Apostila II

Estatística Espacial

Alexandre Diniz

Estatística Espacial

Análise espacial x não espacial

Em termos gerais, pode-se definir análise espacial como o estudo quantitativo de fenômenos posicionados no espaço. Portanto, o interesse é centrado nos processos que ocorrem no espaço e os métodos aqui empregados buscam descrever ou explicar o comportamento desses processos, bem como a sua relação com outros fenômenos espaciais. Os dados portanto, representam, no mais das vezes, amostras desses processos, a partir dos quais se busca fazer inferências sobre o comportamento desses fenômenos. Portanto, em análise espacial, utiliza-se de técnicas que explicitamente incorporam as localizações ou arranjos espaciais dos objetos/fenômenos em questão.

Um exemplo:

Biogeografia – o número de espécies de plantas e um conjunto de pequenas ilhas.

1- Relação entre o número de espécies e o tamanho das ilhas

- Relação positiva – quanto maior a área das ilhas, maior será o número de espécies, uma vez que o tamanho da ilha influencia a probabilidade de se encontrar uma maior diversidade de hábitats.
- Análise espacial não está particularmente envolvida neste estágio. O simples fato de que se está lidando com um conjunto de áreas que têm posição geográfica definida não torna o estudo espacial.

2- Distância – isolamento das ilhas em relação a outras ilhas e ao continente e o número de espécies.

- Relação negativa – evidência empírica demonstra que a distribuição de espécies de pássaros entre ilhas do Pacífico, demonstram que a distância da Nova Guiné reduz o número de espécies.
- Análise espacial mais claramente presente, uma vez que a localização relativa das ilhas (proximidade umas das outras e do continente) influenciam o número de espécies e estão sendo exploradas na análise.

O fato é que a análise espacial produz resultados diferentes daqueles das abordagens não espaciais, uma vez que os seus resultados são geralmente mais robustos por incorporarem a dimensão espacial.

Tipos de fenômenos e relações espaciais

Discretos – fenômenos espaciais são constituídos por pontos, linhas e áreas. Ex: plantas, pessoas, lojas, epicentros de terremotos.

Contínuos – ênfase é na continuidade de fenômenos espaciais. Exemplos clássicos são os fenômenos no ambiente natural tais como a temperatura, elevação, pressão atmosférica, solo, etc. Portanto, tais camadas de informação, em princípio, podem ser observadas e medidas em todos os pontos da terra.

Atributos - Os fenômenos discretos têm características ou atributos associados diretamente a eles. Tais atributos podem ser medidos a partir das escalas clássicas : nominal, ordinal ou intervalar. Pontos podem ser distinguíveis por classes de fenômenos (tipos de espécies de plantas), ordem (vilas, povoados, cidades, metrópoles), ou ao longo de uma escala contínua de medição (profundidade). O mesmo se aplica às linhas. Cursos d'água , por exemplo, podem ser limpos ou poluídos; de primeira ordem, segunda ordem ou terceira ordem. Ou ainda podem ter dados em relação ao seu volume de descarga. As áreas têm usos de solo específicos (nominal), mas essas mesmas áreas podem ser classificadas de acordo com alguma ordenação. Ou ainda, os setores censitários podem ser dados contínuos como densidade populacional, produção econômica, etc.

Porém se lidarmos exclusivamente com os atributos, ignorando as relações espaciais entre os pontos de amostragem, não se pode afirmar que estamos fazendo análise espacial. Mesmo que nossas unidades de

observação sejam espacialmente definidas. Portanto, apesar das variáveis serem importantes, esses valores divorciados do seu contexto espacial perdem valor e significado. Para a produção da análise espacial necessita-se, pelo menos, informação sobre a localização e atributos, independentemente de como esses atributos sejam mensurados.

Caso desejamos estudar apenas o arranjo espacial ou padrão espacial de entidades, então esta é essencialmente uma questão geométrica e nós apenas coletamos dados sobre as localizações das entidades. Mas, caso queiramos comparar os arranjos espaciais de diferentes entidades ou o estudo dos padrões espaciais de medidas tomadas a partir de locais específicos, então precisaremos fazer uso de informações sobre os atributos também.

