

5 - Dados de área (*areal data*).

Prof. Vinícius D. Mayrink

EST171 - Estatística Espacial

Sala: 4073

Email: vdm@est.ufmg.br

1º semestre de 2024

Apresentamos aqui ferramentas exploratórias e modelagens desenvolvidas para dados coletados em unidades de área.

Os dados de área, em geral, representam unidades geográficas com formato irregular. Entretanto, podemos também trabalhar com o caso particular de *Lattices* (grid regular).

No contexto dos dados de área, as questões de inferência mais comuns são:

(1) Existe padrão espacial? Ele é forte?

“Padrão espacial”: Medidas similares para unidades de área próximas umas das outras, medidas diferentes para unidades de área distantes umas das outras. Algumas vezes, esse padrão pode ser identificado visualmente, entretanto, precisamos saber quantificar sua intensidade. Se não há padrão espacial, as medidas de interesse são independentes da vizinhança (arranjo espacial aleatório dos valores).

(2) Desejamos suavizar os dados? Quanto?

Exemplo: Contagens de casos de câncer em diversas unidades de área. O conjunto de dados pode conter *Outliers*; os valores grandes são maiores do que o esperado? Em uma superfície de contagens esperadas teremos valores altos puxados para baixo e valores baixos puxados para cima (noção da suavização). Suavização máxima (único valor para todas as regiões; exagero). Não há uma regra fixa para determinar a melhor intensidade da suavização.

(3) Realizaremos inferência para novas áreas? Iremos agrupar áreas?

Exemplo: Modificar as unidades de áreas (blocar/juntar regiões vizinhas); o que dizer sobre as medidas referentes ao bloco tomando como base as medidas individuais originais.

(4) Análise descritiva ou Modelagem?

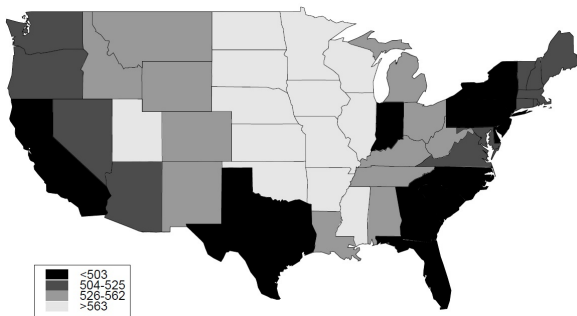


Fig.2 : Pontuação média obtida no exame verbal SAT em 1999.

Estes dados exibem forte padrão espacial; estados da região central (*midwest*) tem melhor desempenho. Algumas covariáveis poderiam ser consideradas para ajudar a explicar esse padrão como, por exemplo, a proporção de estudantes que fazem o exame (historicamente, instituições do *midwest* americano preferem o exame ACT ao invés do SAT; apenas os melhores estudantes desta região tentam o SAT).

Matriz de proximidade.

Dadas as medidas Y_1, Y_2, \dots, Y_n associadas com as unidades de área $1, 2, \dots, n$, especificamos uma matriz W com entrada w_{ij} referente as áreas i e j . Assumimos diagonal principal $w_{ii} = 0$. Algumas possibilidades para w_{ij} são:

- $w_{ij} = 1$ se i, j dividem fronteira comum (0 caso contrário);

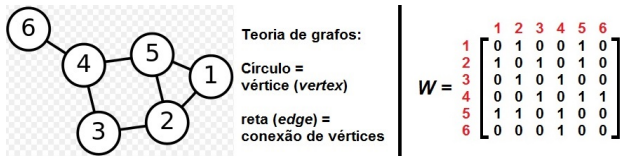


Fig.3 : Exemplo: Grafo com 6 vértices (regiões).

- w_{ij} é a inversa da distância entre as unidades.
- $w_{ij} = 1$ se a distância entre unidades é $\leq k$ (0 c.c.).
- $w_{ij} = 1$ para as m áreas vizinhas com maior proximidade (0 c.c.).

Elementos de W podem ser vistos como pesos;
o peso é maior quando i e j são áreas próximas.

W fornece um mecanismo para introduzir a estrutura espacial na modelagem.

Considere \tilde{W} tal que: a linha i é padronizada por $w_{i+} = \sum_j w_{ij}$,
ou seja, $\tilde{w}_{ij} = w_{ij}/w_{i+}$.

Neste caso, temos: $\tilde{W} \mathbf{1} = \mathbf{1}$ que é um vetor coluna de 1's.

Em palavras, a soma de qualquer linha em \tilde{W} resultará em 1 (\tilde{W} é estocástica nas linhas, pode não ser simétrica).

Defina os seguintes intervalos de distâncias: $(0, d_1]$, $(d_1, d_2]$, $(d_2, d_3]$, etc.
Isto introduz os seguintes conceitos:

- Vizinhança de 1ª ordem da unidade i :
São as unidades com distância $\in (0, d_1]$ em relação a i .
Matriz de proximidade:
 $W^{(1)}$ é tal que $w_{ij}^{(1)} = 1$ se i, j são vizinhos de 1ª ordem (0 c.c.).
- Vizinhança de 2ª ordem da unidade i :
São as unidades com distância $\in (d_1, d_2]$ em relação a i .
Matriz de proximidade:
 $W^{(2)}$ é tal que $w_{ij}^{(2)} = 1$ se i, j são vizinhos de 2ª ordem (0 c.c.).
- Vizinhança de 3ª ordem da unidade i :
São as unidades com distância $\in (d_2, d_3]$ em relação a i .
Matriz de proximidade:
 $W^{(3)}$ é tal que $w_{ij}^{(3)} = 1$ se i, j são vizinhos de 3ª ordem (0 c.c.).
- e assim por diante...

Medidas de associação espacial.

Duas estatísticas usadas para medir a intensidade da associação espacial entre unidades de área são conhecidas por “*I de Moran*” e “*C de Geary*”.

Elas são extensões (para o contexto espacial) de estatísticas que medem associação entre observações de séries temporais. Sejam z_1, \dots, z_n observações de uma série temporal, temos:

- Autocorrelação de ordem k (caso particular do *I de Moran*)

$$\hat{\rho}_k = \frac{\sum_{i=k+1}^n (z_i - \bar{z}) (z_{i-k} - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2}.$$

- Estatística Durbin-Watson (caso particular do *C de Geary*).

$$DW = \frac{\sum_{i=1}^n (z_i - z_{i-1})^2}{\sum_{i=1}^n (z_i - \bar{z})^2}.$$

Esta estatística é comumente usada para testar a presença de autocorrelação entre os resíduos da regressão. Se z_i é um resíduo, temos $\bar{z} \approx 0$ na formulação acima.

Considere as medidas Y_1, \dots, Y_n associadas com as unidades de área $1, 2, \dots, n$. A estatística *I de Moran* tem a seguinte forma.

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_i \sum_j w_{ij}) \sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Veja que esta estatística é uma razão de expressões quadráticas envolvendo \mathbf{Y} .

É possível mostrar que se Y_i 's são i.i.d., então I será assintoticamente Normal com média $-1/(n-1) \approx 0$ e variância

$$\text{Var}(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2},$$

sendo

$$S_0 = \sum_i \sum_j w_{ij}, \quad S_1 = (1/2) \sum_i \sum_j (w_{ij} + w_{ji})^2 \quad \text{e}$$

$$S_2 = \sum_{k=1}^n [(\sum_{j=1}^n w_{kj}) + (\sum_{i=1}^n w_{ik})]^2.$$

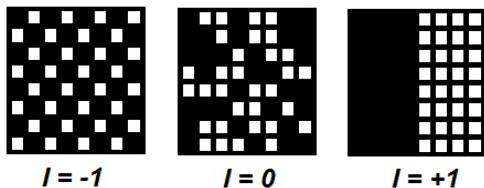


Fig.4 : Exemplo: Configuração espacial simples e a estatística I de Moran. Em geral temos $I \in [-1, 1]$ (é possível obter $|I| > 1$); -1 = dispersão perfeita, 0 = padrão espacial aleatório e $+1$ = correlação espacial perfeita.

A estatística C de Geary é dada por

$$C = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n w_{ij} (Y_i - Y_j)^2}{2(\sum_i \sum_j w_{ij}) \sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Temos $C \geq 0$ com média 1.

