e variâncias zonais do risco, dependendo das posições $\mathbf{u}_i$ e $\mathbf{u}_j = \mathbf{u}_i + \mathbf{h}$; onde $w = 1, ..., W$ representa as zonas de risco e n_w refere-se ao número de observações $z(\mathbf{u}_i)$ contidos em *w*. Neste trabalho, o critério adotado para a aplicação das médias e das variâncias zonais é estabelecido da seguinte forma: para um certo vetor distância **h** de análise, verifica-se em qual zona de risco a maioria dos pares de pontos com localizações em $\mathbf{u}_i$ e $\mathbf{u}_i + \mathbf{h}$ se encontra; em seguida, aplica-se a correspondente média e variância zonal estimadas por $\hat{\mu}^*_w$ e $\hat{\sigma}^{2*}_{R_w}$, respectivamente. Deve ser realçado que o critério adotado é empírico, portanto não deve ser considerado único. É importante investigar

formas alternativas e avaliá-las, um tema que será conduzido num trabalho futuro.

- $\left[\overline{\frac{n(\mathbf{u}_i)+\ n(\mathbf{u}_i+\mathbf{h})}{n(\mathbf{u}_i)\,n(\mathbf{u}_i+\mathbf{h})}}\right]$: conforme definido anteriormente na página 29.

3.6 Componentes do estimador proposto para o semivariograma do risco

A formulação proposta para o estimador do semivariograma do risco, Equação (3.12), pode ser vista como sendo a diferença entre dois componentes com funções distintas:

1) O primeiro componente refere-se ao termo $\hat{\gamma}^{*Z}_{(\mathbf{h})}$, o qual rege a estrutura de correlação espacial do risco;
2) O segundo componente refere-se ao termo

   $-\frac{1}{2}\left[\hat{\mu}_w(1-\hat{\mu}_w)-\hat{\sigma}^2_{R_w}\right]\left[\overline{\frac{n(\mathbf{u}_i)+n(\mathbf{u}_j)}{n(\mathbf{u}_i)\ n(\mathbf{u}_j)}}\right]$, que age como um fator de correção (um *offset* negativo) para cada distância (*lag*) investigada.

Para elucidar melhor o papel desses dois componentes considere o exemplo ilustrado na Figura 3.1.

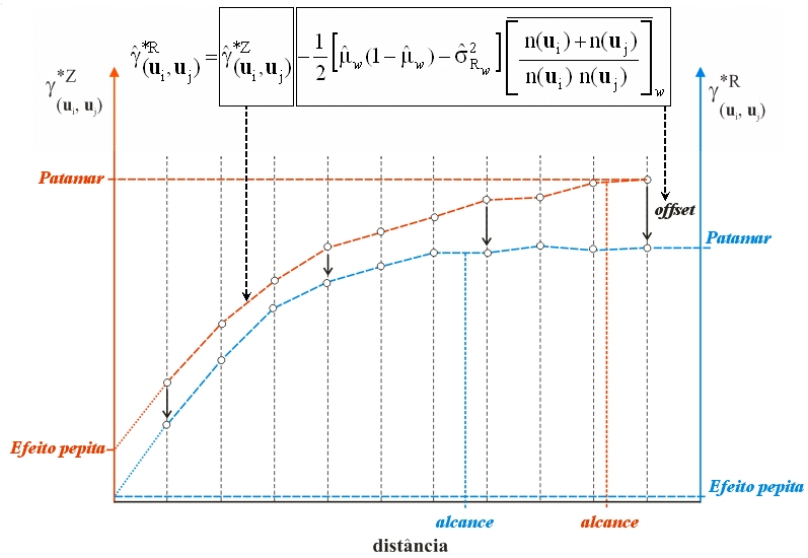


FIGURA 3.1 - Componentes do estimador proposto para o semivariograma do risco.

Tomando como referência a Figura 3.1, os efeitos esperados decorrentes das correções impostas para o estimador proposto, $\hat{\gamma}^{*R}_{(\mathbf{h})}$, são:

1) Uma diminuição acentuada do efeito pepita (C_0). O efeito pepita representa a quantidade de variação ao acaso de uma localização para outra. Mais precisamente, quanto menor for a proporção do efeito pepita em relação ao patamar do semivariograma do risco, maior a continuidade do fenômeno investigado (Chilès e Delfiner, 1999). Assim, a diminuição do efeito pepita pode representar que o modelo conseguiu filtrar alguns aspectos da instabilidade presente em populações (principalmente se a configuração geométrica da região de estudo é constituída de áreas de pequenas populações próximas).

2) Espera-se também que em relação às estimativas obtidas pelo estimador de Oliver et al. (1998), Equação (3.11), haja uma diminuição do valor estimado do patamar (C) ($C \cong$ à variância dos dados observados) e um valor estimado de alcance (a) condizente com as dimensões da área de estudo, caso uma tendência expressa nos dados seja observada.

Uma vez estabelecido $\hat{\gamma}^{*R}_{(\mathbf{h})}$, o mesmo é aplicado sobre a informação disponível para a análise da dependência espacial do risco. O próximo passo é a identificação de um modelo teórico que melhor se ajusta a $\hat{\gamma}^{*R}_{(\mathbf{h})}$. Entre os modelos teóricos de semivariogramas mais freqüentemente utilizados na geoestatística destacam-se: o Gaussiano, o Exponencial, o Esférico e o Linear, os quais podem ser vistos em Journel e Huijbregts (1978), Chilès e Delfiner (1999), Deutsch e Journel (1998) e outras literaturas correlatas. Neste trabalho, o modelo teórico de ajuste aplicado ao semivariograma do risco é denotado por "g". Assim, $g^{*R}_{(\mathbf{h})}$, refere-se a um modelo teórico de ajuste aplicado ao semivariograma do risco decorrente do estimador proposto, Equação (3.12). De maneira análoga, $g^{R}_{(\mathbf{h})}$, refere-se a um modelo de ajuste aplicado ao semivariograma do risco decorrente do estimador de Oliver et al. (1998), Equação (3.11).

Seguindo, estimativas do risco em localizações dentro da região de estudo são estabelecidas, conforme descrito na próxima seção.

3.7 Co-krigeagem binomial

As diversas técnicas do âmbito da geoestatística que possibilitam estimar fenômenos regionalizados são conhecidas sob a designação genérica de krigeagem, nome que foi dado por Matheron (1963) em homenagem ao matemático sul-africano Daniel G. Krige.

Objetiva-se estimar o risco, R(**u**), a partir da informação disponível Z($\mathbf{u}_i$). Na geoestatística, a estimação de um atributo à custa de outro atributo se insere no âmbito da co-krigeagem. Neste trabalho, e de acordo com Oliver et al. (1998) e Webster et al. (1994), denomina-se de co-krigeagem binomial, decorrente do caráter binomial dos dados, para a qual são necessários as covariâncias diretas entre Z($\mathbf{u}_i$) e Z($\mathbf{u}_j$), e as covariâncias cruzadas entre Z($\mathbf{u}_i$) e o risco R(**u**).

Na co-krigeagem binomial o risco em uma localização desconhecida $\mathbf{u}_0$ pode ser estimado por uma combinação linear de *k* taxas vizinhas z($\mathbf{u}_i$), da seguinte forma [Oliver et al. (1998)]:

$$\hat{R}(\mathbf{u}_0) = \sum_{i=1}^{k} \lambda(\mathbf{u}_i)\, z(\mathbf{u}_i) \qquad (3.13)$$

em que:

- z($\mathbf{u}_i$): é a taxa observada no i-ésimo suporte geográfico com centróide em $\mathbf{u}_i$;
- *k*: refere-se ao número de centróides considerados no cálculo de $\hat{R}(\mathbf{u}_0)$. Na prática, somente os centróides mais próximos de $\mathbf{u}_0$ são considerados;
- $\lambda(\mathbf{u}_i)$: é o peso atribuído à i-ésima observação z($\mathbf{u}_i$).

Na Equação (3.13), os pesos $\lambda(\mathbf{u}_i)$ são calculados considerando-se a estrutura de correlação espacial do risco e duas propriedades básicas:

1) Não tendência: significa que, em média, a diferença entre valor estimado e o verdadeiro valor para o mesmo ponto deve ser nula, então:

$$E\left[\hat{R}(\mathbf{u}_0) - R(\mathbf{u}_0)\right] = 0 \tag{3.14}$$

$$E\left[\sum_{i=1}^{k}\lambda(\mathbf{u}_i)\,z(\mathbf{u}_i) - R(\mathbf{u}_0)\right] = 0 \tag{3.15}$$

$$\sum_{i=1}^{k}\lambda(\mathbf{u}_i)\,E\left[z(\mathbf{u}_i)\right] - E\left[R(\mathbf{u}_0)\right] = 0 \tag{3.16}$$

$$\sum_{i=1}^{k}\lambda(\mathbf{u}_i)\,\mu - \mu = 0 \tag{3.17}$$

$$\mu\left[\sum_{i=1}^{k}\lambda(\mathbf{u}_i) - 1\right] = 0 \;\Rightarrow\; \sum_{i=1}^{k}\lambda(\mathbf{u}_i) = 1 \tag{3.18}$$

Portanto, para que a estimativa não tenha tendência, é necessário que a soma dos pesos seja igual a um.

2) Variância mínima: significa que o estimador possui a menor variância dentre todos os estimadores lineares; assim:

$$\sigma^2_{R(\mathbf{u}_0)} = E\left\{\left[\hat{R}(\mathbf{u}_0) - R(\mathbf{u}_0)\right]^2\right\} \text{ é mínima} \tag{3.19}$$

Seguindo, a Equação (3.19) deve ser minimizada sob a restrição de que $\sum_{i=1}^{k}\lambda(\mathbf{u}_i) = 1$. O processo de minimização é estabelecido através de técnicas de lagrange, que são suficientemente explicadas em livros de cálculo avançado, e pode ser visto, por

exemplo, em Journel (1998). Minimizando $\sigma^2_{R(\mathbf{u}_0)}$ sob a restrição de que $\sum_{i=1}^{k}\lambda(\mathbf{u}_i) = 1$, os pesos $\lambda(\mathbf{u}_i)$ são obtidos a partir do sistema de equações (Lajaunie, 1991):

$$\begin{cases} \sum_{i=1}^{k}\lambda(\mathbf{u}_i)\,C^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} + \varphi = C^{ZR}_{(\mathbf{u}_j,\mathbf{u}_0)} \quad \text{para todo } j = 1,...,k \\ \sum_{i=1}^{k}\lambda(\mathbf{u}_i) = 1 \end{cases} \tag{3.20}$$

em que:

- $C^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)}$: são as covariâncias diretas entre as V.A. $Z(\mathbf{u}_i)$ e $Z(\mathbf{u}_j)$;
- $C^{ZR}_{(\mathbf{u}_j,\mathbf{u}_0)}$: são as covariâncias cruzadas entre as V.A. $Z(\mathbf{u}_j)$ e o risco $R(\mathbf{u}_0)$ no ponto a ser estimado;
- φ : é o multiplicador de Lagrange (*empregado no processo de minimização*);
- k e $\lambda(\mathbf{u}_i)$: conforme definidos anteriormente na Equação (3.13).

O sistema de Equações (3.20) pode ser escrito em notação matricial como:

$$\mathbf{T}.\boldsymbol{\lambda} = \mathbf{t} \quad \Rightarrow \quad \boldsymbol{\lambda} = \mathbf{T}^{-1}.\mathbf{t} \tag{3.21}$$

em que:

- $\mathbf{T} = \begin{vmatrix} C^{Z}_{(\mathbf{u}_1,\mathbf{u}_1)} & C^{Z}_{(\mathbf{u}_1,\mathbf{u}_2)} & \cdots & C^{Z}_{(\mathbf{u}_1,\mathbf{u}_k)} & 1 \\ C^{Z}_{(\mathbf{u}_2,\mathbf{u}_1)} & C^{Z}_{(\mathbf{u}_2,\mathbf{u}_2)} & \cdots & C^{Z}_{(\mathbf{u}_2,\mathbf{u}_k)} & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ C^{Z}_{(\mathbf{u}_k,\mathbf{u}_1)} & C^{Z}_{(\mathbf{u}_k,\mathbf{u}_2)} & \cdots & C^{Z}_{(\mathbf{u}_k,\mathbf{u}_k)} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{vmatrix}$;

- $\boldsymbol{\lambda} = \begin{vmatrix} \lambda(\mathbf{u}_1) \\ \lambda(\mathbf{u}_2) \\ \vdots \\ \lambda(\mathbf{u}_k) \\ \varphi \end{vmatrix}$ e $\mathbf{t} = \begin{vmatrix} C^{ZR}_{(\mathbf{u}_1,\mathbf{u}_0)} \\ C^{ZR}_{(\mathbf{u}_2,\mathbf{u}_0)} \\ \vdots \\ C^{ZR}_{(\mathbf{u}_k,\mathbf{u}_0)} \\ 1 \end{vmatrix}$

A solução do sistema de Equações (3.20) e/ou (3.21) requer o conhecimento prévio das covariâncias $C^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)}$ e $C^{ZR}_{(\mathbf{u}_i,\mathbf{u}_0)}$. Lajaunie (1991) demonstra que as covariâncias das taxas observadas $C^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)}$ e as covariâncias cruzadas $C^{ZR}_{(\mathbf{u}_i,\mathbf{u}_0)}$, entre a taxa observada e o risco, dependem da estrutura da covariância do risco $C^{R}_{(\mathbf{u}_i,\mathbf{u}_j)}$, ou seja:

$$C^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} = \begin{cases} C^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} & \text{se } i \neq j \\ C^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} + \dfrac{1}{n(\mathbf{u}_i)}\mu(1-\mu) & \text{se } i = j \end{cases} \tag{3.22}$$

$$C^{ZR}_{(\mathbf{u}_j,\mathbf{u}_0)} = C^{R}_{(\mathbf{u}_j,\mathbf{u}_0)} \tag{3.23}$$

em que:

- μ refere-se ao risco médio, estimado a partir das taxas $Z(\mathbf{u}_i)$;

- $C^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} = C^{R}_{(\mathbf{h})} = C^{R}_{(\mathbf{0})} - \gamma^{*R}_{(\mathbf{h})} = \sigma^{2}_{R} - \gamma^{*R}_{(\mathbf{h})}$, onde σ^{2}_{R} e $\gamma^{*R}_{(\mathbf{h})}$, são estimados a partir do modelo teórico para o semivariograma do risco.

Além da estimativa do risco em uma determinada localização **u**, o sistema de Equações (3.20) fornece também a dispersão em torno da estimativa resultante do risco, denominada neste trabalho de variância de co-krigeagem, calculada da seguinte forma (Webster et al., 1994):

$$\hat{\sigma}^{2}_{R}(\mathbf{u}_0) = C^{Z}_{(\mathbf{0})} - \sum_{i=1}^{k} \lambda(\mathbf{u}_i) C^{ZR}_{(\mathbf{u}_i,\mathbf{u}_0)} + \varphi = C^{R}_{(\mathbf{0})} - \sum_{i=1}^{k} \lambda(\mathbf{u}_i) C^{R}_{(\mathbf{u}_i,\mathbf{u}_0)} + \varphi \quad \text{se } i \neq 0 \quad (3.24)$$

A Equação (3.24) em termos matriciais, fica:

$$\hat{\sigma}^{2}_{R}(\mathbf{u}_0) = C^{Z}_{(\mathbf{0})} - \boldsymbol{\lambda}^{t}\,\mathbf{t} = C^{R}_{(\mathbf{0})} - \boldsymbol{\lambda}^{t}\,\mathbf{t} \quad (3.25)$$

A precisão dos resultados decorrentes da co-krigeagem binomial depende substancialmente da estrutura de correlação espacial imposta pelo estimador do semivariograma do risco. Neste sentido, é importante verificar a adequação do estimador proposto, $\hat{\gamma}^{*R}_{(\mathbf{h})}$. Para isto, um estudo de simulação é conduzido, conforme descrito na seção seguinte.

3.8 Simulação da distribuição empírica do semivariograma do risco

Para observar o comportamento do estimador proposto, um estudo de simulação é conduzido com o objetivo de avaliar a estrutura de correlação espacial do risco estimada versus a simulada. O procedimento de simulação considera as duas fontes de aleatoriedade incorporadas no modelo, isto é, o número de casos de ocorrência do evento, $L(\mathbf{u}_i)$, e o risco, $R(\mathbf{u}_i)$, i = 1, ..., *N* [10]. Este procedimento é estabelecido da seguinte forma:

[10] Neste trabalho, considera-se o estimador de co-krigeagem não exato. Isto significa que o risco estimado $R(\mathbf{u}_i) \neq Z(\mathbf{u}_i)$, i =1, ..., *N*.

1) Inicialmente S_1 realizações de R($\mathbf{u}_i$), denotado por $\{ r_{s_1}(\mathbf{u}_i)$, i = 1, …, $N\}$, s_1 = 1, …, S_1, são estabelecidas por simulação de uma distribuição Gaussiana, com parâmetros de média e variância oriundos da co-krigeagem binomial. Caso $r_{s_1}(\mathbf{u}_i)$ seja negativo ou $r_{s_1}(\mathbf{u}_i) > 1$ simula-se novamente $r_{s_1}(\mathbf{u}_i)$;

2) Em seguida, para cada $r_{s_1}(\mathbf{u}_i)$, simulado obtém-se S_2 simulações da V.A. $L(\mathbf{u}_i)$, denotado por $\{ l_{s_2}(\mathbf{u}_i), i = 1, \ldots, N\}$, s_2 = 1, …, S_2, estabelecidas por sorteio aleatório de uma distribuição binomial com parâmetros $r_{s_1}(\mathbf{u}_i)$ e n($\mathbf{u}_i$);

3) Divide-se cada $l_{s_2}(\mathbf{u}_i)$ simulado por n($\mathbf{u}_i$) para se obter o conjunto de valores simulados da V.A. Z($\mathbf{u}_i$), denotado por $\{z_s(\mathbf{u}_i)$ i = 1, ..., $N\}$, s = 1, ..., $S_1.S_2$;

4) Para cada um dos conjuntos de valores simulados, $\{z_s(\mathbf{u}_i)$ i = 1, ..., $N\}$, s = 1, ..., $S_1.S_2$, o semivariograma do risco é estimado por $\hat{\gamma}^{*R}_{(\mathbf{h})}$. Então, para $S_1.S_2$ simulações constitui-se o conjunto $\{ \gamma^{*R}_{(\mathbf{h})_s}$, s = 1, ..., $S_1.S_2\}$.

O procedimento de simulação descrito acima é sintetizado através de um diagrama de blocos, conforme ilustrado na Figura 3.2, em que $S_1 = S_2 = S$.

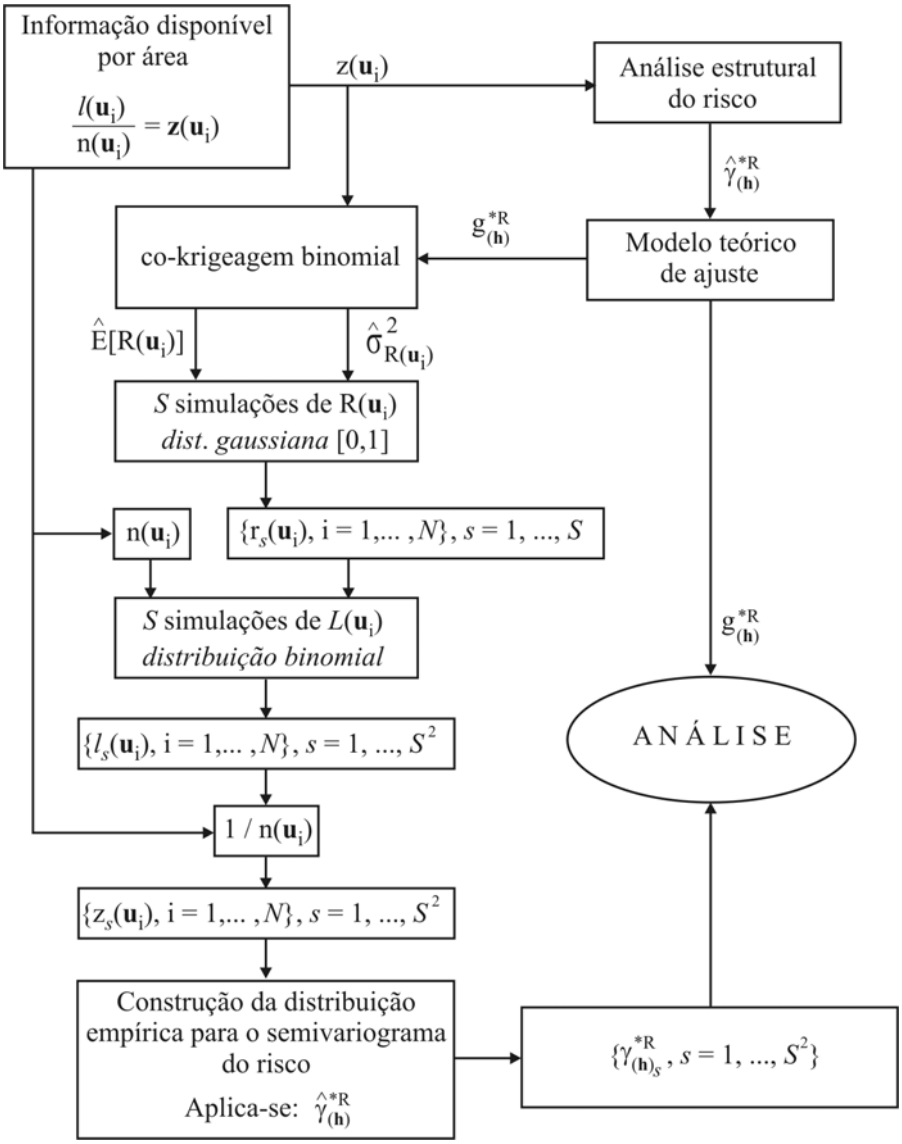


FIGURA 3.2 - Diagrama de simulação para geração da distribuição empírica para o semivariograma do risco.

O diagrama apresentado na Figura 3.2 e a descrição que segue pode ser também aplicado na avaliação de $g^{R}_{(\mathbf{h})}$ em relação à distribuição simulada dada por $\hat{\gamma}^{R}_{(\mathbf{h})_s}$.

Estabelecido a distribuição empírica para o semivariograma do risco, uma análise é conduzida para verificar o comportamento da estrutura de correlação espacial do risco, imposta por $g^{*R}_{(\mathbf{h})}$, em relação à distribuição simulada dada por $\hat{\gamma}^{*R}_{(\mathbf{h})_s}$. Esta análise

pode ser realizada visualmente através de um gráfico, apresentando os valores de $g^{*R}_{(\mathbf{h})}$ conjuntamente com os valores simulados $\gamma^{*R}_{(\mathbf{h})_s}$. É desejável que, para cada distância de análise (*lag*), os desvios produzidos entre $g^{*R}_{(\mathbf{h})}$ e a média dos valores simulados $\gamma^{*R}_{(\mathbf{h})_s}$, denotada por $\overline{\gamma}^{*R}_{(\mathbf{h})_s}$, sejam pequenos. Desta maneira, os desvios em cada *lag*, denotados por D_{nl}, *nl* = 1, ..., nlag (em que: nlag refere-se ao número de *lags*), são dados por:

$$D_{nl} = \left| \overline{\gamma}^{*R}_{(\mathbf{h})} - g^{*R}_{(\mathbf{h})} \right| \tag{3.26}$$

É desejável também que o erro quadrático médio, denotado por EQM, para todas as distâncias de análise, se situe perto de zero. Isto é calculado da seguinte forma:

$$EQM = \frac{\sum_{nl=1}^{nlag} D_{nl}^2}{nlag} \tag{3.27}$$

Os valores de D_{nl} e o valor de EQM, permitem avaliar o comportamento da estrutura de correlação espacial do risco, $g^{*R}_{(\mathbf{h})}$, em relação a $\overline{\gamma}^{*R}_{(\mathbf{h})_s}$.

Uma vez avaliado e aceito $g^{*R}_{(\mathbf{h})}$, superfícies do risco podem ser estimadas por co-krigeagem binomial, conforme Seção 3.7. Essas superfícies representam a média da distribuição do risco, isto é, $\hat{E}[R(\mathbf{u})]$, $\mathbf{u} \in A$. No entanto, existem outras situações, normalmente para fins de planejamento, em que se exige avaliar a probabilidade do campo aleatório em estudo exceder um dado valor de corte (em outras palavras, o que se pretende é avaliar quantos daqueles valores no espaço excedem simultaneamente um dado valor limite e qual a probabilidade de isso acontecer). Para satisfazer a essa necessidade, procedimentos de simulação seqüencial condicionada podem ser empregados, conforme descrito na seção seguinte.

3.9 Construção de cenários do risco por simulação seqüencial condicionada

Esta seção aborda apenas os principais conceitos básicos da simulação seqüencial condicionada e mostra como esta abordagem pode ser utilizada para a construção de cenários do campo aleatório R(**u**). Está fora do âmbito deste trabalho a análise aprofundada dos formalismos e dos métodos apontados, os quais podem ser consultados nas literaturas aqui referenciadas.

Considera-se que cenários são “mapas” (ou imagens) construídos a partir de valores de cortes obtidos da distribuição acumulada de R(**u**). A idéia de se empregar a simulação seqüencial condicionada como instrumento para construção de cenários é relativamente simples. Primeiro um conjunto de imagens equiprováveis ou igualmente representativas de R(**u**) é gerado por simulação. Tomando o conjunto de imagens simuladas, tem-se, para cada localização **u** da malha do mapa espacial, conforme ilustra a Figura 3.3, um conjunto de valores simulados r(**u**) da V.A. R(**u**). A partir deste conjunto, a função de distribuição acumulada de R(**u**) é estabelecida, possibilitando o cálculo de vários valores de cortes e, posteriormente, a construção de cenários. Por exemplo, valores de cortes obtidos de decis 0,1 e 0,9 da distribuição acumulada de R(**u**) podem representar, de acordo com o fenômeno em estudo, cenários otimistas e pessimistas, respectivamente.

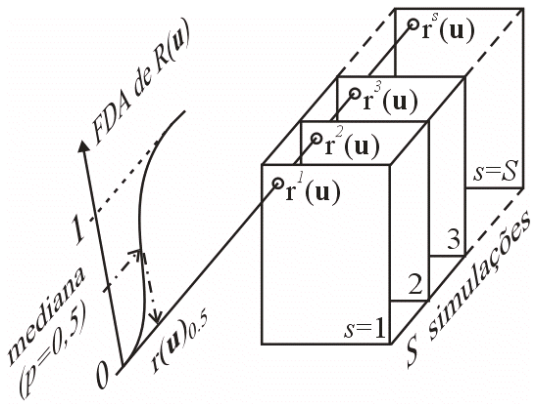


FIGURA 3.3 - Exemplo ilustrativo de um valor de corte obtido da mediana da função de distribuição acumulada de R(**u**).

Neste sentido, os procedimentos de simulações são instrumentos importantes, porque permitem explorar diferentes características do campo aleatório em estudo e podem

auxiliar o planejador que, orientado pelos seus objetivos, tem a possibilidade de escolher cenários mais adequados ao trabalho a ser executado.

Antes de apresentar a solução empregada neste trabalho para construção de cenários, faz-se necessário a introdução de alguns conceitos básicos que envolvem o método de simulação seqüencial condicionada.

Considere uma função conjunta de M V.A. e k amostras condicionantes iniciais:

$$F_{(M)}[(\mathbf{u}_1, ..., \mathbf{u}_m; r_1, ..., r_m) \mid (k)] = Prob\{R(\mathbf{u}_\alpha) \leq r(\mathbf{u}_\alpha), \alpha=1, ..., M \mid (k)\} \quad (3.28)$$

em que:

- $\{\mathbf{u}_\alpha, \alpha=1, ..., M\} \in A \subset \rfloor^2$, estão distribuídos regularmente no espaço, compondo uma estrutura de representação de grade regular (ou malha do mapa espacial);

Função de Distribuição Acumulada Condicionada (FDAC) multivariada expressa na Equação (3.28) é usada para modelar a incerteza conjunta dos M valores $r(\mathbf{u}_\alpha)$, $\alpha =1, ..., M$.

Para se obter um conjunto de estimativas, $\{r(\mathbf{u}_\alpha), \alpha=1, ..., M\}$, através do método de simulação seqüencial condicionada, os seguinte passos são realizados (Deutsch e Journel (1998)):

1) Define-se aleatoriamente uma localização $\mathbf{u}_\alpha$. Em seguida, simula-se um valor de $r(\mathbf{u}_\alpha)$ a partir da $F[\mathbf{u}_\alpha; r \mid (k)] = Prob\{R(\mathbf{u}_\alpha) \leq r(\mathbf{u}_\alpha) \mid (k)\}$, aplicando-se o método de transformação inversa;

2) Uma vez simulado, o valor de $r(\mathbf{u}_\alpha)$ é então considerado como dado condicionante para os subseqüentes passos de simulação, passando os dados condicionantes a $\{k+1\} = \{k\} \cup \{r(\mathbf{u}_\alpha)\}$;

3) Simulação de um novo valor $r(\mathbf{u}_{\alpha'})$, $\mathbf{u}_{\alpha'} \neq \mathbf{u}_\alpha$, a partir da $F[\mathbf{u}_{\alpha'}; r \mid (k+1)]$, com base nos $\{k+1\}$ valores condicionantes;

4) Atualização dos dados condicionantes para $\{k+2\} = \{k+1\} \cup \{r(\mathbf{u}_{\alpha})\}$;

5) Repetição dos dois passos anteriores até que todas as localizações $\mathbf{u}_{\alpha}$ da malha do mapa espacial tenham sido simuladas.