Conceitos gerais em análise de dados espaciais

Análise de dados espaciais envolve uma descrição apurada de dados relacionados à processos operando no espaço, bem como a exploração de padrões e relações em tais dados e a busca de explicação para tais padrões e relações. Cabe portando fazer uma distinção entre os métodos que são essencialmente voltados para a visualização de dados espaciais – aqueles que são exploratórios; aqueles voltados para resumir e investigar padrões e relações espaciais; e aqueles que contam com especificação de um modelo estatístico e a estimação de parâmetros. Esta é uma distinção que auxilia no entendimento global, mas que não é muito precisa.

1. Visualizar dados espaciais

Requerimento essencial em todo tipo de analise de dados é a habilidade de ver a distribuição de dados. Gráficos de dispersão e outras modalidades gráficas são ferramentas importantes para o analista que busca compreender padrões espaciais, formular hipóteses e testa-las. Neste sentido na análise de dados espaciais, o mapeamento cumpre uma função muito importante. Equivalente do gráfico de dispersão.

2. Explorar dados espaciais

Métodos exploratórios de análise buscam realizar boas descrições dos dados, auxiliando o analista a desenvolver hipóteses sobre tais dados. Tais técnicas são marcadas por uma série de afirmações a priori sobre os dados e muitas são desenhadas especificamente para serem robustas, ou seja, resistentes à influencia de valores extremos ou outliers.

No contexto da análise de dados espaciais, os métodos exploratórios podem ser empregados no formato de mapas, enquanto outros podem envolver gráficos convencionais. Portanto, a diferença do agrupamento anterior se dá pelo nível de sofisticação empregado.

3. Modelar dados espaciais

Utilizados quando se busca testar certas hipóteses, ou estimar, com alguma precisão, a extensão e a forma que certas relações estabelecem. Modelos estatísticos estão implícitos em todas as formas de inferência estatística e teste de hipóteses, apesar do termo modelo não se utilizado explicitamente nos textos de estatística elementar. Como os modelos estatísticos estão voltados para fenômenos que são estocásticos (que são sujeitos a incerteza e variabilidade, ou governados pelas leis da probabilidade), utiliza-se uma linguagem que nos permite representar tal incerteza matematicamente. Por isso se faz uso do conceito de variável aleatória e a sua distribuição de probabilidade.

Existe uma miríade de métodos e técnicas voltadas para o estudo espacial de um conjunto de pontos, linhas e áreas. Aqui, porém, explorar-se-á algumas técnicas voltadas ao estudo de pontos e áreas

Métodos para distribuição de pontos

Aqui os dados consistem de uma série de pontos locacionais (s_1, s_2, ...s_n) em uma dada área ou região R, na qual eventos de interesse aconteceram. O termo evento é utilizado em um sentido mais geral, uma vez que os eventos em questão podem estar relacionados à uma grande variedade de fenômenos espaciais que podem ocorrer em cada ponto.

Ex: localizações de núcleos celulares em um a parcela de tecido microscópico, certos tipos de árvores em uma floresta, casos de doenças ou tipos de crime em uma região geográfica.

Um padrão espacial de pontos é um exemplo simples de dados espaciais, pois os dados são compostos basicamente pelas coordenadas dos eventos. Isto não significa necessariamente que a análise será mais fácil do que outros tipos de análise. Na verdade, do ponto de vista estatístico, padrões de pontos podem muitas vezes ser matematicamente mais complexos.

No entanto, muitas vezes os dados não são compostos exclusivamente pela localização de eventos. Pode-se também ter outros atributos associados a estes eventos que podem ser incorporados na análise. Tais atributos podem ser classes de fenômenos (tipos de crimes, espécies de plantas), ou o tempo da ocorrência de um dado evento (data).

O objetivo básico na análise espacial de padrões de pontos é examinar se um conjunto de eventos apresenta um padrão sistemático, ou aleatório. As alternativas são a formação de agrupamentos ou regularidade na distribuição de dados. Caso exista algum tipo de padrão sistemático, pode-se buscar compreender em qual escala este padrão ocorre e se algum tipo de padrão de concentração ocorre nas proximidades de algum tipo de fator ou entidade. A partir desta constatação, uma série de hipóteses são passíveis de serem formuladas.

