ões649

UNIVERSIDADE FEDERAL DE MINAS GERAIS - UFMG
INSTITUTO DE CIÊNCIAS EXATAS - ICEx
DEPARTAMENTO DE ESTATÍSTICA

Dissertação de Mestrado:

DESENVOLVIMENTO DE UMA MEDIDA DE ASSOCIÇÃO ENTRE ESPAÇO E TEMPO

Fábio Rocha da Silva
Orientador: Renato Martins Assunção

FÁBIO ROCHA DA SILVA

DESENVOLVIMENTO DE UMA MEDIDA DE ASSOCIÇÃO ENTRE ESPAÇO E TEMPO

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial a obtenção do título de Mestre em Estatística.

Orientadora: Prof. Dr. Renato Martins Assunção.

Universidade Federal de Minas Gerais
Belo Horizonte
Fevereiro de 2011

Dedico este trabalho
a minha família, Maria, Manoel e Heraldo.

Agradecimentos

Agradeço a Deus pelas conquistas que tem me proporcionado, pelas pessoas que tem colocado na minha vida e pela minha saúde.

Aos meus pais, Maria Rocha e Manoel de Fátima e ao meu irmão Heraldo que são mais que tudo na minha vida. Agradeço a eles por terem me ensinado a forma mais sublime do amor e por nunca faltarem com seu carinho e atenção.

À minha família e aos amigos pela confiança e apoio.

Aos amigos que fiz na UFMG, durante a graduação e o mestrado, pelas discussões estatísticas que, de forma direta ou indireta, me ajudaram nesta dissertação, pelos dias de estudo em conjunto para as inúmeras provas e listas de exercícios e, principalmente, pelos momentos de descontração tão importantes e necessários. Agradeço principalmente ao amigo José Luiz e ao amigo Cristiano Carvalho por terem sido muito mais do que amigos em todo este período.

À minha orientadora, Professor Renato M Assunção pela confiança, paciência, dedicação e comprometimento ao me orientar durante vários anos de estudos na UFMG.

Aos membros da banca examinadora, professores Marcelo Azevedo Costa (UFMG), (UFES) pelas contribuições, correções e sugestões na dissertação.

Aos professores do departamento de Estatística pelos conhecimentos compartilhados durante estes seis anos de convívio.

À CAPES e à FAPEMIG pela bolsa de mestrado, uma vez que iniciei recebendo bolsa da primeira instituição e depois, para atender os interesses do departamento, passei a receber bolsa da outro instituição. À FAPEMIG por diversos apoios finaceiros prestados para a participação em eventos e ao CNPq pela bolsa de iniciação cientifica que me colocou em contato com pesquisas e despertou meu interesse pela área acadêmica.

Enfim, peço a Deus que abençoe sempre aqueles que estiveram sempre do meu lado me dando apoio, peço que o homem lá de cima nunca lhes deixem faltar o que vocês não me deixaram faltar a "FÉ".

Muito Obrigado!

Fábio Rocha da Silva

Resumo

Existem varias técnicas estatísticas para testar a hipótese de que os tempos e posições de eventos pontuais em R^3 são independentes. Isto é, testar se casos que estão próximos no tempo tendem a estar próximos no espaço também. Estes testes sofrem de um problema típico dos testes de hipótese: se existir muitos eventos, o teste pode ser significativo mesmo se a associação entre tempo e espaço for fraca.

Existem propostas de medidas de associação para tabelas de contingência que procuram corrigir este problema. Neste trabalho, tentaremos adaptar estas idéias e introduzir uma nova medida de associação para dados contínuos. Esta medida pode ser usada em estudos de processos pontuais espaço-temporais. O objetivo deste trabalho é o desenvolvimento de metodologia estatística para medir o grau de associação entre as coordenadas espaciais e as coordenadas temporais de eventos pontuais.

Palavras-chave: Medida de associação, espaço-tempo, τ de Goodman e Kruskal

Sumário

Página

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Introdução

A consideração simultânea dos padrões espaciais e temporais da ocorrência dos eventos é importante para identificar clusters ou conglomerados espaços-temporais. Definimos o cluster espaço-temporal como uma região geograficamente pequena em relação à região em estudo e que concentra um número excessivo de eventos durante um período limitado de tempo.

O teste de detecção de conglomerados espaços-temporais mais popular foi desenvolvido por Knox (1964). Especificando-se distâncias críticas temporais e espaciais é possível determinar se um par de eventos está próximo no tempo e no espaço. O teste baseia-se no número X de pares de eventos que estão simultaneamente próximos no espaço e no tempo. Um alto valor X seria uma indicação de que há uma tendência de casos próximos no tempo serem também próximos no espaço, retratando a interação espaço-tempo.

O teste de Knox, assim como as outras técnicas para testar a hipótese de independência entre espaço e tempo, sofre de um problema típico dos testes de hipóteses: se existirem muitos eventos, o teste pode ser significativo mesmo se a associação entre espaço tempo for fraca.

Seria de grande valia termos uma medida de associação que possa ser usada em conjunto com o teste de hipóteses e que mensure a magnitude da possível relação entre as variáveis. Existem diversas propostas de medidas de associação para tabelas de contingência que procuram complementar o teste qui-quadrado de Pearson. Uma destas propostas foi proposta por Goodman e Kruskal (1954) denominada tau de Goodman e Kruskal.

O tau de Goodman e Kruskal (notação: τ_{GK}) é obtido usando o principio da redução proporcional dos erros. O coeficiente tem por objetivo responder à questão: Em que medida o fato de conhecermos a classificação de uma das variáveis (por exemplo, a linha da tabela em que a observação se encontra) nos torna mais hábeis para prevermos a classificação da outra variável (a coluna na qual cai a observação)?

Esta estatística tem algumas propriedades desejáveis, tais como ter limites zero (nenhuma associação) e um (completa associação) e não mudar o seu valor com a permutação de linhas e colunas.

Além disso, τ_{GK} tem uma interpretação muito clara: mede o decréscimo relativo na pro-

babilidade de errar a previsão da variável linha ao conhecer a variável coluna (ou vice versa). Por exemplo, se $\tau_{GK} = 0.8$, isto significa que temos uma redução de 80% na probabilidade de errar a previsão de uma das variáveis, quando se usa a informação sobre a outra variável.

A presente dissertação tem por objetivo adaptar as idéias da medida de associação de Goodman e Kruskal e introduzir uma nova medida de associação para dados contínuos. Esta medida pode ser usada em estudos de processos pontuais espaços-temporais. O objetivo deste trabalho é o desenvolvimento de metodologia para medir o grau de associação entre as coordenadas espaciais e as coordenadas temporais de eventos pontuais e o estudo de suas possíveis propriedades.

Um objetivo adicional deste trabalho é utilizar a nossa medida de associação em vetores aleatórios contínuos. Nossa intenção é ter uma medida de associação capaz de captar relações não-lineares entre variáveis aleatórias.

Capítulo 2

Conceitos e Definições Gerais

2.1 Processos Pontuais

Um conceito de grande importância na análise de fenômenos espaciais é a dependência espacial entre as observações. As inferências nesse tipo de dado não são tão eficientes como em amostras independentes. Existe uma perda de poder explicativo, dado que as variâncias maiores para as estimativas levam a níveis menores de significância em testes de hipóteses e a um pior ajuste para os modelos estimados. Assim, considera-se os dados espaciais não como um conjunto de amostras independentes, mas como uma realização de um processo estocástico. Nesse processo, todas as observações são utilizadas conjuntamente para descrever o padrão espacial do fenômeno estudado.

Usualmente, os dados espaciais podem ser classificados em três grandes grupos: processos pontuais (eventos ou padrões pontuais); variação contínua (superfícies contínuas); e variação discreta (áreas com contagens e taxas agregadas). Em particular, eventos ou padrões pontuais são fenômenos cujas ocorrências são identificadas como pontos localizados no espaço. São exemplos de processos pontuais a localização de crimes, a ocorrência de crimes, a ocorrência de doenças e a localização de espécies vegetais.

Tecnicamente, processos pontuais são definidos como um conjunto de pontos distribuídos em uma área, cuja localização foi gerada por um mecanismo estocástico (Diggle (2003)). O conjunto desses pontos é denominado padrão espacial de pontos e um ponto em particular é denominado evento. O objetivo é estudar a distribuição espacial dos eventos, testando hipóteses sobre o padrão observado: se ele é aleatório, se apresenta aglomerados, regularidade na distribuição ou outras hipóteses de interesse.

Na Figura 2.1 existem dois padrões espaciais de pontos que parecem ser diferentes. A primeira figura não mostra nenhuma estrutura óbvia e deve ser considerada como um padrão completamente aleatório. Por outro lado, a segunda figura evidencia uma clara formação de aglomerados, que requer alguma explicação apropriada.

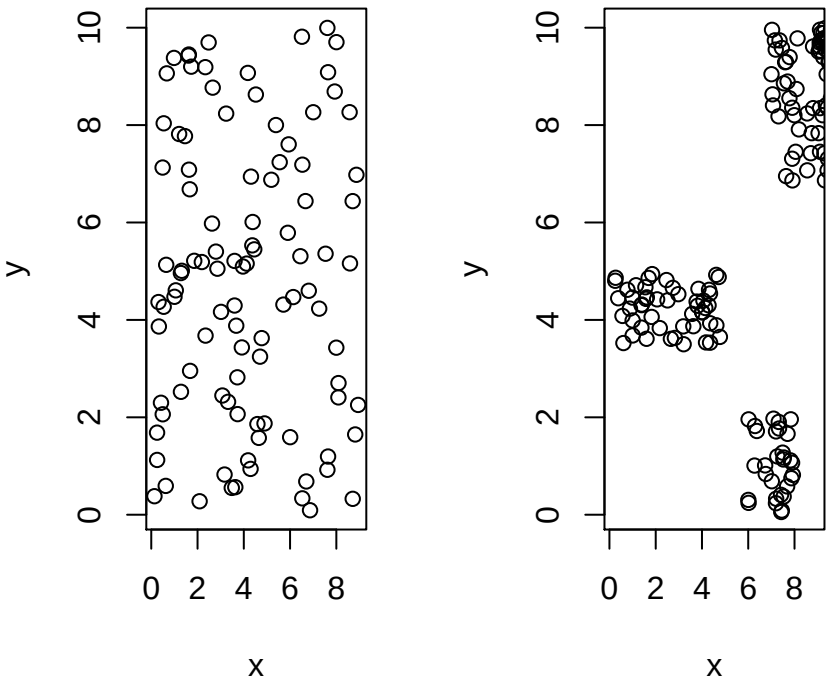

Figura 2.1: Exemplos de Processos Pontuais.

O interesse maior é encontrar sub-regiões de uma área em estudo com maior probabilidade de ocorrência de eventos ou de maior intensidade. No modelo de completa aleatoriedade, considera-se que os eventos têm igual probabilidade de ocorrência em toda a região e que suas posições são independentes umas das outras. Essa formulação permite estabelecer uma base de comparação entre uma distribuição completamente aleatória e os eventos observados.

2.2 Conglomerados Espaço-Temporais

Uma situação de grande interesse prático é quando o tempo de ocorrência dos eventos é registrado. Assim, o interesse é verificar se espaço e tempo interagem, isto é, se eventos aglomeram no espaço e no tempo simultaneamente.

Muitos estudos avaliam se existe correlação puramente espacial ou puramente temporal de eventos. É comum encontrar substancial variação espacial refletindo a distribuição geográfica não-uniforme da população de risco ou dos fatores ambientais, assim como é usual encontrar aglomerados temporais devido a efeitos sazonais ou tendências de crescimento ou decrescimento acentuado da taxa de ocorrência dos eventos, ao longo do tempo. No entanto, quando as informações tanto de espaço quanto de tempo estão disponíveis, pode-se testar a existência de aglomerados no espaço e no tempo simultaneamente, após ajustar por possíveis

variações puramente espaciais ou puramente espaciais. O objetivo é testar se casos próximos no espaço tendem a estar também próximos no tempo. Se isto ocorre, pode-se dizer que existem aglomerados espaço-temporais ou que os dados exibem interação espaço-tempo.

