Mining Model Trees from Spatial Data

Donato Malerba, Michelangelo Ceci, and Annalisa Appice

Dipartimento di Informatica, Università degli Studi di Bari,
via Orabona, 4 - 70126 Bari - Italy
{malerba, ceci, appice}@di.uniba.it

Abstract. Mining regression models from spatial data is a fundamental task in Spatial Data Mining. We propose a method, namely Mrs-SMOTI, that takes advantage from a tight-integration with spatial databases and mines regression models in form of trees in order to partition the sample space. The method is characterized by three aspects. First, it is able to capture both spatially global and local effects of explanatory attributes. Second, explanatory attributes that influence the response attribute do not necessarily come from a single layer. Third, the consideration that geometrical representation and relative positioning of spatial objects with respect to a reference system implicitly define both spatial relationships and properties. An application to real-world spatial data is reported.

1 Introduction

The rapidly expanding market for spatial databases and Geographic Information System (GIS) technologies is driven by the pressure from the public sector, environmental agencies and industries to provide innovative solutions to a wide range of data intensive applications that involve spatial data, that is, a collection of (spatial) objects organized in thematic layers (e.g., enumeration districts, roads, rivers). A thematic layer is characterized by a geometrical representation (e.g., point, line, and polygon in 2D) as well as several non-spatial attributes (e.g., number of inhabitants), called thematic attributes. A GIS provides the set of functionalities to adequately store, display, retrieve and manage both geometrical representation and thematic attributes collected within each layer and stored in a spatial database. Anyway, the range of GIS applications can be profitably extended by adding spatial data interpretation capabilities to the systems. This leads to a generation of GIS including Spatial Data Mining (SDM) facilities [11].

Spatial Data Mining investigates how interesting and useful but implicit knowledge can be extracted from spatial data [8]. Regression is a fundamental task of SDM where the goal is to mine a functional relationship between a continuous attribute Y_i (*response attribute*) and m continuous or discrete attributes $X_{j,i}$ $j = 1, ..., m$ (*explanatory attributes*). The training sample consists of spatial objects. For instance, for UK census data available at the level of Enumeration Districts (EDs), a possible goal may be estimating the response attribute "number of migrants" associated to each ED i on the basis of explanatory attributes $X_{j,i}$ (e.g., "number of inhabitants") associated to EDs.

A. Jorge et al. (Eds.): PKDD 2005, LNAI 3721, pp. 169–180, 2005.
© Springer-Verlag Berlin Heidelberg 2005

The simplest approach to mine regression models from spatial data, is based on standard regression [18] that models a functional relationship in the form: $Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i}$, where i is each ED area. The main problem with this model is that it disregards the arrangement properties due to spatial structure of data [5] (e.g., the phenomenon of migration is typically stronger in peripheral EDs). When spatially-dependent heterogeneity of the model can be anticipated by the analyst, the model can be improved by introducing a dummy variable $D_i \in \{0, 1\}$, which differentiate the behavior of the model according to a predefined partitioning of areas in two groups. In this way, the model is either $Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} + \gamma D_i$ (constant spatial variation) or $Y_i = \beta_0 + (\beta_1 + \gamma D_i) X_{1,i} + \ldots + \beta_k X_{k,i}$ (regression parameter spatial variation). However, when the areas of homogeneous dependence cannot be anticipated by the expert, a solution is represented by model trees [16] that approximate a piece-wise (linear) function by means of a tree structure, where internal nodes partition the sample space (as decision trees), while leaves are associated to (linear) functions. In this way, it is possible to automatically determine different regression model for different areas.

In this paper, we propose the model tree induction method, namely Mrs-SMOTI (Multi-relational Spatial Stepwise Model Tree Induction), that faces several degrees of complexity which characterize the regression problem from spatial data. In the next section, we discuss these problems and introduce our solution. Section 3 presents a stepwise approach to mine spatial regression models. Section 4 focuses on spatial database integration. Finally, an application is presented in Section 5 and some conclusions are drawn.

