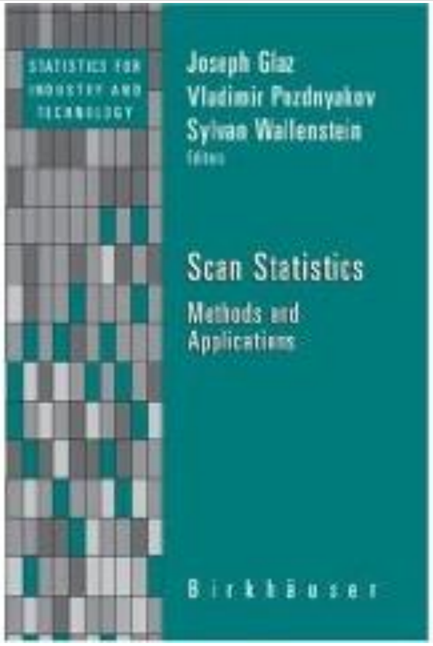
Detecção de conglomerados espaciais e espaço-temporais: uma revisão da teoria e aplicações.

Prof. Marcelo Azevedo Costa
Departamento de Engenharia de Produção
Universidade Federal de Minas Gerais

Resumo

Os métodos de detecção de conglomerados (clusters) espaciais e espaço-temporais têm a sua origem no trabalho de Joseph Nauss (1965) - Clustering of random points in two dimensions. Posteriormente, foram desenvolvidos os métodos de varredura espacial (Kulldorff, 1997) que pré-definem uma geometria de busca para detectar conglomerados. Inferência estatística é obtida a partir de simulações de Monte Carlo. Como alternativa à busca com geometrias pré-definidas, foram desenvolvidos os métodos de busca com geometria arbitrária. Em geral, geometrias mais flexíveis demandam maior tempo computacional e reduzem o poder de detecção de conglomerados. Uma segunda classe de modelos de detecção de conglomerados utilizam a abordagem Bayesiana. Esses métodos permitem definir, a priori, o número de conglomerados que se deseja encontrar. Utilizando métodos de cadeias de Markov (MCMC) é possível amostrar das distribuições a posteriori e estimar o número e as localizações dos conglomerados. Estudos de casos para as diferentes abordagens são apresentadas em contextos epidemiológicos e industriais.

Applications of Spatial Scan Statistics: A Review

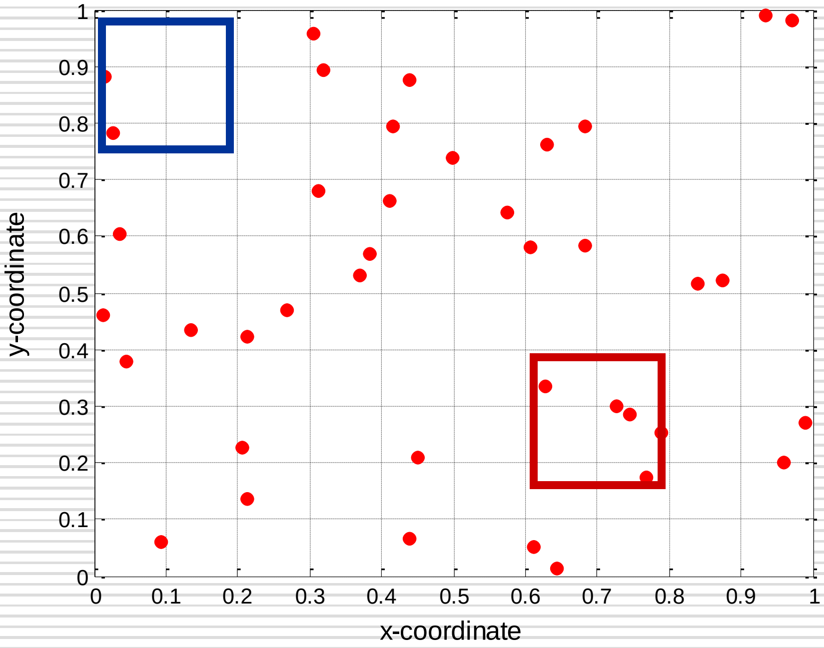


- Em 1965, Joseph Naus publicou o primeiro trabalho sobre a estatística espacial *scan*, intitulado "Agrupamento aleatório de pontos em duas dimensões". Este trabalho deu origem a uma importante área da estatística espacial e proporcionou o surgimento de uma variedade de técnicas e aplicações, propostas em diversas áreas do conhecimento, incluindo a arqueologia, astronomia, imagens cerebrais, criminologia, demografia, a detecção precoce de surtos de doenças, ecologia, epidemiologia, florestal, geologia, história, psicologia e medicina veterinária.

How to exterminate a bug?

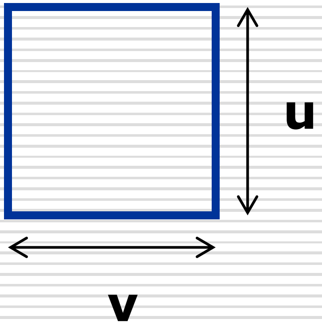
Clustering of random points in two dimensions

by J. I. Naus



Objetivo: obter os limites superiores e inferiores da probabilidade de encontrar ao menos um cluster de dimensões **v** e **u** contendo pelo menos **n** pontos,

$$P(n \mid N, u, v).$$



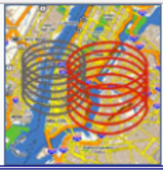
Clustering of random points in two dimensions. *Biometrika* **52** (1965), 263-267.

Kulldorff M. A spatial scan statistic. Communications in Statistics: Theory and Methods, 1997; 26:1481-1496.

Seguro | https://www.satscan.org

SaTScan™

Software for the spatial, temporal, and space-time scan statistics



- Home
- Download [SaTScan v9.4.4 *August 16 2016*]
- Technical Documentation
- Bibliography
- Tutorials
- Data Sets
- Related Software
- Contact Us

Other Scan Statistics Software

TreeScan

Follow @martinkulldorff

Purpose

SaTScan™ is a free software that analyzes spatial, temporal and space-time data using the spatial, temporal, or space-time scan statistics. It is designed for any of the following interrelated purposes:

- Perform geographical surveillance of disease, to detect spatial or space-time disease clusters, and to see if they are statistically significant.
- Test whether a disease is randomly distributed over space, over time or over space and time.
- Evaluate the statistical significance of disease cluster alarms.
- Perform repeated time-periodic disease surveillance for early detection of disease outbreaks.

The software may also be used for similar problems in other fields such as archaeology, astronomy, botany, criminology, ecology, economics, engineering, forestry, genetics, geography, geology, history, neurology or zoology.

Data Types and Methods

SaTScan uses either a Poisson-based model, where the number of events in a geographical area is Poisson-distributed, according to a known underlying population at risk; a Bernoulli model, with 0/1 event data such as cases and controls; a space-time permutation model, using only case data; an ordinal model, for ordered categorical data; an exponential model for survival time data with or without censored variables; or a normal model for other types of continuous data. The data may be either aggregated at the census tract, zip code, county or other geographical level, or there may be unique coordinates for each observation. SaTScan adjusts for the underlying spatial inhomogeneity of a background population. It can also adjust for any number of categorical covariates provided by the user, as well as for temporal trends, known space-time clusters and missing data. It is possible to scan multiple data sets simultaneously to look for clusters that occur in one or more of them.

Developers and Funders

The SaTScan™ software was developed by Martin Kulldorff together with Information Management Services Inc. Financial support for SaTScan has been received from the following institutions:

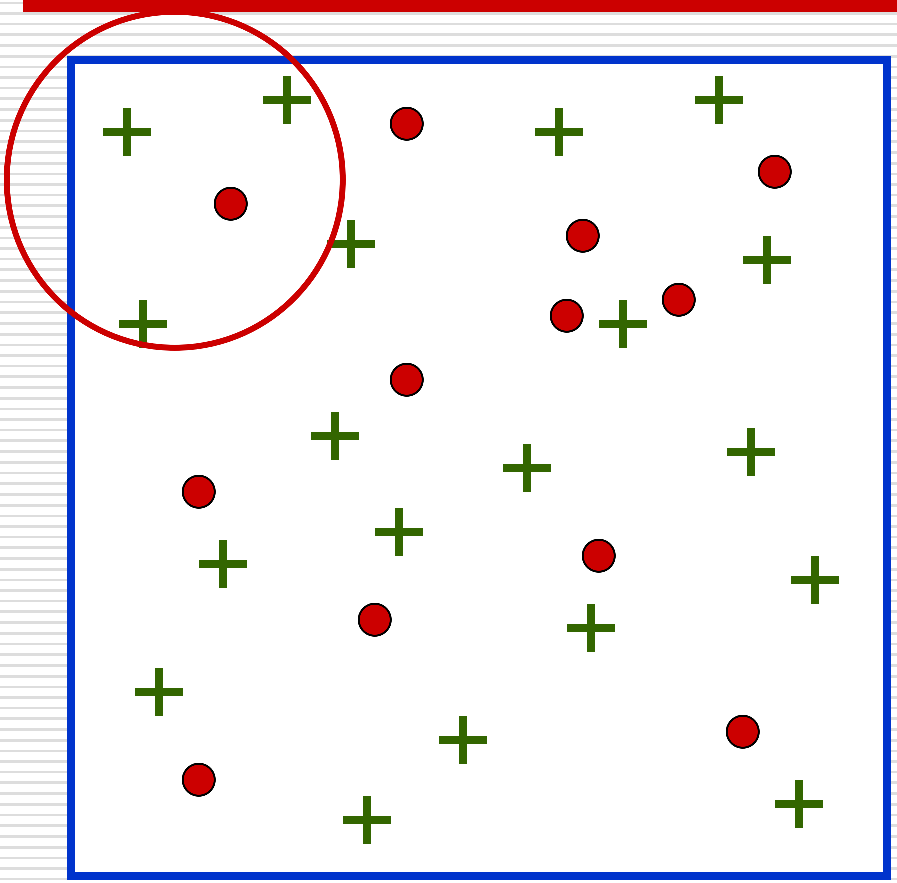
- National Cancer Institute, Division of Cancer Prevention, Biometry Branch [v1.0, 2.0, 2.1]

- National Cancer Institute, Division of Cancer Control and Population Sciences, Statistical Research and Applications Branch [v3.0 (part), v6.1 (part), 8.0 (part), v9.0 (part)]), 9.2-9.4
- Alfred P. Sloan Foundation, through a grant to the New York Academy of Medicine (Farzad Mostashari, PI) [v3.0 (part), 3.1, 4.0, 5.0, 5.1]
- Centers for Disease Control and Prevention, through Association of American Medical Colleges Cooperative Agreement award number MM-0870 [v6.0, 6.1 (part)].
- National Institute of Child Health and Development, through grant #RO1HD048852 [7.0, 8.0, 9.0 (part)]
- National Cancer Institute, Division of Cancer Epidemiology and Genetics [v9.0 (part)]
- National Institute of General Medical Sciences, through a Modelling Infectious Disease Agent Studies grant #U01GM076672.[v9.0 (part),9.1]

Their financial support is greatly appreciated. The contents of SaTScan are the responsibility of the developer and do not necessarily reflect the official views of the funders.

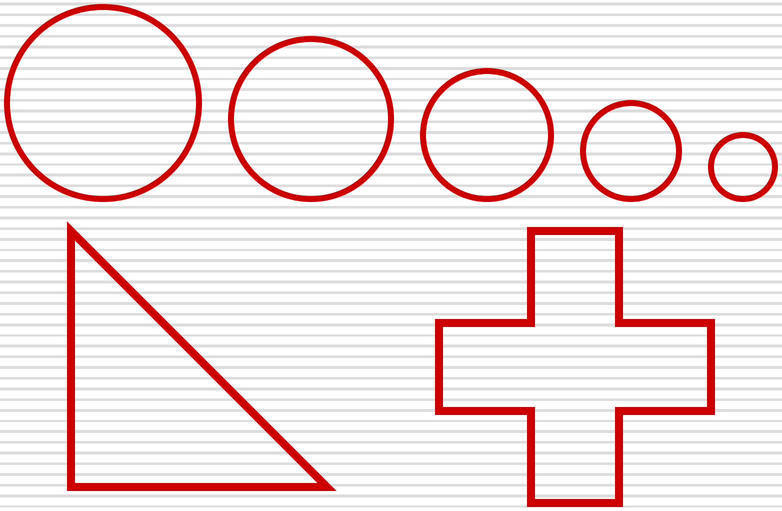
www.satscan.org

Entendendo a Estatística de Varredura (Scan)

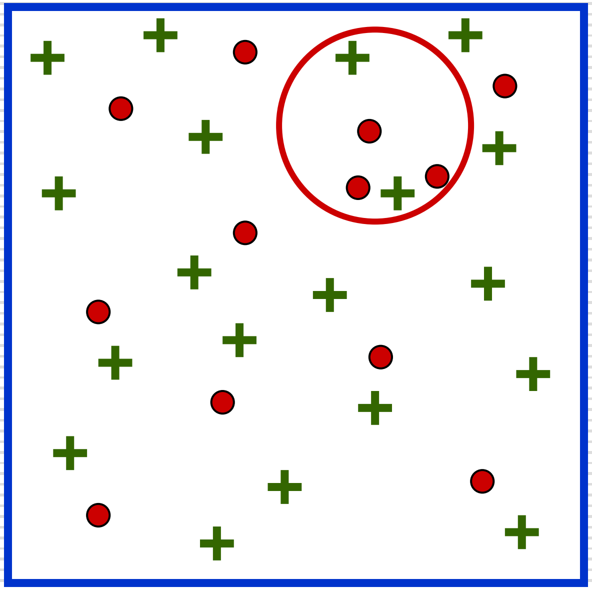


- Casos
- Controle
- Janela de Varredura (z)

Mais janelas:



O que é um *cluster*? (conglomerado)