As ocorrências de eventos no tempo e no espaço são registradas em vários problemas aplicados. Na área da saúde, por exemplo, pode-se observar o dia e a localização da morte de um indivíduo ou o dia da eclosão e a área geográfica de novos casos de certa doença. Na análise de crimes, registram-se delitos pela sua hora, data de ocorrência e pela área onde eles ocorreram dentro de uma cidade. Em ecologia há interesse no padrão espacial de espécies de fauna e flora, e também onde e como distribuições geográficas particulares mudam com o tempo. Em astronomia, existe interesse na distribuição espacial de estrelas e galáxias, bem como na questão de onde e quando esses padrões espaciais mudam com o tempo.

Na análise de conglomerados espaço-temporais, os dados em geral consistem de um conjunto de eventos no espaço euclidiano bidimensional, dentro de uma região poligonal no espaço e entre limites temporais superiores e inferiores. De uma forma geral, busca-se detectar a existência de pequenas regiões, em breves momentos do tempo, em que existe um número acima do esperado de casos excessivamente próximos no espaço e no tempo, considerando uma distribuição de referência. Se um padrão de eventos pontuais observados apresentar desvios significativos do comportamento esperado para essa distribuição, há a indicação da existência de uma distribuição espaço-tempo diferente da distribuição de referência, que merece ser objeto de maior análise.

O teste de detecção de conglomerados espaços-temporais mais popular foi desenvolvido por Knox (1964). Especificando-se distâncias críticas temporais e espaciais é possível determinar se pares de eventos estão próximos no tempo e no espaço. O teste baseia-se no número X de pares de eventos que estão simultaneamente próximos no espaço e no tempo. Um alto valor X seria uma indicação de que há uma tendência de casos próximos no tempo serem também próximos no espaço, retratando a interação espaço-tempo.

2.3 Teste de Knox

O teste proposto por Knox (1964) é um método puramente retrospectivo, voltado para testar globalmente a presença de conglomerados espaço-temporais em processos pontuais. As hipóteses a serem testadas pelo teste de Knox são:

H_0 : Os tempos de ocorrência dos eventos são distribuídos aleatoriamente através das localizações. Isto é, as distâncias no tempo entre pares de eventos são independentes de suas distâncias no espaço entre pares de eventos.

H_1 : Pares de eventos próximos no espaço tendem a ser próximos no tempo.

Este teste baseia-se na contagem do número de pares de eventos que ocorrem em intervalos críticos pré-especificados de tempo e distância. Considerando-se n pontos existem $n(n-1)/2$ pares de pontos distintos. Seja n_s o número observado de pares de eventos que são próximos

no espaço (ou seja, pares separados por uma distância espacial menor que a distância crítica espacial). Seja n_t o número observado de pares de eventos que são próximos no tempo (ou seja, pares separados por uma distância temporal menor que a distância crítica temporal). As distâncias críticas devem ser definidas pelo usuário de acordo com o seu conhecimento sobre o processo.

A estatística de teste é simplesmente n_{st}, o número observado de pares de eventos que são próximos no espaço e no tempo simultaneamente. A estatística de teste excede seu valor esperado $2n_s n_t/n(n-1)$ quando pontos que são próximos no espaço são mais próximos no tempo que o esperado.

Esta estatística é comparada com uma distribuição de referência (sob a hipótese nula de que o processo não apresenta interação espaço-tempo) que é obtida através de permutações aleatórias dos índices de tempo dos eventos originais. Portanto, um valor-p pequeno é uma evidência a favor da hipótese de interação espaço-tempo.

2.4 Tabelas de Contingência

Se os dados forem classificados de acordo com dois critérios (isto é, com duas classificações marginais) estamos tratando das clássicas *tabelas de contingência*(Miles e Huberman (1984)). As tabelas de contingência são utilizadas para estudar a relação entre duas variáveis categóricas descrevendo a frequências das categorias de uma das variáveis relativamente às categorias de outra.

A forma geral de uma tabela de contingência é exemplificada na Tabela 5.1 na qual uma amostra das n observações é classificada relativamente a duas variáveis qualitativas, uma com r categorias e outra com c categorias. É designada por tabela de contingência $r \times c$. A freqüência observada ou contagem na categoria i da variável linha e na categoria j da variável coluna, é representada por n_{ij}. O total de observações na categoria i da variável linha é $n_{i.}$ e o total de observações na categoria j da variável coluna é $n_{.j}$. Estes são designados por totais marginais, e em termos das freqüências das caselas, n_{ij} , são expressos por:

Tabela 2.1: Forma geral de uma tabela de contingência

Linhas	Colunas (Variável C)				
(Variável R)	1	2	$\cdots$	c	Total
1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	n

2.4.1 Hipótese de independência dos critérios de classificação

A independência dos critérios de classificação significa a independência das variáveis R e C. Esta independência é expressa probabilisticamente pelo conjunto das seguintes igualdades:

$$p_{ij} = p_{i.} \times p_{.j} \qquad i = 1, \cdots, r \quad j = 1, \cdots, c$$

na qual $p_{i.} = P$(uma observação ser classificada na categoria i da variável R) $i = 1, \cdots, r$ e $p_{.j} = P$(uma observação ser classificada na categoria j da variável R) $j = 1, \cdots, c$.

Evidentemente que:

$$p_{i.} = \sum_{j=1}^{c} p_{ij} \quad i = 1, \cdots, r \quad \text{e,}$$

$$p_{.j} = \sum_{i=1}^{r} p_{ij} \quad j = 1, \cdots, c$$

Em resumo se queremos testar a independência das variáveis R e C, as hipóteses que estão em jogo serão:

$$\begin{aligned} H_0 &: p_{ij} = p_{i.} \cdot p_{.j} \quad i = 1, \cdots, r \quad j = 1, \cdots, c \\ H_1 &: \exists (i, j) : p_{ij} \neq p_{i.} \cdot p_{.j} \end{aligned}$$

2.4.2 Teste do qui-quadrado para independência

Os estimadores de máxima verosimilhança de $p_{i.}$ e de $p_{.j}$ são naturalmente:

$$\hat{p}_{i.} = \frac{n_{i.}}{n} \quad i = 1, \cdots, r$$

$$\hat{p}_{.j} = \frac{n_{.j}}{n} \quad j = 1, \cdots, c$$

Então se a hipótese H_0 (hipótese de independência) é válida, os estimadores de máxima verosimilhança de p_{ij} serão:

$$\hat{p}_{ij} = \hat{p}_{i.} \times \hat{p}_{.j} = \frac{n_{i.}}{n} \frac{n_{.j}}{n}$$

e os estimadores de máxima verosimilhança das frequências esperadas E_{ij}, sob H_0 serão:

$$E_{ij} = n \hat{p}_{i.} \hat{p}_{.j} = n \frac{n_{i.}}{n} \frac{n_{.j}}{n} = \frac{n_{i.} n_{.j}}{n} \quad i = 1, \cdots, r \quad j = 1, \cdots, c$$

De acordo com a distribuição multinomial do vetor de frequências $(n_{11}; n_{12}; \cdots; n_{rc})$ e admitindo que é válida a hipótese de independência dos critérios de classificação, a estatística

$$\chi^2_{rc} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

tem distribuição assintótica do qui-quadrado com $(r-1)(c-1)$ graus de liberdade.

Pela expressão da estatística χ^2_{rc} podemos entender qual a região crítica do teste de independência. Quando não ocorre independência é natural que as frequências observadas n_{ij} sejam substancialmente diferentes das frequências que esperamos observar quando a independência ocorre, E_{ij}. Então devemos rejeitar a hipótese de independência dos critérios de classificação quanto χ^2_{rc} tem um valor bastante grande, isto é, quando $\chi^2_{rc} > k$.

Dado um nível de significância $\alpha = P(\chi^2_{rc} > k|H_0 \text{ verdadeira})$ conclui-se que $k = \chi^2_{(r-1)(c-1):(1-\alpha)}$. Em resumo, devemos rejeitar a hipótese de independência dos critérios de classificação, ao nível de significância α, se

$$\chi^2_{rc} > \chi^2_{(r-1)(c-1):(1-\alpha)}$$

Uma decisão em termos de P-value, será: Rejeitar a hipótese nula dos critérios de classificação, ao nível de significância α, se

$$P - valor = P\left(\chi^2_{(r-1)(c-1):(1-\alpha)} > \chi^2_{rc}\right) < \alpha$$

Problemas com o teste qui-quadrado

Consideremos a seguinte classificação da atividade cerebral (em " retardada"e "não retardada") de 100 doentes psiquiátricos (classificados em "com desordens afetivas", e "neuróticos").

Tabela 2.2: Exemplo de uma tabela de contingência

	Desordens afetivas	Neuroses	Total
Atividade retardada	26	24	50
Atividade não-retardada	24	26	50
Total	50	50	100

Se quisermos testar se o tipo de atividade não é influenciada pelo tipo de desordem psiquiátrica, então devemos estimar as frequências esperadas e após o cálculo da estatística de teste, $\chi^2_{rc} = \sum \frac{(n_{ij}-E_{ij})^2}{E_{ij}} = 0.16$ e como $P-valor = P\left(\chi^2_{1:(0.95)} > 0.16\right) = 0.6892$ concluímos, com 5% de significância, que existem evidência de que as variáveis atividade e desordem psiquiátrica são independentes uma vez que o p-valor inferior ao nível de significância.

Suponhamos agora que iremos multiplicar o tamanho amostral por 100, mantendo as mesmas proporções em cada célula da tabela de contingência:

Tabela 2.3: Exemplo de uma tabela de contingênciacom o tamanho amostral multiplicado por 100

	Desordens afetivas	Neuroses	Total
Atividade retardada	2600	2400	5000
Atividade não-retardada	2400	2600	5000
Total	5000	5000	10000

Calculando a estatística χ^2_{rc}, obtemos o valor igual a 16, com $P-valor = P\left(\chi^2_{1:(0.95)} > 16\right) = 0.0000633$ o que nos leva a rejeição da hipótese de independência entre as variáveis. Quando multiplicamos o tamanho amostral por uma constante k, mantendo as mesmas proporções em cada célula da tabela de contingência, o valor esperado também fica multiplicado por uma constante k e a estatística de teste χ^2_{rc} também fica multiplicada por uma constante k (Goodman (1964)), quer dizer:

$$\begin{aligned}(\chi^2_{rc})' &= \sum \frac{(kn_{ij} - kE_{ij})^2}{kE_{ij}} = \sum \frac{k^2(n_{ij} - E_{ij})^2}{kE_{ij}} = \\ &= k\sum \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = k * \chi^2_{rc}\end{aligned}$$

A situação acima acontece porque o teste qui-quadrado de independência é um método global que se limita a testar se existe ou não associação entre variáveis, sem identificar a magnitude desta possível associação.

Por estas razões põe-se a questão de medir o grau de associação entre as variáveis linha e coluna. Tem sido proposto um grande numero de diferentes coeficientes de associação. A razão dessa variedade reside nos diferentes aspectos de associação que medem. Em certas circunstâncias uns são mais apropriados que outros.

Porém a maioria destes coeficientes sofrem uma grande desvantagem: não permitem interpretações probabilísticas. Por essa razão Goodman e Kruskal propuseram medidas interpretáveis num sentido preditor.

2.5 Coeficiente de associação de Goodman e Kruskal

O tau de Goodman e Kruskal (notação τ_{GK}) é obtido usando o principio da redução proporcional dos erros na predição de uma das variáveis. São calculados dois erros de predição, um é calculado o erro ao se tentar alocar os elementos de uma variavel ao seu respectivo nível na ausência de informações sobre de outra variável, e o segundo tipo é calculado erro ao se tentar alocar os elementos de uma variavel ao seu respectivo nível mas dessa vez com o conhecimento prévio do valor da outra variável.