Observe que: i) o procedimento descrito acima requer a FDAC univariada, $F[\mathbf{u}_{\alpha}; r \mid (k)]$, para obter valores de $r(\mathbf{u}_{\alpha})$ da V.A. $R(\mathbf{u}_{\alpha})$, em cada localização $\mathbf{u}_{\alpha}$ a simular, condicionada às k amostras mais correlacionadas e aos valores pré-simulados, ambos dentro da vizinhança de $\mathbf{u}_{\alpha}$; ii) o conjunto resultante de valores simulados $\{r(\mathbf{u}_{\alpha}), \alpha=1, ..., M\}$ representam uma realização do campo aleatório $R(\mathbf{u}_{\alpha})$, ou simplesmente $R(\mathbf{u})$, que é o campo aleatório em estudo. iii) os passos de 1 a 5, devem ser repetidos várias vezes para a obtenção de um conjunto de imagens equiprováveis ou igualmente representativas do campo aleatório $R(\mathbf{u})$.

A geoestatística oferece basicamente duas abordagens de simulação seqüencial condicionada: i) a simulação seqüencial paramétrica; ii) a simulação seqüencial não-paramétrica. Do ponto de vista do método de simulação seqüencial condicionada, a diferença básica entre essas duas abordagens consiste na forma como a FDAC univariada, $F[\mathbf{u}_{\alpha}; r \mid (k)]$, é estimada.

Na simulação seqüencial condicionada paramétrica, a FDAC é estimada a priori por um conjunto limitado de parâmetros. Um exemplo típico é o modelo de distribuição Gaussiano que é totalmente determinado pelos valores da média e da variância da distribuição. Sob esta abordagem, a GSLIB (Deutsch e Journel, 1998) disponibiliza o método de Simulação Seqüencial Gaussiana (SSG). Neste caso, para cada localização $\mathbf{u}_{\alpha}$ a simular, a FDAC, $F[\mathbf{u}_{\alpha}; r \mid (k)]$, é estimada pela média e variância obtidas diretamente dos estimadores de krigeagem (neste trabalho, da co-krigeagem binomial). Trata-se de um modo bastante simples de estimar as FDAC, $F[\mathbf{u}_{\alpha}; r \mid (k)]$. No entanto, este método requer a hipótese de multi-normalidade dos dados, uma suposição de difícil verificação e extremamente forte, que pode não ser apropriada, para modelagem da V.A. $R(\mathbf{u})$. Mais detalhes desta abordagem podem ser vistos em Deutsch e Journel (1998) e Goovaerts (1997).

Na simulação seqüencial condicionada não-paramétrica, a FDAC não é estimada a priori e, portanto, não pode ser determinada por um conjunto limitado de parâmetros. Sob esta abordagem, a GSLIB (Deutsch e Journel, 1998) disponibiliza o método de Simulação Seqüencial por Indicação (SSI). Neste caso, para cada localização $\mathbf{u}_\alpha$ a simular, a FDAC, $F[\mathbf{u}_\alpha; r \mid (k)]$, é obtida por um conjunto de valores estimados que representam uma aproximação discretizada do modelo de distribuição. Isto é realizado empregando-se o formalismo por Indicação. Este formalismo está sintetizado no Apêndice F. Mais detalhes desta abordagem podem ser vistos em Felgueiras (1999), Deutsch e Journel (1998) e Goovaerts (1997).

Apresentadas as duas abordagens de simulação seqüencial condicionada, paramétrica e não-paramétrica, e os respectivos métodos SSG e SSI, optou-se por empregar o método SSI para a construção de cenários. A razão desta escolha é que este procedimento de simulação, diferente do método SSG, não impõe nenhum tipo de distribuição de probabilidade a priori para a V.A. R($\mathbf{u}$). Assim, cenários são construídos conforme descrito abaixo:

1) Estimação da V.A. R($\mathbf{u}$) por co-krigeagem binomial no centróide de cada área componente da região de estudo, denotada por R($\mathbf{u}_i$). Isto resulta um conjunto de estimativas, $\{r(\mathbf{u}_i), i = 1, ..., N\}$, que é a informação disponível para o método SSI;

2) Definição de valores de cortes, denotado por r_c, *c = 1, ..., número de cortes*, sobre o conjunto de dados $\{r(\mathbf{u}_i), i = 1, ..., N\}$. Por exemplo, para 3 valores de cortes poderia ser: r_1 = primeiro quartil, r_2 = segundo quartil e r_3 = terceiro quartil. *Esta etapa é um pré-requisito do método SSI;*

3) Para cada valor de corte r_c estabelecido, transforma-se a V.A. R($\mathbf{u}_i$) numa variável indicadora, denotada por $I(\mathbf{u}_i; r_c)$, do tipo:

$$I(\mathbf{u}_i; r_c) = \begin{cases} 1 \text{ se } R(\mathbf{u}_i) \le r_c \\ 0 \text{ se } R(\mathbf{u}_i) > r_c \end{cases}, \qquad i = 1, ..., N \tag{3.29}$$

Isto resulta um conjunto de valores $\{i(\mathbf{u}_i), i = 1, ..., N\}$ da V.A. $I(\mathbf{u}_i; r_c)$;

4) Para cada conjunto $\{i(\mathbf{u}_i), i = 1, ..., N\}$ da V.A. $I(\mathbf{u}_i; r_c)$, define-se um modelo teórico de semivariograma – *segundo pré-requisito do método SSI*;

5) Aplicação do método SSI para obtenção de um conjunto de imagens equiprováveis ou igualmente representativas do campo aleatório $R(\mathbf{u})$;

6) Construção da FDA de $R(\mathbf{u}_\alpha)$, em cada localização $\mathbf{u}_\alpha$, a partir do conjunto de imagens simuladas. Por fim, a construção de cenários a partir de valores de cortes obtidos da distribuição acumulada de $R(\mathbf{u}_\alpha)$.

O procedimento descrito acima é sintetizado através de um diagrama de blocos, conforme ilustra a Figura 3.4.

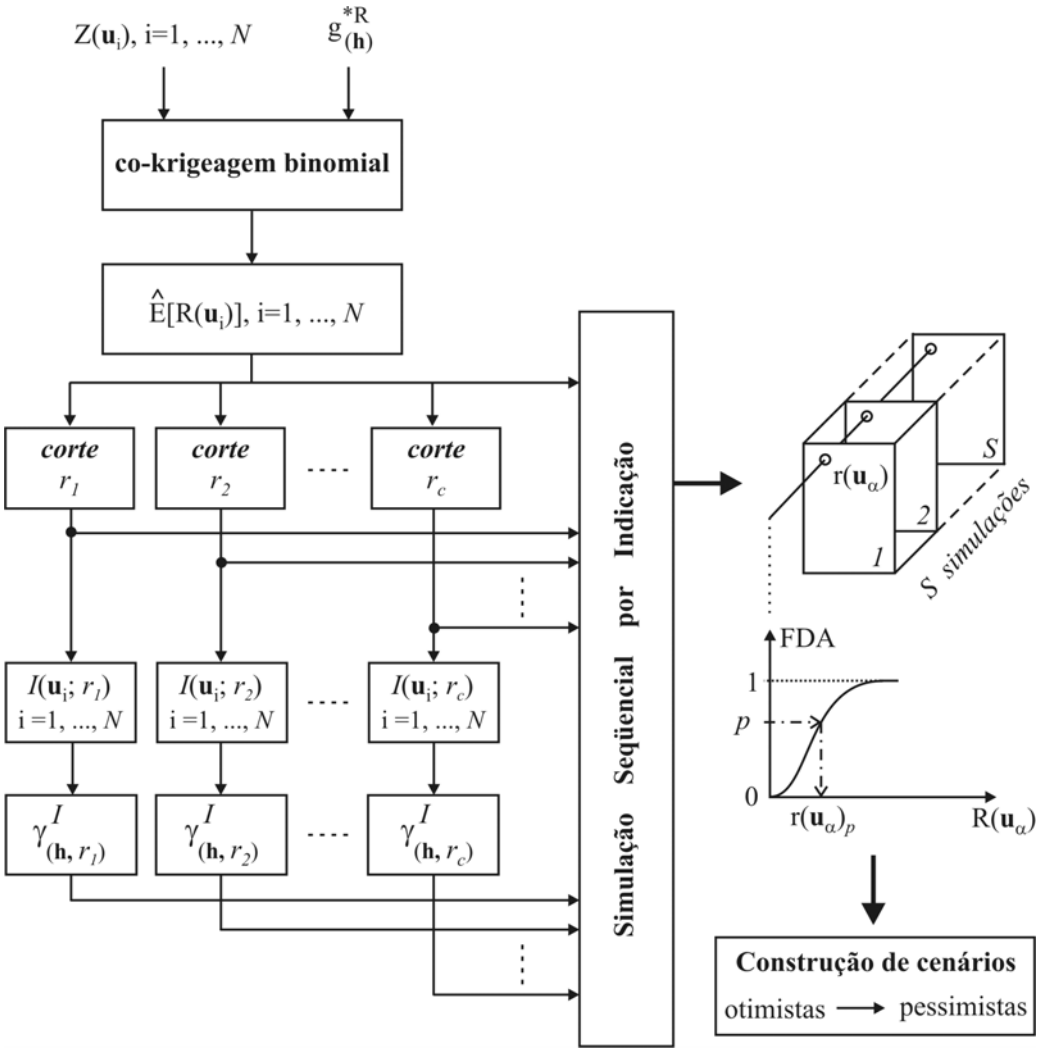


FIGURA 3.4 - Síntese do procedimento para construção de cenários por simulação seqüencial por indicação.

O capítulo seguinte apresenta uma aplicação do modelo geoestatístico binomial para a geração de cenários do risco de homicídio na cidade de São Paulo, no triênio 2002 - 2004.

CAPÍTULO 4

ESTUDO DE CASO: CENÁRIOS DO RISCO DE HOMICÍDIO NA CIDADE DE SÃO PAULO NO TRIÊNIO 2002 - 2004

4.1 Introdução

As mortes violentas no Brasil ao longo destas últimas décadas vêm assumindo proporções cada vez maiores, o que tem gerado um intenso debate nos mais variados setores da sociedade. Embora não seja um fenômeno exclusivo da sociedade brasileira, uma vez que atinge vários países com diferentes níveis de desenvolvimento, variando apenas de intensidade e padrão, muitos estudos têm se dedicado a este problema e apontam os homicídios como sendo a principal causa de mortes violentas dentro do capítulo das "causas externas" (CID-10), evidenciando ser este um problema de grande importância para a Saúde Pública, devido à sua magnitude e gravidade [Akerman e Bousquat (1999), Maia (2000), Gawryszewski e Jorge (2000), Caldeira (2000) e Beato et al. (2003)].

No caso específico da cidade de São Paulo, o aumento considerável da mortalidade por homicídios, ocorrida a partir da década de 80 e exacerbada a partir de 1996, coloca o município numa situação preocupante. Segundo os dados divulgados pelo Programa de Aprimoramento das Informações de Mortalidade (São Paulo. PROAIM, 2005), a evolução anual das taxas de homicídios na cidade de São Paulo nos últimos dez anos, de 1996 a 2005, sugere a distinção de três períodos: no primeiro, que compreende de 1996 a 1999, a taxa de homicídios apresenta um crescimento da ordem de 18,5%. Em 1996 foi registrada na capital paulista uma taxa média de 42 homicídios por cem mil habitantes, contra uma taxa média de 51,5 homicídios por cem mil habitantes em 1999. No segundo período, que vai de 1999 a 2001, a taxa média de homicídios se estabiliza num patamar expressivo, da ordem de 52 homicídios por cem mil habitantes. No terceiro que compreende de 2001 a 2005, a tendência ascendente é rompida, dando lugar a um movimento de queda a partir de 2001. No final de 2005 foi registrada uma

taxa média de 24 homicídios por cem mil habitantes, um decréscimo em torno de 54% em relação à taxa média constatada no segundo período. Esta evolução é ilustrada na Figura 4.1, conforme segue.

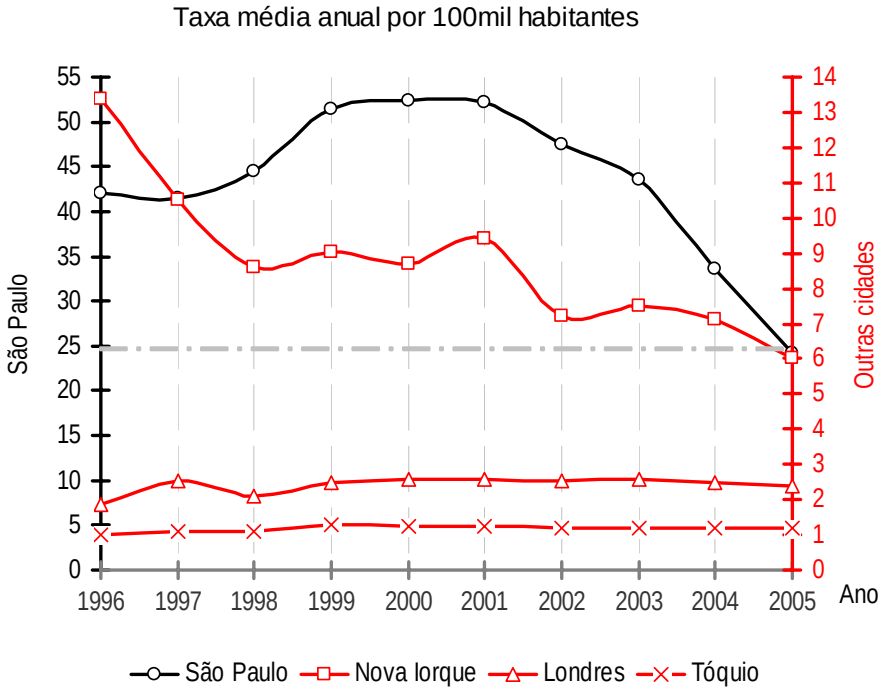


FIGURA 4.1 - Evolução da taxa média de homicídios na cidade de São Paulo de 1996 a 2005.

O retrato expresso na Figura 4.1, mostra que a magnitude dos índices de homicídios na cidade de São Paulo são alarmantes quando comparados aos de outras cidades de países desenvolvidos de mesma grandeza populacional. Por exemplo, tomando os valores da Figura 4.1, constata-se que a taxa média de homicídios no município de São Paulo para o período de 1996 a 2005 é da ordem de 43 mortes por cem mil habitantes, aproximadamente 5 vezes maior que a taxa média de homicídios observada no mesmo período para a cidade de Nova Iorque (8,74); 18 vezes maior que a de Londres (2,41); e 37 vezes maior que a de Tóquio (1,18).

Apesar do risco de homicídio na capital paulistana seguir um comportamento descendente nos últimos quatro anos, a sua distribuição pela cidade ocorre de modo desigual – *estimar "quanto" não basta; é preciso saber "onde" os homicídios ocorrem*. Neste sentido, ferramentas de análise que permitam produzir uma avaliação do risco de

homicídio e de sua distribuição espacial potencializam os meios de vigilância e, conseqüentemente, possibilitam fornecer informações importantes para o desenho de políticas de promoção da saúde considerando novas estratégias de controle e prevenção.

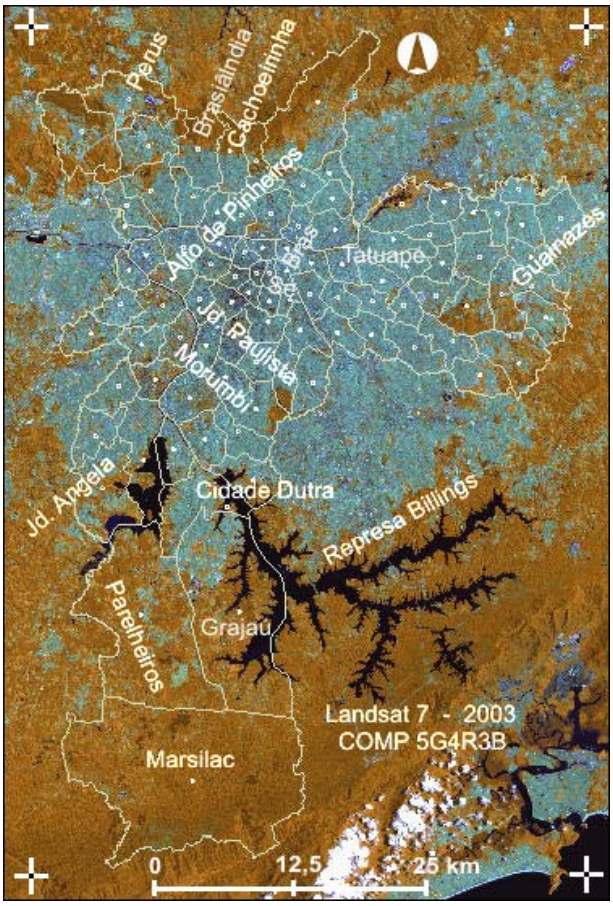
A reflexão brasileira sobre o impacto da violência na saúde, mesmo que ainda incipiente, avançou muito nos últimos 12 anos (Minayo, 1994). Os novos métodos que têm incorporado o espaço e o tempo como variáveis de análise, a partir da possibilidade de localização para os dados de homicídios, sistematizados através da classificação "causas externas" [Cruz (1996), Carvalho (1997), Assunção et al. (1998), Santos et al. (2001)] e vários outros trabalhos, modelos explicativo - determinantes Beato (1998), Lima et al. (2005), Nery e Monteiro (2006), vêm contribuindo de maneira consistente e firme para melhorar a fase de diagnóstico de situações para estabelecer o avanço necessário, em bases operativas, para o apoio aos serviços de saúde e ao planejamento das ações. Visando contribuir nesta direção, este capítulo emprega o modelo de risco geoestatístico binomial e tem sua aplicação avaliada na geração de cenários do risco de homicídio para a cidade de São Paulo no triênio 2002 – 2004.

O presente capítulo apresenta a seguinte organização: na Seção 4.2 apresenta-se a área de estudo e suas principais características; na Seção 4.3 descreve-se a fonte dos dados e como a informação é estabelecida; a Seção 4.4 fornece uma síntese do procedimento adotado para a elaboração de cenários do risco de homicídio na cidade de São Paulo, durante o período de 2002 a 2004; a Seção 4.5 estabelece uma análise preliminar dos dados, inicialmente através de estatísticas descritivas (sub-Seção 4.5.1), depois se verifica o problema da instabilidade na informação decorrente de áreas com pequenas populações (sub-seção 4.5.2), e a variação da tendência espacial dos dados (sub-Seção 4.5.3); na Seção 4.6 definem-se as zonas de risco de homicídio sobre a área de estudo; a Seção 4.7 dedica-se à análise da estrutura de correlação espacial do risco de homicídio; na sub-Seção 4.71 verifica-se o impacto da estrutura de correlação espacial, imposta pelo estimador do semivariograma do risco, sobre as estimativas do risco de homicídio; na sub-Seção 4.72 apresenta-se um estudo de simulação para verificar o comportamento do estimador proposto para o semivariograma do risco; na Seção 4.8 aplica-se o procedimento de co-krigegagem binomial para obtenção de estimativas do risco de

homicídio e na Seção 4.9 aplica-se o procedimento de simulação seqüencial por indicação para construção de cenários.

4.2 Área de estudo

A região de estudo refere-se à cidade de São Paulo, a qual ocupa uma área de apenas 0,02% do território Brasileiro. Segundo o censo demográfico de 2000 realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), o município apresenta alguns números que o colocam entre as maiores cidades do mundo, tais como: área total de 1.524 km^2, população de aproximadamente 10,5 milhões de habitantes, divisão territorial composta por 96 distritos, densidade populacional 6.885 habitantes/km^2, número de ruas e avenidas igual a 48.840, bairros e vilas perfazendo 1.544. Sua localização está compreendida sob as seguintes coordenadas geográficas: LONG de -46^o 50′ a -46^o 21′ e LAT de -23^o 21′ a -24^o 00′, conforme ilustra a Figura 4.2.



ÁREA DE ESTUDO

Retângulo Envolvente

Norte-Sul	78 km
Leste-Oeste	52 km

GEOMETRIA

Distâncias entre centróides

Mínima	1,15 km
Média	17,80 km
Máxima	59,78 km

FIGURA 4.2 - Mapa da cidade de São Paulo com destaque dos 96 distritos.

4.3 Os dados de homicídios

Os registros dos homicídios foram disponibilizados pelo PROAIM (São Paulo. PROAIM, 2005) no Município de São Paulo, que foi criado pela Prefeitura em 1989 com o objetivo de fornecer as informações de mortalidade necessárias ao diagnóstico de saúde, a vigilância epidemiológica e a avaliação dos serviços de saúde.

O programa é coordenado pela Secretaria Municipal da Saúde (SMS) e executado em conjunto com o Serviço Funerário do Município de São Paulo (SFMSP) e a Companhia de Processamento de Dados do Município de São Paulo (PRODAM). O SFMSP, autarquia responsável pelo encaminhamento do registro e sepultamento dos óbitos ocorridos na cidade de São Paulo, permite acesso oportuno às declarações destes óbitos. O PROAIM realiza o processamento, a análise e a divulgação das informações de mortalidade em nível municipal. As informações processadas consideram a localização de ocorrência do óbito a residência da vítima.

Por razões confidenciais, esses dados são agregados por distritos e disponibilizados na forma de contagens (número de ocorrências). Os dados de contagem são utilizados conjuntamente com dados de pessoas (número de habitantes por área), sendo possível calcular as taxas de óbitos por cem mil habitantes. Por fim, estas taxas foram associadas aos centróides dos distritos, constituindo a informação observada do fenômeno em estudo. Os dados utilizados estão disponibilizados nos Anexos B, C e D.

4.4 Fluxograma de trabalho para geração de cenários do risco de homicídio

Esta seção fornece uma síntese do procedimento adotado para a elaboração de cenários do risco de homicídio na cidade de São Paulo, durante o período de 2002 a 2004. A informação disponível são as taxas de homicídios agregadas aos 96 distritos que compõem o município. Para cada ano investigado: 1) uma análise preliminar é conduzida para detectar as principais características dos dados observados. Essa análise é realizada através de estatísticas descritivas, de gráficos que possibilitam verificar a instabilidade da informação decorrente de pequenas populações, e do emprego de técnicas de análise espacial que permitem detectar possíveis tendências presentes nos

dados; 2) aplicação do estimador proposto para o semivariograma do risco, conforme Seção 3.5, para a análise da dependência espacial do risco de homicídio; 3) definição do modelo de ajuste aplicado ao semivariograma do risco; 4) estimação do risco de homicídio nos centróides das áreas componentes por co-krigeagem binomial; 5) construção da distribuição empírica do semivariograma do risco por simulação; 6) análise para verificar o comportamento da estrutura de correlação espacial do risco em relação à distribuição simulada do semivariograma do risco; 7) estimação do risco de homicídio por co-krigeagem binomial; 8) aplicação do procedimento de simulação seqüencial por indicação para geração de realizações equiprováveis do risco de homicídio e 9) construção de cenários através de valores de cortes estabelecidos da distribuição acumulada do risco, possibilitando a criação desde cenários otimistas a pessimistas. Este procedimento é ilustrado, conforme a Figura 4.3.

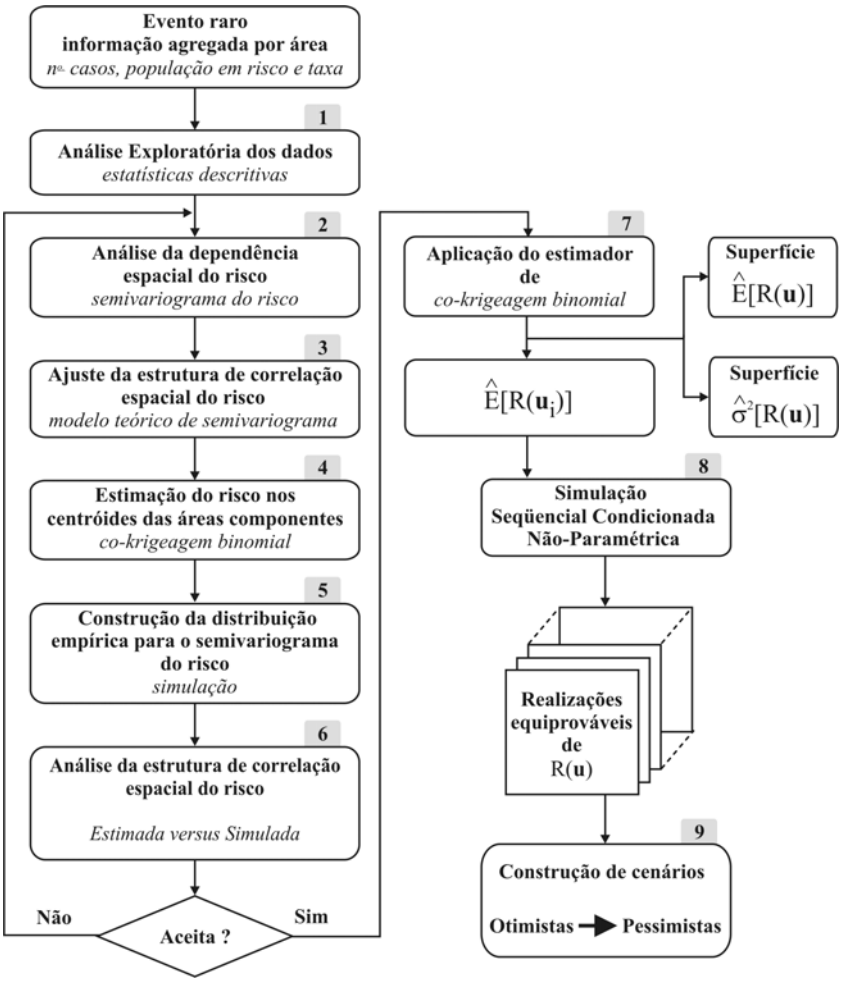


FIGURA 4.3 - Fluxograma de trabalho para geração de cenários do risco de homicídio.

Neste trabalho, parte dos algoritmos geoestatísticos necessários à modelagem do fenômeno investigado foram implementados no programa MATLAB-6.5 e utilizados conjuntamente com a biblioteca de programas geoestatísticos – GSLIB (Deutsch e Journel, 1998). Os resultados obtidos foram exportados para o Sistema de Processamento de Informações Georeferenciadas – SPRING (Câmara et al., 1996), que conta com bases cartográficas dos distritos da cidade de São Paulo, para visualização e produção de mapas e cenários do risco de homicídio. Seguindo a recomendação de vários estudos sobre esquemas de cores em mapas (Grauman et al., 2000) e (Brewer e Pickle, 2002), um esquema de cor *double-ended* foi utilizado: um gradiente de vermelho é usado para indicar áreas de alto risco, maior que a média, e um gradiente de azul é usado para indicar áreas de baixo risco, conforme ilustrado na Figura 4.4.

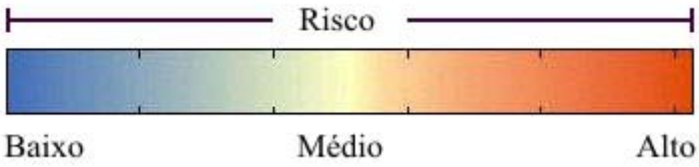


FIGURA 4.4 - Esquema de cor *double-ended.*

4.5 Análise preliminar dos dados

4.5.1 Estatísticas descritivas das taxas de homicídios

Uma análise preliminar através de estatísticas descritivas das taxas de homicídios, agregadas aos 96 distritos da cidade de São Paulo, mostra que a taxa média de homicídios por cem mil habitantes decresce no período 2002 – 2004, conforme sintetizado na Tabela 4.1. Em 2002 a taxa média de homicídios registrada na cidade de São Paulo foi da ordem de 41 homicídios por cem mil habitantes, contra uma taxa de 38 homicídios por cem mil habitantes em 2003 e 30 homicídios por cem mil habitantes em 2004, portanto um decréscimo da ordem de 27%. Observa-se também, para o mesmo período de investigação, um declínio nas variâncias e/ou desvios padrões. Em 2002 o coeficiente de variação foi da ordem de 58%, contra 54% em 2003 e 56% para 2004.

Isto pode ser um indicativo que aponta que as diferenças entre os distritos com valores extremos de taxas de homicídios se mantiveram estáveis de 2002 a 2004.

TABELA 4.1 - Estatísticas das taxas de homicídios no triênio 2002 - 2004.

Valores por 100 mil habitantes			
Estatística	**2002**	**2003**	**2004**
Tamanho da amostra	96	96	96
Média	41,09	37,98	30,25
Variância	585,34	432,78	289,73
Desvio padrão	24,19	20,80	17,02
Coeficiente de variação (%)	58%	54%	56%

Neste caso, tanto a média como o desvio padrão podem não ser medidas adequadas para representar o conjunto de valores, pois podem ser afetados por valores extremos. Para avaliar a forma da distribuição dos dados Tukey (1977) sugere três medidas, que são: i) o quartil inferior ou primeiro quartil, refere-se a um valor que deixa um quarto dos valores abaixo e três quartos acima dele; ii) a mediana ou segundo quartil, é um valor que deixa metade dos dados abaixo e metade acima dele; iii) o terceiro quartil ou quartil superior, é um valor que deixa três quartos dos dados abaixo e um quarto acima dele.

Diferente da média e do desvio padrão, essas três medidas são resistentes de posição de uma distribuição (Bussad e Morettin, 1995), porque em geral são pouco afetadas por mudanças de uma pequena porção dos dados. Essas medidas podem ser traduzidas graficamente num desenho esquemático, denominado *box plot*, conforme ilustrado nas Figuras 4.5(a), 4.5(b) e 4.5(c).

