Universidade Federal de Minas Gerais

Modelos Matemáticos de Propagação de Epidemias Baseados em Redes Sociais e Detecção de Clusters de Doenças

Tese apresentada ao Departamento de Engenharia Elétrica da Universidade Federal de Minas Gerais
como requisito final para a obtenção do título de
Doutor em Engenharia Elétrica.

Aluno: **Alexandre Celestino Leite Almeida**

Orientador: **Ricardo Hiroshi Caldeira Takahashi (MAT/UFMG)**

Co-Orientador: **Luiz Henrique Duczmal (EST/UFMG)**

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da FAPEMIG/PMCD.

Belo Horizonte, setembro de 2011

Tese de doutorado defendida e aprovada em 16 de setembro de 2011, pela Banca Examinadora constituída por:

Prof. Dr. Ricardo Hiroshi Caldeira Takahashi
Orientador

Prof. Dr. Luiz Henrique Duczmal
Co-orientador

Prof. Dr. André Luiz Fernandes Cançado

Prof. Dr. Eduardo Mazoni Andrade Marçal Mendes

Prof. Dr. Frederico Rodrigues Borges da Cruz

Prof. Dr. Marcelo Martins de Oliveira

Dedico esta tese

à Deus,
à minha esposa Petrusca
e à minha mãe Sara.
Pelo apoio incondicional,
força e incentivo.

Agradecimentos

Agradeço a Deus, por uma vida cheia de oportunidades,
pela extrema generosidade e por me permitir compreender algumas coisas;

à minha esposa Petrusca, pelo amor, paciência, apoio e dedicação;

à minha família, em especial à minha mãe Sara, por sua dedicação e apoio;

ao meu orientador, o professor Ricardo Takahashi, pela sabedoria, apoio, dedicação, amizade e grande paciência;

ao meu co-orientador, o professor Luiz Duczmal, pela sabedoria, amizade, apoio, dedicação e por me deixar ganhar no boliche;

aos amigos Anderson Duarte, Fernando Oliveira, Emerson Bodevan, Spencer Barbosa, Flávio Moura, João e Juliana Zuliani, André Cruz, Luciana Rocha, Fabrício e Camila Ceolin, Werley Facco, Alex Moura... e já me desculpo pelos que possa ter esquecido;

aos colegas, técnicos e professores do Campus Alto Paraopeba da UFSJ,
em especial aos companheiros do DeFiM pelo grande apoio;

aos colegas e professores do PPGEE da UFMG, em especial aos companheiros do GOPAC pelas discussões científicas e momentos de lazer;

aos colegas do grupo de otimização, pelos ótimos encontros, excelentes discussões e pela amizade;

à banca examinadora deste trabalho, André Cançado (UnB), Eduardo Mazoni (UFMG), Frederico Cruz (UFMG) e Marcelo Oliveira (UFSJ) por terem paciência e dedicação em ler este trabalho, pelas sugestões, pelos elogios e em especial, é claro, por terem aprovado esta tese;

ao professor Martin Kulldorff, pelas críticas e sugestões;

e à Fapemig, pelo apoio financeiro.

"Se enxerguei mais longe é porque
me apoiei em ombros de gigantes."
Isaac Newton

Resumo

Estudos para compreender melhor epidemias são atualmente feitos em três áreas distintas: a matemática com modelos de equações diferenciais, a física com seus modelos de propagação em redes e a estatística na detecção de clusters. Apesar do objetivo comum, estas áreas tem pouca ou nenhuma intersecção, o que torna o conhecimento no assunto um pouco fragmentado. Nossos estudos são no sentido de buscar ferramentas na intersecção das áreas. Simulações computacionais de epidemias em cenários hipotéticos são ferramentas valiosas para entender e predizer o comportamento de epidemias. Desta forma, pesquisadores tem estudado modelos matemáticos capazes de descrever a dinâmica de epidemias em diversos cenários. Doenças de transmissão direta pessoa a pessoa (tais como gripe, HIV, varíola, etc.) dependem de como os indíviduos iteragem entre si. Além disso, doenças infecciosas se comportam de formas diferentes em populações com estruturas sociais diferentes. Modelos matemáticos de equações diferenciais não levam em consideração estes aspectos e, por isso, propomos um novo modelo baseado em indivíduos (MBI), de forma a considerar diferentes organizações da sociedade como redes complexas, cujos estudos indicaram dificuldades no ajuste dos parâmetros do modelo SIR para redes. Propomos, então, um modelo de equações diferenciais, o μSIR, capaz descrever o modelo SIR em uma rede qualquer. O μSIR necessita de toda a informação da rede, o que pode inviabilizar seu uso do ponto de vista prático. Devido a isso, desenvolvemos o modelo HMF-MC, um modelo multi-comunidades cujas informações necessárias são as frequências de conectividades em cada comunidade e o número de arestas entre elas. Detectar rapidamente conglomerados espaciais, ou espaço-temporais, de casos ou sintomas de epidemias é de grande utilidade para que os órgãos públicos de saúde possam agir rapidamente e controlar surtos epidêmicos. Apresentamos, então, as estatísticas: espacial *scan* de Kulldorf e espacial para fluxo de indívíduos (*Workflow Scan Statistic*). Ambas as estatísticas não são ideais para serem usadas em epidemias modeladas por equações diferenciais. Propomos, então, a estatística WEB, baseada no valor esperado do número de casos e que nossos experimentos mostraram bons resultados para utilização em conjunto com modelos de equações diferenciais. A estatística espacial *scan* de Kulldorf, apesar de bem consolidada, ainda é passível de melhorias e, por isso, propomos uma nova técnica de inferência, a Data-Driven. Com este conjunto de modelos e estatísticas propostos, esperamos ajudar na desfragmentação do estudo de epidemiologia.

Palavras-chave: Epidemiologia, simulação de epidemias, modelo SIR, modelo baseado em indivíduos, MBI, redes complexas, vigilância, detecção de conglomerados, detecção de *clusters*, estatística *scan*, *workflow*.

Abstract

Theoretical studies to better understand mathematical epidemics are currently made in at least three distinct areas: the mathematical models of differential equations, the physics of their propagation models on networks and the statistics of disease cluster detection. Despite the common goal, those areas have little or no intersection. Our studies are directed at seeking tools within the intersection of those areas. Computer simulations of epidemics on hypothetical scenarios are valuable tools for understanding and predicting the behavior of epidemics. Thus, we have studied mathematical models that are able to describe the dynamics of epidemics in various scenarios. Direct transmission diseases from person to person (such as influenza, HIV, smallpox, etc.) depends on how individuals interact with each other. Moreover, infectious diseases behave differently in populations with different social structures. Mathematical models of differential equations usually do not take into account these aspects and, therefore, we propose a new individual based model (MBI), in order to consider different organizations within the community as complex networks. Studies of this kind of interactions have indicated difficulties when adjusting the parameters of the SIR to networks. We propose then a differential equation model, the μSIR, able to describe the usual SIR model in any network. However, the μSIR models needs complete information from the network, which could result in an impractical tool. Because of this, we developed the HMF-MC model, a multi-communities model whose required information consists of the frequencies of connectivity within each community and the number of edges between them. Methods for the early detection of geographic disease clusters of cases or symptoms are important tools in disease surveillance, endowing to the public health agencies the ability to intervene quickly in order to control outbreaks. We discuss two statistics: Kulldorff's spatial scan and the Workflow scan statistic. Both statistics are not ideally fitted for use in epidemics models employing differential equations. We thus propose the WEB statistic, based on the expected number of cases; our experiments showed good results for use in conjunction with models of differential equations. Kulldorff's spatial scan statistic, although well established, still requires improvement, and therefore, we propose a new technique, the data-driven inference. With this set of proposed tools, statistics and models in epidemiology, we hope to contribute to unify some of the current practices of the area.

Keywords: Epidemiology, mathematical models, simulation of epidemics, SIR model, individual based model, IBM, agent based model, complex networks, surveillance, cluster detection, workflow scan statistic.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos e Abreviações

N_Δ	número de triângulos
N_3	número de triplas de vértices conectados
BA	rede de Barabàsi-Albert
DM	rede Dorogovtsev-Mendes
β	taxa de contato
γ	taxa de recuperação
μ	taxa de renovação
Δ_t	intervalo de tempo
$\langle k \rangle$	conectividade média
Θ	prob. de uma aresta apontar para um infectado
$\mathcal{A}$	matriz de contato
SIR	modelo suscetível-infectado-recuperado
MBI	modelo baseado em indivíduos
μSIR	modelo micro SIR (micro analítico)
MKV	modelo de cadeias de Markov
GG	modelo MKV com $\Delta_t = 1$ de Guerra e Gomez-Gardenes
HMF	heterogeneous mean field
HMF-MC	modelo HMF multi-comunidades
$LR(.)$	razão de máxima verossimilhança
$LLR(.)$	logaritmo da razão de máxima verossimilhança
c_z	número de casos encontrados na região z
μ_z	número de casos esperados na região z
EBSS	expectation based scan statistic
WEB	workflow expectation based
$r_{\max}$	raio máximo do cluster trabalho no workflow
scan_k	distribuição empírica dos clusters de tamanho k
Gumbel_k	distribuição Gumbel ajustada aos clusters de tamanho k
CMV	cluster mais verossímil

Capítulo 1

Introdução

1.1 Relevância

Doenças infecciosas são motivo de pânico desde a antiguidade, dizimando populações e causando grandes prejuízos econômicos. Podemos encontrar referências às epidemias e ao grande impacto social por elas causado desde a antiguidade até os dias atuais. Aproximadamente um terço da população de Atenas morreu em função da epidemia conhecida como a Praga de Atenas, bem descrita por uma das vítimas, Thucydides, no período de 430 a 427 a.C.. Cicatrizes de lesões, possivelmente provenientes de varíola, foram encontradas em múmias do período 1570 a 1085 a.C. e também na múmia de Ramsés V, que morreu em 1157 a.C. [73]. A peste negra levou a morte de um quarto da população da Europa durante os anos de 1347 a 1350. A epidemia mundial da gripe causou cerca de 20 milhões de óbitos [22]. A mortalidade causada pelas grandes epidemias é muito maior que a causada por todas as guerras juntas [4].

Além do grande prejuízo em vidas humanas, as doenças infecciosas também atacam animais de valor econômico, podendo causar desde a redução de sua produtividade até sua mortalidade. As epidemias, em seus nomes populares, da *vaca louca* e da *gripe aviária* são exemplos de epidemias que geraram grandes prejuízos econômicos [22].

Epidemias também podem ser induzidas pelo próprio ser humano por meio da liberação ou disseminação intencional de agentes biológicos. Na antiguidade e na idade média a guerra biológica era praticada utilizando substâncias tóxicas provenientes de organismos vivos. Os exércitos usavam corpos em decomposição para contaminar o abastecimento de água de uma cidade, ou jogavam dentro das muralhas inimigas cadáveres em decomposição de vítimas de doenças como varíola ou peste bubônica. Durante a Guerra Fria, os Estados Unidos e a ex-União Soviética desenvolveram pesquisas voltadas para a guerra bacteriológica. No século XXI, o terrorismo encontrou uma nova modalidade para gerar pânico: o bioterrorismo.

Como se não bastasse o ataque feito pelas epidemias aos seres humanos e animais, com o aumento na utilização dos computadores no final do século XX, surgiu um novo problema: o *vírus de computador*[1]. E com o crescente uso

[1]Programa de computador malicioso desenvolvido por programadores que, assim como vírus biológico, infecta o computador, prolifera para outros e pode levar desde o aparecimento

dos computadores pessoais e da rede mundial de computadores, a *internet*, a proliferação desses vírus ganhou grande velocidade e com isso surgiu um novo problema: os *cavalos de Tróia*[2]. Juntas, essas duas novas ameaças geram insegurança e prejuízos que vão desde o usuário doméstico a grandes empresas multinacionais [85].

Os problemas acima citados mostram a relevância do estudo do comportamento das doenças infecciosas e são objetos de estudo da chamada *Epidemiologia*, a qual é responsável por desenvolver métodos de detecção, prevenção, controle e erradicação de doenças.

1.2 Motivação

Em 1760, Bernoulli [13] estudou o comportamento matemático da varíola, e com ele surgiu uma nova área da epidemiologia: a *Epidemiologia matemática* [105]. Atualmente o interesse em modelar doenças infecciosas tem crescido muito e parece ter uma infinidade de problemas a serem entendidos e resolvidos.

Entender o comportamento dinâmico de uma epidemia é o primeiro passo para controlá-la. Modelos matemáticos e simulações computacionais de epidemias em cenários hipotéticos são ferramentas valiosas para entender seu comportamento. Modelos e simulações podem antecipar possíveis surtos epidêmicos e ajudar a desenvolver intervenções que controlem, ou previnam, a epidemia.

Conhecer o comportamento da dinâmica epidemiológica em diferentes organizações de sociedades é um fator importante a ser considerado na modelagem de epidemias e é um desafio que tem despertado o interesse de alguns estudiosos [81, 92, 67, 79]. Organizações diferentes da sociedade podem determinar, por exemplo, campanhas de vacinação diferentes [84, 96].

Além disso, detectar o surgimento de uma nova infecção, ou mesmo definir um comportamento epidemiológico diferenciado de certas regiões em relação a outras, é muito importante e define uma área de epidemiologia chamada *vigilância epidemiológica*. Estudos dessa natureza podem definir quais regiões devem ter intervenções na área de saúde. Quanto antes for detectado o surgimento de uma epidemia, mais rápido pode-se tentar controlá-la, aumentando a probabilidade de sucesso dessa ação.

Classificar alterações de características em certo número de indivíduos como epidemia, significa dizer se o número de indivíduos alterados é ou não significativo do ponto de vista estatístico. Essa decisão de significância deve ser feita quando se detecta espaciais. Estudos de detecção de clusters espaciais são procedimentos importantes na área de saúde pública [56, 26].

O diagnóstico preciso sobre a característica aleatória ou não de um determinado evento espacial como, por exemplo, uma doença contagiosa, e a delimitação da região geográfica de ocorrência possibilitam aos órgãos competentes a elaboração de políticas eficientes de controle e combate.

Uma vez conhecidas as duas ferramentas acima (modelagem de epidemias

de simples mensagens até danos físicos nas máquinas.

[2]Assim como o vírus de computador, são programas maliciosos, desenvolvidos com a finalidade de dar acesso, a usuários não autorizados, ao sistema do computador infectado, ou de roubar informações contidas no mesmo.

e detecção de clusters), por um lado temos a capacidade preditiva dos modelos matemáticos e, por outro, uma metodologia estática da estatística de detecção de clusters. Desta forma, o uso das duas técnicas pode ser a solução para que um cluster não detectado com dados atuais possa ser previsto num determinado tempo futuro. Estas ferramentas acima não são capazes de fazer tal previsão em separado. Tal previsão é de suma importância para que as Secretarias de Saúde tomem as precauções necessárias.

1.3 Objetivos

Muitas contribuições surgiram em epidemiologia matemática nos últimos anos, mas ainda existem diversos problemas a serem entendidos e resolvidos. Atualmente temos três grandes frentes de pesquisas que tentam entender o comportamento de epidemias:

- Modelagem por equações diferenciais: geralmente é feita ou por pesquisadores da área de matemática que desconsideram o uso de redes no processo de contato, ou por pesquisadores da área de física que utilizam-se de um número reduzido de informações da rede na criação de uma aproximação para o processo de difusão;

- Processo de difusão em redes: feita por pesquisadores da área de física, que geralmente utilizam-se de simulações, aproximações e modelos simplificados, não havendo uma equação diferencial universal (independente do tipo de rede) para o processo;

- Detecção de clusters: feita por pesquisadores da área de estatística, considerando modelos de comportamento simplificado para as epidemias, nos quais as doenças seguem alguma distribuição de probabilidade conhecida, ignorando quaisquer modelos de equações diferenciais ou de propagação em redes sociais.

Desta forma nota-se que, apesar do objetivo comum em se entender o comportamento de epidemias, estas áreas possuem nenhuma, ou muito pouca, intersecção. O desafio que propomos neste trabalho é o de unir diferentes técnicas no intuito de tentar iniciar uma desfragentação do conhecimento epidemiológico. Explicamos abaixo mais especificamente o que foi feito.

A modelagem de sistemas epidemiológicos é comumente feita pela classificação dos indivíduos em compartimentos distintos. Um dos modelos mais utilizados é o modelo de equações diferenciais não-lineares desenvolvido em [55], o qual relaciona os indivíduos classificados em três compartimentos: suscetíveis, infectados e recuperados. Este modelo é chamado modelo SIR, e será descrito com mais detalhes na Seção 4.2. Baseado nas premissas deste modelo, um modelo estocástico baseado em indivíduos é desenvolvido em [74], denominado MBI, e será descrito na Seção 4.3. Desta forma:

- propomos um MBI para redes complexas (Seção 5.1) e, por meio dele, mostramos dificuldades na adequação do modelo SIR às redes sociais por meio do ajuste de parâmetros (Seção 5.2).

Diante das dificuldades encontradas nos ajustes dos parâmetros do modelo SIR, notamos a necessidade de modelos mais complexos. Desta forma, apresentamos modelos já utilizados em outros trabalhos e:

- propomos o modelo μSIR (Seção 5.5) de forma que, independente do tipo de rede utilizada, seja um modelo de equações diferenciais que modele o comportamento de epidemias em redes de forma equivalente ao modelo SIR, sem nenhuma perda de informação da rede utilizada.

Apesar dos bons resultados do μSIR, pode ser impraticável o conhecimento total da rede para grandes populações no mundo real. Uma vez que a estrutura de comunidades está presente em diversos tipos de sociedades, tomando como base o μSIR:

- propomos o modelo de equações diferenciais HMF-MC (Seção 5.7), capaz de, uma vez dividida a sociedade em comunidades, descrever o comportamento epidemiológico levendo-se em conta a distibuição dos números de conexões dentro de cada comunidade e o número de conexões entre cada par de comunidades.

Por meio destes modelos esperamos ter dado um primeiro passo para contribuir na união entre os estudos epidemiológicos em redes e as equações diferenciais. Do ponto de vista de detecção de clusters, apresentamos as técnicas mais utilizadas atualmente e:

- propomos a estatística WEB (Seção 6.5), que nossos estudos mostraram ser uma boa ferramenta na detecção de clusters em epidemias geradas por meio de equações diferenciais que modelem epidemias com contatos entre indivíduos de diversas cidades.

No intuito de tentar melhorar técnicas de detecção de clusters já existentes:

- propomos uma nova inferência, a *data-driven* (Capítulo 7), como uma inferência de p-valor mais precisa para a estatística de detecção mais consolidada, a estatísitca scan de Kulldorf (e suas variações).

Desta forma, esperamos contribuir como uma primeira união entre as técnicas de modelagem por equações diferenciais e detecção de clusters. Na próxima seção apresentamos a estrutura do texto.

1.4 Estrutura do texto

Este trabalho está organizado em capítulos, e os resultados para os modelos e estatísticas propostos encontram-se ao longo destes capítulos.

No Capítulo 2 apresentamos uma pequena introdução à epidemiologia, noções básicas e vigilância.

No Capítulo 3 apresentamos uma introdução às redes sociais, principais características e algumas das construções mais utilizadas em epidemiologia.

Apresentamos também as redes com estrutura de comunidades, bem como o método de detecção espectral para tal estrutura.

No Capítulo 4 apresentamos o modelo de equações diferenciais SIR e um modelo baseado em indivíduos (MBI), modelo estocástico equivalente ao SIR. Estes modelos servirão de base para o desenvolvimento dos modelos propostos. Apresentamos também um modelo de equações diferenciais para múltiplas cidades interconectadas bem como um modelo de equações diferenciais estocásticas, o qual será usado para testar a eficiência da estatística WEB que propomos no Capítulo 6.

No Capítulo 5 apresentamos os modelos mais utilizados atualmente para modelar epidemias baseados no modelo SIR para uma sociedade organizada em rede. Primeiramente propomos um MBI para redes na Seção 5.1 e o utilizamos em experimentos para mostrar dificuldades na adequação dos parâmetros do modelo SIR em redes na Seção 5.2. Apresentamos, então, os modelos HMF e baseados em cadeias de Markov e em seguida, na Seção 5.5, propomos o modelo μSIR como um modelo analítico de equações diferenciais equivalente ao modelo SIR em redes. Os resultados e comparações com os demais modelos são apresentados na Seção 5.6. Por fim, com a intenção de modelar especificamente epidemias em redes com estrutura de comunidades e encontrar um modelo que possa ser usado futuramente em casos reais, na Seção 5.7 propomos um modelo similar ao HMF para múltiplas comunidades, o HMF-MC. Os resultados dos experimentos para este modelo são apresentados na Seção 5.8.

No Capítulo 6 apresentamos os métodos estatísticos para detecção de clusters mais utilizados em vigilância epidemiológica. Apresentamos a estatística espacial *scan* de Kulldorf na Seção 6.2, bem como a estatística espacial para fluxo de indivíduos conhecida como *workflow* na Seção 6.3. Apresentamos, ainda, a estatística espacial baseada no valor esperado na Seção 6.4 e, por fim, propomos a estatística WEB na Seção 6.5, cujos resultados são apresentados na Seção 6.6.

No Capítulo 7 mostramos, através de experimentos, uma possível imprecisão da inferência do p-valor calculado empiricamente no algoritmo do scan circular da estatística espacial de Kulldorf. Propomos, então, uma nova inferência, a data-driven.

No Capítulo 8 fazemos as considerações finais e apresentamos os tabalhos futuros.

A Figura 1.1 ilustra o fluxograma dos assuntos abordados, bem como os modelos e metodologias propostos.

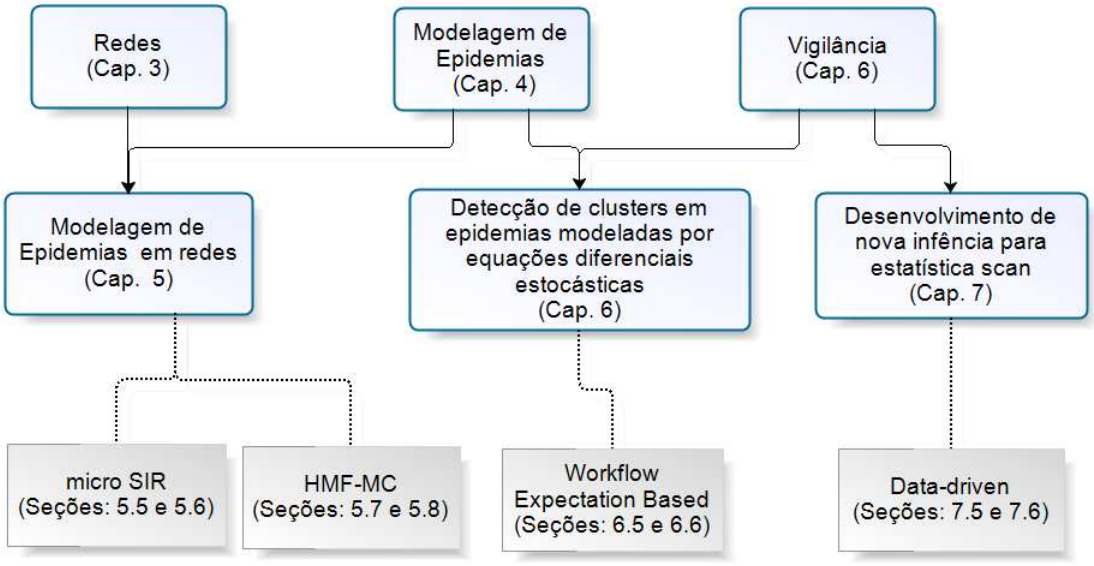


Figura 1.1: Fluxograma de assuntos abordados. Na última linha indicamos os modelos e metodologias propostos.

Capítulo 2

Epidemiologia Matemática

2.1 Introdução

A palavra *epidemiologia* é derivada do grego (*epi* = sobre; *demos* = população, povo; *logos* = estudo). Desta forma, epidemiologia etimologicamente significa *estudo do que ocorre em uma população.*

Em [3], a epidemiologia é definida como:

> *ciência que estuda o processo saúde-enfermidade na sociedade, analisando a distribuição populacional e os fatores determinantes dos riscos de doenças, agravos e eventos associados à saúde, propondo medidas específicas de prevenção, controle ou erradicação de enfermidades, danos ou problemas de saúde e de proteção, promoção ou recuperação da saúde individual e coletiva, produzindo informação e conhecimento para apoiar a tomada de decisão no planejamento, administração e avaliação de sistemas, programas, serviços e ações de saúde.*

Entende-se por doença infecciosa, ou doença transmissível, aquela causada por um agente biológico (por exemplo: vírus ou bactéria). Doenças infecciosas não afetam apenas o ser humano, mas também animais domésticos e animais de valor comercial. Com o desenvolvimento crescente dos computadores e da internet, epidemias de vírus de computador e dos chamados cavalos de tróia tornam-se um problema a ser estudado pela epidemiologia [85].

O estudo matemático da epidemiologia começou a ser realizado em 1760 por Bernoulli [13] no estudo da varíola. Somente a partir da metade do século XIX, com o avanço do conhecimento médico sobre microorganismos e doenças infecciosas, começam a surgir teorias matemáticas para fenômenos epidemiológicos.

O epidemiologista inglês Sir William Heaton Hamer, professor emérito da primeira geração de epidemiologistas da London School, construiu em 1906 a curva epidêmica do sarampo. Hamer desenvolveu a expressão matemática do sarampo tendo por base a sua *teoria mecânica de números e densidade.* Hamer postulou que o comportamento epidêmico podia ser entendido, e formalizado matematicamente, como uma dinâmica dos contatos entre indivíduos sadios e infectados. Este postulado tornou-se um importante conceito na aplicação da matemática em epidemiologia, conhecido como *princípio de ação de massas.* O

princípio de ação de massas define que a taxa de transmissão da doença é proporcional ao produto da densidade de indivíduos não infectados e infectados. Em 1927, Kermack e McKendrick [55] desenvolveram uma teoria relacionando o surgimento de uma epidemia a um valor crítico do número de suscetíveis. O princípio de ação de massas e a teoria do valor crítico são os dois marcos nos estudos da epidemiologia moderna.

2.2 Conceitos Básicos

Epidemia é a alteração, espacial e cronologicamente delimitada, do estado de saúde-doença de uma população, caracterizada por uma elevação inesperada e descontrolada da incidência de casos de determinada doença, ultrapassando valores da faixa de variação da prevalência da doença ou agravo da saúde preestabelecidos para aquela circunstância e doença [35]. Desta forma, em uma população que não esteja sujeita ao agravo da saúde, um pequeno número de casos já seria o suficiente para caracterizar uma epidemia.

Segundo [3], um indivíduo suscetível é *aquele que não possui a resistência a determinado agente patogênico e que, por esta razão, pode contrair a doença se posto em contato com o mesmo.*

Ainda segundo [3], um sistema epidemiológico é *o conjunto formado por agente, suscetível e pelo meio ambiente, dotado de uma organização interna que regula as interações determinantes da produção de doença, juntamente com os fatores vinculados a cada um dos elementos do sistema.*

A epidemiologia matemática é fundamentada em conjunto de premissas capazes de quantificar alguns dos aspectos biológicos do sistema epidemiológico.

Neste trabalho, todas as premissas utilizadas são baseadas em doenças infecciosas, ou seja, aquelas cujo agente de transmissão da doença (vírus, bactérias, etc) pode ser passado de um indivíduo hospedeiro (ser humano, animal ou computador) infectado para um suscetível.

No estudo de doenças infecciosas, a abordagem mais utilizada, e que trataremos neste trabalho, é a modelagem compartimental, que consiste em dividir a população de hospedeiros em classes, ou compartimentos [4, 45, 89], sendo os modelos mais comuns listados na Tabela 2.1. Neste trabalho estudaremos o modelo SIR cujas classes são:

- suscetíveis: indivíduos que não estão infectados, mas podem vir a estar;
- infectados: indivíduos que possuem o agente patogênico e podem transmitilo a outros indivíduos;
- recuperados: indivíduos que se recuperaram da infecção e adquiriram imunidade, temporária ou não, à doença.

A *Força de infecção* (λ) é o número de novos casos por unidade de tempo de uma doença. É um parâmetro epidemiológico cuja estimativa é de grande importância, pois determina não somente a dimensão da epidemia, mas também o esforço para combatê-la.

A *Reprodutibilidade basal* (R_0) é o número de casos da doença produzidos apenas pelo primeiro indivíduo infectado em uma população completamente

SAIR	suscetível, antídoto, infectado, recuperado
SEIR	suscetível, exposto, infectado, recuperado
SI	suscetível, infectado
SIIs	suscetível, infectado, isolado
SIR	suscetível, infectado, recuperado
SIRIs	suscetível, infectado, recuperado, isolado
SIRS	suscetível, infectado, recuperado, suscetível
SIS	suscetível, infectado, suscetível
SIV	suscetível, infectado, vacinado
SVI	suscetível, vacinado, infectado
TSIR	SIR baseado em série temporal (time-series SIR model)

Tabela 2.1: Alguns modelos compartimentais encontrados na literatura.

suscetível [45]. Este número serve para medir se uma doença permanecerá em uma população ou se será erradicada.

Umas das estratégias de controle de epidemias mais utilizadas é a *vacinação*, na qual se força a mudança de classe do indivíduo de suscetível para imune (recuperado), sem que ele passe pela classe de infectado. A diminuição do número de indivíduos suscetíveis implica na diminuição da força de infecção e da reprodutibilidade basal [4].

2.3 Vigilância Epidemiológica

Segundo a lei brasileira 8.080/90, vigilância epidemiológica é definida como:

> *um conjunto de ações que proporcionam o conhecimento, a detecção ou prevenção de qualquer mudança nos fatores determinantes e condicionantes de saúde individual ou coletiva, com a finalidade de recomendar e adotar as medidas de prevenção e controle das doenças ou agravos.*

A vigilância epidemiológica é hoje a ferramenta metodológica mais importante para a prevenção, controle ou mesmo erradicação de doenças.

Suas principais aplicações são a detecção de epidemias, aglomerados específicos (infecções, sintomas, defeitos congênitos, etc.), doenças emergentes, etc. Abrange a capacidade de identificar padrões de comportamento de eventos adversos à saúde, desenvolver investigações epidemiológicas complementares e garantir a rápida identificação e controle de ameaças emergentes à saúde [97].

O principal objetivo da vigilância é detectar rapidamente um inesperado aumento na incidência de certa doença, de modo a se poder efetuar ações de prevenção e tratamento mais rapidamente. Quanto mais cedo se detecta o início de uma epidemia, maiores as chances de se prevenir o espalhamento da doença e mortalidade dos indivíduos.

O ponto fraco dos sistemas públicos de vigilância em saúde existentes está na coleta e divulgação de dados sobre sintomas e infecções das doenças. Por esta razão, a modelagem matemática de epidemias é uma ferramenta indispensável na criação de cenários hipotéticos que ajudarão no entendimento e

desenvolvimento de ferramentas de vigilância. Fazemos tal modelagem no Capítulo 4 e no Capítulo 6 abordaremos como método de vigilância a detecção de clusters espaciais de sintomas ou infecções.

Capítulo 3

Redes Sociais

Segundo [75], uma rede é um conjunto de itens (chamados vértices ou nós) com conexões entre eles (chamadas arestas ou ligações). Na literatura matemática as redes são também chamadas de grafos. As arestas (conexões) são usadas para indicar alguma espécie de relação entre os vértices (itens) que conectam, em conformidade com o problema modelado. O estudo de redes, feito na forma de teoria dos grafos, é um dos pilares fundamentais da matemática discreta.

As redes fazem parte do cotidiano humano e podem ser facilmente encontradas. Como exemplos temos a distribuição de água, energia elétrica, telecomunicações, estradas, rotas marítimas e aéreas, bancos de dados, páginas web, citações bibliográficas, redes sociais de relacionamento (Orkut, Twiter, etc), redes de reações químicas e de interações proteicas, etc.

As primeiras análises de estruturas em redes sociais foram introduzidas por Erdös e Rényi [31, 32, 33]. Eles propuseram um modelo que consistia de uma probabilidade p de dois vértices serem ligados por uma aresta. Esta consideração gerava uma rede aleatória que segue uma distribuição de Poisson, fazendo com que seja raro encontrar grandes diferenças na concentração do número de arestas presentes em um vértice. Este tipo de rede é comumente chamado de rede aleatória clássica, ou simplesmente rede aleatória.

Muitas redes observadas na natureza e nas relações humanas possuem características diferentes das encontradas em redes aleatórias. Geralmente possuem uma grande diferença na concentração do número de arestas presentes em um vértice.

Em 1967, Stanley Milgram procurou estimar quantas etapas são necessárias para que dois estranhos estabeleçam um vínculo. Milgram enviou cartas para pessoas em Boston e Nebraska solicitando que reenviassem as cartas a um destinatário em Massachusetts. Contudo, ele não forneceu o endereço desse destinatário, apenas o primeiro nome e algumas informações pessoais. As pessoas enviaram as cartas a pessoas que acreditavam estar mais próximas do destinatário final. Como resultado, as cartas chegaram ao destinatário final após cinco postagens em média. Isso significava que com uma postagem a mais seria possível chegar a qualquer destinatário. Surgia, então, a hipótese dos *seis graus de separação*, na qual qualquer pessoa está indiretamente ligada a outra do planeta por apenas 6 pessoas entre elas. Surge então a idéia de *mundo pequeno*.

Somente nos últimos anos, com o avanço tecnológico dos sistemas de aquisição de dados e o aumento do poder computacional, pesquisas nesta área puderam se desenvolver.

Duncan Watts e Steven Strogatz [100, 101] investigaram as redes de mundo pequeno. Contudo, estas redes ainda não modelavam bem muitas redes encontradas, surgindo o conceito de *redes livres de escala* proposto por Barabási e Albert [10], nas quais as redes teriam um pequeno número de vértices com grande número de arestas e um grande número de vértices com um pequeno número de arestas. A partir de então, novas formulações baseadas neste modelo vêm sendo propostas para modelar diversas estruturas encontradas no mundo real.

3.1 Conceitos Básicos e Propriedades

Uma rede complexa é definida como uma rede cuja estrutura não segue um padrão regular. Porém, é difícil encontrar na literatura uma conceituação clara e universalmente aceita de padrão regular. Por questões de simplicidade, consideraremos as redes regulares como sendo aquelas em que todos os vértices possuem o mesmo número de arestas ligadas a eles. Consideraremos, então, que redes complexas são redes que não são regulares.

Chamamos de rede direcional aquela que possui um sentido de ligação, ou seja, existe um vértice inicial e um final ordenados. Neste caso chamaremos as ligações entre os vértices de arcos. Redes não direcionais, ou bidirecionais, são aquelas cujas ligações não possuem sentido de ligação, sendo suas ligações chamadas de arestas.