2 Spatial Regression: Background and Motivations

While model tree learning has been widely investigated in the data mining literature [16,10,17], as far as we know, no attention has been given to the problem of mining model trees from spatial data. Model tree induction from spatial data raises several distinctive problems: i) some explanatory attributes can have spatially global effect on the response attribute, while others have only a spatially local effect; ii) explanatory attributes that influence the response attribute not necessarily come from a single layer, but in most of cases they come from layers possibly spatially related with the layer that is the main subject of the analysis; iii) geometrical representation and relative positioning of spatial objects with respect to some reference system implicitly define both spatial relationships (e.g., "intersects", "distance") and spatial attributes (e.g., "area", "direction").

Concerning the first point, it would be useful to identify the global effect of some attributes (possibly) according to the space arrangement of data. Indeed, in almost all model trees induction methods, the regression model associated with a leaf is built on the basis of those training cases falling in the corresponding partition of the feature space. Therefore, models in the leaves have only a local validity and do not consider the global effects that some attributes might have in the underlying model. In model trees, global effects can be represented by

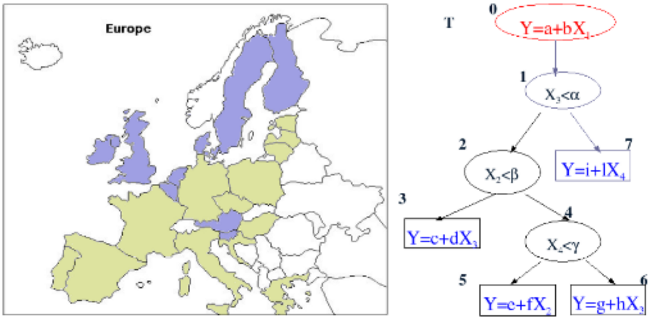


Fig. 1. An example of spatial model tree with regression and splitting nodes. Node 0 is a regression node that captures a global effect between the unemployed rate (Y) and the GDP per capita (X_1). It is associated to all countries. Node 1 splits the sample space as depicted in the map. Functions at leaves only capture local effects.

attributes that are introduced in the linear models at higher levels of the tree. This requires a different tree-structure where internal nodes can either define a partitioning of the sample space or introduce some regression attributes in the linear models to be associated to the leaves. In our previous work [10], we proposed the method SMOTI whose main characteristic is the construction of trees with two types of nodes: splitting nodes, which partition the sample space, and regression nodes, which perform only straight-line regression. The multiple model associated to a leaf is built stepwise by combining straight-line regressions along the path from the root to the leaf. In this way, internal regression nodes contribute to the definition of multiple models and capture global effects, while straight-line regressions at leaves capture only local effects. Detecting global and local effects over spatial data, allows to model phenomena, that otherwise, would be ignored. As an example we show a simplistic case: suppose we are interested in analyzing the unemployed rate in EU. In this case, it may be found that the unemployed rate of each country is proportional to its GDP (Gross Domestic Product) per capita. This behavior is independent of the specific country and represents a clear example of global effect. This global effect corresponds to a regression node in higher levels of the tree (see Fig. 1).

The second point enlightens that the value of the response attribute may go beyond the values of explanatory attributes of the spatial object to be predicted. In particular, it is possible that the response attribute depends on the attribute values of objects spatially-related to the object to be predicted and possibly belonging to a different layer. In this point of view, the response attribute is associated to the spatial objects that are the main subjects of the analysis (target objects) while each explanatory attribute refers either to the target objects to be predicted or to the spatial objects that are relevant for the task in hand and are spatially related to the target ones (non-target objects). This is coherent with the idea of exploiting intra-layer and inter-layer relationships when mining spatial data [1]. Intra-layer relationships describe a spatial interaction between two spatial objects belonging to the same layer, while inter-layer relationships describe a spatial interaction between two spatial objects belonging to different layers. According to [5], intra-layer relationships make available both spatially-lagged explanatory attributes useful when the effect of an explanatory attribute at any site is not limited to the specified site (e.g., the proportion of people suffering from respiratory diseases in an ED also depends on the high/low level of pollution of EDs where people daily move) and spatially lagged response attribute, that is, when autocorrelation affects the response values (e.g., the price for a good at a retail outlet in a city may depend on the price of the same good sold by local competitors). Differently, inter-layer relationships model the fact that the response attribute value observed from some target object may depend on explanatory attributes observed at spatially related non target objects belonging to different layers. For instance, if the "EDs" layer is the subject of the analysis and the response attribute is the mortality rate associated to an ED, mortality rate may depend on the air-pollution degree on crossing roads. Although spatial regression systems (such as the R spatial project - http://sal.uiuc.edu/csiss/Rgeo/index.html) are able to deal with user defined intra-layer spatial relationships, they ignore inter-layer relationships that can be naturally modeled by resorting to the multi-relational setting [4].