Medidas de tendência central e dispersão espacial

1.0 Centro Médio

- Análogo à média aritmética.
- Definido como o ponto de um plano que minimiza a soma das distâncias quadráticas a todos os outros pontos do plano.
- Também pode ser encarado como o ponto de equilíbrio de um dado plano
- A posição é construída com base na média aritmética dos valores de X e de Y, tomados de maneira independente

Para o cálculo deve-se:

- Estabelecer um sistema de coordenadas
- Enumerar as observações
- Fazer a leitura das coordenadas de cada ponto
- Aplicar as seguintes fórmulas aos valores observados

Centro médio

$$X = \frac{\sum x_i}{n} \qquad Y = \frac{\sum y_i}{n}$$

2.0 Centro Médio Ponderado

No cálculo do centro médio, atribuímos o mesmo peso a todos os valores, além de levarmos em consideração apenas a localização dos pontos. O centro médio representa o centro gravitacional de um conjunto de pontos, independente da intensidade de ocorrência dos pontos. Entretanto, em certas ocasiões é imperativo levar em consideração os pesos dos diversos fenômenos associados à áreas específicas.

Ex_1: não basta saber a localização das indústrias, mas é importante levar em conta a sua produção;
Ex_2: não basta saber a localização das cidades, mas é importante levar em conta a sua população;
Ex_3: não basta saber a localização dos tornados ou terremotos, mas é importante levar em conta a sua magnitude ou intensidade.

Portanto, o valor de ponderação, em geral, é a intensidade de ocorrência de um determinado fenômeno.

Procedimentos

- Replicar as etapas 1-3 do cálculo do centro médio
- Aplicar as seguinte fórmulas

$$\mathbf{X_w = \frac{\sum w_i x_i}{\sum w_i}} \qquad \mathbf{Y_w = \frac{\sum w_i y_i}{\sum w_i}}$$

Obs: a localização do ponto médio e do ponto médio ponderado é afetada pela localização de cada ponto em particular, assim como o peso de cada localização. Pontos com localizações extremas, ou com pesos altos atraem para si o centro de distribuição.

3.0 Distância Padrão ou Raio Padrão

Não basta saber a média central de uma distribuição. Assim como na estatística descritiva não espacial, distribuições distintas podem apresentar o mesmo ponto central. Por isso, as medidas de variabilidade ou dispersão, tendo como base o ponto central são úteis no estudo de distribuições geográficas.

Enquanto na estatística não espacial a dispersão é medida acima e abaixo do ponto central (média), na estatística espacial a dispersão é medida em torno do ponto central. A distância padrão, porém representa o raio dinâmico, ou raio padrão, que representa a variabilidade de um conjunto de pontos em torno de um valor médio central. Ao logo do processo, obtém-se um círculo, centrado no centro médio, cujo raio é a distância padrão.

A distância padrão é equivalente ao desvio padrão. No entanto, o desvio padrão unidimensional tem por base diferenças ou distâncias quadráticas de cada valor de X à média do conjunto. A distância padrão terá a mesma fundamentação, porém em relação a dois eixos: X e Y.

.

$$\text{Distância padrão} = \sqrt{\frac{\sum(X_i - \text{Média de X})^2 + \sum(Y_i - \text{Média de Y})^2}{N}}$$

Obs_1: Assim como o desvio padrão é mensurado na unidade de medida dos dados originais, também a distância padrão é expressa nas unidades de medida de X e Y (Cm, Km, Milhas, Graus, etc.)

Obs_2:Valores muito discrepantes do conjunto de dados costumam influenciar o valor da distância padrão, uma vez que as suas distâncias são elevadas ao quadrado.

4.0 Distância padrão ponderada

O cálculo anterior não levou em consideração a magnitude, ou peso das localidades. Utilizou-se apenas a localização dos pontos.

Quando a magnitude do fenômeno localizado nos pontos for relevante, deve-se utilizar para o cálculo a fórmula:

$$\text{Distância padrão} = \sqrt{\frac{\sum w(X_i - \text{Média de Xw})^2 + \sum (Y_i - \text{Média de Yw})^2}{\sum w}}$$

Onde o centro médio ponderado é dado por:

$$X_w = \frac{\sum w_i x_i}{\sum w_i} \qquad Y_w = \frac{\sum w_i y_i}{\sum w_i}$$

Obs: As distâncias ponderadas e não ponderadas nem sempre são coincidentes. O tamanho do raio é proporcional ao grau de dispersão das distribuições e ao peso dos fenômenos nos pontos em questão.