Neste trabalho, consideraremos somente as redes bidirecionais, sem ciclos unitários, ou seja, arestas que iniciam e terminam em um mesmo vértice, e sem múltiplas arestas, ou seja, um mesmo par de vértices é ligado por no máximo uma única aresta.

A forma mais utilizada para descrever uma rede é através de sua *matriz de adjacência* $\hat{A}$, cujos elementos são 0 ou 1. Um elemento $\hat{a}_{ij}$ da matriz de adjacência de uma rede assume o valor 1 se existe uma ligação entre os vértices i e j e assume o valor 0 caso contrário. Redes bidirecionais possuem sua matriz de adjacência simétrica. A Figura 3.1 mostra uma rede bidirecional composta por 4 vértices e 3 arestas, e sua respectiva matriz de adjacência.

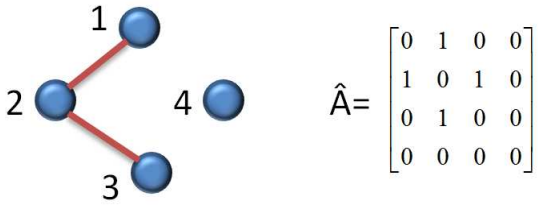


Figura 3.1: *Exemplo de uma rede com 4 vértices, 2 arestas e sua matriz de adjacência.*

3.1.1 Densidade

Definimos como a densidade de uma rede a razão entre o número de arestas presentes no grafo e o número total de arestas possíveis no mesmo. Uma rede com N vértices tem $N(N-1)/2$ arestas possíveis. Se N_a é o número de arestas presentes no grafo, então a densidade d da rede é dada por:

$$d = \frac{2N_a}{N(N-1)}.$$

3.1.2 Distribuição de Graus

O grau, ou conectividade, de um vértice é definido como o número de arestas a ele ligadas. Em redes geradas de forma aleatória, a cada vértice i podemos associar a probabilidade $p(i, k, N)$ deste, em uma rede de tamanho N, ter k conexões. A distribuição de graus da rede é então dada por:

$$P(k, N) = \frac{1}{N} \sum_{i=1}^{N} p(i, k, N).$$

Se todos os vértices da rede são estatisticamente equivalentes, eles possuem a mesma distribuição de graus $P(k, N)$, e passaremos a omitir o tamanho da rede, denotando a probabilidade de um vértice qualquer ter grau k por $P(k)$.

O grau médio de uma rede é dado por

$$\bar{k} = \sum_{k} kP(k).$$

Para construir uma rede com determinada distribuição de graus, proseguimos da seguinte forma:

1. enumere cada vértice;
2. para cada vértice j atribua k_j meia-arestas (aresta ligada a apenas um vértice) de acordo com a distribuição de graus $P(k, N)$;
3. aleatoriamente substitua pares de meia-arestas, de vértices diferentes, por uma aresta ligando-os.

Os seguintes exemplos mostram distribuições de graus típicas em redes:

- *Distribuição de Poisson:*

$$P(k) = \frac{e^{-\bar{k}}\bar{k}^k}{k!}$$

  As redes aleatórias clássicas possuem esta distribuição.

- *Distribuição Exponencial:*

$$P(k) \propto e^{-k/\bar{k}}$$

- *Distribuição por Lei de Potência:*

$$P(k) \propto k^{-\gamma}$$

em que $k \neq 0$ e γ é expoente da lei de potência.

Ao contrário das distribuições de Poisson e exponencial, a distribuição por lei de potência não tem uma escala natural e, por este motivo, redes que tem esta ditribuição são chamadas *redes livres de escala*.

3.1.3 Aglomeração (Clustering)

Um importante parâmetro para caracterizar a topologia de uma rede é o *coeficiente de aglomeração*[1] C, também chamado de transitividade. Existem várias maneiras para medir este coeficiente [78, 11], que geralmente levam em consideração a probabilidade média de conexão de dois vértices, dado que eles tem ligação com um vértice em comum. Matematicamente, o coeficiente de aglomeração é três vezes a razão entre o número de triângulos[2] presentes na rede, N_Δ, e o número de triplas de vértices conectados[3], N_3, ou seja:

$$C = \frac{3N_\Delta}{N_3}.$$

A Figura 3.2 ilustra 3 diferentes redes cujos cálculos de N_Δ, N_3 e C são mostrados ao lado direito de cada rede.

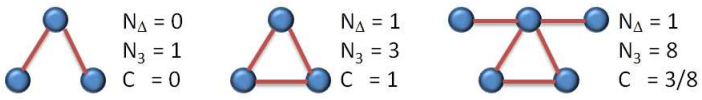


Figura 3.2: *Exemplos de redes e respectivos número de triângulos, número de triplas conectadas e coeficiente de aglomeração.*

3.2 Modelos e Construções

3.2.1 Redes regulares

Como foi dito anteriormente, redes regulares são redes em que todos vértices possuem o mesmo grau. Definimos que um vértice é vizinho de outro se eles possuem um aresta entre eles. Um rede regular pode ser facilmente construída se imaginarmos a disposição dos vértices de forma matricial. A Figura 3.3 ilustra os vizinhos de um vértice escolhido. Na figura temos três exemplos de vizinhanças: A possui 4 vizinhos, sendo esta vizinhança chamada de *vizinhança de Von Newman*, B possui 8 vizinhos, sendo esta chamada de *vizinhança de Moore*, por fim, o vértice C com 24 vizinhos.

[1] Do inglês *cluster coefficient.*

[2] Tripla de vértices conectados entre si.

[3] Conjunto de três vértices não ordenados com duas arestas entre eles. Um triângulo, por exemplo, possui 3 triplas conectadas.

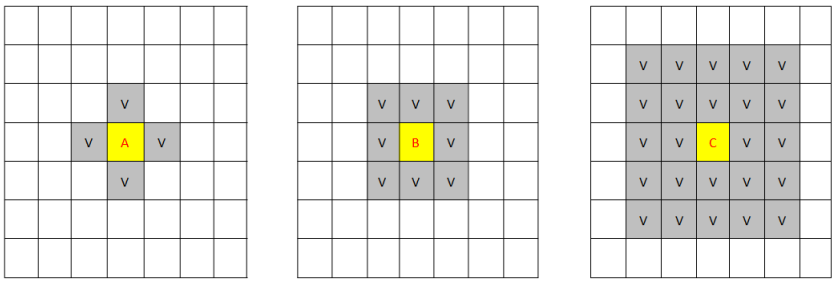


Figura 3.3: *Exemplos de vizinhanças. A, B e C possuem 4, 8 e 24 vizinhos respectivamente.*

Ao imaginarmos as vizinhanças na forma matricial nos deparamos com um problema: os vértices dos extremos da matriz possuem menos arestas que os centrais, o que, pela definição utilizada, faz com que a rede não seja regular. Para contornar este problema, a construção desta rede deve ser interpretada na superfície de um toro, conforme Figura 3.4, de forma que não existam vértices nos extremos.

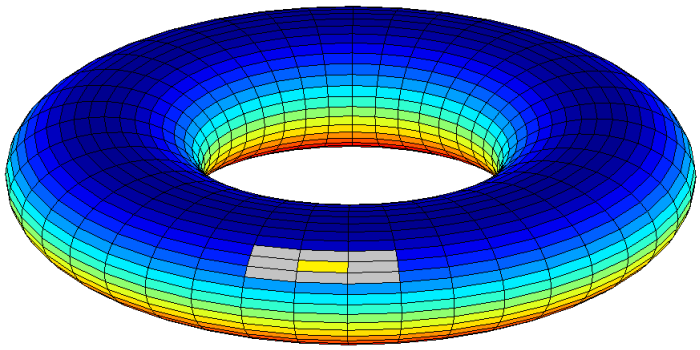


Figura 3.4: *Visualização em 3 dimensões para vizinhanças na superfície de um toro.*

Outras construções de redes regulares são possíveis, sendo estas exemplificadas aqui por serem as de mais simples visualização.

3.2.2 Redes gaussianas

Redes gaussianas são redes complexas nas quais a probabilidade da ocorrência de uma aresta entre quaisquer dois vértices é a mesma. A função distribuição de probabilidade dos graus dos indivíduos de uma rede gaussiana é aproximadamente uma curva gaussiana, conforme a Figura 3.5. Neste tipo de rede, os vértices tem aproximadamente o mesmo número de arestas. Desta forma, não existem subredes de vértices densamente interconectados.

A construção de uma rede gaussiana consiste em repetir o seguinte procedimento até se atingir o número desejado de arestas:

1. sortear pares de vértices escolhidos com igual probabilidade;

2. adicionar uma aresta ligando estes vértices;

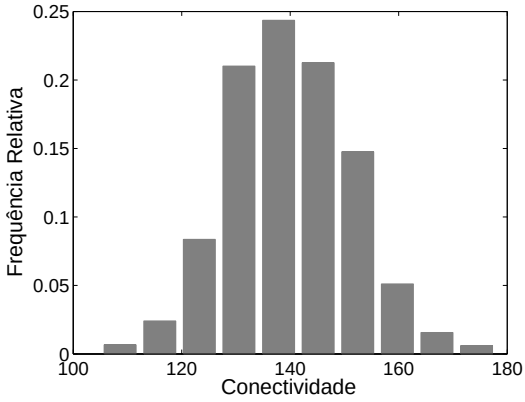


Figura 3.5: *Histograma da frequência relativa da conectividade dos vértices em uma rede gaussiana com* 2.000 *vértices e densidade* $0,07$.

3.2.3 Redes livres de escala

Redes livres de escala[4] são redes complexas nas quais a probabilidade de um vértice ter grau k segue a lei de potência $P(k) \propto k^{-\gamma}$, em que γ é um número real maior que zero. A distribuição da conectividade dos vértices em uma rede livre de escalas é exemplificada na Figura 3.6. A característica principal destas redes é possuir um pequeno número de vértices com elevado número de arestas e um grande número de vértices com baixo número de arestas. Esta propriedade é comumente observada em diversas relações humanas.

- ***Rede Barabási-Albert***

A primeira construção de uma rede aleatória livre de escala foi proposta por Barabási-Albert (BA) em [10], na qual os autores verificam a necessidade deste tipo de rede ser construída por meio de um processo de crescimento (acréscimo de vértices com o tempo) e ligação preferencial entre os vértices. Tal formulação gera uma distribuição seguindo a lei de potência com $\gamma = 2,9 \pm 0,1$. Ainda segundo os autores, estudos empíricos mostram que diversas relações humanas possuem uma lei de potência com γ variando de $2,1$ a 3, em grande parte próximas de $2,1$. Na construção da rede BA a conectividade média converge para $\langle k \rangle = 2m$. A construção de uma rede BA consiste em:

1. começamos a rede com um pequeno número de vértices (m_0);
2. (crescimento) adicionamos um novo vértice à rede;
3. (ligação preferencial) ligadas ao último vértice incluído, adicionamos $m(\leq m_0)$ novas arestas ligadas a diferentes vértices já existentes na rede escolhidos com probabilidade $P(k_i) = k_i / \sum_j (k_j)$, em que k_i é o número de conexões (conectividade) do vértice i e o somatório corresponde à conectividade total da rede;

[4] Do inglês: *scale free networks*.

4. voltamos ao passo 2 e repetimos o processo até completarmos o número de vértices desejados.

A maioria dos artigos que explicam a construção da rede BA afirmam começar com uma rede inicial totalmente desconectada, mas note que este fato impossibilita iniciar o procedimento devido ao cálculo de $P(k_i)$. Ainda que a rede inicial comece com um certo número de arestas, um vértice sem conexões nunca receberá novas arestas. Em [49], o autor inicia a rede com m vértices totalmente conectados entre si, que será a rede inicial utilizada neste trabalho.

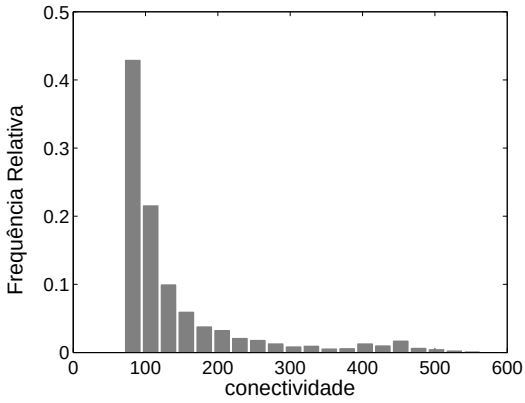


Figura 3.6: *Histograma da frequência relativa da conectividade dos vértices em uma rede livre de escala do tipo BA com* 2.000 *vértices e densidade* 0,07.

- ***Rede Dorogovtsev-Mendes***

Partindo do modelo BA, Dorogovtsev e Mendes (DM) [23, 25] modificam o modelo BA para que cada passo no tempo inclua o seguinte mecanismo de crescimento:

- desenvolvimento de redes: considera a possibilidade de conexão entre vértices existentes e consiste em adicionar cm novas arestas entre os vértices existentes, com probabilidade proporcional ao produto de suas conectividades, em que c é uma constante não negativa.

O modelo DM segue a mesma lei de potência do modelo BA com expoente dado por $\gamma = 2 + \dfrac{1}{1+2c}$, com a conectividade média convergindo para $\langle k \rangle = 2(m + mc)$.

- ***Rede Preferencial***

Construímos uma rede com ligações preferenciais sem a opção de crescimento com a finalidade de modelar o fato de que pessoas que têm contato com muitas pessoas têm maior probabilidade de terem contato entre si, e consideramos neste trabalho mais um tipo de rede, que denominamos *rede preferencial*. Construímos esta rede repetindo o seguinte procedimento até atingirmos o número de arestas desejadas:

1. escolhe-se um par de vértices, cada vértice com probabilidade $P(k_i) = k_i / \sum_j (k_j)$;

2. cria-se uma aresta entre eles;

A Figura 3.7 ilustra a distribuição de conexões para esta rede.

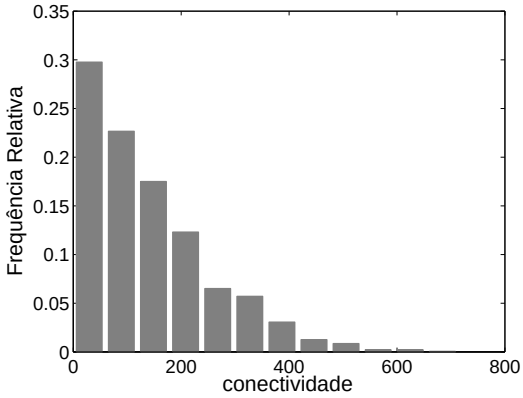


Figura 3.7: *Histograma da frequência relativa da conectividade dos vértices em uma rede livre de escala preferencial com* 2.000 *vértices e densidade* $0,07$.

3.2.4 Redes espaciais

Redes espaciais são grafos cujos vértices possuem uma localização espacial e uma métrica associada. Em [28] a construção destas redes é feita da seguinte forma:

1. a rede inicia com os vértices com localização fixa em algum espaço com métrica associada;

2. com igual probabilidade, escolhemos um vértice N_i;

3. com probabilidade que decai com a distância ao primeiro vértice, escolhemos um outro vértice, N_j;

4. criamos uma aresta entre os vértices escolhidos;

5. retornamos ao passo 2 e repetimos o processo até que se atinja o número de arestas desejadas.

Consideraremos esta construção no restante do texto. Utilizaremos como a probabilidade de sortear o segundo vértice, N_j, após termos sorteado o primeiro vértice, N_i, no processo acima a seguinte função de probabilidade:

$$P_{N_i}(N_j) = \beta e^{-\alpha d_{ij}},$$

em que d_{ij} é a distância espacial entre os vértices N_i e N_j, α é o coeficiente não negativo que regula a dependência das arestas e as distâncias entre os vértices, e β ajusta a função de probabilidade no intervalo $[0, 1]$. A Figura 3.8 ilustra a

distribuição de conexões para uma rede espacial com $\alpha = 30$ e $\beta = \sum_j e^{-\alpha d_{ij}}$ para cada vértice N_i.

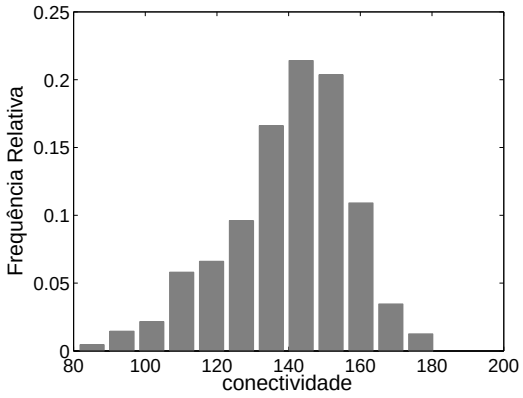


Figura 3.8: *Histograma da frequência relativa da conectividade dos vértices em uma rede espacial com* 2.000 *vértices e densidade* 0,07.

Alguns autores, similarmente ao modelo BA, constroem a rede por crescimento com preferencialidade que decai com a distância entre os vértices. Não trataremos deste tipo de rede neste trabalho.

Redes espaciais geralmente possuem coeficiente de aglomeração superior às redes geradas com ligação preferencial dependente do número de conexões. Outras construções de redes espaciais podem ser encontradas em [104, 52, 53]. Em [103], é apresentada uma rede complexa construída com ligação preferencial que leva em consideração tanto o aspecto espacial quanto as conectividades dos vértices.

Redes livres de escala possuem coeficiente de aglomeração levemente superior às redes gaussianas. Redes reais demonstram ter coeficiente de aglomeração ainda maior que os esperados nas redes gaussiana, BA, DM e preferencial. Desta forma, [30] propõe um novo tipo de rede: redes livres de escala altamente aglomeradas, que são baseadas no modelo BA, porém com um processo de ativação e desativação de vértices. Um estudo das propriedades das redes altamente aglomeradas é feito em [75], no qual o autor propõe a construção da rede através da divisão dos vértices em grupos.

Alguns autores têm trabalhado em métodos que possibilitem construir redes que, além de seguirem uma distribuição de conectividade dada, tenham um coeficiente de aglomeração desejado. Em [47], o autor estende o modelo BA acrescentando um passo no qual força o fechamento de triângulos. Em [77] o autor constrói a rede com um número de triângulos e vértices especificados a priori. Um estudo formal do processo de aglomeração em redes complexas pode ser encontrado em [91].

3.2.5 Redes com estrutura de comunidades

Se considerarmos uma rede social que represente o contato de pessoas de uma pequena cidade, observaremos que os familiares apresentam-se mais den-

samente conectados entre si e que as conexões entre membros de famílias difrerentes são mais esparsas.

Devido à variedade de definições existentes, a definição exata de comunidade é uma tarefa complicada. Para cada definição, várias decisões podem ser tomadas: se um mesmo nó pode ou não fazer parte de mais de uma comunidade, se nas ligações com pesos serão usadas com ou sem estes pesos, etc. Para este trabalho, uma noção intuitiva será suficiente: uma comunidade (ou módulo) pode se vista intuitivamente como uma sub-rede de nós densamente conectados entre si com relativamente poucas conexões com o resto da rede.

Uma mesma sociedade tem diversas possibilidades de organização em comunidades: grupos de trabaho, círculo de amigos, familiares, etc. Com a difusão da internet vemos o surgimento de novas comunidades, as chamadas comunidades virtuais.

Comunidades são importantes devido a poderem ser relacionadas com diversos sistemas, como por exemplo proteínas relacionadas à metástase de câncer [50], páginas da internet relacionadas por tópicos similares [34], grupos de indivíduos que interagem entre si [39, 5], redes de interação proteína-proteína classificadas por suas funções específicas dentro da célula [87, 18], etc.

A Figura 3.9 ilustra um grafo com 4 comunidades distintas, nas quais seus indivíduos estão mais conectados entre si e com uma densidade menor de conexões entre indivíduos de comunidades diferentes.

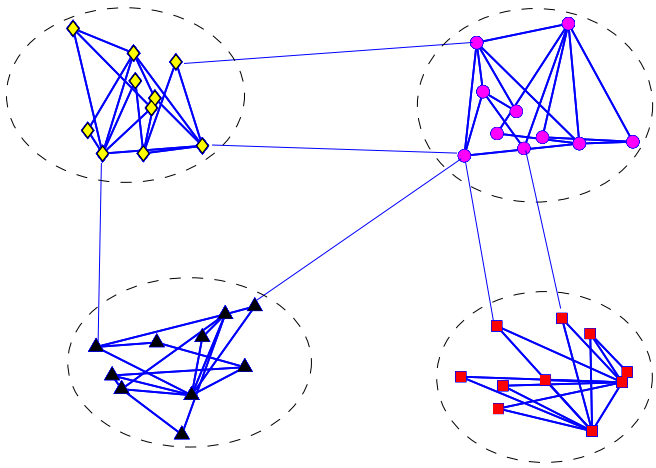


Figura 3.9: Exemplo de rede com estrutura de comunidade.

Às redes com duas ou mais comunidades diremos ter estrutura de comunidade. Neste trabalho, a construção destas redes será feita pela construção independente de redes (as comunidades) conforme apresentadas anteriormente (gaussiana, BA, DM, etc.), e a posteriori serão feitas ligações entre estas redes por alguma regra específica, formando assim uma rede maior.

3.3 Detecção de Comunidades em Redes

Nos últimos anos, o problema de detecção de comunidades tem recebido considerável atenção (apesar de não se ter uma definição formal única).

A capacidade de se detectar estrutura de comunidade em uma rede tem claramente aplicações práticas, como por exemplo na identificação de comunidades em redes de citações em artigos e periódicos que podem representar um mesmo assunto [86]. Identificar comunidades pode ajudar a entender as redes com maior eficiência.

A detecção da comunidade também pode fornecer informações sobre as funções dos nós individuais. Por exemplo, um nó mais exterior de uma comunidade pode funcionar como um importante mediador entre as comunidades enquanto um nó central proporciona maior controle e estabilidade à comunidade [43].

Uma dificuldade que pode ocorrer é que alguns nós pertençam simultaneamente a várias comunidades, caso em que as comunidades são ditas sobrepostas. Tal sobreposição de comunidades são abundantes, por exemplo, em redes espaciais.

Outra dificuldade na detecção da comunidades é que as redes podem conter estruturas hierárquicas, o que significa que as comunidades podem ser partes de comunidades ainda maiores. Isso leva ao problema da escolha da melhor partição dentre as possíveis.

Por não ser o foco deste trabalho, apresentamos um único método para detecção de comunidades.

3.3.1 Detecção e estimação de comunidades por análise espectral

Os métodos mais utilizados para detecção de comunidades em redes têm sido desenvolvidos na última década [36]. Um grande número de definições, métodos e algoritmos relacionados foram propostos ao longo dos anos.

Os métodos tradicionais utilizados na teoria dos grafos normalmente são ruins para grandes redes porque restringem fortemente as soluções. Métodos de otimização global têm sido amplamente utilizados e têm se mostrado eficientes para muitas redes testadas. No entanto, o trabalho [37] demonstra que o método mais popular, otimização da modularidade, sofre de limitações sérias.

Ao longo dos anos, um grande número de algoritmos foi introduzido, por exemplo, o método baseado na centralidade [39], fatoração da matriz não-negativa [98], aglomeração espectral [76, 106], busca local [9], etc.

Algoritmos espectrais têm se mostrado uma ferramenta poderosa. Os resultados obtidos por métodos espectrais, muitas vezes superam os de outras abordagens. A abordagem espectral é simples de implementar e pode ser feita por métodos padrão de álgebra linear. Por estes motivos, utilizaremos neste trabalho tal metodologia na detecção de comunidades.

A análise espectral tem tido grande sucesso na descoberta da estrutura da comunidade, sendo baseada em diversas matrizes: na matriz de adjacência [17], matriz Laplaciano padrão [6], matriz Laplaciano normalizada [19], matriz de modularidade [76], matriz de correlação [93] e várias outras variantes destas matrizes. Neste trabalho, utilizaremos a matriz Laplaciano normalizada.

- **Matriz Laplaciano normalizada:** Seja A a matriz de adjacência e D a matriz diagonal qua contem a conectividade dos nós. A matriz Laplaciano normalizada N é definida por $N = I - T$, sendo I a matriz identidade e T a matriz de transição (probabilidade de em um passeio aleatório se mover do nó i para o nó j) e é calculada por $T = D^{-1}A$.

A matriz Laplaciano normalizada está estreitamente correlacionada à dinâmica de difusão das redes. De forma a investigar a dinâmica de difusão das redes, em [19] a estrutura da comunidade é identificada através dos autovalores e autovetores da matriz Laplaciano normalizada. Se λ é um autovalor da matriz de transição, então $\lambda^N = 1 - \lambda$ é um autovalor da matriz Laplaciano normalizada com os mesmos respectivos autovetores.

Do ponto de vista prático, a detecção de comunidades consiste em ordenar os m autovalores em ordem crescente e o comprimento da $i-$ésima auto distância[5] é definido por $d_i = \lambda^N_{i+1} - \lambda^N_i$, $(1 \leq i \leq m-1)$. Desta forma, número de comunidades será dado por n, sendo que a $n-$ésima auto distância é maior de todas. A Figura 3.10 ilustra os autovalores ordenados encontrados para a matriz Laplaciano normalizada para o grafo referente à Figura 3.9, bem como indica que a maior auto distância é d_4, significando a existência de 4 comunidades.

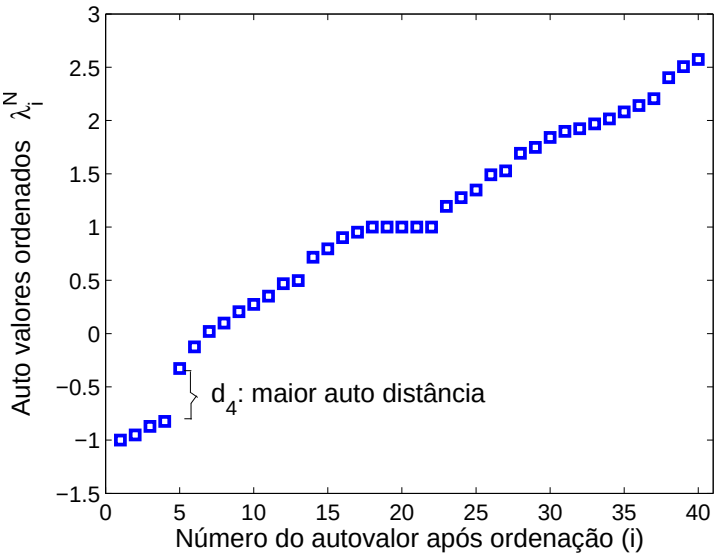


Figura 3.10: Auto valores ordenados da matriz laplaciano normalizada obtida a partir da rede da Figura 3.9 com 4 comunidades. A maior auto distância encontrada é para $i = 4$, indicando a existência de 4 comunidades.

Uma vez descoberto o número de comunidades é necessário um processo que particione a rede e responda quais nós pertencem a qual comunidade. Para redes grandes pode ser impraticável tentar fazer esta partição devido à elevada dimensão do problema. Porém, uma vez conhecido o número de comunidades

[5] do inglês: eigengap

n, seleciona-se n autovetores correspondentes aos n menores autovalores da matriz Laplaciano normalizada. Estes autovetores são colocados nas colunas de uma nova matriz e o transposto da $i-$ésima linha desta matriz é tomada como a projeção do vetor correspondente ao nó i. A partição em comunidades é então realizada nestas projeções (linhas da nova matriz) tomadas como pontos no espaço n-dimensional.

A Figura 3.11 ilustra a projeção do auto-espaço coluna no espaço gerado pelos autovetores da matriz Laplaciano normalizada referente à rede da Figura 3.9 correspondentes aos autovalores λ_2^N, λ_3^N e λ_4^N. Nota-se as quatro partições para as comunidades encontradas.

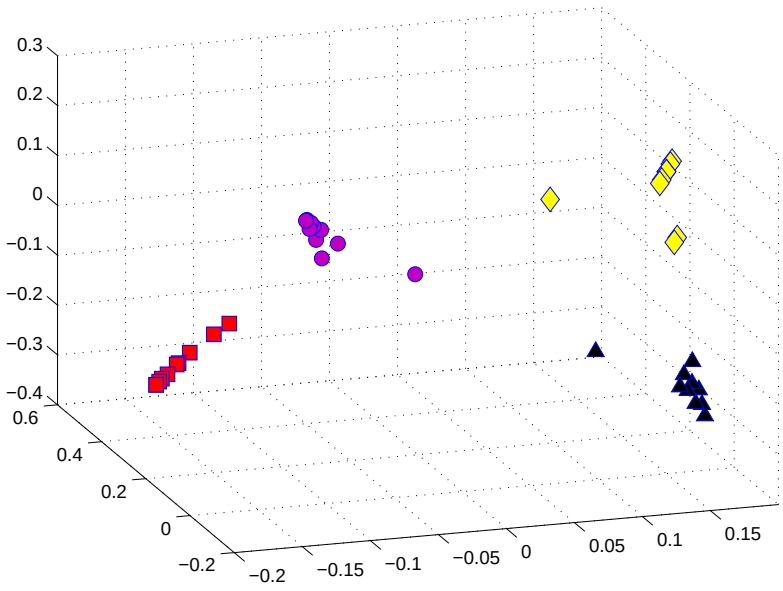


Figura 3.11: Projeção dos autovetores encontrados da matriz Laplaciano normalizada referente à rede da Figura 3.9.

É útil observar que a dimensão do problema em particionar a rede por meio destes novos vetores projetados é drasticamente menor que o problema original, sendo agora $n-$dimensional. Neste trabalho usaremos o método k-médias[6] [63, 90, 95] para realizar esta partição no novo espaço, uma vez que já é um método bem conhecido.

3.3.2 Redes espaciais possuem estrutura de comunidades?

Nem sempre a decisão pelo número da maior auto distância pode ser a mais adequada. A Figura 3.12 ilustra a dificuldade na tomada de decisão por estrutura de comunidade. Nota-se claramente a ausência de comunidades na rede BA (gerada como apresentado na Seção 3.2.3), uma vez que possui um único autovalor isolado dos demais. Na rede espacial (gerada como apresentado na Seção 3.2.4), apesar de não ter sido gerada com estrutura de comunidade, nota-se que os primeiros autovalores seguem um padrão de crescimento bem próximos, e auto distâncias relativamente menores que os demais. Uma

[6] do inglês: k-means

possibilidade para a escolha do número de comunidades nestes casos é ignorar a auto distância d_1, escolhendo então a maior auto distância a partir de d_2, considerando assim a existência de no mínimo 2 comunidades, porém tal arbitrariedade pode levar a conclusões errôneas. Uma outra visão para os autovalores da matriz Laplaciano normalizada é que eles também podem indicar uma estruturação intrínseca da rede.

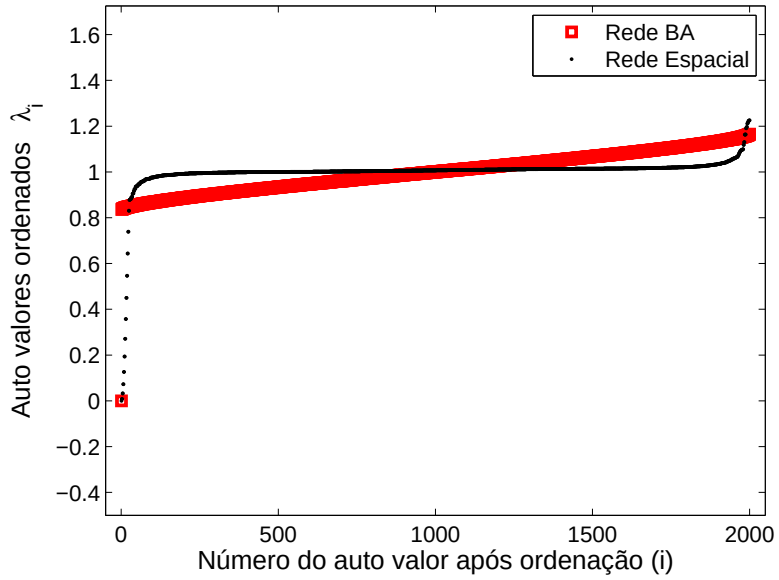


Figura 3.12: Autovalores encontrados para redes BA e espacial com 2000 vértices. Na rede BA nota-se claramente a indicação de apenas uma única comunidade. Na rede espacial, o número de comunidades é mais difícil de se notar, podendo indicar que as comunidades tem intersecções entre si.

A maior auto distância encontrada para os autovalores da rede espacial analisada na Figura 3.12 é d_8, indicando a existência de 8 comunidades. Após classificação e reordenamento dos nós em comunidades, obtemos a rede[7] mostrada na Figura 3.13 (esquerda) e os autovalores ordenados (direita) da matriz Laplaciano normalizada para a primeira comunidade (no sentido de cima para baixo). Apesar de ficar claro quem são as possíveis comunidades, nota-se um grande número de conexões entre alguns pares de comunidades. Além disso, as comunidades encontradas também têm estrutura de comunidades, como pode ser visto pelos autovalores encontrados. Este aspecto recursivo deste tipo de rede impossibilita, em geral, sua classificação como estrutura de comunidades.

[7] nesta figura, cada ponto (a, b) do gráfico representa a existência de uma aresta entre os nós a e b.

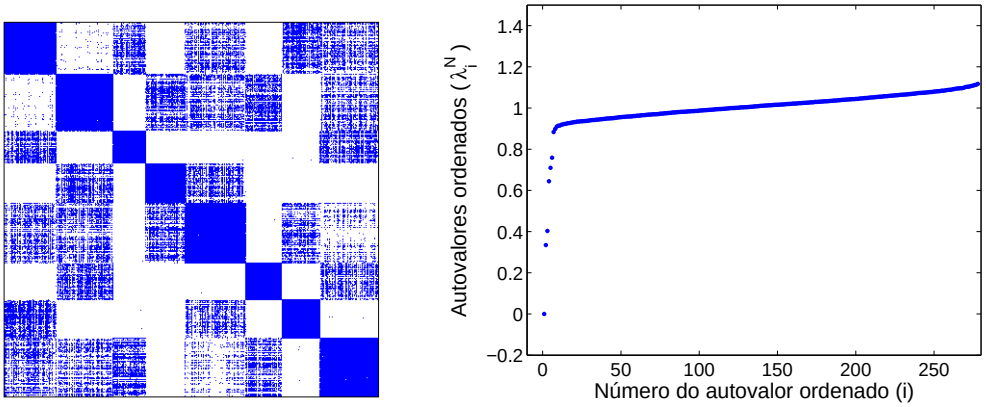


Figura 3.13: Rede espacial após reordenamento pelas comunidades (esquerda) e autovalores ordenados da primeira comunidade (direita).

