MARCOS OLIVEIRA PRATES

MÉTODO SCAN FLEXÍVEL PARA DETECÇÃO EM ÁRVORES HIERÁRQUICAS

Belo Horizonte
16 de junho de 2008

MARCOS OLIVEIRA PRATES
Orientador: Renato Martins Assunção

MÉTODO SCAN FLEXÍVEL PARA DETECÇÃO EM ÁRVORES HIERÁRQUICAS

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Estatística.

Belo Horizonte
16 de junho de 2008

MARCOS OLIVEIRA PRATES
Advisor: Renato Martins Assunção

FLEXIBLE SCAN METHOD FOR DETECTING CLUSTERS IN A HIERARCHICAL TREE

Dissertation presented to the Graduate Program in Statistics of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Statistics.

Belo Horizonte
June 16, 2008

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Método Scan Flexível para Detecção em Árvores Hierárquicas

MARCOS OLIVEIRA PRATES

Dissertação defendida e aprovada pela banca examinadora constituída por:

Ph. D. Renato Martins Assunção – Orientador
Federal University of Minas Gerais

Ph. D. Marcelo Azevedo Costa – Co-orientador
Federal University of Minas Gerais

Ph. D. Sabino José Ferreira Neto
Federal University of Minas Gerais

Ph. D. Fábio Prates Machado
University of São Paulo

Belo Horizonte, 16 de junho de 2008

Dedico esse trabalho ao meu irmão Henrique Oliveira Prates que esteve presente em minha vida da melhor maneira possível, com ensinamentos, amizade, carinho e exemplos. E que até hoje acredito e sinto sua presença ao meu lado.

Acknowledgments

Primeiramente agradeço aos meus familiares pelo apoio. Não só na caminhada ao mestrado, mas por terem me ensinado muito além disso, durante todo o meu percurso na vida. Meus pais, Antonio Augusto e Maria Helena, pelo exemplo, carinho, força e ensinamentos. Meus irmãos, Raquel e Henrique, que sempre se acharam um pouco meus pais, e sempre significaram muito em amizade, afeto, ajuda. A Bruna, que de maneira inesperada e nova entrou na minha vida e participa comigo dessa caminhada e contribui para meu crescimento amoroso, pessoal e acadêmico. Aos meus cunhados Elaine e João Luiz, membros queridos da família, que dividem comigo tanto os momentos bons quanto os complicados. Aos meus sobrinhos e afilhados João Augusto, Isabela e Celina pela alegria, energia, astral e carinho, com que preenchem nossas vidas.

Agradeço aos meus orientadores, Renato Assunção e Marcelo Costa, pelos ensinamentos estatísticos e, muito além deles, pelo empenho na pesquisa, o incentivo, a amizade no trabalho (e fora dele), por acreditarem em mim e ajudarem-me na realização de sonhos profissionais. Abro aqui um agradecimento especial para Andrea Iabrudi que me introduziu à estatística e abriu as portas para um novo mundo que aprendi a apreciar e gostar muito.

Não quero esquecer-me de agradecer aos que, além dos meus orientadores, contribuíram para que o meu trabalho de dissertação se tornasse possível: João Luiz, por ter perdido algumas horas de sua vida programando comigo quando os problemas do meu algoritmo pareciam insolucionáveis. Camilla, pelas aulas de latex extremamentes necessárias na reta final. Os companheiros do LESTE, em especial, Danilo, Alexandre, João, Aline e Simone, por estarem sempre presentes, ajudando, solucionando dúvidas ou simplesmente descontraindo o ambiente quando as coisas ficavam difíceis.

Por último, mas não menos importantes, agradeço aos amigos que participam da minha vida de forma influente e, em muitos momentos, ajudam-me a superar os obstáculos que surgem no percurso. Aos amigos de CP e COLTEC, com destauqe para Lívia, Luiz, Tadeu, Thiago, Fernando, Daniel, Júlia, Letícia, Dani, Ana Paula, Lucão, Rudá. Aos "agregados" do CP e COLTEC, Camilla, Bob, Maíra, Gustavo Gazzola. A todos os meus primos e, em especial, Dani(minha madrinha), Mateus, Fernanda, Bel,

Lu, Gustavo, Ian, João Paulo. Aos tios e tias, sempre presentes.

Resumo

Esse trabalho apresenta um eficiente algoritmo de varredura para bancos de dados hierárquicos que podem ser representados na forma de árvores. O algoritmo procura através dos galhos da árvore e é capaz de agregar folhas em diferentes galhos. A varredura procura por um cluster candidato através da estatística Minimum Description Length (MDL). A estatística de teste combina o logaritmo da razão de verossimilhança e a quantidade de informação necessária para representar internamente o cluster. Esse segundo termo controla os graus de liberdade do algoritmo de busca. Fazendo isso, a metodologia previne o acréscimo de folhas que desnecessariamente aumentem o termo do logaritmo da razão de verossimilhança. Resultados mostram que a metodologia MDL é um algoritmo flexível capaz de detectar clusters em bancos de dados hierárquicos nos quais os elementos do cluster estão distribuídos pela árvore. Dessa forma, o algoritmo explora grupos de clusters que não são explícitos simplesmente olhando para cortes nos galhos ou em análises combinatórias dos dados fornecidos.

Abstract

This work presents an efficient scan algorithm for hierarchical data sets that can be represented as a tree structure. The algorithm searches through the branches of the tree and it is able to aggregate leaves located in different branches. The scan search aims at detecting a cluster candidate with Minimum Description Length (MDL) statistic. The test statistic combines the log-likelihood and the amount of information necessary to represent internally the cluster. This second term controls the degree of freedom of the search algorithm. By doing this, the methodology prevents the algorithm to add leaves that unnecessarily increase the log-likelihood term. Results show that the MDL methodology is a flexible algorithm which detects clusters in hierarchical data sets whose cluster elements are partially scattered in the tree. Therefore, it explores potential groups of clusters that are not explicit by just looking at cuts of branches or combinatorial analysis of the given data.

