

Tabasso, Myriam; Arbia, Giuseppe

Conference Paper

Spatial Econometric Modelling Of Massive Datasets: The Contribution Of Data Mining

53rd Congress of the European Regional Science Association: "Regional Integration: Europe, the Mediterranean and the World Economy", 27-31 August 2013, Palermo, Italy

Provided in Cooperation with:

European Regional Science Association (ERSA)

Suggested Citation: Tabasso, Myriam; Arbia, Giuseppe (2013) : Spatial Econometric Modelling Of Massive Datasets: The Contribution Of Data Mining, 53rd Congress of the European Regional Science Association: "Regional Integration: Europe, the Mediterranean and the World Economy", 27-31 August 2013, Palermo, Italy, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:

<https://hdl.handle.net/10419/124093>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

SPATIAL ECONOMETRIC MODELLING OF MASSIVE DATASETS: THE CONTRIBUTION OF DATA MINING

G. Arbia,¹ M. Tabasso²

Abstract

In this paper we provide a brief overview of some of the most recent empirical research on spatial econometric models and spatial data mining. Data mining in general is the search for hidden patterns that may exist in large databases. Spatial data mining is a process to discover interesting, potentially useful and high utility patterns embedded in large spatial datasets. The field of spatial data mining has been influenced by many other disciplines: databases technology, artificial intelligence, machine learning, probabilistic statistics, visualization, information science, and pattern recognition. This process is more complex than conventional data mining because of the complexities inherent in spatial data. Spatial data are multi-sourced, multi-typed, multi-scaled, eterogeneous, and dynamic. The main difference between data mining and spatial data mining is that in spatial data mining tasks we use not only non-spatial attributes (as it is usual in data mining in non-spatial data), but also spatial attributes. We suggest some directions along which spatial econometric modeling could benefit from the cross-fertilization spatial data mining techniques such as Classification and Regression Trees (CART). We use the CART algorithm to fit empirical data and produce a tree with optimal tree size for different specifications of econometric models. We also examine some diagnostic measures to evaluate the spatial autocorrelation of the pseudo-residuals obtained from the regression tree analysis and we compare the accuracy and performance of different versions of CART that take into account the effects of spatial dependence. To address this issue, we start examining a non-spatial regression tree, then we include the geographical coordinates of data in the covariate set and finally, we consider one of the most common spatial econometric models: Spatial Lag combined with two versions of regression trees: non-spatial regression tree and geographical coordinates based regression tree. This allows us to determine the strength and the possible role of spatial arrangement on the variables in the predictive model and reduce the effect of spatial autocorrelation on prediction errors. In particular, we test the sensibility of various regression trees with different spatial weights matrix specifications such that to remove the spatial autocorrelation on pseudo-residuals and to improve the accuracy of spatial predictive models.

Keywords: spatial econometric models, spatial data mining, CART, spatial autocorrelation

JEL Classification: C14, C31, C52, C81, R10

¹University “Cattolica del Sacro Cuore”, Faculty of Economics- Department of Statistics and Institute of Hygiene , Via F. Vito, 1 - 00168 Roma - email: giuseppe.arbia@rm.unicatt.it

²University of Rome “Sapienza”, Doctoral School of Economics- Department of Economic and Social Analysis, P.le A. Moro, 5 - 00161 Roma - email: myriam.tabasso@uniroma1.it

1 Introduction

In this paper we present the contribution of the spatial data mining on spatial econometric models of massive datasets. We propose a data mining methodology that explicitly considers the phenomenon of spatial autocorrelation on prediction errors. This paper is organized as follows. Section 2 describes the main concepts and challenges of data mining, in particular the Classification and Regression Trees (CART) as an important data mining methodology for the analysis of large data sets via binary partitioning procedure. Section 3 introduces the principal differences of *spatial data mining* with respect to *classical data mining*. The focus of Section 4 is the analysis of different versions of CART to compare the performance and evaluate the spatial autocorrelation of prediction errors of the regression trees (pseudo-residuals). Finally, Section 5 reports some concluding remarks and future works.

2 Spatial Linear Regression Models and data mining

2.1 Spatial regression

In this section, we present in brief a general framework that allows to incorporate spatial correlation structures into a linear regression model. The focus is how one can to incorporate spatial effects into a linear regression models by considering the following general specification (Anselin,1988):

$$y = \rho W_1 y + X\beta + \varepsilon \quad (1)$$

$$\varepsilon = \lambda W_2 \varepsilon + \mu \quad (2)$$

with $\mu \sim N(0, \Omega)$ and the diagonal elements of the error covariance matrix Ω as: $\Omega_{ii} = h_i(z\alpha)$, $h_i > 0$.

In this specification, β is $K \times 1$ vector of parameters associated with exogenous variable X ($N \times K$ matrix), ρ is the coefficient of the spatially lagged dependent variable, and λ is the coefficient in a spatial autoregressive structure for the disturbance ε . The disturbance μ is normally distributed with a general diagonal covariance matrix Ω . The diagonal elements allow for heteroschedasticity as a function of $P + 1$ exogenous variables z , which include a constant term. The P parameters α are associated with nonconstant terms, such that, for $\alpha = 0$, it follows that $h = \sigma^2$ (the classic homoskedasticity situation). The two $N \times N$ matrices W_1, W_2 are standardized spatial weight matrices, associated with a spatial autoregressive process in the dependent variable and the disturbance term respectively. In total, the model has $3 + k + p$ unknown parameters, in vector form:

$$\theta = [\rho, \beta', \lambda, \sigma^2, \alpha']' \quad (3)$$

When subvectors of the parameter vector (3) are set to zero, specifically we have the following situations which correspond to four traditional spatial autoregressive models commonly discussed in the literature (see e.g. Hordijk, 1979; Anselin 1980, 1988; Bivand 1984):

1^oCase: $\rho = 0, \lambda = 0, \alpha = 0$ (P+2 constraints):

$$y = X\beta + \varepsilon \quad (4)$$

that is the classical linear regression model.

2^oCase: $\lambda = 0, \alpha = 0$ (P+1 constraints):

$$y = \rho W_1 y + X\beta + \varepsilon \quad (5)$$

that is the mixed regressive spatial-autoregressive model: (**SAR** or Spatial Lag Model) (which includes the common factor specifications, i.e. with WX , as special case).