3.4 O uso de Redes em Epidemiologia

O estudo de epidemias em redes livres de escalas baseadas no modelo BA pode ser encontrado em [81, 82, 83, 12, 49] e um estudo do processo de imunização em [84]. Estes estudos mostram uma diferença na evolução de epidemias em redes livres de escala quando comparadas a redes puramente aleatórias.

Estudos mais recentes de epidemias em redes livres de escala levam em consideração o coeficiente de aglomeração como um aspecto importante para tratar o assunto [92, 29, 67, 79, 77, 8]. Segundo [67], o coeficiente de aglomeração é o fator dominante no controle da taxa de crescimento da epidemia.

Capítulo 4

Modelagem de Epidemias

4.1 Introdução

Um modelo matemático em epidemiologia deve ser capaz, inicialmente, de descrever a situação atual de uma epidemia em uma população. Uma vez que tenha esta capacidade, deve ser capaz, ainda, de predizer futuras alterações no sistema epidemiológico.

Modelos matemáticos de doenças infecciosas são baseados principalmente na condição inicial da epidemia, em taxas de transferência e em uma estrutura qualitativa básica. Eles são a interface entre problemas reais da epidemiologia e a formulação teórica que pode ser manipulada matemática e computacionalmente.

O modelo mais utilizado é o SIR (suscetível-infectado-recuperado), que divide a população em três classes disjuntas: (S) suscetíveis, (I) infectados e (R) recuperados. Este modelo é encontrado na literatura também com o nome K-M, em homenagem a seus primeiros proponentes Kermack e McKendrick [55], e descreve a evolução temporal do tamanho de cada classe epidemiológica por meio de equações diferenciais. O modelo clássico SIR considera que a distribuição de indivíduos é espacial e temporalmente homogênea [45]. Este pode ser um motivo pelo qual ele não é capaz de explicar a persistência, ou erradicação, de algumas doenças infecciosas [81, 82].

Uma nova abordagem para a modelagem de sistemas ecológicos surge nas últimas décadas: os modelos baseados em indivíduo (MBI). Os MBIs têm sido utilizados há mais de trinta anos [51], porém, tornaram-se mais conhecidos em 1988 com o trabalho de Houston *et al.* [48]. A partir de então, os MBIs têm tido crescente utilização [41, 54]. Seu uso é promissor dentro de diversas áreas, apresentando implicações teóricas importantes e mostrando ser uma ferramenta poderosa para contornar dificuldades presentes em epidemiologia.

Em seu trabalho, Henrique Giacomini [38] discute as seguintes sete motivações para o uso do MBI em ecologia:

1. a grande complexidade de sistemas ecológicos, com consequentes dificuldades em sua análise matemática formal;

2. emergência de processos populacionais, resultando das interações entre indivíduos e destes com o meio;

3. poder de predição;

4. a adoção de uma visão evolutiva;

5. indivíduos são tratados de forma discreta;

6. interações são localizadas no espaço e

7. indivíduos podem possuir diferenças entre si.

Segundo [38], devido ao MBI ser baseado em um conjunto de regras matemáticas e algoritmos computacionais, *cada MBI pode ser desenvolvido, descrito e entendido de forma diferente por diferentes pessoas. Em trabalhos publicados, a descrição de um MBI é em maior parte verbal, e pela limitação de espaço nos textos dos periódicos, quase nunca contempla todos os detalhes.*

Desta forma, apresentamos na Seção 4.2 o modelo SIR. Na Seção 4.3, apresentamos o MBI e demonstramos sua equivalência, em média, ao modelo SIR. Os modelos SIR e MBI servem de base para o desenvolvimento de modelos de epidemias em redes neste trabalho. Apresentamos, também, modelos multicidades na Seção 4.4, que servirão para estudos de vigilância do Capítulo 6.

4.2 Modelo SIR

O modelo SIR, considerando uma população de tamanho constante, é desenvolvido baseado nas seguintes definições:

- S, I e R: número de indivíduos suscetíveis, infectados e recuperados respectivamente;
- N: tamanho total da população;
- β: *taxa de contato* - também chamado coeficiente de transmissão, é o número médio de contatos, de um indivíduo por unidade de tempo, em que há a transmissão da doença;
- γ: *taxa de recuperação* - γ^{-1} representa o tempo médio que um indivíduo fica na classe de infectado;
- μ: *taxa de renovação* - número de indivíduos, por unidade de tempo, que morrem e, neste mesmo número, nascem outros suscetíveis;

O fluxograma da Figura 4.1 ilustra o modelo SIR. Neste, o número total de indivíduos que morrem, $\mu S(t)+\mu I(t)+\mu R(t)$ (setas saindo de S, I e R), é igual ao número de indivíduos que são acrescidos, $\mu N(t)$, na classe de suscetíveis (seta que chega em S), mantendo a população com tamanho constante N. Pela definição de β, $\beta I(t)/N$ é o número médio de contatos, de indivíduos infectados com suscetíveis por unidade de tempo, desta forma, $\beta I(t)S(t)/N$ é o número de novas infecções por unidade de tempo (seta que chega em I). Após a duraçao da infecção, γ^{-1}, o indivíduo infectado se recupera e passa para a classe recuperado (seta que chega em R).

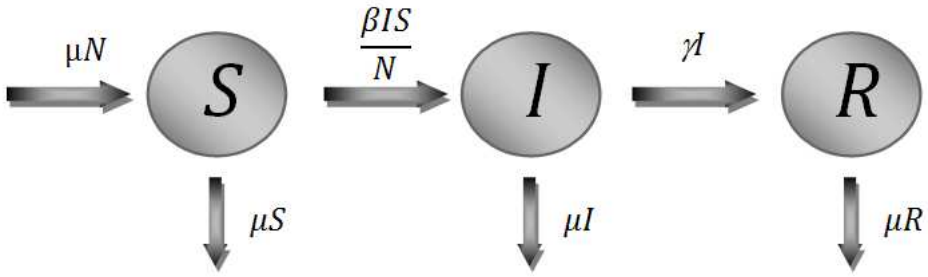


Figura 4.1: *Representação esquemática do modelo SIR*

Desta forma, o modelo SIR é definido pelo seguinte sistema de equações diferenciais:

$$\begin{aligned} \frac{dS}{dt} &= \mu N - \mu S - \frac{\beta IS}{N}, \quad S(0) = S_o \geq 0, \\ \frac{dI}{dt} &= \frac{\beta IS}{N} - \gamma I - \mu I, \quad I(0) = I_o \geq 0, \\ \frac{dR}{dt} &= \gamma I - \mu R, \quad R(0) = R_o \geq 0, \end{aligned} \tag{4.1}$$

em que $S(t) + I(t) + R(t) = N$ para todo $t \geq 0$.
Para o modelo SIR a reprodutibilidade basal é dada por:

$$R_0 = \frac{\beta}{\mu + \gamma}.$$

A força de infecção é dada por:

$$\lambda = \beta I.$$

É possivel mostrar que, se $R_0 \leq 1$, então, a partir de algum instante no tempo, o número de infectados será decrescente e a epidemia se erradicará. No caso em que $R_0 > 1$ a epidemia persistirá na população [45].

4.3 Modelo Baseado em Indivíduos (MBI)

Um modelo baseado em indivíduos (MBI) consiste de uma estrutura na qual ocorrem os relacionamentos de um certo número de indivíduos, cujo comportamento é determinado por um conjunto de características. À medida que os relacionamentos ocorrem, as características de cada indivíduo podem mudar. Neste trabalho, utilizaremos o MBI desenvolvido em [74], formulado de forma a reproduzir as premissas utilizadas na formulação do modelo SIR, que considera um tratamento estocástico dos contatos entre os indivíduos da população. O modelo MBI é desenvolvido a partir das seguintes premissas epidemiológicas:

1. *Tamanho da população*: a população tem tamanho constante N;

2. *Características do indivíduo*: cada indivíduo possui como única característica seu estado com relação à epidemia, e assume um e somente um dos estados;

3. *Distribuição estatística*: a mortalidade e recuperação dos indivíduos seguem uma distribuição uniforme em cada intervalo de tempo e, além disso, todos os indivíduos têm a mesma probabilidade de contato;

4. *Regras para mudança de estado*: em cada instante de tempo, cada indivíduo pode mudar de estado, conforme as seguintes regras:

   - $S, I, R \longrightarrow S$: ocorre a renovação dos indivíduos, ou seja, eles morrem e, para que a população permaneça de tamanho constante, um novo indivíduo suscetível nasce;
   - $S \longrightarrow I$: ocorre quando um indivíduo suscetível tem contato com um indivíduo infectado, adquirindo a doença e passando para o estado I;
   - $I \longrightarrow R$: ocorre a recuperação de um indivíduo infectado, tornando-se imune a uma nova infecção.

O procedimento para a evolução de um passo no tempo Δ_t, para cada indivíduo da população, se resume a:

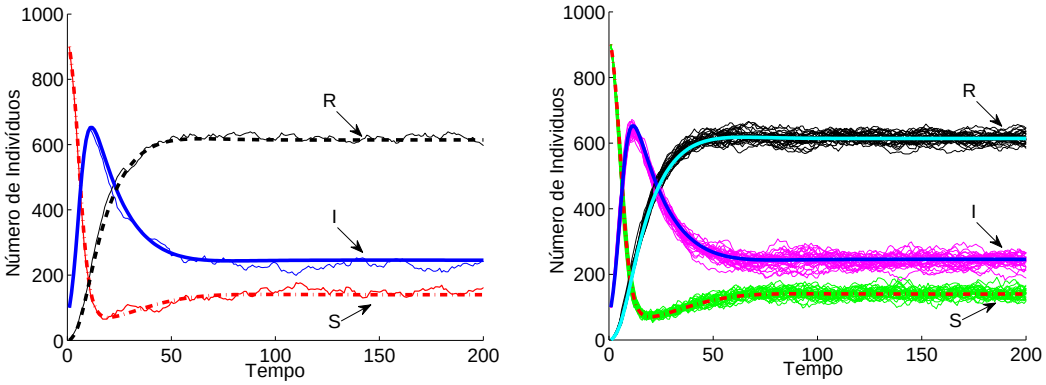
- sorteia-se aleatoriamente sua renovação (sua morte e nascimento de um indivíduo suscetível) com probabilidade $\mu\Delta_t$;
- se o indivíduo estiver infectado e caso não tenha ocorrido sua renovação, sorteia-se com probabilidade $\frac{\beta\Delta_t}{N-1}$ dentre os indivíduos suscetíveis aqueles que serão infectados por ele;
- ainda se o indivíduo estiver infectado, com probabilidade $\gamma\Delta_t$ sorteia-se aleatoriamente sua recuperação.

4.3.1 Equivalência entre os modelos SIR e MBI

A Figura 4.2 ilusta a equivalência entre o modelo de equações diferenciais SIR e o MBI.

Nesta seção, mostraremos que, em média, o MBI converge para o modelo SIR. Definimos os seguintes eventos ocorridos no intervalo de tempo $(k\Delta_t, (k+1)\Delta_t)$:

- $\Delta(k)$: o evento de um indivíduo da classe I se recuperar, passando para a classe S;
- $\Omega(k)$: o evento de um indivíduo morrer e, consequentemente, nascer outro na classe S;
- $\Gamma(k)$: o evento de se sortear um indivíduo da classe S;



(a) Execução do MBI uma vez e solução pelo modelo SIR

(b) Superposição da execução do MBI 30 vezes e solução pelo modelo SIR

Figura 4.2: *Equivalência SIR/MBI, considerando* $N = 1000$, $\mu = \frac{1}{50}$, $\gamma = \frac{1}{20}$, $\beta = \frac{1}{2}$, $S_0 = 900$, $I_0 = 100$ *e* $R_0 = 0$. *As linhas mais grossas representam a solução do modelo SIR e as linhas mais finas, o modelo MBI.*

- $\Phi(k)$: o evento de um indivíduo ter um contato com um indivíduo da classe I;
- $\Lambda(k)$: o evento de um indivíduo da classe S passar para a classe I, devido ao contato com outro indivíduo da classe I.

A probabilidade de ocorrer a recuperação de um indivíduo é dada por:

$$P(\Delta(k)) = \gamma\Delta_t. \tag{4.2}$$

A probabilidade de ocorrer a renovação de um indivíduo é dada por:

$$P(\Omega(k)) = \mu\Delta_t. \tag{4.3}$$

A probabilidade da ocorrência de novas infecções é a probabilidade de ocorrer simultanemente os eventos Γ e Φ, ou seja:

$$P(\Lambda(k)) = P(\Gamma(k) \cap \Phi(k)). \tag{4.4}$$

Por serem eventos independentes, podemos reescrever:

$$P(\Lambda(k)) = P(\Gamma(k))P(\Phi(k)). \tag{4.5}$$

A probablidade de sorteio de um indivíduo suscetível é dada por:

$$P(\Gamma(k)) = \frac{S(k)}{N}. \tag{4.6}$$

A probabilidade de acontecer um contato com um indivíduo infectado é proporcional ao número de infectados presentes naquele momento, sendo dada por:

$$P(\Phi(k)) = \frac{\beta\Delta_t}{N-1} I(k). \tag{4.7}$$

Das equações (4.5), (4.6) e (4.7) concluímos que:

$$P(\Lambda(k)) = \frac{\beta\Delta_t}{N-1} \frac{S(k)I(k)}{N}. \tag{4.8}$$

Calculemos, agora, os números médios de mudanças de classes ocorridos. O número médio de indivíduos, por unidade de tempo, da classe I que se recuperam da infecção e passam para a classe R é calculado a partir da equação (4.2) e é dado por:

$$\gamma\Delta_t I(k). \tag{4.9}$$

Os números médios de indivíduos, por unidade de tempo, das classes S, I e R que morrem e são substituídos por outros da classe S são calculados a partir da equação (4.3), sendo dados respectivamente por

$$\mu\Delta_t S(k) \tag{4.10}$$
$$\mu\Delta_t I(k) \tag{4.11}$$
$$\mu\Delta_t R(k) \tag{4.12}$$

Se um experimento é realizado X vezes, X suficientemente grande, e verifica-se a ocorrência de certo evento E vezes, então, segundo [80], a probabilidade P_e deste evento ocorrer se aproxima de:

$$\frac{E}{X}. \tag{4.13}$$

No MBI, em cada intervalo de tempo, executamos o algoritmo descrito uma vez para cada indivíduo, ou seja, N vezes. A probabilidade da infecção ocorrer é dada por $P(\Lambda(k))$ na equação (4.8). Fazendo $X = N$ e $P_e = P(\Lambda(k))$ na equação (4.13), concluímos que o número médio de indivíduos, por unidade de tempo, que passam da classe S para a classe I é dado por:

$$N\frac{\beta\Delta_t}{N-1} \frac{S(k)I(k)}{N}. \tag{4.14}$$

Para valores de N razoavelmente grandes podemos considerar $N/(N-1) \approx 1$, e reescrevermos o número médio de novas infecções como:

$$\frac{\beta\Delta_t S(k)I(k)}{N}. \tag{4.15}$$

As equações (4.9), (4.10), (4.11), (4.12) e (4.15) permitem descrever a dinâmica do MBI:

- suscetíveis: recebe o acréscimo de todos os indivíduos renovados ($\mu\Delta_t S + \mu\Delta_t I + \mu\Delta_t R$), e lembrando que $S(k) + I(k) + R(k) = N$, é dado por $\mu\Delta_t N$. Terá, ainda, a perda dos seus próprios indivíduos renovados

$(\mu\Delta_t S)$ e dos indíviduos que foram infectados $(\beta\Delta_t I(k)S(k)/N)$;

- infectados: terá a perda dos indivíduos recuperados $(\gamma\Delta_t I)$ e renovados $(\mu\Delta_t I)$ e o acréscimo das novas infecções $(\beta\Delta_t I(k)S(k)/N)$;
- recuperados: terá a perda dos indivíduos renovados $(\mu\Delta_t I)$ e o acréscimo dos indivíduos que se recuperaram da infecção $(\gamma\Delta_t I)$.

Matematicamente podemos, então, escrever a dinâmica associada ao comportamento médio do MBI:

$$\begin{aligned} S(k+1) = &\quad S(k) + \mu\Delta_t N - \mu\Delta_t S(k) - \frac{\beta\Delta_t I(k)S(k)}{N}, \\ I(k+1) = &\quad I(k) - \mu\Delta_t I(k) - \gamma\Delta_t I(k) + \frac{\beta\Delta_t I(k)S(k)}{N}, \\ R(k+1) = &\quad R(k) + \gamma\Delta_t I(k) - \mu\Delta_t R(k). \end{aligned} \tag{4.16}$$

Lembrando que k representa o instante de tempo $k\Delta_t$ e que

$$\frac{df}{dt} = \lim_{\Delta_t \to 0} \frac{f(t+\Delta_t) - f(t)}{\Delta_t}$$

notamos que a equação (4.16) coincide com a equação de diferenças da equação (4.1) utilizada na obtenção de soluções do modelo SIR por integração numérica. Assim, o modelo MBI converge em média para o modelo SIR.

A maior motivação para o uso do MBI é a facilidade em se agregar novas características aos indivíduos, como por exemplo sexo, idade e em especial o uso de redes que representem seu contatos, conforme mostraremos na Seção 5.1.

4.3.2 Estimação dos Parâmetros de uma Epidemia

Um vez conhecida uma série temporal que represente o número de indivíduos infectados, estimaremos os parâmetros que melhor adequam o modelo SIR aos dados observados.

Dada uma série temporal de indivíduos infectados I_m e um conjunto de parâmetros $\wp$, definiremos como função custo $C(I_m, \wp)$ o erro quadrático entre I_m e a curva de infectados obtida pelo modelo SIR com os parâmetros $\wp$, que denotaremos por $I_\wp$. Assim:

$$C(I_m, \wp) = \sum_i (I_m(i) - I_\wp(i))^2 \tag{4.17}$$

em que i representa o instante de tempo. O conjunto de parâmetros $\widehat{\wp}$, cujo

modelo SIR melhor representa I_m, será dado por:

$$\widehat{\wp} = \arg\min_{\wp} C(I_m, \wp) \tag{4.18}$$

Na equação (4.18), $\widehat{\wp}$ poder ser obtido por vários métodos de otimização. Neste trabalho optamos pelo método do gradiente. Utilizaremos esta estimação de parâmetros para analisar os resultados obtidos na Seção 5.2.

Epidemias que ocorrem em cidades diferentes podem ter comportamentos diferentes, seja por questões sanitárias, econômicas, imunológicas, alimentação ou mesmo hábitos diversos. Além disso, se existem indivíduos infectados em uma cidade, existe a chance da infecção se espalhar para as cidades em que seus indivíduos tenham contato com tais indivíduos infectados, e assim por diante. Tais premissas dificultam encontrar um modelo SIR que leve em consideração toda a região de estudo como única. Uma proposta para tal modelo é a modelagem multi-cidades e que trate cada cidade independentemente e os contatos entre os indivíduos das diversas cidades.

4.4 Modelos Multi-Cidades Baseados no SIR

Os modelos apresentados nas Seções 4.2 e 4.3 são formulados levando-se em consideração que toda a população de estudo encontra-se numa mesma região, não havendo distinção quanto à posição espacial dos indivíduos.

O crescimento das grandes cidades tem como efeitos aspectos que devem ser levados em consideração nos modelos epidemiológicos. Indivíduos das classes mais baixas, geralmente trabalhadores nos centros das grandes cidades, devido ao alto custo de vida, são obrigados a morarem em outras cidades, locomovendo-se diariamente à cidade de trabalho. Isso gera um efeito interessante: os grandes centros tem um população oscilante muito grande durante o dia e uma pequena população durante a noite. As cidades próximas têm um efeito contrário.

Vários estudos e modelos podem ser encontrados na literatura [99, 7, 20, 15, 21, 96]. Nas próximas seções apresentamos um modelo que leve em consideração os aspectos acima.

4.4.1 Modelo SIR Multi-Cidades

O modelo que trataremos baseia-se nas mesmas premissas do modelo SIR, acrescido das premissas a seguir:

- Cada cidade i possui população residente constante N_i;
- Infecções ocorrem na cidade de trabalho e acontecem segundo o modelo SIR;
- Cada cidade possui taxas de infecção que podem ser distintas entre sí.

Definimos como:

- S_i, I_i e R_i respectivamente os números de indivíduos suscetíveis, infectados e recuperados da cidade i;

- γ_i e ρ_i respectivamente as taxas de recuperação e renovação da cidade i;
- β_i a taxa de infecção na cidade i;
- p_{ij} a fração de indivíduos da cidade i que trabalha na cidade j.

Como cada cidade tem sua dinâmica epidêmica modelada por um modelo SIR com seu parâmetros distintos, uma vez que cada modelo tem três equações diferenciais, um modelo para n cidades será um sistema de equações diferenciais com $3n$ equações.

Para o caso de n cidades com fluxo de indivíduos:

$$\begin{aligned}
\frac{dS_i}{dt} &= \rho_i N_i - \rho_i S_i - S_i \textstyle\sum_{j=1}^{n} I_j A_{ij}, \quad S_i(0) = S_i^o \geq 0, \\
\frac{dI_i}{dt} &= S_i \textstyle\sum_{j=1}^{n} I_j A_{ij} - \gamma_i I_i - \rho_i I_i, \quad I_i(0) = I_i^o \geq 0, \\
\frac{dR_i}{dt} &= \gamma_i I_i - \rho_i R_i, \quad R_i(0) = R_i^o \geq 0.
\end{aligned}$$

para $i = 1, 2, ..., n$, no qual $A_{ij} = \displaystyle\sum_{k=1}^{n} \left(\beta_k \frac{p_{ik} p_{jk}}{\sum_{l=1}^{n} N_l p_{lk}} \right) = A_{ji}$.

A Figura 4.3 ilustra a solução do sistema acima para o número de infectados de 20 cidades com matrizes de fluxo p_{ij} distintas mostradas abaixo, e com parâmetros idênticos para todas as cidades: $N_i = 1000$, $\gamma_i = 0,05$, $\rho_i = 0,01$ e $\beta_i = 0,3$ para $i = 1, ..., 20$ e valores iniciais de $S_1^o = 900$, $I_1^o = 100$ e para $i = 2, ..., 20$ com $S_i^o = 1000$ e $I_i^o = 0$.

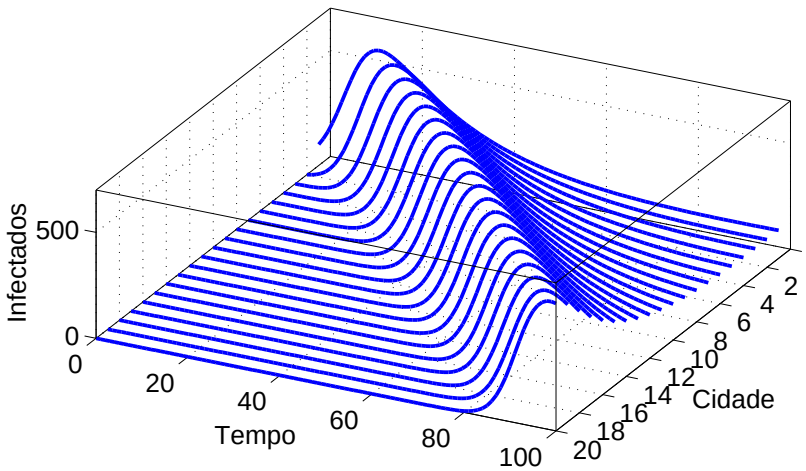


Figura 4.3: Solução particular para o número de infectados. O atraso do início da epidemia nas cidades mais distantes é claramante notado.

A matriz de fluxo com a qual foi gerada a Figura 4.3 é composta por $0,9$ na diagonal principal e $0,1$ na diagonal superior, representando o caso em que as cidades estão alinhadas (do ponto de vista do fluxo de indivíduos), em que cada cidade durante o horário de trabalho permanece com 90% de sua população

original e recebe 10% dos indivíduos da cidade anterior, exceto pela cidade 20 que não possui indivíduos que se locomovem para outras cidades.

4.4.2 Modelo SIR Multi-Cidades Estocástico

Equações diferenciais estocásticas[1] (SDE) têm sido usadas em modelos em diversas áreas de aplicação incluindo biologia, química, epidemiologia, mecânica, microeletrônica, economia e finanças.

Do ponto de vista de simulação, vários modelos epidemiológicos têm a transição de estados dos indivíduos baseada em sorteios, nos quais, em caso de sucesso do sorteio, o indivíduo muda de estado. Então, o número de indivíduos que mudam de estado, para cada tipo de transição, é uma variável aleatória com distribuição binomial $b(M, u)$, onde M é o número de indivíduos um passo antes do tempo em que se deseja calcular e u a probabilidade de sucesso da transição de estado. Uma vez que tal distribuição pode ser aproximada por $b(M, u) \approx Mu + \sqrt{Mu(1-u)}N(0,1)$ para M suficientemente grande, definimos

$$G(M, u) = \sqrt{Mu(1-u)}$$

e descrevemos nosso modelo epidemilógico pelo seguinte sistema de equações diferenciais estocásticas:

$$\begin{aligned}
dS_i = \; & (\rho_i N_i - \rho_i S_i - S_i \textstyle\sum_{j=1}^{n}(I_j A_{ij}))dt + \\
& + G(I_i, \rho_i)dW_i^2 + G(R_i, \rho_i)dW_i^3 - (\textstyle\sum_{k=1}^{n} G_{ik})dW_i^4, \\
dI_i = \; & (S_i \textstyle\sum_{j=1}^{n}(I_j A_{ij}) - \gamma_i I_i - \rho_i I_i)dt - \\
& - G(I_i, \rho_i)dW_i^2 - G(i_i, \gamma_i)dW_i^1 + (\textstyle\sum_{k=1}^{n} G_{ik})dW_i^4, \qquad (4.19) \\
dR_i = \; & (\gamma_i I_i - \rho_i R_i)dt + \\
& + G(i_i, \gamma_i)dW_i^1 - G(R_i, \rho_i)dW_i^3, \\
& S_i(0) = S_i^o \geq 0, I_i(0) = I_i^o \geq 0 \text{ and } R_i(0) = R_i^o \geq 0.
\end{aligned}$$

no qual

$$G_{ik} = G\left(S_i p_{ik} \sum_{j=1}^{n}(I_j p_{jk}), \frac{\beta_k}{\sum_{l=1}^{n} N_l p_{lk}}\right)$$

W_i^k é um movimento Browniano padrão (ou um processo de Wiener), que do ponto de vista de discretização para cálculos computacionais resume-se a:

$$W_i^k(t) = W_i^k(t - \Delta t) + dW_i^k(t)$$

em que cada $dW_i^k(t)$ é uma variável aleatória independente da forma $\sqrt{\Delta t}N(0,1)$. O sistema é resolvido pela integral de Itô, que computacionalmente pode ser visto como o método de Euler explícito. Uma introdução prática e acessível à

[1] Do inglês: Stochastic differential equation

simulação numérica de equações diferenciais estocásticas pode ser encontrada em [46].

Em sua formulação original, os modelos MBI e SIR possuem uma representação da população e suas relações dada por um grafo completo, ou seja, todos os indivíduos podem se relacionar entre si. Por não ser uma característica observada na maioria das relações de uma sociedade, utilizaremos as redes sociais apresentadas no Capítulo 3, que modelam tais relações e serão usadas nos modelos que apresentamos a seguir.

Capítulo 5

Modelos Epidemiológicos em Redes

Um fator importante a se levar em consideração na modelagem de epidemias é o comportamento social dos indivíduos. O modelo SIR considera que os contatos entre os indivíduos são puramente aleatórios, ou seja, todos os indivíduos tem a mesma probabilidade de relacionarem entre si. Esta não é uma característica comumente encontrada nas sociedades atuais, onde encontramos indivíduos com os mais variados números e tipos de relacionamentos. Tais relacionamentos são, geralmente, modelados por meio de grafos (também denominados redes sociais, ou simplesmente redes), sendo que o comportamento de epidemias em sociedades com diferentes estruturas tem sido modelado por redes sociais em vários trabalhos [81, 82, 12, 49, 92, 29, 67, 79, 77, 8]. Apesar de ser complexa a inserção de redes sociais no modelo SIR, é extremamente simples no MBI, o que faz deste uma ferramenta ainda mais valiosa.

Neste capítulo propomos um modelo MBI sobre redes (Seção 5.1), utilizando-o em experimentos numéricos que sugerem dificuldades na adequação do modelo SIR clássico em redes sociais através de simples ajuste de parâmetros (Seção 5.2), indicando a necessidade de modelos mais complexos. Apresentamos os modelos atualmente utilizados para modelagem de epidemias em redes (Seções 5.3 e 5.4), os quais tem limitações para redes com algum tipo de estruturação, como por exemplo a estrutura em comunidades. A fim de se obter um modelo capaz de descrever o comportamento de epidemias em redes de forma equivalente ao modelo SIR no caso de redes quaisquer propomos o modelo analítico μSIR na Seção 5.5 e mostramos seus resultados na Seção 5.6. Propomos, também, um modelo que necessite de menos informações que toda a rede, o HMF-MC, na Seção 5.7 e mostramos alguns experimentos na Seção 5.8.

5.1 MBI sobre Redes

Os modelos SIR e MBI em sua formulação original têm as relações entre indivíduos representadas por uma grafo completo[1]. Contudo, buscamos modelos em que as relações (arestas) entre os indivíduos (vértices) da população são representadas por uma rede qualquer.

Propomos, então, uma nova versão do MBI, que considera que os indivíduos do MBI se organizam em uma rede, ou seja, para cada indivíduo temos um

[1]Grafo completo: todo vértice possui arestas ligando-o com todos os outros vértices do grafo.

vértice correspondente associado e a cada par de indivíduos associamos uma aresta se e somente se eles tem contato entre si. Do ponto de vista do algoritmo, no segundo passo do algoritmo do MBI apresentado na Seção 4.3, mesmo que um indivíduo seja sorteado para ter contato com um infectado, a infecção não ocorrerá caso não exista uma aresta entre eles.

5.2 Estudo da adequação do modelo SIR clássico às redes

No MBI original e no modelo SIR, a taxa de contato β pressupõe que todos os indivíduos têm contato entre si. Dada uma rede complexa com densidade d, $0 < d \leq 1$ e taxa de contato entre indivíduos β, é intuitivo esperar que os resultados das simulações pelo MBI sobre tal rede nos forneçam uma dinâmica que possa ser descrita por uma nova taxa de contato $\overline{\beta} = d\beta$ e pelos parâmetros originais γ e μ, uma vez que estes dois útimos não têm relação com a estrutura social na qual a epidemia está inserida.

Para verificarmos os resultados pretendidos fizemos seis tipos de simulações:

1. MBI sobre de uma rede complexa gaussiana;
2. MBI sobre uma rede complexa BA;
3. MBI sobre uma rede complexa DM;
4. MBI sobre uma rede complexa preferencial;
5. MBI sobre uma rede complexa espacial.

Para cada tipo de simulação na qual o MBI foi executado sobre uma rede complexa, utilizamos uma única rede correspondente. Nas redes BA e DM geradas encontramos leis de potência com expoente $3,01$ e $2,27$ respectivamente.

Para cada simulação fizemos 1000 execuções. Estimamos para cada execução os parâmetros γ, β e μ, conforme equação (4.18), e fizemos a análise estatística sobre a hipótese nula de que a média dos valores encontrados seja o valor teórico e cuja hipótese alternativa seja que os valores encontrados sejam diferentes do teórico.

Além da análise anterior, no intuito de verificar qual o melhor modelo SIR que se adequa às execuções do MBI, para cada tipo de simulação feita, obtivemos as curvas médias de infectados e suscetíveis. A partir da curva média de infectados, estimamos os parâmetros do modelo SIR. Fizemos, então, a comparação entre as curvas médias obtidas pelas execuções do MBI e as curvas obtidas pelo modelo SIR com os parâmetros estimados. Uma vez que a população tem tamanho constante, a análise da curva de indivíduos recuperados foi omitida pois podemos encontrar R a partir da equação $R(t) = N - I(t) - S(t)$.

Os valores utilizados nas simulações foram: $N = 2000$, $S_0 = 1990$, $I_0 = 10$, $R_0 = N - S_0 - I_0$, $\gamma = 0,05$, $\beta = 1,5$ e $\mu = 0,01$. Utilizamos redes com densidade aproximada $d = 0,07$.

5.2.1 MBI sobre redes complexas

Análise estatística dos parâmetros.

Os gráficos da Figura 5.1 foram gerados após execução do MBI sobre as redes gaussiana, BA, DM, preferencial e espacial respectivamente, e mostram os histogramas de frequência relativa para $\hat{\gamma}$, $\hat{\beta}$ e $\hat{\mu}$, encontrados conforme a estimação dos parâmetros γ, β e μ, respectivamente. Contém também, sobrepostas aos histogramas, as gaussianas de mesmas médias e desvios padrão dos parâmetros estimados, seu valor teórico e os valores críticos para 5% de significância dos testes.

Assim como feito para o MBI em sua formulação original, definimos para as simulações feitas as seguintes hipóteses nulas:

$$\begin{aligned} H_0^{\gamma}\colon &\quad \hat{\gamma} = \gamma \\ H_0^{\beta}\colon &\quad \hat{\beta} = \overline{\beta} = d\beta \\ H_0^{\mu}\colon &\quad \hat{\mu} = \mu\ . \end{aligned}$$

A partir dos valores estimados verificamos que:

- considerando o MBI sobre as redes gaussiana e espacial: não rejeitamos nenhuma das hipóteses H_0^{γ}, H_0^{β} e H_0^{μ} com nível de significância de 5%;
- considerando o MBI sobre as redes BA, DM e preferencial: não rejeitamos a hipótese H_0^{μ}, contudo rejeitamos H_0^{γ} e H_0^{β} com nível de significância de 5%.

Dinâmica. Os gráficos da Figura 5.2 ilustram as curvas de infectados e suscetíveis obtidas com a média das evoluções do MBI e as curvas do modelo SIR correspondente. O coeficiente de aglomeração da rede, a conectividade média $\langle k_{I_0} \rangle$ dos indivíduos infectados no tempo inicial e os parâmetros estimados são mostrados na Tabela 5.1.

Tipo	C	$\langle k_{I_0} \rangle$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\mu}$
Gaussiana	$0,07$	$140,1$	$0,0517$	$0,1056$	$0,0104$
BA	$0,148$	$136,7$	$0,0673$	$0,1504$	$0,0117$
DM	$0,28$	$136,7$	$0,0846$	$0,2029$	$0,0134$
Preferencial	$0,19$	$139,9$	$0,0739$	$0,1741$	$0,0121$
Espacial	$0,52$	$139,9$	$0,0518$	$0,1011$	$0,0112$

Tabela 5.1: Coeficientes de aglomeração, conectividade média da população inicial de infectados e parâmetros encontrados para as redes analisadas.