Veja que C também é uma razão de expressões quadráticas envolvendo \mathbf{Y} .

Se Y_i 's são i.i.d., então C também será assintoticamente Normal; ver Ripley (1981) para detalhes sobre a variância.

Valores baixos de C (i.e. entre 0 e 1) indicam presença de associação espacial.

Para realizar um teste de significância da associação espacial via estatística I ou C , recomenda-se a abordagem Monte Carlo.

Sob a hipótese nula de " Y_i 's i.i.d.", a distribuição de I ou C deverá ser invariante a permutações dos Y_i 's.

O teste exato requer a avaliação de I ou C para todas as $n!$ permutações dos Y_i 's; isto será computacionalmente caro se n for grande.

Estratégia: trabalhar com, por exemplo, 1000 permutações aleatórias dos Y_i 's (amostras Monte Carlo). Devemos comparar o valor de I ou C (da amostra original) com aqueles calculados para as amostras com permutação; verifique se o valor de I ou C (original) aparece raramente ou com frequência entre os I 's ou C 's das amostras com permutação (calcule um p-valor empírico, *script R*).

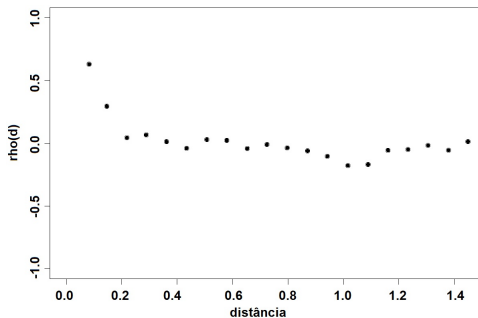


Fig.5 : Correlograma baseado na estatística I de Moran (script R).

- Calcule $I^{(1)}$: Na expressão de I , substitua w_{ij} por $w_{ij}^{(1)}$ de $W^{(1)}$.
- Calcule $I^{(2)}$: Na expressão de I , substitua w_{ij} por $w_{ij}^{(2)}$ de $W^{(2)}$, e assim por diante...
- Gráfico de $I^{(r)}$ vs. r . Conforme r cresce, observar um declínio inicial seguido por uma oscilação em 0. Isso indica "padrão espacial".
- Análogo espacial do gráfico de autocorrelação temporal.

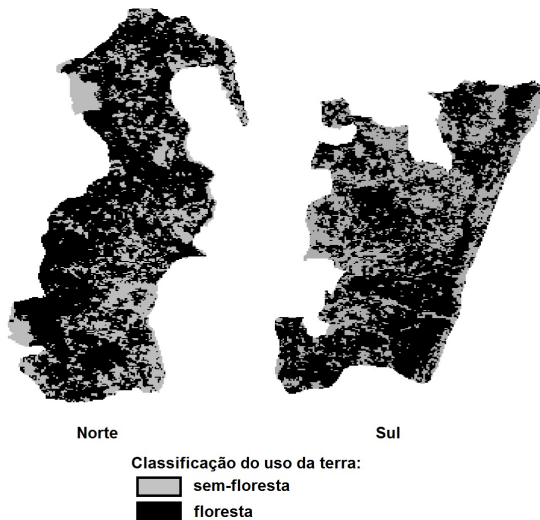


Fig.6 : Mapa com classificação binária (uso da terra) para aproximadamente 25000 pixels ($1km \times 1km$) representando a região leste de Madagascar.

O mapa binário anterior mostra um padrão espacial do uso da terra.

O correlograma para dados binários, apresentado a seguir, pode ser usado como alternativa para avaliar a associação espacial em uma direção particular: Oeste-Leste (W-E), Sul-Norte (S-N), Sudoeste-Nordeste (SW-NE), Sudeste-Noroeste (SE-NW).

Neste caso, os componentes espaciais são reduzidos para 1 dimensão; podemos então calcular autocorrelações com diferentes *lags* para as direções. Veja os passos a seguir...

Dado um pixel de referência, identifique outros pixels em uma direção de interesse.

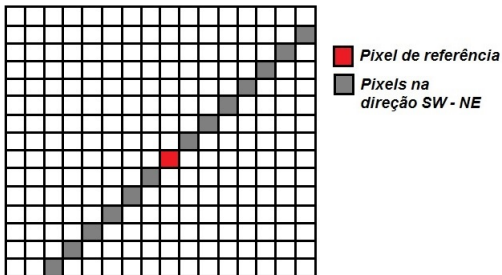


Fig.7 : Construindo o correlograma para dados binários.

Calcule as distâncias (e.g. Euclidiana) até o pixel de referência.

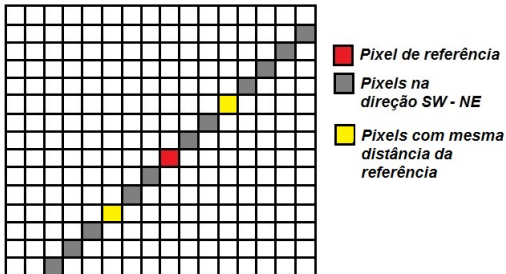


Fig.8 : Construindo o correlograma para dados binários.

Notação:

- $X = 0$ ou 1 indica a resposta observada do pixel de referência.
- $Y = 0$ ou 1 indica a resposta observada de um pixel localizado a uma distância d (direção fixada) do pixel de referência.

Avalie diversos pixels de referência na região de interesse.

Para cada pixel de referência e para uma direção e distância d fixadas, obtenha todos os pares (X, Y) e com eles construa uma tabela (2×2) como segue:

	$Y = 1$	$Y = 0$
$X = 1$	p_{11}	p_{10}
$X = 0$	p_{01}	p_{00}

sendo p_{11} a proporção de resultados $(X = 1, Y = 1)$.

Calcule $\log(odds) = \log\left(\frac{p_{11}p_{00}}{p_{10}p_{01}}\right)$ para cada combinação direção/distância.

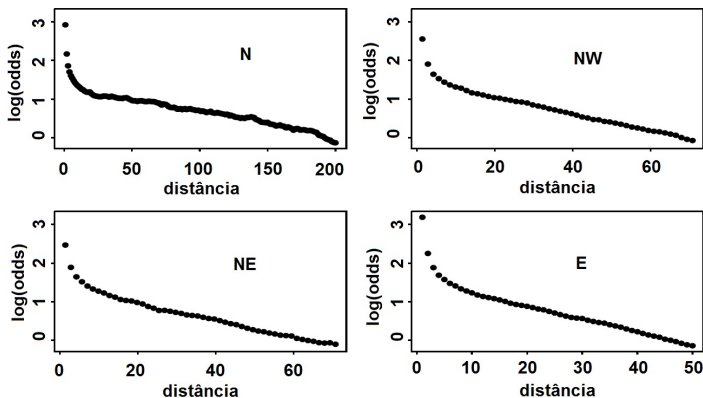


Fig.9 : Versão da autocorrelação amostral para um mapa binário. *Log-odds* para o uso da terra vs. distância em 4 direções (N, NW, NE, E).

Note que o padrão é forte, sendo necessário uma distância de pelo menos 40km antes de cair para aproximadamente 0. Perderíamos pouca informação espacial se trabalharmos com uma resolução menor ($4km \times 4km$).

Suavizadores espaciais.

Um dos objetivos da ilustração gráfica dada por um gráfico *choropleth map* para os Y_i 's é a suavização. Ao escolhermos o número de classes (tonalidades de cores) usadas no gráfico, estamos inserindo a ideia da suavização (não espacial).

A matriz de proximidade W fornece um suavizador espacial. Para suavizar, substitua Y_i por $\hat{Y}_i = \sum_j w_{ij} Y_j / w_{i+}$. Isto assegura que o valor registrado para a unidade de área i será “similar” aos valores de seus vizinhos. Quanto mais vizinhos são usados no cálculo de \hat{Y}_i , maior suavização iremos atingir.

\hat{Y}_i pode ser visto como um suavizador não-usual, pois ignora o valor Y_i observado para a unidade i . Alternativamente, podemos incluir a observação Y_i usando o seguinte suavizador: $\hat{Y}_i^* = (1 - \alpha) Y_i + \alpha \hat{Y}_i$, sendo $0 < \alpha < 1$.