3^oCase: $\rho = 0, \alpha = 0$ (P+1 constraints):

$$y = X\beta + (I - \lambda W_2)^{-1} \mu \quad (6)$$

that is the linear regression model with a spatial autoregressive disturbance: **Spatial Error Model (SEM)**.

4^oCase: $\alpha = 0$ (P constraints):

$$y = \rho W_1 y + X\beta + (I - \lambda W_2)^{-1} \mu \quad (7)$$

that is the mixed regressive-spatial autoregressive model with a spatial autoregressive disturbance.

A variant of the spatial lag model that include spatially lagged independent variables is known as the **Spatial Durbin Model (SDM)**, LeSage and Pace 2009):

$$y = \rho W y + X\beta + WX\lambda + \varepsilon \quad (8)$$

where λ is the vector of coefficients for spatially lagged independent variables WX . The use of this model instead of the spatial lag model in (5) can potentially remove omitted variable bias, as discussed in detail in LeSage and Pace (2009). An alternative model with respect to SEM is the spatial moving average (SMA) model (Fingleton 2008):

$$y = \rho W y + X\beta + (I + \lambda W)\mu \quad (9)$$

As can be observed by comparing (6) and (9), the spatial multiplier is not present in the

SMA model. The SMA model is used to model localized effects. By its specification, spatial effects will affect only the first-order neighbors as defined by the weights matrix. In particular, this can be seen by considering the expanded form of $(I - \lambda W_2)^{-1}$.

A Leontief expression of the last matrix, under the assumption that $|\lambda| < 1$ is given by

$$(I - \lambda W_2)^{-1} = I + \lambda W + \lambda^2 W^2 + \dots \quad (10)$$

As argued by Anselin (2003), the complete structure of the variance-covariance matrix then follows as the product of the (10) with its transpose, yielding a sum of terms containing matrix powers and products of W , scaled by powers of λ . Specifically the lowest order term is I , followed by λW and $\lambda W'$, $\lambda^2(W^2 + WW' + W'^2)$ and so on. For a spatial weights matrix corresponding to first-order contiguity, each of the powers involves a higher order of contiguity, in effect creating band of every larger reach around each location, relating every location to every other one. Moreover the powers of the autoregressive parameter (with $|\lambda| < 1$) ensure that the covariance decreases with higher orders of contiguity.

Instead, the only diagonal non zero elements in the variance-covariance matrix are those corresponding to nonzero elements in W elements in W (or, equivalently, W') and $W W'$.

For W defined as first-order contiguity, such elements consist of location pairs that are first- and second-order neighbors, but no higher orders of contiguity. Consequently, the range of the effect of the spatial multiplier is much smaller than for a corresponding SAR model.

Several authors have suggested to combine spatial lag with spatial error dependence. The most general form is the spatial autoregressive, moving-average (SARMA) processes outlined by Huang(1984). Formally, a SARMA (p,q) process can be expressed as (Anselin and Bera, 1998)

$$y = \rho_1 W_1 y + \rho_2 W_2 y + \dots + \rho_p W_p y + \varepsilon \quad (11)$$

$$\varepsilon = \lambda_1 W_1 \mu + \lambda_2 W_2 \mu + \dots + \lambda_q W_q \mu + \varepsilon \quad (12)$$

A different specification that combines spatial-autoregressive model with spatial-autoregressive disturbances is often referred to as a SARAR(p,q) model, see Anselin and Florax (1995). In modeling the outcome for each unit as dependent on a weighted average of the outcomes of other units, SARAR models determine outcomes simultaneously. Formally a SARAR (1,1) process can be expressed in (1). These various specifications are the most important to analyse global and local externalities in spatial econometric models (see Anselin, 2003).

2.2 Data mining and KDD

Several authors have observed that the term “data mining” has had a varied history (Fayyad, Piatetsky-Shapiro, and Smyth 1996; Smyth, 2000). It can be considered as a single step in the

multi-step process of Knowledge Discovery in databases (KDD), where KDD is defined as the “*non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*” (Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy, 1996). The term process implies that KDD comprises many steps, which involve data preparation, search for patterns, knowledge evaluation, all repeated in multiple iterations. In alternatively, data mining is the “*process of extracting valid, previously unknown, comprehensible and actionable information from large database and using it to make crucial business decisions*”(Simoudis, 96). In this case data mining, not KDD, is viewed as the overall process of extracting high-level knowledge from low-level data. Some authors underline the difficulty to isolate a core set of fundamental techniques that clearly distinguish data mining from any single component discipline: in some way it is a uniquely powerful combination of individual techniques from each discipline associated with analyzing massive data sets. In particular, it is a multidisciplinary field and it includes: *machine learning, statistics, database technology, high performance computing, data visualization, image processing*. (Behnke and Dobinson, 2000). According to Weiss and Davison (2010) data mining can be considered a possible response to many problems like the *scalability* of traditional statistical techniques, which often cannot handle data sets with milion or billions of records and hundreds or thousands of variables; *highly unstructured* (non-numeric) data: text, audio, video, images. This data cannot easily be analyzed using traditional statistical tecniques and the number of data analysts has not matched the exponential growth in the amount of data, which has caused much of this data to remain unanalyzed in a “*data tomb*” (Fayyad, 2003). In data mining the analyst does not need to make specific assumptions about the data nor formulate a specific hypothesis to test. The data mining process is typically **data-driven** and inductive rather than hypothesis-driven or deductive process used by statisticians.

The data mining tasks can be categorized in **predictive** tasks and **descriptive** tasks (Weiss and Davison, 2010). The predictive tasks allow to predict the value of a variable based on other existing information, while the descriptive tasks summarize the data in some manner. We briefly describe the principal predictive and descriptive data mining tasks. **Classification and regression** tasks are predictive tasks that involve building a model to predict a target, or dependent variable, from a set of explanatory or independent variables. **Association rule analysis** is a descriptive data mining task that involves discovering patterns, or associations, between elements in a data set. The associations are represented in the form of rules, or implications. The most common association rule task is *market basket analysis*. **Cluster analysis** is a descriptive data mining task where the goal is to group similar objects in the same cluster and dissimilar objects in different clusters. **Text mining**: the unstructured nature of text require special consideration. Example applications of text mining includes the identification of specific noun phrases such as people, products and companies, which can then be used in more sophisticated co-occurrence analysis to find nonobvious relationships among people or organizations. A second application area that is growing in importance is sentiment analysis, in which blogs, discussion boards, and reviews are analyzed for opinions about products or brands. **Link Analysis**: is a form of net-

work analysis that examines associations between objects. For example, given a graph showing relationships between objects, link analysis can find particularly important or well-connected objects and show where networks may be weak (e.g., in which all paths go through one or a small number of objects).

