

Uma análise da variação do mercado de energia da CEMIG nos anos de 2013 e 2018 utilizando a estatística de varredura espacial - *scan*

Felipe Ferreira Lopes, Igor Galvão Carvalho Dias, Marcelo Azevedo Costa

Departamento de Engenharia de Produção

Universidade Federal de Minas Gerais

felipeferr13@gmail.com, galvaoigor95@gmail.com, macosta@ufmg.br

RESUMO

A estatística de varredura espacial *scan* é amplamente utilizada para análise da ocorrência de conglomerados espaciais. Além da detecção de conglomerados onde a incidência média de um evento está significativamente acima ou abaixo da média nas demais regiões, o método permite a inferência estatística utilizando método de simulação de Monte Carlo. Neste trabalho, o método *scan* é aplicado à análise da variação do mercado de energia considerando os alimentadores de uma distribuidora de energia. Foram considerados os dados referentes aos anos de 2013 e 2018. Foram identificados conglomerados próximos à região metropolitana de Belo Horizonte e na região do Triângulo Mineiro e centro-oeste. Embora o mercado de energia de 2013 e 2018 sejam muito similares, há evidência estatística de conglomerados onde ocorreu a redução de mercado e conglomerados onde ocorreu o aumento do mercado. Esta informação é extremamente relevante para o planejamento futuro da expansão das atividades da distribuidora.

Palavras chave: *estatística de varredura, estatística espacial, métodos de monte carlo.*

ABSTRACT

The spatial scan statistic is a widely used statistical method to evaluate the occurrence of spatial clusters. In addition to detecting clusters where the mean incidence of an event is significantly above or below the mean in the remaining regions, the scan method allows statistical inference by using a Monte Carlo simulation approach. In this work, the scan method is applied to evaluate the electricity energy variation considering the electricity feeders of a power distributor. Data for 2013 and 2018 were considered. Significant clusters were identified near the metropolitan region of Belo Horizonte and in the region of Triângulo Mineiro and Midwest. Although the 2013 and 2018 total energy markets are very similar, there is statistical evidence of clusters in which market downturns were observed and conglomerates in which market upturns were observed. This information is extremely relevant for future planning of the distributor's activities.

Keywords: *scan statistics, spatial statistis, Monte Carlo methods.*

1. Introdução

Estudos de detecção de conglomerados espaciais têm sido utilizados em diversas áreas do conhecimento, devido à frequente necessidade de avaliação se a ocorrência de determinado evento espacial tem característica aleatória ou não. Exemplos de aplicações são: na silvicultura, o descobrimento se a proporção de determinada espécie de árvore é maior em certa região, ou se uma característica específica é mais incomum; na astronomia, a detecção de clusters de estrelas similares na distribuição irregular de todas estrelas no espaço; e, na saúde, a avaliação de concentrações geográficas de doenças (Kulldorff, 1997).

Define-se como conglomerado um conjunto de áreas que apresentam características específicas, quando considerada a hipótese nula (H_0) de que os eventos são gerados aleatoriamente sobre a

região de estudo. Um conglomerado pode ser caracterizado como temporal, espacial ou espaço-temporal, dependendo da variável (espaço e/ou tempo) sobre a qual é realizada a análise de contagem dos eventos (Costa et al, 2005). Existem diversas abordagens para a delimitação de conglomerados. Métodos computacionais puramente gráficos utilizam a sobreposição de círculos para a identificação de regiões críticas, não oferecendo uma medida de significância para a região delimitada (Costa et al, 2005).

Uma alternativa é a utilização do método *scan* (varredura), proposto por Kulldorff (1997), que atribui um nível de significância à estatística de teste via simulação de Monte Carlo, o que permite a redução do erro do tipo I. Em sua proposta inicial, o método é restrito à busca de conglomerados que apresentam geometria circular. Sendo assim, quando o conglomerado real apresenta uma geometria irregular, o método apresenta limitações, pois detecta nenhuma ou pequenas áreas do mesmo. Dessa forma, a principal vantagem do método para conglomerados de geometria arbitrária é sua capacidade de se aproximar mais da região crítica, que é incorporada ao resultado final (Costa et al, 2005). Entretanto, o método circular possui um custo computacional menor, uma vez que na detecção de geometria arbitrária, a busca exaustiva por todos possíveis clusters críticos em uma região dividida em n subáreas, acarreta uma varredura sobre 2^n candidatos.