The third point is due to the fact that geometrical representation and relative positioning of spatial objects with respect to some reference system implicitly define both spatial relationships and spatial attributes . This implicit information is often responsible for the spatial variation over data and it is extremely useful in modelling [15]. Hence, spatial regression demands for the development of specific methods that, differently from traditional ones, take the spatial dimension of the data into account when exploring the spatial pattern space. In this way, thematic and spatial attribute values of target objects and spatially related non-target objects are involved in predicting the value of the response attribute.

The need of extracting and mining the information that is implicitly defined in spatial data motivates a tight-integration between spatial regression method and spatial database systems where some sophisticated treatment of real-world geometry is provided for storing, indexing and querying spatial data. This is confirmed by the fact that spatial operations (e.g., computing the topological relationships among spatial objects) are available free of charge for data analysts in several spatial database advanced facilities [6].

In this work, we present Mrs-SMOTI that extends SMOTI by taking advantage of a tight integration with a spatial database in order to mine stepwise a spatial regression model from multiple layers. The model is built taking into account all three degrees of complexity presented above.

3 Stepwise Mining of a Spatial Regression Model

Mrs-SMOTI mines a spatial regression model by partitioning training spatial data according to intra-layer and inter-layer relationships and associating different regression models to disjoint spatial areas. In particular, it mines spatial data and performs the stepwise construction of a tree-structured model with both splitting nodes and regression nodes until some stopping criterion is satisfied. In this way, it faces the spatial need of distinguishing among explanatory attributes that have some global effect on the response attribute and others that have only local effect. Both splitting and regression nodes may involve several layers and spatial relationships among them.

Spatial split. A spatial splitting test involves either a *spatial relationship condition* or a *spatial attribute condition* on some layer from S. The former partitions target objects according to some spatial relationship (either intra-layer or inter-layer). For instance, when predicting the proportion of people suffering from respiratory diseases in EDs, it may be significant to mine a different regression function according to the presence or absence of main roads crossing the territory. An extra-consequence of performing such spatial relationship condition concerns the introduction of another layer in the model. The latter is a test involving a boolean condition ("$X \leq \alpha$ vs. $X > \alpha$" in the continuous case and "$X \in \{x_1, \ldots, x_k\}$ vs. $X \notin \{x_1, \ldots, x_k\}$" in the discrete one) on a thematic attribute X of a layer already included in the model. In addition to thematic attributes, an attribute condition may involve a spatial property (e.g., the area for polygons and the extension for lines), that is implicitly defined by the geometrical structure of the corresponding layer in S. It is noteworthy that only spatial relationship conditions add new layers of S to the model. Consequently, a split on a thematic attribute or spatial property involves a layer already introduced in the model. However, due to the complexity of computing spatial relationships, we impose that a relationship between two layers can be introduced at most once in each unique path connecting the root to the leaf.

Coherently with [10], the validity of a spatial splitting test is based on an heuristic function $\sigma(t)$ that is computed on the attribute-value representation of the portion of spatial objects in S falling in t_L and t_R, that is, the left and right child of the splitting node t respectively. This attribute-value representation corresponds to the tuples of S derived according to both spatial relationship conditions and attribute conditions along the path from the root of the tree to the current node. We define $\sigma(t) = (n(t_L)/(n(t_L) + n(t_R)))R(t_L) + (n(t_R)/(n(t_L) + n(t_R)))R(t_R)$, where $n(t_L)$ $(n(t_R))$ is the number of attribute-value tuples passed down to the left (right) child. Since intra-layer and inter-layer relationships lead to a regression model that may include several layers (not necessarily separate), it may happen that $n(t) \neq n(t_L) + n(t_R)$ although the split in t satisfies the mutual exclusion requirement. This is due to the many-to many nature of intra-layer and inter-layer relationships. In fact, when several spatial objects are spatially related to the same object (e.g., a single ED may be intersected by zero, one or more roads), computing spatial relationships may return a number of attribute-value tuples greater than one. $R(t_L)$ $(R(t_R))$ is the Minimum Squared Error (MSE) computed on the left (right) child t_L (t_R) as follows:

$$R(t_L) = \sqrt{\frac{1}{n(t_L)} \sum_{i=1 \ldots n(t_L)} (y_i - \widehat{y}_i)^2} \quad \left(R(t_R) = \sqrt{\frac{1}{n(t_R)} \sum_{i=1 \ldots n(t_R)} (y_i - \widehat{y}_i)^2}\right),$$

such that $\widehat{y_i}$ is the response value predicted according to the spatial regression model built by combining the best straight-line regression associated to t_L (t_R), with all straight-line regressions in the path from the root to t_L (t_R) [3].

Spatial regression. A spatial regression node performs a straight-line regression on either a continuous thematic attribute or a continuous spatial property not yet introduced in the model currently built. Coherently with the stepwise procedure [3], both response and explanatory attributes are replaced with their residuals. For instance, when a regression step is performed on a continuous attribute X, the response attribute is replaced with the residual $Y' = Y - \hat{Y}$, where $\hat{Y} = \hat{\alpha} + \hat{\beta}X$. The regression coefficients $\hat{\alpha}$ and $\hat{\beta}$ are estimated on the attribute-value representation of the portion of S falling in the current node.

According to the spatial structure of data, the regression attribute comes from one of the layers already involved in the model. Continuous thematic and spatial attributes of these layers, which have not yet been introduced in the model, are replaced with the corresponding residuals in order to remove the effect of the regression attribute. Whenever a new layer is added to the model (by means of a spatial relationship condition), continuous thematic and spatial attributes, introduced with it, are replaced with the corresponding residuals. Residuals are contextually computed on the attribute-value representation of the portion of S falling in the current node. In this way, the effect of regression attributes previously introduced in the model by regression steps is also removed by introduced attributes.

The evaluation of a spatial regression step $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$ is based on the heuristic function $\rho(t)$, that is: $\rho(t) = \min\{R(t), \sigma(t')\}$, where t' is the best spatial splitting node following the regression step in t. This look-ahead step involved in the heuristic function above depends on the fact that spatial split looks for best straight-line regression after the split condition is performed, while the regression step does not. A fairer comparison would be growing the tree at a further level to base the computation of $\rho(T)$ on the best multiple linear regressions after the regression step on X_i is performed [10].

Stopping criteria. Three different stopping criteria are implemented. The first requires that a minimal number of target objects fall in current node. The second stops the induction process when the coefficient of determination is greater than a threshold [18]. This coefficient is a scale-free one-number summary of the strength of the relation between explanatory attributes in the actual multiple model and the response attribute. Finally, the third stops the induction process when no further regression step can be performed (i.e. all continuous attributes are included in the current model) also after introducing some new layer.

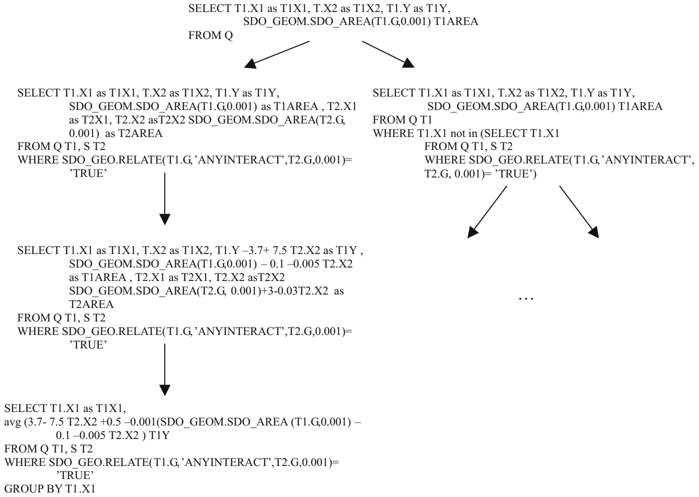
4 Spatial Database Integration

Most spatial data mining systems process data in main memory. This results in high performance for computationally intensive processes when enough memory is available to store all necessary data. However, in spatial data intensive processes it is important to exploit powerful mechanisms for accessing, filtering and indexing data, such as those available in spatial DBMS (DataBase Management Systems). For instance, spatial operations (e.g., computing the topological relationships among spatial objects) supported by any spatial DBMS take advantage from spatial indexes like Quadtrees or Kd-tree [14]. This motivates a tight integration of spatial data mining systems and spatial DBMS in order to i) guarantee the applicability of spatial data mining algorithms to large spatial datasets; ii) exploit useful knowledge of spatial data model available, free of charge, in the spatial database, iii) specify directly what data stored in a database have to be mined, iv) avoid useless preprocessing leading to redundant data storage that may be unnecessary when part of space of the hypothesis may be never explored.