According to Mitra et al. (2002) the main challenges in the data mining procedure are: *massive data sets and high dimensionality* (huge data sets increases the size of the space of patterns); *user interaction and prior knowledge*: data mining is inherently an interactive and iterative process; *overfitting and assessing the statistical significance*: regularization and resampling methodologies need to be emphasized for model design; *understandability of patterns*: rule structuring, natural language representation, and the visualization of data and knowledge; *non-standard and incomplete data*: the data can be missing and/or noisy; *mixed media data*: learning from data that is represented by a combination of various data (media, like numeric, symbolic, images and text); *management of changing data and knowledge*: rapidly changing data (nonstationary), in a database that is modified, deleted, augmented, may make previously discovered patterns invalid (incremental methods for updating the patterns); *integration*: data mining tools are often only a part of the entire decision making system.

2.3 CART: classification and regression trees

We briefly recall some general background on Classification and Regression Trees (CART). Classification and regression tree has been an important data mining methodology for the analysis of large data sets via binary partitioning procedure (Breiman et al., 1984). It consists in recursive division of N cases on which a response variable and a set of predictors are observed. Such a partitioning procedure is known as **regression tree** when the response variable is continuously valued and as a **classification tree** when the response variable is categorical. A classification tree procedure provides not only a classification rule for new cases of unknown class, but also an analysis of the dependence structure in large data sets. Figure 1 depicts a simple tree structure with tree layers of nodes.

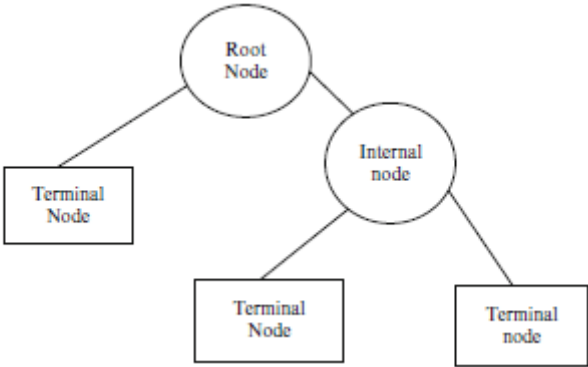


Figure 1: A simple tree structure (Y. Leung, 2010).

The root node contains the entire learning sample and the other nodes correspond to subgroups of the learning sample. The two subgroups in the left and right offspring nodes are disjoint, and their union comprises the subgroups for the parent node. A critical step of the tree-based technique is to determine the split from one parent node to two offspring nodes.

In a tree structured predictor the space X is partitioned by a sequence of binary splits into terminal nodes. In each terminal node t , the predicted response value $y(t)$ is constant. Starting with a learning sample \mathcal{L} , three elements are necessary to determine a tree predictor:

1. a way to select a split at every intermediate node;
2. a rule for determining when a node is terminal;
3. a rule for assign a value $y(t)$ to every terminal node t .

It is therefore necessary first to define a criterion of accuracy of the rule prediction; to this end it is typically used the **Mean squared error** $R(\mathbf{d})$ of the predictor \mathbf{d} that can be estimated according to following criterion.

Definition 1 (Breiman et al., 1984) Define the mean squared error $R^*(d)$ of the predictor d as

$$R^*(d) = E(Y - d(\mathbf{X}))^2 \quad (13)$$

where: $R^*(d)$ is the expected squared error using $d(\mathbf{X})$, $d(\mathbf{X})$ is a predictor of Y , $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$. The optimal predictor has a simple form:

Proposition 1 (Breiman et al., 1984) The predictor d_B which minimizes $R^*(d)$ is

$$d_B(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) \quad (14)$$

$d_B(\mathbf{x})$ is the conditional expectation of the response, given that the measurement vector is \mathbf{x} .

Given a learning sample \mathcal{L} consisting of $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n), \dots, (\mathbf{x}_N, y_N)$ to construct a predictor $d(\mathbf{x})$ and to estimate its MSE $R^*(d)$, if we use as accuracy criterion the **resubstitution estimate** for $R^*(d)$ we have:

$$R(d) = \frac{1}{N} \sum_{n=1}^N (y_n - d(\mathbf{x}_n))^2 \quad (15)$$

as the optimal predictor $y(t)$ that minimizes $R(d)$.

Proposition 2 (Breiman et al., 1984) The value of $y(t)$ that minimizes $R(d)$ is the average of y_n for all cases (\mathbf{x}_n, y_n) falling into t ; that is, the minimizing $y(t)$ is

$$\bar{y}(t) = \frac{1}{N(t)} \sum_{\mathbf{x}_n \in t} y_n \quad (16)$$

where the sum is over all y_n such that $\mathbf{x}_n \in t$ and $N(t)$ is the total number of cases in t .

So the problem of assigning a value to each node is solved by replacing the values in the node with their arithmetic mean, which represents the best forecast if you choose to resubstitution estimate of $R(d)$ as a measure of the accuracy of predictor.

If the optimal $\bar{y}(t)$ (16) represents the prediction of Y for node t and by using the notation $R(T)$ instead $R(d)$, where T is a generic regression tree we define

$$R(t) = \frac{1}{N} \sum_{\mathbf{x}_n \in t} (y_n - \bar{y}(t))^2 \quad (17)$$

and

$$R(T) = \sum_{t \in \tilde{T}} R(t) \quad (18)$$

where \tilde{T} is the set of terminal nodes of T .