5.0 Dispersão relativa

Quando um fenômeno é estudado em áreas de tamanho diferente, deve-se resistir à tentação de comparar os resultados relativos à dispersão. A técnica é muito sensível às distâncias, e consequentemente ao tamanho da área de estudo. A dispersão relativa, no entanto, oferece uma solução para este problema. A técnica assemelha-se ao coeficiente de variação. O seu cálculo envolve além da distância padrão, a área efetiva na qual a distância padrão foi mensurada, que pode também ser representada por um raio ou círculo. Portanto,

$$\text{Dispersão relativa} = \frac{\text{Distância padrão}}{ra}$$

Onde,

ra – Raio da área efetiva de mensuração do fenômeno estudado

Em IDRISI, CENTER calcula a média ponderada e não ponderada e o raio padrão de um conjunto de pontos expressos como freqüências de células. No processo, o coeficiente de dispersão relativa é também calculado. A rotina CENTER requer que se revele o nome da imagem a ser avaliada. No processo, a rotina assume que os valores presentes na imagem representem o número de pontos em cada célula (para o cálculo da média central), ou pesos a serem aplicados à posição da célula (para o cálculo da média central ponderada).

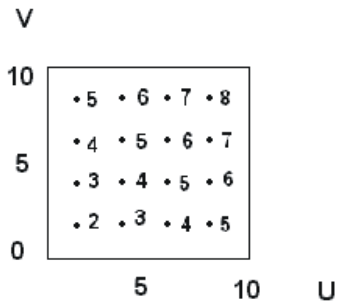
Os resultados trazem o centro médio ponderado pelos valores das células, levando-se em consideração as sua posição em relação às linhas e colunas, bem como em relação ao sistema de coordenadas utilizados. Da mesma maneira, o raio padrão é expresso em termos de células e unidades de referência. CENTER também indica o coeficiente de dispersão relativa (CRD). O CRD expressa o grau de dispersão relativa em uma dada área. Ele é calculado a partir da razão entre o raio padrão e o raio de um círculo que contenha a mesma área sendo estudada, vezes 100.

Análise de Tendência de Superfície

Técnica similar à análise de regressão múltipla, na qual as variáveis independentes são compostas pelas coordenadas espaciais dos diversos pontos em questão. O objetivo é identificar a tendência distributiva de um dado fenômeno em uma determinada região, a partir da identificação de tendências relativas à latitude e longitude. Tal objetivo é atingido a partir do cálculo de equações polinomiais lineares, quadráticas e cúbicas para um conjunto de pontos.

Exemplo:

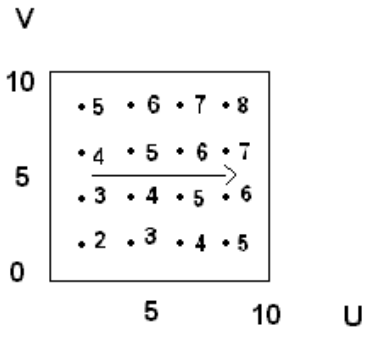
O tamanho médio dos seixos encontrados a partir de pontos amostras em uma praia. Existe uma tendência espacial em relação a esta distribuição?

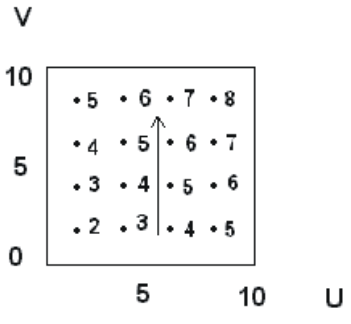


Ao produzirmos uma análise de regressão simples, utilizando o tamanho médio dos seixos como variável dependente (Y) e as duas coordenadas dos pontos (U) e (V) teremos:

$Y = 1.45 + 0.71U$ $r_{YU} = 0.85$
$Y = 1.45 + 0.71V$ $r_{YV} = 0.85$

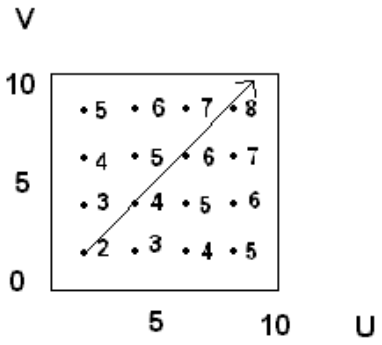
A primeira dessas equações indica uma forte tendência de crescimento no tamanho médio dos seixos em relação ao leste ao longo da praia. Por outro lado, a segunda equação sugere a presença de uma forte tendência ao aumento no tamanho médio dos seixos em direção ao norte da praia.





No entanto, ao analisarmos mais detidamente a distribuição de pontos, percebe-se um incremento no tamanho médio dos seixos na direção sudoeste – nordeste. Introduzindo-se as duas medidas simultaneamente na equação tem-se:

Y=0.0 + 0.5U + 0.5V RY.UV = 1.0



Obs: Como se está utilizando a mesma unidade de medida para U e V (graus), pode-se comparar os coeficientes de regressão.