Isto é, este coeficiente tem por objetivo responder à questão: Em que medida o fato de conhecermos a classificação de uma das variáveis (seja ela linha ou coluna) nos torna mais hábeis para prevermos a classificação da outra variável ?

O método consiste em apagar a informação de que linha e de que coluna um elemento pertence, e, logo depois , tentar recolocar este elemento a sua respectiva linha seguindo duas regras:

A regra 1 consiste em tentar realocar o elemento a sua verdadeira linha usando apenas a informação do total marginal das linhas. Melhor dizendo, vamos supor que se o elemento pertence à linha i ele recebe uma cor. Suponha agora que a cor de todos os elementos foram apagadas e que a única informação de que dispomos para recolocar estes indivíduos a sua verdadeira linha seja o número total de indivíduos pertencentes a cada linha.

- tentaremos realocar os indivíduos
  sua respectiva linha

Tabela 2.4: Regra 1

Variável Linha	Variável Coluna			
	Nível 1	Nível 2	Nível 3	Total
Nível 1				$n_{1.}$
Nível 2				$n_{2.}$
Nível 3				$n_{3.}$
Total				N

É claro que ao tentarmos realocar os elementos a sua respectiva linha cometermos erros. Denotamos este primeiro conjunto de erros por A, sendo calculado pela seguinte fórmula:

$$A = N \sum_{i=1}^{n} {}_{l}p_{i.}(1 - p_{i.}),$$

em que N é o total de elementos na tabela de contingência, n_lé o total níveis da variável linha, e $p_{i.}$ é a probabilidade marginal da linha i.

A regra 2 consiste em usar a informação sobre os totais marginais da variável coluna para tentar realocar cada elemento a sua verdadeira linha. Melhor explicando, suponhamos novamente que se o elemento pertence à linha i ele recebe uma cor e que por algum motivo a cor de todos os elementos foram apagadas . A diferença é que agora para realocá-los à sua verdadeira linha ,temos além da informação do número total de indivíduos pertencentes a cada linha, temos também a informação do número total de indivíduos pertencentes a cada coluna.

- realocar os elementos
  a suas respectivas linhas

Tabela 2.5: Regra 1

Variável Linha	Variável Coluna			
	Nível 1	Nível 2	Nível 3	Total
Nível 1	n_{11}	n_{12}	n_{13}	$n_{1.}$
Nível 2	n_{21}	n_{22}	n_{23}	$n_{2.}$
Nível 3	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

- conhecendo o total de
  elementos na coluna

O número de erros obtidos por esta regra é denotado por B e é calculado da seguinte forma:

$$B = \sum_j n_{.j} \sum_i \frac{p_{ij}}{p_{j.}} \left(1 - \frac{p_{ij}}{p_{j.}}\right),$$

na qual $n_{.j}$ representa o total marginal da coluna j, p_{ij} é a probabilidade conjunta , e $p_{.j}$ é a probabilidade marginal da coluna j.

Então a estatística de Goodman e Kruskal é definida como:

$$\tau_{GK} = \frac{A - B}{A}$$

Esta estatística tem algumas propriedades desejáveis, tais como ser uma medida com limites zero (nenhuma associação) e um (completa associação) e não mudar com a permutação de linhas e colunas.

Além disso τ_{GK} tem uma interpretação muito clara: mede o decréscimo relativo na probabilidade de errar a previsão da variável linha ao conhecer a variável coluna (ou vice versa). Por exemplo, se $\tau_{GK} = 0.8$ uma redução de 80% na probabilidade de errar a previsão de uma das variáveis, quando se usa a informação sobre a outra variável.

Consideremos a seguinte tabela de contingência relativa à classificação hipotética de 284 doentes psiquiátricos segundo o diagnóstico da sua doença e o seu estado social. (Des. personal. significa Desordem personalidade)

Tabela 2.6: Doente psiquiátricos classificados segundo o diagnósticoda sua doença e o seu estrato social

Classe	Diagnóstico				
Social	Neurótico	Depressivo	Des. personal	Esquizofrênico	Total
1	45	25	21	18	109
2	10	45	24	22	101
3	17	21	18	18	74
Total	72	91	63	284	

Queremos saber como é que o conhecimento da classe social pode ajudar a predizer o diagnóstico. Se quisermos predizer qual o diagnóstico, sem termos conhecimento da classe social, iremos errar certo número de vezes. O número esperado de erros ao tomarmos uma decisão errada é calculado por:

$$A = N \sum_i p_{i.}(1 - p_{i.}) = 284 * (0.24 + 0.23 + 0.19) = 186.96$$

Se tivermos a informação de que classe social o doente pertence, podemos agora calcular o numero esperado de erros de previsão do diagnóstico, quando tenho conhecimento da classe social:

$$B = \sum_j n_{.j} \sum_i \frac{p_{ij}}{p_{j.}} \left(1 - \frac{p_{ij}}{p_{j.}}\right) = 175.7$$

Assim o conhecimento da classe social do doente fez diminuir a probabilidade de errar a previsão do diagnóstico.

O coeficiente de Godman-Kruskal

$$\tau_{GK} = \frac{A - B}{A} = \frac{186.96 - 175.7}{186.96} = 0.0602$$

ou seja temos uma redução de 10% na probabilidade de errar, quando se conhece a classe social.

Capítulo 3

Medida de Associção entre Espaço Tempo

3.1 Motivação

Encontramos na literatura vários aplicações do teste de Knox em que o teste sugeriu que existiam conglomerados espaço-temporais, porém estes conglomerados tinham pouca significância prática.

Por exemplo uma aplicação do teste knox foi feita aos dados de arrombamento a residências na cidade de Belo Horizonte no período de 2003 a 2005 feita pelo *Centro de Criminalidade e Segurança Pública* da UFMG feita por Gneiting e A.E. (2004) o teste foi significativo. No entanto, existem conglomerados fracos, ou seja, existem apenas poucos conglomerados com um número pequeno de eventos dentro de cada um deles.

Os mapas exibidos na Figura 3.1 mostram os casos de roubos à residência em Belo Horizonte para os anos de 2004 e 2005. Estes mapas ilustram o problema descrito acima. As regiões em vermelho no mapa indicam conglomerados em que houveram grande número de roubos à residência em um curto espaço de tempo e distância. Podemos perceber o que é predominante no mapa são conglomerados com apenas um caso, representados pelos pontos na cor preta, e que existem poucos conglomerados com um número grande de casos.

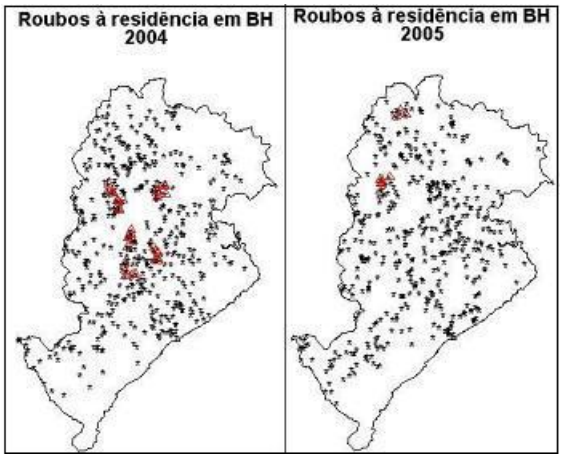

Figura 3.1: Mapas dos roubos em residências em BH em 2004 e 2005, com a localização, em vermelho, dos grupos de casos com uma maior interação espaço-temporal.

Esta situação ocorre porque o teste de knox, assim como os outros métodos utilizados para testar a hipótese de interação espaço-tempo, se limitam a verificar a existência ou não de uma interação espaço-tempo, sem identificar a magnitude desta interação.

É importante fazer a distinção entre significância estatística e importância prática. A significância estatística simplesmente significa que nós rejeitamos a hipótese nula. A capacidade do teste para detectar diferenças que levam à rejeição da hipótese nula depende do tamanho da amostra. Por exemplo, para uma grande amostra em particular, o teste pode rejeitar a hipótese nula de que duas variáveis são independentes. No entanto, na prática, a dependência entre as duas variáveis pode ser relativamente pequena a ponto de não ter nenhum significado real. Da mesma forma, se o tamanho da amostra é pequeno, uma diferença que é grande em termos práticos pode não levar à rejeição da hipótese nula de independência entre as duas variáveis.

Além do problema descrito acima, sabemos que o tipo de relação que pode existir entre espaço e tempo não é monótona. Isto quer dizer que não poderemos ter uma interpretação direta sobre a associação como temos como coeficiente de correlação de Pearson, por exemplo.

O que buscamos obter é uma medida de associação que consiga captar relacionamentos não monótonos como na Figura 3.2 e, que tenha boas bases teóricas.

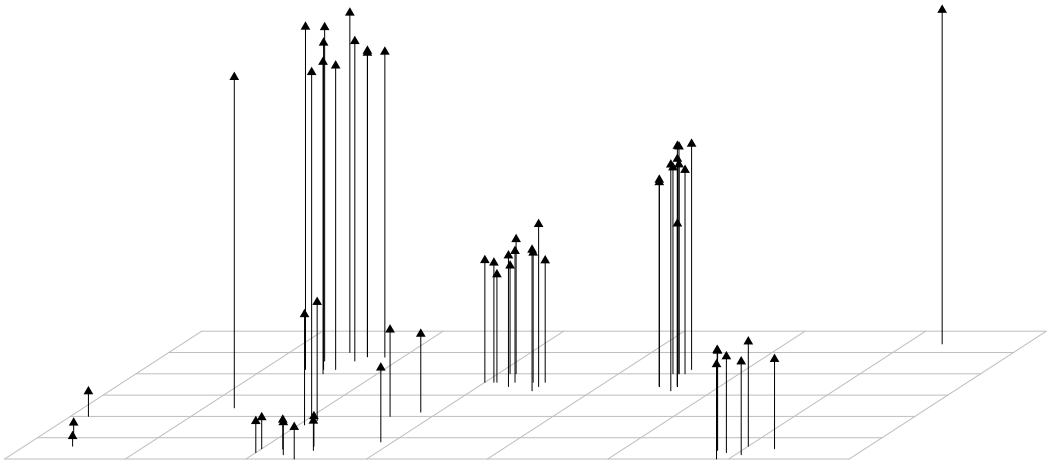

Figura 3.2: Gráfico de coordenadas espaciais e temporais em R^3

Nesta dissertação de mestrado usamos as idéias de Goodman e Kruskal para construir uma medida de associação entre espaço e tempo que permita dizer a magnitude de uma possível associação espaço-tempo que não seja afetada pelos problemas descritos acima.

3.2 Medida de Associação

Inicialmente iremos mostrar como a construção da medida de associação para duas variáveis aleatórias contínuas X e Y quaisquer e, logo depois, estender para as variáveis espaço tempo.

Gráficos de Dispersão são comumente usados para exibir e comparar valores numéricos, como dados científicos, estatísticos e de engenharia. Gráficos de Dispersão têm dois eixos de valores, mostrando um conjunto de dados numéricos ao longo do eixo horizontal e outro ao longo do eixo vertical.