Análise estatística das curvas de suscetíveis.

Nesta análise, faremos um estudo sobre as curvas de suscetíveis da Figura 5.2 de forma a validar as diferenças encontradas pelo SIR com os parâmetros ajustados e as médias do MBI, obtendo uma confirmação do que os gráficos nos sugerem. Fizemos, então, como segue.

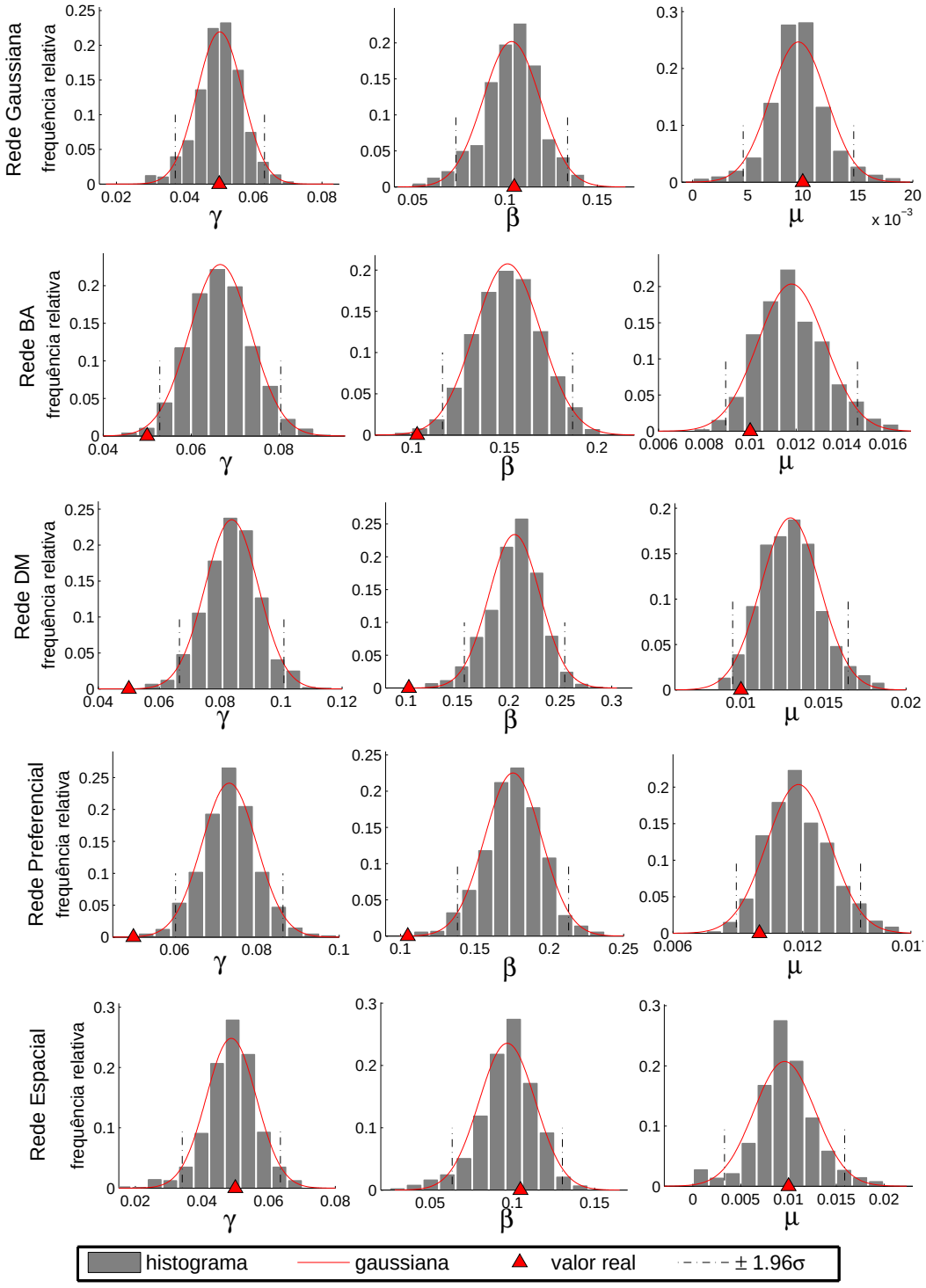


Figura 5.1: Histogramas dos parâmetros estimados para o MBI sob redes complexas. Curvas gaussianas de mesmas médias e desvios padrão são mostradas para efeito de visualização.

A partir dos parâmetros estimados da curva média de infectados do MBI, para cada tipo de simulação obtivemos a curva de suscetíveis do modelo SIR

estimado. A partir da Figura 5.2 notamos que, apesar do ajuste razoável das curvas do número de infectados do modelo SIR à média do MBI, as curvas dos números de suscetíveis tem diferenças notáveis nas redes BA, DM e preferencial. Este fato indica certa dificuldade no ajuste dos parâmetros do modelo SIR para estas redes.

Para cada curva de suscetíveis do MBI, definimos o resíduo[2] da curva como a soma dos resíduos em cada ponto da curva, ou seja:

$$R(S) = \sum S - S_{SIR}.$$

Com os mesmos parâmetros estimados do modelo SIR, fizemos simulações de Monte Carlo utilizando-se do MBI em sua formulação original para gerar a hipótese nula de igualdade da média dos resíduos. Denotaremos por R_R e R_M respectivamente como o resíduo do MBI em rede e do MBI em sua formulação original. As hipóteses alternativas utilizadas foram construídas de forma a verificar as diferenças encontradas nas curvas de suscetíveis da Figura 5.2.

Rede	$\mathbf{H_0}$	$\mathbf{H_1}$	**p-valor empírico**
Gaussiana	$R_R = R_M$	$R_R \neq R_M$	$0,85$
BA	$R_R = R_M$	$R_R > R_M$	$0,10$
DM	$R_R = R_M$	$R_R > R_M$	$0,00$
Preferencial	$R_R = R_M$	$R_R > R_M$	$0,00$
Espacial	$R_R = R_M$	$R_R < R_M$	$0,48$

Tabela 5.2: Análise estatística das curvas do número de suscetíveis obtidas pelo MBI em redes. Os p-valores indicam a percentagem de execuções do MBI em redes que levam a não rejeitar a hipótese nula H_0, desfavorável à hipótese alternativa H_1.

Na Tabela 5.2 mostramos os resultados obtidos desta análise. Os p-valores empíricos encontrados corroboram com os gráficos de suscetíveis da Figura 5.2. Estes resultados indicam dificuldades em se ajustar o modelo SIR para propagação de epidemias em algumas redes, o que nos leva a uma busca por modelos mais complexos.

[2]Definimos o resíduo desta forma na intenção de conseguirmos detectar curvas de suscetíveis sistematicamente acima ou abaixo das do modelo SIR.

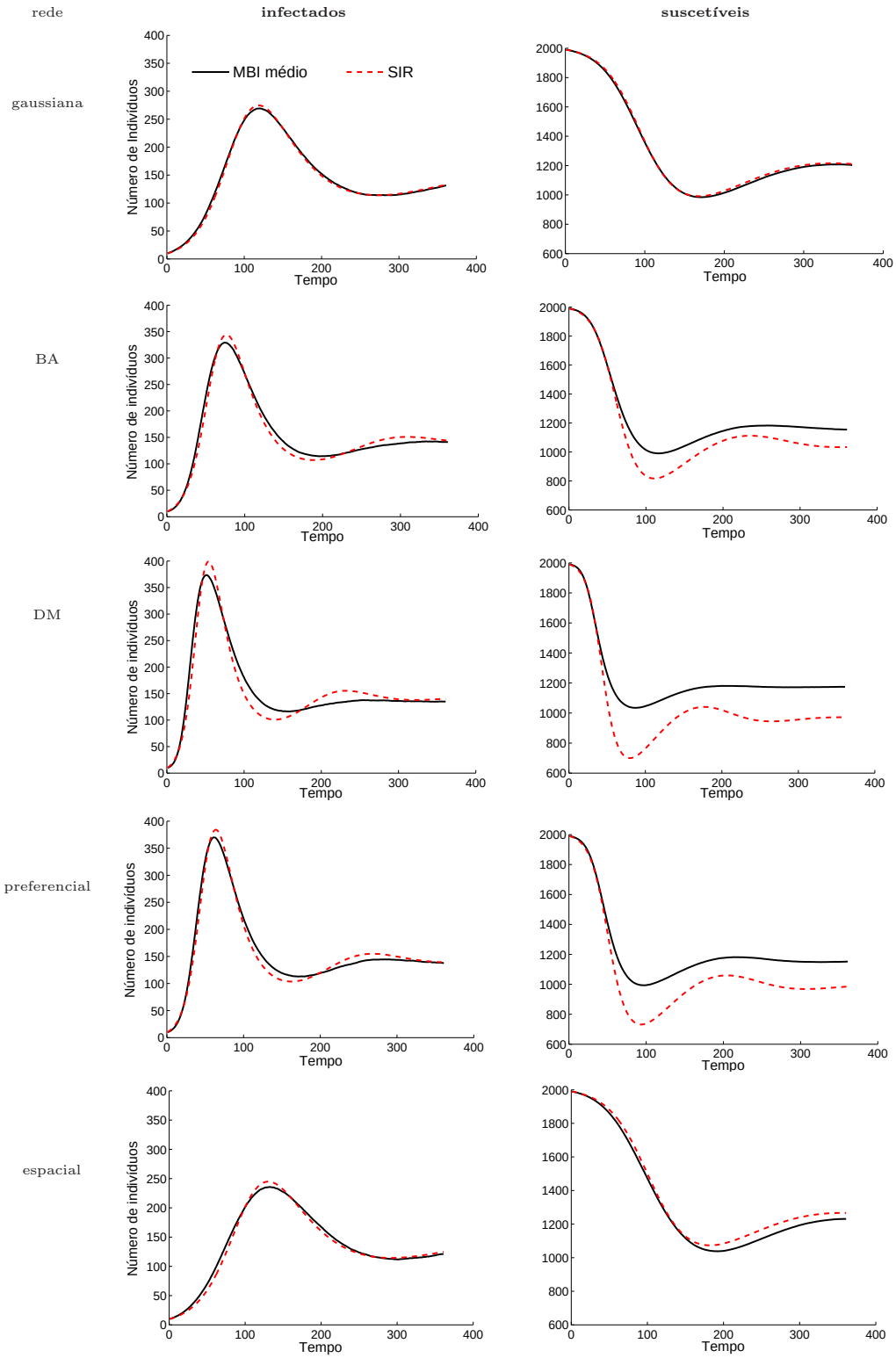


Figura 5.2: Comparação das curvas médias do MBI em rede.

5.2.2 Discussão

Utilizamos o MBI proposto sobre redes como uma alternativa ao modelo epidemiológico de equações diferenciais SIR, possibilitando o tratamento dos indivíduos de forma discreta. A utilização do MBI possibilitou, também , o estudo da propagação de uma epidemia relativamente ao tipo de rede complexa na qual a sociedade está organizada.

- as simulações feitas indicam que, caso uma sociedade se organize conforme uma rede complexa gaussiana, então o MBI agregado desta rede se comporta, em todos os tipos de simulações feitas, conforme um modelo SIR, com uma nova taxa de contato $\overline{\beta} = d\beta$. Este era um resultado esperado, e que novamente valida a metodologia empregada.

- as análises estatísticas feitas indicam que, caso uma sociedade se organize conforme uma rede complexa livre de escalas ou preferencial, o comportamento do MBI agregado desta rede encontra dificuldades em ser modelado pelo modelo SIR com apenas uma nova taxa de contato que preserve os significados de seus parâmetros;

- os resultados indicam que o modelo SIR dado pela equação (4.1) encontra dificuldades em modelar um epidemia em uma sociedade organizada como rede livre de escalas, preferencial ou espacial. Os gráficos da Figura 5.2 mostram diferenças, nas curvas de suscetíveis, entre a dinâmica da epidemia e o modelo SIR com os parâmetros estimados.

- uma comparação da distribuição de graus das redes gaussiana e espacial, Figuras 3.6 e 3.8, poderiam nos levar ingenuamente a crer que o comportamento de epidemias nestas redes fosse equivalente. A análise estatística de seus parâmetros, Figura 5.1, reforçaria esta hipótese. A análise das dinâmicas da epidemia nestas redes, Figura 5.2, ilustram uma pequena diferença na dinâmica entre as curvas e um pequeno distanciamento entre a curva média de suscetíveis da epidemia na rede espacial e o modelo SIR estimado. A maior diferença entre as duas redes está em seu coeficiente de aglomeração, Tabela 5.1, o que está de acordo com a afirmação, feita por [67], que o coeficiente de aglomeração seria o fator dominante no controle da taxa de crescimento da epidemia.

Nossos estudos numéricos indicam dificuldades em ajustar o modelo SIR para epidemias em algumas redes, o que poderia gerar campanhas de vacinação ineficazes ou com gastos desnecessários. Apresentamos nas seções que seguem os resultados obtidos para modelos mais complexos de epidemias em redes, de forma a se buscar um modelo equivalente ao SIR para redes.

5.3 Aproximação heterogênea de campo médio (HMF)

A aproximação heterogênea de campo médio[3] (HMF) classifica os nós da rede de acordo com sua conectividade (número de vizinhos) e considera que a conectividade média $\langle k \rangle$ dos nós é finita. Estudos do processo epidêmico em redes utilizando o HMF são encontrados em [81, 82, 68] e no estudo de fenômenos críticos em [24].

Denotamos por $s_k(t)$, $i_k(t)$ e $r_k(t)$ a respectiva densidade de nós suscetíveis, infectados e recuperados de conectividade k no instante de tempo t. Considerando redes estáticas, temos que $s_k(t) + i_k(t) + r_k(t) = 1 \ \forall t$ e $\forall k$.

Seja $P(k)$ a distribuição de conectividade dos nós de uma rede. Assim, a conectividade média é dada por $\langle k \rangle = \sum_k kP(k)$. As proporções globais de nós em cada estado são dadas por $s(t) = \sum_k P(k)s_k(t)$, $i(t) = \sum_k P(k)i_k(t)$ e $r(t) = \sum_k P(k)r_k(t)$.

A probabilidade escolher aleatoriamente um nó com s arestas é proporcional a $sP(s)$. Desta forma, a probabilidade de aleatoriamente uma aresta chegar a um indivíduo infectado é dada por

$$\Theta(t) = \frac{\sum_k kP(k)i_k(t)}{\sum_s sP(s)} = \frac{\sum_k kP(k)i_k(t)}{\langle k \rangle}. \tag{5.1}$$

Assim, temos o seguinte sistema de equações diferenciais acoplado:

$$\begin{aligned}
\frac{ds_k(t)}{dt} &= \mu - \mu s_k(t) - \tfrac{\beta}{N} k s_k(t)\Theta(t), & s_k(0) = s_k^o, \\
\frac{di_k(t)}{dt} &= \tfrac{\beta}{N} k s_k(t)\Theta(t) - \gamma i_k(t) - \mu i_k(t), & i_k(0) = i_k^o, \\
\frac{dr_k(t)}{dt} &= \gamma i_k(t) - \mu r_k(t), & r_k(0) = r_k^o,
\end{aligned} \tag{5.2}$$

$0 \leq s_k^o \leq 1$, $0 \leq i_k^o \leq 1$, $0 \leq r_k^o \leq 1$.

Este modelo considera que os nós de uma mesma classe de conectividade tem as mesmas propriedades dinâmicas. É util observar que esta formulação negligencia diversos aspectos da rede, como por exemplo o coeficiente de aglomeração, a conexidade etc.

5.4 Modelagem por Cadeias de Markov (MKV)

Neste modelo, trataremos o problema de forma microscópica ao nível do indivíduo, ou seja, consideraremos as probabilidades S_i, I_i e R_i de cada indivíduo i estar respectivamente em um dos estados suscetível, infectado e recuperado. Desta forma, para cada indivíduo i temos $S_i(t) + I_i(t) + R_i(t) = 1 \ \forall t$.

Chamaremos de matriz de contato $\mathcal{A}$ a matriz cujo elemento $\mathcal{A}_{ij}$ contem a probabilidade do indivíduo i ter contato com indivíduo j. Em redes sem peso e não direcionadas esta matriz se resume à matriz de adjacência ($\mathcal{A}_{ij} = 1$ se $i \leftrightarrow j$ e $\mathcal{A}_{ij} = 0$ caso contrário).

[3] Sigla proveniente do inglês: Heterogeneous Mean-Field (HMF)

Estes modelos são baseados no formalismo de cadeias de Markov em tempo discreto. A probabilidade do indivíduo i para ter um contato com o indivíduo contagioso j no intervalo de tempo Δ_t é dada por $\frac{\Delta_t \beta}{N}\mathcal{A}_{ij}I_j$. Considerando eventos independentes, a probabilidade do indivíduo i não ser infectado é dada por $\prod_j (1 - \frac{\Delta_t \beta}{N}\mathcal{A}_{ij}I_j)$. Portanto, a probabilidade de infecção do indivíduo suscetível i é dada por:

$$Y_i(t, \Delta_t) = 1 - \prod_j (1 - \frac{\Delta_t \beta}{N}\mathcal{A}_{ij}I_j). \tag{5.3}$$

O seguinte sistema de equações dinâmicas pode então ser escrito:

$$\begin{aligned} S_i(t+\Delta_t) &= S_i(t) + \Delta_t\mu - \Delta_t\mu S_i(t) - S_i(t)Y_i(t,\Delta_t), \\ I_i(t+\Delta_t) &= I_i(t) - \Delta_t\gamma I_i(t) - \Delta_t\mu I_i(t) + S_i(t)Y_i(t,\Delta_t), \\ R_i(t+\Delta_t) &= R_i(t) + \Delta_t\gamma I_i(t) - \Delta_t\mu R_i(t), \end{aligned} \tag{5.4}$$

$S_i(0) = S_i^o$, $I_i(0) = I_i^o$, $R_i(0) = R_i^o$.

Comumente cadeias de Markov são utilizadas com $\Delta_t = 1$. Distinguiremos este caso nomeando-o modelo GG , que foi utilizado como modelo SIS em [40, 42], porém mostraremos que tal escolha de Δ_t induziu resultados pouco satisfatórios. Apesar de em [42] serem usadas *ensembles*[4] e neste trabalho utilizarmos uma matriz de adjacência, afirmamos que os resultados são similares para os experimentos feitos. Para escolhas de $\Delta_t < 1$ denominaremos como modelo MKV. No trabalho [16] é considerado um passo de tempo infinitesimal Δ_t, porém, a escolha Δ_t depende dos parâmetros do problema, podendo tornar o problema intratável do ponto de vista computacional caso se necessite de valores pequenos demais. O procedimento é ilustrado no Algoritmo 1.

$(S(0), I(0), R(0)) \leftarrow (S_0, I_0, R_0)$
$(\bar{\gamma}, \bar{\mu}) \leftarrow (\Delta_t.\gamma, \Delta_t.\mu)$
for $k = 0 : \Delta_t : T_f - 1$ **do**
  for cada indivíduo i **do**
    $Q \leftarrow 1 - \prod_j (1 - \Delta_t \frac{\beta}{N}\mathcal{A}_{ij}I_j(k))$
    $S_i(k+\Delta_t) \leftarrow S_i(k) + \bar{\mu} - \bar{\mu}S_i(k) - QS_i(k)$
    $I_i(k+\Delta_t) \leftarrow I_i(k) + QS_i(k) - \bar{\gamma}I(k) - \bar{\mu}I_i(k)$
    $R_i(k+\Delta_t) \leftarrow R_i(k) + \bar{\gamma}I_i(k) - \bar{\mu}R_i(k)$
  end for
  $S(k+\Delta_t) \leftarrow \sum_i S_i(k+\Delta_t)$
  $I(k+\Delta_t) \leftarrow \sum_i I_i(k+\Delta_t)$
  $R(k+\Delta_t) \leftarrow \sum_i R_i(k+\Delta_t)$
end for

Algoritmo 1: MKV(μ, β, γ, N,S_0,I_0,R_0,T_f,$\mathcal{A}$,Δ_t)

5.5 O modelo μSIR

O uso de cadeias de Markov no estudo de propagação de epidemias em redes é relativamente novo. O seu uso para valores de Δ_t muito pequenos

[4]Uma ensemble é uma matriz que contem as probabilidades de ocorrências das arestas

pode tornar o problema intratável. Além disso, diante das dificuldades em adequar os parâmetros do modelo SIR clássico que apresentamos na Seção 5.2 é desejável se encontrar um modelo analítico que explique o comportamento de epidemias em redes de forma equivalente ao modelo SIR clássico.

Diante disso, propomos o modelo μSIR[5] com uma formulação por equações diferenciais que, assim como na modelagem por cadeias de Markov, trata os indivíduos de forma microscópica com a conveniência de não ser necessária a escolha a priori de um parâmetro Δ_t.

Seja Y_i a probabilidade do indivíduo i passar do estado suscetível para infectado. Sejam S_i, I_i e R_i as probabilidades de cada indivíduo i estar respectivamente em um dos estados suscetível, infectado e recuperado. Assim, podemos escrever o seguinte sistema de equações diferenciais acoplado:

$$\begin{aligned}
\frac{dS_i(t)}{dt} &= \mu - \mu S_i(t) - S_i(t)Y_i(t), \qquad S_i(0) = S_i^o, \\
\frac{dI_i(t)}{dt} &= S_i(t)Y_i(t) - \gamma I_i(t) - \mu I_i(t), \quad I_i(0) = I_i^o, \\
\frac{dR_i(t)}{dt} &= \gamma I_i(t) - \mu R_i(t), \qquad\qquad R_i(0) = R_i^o,
\end{aligned} \tag{5.5}$$

$0 \leq S_i^o \leq 1$, $0 \leq I_i^o \leq 1$, $0 \leq R_i^o \leq 1$.

Do ponto de vista macroscópico, os números de indivíduos suscetiveis $S(t)$, infectados $I(t)$ e recuperados $R(t)$ são dados respectivamente por: $S(t) = \sum_i S_i(t)$, $I(t) = \sum_i I_i(t)$ e $R(t) = \sum_i R_i(t)$.

Alguns trabalhos sobre modelos epidemiológicos estocásticos tratam as constantes γ, β e μ como probabilidades da ocorrêcias de seus respectivos eventos mas, do ponto de vista de equações diferenciais, são tratadas como taxas de transição por unidade de tempo. Desta forma, para um intervalo de tempo Δ_t, as taxas de renovação e recuperação são dadas respectivamente por $\Delta_t\mu$ e $\Delta_t\gamma$. A probabilidade do individuo i ter um contato infeccioso com o indivíduo j, em um intervalo Δ_t de tempo, é dada por $\Delta_t \frac{\beta}{N}\mathcal{A}_{ij}I_j$. Considerando eventos independentes, a probabilidade do indivíduo i não ser infectado é dada por $\prod_j (1 - \Delta_t \frac{\beta}{N}\mathcal{A}_{ij}I_j)$. A taxa $P_i(t, \Delta_t)$ de contatos infecciosos do indivíduo i será dada por:

$$P_i(t, \Delta_t) = 1 - \prod_j (1 - \Delta_t \frac{\beta}{N}\mathcal{A}_{ij}I_j(t))$$

Após cálculos algébricos podemos reescrever $P_i(t, \Delta_t)$:

$$P_i(t, \Delta_t) = \Delta_t \frac{\beta}{N} \sum_j (\mathcal{A}_{ij}I_j(t)) + \Delta_t^2 \Gamma(\mathcal{A}_{ij}, I_j(t), \Delta_t)$$

onde Γ é um polinômio em Δ_t. Podemos então escrever:

[5] lê-se: micro SIR

$$\begin{aligned}
\frac{S_i(t+\Delta_t)-S_i(t)}{\Delta_t} &\approx \quad \mu - \mu S_i(t) - S_i(t)\tfrac{P_i(t,\Delta_t)}{\Delta_t} \\
\frac{I_i(t+\Delta_t)-I_i(t)}{\Delta_t} &\approx \quad S_i(t)\tfrac{P_i(t,\Delta_t)}{\Delta_t} - \gamma I_i(t) - \mu I_i(t) \\
\frac{R_i(t+\Delta_t)-R_i(t)}{\Delta_t} &\approx \quad \gamma I_i(t) - \mu R_i(t), \qquad (5.6)
\end{aligned}$$

uma vez que $\lim_{\Delta_t \to 0} \frac{P_i(t,\Delta_t)}{\Delta_t} = \frac{\beta}{N}\sum_j \mathcal{A}_{ij} I_j$, passamos o limite quando Δ_t tende a zero, encontrando a equação (5.5) com

$$Y_i(t) = \frac{\beta}{N}\sum_j \mathcal{A}_{ij} I_j(t). \qquad (5.7)$$

Esta formulação, por considerar todas as informações da matriz de contatos $\mathcal{A}$, permite a modelagem da dinâmica da epidemia, por exemplo, em redes conexas, desconexas e com diferentes estados iniciais. A vantagem deste modelo em relação ao HMF é que leva em consideração todas as propriedades da rede, não precisando da premissa de homogeneidade da dinâmica nos nós de mesma conectividade.

Desenvolvemos o modelo μSIR de forma a se ter um modelo de equações diferenciais apto a descrever o comportamento de uma epidemia em uma rede qualquer e no intuito de modelar tais epidemias em redes em que os modelos conhecidos ou falham, ou são computacionalmente inviáveis. Simulações e resultados deste modelo são apresentados na Seção 5.6.

5.6 Resultados para o μSIR

Experimentos numéricos foram realizados, a fim de comparar o comportamento do modelo proposto com outros modelos. Os modelos considerados foram:
[SIR]: O modelo SIR clássico, conforme equação (4.1).
[MBI]: O modelo baseado em indivíduos.
[HMF]: O modelo HMF, conforme equação (5.2).
[GG]: O modelo proposto em [42], que é uma cadeia de Markov com $\Delta_t = 1$.
[MKV]: Modelo baseado em uma Cadeia de Markov, conforme equação (5.4), com $\Delta_t < 1$.
[μSIR]: O modelo Proposto, conforme equação (5.5.)

Nos experimentos, a matriz de adjacência da rede foi diretamente empregada nos modelos MBI, GG, MKV e μSIR. A mesma matriz de adjacência foi utilizada ao invés da matriz *annealed* utilizada no modelo GG em [42], o que significa que o conjunto de redes assumidas por esse modelo, neste caso, tem uma única rede. Os parâmetros epidemiológicos dos experiemrentos foram escolhidos, em sua maioria, de forma a permitir uma comparação com os resultados apresentados em [42]. A contagem de graus dos nós foi realizada, a fim de fornecer as informações utilizadas no modelo HMF. Nenhuma informação conectividade é passada ao modelo SIR, que implicitamente assume um gráfico

totalmente conectado. Os resultados apresentados aqui correspondem, no caso do MBI, às médias de 1000 simulações Monte Carlo.

5.6.1 Rede totalmente conectada

O primeiro experimento considerado foi feito de forma a validar o μSIR e demais modelos com o modelo SIR clássico. Desta forma, consideramos uma rede totalmente conectada (significando que quaisquer dois indivíduos podem ter contato entre si, correspondendo ao SIR clássico) de um sistema correspondendo ao SIR com parâmetros $\gamma = 0$, $\mu = 0,3$, $\beta = 0,7$ e $N = 5.000$ (também comnhecido como modelo SIS). A Figura 5.3 ilustra as respostas temporais dos modelos. Neste caso, os modelos SIR e HMF têm expressões analíticas idênticas. Torna-se evidente que os modelos SIR, HMF, MBI, MKV e μSIR têm respostas semelhantes, e a resposta do modelo GG é claramente diferente dos demais. A Figura 5.4 ilustra, ainda para redes totalmente conectadas, uma comparação dos níveis endêmicos para valores diferentes de β. Deve ser notado que mais uma vez dois grupos de respostas aparecem: os modelos SIR, MBI, MKV e μSIR levam a um grupo de valores semelhantes, e o modelo GG leva para outros valores. Estes resultados sugerem que o modelo GG sofre de uma espécie de efeito sub-amostragem ou de acúmulo de erros. Isto é causado pela suposição de que os eventos de infecção de vários indivíduos são independentes, como indicado na equação (5.3). Embora isso possa ser assumido para pequenos intervalos de tempo, isso se torna cada vez mais impreciso à medida que o intervalo de tempo cresce, pois para intervalos maiores uma dependência entre os eventos emerge. Como resultado o modelo GG subestima o número de infecções, pois ignora as infecções causadas por indivíduos que foram infectados mais cedo no mesmo intervalo de tempo. A capacidade do modelo μSIR em recuperar os modelos clássicos, por outro lado, é suportada pelos resultados.

5.6.2 Rede gaussiana

Um experimento foi realizado considerando-se uma rede gaussiana (uma rede construída com a mesma probabilidade de conexão entre quaisquer dois nós), também chamado de Erdos-Renyi. Pode ser demonstrado que as redes gaussianas representam instâncias de redes complexas que podem ser representados pelo modelo SIR clássico (corrigindo-se apenas o parâmetro β pela densidade da rede). A Figura 5.5 ilustra o comportamento assintótico do número de indivíduos infectados para os modelos SIR, GG, MKV e μSIR, para uma rede gaussiana com 10.000 nós, alguns valores de $\beta = \lambda N$ e conectividade média de $\langle k \rangle = 50$. Novamente nota-se um comportamento singular do modelo GG em relação aos demais modelos. Este experimento juntamente com o anterior ajuda a validar o modelo μSIR nas situações onde se conhece analiticamente a resposta do problema.

5.6.3 Rede livre de escalas

Outro experimento foi realizado considerando-se uma rede complexa como descrito em [10], que segue a lei de potência $P(k) \sim k^{-\alpha}$ com $\alpha = 3,0$ e conectividade média $\langle k \rangle = 4$. A população com $N = 5.000$ indivíduos e parâmetros

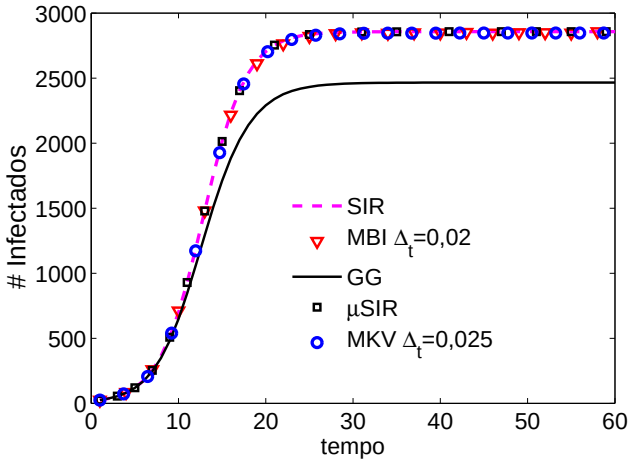


Figura 5.3: Comportamento temporal do número de indivíduos infectados dos modelos em uma rede completa. A linha tracejada ilustra o comportamento temporal do modelo GG, que difere do comportamento dos demais modelos, inclusive do modelo teórico SIR, demonstrando que uma escolha indevida Δ_t pode acarretar em resultados pouco satisfatórios.

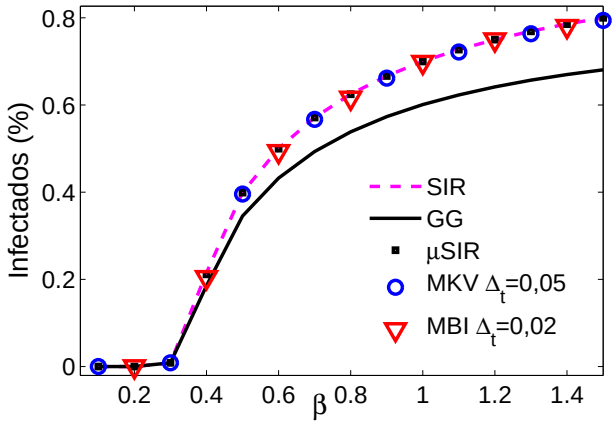


Figura 5.4: Comportamento assintótico do número de indivíduos infectados em uma rede completa. O gráfico ilustra o nível assintótico dos modelos para diferentes valores de β. Novamente notamos uma discrepância nos resultados do modelo GG em relação aos demais.

$\beta = \lambda N$, $\gamma = 0$ e $\mu = 0,3$ foram considerados. Os resultados são apresentados na Figura 5.6. Este experimento sugere que o modelo GG subestima sistematicamente o número de indivíduos infectados. Por outro lado, os modelos HMF e μSIR ambos concordam com os modelos MBI e MKV: esta pode ser interpretada como uma evidência da validade desses quatro modelos, que fornecem suporte mútuo.

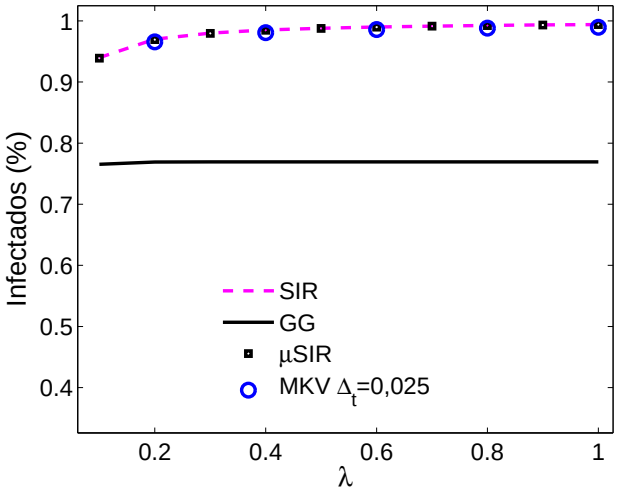


Figura 5.5: Comportamento assintótico do número de infectados em uma rede gaussiana. Em uma rede gaussiana as premissas do modelo SIR ainda o permitem modelar a epidemia nesta rede. Para diferentes v alores de $\beta = \lambda N$ foram encontrados os valores endêmicos pra os modelos SIR, GG, μSIR e MKV. Este experimento ajuda a validar os modelo MKV (para pequenos valores de Δ_t) e μSIR, porém ilustra uma falha no modelo GG.