Resumo Estendido

Muitas informações biométricas podem ser organizadas de forma hierárquica. O exemplo mais antigo é a classificação de Linnaean para seres vivos. O sistema classifica todos os organismos de uma forma hierárquica, sucessivamente selecionando domínio, reino, filo, classe, ordem, família, gênero e espécie para cada organismo vivo. Classificações hierárquicas também podem ser aplicadas para coisas não vivas, por exemplo, organizações de saúde podem ser classificadas de acordo com o aumento do nível da escala geográfica, tais como hospitais classificados por estado, cidade, distritos e bairro.

Profissões também podem ser naturalmente organizadas de uma maneira hierárquica, agrupando profissões relacionadas por áreas de trabalho. Estas áreas, por sua vez, são aninhadas em áreas de trabalhos mais gerais, determinando assim, os diversos níveis da árvore. citeKulldorffFangWalsh desenvolveram um método de vigilância para detecção de clusters em árvores hierárquicas e aplicaram-no na busca por sub-conjuntos de profissões com evidências incomuns de taxas de mortes por doenças relacionadas à profissão. Sem idéias preconcebidas sobre qual profissão específica ou grupo de profissões podem estar relacionadas com o maior risco, eles estavam interessados em decidir qual era o cluster de profissões onde o número de casos era particularmente freqüente. Ao invés de procurarem entre todas as possíveis combinações, eles consideraram uma menor classe de clusters possíveis. De maneira geral, os clusters são formados por apenas uma profissão ou todas as profissões pertencentes a um determinado galho. Isso reduz substancialmente o número de clusters a serem procurados. Ao mesmo tempo, essa classe é grande o suficiente para conter os muitos clusters que naturalmente ocorrem.

A motivação do nosso trabalho é propor um método de varredura mais flexível que nos permita que uma maior classe de clusters seja analisada. Essa classe é maior do que a proposta por Kulldorff et al. (2003) e inclui casos onde apenas algumas profissões específicas, pertencentes a dois ou mais galhos, possuem maior risco do que o restante das profissões. E estas profissões de maior risco, por sua vez, estão aninhadas em diferentes galhos da árvore. Nosso método de varredura é capaz de detectar tais clusters (profissões de maior risco), enquanto o método de Kulldorff et al. (2003) necessariamente terá de incluir todas as profissões dos galhos e não somente as profissões que

especificamente possuem maior risco.

Em nosso método, utilizamos o princípio de seleção de modelos Minimum Description Length (MDL) (Rissanen (1989)). No nosso caso, o princípio MDL combina o logaritmo da razão de verossimilhança como uma medida de adequação e a quantidade de informação necessária para determinar os caminhos na árvore, que especificam o cluster. Essa quantidade é baseada no número de bits necessários para descrever o cluster candidato (o que é analisado no momento) na árvore. Esse segundo termo controla os graus de liberdade no algoritmo de busca, prevenindo que o método adicione profissões desnecessárias que aumentem o logaritmo da razão de verossimilhança.

Resultados mostram que o poder do método aumenta quando os riscos relativos dos clusters também aumentam. Além disso, os clusters detectados normalmente incluem elementos pertencentes aos clusters reais (clusters criados para a simulação). À medida que o risco relativo aumenta, o método detecta uma maior quantidade de elementos do cluster real. Apesar desse fato, quando o cluster criado está espalhado pela árvore, a performance do método é comprometida.

Aplicando o método a dados reais(i.é, dados coletados por algum órgão oficial), usamos, como exemplo, dados onde mortes por silicose eram os casos observados e divididos por um fator de risco (onde profissão é o fator de risco). O método encontrou um grupo de profissões que provadamente possuem características de maior risco da doença.

Dessa forma, esse trabalho mostra que a metodologia MDL é um novo e flexível método para detectar clusters em bancos de dados hierárquicos, nos quais os elementos do cluster estão parcialmente espalhados pela árvore. Portanto, o método potencialmente explora grupos de clusters que não são explícitos simplesmente observando cortes nos galhos da árvore em estudo ou em uma análise combinatória dos dados fornecidos.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Much biometric information can be organized in a hierarchical way. The oldest example is the Linnaean classification of all living things. The system classify all the organisms in a hierarchical way by successively selecting a domain, kingdom, phylum, class, order, family, genus and species for each living organism. Hierarchical classification can also be applied to non-living things for example health organizations can be classified according to increasing levels of geographical scale such as the health care providers classified by their state, county, towns and neighborhood locations.

Occupational is also naturally organized as a hierarchical tree by grouping related occupations in ever increasing levels of aggregation. For example, one can cluster together all types of therapists specializations (Respiratory therapists, Occupational therapists, Physical therapists, Speech therapists, Therapist and Physician Assistant) into a single class called Therapists. Next, one groups the *Therapist* class with the *Registered nurses*, *Pharmacists* and *Dietitians* classes to form the *Health Assessment and Treating Occupations* class, and so on into larger groups.

Kulldorff et al. (2003) developed a surveillance method to detect clusters located in a hierarchical tree and applied it to search for subset of occupations with unusual evidence of death rates for occupationally related diseases. In each occupation, the number of deaths in a certain period are separated into two types: those caused by *silicosis* and all the other deaths. Without preconceived ideas about what specific occupation or group of occupations may be related to increased risk, he was interested in deciding if there was a cluster of occupations where death from *silicosis* was particularly frequent. In his work, Kulldorff et al. (2003) used the United States Census Bureau (1982) classification for occupations as a hierarchical tree in 503 categories. An exhaustive search of all possible combination of occupation is not computationally feasible. It is necessary to create a class of combinations that is large enough to contain combinations substantially interesting and, at the same time, this class must be small enough to be

scanned in practical terms.

