Revista Brasileira de Estatística(2002)

Análise de Conglomerados Espaciais Via Árvore Geradora Mínima

Renato M. Assunção; Juliano P. Lage; Edna A. Reis

ABRIL/2010

SUMÁRIO

INTRODUÇÃO

A **estatística espacial** consiste em realizar uma análise de conglomerados quando esses objetos possuem uma localização espacial. Para cada área tem-se as medidas de um conjunto de variáveis de interesse que formam o seu perfil. Deseja-se reunir essas pequenas áreas em regiões diferentes que atendam a duas condições simultâneamente:

- áreas de uma mesma região devem ser similares com relação as variáveis do perfil.
- áreas de regiões diferentes devem ser dissimilares

OBJETIVO

O autor se baseou no trabalho de Maravalle e Simone(1995), transformando o mapa num **grafo** e reduzindo-o a uma Árvore Geradora. O principal foco buscado pelo autor, foi a partir da Árvore Geradora Mínima (AGM), particionar sucessivamente a árvore geradora para obter a **regionalização**.

CONGLOMERADOS ESPACIAIS

Considerando **n** áreas (municípios). A cada área i, sendo i=1,2,...,n tem um vetor $x_i = x_{i1}, x_{i2}, \ldots, x_{im}$ de m características quantitativas, constituindo o perfil da área.

Duas áreas são consideradas vizinhas quando possuem uma fronteira em comum.

Um conglomerado é qualquer subconjunto de área.

Uma região está particionada em **conglomerados espaciais** quando as áreas que formam esta região, estiverem agrupadas em conglomerados disjutos e conectados.

Em uma **Árvore geradora**, quaisquer dois nós são unidos por um único caminho.

O que é uma Árvore Geradora Mínima?

- Possui n nós
- Possui n-1 arestas
- Ao cortar 1 aresta qualquer, temos dois subgrupos.

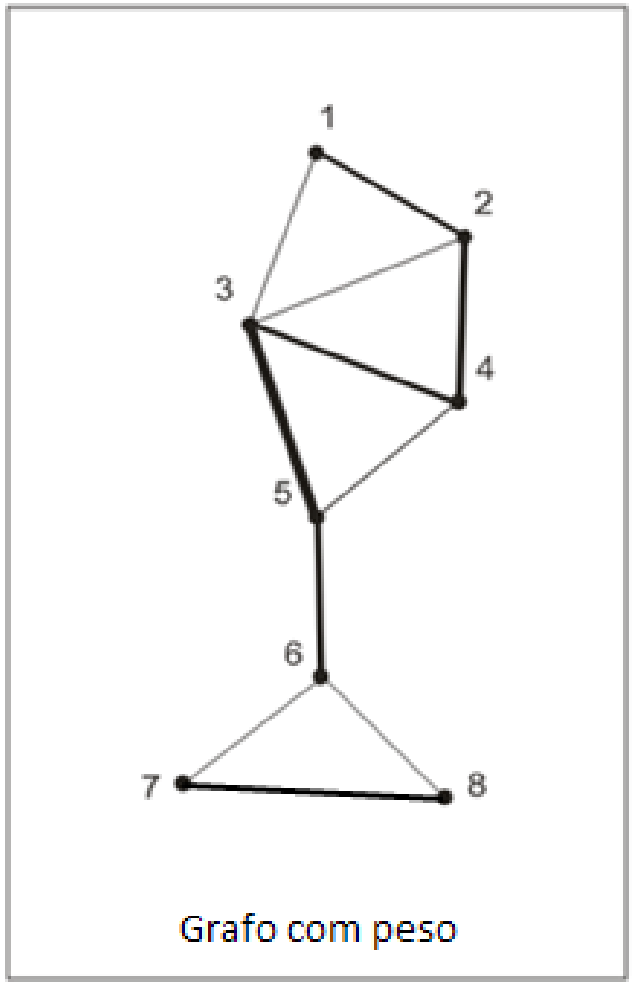
A cada aresta é associado um custo, em que está medindo o grau de dissimilaridade entre duas áreas vizinhas. Este custo pode ser obtido utilizando a **Distância Euclidiana ou a Distância de Marralanobis**.