Este trabalho abrange a detecção de conglomerados espaciais na variação de consumo energético em alimentadores de energia, propondo uma adaptação do método *scan* introduzido por Kulldorff (1997). O trabalho objetiva a identificação de clusters geográficos de alimentadores que apresentam variação na quantidade de energia consumida significativamente maior que o restante, sendo essa variação positiva ou negativa.

Similar ao feito por Kulldorff (2009), é desenvolvida uma análise de varredura para dados contínuos baseada na distribuição Normal (Gaussiana). Sob a hipótese nula, todas as observações têm a mesma distribuição e compartilham o mesmo parâmetro de média e variância. Sob a hipótese alternativa, existe um conglomerado onde as observações têm uma média maior ou menor que as observações fora do conglomerado. Uma característica importante do método é o fato da inferência estatística permanecer válida mesmo se a verdadeira distribuição não é Normal, garantindo que o erro tipo I é mantido. Isso é feito avaliando-se a significância estatística dos conglomerados por meio de um procedimento de teste de Monte Carlo baseado em permutação. O método é então aplicado aos dados pontuais dos alimentadores da área de concessão da Companhia Energética de Minas Gerais (CEMIG).

2. O método de varredura espacial - *scan*

2.a. Observações e localizações

Os dados consistem em observações contínuas, com valores x_i , $i = 1, \dots, N$. Cada observação ocorre em um local no espaço s , $s = 1, \dots, S$, com coordenadas de latitude e longitude $lat(s)$ e $lon(s)$. Cada localização tem uma ou mais observações, de forma que $S \leq N$.

Para cada local s , a soma dos valores observados é $\sum_{i \in s} x_i$ e o número de observações no local como n_s . A soma de todos os valores é $X = \sum_i x_i$.

2.b. Janela de varredura

O método de varredura *scan* com geometria circular é definido por meio de um número grande de círculos sobrepostos (Kulldorff, 1997). Para cada círculo z a razão log verossimilhança $LLR(z)$ é calculada e a estatística de teste é definida como a estatística LLR máxima em todos os círculos.

A janela de varredura pode variar de acordo com a aplicação, sendo o mais comum defini-la como círculos centrados em uma observação e com o raio variando continuamente entre zero e o limite superior. Esse limite superior é normalmente definido de forma que o círculo contenha, no máximo, 50 por cento de todas as observações, para garantir que conglomerados pequenos e grandes possam ser encontrados (Kulldorff et al, 2009). O limite não é definido acima desse valor. Por exemplo, um conglomerado circular de valores elevados com 70 por cento de todas as observações pode ser melhor interpretado como um conglomerado de valores baixos espacialmente desconectados cobrindo 30 por cento das observações que estão fora do círculo,

uma vez que o grupo menor difere da maioria das observações, e não o grupo dos 70 por cento. O tamanho máximo do conglomerado também pode ser definido usando unidades específicas de distância (5 km, por exemplo). Círculos com uma observação são ignorados. O número de observações no círculo z é $n_z = \sum_{s \in z} n_s$ e a soma dos valores observados no círculo z é $x_z = \sum_{s \in z} x_s$

2.c. Cálculo da verossimilhança

Sob a hipótese nula, o estimador da máxima verossimilhança da média e da variância é $\mu = X/N$ e $\sigma_z^2 = \frac{\sum_i (\mu - x_i)^2}{N}$ respectivamente. Dessa forma, a função verossimilhança sob a hipótese nula é:

$$L_0 = \prod_i \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \quad (1)$$

e a função log verossimilhança é:

$$\ln L_0 = -N \ln(\sqrt{2\pi}) - N \ln(\tau) - \sum_i \frac{(x_i - \mu)^2}{2\tau^2} \quad (2)$$

Sob a hipótese alternativa, calculamos primeiro os estimadores da verossimilhança que são específicos para cada círculo z , que é $\mu_z = x_z/n_z$ para a média dentro do círculo e $\lambda_z = (X - x_z)/(N - n_z)$ para a média fora do cluster. O estimador de máxima verossimilhança para a variância comum é:

$$\sigma_z^2 = \frac{1}{N} (\sum_{i \in z} x_i^2 - 2x_z \mu_z + n_z \mu_z^2 + \sum_{i \in z} x_i^2 - 2(X - x_z) \lambda_z + (N - n_z) \lambda_z^2) \quad (3)$$

