GIS/data mining applied for identification of environmental risk factors for diseases

S. Anno

School of Engineering, Shibaura Institute of Technology, Japan

Abstract

This study has tried to identify environmental risk factors for respiratory system diseases in Tokyo metropolitan area using GIS and data mining. GIS were applied to the analysis of spatial relationships between the distribution of the diseases and SPM exposure levels and distances from the national roads as environmental factors. A CART model as a data mining tool was applied for assessment of habitat types where there was a potential risk of the incidence of the diseases with the databases obtained from GIS analysis while identifying environmental risk factor levels concerned with the incidence of the diseases. Another purpose of this study was to provide a comprehensive outline of the methods of data mining, which have wide applications in a GIS and other areas where spatial data are used, whilst indicating its strengths and weaknesses, how and when to apply the methods and their role when used.

Keywords: geographic information systems (GIS), spatial analysis, data mining, classification and regression trees (CART), decision tree.

1 Introduction

Human respiratory effects of environmental factors are well documented; numbers of epidemiological studies have examined affect of environmental factors on respiratory health [1, 2, 3, 4, 5]. However, those studies have not analyzed spatial associations between respiratory system diseases and environmental exposures; moreover, those studies have not identified environmental risk factor levels concerned with the incidence of the diseases. Then, our study tried to analyze relationships between spatial pattern of the diseases and environmental factors; and identify environmental risk factor levels concerned with the incidence of the diseases using spatial analysis in geographic
information systems (GIS) and data mining. Another purpose of this study was
to provide a comprehensive outline of the methods of data mining, which have
wide applications in a GIS and other areas where spatial data are used while
indicating its strengths and weaknesses, how and when to apply and its role to be
used.

2 Material and methods

2.1 Spatial and attribute data

The study site was focused on Tokyo metropolitan area. The digital map for the
area with a scale of 1:2,500, which contains the national roads, was used as
spatial data. As for attribute data, data on a hundred of the subjects diagnosed
with typical symptoms of respiratory system diseases, which were collected from
a health survey by the agency of Aozora, were used for the study. Also, data on
the average SPM concentrations for 22 years (i.e., from 1978 through 1999)
measured along the major roadways and residential districts of Tokyo consisting
of 70 observation sites in total were used as attribute data [6]. Those spatial and
attribute data were integrated within a GIS database of the study area and used
for spatial analysis.

2.2 Spatial analysis

Spatial analysis in a GIS was conducted in the following procedures. First of all,
SPM data plotted as points on a layer in a GIS were interpolated using the
method of Inverse Distance Weighted (IDW) to estimate spatial distribution of
SPM concentrations throughout the study area and converted into raster data,
which divides space into a mesh unit. There are some intervening factors at the
individual scale that may influence normal process of environment and diseases
phenomena. Within the raster data, the spatial resolutions of the data were set to
100 m² and 500 m² respectively for addressing health events and environmental
exposures at different spatial scales. The data sets consisted of 97,500 meshes
for 100 m² and 5,246 meshes for 500 m² respectively. Secondary, the patients’
addresses were converted into longitude and latitude by using geocoding service
for CSV formatted file on WWW [7]; and geocoded to the digital map, which
represented by points on a layer. Each data set including SPM concentrations,
the national roads and the patients was represented by a separate layer. Those
layers were overlaid and outputted as the raster databases. Each mesh of the
raster databases represents the number of the patients, the value for SPM
concentrations and the attribute value for distances from the national roads.
Patient data within a mesh were maintained at the individual level and the data
were aggregated by each mesh. In other words, the meshes that represent a
patient/patients were assigned an attribute value of present; and the meshes that
did not represent a patient/patients were assigned an attribute value of absent.
The raster databases obtained from the spatial analysis in a GIS were exported as
databases and used for further processes of data mining.
2.3 Data mining

Data mining methodologies have become quite popular in recent years as computing speed and storage capabilities have increased. There are a number of classification and prediction algorithms that fall under the rubric of data mining. These include neural networks, clustering programs, association rules and classification and regression trees (CART). The ultimate objective of data mining is knowledge discovery within a large database. In other words, using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in a large database and infers rules that allow the prediction of future results. This usually involves the prediction of some criterion variable value (i.e., outcome or dependent variable value) from a number of predictor or independent variable values. Data mining techniques have wide applications in a GIS and other areas where spatial data are used [8, 9, 10]. In our study, a CART approach [11] was applied for analyzing the databases, which were exported from the raster databases.