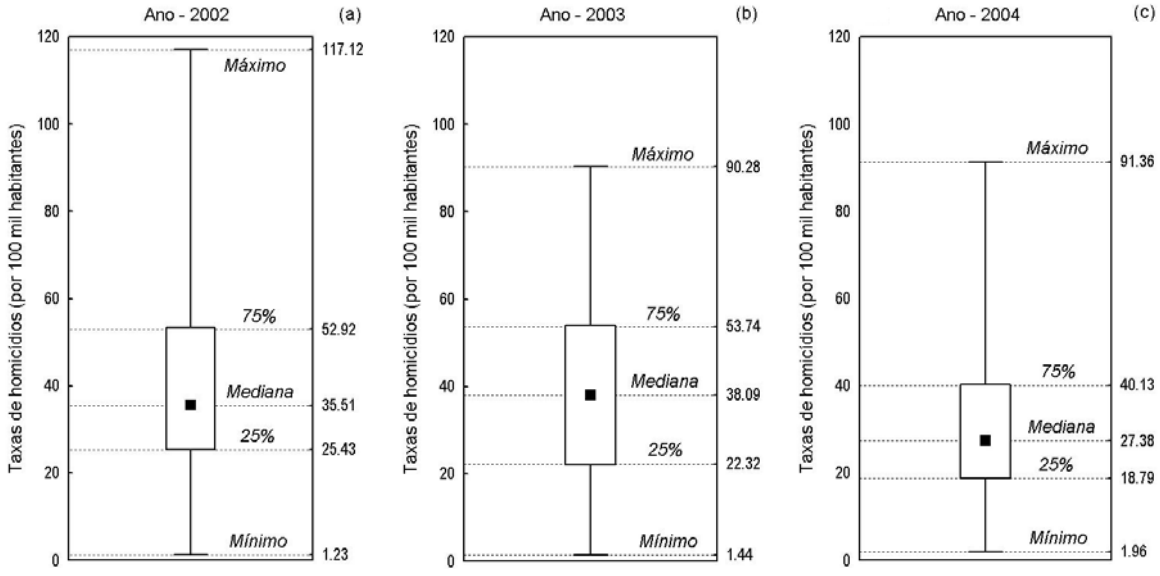


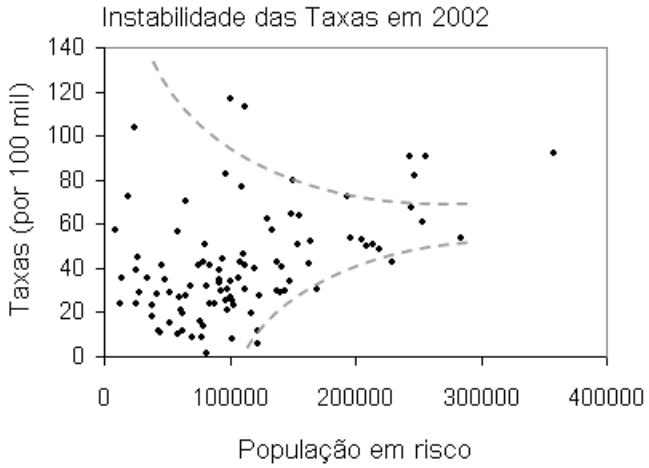
FIGURA 4.5 - *Box plot* das taxas de homicídios: (a) 2002, (b) 2003 e (c) 2004.

As Figuras 4.5(a), 4.5(b) e 4.5(c) dão uma idéia da posição, dispersão, assimetria e caudas das distribuições das taxas de homicídios em 2002, 2003 e 2004, respectivamente. A posição central dos valores é dada pela mediana, e a dispersão pela diferença entre o quartil superior e o quartil inferior. As posições relativas do quartil inferior, mediana e quartil superior dão uma noção da asssimetria da distribuição. Os comprimentos das caudas são dados pelas diferenças entre o quartil inferior e o valor mínimo, e entre o valor máximo e o quartil superior. Para cada ano investigado, os resultados obtidos mostram que as taxas de homicídios possuem distribuições assimétricas à esquerda, sendo mais acentuadas em 2002 e 2004 e menos em 2003. Para o ano de 2002, o coeficiente de assimetria calculado foi igual a 1, em 2003 igual a 0,37, e em 2004 igual a 0,8.

4.5.2 A instabilidade das taxas observadas

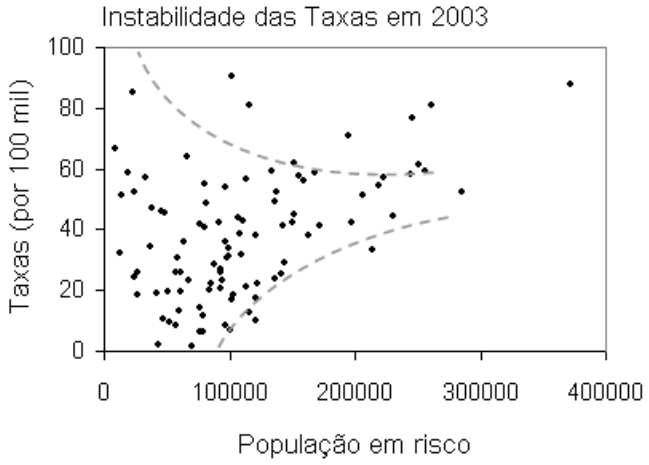
Para verificar a dependência da variância das taxas de homicídios ao tamanho da população, a análise é conduzida através de um gráfico que mostra os valores das taxas de homicídios versus o número de habitantes, para os distritos da cidade de São Paulo. Para cada ano investigado, os resultados obtidos mostram que quanto menor o tamanho

da população, maior a variabilidade da taxa observada. Conforme vários estudos (Assunção et al., 1998), (Beato et al., 1997) e outros, este comportamento é típico de taxas em geral. Deste modo, as flutuações extremas, isto é, os valores mais altos e mais baixos, ocorrem nas áreas de menores populações, sem ter associação com os riscos associados àquelas áreas. Isto tende a produzir um efeito, o qual já foi denominado de "efeito funil", e que está realçado por linhas pontilhadas nas Figuras 4.6, 4.7 e 4.8.



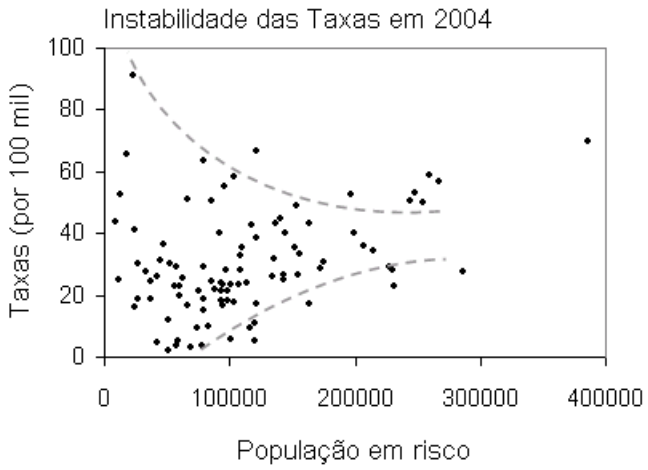
População	
Média	109.920,00
Desvio padrão	67.546,00
Taxa observada (por 100mil)	
Média	41,09
Desvio padrão	24,19

FIGURA 4.6 - Instabilidade das taxas de homicídios versus a população em risco em 2002.



População	
Média	110.582,00
Desvio padrão	69.019,00
Taxa observada (por 100mil)	
Média	37,98
Desvio padrão	20,80

FIGURA 4.7 - Instabilidade das taxas de homicídios versus a população em risco em 2003.



População	
Média	111.248,00
Desvio padrão	70.549,00

Taxa observada (por 100mil)	
Média	30,25
Desvio padrão	17,02

FIGURA 4.8 - Instabilidade das taxas de homicídios versus a população em risco em 2004.

4.5.3 A tendência espacial dos dados de homicídios

Uma forma de explorar a variação da tendência espacial dos dados é através do emprego do estimador de média móvel espacial [Câmara et al. (2004)]. A idéia básica é estimar para cada área componente da região de estudo um novo valor, que corresponde à média dos valores das áreas vizinhas. Isto tende a produzir resultados com menor flutuação que os dados originais, e possibilita uma nova visão do fenômeno investigado. O emprego deste estimador depende da definição de vizinhança adotada, que geralmente é estabelecida por uma matriz de proximidade espacial (também chamada matriz de vizinhança), a qual pode ser calculada a partir de vários critérios (Bailey e Gatrell (1995) e Câmara et al. (2004)]. Nesta análise o critério adotado para a construção da matriz de proximidade espacial foi o de vizinhança entre as áreas componentes.

Os resultados obtidos são descritos a seguir. Para 2002, a Figura 4.9(a) apresenta a variabilidade espacial das taxas de homicídios observadas, e a Figura 4.9(b) apresenta o resultado da média móvel espacial sobre os 96 distritos da cidade de São Paulo. Com a média local, há um alisamento: o valor mínimo é de 10,33 (homicídios por 100 mil habitantes) e o máximo é reduzido da ordem de 19,7%.

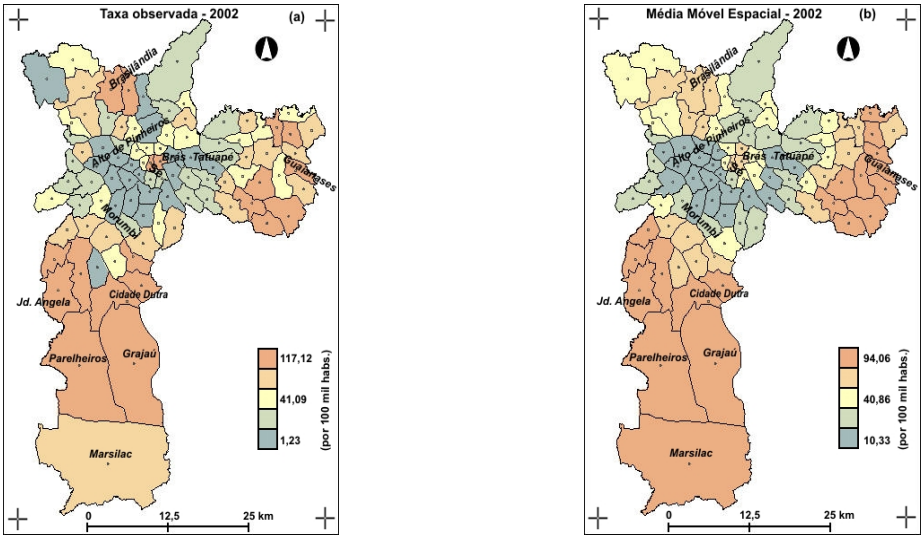


FIGURA 4.9 - Agrupamento estatístico por quintil em 2002: (a) taxas observadas de homicídios e (b) estimador de média móvel espacial.

Comparando os dois mapas da Figura 4.9, observa-se que a média móvel local fornece uma visão mais bem definida das grandes tendências do fenômeno em estudo, e no caso do risco de homicídio, mostra um forte gradiente centro-periferia.

De maneira análoga, para 2003, as Figuras 4.10(a) e 4.10(b) apresentam a variabilidade espacial das taxas de homicídios observadas e da média móvel espacial, respectivamente. Neste caso, o efeito de alisamento conduz o valor mínimo para 9,55 (homicídios por 100 mil habitantes) e o valor máximo é reduzido da ordem de 13%.

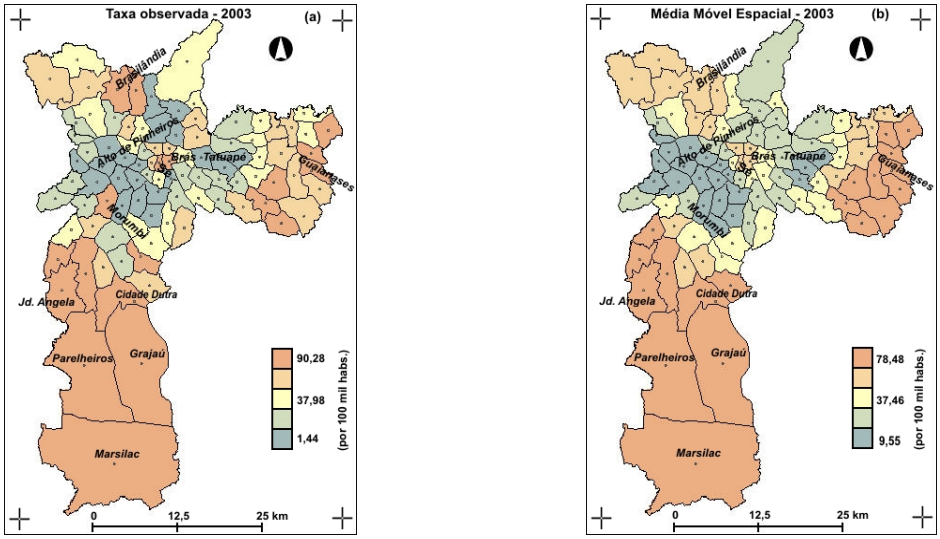


FIGURA 4.10 - Agrupamento estatístico por quintil em 2003: (a) taxas observadas de homicídios e (b) estimador de média móvel espacial.

Para 2004, a variabilidade espacial das taxas de homicídios observadas é conforme a Figura 4.11(a), e a média móvel espacial é conforme a Figura 4.11(b). Neste caso, o efeito de alisamento conduz o valor mínimo para 6,87 (homicídios por 100 mil habitantes) e o máximo é reduzido da ordem de 34%.

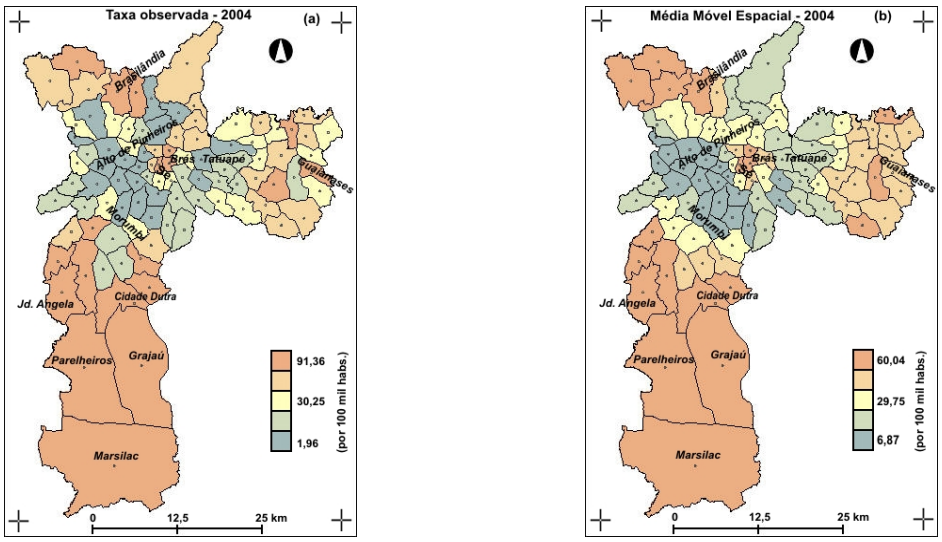


FIGURA 4.11 - Agrupamento estatístico por quintil em 2004: (a) taxas observadas de homicídios e (b) estimador de média móvel espacial.

Comparando as Figuras 4.10(a) versus 4.10(b) e 4.11(a) versus 4.11(b), primeiro observa-se novamente que os resultados oriundos da média móvel espacial fornecem uma visão mais bem definida das grandes tendências do fenômeno investigado e, segundo, que o risco de homicídio segue o mesmo padrão de variabilidade centro-periferia observado em 2002.

4.6 Definição das zonas de risco

Inicialmente cabe lembrar que a definição das zonas de risco faz-se necessária, uma vez que a configuração heterogênea da área de estudos apresenta médias e variâncias zonais distintas. Neste estudo de caso, para cada ano investigado, zonas de risco foram estabelecidas empiricamente a partir de um valor de corte da distribuição acumulada das taxas observadas, como sendo o percentil de 75%. Este valor de corte reflete a configuração de tendência observada nos dados, possibilitando a produção de dois

estratos de risco diferenciado, em que: 72 distritos em torno da região central foram classificados como de baixo risco e 24 distritos do centro e periferia de alto risco. A Tabela 4.2 sumariza as principais características dessas duas zonas de risco.

TABELA 4.2 - Sumário das zonas de risco em 2002, 2003 e 2004.

Classificação dos 96 distritos político-administrativos							
Zona de baixo risco (72 distritos)				Zona de alto risco (24 distritos)			
Taxas/100mil	2002	2003	2004	Taxas/100mil	2002	2003	2004
Mínima	1,2	1,4	1,9	Mínima	53,3	54,0	40,2
Média	29,8	28,8	22,4	Média	74,8	65,3	54,0
Máxima	52,2	52,5	40,0	Máxima	117,1	90,2	91,3
Desvio Padrão	12,5	11,2	9,6	Desvio Padrão	18,6	14,6	11,5

Em complemento à Tabela 4.2, um modo simples e ilustrativo de verificar a estratificação imposta aos 96 distritos é a visualização dos valores das taxas de homicídios na forma de mapa temático. Para isto, distritos que possuem valores de taxa menor ou igual ao valor de corte estabelecido são demarcados na cor azul, caso contrário, são demarcados na cor vermelha. Vale ressaltar que o uso de diferentes pontos de corte da variável induz à visualização de diferentes aspectos. Os resultados desta estratificação são ilustrados nas Figuras 4.12 (a, b, c), nas quais as zonas de alto risco são realçadas por elipses na cor vermelha, reforçando a tendência observada nos dados.

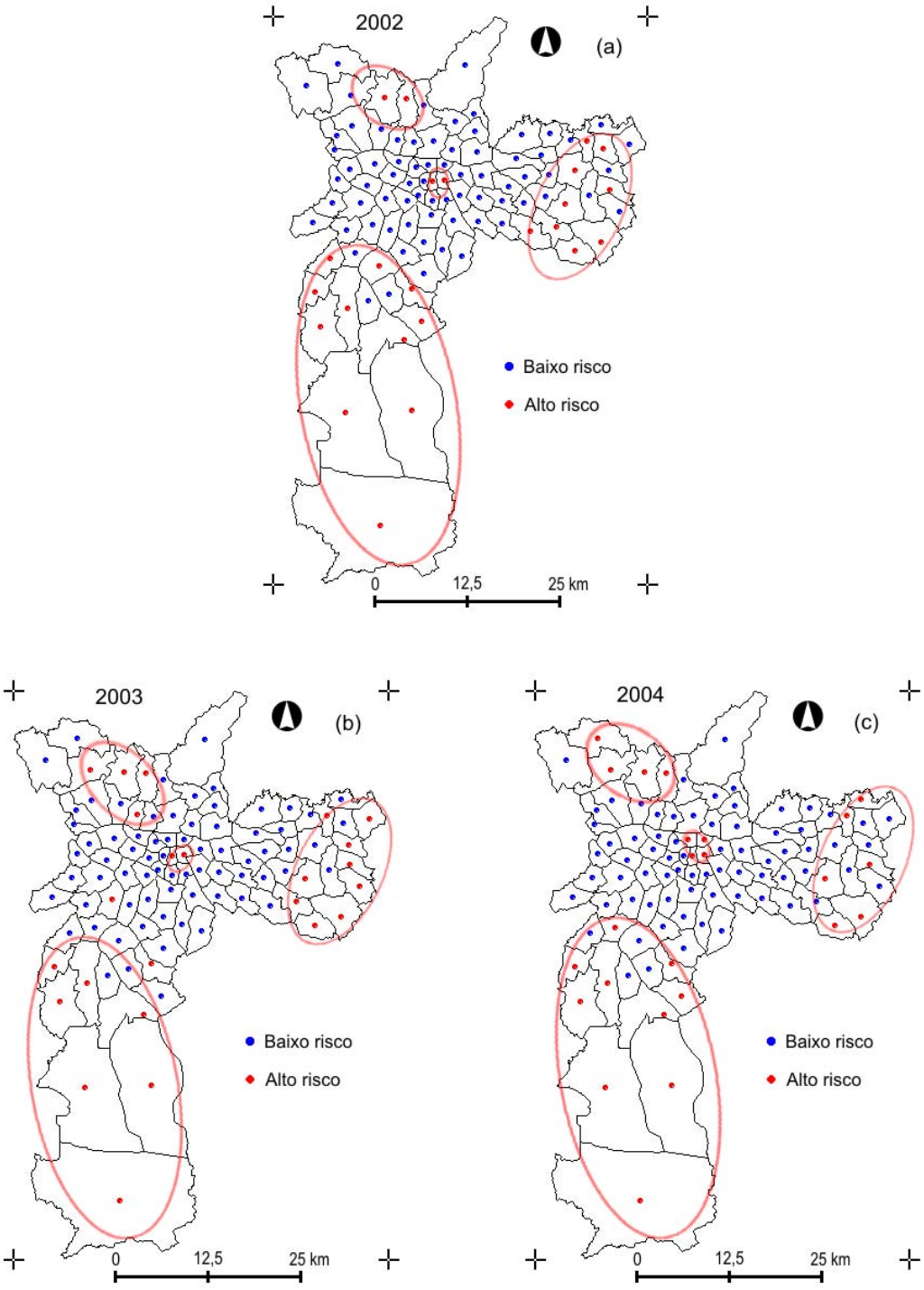


FIGURA 4.12 - Estratificação e realce das zonas de risco: (a) 2002, (b) 2003 e (c) 2004.

Estabelecidas as zonas de risco, a próxima etapa refere-se à análise da estrutura de correlação espacial do risco de homicídio, conforme descrito a seguir.

4.7 Análise da estrutura de correlação espacial do risco de homicídio

O estudo da estrutura de correlação espacial do risco de homicídio dada pela análise do estimador de semivariograma do risco não deve constituir, e não constitui normalmente, o objetivo final da análise espacial. Na realidade é necessário estimar valores do risco em locais não-amostrados. Desta forma, esta etapa do processo de modelagem do risco deve ser vista como um passo fundamental, mas não final, que precede o método de co-krigeagem .

Para verificar o comportamento do estimador proposto e a influência da estimação de seus parâmetros na estrutura de correlação espacial do risco de homicídio, esta análise é conduzida, para cada ano investigado (2002, 2003 e 2004), sob o enfoque de três alternativas:

1) a análise é conduzida utilizando o estimador para o semivariograma do risco conforme Oliver et al. (1998), $\hat{\gamma}^{R}_{(\mathbf{h})}$.

2) a análise é conduzida a partir do estimador proposto para o semivariograma do risco, $\hat{\gamma}^{*R}_{(\mathbf{h})}$, considerando apenas uma zona de risco (W = 1). Neste caso supõem-se os parâmetros de média e variância do risco constantes, os quais são estimados a partir dos dados observados;

3) a análise é conduzida também segundo o estimador proposto para o semivariograma do risco, mas conforme a estratificação estabelecida para a região de estudo em duas zonas de risco (W = 2). Sob esta alternativa, observou-se empiricamente, que até a uma certa distância d de análise ($d \cong 20$ km), a maior parte dos pares de pontos, com localizações nos centróides dos distritos, pertencem à zona de baixo risco. Assim, para distâncias menor ou igual a d, empregou-se no cálculo de $\hat{\gamma}^{*R}_{(\mathbf{h})}$, a média e a variância decorrentes da zona de baixo risco. Para distâncias maiores que d, observou-se grande influência de pontos que pertencem à zona de alto risco sobre

áreas de baixo risco. Neste caso, empregou-se a média e a variância decorrentes da zona de alto risco.

Para cada ano investigado, os resultados são apresentados conforme a ordem das alternativas expostas anteriormente. Assim, para o ano de 2002 a estrutura de correlação espacial do risco de homicídio ilustrada na Figura 4.13 (a) é decorrente do estimador segundo Oliver et al. (1998), na Figura 4.13 (b) segundo o estimador proposto com $W = 1$, e na Figura 4.13 (c) oriundo do estimador proposto com $W = 2$. Junto aos semivariogramas do risco estimados segue também seus respectivos modelos de ajuste, representados por uma linha contínua de cor vermelha, cujos parâmetros [efeito pepita (C_0), contribuição (C_1) e alcance (a)] estão sumarizados na Tabela 4.3.

Para o ano de 2003 a estrutura de correlação espacial do risco de homicídio apresentada na Figura 4.14 (a) é decorrente do estimador segundo Oliver et al. (1998), na Figura 4.14 (b) segundo o estimador proposto com $W = 1$, e na Figura 4.14 (c) oriundo do estimador proposto com $W = 2$. A Tabela 4.4 sintetiza os parâmetros para os modelos de ajuste em 2003.

Para o ano de 2004 a estrutura de correlação espacial do risco de homicídio mostrada na Figura 4.15 (a) é decorrente do estimador segundo Oliver et al. (1998), na Figura 4.15 (b) segundo o estimador proposto com $W = 1$, e na Figura 4.15 (c) oriundo do estimador proposto com $W = 2$. A Tabela 4.5 sintetiza os parâmetros para os modelos de ajuste em 2004.

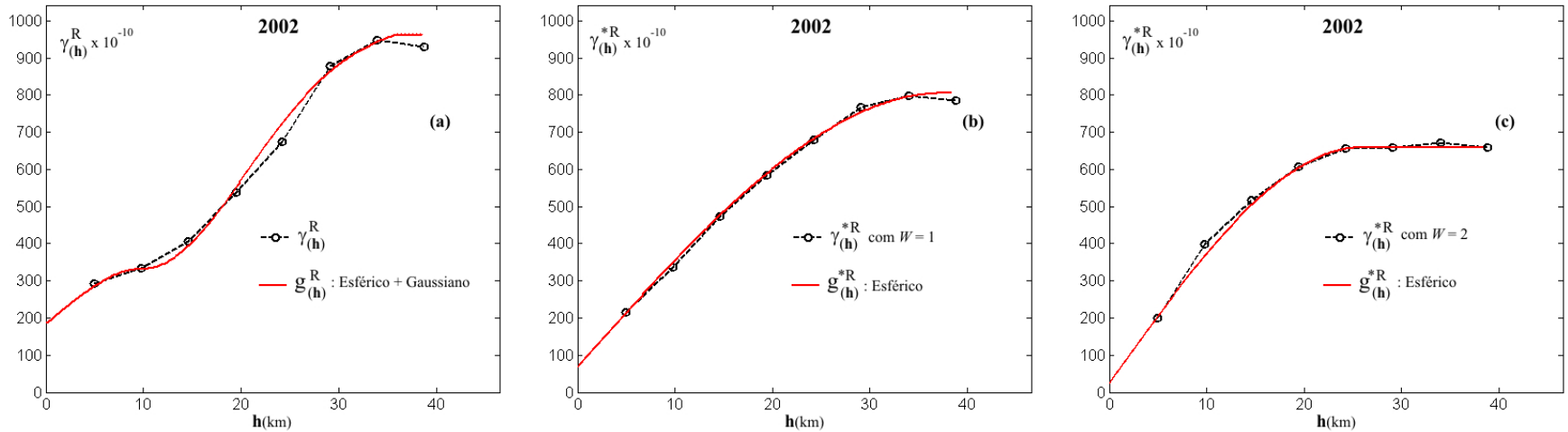


FIGURA 4.13 - Dependência espacial do risco de homicídio em 2002. Estimador: (a) Oliver et al. (1998); (b) proposto com $W = 1$ e (c) proposto com $W = 2$.

TABELA 4.3 - Modelos teóricos de semivariogramas do risco, em 2002.

Parâmetros		(a) Oliver et al. (1998)	(b) proposto com $W = 1$	(c) proposto com $W = 2$
efeito pepita (C_0)		187,30 x 10^{-10}	67,89 x 10^{-10}	23,41 x 10^{-10}
patamar (C)		997,70 x 10^{-10}	805,41 x 10^{-10}	659,09 x 10^{-10}
1ªestrutura:	tipo	esférico	esférico	esférico
	contribuição (C_1)	144,87 x 10^{-10}	737,52 x 10^{-10}	635,68 x 10^{-10}
	alcance (a_1)	9,63 km	37 km	26 km
2ªestrutura:	tipo	gaussiano	- x -	- x -
	contribuição (C_2)	665,53 x 10^{-10}	- x -	- x -
	alcance (a_2)	26,32 km	- x -	- x -

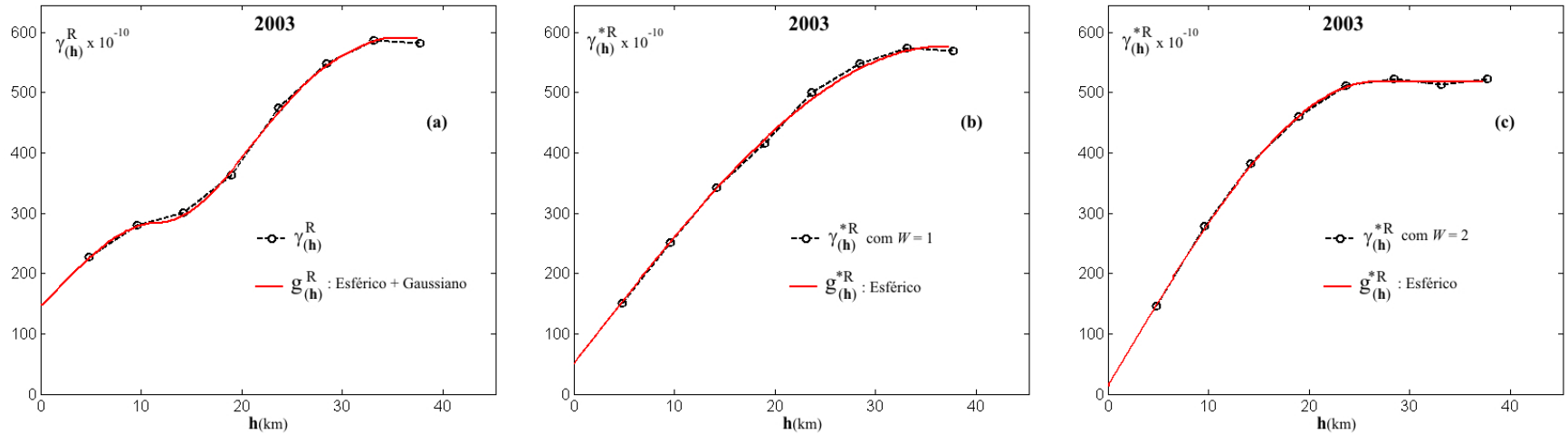


FIGURA 4.14 - Dependência espacial do risco de homicídio em 2003. Estimador: (a) Oliver et al. (1998); (b) proposto com $W = 1$ e (c) proposto com $W = 2$.

TABELA 4.4 - Modelos teóricos de semivariogramas do risco, em 2003.