Distância Euclidiana

$$Custo(i,k) = \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2}$$

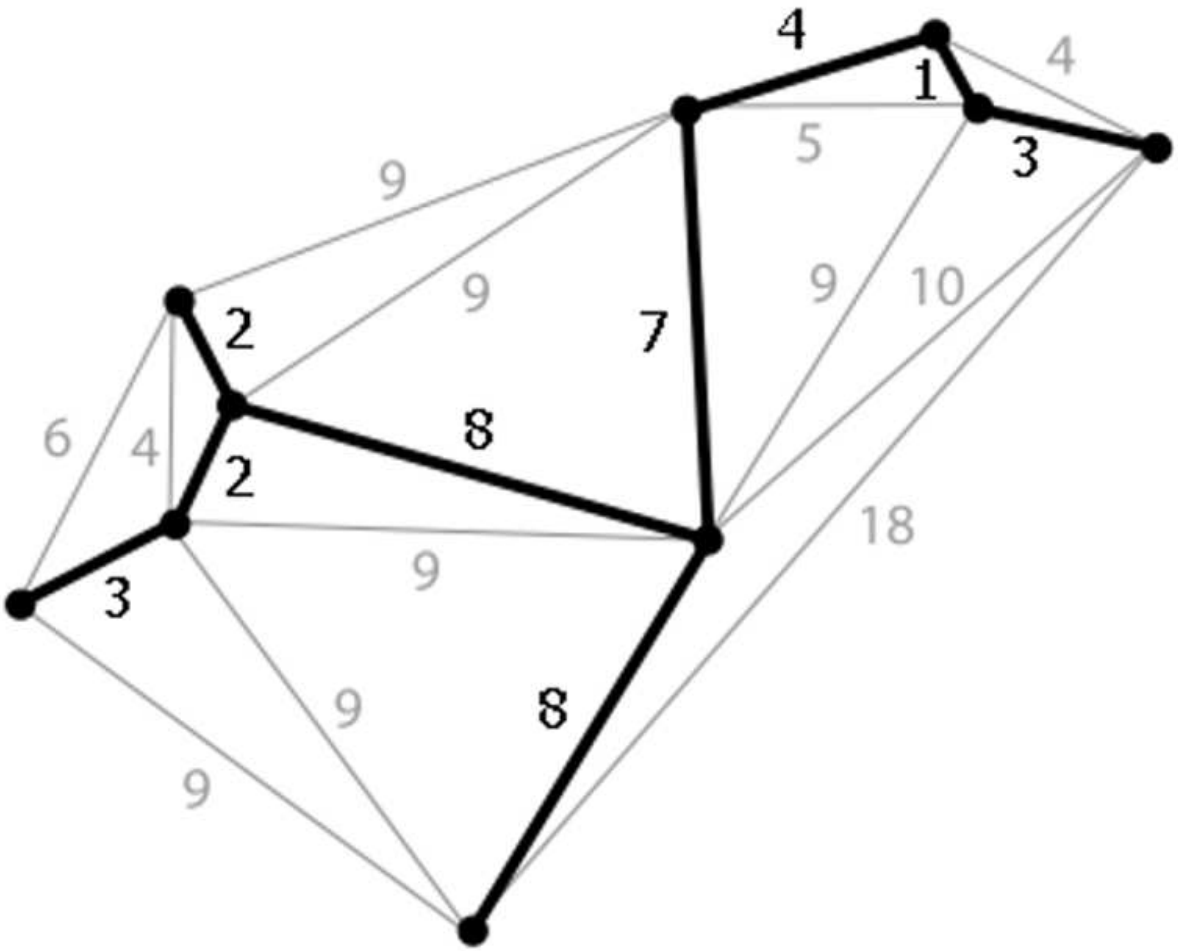
O MÉTODO DA ÁRVORE GERADORA MÍNIMA

A idéia é simplificar o grafo para ficar o número mínimo de arestas. Isto é obtido, quando excluído os grafos de maior custo.



Algorítimo de prim para AGM

Algorítimo de Prim foi o método utilizado pelos autores, afim de construir uma AGM. Pois é possível obter um grafo e possuir custos associados a este grafo.



Critérios de poda da AGM

Após criada a AGM, o passo seguinte será particiona-la, pois, assim obtem os conglomerados espaciais.

- Verifica-se as n arestas da árvore
- Apagar uma das arestas para, com isso, obter dois subgrafos desconectados. Para atender as condições acima, a aresta a ser apagada será a que possuir maior custo.

Critérios de poda da AGM

A definição alternativa de custo utilizada pelo autor para apagar arestas sucessivamente do grafo da AGM, foi:

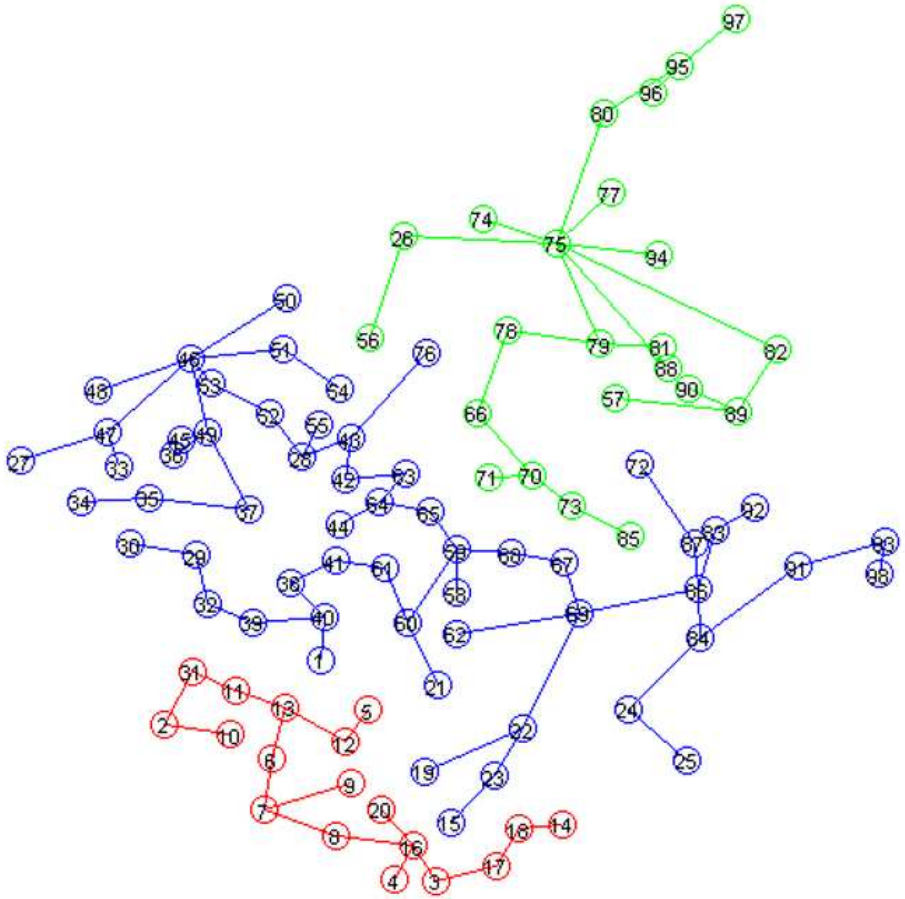
Soma de Quadrados dos desvios no espaço

$$SSTO = \sum_{j=1}^{m} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \ , \ \bar{x}_j = \sum_{i=1}^{m} \frac{x_{ij}}{n}$$

Uma outra medida utilizada é a soma das somas de quadrados de cada conglomerado (SSA). O autor definiu o **custo de apagar a aresta** como sendo a diferença, SSTO-SSA.

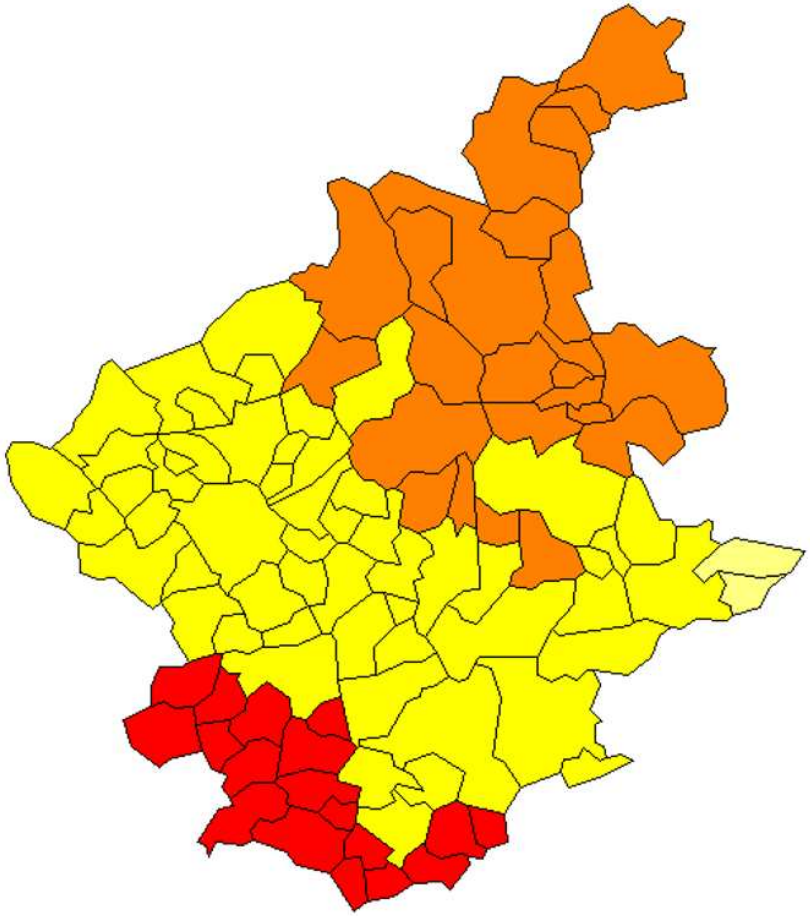
Critérios de poda da AGM

Então será apagada a aresta de maior custo. O processo termina quando todas as arestes forem apagadas, gerando **n** conglomerados.



APLICAÇÃO DO MÉTODO

Para escolher o número de cluster (parar a poda da AGM) um critério pode ser observar o gráfico da SSTO-SSA versus o número de cluster.



CONCLUSÕES

O uso do método torna-se fácil com sua implementação no software SKATER.

Possibilidade de adaptar o método à situção em que as variáveis medidas nas áreas são taxas calculadas com base nos diferentes tamanhos de população entre áreas.

critério de poda

Se usar o critério de poda usando a medida de dissimilaridade tem desvantagens: as últimas arestas a serem adicionadas na AGM tendem a ter os maiores custos (isto tenderá a quebrar o grafo nas últimas arestas que foram adicionadas a árvore). Alem disso, O custo de uma aresta é a medida de dissimilaridade entre duas arestas