Finalmente, poderíamos tentar uma modelagem suavizadora, isto é, uma suavização baseada em $E(Y_i | \text{Dados})$ que representa a média da distribuição preditiva. Suavizadores surgem então como resultados de modelos espaciais hierárquicos que usamos para explicar Y_i 's.

Campos Aleatórios Markovianos (*MRF - Markov Random Fields*).

Considere $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)'$ e o conjunto $\{p(Y_i | Y_{j \neq i})\}$.

Sabemos que as distribuições condicionais completas em $\{p(Y_i | Y_{j \neq i})\}$ podem ser determinadas a partir da conjunta $p(\mathbf{Y})$.

Questão: Por acaso $\{p(Y_i | Y_{j \neq i})\}$ determina $p(\mathbf{Y})$?

Um resultado técnico que permite recuperar a distribuição conjunta $p(\mathbf{Y})$ a partir das condicionais completas é conhecido como Lema de Brook (*Brook's Lemma*).

Cuidado! Não podemos utilizar um conjunto arbitrário de distribuições condicionais completas para determinar unicamente a conjunta.

Ex.: Sejam $Y_1|Y_2 \sim N(\alpha_0 + \alpha_1 Y_2, \sigma_1^2)$ e $Y_2|Y_1 \sim N(\beta_0 + \beta_1 Y_1^3, \sigma_2^2)$. Veja que:

- $E(Y_1) = E[E(Y_1|Y_2)] = E[\alpha_0 + \alpha_1 Y_2] = \alpha_0 + \alpha_1 E(Y_2)$, ou seja, $E(Y_1)$ e $E(Y_2)$ são linearmente relacionados.
- $E(Y_2) = E[E(Y_2|Y_1)] = E[\beta_0 + \beta_1 Y_1^3] = \beta_0 + \beta_1 E(Y_1^3)$.

Estes dois resultados serão válidos simultaneamente apenas em casos simples [ex.: $E(Y_1) = E(Y_1^3)$]. Muitas vezes teremos $p(Y_1|Y_2)$ e $p(Y_2|Y_1)$ incompatíveis em relação à determinação de $p(Y_1, Y_2)$.

Nosso objetivo aqui não é examinar condições de compatibilidade.

A conjunta $p(\mathbf{Y})$ pode ser imprópria, mesmo que as condicionais completas $p(Y_i|Y_{j \neq i})$ sejam próprias para todo i .

Ex.: Considere $p(Y_1, Y_2) \propto \exp\{-\frac{1}{2}(Y_1 - Y_2)^2\}$; temos $(Y_1|Y_2) \sim N(Y_2, 1)$ (própria) e $(Y_2|Y_1) \sim N(Y_1, 1)$ (própria), mas $p(Y_1, Y_2)$ é imprópria.

Lema de Brook: Considere a seguinte identidade:

$$\begin{aligned} p(Y_1, \dots, Y_n) &= \frac{p(Y_1|Y_2, \dots, Y_n)}{p(Y_{1,0}|Y_1, \dots, Y_n)} \frac{p(Y_2|Y_{1,0}, Y_3, \dots, Y_n)}{p(Y_{2,0}|Y_{1,0}, Y_3, \dots, Y_n)} \\ &\times \dots \times \frac{p(Y_n|Y_{1,0}, \dots, Y_{n-1,0})}{p(Y_{n,0}|Y_{1,0}, \dots, Y_{n-1,0})} p(Y_{1,0}, \dots, Y_{n,0}). \end{aligned}$$

Neste caso, $\mathbf{Y}_0 = (Y_{1,0}, \dots, Y_{n,0})'$ é qualquer ponto fixo no suporte de $p(Y_1, \dots, Y_n)$; portanto, a conjunta $p(Y_1, \dots, Y_n)$ é determinada pelas distribuições condicionais completas, visto que à direita da igualdade [desconsidere a constante $p(Y_{1,0}, \dots, Y_{n,0})$] temos apenas as condicionais completas.

Apenas o núcleo da conjunta é determinado via Lema de Brook.

Falta a constante normalizadora.

- Se $p(\mathbf{Y})$ é imprópria, não poderemos obter a constante normalizadora.
- Se $p(\mathbf{Y})$ é própria, então o fato de integrar 1 determinará essa constante.

Assuma que $y_0 = (y_{10}, y_{20}, y_{30})$ é um ponto fixo qualquer no suporte (domínio) de $p(Y_1, Y_2, Y_3)$.

$$\begin{aligned} p(Y_1, Y_2, Y_3) &= p(Y_1 | Y_2, Y_3) p(Y_2, Y_3) \\ &= p(Y_1 | Y_2, Y_3) \frac{p(Y_1 = y_{10}, Y_2, Y_3)}{p(Y_1 = y_{10} | Y_2, Y_3)} = \frac{p(Y_1 | Y_2, Y_3)}{p(Y_1 = y_{10} | Y_2, Y_3)} p(Y_2, Y_1 = y_{10}, Y_3) \\ &= \frac{p(Y_1 | Y_2, Y_3)}{p(Y_1 = y_{10} | Y_2, Y_3)} p(Y_2 | Y_1 = y_{10}, Y_3) p(Y_1 = y_{10}, Y_3) \\ &= \frac{p(Y_1 | Y_2, Y_3)}{p(Y_1 = y_{10} | Y_2, Y_3)} p(Y_2 | Y_1 = y_{10}, Y_3) \frac{p(Y_1 = y_{10}, Y_2 = y_{20}, Y_3)}{p(Y_2 = y_{20} | Y_1 = y_{10}, Y_3)} \\ &= \frac{p(Y_1 | Y_2, Y_3)}{p(Y_1 = y_{10} | Y_2, Y_3)} \frac{p(Y_2 | Y_1 = y_{10}, Y_3)}{p(Y_2 = y_{20} | Y_1 = y_{10}, Y_3)} p(Y_3, Y_1 = y_{10}, Y_2 = y_{20}) \\ &= \frac{p(Y_1 | Y_2, Y_3)}{p(Y_1 = y_{10} | Y_2, Y_3)} \frac{p(Y_2 | Y_1 = y_{10}, Y_3)}{p(Y_2 = y_{20} | Y_1 = y_{10}, Y_3)} p(Y_3 | Y_1 = y_{10}, Y_2 = y_{20}) p(Y_1 = y_{10}, Y_2 = y_{20}) \\ &= \frac{p(Y_1 | Y_2, Y_3)}{p(Y_1 = y_{10} | Y_2, Y_3)} \frac{p(Y_2 | Y_1 = y_{10}, Y_3)}{p(Y_2 = y_{20} | Y_1 = y_{10}, Y_3)} \frac{p(Y_3 | Y_1 = y_{10}, Y_2 = y_{20})}{p(Y_3 = y_{30} | Y_1 = y_{10}, Y_2 = y_{20})} p(Y_1 = y_{10}, Y_2 = y_{20}, Y_3 = y_{30}) \end{aligned}$$

Quando o número de unidades de área é grande, será complicado escrever a conjunta $p(\mathbf{Y})$. Iremos preferir trabalhar exclusivamente com as n distribuições condicionais completas.

Do ponto de vista espacial, devemos pensar que a distribuição condicional completa de Y_i dependerá apenas dos vizinhos de i . Adotando alguma estrutura de vizinhança (ex.: $w_{ij} = 1$ se i e j são adjacentes, 0 c.c.) denote:

- ∂_i o conjunto dos índices referentes às regiões vizinhas de i .
- $i \sim j$ indica que i é vizinho de j (j é vizinho de i).

Modelagem local:

Considere a seguinte especificação $p(Y_i | Y_{j \neq i}) = p(Y_i | Y_{j \in \partial_i})$.

Pergunta: $\{p(Y_i | Y_{j \in \partial_i})\}$ determina $p(\mathbf{Y})$? Queremos ter certeza de que ao implementarmos um *Gibbs Sampling* para simular realizações de $p(\mathbf{Y})$, existirá de fato um única distribuição estacionária. Resposta = Sim.

O fato de usarmos uma especificação “local” para determinar uma distribuição conjunta “global” é conhecido na literatura como *Markov Random Field* (MRF).

Algumas definições importantes:

- clique: é um conjunto de células (ou índices) tais que cada elemento é vizinho de todos os outros. Em um espaço com n células, dependendo da definição da estrutura de vizinhança, cliques podem ter tamanho $1, 2, \dots, n$.