So that

$$R(T) = \frac{1}{N} \sum_{t \in \tilde{T}} \sum_{\mathbf{x}_n \in t} (y_n - \bar{y}(t))^2 \quad (19)$$

where for every node t , $\sum_{\mathbf{x}_n \in t} (y_n - \bar{y}(t))^2$ is the *within node* of squares and it is the total squared deviations of the y_n in t from their average. By summing over $t \in \tilde{T}$ one obtains the total within node sum of squares, and dividing by N one provides the average. Given any set \mathcal{S} of splits of a current terminal node t in \tilde{T} ,

Definition 2 (Breiman et al., 1984) *The best split s^* of t is that split in \mathcal{S} which produces the largest reduction of $R(T)$. More precisely, for any split s of node t into t_L and t_R , let*

$$\Delta R(s, t) = R(t) - R(t_L) - R(t_R) \quad (20)$$

Take the best split s to be a split such that

$$\Delta R(s^*, t) = \max_{s \in \mathcal{S}} \Delta R(s, t) \quad (21)$$

Thus, a regression tree is constructed iteratively dividing the nodes in order to produce the maximum decrease of $R(T)$. This criterion identifies the breakdown threshold of the space of explanatory variables that most effectively separates the high response values from the low ones. Let us defined the tree thus obtained as T_{max} . To select the optimal sequence we consider the cost-complexity *pruning*:

Definition 3 (Breiman et al., 1984) *For any subtree $T \leq T_{max}$, define its **complexity** as $|\tilde{T}|$, the number of terminal nodes in T . Let $\alpha \geq 0$ be a real number called the **complexity parameter** and define the cost-complexity measure $R_\alpha(T)$ as*

$$R_\alpha(T) = R(T) + \alpha |\tilde{T}| \quad (22)$$

For each value of α , find that subtree $T(\alpha) \preceq T_{max}$ which minimizes $R_\alpha(T)$:

$$R_\alpha(T(\alpha)) = \min_{T \preceq T_{max}} R_\alpha(T) \quad (23)$$

The result is a decreasing sequence of trees $T_1 > T_2 > \dots > \{t_1\}$ with $T_1 \preceq T_{max}$ and a corresponding increasing sequence of α values $0 = \alpha_1 < \alpha_2 < \dots$ such that for $\alpha_k \leq \alpha < \alpha_{k+1}$, where $k = 1, \dots, K$ and T_k is the smallest subtree of T_{max} minimizing $R_\alpha(T)$.

This criterion leads to a cross-validation estimate of the relative error that can be used to judge the goodness of the partition tree.

To select the right sized tree from the sequence $T_1 > T_2 > \dots$ estimates of $R(T_k)$ are needed.

Let us randomly divided \mathcal{L} into V -fold cross validation $\mathcal{L}_1, \dots, \mathcal{L}_V$ such that each sub sample \mathcal{L}_v , $v = 1, \dots, V$, has the same number of cases (as nearly as possible).

For each v , this produces the trees $T^{(v)}(\alpha)$ which are the minimal error-complexity trees for the parameter value α . Grow and prune using all of \mathcal{L} , getting the sequence $\{T_k\}$ and $\{\alpha_k\}$.

The **cross-validation estimates** are given by

$$R^{CV}(T_k) = \frac{1}{N} \sum_{v=1}^V \sum_{(x_n, y_n) \in \mathcal{L}_v} (y_n - d_k^{(v)}(\mathbf{x}_n))^2 \quad (24)$$

and the corresponding relative error estimate

$$RE^{CV}(T_k) = R^{CV}(T_k)/R(\bar{y}) \quad (25)$$

$$R(\bar{y}) = \frac{1}{N} \sum_{n=1}^N (y_n - \bar{y})^2 \quad (26)$$

where $d_k^{(v)}(\mathbf{x})$ is the predictor corresponding to the tree $T^{(v)}(\alpha'_k)$ with $(\alpha'_k) = \sqrt{\alpha_k \alpha_{k+1}}$. The tree selected is T_K where K is the maximum k such that

$$R^{CV}(T_k) \leq R^{CV}(T_{k0}) + SE \quad (27)$$

where

$$R^{CV}(T_{k0}) = \min_k R^{CV}(T_k) \quad (28)$$

It is called **1-SE rule**.

In conclusion, the tree structured approach presents many advantages: it needs of only a few elements: the set of questions, a rule for selecting the best split at any node, a criterion for choosing the right-sized tree; it a powerful and flexible classification tool: it can be applied to any data structured and the final classification has a simple form which can be compactly stored and that efficiently classifies new data; it makes powerful use of conditional information in handling nonhomogeneous relationships; it does automatic stepwise variable selection and

complexity reduction; it gives not only the predicted classification but also it estimates the misclassification probability for the object; it is invariant under all monotone transformations of individual ordered variables; it is extremely robust with respect to outliers and misclassified point in the sample; it provides easily understood and interpreted information regarding the predictive structure of the data.

3 Spatial data mining and our contribution

Let us start introducing the following definition:

Definition 4 *Spatial data mining and knowledge discovery (SDMKD) is the efficient extraction of hidden, implicit, interesting, previously unknown, potentially useful, ultimately understandable, spatial or non-spatial knowledge (rules, regularities, patterns, constraints) from incomplete, noisy, fuzzy, random and practical data in large spatial databases (Deren and Shuliang, 2005).*

A spatial pattern expresses a spatial relationship among spatial objects and to extract spatial patterns from spatial data sets it is important to identify the relevant spatial objects and the properties of, and relationships between, relevant spatial objects (Malerba, 2007). We observe three principal differences with respect to *classical data mining*. First, classical data mining treats each input as independent of other inputs, whereas spatial patterns often must satisfy the constraints of continuity and high autocorrelation among nearby features. Second, classical data mining deals with numbers and categories. In contrast, spatial data is more complex and includes extended objects such as points, lines, and polygons. Spatial objects have a geometry which need to be represented. In spatial data bases, object of the same type are organized in layers, each of which can have its own set of attributes and at most one geometry attribute. Third, classical data mining works with explicit inputs, whereas spatial predicates (e.g., overlap) and attributes (e.g., distance, spatial autocorrelation) are often implicit. Spatial objects have a locational property which implicitly defines spatial relationships between objects: topological, distance and direction relations.

SDM is a confluence of databases technology, artificial intelligence, machine learning, probabilistic statistics, visualization, information science, pattern recognition and other disciplines. The specificity of SDM lies in its interaction with space. In effect, a geographical database constitutes a spatio-temporal continuum in which properties concerning a particular place are generally linked and explained in terms of the properties of its neighborhood. We can thus see the great importance of spatial relationships in the analysis process. Temporal aspects for spatial data are also a central point but are rarely taken into account (Zeitouni, 2000). It is necessary to develop new methods that consider the huge volume of data (e.g. encoding geometric location), the time consuming and the complexity of spatial relationships and spatial data handling. Basic tasks of spatial data mining are: *a) spatial classification*: finds a set of

rules which determine the class of the classified object according to its attributes; *b) spatial regression or prediction model*: the response attribute depends on the attribute values of objects spatially-related to the object to be predicted; *c) spatial association rules*: find (spatially related) rules from the database. Association rules describe patterns, which are often in the database. The association rule has the following form: $A \rightarrow B(s\%; c\%)$, where “s” is the *support* of the rule (the probability, that A and B hold together in all the possible cases) and “c” is the *confidence* (the conditional probability that B is true under the condition of A); *d) spatial clustering*: groups the object from database into clusters in such a way that object in one cluster are similar and objects from different clusters are dissimilar (*partitioning method, hierarchical method, density based method and grid-based method*); *e) spatial trend detection*: finds trends in database. A trend is a temporal pattern in some time series data. A spatial trend is defined as a pattern of change of a non-spatial attribute in the neighborhood of a spatial object.