Another difficulty is the multiple testing problem. There is a need to control the error Type I in a statistical analysis that deals with a large number of comparisons between different groups. These problems led him to consider a scan statistics methodology to detect clusters in the hierarchical tree. The method is adapted from the spatial scan method proposed by Kulldorff (1997). Rather than scanning all possible combinations, he considers a smaller group of possible clusters. In a broad sense, these clusters are those formed by single occupations or all occupations belonging to a specific branch in the hierarchical tree. This reduces substantially the number of clusters to be searched and, at the same time, is large enough to contain many naturally occurring clusters. To control the error Type I, the statistical inference is based on Monte Carlo simulation under the hypothesis that there is no cluster present in the tree. The test statistics null distribution takes into account the multiple testing involved in the analysis.

However, there are situations where a more flexible scan procedure can be important. Consider for example the case of latex allergy. This disease became common among doctors and nurses after the appearance of HIV, when physicians and nurses started to adopt latex gloves to protect from infection. Another group of occupation that share the same problem are cleaners, who use the latex gloves to be protected from the chemical products used in their work. Although both occupations share a higher risk of the same disease they are not close in the occupational tree structure. While doctors and nurses have similar job characteristics, cleaners carry out activities that put them into a different position in the tree structure.

Applying the Kulldorff et al. (2003) scan method, one is likely to miss situations such as a cluster of latex allergy. This is due to the fact that the class of candidate clusters proposed by Kulldorff et al. (2003), include groups located in different positions of the hierarchical tree only by including many other occupations not associated with high risk. These additional occupations will lower the evidence to detect the clusters and the final result is a lower power test.

The motivation for our work is to propose a more flexible scan statistics method that allows for a larger class of possible cluster candidates. The class is larger than that proposed by Kulldorff et al. (2003) and includes clusters similar to that described in the latex allergy situation. Another possibility that is covered in our model but not in Kulldorff et al. (2003) is when only a few occupations within two or more classes have higher risks than the rest and these classes are nested into a larger class. Our scan method can detect such a cluster while for the Kulldorff et al. (2003) scan method will necessarily include all the occupations in the classes, not only the specific occupations with higher risks.

The spatial cluster detection techniques based on the scan statistic has faced a similar challenge. Several authors adopt a more flexible cluster definition by considering non-circular shapes for the candidate clusters (Duczmal and Assunção (2004); Patil and Taillie (2003); Patil and Taillie (2004); Tango and Takahashi (2005); Assunção et al. (2006)). In this spatial context, a large problem is the octopus-shaped solutions the scan method finds. Significant clusters tend to be much larger and oddly shaped as a result of the test, as pointed out by Tango and Takahashi (2005) Any reasonable solution requires some kind of penalized likelihood approach where very peculiar shaped clusters has less interest. Ad hoc solutions has been proposed and no consensus has been reached yet (see Costa et al. (2008); Costa et al. (2005); Tango and Takahashi (2005); Duczmal et al. (2006)).

In our hierarchical tree problem there is a natural solution based on the selection model approach using the minimum description length (MDL) principle (Rissanen (1989)). In our specific case the MDL principle combines a likelihood-based measure of fitness with the amount of information to account for the path in the tree, that specifies the cluster. This amount is based on the number of bits needed to describe a candidate cluster in the tree.

This paper is organized in the following way. In section 2 we describe the scan method proposed by Kulldorff et al. (2003). In section 3 we derive an expression for the MDL in the hierarchical tree approach. In section 4 there is a brief explanation about Bayesian smoothing rates a necessary preliminary step in our procedure.In section 5 we describe the use of the MDL scan method in the hierarchical tree problem. In section 6 is presented a detailed algorithm of the MDL scan method. In section 7 we show some results in different cluster scenarios for the method. Finally, in section 8 we summarize our findings and discuss the relative merits of the MDL scan method.

Chapter 2

The Tree-Based Scan Statistic Method

To construct a hierarchical tree variable, we start defining the leaves, which each leaf contain all the information of the data set as the number of death by *silicosis*, the total number of death and any co-variates. Each leaf belong to a branch and branches with related leaves connect to the same node. Each node is also in the end of a branch and connects it with a higher level node. The process repeats until we reach the highest level node, the root. A example of a hierarchical tree is presented in Figure 2.1.

The Tree-Based scan statistic is based on a hierarchical structured data such that all observed data are on the leaves. Each leaf corresponds to an occupation with the following information: total number of people with that occupation and the number of *silicosis* death by occupation. The initial tree is defined a priori and its structure is fixed, this structure describes the proximity of one occupation to another. In the Tree-

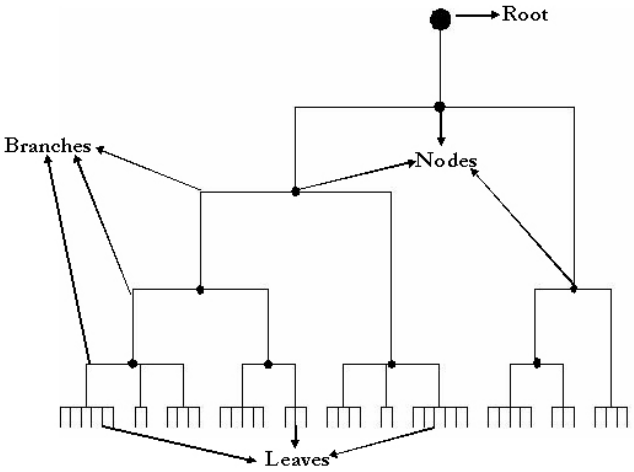

Figure 2.1: Example of Hierarchical tree with definitions of leaves, branches and nodes.