Portanto, neste exemplo acabamos de demonstrar a utilização de uma equação polinomial linear, na qual

$Z = b_0 + b_1x + b_2y$

Ou seja, trabalhamos com tendências unidirecionais (Norte-Sul x Leste-Oeste). Entretanto, pode-se trabalhar com padrões espaciais mais complexos com a ajuda de outras equações. Neste sentido, poder-se-ia ter relações como essas:

$Y= a + b_1U – b_2U^2 +/-$ e

Indicando que o tamanho médio dos seixos aumentaria ao nos distanciarmos da paria no sentido leste e depois voltaria a diminuir. Ou ainda,

$Y= a - b_1V – b_2V^2 +/-$ e

Indicando que o tamanho médio dos seixos diminuiria e depois aumentaria no sentido norte.

Adicionando-se as duas fórmulas demanda a incorporação de um termo de interação entre U e V de modo que se tem uma relação quadrática:

$Y = a + b_1U + b_2U^2 + b_3UV + b_4V + b_5V^2 +/-$ e

Tal equação produz uma tendência de superfície quadrática, cuja forma se assemelha a de um domo.

Superfícies de grandeza maior devem ser produzidas onde os padrões espaciais forem excessivamente complexos. Na maioria dos casos tais superfícies são utilizadas para descrever padrões de maneira mais apurada, já que o significado dos coeficientes de regressão são mais difíceis de serem interpretados.

Em IDRISI o módulo TREND calcula e produz superfícies de tendência com base em equações polinomiais lineares, quadráticas e cúbicas. O módulo TREND, além de produzir a imagem, também traz informações sobre a equação polinomial produzida, bem como o seu poder de predição. O número de células utilizadas para determinar a imagem de tendência é informada, seguido dos coeficientes da equação polinomial. O

coeficiente "b0" representa o termo interceptor, e os coeficientes "bn" são os coeficientes de inclinação (b) associados à ordem da equação polinomial selecionada. O grau de predição do modelo é expresso em percentagem e é acompanhado pelo valor de F e os graus de liberdade.

Métodos para o estudo de dados relativos à áreas

Na parte anterior nós estávamos preocupados com a variação espacial de um atributo localizado em pontos específicos da superfície, ou com atributos que variavam continuamente no espaço, mas que foram amostrados em pontos específicos da superfície. Neste módulo, porém, voltaremos a atenção para atributos que não variam continuamente, mas que têm valores específicos para sub-áreas que compõem uma dada região de estudo. Essas entidades fixas podem constituir unidades de área como estados, municípios, distritos, setores censitários, ou até mesmo pixels.

Neste sentido, não estamos interessados em estimar valores de atributos vinculados a certos pontos no espaço nos quais observações são feitas. Primeiro, porque não existem valores entre as áreas de análise, e depois porque os valores, no mais das vezes, já foram observados em todas as áreas possíveis. Aqui estamos mais interessados na detecção e possíveis explicações para o padrão espacial ou tendência de distribuição para esses valores de área.

Cruzamento de dados categóricos

1.0 O teste de Qui-quadrado

Avalia se freqüências obtidas empiricamente diferem significativamente daquelas esperadas com base em suposições teóricas. Testa a relação entre duas variáveis nominais com base em uma tabela de contingências.

Ex: relação entre a localização residencial e modalidade de domicílio.

X1 = centro x periferia
X2 = domicílios chefiados por homens x domicílios chefiados por mulheres

Tipo de domicílio	Centro	Periferia	Total
Mulheres	60	40	100
Homens	175	250	425
Total	235	290	525

H1: Existe uma relação entre o tipo predominante de domicílio e a localidade no âmbito da cidade
H0: Não existe relação alguma entre o tipo predominante de domicílio e a localidade no âmbito da cidade

Com base nos dados acima, sabe-se que:

25,53% dos domicílios do centro são chefiados por mulheres, contra 13,79% daqueles situados na periferia. Esta diferença é estatisticamente significativa? Através do teste de Qui-quadrado pode-se testar esta suspeita.

$$X2 = \frac{\sum^{c}\sum^{l} (O_{ij}-E_{ij})^2}{E_{ij}}$$

Onde,
$\sum^{c}$ – Somatório das colunas
$\sum^{l}$ - Somatório das linhas
O_{ij} – Valores observados
E_{ij} - Valores esperados caso nenhuma relação existisse