Parâmetros		(a) Oliver et al. (1998)	(b) proposto com $W = 1$	(c) proposto com $W = 2$
efeito pepita (C_0)		147,14 x 10^{-10}	50,20 x 10^{-10}	12,98 x 10^{-10}
patamar (C)		607,39 x 10^{-10}	576,03 x 10^{-10}	518,68 x 10^{-10}
1ªestrutura:	tipo	esférico	esférico	esférico
	contribuição (C_1)	136,97 x 10^{-10}	525,83 x 10^{-10}	505,70 x 10^{-10}
	alcance (a_1)	11,4 km	36,5 km	26,5 km
2ª estrutura:	tipo	gaussiano	- x -	- x -
	contribuição (C_2)	323,28 x 10^{-10}	- x -	- x -
	alcance (a_2)	22,6 km	- x -	- x -

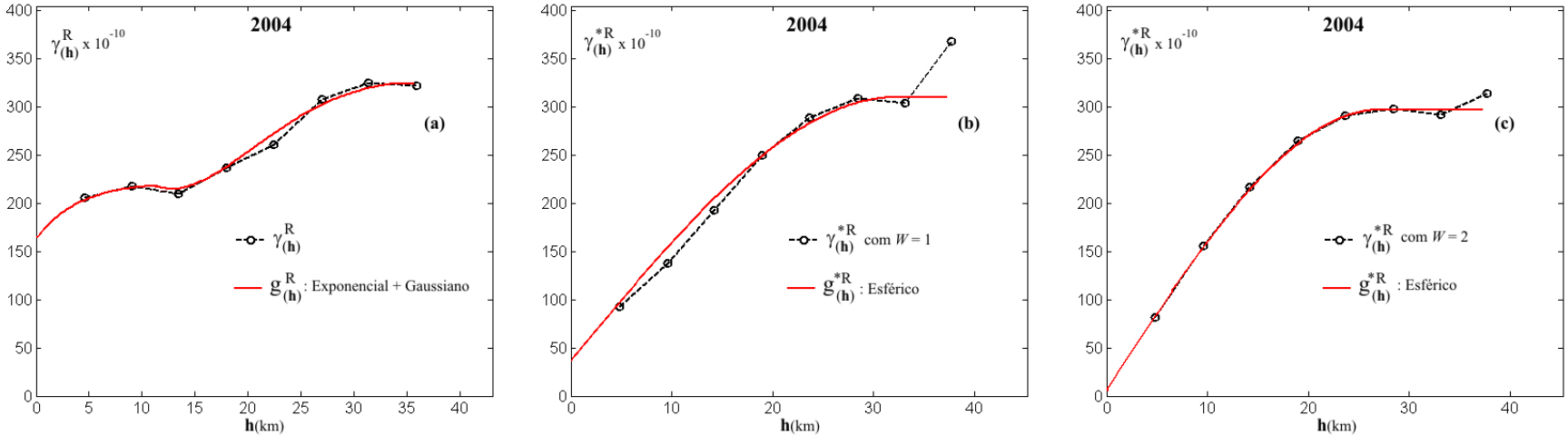


FIGURA 4.15 - Dependência espacial do risco de homicídio em 2004. Estimador: (a) Oliver et al. (1998); (b) proposto com $W = 1$ e (c) proposto com $W = 2$.

TABELA 4.5 - Modelos teóricos de semivariogramas do risco, em 2004.

Parâmetros		(a) Oliver et al. (1998)	(b) proposto com $W = 1$	(c) proposto com $W = 2$
efeito pepita (C_0)		165,14 x 10^{-10}	36,50 x 10^{-10}	5,79 x 10^{-10}
patamar (C)		338,97 x 10^{-10}	310,29 x 10^{-10}	296,97 x 10^{-10}
1ªestrutura:	tipo	exponencial	esférico	esférico
	contribuição (C_1)	56,49x 10^{-10}	273,79 x 10^{-10}	291,18 x 10^{-10}
	alcance (a_1)	11,5 km	36,5 km	26,9 km
2ªestrutura:	tipo	gaussiano	- x -	- x -
	contribuição (C_2)	117,34 x 10^{-10}	- x -	- x -
	alcance (a_2)	22,5 km	- x -	- x -

Para todo período investigado (de 2002 a 2004), observa-se que as estruturas de correlação espacial resultantes são distintas, conduzindo a diferentes interpretações sob a variabilidade espacial do risco de homicídio:

1) Sob a primeira alternativa, as estruturas de correlação espacial decorrentes do estimador de Oliver et al. (1998), $\hat{\gamma}^{R}_{(\mathbf{h})}$, são do tipo aninhada. Para 2002 e 2003 são estabelecidas por um modelo analítico tipo esférico, seguido de um modelo gaussiano, conforme ilustram as Figuras 4.13 (a), 4.14 (a), respectivamente. Para 2004 a estrutura aninhada é composta de um modelo analítico tipo exponencial, seguido de um modelo gaussiano, conforme ilustra a Figura e 4.15 (a). Essas estruturas revelam que o risco de homicídio, durante o período de investigação, pode estar auto-correlacionado numa distância, alcance (*a*), de até aproximadamente 40 km. Quando se observam as dimensões da área de estudo (largura = 52 km e altura = 78 km - vide Figura 4.2), distâncias da ordem de 30 km a 40 km são extremamente elevadas, porque apresentam uma dissimilaridade muito grande observada entre pares que compreendem a periferia da cidade com altas taxas de homicídios contra áreas que agregam baixas taxas de homicídios, gerando semivariogramas com valores elevados de patamar ($C = C_0 + C_1$); isto é um efeito de tendência. A literatura geoestatística sugere como distância máxima de análise a metade da menor dimensão (largura ou altura) que compõe a área de estudo. Isto nos leva a pensar que não há informação nesse dado para estimar os parâmetros de correlação para distâncias maiores que aproximadamente 25 km, ou porque o processo não é estacionário, ou porque o estimador é inadequado. Portanto, do ponto de vista da geoestatística convencional, essas estruturas não são apropriadas.

2) Sob a segunda alternativa, as estruturas de correlação espacial decorrente do estimador proposto $\hat{\gamma}^{*R}_{(\mathbf{h})}$ com $W = 1$ são completamente diferentes das anteriores. Este fato se deve principalmente à correção imposta no estimador do semivariograma empírico (ponderado pela população). Neste caso, as estruturas de correlações espaciais do risco de homicídio são estabelecidas por um único modelo analítico do tipo esférico, conforme ilustradas nas Figuras 4.13 (b), 4.14 (b) e 4.15 (b), respectivamente. Essas estruturas revelam que o risco de homicídio durante o período de investigação segue

auto-correlacionado numa distância de até aproximadamente 35 km. Conforme já discutido anteriormente, distâncias dessa ordem de grandeza são elevadas para a dimensão da região de estudo. Isto nos leva a concluir que sob esta alternativa as correções sugeridas para os parâmetros do estimador do semivariograma do risco não produzem um modelo compatível com a estrutura de correlação espacial imposta pela geometria dos dados.

3) Sob a terceira alternativa, as estruturas de correlação espacial decorrentes do estimador proposto $\hat{\gamma}^{*R}_{(\mathbf{h})}$ com $W = 2$, são estabelecidas por um único modelo analítico do tipo esférico, conforme ilustradas nas Figuras 4.13 (c), 4.14 (c) e 4.15 (c), respectivamente. Essas estruturas revelam que o risco de homicídios pode estar auto-correlacionado numa distância de até aproximadamente 27 km. Sob esta alternativa, os resultados obtidos são mais condizentes com as dimensões da área de estudo, com a geometria de amostragem, e mais coerentes sob a ótica da geoestatística convencional. Além disso, observa-se também uma redução acentuada do efeito pepita (C_0). A proporção deste valor para o patamar do semivariograma ($C = C_0 + C_1$), é um indicativo da quantidade de variação ao acaso de um ponto para outro, e quanto menor seu valor, mais forte é a dependência espacial do risco de homicídio. A Tabela 4.6 sintetiza as percentagens das proporções, dadas por $[C_0 / (C_0 + C_1)] * 100$, para cada uma das alternativas apresentadas.

TABELA 4.6 - Síntese das proporções obtidas sob cada alternativa.

		Proporção $[C_0 / (C_0 + C_1)] * 100$		
Alternativa		**2002**	**2003**	**2004**
1	$\hat{\gamma}^{R}_{(\mathbf{h})}$	18,77 %	24,22 %	48,72 %
2	$\hat{\gamma}^{*R}_{(\mathbf{h})}$ com $W = 1$	8,43 %	8,71 %	11,76 %
3	$\hat{\gamma}^{*R}_{(\mathbf{h})}$ com $W = 2$	3,55 %	2,50 %	1,94 %

A próxima etapa de análise objetiva verificar o impacto de cada uma dessas estruturas espaciais nas estimativas do risco de homicídio, conforme descrita a seguir.

4.7.1 Impacto da estrutura de correlação espacial nas estimativas do risco de homicídio

Na geoestatística as estimativas obtidas através dos métodos de krigeagem e/ou co-krigeagem dependem substancialmente da estrutura de correlação espacial imposta pelo semivariograma. Para isto, uma análise das estimativas do risco de homicídio, nos centróides de cada distrito, é conduzida a partir das estruturas de correlações impostas por $\hat{\gamma}^{R}_{(\mathbf{h})}$, $\hat{\gamma}^{*R}_{(\mathbf{h})}$ com $W = 1$ e $\hat{\gamma}^{*R}_{(\mathbf{h})}$ com $W = 2$.

Para cada ano investigado, os resultados obtidos revelam que as estimativas do risco de homicídio nos centróides são aproximadamente iguais. Entretanto, as variâncias das estimativas são significativamente menores quando $\hat{\gamma}^{*R}_{(\mathbf{h})}$ com $W = 2$ é utilizado. Isto pode ser constatado através das Figuras 4.16, 4.17 e 4.18, conforme segue.

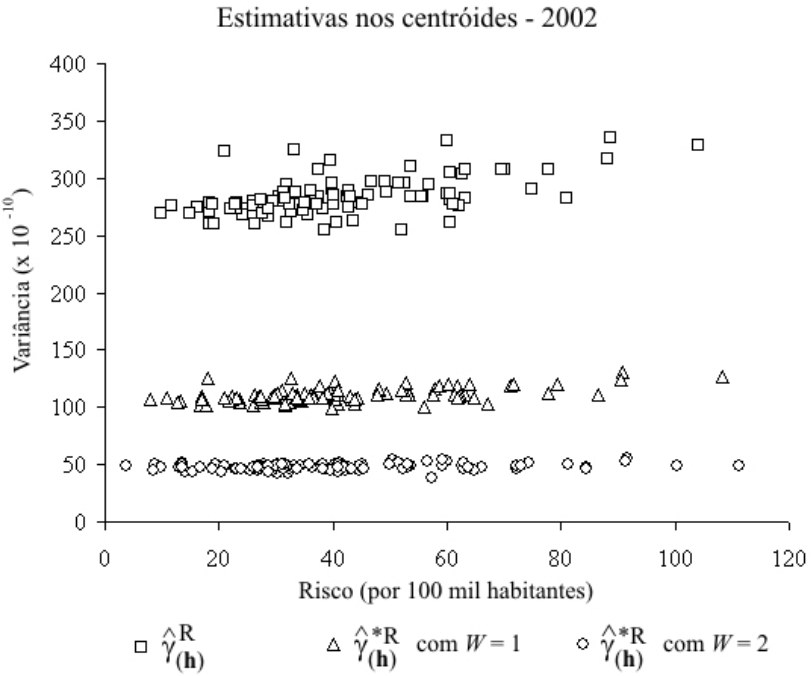


FIGURA 4.16 - Estimativas do risco de homicídio e variâncias das estimativas nos centróides dos distritos, em 2002.

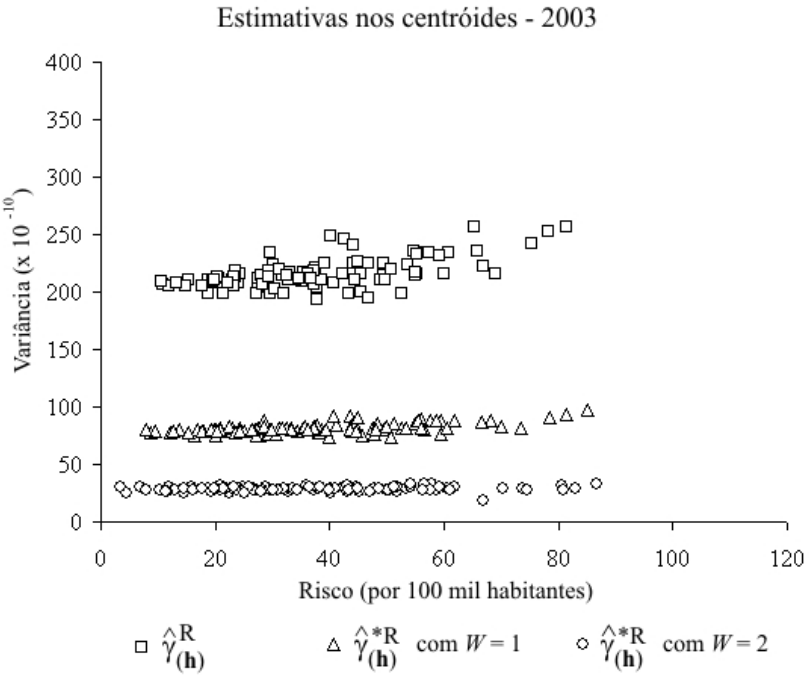


FIGURA 4.17 - Estimativas do risco de homicídio e variâncias das estimativas nos centróides dos distritos, em 2003.

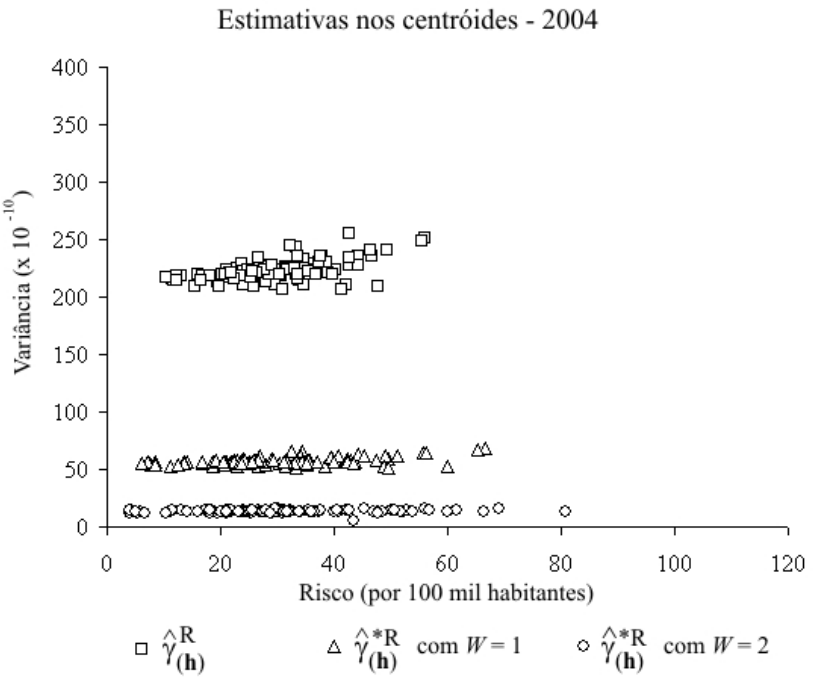


FIGURA 4.18 - Estimativas do risco de homicídio e variâncias das estimativas nos centróides dos distritos, em 2004.

Prosseguindo, a Tabela 4.7 apresenta uma síntese das médias das estimativas obtidas nos 96 centróides dos distritos que compõem a cidade de São Paulo, em que: MRE refere-se à média dos riscos estimados e MVE à média das variâncias estimadas.

TABELA 4.7 - Síntese das médias das estimativas nos 96 centróides, valores expressos por 100 mil habitantes.

		2002		2003		2004		MVE
Alternativa		MRE	MVE	MRE	MVE	MRE	MVE	**Acumulado**
1	$\hat{\gamma}^{R}_{(\mathbf{h})}$	41,07	283,95	37,83	216,09	29,71	222,79	722,83
2	$\hat{\gamma}^{*R}_{(\mathbf{h})}$ com $W = 1$	41,16	110,59	38,01	81,62	30,24	57,38	249,59
3	$\hat{\gamma}^{*R}_{(\mathbf{h})}$ com $W = 2$	41,10	47,92	37,96	28,52	30,20	13,48	89,92

Como pode ser observado, para cada ano investigado e independente da alternativa empregada, as médias das estimativas do risco possuem valores bem próximos. Entretanto, o mesmo não é verdade com relação aos valores das médias das variâncias estimadas. Tomando as médias acumuladas das variâncias estimadas, no período investigado (de 2002 a 2004), verifica-se que a terceira alternativa é a que produz menor valor, uma redução da ordem de 64% em relação à segunda alternativa e próxima de 87% em relação à primeira.

Os resultados desta análise evidenciam a importância da estimação dos parâmetros para o estimador do semivariograma do risco e mostram que, com a estratificação da região de estudo em duas zonas de risco ($W = 2$), o procedimento de co-krigeagem binomial se tornou mais eficiente.

A próxima etapa de análise objetiva verificar o comportamento da estrutura de correlação espacial do risco com relação à sua distribuição simulada, conforme descrita a seguir.

4.7.2 Simulação da distribuição do semivariograma do risco

Dando continuidade à análise da estrutura de correlação espacial do risco de homicídio, uma vez estimado o risco e a variância da estimativa, nos centróides dos 96 distritos político-admnistrativos, um estudo de simulação é conduzido para a construção da distribuição empírica do semivariograma do risco. Este procedimento é realizado conforme descrito anteriormente na Seção 3.8 e objetiva verificar o comportamento da estrutura de correlação espacial do risco estimada em relação à sua distribuição simulada. Conforme já mencionado, é desejável que os desvios produzidos, entre a estrutura de correlação espacial do risco estimada e a média da distribuição empírica do semivariograma do risco, sejam muito pequenos e que sua média se situe perto de zero.

Para cada ano investigado, foram realizadas 100 simulações sob o enfoque de três alternativas:

1) $g^{R}_{(\mathbf{h})}$ versus $\gamma^{R}_{(\mathbf{h})_S}$;

2) $g^{*R}_{(\mathbf{h})}$ versus $\gamma^{*R}_{(\mathbf{h})_S}$ com $W = 1$;

3) $g^{*R}_{(\mathbf{h})}$ versus $\gamma^{*R}_{(\mathbf{h})_S}$ com $W = 2$.

Para o ano de 2002, o resultado decorrente da primeira alternativa, Figura 4.19(a), revela que a estrutura de correlação espacial do risco de homicídio $g^{R}_{(\mathbf{h})}$ apresenta desvios acentuados em relação à média da distribuição empírica dada por $\hat{\gamma}^{R}_{(\mathbf{h})_S}$. Sob a segunda alternativa, Figura 4.19(b), verifica-se que $g^{*R}_{(\mathbf{h})}$ apresenta desvios moderados em relação à média da distribuição empírica dada por $\hat{\gamma}^{*R}_{(\mathbf{h})_S}$ com $W = 1$. O resultado decorrente da terceira alternativa, Figura 4.19(c), mostra que os desvios de $g^{*R}_{(\mathbf{h})}$ em relação à média da distribuição empírica dada por $\hat{\gamma}^{*R}_{(\mathbf{h})_S}$ com $W = 2$, são menores.

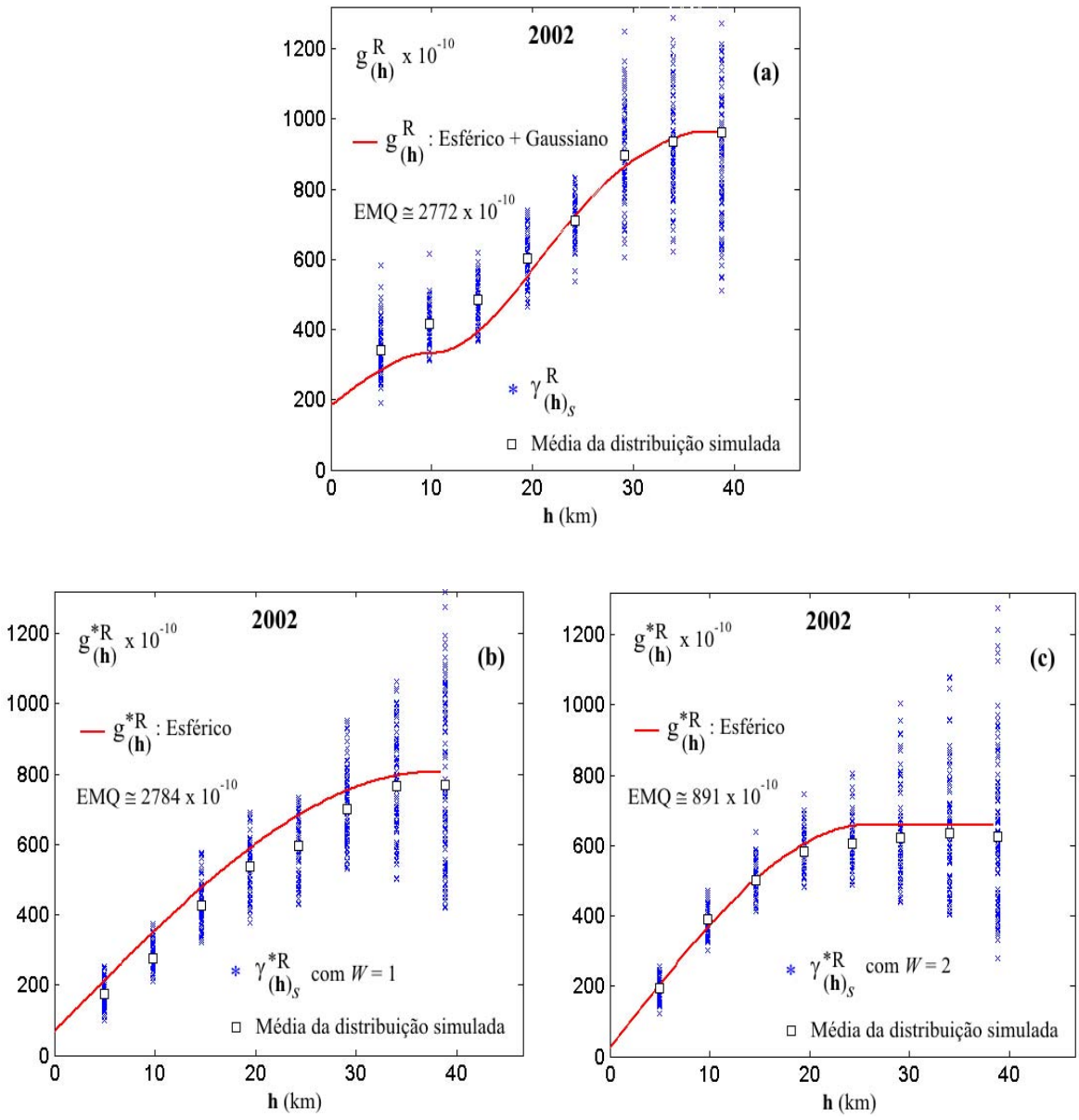


FIGURA 4.19 - Comportamento da estrutura de correlação espacial do risco de homicídio em relação à distribuição simulada do semivariograma do risco, para 2002: (a) segundo Oliver et al. (1998), (b) oriundo do estimador proposto com $W = 1$ e (c) conforme estimador proposto com $W = 2$.

Prosseguindo, para o ano de 2003, as estruturas de correlação espacial do risco de homicídio em relação às suas respectivas distribuições simuladas seguem o mesmo comportamento observado em 2002. Os resultados obtidos são apresentados nas Figuras 4.20(a), 4.20(b) e 4.20(c), conforme segue.

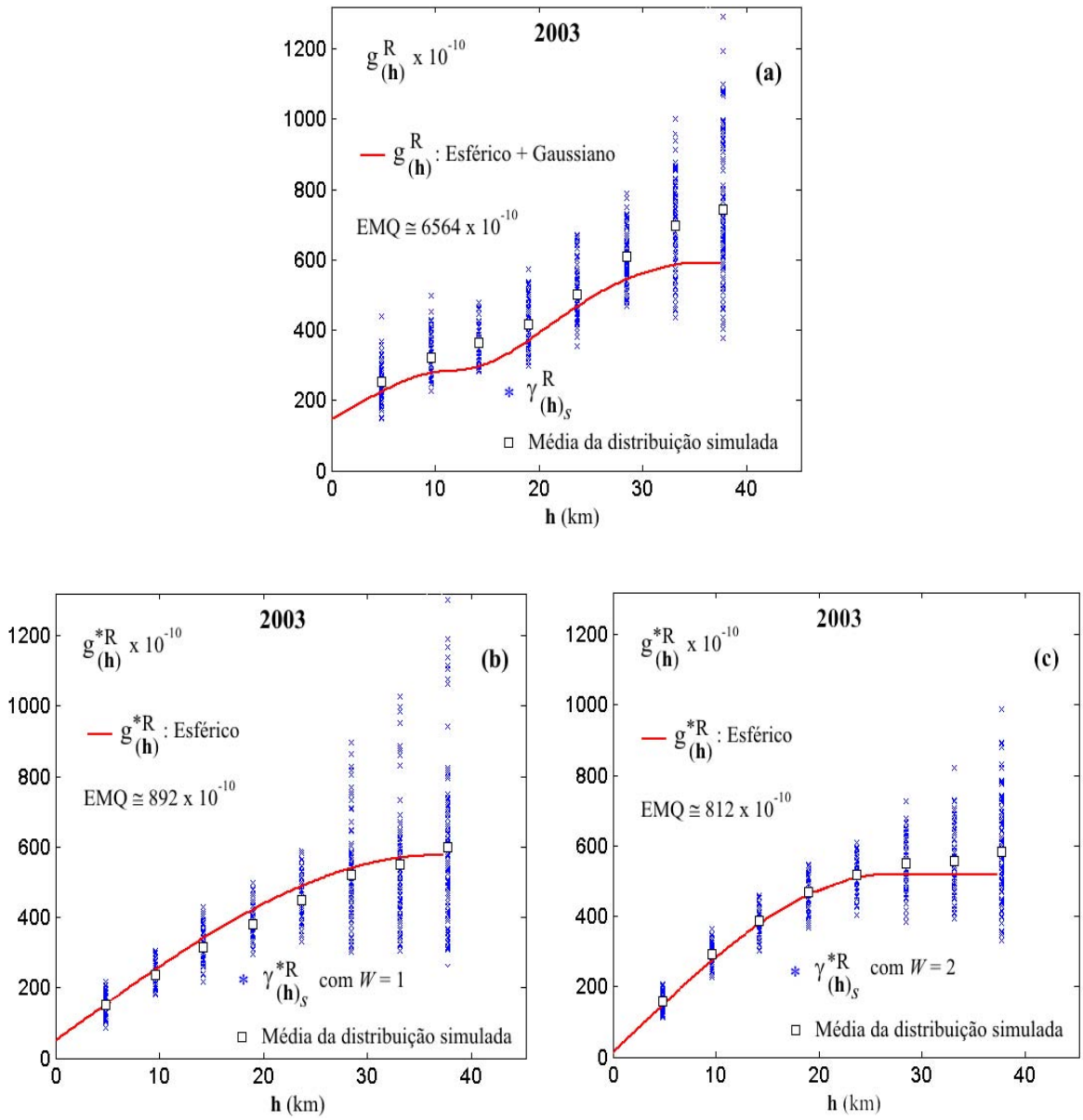


FIGURA 4.20 - Comportamento da estrutura de correlação espacial do risco de homicídio em relação à distribuição simulada do semivariograma do risco, para 2003: (a) segundo Oliver et al. (1998), (b) oriundo do estimador proposto com $W = 1$ e (c) conforme estimador proposto com $W = 2$.

Para o ano de 2004, os resultados obtidos são ilustrados conforme as Figuras 4.21(a), 4.21(b) e 4.21(c), apresentando o mesmo comportamento observado em 2002 e 2003.

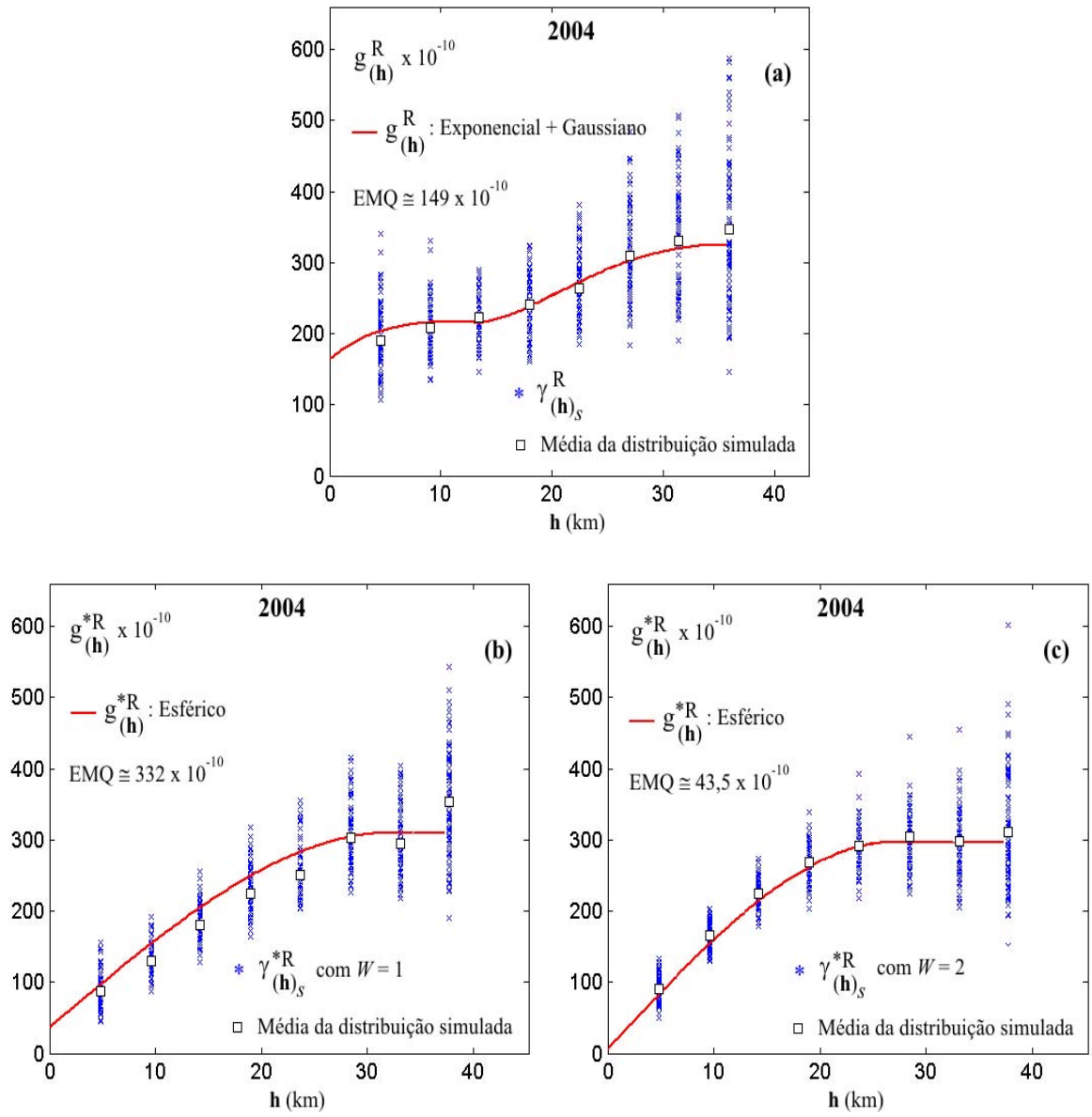


FIGURA 4.21 - Comportamento da estrutura de correlação espacial do risco de homicídio em relação à distribuição simulada do semivariograma do risco, para 2004: (a) segundo Oliver et al. (1998), (b) oriundo do estimador proposto com $W = 1$ e (c) conforme estimador proposto com $W = 2$.