Based Scan statistics the search window is defined as cuts on any branch of the tree. To analyze the method all possible cuts in any branch of the tree is defined as Simple Cuts. Since for each cut the number of total death differ, it is not possible to take the number of *silicosis* death as a statistic. Instead a likelihood function (Loader (1991), Kulldorff (1997)) is proposed that estimates the risk inside and outside the cut. The likelihood function applied for the Tree-Based Scan is the same explained in Section 5. To analyze the proposed method two more complex cuts are also evaluated. For any node with two or more branches it is possible to have the Combinatorial and Ordinal Cuts. Suppose that we have a node with branches A, B, C and D. The Combinatorial Cut is defined as all possible combinations of the four branches, therefore besides the simple cuts the additional cuts are also evaluated [A,B] [A,C], [A,D], [B,C], [B,D], [C,D] [A,B,C], [A,B,D] ,[A,C,D] and [B,C,D]. Now, assume that for some node a previous order is define, take as examples node with branches teachers in kindergarten (A), elementary schools (B), high schools (C) and colleges (D). There is an ordered list associated with the age of the students, therefore it is plausible to believe that for A to be combined with D all others must be included, in the sense that A has no direct connection with D. So the Ordinal Cuts are defined as [A,B] [B,C], [C,D], [A,B,C], and [B,C,D], but not [A,D] for example. One can observed that Simple Cuts are subsets of Ordinal Cuts that are subsets of Combinatorial Cuts. After defining the three types of cuts analysis are made using the method for each cut type and results are obtained with the cuts that return the smallest p-value based on a Monte Carlos simulation.

Chapter 3

Minimum Description Length

The Minimum Description Length (MDL) principle is a general method for inductive inference, based on the idea that more regularities we find on the data set, more we are able to compress it and therefore more we can learn from the data. For our study the MDL principle will be used to select a best fitting model, suppose a class of models denoted by $\mathcal{H}$ and any model belonging H $\in \mathcal{H}$. In this context the MDL defines the best fitting model H $\in \mathcal{H}$ the one such that the shortest code length is necessary to describe the data D. The MDL statistics, also known as stochastic complexity of D, is given as:

$$MDL = L(\hat{e} \mid \hat{H}) + COMP(\hat{H})$$

where $L(\hat{e} \mid \hat{H})$ is the code length of the lack of fit conditioned on the fitted model $\hat{H}$, as demonstrated by Rissanen (1989) the lack of fit $\hat{e}$ is given by the negative of the log-likelihood of the fitted model $\hat{H}$. The other parameter, $COMP(\hat{H})$ is the code length of the fitted model $\hat{H}$, as explained by Grünwald (2005) for counting problems is the logarithm of distinguishable point hypotheses given to the model. The MDL principle selects the simplest model, in the sense that it allows the shorter description of the data given that all models equally fit the data.

Chapter 4

The Bayesian Smoothing Rates

Dynamic methods for detecting irregular spatial clusters often produce large and strongly irregular solutions when compared to the true simulated cluster (Duczmal and Assunção (2004), Assunção et al. (2006)). This behavior is due to the dynamic cluster growth procedure which searches for candidates that maximizes the likelihood statistic. Consequently, the simple aggregation of adjacency areas with high rates might produce cluster with maximum likelihood. However, areas in rural regions with lower population might randomly present high rates. Therefore, selecting areas that purely maximize the likelihood may randomly find a significant irregular cluster consequently reducing the power of the methodology.

We propose the use of empirical Bayes estimators (Marshall (1991), Santos et al. (2005)) to smooth the rates in the leaves and to avoid the selection of odd structure but keeping the likelihood maximization principle as part of the criterion to select branches and leaves in the tree. In our proposal, the rates are adjusted as a combination of the leaf rate and the branch rate where that particular leaf is located. Let B_i be the Bayesian rate estimate for the leaf i, l_i the observed rate and m the mean rate of the leaves in a particular branch. The smoothed Bayesian rate is

$$B_i = C_i l_i + (1 - C_i)m$$

where C_i is a number between 0 and 1 and is given

$$C_i = \frac{s^2 - \frac{m}{\bar{n}}}{s^2 - \frac{m}{\bar{n}} + s^2 - \frac{m}{n_i}}$$

$\bar{n}$ is the branch population average in risk, n_i is the population in the leaf i, n is the

total population of the branch and

$$s^2 = \sum \frac{n_i(l_i - m)^2}{n}$$

The Bayesian rate estimate is a weighted average between the leaf rate and the rate of the particular branch. If the population in the leaf is large, the adjusted rate is closer to the observed one otherwise, the adjusted rate is closer to the mean rate in the branch.

Chapter 5

The MDL Tree-Based Scan Statistic Method

Let the hierarchical data be represented as a tree, like shown in Figure 5.1. All the input attributes are inside the leaves. Each leaf has the data from a specific occupation, the total number of death and the number of death by *silicosis*. Figure 1 illustrates an hierarchical tree structure. The scan window represents the set of braches and leaves below a cut in any branch of the tree. As the total number of death in each cut varies, defining the number of death by *silicosis* doesn't represent an appropriate test statistics. Another approach is defining a likelihood function that distinct the risk inside the group of leaves inside the generated sub-tree, created by the cut in some branch, and leaves inside the other sub-tree.

Let use a Minimum Description Length (MDL) as our model selection criteria. The MDL model consists in combine the accuracy of a test statistics and the amount of necessary information given for the test be capable of get that accuracy level. So the MDL test statistics will consist in a:

$$MDL = I - T$$

where T is the test statistics (log likelihood) and I in the amount of information passed.

Based on that, let c_i be the number of death by *silicosis* in a leaf i. This number can be approximated by a Poisson $(\lambda_i n_i)$ distribution, where n_i is the total number of death in the leaf i, and λ_i is the probability that the death is caused by *silicosis* in the occupation i.