In our approach we extend the methodology of CART in the framework of spatial econometric models in large datasets. The contribution of this work is to evaluate the effect of including spatially lagged variables, geographical coordinates or a combination of them in the set of predictors of regression tree, in terms of spatial autocorrelation among pseudo-residuals. To this end we test several versions of CART and we compare the accuracy and performance of non-spatial and spatial regression tree to predict the response variable in the context of spatial database. In particular, we assess the sensibility of various predictive models with different spatial weights matrix specifications such that to remove the spatial autocorrelation on pseudo-residuals.

The implementation is based on the package “*rpart*” (Therneau et al., 2012) in R version 2.15.1, to build a decision tree on data with minimum prediction error. Pruning for the overfit regression tree used the highest cross-validation error less than one standard error above the minimum cross-validation error. The minimum “xerror” or cross validation error was added to the “xstd” (standard deviation) creating the one standard error (1-SE) bar. The resulting value was then to determine the proper number of splits of optimal tree. In addition to this value was also determined by plotting the cross-validation relative error against the cost-complexity parameter (cp-value). To evaluate the accuracy of the fit it was determined the apparent and X-relative R^2 , where the first is derived by subtracting the relative error by one and the second is determined by subtracting one from the cross-validation error.

Finally, we calculate for different versions of CART the pseudo-residuals by function *residuals.rpart* (residuals from a fitted Rpart object).

4 Empirical Analysis

In this section we present several versions of non-spatial and spatial regression trees based on geographical coordinates and spatially lagged variables. Our approach to spatial prediction is based on both non-spatial properties of CART and on attributes and function describing spatial

relations and spatial proximity between the objects. We compare the performances of different versions of CART taking into account the effect of spatial dependence. In this empirical part our contribution is *analyzing the pseudo-residuals of regression tree looking at their spatial features* (like, e.g. spatial autocorrelation) to see whether they contain some addition hidden information. In order to deal this phenomenon, we test the procedure by considering a subset of dataset of US Southern county homicides used by Anselin (2007). The dataset is composed by 1412 Souther US counties (Washington D.C., Texas, Oklahoma, Arkansas, Louisiana, Mississippi, Alabama, Tennessee, Kentucky, Georgia, South Carolina, North Carolina, Florida, Virginia, West Virginia, Maryland and Delaware) and 7 variables (pertaining to 1960) as follow:

Name	Description
FIPSNO	Code
HR60	Homicide Rate per 100,000
RD60	Resource Deprivation/Affluence Component (principal component: percent black, log of median family income, gini index of family income inequality, percent of families female headed (percent of families single parent for 1960) and percent of families below poverty (percent of families below 3,000 dollars for 1960)
PS60	Population Structure Component (principal component: log of population and the log of population density)
UE60	Percent of civilian labor force that is unemployed
DV60	Percent of males 14 and over who are divorced
MA60	Median age

Source: <https://geodacenter.asu.edu/sdata>

The spatial distribution of the homicide rate is showed in the following map.

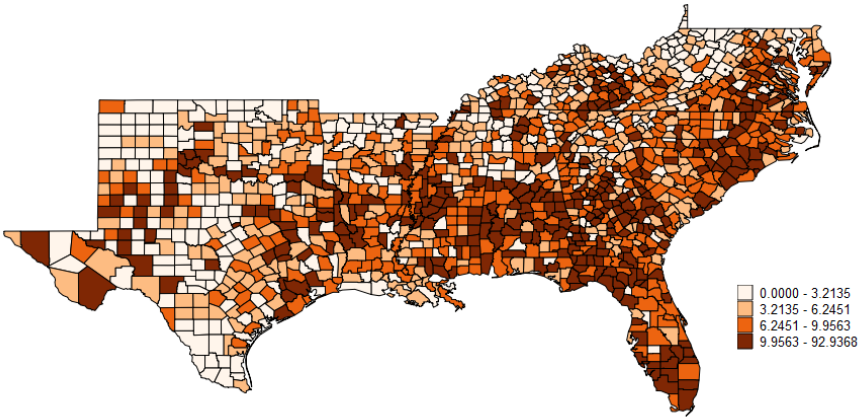


Figure 2: Map of homicide rate (HR60)

In particular we test four different versions of regression tree (RT) to predict the response variable: Homicide Rate (HR60).

Model	Set of predictors
Non-Spatial	resource deprivation, population structure, labour force unemployed, divorced rate, median age
Geographical Coordinates-based Spatial RT	resource deprivation, population structure, labour force unemployed, divorced rate, median age, coordx, coordy
W-based Spatial RT	resource deprivation, population structure, labour force unemployed, divorced rate, median age, spatially-lagged homicide rate
Geographical coordinates and W-based Spatial RT	resource deprivation, population structure, labour force unemployed, divorced rate, median age, spatially-lagged homicide rate, coordx, coordy

In the W-based spatial regression tree, to construct spatially lagged response variable, we consider different spatial weights matrices in order to check the “robustness” of pseudo-residuals spatial autocorrelation for each model. In particular we compute the following spatial weights matrices (row-standardization):

1. first-order contiguity (*rook*): the elements of which are $w_{ij} = 1$ when i and j share common border;
2. first and second order contiguity (*rook1 – 2*): it is a cumulative matrix that includes first and second order contiguity;
3. queen contiguity (*queen*): the elements of which are $w_{ij} = 1$ when i and j share common borders and common corners;
4. distance based contiguity: $dk1, dk2, dk3, dk4, dk5$ based on the minimum distance needed to make sure that all the areas are linked to at least k neighbours $\{k = 1, 2, 3, 4, 5\}$.