Finalizando, os resultados obtidos da análise da estrutura de correlação espacial do risco de homicídio apontam que o estimador proposto com $W = 2$ pode ser mais adequado à análise do fenômeno investigado, visto que: 1) proporciona estrutura de correlação espacial com menor valor de efeito pepita e distância de dependência espacial mais condizente com as dimensões da área de estudo; 2) torna o procedimento de

co-krigeagem binomial mais eficiente; 3) os desvios produzidos, entre a estrutura de correlação espacial do risco estimada e a média da distribuição empírica do semivariograma do risco, são menores.

4.8 Estimação da superfície do risco de homicídio

Após a análise e seleção do modelo de estrutura de correlação espacial, estimativas do risco de homicídio foram estabelecidas por co-krigeagem binomial. Optou-se por calculá-las em várias localizações dispostas numa grade regular, com resolução de 150 metros na direção Norte-Sul e 98 metros na direção Leste-Oeste. A escolha desta resolução foi apenas para manter a mesma relação de aspecto da área de estudo (Norte-Sul 78 km / Leste-Oeste 52 km) e proporcionar grades densas de valores (520 linhas por 530 colunas).

As superfícies geradas representam a evolução média da distribuição do risco de homicídio, para o período investigado, na cidade de São Paulo. Os resultados podem ser vistos[11] nas Figura 4.22 (a), 4.22 (b) e 4.22 (c), conforme segue.

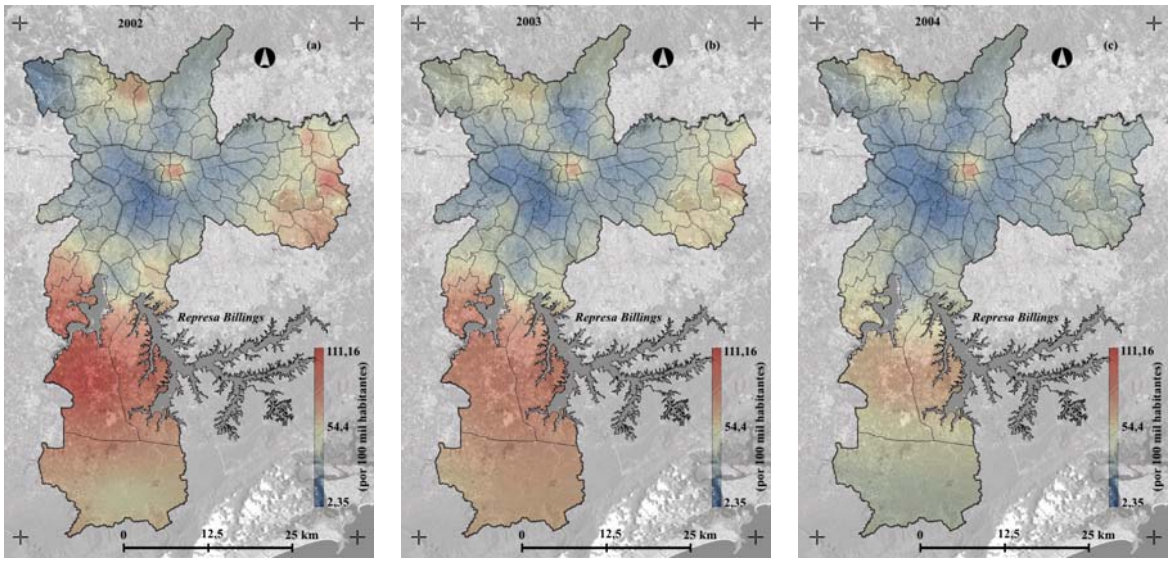


FIGURA 4.22 - Evolução da média da distribuição do risco de homicídio, com detalhe da mancha urbana na imagem de fundo: (a) 2002, (b) 2003 e (c) 2004.

[11] Para melhor visualização dessas imagens e de outras que surgirão até o final deste Capítulo, consulte o CDROM em anexo.

A seguir, apresentam-se as principais observações derivadas dos três mapas, não à guisa de conclusão, mas como uma contribuição para o entendimento de como esta "epidemia" se espalha por toda a cidade.

Quando analisamos a evolução do padrão espacial do risco de homicídio na cidade de São Paulo, de 2002 a 2004, verifica-se facilmente que algumas áreas concentram grande parte dos homicídios, enquanto em outras as mortes ocorrem em menores proporções (vide Apêndice E para identificação dos distritos).

Partindo da região central da cidade, observa-se a ocorrência de um foco de homicídios, também denominado "de área quente" [em inglês *hot zones* ou *hot spot* (Harries, 1999)], que abarca principalmente os distritos da Sé, Brás e imediações. Conforme estudo recente (Ceccato et al., 2004), boa parte desta região é caracterizada por prédios comerciais, hotéis, restaurantes, cinemas e hospitais, são lugares que concentram muitas pessoas durante o dia. Estas áreas possuem ou estão localizadas próximas a terminais de transportes (metrôs e ônibus) e possuem uma atividade noturna intensa através de bares, casas de bingos, casas de prostituição (saunas); além disso, está rodeada de cortiços e concentram um grande mercado de drogas. Um fato que chama a atenção nos mapas é que esta área praticamente se mantém estável de 2002 a 2004. Para se ter uma idéia aproximada do risco nesta área, calculou-se, para cada ano, a média do risco de homicídio sobre os distritos envolvidos[12]. Conforme mostra a Tabela 4.8, em 2002, esta área apontava para um risco médio da ordem de 128 mortes por cem mil habitantes passando para cerca de 116 mortes por cem mil habitantes em 2003, e depois para quase 124 mortes por cem mil habitantes em 2004, portanto uma redução de apenas 3,1% em relação a 2002.

TABELA 4.8 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos do centro da cidade de São Paulo.

Distrito	**2002**	**2003**	**2004**
Sé	57,90	53,02	54,74
Brás	70,34	63,71	68,86
Total	128,24	116,73	123,60

[12] Resultados baseados em operações zonais do módulo LEGAL (Linguagem Espacial de Processamento Algébrico) do sistema SPRING (http://www.dpi.inpe.br/spring).

Deslocando-se do centro em direção às periferias, observa-se que o *hot spot* central vai se dissolvendo, até que surge uma área significativa que engloba vários distritos que são menos vulneráveis à criminalidade. Boa parte desta área é dotada de melhor infra-estrutura, nível econômico mais elevado e qualidade de vida melhor comparado com o resto da cidade. Como exemplo, os distritos de Jardim Paulista, Alto de Pinheiros, Morumbi, Tatuapé e Perdizes. Conforme indica a Tabela 4.9, somente essas áreas tiveram seus índices reduzidos de 2002 para 2004 da ordem 32%, cerca de 10 vezes mais que no centro da cidade. Um fato que chamou atenção, vide Figura 4.22, refere-se ao distrito do Morumbi. Em 2002 apresentava um índice da ordem de (29,29), em 2003 ao invés de diminuir, aumenta para (37,76) e, em 2004 decai para (22,85). Uma possível explicação para esta flutuação é porque ali convivem extremos de riqueza e pobreza (favela de Heliópolis, a maior da capital com aproximadamente 100 mil habitantes).

TABELA 4.9 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos em torno do centro da cidade de São Paulo.

Distrito	**2002**	**2003**	**2004**
Jardim Paulista	7,75	7,15	4,55
Alto de Pinheiros	17,76	8,60	7,13
Morumbi	29,29	37,76	22,85
Tatuapé	20,46	19,51	18,51
Perdizes	12,41	10,23	6,43
Total	87,67	83,25	59,47

A medida em que se avança das áreas de baixo risco em direção às periferias, estas vão se dissolvendo até que surgem três áreas de alto risco, uma localizada na periferia Sul da cidade e as outras duas, nas periferias Leste e Norte, respectivamente. Esta concentração espacial da violência nas regiões periféricas da cidade certamente não se deve a "maus fluídos" provenientes do subsolo.

A periferia Sul teoricamente é considerada área de proteção ambiental por englobar as represas de Guarapiranga e Billings. Distritos como Capão Redondo, Jardim Ângela, Jardim São Luis, Cidade Dutra, Parelheiros, Grajaú e Marsilac, são considerados lugares de altíssimo risco do município; boa parte da população que não tem para onde

ir se desloca para esta região. São distritos que possuem grandes densidades populacionais com ocupação totalmente desordenada e irregular nas bordas das represas, falta de infra-estrutura básica, alta proporção de pessoas jovens (de 10 a 25 anos) e uma grande parcela de adultos com baixo grau de escolaridade. Somente em Parelheiros e Grajaú a população cresceu quase 40 mil pessoas, de 2002 para 2004, acompanhada de um índice que oscilou entre 90 a 60 homicídios por cem mil habitantes. Apesar deste quadro trágico, houve uma melhora de 2002 para 2004. Observe nos mapas, Figura 4.22, que o *hot spot* sofre uma atenuação gradual. A Tabela 4.10 nos dá uma idéia aproximada desta diminuição, que foi da ordem de 30%. No entanto, essas áreas ainda são de alto risco quando comparadas a outros locais da cidade. Por exemplo, tomando os dados das Tabelas 4.9 e 4.10 constata-se que em 2004 a chance de morrer assassinado em Parelheiros era cerca de treze vezes maior que no Jardim Paulista, em Grajaú da ordem de oito vezes em relação a Alto de Pinheiros, ou então, no Jardim Ângela cerca de nove vezes maior que em Perdizes.

TABELA 4.10 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos da periferia Sul da cidade de São Paulo.

Distrito	**2002**	**2003**	**2004**
Capão Redondo	70,34	63,71	68,86
Jardim São Luiz	72,60	65,78	47,23
Jardim Ângela	91,52	78,29	56,44
Cidade Dutra	72,33	69,17	49,19
Parelheiros	96,78	79,28	61,21
Grajaú	85,22	80,22	62,02
Marsilac	64,55	66,40	46,18
Total	553,34	502,85	391,13

A Periferia Leste da cidade concentra aproximadamente 20% da população do município (são quase 2 milhões de habitantes). É uma região marcada pela presença de loteamentos precários e favelas, baixa renda familiar, falta de infra-estrutura urbana, etc. É uma região também estigmatizada pela violência, a qual se espalha por cerca de dez distritos. Conforme ilustra a Figura 4.22 (a), em 2002 observou-se a ocorrência de três *hot spots*: o primeiro que englobava parte dos distritos de São Miguel e Vila Curuçá,

cujo risco médio de homicídios foi da ordem de 60 mortes por cem mil habitantes; o segundo, um pouco maior, que englobava boa parte dos distritos de Lajeado, Guaianazes, Cidade Tiradentes e José Bonifácio, cujo risco médio de homicídios foi da ordem de 66 mortes por cem mil habitantes; e o terceiro, que abarcava parte dos distritos de Parque do Carmo, Iguatemi, São Rafael e São Mateus, com risco médio de homicídios próximo de 63 mortes por cem mil habitantes. Somente em 2002, estes distritos atingiram um total de aproximadamente 640 mortes por cem mil habitantes. No entanto, de 2002 para 2004, conforme ilustram os mapas das Figuras 4.22 (a), (b) e (c), houve uma queda acentuada da violência nesta região. De 2002 para 2003 foi da ordem 10,77 % e de 2002 para 2004 de aproximadamente 40,4%, conforme os dados da Tabela 4.11.

TABELA 4.11 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos da periferia Leste da cidade de São Paulo.

Distrito	**2002**	**2003**	**2004**
São Miguel	60,98	49,01	42,60
Vila Curuçá	60,87	48,72	37,02
Lajeado	64,40	61,03	34,67
Guaianazes	79,20	71,03	41,84
Cidade Tiradentes	65,18	60,86	34,97
José Bonifácio	57,93	55,31	34,32
Parque do Carmo	60,07	55,79	40,92
Iguatemi	68,08	57,27	39,78
São Rafael	64,42	58,50	41,01
São Mateus	59,72	54,34	34,80
Total	640,85	571,86	381,93

Na periferia Norte da cidade constatou-se, em 2002, a presença de um *hot spot* que englobava parte dos distritos de Cachoeirinha e Brasilândia. Esses dois distritos juntos possuíam uma aglomeração populacional da ordem de 400 mil habitantes. Parte dessa região é considerada área de conservação e recuperação ambiental, e tem características similares às outras regiões periféricas da cidade. Um fato que chamou atenção nos mapas, Figuras 4.22 (a), (b) e (c), foi o deslocamento do foco de homicídios nesta região. Inicialmente, em 2002, estava concentrado entre os distritos de Cachoeirinha e

Brasilândia, atingindo em 2004 o distrito de Perus. Em 2002 o distrito de Perus apresentava um risco da ordem 40 mortes por cem mil habitantes, passando para cerca de 56 mortes por cem mil habitantes em 2004, um aumento razoável, da ordem de 40%.

TABELA 4.12 - Risco médio de homicídios por cem mil habitantes sobre alguns distritos da periferia Norte da cidade de São Paulo.

Distrito	**2002**	**2003**	**2004**
Brasilândia	57,35	53,89	49,12
Cachoeirinha	55,28	48,07	39,24
Perus	40,22	46,42	56,37
Total	152,85	148,38	144,73

As observações que seguem são relativas às estimativas obtidas. Primeiro, observa-se com clareza através dos mapas apresentados nas Figuras 4.22(a), 4.22(b) e 4.22(c), que a média da distribuição do risco de homicídio sofre um enorme efeito de suavização na região Sul da cidade, mais especificamente sobre os distritos de Parelheiros, Grajaú e Marsilac. Este efeito é decorrente da pouca quantidade de informação disponível nesta região, a qual se restringe em apenas três amostras localizadas nos respectivos centróides dos distritos mencionados. Somente esta região ocupa uma área de aproximadamente 30% da área total da cidade (450 km^2 / 1524 km^2), portanto uma amostragem mais representativa faz-se necessário. Segundo, as geometrias impostas por estes distritos influenciam também nas variâncias das estimativas do risco. As variâncias são maiores aonde a informação é escassa e vice-versa, conforme ilustram as Figuras 4.23 (a), 4.23 (b) e 4.23 (c).

Uma solução para amenizar os problemas mencionados acima, seria refazer estas áreas, por exemplo, com auxílio de imagens de satélite de alta resolução, separando as áreas de concentração populacional, onde supostamente há maior incidência de homicídios, das áreas de reservas ambientais. Ainda seria possível subdividir esses distritos em áreas menores, de forma a homogeneizar com os demais distritos da cidade. Isto poderia ser realizado utilizando como base o mapa de setores censitários fornecido pelo IBGE e depois estimar o número de homicídios para cada nova subdivisão. Isto aumentaria a

quantidade de informação para estimar o risco nesta região, o erro de predição tenderia a ser menor e, conseqüentemente, o efeito de suavização seria menos acentuado.

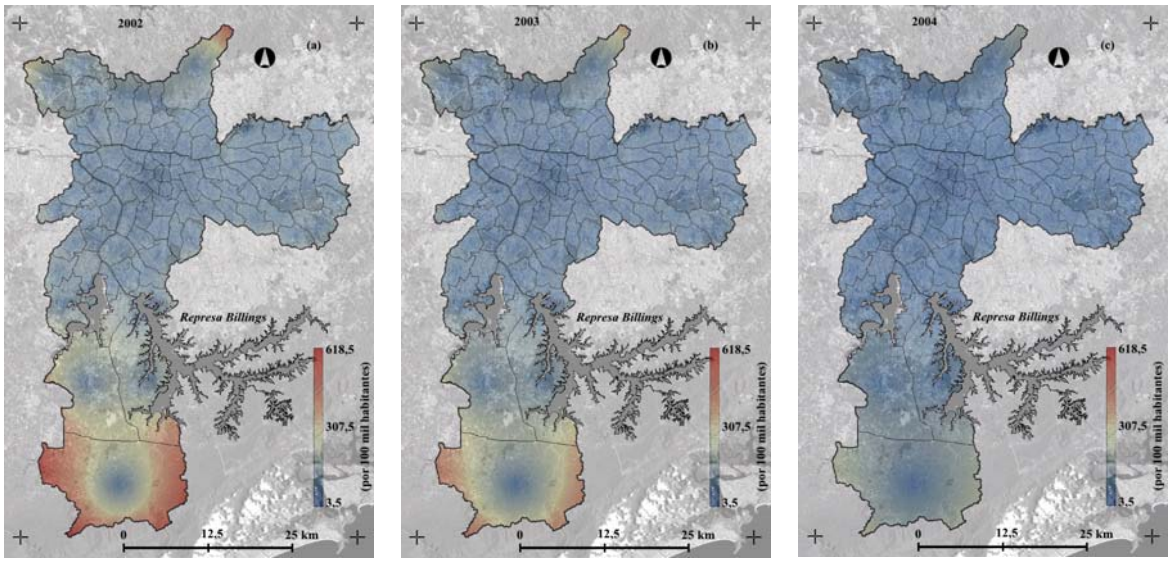


FIGURA 4.23 - Evolução das variâncias das estimativas do risco de homicídio: (a) 2002, (b) 2003 e (c) 2004.

Esta análise teve por objetivo mostrar as principais tendências da distribuição espacial do risco de homicídio na cidade de São Paulo, durante os anos de 2002 a 2004. Os resultados apresentados são condizentes com a estrutura urbana da cidade e trabalhos recentes [Ceccato et al. (2004), Nery e Monteiro (2006)]. Para complementar, a seção seguinte apresenta através de procedimentos de simulação seqüencial condicionada a construção de cenários para o risco de homicídio.

4.9 Construção de cenários do risco de homicídio

A construção de cenários para o risco de homicídio, R(**u**), foi estabelecida por simulação seqüencial condicionada por indicação, conforme descrito na Seção 3.9. Para cada ano investigado (2002, 2003 e 2004) os seguintes passos foram realizados:

1) Para cada um dos 96 centróides das áreas componentes da região de estudo, uma estimativa do risco de homicídio foi obtida por co-krigeagem binomial, conforme descrito na Seção 3.7. Isto resultou num conjunto de valores estimados, denotado por $\{r(\mathbf{u}_i), i = 1, ..., 96\}$ da V.A. $R(\mathbf{u}_i)$, cujas estatísticas estão sumarizadas na Figura 4.24.

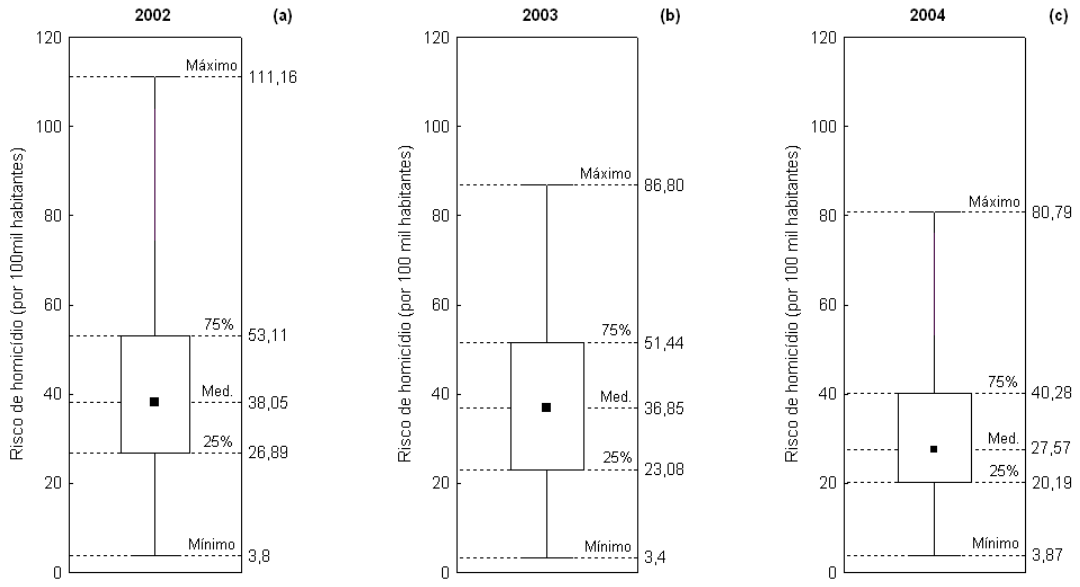


FIGURA 4.24 - *Box plot* das estimativas do risco de homicídio nos centróides: (a) 2002, (b) 2003 e (c) 2004.

2) Para estabelecer os valores de cortes, requisito necessário do formalismo por indicação (Felgueiras (1999), Deutsch e Journel (1998)), dividiu-se o conjunto de dados $\{r(\mathbf{u}_i), i = 1, ..., 96\}$ em 5 subconjuntos selecionando-se 4 valores de cortes. Cada um dos 5 subconjuntos foi definido com, aproximadamente, a mesma cardinalidade. Os valores de cortes estabelecidos, para cada ano investigado, estão apresentados na TABELA 4.13.

TABELA 4.13 - Valores de cortes (por 100 mil habitantes) para os anos de 2002, 2003 e 2004.

Valores de cortes	**2002**	**2003**	**2004**
1º Quintil *(p=0,2)*	25,48	20,95	18,21
2º Quintil *(p=0,4)*	31,66	29,60	24,52
3º Quintil *(p=0,6)*	41,11	43,02	31,08
4º Quintil *(p=0,8)*	59,11	54,22	42,70

3) Para cada valor de corte, realizou-se a transformação por indicação do conjunto de dados $\{r(\mathbf{u}_i), i = 1, ..., 96\}$; isto é, se $r(\mathbf{u}_i) \leq$ valor de corte $\Rightarrow r(\mathbf{u}_i) = 1$, caso contrário, $r(\mathbf{u}_i) = 0$. Criaram-se, assim, para cada ano investigado, 4 conjuntos de variáveis por indicação. Para cada um desses conjuntos, os semivariogramas por

indicação foram estabelecidos, conforme apresentados nas Figuras 4.25, 4.26 e 4.27, e depois ajustados com modelos teóricos tipo esférico, cujos parâmetros estão sintetizados nas TABELAS 4.14, 4.15, e 4.16.

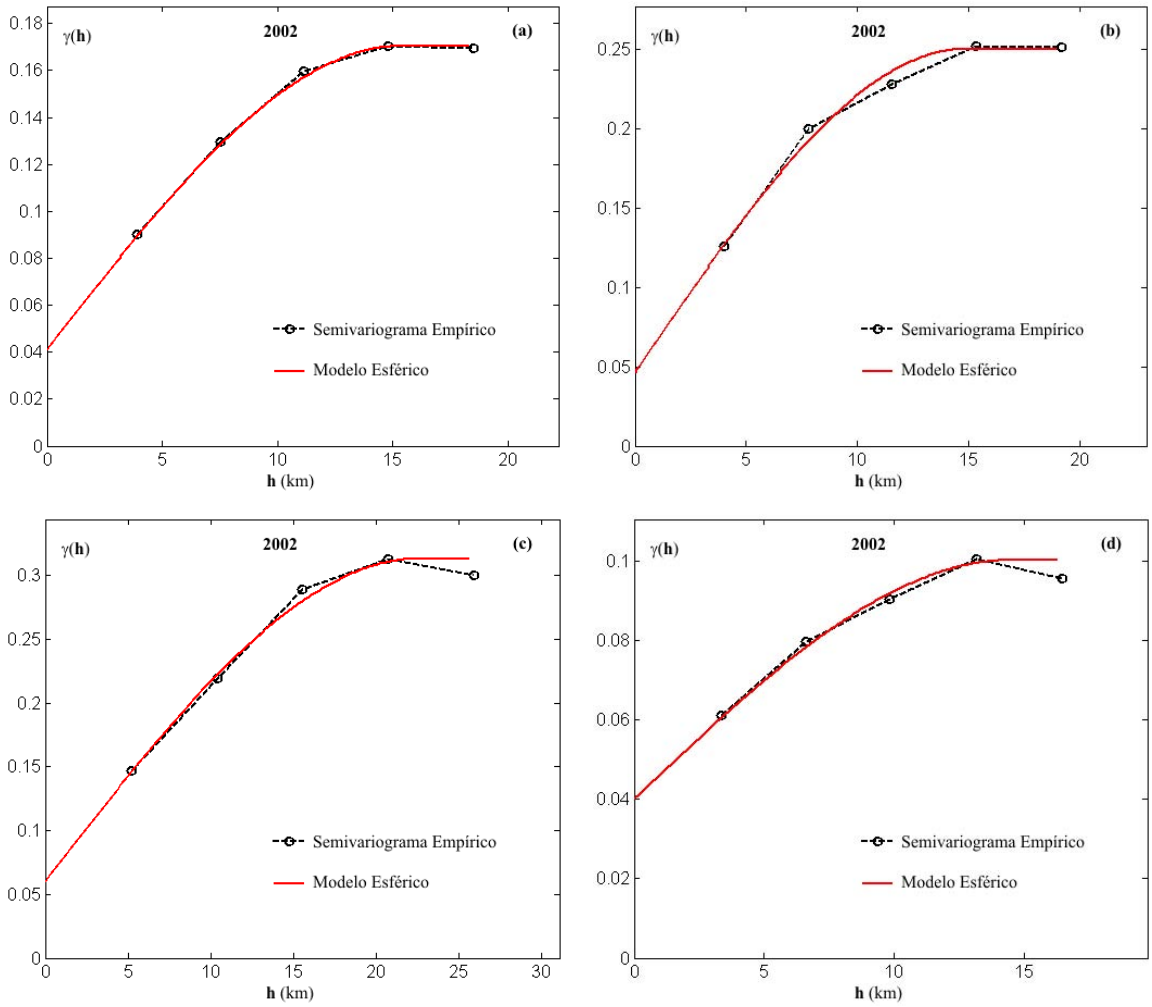


FIGURA 4.25 - Semivariogramas por indicação segundo valores de cortes em 2002: (a) 1º quintil, (b) 2º quintil, (c) 3º quintil e (d) 4º quintil.

TABELA 4.14 - Parâmetros dos modelos esféricos de semivariogramas segundo os valores de cortes, para o ano de 2002.

Valores de cortes	efeito pepita (C_0)	contribuição (C_1)	alcance (a) km
1º quintil = 25,48	0,041	0,129	15355,0
2º quintil = 31,66	0,046	0,203	14822,5
3º quintil = 41,11	0,060	0,253	22620,0
4º quintil = 59,11	0,040	0,060	14498,0

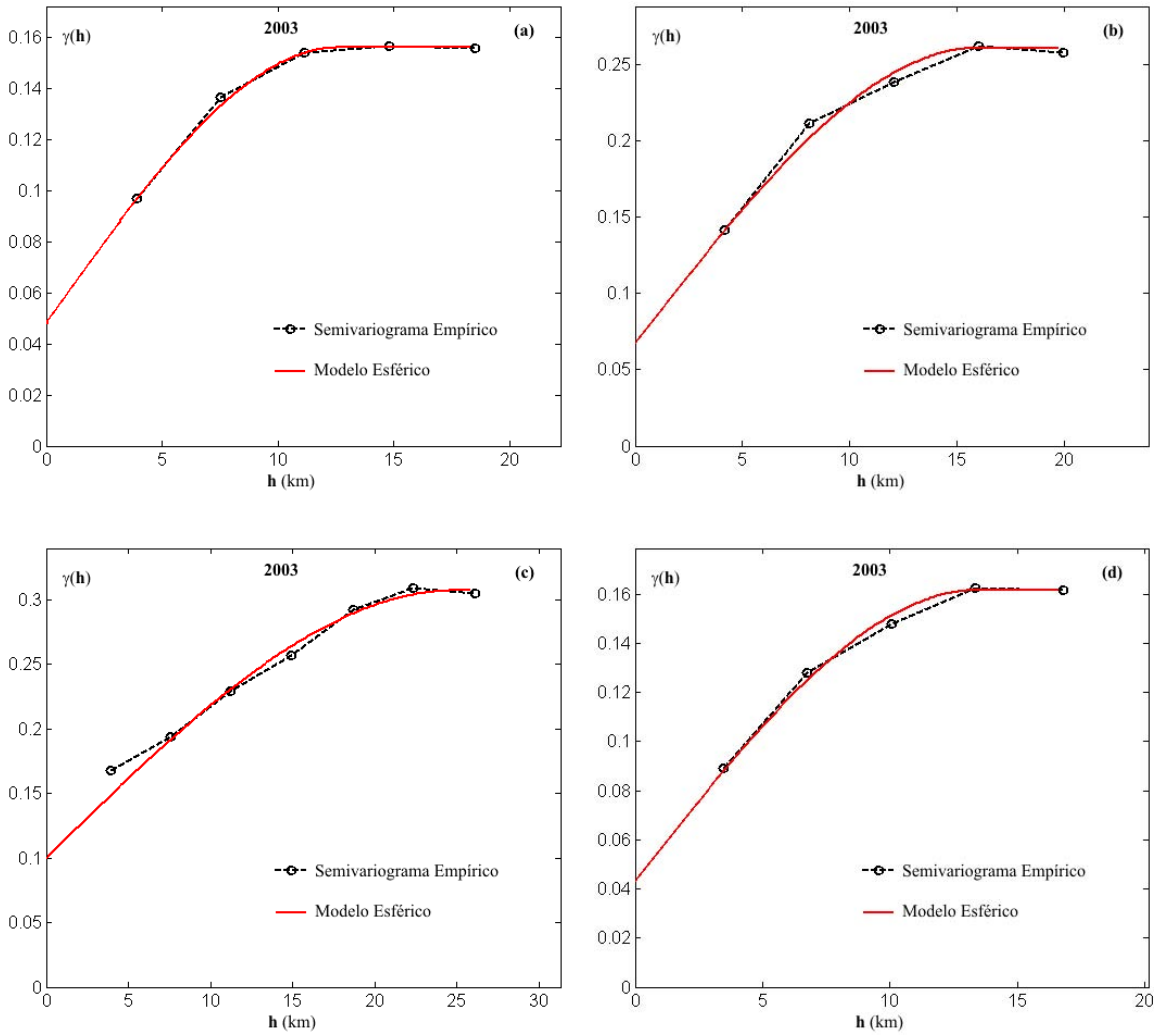


FIGURA 4.26 - Semivariogramas por indicação segundo valores de cortes em 2003: (a) 1º quintil, (b) 2º quintil, (c) 3º quintil e (d) 4º quintil.