Define $C = \sum_i c_i$ and $N = \sum_i n_i$ as the total number of deaths by *silicosis* and the total number of death respectively in the tree.

Conditional to the total number of cases, the relative distribution among different

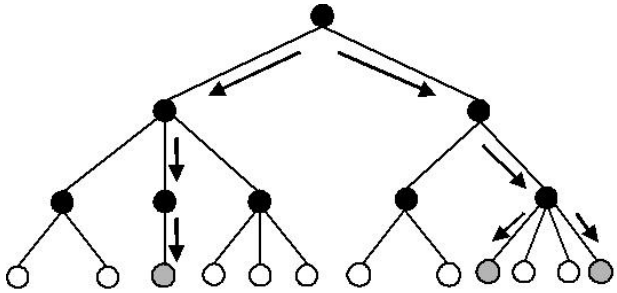

Figure 5.1: Path for the leaves belonging to the G group.

leaves follows a multinomial distribution, with likelihood:

$$L(\lambda, c) = \prod_i \left(\frac{\lambda_i n_i}{\sum_j \lambda_j n_j} \right)^{c_i}$$

The null hypothesis is $\lambda_i = \lambda$ for all leaves i. The alternative hypothesis is defined as, there is a group G of leaves in which $\lambda_i = \lambda_G$ and $\lambda_i = \lambda_R < \lambda_G$ to all the remaining leaves, $i \notin G$. It is possible to define a log-likelihood ratio test statistics.

$$T = max_G \left\{ c_G log\left(\frac{c_G}{n_G}\right) - (C - c_G) log\left(\frac{C - c_G}{N - n_G}\right) \right\} I\left(\frac{c_G}{n_G} > \frac{C - c_G}{N - n_G}\right)$$

where $c_G = \sum_{i \in G} c_i$ and $n_G = \sum_{i \in G} n_i$, e $I(.)$ is the indicator function.

After decided which will be our test statistics, it is necessary to define numerically the amount of information given for the T statistics. Since, we get a tree structure in our problem it is possible to visualize (Figure 5.1) that the path for all leaves belonging to the G group as the necessary amount of information that the T statistics needs to get it results.

However if two or more leaves share some similar path is not necessary to repeat the information given previously to the T statistics, it is only necessary to count the number of similar paths we have in the leaves of the candidate cluster. For example assume we have in our candidate cluster the leaves shown in Figure 5.1, therefore instead of defining the information as the three full paths: $2 \rightarrow 3 \rightarrow 1 + 2 \rightarrow 2 \rightarrow 4 + 2 \rightarrow 2 \rightarrow 4$, one can count the number of similar path and specify the leaves in the model as: $2 \rightarrow 3 \rightarrow 1 + 2 \rightarrow 2 \rightarrow \{\frac{4}{4}$.

Based in what is said in Grünwald (2005) the maximum amount of information in a path can be defined as:

$$I = \sum_1^k log_2 j + m log_2 dt$$

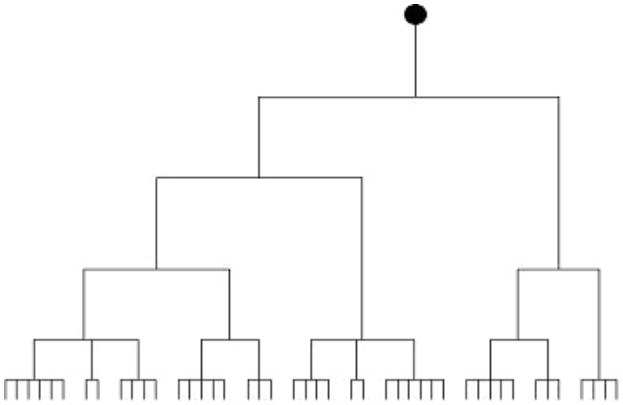

Figure 5.2: Hierarchical tree.

where k is the total number of elements in the path, j is the value of each element (numbers of brothers that the node have), m is the number of similar paths and dt is the tree depth.

If you path is: $2, 3, 1, 2, 2, 4, 4 + 1$ similar path shared by the leaves represented in the right side of the tree, the transformation to the numerical representation will be: $log_2 2 + log_2 3 + log_2 1 + log_2 2 + log_2 2 + log_2 4 + log_2 4 + 1 log_2 3$.

With both parameter well defined our cluster information will be the group of leaves that minimize the MDL statistics.

The created method is a variation of the scan method based in a Tree-based statistics proposed by Kulldorff et al. (2003).The new model consists of making a search in a more complete way being capable to find relations between occupations even though these are not related directly.

From a database, it is possible to define a hierarchical tree, that mean, each node has 1 to n branches. So the objective is to detect any pattern or correlation between the leaves of this tree by some stipulated factor.

The proposed model makes a scan, based on the MDL criteria to find the group of occupations that has the greater discrepancies compared to the null hypothesis, without making an exhausting search. To do the scan capable of finding this group of leaves inside the tree we proposed an algorithm that uses the lesser number of assumptions a priori, being capable to be executed with efficiency and producing a satisfactory result.

The execution of the algorithm consists in the following steps, given a tree, are created the groups G_{1i} related to the A_i group, where the A_i is the group of all neighbors leaves. After find the group with maximum likelihood (G_{1i}) for each A_i the branches are populated with the number of cases occurred by *silicosis* and the total number

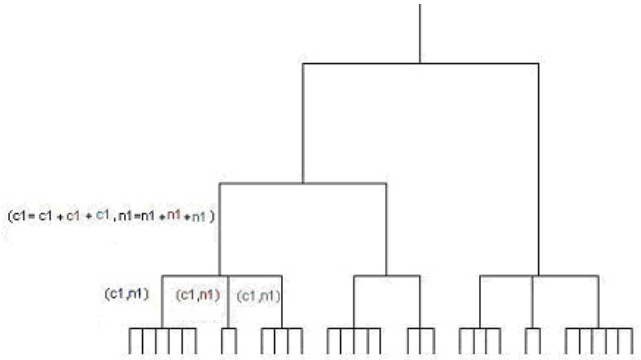

Figure 5.3: Filling the branches with population and cases.