Seja o seguinte o seguinte diagrama de dispersão:

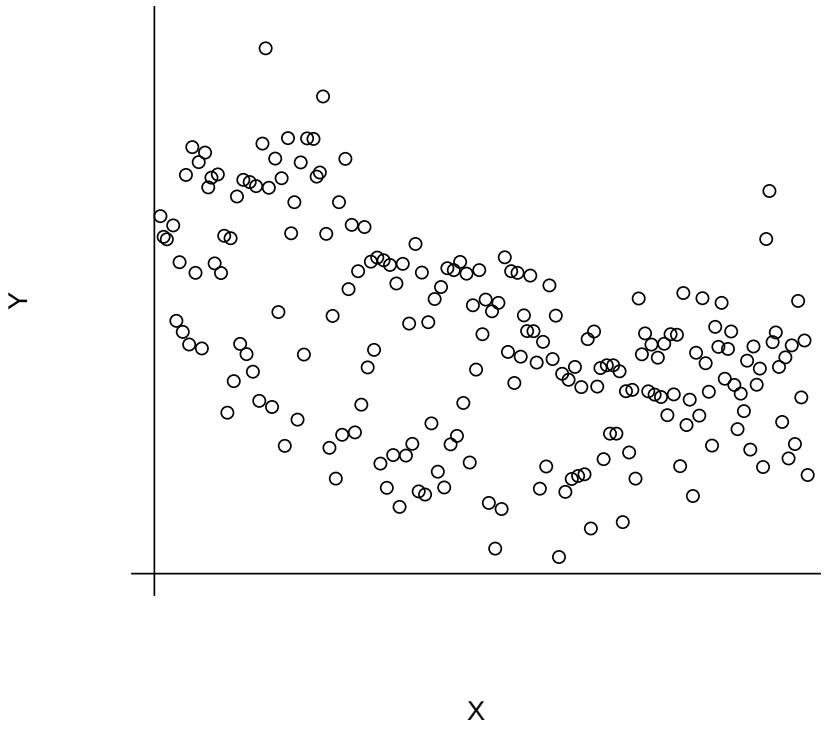

Figura 3.3: Gráfico de pontos com coordenadas em R^2

Suponha que iremos traçar uma grade, com quadrados de lado de lados Δ_i e Δ_j. Observe que agora temos uma estrutura semelhante a uma tabela de contingência, quer dizer cada celular seria correspondente a um quadrado de lados Δ_i e Δ_j e os elementos desta célula seria o número de pontos dentro deste quadrado.

Ao encararmos um gráfico de dispersão em R^2 como uma tabela de contingência iremos adaptar a medida de Goodman e Kruskal para dados contínuos. Esta medida usa duas quantidades em sua construção:

A primeira é **A** e é a taxa de acertos na predição de cada valor da variável X (que são as coordenadas no espaço R^2) . Ou seja, ao invés de procurar obter a taxa esperada de erros, como fizeram Goodman e Kruskal, calcularemos a taxa esperada de acertos na predição de cada valor da variável X.

Taxa de Classificações corretas na classe de tamanho Δ:

$$\frac{1}{\Delta}P\left(\begin{array}{c}\text{Classificar}\\ \text{corretamente}\\ \text{em classe de}\\ \text{tamanho } \Delta\end{array}\right) = \frac{1}{\Delta}\sum_{\text{Classe } i} P\left(\begin{array}{cc}\text{Classificar} & \text{e pertencer}\\ \text{corretamente} & \text{classe } i\end{array}\right)$$

$$= \frac{1}{\Delta} \sum_i P(\in \text{ classe } i) \cdot \ P \left(\begin{array}{c} \text{classificar} \\ \text{corretamente} \end{array} \middle| \in \text{classe i} \right)$$

$$= \frac{1}{\Delta} \sum_i P\left(\in \text{Classe } i\right) \cdot \ P \left(\begin{array}{c} \text{classificar} \\ \text{na classe } i \end{array} \middle| \in \text{classe } i \right)$$

$$= \frac{1}{\Delta} \sum_i P\left(\in \text{Classe } i\right) \cdot \ P \left(\begin{array}{c} \text{classificar} \\ \text{na classe } i \end{array} \right)$$

$$= \frac{1}{\Delta} \sum_i f(x_i) \cdot \ \Delta f(x_i)\Delta + \frac{1}{\Delta} O(\Delta^2)$$

$$\sum_i f^2(x_i)\Delta + O(\Delta)$$

$$\longrightarrow \int f^2(x)dx = \int f(x)f(x)dx =$$

$$= E\left[f(X)\right] = A$$

Ou seja, A é o valor esperado da densidade de X, em um ponto X que segue a distribuição f.

Para a quantidade **B** é usada a informação sobre a variável tempo para predizer os valores, em cada quadrado de tamanho Δ, da variável espacial. O valor esperado da taxa de acertos da variável espacial X dado o valor da variável tempo é o que chamaremos de B.

Taxa de Classificações corretas na classe de tamanho Δ dado que tempo é $t \in (t_j, t_{j+\Delta_j})$=

$$\frac{1}{\Delta} \sum_{t_j} P \left(\begin{array}{c} \text{Classificar} \\ \text{corretamente} \\ \text{em classe de} \\ \text{tamanho } \Delta \end{array} \middle| \in \text{ tempo } t_j \right) P\left(\in \text{tempo } t_j\right) =$$

$$= \frac{1}{\Delta} \sum_{\text{Classe } i} \sum_{\text{Classe } j} P \left(\begin{array}{ccc} \text{Classificar} & \text{e} & \text{pertencer} \\ \text{corretamente} & & \text{classe } i \end{array} \middle| \in \text{classe } j \right) P\left(\in classej\right)$$

$$= \frac{1}{\Delta} \sum_j P(\in \text{ classe } j) \sum_i P(\in \text{ classe } i) \cdot \ P \left(\begin{array}{c} \text{classificar} \\ \text{corretamente} \end{array} \middle| \in \text{classe i e } \in \text{ classe j} \right)$$

$$= \frac{1}{\Delta} \sum_j P\left(\in \text{Classe } j\right) \sum_i P\left(\in \text{Classe } i\right) \cdot \ P \left(\begin{array}{c} \text{classificar} \\ \text{na classe } i \end{array} \middle| \in \text{classe i e } \in \text{classe j} \right)$$

$$= \frac{1}{\Delta} \sum_j P\left(\in \text{Classe } j\right) \sum_i P\left(\in \text{Classe } i\right) \cdot \ P \left(\begin{array}{c} \text{classificar} \\ \text{na classe } i \end{array} \middle| \in \text{classe j} \right)$$

$$= \frac{1}{\Delta_i} \sum_j f(t_j)\Delta_j \sum_i f(x_i|t_j)\Delta_i f(x_i|t_j)\Delta_i + \frac{1}{\Delta_i} O(\Delta_i^2) + \frac{1}{\Delta_j} O(\Delta_j^2) =$$

$$= \sum_j f(t_j)\Delta_j \left[\sum_i f^2 \left(x_i | T \in t_j \pm \frac{\Delta_j}{2} \right) \Delta_i + O(\Delta_i) \right] + O(\Delta_j) =$$

$$\longrightarrow \sum_j f(t_j)\Delta_j \cdot \int f^2 \left(X | T \in t_j \pm \frac{\Delta_j}{2} \right) dx + O(\Delta_j) =$$

$$= \sum_j f(t_j)\Delta_j \cdot E \left(f\left(X|T \in t_j\right) | T \in t_j \pm \frac{\Delta_j}{2} \right) + O(\Delta_j)$$

$$\longrightarrow \int E_{X|T=t} \left(f(X|T=t) \right) f(t) dt = E_T \left[E_{X|T=t} \left(f(X|T=t) \right) \right] =$$

$$= E_T \left[E_{X|T=t} \left(f(X|T=t) \right) \right] = B$$

Onde E_T é o valor esperado com relação à variável tempo; $E_{X|T=t}$ é o valor esperado da variável espaço dado o tempo t; $f(X|T=t)$ é a densidade condicional da variável espaço dado o tempo t.

A nossa medida de associação se propõe a responder a seguinte pergunta: "Em que proporção a informação de uma das variáveis nos ajuda a acertar a predição de cada valor da outra variável?"Sendo assim definimos a nossa medida de associação como sendo a razão:

$$\Psi = \frac{B - A}{B} \tag{3.1}$$

Além da interpretação de aumento da capacidade preditiva, esta medida de associação possui três propriedades muito claras:

1. Os valores do índice estão entre 0 e 1.

2. Independência:
   Se as variáveis são independentes, o índice é zero.

3. Coerência:
   O índice aumenta com o aumento da dependência, sendo igual 1, quando uma variável é totalmente dependente da outra.

3.3 Exemplos

Exemplo 1 Caso Independente

Seja:

$$f(x,t) = \begin{cases} e^{-(x+t)} & 0 \leq x \leq \infty \quad 0 \leq t \leq \infty \\ 0 & cc \end{cases}$$

$$f_X(x) = e^{-x} \quad 0 \leq x \leq \infty$$

$$f_T(x) = e^{-t} \quad 0 \leq t \leq \infty$$

$$f_{X|Y}(x|y) = e^{-x}$$

$$A = E_X[f_X(X)] = \int_0^\infty e^{-2x}dx = \left.\frac{-e^{-2x}}{2}\right|_0^\infty = \frac{1}{2}$$

$$E_{X|T}(f_{X|Y}(X|Y)) = g(y) = \int_0^\infty e^{-2x} = \frac{1}{2}$$

$$E_Y[g(Y)] = \int_0^\infty \frac{1}{2}e^{-y}dy = \frac{1}{2} = B$$

$$GK = \frac{B-A}{B} = 0$$

Exemplo 2: A Distribuição Normal Bivariada

$$f(x,t) = \begin{cases} \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot exp\left\{\frac{-1}{2(1-\rho^2)}\left[x^2 - 2\rho(x\cdot t) + t^2\right]\right\}, \\ , -\infty < x < \infty, \ -\infty < t < \infty, \ -1 \leq \rho \leq 1 \\ 0 \qquad c.c \end{cases}$$

onde $X \sim N(0,1)$ e $T \sim N(0,1)$.

$$f(X|T) = \frac{1}{\sqrt{2\pi(1-\rho^2)}} exp\left(\frac{-1}{2}\frac{(x-\rho t)^2}{(1-\rho^2)}\right)$$

$X \sim N(\mu_X, \sigma_X^2)$ e $T \sim N(\mu_T, \sigma_T^2)$

$$f(x,t) = \frac{1}{2\pi\sigma_X\sigma_T\sqrt{1-\rho^2}} \cdot exp\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - \frac{2\rho(x-\mu_X)(t-\mu_T)}{\sigma_X\sigma_T} + + \left(\frac{t-\mu_T}{\sigma_T}\right)^2\right\}$$

$$A = \frac{1}{2\sigma_X\sqrt{\pi}}$$

$$B = \frac{1}{2\sigma_X\sqrt{\pi(1-\rho^2)}}$$

$$\begin{aligned} GK &= \frac{B-A}{B} = \frac{1-\sqrt{1-\rho^2}}{2\sigma_X\sqrt{\pi}\sqrt{1-\rho^2}} \cdot 2\sigma_X\sqrt{\pi(1-\rho^2)} = \\ &= 1 - \sqrt{1-\rho^2} \end{aligned}$$

Gráfico do Comportamento de GK de acordo com ρ

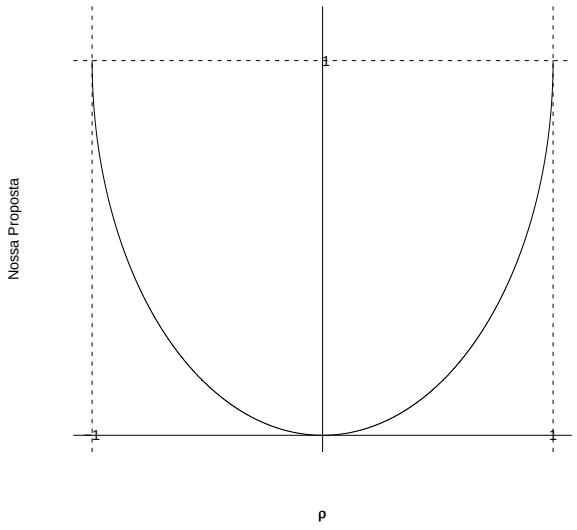

Figura 3.4: Gráfico do Comportamento de GK de acordo com ρ:

Exemplo 3 Anel

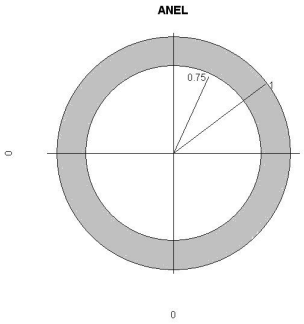

Figura 3.5: Exemplo Anel

Seja R, a região limitada pelos círculos$x^2 + y^2 = 3/4$ e $x^2 + y^2 = 1$.
R pode ser escrita como: $\{(x; y)|0, 75 \leq x^2 + y^2 \leq 1\}$.

$$GK = \frac{(1, 455131 - 1, 188801)}{1, 455131} = 0, 1830282$$

Capítulo 4

Método de Estimação por Kernel (Núcleo Estimador)

A medida de associação proposta nesta dissertação depende de densidades de probailidade. Sabemos que ao fazermos inferência de um modelo específico é possível obter um ganho muito grande em eficiência, mas somente se o modelo de probabilidade assumido for pelo menos aproximadamente verdadeiro.