O que está ocorrendo no cluster z?
1: caso
0: controle

Dentro = {1,1,1,0,0}
Fora = {1,...,1,0,...,0}

Estatística de Teste

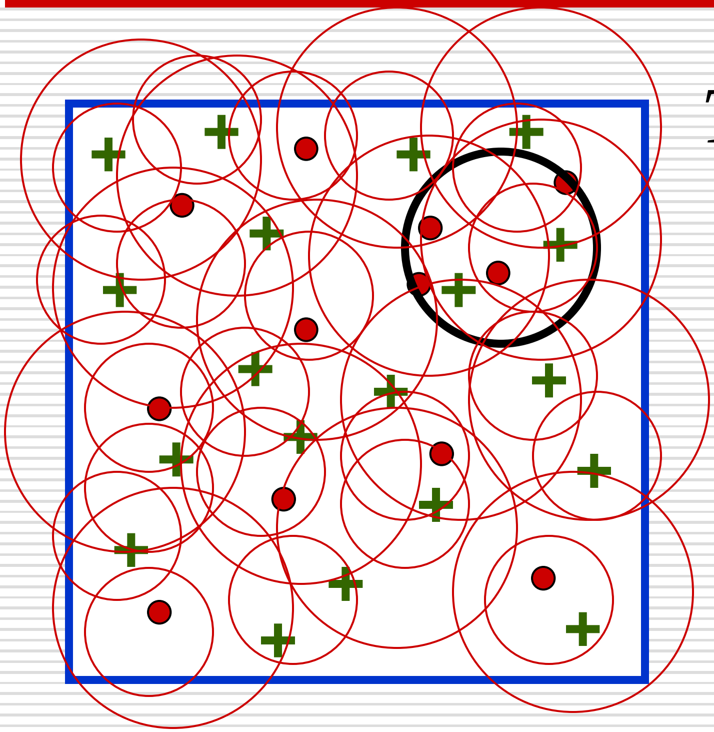
$$P(Cenário) = P(Dentro \cap Fora)$$

$$P(Cenário) = P(Dentro) \cdot P(Fora)$$

$$P(Dentro) = P(\{1, 1, 1, 0, 0\}) \quad = q^3(1-q)^2$$

q é a probabilidade de um indivíduo em (dentro) z ser um caso

A estatística de teste



$$T_z = q^{c_z}(1-q)^{n_z}\, r^{C-c_z}(1-r)^{N-n_z}$$

$$\kappa = \underset{z}{\arg\max}\,\{T_z\}$$

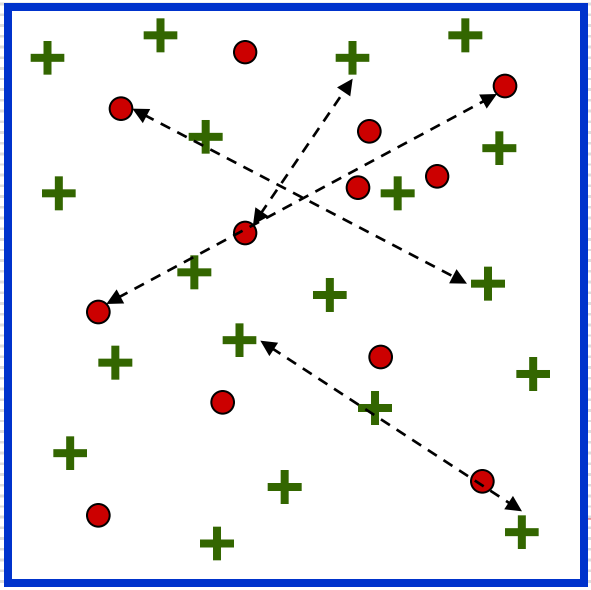
Sob a restrição de que:

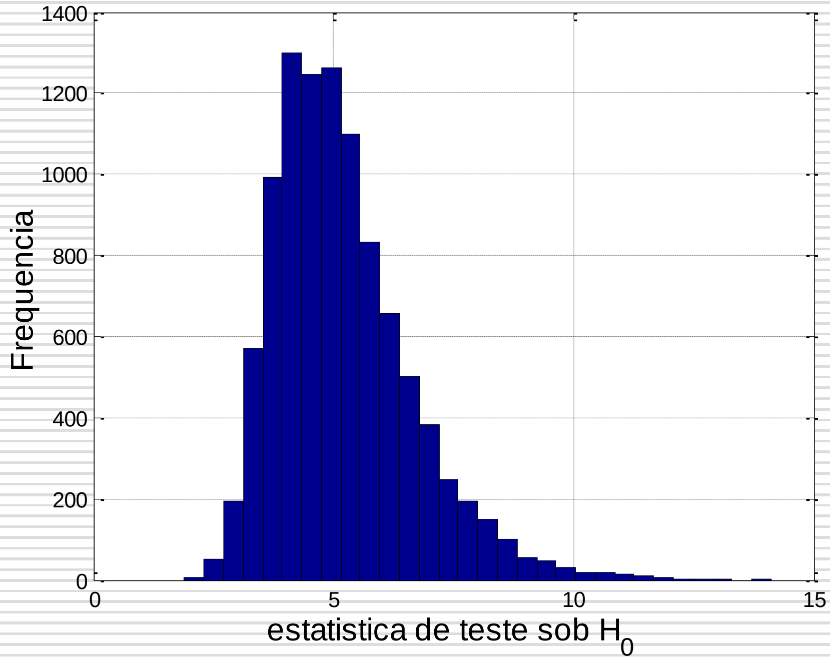
$$q > r$$

Como saber se o cluster z é crítico?

- Sob a hipótese da **NÃO** existência de um cluster espacial qual a distribuição da estatística de teste?

Solução: Simulação





Componentes necessárias para uso da estatística espacial de varredura

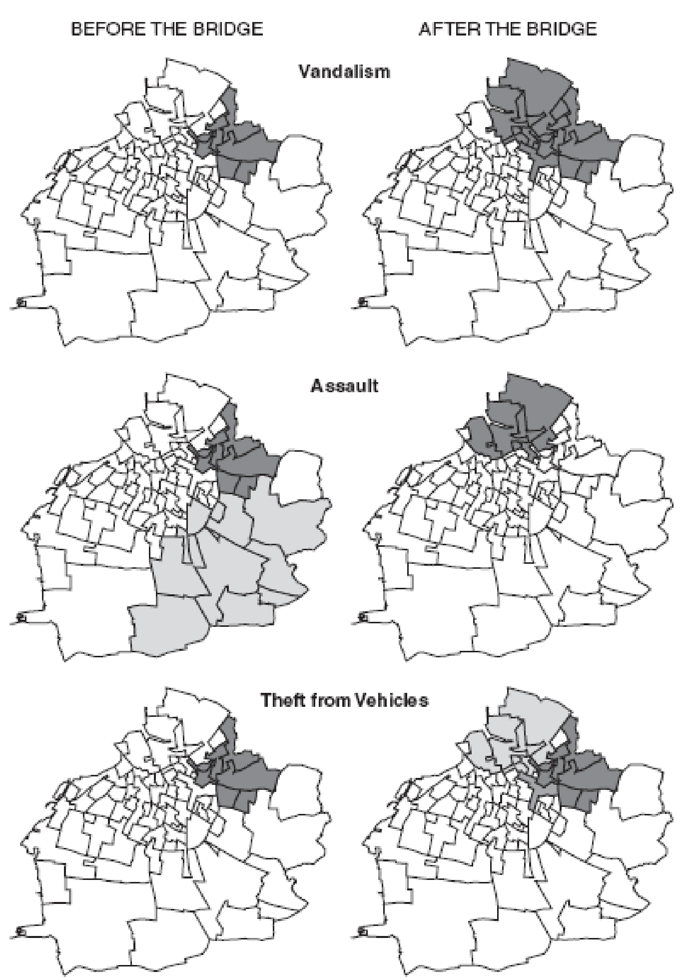
- Casos
- Controle/População

Algumas Áreas de Aplicação:

- Geology
- Medical Imaging
- Chronic Disease Epidemiology
- Infectious Disease Epidemiology
- Parasotology
- Psycology
- Veterinary Medicine
- Accidents and Crime
- Humanities and Social Sciences

Applications to Accidents and Crime

Ceccato, V. and Heining, R. (2004). Crime in border regions: The Scandinavian case of Öresund, 1998-2001, *Annals of the Association of American Geographers*, **94**, 807-826.

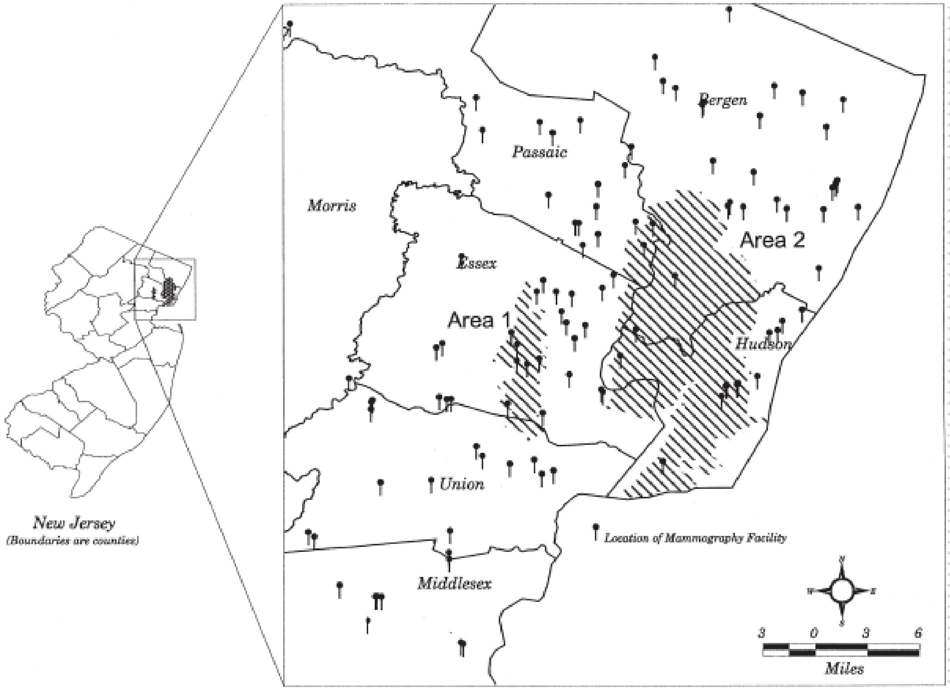


No significant clusters were found closer to the vicinity of the terminal of the bridge but notably shifts in geographical locations of previous clusters and emerged clusters were reported for some of the offences

Applications in Chronic Disease Epidemiology

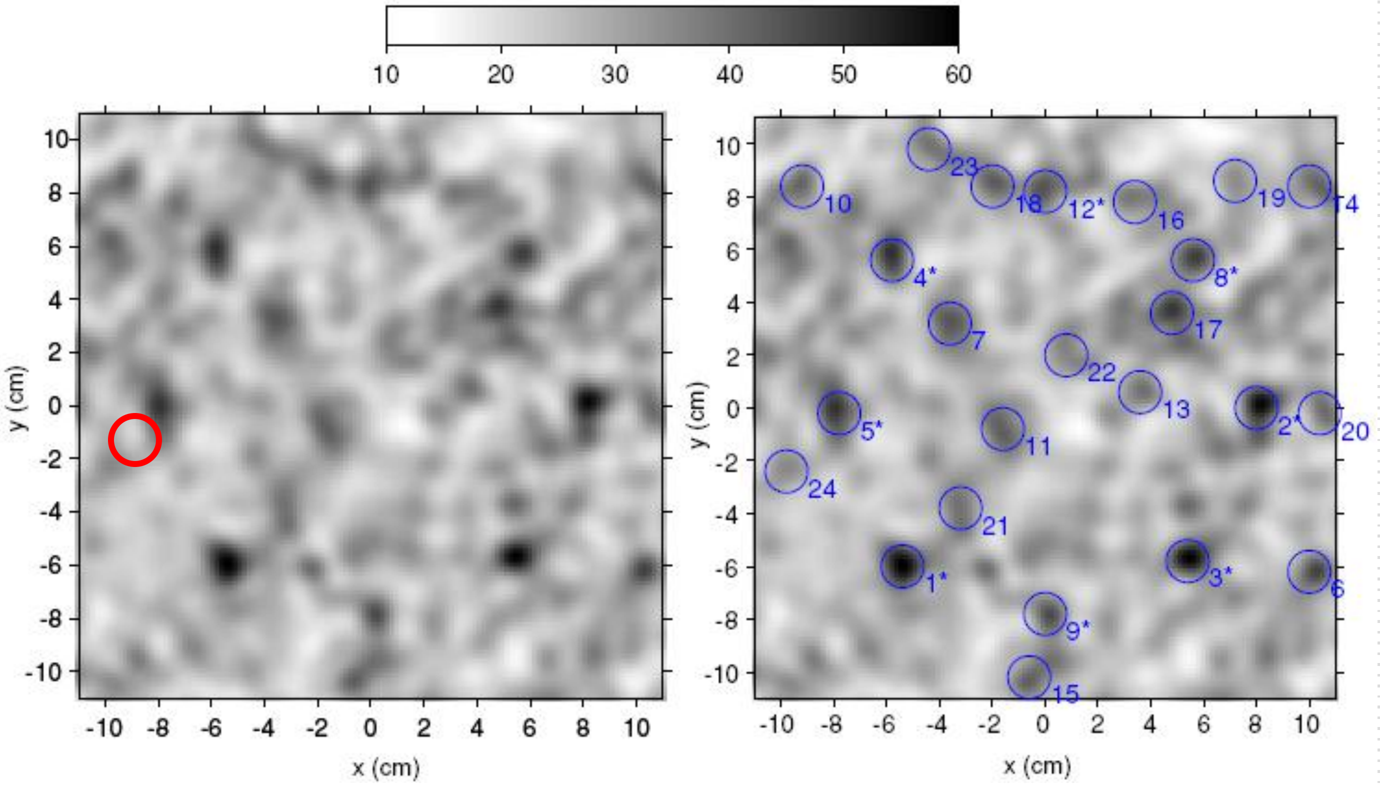
Roche, L.M., Skinner, R. and Weinstein, R.B. (2002). Use of a geographic information system to identify and characterize areas with high proportions of distant stage breats cancer, *Journal of Public Health Management and Practice*, **8**, 26-32.

- Instead of population at risk, the total number of diagnosed cases of breast cancer is the “population” and “cases” corresponds to a fraction of breast cancer cases at distance stage.