Some examples of integrating spatial data mining and spatial database system are presented in [11] for classification tasks and in [1] for association rules discovery tasks. In both cases, a data mining algorithm working in first-order logic is only loosely integrated with a spatial database by means of some middle layer module that extracts spatial attributes and relationships independently from the mining step and represents these features in a first-order logic formalism. Thus, data mining algorithms are practically applied to preprocessed data and this preprocessing is user-controlled. Conversely, in [6] a spatial data mining system, named SubgroupMiner, is proposed for the task of subgroup discovery in spatial databases. Subgroup discovery is here approached by taking advantage from a tight integration of the data mining algorithm with the database environment. Spatial relationships and attributes are then dynamically derived by exploiting spatial DBMS extension facilities (e.g., packages, cartridges or extenders) and used to guide the subgroup discovery.

Following the inspiration of SubgroupMiner, we assume an object-relational (OR) data representation, such that spatial patterns representing both splitting and regression nodes are expressed with spatial queries. These queries include spatial operators based on the non-atomic data type for geometry consisting in an ordered set of coordinates (X, Y) representing points, lines and polygons. Since no spatial operator is present in basic relational algebra or Datalog, we resort to an extension of the OR-DBMS Oracle Spatial Cartridge $9i$ where spatial operators to compute spatial relationships and to extract spatial attributes are made available free of charge [6]. These operators can be called in SQL queries. For example: *SELECT * FROM EDs x, Roads y WHERE SDO_GEOM.*
RELATE(x.geometry,’ANYINTERACT’,y.geometry, 0.001) = ’TRUE’
This spatial query retrieves the pairs ⟨ED, Road⟩ whose topological relationship is “not disjoint” by means or the Oracle operator ”RELATE”. It is noteworthy that, the use of such SQL queries, appears to be more direct and much more practical than formulating non-trivial extension of relational algebra or Datalog such that those provided in constraint database framework [9].

When running a spatial query (associated to a node of the tree), the result is a set of tuples describing both thematic attributes and spatial attributes of involved layers. The FROM clause includes layers (not necessarily different) in the model at the current node. The WHERE clause includes split conditions found along the path from the root to the current node. The negation of either a



```
SELECT T1.X1 as T1X1, T.X2 as T1X2, T1.Y as T1Y,
            SDO_GEOM.SDO_AREA(T1.G,0.001) T1AREA
FROM Q
```

```
SELECT T1.X1 as T1X1, T.X2 as T1X2, T1.Y as T1Y,
            SDO_GEOM.SDO_AREA(T1.G,0.001) as T1AREA , T2.X1
            as T2X1, T2.X2 asT2X2 SDO_GEOM.SDO_AREA(T2.G,
            0.001)  as T2AREA
FROM Q T1, S T2
WHERE SDO_GEO.RELATE(T1.G,'ANYINTERACT',T2.G,0.001)=
            'TRUE'
```

```
SELECT T1.X1 as T1X1, T.X2 as T1X2, T1.Y as T1Y,
            SDO_GEOM.SDO_AREA(T1.G,0.001) T1AREA
FROM Q T1
WHERE T1.X1 not in (SELECT T1.X1
            FROM Q T1, S T2
            WHERE SDO_GEO.RELATE(T1.G,'ANYINTERACT',
            T2.G, 0.001)= 'TRUE')
```

```
SELECT T1.X1 as T1X1, T.X2 as T1X2, T1.Y –3.7+ 7.5 T2.X2 as T1Y ,
            SDO_GEOM.SDO_AREA(T1.G,0.001) – 0.1 –0.005 T2.X2
            as T1AREA , T2.X1 as T2X1, T2.X2 asT2X2
            SDO_GEOM.SDO_AREA(T2.G, 0.001)+3-0.03T2.X2  as
            T2AREA
FROM Q T1, S T2
WHERE SDO_GEO.RELATE(T1.G,'ANYINTERACT',T2.G,0.001)=
            'TRUE'
```