TABELA 4.15 - Parâmetros dos modelos esféricos de semivariogramas segundo os valores de cortes, para o ano de 2003.

Valores de cortes	efeito pepita (C_0)	contribuição (C_1)	alcance (a) km
1º quintil = 20,95	0,048	0,108	12580,0
2º quintil = 29,60	0,067	0,193	16000,0
3º quintil = 43,02	0,100	0,206	24864,3
4º quintil = 54,22	0,043	0,118	13420,0

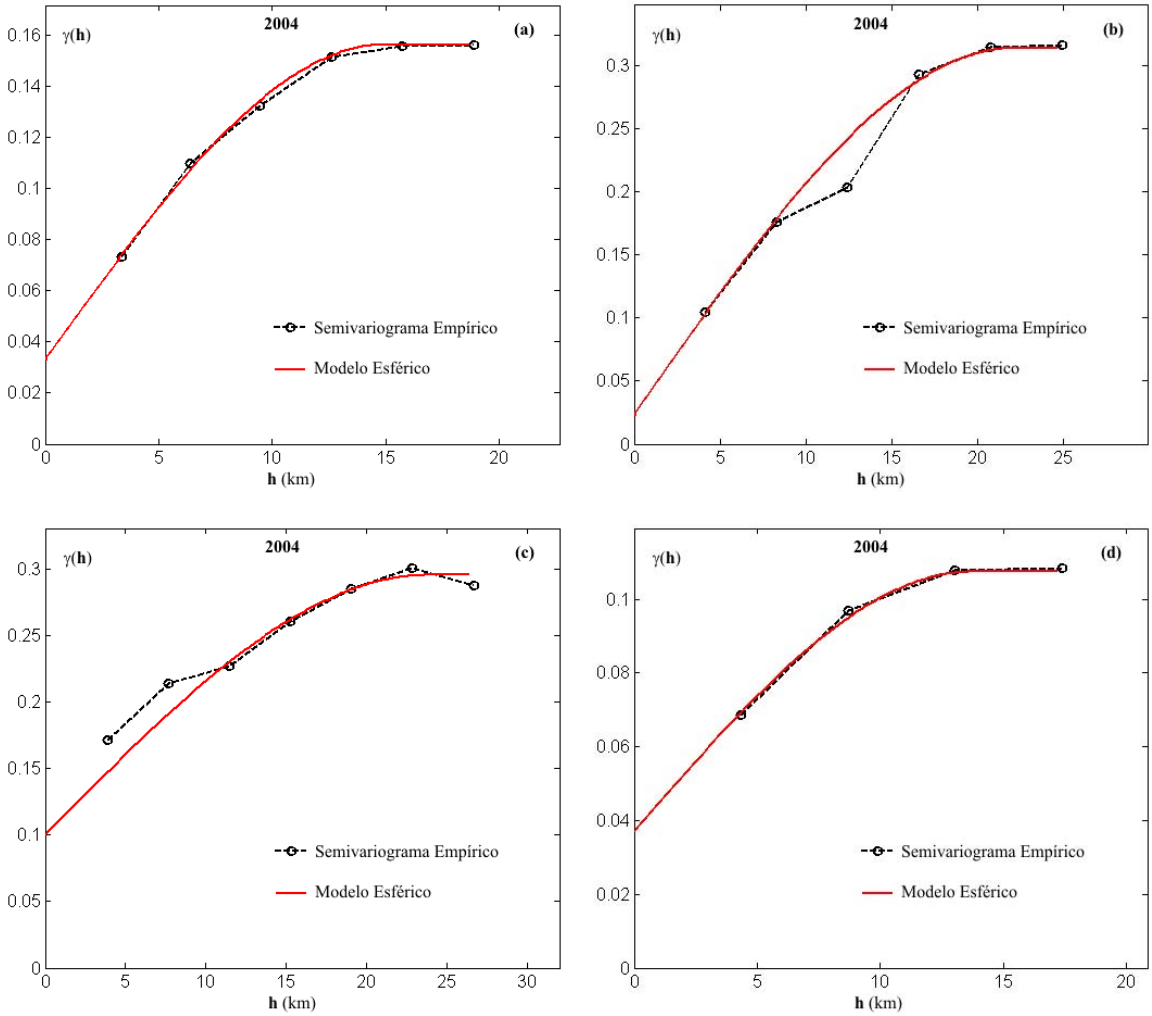


FIGURA 4.27 - Semivariogramas por indicação segundo valores de cortes em 2004: (a) 1º quintil, (b) 2º quintil, (c) 3º quintil e (d) 4º quintil.

TABELA 4.16 - Parâmetros dos modelos esféricos de semivariogramas segundo os valores de cortes, para o ano de 2004.

Valores de cortes	efeito pepita (C_0)	contribuição (C_1)	alcance (a) km
1º quintil = 18,21	0,033	0,123	14931,0
2º quintil = 24,52	0,023	0,290	22134,3
3º quintil = 31,08	0,100	0,195	23811,0
4º quintil = 42,70	0,037	0,070	13714,4

4) O procedimento de simulação seqüencial por indicação foi aplicado, utilizando-se o conjunto de dados {r($\mathbf{u}_i$), i = 1, ..., 96}, os valores de cortes apresentados na TABELA 4.13 e os modelos teóricos de semivariogramas apresentados nas

TABELAS 4.14, 4.15 e 4.16, respectivamente. Foram geradas 200 realizações equiprováveis de R(**u**), cada uma com resolução de 150 metros na direção Norte-Sul e 98 metros na direção Leste-Oeste, proporcionando grades densas de valores (520 linhas por 530 colunas). O procedimento de simulação seqüencial por indicação foi executado com o programa *sisim.exe* da GSLIB [Deutsch e Journel (1998)]. A Figura 4.28 ilustra um exemplo, para o ano de 2002, em que se mostra o mapa do risco de homicído oriundo da co-krigeagem binomial, Figura 4.28 (a), e 3 realizações equiprováveis de R(**u**), Figuras 4.28 (b), (c) e (d), escolhidas aleatoriamente da simulação.

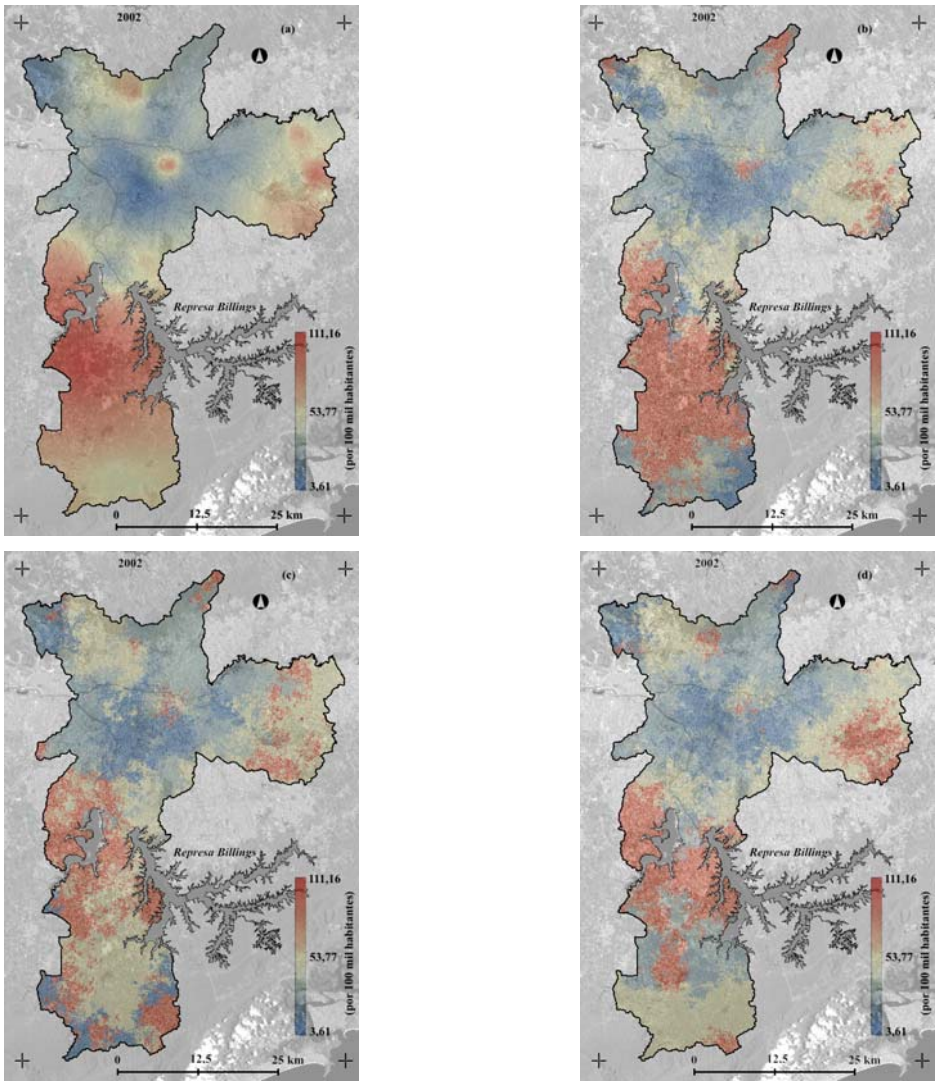


FIGURA 4.28 - Em 2002: (a) Mapa do risco de homicídio por co-krigeagem binomial; (b), (c) e (d) realizações equiprováveis de R(**u**) oriundas da simulação.

5) Uma vez obtido o conjunto de 200 realizações alternativas equiprováveis do risco de homicídio, R(**u**), por simulação estocástica, tem-se, para cada localização **u** da malha do mapa espacial, um conjunto de 200 valores simulados r(**u**) da V.A. R(**u**). A partir deste conjunto, a função de distribuição acumulada de R(**u**) foi estabelecida, possibilitando o cálculo de vários valores de cortes e, posteriormente, a construção de cenários, conforme segue.

Para cada ano investigado, foram selecionados três valores de cortes tomados da distribuição acumulada de R(**u**): dois extremos e um mediano.

As Figuras 4.29 (a), (b) e (c) apresentam a evolução de cenários do risco de homicídio de 2002 a 2004, respectivamente, para valores de cortes obtidos do segundo decil da função de distribuição acumulada de R(**u**). Referem-se a representações de cenários otimistas, em que cada localização **u** da malha do mapa espacial é excedida por 80% dos valores simulados. Neste caso, as áreas do mapa que apresentam valores elevados indicam que o risco de homicídio nessas localizações é muito alto, mesmo em cenários otimistas.

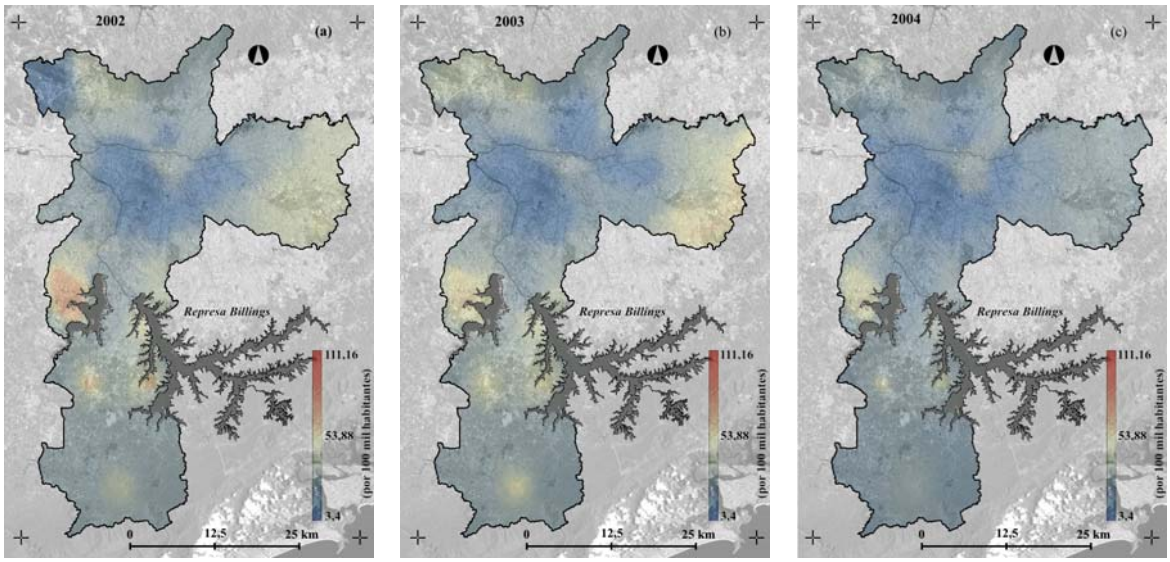


FIGURA 4.29 - Evolução de cenários otimistas do risco de homicídio, com detalhe da mancha urbana na imagem de fundo: (a) 2002, (b) 2003 e (c) 2004.

Uma forma alternativa de visualizar os resultados da Figura 4.29, é apresentar nos mapas somente as áreas consideradas de alto risco. A Figura 4.30 ilustra um exemplo, da evolução de cenários otimistas (de 2002 a 2004), em que são apresentados somente as áreas em que o risco de homicídio excede 40 mortes por cem mil habitantes.

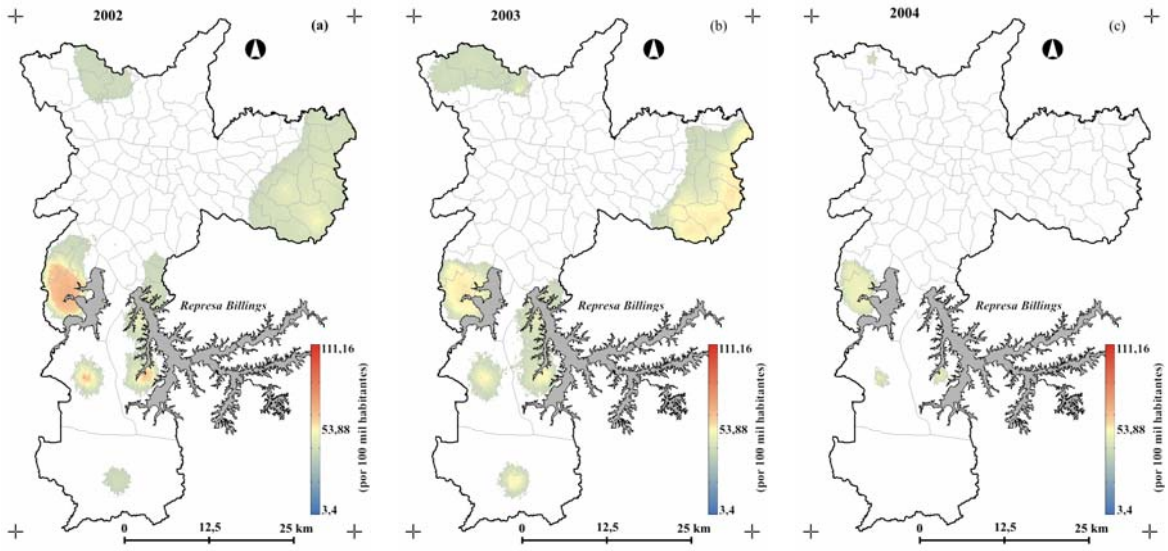


FIGURA 4.30 - Evolução de cenários otimistas, com exposição somente das áreas em que o risco de homicídio excede 40 mortes por cem mil habitantes: (a) 2002, (b) 2003 e (c) 2004.

A situação adversa dos cenários otimistas são os cenários pessimistas. As Figuras 4.31 (a), (b) e (c) apresentam a evolução de cenários do risco de homicídio de 2002 a 2004, respectivamente, para valores de cortes obtidos do oitavo decil da função de distribuição acumulada de R(**u**). Referem-se a representações de cenários pessimistas, em que cada localização **u** da malha do mapa espacial é excedida apenas por 20% dos valores simulados. Neste caso, as áreas do mapa que apresentam valores baixos indicam que o risco de homicídio nessas localizações é baixo, mesmo em cenários pessimistas.

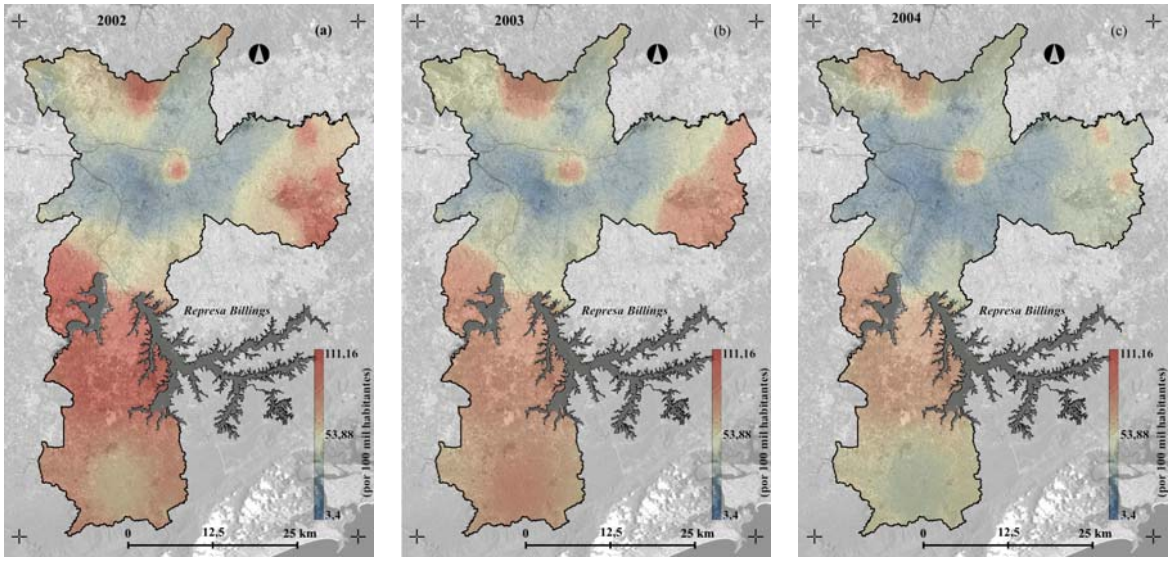


FIGURA 4.31 - Evolução de cenários pessimistas do risco de homicídio, com detalhe da mancha urbana na imagem de fundo: (a) 2002, (b) 2003 e (c) 2004.

De maneira análoga à Figura 4.30, segue um exemplo da evolução de cenários pessimistas (de 2002 a 2004), no qual são apresentados somente as áreas em que o risco de homicídio excede 40 mortes por cem mil habitantes, conforme ilustrado na Figura 4.32. Comparando-se os resultados das Figuras 4.30 e 4.32, observa-se com clareza o crescimento das áreas que são mais vulneráveis a violência.

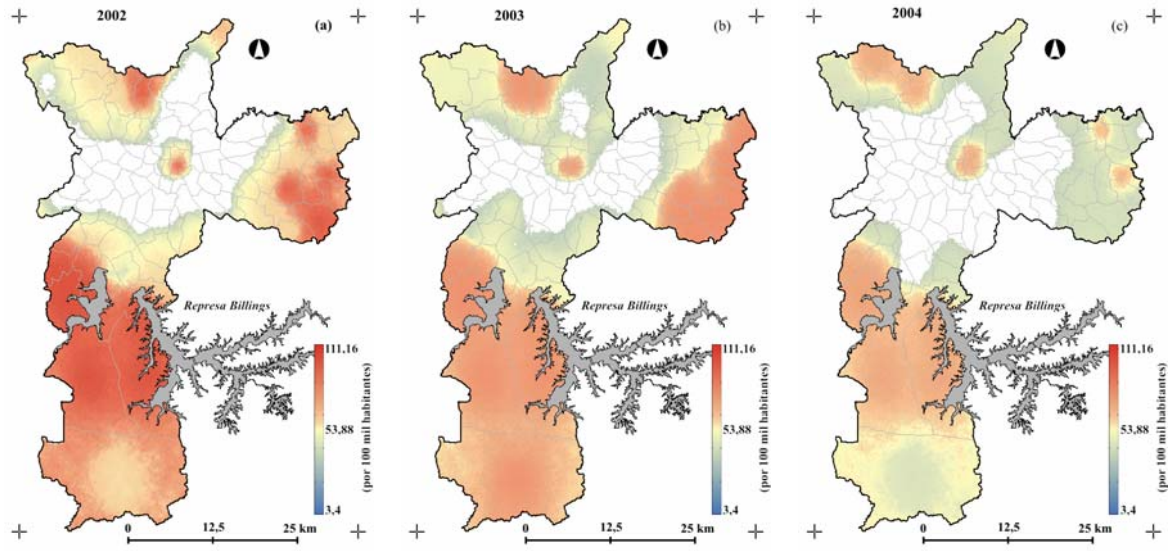


FIGURA 4.32 - Evolução de cenários pessimistas, com exposição somente das áreas em que o risco de homicídio excede 40 mortes por cem mil habitantes: (a) 2002, (b) 2003 e (c) 2004.

Seguindo, os resultados apresentados nas Figuras 4.39 (a), (b) e (c), referem-se a representações de cenários intermediários do risco de homicídio, obtidos da mediana da distribuição acumulada de R(**u**).

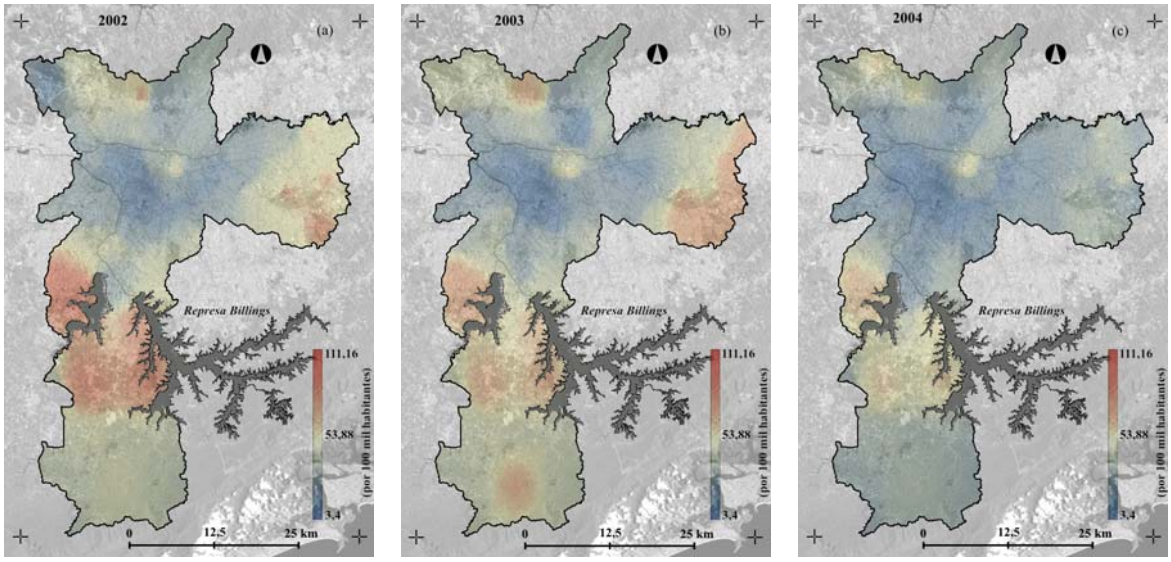


FIGURA 4.33 - Evolução de cenários medianos do risco de homicídio, com detalhe da mancha urbana na imagem de fundo: (a) 2002, (b) 2003 e (c) 2004.

De maneira análoga às Figuras 4.30 e 4.32 , segue um exemplo da evolução de cenários medianos, no qual são apresentados somente as áreas em que o risco de homicídio excede 40 mortes por cem mil habitantes, conforme ilustrado na Figura 4.34.

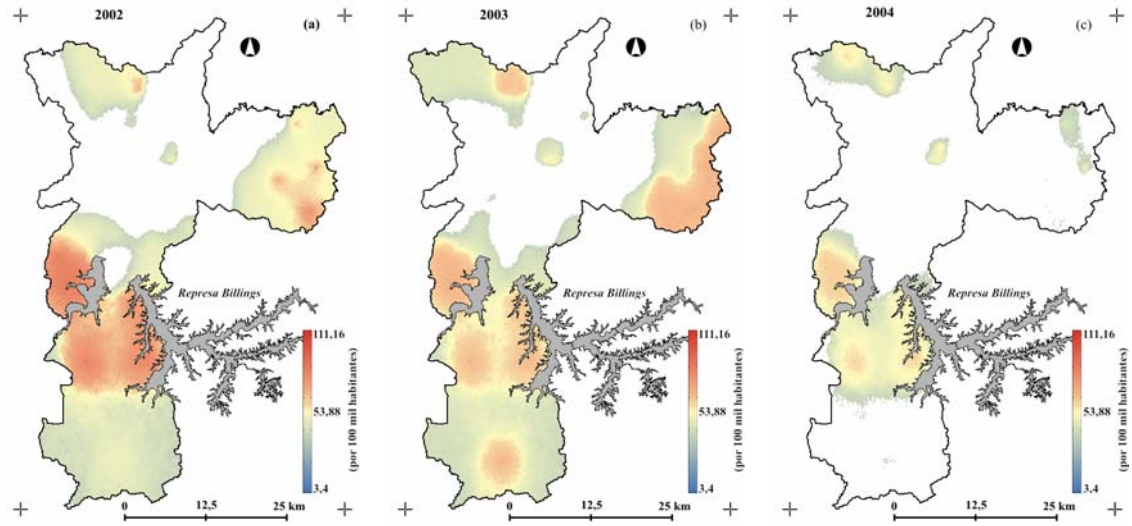


FIGURA 4.34 - Evolução de cenários medianos, com exposição somente das áreas em que o risco de homicídio excede 40 mortes por cem mil habitantes: (a) 2002, (b) 2003 e (c) 2004.

Para complementar, campos de incerteza foram construídos pela diferença entre quartis (3º quartil - 1º quartil) da distribuição acumulada de R(**u**). Os resultados são mostrados nas Figuras 4.35 (a), (b) e (c), os quais apresentam variações proporcionais ao comportamento do risco de homicídio na região de estudo. Tomando como referência as Figuras 4.35 (d), (e) e (f), nas áreas em que existe uma variação maior do valor do risco de homicídio, os valores de variância são maiores, como mostram as Figuras 4.35 (a), (b) e (c). Por outro lado, nas áreas em que o atributo tem variação mais suave, ou não varia, observam-se valores de incertezas menores.