Para gerar a tabela de freqüências esperadas:

A	B	A+B
C	D	C+D
A+ C	B+D	A+B+C+D = N

Número esperado na célula A

$$\frac{(A+B)\,(A+C)}{N} = \frac{(235)\,(100)}{525} = 44.76$$

Número esperado na célula B

$$\frac{(A+B)\,(B+D)}{N} = \frac{(290)\,(100)}{525} = 55.24$$

Número esperado na célula C

$$\frac{(C+D)\,(A+C)}{N} = \frac{(235)\,(425)}{525} = 190.24$$

Número esperado na célula D

$$\frac{(C+D)\,(B+D)}{N} = \frac{(425)\,(290)}{525} = 234.76$$

Valores esperados

Tipo de domicílio	Centro	Periferia	Total
Mulheres	44.76	55.24	100
Homens	190.24	234.76	425
Total	235	290	525

Inserindo os valores na fórmula:

$$X^2 = \frac{(60 - 44.76)^2}{44.76} + \frac{(40-55.24)^2}{55.24} + \frac{(175-190.24)^2}{190.24} + \frac{(250-234.76)^2}{234.76} =$$

$X^2 = 5.19 + 4.20 + 1.22 + 0.99$

$X^2 = 11.6$

Graus de liberdade:

C e L > 1 GL= (C-1) (L-1)
C = 1 e L ≠ 1 GL= (L-1) (1)
C ≠ 1 e L = 1 GL= (C-1) (1)

C = Nº de colunas
L = Nº de linhas

Comparar valor computado com valor crítico

X^2 crítico = 6.635
X^2 computado = 11.60

Conclusão: Confirma-se a presença de uma relação estatística
Obs:
Amostras devem ser independentes
N por célula deve ser maior ou igual a 5

2.0 Coeficiente de V de Cramer

Medida de grau de associação entre variáveis nominais, configurando-se como um coeficiente de correlação que oscila entre 0, indicando ausência de correlação, e 1 ,indicando uma correlação perfeita. Utilizado quando se trabalha com tabelas retangulares (4X; 5X7).

Produzido a partir da seguinte fórmula:

$$\text{Cramer's V} = \frac{X^2 .}{N\,(k-1)}$$

Sendo que o valor mínimo se refere ao menor número de colunas e linhas presentes na análise.

Com base nos dados acima:

$$\text{Cramer's V} = \frac{11.6 .}{525\,(1)}$$

Cramer´s V = 0.1486

3.0 Kappa

Permite confirmar a relação entre **duas variáveis categóricas que possuem classes similares**. O coeficiente de Kappa oscila entre 0 e 1, com interpretação similar à de Cramer's V.

Ex: Dois supervisores classificaram o desempenho em sala de aula de 72 monitores de acordo com os seguintes critérios: autoritário, democrático e permissivo. Os resultados finais são apresentados abaixo:

Supervisor 2 ►

Supervisor 1▼	Autoritário	Democrático	Permissivo	Total
Autoritário	17 (23.6)	4 (5.6)	8 (11.1)	29 (40.3)
Democrático	5 (6.9)	12 (16.7)	-	17 (23.6)
Permissivo	10 (13.9)	3 (4.2)	13 (18.1)	26 (36.1)
Total	32 (44.4)	19 (26.4)	21 (29.2)	72 (100.0)

A media mais simples de concordância entre as duas avaliações seria simplesmente a proporção de avaliações que são comuns entre os dois supervisores. Neste caso o resultado seria 58.3%. Entretanto, esta metodologia não leva em consideração o fator chance no processo de coincidência entre as avaliações.

Para corrigir este erro, pode-se computar a proporção de casos coincidentes que se esperaria encontrar caso as avaliações fossem independentes. Por exemplo, o supervisor 1 classificou 40.3% dos monitores como autoritários, enquanto o supervisor 2 classificou 44.4% dos monitores como autoritários. Caso as avaliações fossem independentes, você esperaria que 17.9% (40.3% x 44.4%) dos monitores fossem classificados como

autoritários por ambos supervisores. Da mesma forma 6.2% (23.6% x 26.4%) seriam classificados como democráticos e 10.5% (36.5% x 29.2%) como permissivos. Portanto, 34.6% de todos os monitores seriam classificados da mesma forma por pura chance.