...

```
SELECT T1.X1 as T1X1,
avg (3.7- 7.5 T2.X2 +0.5 –0.001(SDO_GEOM.SDO_AREA (T1.G,0.001) –
            0.1 –0.005 T2.X2 ) T1Y
FROM Q T1, S T2
WHERE SDO_GEO.RELATE(T1.G,'ANYINTERACT',T2.G,0.001)=
            'TRUE'
GROUP BY T1.X1
```

Fig. 2. An example of spatial model tree with regression, splitting and leaf nodes expressed by means of spatial queries assuming that training data are stored in spatial layers (e.g., Q and R) of a spatial database

spatial relationship condition or an attribute condition involving some attribute of a non-target layer is transformed into a negated nested spatial sub-query. This is coherent with the semantic of tests involving multiple tables of a relational database [2]. Finally, the SELECT clause includes thematic and spatial attributes (or their residuals) from the layers involved in the WHERE clause.

Leaf nodes are associated with aggregation spatial queries, that is, spatial queries where all tuples referring the same target object are grouped together. In this way, the prediction of the response variable is the average response value predicted on the set of attribute-value tuples describing the unique target object to be predicted. This means that spatial model trees can be expressed in form of a set of SQL spatial queries (see Fig. 2). Queries are stored in XML format that can be subsequently used for predicting (unknown) response attributes.

5 Spatial Regression on Stockport Census Data

In this section we present a real-world application concerning the mining of spatial regression models. We consider both 1991 census and digital map data provided in the context of the European project SPIN! (Spatial Mining for Data of Public Interest) [12]. This data concerns Stockport, one of the ten metropolitan districts in Greater Manchester (UK) which is divided into twenty-two wards for a total of 589 census EDs. Spatial analysis is enabled by the availability of vectorized boundaries for 578 Stockport EDs as well as by other Ordnance Survey digital maps of UK. Data are stored in an Oracle Spatial Cartridge 9i database.

The application in this study investigates the number of unemployed people in Stockport EDs according to the number of migrant people available for each ED in census data as well as geographical factors represented in topographic maps stored in form of layers. The target objects are the Stockport EDs, while other layers, such as, shopping (53 objects), housing (9 objects) and employment areas (30 objects) are the non target objects. The EDs play the role of both target objects and non target objects when considering intra-layer relationship on EDs.

Two experimental settings are defined. The first setting (BK_1) is obtained by exclusively considering the layer representing EDs. The second setting (BK_2) is obtained by considering all the layers. In both settings, intra-layer relationships on EDs make possible to model the unemployment phenomenon in Stockport EDs by taking into account the self-correlation on the spatially lagged explanatory attributes of EDs. The auto-correlation on the spatially-lagged response attribute can be similarly exploited during the mining process. In this study, we consider (intra-layer and inter-layer) spatial relationships that describe some (non disjoint) topological interaction between spatial objects. Furthermore, we consider area of polygons and extension of lines as spatial properties.

In order to prove the advantage of using intra-layer and inter-layer relationships in the mining process, we compare the spatial regression model mined by Mrs-SMOTI with the regression models mined by SMOTI and M5'[17]. Since SMOTI and M5' work under single table assumption, we transform the original object-relational representation of Stockport data in a single relational table format. Two different transformations are considered. The former (P1) creates a single table by deriving all thematic and spatial attributes from layers according to all possible intra-layer and inter-layer relationships. This transformation leads to generate multiple tuples for the same target object. The latter transformation (P2) differs from the previous one because it does not generate multiple tuples for the same target object. This is obtained by including aggregates (i.e., the average for continuous values and the mode for discrete values)[7] of the attributes describing the non target objects referring to the same target object[1].

Model trees are mined by requiring that the minimum number of spatial target objects falling in an internal node must be greater than the square root of the number of training target objects, while the coefficient of determination must be below 0.80. Comparison is performed on the basis of the average MSE, number of regression nodes and leaves obtained by means of the same five-fold cross validation of Stockport data. Results are reported in Table 1.