A log-verossimilhança para o círculo z é:

$$\begin{aligned} \ln L(z) = & -N \ln(\sqrt{2\pi}) - N \ln(\sqrt{\sigma_z^2}) \\ & - \frac{1}{2\sigma_z^2} (\sum_{i \in z} x_i^2 - 2x_z \mu_z + n_z \mu_z^2) \\ & + \sum_{i \in z} x_i^2 - 2(X - x_z) \lambda_z + (N - n_z) \lambda_z^2 \end{aligned} \quad (4)$$

2.d. Método de Monte Carlo

A significância estatística do conglomerado mais provável é avaliada usando o teste de hipóteses de Monte Carlo (Dwass, 1957). Em vez de gerar dados aleatórios a partir da distribuição normal, um grande conjunto de conjuntos de dados aleatórios é criado permutando aleatoriamente os valores observados x_i e seus locais correspondentes s . Neste caso, a análise é condicionada às observações contínuas que foram observadas, assim como aos locais onde ocorreram as observações. A partir do procedimento de aleatorização (ou randomização), o nível α (erro tipo I)

será mantido, mesmo que as observações não venham realmente de uma distribuição Normal. São as observações individuais que são permutadas.

Para cada conjunto de dados aleatorizados, a log-verossimilhança é calculada para cada um dos possíveis círculos. O conglomerado, ou círculo, mais provável é encontrado e seu valor para o log da razão de verossimilhança é registrado. Se o log da razão de verossimilhança do conjunto de dados reais está entre os 5 por cento maiores de todos os conjuntos de dados simulados, então o conglomerado mais provável do conjunto de dados reais é estatisticamente significativo considerando um nível α de 0,05 (5%). Portanto, se existem M conjunto de dados aleatórios, então o valor-P do cluster mais provável é igual a $R/(M + 1)$, em que R é o número de simulações com valores de log-verossimilhança maiores que o valor real observado. Com o objetivo de obter valores com alta resolução para o valor-P com um número finito de decimais, o número de simulações M deve ser, por exemplo de 999, 9999, ou maiores.

Observe que esses valores-P baseados em simulações de Monte Carlo são exatos no sentido de que, sob a hipótese nula, a probabilidade de observar um valor-P menor ou igual a p é exatamente p . Isso é verdade independentemente do número de conjuntos de dados aleatórios M , mas um M maior fornecerá maior poder estatístico.

2.e. Varredura para valores altos e baixos

A estatística de varredura utilizando a distribuição Normal procurará por clusters cujas médias tenham valores excepcionalmente altos, bem como clusters cujas médias sejam excepcionalmente baixas. Em alguns casos, faz mais sentido procurar apenas clusters com valores elevados. O primeiro é facilmente realizado adicionando uma função indicadora no processo de varredura. Se o analista está interessado apenas em cluster com valores baixos, a função do indicador é $I(\mu_z < \lambda_z)$.

3. Aplicação do método *scan* para a detecção de conglomerados de alimentadores com variação de consumo de energia em Minas Gerais

3.a. Adaptação método *scan*

Neste artigo, adaptamos o método *scan* apresentado na seção 2 para análise de dados de consumo de alimentadores de energia. Cada alimentador tem registrado o valor de consumo para o ano a igual a x_i^a com $i = 1, \dots, N$, sendo N o número total de alimentadores, e está em um local no espaço $s = 1, \dots, S$, com coordenadas de latitude e longitude $lat(s)$ e $lon(s)$, de forma que $S = N$, uma vez que cada alimentador i possui apenas uma observação por ano a .

O objetivo desta análise é verificar a existência de conglomerados de alimentadores que tiveram variação absoluta do consumo de energia significativamente maior/menor que os alimentadores restantes. Para isso, é utilizada a variação do consumo no alimentador i entre os anos a e b , dada por $y_i^a = x_i^a - x_i^b$. A variação total de consumo de energia entre os anos a e b é dada por $Y = \sum_i y_i$.

A janela de varredura é feita com o tamanho do conglomerado sendo estabelecido *a priori*, de forma que são analisados conglomerados com k alimentadores. O conglomerado z_i é composto pelos k alimentadores mais próximos do alimentador i . A distância d_{ij} entre os alimentadores i e j é calculada como sendo a distância euclidiana entre as posições s dos alimentadores, dada por

$$d_{ij} = \sqrt{(s_x^i - s_x^j)^2 + (s_y^i - s_y^j)^2}, \text{ sendo } s_x^i \text{ e } s_y^i \text{ as coordenadas cartesianas do alimentador } i.$$

Para cada conglomerado z , é calculada a razão da log-verossimilhança $LLR(z)$, e a estatística-teste é definida como sendo a máxima LLR em todos os conglomerados.