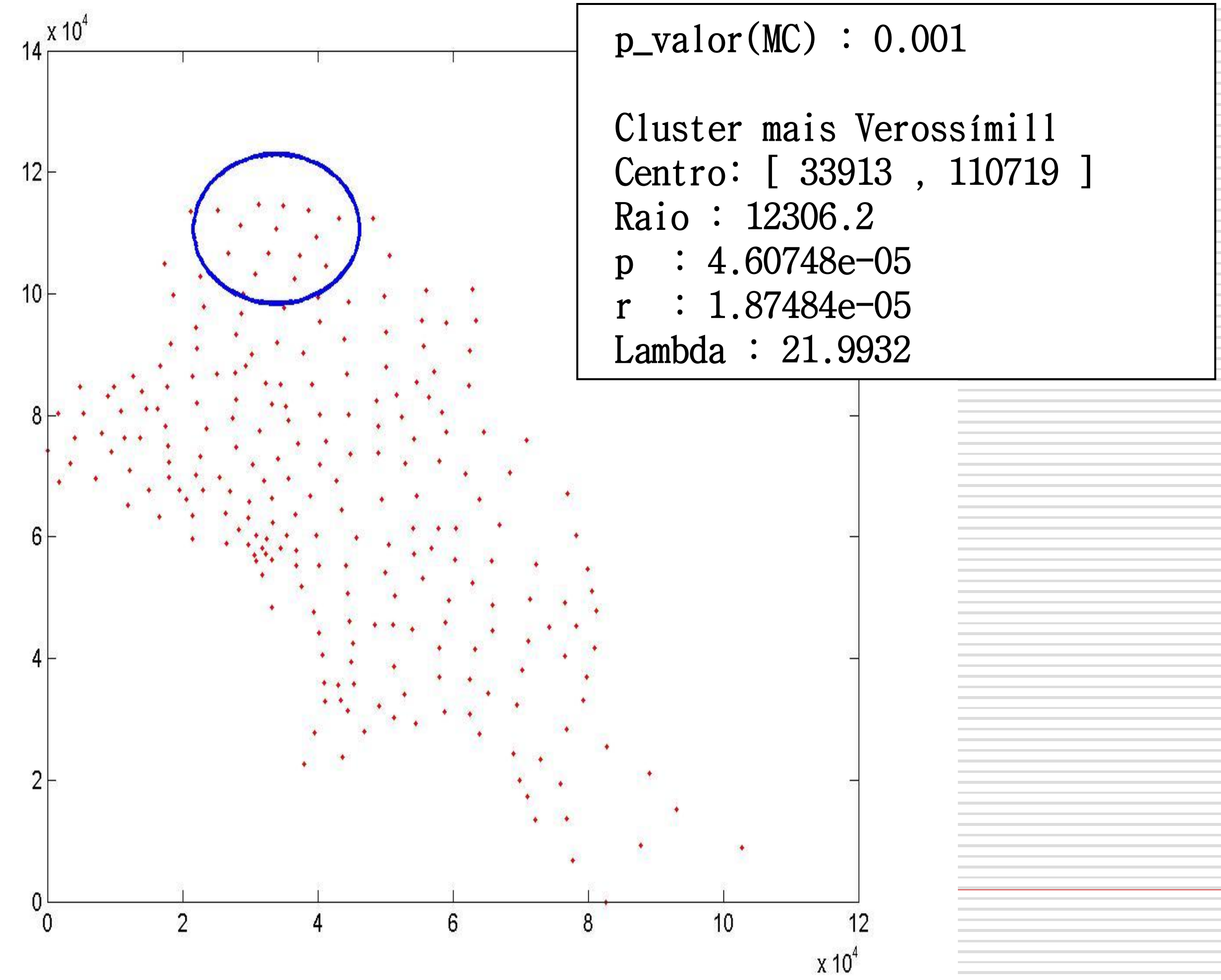


Medical Imaging

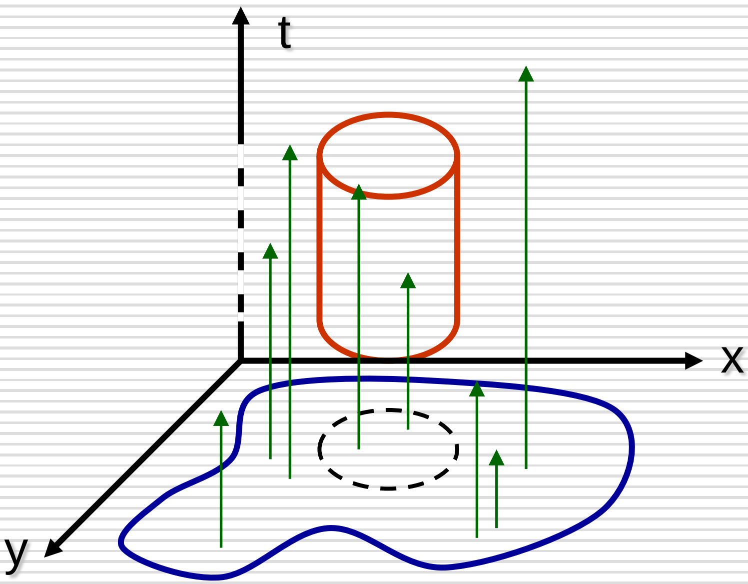
Popescu, L.M. amd Lewitt, R. M. (2006) Small nodule detectability evaluation using a generalized scan-statistic model. *Physics in Medicine and Biology*, **51**, 6225-6244.



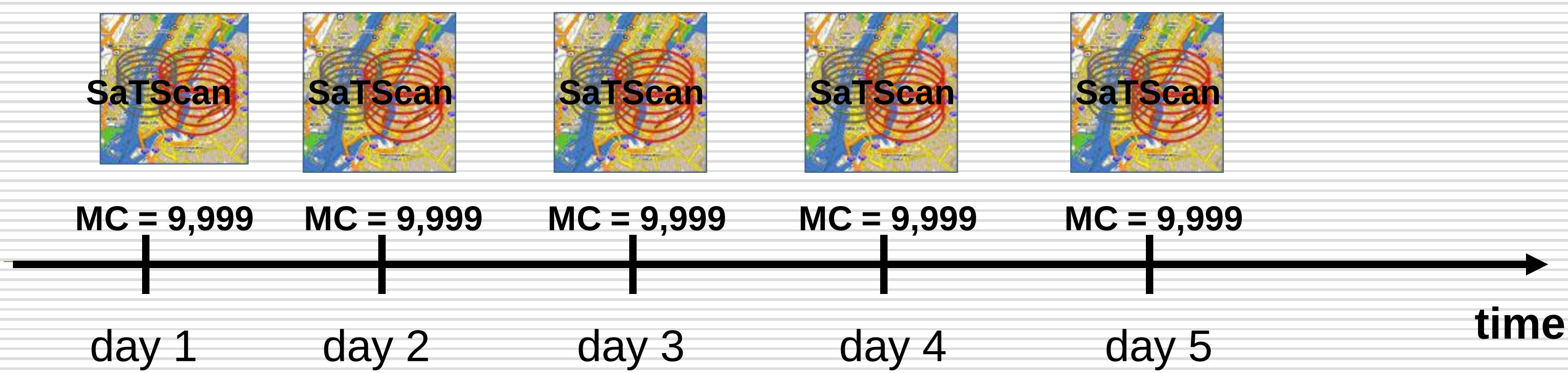
Um exemplo do método de Scan (dados de área)



O Problema da Simulação de Monte Carlo na Varredura ***Scan.***

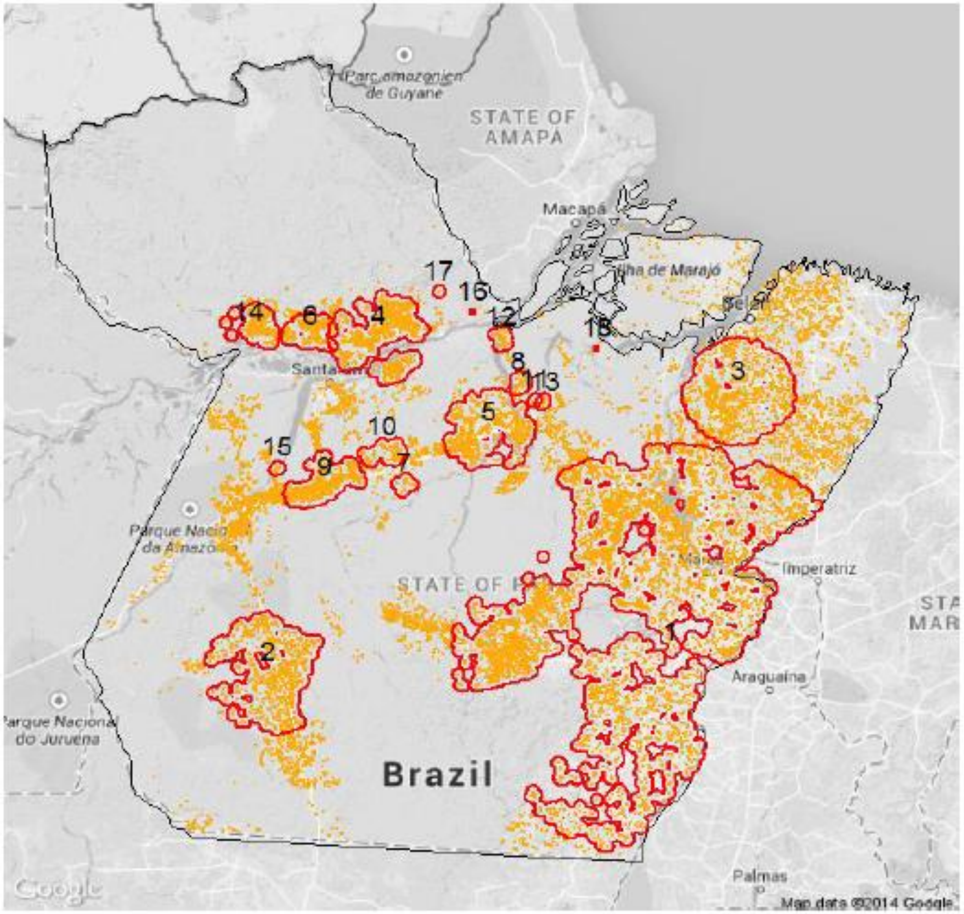


- Em algumas aplicações como em vigilância epidemiológica, os dados são atualizados diariamente.
- O método de varredura é executado diariamente e para uma variedade de doenças.

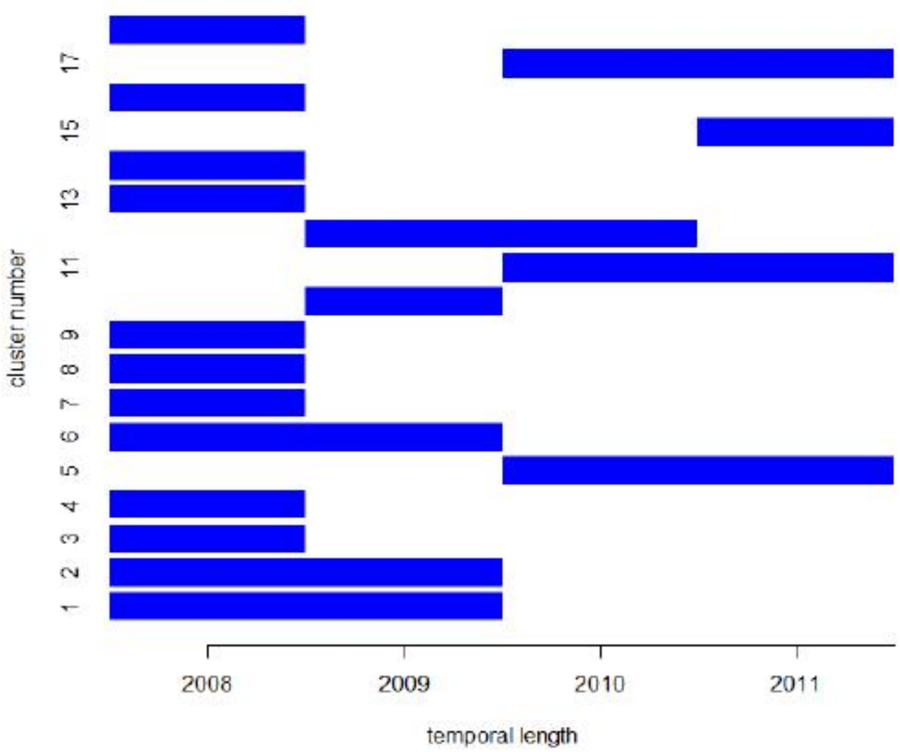


Epidemiologically inspired approaches to land-use policy evaluation: lessons from the Rural Environmental Registry (CAR) in the Brazilian Amazon (2017)

Marcelo A. Costa, Raoni Rajão, Marcelo C. C. Stabile, Andrea A. Azevedo, Juliano Correa



(a) Geographical location of detected clusters.



(b) Time range of detected clusters.

Figure 1 Spatial clustering analysis for the state of Pará (PA)

ANÁLISE DE CONGLOMERADOS DE ACIDENTES DE TRÂNSITO UTILIZANDO GOOGLE MAPS E ESTATÍSTICA ESPACIAL

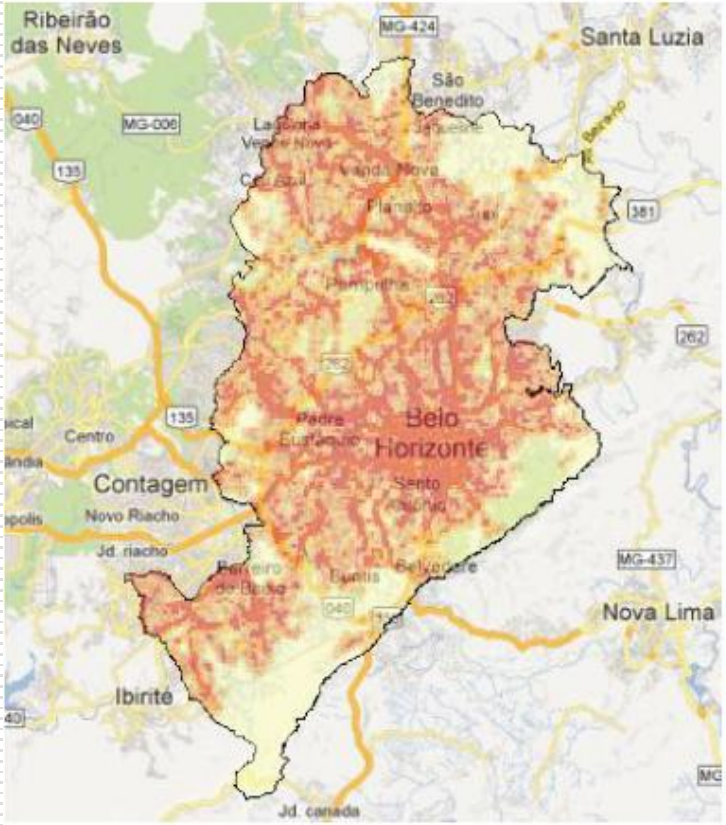
Marcos Antônio da Cunha Santos
Universidade Federal de Minas Gerais
msantos@est.ufmg.br