In order to check the influence of these matrices, in Table 1 we present the summary measures for spatial weights matrices: *number of regions, total number of links and average number of links*:

Table 1: Summary measures for different spatial weights matrices

Weigths matrix	n	total links	average number of links
rook	1412	7700	5.45
rook1-2	1412	23768	16.83
queen	1412	8096	5.73
dk1	1412	27648	19.58
dk2	1412	78432	55.55
dk3	1412	142394	100.85
dk4	1412	159558	113.00
dk5	1412	165048	116.89

The first version of regression tree is the “non-spatial regression tree” (Figure 3):

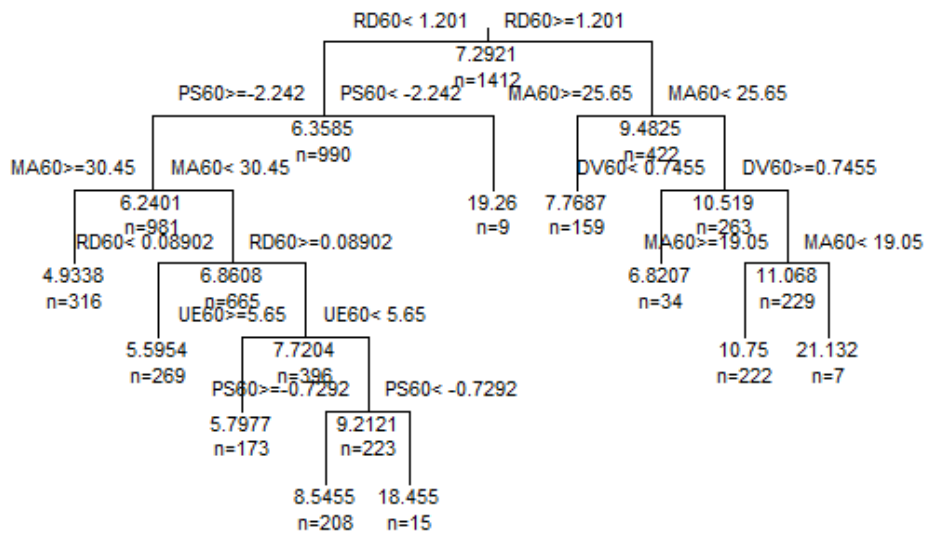


Figure 3: The non-spatial regression tree

The following plots show respectively the cross validation results and the “pseudo R-square” for different splits (Figure4):

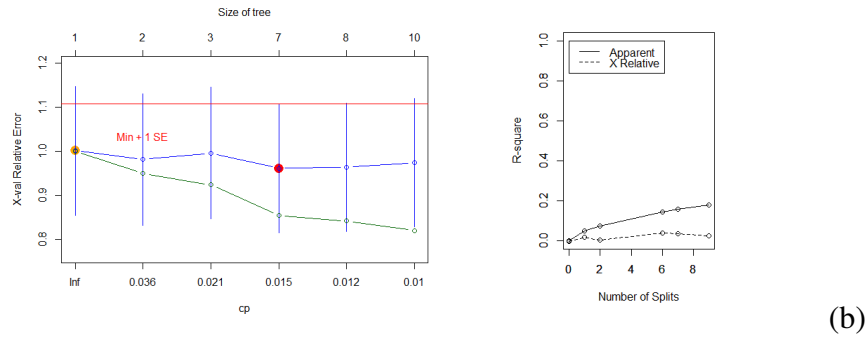


Figure 4: (a) Cross validation results of Non-spatial Regression Tree (blu line: trend of xerror, green line: trend of relative error, red line: 1-SE bar) ; (b) Apparent, X-Relative R-Square and Cross Validation Relative Error graphs of Non-spatial Regression Tree (Apparent $R^2=1$ -relative error; X relative $R^2=1$ - xerror)

The quantile map of pseudo-residuals suggests the possible presence of spatial clusters.

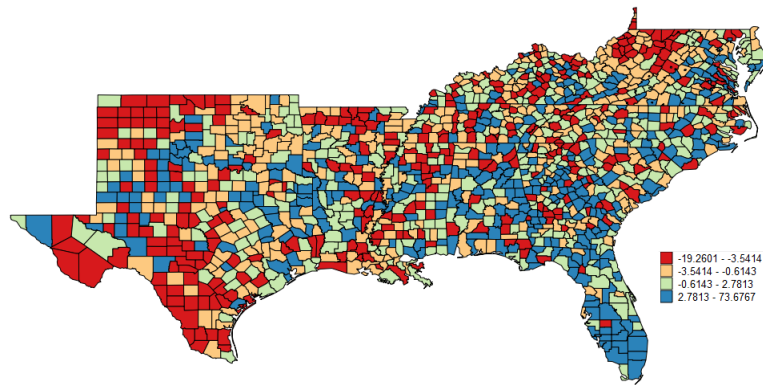


Figure 5: Quantile map of pseudo-residuals of Non-spatial Regression Tree

We also note that in geographical coordinates-based spatial regression tree, the quantile map of pseudo-residuals shows still a spatial structure of pseudo-residuals.

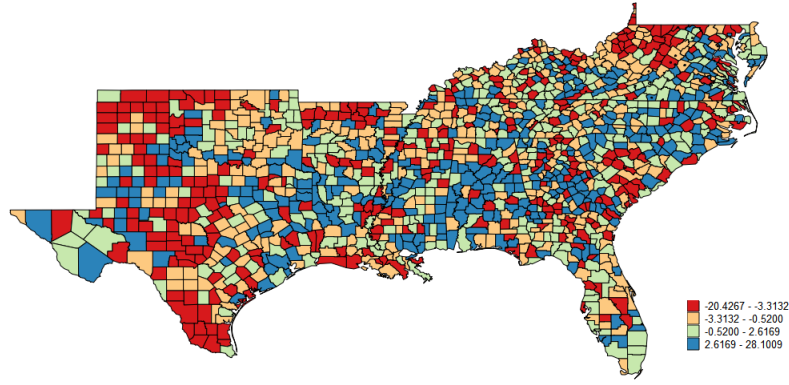


Figure 6: Quantile map of pseudo-residuals of geographical coordinates-based Spatial Regression Tree

We summarize the performance of different versions and the presence of pseudo-residuals spatial autocorrelation of regression tree. The Table 2 compares the values of permutational Moran’s I on pseudo-residuals of non spatial regression tree (without geocoords) and regression tree based on geographical coordinates (with geocoords) using different spatial weights.