According to Breiman et al [11], a CART is an algorithm that learns binary decision tree representations. Decision tree models classify data using a series of if-then rules depicted in a tree representation. The basis of decision tree algorithms is the recursive partitioning of the data into more homogenous subsets. A CART model was trained in a round-robin fashion (i.e., leave-one-out or k-fold cross-validation with k=N) in order to assess the model performance [11, 12].

A CART model was implemented for assessment of habitat types where there is a potential risk of the incidence of respiratory system diseases using the databases in terms of the mesh size of 100 m² and 500 m² respectively as the estimated suitable area for population at risk. The following variables were considered as possible predictors of the presence/absence of patients in the modeling. The outcome variable has two categorical variable values: present and absent. The two predictor variables are measurements of distances from the national roads and SPM concentrations. A CART model was built using the databases in order to interpret the differences in habitat types between the patients’ presence and the patients’ absence; and developed to predict the presence/absence of patients allowing the ranking of variables. The CART approach produced a decision tree that could form the basis of simple if-then rules that could be used to predict the outcome of interest. The tree can be viewed as providing a probability model, with partition defined by the overall probability of misclassification. More, a chi-square test on a two-dimensional contingency table was used to examine strength of relations between variables in order to evaluate the CART model. Based on the threshold on the output decision variable obtained from the results of data mining, two groups were established: under and over the threshold; and a chi-square test was performed to examine if there were strength of relationships between patients’ presence and the predictor variables.
3 Results

The decision tree in case of mesh size of 100 m² is shown in figure 1. The hierarchical structure of the tree can be directly used to evidence the predictor variables, which appear to have effects on the distribution of patients in the study area. The tree can be interpreted that habitat types and areas at high risk of the diseases are locations of: 1) less than 1509.25m from the roads and 2) 1509.25m and over away from the roads with SPM concentrations of 0.19255mg/m³ and over. Of particular interest is the predictor variable value of SPM concentrations selected by the model exceeds the threshold (0.10mg/m³) that is defined by the Ministry of Environment as a critical value for the notification of the citizens based on long-term evaluation. The results go for the case of mesh size of 500 m². The result obtained from a qui-square test suggests that, in case of mesh size of 100 m², there were significant differences (p < 0.01) between presence of the patients and distances from the roads; and between presence of the patients and SPM concentrations. In case of mesh size of 500 m², there were significant differences (p < 0.05) between presence of the patients and distances from the roads; and between presence of the patients and SPM concentrations. Especially, the diseases were more significantly associated with distances from the roads compared to SPM concentrations.

Figure 1: The decision tree in case of mesh size of 100 m².

4 Discussion

The resulting decision tree was clearly illustrated that the most important predictor was a distance from the roads; in other words, the incidence of the diseases in the study area were strongly influenced by a distance from the roads. It suggested that automobile exhaust could be considered as the environmental risk factor and affect the distribution of patients of the diseases in the area. Mesh size may affect results of analysis and should be based on the minimum mapping unit because using too large a cell size will cause some information to be lost. For addressing health and environmental problems in this study, mesh size of
100 m² was a suitable spatial scale to represent and analyze the spatial variation of the diseases and the environmental factors while taking the results of a chi-square test into consideration. Methods of data mining are useful for not only finding patterns and subtle relationships in a large database but inferring rules that allow the prediction of future results. A CART methodology as a data mining tool, which yields a set of decision trees, allows the accurate prediction of outcome for future patients based on the values of their predictor variables and can be used as a new method of analyzing health data within the context of public health concerns. Furthermore, the methodology of data mining used in this study can be expected to make an approach to an effective application of GIS.

Acknowledgements

The author would like to thank Dr. Sagara at Research Center for Information Sciences, University of Tokyo for help with providing the geocoding service for this study.

References


[7] CSV address matching service; the Center for Spatial Information Science, the University of Tokyo, Online. [http://fujieda.csis.u-tokyo.ac.jp/cgi-bin/geocode.cgi](http://fujieda.csis.u-tokyo.ac.jp/cgi-bin/geocode.cgi).