Se o modelo de probabilidade assumido não estiver correto, as inferências podem ser piores e inúteis, levando a enganos grosseiros na interpretação dos dados. Os métodos de suavização oferecem uma ponte entre não estabelecer nenhuma hipótese na estrutura formal (abordagem puramente não-paramétrica) e estabelecer hipóteses muito fortes (abordagem paramétrica). Ao adotar uma hipótese relativamente fraca de que a verdadeira densidade é suave permite que os dados contem ao analista qual é seu verdadeiro padrão.

Usando o método de Estimação de Densidades via núcleo estimador 4.0.1 e a igualdade que será vista no capítulo 4.1 fornecemos um método bootstrap para a estimação de nossa medida 4.1.1

4.0.1 Estimação de Densidades de Probabilidade via Nucleo Estimador

Um Nucleo Estimador (Simonoff (1996)) é uma função ponderada usada em técnicas de estimação não paramétrica. Núcleo estimadores são usados na estimação de densidades, na estimação de funções de intesidade e em funções de regressão.

O núcleo estimador também é usado em séries temporais, no uso do periodograma para estimar a densidade espectral.Uma aplicação adiconal está na estatística espacial, onde a nossa função é usada para estimação de processos pontuais variando com o tempo.

O nucleo é uma função K não negativa e integravel que satisfaz as duas suposições seguintes :

$$\int_{-\infty}^{\infty} K(u)du = 1,$$

$$K(-u) = K(u) \quad para \quad todos \quad os \quad valores \quad de \quad u.$$

A primeira suposição diz que a densidade do nucleo estimador resulta em uma função densidade de probabilidade.

Alguns tipos de Funções Nucleo são comomumente usadas: Uniforme, Epanechnikov e Gaussiana.

Uniforme:

$$K(u) = \frac{1}{2} \quad I_{|u|<1},$$

Epanechnikov:

$$K(u) = \frac{3}{4}(1-u^2) \quad I_{|u|<1},$$

Gaussiano

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}u^2}.$$

4.0.2 Função de Densidade

Nesta seção, faremos uma breve abordagem sobre o núcleo-estimador para função de densidade, a qual será utilizada nos estimadores de regressão não paramétricos. Então, dada uma amostra aleatória $(x_1, x_2, \ldots, x_n)$, a estimação de uma função de densidade, avaliada no ponto x, é definida por:

$$\hat{f}(x) = \frac{1}{nh} = \sum_{i=1}^{n} \left(\frac{x - x_i}{h} \right),$$

onde K é a função núcleo e h é o parâmetro de suavização que é conhecido como janela Fan (1995).

4.1 Uma Igualdade para A

Sejam X_1 e X_2 variáveis aleatórias independentes e identicamente distribuídas com função de distribuição acumulada F_x. Seja $Z = X_1 - X_2$. A função densidade da variável Z é dada por:

$$f_Z(z) = \int f_{X_1,X_2}(x_1, x_1 - z)|J(x_1, x_2)|dx_1$$

sabemos que o determinante do Jacobiano da transformação $Z = X_1 - X_2$ é dado por:

$$J(x_1, x_2) = \begin{vmatrix} \frac{\partial x_1}{\partial x_1} & \frac{\partial x_1}{\partial z} \\ \frac{\partial x_1 - z}{\partial x_1} & \frac{\partial x_1 - z}{\partial z} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 1 & -1 \end{vmatrix} = -1$$

portanto a função densidade de Z é dada por:

$$f_Z(z) = \int f_{X_1,X_2}(x_1, x_1 - z)|J(x_1, x_2)|dx_1 = \int f_{X_1,X_2}(x_1, x_1 - z)dx_1$$

Como X_1 e X_2 são independentes e identicamente distribuídos temos que:

$$f_Z(z) = \int f_X(x_1)f_X(x_1 - z)dx_1$$

Avaliando a densidade da variável aleatória Z em $z = 0$ temos :

$$\begin{aligned} f_Z(0) &= \int f_X(x_1)f_X(x_1)dx_1 \\ &= \int f_X(x)f_X(x)dx \\ &= \int f_X^2(x)dx \\ &= E[f(X)] \end{aligned}$$

Ou seja, $A = E\left[f(X)\right]$ pode ser visto como $f_Z(0)$, onde $Z = X_1 - X_2$ e X_1 e X_2 são independentes e identicamente distribuídos ($X_1, X_2 \sim F_X$).

4.1.1 Procedimento bootstrap

Baseado na igualdade $E[f(X)] = f_Z(0)$ e usando o método de estimação de densidade via nucleo estimador, temos o seguinte procedimento bootstrap (Efron (1982)):
Passo 1: obtenção de A:

1. Serão selecionadas duas amostras de forma independente, com reposição, cada uma com tamanho m (suficientemente grande) da variável aleatória X
2. posteriormente será construída um vetor composta pelas diferenças entre cada um dos elementos da primeira amostra com os elementos da segunda amostra.
3. Estimaremos a densidade deste vetor usando o método de Kernel para estimativas de densidade.
4. Avaliaremos esta densidade estimada no ponto zero, obtendo assim, uma estimativa para A.

Passo 2 : Estimação de B:

1. Inicialmente iremos dividir os possíveis valores de Y em k intervalos.
2. Para cada um destes k intervalos faça:

(a) selecionamos duas amostras e calcularemos as diferenças entre os valores das duas amostras conforme o Passo 1 .

(b) Estimaremos a densidade deste vetor usando o método de Kernel .

(c) Avaliaremos esta densidade estimada no ponto zero.

(d) Ainda neste intervalo, multiplicaremos o valor obtido da densidade do vetor de diferenças no ponto zero pelo valor da probabilidade de selecionarmos um valor de Y neste intervalo considerado.

3. Após repetir este procedimento para os k intervalos teremos que o valor de B,que é a soma dos k valores obtidos por o procedimento mencionado acima

Passo 3 : Calcularemos a nossa medida:

$$\frac{B-A}{B}$$

Nos investigamos a aproximação do calculo de nossa medida pelo bootstrap citado acima. Geramos 20 pares X e Y de uma normal bivariada com os seguintes parâmetros:
$\mu_X = \mu_Y = 0$, $\sigma_X^2 = \sigma_Y^2 = 5$ e $\rho = 0.5$

Nesta investigação reamostramos 200 vezes, a distribuição das quantidades A, B e da nossa medida. Na figura 4.1, temos o gráfico de uma normal bivariada em uma destas gerações:

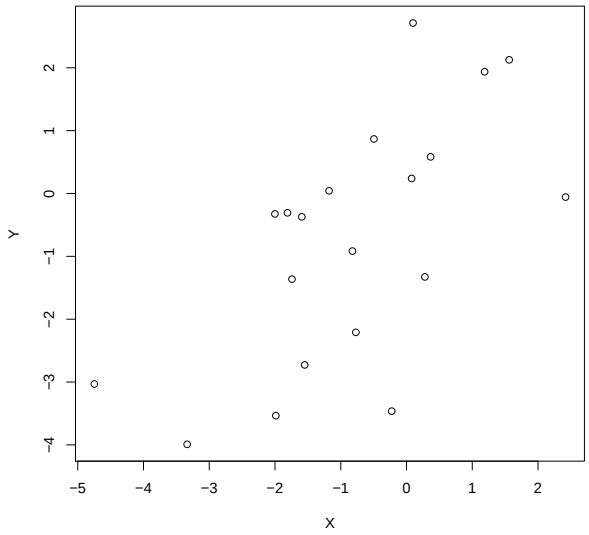

Figura 4.1: Gráfico dos 20 pontos gerados de uma normal bivariada com os parâmetros:$\mu_X = \mu_Y = 0$, $\sigma_X^2 = \sigma_Y^2 = 5$ e $\rho = 0.5$.

Para uma normal bivariada com os parâmetros do item anterior temos que a nossa medida assume o valor:

$$\begin{aligned}\Psi = GK = \frac{B-A}{B} &= 1-\sqrt{1-\rho^2} \\ &= 1-\sqrt{1-0.5^2} \\ &= 0.134\end{aligned}$$

Sendo

$$A = \frac{1}{2\sigma_X\sqrt{\pi}} = \frac{1}{2\sqrt{5\pi}} = 0.126, \text{ e:}$$

$$B = \frac{1}{2\sigma_X\sqrt{\pi(1-\rho^2)}} = \frac{1}{2\sqrt{5\pi(1-0.5^2)}} = 0.146$$

Os resultados da simulação estão sintetizados na figura 4.2.Observando esta figura podemos dizer que o procedimento bootstrap nos fornece uma boa ferramenta de estimação para a nossa medida de associação. Isto porque é perceptível, no histograma à esquerda, que as medidas obtidas via simulação bootstrap ficaram em torno do valor calculado no exemplo(0.134).

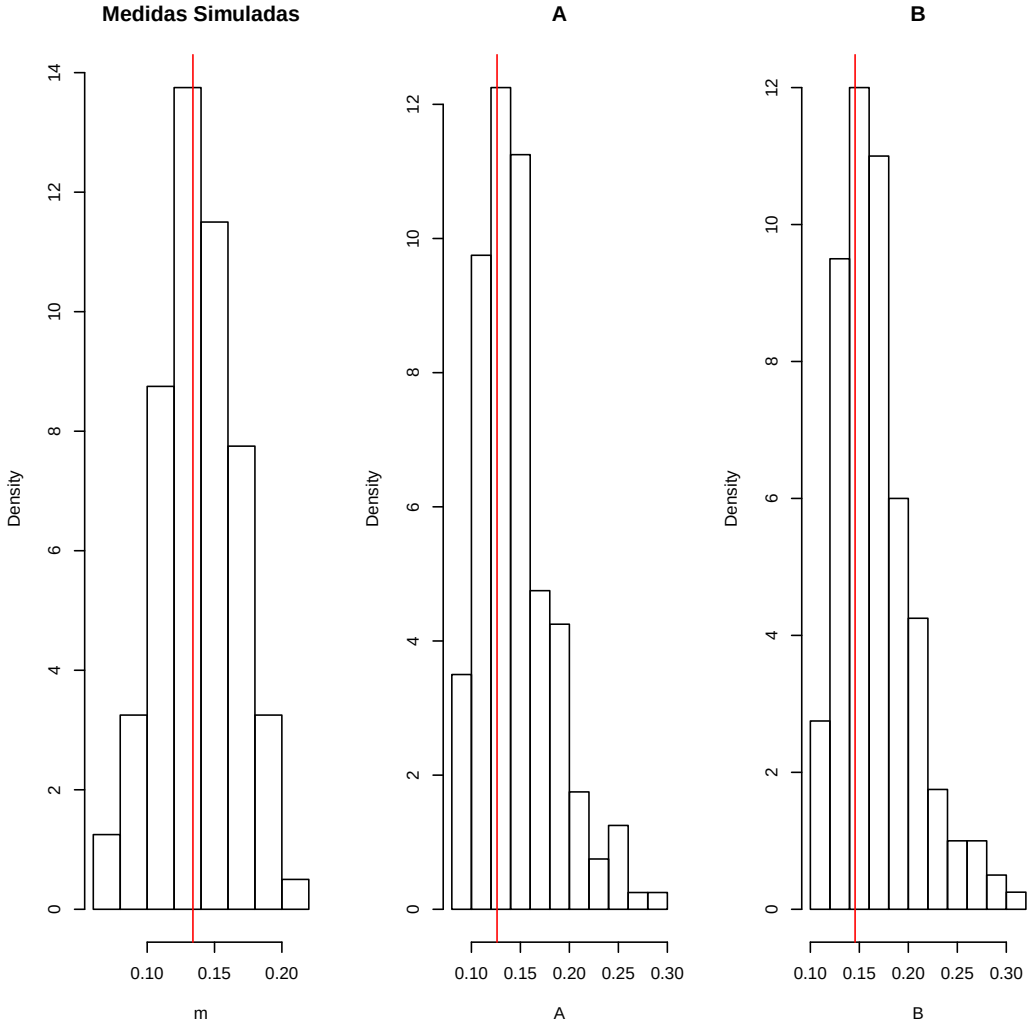

Figura 4.2: Gráfico das medidas de associação simuladas com o procedimento bootstrap, nos quais a linha vermelha representa a quantidade real: Gráfico à esquerda: medidas de associação obtidas via bootstrap, gráfico do centro: medidas obtidas com o bootstrap da quantidade A e o gráfico à direita são as medidas da quantidade B.