O consumo total de um conglomerado é dado por $x_z = \sum_{s \in z} s_s$. A função verossimilhança é calculada da mesma forma das equações (1), (2), (3) e (4) apresentadas por Kulldorff (2009).

Para a implementação dos algoritmos, foi utilizado o R project, software gratuito disponível em: <https://www.r-project.org/>.

4. Metodologia: detecção de conglomerados de consumo de energia em Minas Gerais

Foi feita uma análise da variação de consumo de energia nos alimentadores da CEMIG entre os anos de 2013 e 2018. No ano de 2018, foram identificados 2179 alimentadores, 11,4% a mais que os 2427 alimentadores identificados em 2013. Desses, 2027 alimentadores são comuns aos dois anos. Somente 152 alimentadores são específicos para o ano de 2013 e 400 alimentadores são específicos para o ano de 2018, como mostrado na Figura 1.

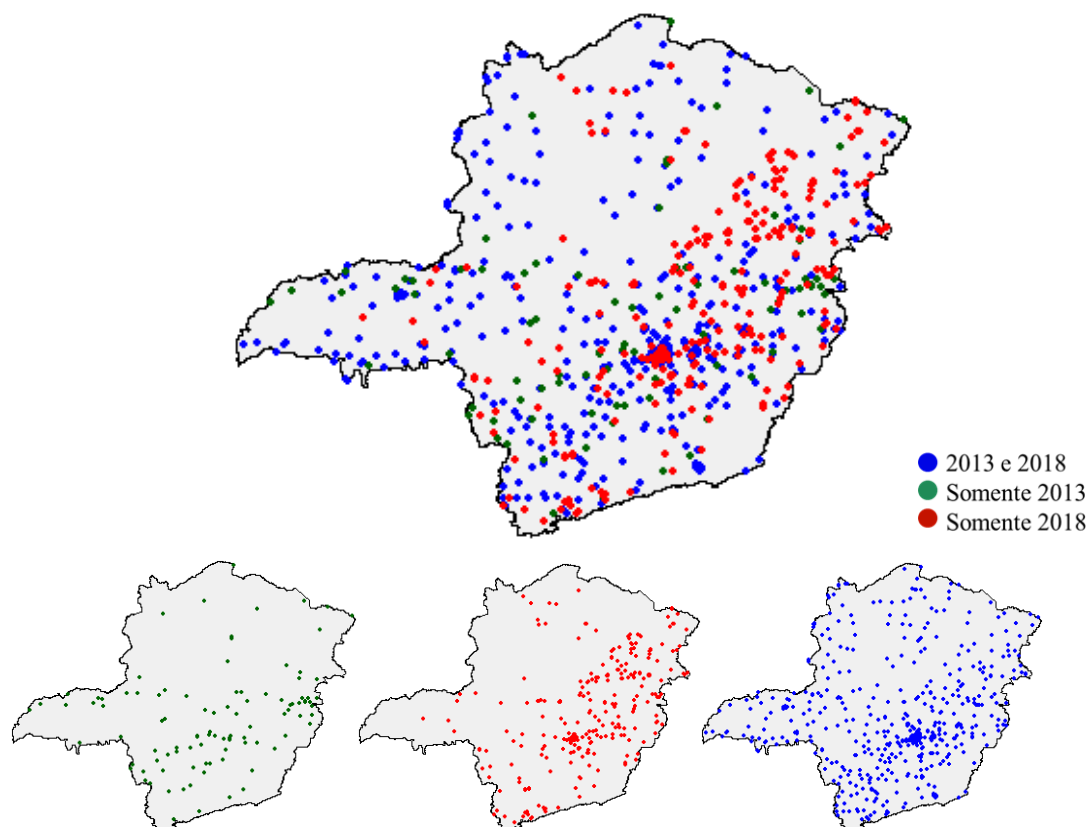


Figura 1. Comparação da distribuição espacial e disponibilidade dos alimentadores para os anos de 2013 e 2018.