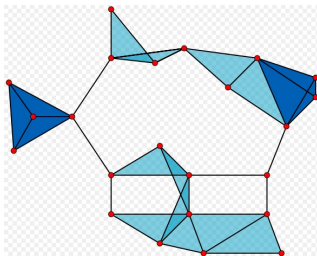


Fig.10 : Grafo com 23 cliques tamanho 1 (vértices em vermelho), 42 cliques tamanho 2 (cada *edge*), 19 cliques tamanho 3 (triângulos em azul claro ou escuro) e 2 cliques tamanho 4 (áreas azul escuro).

Algumas definições importantes (continuação):

- função potencial ou potencial de ordem k : é uma função com k argumentos permutáveis. Os argumentos seriam os valores obtidos de variáveis associadas às células de um “clique” de tamanho k .

Exemplos: Para Y_i contínuo e $k = 2$, um potencial poderia ser

$Q(Y_i, Y_j) = Y_i Y_j$ para o “clique” $\{i, j\}$ de tamanho 2. Para Y_i binário e $k = 2$, poderíamos ter $Q(Y_i, Y_j) = I(Y_i = Y_j) = Y_i Y_j + (1 - Y_i)(1 - Y_j)$.

Na prática, apenas cliques ≤ 2 são considerados.

Distribuição de Gibbs: $p(\mathbf{Y})$ representa uma distribuição de Gibbs se for uma função de Y_i apenas através de potenciais em cliques. Isto é:

$$p(\mathbf{Y}) \propto \exp \left\{ \gamma \sum_k \sum_{\alpha \in \mathcal{M}_k} \phi^{(k)}(Y_{\alpha_1}, Y_{\alpha_2}, \dots, Y_{\alpha_k}) \right\},$$

sendo $\phi^{(k)}$ um potencial de ordem k , \mathcal{M}_k uma coleção de subconjuntos (tamanho k) dos índices $\{1, 2, \dots, n\}$, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)'$ um subconjunto dessa coleção e $\gamma > 0$ um parâmetro de escala.

Teorema de Hammersley-Clifford: Se $\{p(Y_i | Y_{j \in \partial_i})\}$ determina uma única $p(\mathbf{Y})$ (i.e. temos um MRF), então $p(\mathbf{Y})$ terá o formato da distribuição de Gibbs acima.

A prova do Teorema de Hammersley-Clifford pode ser encontrada em Cressie (1993, pag. 417, 418). Sua importância para a modelagem espacial está no fato de reduzir a complexidade das distribuições condicionais completas (serão condicionadas em ∂_i apenas).

Geman e Geman (1984) fornecem o sentido contrário deste teorema (eles iniciam com a distribuição de Gibbs e determinam um MRF). A conclusão é que poderíamos usar as condicionais completas para amostrar da distribuição de Gibbs original (daí o nome “*Gibbs Sampling*”).

Se usarmos apenas cliques de tamanho 1, então Y_i 's são independentes.

Se usarmos cliques de tamanho 2 (comum na prática), teremos $p(\mathbf{Y})$ com um formato de diferença pareada (*pairwise difference*). Para dados contínuos em \mathbb{R} , podemos usar:

$$p(Y_1, \dots, Y_n) \propto \exp \left\{ -\frac{1}{2\tau^2} \sum_{i,j} (Y_i - Y_j)^2 I(i \sim j) \right\}.$$

Note que, temos acima o formato da distribuição de Gibbs com potenciais de ordem 1 e 2: $(Y_i - Y_j)^2 = Y_i^2 - 2Y_i Y_j + Y_j^2$.

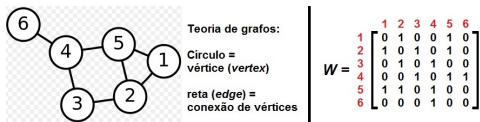
Além disso:

$$p(Y_i | Y_{j \neq i}) = p(Y_i | Y_{j \in \partial_i}) = N \left[\sum_{j \in \partial_i} Y_j / m_i, \tau^2 / m_i \right],$$

sendo m_i o número de vizinhos de i .

Veja que $E(Y_i | Y_{j \in \partial_i})$ é a média dos Y_j 's vizinhos.

Exemplo: Considere novamente o seguinte grafo com 6 regiões:



$$p(Y_1, \dots, Y_6) \propto \exp \left\{ -\frac{1}{2\tau^2} [(Y_1 - Y_2)^2 + (Y_2 - Y_3)^2 + (Y_2 - Y_5)^2 + (Y_3 - Y_4)^2 + (Y_4 - Y_5)^2 + (Y_1 - Y_5)^2 + (Y_4 - Y_6)^2] \right\}$$

$$p(Y_1 | Y_2, \dots, Y_6) \propto \exp \left\{ -\frac{1}{2\tau^2} [(Y_1 - Y_2)^2 + (Y_1 - Y_5)^2] \right\}$$

$$p(Y_1 | Y_2, Y_5) \propto \exp \left\{ -\frac{1}{2\tau^2} [Y_1^2 - 2Y_1 Y_2 + Y_2^2 + Y_1^2 - 2Y_1 Y_5 + Y_5^2] \right\}$$

$$\propto \exp \left\{ -\frac{1}{2\tau^2} [Y_1^2 - 2Y_1 Y_2 + Y_1^2 - 2Y_1 Y_5] \right\}$$

$$\propto \exp \left\{ -\frac{1}{2\tau^2} [2Y_1^2 - 2Y_1(Y_2 + Y_5)] \right\}$$

$$\propto \exp \left\{ -\frac{1}{2\tau^2} 2 \left[Y_1^2 - 2Y_1 \frac{Y_2 + Y_5}{2} \right] \right\}.$$

Temos aqui o núcleo da $N \left[\frac{Y_2 + Y_5}{2}, \frac{\tau^2}{2} \right]$.

Modelo Autoregressivo Condicional (CAR).

Introduzido por Besag (1974), este tipo de modelo pode ser convenientemente empregado no contexto do *Gibbs Sampling* e de outros métodos MCMC para ajustar certas classes de modelos hierárquicos.

Caso Gaussiano (autonormal).

$$(Y_i | Y_{j \neq i}) \sim N \left[\sum_j b_{ij} Y_j, \tau_i^2 \right] \quad \text{para } i = 1, \dots, n.$$

Usando o Lema de Brook, temos: $p(Y_1, \dots, Y_n) \propto \exp \left\{ -\frac{1}{2} \mathbf{Y}' D^{-1} (I - B) \mathbf{Y} \right\}$, sendo $B = \{b_{ij}\}$ (matriz possivelmente não simétrica) e $D = \text{diag}\{\tau_1^2, \dots, \tau_n^2\}$.

Esta expressão sugere a distribuição Normal Multivariada com $\mu_Y = \mathbf{0}$ e matriz de covariâncias $\Sigma_Y = (I - B)^{-1} D$.

Para $D^{-1}(I - B)$ ser simétrica devemos ter $b_{ij}/\tau_i^2 = b_{ji}/\tau_j^2$ para todo i, j .