A diferença entre a proporção observada de casos nos quais as avaliações coincidem e aquela que ocorre ao acaso é 0.237 (0.583-0.346). O coeficiente de Kappa nomaliza esta diferença ao dividi-la pela diferença máxima possível aos totais marginais. Neste exemplo, o maior valor possível de concordância real é de 1-0.346 (nível de chance). Portanto,

Kappa = 0.237/ (1-0.346) = 0.362

Kappa = diferença entre concordância observada e concordância ao acaso
(1-concordância ao acaso)

Em Idrisi, a opção CROSSTAB executa duas operações. Primeiramente, o programa executa um cruzamento de imagens na qual as categorias de uma imagem são comparadas às categorias de uma segunda imagem em forma de tabela. Nesta tabela são apresentadas todas as formas de combinação. O resultado desta operação é uma tabela que traz os totais tabulados, bem como duas medidas de associação entre as imagens.

A primeira desta medidas é o Cramer's V. Além disso, a estatística de Chi-quadrado é também produzida juntamente com os graus de liberdade apropriados para que a significância de Cramer's V possa ser testada. Se o Chi-quadrado é significante, o mesmo vale para Cramer's V.

Caso as duas imagens tenham o mesmo número de categorias, uma outra medida de grau de associação está disponível: Kappa. Esta medida também varia entre 0.0 e 1.0 com a mesma interpretação. Entrementes, Kappa só pode ser utilizada quando as categorias atinentes às imagens em questão expressem o mesmo tipo de dados, com o mesmo número de classes.

A Segunda operação é a cross-classificação de imagens. Tal operação pode ser definida como uma justaposição (OVERLAY) múltipla, demonstrando todos os tipos de combinação. O resultado é uma nova imagem que demonstra os locais de todas as combinações entre categorias presentes nas imagens originais.
No processo, uma legenda é automaticamente produzida demonstrando essas combinações.

Obs_1. As imagens de entrada não podem possuir mais do que 128 categorias.
Obs_2. A imagem de saída não pode possuir mais do que 256 categorias.

Análise de Regressão Linear Múltipla

Ver apostila de estatística básica

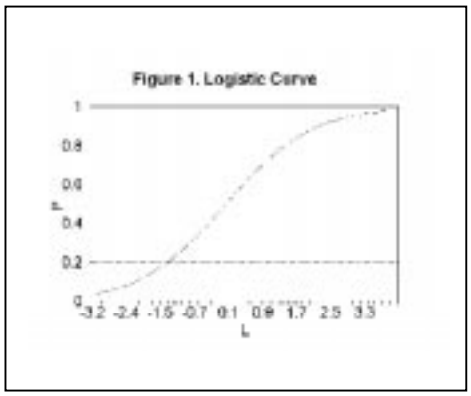
Análise de Componentes Principais

Ver apostila de estatística básica

Análise de Regressão Logística

Em regressão logística, a variável dependente é discreta, tal qual o uso e ocupação do solo, por exemplo. Se a variável dependente é dicotômica, o valor de Y toma apenas dois valores 1 e 0. Ao estimarmos mudanças na cobertura florestal, por exemplo, Y=1 representaria eventos nos quais a floresta se modificou, enquanto Y=0 representaria eventos nos quais as florestas permaneceram da mesma forma no período em questão.

No caso de três variáveis independentes, a regressão logística é definida da seguinte forma:



logit(p)=a+b1*x1+b2*x2+b3*x3

Onde,
logit(p) = ln(p/(1-p)) (L na figura acima), p é a variável dependente, a probabilidade de que um determinado evento tenha ocorrido. Os demais componentes têm o mesmo significado que em regressão linear múltipla.

A relação entre uma variável dependente e outra independente segue uma curva logística como representado a figura acima.

Note-se que a transformação logística de dados dicotômicos garante que a variável dependente da regressão seja contínua, e que a nova variável (a probabilidade de ocorrência de um determinado fenômeno) seja irrestrita, sem fronteiras. Além do mais, a transformação também garante que as estimativas de probabilidade sejam contínuas oscilando entre 0-1.

O método para determinar a relação logística utilizado no modelo LOGITREG é o método nos quadrados mínimos com introdução de pesos, que segue os seguintes procedimentos:

(1). LOGITREG transforma a probabilidade de se produzir uma nova variável dependente L (também chamada de razão de probabilidades):

$Li = \text{logit}(p_i) = \ln(p_i/(1-p_i))$

Onde p_i é a probabilidade para observação $_i$ (note-se que LOGIT significa unidades logísticas)

(2). Introduz pesos para cada observação (para variáveis dependentes e independentes) usando:

$W_i = \sqrt{N_i} * p_i*(1-p_i)$

Onde,
W_i é o peso para cada observação i,
N_i é o tamanho da amostra a partir da qual a probabilidade para cada observação foi calculada.