No ano de 2013, o consumo médio de energia por alimentador foi de 14.057.962 KWh, sendo o consumo máximo 260.748.183 KWh do alimentador *BHCN01*. Já no ano de 2018, o consumo médio de energia por alimentador foi de 15.070.366 KWh (+7,2%), sendo o consumo máximo 120.254.369 KWh, do mesmo alimentador *BHCN01*.

A variação média da energia consumida por alimentador entre 2013 e 2018 foi de 623.542 KWh. A variação negativa mais acentuada foi de -140.493.814 KWh e a positiva foi de 51.498.168 KWh. A distribuição das variações é apresentada na Figura 2.

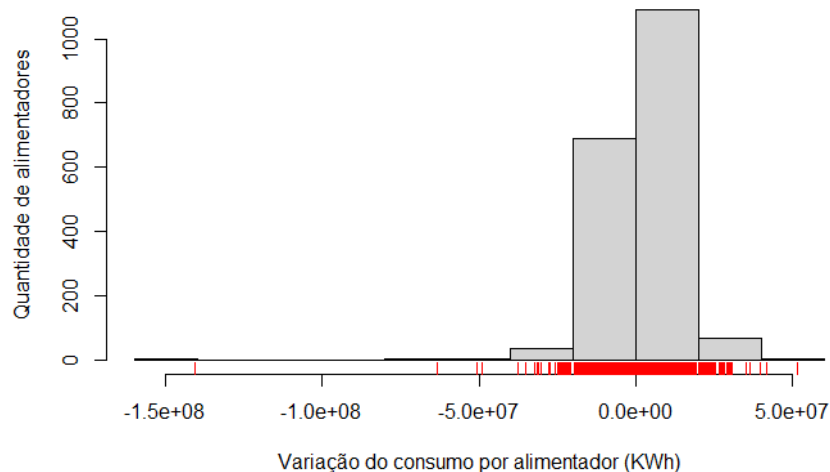


Figura 2. Distribuição das variações do consumo entre 2013 e 2018 por alimentador

Analisou-se o consumo dos alimentadores para os anos de 2013 e 2018. No ano de 2013, foi medido um consumo total de 44.390.394.395 KWh (44.390 GWh). Desse valor, 16.874.596.313 KWh (38,01%) possui ausência de informação do alimentador. Considerando as informações de localização dos alimentadores, 204 alimentadores não possuem coordenadas geográficas. Esse número inclui alimentadores com registro de identificação ausente. Os 204 alimentadores sem coordenadas (e/ou sem nome) representam 38,01% do mercado total de energia em 2013. Os alimentadores que possuem coordenadas geográficas e informação de mercado representam 61,99% do mercado total de energia em 2013.

Considerando o ano de 2018, foi medido um consumo total de 44.134.022.936 KWh (44.134 MWh). Ou seja, houve uma redução de 0,58% em comparação com o ano de 2013. Desse valor, um mercado de 15.372.021.350 KWh (34,94%) possui ausência de informação do alimentador e 49 alimentadores não estão vinculados a coordenadas geográficas, totalizando 34,94% do mercado de energia no respectivo ano. Em suma, os alimentadores que possuem coordenadas geográficas e informação de mercado representam 65,06% do mercado total de energia em 2018.

Foi testada a existência de conglomerados para diferentes valores de k , referente à quantidade de alimentadores que compõe o conglomerado. Como abordado por Kulldorff (2009), o conglomerado não deve ter mais que 50% de todas as observações, pois, caso isso ocorra, o grupo fora do conglomerado seria o que difere da maioria das observações e não o conglomerado em si. Limitamos, dessa forma, o maior número de k analisado a 500 alimentadores.

5. Resultados

A síntese dos resultados é mostrada na Figura 3 e na Tabela 1. Foram encontrados conglomerados significativos para valores de $k = 10, 150, 200, 250$ e 500. Não foram encontrados conglomerados significativos para os valores de $k = 50$ e 100. Os conglomerados muitas vezes apresentam alimentadores em comum, uma vez que foi estabelecido que seriam considerados os k alimentadores mais próximos do alimentador i para a formação do conglomerado z_i , centrado em i . Dessa forma, existem sobreposições na medida que um alimentador pode pertencer a conglomerados significativos “centrados” em outros alimentadores, do qual ele faz parte dos k mais próximos, bem como ser “centro” de um conglomerado significativo junto a seus k alimentadores mais próximos.