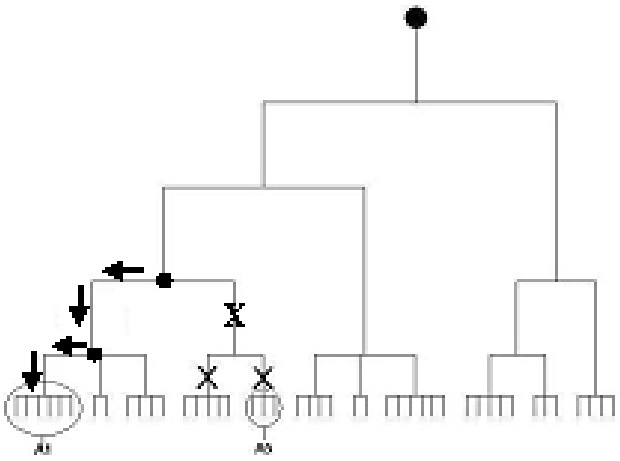

Figure 5.4: Returned group with the elements G_{11} e G_{15} .

of cases inside of each group G_{1i} inside the sub-tree below this branch, like shown in Figure 5.3.

After filling the leaves of the tree with the data, the algorithm is executed identifying the way in the tree that return the group of leaves with the minimum MDL among all the options, like show in Figure 5.4.

To find the p-value associated to the test statistics on the null hypothesis, a Monte Carlo procedure is executed.

Chapter 6

The MDL Tree Scan algorithm

The MDL Tree Scan algorithm starts by using a Bayesian smoothing to the leaves in the same branch and selects the leaves with higher rates as initial cluster candidates. In sequence, the algorithm gradually scans the nodes of the tree using the information of the leaves aggregated to the node level. Succinctly, each node stores the information of the total population and the total number of cases of the leaves below. The algorithm compares de MDL statistic of the candidate cluster with the statistics when an entire node is aggregated. If the MDL statistic decreases then the algorithm searches for the group of leaves inside of the node which contributes to minimize the MDL. By doing this, the algorithm avoids searching in leaves whose nodes do not indicate that the test statistic will be minimized and therefore saves computing time and improve power. The outline of the algorithm is shown below.

MDL Hierarchical Scan Algorithm

1. Create a Tree from the Hierarchical database

2. Fulfill the Tree leaves with the given attributes, cases and population.

3. Fulfill the nodes of the Tree with the total number of cases and population of the leaves below.

4. Define A_i as the group of leaves in the same final branch.

For each group A_i **do**

(a) Smooth the rates of each leaf using the bayesian estimator.

(b) Select the leaf with the highest smoothing rate and create G_i.

(c) Add new leaves from A_i to G_i if they increase the likelihood.

End For

For each group G_i **do**

(a) Start the cluster candidate G with G_i

(b) Go to the immediate higher node

For each group G_j, $j \neq i$ below the node **do**

(i) Test if the MDL from $G + G_j$ decreases.

(ii) If yes, update G with G_j

End For

Repeat

(c) Go to the next higher node.

(d) Select the nodes below that do not contain the cluster G.

If there is one or mode nodes capable of reducing the MDL **do**

(i) Explore the lower levels of the nodes that minimizes the MDL.

(ii) Select the groups G_* of the nodes that minimizes the MDL.

(iii) Update cluster candidate G with the new groups.

End If

Stop if the root node is reached

End Repeat

End For

5. Store the cluster candidate with the minimum MDL.

Chapter 7

Results

A simulated example was created to evaluate the performance of the algorithm in different conditions for the cluster. Figure 7.1 show the corresponding tree. A cluster with three leaves was defined so that the number of cases in these leaves were much higher than the remaining leaves.

Using this example, Figure 7.2 shows the detected cluster. In this particular example, the method was able to find the exact true cluster.

In sequence, we used the structure proposed by The United States Bureau of Census for occupations classifications to create the hierarchical tree. Similarly, we generate synthetic data by fixing the number of deaths in each leaf and we randomly distributed the total number of deaths by *silicosis* selecting some leaves with a higher relative risk than the remaining leaves. Since we assumed a Poisson distribution for our model and conditioning to the total number of *silicosis* deaths, the distribution of the cases among the occupations follows a multinomial distribution. We simulated 5 different clusters, these clusters were created in order to cover the most important types to analyze.

1. A cluster with only 1 leaf with fixed and known position.

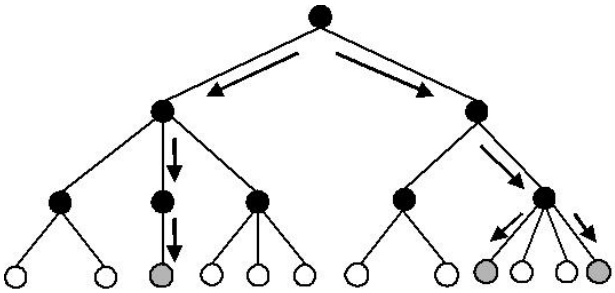

Figure 7.1: Generated tree. The three gray leaves represent the cluster.

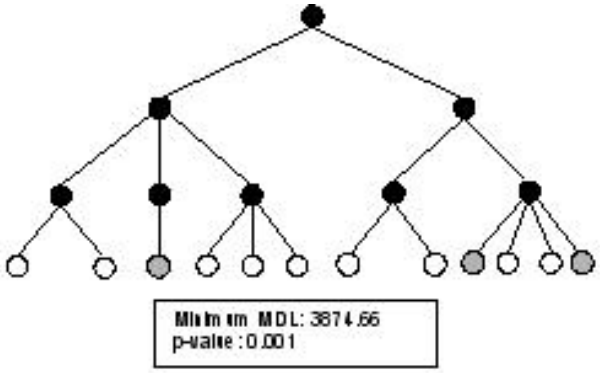

Figure 7.2: Returned tree, where the three gray leaves is the candidate cluster.