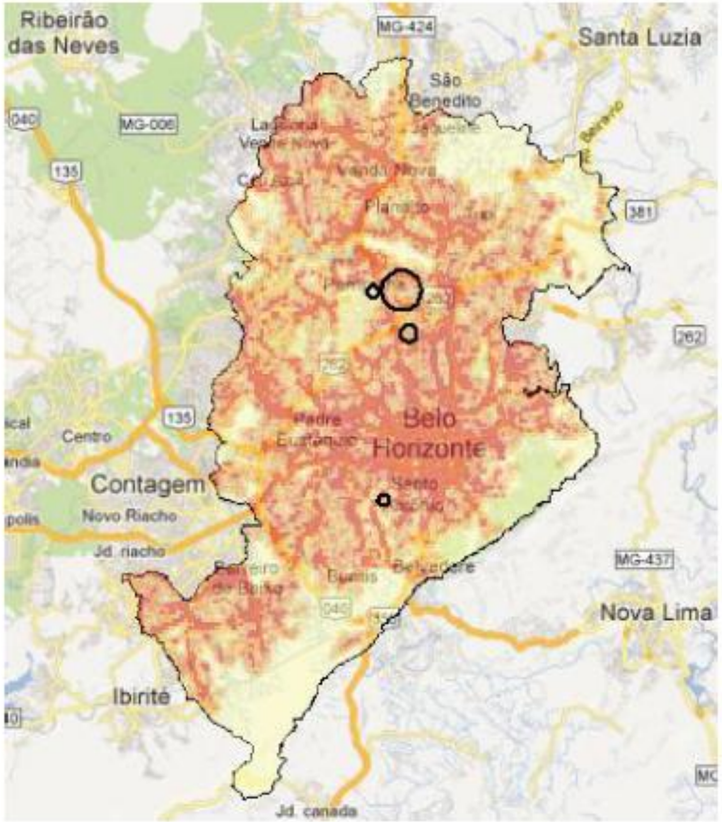
Marcelo Azevedo Costa
Universidade Federal de Minas Gerais
macosta.est@gmail.com

Marcos Oliveira Prates
Universidade Federal de Minas Gerais
marcosop@est.ufmg.br

Cadernos do IME. Série Estatística, v. 35, p. 1-15, 2013.

Clusters identificados utilizando o método SaTScan espacial

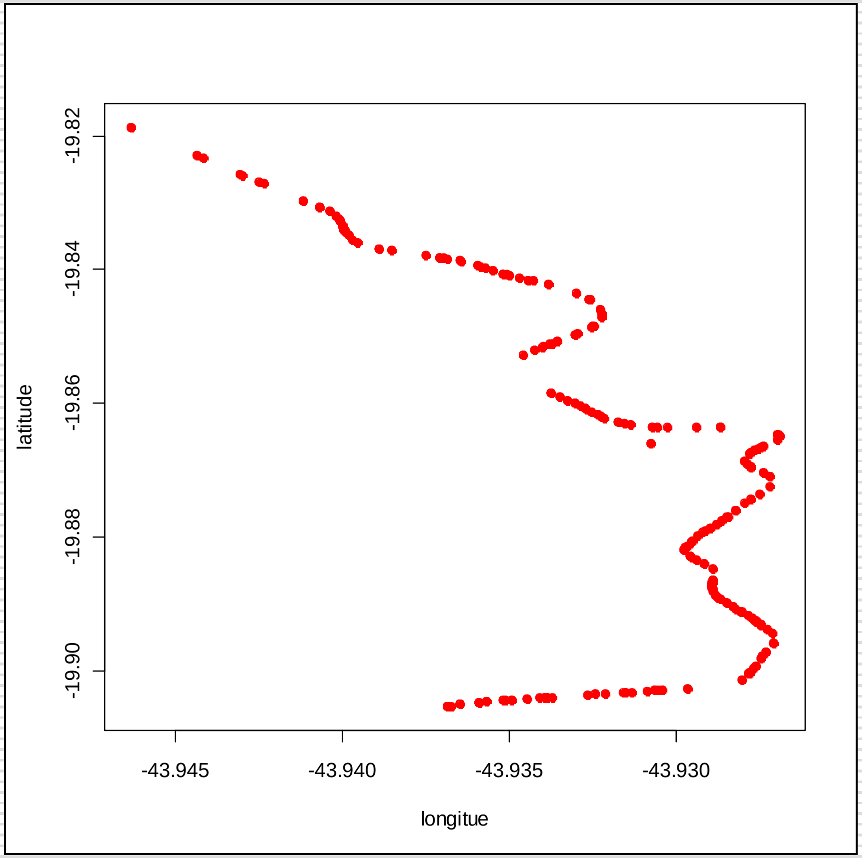




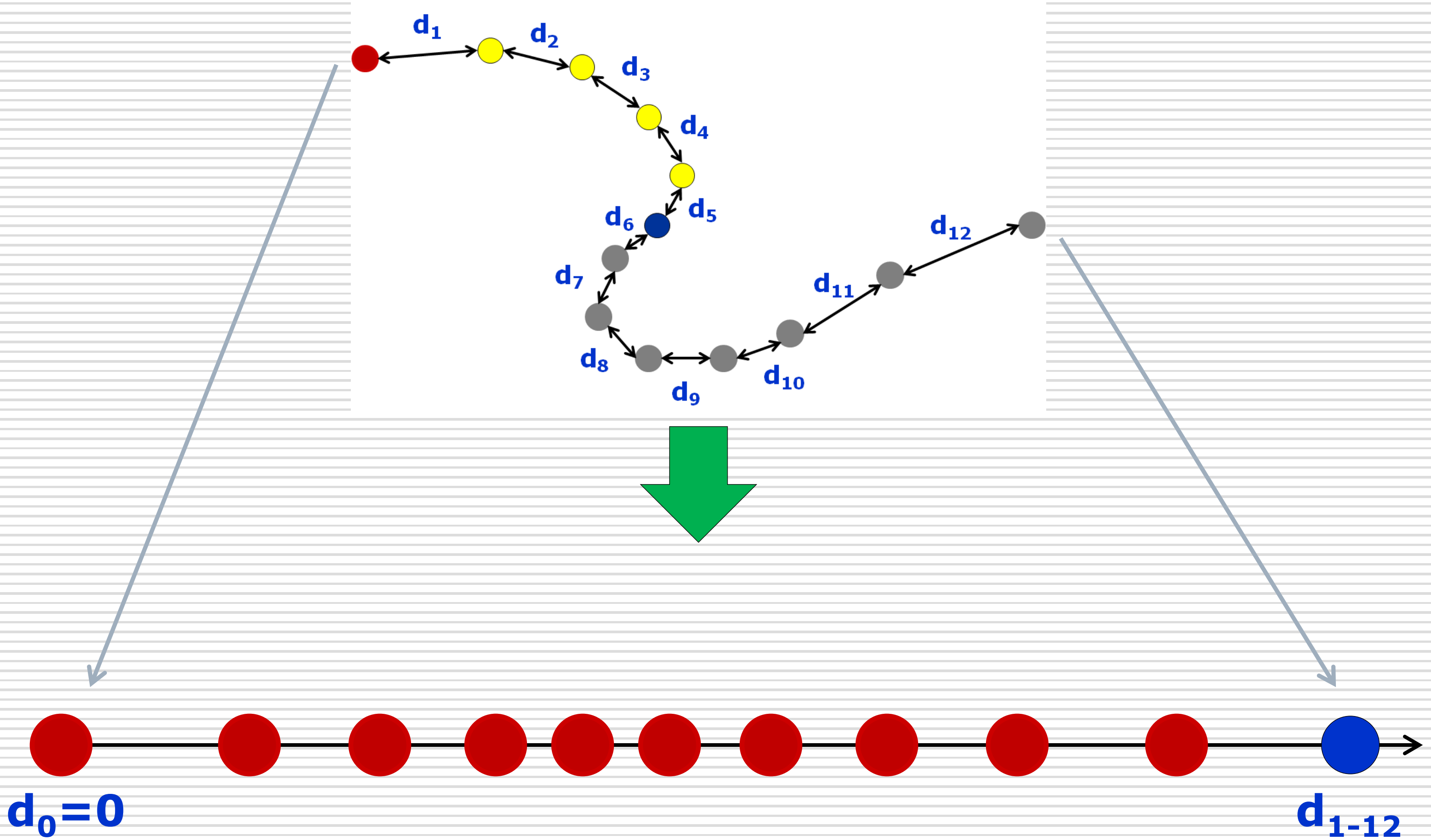
Eventos localizados dentro do círculo e fora do arruamento de interesse são considerados...

Street Scan

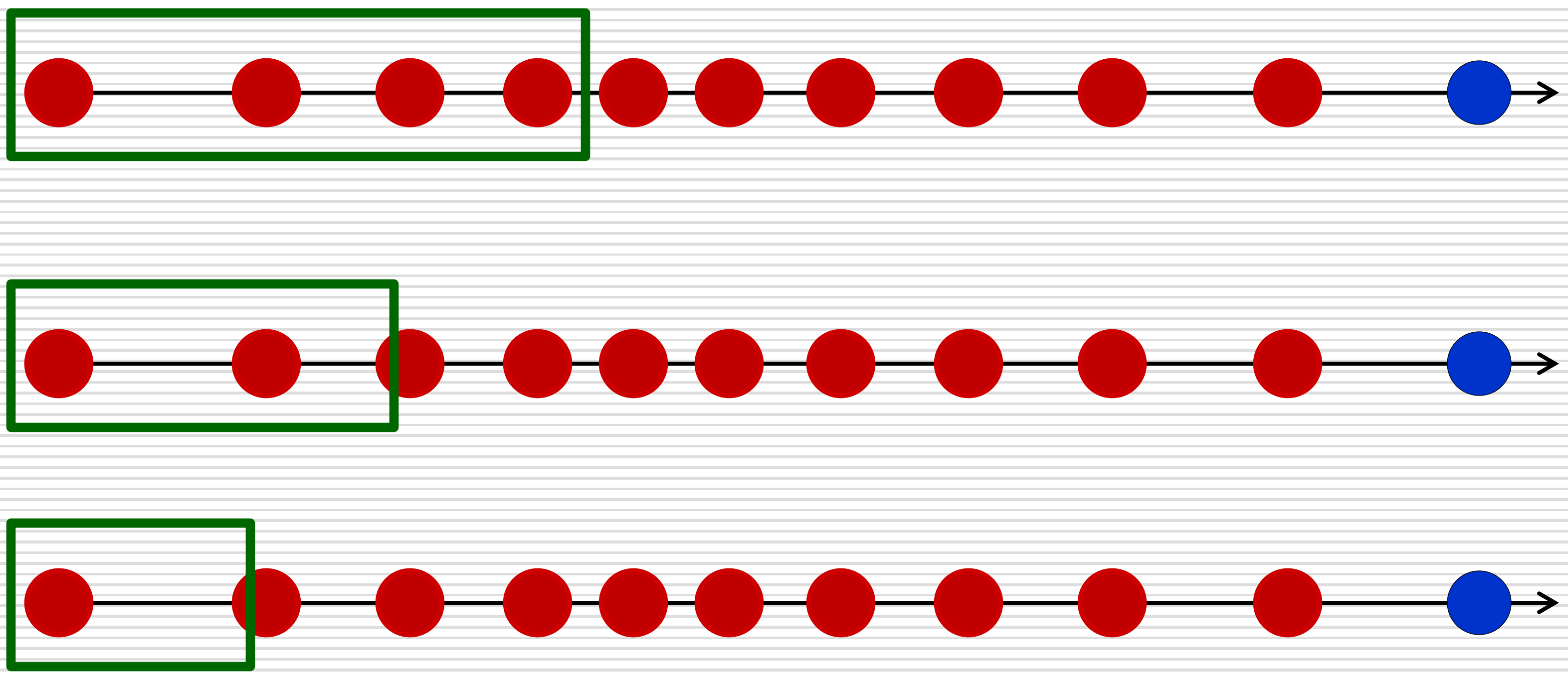

Ribeirão
das Neves
Santa Luzia
Belo
Horizonte
Contagem
Sabará
Nova Lima
Ibirité
Sarzedo
Raposos
Map data ©2012 Google, MapLink


latitude
-19.82
-19.84
-19.86
-19.88
-19.90
longitue
-43.945
-43.940
-43.935
-43.930

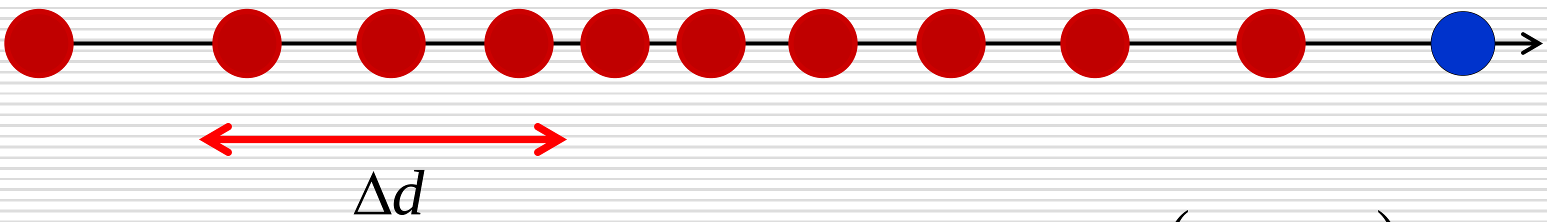
Projetando os pontos da Rua em uma Reta



Street Scan



Street Scan



$$N_{\Delta d} \sim Poisson(\lambda \cdot \Delta d)$$

- O estimador de máxima verossimilhança para λ é:

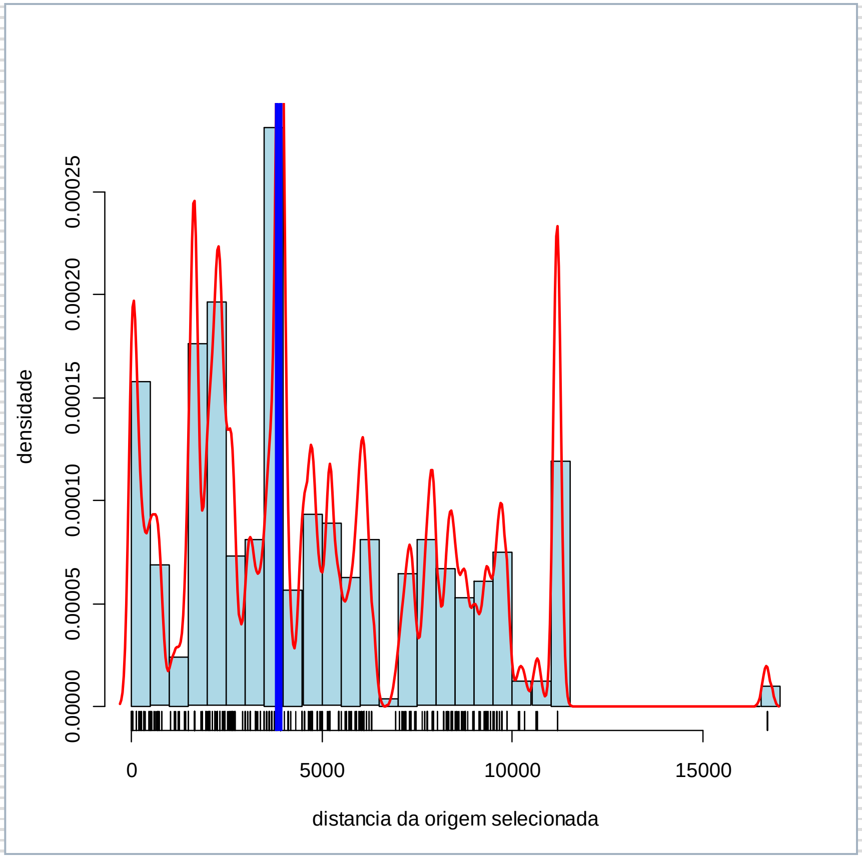
$$\hat{\lambda} = \frac{C-2}{D}$$

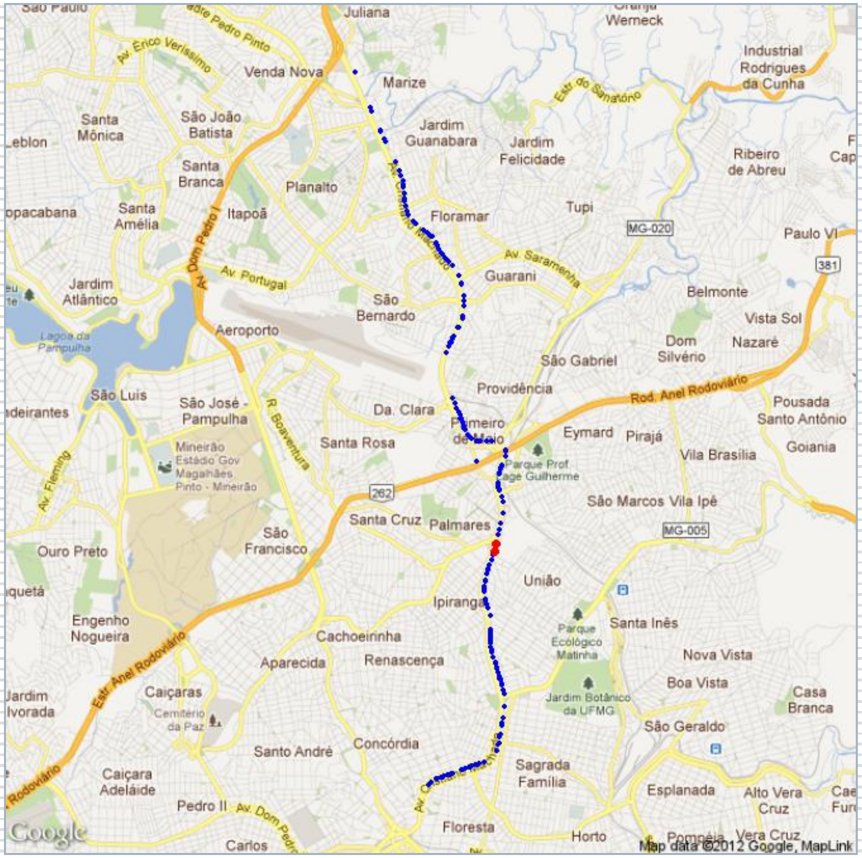
onde **C** é o número total de pontos e **D** é a distância entre o ponto origem e o ponto mais distante na reta.

Simulação de Monte Carlo

- A hipótese nula é a de que os C-2 pontos estão ocorrendo de forma homogênea entre os pontos inicial e final.
  - São gerados **C-2** pontos uniformemente distribuídos ao longo da rua/avenida.
  - A estatística de teste é calculada.
  - O procedimento é repetido inúmeras vezes. (99 ou 999 vezes)

Avenida Cristiano Machado
Belo Horizonte, Minas Gerais





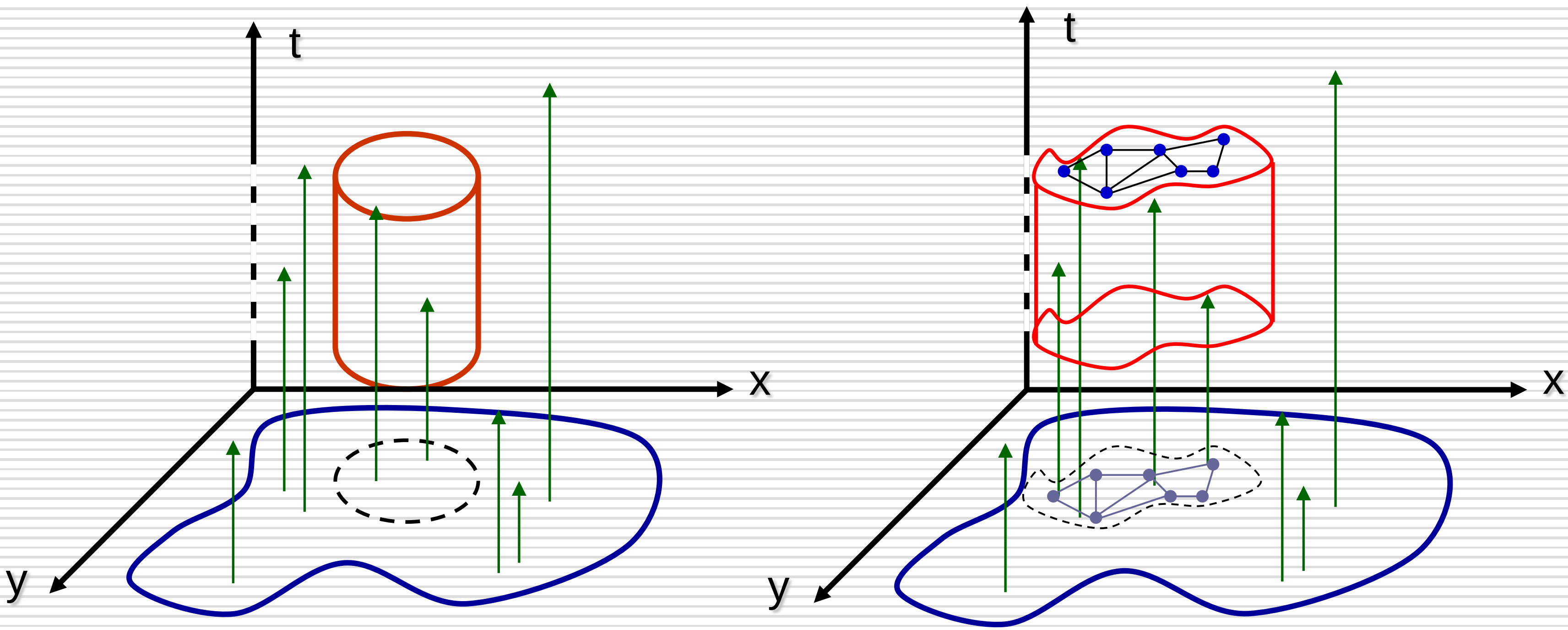
Visualizando os resultados Google Maps

Data: pontos • Chart ID: MapID62e655cd
R version 2.13.1 (2011-07-08) • googleVis-0.2.14 • Google Terms of Use • Data Policy

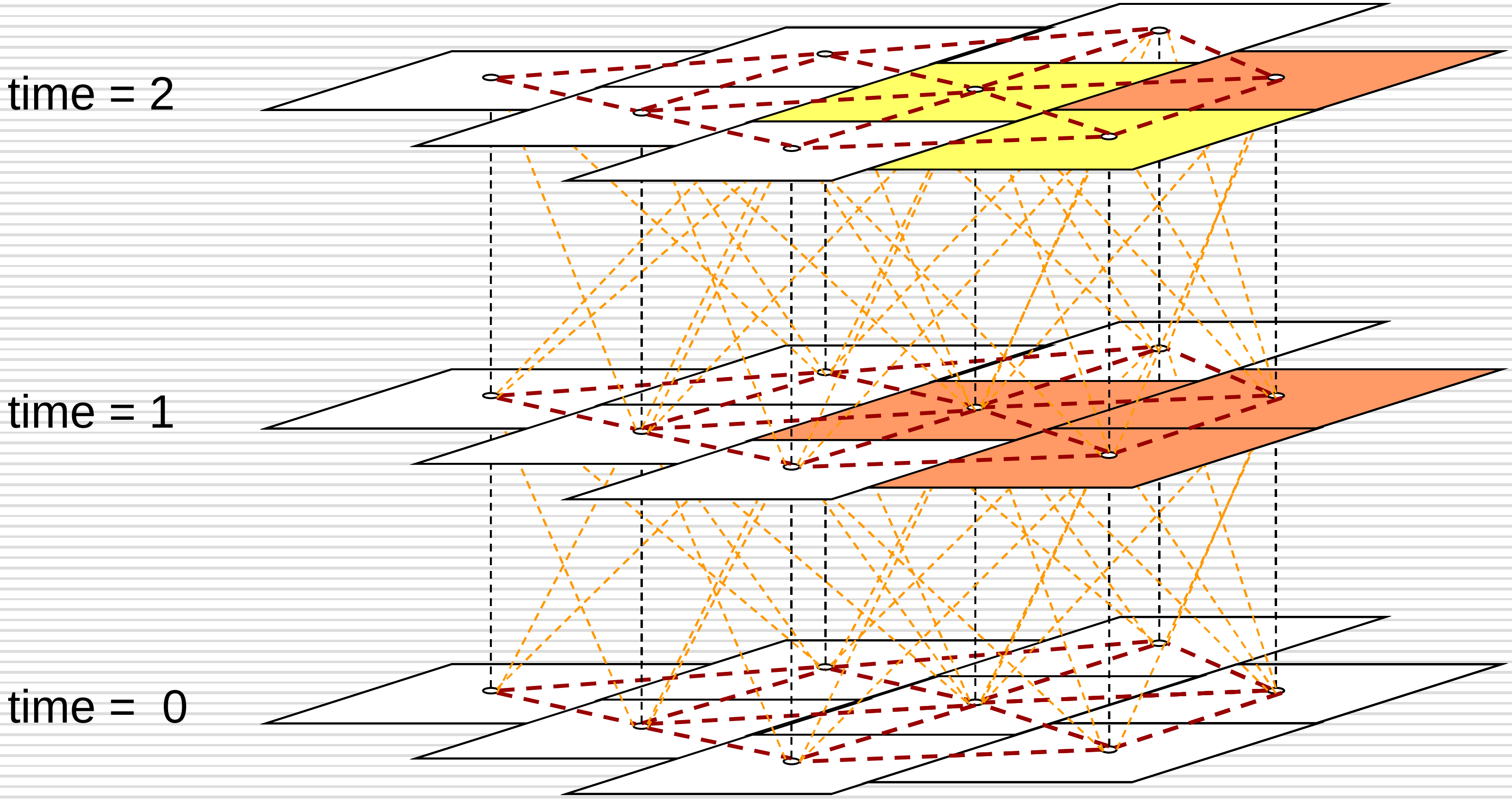
Street View

Data: pontos • Chart ID: MapID3a1c7e5e
R version 2.13.1 (2011-07-08) • googleVis-0.2.14 • Google Terms of Use • Data Policy

Flexible Shapes for Scanning Window



Scanning Graph Structures



Costa and Kulldorff *International Journal of Health Geographics* 2014, **13**:20
http://www.ij-healthgeographics.com/content/13/1/20

INTERNATIONAL JOURNAL OF HEALTH GEOGRAPHICS

METHODOLOGY

Maximum linkage space-time permutation scan statistics for disease outbreak detection

Marcelo A Costa[1*] and Martin Kulldorff[2]

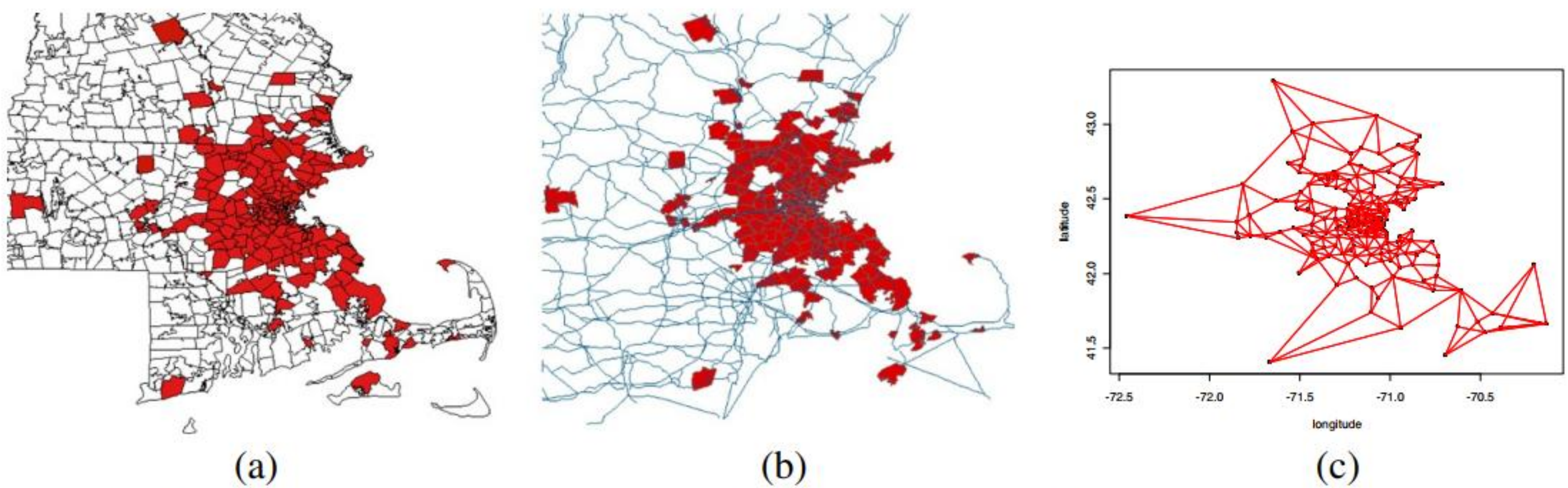
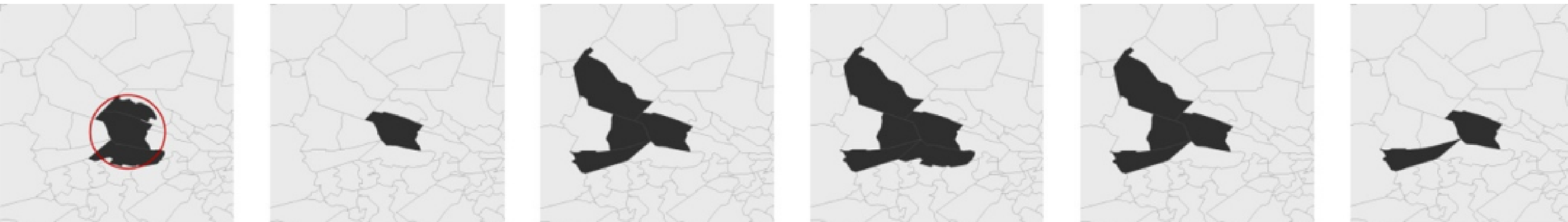


Figure 2 Final graph structure (c) generated using geographic adjacency information (a) and main transportation routes (b).