Como mostrado na Tabela 1, para $k=10$, foram encontrados 8 conglomerados significativos. Para $k = 50$ e $k=100$, não foram encontrados conglomerados. O número de conglomerados encontrados aumenta de maneira significativa quando k aumenta, o que é intuitivo dado a característica de sobreposição de conglomerados, com alimentadores que tiveram grande variação pertencendo a mais de um conglomerado significativo.

Percebe-se uma tendência de concentração dos alimentadores pertencentes a conglomerados significativos na área próxima à região metropolitana de Belo Horizonte, os quais estão presentes em todos os valores de k analisados que obtiveram conglomerados significativos. Com o aumento de k , aparecem conglomerados que englobam alimentadores na região central/noroeste do estado, como mostrado na Figura 3 com $k = 500$. Essa nova região tem menor densidade de alimentadores do que a região leste do estado, como mostrado na Figura 1, e também não foi a região que tem mais dados de novos alimentadores em 2018, como também mostrado na Figura 1.

k	Clusters significativos	-ln Verossimilhança			p-valor		
		min.	med.	max.	min.	med.	max.
10	8	-16,04064	-16,04064	-16,0445	0,001	0,001	0,010
50	na	na			na		
100	na	na			na		
150	4	-16,05495	-16,05495	-16,05495	0,051	0,051	0,051
200	18	-16,05514	-16,05508	-16,05503	0,045	0,046	0,049
250	20	-16,0551	-16,05548	-16,05542	0,016	0,032	0,049
500	379	-16,05562	-16,05545	-16,05439	0,008	0,025	0,051

Tabela 1: Resultados obtidos com $k = 50, 100, 150, 200, 250$ e 500 .