Exemplo (Caso $n = 2$)

$$B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \quad I - B = \begin{bmatrix} 1 - b_{11} & -b_{12} \\ -b_{21} & 1 - b_{22} \end{bmatrix}$$

$$D = \begin{bmatrix} \tau_1^2 & 0 \\ 0 & \tau_2^2 \end{bmatrix} \quad D^{-1} = \begin{bmatrix} 1/\tau_1^2 & 0 \\ 0 & 1/\tau_2^2 \end{bmatrix}$$

$$D^{-1}(I - B) = \begin{bmatrix} 1/\tau_1^2 & 0 \\ 0 & 1/\tau_2^2 \end{bmatrix} \begin{bmatrix} 1 - b_{11} & -b_{12} \\ -b_{21} & 1 - b_{22} \end{bmatrix} = \begin{bmatrix} \frac{1 - b_{11}}{\tau_1^2} & -\frac{b_{12}}{\tau_1^2} \\ -\frac{b_{21}}{\tau_2^2} & \frac{1 - b_{22}}{\tau_2^2} \end{bmatrix}$$

Se $\tau_1^2 \neq \tau_2^2$, precisamos ter $b_{12} \neq b_{21}$ para que $D^{-1}(I - B)$ seja simétrica.

$$p(Y_1, \dots, Y_n) \propto \exp \left\{ -\frac{1}{2} Y' D^{-1} (I - B) Y \right\} \\ \propto \exp \left\{ -\frac{1}{2} Y' D^{-1} Y \right\} \exp \left\{ +\frac{1}{2} Y' D^{-1} B Y \right\}$$

caso particular ($n=2$)

$$Y' D^{-1} Y = [Y_1 \ Y_2] \begin{bmatrix} 1/\tau_1^2 & 0 \\ 0 & 1/\tau_2^2 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \frac{Y_1^2}{\tau_1^2} + \frac{Y_2^2}{\tau_2^2}$$

Generalizando: $Y' D^{-1} Y = \sum_{i=1}^n \frac{Y_i^2}{\tau_i^2}$

$$Y' D^{-1} B Y = [Y_1 \ Y_2] \begin{bmatrix} \frac{b_{11}}{\tau_1^2} & \frac{b_{12}}{\tau_1^2} \\ \frac{b_{21}}{\tau_2^2} & \frac{b_{22}}{\tau_2^2} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \frac{Y_1^2 b_{11}}{\tau_1^2} + \frac{Y_1 Y_2 b_{21}}{\tau_2^2} + \frac{Y_1 Y_2 b_{12}}{\tau_1^2} + \frac{Y_2^2 b_{22}}{\tau_2^2}$$

Generalizando: $\sum_{i=1}^n \sum_{j=1}^n \frac{Y_i Y_j b_{ij}}{\tau_i^2}$

$$p(y_1, \dots, y_n) \propto \exp \left\{ -\frac{1}{2} \left[\sum_{i=1}^n \frac{y_i^2}{\tau_i^2} - \sum_{i=1}^n \sum_{j=1}^n \frac{y_i y_j b_{ij}}{\tau_i^2} \right] \right\}$$

Condicional completa

$$p(y_i | y_{j \neq i}) \propto \exp \left\{ -\frac{1}{2} \left[\frac{y_i^2}{\tau_i^2} - \sum_{j=1}^n y_i y_j b_{ij} / \tau_i^2 \right] \right\}$$

$$\propto \exp \left\{ -\frac{1}{2\tau_i^2} \left[y_i^2 - y_i \sum_{j=1}^n y_j b_{ij} \right] \right\}$$

se fizermos $b_{ij}^* = \frac{1}{2} b_{ij}$ teremos

$$\propto \exp \left\{ -\frac{1}{2\tau_i^2} \left[y_i^2 - 2 y_i \sum_{j=1}^n y_j b_{ij}^* \right] \right\}$$

Núcleo da $N \left[\sum_{j=1}^n y_j b_{ij}^*, \tau_i^2 \right]$

Retornando a W (matriz de proximidade), seja $b_{ij} = w_{ij}/w_{i+}$ e $\tau_i^2 = \tau^2/w_{i+}$ então

$$p(Y_1, \dots, Y_n) \propto \exp \left\{ -\frac{1}{2\tau^2} \mathbf{Y}'(D_w - W)\mathbf{Y} \right\}$$

sendo $D_w = \text{diag}\{w_{1+}, \dots, w_{n+}\}$. Podemos reescrever:

$$p(Y_1, \dots, Y_n) \propto \exp \left\{ -\frac{1}{2\tau^2} \sum_{i \neq j} w_{ij} (Y_i - Y_j)^2 \right\}.$$

Esta especificação é denominada Modelo Autoregressivo Intrínseco (IAR).

Esta formulação é imprópria visto que $(D_w - W) \mathbf{1} = \mathbf{0}$. Temos um exemplo de uma distribuição conjunta imprópria, mas com condicionais completas próprias.

Os Y_i 's não estão centrados em nenhum ponto do espaço; veja que, para uma constante C , na expressão acima temos $p(Y_1, \dots, Y_n) = p(Y_1 + C, \dots, Y_n + C)$. A restrição $\sum_i Y_i = 0$ resolverá este problema de centralização.

$$Y'(D_W - W)Y$$

Caso particular $n=2$

$$\begin{bmatrix} Y_1 & Y_2 \end{bmatrix} \begin{bmatrix} W_{1+} - W_{11} & -W_{12} \\ -W_{21} & W_{2+} - W_{22} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} =$$

$$\begin{bmatrix} Y_1(W_{1+} - W_{11}) - Y_2 W_{21} & Y_2(W_{2+} - W_{22}) - Y_1 W_{12} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} =$$

$$Y_1^2(W_{1+} - W_{11}) - Y_2 Y_1 W_{21} + Y_2^2(W_{2+} - W_{22}) - Y_1 Y_2 W_{12} =$$

$\underbrace{\quad}_{\rightarrow = 0} \quad \underbrace{\quad}_{\rightarrow = 0}$
 $\rightarrow W_{11} + W_{12} = 0 + W_{12} = W_{12} \quad \rightarrow W_{22} + W_{21} = 0 + W_{21} = W_{12}$

$$= W_{12}(Y_1^2 - 2Y_1Y_2 + Y_2^2) = W_{12}(Y_1 - Y_2)^2$$

A conjunta $p(\mathbf{Y})$ não pode ser usada como um modelo para dados; dados não surgem sob um mecanismo estocástico impróprio. Além disso, não podemos impor uma centralização para realizações de medidas aleatórias. O modelo autonormal impróprio deve ser tratado como uma especificação de distribuição *a priori* para efeitos aleatórios espaciais inseridos no 2º nível da modelagem hierárquica.

Estratégia simples para tornar $p(\mathbf{Y})$ própria: Substitua $\Sigma_Y^{-1} = D_w - W$ por $D_w - \rho W$, escolhendo ρ de forma que Σ_Y exista. Isto implica que:

$$Y_i | Y_{j \neq i} \sim N \left(\rho \sum_j \frac{w_{ij} Y_j}{w_{i+}}, \frac{\tau^2}{w_{i+}} \right),$$

logo estamos modelando $E(Y_i | Y_{j \neq i}) =$ proporção da média dos valores vizinhos.

Veja que: $\Sigma_Y^{-1} = D_w - \rho W = D_w^{1/2} (I - \rho D_w^{-1/2} W D_w^{-1/2}) D_w^{1/2}$, ou seja, a escolha de ρ está ligada à matriz $D_w^{-1/2} W D_w^{-1/2}$.

Σ_Y^{-1} é não-singular se $\rho \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$ sendo $\lambda_{(1)} < \lambda_{(2)} < \dots < \lambda_{(n)}$ os autovalores ordenados de $D_w^{-1/2} W D_w^{-1/2}$. Visto que $\text{tr}(D_w^{-1/2} W D_w^{-1/2}) = 0 = \sum_{i=1}^n \lambda_{(i)}$, temos $\lambda_{(1)} < 0$, $\lambda_{(n)} > 0$ e $0 \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$ (script R).

Vantagens de ρ :

- Torna a distribuição própria;
- Adiciona flexibilidade paramétrica;
- $\rho = 0$ implica em Y_i 's independentes $N(0, \tau^2/w_{i+})$.
Sem ρ não teríamos esta interpretação.

Desvantagens de ρ :

- Seria razoável esperar que Y_i é a proporção da média dos valores vizinhos?
- ρ pode ser enganoso quanto a força da associação espacial.
Banerjee et al. (2004) menciona um estudo simulado para calibrar ρ em relação à medida de associação espacial *I de Moran*. Para $\tau^2 = 1$ fixo, simulações em um grid 10×10 e usando vizinhança de 1ª ordem:
quando $\rho = 0.80$, obteve-se $0.1 \leq I \leq 0.15$,
quando $\rho = 0.90$, obteve-se $0.2 \leq I \leq 0.25$,
quando $\rho = 0.99$, obteve-se $I \leq 0.50$; ou seja, ρ alto e I baixo.
No contexto Bayesiano, uma distribuição *a priori* para ρ indicando associação espacial deverá concentrar massa de probabilidade próximo de 1.

Se $\rho = 0$, temos Y_i 's independentes $N(0, \tau^2/w_{i+})$; isso sugere que τ^2 não deve ser visto como uma quantificação da associação espacial. Se c é uma constante, temos $cY_i \sim N(0, c^2\tau^2/w_{i+})$; ou seja, τ^2 se torna $c^2\tau^2$, mas a associação espacial entre os Y_i 's permanece a mesma.