Table 2: Permutational Moran’s I on pseudo-residuals of non-spatial regression tree and spatial regression tree based on geographical coordinates

Permutational Moran’s I		
Weights matrix	Without geocoords	With geocoords
rook	0.1452 (0.001)	0.0989 (0.001)
rook1-2	0.134 (0.001)	0.0998 (0.001)
queen	0.1357 (0.001)	0.0971 (0.001)
dk1	0.1273 (0.001)	0.0897 (0.001)
dk2	0.1022 (0.001)	0.0579 (0.001)
dk3	0.0838 (0.001)	0.0406 (0.001)
dk4	0.0776 (0.001)	0.0366 (0.001)
dk5	0.0761 (0.001)	0.0356 (0.001)

Notes: number of simulations=999, pseudo-pvalue in brackets, “*” statistically significant at 0.5 level.

Now, we check and compare the critical threshold distance that allows to remove the spatial autocorrelation on pseudo-residuals of non-spatial regression tree for different distance that includes at least k neighbours ($k = 1, 2, 3, 4, 5$) and we show the trend of pseudo-pvalue with respect to critical distance.

Table 3: Critical threshold distance such that to remove the spatial autocorrelation on pseudo-residuals of non-spatial regression tree

k	Threshold distance	Average number of links	Permutational Morans' I
≥ 1	2147.491	1383.572	-0.000471 (0.052)
≥ 2	2167.518	1386.317	-0.000505 (0.057)
≥ 3	2231.966	1394.048	-0.000538 (0.062)
≥ 4	2247.105	1395.623	-0.000536 (0.056)
≥ 5	2151.765	1384.183	-0.000513 (0.084)

Notes: number of simulations=999, pseudo-pvalue in brackets, “*” statistically significant at 0.5 level

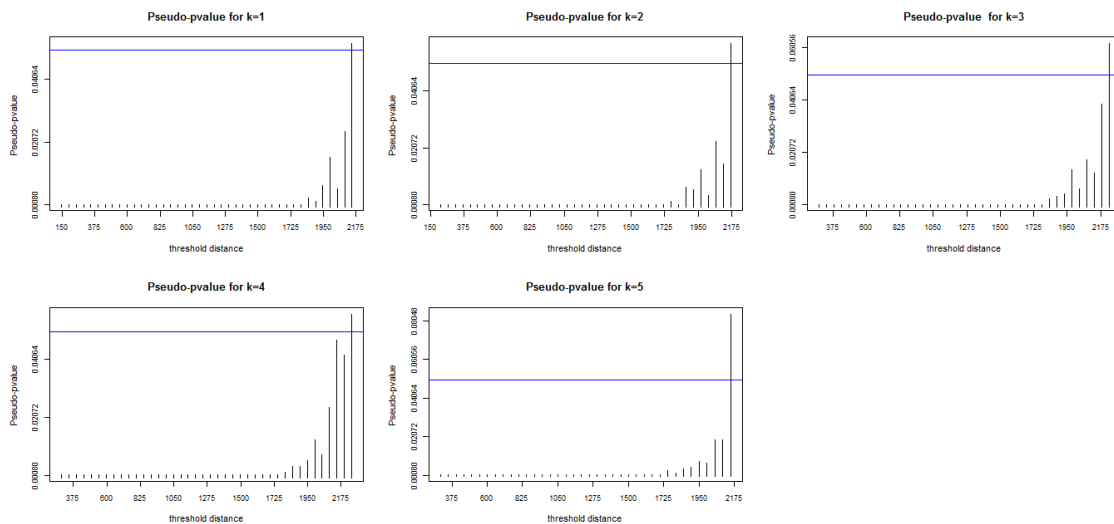


Figure 7: The trend of pseudo-pvalue on threshold distance of non-spatial tree(the blue bar indicates the statistically significant level)

We also note that the inclusion of geographical coordinates in non-spatial version of regression tree leads to a decrement of Permutational Moran's I for any spatial matrix and an improvement of accuracy, in particular the Apparent Rsquare increases from 0.180 to 0.385. The Table 4 evaluates the permutational Moran's I on the pseudo-residuals of non-spatial regression tree

and spatial regression tree based on the geographical coordinates, when we include in the set of predictors a specific lag, using different spatial weights.

Table 4: Comparison of Permutational Moran’s I on pseudo-residuals of spatial lag combined with geographical coordinates regression tree

Spatial lag	Permutational Moran’s I		Apparent Rsquare	
	without geocoords	with geocoords	without geocoords	with geocoords
rook	-0.0849 (0.001*)	-0.0849 (0.001*)	0.275	0.275
rook1-2	-0.0361 (0.001*)	-0.0361 (0.001*)	0.447	0.452
queen	-0.0891 (0.001*)	-0.0871 (0.001*)	0.267	0.287
dk1	-0.0301 (0.001*)	-0.0424 (0.001*)	0.391	0.421
dk2	-0.01 (0.032*)	-0.0086 (0.066)	0.426	0.469
dk3	0.0041 (0.885)	0.0054 (0.922)	0.388	0.464
dk4	0.0082 (0.983)	0.0076 (0.964)	0.459	0.461
dk5	0.0024 (0.802)	0.0021 (0.788)	0.350	0.433

Notes: number of simulations=999, pseudo-pvalue in brackets, “*” statistically significant at 0.5 level.

As can be seen in Table 4, in geographical coordinates-based spatial regression tree, the inclusion of spatially lagged response variable by using spatial weights matrix that includes at least two neighbours (dk2), allows to remove the presence of pseudo-residuals spatial autocorrelation. We can also, note that the critical threshold distance such that to remove the spatial autocorrelation on pseudo-residuals is 167.518 and average numbers of links is 55.546, much lower than the threshold distance in the case of non-spatial regression tree (Table 3).