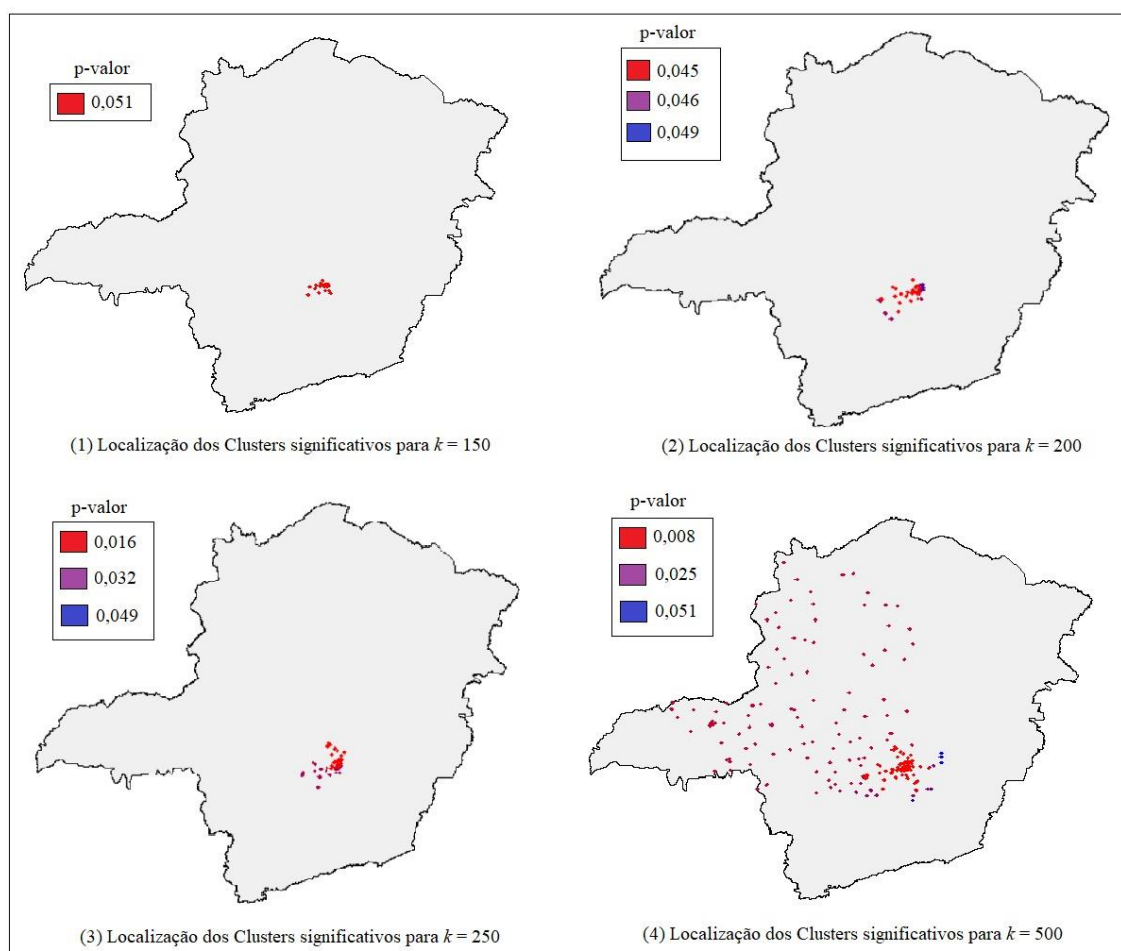


Figura 3. Conglomerados significativos a um nível de 5% para $k = 150, 200, 250$ e 500

Na análise com $k = 10$, foi notado uma peculiaridade que vale ser ressaltada. Os 8 conglomerados significativos encontrados (p-valores entre 0,002 e 0,008) são formados por 12 alimentadores situados na cidade de Belo Horizonte, como indicado na Figura 4.1. As coordenadas dos alimentadores remetem a duas subestações, sendo 8 deles na *Subestação Gutierrez* (Figura 4.2) e 4 deles na *Subestação Centro 1* (Figura 4.3). As outras subestações da capital mineira, como a *Subestação Carlos Prates* e a *Subestação Barro Preto*, por exemplo, não apresentam alimentadores em conglomerados significativos para o valor de k mais baixo analisado.



Figura 4.1 Conglomerados significativos a um nível de 5% para $k = 10$. Mapa ampliado de Belo Horizonte. Fonte: Google Earth



Figuras 4.2 Alimentadores da *Subestação Centro 1* e **4.3** Alimentadores da *Subestação Gutierrez*

Para a análise mais aprofundada, considerou-se o cluster mais significativo para conglomerados cujo $k = 150$. Essa escolha se deu após avaliar que para valores menores de k , como $k = 10$, a distância dos alimentadores seria pequena a ponto de um conglomerado poder ser composto por poucos centros de distribuição, representando o comportamento de consumo de uma região diminuta.

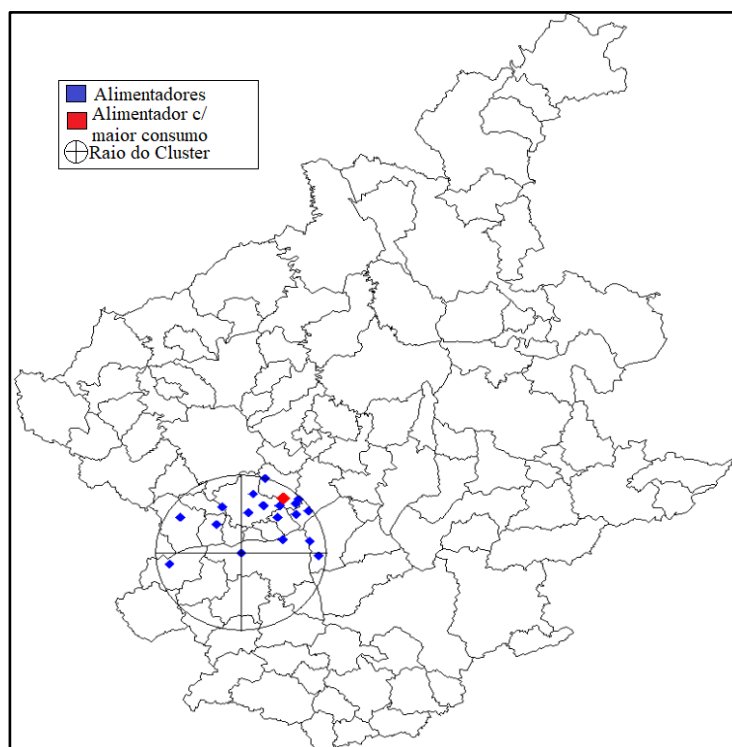


Figura 4: Localização a dos alimentadores do cluster mais significativo para $k = 150$, plotados na região metropolitana de Belo Horizonte

O conglomerado mais significativo é centrado no alimentador *BMO02*, localizado no Município de Brumadinho. O conglomerado contempla também os municípios de Belo Horizonte, Betim, Contagem, Nova Lima, Brumadinho, Ibirité, Igarapé, Itatiaiuçu, Mateus Leme e Juatuba, todos municípios da região metropolitana de Belo Horizonte. O alimentador de maior consumo desse cluster também é o alimentador de maior consumo do estado, no qual se localiza no Parque Industrial de Contagem.