Se $p(\mathbf{Y})$ é própria, a variedade de padrões espaciais fica limitada. No caso em que o modelo CAR é assumido para efeitos aleatórios, a especificação imprópria pode permitir mais opções de padrões espaciais *a posteriori*. Podemos aplicar o modelo CAR com a especificação própria ou imprópria.

Limites mais simples para ρ podem ser obtidos se substituirmos a matriz de vizinhança W por $\tilde{W} \equiv \text{diag}(1/w_{1+}, \dots, 1/w_{n+})W$. A matriz \tilde{W} é estocástica nas linhas (linhas somam 1); ela não é simétrica. Poderemos escrever:

$$\Sigma_Y^{-1} = M^{-1}(I - \alpha\tilde{W}) \quad \text{ou} \quad \Sigma_Y = (I - \alpha\tilde{W})^{-1}M$$

sendo M diagonal. Neste caso, $I - \alpha\tilde{W}$ é não-singular se $|\alpha| < 1$ (mais detalhes serão vistos na descrição do modelo SAR).

Considere novamente a seguinte expressão do modelo autonormal:

$$p(Y_1, \dots, Y_n) \propto \exp \left\{ -\frac{1}{2} \mathbf{Y}' D^{-1} (I - B) \mathbf{Y} \right\}.$$

Podemos representar o sistema de variáveis aleatórias como: $\mathbf{Y} = B\mathbf{Y} + \epsilon$ ou equivalentemente $(I - B)\mathbf{Y} = \epsilon$.

Uma distribuição para \mathbf{Y} induz uma distribuição para ϵ .

Se $p(\mathbf{Y})$ é própria, temos (acima) $\mathbf{Y} \sim N[\mathbf{0}, (I - B)^{-1}D]$ e então

$$\epsilon = (I - B)\mathbf{Y} \sim N[(I - B)\mathbf{0}, (I - B)(I - B)^{-1}D(I - B)'] = N[\mathbf{0}, D(I - B)'],$$

isto é, os componentes de ϵ não são independentes.

Além disso,

$$\text{Cov}(\epsilon, \mathbf{Y}) = \text{Cov}[(I - B)\mathbf{Y}, \mathbf{Y}] = (I - B)\text{Cov}(\mathbf{Y}, \mathbf{Y}) = (I - B)(I - B)^{-1}D = D.$$

O modelo CAR especifica Σ_Y^{-1} e não Σ_Y (não modela diretamente a associação). Quando $p(\mathbf{Y})$ é própria, podemos recorrer a resultados teóricos da Normal Multivariada para interpretar as entradas em Σ_Y^{-1} .

- Se $\Sigma_Y^{-1} = D^{-1}(I - B)$, temos na diagonal $(\Sigma_Y^{-1})_{ii} = 1/\tau_i^2$.
- $\text{Var}(Y_i | Y_{j \neq i}) = 1/(\Sigma_Y^{-1})_{ii}$.
- Se $(\Sigma_Y^{-1})_{ij} = 0$, então Y_i e Y_j são condicionalmente independentes dado $Y_{k \neq i, j}$.

$b_{ij} = 0 \Rightarrow$ independência condicional entre Y_i e Y_j .

A escolha da estrutura de vizinhança (ex.: w_{ij} em b_{ij}) implica em uma coleção de independências condicionais.

Assumindo estrutura de vizinhança de 1ª ordem estaremos trabalhando com propriedades locais Markovianas.

Podemos introduzir covariáveis (termo $\mathbf{X}\beta$) no modelo CAR. Condicional em $\beta = (\beta_0, \beta_1, \dots, \beta_d)$, esse termo irá alterar apenas a estrutura da média. Podemos usar o modelo autonormal CAR como distribuição para efeitos aleatórios espaciais; estes efeitos são adicionados à estrutura de regressão para a média $\mathbf{X}\beta$.

Modelo autonormal (com covariáveis):

$$p(\mathbf{Y}) \propto \exp \left\{ -\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)' D^{-1} (I - B)(\mathbf{Y} - \mathbf{X}\beta)' \right\}.$$

Representação do sistema de variáveis aleatórias: $\mathbf{Y} = B\mathbf{Y} + (I - B)\mathbf{X}\beta + \epsilon$.
Equivalentemente, $(I - B)\mathbf{Y} = (I - B)\mathbf{X}\beta + \epsilon$ então $(I - B)(\mathbf{Y} - \mathbf{X}\beta) = \epsilon$.

Se $p(\mathbf{Y})$ é própria, temos

$$\mathbf{Y} \sim N[\mathbf{X}\beta, (I - B)^{-1}D]$$

$$\mathbf{Y} - \mathbf{X}\beta \sim N[\mathbf{0}, (I - B)^{-1}D]$$

$$\epsilon = (I - B)(\mathbf{Y} - \mathbf{X}\beta) \sim N[(I - B)\mathbf{0}, (I - B)(I - B)^{-1}D(I - B)'] = N[\mathbf{0}, D(I - B)'],$$

$$\text{Cov}(\epsilon, \mathbf{Y}) = \text{Cov}[(I - B)(\mathbf{Y} - \mathbf{X}\beta), \mathbf{Y}] = (I - B)\text{Cov}(\mathbf{Y} - \mathbf{X}\beta, \mathbf{Y}) = (I - B)\text{Cov}(\mathbf{Y}, \mathbf{Y}) = D.$$

- Um modelo CAR próprio pode ser usado para dados pontuais:

Faça w_{ij} = inversa da distância entre os pontos i e j .

Entretanto, a previsão espacial (antes desenvolvida via *Kriging*) se torna *ad hoc* através do CAR próprio; isto é, para prever em um novo sítio Y_0 precisamos especificar uma Normal para $(Y_0|\mathbf{Y})$.

$$(Y_0|Y_1, \dots, Y_n) \sim N \left[\rho \sum_j w_{0j} Y_j / w_{0+}, \tau^2 / w_{0+} \right].$$

Esta distribuição é bem definida e pode ser usada para determinar a conjunta (própria) $p(Y_0, Y_1, \dots, Y_n)$ que não será a distribuição CAR (imprópria) obtida pela especificação das condicionais completas (sem ρ) para Y_0, Y_1, \dots, Y_n via Lema de Brook.

Caso não-Gaussiano.

Em muitos casos a distribuição Normal não será apropriada aos dados
Exemplo: dados binários ou contagens.

Nosso ponto de partida será modelar utilizando alguma distribuição membro da família exponencial. Veja que a Normal é membro da família exponencial; portanto, iremos explorar propriedades observadas na modelagem Gaussiana que são comuns a outros integrantes da família exponencial.

Uma família de p.m.f.'s ou p.d.f.'s é chamada de família exponencial se ela pode ser expressa como segue:

$$\begin{aligned} p(Y_i|\theta_i) &= G(Y_i) H(\theta_i) \exp\{A(\theta_i) B(Y_i)\} \\ &= \exp\{A(\theta_i) B(Y_i) + \log[G(Y_i)] + \log[H(\theta_i)]\} \\ &= \exp\{A(\theta_i) B(Y_i) + C(Y_i) + D(\theta_i)\} \end{aligned}$$

Lembrete: no caso autonormal, desejamos propor $p(Y_i|Y_{j \neq i})$ com formato de diferenças pareadas. Considere o seguinte resultado apresentado na Página 420 de Cressie (1993):

Teorema (Besag, 1974): Assumindo o formato da família exponencial (rever slide), teremos a configuração de diferenças pareadas se:

$$A(\theta_i) = \mu_i + \sum_{j=1}^n c_{ij} B(Y_j), \quad \text{para } i = 1, \dots, n,$$

sendo c_{ij} tal que: $c_{ij} = c_{ji}$, $c_{ii} = 0$ e $c_{ij} = 0$ se $j \notin \partial_i$.