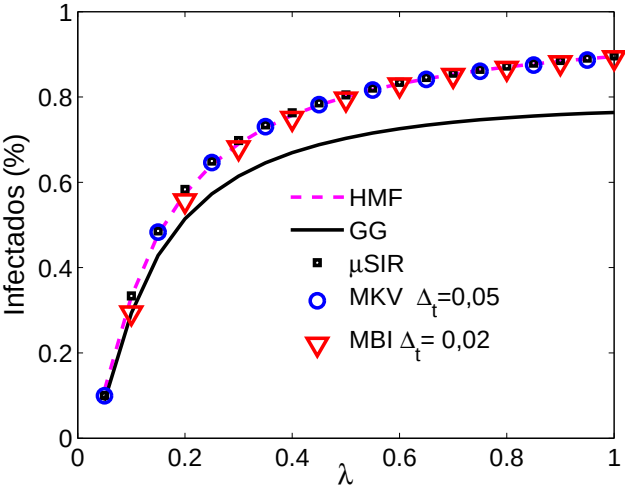


Figura 5.6: Comportamento assintótico do número de infectados em uma rede livre de escalas. Esta figura ilustra os níveis endêmicos para diferentes valores de $\beta = \lambda N$ em uma rede livre de escalas de Barabasi. Foram comparados os modelos HMF, MBI, GG, μSIR e MKV, sendo que apenas o modelo GG difere dos demais. Este experimento refaz um dos experimentos obtidos em [42], mostrando que seus resultados foram pouco satisfatórios.

5.6.4 Rede regular

Fizemos um experimento em uma rede regular com 2.000 indivíduos, cada um com 24 vizinhos. A Figura 3.3 ilustra quais indivíduos são considerados vizinhos (N) para serem conectados a um indivíduo (I). A fim de que todos

os indivíduos tivessem o mesmo número de contatos, a relação de vizinhança foi construida na superfície de um toro. É interessante observar que, numa situação hipotética em que a epidemia se iniciasse com 25 indivíduos infectados (indivíduo (I) e seus 24 vizinhos (N)), apenas os 16 indivíduos mais externos poderiam inicialmente infectar novos indivíduos, o que pode torar o início da epidemia um pouco mais lento. Além disso, esta topologia de rede aumenta o caminho médio entre os nós. Estes dois fatos tendem a influenciar em uma menor velocidade de propagação da epidemia.

As Figuras 5.7 e 5.8 ilustram o comportamento temporal dos modelos μSIR, SIR, MKV, HMF e MBI nesta rede. As curvas do MBI nas duas figuras são relativas à média das curvas de infectados de 1.000 simulações de Monte Carlo. Os modelos SIR e HMF não foram desenvolvidos levando em consideração redes com propriedade de forte correlação como da rede regular, e ilustramos seus comportamentos apenas para se ter uma ideia da diferença quando comparados com os demais modelos. Uma vez que todos os indivíduos têm a mesma conectividade, o modelo HMF coincide com o modelo SIR com novo parâmetro $\bar{\beta} = 0,012\beta$, que é um comportamento que espera-se ser diferente para este tipo de rede devido à sua estrutura rígida. Os modelos μSIR e MKV novamente seguem o mesmo comportamento nas duas figuras. O MBI simula a epidemia como um processo de contato. Processos de contato em redes com fortes correlações, em certas redes e sob certos parâmetros, não são bem descritos pelos resultados de campo médio. A Figura 5.7 ilustra o caso em que o MBI, apesar da epidemia ser mais lenta conforme a intuição prévia, tem uma aproximação ruim comparada ao μSIR para $\beta = 30$. Contudo, na Figura 5.8, para $\beta = 100$, a curva de infectados do MBI é melhor descrita pela curva do μSIR. Estes fatos nos levam a acreditar que o μSIR coincida com o campo médio para este tipo de rede.

5.6.5 Rede com estrutura de comunidades

Indivíduos de uma sociedade costumam se organizar em subgrupos dentro dos quais se tem maior possibilidade de contato que no restante da sociedade. Estudos recentes têm levado estas redes em consideração [44], em especial do ponto de vista do controle de epidemias nestas redes [66, 88]. Um experimento foi feito em uma rede contruída da seguinte forma: foram construídas 10 redes conforme [10] com 200 indivíduos cada. Uma rede maior foi criada interligando as redes anteriores em série conforme a Figura 5.9 com um único link aleatoriamente escolhido entre 2 redes. A Figura 5.10 ilustra o comportamento temporal dos modelos HMF e μSIR para os parâmetros $\gamma = 0,05$, $\mu = 0,01$ e $\beta = 25$, com indivíduos infectados inicialmente localizados em uma das comunidades extremas. O modelo HMF, que não foi desenvolvido para redes fixas com estas particularidades, sendo ilustrado apenas para se ter uma noção de distanciamento dos modelos, e parece se comportar como se estivesse em uma única rede com propriedades idênticas às redes menores. Intuitivamente, poderíamos esperar um comportamento como segue: cada comunidade após infectada se comporta aproximadamente como um cidade isolada devido à estreita ligação com as demais, porém o início de cada infecção acontece em

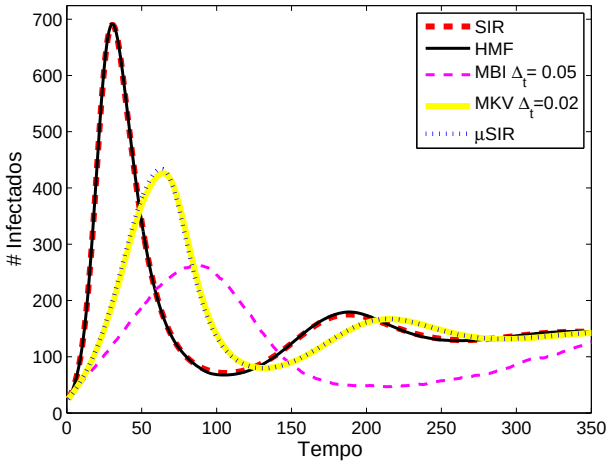


Figura 5.7: Comportamento temporal do número de indivíduos infectados em uma rede regular. Os parâmetros usados foram $\gamma = 0,1$, $\mu = 0,01$, $\beta = 30$. Ilustramos as curvas dos os modelos SIR, HMF, μSIR, MKV, e MBI. O comportamento em redes regulares é conhecido não ser igual ao modelo SIR e HMF. Os modelos MKV e μSIR convergem para outros valores. Neste caso o MBI, que simula um processo de contato, não parece coincidir com nenhum modelo.

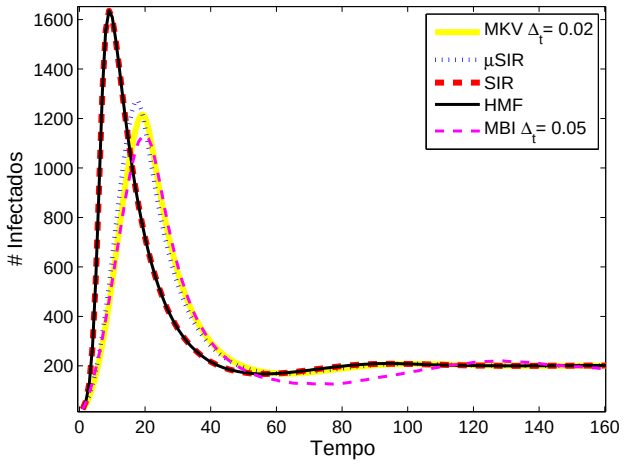


Figura 5.8: Comportamento temporal do número de indivíduos infectados em uma rede regular. Os parâmetros usados foram $\gamma = 0,1$, $\mu = 0,01$, $\beta = 100$. Ilustramos as curvas dos os modelos SIR, HMF, μSIR, MKV, e MBI. Neste caso, o MBI, que simula a epidemia como processo de contato, tem uma resposta parecida com a do μSIR.

tempos distintos e de forma sequencial. Este comportamento pode ser visto no comportamento do modelo μSIR da Figura 5.10 nesta rede.

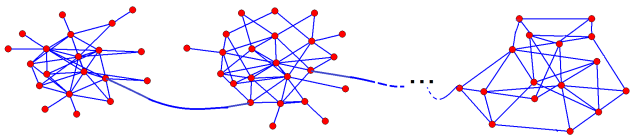


Figura 5.9: Rede com estrutura de comunidade.

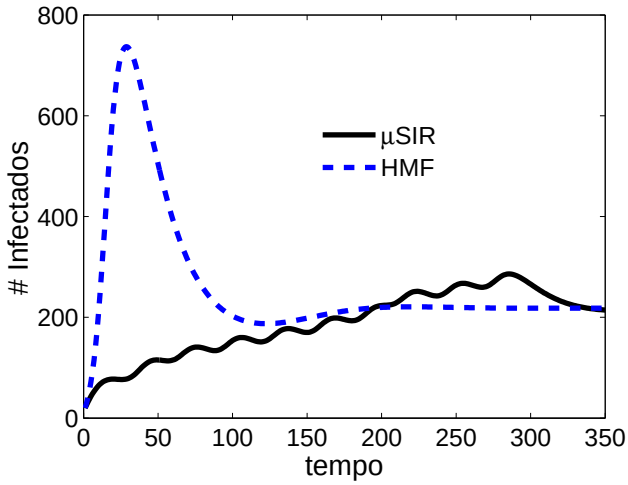


Figura 5.10: Comportamento temporal do número de indivíduos infectados em rede com estrutura de comunidade. A rede foi construída da ligação de 10 comunidades, com 200 indivíduos cada, cada comunidade organizada em uma rede livre de escalas de Barabasi, e são interligadas em série por uma única aresta entre dois nós escolhidos aleatoriamente, conforme Figura 5.9. Intuitivamente espera-se um comportamento similar ao apresentado pelo modelo μSIR: infecções ocorrendo nas comunidades em cascata. O modelo HMF apresenta um comportamento similar a uma rede que tem as mesmas propriedades de cada comunidade, ignorando a estrutura de comunidade.

5.6.6 Rede espacial

A Figura 5.11 ilustra a topologia de uma rede espacial construída com 1.000 nós, $\langle k \rangle = 10$, e α=100. Apesar da informação espacial ser perdida ao tratarmos apenas a matriz de contatos, esta informação fica implícita na distribuição das ligações entre os nós da rede. Podemos notar na figura que as ligações são geralmente curtas e entre indivíduos muito próximos. Esta configuração pode gerar uma velocidade menor de propagação da epidemia, uma vez que o caminho médio entre dois nós aumenta em relação às redes gaussianas.

Um experimento com uma rede espacial composta por meio milhão de nós, $\langle k \rangle = 10$, e α=100 foi executado. A Figura 5.12 ilustra o comportamento temporal dos modelos HMF, μSIR e SIR nesta rede. A distribuição de conectividades deste tipo de rede é parecida com a de uma rede Gaussina. Sendo a

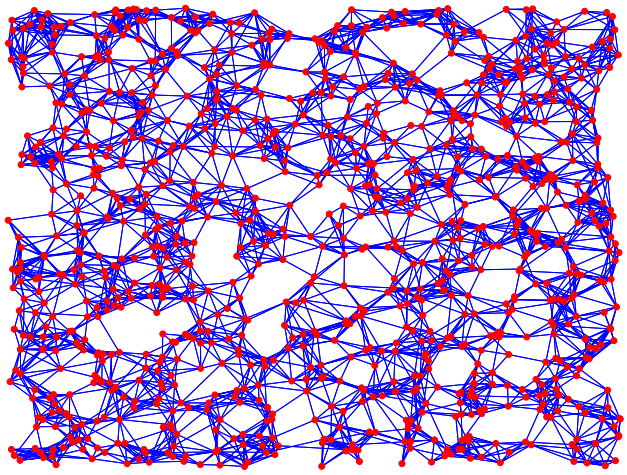


Figura 5.11: Rede espacial. A rede foi construída com 1.000 nós, $\langle k \rangle = 10$, e α=100. Somente links curtos são observados, o que diminui a velocidade de propagação de uma epidemia.

distribuição de conectividades a única informação levada em conta pelo HMF, este modelo tem como resposta o equivalente a uma rede Gaussiana, ou seja, se aproxima do modelo SIR. Uma vez que o modelo μSIR não perde informações da matriz de contatos, o mesmo consegue nos dar uma resposta compatível com a esperada, i.e., uma epidemia mais lenta que em uma rede Gaussiana.

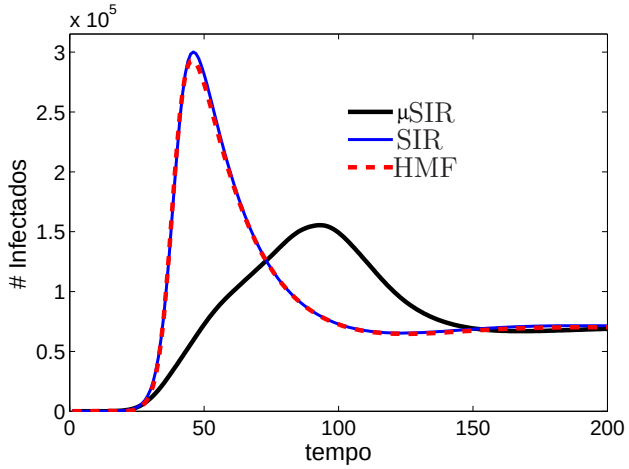


Figura 5.12: Comportamento temporal do número de indivíduos infectados em uma rede espacial. A rede utilizada foi contruída com 500.000 nós, $\langle k \rangle = 10$, e $\alpha = 100$. Os parâmetros epidêmicos utilizados foram $\beta = 20.000$, $\gamma = 0,05$, $\mu = 0,01$ iniciando a a epidemia com apenas um único indivíduo infectado e todos os demais suscetíveis.

5.6.7 Discussão

Nos modelos MBI e MKV, a escolha de valores menores de Δ_t melhora significativamente a precisão dos modelos, porém aumenta o custo computacional dos mesmos, podendo ser inviável seu uso. A necessidade de uma escolha

arbitrária nestes modelos é um inconveniente, uma vez que tal escolha depende dos parâmetros do problema em questão. Além disso, os resultados com o MBI são obtidos por meio de simulações de Monte Carlo, o que onera ainda mais a obtenção de resultados com o algoritmo.

Nesta seção apresentamos os resultados para uma nova formulação da descrição da dinâmica de epidemias em redes por um sistema de equações diferenciais, o μSIR. A formulação proposta fornece uma descrição analítica da evolução média de cada indivíduo em uma população que está conectada de acordo com qualquer topologia de rede arbitrária. O modelo μSIR recupera o modelo SIR tradicional, nos casos em que os pressupostos do modelo SIR valem, e também está de acordo com simulações do MBI, MKV e com os resultados da abordagem HMF recente.

O uso de redes dinâmicas no modelo μSIR é direto, bastando para isso utilizarmos um matriz de contato $\mathcal{A}(t)$ cujas probabilidades de contato $\mathcal{A}_{ij}(t)$ são dependentes do tempo, incluindo aqui as redes *annealed*. Esta é mais uma importância do modelo, uma vez que permite o estudo analítico do fenômeno para redes dinâmicas.

O modelo proposto permite estudos inviáveis no modelo HMF como por exemplo: redes desconexas, redes direcionadas, com estrutura de comunidades, com diferentes níveis de coeficientes de clusterização, com propriedades geográficas além de permitir o estudo de processos de vacinação ao nível do indivíduo.

O modelo μSIR foi implementado em Matlab® (Versão 7.6.0.324 R2008.a). O computador utilizado em todas as simulações possuia um processador Intel Core 2 duo T6500 2,1GHz e 4GB de Ram. Em uma rede Espacial com meio milhão de indivíduos e conectividade média $\langle k \rangle = 10$ executou 200 unidades de tempo em aproximadamente 3 minutos. O modelo HMF, uma vez que possui um número extremamente reduzido de equações (neste mesmo experimento foram encontradas apenas 29 conectividades diferentes) executou em menos de 3 segundos, porém obteve resultados pouco satisfatórios em alguns tipos de redes. Acreditamos que o modelo HMF continue sendo um boa ferramenta para o estudo de comportamentos assintóticos, porém, além do μSIR também ser útil neste caso, para a necessidade do estudo temporal (no estudo de campanhas de vacinação por exemplo) o modelo μSIR parece ser o mais adequado.

A principal vantagem do modelo μSIR em relação à modelagem por Cadeias de Markov está no fato de que o μSIR herda todo o conhecimento de resoluções de equações diferenciais desenvolvido ao longo dos anos, enquanto a solução por Cadeias de Markov representa um método de integração númerica por retângulos cuja base Δ_t necessita ser escolhida antes do início da solução do problema.

O modelo μSIR proposto foi desenvolvido de forma a ser um modelo de propagação de epidemias em redes que seja equivalente ao modelo SIR clássico devidamente ajustado para que a propagação da epidemia ocorra sobre essas redes. Os testes feitos indicam que o modelo é satisfatório como um modelo para redes equivalente ao SIR.

5.7 HMF Multi-Comunidades (HMF-MC)

O μSIR é um modelo teórico e necessita de um número de informações muito grande para ser usado em um problema prático. Diante disso, desenvolvemos nesta seção um modelo nos moldes do HMF, o HMF-MC, de forma a se modelar a epidemia a partir de um número menor de informações e, em especial, para ser usado em conjunto com a análise espectral apresentada na Seção 3.3.1.

Consideraremos o caso em que existe algum tipo de estruturação da rede em subredes mais densas, ou seja, de forma que possa ser entendida como uma rede com estrutura de comunidades. Suporemos conhecidos apenas a distribuição de conectividades dentro de cada comunidade (conforme já utilizado no HMF) e o número de conexões existentes entre cada par de comunidades .

O HMF em sua formulação original utiliza como única informação os dados do histograma de conectividade da rede. Uma vez que para redes sem estruturas rígidas ele funciona bem, dentro de cada comunidade o modelo HMF-MC irá considerar que os contatos internos seguem a formulação do HMF. Os contatos entre indivíduos de comunidades diferentes serão feitos por meio da matriz de pesos C. A Figura 5.13 ilustra de forma intuitiva o funcionamento do modelo: dentro de cada comunidade um modelo HMF e os pesos das ligações se referem ao número de ligações entre os indivíduos das comunidades.

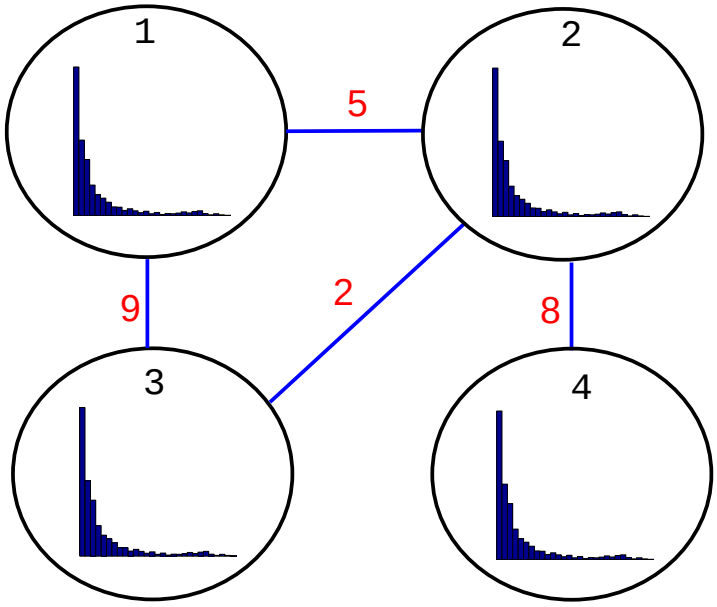


Figura 5.13: Ilustração do funcionamento do HMF-MC. Dentro de cada comunidade somente a contagem das conectividades é necessária, calculando as infecções internas de forma idêntica ao HMF original. As conexões existentes entre comunidades são agrupadas em uma matriz de pesos (ligações em azul com valores em vermelho), e as infecções entre indivíduos de comunidades diferentes é calculada pressupondo homogeneidade (puramente aleatória) das conexões.

A matriz C abaixo ilustra o número de conexões entre as comunidades da Figura 5.13. Esta matriz é usada, juntamente com as informações das freqûências das conectividades internas de cada comunidade, no modelo HMF-MC.

$$C = \begin{bmatrix} 0 & 5 & 9 & 0 \\ 5 & 0 & 2 & 8 \\ 9 & 2 & 0 & 0 \\ 0 & 8 & 0 & 0 \end{bmatrix}$$

Seja C_{ij} número de arestas entre as comunidades i e j (define-se $C_{ii} = 0$ $\forall i$), $S_{ik}(t)$, $I_{ik}(t)$, $R_{ik}(t)$ as respectivas densidades de nós suscetíveis, infectados e recuperados da comunidade i com conectividade k no instante de tempo t. Para cada comunidade i temos que $S_{ik}(t) + I_{ik}(t) + R_{ik}(t) = 1$ $\forall t$ e $\forall k$. Seja ainda $P_i(k)$ fração dos indivíduos da comunidade i que tem conectividade k e seja ainda $\langle k \rangle_i$ a conectividade média da comunidade i.

O modelo será dado pelo seguinte sistema de equações diferenciais:

$$\begin{aligned} \frac{dS_{ik}(t)}{dt} &= \mu - \mu S_{ik}(t) - \tfrac{\beta}{N} k S_{ik}(t)\Theta_i(t) - \tfrac{\beta}{N} S_{ik}(t) \textstyle\sum_j \left(C_{ij} \sum_{\bar{k}} \frac{P_j(\bar{k}) I_{j\bar{k}}(t)}{N_j} \right), \\ \frac{dI_{ik}(t)}{dt} &= \tfrac{\beta}{N} S_{ik}(t) \textstyle\sum_j \left(C_{ij} \sum_{\bar{k}} \frac{P_j(\bar{k}) I_{j\bar{k}}(t)}{N_j} \right) + \tfrac{\beta}{N} k S_{ik}(t)\Theta_i(t) - \gamma I_{ik}(t) - \mu I_{ik}(t), \\ \frac{dR_{ik}(t)}{dt} &= \gamma I_{ik}(t) - \mu R_{ik}(t), \\ & S_{ik}(0) = S^o_{ik}, \quad I_{ik}(0) = I^o_{ik}, \quad R_{ik}(0) = R^o_{ik} \end{aligned} \tag{5.8}$$

em que

$$\Theta_i(t) = \frac{\sum_k k P_i(k) I_{ik}(t)}{\sum_s s P_i(s)} = \frac{\sum_k k P_i(k) I_{ik}(t)}{\langle k \rangle_i} \tag{5.9}$$

similarmente ao HMF. O termo $\sum_{\bar{k}} \frac{P_j(\bar{k}) I_{j\bar{k}}(t)}{N_j}$ é a probabilidade de se escolher aleatoriamente na comunidade j um indivíduo infectado. Desta forma, estamos considerando que a distribuição de arestas entre nós de comunidade diferentes é feita de forma puramente aleatória, diferentemente do cálculo de Θ_i onde se considera uma maior probabilidade de receber uma aresta para indivíduos com maior conectividade.

É interessante notar que caso haja apenas uma comunidade, o modelo se resume ao HMF e, caso escolhamos um número de comunidades igual ao número de indivíduos, a matriz C será idêntica à matriz de contatos $\mathcal{A}$ e se resume ao modelo μSIR. Isto demonstra a importância em uma boa escolha do número de comunidades, sendo que esta decisão pode ser tomada conforme apresentado na Seção 3.3. Desta forma, o HMF-MC foi proposto como um modelo intermediário entre o HMF e o μSIR.

O modelo HMF-MC considera dois tipos de iterações de formas distintas: contatos internos de uma mesma comunidade são calculados conforme o HMF

padrão e os contatos entre indivíduos de comunidades diferentes são calculados similarmente aos contatos do sistema Multi-Cidades apresentado na Seção 4.4.1. Simulações e resultados deste modelo são apresentados a seguir.

5.8 Resultados para o HMF-MC

Nesta seção apresentamos alguns resultados para o modelo HMF-MC proposto. Utilizaremos para comparação a resposta do modelo μSIR proposto na Seção 5.5. Intuitivamente, o caso em que a epidemia se propaga mais rapidamente (lentamente) acontece quando as ligações entre as comunidades ocorrem entre nós de alta (baixa) conectividade. O HMF-MC é proposto como uma aproximação para o comportamento da epidemia e os experimentos mostram uma curva de infectados entre os casos mais lento e mais rápido. Em todos os experimentos mantivemos constante a taxa de infecção global, ou seja, β/N foi mantido constante e igual a $0,0125$. Os demais parâmetros foram $\gamma = 0,05$ e $\mu = 0,01$.

5.8.1 Comunidades em série

Fizemos um experimento para a rede utilizada na Seção 5.6.5 conforme a Figura 5.9. Variaremos apenas o número de conexões entre as cidades.

A Figura 5.14 ilustra o comportamento da epidemia se propagando em uma rede com uma única conexão entre as comunidades se iniciando com apenas um único indivíduo infectado localizado em uma das comunidades extremas. Esta rede é um caso extremo para estudos de propagação de epidemias e a maioria dos modelos é incapaz de descrever o comportamento da epidemia de forma satisfatória.

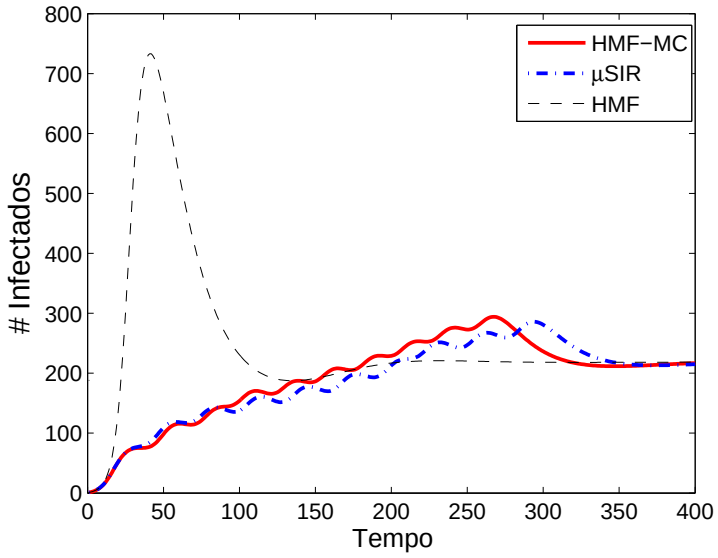


Figura 5.14: Comparação do comportamento dos modelos HMF, HMF-MC e μSIR em uma rede de comunidades conforme Figura 5.9.

Nota-se que a velocidade do HMF-MC é maior que do μSIR na Figura 5.14. Tal fato pode intuitivamente ser explicado pelo fato de o HMF-MC trocar uma

única conexão por um processo homogêneo de conexões entre os indivíduos de comunidade diferentes. Na Figura 5.15 apresentamos o comportamento do μSIR (em verde) em redes nas quais a conexão entre as comunidades foi gerada aleatoriamente, bem como sua linha média. Verifica-se que realmente existe uma diferença de velocidade entre os modelos.

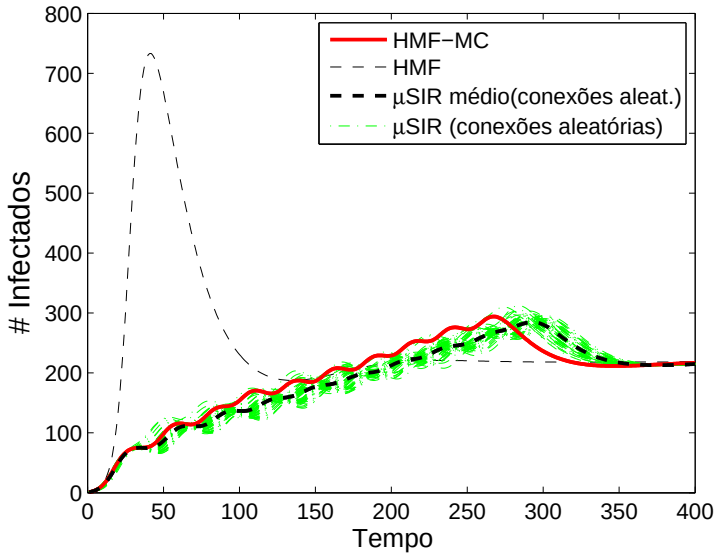


Figura 5.15: Comparação entre o HMF-MC e o comportamento médio do μSIR para conexões geradas aleatoriamente entre as comunidades.

A Figura 5.16 ilustra o caso em que as conexões entre os indivíduos de comunidades diferentes são feitas com preferencialidade conforme Seção 3.2.3, o que ilustra que a velocidade do HMF-MC para esta rede é comparável com o caso mais rápido para a epidemia.

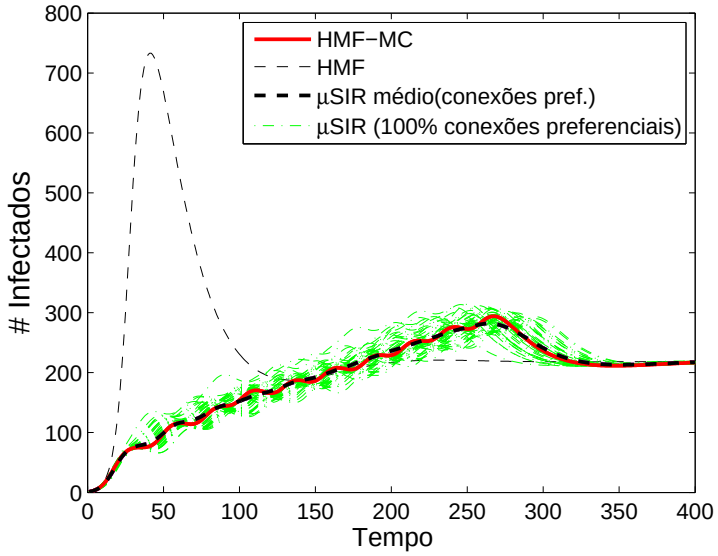


Figura 5.16: Compraração entre o HMF-MC e redes com ligações entre comunidade geradas com preferencialidade.

A Figura 5.17 ilustra o caso em que as comunidades conectadas têm 5 conexões entre si e as ligações preferenciais tem 50% de chance de ocorrer. Isso ilustra o fato de um maior número de ligações entre as comunidades ser mais próximo do caso homogêneo modelado pelo HMF-MC. Além disso, nesta rede o HMF-MC já não está mais tão próximo do caso extremo de 100% de preferencialidade.

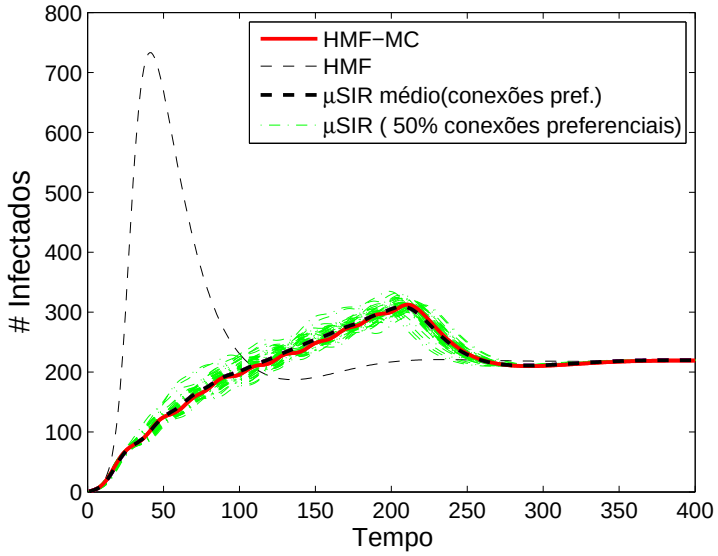


Figura 5.17: Comparação entre o HMF-MC e do μSIR em uma rede em série com 5 ligações entre as comunidades conectadas.

A Figura 5.18 ilustra o caso em que a rede foi construída com 10 ligações

entre as comunidades conectadas. Neste experimento as ligações foram geradas de forma puramente aleatória. Notamos novamente nesta figura que o número de conexões entre as comunidades diminui o erro do modelo HMF-MC.

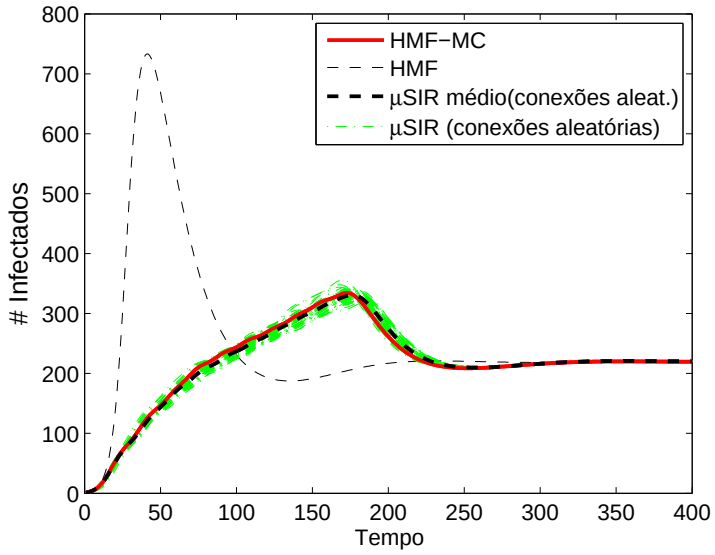


Figura 5.18: Comparação entre o HMF-MC e do μSIR em uma rede em série com 10 ligações entre as comunidades conectadas.

5.8.2 Rede de comunidades completa

Um novo tipo de rede foi considerado em que todas as comunidades estão interligadas entre si. Esta topologia se assemelha mais com a realidade que a topologia peculiar da rede apresentada na Seção 5.8.1.

A Figura 5.19 ilustra o comportamento do HMF-MC e μSIR para o caso de apenas uma única ligação para cada par de comunidades.

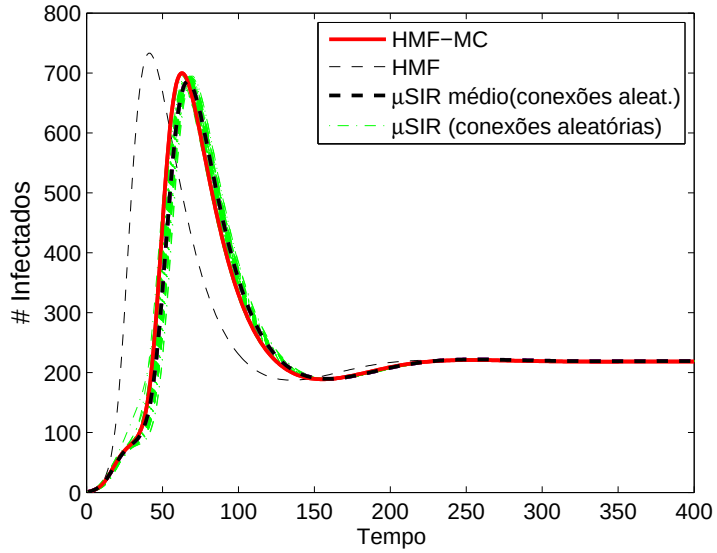


Figura 5.19: Comportamento do HMF-MC e μSIR numa rede completa de comunidades com 1 ligação entre cada par de comunidades.

A Figura 5.20 ilustra o comportamento do HMF-MC e μSIR para o caso de 5 ligações para cada par de comunidades.