(3). Aplica-se o método dos quadrados mínimos à seguinte equação linear:

$W_i * L_i = \sum(W_i * b_k * X_{ik})$

Onde,

K é o número de variáveis independentes

X_{ik} é o valor da variável Kn para cada observação i
b_k é o coeficiente para a variável Kn

(4) Para estimar pi a partir do Li predito para cada observação i:

$P_i = \exp(L_i)/1+\exp(L_i)$

A regressão logística tem os mesmo pré-requisitos em relação às variáveis do que a regressão linear múltipla. Portanto, quando se utilizar de imagens neste módulo, verifique se estas não são espacialmente autocorrelacionadas. Por esta razão, os graus de liberdade verdadeiros não são calculáveis (por isto os graus de liberdade aparentes são computados. Para reduzir o impacto da dependência espacial entre observações, pode-se tratar as imagens antes de se gerar o modelo LOGITREG.

Neste sentido, pode-se
(1) usar o refinamento de pixels, a partir do módulo CONTRACT para se tomar apenas porções de amostras
(2) produza amostras de uma imagem e extraia os valores das variáveis dependente e independente para a produção de arquivos de dados
(3) usar uma imagem máscara para eliminar certos pixels da análise

Em Idrisi o módulo LOGITREG produz regressão logística e estimativas a partir de imagens ou arquivos de dados. LOGITREG demanda que se indique o tipo de regressão a ser produzido, entre imagens ou arquivos de dados.

A variável dependente deve ser uma imagem cujo formato é real binário, demonstrando probabilidades que oscilam entre 0 e 1. Já as imagens independentes devem estar em formato real. A imagem máscara, quando utilizada, deve estar m formato byte-binário com 1's em todas as células que deverão compor a regressão e 0 nos demais locais.

É também necessário estabelecer um nível de confiança que oscila entre 0 e 1 (a medida em que eventos futuros corresponderão à imagem indicada). Uma outra tarefa importante é definir o tamanho da amostra para o cálculo de probabilidade (número inteiro) a partir dos quais serão calculadas probabilidades associadas a todos os pontos.

Já a rotina LOGITREG, demanda como variável dependente uma imagem probabilística, representando a probabilidade de ocorrência de um certo evento. A probabilidade pode ser obtida a partir da freqüência na qual o evento ocorreu em uma amostra de células, ou ao combinar-se evidência indireta com julgamentos subjetivos do tipo Bayesian and Dempster-Shafer uncertainty.

O tamanho da amostra é um parâmetro demandado pela rotina LOGITREG. Caso você tenha apenas uma imagem Booleana para um determinado evento, pode-se produzir uma imagem de probabilidade utilizando-se a rotina FILTER a partir de um tamanho de kernel definido. Isto assume que cada pixel no filtro de kernel é uma amostra para o pixel central, e que o tamanho da amostra é o conjunto de pixels inseridos no filtro de kernel.

Bibliografia Básica

Bailey, T and Gatrell, A. **Interactive Spatial Data Analysis**. New York – Longman Scientific & Technical – 1995.

Barber, G. **Elementary Statistics of Geographers** New York – The Guilford Press – 1988.

Clif, A and Ord, J. **Spatial Autocorrelation** – London – Pion Limited - 1973

Collett, D. **Modelining Binary Data** – London – Chapman & Hall - 1991

Eastman, J. **Idrisi 32 Tutorial** – Worcester, Ma. Clark Labs – 1999.

Hammond, R. and Mccullagh, P. **Quantitative Techniques in Geography: an introduction**. London, Oxford University-1975.

Gerardi, L. Silva, B. **Quantificação em Geografia** – São Paulo – Difel - 1981

Johnston, R.J. **Multivarite Statistical Analysis in Geography**. New York – Longman Scientific & Technical. 1978.

King, L. **Statistical Analysis in Geography**. Englewood Cliffs, N.J. Prentice –Hall Inc.- 1969.

Mongomery, D. and Peck E. **Introduction to Linear Regression Analysis**. New York – John Wiley & Sons, INC. 1992.

Wrigley, N. **Statistical Applications in the Spatial Sciences**. London – Pion Limited – 1979.