- Exemplo 1 (CAR Gaussiano, Lembrete $(Y_i | Y_{j \neq i}) \sim N[\sum_j b_{ij} Y_j, \tau_i^2]$):

$$\begin{aligned} p(Y_i | Y_{j \neq i}) &= (2\pi\tau_i^2)^{-1/2} \exp \left\{ -\frac{1}{2\tau_i^2} [Y_i - \theta_i]^2 \right\} \\ &= (2\pi\tau_i^2)^{-1/2} \exp \left\{ -\frac{1}{2\tau_i^2} [Y_i^2 - 2Y_i\theta_i + \theta_i^2] \right\} \\ &= \exp \left\{ -\frac{1}{2} \log(2\pi\tau_i^2) - \frac{Y_i^2}{2\tau_i^2} + \frac{Y_i\theta_i}{\tau_i^2} - \frac{\theta_i^2}{2\tau_i^2} \right\} \end{aligned}$$

sendo: $A(\theta_i) = \theta_i$ (link canônico), $B(Y_i) = Y_i/\tau_i^2$,
 $C(Y_i) = -\log(2\pi\tau_i^2)/2 - Y_i^2/(2\tau_i^2)$ e $D(\theta_i) = -\frac{\theta_i^2}{2\tau_i^2}$.

Reveja a conjunta $p(\mathbf{Y})$ e note que $\mu_i = 0 \forall i$.

De acordo com o Teorema de Besag (1974),

$$A(\theta_i) = \theta_i = \mu_i + \sum_j c_{ij} Y_j / \tau_j^2 = 0 + \sum_j (c_{ij} / \tau_j^2) Y_j = \sum_j b_{ij} Y_j.$$

- Exemplo 2 (CAR Poisson, dados de contagem):

$$\begin{aligned} p(Y_i | Y_{j \neq i}) &= \exp\{-\theta_i\} \theta_i^{Y_i} / (Y_i!) \\ &= \exp\{-\theta_i + Y_i \log(\theta_i) - \log(Y_i!)\} \end{aligned}$$

sendo $A(\theta_i) = \log(\theta_i)$ (log link), $B(Y_i) = Y_i$, $C(Y_i) = -\log(Y_i!)$ e $D(\theta_i) = -\theta_i$.

De acordo com o Teorema de Besag (1974),

$$\exp\{A(\theta_i)\} = \theta_i = \exp\left\{\mu_i + \sum_j c_{ij} Y_j\right\} = \exp\left\{\mu_i + \psi \sum_j b_{ij} Y_j\right\}.$$

sendo $\psi > 0$ um parâmetro de dispersão.

- Exemplo 3 (CAR Bernoulli, dados binários, Modelo autologístico): $(Y_i | Y_{j \neq i}) \sim \text{Bernoulli}(\theta_i)$ que é da família exponencial.

$$\begin{aligned} p(Y_i | Y_{j \neq i}) &= \theta_i^{Y_i} (1 - \theta_i)^{1 - Y_i} \\ &= \exp \{ Y_i \log(\theta_i) + \log(1 - \theta_i) - Y_i \log(1 - \theta_i) \} \\ &= \exp \left\{ Y_i \log \left(\frac{\theta_i}{1 - \theta_i} \right) + \log(1 - \theta_i) \right\} \end{aligned}$$

sendo $A(\theta_i) = \log \left(\frac{\theta_i}{1 - \theta_i} \right)$ (link logístico), $B(Y_i) = Y_i$, $C(Y_i) = 0$ e $D(\theta_i) = \log(1 - \theta_i)$.

De acordo com o Teorema de Besag (1974),

$$\begin{aligned} \frac{p(Y_i = 1 | Y_{j \neq i})}{p(Y_i = 0 | Y_{j \neq i})} &= \frac{\theta_i}{1 - \theta_i} = \exp \{ A(\theta_i) \} \\ &= \exp \left\{ \mu_i + \sum_j c_{ij} Y_j \right\} = \exp \left\{ \mu_i + \psi \sum_j b_{ij} Y_j \right\}. \end{aligned}$$

sendo $\psi > 0$ um parâmetro de dispersão.

Função logística: $p(Y_i = 1 | Y_{j \neq i}) = \theta_i = \frac{\exp \{ A(\theta_i) \}}{1 + \exp \{ A(\theta_i) \}}$

Quando covariáveis estão disponíveis, podemos introduzir um componente não-autoregressivo de regressão linear. Faça $\mu_i = \mathbf{X}_i' \boldsymbol{\beta}$ para algum conjunto de d covariáveis $\mathbf{X}_i = (1, X_{1i}, \dots, X_{di})'$.

No modelo autologístico:

$\log \frac{P(Y_i=1)}{P(Y_i=0)} = \mathbf{X}_i' \boldsymbol{\beta} + \psi \sum_j w_{ij} Y_j$, sendo $w_{ij} = 1$ se $i \sim j$ (0 c.c.).

$$\begin{aligned} p(Y_i | Y_{j \neq i}) &= \exp \left\{ Y_i \log \left(\frac{\theta_i}{1 - \theta_i} \right) + \log(1 - \theta_i) \right\} \\ &\propto \exp \left\{ Y_i \log \left(\frac{\theta_i}{1 - \theta_i} \right) \right\} \\ &\propto \exp \left\{ Y_i \left(\mathbf{X}_i' \boldsymbol{\beta} + \psi \sum_j w_{ij} Y_j \right) \right\} \\ &\propto \exp \left\{ Y_i \mathbf{X}_i' \boldsymbol{\beta} + \psi \sum_j w_{ij} Y_i Y_j \right\} \end{aligned}$$

Usando o Lema de Brook: $p(Y_1, \dots, Y_n) \propto \exp \left\{ \boldsymbol{\beta}' (\sum_i Y_i \mathbf{X}_i) + \psi \sum_i \sum_j w_{ij} Y_i Y_j \right\}$.

A última expressão tem formato da distribuição de Gibbs e é atraente em termos de modelagem. Entretanto, para inferência via verossimilhança ou Bayesiana, a constante normalizadora é necessária, visto que esta será uma função de β e ψ .

O cálculo dessa contante normalizadora requer uma soma para todos os 2^n possíveis valores que (Y_1, \dots, Y_n) pode assumir. Mesmo para tamanhos amostrais moderados, isto representará um desafio computacional.

O caso em que Y_i assume uma de várias categorias representa uma extensão natural do modelo autológicoístico binário. Sejam $1 = 1, \dots, L$ índices representando as L possíveis categorias, podemos definir:

$$P(Y_i = 1 | Y_{j \neq i}) \propto \exp \left\{ \psi \sum_{j \neq i} w_{ij} I(Y_j = 1) \right\} \text{ sendo } w_{ij} = 1 \text{ se } i \sim j \text{ (0 c.c.)}.$$

A distribuição resultante acima é denominada "*Potts Model*"; ela também terá o problema da constante normalizadora. Podemos usar este modelo para efeitos aleatórios (alternativa ao modelo autonormal).

Modelo Autoregressivo Simultâneo (SAR).

Considere novamente a especificação: $\mathbf{Y} = B\mathbf{Y} + \epsilon$ equivalente a $(I - B)\mathbf{Y} = \epsilon$.

Ao invés de “ \mathbf{Y} induz a distribuição de ϵ ”, faremos “ ϵ induz a distribuição de \mathbf{Y} ”.

Assim como na modelagem autoregressiva de séries temporais, assuma inovações ϵ_i independentes.

Generalizando: $\epsilon \sim N(\mathbf{0}, \tilde{D})$ sendo $\tilde{D} = \text{diag}\{\sigma_1^2, \dots, \sigma_n^2\}$.

Ressaltando que:

- \tilde{D} não tem conexão com D do modelo CAR;
- a matriz B (a seguir) pode ser ou não a mesma do modelo CAR.

No modelo SAR temos: $Y_i = \sum_j b_{ij} Y_j + \epsilon_i$ com $\epsilon_i \sim N(0, \sigma_i^2)$, $i = 1, \dots, n$.

Se $(I - B)$ possui posto completo (*full rank*), então esta matriz será inversível e $p(\mathbf{Y})$ será própria.

Veja que: $(I - B)\mathbf{Y} = \epsilon$ logo $\mathbf{Y} = (I - B)^{-1}\epsilon$.

$\epsilon \sim N_n(\mathbf{0}, \tilde{D})$ então $\mathbf{Y} \sim N\left[\mathbf{0}, (I - B)^{-1}\tilde{D}[(I - B)^{-1}]'\right]$.