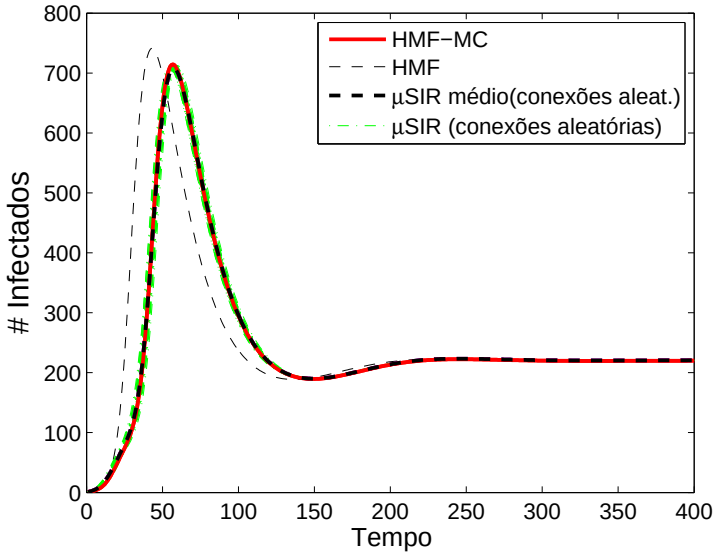


Figura 5.20: Comportamento do HMF-MC e μSIR numa rede completa de comunidades com 5 ligações entre cada par de comunidades.

5.8.3 Rede Livre de Escalas de Comunidades Livre de Escalas

A Figura 5.21 ilustra o resultado do HMF-MC para uma rede livre de escalas composta por 20 comunidades, em que cada comunidade é composta por uma rede livre de escalas compostas de 200 indivíduos. Cada par de comunidades interligadas possui 5 conexões entre si dispostas aleatoriamente.

Esta rede representa uma possível instância de uma sociedade em que, além dos indivíduos se ligarem com preferencialidade, existem comunidades mais conectadas que as demais.

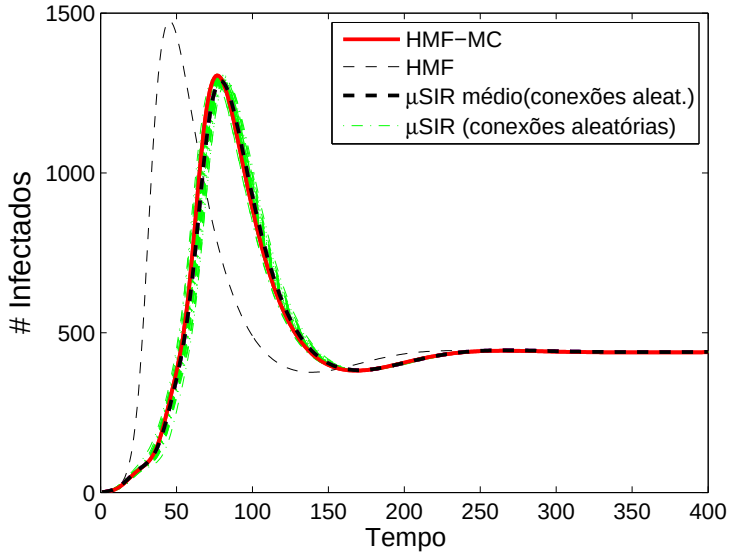


Figura 5.21: Comportamento do HMF-MC e μSIR numa rede livre de escalas de comunidades com 5 ligações entre cada par de comunidades.

5.8.4 Discussão

Para redes sem nenhum tipo de estruturação, modelos mais simples como o HMF podem ser utilizados. Para redes espaciais entendemos que o μSIR seja o ideal. Para redes com estrutura de comunidades, em que se deseja utilizar um modelo para o mundo real inferindo menos informações da rede, desenvolvemos o HMF-MC como uma alternativa.

Apesar de provavelmente ser um evento raro a topologia das comunidades em série, fizemos os experimentos nestas redes de forma a testar o HMF-MC em um caso extremo, tanto pela topologia das comunidades quanto pelas estreitas ligações entre elas. Apesar do viés encontrado, o HMF-MC demonstrou bons resultados.

Para casos menos extremos, tanto do ponto de vista da topologia da rede de comunidades quanto das ligações entre elas, o HMF-MC demonstrou ter um viés menor. Estes casos são mais prováveis de serem encontrados no mundo real, o que torna seu uso mais atrativo do ponto vista prático.

O número de informações a serem utilizadas pelo HMF-MC é muito menor que pelo modelo μSIR, e estas, do ponto de vista prático, podem ser inferidas por estudos estatísticos das sociedades em questão. A inferência completa da rede, porém, é um processo de engenharia reversa difícil, sendo que o único trabalho que aborda esse problema parece ser [64]. A metodologia ali proposta não parece, no entanto, fácil de ser aplicada na maioria das situações.

Capítulo 6

Vigilância de Epidemias

6.1 Introdução

A vigilância em saúde pública tem sido objeto de constante discussão entre acadêmicos, pesquisadores e profissionais de serviços.

O processo de vigilância pode ser resumido como segue: monitoramento de uma população, definida por suas características, ocupando um especificado espaço geográfico e, ou, temporal, constituída de X indivíduos, dos quais temos Y ocorrências de algum evento. O interesse está em descobrir se essas Y ocorrências são aleatórias ou não.

Um conglomerado (ou aglomerado, ou do inglês *cluster*) é uma agregação incomum de eventos que são agrupados no tempo e, ou, no espaço. No restante do texto utilizaremos o termo *cluster* por já ser de uso comum. Mais especificamente, é uma subregião no espaço e, ou, tempo em que a ocorrência de casos de um fenômeno de interesse é discrepante da região na qual a subregião está inserida, isto é, muito mais alta ou muito mais baixa. Esse fenômeno pode ser a infecção por alguma doença, ocorrência de crimes, vírus de computador, etc.

O objetivo da análise estatística em vigilância é detectar a presença de clusters de casos e descobrir se tais clusters têm alta probabilidade de não serem aleatórios. Do ponto de vista de vigilância, a única hipótese estatística pertinente é de aleatoridade ou não das ocorrências.

Dados históricos são examinados a fim de evidenciar tendências ou detectar clusters de casos. Um procedimento proposto para monitoramento finito é a denominada estatística scan.

A estatística scan teve suas origens nos trabalhos de Naus [69, 71, 70]. Supõe-se que N eventos ocorrem em um espaço e, ou, tempo e assume-se que os eventos observados são distribuídos de forma independente e uniforme. A hipótese a ser testada é a de aleatoridade das ocorrências neste espaço contra a hipótese alternativa de ocorrência de clusters no mesmo espaço.

Tanto em epidemiologia quanto em vigilância, a detecção de clusters é uma ferramenta muito utilizada, especialmente na detecção precoce de manifestações epidêmicas [26, 27, 58, 59, 60].

Estes clusters são chamados puramente espaciais quando a ocorrência do evento é mais alta em algumas áreas do que em outras. Quando a incidência de eventos é mais alta durante um determinado intervalo de tempo, esses clusters são puramente temporais. Quando levamos em conta tanto o espaço quanto

o tempo, ou seja, a ocorrência dos eventos é maior em determinado espaço durante um certo intervalo de tempo, dizemos que estes clusters são espaço-temporais.

Métodos de vigilância temporal são usados para monitorar a incidência de eventos para um certa região de forma isolada. Uma discussão destes métodos pode ser encontrada em [94, 102, 72]. Estes métodos são baseados nos dados de incidência sobre o tempo e em um valor predeterminado que, caso excedido, determina um existência do aglomerado temporal.

No caso de vigilância espacial, os métodos são desenvolvidos para detectar a formação de clusters quando se tem a quantidades de ocorrências em diversas regiões que, juntas, formam uma área geográfica de estudo de grande tamanho. Métodos de vigilância espacial são discutidos em [65, 62].

Os métodos de vigilância espaço-temporal são desenvolvidos de forma a incorporar tanto a informação temporal como espacial. Estes métodos são discutidos em [94, 102]. Tais métodos consideram o número de casos e população que ocorrem dentro um intervalo espaço-tempo.

Dado o mapa de uma região de estudo, subdividido em regiões com populações de diferentes tamanhos, não faz sentido encontrar clusters com contagem de casos discrepante. Deve-se levar em consideração a população de cada região e buscar um cluster que tenha um número de casos discrepante do esperado (razão entre o número de casos observados e a população). Desta forma, [14] desenvolveu um método que consistia de janelas circulares com centros em cada região do mapa, que cresciam continuamente até atingir um número crítico de casos. Para cada centro e raio, o número de casos encontrados na janela é comparado ao número de casos esperados na mesma.

Levar em consideração apenas o número de casos esperado gera o problema de encontrar clusters com pouca significância do ponto de vista estatístico. Este problema pode ser melhor entendido no seguinte exemplo: suponha um mapa (conjunto de regiões no espaço) com população total de 100.000 habitantes e duas regiões A e B deste mapa com populações $N_A = 100$ e $N_B = 10.000$ habitantes respectivamente. Se o número de casos observados de certa doença é $C = 1.000$, então esperamos, caso não haja nenhum cluster no mapa, 1 caso a cada 100 habitantes por região. Desta forma, o número de casos esperados nas regiões A e B são respectivamente $\mu_A = 1$ e $\mu_B = 100$. Se o número de casos observados nas regiões A e B são respectivamente $C_A = 2$ e $C_B = 200$, então as razões entre os valores observados e esperados nas duas cidades são os mesmos $C_A/\mu_A = C_B/\mu_B = 2$. Contudo, a probabilidade de ter acontecido por mero acaso mais 1 caso na cidade A é muito maior que ter acontecido mais 100 casos na cidade B.

Neste trabalho, estamos interessados na estatística espacial scan desenvolvida por Kulldorff e Nagarwalla em [61], que contorna o problema acima mencionado em [56], sendo baseada na razão de verossimilhança e que utiliza uma estatística de varredura multidimensional.

6.2 Estatística Espacial Scan de Kulldorff

A estatística espacial scan desenvolvida por Kulldorff e Nagarwalla [61] foi aprimorada em [56], tratando o problema do risco relativo, apresentado no exemplo da seção anterior, de forma a considerar o fato de que o risco relativo, em cidades com proporção de casos idênticos, é mais significativo em populações maiores.

Neste trabalho, consideraremos que a distribuição de casos em cada região do mapa segue uma distribuição da Poisson, com valores esperados proporcional ao tamanho da população desta região. Matematicamente isso significa que, dada uma região R_i, o número de casos C_i, com valor esperado λ_i, tem função de probabilidade

$$f_i(c) = \begin{cases} \dfrac{e^{-\lambda_i}\lambda_i^c}{c!} & \text{se } \ c > 0 \\ 0 & \text{caso contrário,} \end{cases} \tag{6.1}$$

ou seja, $f_i(c)$ é a probabilidade da variável aleatória C assumir o valor c_i. Se N_i e p_i são respectivamente o tamanho da população da região R_i e a probabilidade de um indivíduo desta região ser um caso, então $\lambda_i = p_i N_i$. Denotamos, então, a distribuição de Poisson por $C_i \sim Po(p_i N_i)$.

Chamaremos de *zona* qualquer conjunto conexo de regiões em um mapa. Denotaremos N_z o tamanho da população e por C_z o número de casos na zona z. Supondo que todas as regiões tem a mesma probabilidade de um indivíduo ser um caso, ou seja, $p_i = p \ \forall i$, obtemos $N_z = \sum_{i \in z} N_i$ e $C_z = \sum_{i \in z} C_i$. A distribuição de Poisson possui como propriedade que a soma de variáveis aleatorias independentes com tal distribuição é ainda uma variável aleatória com distribuição de Poisson, cujo parâmetro é a soma dos parâmetros da distribuições. No caso em questão temos que $C_z \sim Po(\sum_{i \in z} pN_i) = Po(pN_z)$.

Denotaremos por Z o conjunto de todas as zonas possíveis do mapa. Trabalharemos com a hipótese nula H_0 de que não exista um cluster no mapa, ou seja $C_z \sim Po(pN_z) \ \forall z \in Z$, e com a hipótese alternativa H_1 de existência de uma zona $\bar{z}$ que é um cluster, ou seja $C_{\bar{z}} \sim Po(pN_{\bar{z}})$ e $C_z \sim Po(qN_z)$ $\forall z \neq \bar{z} \in Z$, com $p \neq q$. Interessados em clusters que se destacam por um número de casos superior ao esperado, definimos a hipótese alternativa como sendo a existência de um cluster onde $p > q$, e o teste de hipóteses pode ser escrito como:

$$\begin{cases} H_0 : & p = q \\ H_1 : & p > q. \end{cases} \tag{6.2}$$

Seja $L(z)$ a função de verossimilança sob a hipótese alternativa H_1, e L_0 a verossimilança sob a hipótese nula H_0. Em [56] encontra-se a demonstração de que o modelo de Poisson tem sua razão de verossimilança dada por:

$$LR(z) = \frac{L(z)}{L_0} = \begin{cases} \left(\frac{c_z}{\mu_z}\right)^{c_z} \left(\frac{C-c_z}{C-\mu_z}\right)^{C-c_z} & \text{se } c_z > \mu_z \\ 1 & \text{caso contrário} \end{cases} \tag{6.3}$$

em que C é o número total de casos no mapa, c_z e μ_z são, respectivamente, o número de casos encontrados e o número de casos esperados em uma zona z. A estatística de teste é $\max_z LR(z)$ que, maximizada sobre todas as zonas do mapa, indentifica o cluster $\bar{z}$ mais verossímil.

Intuitivamente a equação (6.3) poder ser interpretada em função dos riscos relativos dentro ($I(z) = c_z/\mu_z$) e fora ($O(z) = (C-c_z)/(C-\mu_z)$) da zona z. Desta forma, podemos reescrever a função $LR(z)$ da forma:

$$LR(z) = I(z)^{c_z} O(z)^{C-c_z}. \tag{6.4}$$

Computacionalmente, a função LR cresce muito rápido. Uma vez que a função logaritmo é estritamente crescente, se $\bar{z}$ maximiza LR então também maximiza seu logaritmo. Na prática, se trabalha com a maximização de:

$$LLR(z) = \begin{cases} c_z \log\left(\frac{c_z}{\mu_z}\right) + (C-c_z)\log\left(\frac{C-c_z}{C-\mu_z}\right) & \text{se } c_z > \mu_z \\ 0 & \text{caso contrário.} \end{cases} \tag{6.5}$$

Maximizar $LLR(z)$ sobre todas as zonas possíveis do mapa por busca exaustiva tem alta complexidade computacional. Para contornar este problema, em geral, duas técnicas tem sido utilizadas:

- Redução do espaço de parâmetros. Este método consiste em reduzir o espaço de parâmetros Z em um espaço $\bar{Z} \subset Z$, tal que seu tamanho permita uma busca exaustiva. A escolha de $\bar{Z}$ deve ser tal que contenha a zona que maximize $LLR(z)$, ou uma aproximação para esta zona.

- Otimização estocástica. Neste método o espaço de parâmetros não é completamente analisado, mas pode convergir para o máximo global.

Este trabalho se limitará ao primeiro método. Na próxima seção apresentamos o principal método de detecção de clusters por redução do espaço de parâmetros.

6.2.1 Algoritmo Scan Circular

O algoritmo *Scan* Circular proposto por [56] é eficiente, com baixa complexidade computacional, facilmente implementável e, por estes motivos, é amplamente utilizado. Este método é similar ao apresentado por [14], porém, utiliza-se da estatística de maximizar a equação (6.5) para encontrar o cluster mais verossímil.

Este método se baseia em uma janela de forma, tamanho e localização que se modifica sobre uma área geográfica. Para cada janela é calculada a verossi-

milhança com base no número esperado de eventos dentro e fora desta janela. As regiões contidas na janela de maior verossimilhança definem o cluster mais provável. A significância do teste é feita pelo método de Monte Carlo, sob a hipótese nula de não existência do cluster, sobre a distribuição da máxima verossimilhança dos dados aleatórios gerados. A hipótese alternativa é de existência do cluster. Uma escolha natural para a forma da janela é a circular [56], a qual será usada no algoritmo.

Consideraremos que o mapa é dividido em n regiões distintas $R_1, R_2, ..., R_n$. Para cada região R_i definimos um ponto arbitrário em seu interior C_i, que chamaremos centróide. Cada janela circular define uma zona, que é o conjunto de regiões cujos centróides estão contidos na janela. Consideraremos janelas de raio limitado a $r_{\max}$. Para cada região R_k, tomamos todas as janelas circulares de centro C_k e raios possíveis $r < r_{\max}$. Para cada zona distinta definida por estas janelas, avaliamos seu LLR pela equação (6.5). O cluster mais verossímil é aquele de maior LLR.

Para cada um dos n centróides, por maior que seja $r_{\max}$, avaliamos no máximo n zonas. Desta forma, o máximo de zonas a serem avaliadas é n^2, que do ponto de vista computacional é relativamente simples.

O método sempre encontra um cluster mais verossímil, o que não significa que este cluster não seja um evento que ocorreu por mero acaso. Desta forma, a significância estatística do cluster pode ser obtida pelo método de Monte Carlo. Em resumo, consiste em gerar casos sob a hipótese nula H_0 nas regiões do mapa e calcular o cluster mais verossímil. O procedimento é realizado um grande número de vezes, obtendo assim uma distribuição empírica para o LLR. Esta distribuição empírica é comparada com o LLR da solução obtida para os casos observados no mapa e é, então, estimado o p-valor do cluster encontrado.

6.3 Estatística Espacial para Fluxo de Indivíduos (Workflow)

Nos grandes centros urbanos, por motivos econômicos, a grande massa trabalhadora dos centros comerciais moram nas periferias das grandes cidades. Por passarem grande parte do seu tempo em seus locais de trabalho, pressupõe-se que os indivíduos ficam expostos a infecções, na maior parte do tempo, em seu locais de trabalho. Isso geraria uma diluição da contagem de casos dos grandes centros, enfraquecendo a estatística Espacial *Scan* de Kulldorf.

Uma alternativa para tratar este problema é desenvolvido em [26], do inglês chamada *Workflow Spatial Scan Statistic*, onde é proposta uma modificação na estatística espacial *scan* levando-se em conta o fluxo de trabalhadores entre as cidades e que as infecções ocorrem nas cidades de trabalho. Descreveremos abaixo a estatística.

Consideraremos uma região composta por n cidades (ou qualquer subdivisão do mapa), $Z_1, ..., Z_n$. Seja p_{ki} a proporção de indivíduos que reside na cidade Z_k e trabalha na cidade Z_i, $k, i = 1, ..., n$. Seja, ainda, C o número total de casos de certa infecção, N a população na região de estudo e $r_{\max}$ o maior número de cidades a ser considerado dentro do cluster procurado. O processo de cálculo da estatística proposta em [26] se resume a:

1. Para cada $i = 1, ..., n$, ordene em ordem crescente de distância à cidade Z_i as cidades $Z_1, ..., Z_n$, denotando por $Z_{i_1}, ..., Z_{i_n}$ as cidades nesta nova ordem;

2. Para cada $r = 1, ..., r_{\max}$ e cada $k = 1, ..., n$ defina

$$A(k, i, r) = \sum_{s=1}^{r} p_{ki_s};$$

3. Para cada $r = 1, ..., r_{\max}$ e cada $i = 1, ..., n$ ordene $A(k_j, i, r)$ tal que

$$A(k_1, i, r) \geq ... \geq A(k_n, i, r)$$

definindo uma nova ordem de cidades $Z_{k_1}, ..., Z_{k_n}$ respectivamente denotadas por $Z(1, i, r), ..., Z(n, i, r)$;

4. Para cada $r = 1, ..., r_{\max}$ e cada $i = 1, ..., n$, construa os conjuntos:

$$\begin{aligned} Y(1, i, r) &= Z(1, i, r) \\ Y(2, i, r) &= Z(1, i, r) \cup Z(2, i, r) \\ &\vdots \\ Y(n, i, r) &= Z(1, i, r) \cup \cdots \cup Z(n, i, r); \end{aligned}$$

5. Defina $c(s, i, r)$ e $P(s, i, r)$, respectivamente, como o número de casos observados e a população em $Z(s, i, r)$;

6. Defina

$$u(k, i, r) = \sum_{s=1}^{k} A(k_s, i, r) c(s, i, r),$$

$$e(k, i, r) = \frac{C}{N} \sum_{s=1}^{k} A(k_s, i, r) P(s, i, r);$$

7. A estatística para fluxo de indivíduos é, então, calculada por:

$$W(Y(k, i, r)) = \left(\frac{u(k, i, r)}{e(k, i, r)} \right)^{u(k,i,r)} \left(\frac{C - u(k, i, r)}{C - e(k, i, r)} \right)^{C - u(k,i,r)}$$

se $u(k, i, r) > e(k, i, r)$ e 1 caso contrário.

O cluster mais verossímil será a região $Y(K, I, R)$ que maximiza W. O conjunto $Y(K, I, R)$ representa um conjunto de K cidades cujos indivíduos trabalham nas R cidades mais próximas de Z_I (em Z_I e nas $R - 1$ cidades mais próximas dela). Como era de se esperar, as regiões $Y(k, i, r)$ podem ser possivelmente não conexas. Para melhor entendimento do processo acima explicado, damos a interpretação dos cálculos executados:

- $A(k, i, r)$: proporção da população que vive em Z_k e trabalha nas r cidades mais próximas de Z_i;

- $u(k, i, r)$: número de casos observados na região $Y(k, i, r)$ infectados nas r cidades mais próximas de Z_i;

- $e(k, i, r)$: número de casos esperados na região $Y(k, i, r)$ devido à infecção nas r cidades mais próximas Z_i.

A estatística $W(Y(k, i, r))$ é calculada para $i = 1, ..., n$, $k = 1, ..., n$ e $r = 1, ..., r_{\max}$, ou seja, $n^2 r_{\max}$ vezes. Do ponto de vista computacional a estatística é um pouco mais onerosa que a apresentada na Seção 6.2.1, mas ainda assim podemos considerá-la de baixo custo.

A significância estatística do teste é feita pelo método de Monte Carlo, conforme já explicado na Seção 6.2.1.

6.4 Estatística espacial baseada no valor esperado (EBSS)

As estatísticas das Seções 6.2 e 6.3 são baseadas no tamanho da população, e eficientes para detecção de clusters no caso de eventos raros. No caso de doenças contagiosas (tais como gripe, tuberculose, etc), nas quais o número de casos é relativamente elevado em relação à população e possuem comportamentos diferentes para diferentes localizações no espaço e tempo, estas ferramentas podem não ser as mais adequadas.

O conhecimento do histórico do evento, ou mesmo de um modelo estocástico para ele, pode ser usado para a construção de ferramentas de detecção de clusters. Desta forma, consideraremos conhecidos a média e desvio padrão amostrais do evento. Consideraremos, ainda, que o número de casos de cada região em um fixado momento do tempo segue uma distribuição normal. Desta forma, a estatística espacial baseada no valor esperado (EBSS)[1] [72] trabalha com o seguinte teste de hipóteses:

$$\begin{array}{ll} H_0 & : c_i \sim N(\mu_i, \sigma_i^2) \text{ para toda localização espacial } z_i. \\ H_1(Z) & : c_i \sim N(q\mu_i, \sigma_i^2) \text{ para toda localização } z_i \in Z, \text{ e } c_i \sim N(\mu_i, \sigma_i^2) \text{ para} \\ & \text{toda localização espacial } z_i \notin Z, \text{ para alguma constante } q > 1. \end{array}$$

no qual H_0 é a hipótese nula de não ocorrência de cluster no mapa, e $H_1(Z)$ é a hipótese alternativa de que Z é um cluster e não existência de clusters no restante do mapa. Para este teste, obtemos a seguinte expressão para a função razão de verossimilhança:

$$LR(Z) = \frac{\max_{q>1} \prod_{z_i \in Z} P(c_i \sim N(q\mu_i, \sigma_i^2)) \prod_{z_i \in Z} P(c_i \sim N(\mu_i, \sigma_i^2))}{\prod_{z_i} P(c_i \sim N(\mu_i, \sigma_i^2))}$$

de onde, após alguns cálculos algébricos, obtém-se:

$$LR(Z) = \max_{q>1} \exp\left(\frac{1 - q^2}{2} \sum_{z_i \in Z} \frac{\mu_i^2}{\sigma_i^2} + (q - 1) \sum_{z_i \in Z} \frac{c_i \mu_i}{\sigma_i^2} \right)$$

[1] Sigla proveniente do inglês: Expectation-based Scan Statistic (EBSS)

fazendo $B = \sum_{z_i \in Z} \frac{\mu_i^2}{\sigma_i^2}$ e $D = \sum_{z_i \in Z} \frac{c_i \mu_i}{\sigma_i^2}$, obtemos que o valor de q que maximiza a função LR é dado por $q = \max(1, D/B)$. Obtemos então :

$$LR(Z) = \begin{cases} \exp(\frac{D^2}{2B} + \frac{B}{2} - D) & \text{, se } D > B \\ 1 & \text{, caso contrário.} \end{cases} \tag{6.6}$$

Assim como no algoritmo *Scan* Circular proposto por [56], trabalha-se novamente com uma janela de forma circular cujo tamanho e localização se modifica sobre uma área geográfica, calculando para cada janela a verossimilhança dada pela equação acima. As regiões contidas na janela de maior verossimilhança definem o cluster mais provável. A significância do teste é feita pelo método de Monte Carlo, sob a hipótese nula de não existência do cluster, sobre a distribuição da máxima verossimilhança dos dados aleatórios gerados. A hipótese alternativa é de existência do cluster.

6.5 Estatística espacial para fluxo de indivíduos baseada no valor esperado (WEB)

Nesta seção, propomos uma modificação na estatística workflow apresentada na Seção 6.3 de forma a, equivalentemente ao que desenvolvido para a estatística scan de Kulldorf, obtermos uma estatística que leve em consideração o fluxo de trabalho, contudo, baseada no conhecimento de dados históricos, ou modelos estocásticos. Descrevemos a seguir o algoritmo para a estatística espacial para fluxo de indivíduos (WEB)[2]:

Os passos 1 a 4 deste algoritmo são idênticos ao do algoritmo workflow apresentado na Seção 6.3, não sendo necessário repeti-los aqui. Em seguida executam-se os passos:

5. Defina $\mu(k,i,r)$, $\sigma(k,i,r)$ e $c(k,i,r)$ respectivamente como o número de casos esperados, o desvio padrão e o número de casos encontrados da região $Z(k,i,r)$;

6. Defina

$$B(k,i,r) = \sum_{s=1}^{k} A(k_s,i,r)^2 \frac{\mu(k,i,s)^2}{\sigma(k,i,s)^2}$$

$$D(k,i,r) = \sum_{s=1}^{k} A(k_s,i,r)^2 \frac{c(k,i,s)\mu(k,i,s)}{\sigma(k,i,s)^2}$$

7. A estatística WEB calculada para a região $Y(k,i,r)$ será dada por:

$$WEB(Y(k,i,r)) = \exp\left(\frac{D(k,i,r)^2}{2B(k,i,r)} + \frac{B(k,i,r)}{2} - D(k,i,r)\right)$$

se $D(k,i,r) > B(k,i,r)$ e 1 caso contrário.

[2]Sigla proveniente do inglês: Workflow Expectation-based (WEB)

O cluster mais verossímil será a região $Y(K, I, R)$ que maximiza WEB. O conjunto $Y(K, I, R)$ representa um conjunto de K cidades (cluster casa) cujos indivíduos trabalham nas R cidades (cluster trabalho) mais próximas de Z_I (em Z_I e nas $R-1$ cidades mais próximas dela).

Na Seção 4.2 apresentamos um modelo de equações diferenciais estocásticas para epidemias de contato direto. Este tipo de modelo é comumente utilizado em epidemiologia matemática e o utilizaremos para avaliar o desempenho da estatística WEB na Seção 6.6

6.6 Resultados para a estatística WEB

Nesta seção compararemos os resultados obtidos entre as estatísticas EBSS e a proposta WEB.

Chamaremos uma zona z_k de *inflow* se o número de trabalhadores que trabalham em z_k for maior que o número de trabalhadores de z_k que trabalham fora de z_k, e de *outflow* caso contrário. Neste trabalho, estamos interessados em zonas *inflow*, ou seja, em determinar se existe ou não um cluster de cidades onde os indivíduos trabalham e são infectados.

6.6.1 Geração e detecção de epidemias simuladas

Para os casos simulados, consideramos uma região próxima da cidade de Norfolk no estado da Virgínia, Estados Unidos, Figura 6.1. A região inclui 35 cidades, cobrindo aproximadamente 32.000 km^2 e 1,8 milhões de indivíduos. Os dados usados foram retirados do censo[3] dos Estados Unidos feito no ano 2.000. A matriz de workflow foi então calculada por:

$$p_{ik} = \frac{\#(\text{ residentes da cidade } i \text{ que trabalham na cidade } k)}{\#(\text{ residentes da cidade } i)}$$

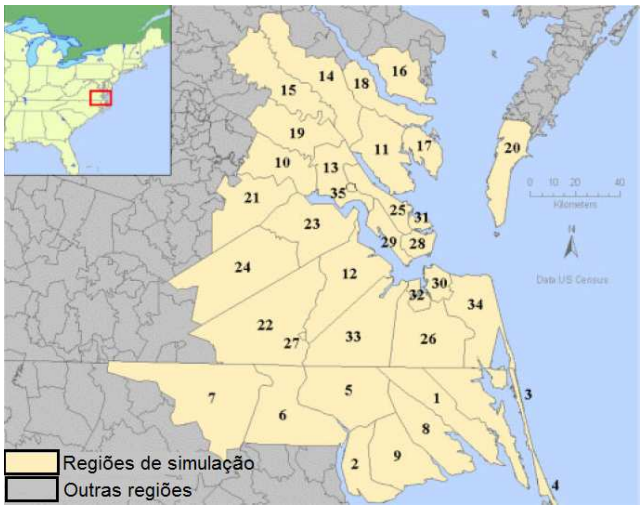


Figura 6.1: Região de estudo para os casos simulados

[3]disponível em www.census.gov

Para os casos simulados o nível de significância utilizado será de $\alpha = 0,05$. Para os testes com a estatística proposta WEB o cluster procurado é o cluster trabalho. Consideramos sucesso, respeitado o nível de significância α, encontrar o cluster nos seguintes casos:

- Tamanho do cluster maior que $r_{\max}$: existir intersecção entre o cluster encontrado e o simulado;
- Tamanho do cluster menor ou igual a $r_{\max}$: o cluster encontrado contiver o cluster simulado.

Calculamos o poder do teste para clusters gerados usando distribuição de probabilidade de cada região e usando a evolução de um passo no tempo por meio de equações diferenciais estocásticas. As distribuições $N(\mu_i, \sigma_i^2)$ são geradas sob H_0 pela equação (4.19) com parâmetros $\beta_i = 0,3$, $\gamma_i = 0,05$ e $\rho_i = 0,01$ $\forall i$, iniciando com 1% da população da cidade 2 infectada e todo o restante do mapa com indivíduos suscetíveis. A geração dos dados sob H_1 para cada teste é explicada a seguir.

Modelo Aleatório

Primeiramente calculamos o poder do teste para o caso mais simples, onde são conhecidos μ_i e σ_i para cada região z_i na unidade de tempo 80 conforme descrito acima , e o cluster é gerado adicionando casos artificialmente. Desta forma, definido um cluster w_0, consistindo de r_0 zonas, e risco relativo RR_{w_0}, calculamos o risco relativo de cada zona z_k no mapa por:

$$RR(z_k) = \begin{cases} \sum_{i=1}^{m} p_{ki} RR_{w_0} & \text{, se } z_k \in w_0 \\ 1 & \text{, caso contrário.} \end{cases} \tag{6.7}$$

Os casos são então distribuidos no mapa seguindo uma distribuição $N(\mu_k \cdot RR(z_k), \sigma_k^2)$. Para cada cluster artificial diferente testado, cada valor de $r_{\max}$ e cada valor do risco relativo RR_{w_0}, fizemos 100.000 simulações de Monte Carlo sob a hipótese nula, obtendo assim uma distribuição empírica para a função densidade de probabilidade da máxima razão de verossililhança. Sob a hipótese alternativa foram feitas 10.000 simulações para cada cluster artificial, $r_{\max}$ e RR_{w_0} diferentes. Efetuamos a busca de cluster também pela estatística scan de Kulldorff, porém nenhum cluster foi encontrado corretamente e por isso omitiremos tal informação nos gráficos e tabelas a seguir.

A Tabela 6.1 contém o cálculo do poder para $RR_{w0} = 1,20$. Somente a detecção de clusters primários foi considerada nesta tabela. Os valores em negrito indicam superioridade de uma estatística em relação à outra. Exceto pelos dois últimos clusters, zonas isoladas *[13]* e *[25]* que são outflow, todas as demais são inflow. Nas regiões inflow notamos claramente a dificuldade das duas estatísticas em detectar corretamente os clusters. Os clusters *[29, 31]* e *[30, 26]* não são possíveis de serem detectados com $r_{\max} = 2$, devido à utilização de janelas circulares na situação de existência de pelo menos uma cidade com distância às duas menor que a distância entre elas. Devido ao formato circular rígido, a utilização de valores pequenos para $r_{\max}$ parece influenciar

negativamente para o workflow. Para $r_{\max} = 4$ e $r_{\max} = 5$ notamos superioridade da estatística com workflow. É útil notar que, fazer $r_{\max} = 5$ não necessariamente implica que o cluster encontrado terá este tamanho, porém, flexibiliza o algoritmo na busca dos clusters cujo formato é desconhecido ao algoritmo.

A Figura 6.2 contém gráficos que ilustram o comportamento do poder do teste em relação ao risco relativo RR_{w_0}. Nas Figuras 6.2(a), 6.2(b) e 6.2(c), fixado $r_{\max} = 5$, notamos a superioridade da estatística workflow para todos os valores de RR_{w_0} (em detectar os clusters como primários). Para $r_{\max} = 4$ os gráficos são similares, e por isso foram omitidos. A Figura 6.2(d) ilustra um comportamento interessante quando detectando o cluster *[30, 26, 32]* como cluster primario com $r_{\max} = 3$. Neste caso, observa-se que a estatística workflow passa a ser inferior à usual à medida que aumentamos o risco relativo. Verificamos que na maioria dos casos em que não houve sucesso em encontrar o cluster como primário, o cluster encontrado foi *[34, 26 ,30]*. Um segundo cálculo foi feito considerando o segundo cluster mais verossimil diferente do primário com mesma significância. Neste caso, considerou-se sucesso o caso em que o cluster procurado foi encontrado como primário ou secundário com significância $\alpha = 0,05$, cujo gráfico também é ilustrado na Figura 6.2(d). Nota-se um ganho de poder das duas estatísticas, porém, o workflow demonstrou superioridade neste caso.