O desempenho do procedimento bootstrap foi investigado usando dois exemplos simulados. Nos dois processos simulados utilizamos como domínio o quadrado unitário, ou seja, o conjunto $\{(x, y) : (x, y) \in [0, 1] \times [0, 1]\}$.

No primeiro exemplo foram gerados 100 pontos uniformemente em $[0, 1] \times [0, 1]$, que pode ser visto na Figura 4.3:

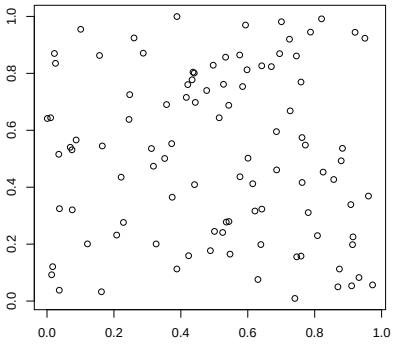

Figura 4.3: Exemplo 1: 100 pontos gerados uniformemente em $[0,1] \times [0,1]$

Na Figura 4.4 podemos observar que o procedimento resultou em valores pouco elevados da nossa medida de associação, isto significa que a provável relação entre as variáveis é fraca, o que já era esperado pois os pontos foram gerados de acordo com um distribuição uniforme no quadrado $[0,1] \times [0,1]$.

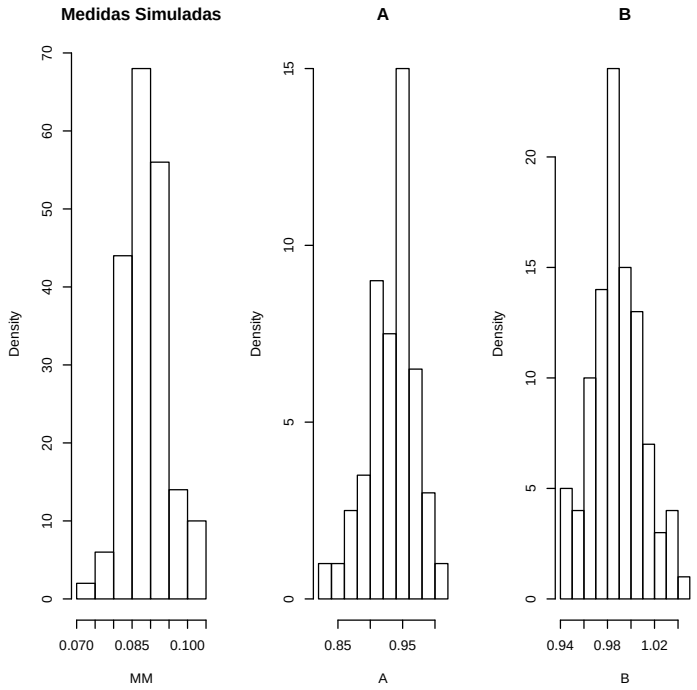

Figura 4.4: Medidas de associação simuladas para o Exemplo 1.

A Figura 4.5 apresenta o segundo exemplo. Neste exemplo foram gerados 100 pontos uniformemente em $[0,1]\times[0,1]$ e 100 pontos distribuídos uniformente no quadrado $[0.5,0.75]\times[0.5,0.75]$.

De acordo com a Figura 4.6, temos que a nossa medida apresenta valores entre 0.31 e 0.50, isto significa que a presença de um cluster nos dados faz com que o nossa medida tenha um aumento substancial.

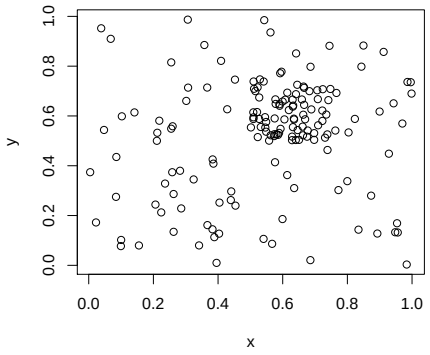

Figura 4.5: Exemplo 2: 100 pontos gerados uniformemente em $[0,1] \times [0,1]$ e 50 pontos gerados uniformemente em $[0.5, 0.75] \times [0.5, 0.75]$.

Considerando as variáveis espaciais X e temporais T, temos o seguinte procedimento bootstrap:

Passo 1: obtenção de A:

1. Como no bootstrap anterior serão selecionadas duas amostras $X_1 e X_2$ de forma independente, com reposição da posições espaciais X , cada uma com tamanho suficientemente grande.

2. costruiremos a variavel Z, composta pelas diferenças entre cada um dos elementos da primeira amostra com cada um dos elementos da segunda amostra ($Z = X_1 - X_2$).

3. Estimaremos a densidade da variável Z usando o método de Kernel para estimativas de densidade bivariadas .

4. Avaliaremos esta densidade estimada no ponto $(0, 0)$ obtendo assim uma estimativa para A.

Passo 2 : Estimação de B:

1. dividir o intervalo de variação de T em k sub-intervalos.

2. Para cada um destes k sub-intervalos faça:

   (a) selecionamos duas amostras de valores da variável X que se encontram no intervalo $j \quad (j = 1, \ldots, k)$,$(X_{1j} \quad e \quad X_{2j})$ e calcularemos a variável Z, composta pelas diferenças entre as amostras de valores da variável X neste intervalo j,$(Z = X_{1j} - X_{2j})$, conforme o Passo 1 .

   (b) Estimaremos a densidade da variável Z usando o método de Kernel para densidades bidimensionais .

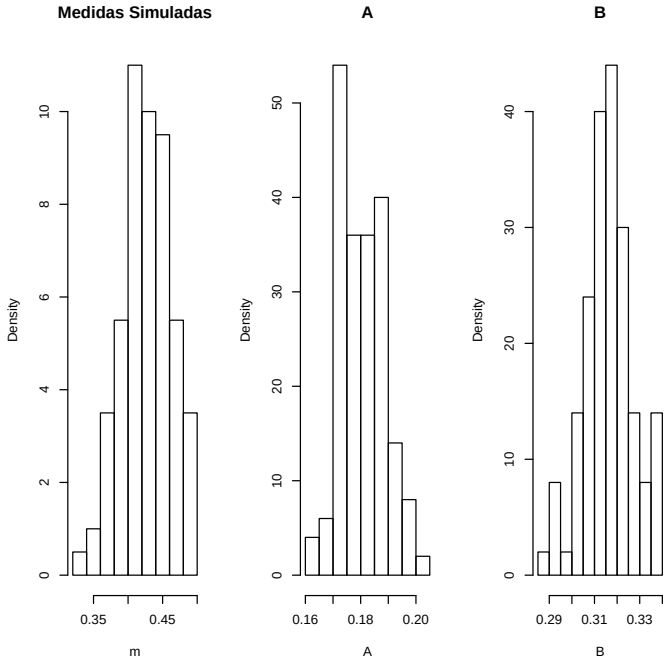

Figura 4.6: Medidas de associação simuladas para o Exemplo 2.

(c) Avaliaremos esta densidade estimada no ponto $(0, 0)$.

(d) Ainda neste intervalo, multiplicaremos o valor obtido da densidade do vetor de diferenças no ponto $(0, 0)$ pelo valor da probabilidade de selecionarmos um valor de T neste intervalo considerado, que é obtida pelo quociente entre o número de pontos no intervalo de tempo j e o número total de pontos.

3. A soma dos k valores obtidos por este procedimento nos dará uma estimativa para B

Passo 3 : Calcularemos a nossa medida:

$$\frac{B - A}{B}$$

O método bootstrap proposto acima foi testado via simulação .O procedimento foi testado em um cenário de completa aleatoriedade (sem conglomerado) e em um cenário onde existe um conglomerado em forma de paralelepípedo. O cenário sem conglomerado e os cenários com conglomerado são descritos nas seções 4.2 e 4.3, respectivamente.

4.2 Sem Conglomerado

Nesta simulação considerou-se a região em forma de um paralelepípedo $10 \times 10 \times 100$ e foram gerados sempre 1000 eventos distribuídos uniformemente nesta região.

A Tabela 4.1 o teste de Knox para os dados simulados . Para cada par de distâncias crítica e tempo crítico considerados há três valores indicados. O primeiro é o índice de Knox

observado, o segundo é o índice de Knox esperado e o terceiro é o p-valor associado ao teste de Knox.

Tabela 4.1: Índice de Knox observado, esperado e p-valor para o teste de Knox realizado com as diferentes distâncias e tempos críticos.

Distancia	Tempo			
	10	25	50	100
0.5	35	91	182	332
	35.64	88.26	172.25	324.21
	0.40	**0.39**	**0.38**	**0.39**
1	135	342	691	1276
	140.09	346.88	677	1274.211
	0.35	**0.38**	**0.31**	**0.32**
2	486	1233	2460	4586
	505.56	1251.81	2443.13	4598.30
	0.33	**0.31**	**0.32**	**0.33**

Podemos observar que em todos os cenários o teste de Knox não foi significativo, ou seja, a hipótese de independência entre as variáveis espaço tempo não foi rejeitada. De acordo com a Figura 4.7 a nossa medida de associação apresenta valores baixos, o que pode significar que os dados apresentam pouca chances de apresentar conglomerados.