2. A cluster with 11 leaves located in the same final node.

3. A cluster with 23 leaves spread in different locations of the tree.

4. A cluster with 22 leaves and it represents a sub-tree.

5. A cluster with 9 leaves and it represents a mixture of models 3 and 4.

For each cluster type described before we simulated 7 distinct scenarios where for each scenarios the relative risk of the cluster was set to 2, 3, 4, 5, 10, 20 and 100, respectively. A data set of $10,000$ trees for each scenario was created in order to test the efficiency of the proposed algorithm. Tables 7.1, 7.2, 7.3, 7.4 and 7.5 present the results for power detection for each value of the relative risk and each cluster type.

Table 7.1: Power results for cluster type 1

RR	$p \leq 0.05$	$p \leq 0.01$	Cluster Size Median	Intersection Median
1	0.0478	0.0088	2	–
2	0.1868	0.1171	2	1
3	0.9917	0.9875	2	1
4	1	1	2	1
5	1	1	2	1
10	1	1	2	1
20	1	1	2	1
100	1	1	2	1

From Tables 7.1 to 7.5, it is possible to see that with a relative risk equal to one (null hypothesis) the proportion of significant cluster for a 5% level of significance is 4.78% and for a significant level of 1% the proportion is about 0.88%, as expected.

Table 7.2: Power results for cluster type 2

RR	$p \leq 0.05$	$p \leq 0.01$	Cluster Size Median	Intersection Median
1	0.0478	0.0088	2	–
2	0.5753	0.4931	7	7
3	0.8799	0.8765	8	8
4	0.9248	0.9235	9	9
5	0.9528	0.9524	10	10
10	0.9921	0.9920	11	11
20	0.9996	0.9996	11	11
100	1	1	11	11

Table 7.3: Power results for cluster type 3

RR	$p \leq 0.05$	$p \leq 0.01$	Cluster Size Median	Intersection Median
1	0.0478	0.0088	2	–
2	0.1960	0.0794	2	1
3	0.5134	0.3459	2	1
4	0.7638	0.6362	2	1
5	0.8902	0.8124	2	1
10	0.9937	0.9877	5	5
20	0.9999	0.9995	8	8
100	1	1	10	10

As the relative risk increases, the detected cluster converges to the true cluster, as expected. Particularly, results for cluster type 1 achieved higher power and in all of the simulated cases, the algorithm was able to include at least one of the true leaves. Results also show that when the cluster is located in the same sub-tree (cluster type 2), the algorithm achieves better detection and power results than the situation where the cluster is spread in the tree (clusters type 3).

For cluster type 3, results get improved if the relative risk is higher than 5. In this particular scenario the cluster is randomly spread in the tree with no association with the leaves in the same branch. Consequently, the method is capable to select leaves in different branches only if the relative risk of the leaves is extremely higher.

Figure 7.3 aims at providing a visual analysis of the sensitivity/specificity in terms of the number of leaves in both detect and true clusters. It also displays the number of leaves in the intersection of the detected and true cluster. The figure show that as the relative risk increases the candidate cluster converges to the exact real cluster.

Table 7.6 presents the proportion of the population of the true cluster inside the detected cluster. Results show that when the cluster is scattered in the tree, the ability

Table 7.4: Power results for cluster type 4

RR	$p \leq 0.05$	$p \leq 0.01$	Cluster Size Median	Intersection Median
1	0.0478	0.0088	2	–
2	0.0466	0.0089	2	0
3	0.0337	0.0086	2	0
4	0.1911	0.1547	9	6
5	0.6697	0.6273	11	9
10	1	1	19	17
20	1	1	21	19
100	1	1	22	20

Table 7.5: Power results for cluster type 5

RR	$p \leq 0.05$	$p \leq 0.01$	Cluster Size Median	Intersection Median
1	0.0478	0.0088	2	–
2	0.0482	0.0099	2	0
3	0.0389	0.0105	2	0
4	0.2222	0.1726	7	4
5	0.7213	0.6792	8	6
10	1	1	11	9
20	1	1	11	9
100	1	1	10	9

Table 7.6: Proportion of the real population in the detected cluster.

	Cluster Size				
Relative Risk	1 leaf (%)	11 leaf (%)	23 leaf (%)	22 leaf (%)	9 leaf (%)
2	100.00	47.39	1.43	0.00	0.00
3	100.00	81.47	6.06	0.44	0.51
4	100.00	87.46	13.24	12.74	15.57
5	100.00	91.84	20.59	48.97	56.95
10	100.00	98.14	49.08	93.45	96.82
20	100.00	99.52	68.60	98.21	99.77
100	100.00	99.87	79.30	99.51	99.99

of the method to correctly identify the true population is low, as mentioned before. Furthermore, the method detects more than 50% of the true population when the relative risk is above 5.