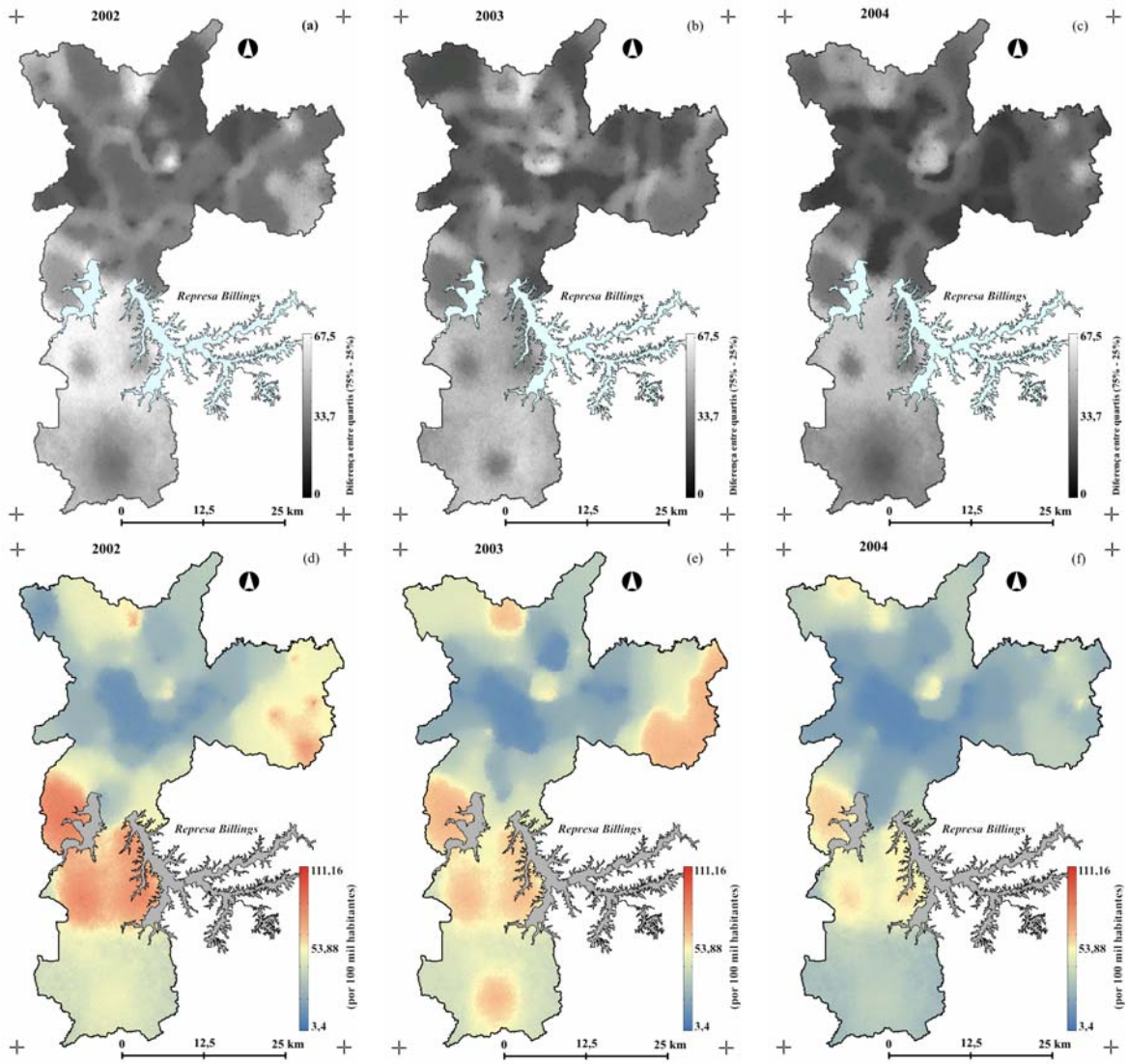


FIGURA 4.35 - Campos de incertezas gerados por simulação seqüencial por indicação: (a) 2002, (b) 2003 e (c) 2004. Cenários medianos do risco de homicídio: (d) 2002, (e) 2003 e (f) 2004.

Esta seção teve por objetivo apresentar o uso da simulação seqüencial condicionada não-paramétrica para a construção de cenários do risco de homicídio na cidade de São Paulo, no triênio 2002 – 2004. Os resultados apresentados elucidam diferentes características do campo aleatório investigado e podem auxiliar o planejador que, orientado pelos seus objetivos, tem a possibilidade de escolher cenários mais adequados ao trabalho a ser executado.

No Capítulo seguinte apresentam-se as conclusões a respeito deste trabalho, com críticas e sugestões para futuros desenvolvimentos.

CAPÍTULO 5

CONSIDERAÇÕES FINAIS

Quanto à metodologia adotada

- A metodologia desenvolvida neste trabalho, para estimação e mapeamento do risco de ocorrência de eventos raros, incorpora os seguintes aspectos:

    1) Considera a natureza binomial dos dados de contagem; isto é, o número de ocorrências associado ao evento é modelado como uma variável aleatória com distribuição binomial;

    2) Emprega o estimador proposto para o semivariograma do risco na análise da dependência espacial do risco associado ao evento em estudo, considerando os problemas da instabilidade que se observa nos dados e sua tendência;

    3) A estrutura de correlação espacial do risco é avaliada através de um esquema de simulação, que possibilita verificar se a mesma é condizente com as dimensões da área de estudo e a geometria imposta pelas áreas componentes da região de estudo;

    4) Emprega os procedimentos: i) de co-krigeagem binomial para obtenção de mapas de média da distribuição do risco; ii) de simulação seqüencial condicionada não-paramétrica para a produção de cenários do risco e de mapas de incerteza;

- A metodologia desenvolvida fornece um ferramental para descrever o processo contínuo do risco. Quando aplicada na investigação do risco de homicídio na cidade de São Paulo, no triênio 2002 – 2004, permitiu a possibilidade de delinear com clareza as regiões da cidade que são mais vulneráveis a esse tipo de violência. Os resultados obtidos (mapas e cenários) são de grande importância

para os órgãos públicos de saúde e segurança, pois, permitem identificar padrões espaciais da violência, podem auxiliar na compreensão dos fatores que atuam no seu desencadeamento, e subsidiar o planejamento de ações que objetivem a sua inibição.

Quanto ao estimador proposto para o semivariograma do risco

- O estimador de Oliver et al. (1998) é construído para casos em que as unidades de área são relativamente homogêneas, como foi o caso de estudo apresentado em seu trabalho. Regiões de estudos que apresentam unidades de área heterogêneas (agregação de grupos sociais distintos, diferenças em população e área), que é o caso das grandes cidades brasileiras, o estimador proposto têm melhor adequação, uma vez que, trabalha a instabilidade do dado decorrente de áreas com pequenas populações, e se ajusta às tendências impostas pela geometria das áreas componentes da região de estudo.

Quanto às limitações do método

- Depende do nível de agregação imposto ao sistema de unidades de área, ou da escala adotada. Isto se reflete na geometria de amostragem, normalmente estabelecida pelos centróides das áreas componentes da região de estudo, e, posteriormente, na definição da estrutura de correlação espacial do risco imposta pelo estimador de semivariograma do risco. Conseqüentemente, as estimativas do risco são válidas somente para o nível de agregação pré-estabelecido.

- Estima o risco de ocorrência do evento somente a partir de dados de taxa agregados por unidade de área. O modelo não emprega o uso de outras informações presente em outros dados.

Sugestões para futuras melhorias e investigações

- Investigar formas alternativas para o cálculo dos parâmetros de média e variância zonal, os quais são empregados na formulação do estimador proposto para o semivariograma do risco. Uma sugestão seria primeiro codificar os dados de acordo com as zonas de risco pré-estabelecidas. Depois, para cada vetor distância **h** de análise calcular: i) o número de pares de pontos correspondentes às respectivas zonas de risco; ii) o número de pares de pontos com dupla identidade; isto é, pares de pontos em que uma das extremidades do vetor distância **h** pertence a uma zona de risco e a outra extremidade de **h** pertence a uma outra zona de risco. De posse deste cálculo poder-se-ia estabelecer um esquema de ponderação, o qual seria aplicado às médias e variâncias zonais. Posteriormente, uma média e uma variância final seriam calculadas e aplicadas na formulação do estimador proposto. Isto talvez contribua para o estabelecimento de uma estrutura de correlação espacial do risco mais condizente com as dimensões da área de estudo, e com as tendências impostas pela geometria das áreas componentes da região de estudo.

- Inclusão no modelo de fatores demográficos e/ou territoriais. Por exemplo, no estudo de caso do risco de homicídio apresentado neste trabalho existem diversos fatores que podem contribuir para explicar a sua distribuição, tais como: índice de concentração de renda, taxa de evasão escolar, crescimento populacional, diferenças nas taxas de desemprego entre os jovens, distribuição de equipamentos públicos como escolas, parques, hospitais e outros. Teoricamente isto poderia ser feito modificando-se o sistema de co-krigeagem, apresentado na Seção 3.7, e o procedimento de simulação, descrito na Seção 3.9, de modo a incorporar outras variáveis ou fatores correlacionados ao risco. A incorporação de alguns desses fatores no procedimento de modelagem poderia tornar as estimativas do risco mais precisas.

- Realizar um estudo comparativo com outras metodologias já bem estabelecidas. Isto é importante, porque permite avaliar as diferenças das estimativas do risco e de sua distribuição espacial.

- Disponibilizar a metodologia apresentada como instrumento de análise, em plataformas de softwares livres. Isto representa uma contribuição adicional desse trabalho que transcende aos resultados aqui apresentados. Essa metodologia implementada poderá ser empregada em outros estudos de mesma natureza.

REFERÊNCIAS BIBLIOGRÁFICAS

Akerman, M.; Bousquat, A. Mapas de risco de violência. **São Paulo em Perspectiva**, v. 13, n. 4, p. 112-120, 1999.

Andrade Neto, P. R.; Ribeiro Junior, P. J.; Fook, K. D. Integration of Statistics and Geographic Information Systems: the R/TerraLib. In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA - GeoInfo, 7., 2005, Campos de Jordão - SP. **Anais...** São José dos Campos - SP: INPE, 2005. p. 139-151. ISBN 85-17-00022-6. Disponível em: < http://www.geoinfo.info/geoinfo2005/papers/P74.PDF >. Acesso em: 05 ago. de 2006.

Anselin, L.; Cohen, J.; Cook, D.; Gorr, W.; Tita, G. Spatial analyses of crime. In: Duffee, D. (Ed.). **Criminal justice 2000**. v. 4. Washington, DC: National Institute of Justice, 2000.

Assunção, R. M.; Barreto, S. M.; Guerra, H. L.; Sakurai, E. Mapas de taxas epidemiológicas: uma abordagem Bayesiana. **Caderno de Saúde Pública**, v. 14, n. 4, p. 713-723, 1998.

Bailey, T. C.; Gatrell, A. C. **Interactive spatial data analysis**. New York: Wiley, 1995. 413 p.

Beato, C. C. F. Determinantes da criminalidade em Minas Gerais. **Revista Brasileira de Ciências Sociais**, v. 13, n. 37, p. 74-89, 1998.

Beato, C. C. F.; Assunção, R. M.; Santos, M. C. **Análise da evolução temporal da criminalidade violenta em Minas Gerais (1986-1997)**. São Paulo: Mimeo, 1997.

Beato, C. C. F.; Souza, R. S. R.; Ottoni, M.; Figueiredo, B.; Silveira, A. M. **Programa Fica Vivo:** ações simples, e resultados efetivos. Belo Horizonte: CRISP, 2003. Disponível em: http://www.crisp.ufmg.br.

Berke, O. Exploratory disease mapping: kriging the spatial risk function from regional count data. **International Journal of Health Geographics**, v. 3, p. 18, 2004.

Best, N.; Richardson, S.; Thomson, A. A comparison of Bayesian spatial models for disease mapping. **Statistical Methods in Medical Research**, v. 14, p. 35-59, 2005.

Bönisch, S. **Geoprocessamento ambiental com tratamento de incerteza:** o caso do zoneamento pedoclimático para a soja no estado de Santa Catarina. 251 p. (INPE-9474-TDI/824). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos - SP, 2001.

Brewer, C. A.; Pickle, L. Evaluation of methods for classifying epidemiological data on choropleth maps in series. **Annals of the Association of American Geographers**, v. 92, n. 4, p. 662-681, 2002.

Bussad, W. O.; Morettin, P. A. **Métodos quantitativos** - estatística básica. 4. ed. São Paulo: Atual Editora LTDA, 1995. 321 p.

Caldeira, T. P. R. **Cidade de muros:** crime, segregação e cidadania em São Paulo. São Paulo: Editora/34 Edusp, 2000. 399 p.

Câmara, G.; Druck, S.; Carvalho, M. S.; Monteiro, A. M. V.; Camargo, E. C. G.; Felgueira, C. A. **Análise espacial de dados geográficos**. Planaltina, DF: Embrapa Cerrados, 2004. 209 p.

Câmara, G.; Souza, R. C. M.; Freitas, U. M.; Garrido, J. SPRING: Integrating remote sensing and GIS by object-oriented data modeling. **Computers & Graphics**, v. 20, n. 3, p. 395-403, 1996.

Camargo, E. C. G. **Desenvolvimento, implementação e teste de procedimentos geoestatísticos (krigeagem) no sistema de processamento de informações georreferenciadas (SPRING)**. 124 p. (INPE-6410-TDI/620). Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos - SP, 1997.

Carvalho, M. S. **Aplicação de métodos de análise espacial na caracterização de áreas de risco a saúde.** Tese de Doutorado. Programa de Engenharia Biomédica, Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1997.

Ceccato, V.; Haining, R.; Kahn, T. The Geography of homicide in São Paulo – Brazil. **Estudos Criminológicos CAP/SSP**, p. 30-61, 2004.

Chilès, J. P.; Delfiner, P. **Geostatistics:** modeling spatial uncertainty. New York: Wiley, 1999. 695 p.

Christensen, O. F.; Waagepetersen, R. Bayesian prediction of spatial count data using generalized linear mixed models. **Biometrics** v. 58, p. 280-286, 2002.

Clayton, D. G.; Kaldor, J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. **Biometrics**, v. 43, p. 671-681, 1987.

Cressie, N. **Statistics for spatial data**. New York: Wiley, 1993. 900 p.

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

Deutsch, C. V.; Journel, A. G. **GSLIB:** Geostatistical Software Library and user's guide. New York: Oxford University Press, 1998. 369 p.

Diggle, P. J.; Tawn, J. A.; Moyeed, R. A. Model based geostatistics. **Applied Statistics**, v. 47, p. 299-350, 1998.

Felgueiras, C. A. **Modelagem ambiental com tratamento de incertezas em sistemas de informação geográfica: o paradigma geoestatístico por indicação**. 171 p. (INPE-8180-TDI/760). Tese (Doutorado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos - SP, 1999.

Gawryszewski, V. P.; Jorge, M. H. P. M. Mortalidade Violenta no Município de São Paulo nos últimos 40 anos. **Revista Brasileira de Epidemiologia**, v. 3, n. 1-3, p. 50-69, 2000.

Goovaerts, P. **Geostatistics for natural resources evaluation**. New York: Oxford University Press, 1997. 483 p.

Goovaerts, P.; Jacquez, G. M.; Greiling, D. Exploring Scale-Dependent Correlations Between Cancer Mortality Rates Using Factorial Kriging and Population-Weighted Semivariograms. **Geographical Analysis**, v. 37, p. 152-182, 2005.

Gotway, C. A.; Young, L. J. Combining Incompatible Spatial Data. **Journal of American Statistical Association**, v. 97, p. 632-648, 2002.

Grauman, D. J.; Tarone, R. E.; Devesa, S. S.; Fraumeni Jr, J. F. Alternate ranging methods for cancer mortality maps. **Journal of the National Cancer Institute**, v. 92, n. 7, p. 534-543, 2000.

Harries, K. **Mapping crime:** principles and practice. Washington: National Institute of Justice (NIJ) of the U.S. Department of Justice, 1999. 195 p.

Isaaks, E. H., Srivastava R. M. **An introduction to applied geostatistics**. New York: Oxford University Press, 1989. 561 p.

Johnson, G. D. Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. **International Journal of Health Geographics**, v. 3, p. 29, 2004.

Journel, A. G. Non-parametric estimation of spatial distribution. **Mathematical Geology**, v. 15, n. 2, p. 445-468, 1983.

_____. **Fundamentals of geostatistics in five lessons**. Stanford: Center for Reservoir Forecasting Applied Earth Sciences Department, 1998.

Journel, A. G.; Huijbreghts, C. **Mining geostatistics**. New York: Academic Press, 1978. 600 p.

Kafadar, K. Choosing among two-dimensional smoothers in practice. **Computational Statistics and Data Analysis**, v. 18, n. 4, p. 419-439, 1994.

Kelsall, J.; Wakefield, J. Modelling spatial variation in disease risk: a geostatistical approach. **Journal of the American Statistical Association**, v. 97, p. 692-701, 2002.

Kyriakidis, P. C. A geostatistical framework for area-to-point spatial interpolation. **Geographical Analysis**, v. 36, n. 3, p. 259-289, 2004.

Lajaunie, C. **Local risk estimation for a rare noncontagious disease based on observed frequencies.** Note N-36/91/G. Fontainebleau: Center de Géostatistique, Ecole des Mines de Paris, 1991.

Lawson, A. B. Tutorial in biostatistics: Disease map reconstruction. **Statistics in Medicine**, v. 20, p. 2183-2204, 2001.

Leyland, A. H.; Davies, C. A. Empirical Bayes methods for disease mapping. **Statistical Methods in Medical Research**, v. 14, p. 17-34, 2005.

Lima, M. L. C.; Ximenes, R. A. A.; Souza, E. R.; Luna, C. F.; Albuquerque, M. F. P. M. Spatial analysis of socioeconomic determinants of homicide in Brazil. **Rev. Saúde Pública**, v. 39, n. 2, p. 176-182, 2005.

Maia, P. B. Vinte Anos de Homicídios No Estado de São Paulo. **Revista São Paulo em Perspectiva**, v. 13, n. 4, p. 121-129, 2000.

Marshall, R. J. Mapping disease and mortality rates using empirical Bayes estimators. **Applied Statistics**, v. 40, n. 2, p. 283-294, 1991.

Martuzzi, M.; Elliott, P. Empirical bayes estimation of small area prevalence of non-rare conditions. **Statistics in medicine**, v. 15, p. 1867-1873, 1996.

Matheron, G. Principles of geostatistics. **Economic Geology**, v. 58, n. 8, p. 1246-1266, 1963.

_____. **The theory of regionalized variables and its applications**. Paris: Les Cahiers du Centre de Morphologie Mathematique de Fontainebleu, 1970. 211 p.

Minayo, M. C. S. Violência social sob a perspectiva da saúde pública. **Cadernos de Saúde Pública**, v. 10, n. 1, p. 7-18, 1994.

Mungiole, M.; Pickle, L. W.; Hansen, S. K. Application of a weighted headbanging algorithm to mortality data maps. **Statistics in Medicine**, v. 18, p. 3201-3209, 1999.

Nery, M. B.; Monteiro, A. M. V. Análise intra-urbana dos homicídios dolosos no Município de São Paulo. In: Encontro Nacional de Estudos Populacionais, 15., 2006, Caxambu - MG**. Anais...** [S. l.]: ABEP 2006. p. 1-16. Disponível em: http://www.abep.org.br/usuario/GerenciaNavegacao.php?caderno_id=504&nivel=2&texto_id=2988. Accesso em: 12 Jul. 2006.

Oliver, M. A.; Webster, R.; Lajaunie, C.; Muir, K. R.; Parkes, S. E.; Cameron, A. H.; Stevens, M. C. G.; Mann, J. R. Binomial Cokriging for Estimating and Mapping the Risk of Childhood Cancer. **IMA Journal of Mathematics Applied in Medicine and Biology**, v. 15, p. 279-297, 1998.

Openshaw, S. Ecological fallacies and the analysis of areal census-data. **Environment and Planning A**, v. 16, p. 17-31, 1984.

Ortiz, J. O.; Felgueiras, C. A.; Druck, S.; Monteiro, A. M. V. Modelagem de fertilidade do solo por simulação estocástica com tratamento de incertezas. **Pesquisa Agropecuária Brasileira-PAB**, v. 39, n. 4, p. 379-389, 2004.

Pickle, L. W. Exploring spatio-temporal patterns of mortality using mixed effects models. **Statistics in Medicine**, v. 19, p. 2251-2263, 2000.

_____. Spatial analysis of disease. In: Beam, C. (Ed.). **Biostatistical applications in cancer research**. v. 7. Boston: Kluwer Academic Publishers, 2002, p. 113-150.

Prefeitura Municipal de São Paulo. PROAIM - **Programa de Aprimoramento das Informações de Mortalidade no Município de São Paulo -** Dados de homicídios. 2005. Disponível em: http://ww2.prefeitura.sp.gov.br//cgi/deftohtm.exe?secretarias/saude/TABNET/SIM/obito.def. Acesso em: 07/03/2005.

Rivoirard, J.; Simmonds, J.; Foote, K.; Fernandes, P.; Bez, N. **Geostatistics for estimating fish abundance**. Oxford, UK: Blackwell Science, 2000. 206 p.

Santos, S. M.; Barcellos, C.; Carvalho, M. S.; Flôres, R. Detecção de aglomerados espaciais de óbitos por causas violentas em Porto Alegre, Rio Grande do Sul, Brasil, 1996. **Caderno Saúde Pública**, v. 17, n. 5, p. 1141-1151, 2001.

Talbot, T. O.; Kulldorff, M.; Forand, S. P.; Haley, V. B. Evaluation of spatial filters to create smoothed maps of health data. **Statistics in Medicine**, v. 19, p. 2399-2408, 2000.

Tukey, J. W. **Exploratory data analysis**. Reading: Addison-Wesley Publishing Company, 1977. 506 p.

Vinhas, L.; Ferreira, K. R. Descrição da TerraLib. In: Casanova, M. A.; Câmara, G.; Davis Jr., C. A.; Vinhas, L.; Queiroz, G. R. (Ed.). **Bancos de dados geográficos**. Curitiba, PR: MundoGeo, 2005. cap. 12, p. 397-439.

Wakefield, J. A critique of statistical aspects of ecological studies in spatial epidemiology. **Environmental and Ecological Statistics**, v. 11, p. 31-54, 2004.

Waller, L. A.; Gotway, C. A. **Applied spatial statistics for public health data**. New Jersey: John Wiley and Sons, 2004. 494 p.

Webster, R.; Oliver, M. A.; Muir, K. R.; Mann, J. R. Kriging the local risk of a rare disease from a register of diagnoses. **Geographical Analysis**, v. 26, p. 168-185, 1994.

APÊNDICE A

DEDUÇÃO DO FORMALISMO PARA O SEMIVARIOGRAMA DO RISCO

Cálculo de $E[Z(\mathbf{u}_i)|R]$

$$E[Z(\mathbf{u}_i)|R] = E\left[\frac{L(\mathbf{u}_i)}{n(\mathbf{u}_i)}\right] = \frac{1}{n(\mathbf{u}_i)} E\left[L(\mathbf{u}_i)\right] = \frac{1}{n(\mathbf{u}_i)}\ n(\mathbf{u}_i).\, R(\mathbf{u}_i) = R(\mathbf{u}_i) \qquad (A.1)$$

Cálculo de $E\left[Z^2(\mathbf{u}_i)|R\right]$

Sabe-se que: $Var\left[Z(\mathbf{u}_i)|R\right] = E\left[Z^2(\mathbf{u}_i)|R\right] - E^2\left[Z(\mathbf{u}_i)|R\right]$

$$E\left[Z^2(\mathbf{u}_i)|R\right] = Var\left[Z(\mathbf{u}_i)|R\right] + E^2\left[Z(\mathbf{u}_i)|R\right]$$

$$E\left[Z^2(\mathbf{u}_i)|R\right] = Var\left[Z(\mathbf{u}_i)|R\right] + R^2(\mathbf{u}_i)$$

$$\begin{aligned} Var[Z(\mathbf{u}_i)\,|\,R(\mathbf{u}_i)] &= Var\left[\frac{L(\mathbf{u}_i)}{n(\mathbf{u}_i)}\right] = \frac{1}{n^2(\mathbf{u}_i)} Var[L(\mathbf{u}_i)] = \\ &= \frac{1}{n^2(\mathbf{u}_i)} n(\mathbf{u}_i) R(\mathbf{u}_i)[1 - R(\mathbf{u}_i)] = \frac{R(\mathbf{u}_i)[1 - R(\mathbf{u}_i)]}{n(\mathbf{u}_i)} \end{aligned} \qquad (A.2)$$

$$E\left[Z^2(\mathbf{u}_i)|R\right] = \frac{R(\mathbf{u}_i)[1 - R(\mathbf{u}_i)]}{n(\mathbf{u}_i)} + R^2(\mathbf{u}_i)$$

$$E\left[Z^2(\mathbf{u}_i)|R\right] = \frac{R(\mathbf{u}_i) - R^2(\mathbf{u}_i) + n(\mathbf{u}_i) R^2(\mathbf{u}_i)}{n(\mathbf{u}_i)}$$

$$E\left[Z^2(\mathbf{u}_i)|R\right] = \frac{R(\mathbf{u}_i) + R^2(\mathbf{u}_i)[n(\mathbf{u}_i) - 1]}{n(\mathbf{u}_i)}$$

$$E\left[Z^2(\mathbf{u}_i)|R\right] = \frac{R^2(\mathbf{u}_i)[n(\mathbf{u}_i) - 1]}{n(\mathbf{u}_i)} + \frac{R(\mathbf{u}_i)}{n(\mathbf{u}_i)}$$

$$E\left[Z^2(\mathbf{u}_i)|R\right] = \frac{n(\mathbf{u}_i) - 1}{n(\mathbf{u}_i)} R^2(\mathbf{u}_i) + \frac{R(\mathbf{u}_i)}{n(\mathbf{u}_i)} \qquad (A.3)$$

Cálculo de $E[Z(\mathbf{u}_i)\,Z(\mathbf{u}_j)\,|\,R]$

$$E[Z(\mathbf{u}_i)\,Z(\mathbf{u}_j)\,|\,R] = E[Z(\mathbf{u}_i)\,|\,R]\;E[Z(\mathbf{u}_j)\,|\,R] = R(\mathbf{u}_i)R(\mathbf{u}_j) \qquad (A.4)$$

A partir das relações (A.3) e (A.4) é possível estabelecer a diferença quadrática condicional de $E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2$, da seguinte forma:

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2 = E\{[Z^2(\mathbf{u}_i) - 2Z(\mathbf{u}_i)Z(\mathbf{u}_j) + Z^2(\mathbf{u}_j)]\,|\,R\}$$

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2 = E[Z^2(\mathbf{u}_i)\,|\,R] - 2E[Z(\mathbf{u}_i)Z(\mathbf{u}_j)\,|\,R] + E[Z^2(\mathbf{u}_j)\,|\,R] \qquad (A.5)$$

Substituindo na Equação (A.5) os termos $E\left[Z^2(\mathbf{u}_i)\,|\,R\right]$ e $E\left[Z^2(\mathbf{u}_j)\,|\,R\right]$ conforme definido na Equação (A.3), e o termo $E\left[Z(\mathbf{u}_i)Z(\mathbf{u}_j)\,|\,R\right]$ conforme definido na Equação (A.4), tem-se:

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2 = \frac{n(\mathbf{u}_i)-1}{n(\mathbf{u}_i)}R^2(\mathbf{u}_i) + \frac{R(\mathbf{u}_i)}{n(\mathbf{u}_i)} - 2R(\mathbf{u}_i)R(\mathbf{u}_j) + \frac{n(\mathbf{u}_j)-1}{n(\mathbf{u}_j)}R^2(\mathbf{u}_j) + \frac{R(\mathbf{u}_j)}{n(\mathbf{u}_j)}$$

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2 = \left(1 - \frac{1}{n(\mathbf{u}_i)}\right)R^2(\mathbf{u}_i) + \frac{R(\mathbf{u}_i)}{n(\mathbf{u}_i)} - 2R(\mathbf{u}_i)R(\mathbf{u}_j) + \left(1 - \frac{1}{n(\mathbf{u}_j)}\right)R^2(\mathbf{u}_j) + + \frac{R(\mathbf{u}_j)}{n(\mathbf{u}_j)}$$

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2 = R^2(\mathbf{u}_i) - \frac{R^2(\mathbf{u}_i)}{n(\mathbf{u}_i)} + \frac{R(\mathbf{u}_i)}{n(\mathbf{u}_i)} - 2R(\mathbf{u}_i)R(\mathbf{u}_j) + R^2(\mathbf{u}_j) - \frac{R^2(\mathbf{u}_j)}{n(\mathbf{u}_j)} + + \frac{R(\mathbf{u}_j)}{n(\mathbf{u}_j)}$$

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2 = \frac{R(\mathbf{u}_i)}{n(\mathbf{u}_i)} - \frac{R^2(\mathbf{u}_i)}{n(\mathbf{u}_i)} + R^2(\mathbf{u}_i) - 2R(\mathbf{u}_i)R(\mathbf{u}_j) + R^2(\mathbf{u}_j) + \frac{R(\mathbf{u}_j)}{n(\mathbf{u}_j)} - - \frac{R^2(\mathbf{u}_j)}{n(\mathbf{u}_j)}$$

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2 = \frac{R(\mathbf{u}_i)}{n(\mathbf{u}_i)}\left[1 - R(\mathbf{u}_i)\right] + \left[R(\mathbf{u}_i) - R(\mathbf{u}_j)\right]^2 + \frac{R(\mathbf{u}_j)}{n(\mathbf{u}_j)}\left[1 - R(\mathbf{u}_j)\right]$$

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2 = \left[R(\mathbf{u}_i) - R(\mathbf{u}_j)\right]^2 + \frac{1}{n(\mathbf{u}_i)} R(\mathbf{u}_i)\left[1 - R(\mathbf{u}_i)\right] + \frac{1}{n(\mathbf{u}_j)} R(\mathbf{u}_j)\left[1 - R(\mathbf{u}_j)\right]$$

Aplicando o operador esperança na igualdade acima, tem-se:

$$E\left\{E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2\right\} = E\left\{\left[R(\mathbf{u}_i) - R(\mathbf{u}_j)\right]^2 + \frac{1}{n(\mathbf{u}_i)} R(\mathbf{u}_i)\left[1 - R(\mathbf{u}_i)\right] + \right.$$
$$\left. + \frac{1}{n(\mathbf{u}_j)} R(\mathbf{u}_j)\left[1 - R(\mathbf{u}_j)\right]\right\}$$

Por definição $E\left\{E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\,|\,R\right]^2\right\} = E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\right]^2$, substituindo tem-se:

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\right]^2 = E\left[R(\mathbf{u}_i) - R(\mathbf{u}_j)\right]^2 + \frac{1}{n(\mathbf{u}_i)} E\left\{R(\mathbf{u}_i)\left[1 - R(\mathbf{u}_i)\right]\right\} +$$
$$+ \frac{1}{n(\mathbf{u}_j)} E\left\{R(\mathbf{u}_j)\left[1 - R(\mathbf{u}_j)\right]\right\}$$

$$E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\right]^2 = E\left[R(\mathbf{u}_i) - R(\mathbf{u}_j)\right]^2 + \frac{E[R(\mathbf{u}_i)]}{n(\mathbf{u}_i)} - \frac{E[R^2(\mathbf{u}_i)]}{n(\mathbf{u}_i)} + \frac{E[R(\mathbf{u}_j)]}{n(\mathbf{u}_j)} - \frac{E[R^2(\mathbf{u}_j)]}{n(\mathbf{u}_j)}$$

Dividindo por 2 a igualdade acima, tem-se:

$$\frac{1}{2} E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\right]^2 = \frac{1}{2} E\left[R(\mathbf{u}_i) - R(\mathbf{u}_j)\right]^2 + \frac{E[R(\mathbf{u}_i)]}{2n(\mathbf{u}_i)} - \frac{E[R^2(\mathbf{u}_i)]}{2n(\mathbf{u}_i)} + \frac{E[R(\mathbf{u}_j)]}{2n(\mathbf{u}_j)} - \frac{E[R^2(\mathbf{u}_j)]}{2n(\mathbf{u}_j)}$$

Na expressão acima os termos:

1) $\frac{1}{2} E\left[Z(\mathbf{u}_i) - Z(\mathbf{u}_j)\right]^2 = \gamma^Z_{(\mathbf{u}_i, \mathbf{u}_j)}$;

2) $\frac{1}{2} E\left[R(\mathbf{u}_i) - R(\mathbf{u}_j)\right]^2 = \gamma^R_{(\mathbf{u}_i, \mathbf{u}_j)}$;

3) $E[R(\mathbf{u}_i)] = \mu$. Sabe-se que: $E\left[Z(\mathbf{u}_i)\right] = E\left\{E\left[Z(\mathbf{u}_i)\,|\,R\right]\right\}$ e $E\left[Z(\mathbf{u}_i)\,|\,R\right] = R(\mathbf{u}_i)$ conforme definido na Equação (A.1), então $E\left[Z(\mathbf{u}_i)\right] = E\left[R(\mathbf{u}_i)\right] = \mu$.