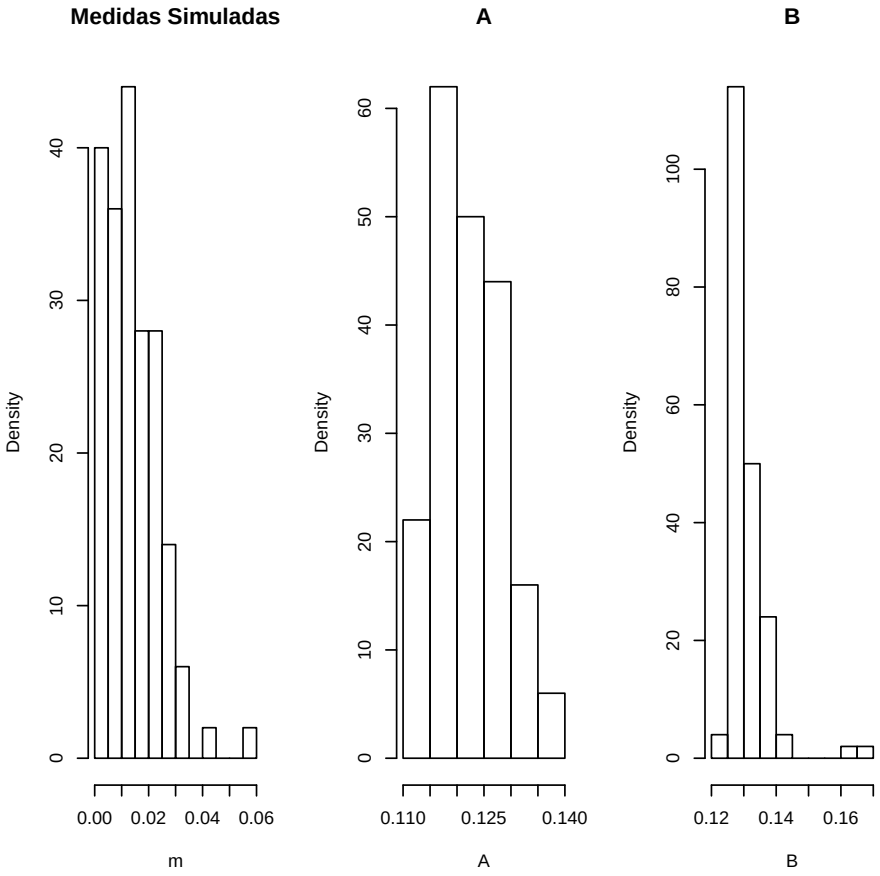

Figura 4.7: Medidas de Associação para os dados simulados uniformemente no paralelepípedo $10 \times 10 \times 100$

4.3 Com Conglomerado

Foram feitas 100 simulações independentes e foram gerados 500 eventos no paralelepípedo $10 \times 10 \times 500$ distribuídos uniformemente. Em seguida foi incluído o conglomerado, também em forma de paralelepípedo $10 \times 10 \times 250$

Para estes dados simulados, aplicamos o teste de Knox obtendo-se a tabela 4.2.

Tabela 4.2: Índice de Knox observado, esperado e p-valor para o teste de Knox realizado com as diferentes distâncias e tempos críticos.

Distancia	Tempo			
	10	25	50	100
0.5	195	259	375	608
	99.22	178.12	305.25	562.88
	0.002	**0.012**	**0.002**	**0.001**
1	698	993	1435	2288
	378.97	678.66	1164.85	2147.80
	0.002	**0.001**	**0.001**	**0.001**
2	2434	3420	5134	8336
	1364.84	2444.11	4195.09	7735.08
	0.001	**0.001**	**0.001**	**0.001**

O teste de Knox foi significativo para todas as configurações apresentadas na tabela 4.2. A Figura 4.8 apresenta valores entre 0.33 e 0.50 evidenciando que a nossa medida consegue captar bem a presença de um cluster com muitos eventos.

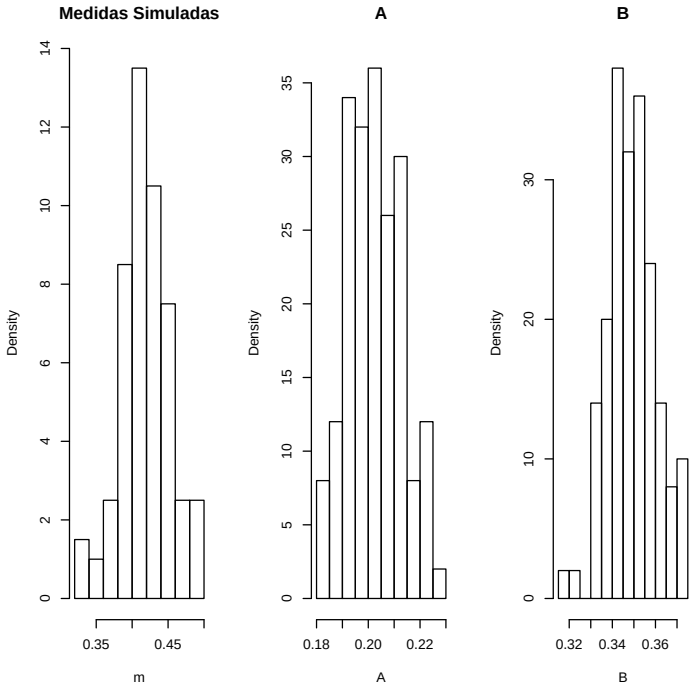

Figura 4.8: Medidas de Associação para os dados simulados com presença de cluster

Capítulo 5

Aplicação aos Dados de Arrombamento

Nesta seção nós focamos na análise espaço-temporal dos arrombamentos a residências em Belo Horizonte. Os dados consistem dos tempos de ocorrência e das coordenadas dos locais onde houve roubos à residência em Belo Horizonte no período de Janeiro de 1995 a Dezembro de 2005. Ao todo foram registrados 2688 casos nesse período considerado. Os dados ao longo dos anos foram os seguintes: 89, 82, 87, 137, 119, 248, 180, 260, 463, 508, 515.

Já não é novidade reconhecer que a ocorrência de crimes é espacialmente organizada. Nem toda área da cidade é igualmente sujeita a incidência de crimes. Este padrão depende do tipo de crime, sendo bastante diferente para os crimes contra o patrimônio e os crimes contra pessoas, por exemplo.

Para tais objetivos, aplicamos o teste de Knox. Considerando os roubos à residência de todos os anos conjuntamente obteve-se a tabela abaixo. Para cada par de distância crítica e tempo crítico considerados há três valores indicados. O primeiro é o índice de Knox observado, o segundo é o índice de Knox esperado e o terceiro é o p-valor associado ao teste de Knox.

Tabela 5.1: Índice de Knox observado, esperado e p-valor para o teste de Knox realizado com as diferentes distâncias e tempos críticos.

Distancia	Tempo			
	15	30	45	60
500	281	450	659	840
	171.73	328.91	484.41	635.32
	0.001	**0.001**	**0.001**	**0.001**
1000	737	1329	1897	2458
	587.01	1124.30	1655.85	2175.11
	0.001	**0.001**	**0.001**	**0.001**
1500	1385	2541	3634	4793
	1208.9	2315.1	3410.1	4479.49
	0.001	**0.001**	**0.001**	**0.001**
2000	2207	4082	5901	7765
	2030.07	3888.18	5726.46	7522.25
	0.001	**0.002**	**0.015**	**0.003**

Pela tabela acima, podemos observar que, para todos os pares de distância e tempo considerados, há evidências de que há interação espaço-temporal no número de roubos à residência em BH.

Após a análise feita acima se repete o teste de Knox para cada ano e cada par de distância e tempo. A tabela abaixo mostra, para cada par de distâncias e tempo críticos, os anos em que os índices de Knox foram significativos a 5%.

Tabela 5.2: Anos para os quais o teste de Knox foram significativos a 5%, dados os pares de distância e tempo críticos.

Distancia	Tempo			
	15	30	45	60
500	2004	2005	2000	2005
	2005		2005	
1000	1999	1996	1999	1999
	2001	1999	2005	2005
	2004	2005		
	2005			
1500	2001	2005	2005	2001
	2004			2003
	2005			2005
2000	1996			2003
	2001			2005
	2005			

A tabela acima mostra que para o ano de 2005 o índice de Knox foi considerado significativo para praticamente todos os pares de distância crítica e tempo crítico. Quando é considerada

uma distância crítica de 1000m e um tempo crítico de 15 dias observa-se que o índice de Knox foi significativo para quatro anos: 1999, 2001, 2004 e 2005. O gráfico abaixo mostra os valores observados e esperados para o índice de Knox levando em conta esse par de distâncias e tempo para cada ano.

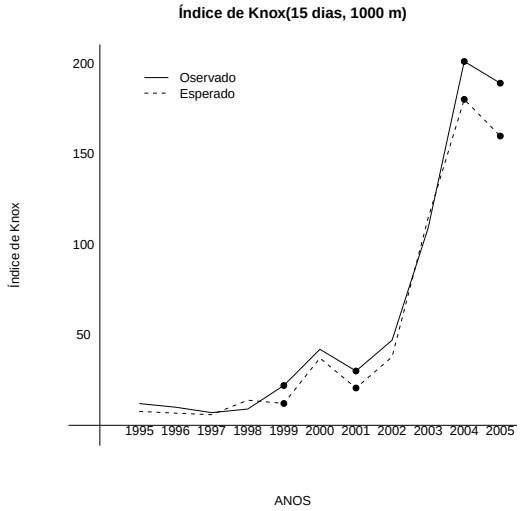

Figura 5.1: Gráfico dos valores observados e esperados do índice de Knox de 1995 a 2005.

Na figura **??** se observa que a partir de 2002 o índice de Knox apresenta um grande aumento. Em 2004 e 2005 podem-se ver as maiores diferenças entre os índices de Knox observado e esperado.

Considerando a distância crítica de 1000m e o tempo crítico de 15 dias para encontrar os vizinhos no espaço e tempo são construídos os mapas dos casos de roubos à residência em BH com indicação, em vermelho, de grupos de casos em que o número de vizinhos no espaço e no tempo simultaneamente são maiores ou iguais à média do número de vizinhos mais dois desvios-padrão. Esses locais são aqueles em que houve grande número de roubos à residência em um curto espaço de tempo e distância.

Considerando esta análise para os anos de 1995 a 2005, os grupos de casos com número alto de vizinhos coincide com as manchas mais intensas obtidas nos mapas de Kernel. Os anos em que se observa um alto número de grupos com casos em que há um valor grande para número de vizinhos são 2001, 2003 e 2004.

Porém para os anos de 1998 e 1999 o teste de knox para detecção de interação espaço-tempo foi significativo, no entanto, não foram encontrados um número grande de vizinhos no espaço e no tempo.

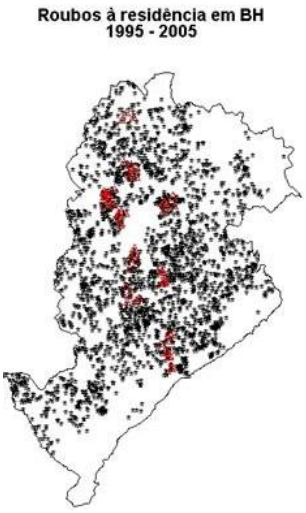

Figura 5.2: Mapas dos roubos em residências em BH de 1995 à 2005, com a localização, em vermelho, dos grupos de casos com uma maior interação espaço-temporal.

Lembrando que a significância estatística simplesmente significa que nós rejeitamos a hipótese nula. A capacidade do teste para detectar diferenças que levam à rejeição da hipótese nula depende do tamanho da amostra. Nos dados de arrombamento o teste de Knox foi significativo . No entanto, na prática, não encontrados um número grande de vizinhos no espaço e no tempo a ponto de ter pouco significado real. Utilizaremos aqui a nossa medida de associação para tentar mensurar a associação entre espaço tempo para os dados de arrombamento de Belo Horizonte. Com este intuito e usando o bootstrap proposto na seção 4.1 reamostramos 200 vezes, a nossa medida. Na figura 5.5, podemos perceber que a nossa medida apresenta valores baixos o que pode ser um indicio que a relação entre espaço tempo seja fraca.