	WEB					EBSS				
zona/ $r_{\max}$	1	2	3	4	5	1	2	3	4	5
28	0,05	**0,33**	**0,31**	**0,38**	**0,43**	**0,13**	0,25	0,28	0,34	0,35
29	0,10	**0,32**	0,28	**0,40**	**0,42**	**0,16**	0,22	**0,30**	0,39	0,40
28, 29	0,56	**0,95**	**0,94**	**0,97**	**0,97**	**0,58**	0,56	0,78	0,89	0,88
29, 31	0,12	0,00	**0,35**	**0,56**	**0,54**	**0,23**	0,00	0,24	0,37	0,37
28, 31	0,06	0,00	**0,41**	**0,51**	**0,53**	**0,21**	**0,26**	0,33	0,41	0,43
28, 29, 31	0,62	**0,97**	**0,91**	**0,99**	**0,98**	**0,72**	0,88	0,75	0,89	0,88
30	**0,40**	**0,40**	**0,64**	**0,67**	**0,65**	0,32	0,37	0,52	0,56	0,57
30, 26	0,58	0,00	**0,92**	**0,94**	**0,94**	**0,58**	0,00	0,77	0,82	0,82
30, 32, 26	0,75	0,97	0,69	**1,00**	**1,00**	**0,86**	0,97	**0,75**	0,89	0,89
13	0,00	0,05	0,07	0,06	0,05	**0,07**	**0,12**	**0,14**	**0,16**	**0,18**
25	0,00	0,00	0,01	0,03	0,03	**0,01**	**0,03**	**0,04**	**0,06**	**0,07**

Tabela 6.1: Cálculo do poder do teste para algumas regiões da Figura 6.1 com clusters gerados pela equação (6.7) com $RR_{w_0} = 1,20$. O nível de significância utilizado foi de $\alpha = 0,05$. Sob a hipótese alternativa, o risco relativo de cada região foi calculado pela equação (6.7) e os casos gerados com distribuição $N(\mu_k \cdot RR(z_k), \sigma_k^2)$. Para cada entrada da tabela foram feitas 10.000 simulações de Monte Carlo.

Modelo SIR multi Cidades Estocástico

Para o caso de estudo de epidemias modeladas por equações diferenciais estocásticas, consideraremos que o estado da epidemia do dia imedatamente anterior não possui clusters. Partindo desta premissa, calcularemos o poder do teste em detectar clusters formados no intervalo de uma unidade de tempo. A propriedade da taxa infecção é proporcional ao parâmetro β_i do modelo SIR

dado pela equação (4.19) e por isso geraremos clusters artificiais mediante a alteração deste parâmetro. Para isso consideraremos a evolução pelo modelo SIR sob H_0 por 79 unidades de tempo. A partir deste último estado da epidemia, evoluímos o modelo SIR multi cidades estocástico uma unidade de tempo sob H_1. Uma vez fixado um cluster w_0 e um valor RR_{w_0}, a evolução do modelo desta última unidade de tempo, para cada zona z_k, é feita pelos parâmetros originais γ_k e ρ_k e por:

$$\bar{\beta_k} = \begin{cases} \beta_k \cdot RR_{w_0} & \text{, se } z_k \in w_0 \\ \beta_k & \text{, caso contrário.} \end{cases} \tag{6.8}$$

A Figura 6.3(a) ilustra a metodologia empregada na geração dos dados. Foram simuladas 10 epidemias pelo modelo SIR estocástico da equação (4.19) sob H_0 por 79 unidades de tempo. Para cada uma destas simulações, sob H_1, simulamos 1.000 evoluções por mais uma unidade de tempo. A significância do teste é feita, para cada uma das 10 simulações, sob 100.000 simulações de Monte Carlo sob H_0. A detecção de clusters é feita na unidade de tempo 80. Efetuamos a busca de clusters também pela estatística scan de Kulldorff, porém nenhum cluster foi encontrado corretamente e por isso omitiremos tal informação nos gráficos e tabelas a seguir.

As Figuras 6.3(b), 6.3(c) e 6.3(d) ilustram o comportamento do poder do teste em relação ao risco relativo RR_{w_0}. Fixado $r_{\max} = 5$, notamos a superioridade da estatística workflow para todos os valores de RR_{w_0} (em detectar os clusters como primários). Para $r_{\max} = 4$ os gráficos são similares, e por isso foram omitidos. Os resultados obtidos são similares aos obtidos pelos gráficos da Figura 6.2, apesar de notarmos um enfraquecimento do poder do teste das duas estatísticas em relação ao teste da Seção 6.6.1. A complexidade do modelo, sendo sua evolução temporal dependente também indivíduos nos estados suscetível e recuperado, pode justificar este aumento na dificuldade em encontrar os clusters.

A Tabela 6.2 contém o cálculo do poder para $RR_{w0} = 1,20$. Somente a detecção de clusters primários foi considerada. Os valores em negrito indicam superioridade de uma estatística em relação à outra. Para os clusters outflow, com $r_{\max} = 4$ e $r_{\max} = 5$, notamos superioridade da estatística com workflow.

6.6.2 Discussão

Os resultados obtidos para a estatística WEB mostrados nas Tabelas 6.1 e 6.2 são similares aos encontrados no trabalho [26]. Isso indica a possibilidade do uso da estatística workflow em conjunto com outras estatísicas espaciais similarmente à estatística WEB, a qual foi construída a partir da estatística EBSS em conjunto com workflow.

Os resultados para o poder dos testes apresentados nas Figuras 6.2 e 6.3 indicam superioridade da estatística proposta para diversos valores risco relativo. Uma vez que o risco relativo em situações reais é desconhecido, estes resultados indicam a que a estística WEB possa ser escolhida independente do quão fraco ou forte seja o evento.

Em especial, os bons resultados obtidos para detecção de clusters no caso

	WEB					EBSS				
zona/ $r_{\max}$	1	2	3	4	5	1	2	3	4	5
28	0,01	**0,18**	**0,14**	**0,17**	**0,23**	**0,06**	0,09	0,09	0,12	0,13
29	0,07	**0,21**	**0,17**	**0,23**	**0,25**	**0,10**	0,14	0,15	0,20	0,20
28, 29	0,24	**0,64**	**0,56**	**0,62**	**0,65**	**0,34**	0,43	0,43	0,50	0,50
29, 31	0,06	0,00	**0,14**	**0,23**	**0,21**	**0,09**	0,00	0,07	0,12	0,12
28, 31	0,02	0,00	**0,13**	**0,18**	**0,16**	**0,06**	**0,04**	0,07	0,11	0,10
28, 29, 31	0,25	**0,64**	**0,45**	**0,63**	**0,56**	**0,35**	0,49	0,32	0,45	0,44
30	**0,30**	**0,29**	**0,53**	**0,54**	**0,54**	0,23	0,23	0,36	0,39	0,39
30, 26	**0,41**	0,00	**0,76**	**0,77**	**0,77**	0,38	0,00	0,58	0,61	0,61
30, 32, 26	**0,48**	0,67	0,28	**0,87**	**0,87**	0,47	0,67	**0,43**	0,71	0,72
13	0,00	0,02	0,02	0,02	0,02	**0,01**	0,02	**0,03**	**0,04**	**0,05**
25	0,00	0,00	0,01	0,03	0,03	0,00	**0,01**	**0,02**	0,03	**0,03**

Tabela 6.2: Cálculo do poder do teste para algumas regiões da Figura 6.1 com clusters gerados com $\bar{\beta}_k$ dado pela equação (6.8) com $RR_{w_0} = 1,20$. O nível de significância utilizado foi de $\alpha = 0,05$. Para cada entrada da tabela foram feitas 10.000 simulações de Monte Carlo.

de epidemias geradas por meio de equações diferenciais estocásticas indicam um caminho (ou pelo menos um primeiro passo) para a união de duas técnicas (detecção de clusters e modelagem), atualmente desconexas, no sentido de entender melhor os fenômenos epidemiológicos.

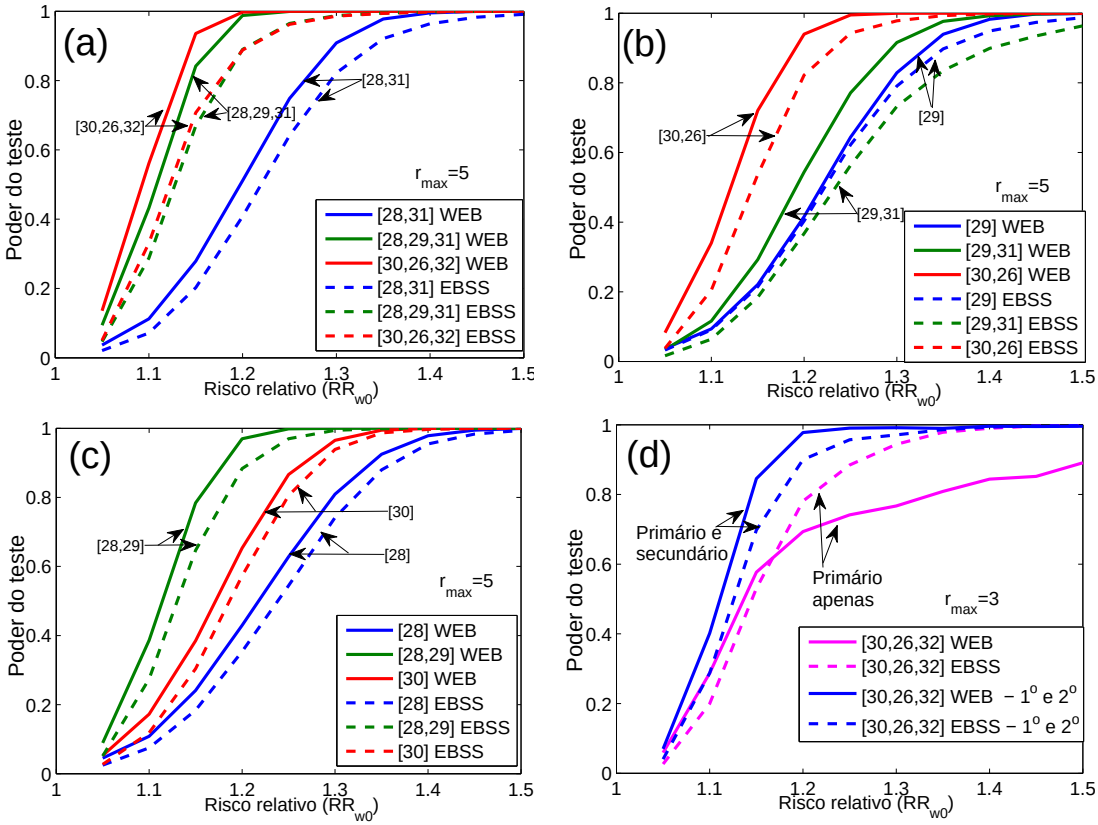


Figura 6.2: Comportamento do poder do teste em relação ao risco relativo RR_{w_o} para as zonas inflow. Os clusters foram gerados a partir da equação (6.7). Os gráficos (a), (b) e (c) ilustram poder do teste em detectar o cluster como primário para $r_{\max} = 5$. Observa-se superioridade da estatística workflow para todos os clusters testados. O gráfico (d) ilustra o comportamento do poder do teste em relação ao cluster *[30,26,32]* pra $r_{\max} = 3$, comparando a detecção apenas como cluster primário e quando consideramos os clusters primários e secundários. O segundo cluster mais verossimil diferente do primário com mesma significância é considerado. Considerou-se sucesso o caso em que o cluster procurado foi encontrado como primário ou secundário com significância $\alpha = 0,05$. O cluster mais encontrado como primário, além do cluster procurado, é o *[34,26,30]*.

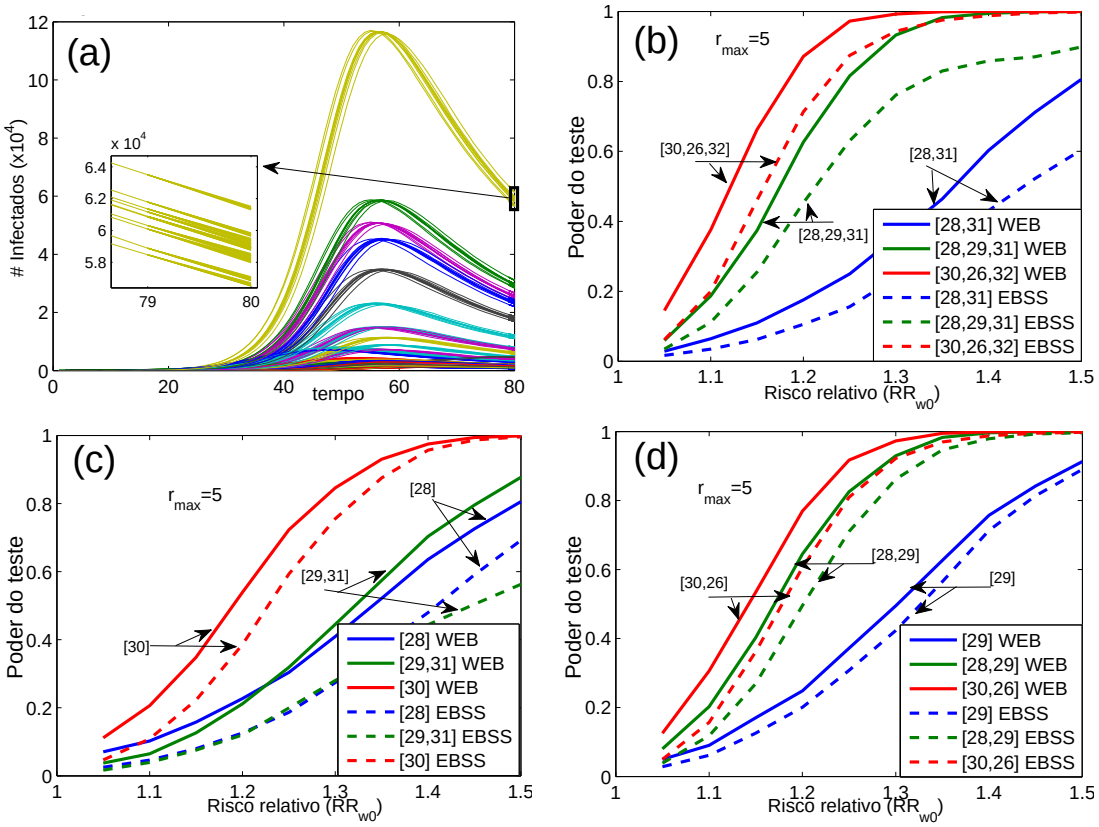


Figura 6.3: Comportamento do poder do teste em relação ao risco relativo RR_{w_o} para as zonas inflow. O gráfico (a) ilustra a geração de clusters: simulação sob H_0 de 10 epidemias por 79 unidades de tempo e, a partir de cada uma destas, 1.000 simulações da evolução de uma unidade de tempo sob H_1 com $\bar{\beta}_k$ dado pela equação (6.8). Os gráficos (b), (c) e (d) ilustram poder do teste em detectar o cluster como primário para $r_{\max} = 5$. Observa-se superioridade da estatística workflow para todos os clusters inflow testados.

Capítulo 7

Inferência Data-Driven (baseada no cluster encontrado)

A estatística scan de Kulldorff para mapas de dados agregados procura por grupos de casos sem especificar o seu tamanho (número de áreas) ou localização geográfica a priori. Sua significância estatística é testada fazendo o ajuste para os testes múltiplos inerentes a tal procedimento. No entanto este ajuste não é feito de forma uniforme para todos os tamanhos de cluster possíveis.

Neste Capítulo nós propomos uma modificação para o teste de inferência do scan usual, incorporando informações adicionais sobre o tamanho do cluster mais verossímil (CMV) encontrado. Fazemos uma nova interpretação dos resultados da estatística scan espacial, colocando uma questão de inferência modificada: qual é a probabilidade da hipótese nula ser rejeitada para o mapa de casos originais observados com um cluster de maior probabilidade de tamanho k, levando em conta apenas os grupos mais prováveis de tamanho k encontrados sob hipótese nula para a comparação? Esta questão é especialmente importante quando o p-valor calculado pelo processo de inferência usual está perto do nível de significância α, tornando difícil a decisão com base na inferência usual.

São apresentados experimentos que justificam a necessidade de uma nova inferência mais apurada para valores próximos dos valores críticos e fornecemos um procedimento prático para fazer inferências mais precisas sobre o cluster mais verossímil (CMV) encontrado pela estatística scan espacial. Resultados para esta nova inferência são apresentados na Seção 7.6

7.1 Aproximações Gumbel

Por meio de extensos testes numéricos foi mostrado [1] que, sob a hipótese nula, a distribuição empírica para clusters do scan circular é aproximadamente a bem conhecida distribuição Gumbel

$$f(x) = \theta^{-1} \exp\{-\exp[(x-\mu)/\theta] - (x-\mu)/\theta\}$$

de parâmetros μ (moda) e θ (escala). Usando essa abordagem semi-paramétrica, a distribuição espacial scan pode ser calculada utilizando um número muito menor de repetições Monte Carlo. Por exemplo, calcular o $\max_{z \in Z} LLR(z)$ sob hipótese nula para apenas 100 mapas aleatórios, e obter sua média e variância para calcular os parâmetros de moda e escala, obtém-se uma distribuição Gum-

bel semi-paramétrica tão precisa quanto uma distribuição puramente empírica produzida após $B = 10.000$ simulações de Monte Carlo [1].

7.2 Distribuição empírica scan$_k$

Uma preocupação pertinente que surge é o fato de ser mais adequado comparar o CMV original apenas com aqueles CMV's de mapas sob a hipótese nula tão semelhantes quanto possível do cluster original, em termos de tamanho da população e localização geográfica. Um exemplo extremo dessa situação seria exigir que os clusters de comparação sejam apenas aqueles (raramente ocorrem) cujo CMV seja exatamente o mesmo CMV original. No entanto, esta tarefa é computacionalmente inviável, pois depende da execução de um número enorme de repetições, a fim de selecionar um número considerável de simulações aleatórias para que os CMV's simulados coincidam com o CMV original. Portanto, é necessário relaxar um pouco estes requisitos. Solicitar que a população seja a mesma do CMV encontrado (independentemente de outros fatores) também pode ser difícil, especialmente para mapas com populações altamente heterogêneas. A possibilidade, então, é permitir que centros de localização e populações sejam diferentes, exigindo, porém, que o número de áreas dentro cluster seja o mesmo.

A estatística espacial scan foi projetada para se ajustar para o teste múltiplo ao avaliar os clusters de diferentes tamanhos e localizações. Este ajuste implicitamente supõe que a distribuição da estatística scan não muda quando restrita para qualquer tamanho fixo de cluster. Como veremos, esta hipótese não pode ser verdadeira. Nós definimos scan$_k$ como a distribuição empírica obtida a partir da distribuição scan empírica que considera apenas os clusters de tamanho k. Mostraremos também que abordagem semi-paramétrica Gumbel pode ser estendida para as distribuições scan$_k$.

7.3 Frequência do tamanho dos clusters

Começamos esta subseção com dois exemplos de mapas com as populações de dados reais. O primeiro mapa é composto por 34 municípios no entorno de Belo Horizonte no Brasil, com 6.262 casos de homicídios durante o período de 1998-2002, para uma população total de 4.357.940 em 2000. O segundo mapa é composto por 245 municípios em 10 estados e no distrito de Columbia, no nordeste dos EUA , com 58.943 mortes de mulheres ajustada por idade, no período de 1988 a 1992, para uma população em risco de 29.535.210 mulheres em 1990 [59].

Para cada mapa, sob hipótese nula foram realizados 1.000.000 de simulações de Monte Carlo e os clusters mais prováveis foram encontrados para cada simulação. O CMV's foram classificados de acordo com seus tamanhos e as freqüências de ocorrência foram exibidas nos histogramas da Figura 7.1. Notamos que para ambos os estudos de caso, representando exemplos típicos de mapas de dados de agregados, a freqüência de tamanhos de cluster varia muito (clusters de tamanho muito pequeno são muito mais comuns). Isso significa que a forma da distribuição empírica scan depende principalmente dos clusters de menor tamanho. Conseqüentemente, o processo de decisão sobre a signi-

ficância do CMV original de casos observados baseia-se principalmente sobre o comportamento de grupos muito pequenos, independentemente do seu tamanho. Alguém poderia argumentar que esta característica, por si só, não deve representar um problema se pudéssemos garantir que as distribuições scan_k fossem quase idênticas para cada valor de k. No entanto, como mostraremos na próxima subseção, existem diferenças significativas entre as diversas distribuições scan_k.

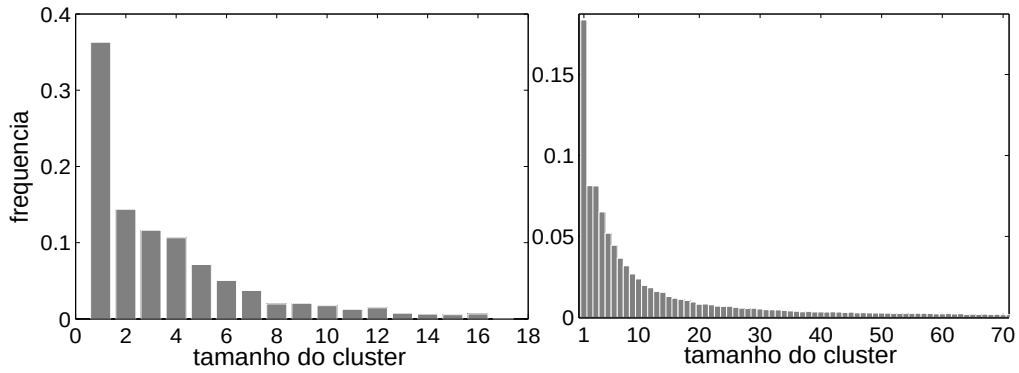


Figura 7.1: Distribuição de freqüência dos tamanhos dos clusters mais verossímeis encontrados sob H_0, com 1.000.000 de simulações de Monte Carlo, para a região metropolitana de Belo Horizonte (esquerda) e do nordeste dos Estados Unidos (direita).

7.4 Ajuste da Gumbel às distribuições scan_k

Como feito anteriormente para a distribuição empírica scan, também definimos a aproximação Gumbel para a distribuição scan_k como a distribuição Gumbel_k. Temos verificado experimentalmente que esse ajuste foi adequado, para todos os tamanhos de cluster em vários mapas diferentes.

Para o mapa de Belo Horizonte, a Figura 7.2 mostra as distribuições scan_k tomadas a partir de 1.000.000 simulações de Monte Carlo e suas respectivas distribuições Gumbel_k ajustadas, para valores de $k = 1$, 6 e 15. Da mesma forma, o mesmo procedimento foi realizado considerando-se o mapa do nordeste os EUA, para valores de $k = 1$, 20 e 80 como visto na Figura 7.3.

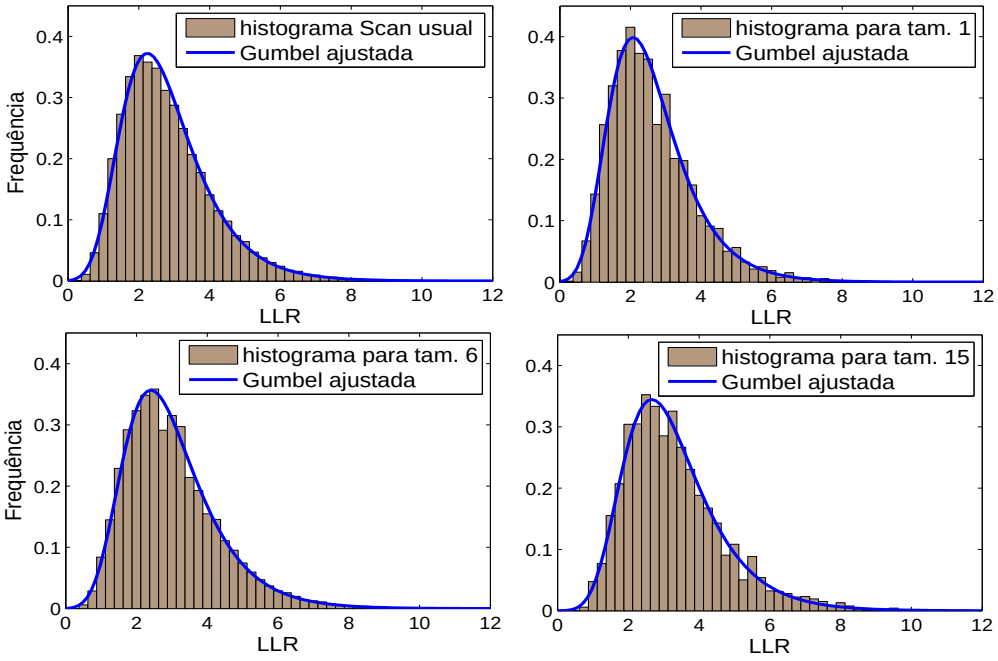


Figura 7.2: Distribuições scan_k obtidas a partir de 1.000.000 de simulações de Monte Carlo no mapa da região metropolitana de Belo Horizonte, e suas respectivas distribuições Gumbel_k ajustadas, para $k = 1$, 6 e 15.

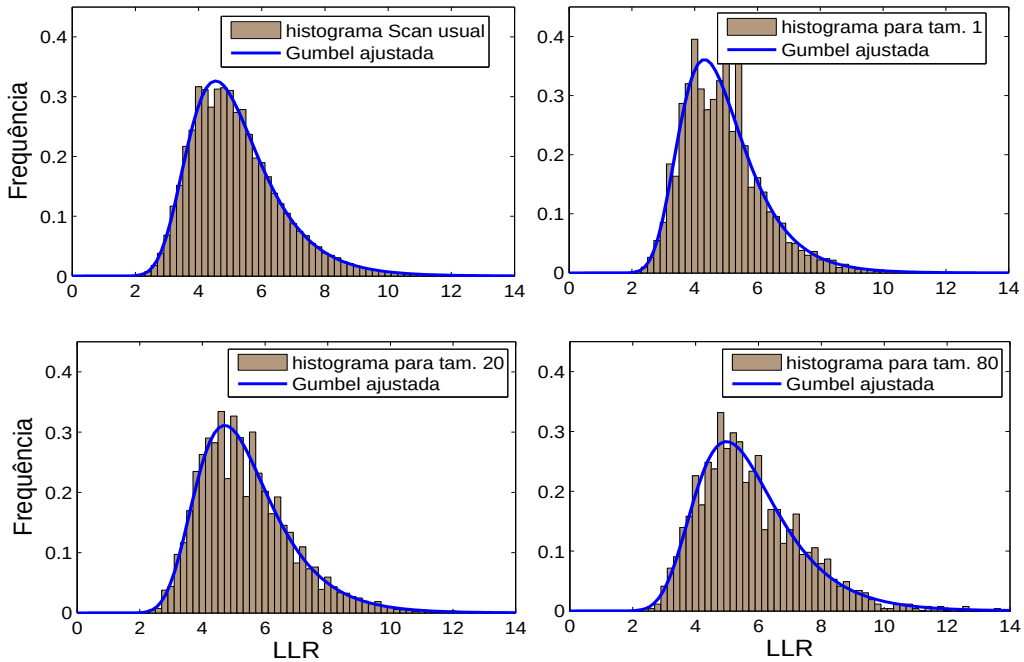


Figura 7.3: Distribuições scan_k obtidas a partir de 1.000.000 de simulações de Monte Carlo no mapa da região do nordeste dos Estados Unidos, e suas respectivas distribuições Gumbel_k ajustadas, para $k = 1$, 20 e 80.

A partir dos resultados das 1.000.000 simulações de Monte Carlo para o

mapa da região metropolitana de Belo Horizonte, observamos que os valores críticos para as distribuições Gumbel$_k$ aumentam monotonicamente com o índice k de 1 para o valor máximo de 17. A distribuição Gumbel ajustada e as distribuições Gumbel$_k$ para $k = 1$, 6 e 15 são exibidos no mesmo gráfico no lado esquerdo da Figura 7.4. Os valores críticos para $\alpha = 0,05$ para as quatro distribuições correspondentes também são mostrados.

Para o mapa do nordeste dos Estados Unidos observamos resultados semelhantes quando colocamos juntas a distribuição Gumbel ajustada e as distribuições Gumbel$_k$ para $k = 1$, 20 e 80 no gráfico à direita da Figura 7.4, com os correspondentes valores críticos para $\alpha = 0,05$.

Esses dois exemplos mostram que as distribuições Gumbel$_k$ mudam com o tamanho k, e são significativamente diferentes da distribuição Gumbel ajustada usual (todos os tamanhos de cluster).

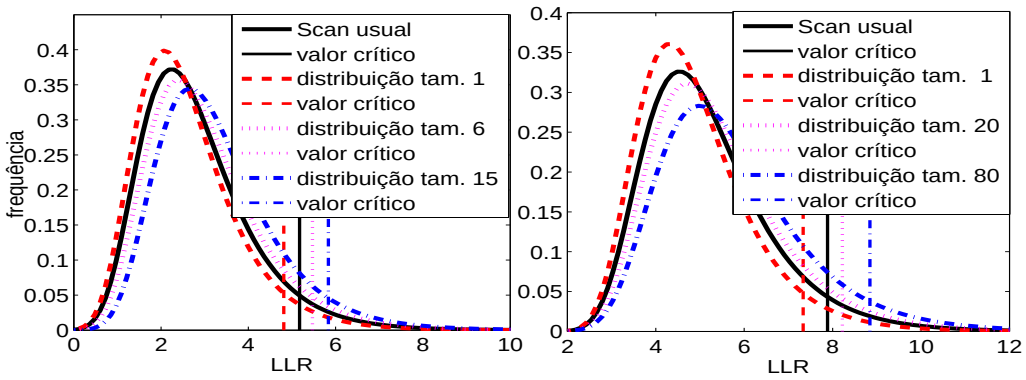


Figura 7.4: A distribuição Gumbel ajustada e as distribuições Gumbel$_k$ para vários valores de k, com os respectivos valores críticos ao nível de significância $\alpha = 0,05$, para a região metropolitana de Belo Horizonte (esquerda) e do nordeste dos Estados Unidos (direita).

7.5 A inferência Data-Driven

Em sua formulação original, o scan circular calcula a significância do cluster mais provável com base na seguinte pergunta: "Levando em conta que o candidato a cluster encontrado tem $LLR = x$, qual é a probabilidade de encontrar um cluster mais provável sob a hipótese nula com $LLR > x$?"

Como visto nesta seção, clusters de pequeno tamanho constituem a maioria entre os CMV's encontrados, e têm a maior influência na determinação da distribuição empírica scan. Na situação em que o CMV tem tamanho grande, a sua inferência será feita basicamente pelo comportamento dos clusters pequenos.

Nesta seção propomos utilizar a informação sobre o tamanho do CMV observado no teste de inferência. Neste caso, estaremos fazendo a seguinte pergunta: "Tendo em vista que o cluster mais provável candidato tem $LLR = x$ e contém k áreas , qual é a probabilidade de encontrar, sob a hipótese nula, um cluster formado por k regiões com $LLR > x$? "

Nossa proposta leva em conta o tamanho do cluster da seguinte forma: dado que o scan circular encontrou o CMV com k regiões , então a sua significância estatística ainda é obtida através de simulações de Monte Carlo, mas selecionando apenas as repetições para as quais as soluções têm CMV com exatamente k áreas. Estamos substituindo a distribuição scan empírica que considera soluções CMV de qualquer tamanho pela distribuição scan_k empírico com CMV's de tamanho exatamente igual a k.

No entanto, quando o p-valor é muito menor ou muito maior do que o nível de significância α, não há nenhuma mudança na decisão de rejeitar, ou não, a hipótese nula. Assim, estamos apenas interessados no processo de decisão quando o p-valor calculado é próximo ao nível de significância α.

Experimentos com a inferência data-driven são apresentados a seguir.

7.6 Resultados para a inferência Data-Driven

Nesta seção apresentamos experimentos numéricos que mostram que a proporção de rejeições da hipótese nula difere notavelmente quando é utilizado o valor crítico usual em comparação com o uso dos valores críticos obtidos pelo data-driven. Mostramos também que o custo computacional de estimar o valor crítico pelo data-driven pode ser reduzido através do uso de filtragem simples.

7.6.1 Variabilidade dos valores críticos

Para avaliar a variabilidade na estimativa dos valores críticos de cada distribuição scan_k, realizamos outro conjunto de simulações de Monte Carlo. Primeiro calculamos um conjunto S_0 de 1.000.000 repetições aleatórias sob hipótese nula e determinamos o valor crítico usual v_0 para o nível de significância $\alpha = 0,05$ empregando a distribuição scan empírica. Em seguida, para cada valor de tamanho k determinamos a distribuição scan_k empírica e os valores críticos correspondentes v_k ao nível $\alpha = 0,05$ de significância.

Em seguida calculamos 100 conjuntos S_i, $i = 1, ..., 100$ com 1.000.000 de repetições aleatórios cada, sob a hipótese nula. Para cada conjunto e para cada valor de tamanho k determinamos scan_{ki}, a distribuição scan_k empírica correspondente para o conjunto S_i. Obtivemos a proporção p_{ki} (respectivamente q_{ki}) de LLR do CMV encontrado com valores maiores que v_0 (respectivamente v_k) na distribuição scan_{ki}, para cada valor de k e $i = 1, ..., 100$. Para cada valor de k, os 100 valores p_{ki} (respectivamente q_{ki}), para $i = 1, ..., 100$, foram usados para construir os intervalos de confiança (95%) U_k (respectivamente D_k), estimando assim a barra de erro na proporção de rejeição da hipótese nula ao longo do processo de inferência usual (respectivamente, data-driven).

Os gráficos na parte esquerda da Figura 7.5 mostram os resultados usando os dados para a área metropolitana de Belo Horizonte (acima) e do nordeste dos Estados Unidos (abaixo). Os gráficos na cor azul mostram as barras de erro para intervalos de confiança não paramétricos de 95% de confiança U_k, para cada tamanho de k, mostrando a média e a variabilidade da proporção de razão de verossimilhança maior que o valor crítico usual v_0, que emprega CMV's de todos os tamanhos k na distribuição scan empírica. Por outro lado, os gráficos em cores vermelhas mostram as barras de erro para os intervalos não-paramétricos de 95% de confiança do data-driven D_k, para cada tamanho de k,

mostrando a média e a variabilidade da proporção de razão de verossimilhança maior que os valores críticos v_k do data-driven, que separa CMV's em diferentes tamanhos k em suas correspondentes distribuições empíricas scan_k .