After evaluating the algorithm performance for synthetic data, we applied our approach to the real data set supplied by NCHS. Table 7.7 shows the occupation code

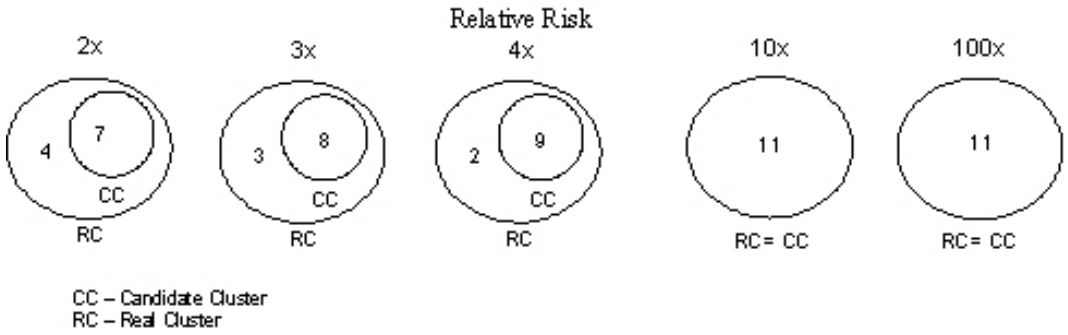

Figure 7.3: Intersection between the cluster type 2 and the candidate cluster.

Table 7.7: Results of occupations in the candidate cluster by the data of the real data set.

Id#[a]	Occupation Category
043	Architect
615	Explosives workers
616	Mining machine operators
617	Mining occupations, n.e.c [b]
703	Lathe and turning machine set-up operators
707	Rolling machine operators
709	Rolling machine operators
719	Molding and casting machine operators
725	Miscellaneous metal and plastic processing machine operators
	Woodworking Machine Operators
768	Crushing and grinding machine operators

[a] Id# = occupational identification code.
[b] n.e.c = not elsewhere classified.

and category for the leaves in the significant primary detected cluster ($p = 0.001$).

The *silicosis* disease is associated with work environments where there is dust of silica in the air. When inhaled, the dust of silica cause damage to the lungs, compromising the ability to absorb oxygen. This disease is commonly related to people who are machine operators in mines or who work with heavy equipments or occupations that deal with dust such as constructions, woodworking, etc. Table 7.7 shows that the detected cluster found leaves whose professions are in fact a critical group with higher chance to develop the disease.

The computing time to run this test was 2 minutes and 20 seconds for a tree with 6 levels and 503 leaves using a Pentium IV 1.6 GHz, 512 Mb RAM and 20 Gb Hard disk.

Chapter 8

Conclusions

This work presents an efficient scan algorithm for hierarchical data sets organized in a tree structure. The algorithm searches through the branches of the tree and it is able to aggregate leaves located in different branches of the tree. The scan search aims at minimizing the minimum description length statistic which combines the log-likelihood and the amount of information necessary to represent internally the cluster. This second term controls the degree of freedom of the search algorithm. The methodology prevents the algorithm to add leaves that unnecessarily increase the log-likelihood term.

Results show that the power of the algorithm is improved when the relative risk of the true clusters increases. Furthermore, the detected cluster normally includes elements of the true cluster and as the relative risk increases the algorithm detects more elements of the true cluster. Nevertheless, if the cluster is scattered in the tree, the performance of the algorithm is compromised.

The method found one significant cluster when it was applied to the *silicosis* data set. The detected group has occupations that are proven to have higher risks for the *silicosis* disease. Comparing our results with those obtained by Kulldorff et al. (2003) 8 professional occupations were common in both results. The MDL method found 2 distinct professions that were located in different branches of the tree. The *Architect* occupation, detected by MDL is compatible with the *silicosis* hazard environment but it is located in a different branch of the tree. Although the *Explosives workers* occupation has a higher rate its population is quite small (385 deaths), this occupation was included in the cluster due to the bayesian smoothing process.

An important issue was to compare both methods, the MDL tree scan statistics and the one proposed by Kulldorff et al. (2003), however we tried to reach the authors and could not get their algorithm to run our data. Because of that was not possible to compare the results obtained by our method and the one proposed by Kulldorff et al. (2003).

This work shows that the MDL methodology is a novel and flexible algorithm that detects clusters in hierarchical data sets whose elements are partially scattered in the tree. Therefore, it explores potential groups of clusters that are not explicit by just looking at cuts of branches or combinatorial analysis of the given data. Furthermore, it provides an important impact in the final analysis.

Bibliography

Assunção, R., Costa, M., Tavares, A., and Ferreira, S. (2006). Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, 25:5:723–742.

Costa, M., Assunção, R., and Kulldorff, M. (2008). Constrained spanning tree algorithm for irregular spatial clustering (submitted). *Journal of Computational & Graphical Statistics*.

Costa, M., Schrerrer, L., and Assunção, R. (2005). Detecção de conglomerados espaciais com geometria arbitrária. *Simpósio Brasileiro de Geoinformática - GEOINFO*.

Duczmal, L. and Assunção, R. A. (2004). Simulated annealing strategty for the detection of arbitrarily shaped spatial cluster. *Computational Statistics and Data Analysis*, 4:269–286.

Duczmal, L., Kulldorff, M., and Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped disease cluster. *Journal of Computational & Graphical Statistics*, 15:2:428–442.

Grünwald, D. P. (2005). *The Minimum Description Length Principle*. The MIT Press, London, England.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26:1481–1496.

Kulldorff, M., Fang, Z., and Walsh, S. J. (2003). A tree-based scan statistic for database disease surveillance. *Biometrics*, 59(2):323–331.

Loader, C. R. (1991). Large-deviation approximations to the distribution of scan statistics. *Adv. in Appl. Probab*, page 751771.

Marshall, R. J. (1991). Mapping disease and mortality rates using empirical bayes estimator. *Applied Statistics*, 40:283–294.

Patil, G. P. and Taillie, C. (2003). Geographic and network surveillance via scan statistics for critical area detection. *Statistical Science*, 18:4:457–465.

Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11:183–197.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.

Santos, A., Rodrigues, A., and Lopes, D. (2005). Aplicações de estimadores bayesianos empíricos para análise espacial de taxas de mortalidade. *Simpósio Brasileiro de Geoinformática - GEOINFO*.

Tango, T. and Takahashi, K. (2005). A flexible shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4:11.