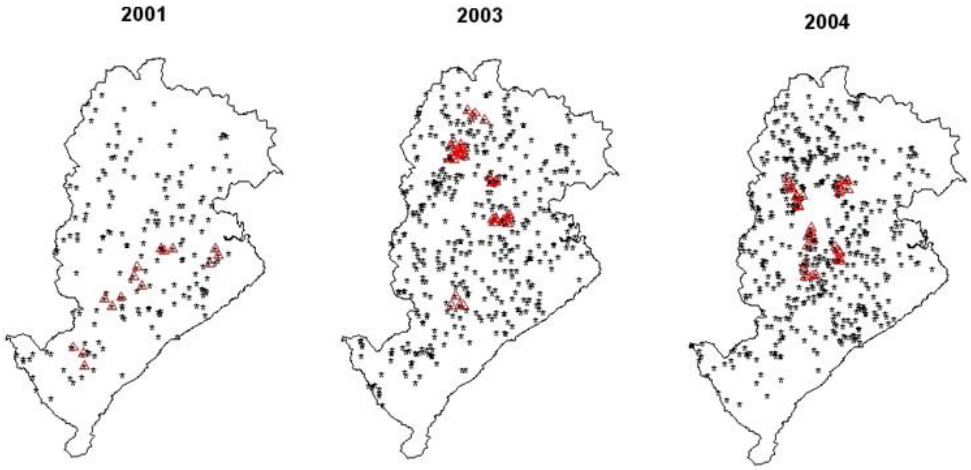

Figura 5.3: Mapas dos roubos em residências em BH para os anos de 2001, 2003, 2004, com a localização, em vermelho, dos grupos de casos com uma maior interação espaço-temporal.

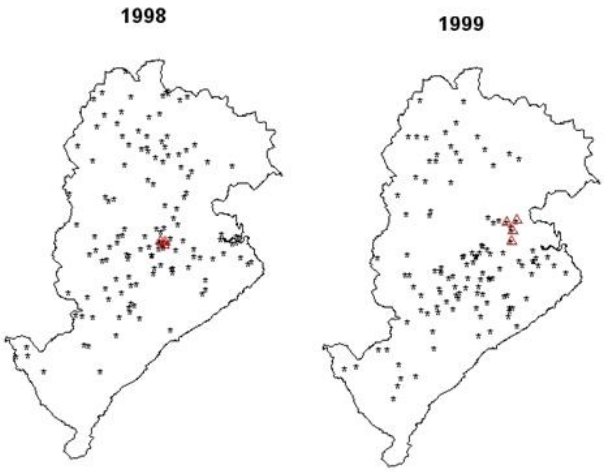

Figura 5.4: Mapas dos roubos em residências em BH para os anos de 2001, 2003, 2004, com a localização, em vermelho, dos grupos de casos com uma maior interação espaço-temporal.

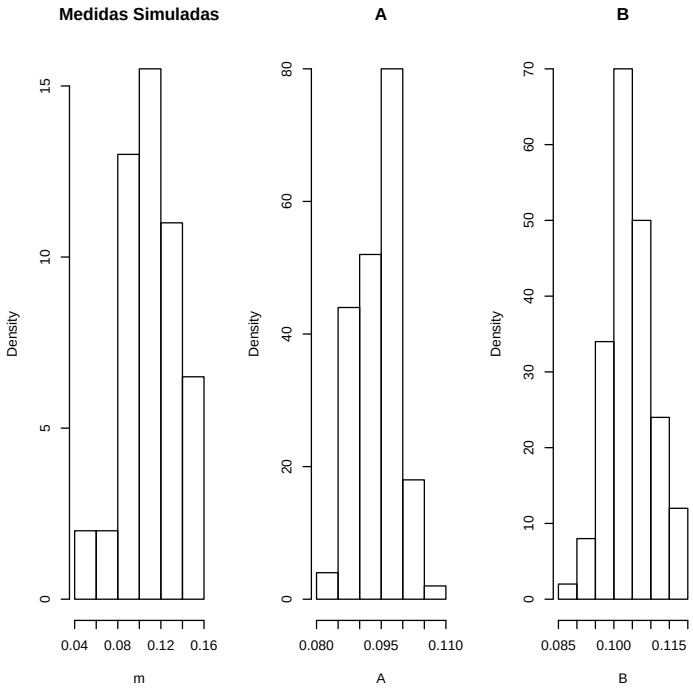

Figura 5.5: Nossa medida para os dados de roubos em residências em BH obtidas através de simulações via bootstrap

Capítulo 6

Conclusões e Trabalhos Futuros

A consideração simultânea dos padrões espaciais e temporais da ocorrência dos eventos é importante para identificar clusters ou conglomerados espaços-temporais. Definimos o cluster espaço-temporal como uma região geograficamente pequena em relação à região em estudo e que concentra um número excessivo de eventos durante um período limitado de tempo.

O teste de detecção de conglomerados espaços-temporais mais popular foi desenvolvido por Knox (1964). Especificando-se distâncias críticas temporais e espaciais é possível determinar se um par de eventos está próximo no tempo e no espaço. O teste baseia-se no número X de pares de eventos que estão simultaneamente próximos no espaço e no tempo. Um alto valor X seria uma indicação de que há uma tendência de casos próximos no tempo serem também próximos no espaço, retratando a interação espaço-tempo.

O teste de Knox, assim como as outras técnicas para testar a hipótese de independência entre espaço e tempo, sofre de um problema típico dos testes de hipóteses: se existirem muitos eventos, o teste pode ser significativo mesmo se a associação entre espaço tempo for fraca.

Seria de grande valia termos uma medida de associação que possa ser usada em conjunto com o teste de hipóteses e que mença a magnitude da possível relação entre as variáveis. Existem diversas propostas de medidas de associação para tabelas de contingência que procuram complementar o teste qui-quadrado de Pearson. Uma das quais foi proposta por Goodman e Kruskal (1964) denominada por tau de Goodman e Kruskal.

O tau de Goodman e Kruskal (notação τ_{GK}) é obtido usando o principio da redução proporcional dos erros.Isto é, o coeficiente tem por objetivo responder à questão: Em que medida o fato de conhecermos a classificação de uma das variáveis (por exemplo, a linha da tabela em que a observação se encontra) nos torna mais hábeis para prevermos a classificação da outra variável (a coluna na qual cai a observação)?

Esta estatística tem algumas propriedades desejáveis, tais como ser uma medida na direção da associação das variáveis com limites zero (nenhuma associação) e um (completa associação) e não mudar o seu valor com a permutação de linhas e colunas.

Além disso, τ_{GK} tem uma interpretação muito clara: mede o decréscimo relativo na probabilidade de errar a previsão da variável linha ao conhecer a variável coluna (ou vice versa). Por exemplo, se $\tau_{GK} = 0.8$, isto significa que temos uma redução de 80% na probabilidade de errar a previsão de uma das variáveis, quando se usa a informação sobre a outra variável.

Nesta dissertação de mestrado usei as idéias de Goodman e Kruskal para construir uma medida de associação entre espaço e tempo para processos pontuais, bem como variáveis aleatórias (X, Y) quaisquer. Esta medida tem boas propriedades tais como: ter os seus valores entre 0 e 1; se

as variáveis são independentes, o índice é zero; tem uma interpretação no sentido do quanto o conhecimento de uma das variáveis nos torna aptos para prever os valores da outra.

O meu interesse é discutir a relação da minha medida de associação entre espaço e tempo com outras medidas conhecidas na literatura tais como o "Scoring Rule". Um Scoring rule é uma medida de desempenho de um mecanismo tomador de decisão quando eles são usados repetidas vezes na tomada de decisões sob incerteza, como explico na próxima seção.

6.1 Scoring Rule

Scoring rules avaliam a qualidade das previsões probabilísticas, atribuindo uma pontuação numérica com base na previsão e sobre o evento que se materializa (Shervish *et al.* (2005)).

Por exemplo, um meteorologista da TV pode dar a probabilidade de chuva, segundo o seu mecanismo preditor, todos os dias. Um espectador pode observar o número de vezes que uma probabilidade de 25% foi citado para a chance de chover em um determinado dia , por um período de dez anos, e comparar com a proporção real de vezes que a chuva caiu. Se a porcentagem de dias chuvosos é substancialmente diferente da probabilidade declarada pelo meteorologista dizemos que o mecanismo de previsão usado pelo meteorologista está mal calibrado. O meteorologista pode ser incentivado a fazer uma melhor previsão, por um sistema de bônus.

Considere o cálculo da probabilidade de um único evento E. Assumimos que um individuo atribui uma probabilidade r para a ocorrência do evento,que provavelmente será diferente da verdadeira probabilidade de ocorrência do evento E ($P(E) = p$) O scoring rule $S(r, \omega)$ dá ao individuo uma bonificação da seguinte forma:
$S(r, \omega) = S_1(r)$, se ω é favorável a ocorrência do evento E e $S(r, \omega) = S_2(r, \omega)$, se ω não é favorável a ocorrência do evento E.

No nosso exemplo, suponha que nós recompensamos um metereologista com um bônus $s(r, \omega)$ quando ele faz uma declaração de chuva com uma probabilidade r e realmente chove. Sendo E o evento chover no dia . Supondo que o nosso meteorologista deseja maximizar sua recompensa esperada, então ele vai escolher uma previsão r que maximiza:

$$E[S(r, \omega)] = p(S_1(r)) + (1 - p)(S_2(r))$$

(nota: Alguns artigos denotam $E[S(r, \omega)]$ por $S(r, p)$ ou $S(r|p)$)

Um scoring rule $S(r, \omega)$ é dito ser próprio se $E[S(r, \omega)]$ é exclusivamente maximizada quando $p = r$ para qualquer valor de p $(0 \leq p \leq 1)$(Gneiting e A.E. (2004)).

E, um scoring rule é definido como estritamente próprio se:

$$E\left[s(r, \omega)\right] \leq E\left[s(p, \omega)\right],$$

sendo que a igualdade se da somente quando $r = p$.

A literatura sobre tais scoring rules é bastante extensa; várias formas scoring rules estritamente próprios foram desenvolvidas sendo o mais conhecido o score de Brier. O score de Brier (Brier (1950)) mede a qualidade de uma previsão de probabilidade . Ele mede o desvio médio quadrado entre a probabilidade prevista de um conjunto de eventos e seus resultados, portanto, um valor

pequeno representa maior precisão . O score de Brier é definido como:

$$s(r, \omega) = 2r_i - \sum_{j=1}^{n} r_j$$

Durante meus trabalhos percebi uma conexão entre o score de Brier e a minha medida de associação entre espaço e tempo.Dentre os meus objetivos futors esta o de estudar as possíveis conexões dos scoring rules com a medida de associação desenvolvida na minha dissertação e suas possíveis propriedades.

Dentre as meus objetivos futuros estão o deu tilizar as possíveis conexões dos scoring rules com a medida de associação desenvolvida na minha dissertação para dados contínuos para a aplicação a distribuições conhecidas, tal como a normal bivariada, e as relações não lineares além utilizar esta conexão dados reais, como por exemplo, aos dados de arrombamento a residências na cidade de Belo Horizonte.

Referências Bibliográficas

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability, *Monthly weather review 78: 13.* **78**: 1–3.

Diggle, P. (2003). *Statistical Analysis of Spatial Point Patterns (second edition)*, Edward Arnold.

Efron, B. (1982). *The Jacknife, the Bootstrap and Other Resampling Plans*, Department of Statistics Stanford University, Philadelphia.

Fan, J. e Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation., *J. Royal Statistical Society* **57(2)**: 371–394.

Gneiting, T. e A.E., R. (2004). Strictly proper scoring rules, prediction, and estimation, *Technical report*, Department of Statistics, University of Washington.

Goodman, L. A. (1964). Some alternatives to ecological correlation, *The American Journal of Sociology* **6**: 610–625.

Goodman, L. A. e Kruskal, W. H. (1954). Measures of association for cross classifications, *American Statistical Association* **49**: 732–764.

Knox, E. (1964). The detection of space-time interactions, *Applied Statistics* **13**: 25–29.

Miles, M. B. e Huberman, A. M. (1984). *Qualitative Data Analysis: A Sourcebook of New Methods*, Sage Publ.

Shervish, M. J., Seidenfeld, T. e Kadane, J. B. (2005). Proper scoring rules, dominated forecasts and coherence, *Decision Analysis* **6**: 202–221.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*, Springer, New York.