Para cada conjunto de dados, as barras de erro azul e vermelho são claramente distintas, mostrando as diferenças entre as inferências usual e data-driven. Um procedimento semelhante foi usado para calcular os intervalos de confiança não paramétricos para as distribuições correspondentes Gumbel_k ajustadas. Os conjuntos correspondentes de barras de erro para as distribuições Gumbel ajustadas são exibidos na parte direita da Figura 7.5, para a região metropolitana de Belo Horizonte (acima) e nordeste do Estados Unidos (abaixo). A partir da Figura 7.5 notamos que os valores obtidos através da inferência data-driven são substancialmente mais próximos do nível $0,05$, para ambos os mapas. A taxa de rejeição no scan usual é cerca de $0,03$-$0,04$ para pequenos clusters, significando que não está rejeitando o suficiente clusters pequenos, retornando assim muitos falsos positivos. O oposto acontece para clusters grandes.

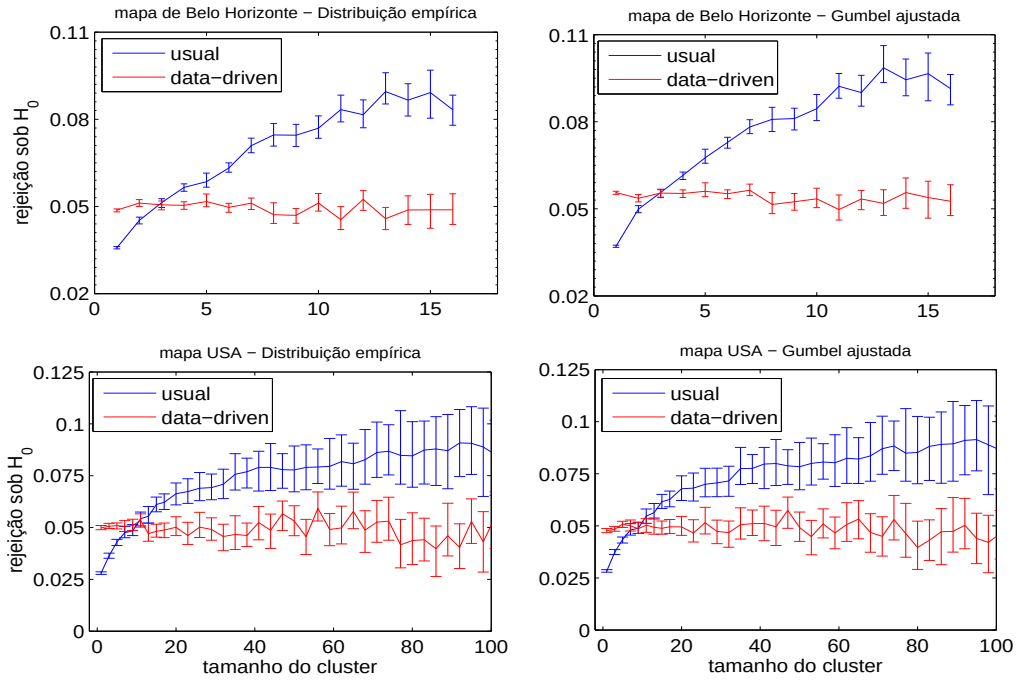


Figura 7.5: Proporção de rejeição da hipótese nula, para as distribuições empírica (esquerda) e Gumbel ajustada (à direita), para a região metropolitana de Belo Horizonte (acima) e do nordeste dos Estados Unidos(abaixo). Os intervalos de 95% de confiança não paramétricos são construídos a partir de 100 experimentos de Monte Carlo com 1.000.000 repetições cada, utilizando os processos usual (azul) e a inferência data-driven (vermelho).

7.6.2 Cálculo de valores críticos na prática

Do ponto de vista prático, deve-se preocupar com o grande número de simulações necessárias para obter um número razoável de repetições para que

seus CMV's tenham exatamente o tamanho do CMV observado, a fim de estimar o valor crítico na inferência data-driven. Mostraremos a seguir que, através de filtragem dos valores críticos v_k para os valores k próximos do tamanho do CMV observado, um número relativamente pequeno de repetições produz uma estimativa consistente do valor crítico.

A Figura 7.6 mostra os valores críticos da data-driven para o mapa do nordeste dos Estados Unidos para os valores de tamanho de k, utilizando 1.000.000 (respectivamente 50.000) repetições, exibidos como pontos vermelhos (respectivamente cruzes azuis). A curva preta sólida representa a média móvel, da janela de tamanho 20, dos valores críticos para cada tamanho $k > 10$ usando as 50.000 repetições (cruzes azuis).

Como pode ser observado na Figura 7.6, as médias móveis caem aproximadamente dentro do conjunto de valores críticos obtidos das 1.000.000 de repetições (pontos vermelhos). A linha horizontal tracejada indica o valor crítico usual . Este esquema é muito simples e suficientemente estável, permitindo a utilização de um pequeno número de repetições de Monte Carlo para estimar os valores críticos pela inferêcia data-driven. Para mapas menores (como o mapa da região metropolitana de Belo Horizonte), o esforço computacional é bem menor, e os valores críticos da data-driven são mais fáceis de calcular.

7.6.3 Discussão

O processo clássico usado na inferência de clusters espaciais utilizando estatística espacial scan de Kulldorff considera todos os clusters mais prováveis encontrados em repetições Monte Carlo sob a hipótese nula, a fim de construir a distribuição empírica da razão de verossimilhança, independentemente do tamanho do cluster e localização. Uma desvantagem potencial desta abordagem é que se assume implicitamente a independência das distribuições do logaritmo da razão de verossimilhança, quando restrita a vários tamanhos do cluster mais provável encontrados. Nós mostramos, por meio de experimentos numéricos, que esta hipótese não é verdadeira para o que ocorre comumente em mapas de dados reais. Dado que o cluster observado provavelmente tem o tamanho k_0, o que significa que o processo de inferência clássica calcula a sua significância com base no comportamento da maioria dos CMV's cujos tamanhos são diferentes de k_0.

Neste capítulo propusemos a inferência alternativa data-driven, que leva em conta apenas os clusters mais prováveis encontrados cujo tamanho é idêntico ao tamanho do cluster mais provável observado. Esta abordagem emprega uma comparação mais específica, evitando assim que o comportamento de clusters de tamanho muito pequeno sejam usados para decidir se um cluster observado grande seja considerado significativo, por exemplo.

Como o número de clusters mais prováveis de certo tamanho específico encontrados em mapas aleatórios é menor do que o número total de repetições Monte Carlo, uma preocupação que pode surgir é o esforço computacional para calcular a significância com o processo de inferência data-driven deve ser muito alto. No entanto, temos mostrado que com o uso da abordagem Gumbel semi-paramétrica e técnicas de filtragem simples, este esforço pode ser atenuado.

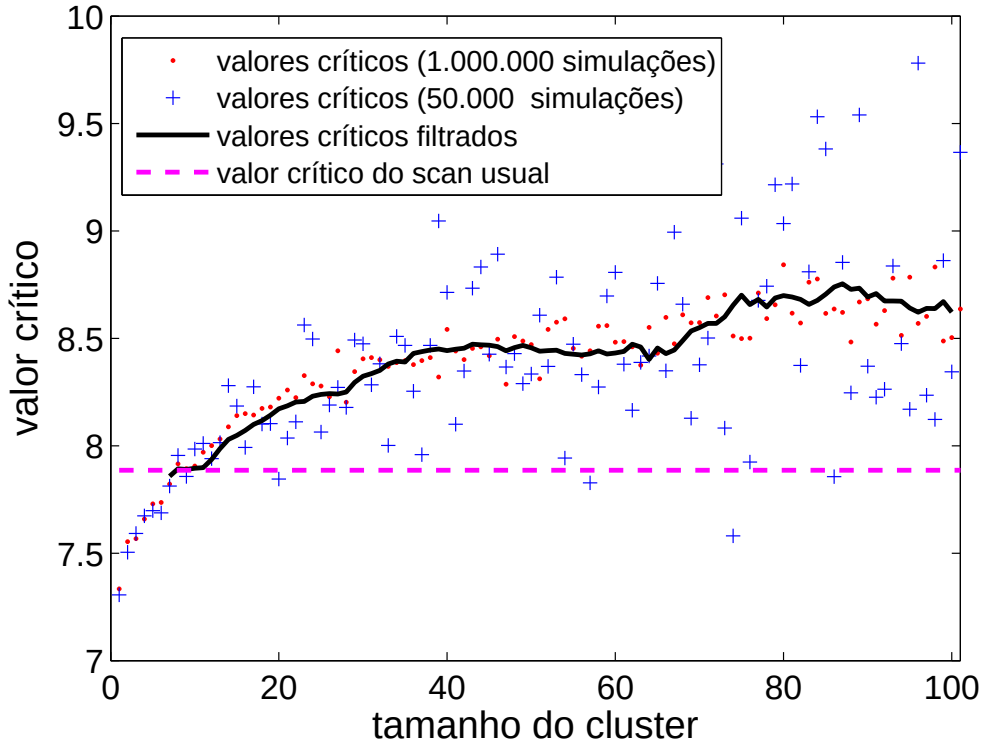


Figura 7.6: Valores críticos da data-driven para o mapa do nordeste do Estados Unidos para os vários valores de tamanho de k, utilizando 1.000.000 (respectivamente 50.000) repetições, exibido como pontos vermelhos (respectivamente cruzes azuis). A curva preta sólida representa a média móvel, para janela de tamanho 20, dos valores críticos associados a cada tamanho de $k > 10$ usando o conjunto de 50.000 repetições (cruzes azuis). A linha tracejada representa o valor crítico do scan usual.

Para cada valor de tamanho k existem diferentes distribuições de valores empíricos; em todos os exemplos estudados, a forma geral dessas distribuições varia suavemente quando k varia, continuando razoável conjecturar que suas médias aumentam monotonicamente com o aumento no valor de k. No entanto, é plausível imaginar uma situação hipotética em que essas mudanças se tornam abruptas com a variação de k, caracterizando instabilidade da hipótese nula. Esta instabilidade deve afetar ainda mais a inferência usual da estatística scan, pois a importância relativa dos diferentes valores de k não seria identificada, como foi feito na presente análise. Este tipo de situação ainda não foi experimentado, embora os experimentos da seção anterior possam servir para avaliar esse tipo de instabilidade em um mapa específico.

Deve-se ressaltar que a estratégia do presente trabalho de empregar o tamanho (número de áreas no interior do cluster) como critério para comparar os clusters mais prováveis é ainda sujeito a viés, devido à distribuição da população heterogênea. A solução para este problema deve ser o uso de ambas as variáveis, tamanho e população, o que seria muito caro, como discutido anteriormente.

O método proposto é mais útil quando o p-valor calculado usando a inferência clássica está próximo do nível de significância α, caso contrário, não haverá mudança no processo de decisão. Nesta situação em que o p-valor está próximo de α, recomendamos que a inferência data-driven deva ser utilizada.

Também fizemos alguns experimentos numéricos utilizando a inferência data-driven para o caso espaço-tempo, considerando não apenas o número de áreas no interior do cluster, mas também a duração do intervalo de tempo dos cilindros [57]. Nossos resultados preliminares sugerem que as diferenças nos valores críticos são ainda mais pronunciadas do que no cenário puramente espacial.

A inferência data-driven poderia ser aplicada ao caso/controle para conjuntos de dados pontuais, levando em consideração o número de casos e o tamanho da população no interior do clusters. Há três opções a considerar para a inferência data-driven neste tipo de conjunto de dados: aqueles com base no número de casos, na população, ou em ambos. Se basearmos no número de casos, desde que o número de casos seja pequeno, a inferência data-driven deve seguir a mesma linha do presente trabalho. Caso contrário, quando o número de casos é grande, ou quando a inferência data-driven for baseada na população, algum tipo de interpolação pode ser necessário.

Os resultados deste capítulo foram publicados em [2].

Capítulo 8

Considerações Finais e Trabalhos Futuros

8.1 Considerações Finais

Este trabalho foi desenvolvido como uma tentativa multidisciplinar de agregar áreas diferentes do conhecimento que buscam entender os fenômenos epidemiológicos. Neste sentido, não apresentamos a união definitiva das áreas, e sim, apresentamos um primeiro passo na desfragmentação do conhecimento na área de epidemiologia.

Apresentamos os modelos epidemiológicos SIR e MBI e as redes sociais mais utilizadas em modelagem de epidemias e propusemos um MBI para simular epidemias nestas redes. O MBI foi de crucial importância para o entendimento da propagação de epidemias em redes, bem como no desenvolvimento das novas ferramentas propostas. Simulações numéricas do MBI proposto sobre as redes sociais apresentadas e as análises de estimação de parâmetros e das curvas obtidas sugerem dificuldades do uso do modelo SIR para modelar epidemias em redes. O fato do coeficiente de aglomeração influenciar nos resultados das simulações indica a importância do modelo ser ao nível do indivíduo, e não da população como um todo.

A aproximação HMF foi apresentada, uma vez que é a técnica mais utilizada em propagação em redes. A modelagam por cadeias de Markov também foi apresentada, sendo esta abordagem mais recente e ao nível do indivíduo. Uma análise bibliográfica sobre tais modelos foi feita e nossos experimentos mostraram que os artigos que usaram a técnica com $\Delta_t = 1$ obtiveram resultados pouco precisos. Ainda ao nível do indivíduo, propusemos o modelo μSIR, cujos resultados demonstraram ser superiores ao demais métodos anteriormente citados. O μSIR foi proposto no sentido de ser um modelo analítico de propagação de epidemias em redes que fosse equivalente ao modelo SIR clássico e, como isso, acreditamos dar uma primeiro passo no sentido de agregar em uma mesma técnica o uso de equações diferenciais[1] e redes. Com a preocupação de desenvolver também uma ferramenta de uso prático, desenvolvemos um novo modelo HMF, o HMF-MC, capaz de modelar epidemias em redes com estrutura de comunidades a partir de dados que possam ser inferidos

[1] Apesar do HMF também ser um sistema de equações diferenciais, é uma aproximação e com uso restrito para redes com propriedades pré-determinadas.

de estudos estatísticos. Neste sentido, obtemos um modelo menos retritivo que o HMF clássico.

Do ponto de vista de vigilância, apresentamos a estatística espacial scan de Kulldorf, método de detecção mais utilizado na prática. Apresentamos ainda as estatísticas Workflow e baseada no valor esperado. Propomos a estatística WEB como a junção das duas técnicas e que demostrou bons resultados tanto para simulações de epidemias com dados aleatórios quanto epidemias modeladas por equações diferenciais. Com a estatística WEB, avançamos um passo no sentido de obter técnicas estatísticas para detecção de clusters espaciais em epidemias modeladas por equações diferenciais e, na linha que propomos, ajuda a desfragmentar os conhecimentos das duas áreas.

A estatística espacial scan de Kulldorf tem resultados muito bons e, por isso, poucas tentativas de melhorá-la do ponto de vista inferencial são encontradas. Neste sentido, propomos uma modificação para o teste de inferência do scan usual, incorporando informações adicionais sobre o tamanho do cluster mais verossímil encontrado. Apresentamos experimentos que justificam a necessidade de uma nova inferência mais apurada para valores próximos dos valores críticos e fornecemos um procedimento prático, data-driven, para fazer inferências mais precisas sobre o cluster mais verossímil encontrado pela estatística scan espacial. Bons resultados foram obtidos com o data-driven, cujos resultados foram publicados em [2], e entendemos que com este estudo não chegamos em uma fronteira final, e sim, a um primeiro passo para futuras melhorias de métodos de detecção de clusters.

O conjunto de propostas de modelos e técnicas apresentados neste trabalho tem sua relevância justificada no sentido de tentar resgatar conhecimentos em áreas distintas para a confecção de novos modelos e técnicas. Além disso, abrem novas perspectivas de caminhos a seguir no estudo epidemiológico, que discutimos na próxima seção.

8.2 Trabalhos Futuros

Um vez que o μSIR foi desenvolvido ao nível do indivíduo, pretendemos usá-lo para modelar e otimizar campanhas de vacinação em redes. Seu uso será de grande valia tanto no sentido de ganho computacional, um vez que o uso do MBI seria feito por simulações de Monte Carlo, quanto no sentido de precisão.

O HMF-MC proposto não é capaz de modelar a propagação de epidemias em redes com estruturas quaisquer, como por exemplo nas redes espaciais. Neste sentido, almejamos encontrar um modelo similar que não dependa da estruturação da rede ou, pelo menos, seja menos restritivo.

Dada uma epidemia que aconteceu em uma população com estrutura social conhecida, pergunta-se como estimar seus parâmetros e, caso a mesma epidemia seja inserida numa população com outra estrutura social conhecida, é desejável descobrir qual seu comportamento.

Dada a dinâmica de uma epidemia de parâmetros conhecidos e estrutura social desconhecida, deseja-se descobrir uma estrutura social na qual o comportamento desta epidemia seja equivalente à estrutura desconhecida, de modo

a poder predizer o comportamento de outra epidemia nesta mesma população. Neste sentido, pretendemos desenvolver métodos práticos de inferência desta estrutura de modo a poder utilizá-la no modelo HMF-MC ou em um modelo similar. Isso levaria a técnicas que poderiam ser usadas na prática pelos órgãos públicos de saúde.

Em uma mesma rede, epidemias diferentes terão diferentes comportamentos. Alguns indivíduos podem, por exemplo, permacer um tempo maior infectados e, com isso, serem grandes propagadores da infecção. Um possível trabalho futuro é encontrar uma estatística para detecção de clusters em redes de forma a encontrar possíveis conjuntos de indivíduos que devam ser isoladamente considerados em campanhas de saúde pública.

Ainda acreditamos ser possível melhorar as técnicas de detecção de clusters espaciais por meio das idéias apresentadas na inferência data-driven. Neste sentido, almejamos encontrar novas possíveis métricas na escolha do cluster como sendo o mais verossímil.

Apesar do data-driven ter sido proposto como uma técnica apenas espacial, sua versão espaço-temporal pode ser desenvolvida de forma natural. Neste sentido, pretendemos fazer experimentos e acreditamos que seus resultados sejam ainda mais expressivos (quando comparados à estatística espaço-temporal scan usual) que no caso espacial.

Apesar de termos conseguido responder a algumas perguntas, o número de questões que surgem com este trabalho é ainda maior. Neste sentido, a cada novo trabalho, novas perguntas surgirão.

Referências Bibliográficas

[1] Abrams, A. M., Kleinman, K., and Kulldorff, M. Gumbel based p-value approximations for spatial scan statistics. *International Journal of Health Geographics 9* (2010), 61. (online version). 79, 80

[2] Almeida, A. C. L., Duarte, A. R., Duczmal, L. H., Oliveira, F. L. P., and Takahashi, R. H. C. Data-driven inference for the spatial scan statistic. *International Journal of Health Geographics 10*, 1 (2011), 47. 88, 90

[3] Almeida Filho, N., and Rouquayrol, M. *Introdução à Epidemiologia*, 4a. ed. Guanabara Koogan, 2006. 7, 8

[4] Anderson, R. M., and May, R. M. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992. 1, 8, 9

[5] Arenas, A., Danon, L., Díaz-Guilera, A., Gleiser, P., and Guimerá, R. Community analysis in social networks. *The European Physical Journal B 38* (2004), 373–380. 20

[6] Arenas, A., Diaz-Guilera, A., and Perez-Vicente, C. J. Synchronization reveals topological scales in complex networks. *Physical Review Letters 96* (2006), 114102. 21

[7] Arino, J., Jordan, R., and van den Driessche, P. Quarantine in a multi-species epidemic model with spatial dynamics. *Mathematical Biosciences 206*, 1 (2007), 46–60. 33

[8] Badham, J., and Stocker, R. The impact of network clustering and assortativity on epidemic behaviour. *Theoretical Population Biology 77*, 1 (2010), 71–75. 25, 37

[9] Bagrow, J. P. Evaluating Local Community Methods in Networks, 2007. Disponível em 01/09/2011 em http://arxiv.org/abs/0706.3880. 21

[10] Barabasi, A., and Albert, R. Emergence of scaling in random networks. *Science 286*, 5439 (1999), 509–512. 12, 16, 48, 51

[11] Barrat, A., and Weigt, M. On the properties of small-world network models. *European Physical Journal B 13*, 3 (2000), 547–560. 14

[12] Barthelemy, M., Barrat, A., Pastor-Satorras, R., and Vespignani, A. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Physical Review Letters 92*, 17 (2004), 178701. 25, 37

[13] Bernoulli, D. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des advantages de l'inoculation pour la prévenir. *Mémoires de Mathématiques et de Physique.* (1760), 1–45. 2, 7

[14] Besag, J., and Newell, J. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society Series A- Statistics in Society 154*, Part 1 (1991), 143–155. 65, 67

[15] Bobashev, G., Morris, R. J., and Goedecke, D. M. Sampling for global epidemic models and the topology of an international airport network. *PLoS ONE 3*, 9 (2008), e3154. 33

[16] Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J., and Faloutsos, C. Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur. 10* (2008), 1:1–1:26. 45

[17] Chauhan, S., Girvan, M., and Ott, E. Spectral properties of networks with community structure. *Phys. Rev. E 80*, 5 (2009), 056114. 21

[18] Chen, J., and Yuan, B. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics 22*, 18 (2006), 2283–2290. 20

[19] Cheng, X.-Q., and Shen, H.-W. Uncovering the community structure associated with the diffusion dynamics on networks. *Journal of Statistical Mechanics: Theory and Experiment 2010*, 04 (2010), P04024. 21, 22

[20] Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.-J., and Vespignani, A. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Med 4*, 1 (2007), e13. 33

[21] Cosner, C., Beier, J. C., Cantrell, R. S., Impoinvil, D., Kapitanski, L., Potts, M. D., Troyo, A., and Ruan, S. The effects of human movement on the persistence of vector-borne diseases. *Journal of Theoretical Biology 258*, 4 (2009), 550–560. 33

[22] Cox, N., Tamblyn, S., and Tam, T. Influenza pandemic planning. *Vaccine 21*, 16 (2003), 1801–1803. 1

[23] Dorogovtsev, S., and Mendes, J. Scaling behaviour of developing and decaying networks. *Europhysics Letters 52*, 1 (2000), 33–39. 17

[24] Dorogovtsev, S. N., Goltsev, A. V., and Mendes, J. F. F. Critical phenomena in complex networks. *Reviews of Modern Physics 80*, 4 (2008), 1275–1335. 44

[25] Dorogovtsev, S. N., and Mendes, J. F. F. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, USA, 2003. 17

[26] Duczmal, L., and Buckeridge, D. A workflow spatial scan statistic. *Statistics in Medicine 25*, 5 (2006), 743–754. 2, 64, 68, 75

[27] Duczmal, L., Cancado, A. L. F., Takahashi, R. H. C., and Bessegato, L. E. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis 52*, 1 (2007), 43–52. 64

[28] Eames, K., and Keeling, M. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proceedings of the National Academy of Sciences of the United States of America 99*, 20 (2002), 13330–13335. 18

[29] Eames, K. T. D. Modelling disease spread through random and regular contacts in clustered populations. *Theoretical Population Biology 73*, 1 (2008), 104–111. 25, 37

[30] Eguiluz, V., and Klemm, K. Epidemic threshold in structured scale-free networks. *Physical Review Letters 89*, 10 (2002). 19

[31] Erdos, P., and Renyi, A. On random graphs. *Publications Mathematicae 6* (1959), 290. 11

[32] Erdos, P., and Renyi, A. On the evolution of random graphs. *Bulletin of the International Statistical Institute 38*, 4 (1960), 343–347. 11

[33] Erdos, P., and Renyi, A. On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary 12* (1961), 261–267. 11

[34] Flake, G. W., Lawrence, S., Giles, C. L., and Coetzee, F. M. Self-organization and identification of web communities. *IEEE Computer 35* (2002), 66–71. 20

[35] Forattini, O. P. *Conceitos básicos de epidemiologia molecular*, 1 ed. EDUSP, 2005. 8

[36] Fortunato, S. Community detection in graphs. *Physics Reports 486*, 3-5 (2010), 75 – 174. 21

[37] Fortunato, S., and Barthelemy, M. Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA 104* (2007), 36. 21

[38] Giacomini, H. C. Sete motivações teóricas para o uso da modelagem baseada no indivíduo em ecologia. *Acta Amazonica 37* (2007), 431–446. 26, 27

[39] Girvan, M., and Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*, 12 (2002), 7821–7826. 20, 21

[40] Gòmez, S., Arenas, A., Borge-Holthoefer, J., Meloni, S., and Moreno, Y. Discrete-time markov chain approach to contact-based disease spreading in complex networks. *EPL (Europhysics Letters) 89*, 3 (2010), 38009. 45

[41] Grimm, V. Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? *Ecological Modelling* (1999), 129–148. 26

[42] Guerra, B., and Gomez-Gardenes, J. Annealed and mean-field formulations of disease dynamics on static and adaptive networks. *Phys. Rev. E 82*, 3, Part 2 (2010). 45, 47, 50

[43] Guimera, R., and Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature 433* (2005), 895. 21

[44] Hébert-Dufresne, L., Noël, P.-A., Marceau, V., Allard, A., and Dubé, L. J. Propagation dynamics on networks featuring complex topologies. *Phys. Rev. E 82*, 3 (2010), 036115. 51

[45] Hethcote, H. W. The mathematics of infectious diseases. *SIAM review 42*, 4 (2000), 599–653. 8, 9, 26, 28

[46] Higham, D. J. An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations. *SIAM Review 43*, 3 (2001), 525–546. 36

[47] Holme, P., and Kim, B. J. Growing scale-free networks with tunable clustering. *Phys. Rev. E 65*, 2 (2002), 026107. 19

[48] Huston, M., Deangelis, D., and Post, W. New computer-models unify ecological theory - computer-simulations show that many ecological patterns can be explained by interactions among individual organisms. *Bioscience 38*, 10 (1988), 682–691. 26

[49] Hwang, D., Boccaletti, S., Moreno, Y., and Lopez-Ruiz, R. Thresholds for epidemic outbreaks in finite scale-free networks. *Mathematical Biosciences and Engineering 2*, 2 (2005), 317–327. 17, 25, 37

[50] Jonsson, P., Cavanna, T., Zicha, D., and Bates, P. Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics 7* (2006), 1–13. 20

[51] Kaiser, H. Dynamics of populations as result of the properties of individual animals. *Fortschritte der Zoologie 25*, 2-3 (1979), 109–136. 26

[52] Kaiser, M. Spatial network growth - generating small world, scale-free, and multi-cluster spatial networks. Tech. rep., School of Engineering and Science, International University Bremen, 2005. 19

[53] Kaiser, M., and Hilgetag, C. Spatial growth of real-world networks. *Phys. Rev. E 69*, 3, Part 2 (2004). 19

[54] Keeling, M., and Grenfell, B. Individual-based perspectives on r-0. *Journal of Theoretical Biology* (2000), 83–92. 26

[55] Kermack, W., and McKendrick, A. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences 115* (1927), 700–721. 3, 8, 26

[56] Kulldorff, M. A spatial scan statistic. *Communications in Statistics-Theory and Methods 26*, 6 (1997), 1481–1496. 2, 65, 66, 67, 68, 71

[57] Kulldorff, M. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal Of The Royal Statistical Society Series A 164*, 1 (2001), 61–72. 88

[58] Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R., and Mostashari, F. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine 2*, 3 (2005), 216–224. 64

[59] Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. An elliptic spatial scan statistic. *Statistics in Medicine 25* (2006), 3929–3943. 64, 80

[60] Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W. K., Kleinman, K., and Platt, R. Multivariate scan statistics for disease surveillance. *Statistics In Medicine 26*, 8 (2007), 1824–1833. 64

[61] Kulldorff, M., and Nagarwalla, N. Spatial disease clusters - detection and inference. *Statistics in Medicine 14*, 8 (1995), 799–810. 65, 66

[62] Lawson, A. B. *Statistical Methods in Spatial Epidemiology*. Wiley, 2006. 65

[63] Lloyd, S. Least squares quantization in PCM. *Information Theory, IEEE Transactions on 28*, 2 (1982), 129 – 137. 23

[64] Maeno, Y. Discovering network behind infectious disease outbreak. *Physica A: Statistical Mechanics and its Applications 389*, 21 (2010), 4755 – 4768. 63

[65] Marshall, R. J. A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 154*, 3 (1991), 421–441. 65

[66] Masuda, N. Immunization of networks with community structure. *New Journal of Physics 11*, 12 (2009), 123018. 51

[67] Miller, J. C. Spread of infectious disease through clustered populations. *Journal of the Royal Society Interface 6*, 41 (2009), 1121–1134. 2, 25, 37, 43

[68] Moreno, Y., Pastor-Satorras, R., and Vespignani, A. Epidemic outbreaks in complex heterogeneous networks. *European Physical Journal B 26*, 4 (2002), 521–529. 44

[69] Naus, J. I. The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association 60* (1965), 532–538. 64

[70] Naus, J. I. A power comparison of two tests on nonrandom clustering. *Technometrics 8* (1966), 493–517. 64

[71] Naus, J. I. Some probabilities expectations and variances of the largest clusters and smallest intervals. *Journal of the American Statistical Association. 61* (1966), 1191–1199. 64

[72] Neill, D. B. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics 8* (2009), 20. 65, 70

[73] Nelson, K. E., Williams, C. M., and Graham, N. M. H. *Infectious Disease Epidemiology: Theory and Practice*. Gaitherseburg, Maryland: Aspen Publication, 2001. 1

[74] Nepomuceno, E. G. *Dinâmica, Modelagem e Controle de Epidemias*. PhD thesis, Universidade Federal de Minas Gerais (UFMG), 2005. 3, 28

[75] Newman, M. E. J. The structure and function of complex networks. *Siam Review 45*, 2 (2003), 167–256. 11, 19

[76] Newman, M. E. J. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E 74* (2006), 036104. 21

[77] Newman, M. E. J. Random Graphs with Clustering. *Physical Review Letters 103*, 5 (2009), 058701. 19, 25, 37

[78] Newman, M. E. J., Strogatz, S. H., and Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E 64*, 2, Part 2 (2001). 14

[79] Noel, P.-A., Davoudi, B., Brunham, R. C., Dube, L. J., and Pourbohloul, B. Time evolution of epidemic disease on finite and infinite networks. *Phys. Rev. E 79*, 2, Part 2 (2009), 026101. 2, 25, 37

[80] Papoulis, A. *Probability, random variables, and stochastic processes*, 3 ed. McGraw-Hill, 1991. 31

[81] Pastor-Satorras, R., and Vespignani, A. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E 63*, 6, Part 2 (2001), 066117. 2, 25, 26, 37, 44

[82] Pastor-Satorras, R., and Vespignani, A. Epidemic spreading in scale-free networks. *Physical Review Letters 86*, 14 (2001), 3200–3203. 25, 26, 37, 44

[83] Pastor-Satorras, R., and Vespignani, A. Epidemic dynamics in finite size scale-free networks. *Phys. Rev. E 65*, 3, Part 2A (2002). 25

[84] Pastor-Satorras, R., and Vespignani, A. Immunization of complex networks. *Phys. Rev. E 65*, 3, Part 2A (2002), 036104. 2, 25

[85] Piqueira, J., Navarro, B., and Monteiro, L. Epidemiological models applied to viruses in computer networks. *Journal of Computer Science 1* (2005), 31–34. 2, 7

[86] Redner, S. How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems 4*, 2 (1998), 131–134. 21

[87] Rives, A. W., and Galitski, T. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America 100*, 3 (2003), 1128–1133. 20

[88] Salathé, M., and Jones, J. H. Dynamics and control of diseases in networks with community structure. *PLoS Comput Biol 6*, 4 (2010), e1000736. 51

[89] Satsuma, J., Willox, R., Ramani, A., Grammaticos, B., and Carstea, A. Extending the SIR epidemic model. *Physica A-statistical Mechanics And Its Applications 336*, 3-4 (2004), 369–375. 8

[90] Seber, G. A. F. *Multivariate Observations*. Wiley, New York, 1984. 23

[91] Serrano, M. A., and Boguna, M. Clustering in complex networks - I. General Formalism. *Phys. Rev. E 74*, 5, Part 2 (2006). 19

[92] Serrano, M. A., and Boguna, M. Percolation and epidemic thresholds in clustered networks. *Physical Review Letters 97*, 8 (2006), 088701. 2, 25, 37

[93] Shen, H.-W., Cheng, X.-Q., and Fang, B.-X. Covariance, correlation matrix and the multi-scale community structure of networks. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics 82*, 1 Pt 2 (2010), 10. 21

[94] Sonesson, C., and Bock, D. A review and discussion of prospective statistical surveillance in public health. *Journal Of The Royal Statistical Society Series A 166*, 1 (2003), 5–21. 65

[95] Spath, H. *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples, translated by J. Goldschmidt.* Halsted Press, New York, 1985. 23

[96] Terry, A. J. Pulse vaccination strategies in a metapopulation sir model. *Mathematical Biosciences and Engineering 7*, 2 (2010), 455–477. 2, 33

[97] Thacker, S., Stroup, D., Parrish, R., and Anderson, H. Surveillance in environmental public health: Issues, systems, and sources. *American Journal of Public Health 86*, 5 (1996), 633–638. 9

[98] Wang, r., Zhang, S., Wang, Y., Zhang, x., and Chen, L. Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures. *Neurocomput. 72*, 1-3 (2008), 134–141. 21

[99] Wang, W., and Zhao, X. An epidemic model in a patchy environment. *Mathematical Biosciences 190*, 1 (2004), 97–112. 33

[100] Watts, D., and Strogatz, S. Collective dynamics of 'small-world' networks. *Nature 393*, 6684 (1998), 440–442. 12

[101] Watts, D. J. *Small worlds: the dynamics of networks between order and randomness.* Princeton University Press, Princeton, NJ, USA, 1999. 12

[102] Woodall, W. H., Marshall, J. B., Jr, M. D. J., Fraker, S. E., and Abdel-Salam, A.-S. G. On the use and evaluation of prospective scan methods for health-related surveillance. *Journal of The Royal Statistical Society Series A 171*, 1 (2008), 223–237. 65

[103] Xulvi-Brunet, R., and Sokolov, I. M. Evolving networks with disadvantaged long-range connections. *Phys. Rev. E 66*, 2 (2002), 026118. 19

[104] Xulvi-Brunet, R., and Sokolov, I. M. Growing networks under geographical constraints. *Phys. Rev. E 75*, 4 (2007), 046117. 19

[105] Yang, H. M. *Epidemiologia matemática: Estudos dos efeitos da vacinação em doenças de transmissão direta.* Athena Scientific, 2001. 2

[106] Zarei, M., Samani, K. A., and Omidi, G. R. Complex eigenvectors of network matrices give better insight into the community structure. *Journal of Statistical Mechanics: Theory and Experiment 2009*, 10, P10018. 21