Finally, the predictive spatial regression tree selected is the following:

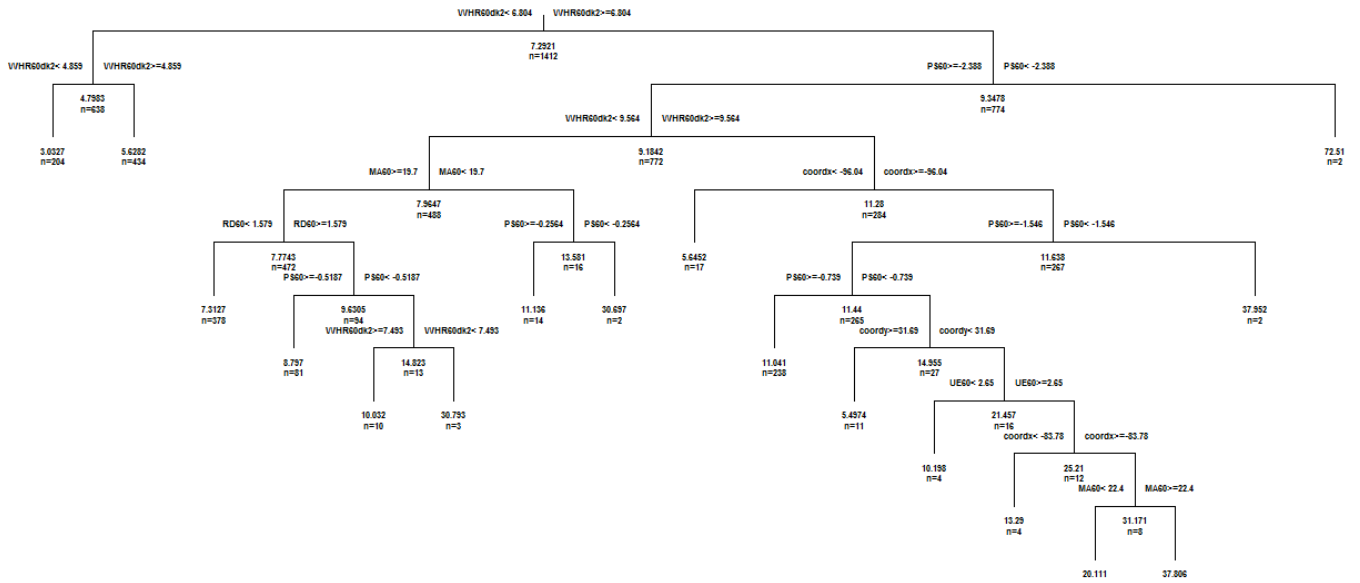


Figure 8: The minimum distance-based spatial regression tree with geographical coordinates allows to remove the presence of pseudo-residuals spatial autocorrelation (minimum distance that all regions are linked at least two neighbours)

5 Final remarks and future works

In this work we integrated some notions of spatial econometrics and spatial data mining, underlining the importance of considering spatial autocorrelation in spatial predictive tasks. In particular we analyzed the pseudo-residuals of the Classification and Regression Trees (CART), in terms of spatial autocorrelation and we showed how “the space” may add significant insights in a regression tree approach. In the presence of pseudo residuals spatial autocorrelation in a structured tree, the introduction of spatially lagged variables and geographical coordinates allows to remove this effect among pseudo residuals.

In future works we would test the procedure in different datasets or simulated data.

Finally, we would extend the approach to different mining techniques: Boosting Trees, SVM (Support Vector Machine), DBSCAN (Density-Based Spatial Clustering of Applications with Noise), Random Forests.

References

- [1] L. Anselin. *Estimation Methods for Spatial Autoregressive Structures*. Regional Science Dissertation and Monography Series, Cornell University, Ithaca, New York, 1980.
- [2] L. Anselin. Spatial externalities, spatial multipliers and spatial econometrics. *International Regional Science Review*, 26(2):153–166, 2003.
- [3] L. Anselin. *Spatial Regression Analysis in R: A Workbook*. Spatial Analysis Laboratory Department of Geography University of Illinois, Urbana-Champaign, 2007.
- [4] L. Anselin and A.K. Bera. *Spatial dependence in linear regression models with an introduction to spatial econometrics*. In Handbook of Applied Economic Statistics, ed. by A. Ullah and D. E. A. Giles. New York: Marcel Dekker, 1998.
- [5] L. Anselin and R. Florax. *New Directions in Spatial Econometrics*. Springer-Verlag, New York, 1995.
- [6] J. Behnke and E. Dobinson. Nasa workshop on issues in the application of data mining to scientific data. *SIGKDD Explor. Newsl.*, 2(1):70–79, June 2000.
- [7] R. Bivand. Regression modeling with spatial dependence: An application of some class selection and estimation methods. *Geographical Analysis*, 16(1):25–37, 1984.
- [8] L. Breiman, G. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. New York: Chapman and Hall, 1984.
- [9] Li Deren and Wang Shuliang. Concepts, principles and application of spatial data mining and knowledge discovery. In *ISSTM*, pages 27–29, August 2005 Beijing, China.
- [10] U. Fayyad. Editorial. *ACM SIGKDD Explorations*, 2(5):1–3, 2003.
- [11] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. MIT Press, 1996.
- [12] B. Fingleton. *Generalized Method of Moments Estimator for a Spatial Model with Moving Average Errors, with Application to Real Estate Prices*. In Arbia. G., and B.H.Baltagi (eds.), *Spatial Econometrics: Methods and Applications*. Physica-Verlag/Springer, 2008.
- [13] L. Hordijk. Problems in estimating econometric relations in space. *Papers of the Regional Science Association*, 42(1):99–115, 1979.
- [14] J.S. Huang. The autoregressive moving average model for spatial analysis. *Australian Journal of Statistics*, 26(2):169–78, 1984.

- [15] J.P. LeSage and R. Kelley Pace. *Introduction to Spatial Econometrics*. Boca Raton: CRC Press / Taylor Francis Group, 2009.
- [16] Y. Leung. *Knowledge Discovery in Spatial Data*. Springer-Verlag Berlin Heidelberg, 2010.
- [17] D. Malerba. Mining spatial data: Opportunities and challenges of a relational approach. In *IASC 07*, August 30th- September 1st, 2007, Aveiro, Portugal.
- [18] S. Mitra, S. K. Pal, and P. Mitra. Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, 1(13), January, 2002.
- [19] E. Simoudis. Reality check for data mining. *IEEE Expert*, 5(11):26–33, October, 1996.
- [20] P. Smyth. Data mining: data analysis on a grand scale? *Statistical Methods in Medical Research*, 4(9):309–327, August, 2000.
- [21] T. M. Therneau, B. Atkinson, and B. Ripley. Rpart: Recursive partitioning, 2012. R package version 4.01-1.
- [22] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The kdd process for extracting useful knowledge from volumes data. *Communications of the ACM, Special Issue on Data Mining*, 39(11):27–34, 1996.
- [23] G.M. Weiss and B.D. Davison. *Data Mining. Handbook of Technology Management*. H. Bidgoli (Ed.), John Wiley and Sons, 2010.
- [24] K. Zeitouni. A survey on spatial data mining methods databases and statistics. In *Point of Views, Information Resources Management Association International Conference (IRMA2000), Data Warehousing and Mining Track*, 2000.