Substituindo os termos acima, tem-se:

$$\gamma^Z_{(\mathbf{u}_i, \mathbf{u}_j)} = \gamma^R_{(\mathbf{u}_i, \mathbf{u}_j)} + \frac{\mu}{2n(\mathbf{u}_i)} - \frac{E[R^2(\mathbf{u}_i)]}{2n(\mathbf{u}_i)} + \frac{\mu}{2n(\mathbf{u}_j)} - \frac{E[R^2(\mathbf{u}_j)]}{2n(\mathbf{u}_j)}$$

Sabe-se que: $E\left[R^2(\mathbf{u}_i)\right] = Var\left[R(\mathbf{u}_i)\right] + E^2\left[R(\mathbf{u}_i)\right]$, substituindo tem-se:

$$\gamma^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} = \gamma^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} + \frac{\mu}{2n(\mathbf{u}_i)} - \frac{1}{2n(\mathbf{u}_i)}\left\{Var[R(\mathbf{u}_i)] + E^2[R(\mathbf{u}_i)]\right\} + \\ + \frac{\mu}{2n(\mathbf{u}_j)} - \frac{1}{2n(\mathbf{u}_j)}\left\{Var[R(\mathbf{u}_j)] + E^2[R(\mathbf{u}_j)]\right\}$$

$$\gamma^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} = \gamma^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} + \frac{\mu}{2n(\mathbf{u}_i)} - \frac{Var[R(\mathbf{u}_i)]}{2n(\mathbf{u}_i)} - \frac{\mu^2}{2n(\mathbf{u}_i)} + \frac{\mu}{2n(\mathbf{u}_j)} - \frac{Var[R(\mathbf{u}_j)]}{2n(\mathbf{u}_j)} - \frac{\mu^2}{2n(\mathbf{u}_j)}$$

$$\gamma^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} = \gamma^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} + \frac{\mu}{2n(\mathbf{u}_i)} + \frac{\mu}{2n(\mathbf{u}_j)} - \frac{Var[R(\mathbf{u}_i)]}{2n(\mathbf{u}_i)} - \frac{\mu^2}{2n(\mathbf{u}_i)} - \frac{Var[R(\mathbf{u}_j)]}{2n(\mathbf{u}_j)} - \frac{\mu^2}{2n(\mathbf{u}_j)}$$

$$\gamma^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} = \gamma^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} + \frac{\mu}{2n(\mathbf{u}_i)} + \frac{\mu}{2n(\mathbf{u}_j)} - \frac{\mu^2}{2n(\mathbf{u}_i)} - \frac{\mu^2}{2n(\mathbf{u}_j)} - \frac{Var[R(\mathbf{u}_i)]}{2n(\mathbf{u}_i)} - \frac{Var[R(\mathbf{u}_j)]}{2n(\mathbf{u}_j)}$$

$$\gamma^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} = \gamma^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} + \frac{\mu}{2n(\mathbf{u}_i)} + \frac{\mu}{2n(\mathbf{u}_j)} - \frac{\mu^2}{2n(\mathbf{u}_i)} - \frac{\mu^2}{2n(\mathbf{u}_j)} - \frac{1}{2}\left\{\frac{Var[R(\mathbf{u}_i)]}{n(\mathbf{u}_i)} + \frac{Var[R(\mathbf{u}_j)]}{n(\mathbf{u}_j)}\right\}$$

$$\gamma^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} = \gamma^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} + \frac{1}{2}\left\{\frac{1}{n(\mathbf{u}_i)} + \frac{1}{n(\mathbf{u}_j)}\right\}\mu(1-\mu) - \frac{1}{2}\left\{\frac{Var[R(\mathbf{u}_i)]}{n(\mathbf{u}_i)} + \frac{Var[R(\mathbf{u}_j)]}{n(\mathbf{u}_j)}\right\}$$

$$\gamma^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} = \gamma^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} + \frac{1}{2}\left\{\frac{1}{n(\mathbf{u}_i)} + \frac{1}{n(\mathbf{u}_j)}\right\}\mu(1-\mu) - \frac{1}{2}\left\{\frac{\sigma^2_{R(\mathbf{u}_i)}}{n(\mathbf{u}_i)} + \frac{\sigma^2_{R(\mathbf{u}_j)}}{n(\mathbf{u}_j)}\right\}$$

Finalmente a expressão teórica do semivariograma do risco, conforme Lajaunie (1991):

$$\gamma^{R}_{(\mathbf{u}_i,\mathbf{u}_j)} = \gamma^{Z}_{(\mathbf{u}_i,\mathbf{u}_j)} - \frac{1}{2}\left\{\frac{1}{n(\mathbf{u}_i)} + \frac{1}{n(\mathbf{u}_j)}\right\}\mu(1-\mu) + \frac{1}{2}\left\{\frac{\sigma^2_{R(\mathbf{u}_i)}}{n(\mathbf{u}_i)} + \frac{\sigma^2_{R(\mathbf{u}_j)}}{n(\mathbf{u}_j)}\right\}$$

APÊNDICE B

TAXA DE HOMICÍDIO NA CIDADE DE SÃO PAULO EM 2002

TABELA B.1 - Taxa de homicídio na cidade de São Paulo em 2002.

No.	Distrito	Número de homicídios	População em risco	Taxa de homicídios por 100mil habitantes
1	Água Rasa	20	84358	23.71
2	Alto de Pinheiros	5	43524	11.49
3	Anhangüera	5	44812	11.16
4	Aricanduva	42	94653	44.37
5	Artur Alvim	51	110328	46.23
6	Barra Funda	3	12547	23.91
7	Bela Vista	7	61837	11.32
8	Belém	9	38259	23.52
9	Bom Retiro	10	25407	39.36
10	Brás	25	24095	103.76
11	Brasilândia	154	253309	60.80
12	Butantã	15	51740	28.99
13	Cachoeirinha	120	150290	79.85
14	Cambuci	8	27620	28.96
15	Campo Belo	18	64968	27.71
16	Campo Grande	31	92341	33.57
17	Campo Limpo	105	195582	53.69
18	Cangaíba	41	140245	29.23
19	Capão Redondo	203	247096	82.15
20	Carrão	12	76681	15.65
21	Casa Verde	26	81683	31.83
22	Cidade Ademar	165	244125	67.59
23	Cidade Dutra	141	193898	72.72
24	Cidade Líder	48	119300	40.23
25	Cidade Tiradentes	105	208704	50.31
26	Consolação	8	52836	15.14
27	Cursino	34	100581	33.80
28	Ermelino Matarazzo	46	108024	42.58
29	Freguesia do Ó	43	144004	29.86
30	Grajaú	330	357903	92.20
31	Guaianazes	118	100748	117.12
32	Iguatemi	84	109074	77.01
33	Ipiranga	21	98578	21.30
34	Itaim Bibi	7	78117	8.96
35	Itaim Paulista	107	219688	48.71

(continua)

TABELA B.1 (continuação)

No.	Distrito	Número de homicídios	População em risco	Taxa de homicídios por 100mil habitantes
36	Itaquera	109	204530	53.29
37	Jabaquara	109	214221	50.88
38	Jaçanã	36	92128	39.08
39	Jaguará	6	25098	23.91
40	Jaguaré	12	42259	28.40
41	Jaraguá	79	154556	51.11
42	Jardim Angela	232	255815	90.69
43	Jardim Helena	58	141623	40.95
44	Jardim Paulista	1	80981	1.23
45	Jardim São Luís	220	243368	90.40
46	José Bonifácio	38	107103	35.48
47	Lajeado	86	164536	52.27
48	Lapa	6	58676	10.23
49	Liberdade	16	59891	26.72
50	Limão	41	80628	50.85
51	Mandaqui	24	103149	23.27
52	Marsilac	5	8768	57.03
53	Moema	6	70193	8.55
54	Mooca	12	61917	19.38
55	Morumbi	12	33763	35.54
56	Parelheiros	126	111498	113.01
57	Pari	5	14054	35.58
58	Parque do Carmo	46	65190	70.56
59	Pedreira	77	133955	57.48
60	Penha	34	123240	27.59
61	Perdizes	8	101677	7.87
62	Perus	31	74620	41.54
63	Pinheiros	13	60872	21.36
64	Pirituba	68	162462	41.86
65	Ponte Rasa	30	97576	30.75
66	Raposo Tavares	32	92021	34.77
67	República	19	46294	41.04
68	Rio Pequeno	34	112571	30.20
69	Sacomã	98	229765	42.65
70	Santa Cecília	22	69098	31.84
71	Santana	14	122461	11.43
72	Santo Amaro	33	58495	56.42
73	São Domingos	35	84350	41.49
74	São Lucas	41	137110	29.90
75	São Mateus	99	155533	63.65
76	São Miguel	80	96706	82.72

(continua)

TABELA B.1 – (conclusão)

No.	Distrito	Número de homicídios	População em risco	Taxa de homicídios por 100mil habitantes
77	São Rafael	81	130365	62.13
78	Sapopemba	153	284600	53.76
79	Saúde	23	117047	19.65
80	Sé	14	19229	72.81
81	Socorro	7	38407	18.23
82	Tatuapé	11	79110	13.90
83	Tremembé	51	169281	30.13
84	Tucuruvi	25	97366	25.68
85	Vila Andrade	34	79110	42.98
86	Vila Curuçá	96	149197	64.34
87	Vila Formosa	28	93375	29.99
88	Vila Guilherme	17	48377	35.14
89	Vila Jacuí	50	148097	33.76
90	Vila Leopoldina	12	26894	44.62
91	Vila Maria	46	112224	40.99
92	Vila Mariana	7	121977	5.74
93	Vila Matilde	26	102206	25.44
94	Vila Medeiros	59	137988	42.76
95	Vila Prudente	27	100126	26.97
96	Vila Sônia	21	87629	23.96

FONTE: São Paulo. PROAIM (2005)

APÊNDICE C

TAXA DE HOMICÍDIO NA CIDADE DE SÃO PAULO EM 2003

TABELA C.1 - Taxa de homicídio na cidade de São Paulo em 2003.

No.	Distrito	Número de homicídios	População em risco	Taxa de homicídios por 100mil habitantes
1	Água Rasa	17	83521	20.35
2	Alto de Pinheiros	1	43019	2.32
3	Anhangüera	22	48619	45.25
4	Aricanduva	22	94520	23.28
5	Artur Alvim	35	109804	31.87
6	Barra Funda	4	12327	32.45
7	Bela Vista	12	61106	19.64
8	Belém	13	37537	34.63
9	Bom Retiro	13	24785	52.45
10	Brás	20	23537	84.97
11	Brasilândia	152	256468	59.27
12	Butantã	10	51244	19.51
13	Cachoeirinha	94	151663	61.98
14	Cambuci	5	27044	18.49
15	Campo Belo	23	64068	35.90
16	Campo Grande	19	92830	20.47
17	Campo Limpo	84	197710	42.49
18	Cangaíba	36	141715	25.40
19	Capão Redondo	154	250435	61.49
20	Carrão	11	75870	14.50
21	Casa Verde	33	80635	40.93
22	Cidade Ademar	142	244441	58.09
23	Cidade Dutra	138	195178	70.70
24	Cidade Líder	46	120590	38.15
25	Cidade Tiradentes	119	218937	54.35
26	Consolação	5	51940	9.63
27	Cursino	34	99750	34.09
28	Ermelino Matarazzo	42	108623	38.67
29	Freguesia do Ó	42	143440	29.28
30	Grajaú	326	371539	87.74
31	Guaianazes	92	101906	90.28
32	Iguatemi	64	113134	56.57
33	Ipiranga	30	98376	30.50
34	Itaim Bibi	5	76364	6.55
35	Itaim Paulista	128	223412	57.29

(continua)

TABELA C.1 – (continuação)

No.	Distrito	Número de homicídios	População em risco	Taxa de homicídios por 100mil habitantes
36	Itaquera	106	206088	51.43
37	Jabaquara	71	214176	33.15
38	Jaçanã	39	92264	42.27
39	Jaguará	6	24766	24.23
40	Jaguaré	8	42120	18.99
41	Jaraguá	89	159324	55.86
42	Jardim Angela	212	261229	81.15
43	Jardim Helena	59	142935	41.28
44	Jardim Paulista	5	79554	6.29
45	Jardim São Luís	188	245554	76.56
46	José Bonifácio	47	107076	43.89
47	Lajeado	99	168202	58.86
48	Lapa	5	57867	8.64
49	Liberdade	18	58838	30.59
50	Limão	44	79852	55.10
51	Mandaqui	19	103114	18.43
52	Marsilac	6	8965	66.93
53	Moema	1	69597	1.44
54	Mooca	16	61181	26.15
55	Morumbi	19	33320	57.02
56	Parelheiros	94	116370	80.78
57	Pari	7	13655	51.26
58	Parque do Carmo	42	65774	63.86
59	Pedreira	72	137524	52.35
60	Penha	27	122619	22.02
61	Perdizes	7	101218	6.92
62	Perus	32	76780	41.68
63	Pinheiros	8	59743	13.39
64	Pirituba	62	162760	38.09
65	Ponte Rasa	35	97240	35.99
66	Raposo Tavares	24	92427	25.97
67	República	21	45536	46.12
68	Rio Pequeno	24	112968	21.24
69	Sacomã	102	230476	44.26
70	Santa Cecília	16	67989	23.53
71	Santana	21	121266	17.32
72	Santo Amaro	15	57413	26.13
73	São Domingos	19	85140	22.32
74	São Lucas	32	135892	23.55
75	São Mateus	90	155817	57.76
76	São Miguel	52	96302	54.00

(continua)

TABELA C.1 – (conclusão)

No.	Distrito	Número de homicídios	População em risco	Taxa de homicídios por 100mil habitantes
77	São Rafael	79	133222	59.30
78	Sapopemba	150	285765	52.49
79	Saúde	15	116441	12.88
80	Sé	11	18764	58.62
85	Vila Andrade	40	82156	48.69
86	Vila Curuçá	64	150609	42.49
87	Vila Formosa	25	93074	26.86
88	Vila Guilherme	5	47525	10.52
89	Vila Jacuí	68	151426	44.91
90	Vila Leopoldina	7	26894	26.03
91	Vila Maria	48	111329	43.12
92	Vila Mariana	12	121033	9.91
93	Vila Matilde	17	101767	16.70
94	Vila Medeiros	67	136585	49.05
95	Vila Prudente	31	99051	31.30
96	Vila Sônia	25	87731	28.50

FONTE: São Paulo. PROAIM (2005)

APÊNDICE D

TAXA DE HOMICÍDIO NA CIDADE DE SÃO PAULO EM 2004

TABELA D.1 - Taxa de homicídio na cidade de São Paulo em 2004.

No.	Distrito	Número de homicídios	População em risco	Taxa de homicídios por 100mil habitantes
1	Água Rasa	8	82668	9.68
2	Alto de Pinheiros	2	42509	4.70
3	Anhangüera	16	52735	30.34
4	Aricanduva	22	94359	23.32
5	Artur Alvim	31	109251	28.38
6	Barra Funda	3	12106	24.78
7	Bela Vista	12	60367	19.88
8	Belém	9	36820	24.44
9	Bom Retiro	10	24172	41.37
10	Brás	21	22986	91.36
11	Brasilândia	153	259596	58.94
12	Butantã	6	50737	11.83
13	Cachoeirinha	75	153009	49.02
14	Cambuci	8	26472	30.22
15	Campo Belo	16	63162	25.33
16	Campo Grande	17	93296	18.22
17	Campo Limpo	80	199806	40.04
18	Cangaíba	38	143158	26.54
19	Capão Redondo	127	253752	50.05
20	Carrão	16	75047	21.32
21	Casa Verde	15	79578	18.85
22	Cidade Ademar	123	244692	50.27
23	Cidade Dutra	103	196416	52.44
24	Cidade Líder	47	121860	38.57
25	Cidade Tiradentes	64	229606	27.87
26	Consolação	1	51046	1.96
27	Cursino	21	98899	21.23
28	Ermelino Matarazzo	36	109195	32.97
29	Freguesia do Ó	36	142841	25.20
30	Grajaú	269	385578	69.77
31	Guaianazes	60	103049	58.22
32	Iguatemi	50	117314	42.62
33	Ipiranga	18	98146	18.34
34	Itaim Bibi	7	74630	9.38
35	Itaim Paulista	66	227137	29.06

(continua)

TABELA D.1 - (continuação)

No.	Distrito	Número de homicídios	População em risco	Taxa de homicídios por 100mil habitantes
36	Itaquera	75	207598	36.13
37	Jabaquara	74	214074	34.57
38	Jaçanã	37	92377	40.05
39	Jaguará	4	24432	16.37
40	Jaguaré	11	41970	26.21
41	Jaraguá	71	164193	43.24
42	Jardim Angela	151	266682	56.62
43	Jardim Helena	58	144220	40.22
44	Jardim Paulista	3	78133	3.84
45	Jardim São Luís	132	247692	53.29
46	José Bonifácio	25	107020	23.36
47	Lajeado	49	171901	28.50
48	Lapa	2	57053	3.51
49	Liberdade	17	57789	29.42
50	Limão	23	79065	29.09
51	Mandaqui	18	103049	17.47
52	Marsilac	4	9165	43.64
53	Moema	2	68988	2.90
54	Mooca	14	60437	23.16
55	Morumbi	9	32875	27.38
56	Parelheiros	81	121422	66.71
57	Pari	7	13264	52.77
58	Parque do Carmo	34	66345	51.25
59	Pedreira	63	141149	44.63
60	Penha	21	121967	17.22
61	Perdizes	6	100733	5.96
62	Perus	50	78978	63.31
63	Pinheiros	3	58623	5.12
64	Pirituba	28	163014	17.18
65	Ponte Rasa	27	96877	27.87
66	Raposo Tavares	22	92809	23.70
67	República	14	44779	31.26
68	Rio Pequeno	27	113336	23.82
69	Sacomã	53	231128	22.93
70	Santa Cecília	11	66881	16.45
71	Santana	13	120050	10.83
72	Santo Amaro	13	56336	23.08
73	São Domingos	21	85913	24.44
74	São Lucas	35	134646	25.99
75	São Mateus	52	156060	33.32
76	São Miguel	53	95874	55.28

(continua)

TABELA D.1 - (conclusão)

No.	Distrito	Número de homicídios	População em risco	Taxa de homicídios por 100mil habitantes
77	São Rafael	59	136104	43.35
78	Sapopemba	79	286857	27.54
79	Saúde	11	115806	9.50
80	Sé	12	18307	65.55
85	Vila Andrade	43	85295	50.41
86	Vila Curuçá	54	151994	35.53
87	Vila Formosa	20	92749	21.56
88	Vila Guilherme	17	46675	36.42
89	Vila Jacuí	41	154786	26.49
90	Vila Leopoldina	5	26887	18.60
91	Vila Maria	39	110411	35.32
92	Vila Mariana	6	120064	5.00
93	Vila Matilde	24	101302	23.69
94	Vila Medeiros	43	135158	31.81
95	Vila Prudente	21	97961	21.44
96	Vila Sônia	19	87810	21.64

FONTE: São Paulo. PROAIM (2005)

APÊNDICE E

DISTRITOS DA CIDADE DE SÃO PAULO

01- Anhanguera
02- Perus
03- Jaragua
04- Brasilândia
05- Cachoerinha
06- Mandaqui
07- Tremembé
08- Tucuruvi
09- Jaçanã
10- São Domingo
11- Pirituba
12- Freguesia do Ó
13- Limão
14- Casa Verde
15- Santana
16- Vila Guilherme
17- Vila Medeiros
18- Vila Maria
19- Penha
20- Cangaiba
21- Ponte Rasa
22- Ermelino Matarazo
23- Vila Jacui
24- Itaquera
25- São Miguel
26- Vila Curuçá
27- Jardim Helena
28- Itaim Paulista
29- Jaguará
30- Vila Leopoldina
31- Lapa
32- Barra Funda
33- Santa Cecília
34- Bom Retiro
35- Parí
36- Brás
37- Belem
38- Mooca
39- Água Rasa
40- Tatuapé
41- Carrão
42- Vila Matilde
43- Vila Formosa
44- Aricanduva
45- Cidade Lider
46- Artur Alvim
47- Parque do Carmo
48- José Bonifácio
49- Lajeado
50- Guaianazes
51- Cidade Tiradentes
52- Iguatemi
53- São Rafael
54- São Mateus
55- Sapopemba
56- São lucas
57- Vila Prudente
58- Ipiranga
59- Cambuci
60- Liberdade
61- Sé
62- República
63- Consolação
64- Bela Vista
65- Jardim Paulista
66- Pinheiros
67- Perdizes
68- Alto de Pinheiros
69- Jaguaré
70- Rio Pequeno
71- Butantã
72- Raposo Tavares
73- Vila Sônia
74- Morumbi
75- Itaim Bibi
76- Moema
77- Vila Mariana
78- Saúde
79- Cursino
80- Sacomã
81- Campo Limpo
82- Vila Andrade
83- Santo Amaro
84- Campo Belo
85- Jabaquara
86- Capão Redondo
87- Jardim São Luis
88- Socorro
89- Campo Grande
90- Cidade Ademar
91- Jardim Ângela
92- Cidade Dutra
93- Pedreira
94- Parelheiros
95- Grajaú
96- Marsilac

FIGURA E.1 - Distritos da cidade de São Paulo.

APÊNDICE F

SÍNTESE DO FORMALISMO POR INDICAÇÃO PARA CONSTRUÇÃO DA FUNÇÃO DE DISTRIBUIÇÃO ACUMULADA CONDICIONADA - FDAC

O formalismo por indicação para variáveis contínuas possibilita estimar um conjunto de valores, em uma localização não amostrada, que representa uma aproximação discretizada da Função de Distribuição Acumulada Condicionada, FDAC, às *k* amostras vizinhas [Journel (1983), Goovaerts (1997) e Felgueiras (1999)].

O primeiro passo, no formalismo por Indicação, é transformar os dados originais em indicadores de probabilidades. Isto é realizado da seguinte forma: considere o conjunto de amostras {r($\mathbf{u}_j$), j = 1, ..., *N* } da V.A. R($\mathbf{u}_j$). Se a variável R($\mathbf{u}_j$) for transformada numa variável indicadora $I(\mathbf{u}_j; r_c)$ com base em um valor de corte r_c, tem-se:

$$I(\mathbf{u}_j; r_c) = \begin{cases} 1 \text{ se } R(\mathbf{u}_j) \le r_c \\ 0 \text{ se } R(\mathbf{u}_j) > r_c \end{cases} \qquad \text{(F.1)}$$

Essa transformação equivale associar a probabilidade 1 (100%) para os valores de R($\mathbf{u}_j$) que são ≤ r_c e 0 caso contrário. O resultado da transformação, expressa em (F.1), é um novo conjunto de dados, composto de 0 e 1, denotado por {$i(\mathbf{u}_j)$, j = 1, ..., *N*} da V.A. $I(\mathbf{u}_j; r_c)$. A partir deste conjunto de dados emprega-se o semivariograma por indicação para a análise da dependência espacial da V.A. $I(\mathbf{u}_j; r_c)$. O estimador para o semivariograma por indicação é definido como [Deutsch e Journel (1998)]:

$$\hat{\gamma}^{I}_{(\mathbf{h}, r_c)} = \frac{1}{2M(\mathbf{h})} \sum_{j=1}^{M(\mathbf{h})} \left[i(\mathbf{u}_j; r_c) - i(\mathbf{u}_j + \mathbf{h}; r_c) \right]^2 \qquad \text{(F.2)}$$

em que: **h** é o vetor distância entre dois pares de pontos; r_c é o valor de corte pré-estabelecido e M(**h**) refere-se ao número de pares de pontos que estão distanciados de **h**.

Como resultado desta análise, tem-se, um modelo teórico de semivariograma, que reflete a continuidade espacial da V.A. $I(\mathbf{u}_j; r_c)$ para o valor de corte pré-estabelecido.

O próximo passo é estimar a V.A. $I(\mathbf{u}_0;\ r_c)$, em uma localização $\mathbf{u}_0$ não amostrada, condicionada às k amostras vizinhas. Em geral, este passo é realizado efetuando-se a krigeagem simples ou ordinária. Como resultado, tem-se, um valor estimado $[i(\mathbf{u}_0;\ r_c) \mid (k)]^*$ entre 0 e 1. Este resultado corresponde à probabilidade de que a V.A. $\mathrm{R}(\mathbf{u}_0)$, na localização $\mathbf{u}_0$, seja menor ou igual ao nível de corte pré-estabelecido. Mais especificamente, neste caso, o que a krigeagem fornece é (Deutsch e Journel (1998)):

$$[i(\mathbf{u}_0; r_c \mid (k)]^* = \mathrm{E}[I(\mathbf{u}_0; r_c) \mid (k)]^* = [Prob\{\mathrm{R}(\mathbf{u}_0) \leq r_c) \mid (k)\}]^* \qquad \text{(F.3)}$$

À medida que se incrementa r_c, obter-se-á outros valores estimados da V.A. $I(\mathbf{u}_0;\ r_c)$. Isto resulta num conjunto de estimativas, $\{[i(\mathbf{u}_0;\ r_c)]^*,\ r_c = 1,...,\ N^{o}\ Cortes\}$, que representa uma aproximação discretizada da FDAC de $\mathrm{R}(\mathbf{u}_0)$, $F[\mathbf{u}_0,\ \mathrm{r} \mid (k)] = Prob\{\mathrm{R}(\mathbf{u}_0) \leq r_c \mid (k)\}$. A Figura F.1 ilustra um exemplo para 4 valores de corte, supondo o mesmo peso para as k amostras vizinhas (k=5) sobre a localização $\mathbf{u}_0$ a estimar. As localizações $\mathbf{u}_j$, j = 1, ..., N (N=9) e $\mathbf{u}_0$ representam sempre a mesma posição geográfica no terreno.

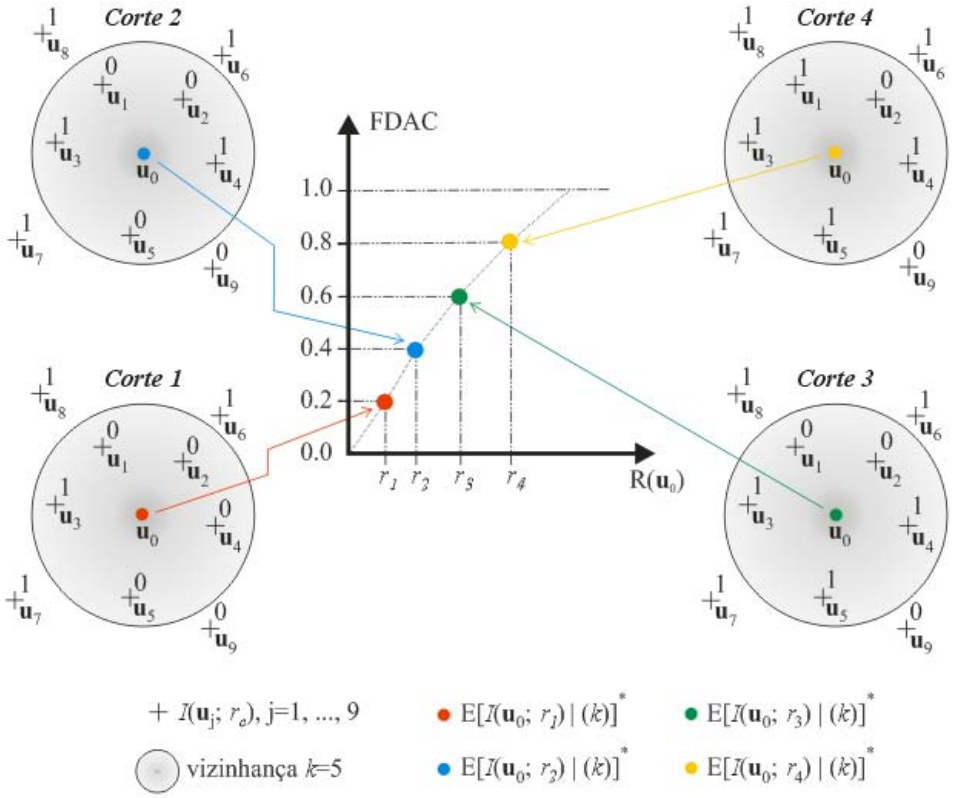


FIGURA F.1 - Exemplo ilustrativo do formalismo por indicação para 4 valores de corte.

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programa de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. São aceitos tanto programas fonte quanto executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.