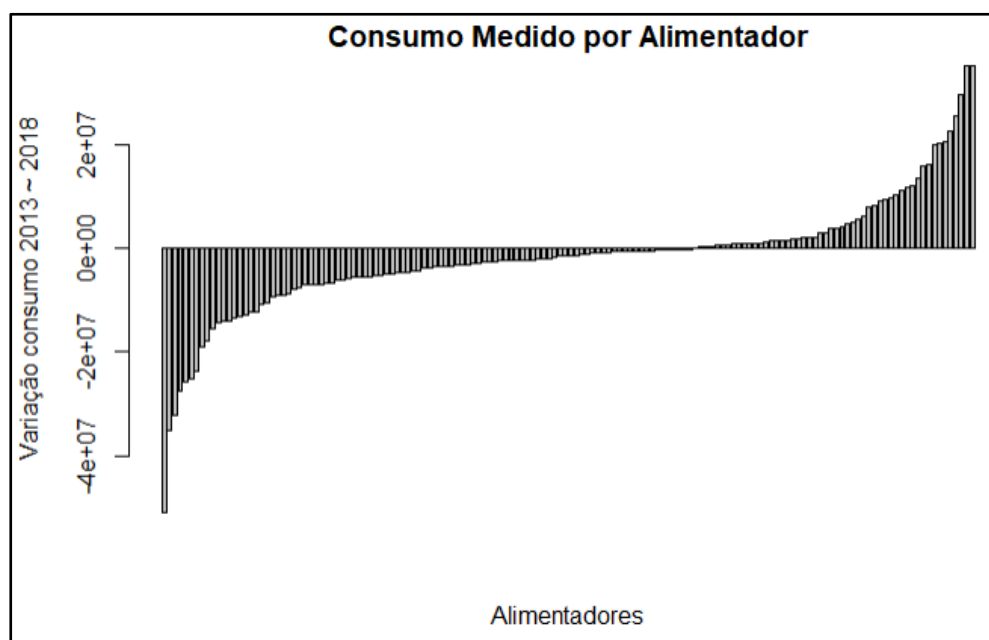


Figura 5: Gráfico da variação do consumo medido do ano de 2013 a 2018 por alimentador do conglomerado centrado no alimentador *BMO02* com $k=150$

Como se pode evidenciar pelo gráfico acima, a maior parte dos alimentadores do conglomerado tiveram uma retração no consumo de energia elétrica do ano de 2013 para o ano de 2018. Essa característica, juntamente com o fato dos alimentadores com as maiores variações terem variação negativa, contribuiu para que esse conglomerado se destacasse como o conglomerado mais significativo.

Durante a análise, percebeu-se que os alimentadores com maior consumo e também maior variação localizam-se em regiões altamente industrializadas, como o Parque Industrial de Contagem e também o Distrito Industrial de Betim, locais onde estão situadas Indústrias Automobilísticas, Refinarias de Petróleo, entre outros empreendimentos.

6. Discussão e conclusão

Este trabalho teve como objetivo descrever, aplicar e analisar os resultados do ajuste do método *scan* para aplicação na detecção de conglomerados de alimentadores de energia com variação de alta de consumo entre anos diferentes. A aplicação se deu com a determinação *a priori* da quantidade k de alimentadores por conglomerado que seria aplicada e o modelo apresentou resultados coerentes para diferentes valores de k , de modo que as regiões significativas encontradas nos testes se sobrepunham em uma região na parte centro-sul do estado de Minas Gerais.

7. Bibliografia

Costa, Marcelo Azevedo; Scherrer, Luciano R.; Assunção, Renato M. Detecção de Conglomerados Espaciais com Geometria Arbitrária. In: **GeoInfo**. 2005. p. 213-222.

DWASS, Meyer. Modified randomization tests for nonparametric hypotheses. **The Annals of Mathematical Statistics**, p. 181-187, 1957.

Kulldorff, Martin. A spatial scan statistic. **Communications in Statistics-Theory and methods**, v. 26, n. 6, p. 1481-1496, 1997.

Kulldorff, Martin; Huang, Lan; Konty, Kevin. A scan statistic for continuous data based on the normal probability model. **International journal of health geographics**, v. 8, n. 1, p. 58, 2009.