Neste caso:

$$\begin{aligned} \text{Cov}(\epsilon, \mathbf{Y}) &= \text{Cov}[(I - B)\mathbf{Y}, \mathbf{Y}] = (I - B)\text{Cov}(\mathbf{Y}, \mathbf{Y}) \\ &= (I - B)(I - B)^{-1}\tilde{D}[(I - B)^{-1}]' = \tilde{D}[(I - B)']^{-1}. \end{aligned}$$

Se $\tilde{D} = \sigma^2 I_n$, teremos

$$\begin{aligned} \mathbf{Y} &\sim N\left[\mathbf{0}, \sigma^2(I - B)^{-1}I_n[(I - B)^{-1}]'\right] \\ &= N\left[\mathbf{0}, \sigma^2[(I - B)]^{-1}[(I - B)']^{-1}\right] \\ &= N\left[\mathbf{0}, \sigma^2[(I - B)'(I - B)]^{-1}\right] \end{aligned}$$

Duas escolhas de B são frequentemente discutidas na literatura:

- $B = \rho W$, sendo W uma matriz de proximidade ($w_{ij} = 1$ se $i \sim j$, 0 c.c.). O parâmetro ρ é um parâmetro espacial autoregressivo e temos $Y_i = \rho \sum_j Y_j I(j \in \partial_i) + \epsilon_i$. Qualquer matriz de proximidade pode ser usada, lembrando que $(I - B) = (I - \rho W)$ será não-singular se $\rho \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$ sendo $\lambda_{(1)} < \dots < \lambda_{(n)}$ os autovalores ordenados de W .
- $B = \alpha \tilde{W}$, sendo que em \tilde{W} a linha i de W foi normalizada para somar 1; $\tilde{w}_{ij} = w_{ij}/w_{i+}$ e \tilde{W} não é simétrica, mas temos $\tilde{W}\mathbf{1} = \mathbf{1}$. O parâmetro α é dito “parâmetro de autocorrelação espacial” e temos $Y_i = \alpha \sum_j Y_j I(j \in \partial_i)/w_{i+} + \epsilon_i$. Na situação de grid regular, w_{i+} são todos iguais $\Rightarrow \alpha$ múltiplo de ρ . Com $\tilde{W}\mathbf{1} = \mathbf{1}$, teremos autovalores tais que $\max|\lambda_i| = 1$; portanto, $(I - B) = (I - \alpha \tilde{W})$ será não singular se $\alpha \in (-1, 1)$, justificando a nomenclatura “parâmetro de autocorrelação”.

Um modelo SAR é usualmente empregado no contexto de regressão; isto é, esse modelo é assumido para os resíduos $\mathbf{U} = \mathbf{Y} - \mathbf{X}\beta$ ao invés dos dados \mathbf{Y} . Se $\mathbf{U} = B\mathbf{U} + \epsilon$ obtemos:

$$\mathbf{Y} = B\mathbf{Y} + (I - B)\mathbf{X}\beta + \epsilon.$$

A expressão acima mostra que \mathbf{Y} é modelado através de um componente “ $B\mathbf{Y}$ ” que estabelece pesos espaciais de vizinhos e um componente “ $(I - B)\mathbf{X}\beta$ ” de regressão linear. Se B é uma matriz de zeros, temos a regressão via Mínimos Quadrados Ordinários. Se $B = I$ obtemos um modelo puramente espacial.

A função log-verossimilhança associada ao modelo SAR (com $\tilde{D} = \sigma^2 I$):

$$\log[L(\beta, \sigma^2; \mathbf{X}, \mathbf{Y})] \propto \log |\sigma^{-1}(I - B)| - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\beta)'(I - B)'(I - B)(\mathbf{Y} - \mathbf{X}\beta).$$

Veja que, se $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \Sigma)$ então:

$$\begin{aligned} p(\mathbf{Y}) &= (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{Y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \right\} \\ \log[p(\mathbf{Y})] &= -\frac{n}{2} \log(2\pi) + \log(|\Sigma|^{-1/2}) - \frac{1}{2} (\mathbf{Y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \end{aligned}$$

SAR com $\tilde{D} = \sigma^2 I_n$: $\Sigma = \sigma^2 [(I - B)'(I - B)]^{-1}$ e $\Sigma^{-1} = \frac{1}{\sigma^2} (I - B)'(I - B)$.

Algumas propriedades de determinantes: $|A^{-1}| = |A|^{-1}$, $|A'| = |A|$ e $|AB| = |A| |B|$.

$$\begin{aligned} |\Sigma|^{-1/2} &= |\sigma^2 [(I - B)'(I - B)]^{-1}|^{-1/2} = \left| \frac{1}{\sigma^2} (I - B)'(I - B) \right|^{1/2} \\ &= |\sigma^{-1}(I - B)|^{1/2} |\sigma^{-1}(I - B)|^{1/2} = |\sigma^{-1}(I - B)| \end{aligned}$$

Apesar de B introduzir um parâmetro de regressão/autocorrelação a forma quadrática na expressão da verossimilhança é rápida de calcular (não requer inversão) e o determinante pode ser resolvido rapidamente usando aproximação via matriz esparsa com diagonal dominante. Portanto, a maximização desta verossimilhança pode ser feita iterativamente (resultado é eficiente).

Podemos também usar uma modelagem Bayesiana com distribuições *a priori* apropriadas. A ausência de um formato hierárquico com efeitos aleatórios implica em um ajuste Bayesiano direto. Entretanto, se introduzirmos efeitos aleatórios no SAR (ao contrário do CAR, onde efeitos aleatórios são definidos via condicionais completas) as correspondentes condicionais completas não terão formatos convenientes; para n grande, o custo computacional será caro.

Note que:

- Modelo CAR: $p(\mathbf{Y}) \propto \exp \left\{ -\frac{1}{2} \mathbf{Y}' D^{-1} (I - B) \mathbf{Y} \right\}.$
- Modelo SAR: $p(\mathbf{Y}) \propto \exp \left\{ -\frac{1}{2} \mathbf{Y}' \left[(I - \tilde{B})^{-1} \tilde{D} [(I - \tilde{B})^{-1}]' \right]^{-1} \mathbf{Y} \right\}.$

Estas duas formas são equivalentes se e somente se

$$(I - B)^{-1} D = (I - \tilde{B})^{-1} \tilde{D} [(I - \tilde{B})^{-1}]',$$

Cressie (1993) mostra que qualquer modelo SAR pode ser representado como um modelo CAR (visto que D é diagonal). O contrário não é verdade.

Modelo Autorregressivo Espaço-Temporal (STAR).

A ideia é trabalhar com matrizes de proximidade definindo vizinhos no espaço e no tempo.

Para ilustrar, considere 9 unidades de área ($i = 1, \dots, 9$), 3 unidades de tempo ($t = 1, 2, 3$). As medidas Y_{it} são espacialmente associadas em cada tempo fixo t . Além disso, podemos querer associar Y_{i2} com Y_{i1} e Y_{i3} .

Seja $Y = (Y_{11}, \dots, Y_{91}, Y_{12}, \dots, Y_{92}, Y_{13}, \dots, Y_{93})'$.
Considere

$$W_1 = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Matriz de proximidade espacial para Y_{it} 's:

$$W_S = \begin{pmatrix} W_1 & 0 & 0 \\ 0 & W_1 & 0 \\ 0 & 0 & W_1 \end{pmatrix}$$

Matriz de proximidade temporal para os Y_{it} 's:

$$W_T = \begin{pmatrix} 0 & W_2 & 0 \\ W_2 & 0 & W_2 \\ 0 & W_2 & 0 \end{pmatrix}, \quad \text{sendo } W_2 = I_9.$$

No modelo SAR, introduzimos associações no espaço e no tempo definindo $B = \rho_S W_S + \rho_T W_T$ ou $B = \rho_{ST} W_S W_T$ sendo:

$$W_S W_T = \begin{pmatrix} 0 & W_1 & 0 \\ W_1 & 0 & W_1 \\ 0 & W_1 & 0 \end{pmatrix}$$

Por exemplo, de acordo com $W_S W_T$, Y_{21} e Y_{41} afetam a média de Y_{12} . De acordo com W_S , Y_{21} e Y_{41} afetam a média de Y_{11} . Muitas outras possibilidades existem...