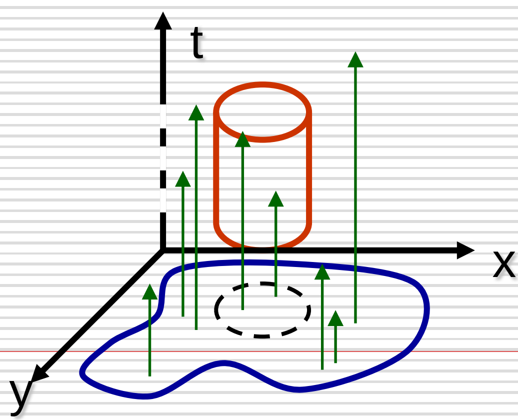
Irregular Clusters…

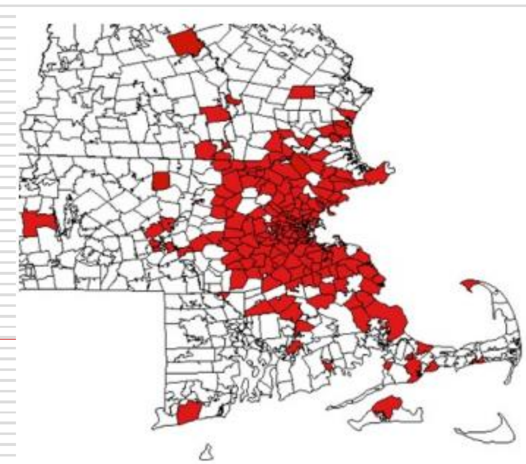
Costa and Kulldorff *International Journal of Health Geographics* 2014, **13**:20 Page 6 of 14
http://www.ij-healthgeographics.com/content/13/1/20



(a) cylindrical scan | (b) *mlink space-time* 28-Aug-2005 | (c) *mlink space-time* 29-Aug-2005 | (d) *mlink space-time* 30-Aug-2005 | (e) *mlink space-time* 31-Aug-2005 to 02-Sep-2005 | (f) *mlink space-time* 03-Sep-2005

Figure 3 Comparison of the clusters detected with the cylindrical and *mlink space-time* permutation scan statistics on 03-Sep-2005. The arbitrary cluster starts with one ZIP code, and it gradually increases until it reaches five ZIP codes. Then the cluster begins to vanishing.





rsatscan

Seguro | https://www.satscan.org/rsatscan/rsatscan.html

rsatscan

Ken Kleinman

2015-02-19

SaTScan is a powerful stand-alone software program that runs spation-temporal scan statistics. It is carefully optimized and contains many tricks to reduce the computational burden of the approach, which is doubly computationaly intensive. First, scanning itself can be costly, particularly in spatio-temporal settings. However, even more difficult, testing involves resampling (Monte Carlo hypothesis testing). For these reasons, it is not worthwhile to attempt replicating SaTScan. In addition, while SaTScan is not open source, it is distributed free of charge.

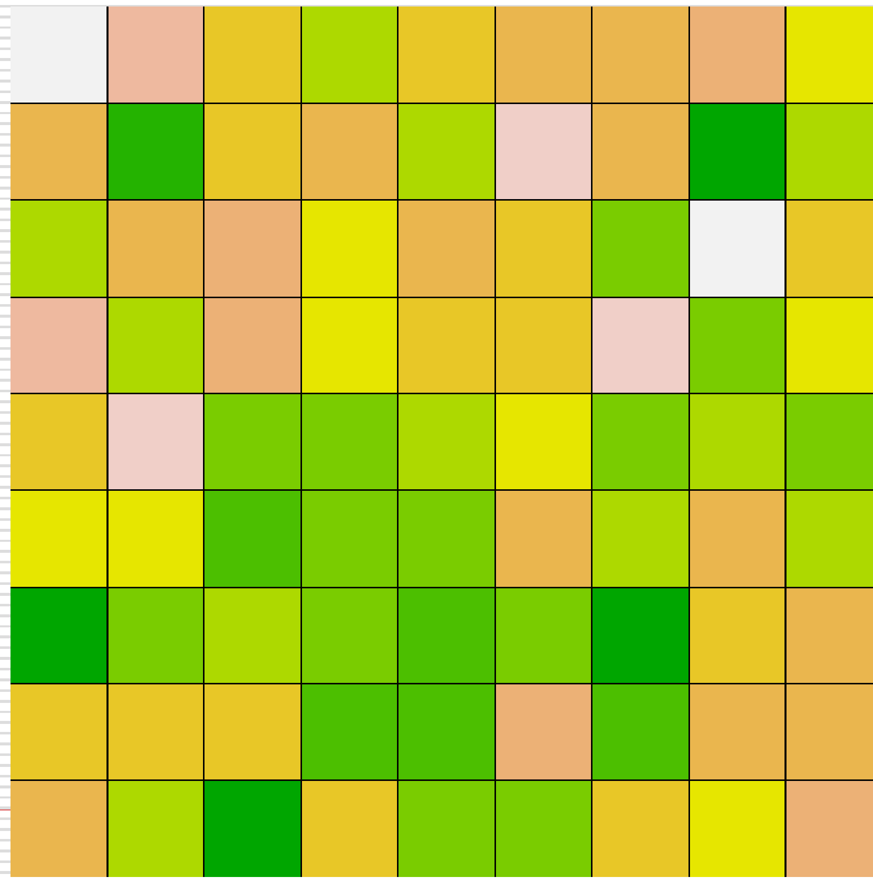
However, use of SaTScan can be cumbersome. There are two means of access: a GUI, and a batch file. The GUI allows complete control, but precludes automated or repeated operation. The batch file allows this, but may be difficult to integrate into other analyses. The `rsatscan` package contains a set of functions and defines a class and methods to make it easy to work with SaTScan from R. This should allow easy automation and integration.

The functions in the package can be grouped into three sets: SaTScan parameter functions that set parameters for SaTScan or write them in a file to the OS; write functions that write R data frames to the OS in SaTScan-readable formats; and the `satscan()` function, which calls out into the OS, runs SaTScan, and returns a `satscan` class object. Successful use of the package requires a fairly precise understanding of the SaTScan parameter file, for which users are referred to the SaTScan manual.

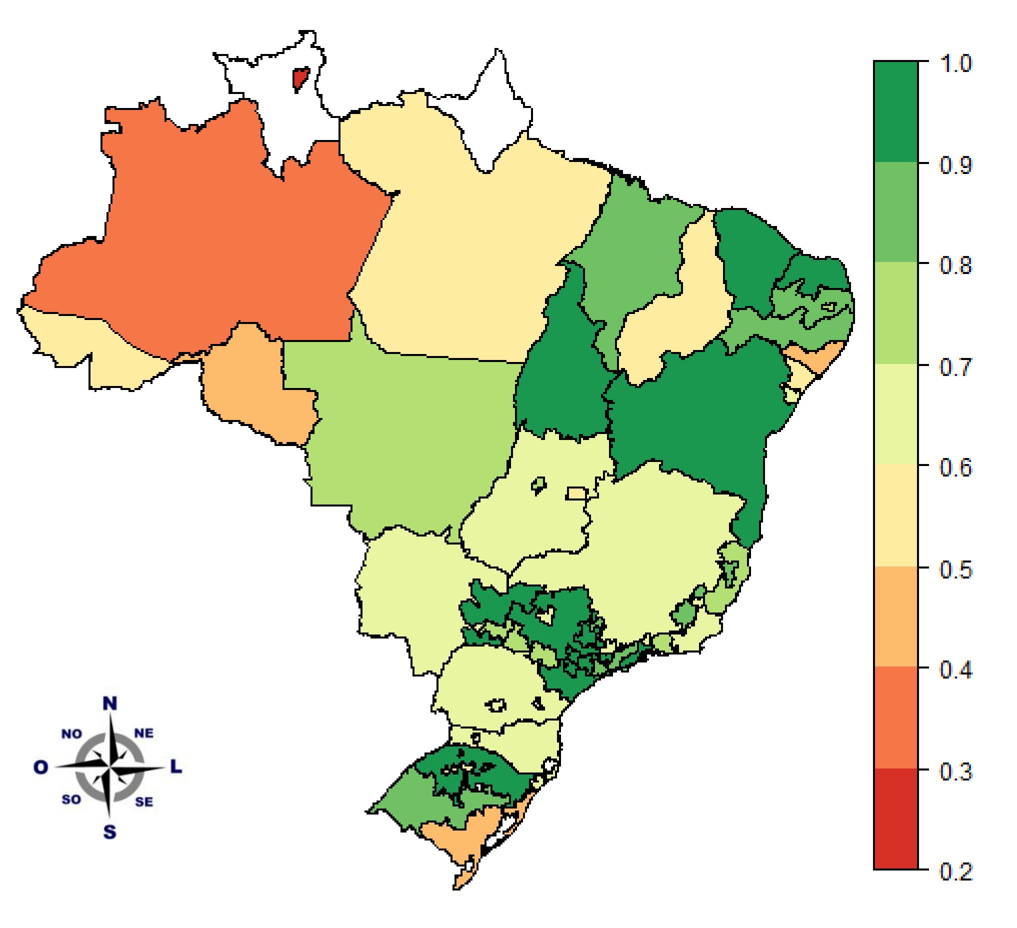
```
library("rsatscan")
```

Novo Problema:
Cluster detection...

Quantos clusters existem neste grid?



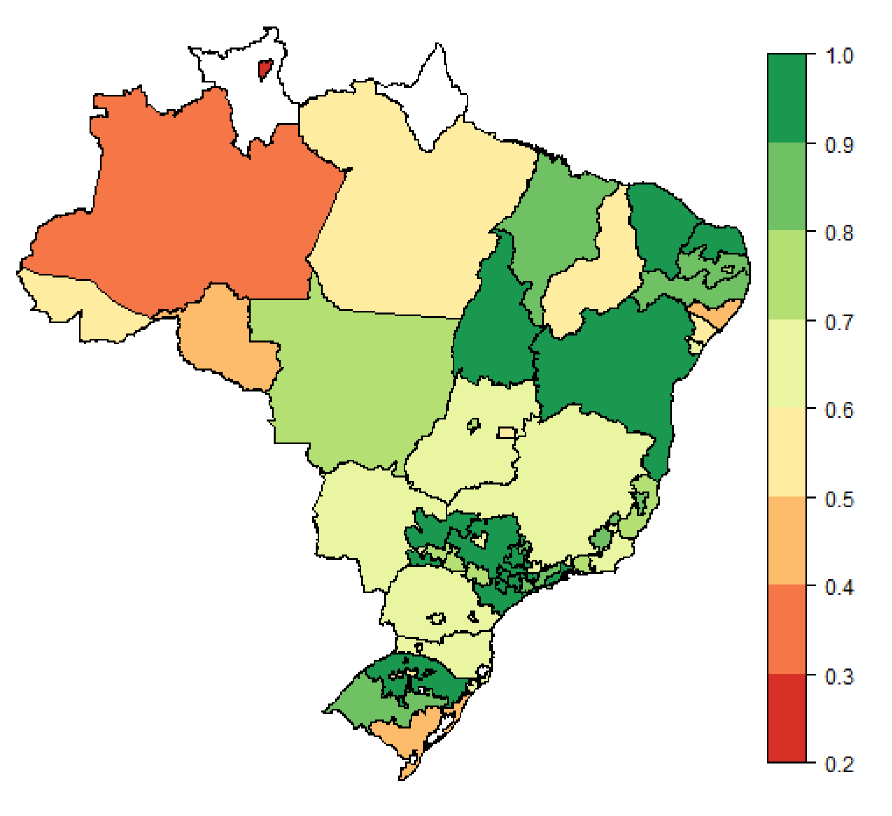
The Estimated efficiency scores across the country



Descriptive statistics of the DEA efficiency scores

Minimum	22.46%
1rst Quantile	56.20%
Median	70.09%
Mean	71.31%
3rd Quantile	86.93%
Maximum	100%

Clustering efficiencies in Brazil



On going work…

- Are the estimated efficiencies grographically clustered?
- How many groups are there?
- Where are the most efficient utilities?

Knorr-Held and Gunter Raber, ***Biometrics*** (2000).

BIOMETRICS 56, 13–21
March 2000

Bayesian Detection of Clusters and Discontinuities in Disease Maps

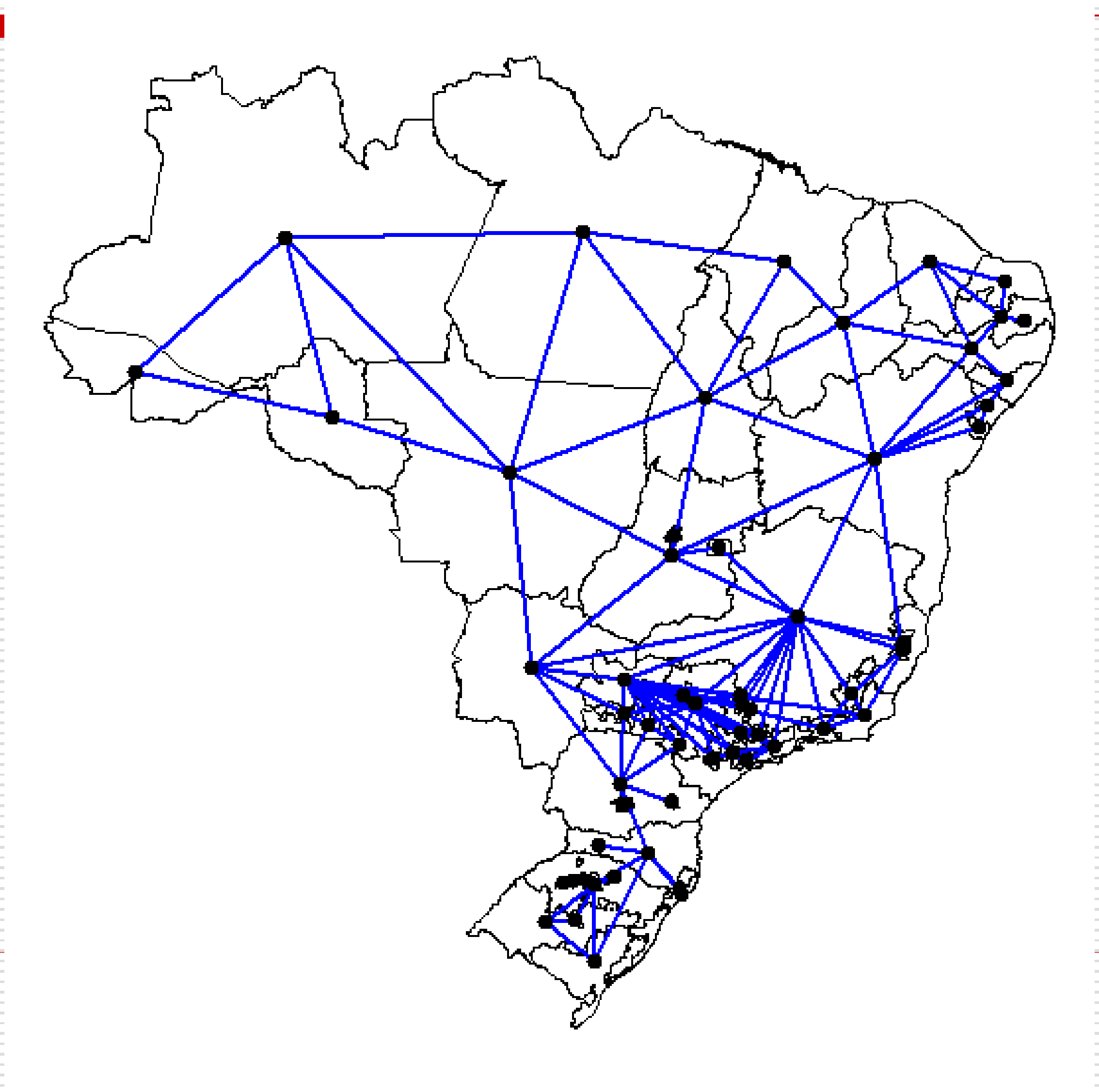
Leonhard Knorr-Held* and Günter Raßer**

Institute of Statistics, University of Munich,
Ludwigstrasse 33, 80539 Munich, Germany
* *email:* leo@stat.uni-muenchen.de
** *email:* rasser@stat.uni-muenchen.de

SUMMARY. An interesting epidemiological problem is the analysis of geographical variation in rates of disease incidence or mortality. One goal of such an analysis is to detect clusters of elevated (or lowered) risk in order to identify unknown risk factors regarding the disease. We propose a nonparametric Bayesian approach for the detection of such clusters based on Green's (1995, *Biometrika* **82,** 711–732) reversible jump MCMC methodology. The prior model assumes that geographical regions can be combined in clusters with constant relative risk within a cluster. The number of clusters, the location of the clusters, and the risk within each cluster is unknown. This specification can be seen as a change-point problem of variable dimension in irregular, discrete space. We illustrate our method through an analysis of oral cavity cancer mortality rates in Germany and compare the results with those obtained by the commonly used Bayesian disease mapping method of Besag, York, and Mollié (1991, *Annals of the Institute of Statistical Mathematics,* **43,** 1–59).

KEY WORDS: Cancer atlas; Clustering; Disease mapping; Oral cavity cancer; Relative risk; Reversible jump MCMC.

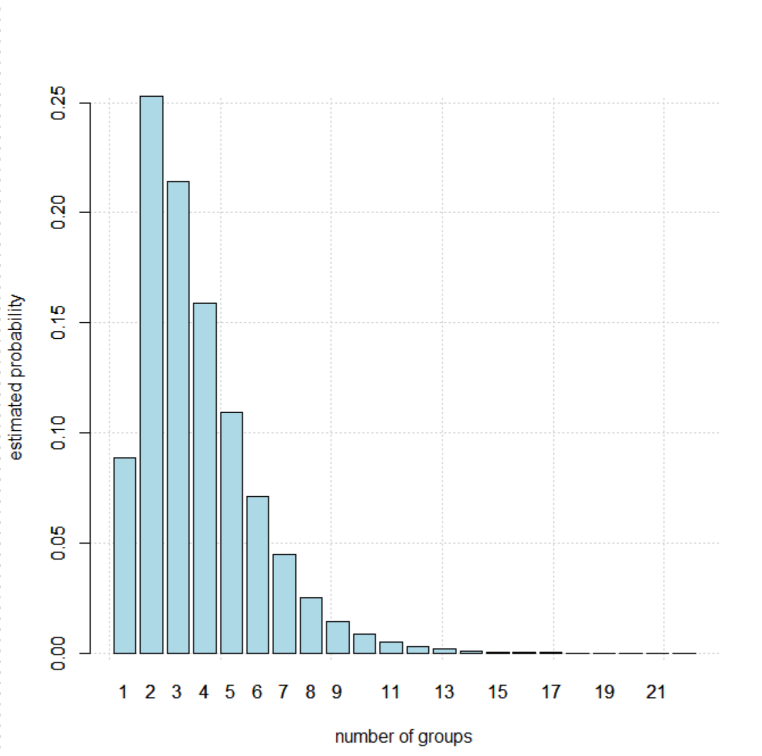
Processo de particionamento do mapa – criação das partições.

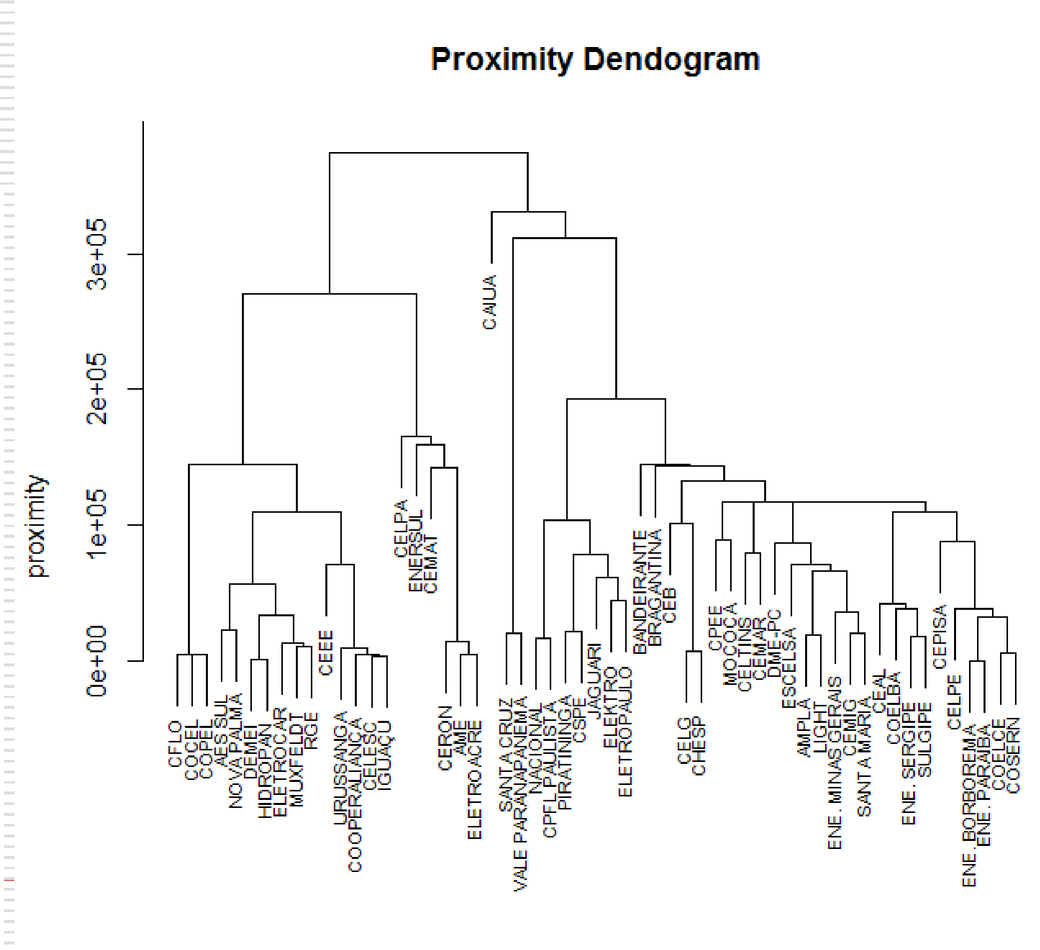


Bayesian detection of clusters in efficiency score maps: an application to energy regulation.

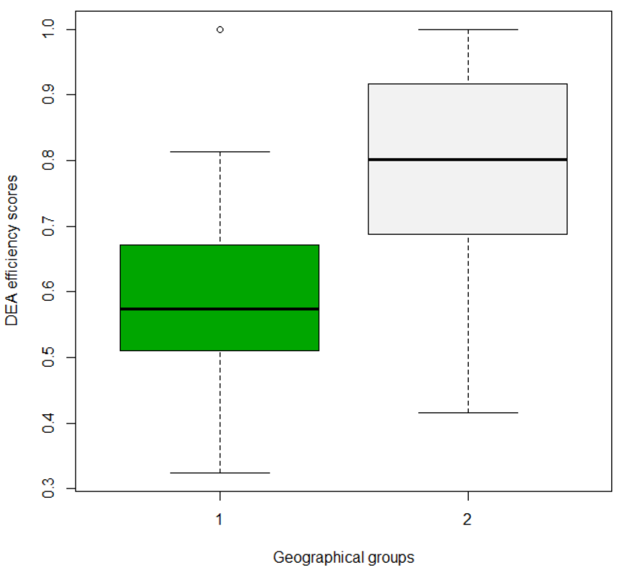
Estimated Probability of the number of groups

Prob(Number of Groups | Data)

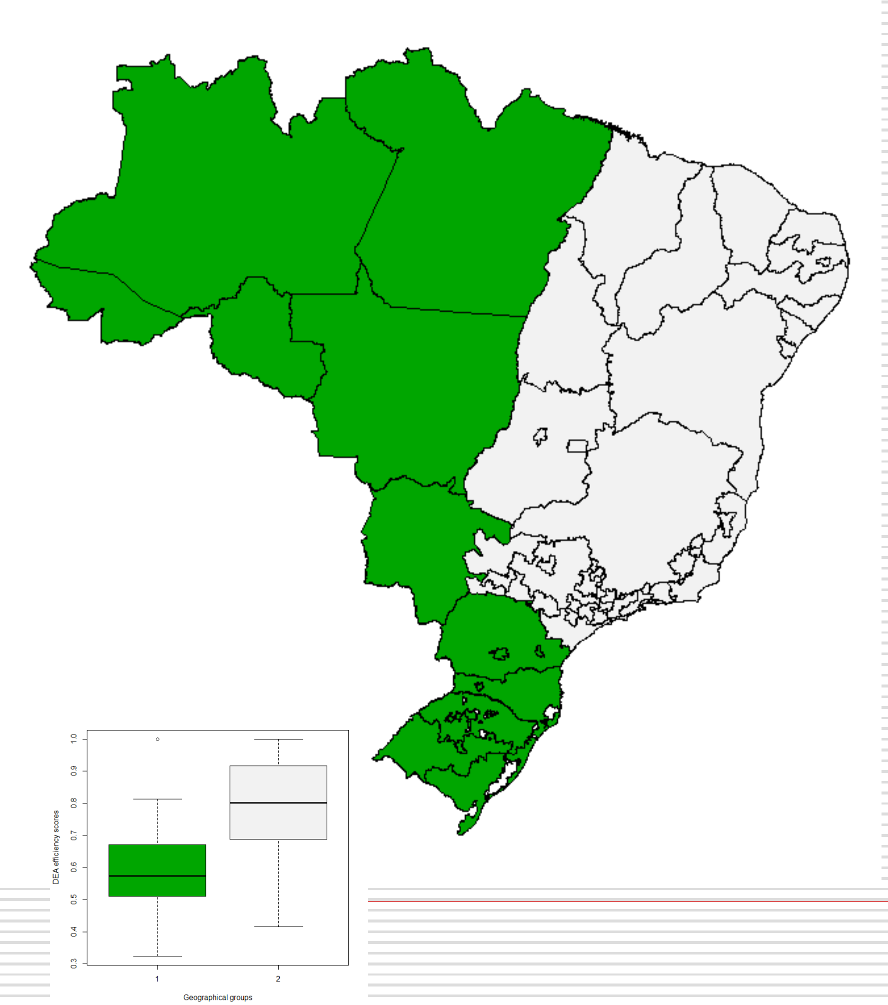


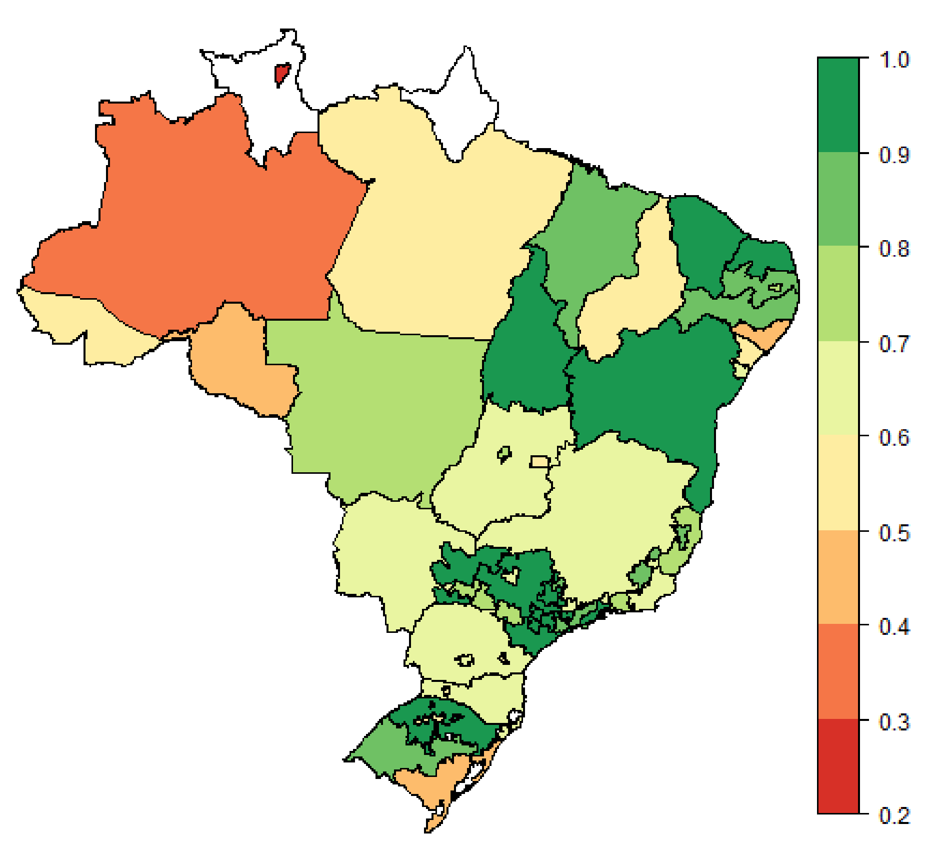


Bayesian detection of clusters in efficiency score maps: an application to energy regulation.

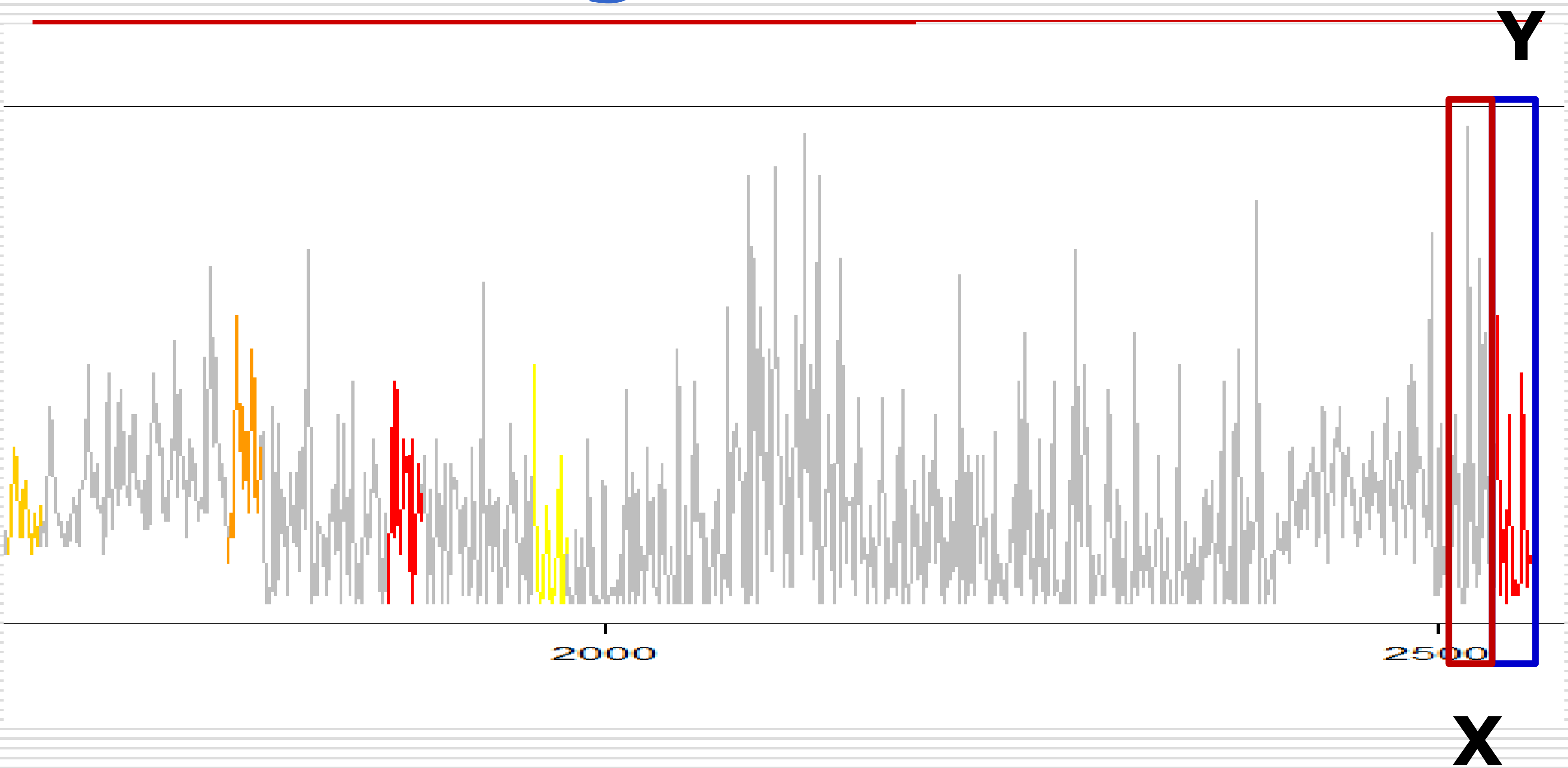


Bayesian detection of clusters in efficiency score maps: an application to energy regulation.



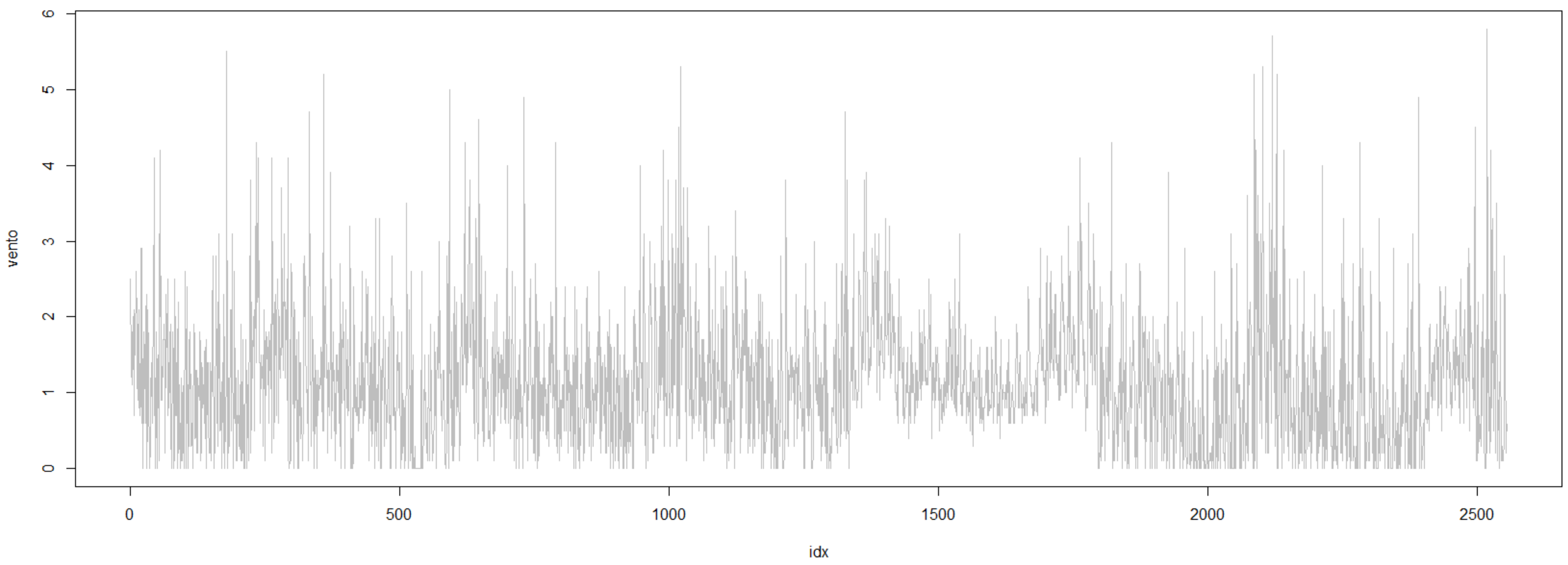


Dynamic Time Scan Forecasting



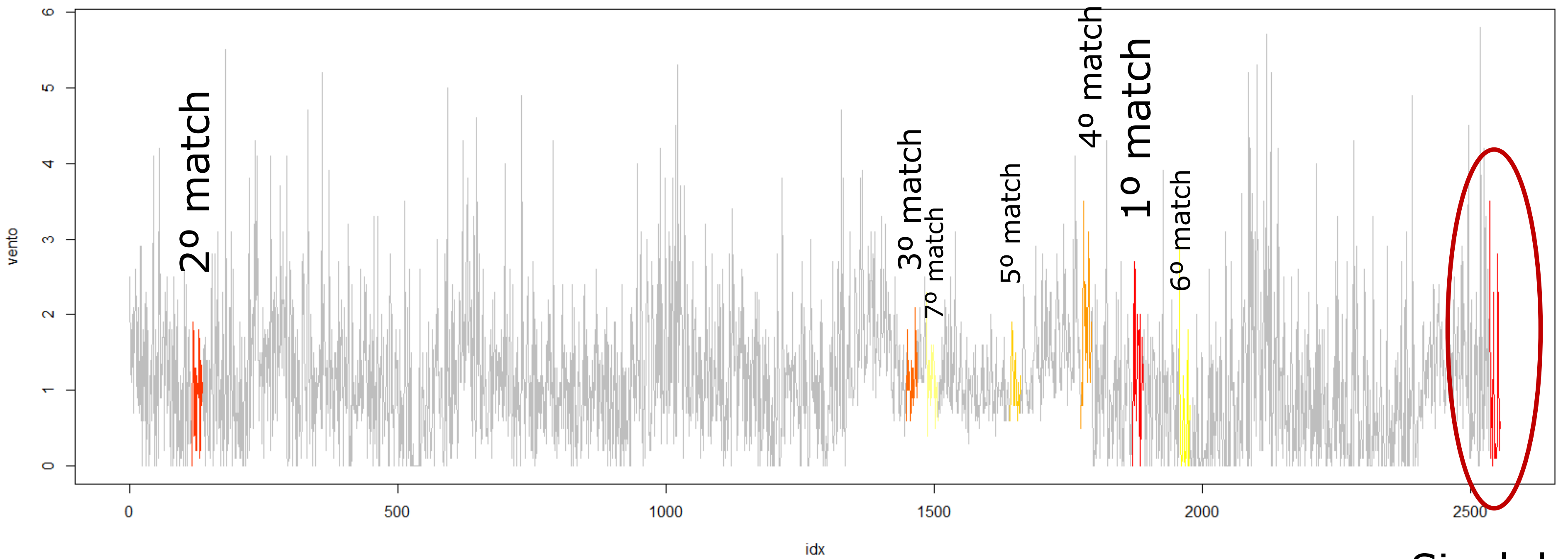
$$Y = f(X)$$

Estudo de caso: Curvelo - MG



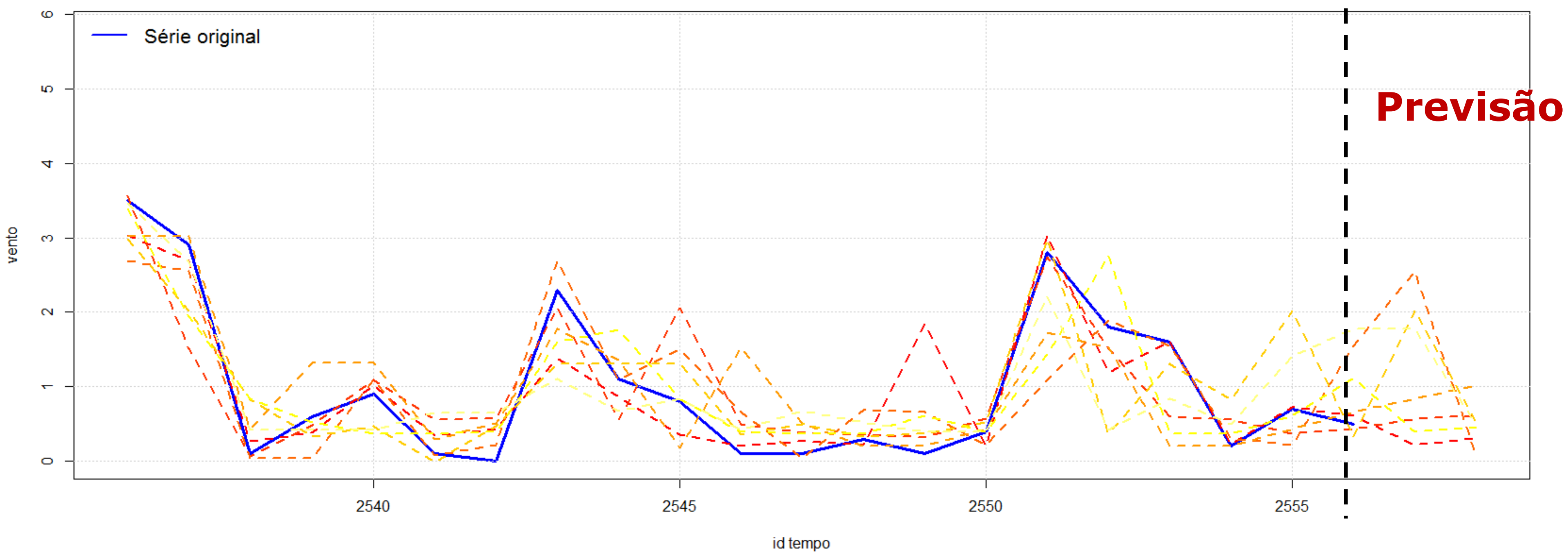

```
janela <- 20      ## Janela de tempo (passado)
best   <- 7       ## Melhores ajustes
k.prev <- 2       ## Passos à frente
```

Estudo de caso: Curvelo - MG



Sinal de interesse

Previsões



$$Y_{t+1} = f(X_{t+1})$$

The 2017 Top Programming Languages…

Language Rank		Types	Spectrum Ranking
1.	Python		100.0
2.	C		99.7
3.	Java		99.5
4.	C++		97.1
5.	C#		87.7
6.	R		87.7
7.	JavaScript		85.6
8.	PHP		81.2
9.	Go		75.1
10.	Swift		73.7