The non-parametric Wilcoxon two-sample paired signed rank test [13] is used for the pairwise comparison of methods. In the Wilcoxon signed rank test, the summations on both positive (W+) and negative (W-) ranks determine the winner. Results of Wilcoxon test are reported in Table 2.

[1] In both P1 and P2 transformations the attribute-value dataset is composed by 5 attibutes for BK_1 (6 when including the lagged response) and 11 for BK_2 (12 when including the lagged response). The number of tuples for P1 is 4033 for BK_1 and 4297 for BK_2. In the case of P2, the number of tuples is 578 in both settings.

Table 1. Average MSE, No. of leaves and regression nodes of trees induced by Mrs-SMOTI, SMOTI and M5'. L1 is "No lagged response", L2 is "Lagged response".

Setting		MSE				Leaves				RegNodes			
		BK1		BK2		BK1		BK2		BK1		BK2	
		L1	L2	L1	L2	L1	L2	L1	L2	L1	L2	L1	L2
Mrs-SMOTI		12.34	13.74	11.99	10.92	19.80	23.40	23.60	23.60	3.4	6.6	3.8	6.2
SMOTI	P1	12.91	10.23	20.11	13.0	101.6	107.6	104.0	111.8	6.2	5.0	15.0	11.4
	P2	11.89	18.17	19.71	15.80	41.00	24.80	42.40	44.20	3.4	4.0	10.2	11.6
M5'	P1	13.52	12.41	12.92	12.30	433.6	872.0	408.6	711.2	-	-	-	-
	P2	12.44	9.19	12.48	9.59	198.0	199.4	199.2	197.4	-	-	-	-

Results confirm that Mrs-SMOTI is better or at worst comparable to SMOTI in terms of predictive accuracy. This result is more impressive when we consider the regression model mined when both intra-layer and inter-layer relationships are ignored. The average MSE of model trees mined by SMOTI taking into account only the number of migrants and the area of EDs is 15.48.

Moreover, when we consider results of SMOTI on data transformed according to P1 and P2, we note that the stepwise construction takes advantage of the tight-integration of Mrs-SMOTI with the spatial DBMS that avoids the generation of useless features (relationships and attributes). The side effect of useless features may lead to models that overfit training data, but fail in predicting new data. In a deeper analysis, we note that even when SMOTI, in average, outperforms Mrs-SMOTI in terms of MSE, the Wilcoxon test does not show any statistically significant difference. Results on the two data settings show that mining the geographical distribution of shopping (housing or employment) areas over EDs (i.e., the spatial relationships between EDs and shopping areas, shopping areas and shopping areas, shopping areas and employment areas, and so on) decreases the average MSE of models mined by Mrs-SMOTI, while no significant improvement is observed in mining the same information with SMOTI. The autocorrelation on the response improves performance of Mrs-SMOTI only for BK_2 level (10.92 vs. 11.99) without significantly increasing tree size.

Table 2. Mrs-SMOTI vs SMOTI and M5': results of the Wilcoxon test on the MSE of trees. If W+$\leq$ W- then results are in favour of Mrs-SMOTI. The statistically significant values ($p \leq 0.1$) are in boldface. L1 is "No lagged response", L2 is "Lagged response".

Setting		Mrs-SMOTI vs. SMOTI P1			Mrs-SMOTI vs. SMOTI P2			Mrs-SMOTI vs. M5' P1			Mrs-SMOTI vs. M5' P2		
		W+	W-	p	W+	W-	p	W+	W-	p	W+	W-	p
BK1	L1	6	9	0.81	9	6	0.81	3	12	0.310	7	8	1.000
	L2	10	5	0.63	6	9	0.81	8	7	1.000	**15**	**0**	**0.060**
BK2	L1	**1**	**14**	**0.125**	**0**	**15**	**0.06**	4	11	0.430	6	9	0.810
	L2	**0**	**15**	**0.06**	3	12	0.31	**0**	**15**	**0.060**	**15**	**0**	**0.060**

- **split** on *EDs'* number of migrants [$\leq$ 47] (578 EDs)
    - **regression** on *EDs'* area (458 EDs)
        - **split** on *EDs - Shopping areas* spatial relationship (458 EDs)
            - **split** on *Shopping areas'* area (94 EDs) ...
            - **split** on *EDs'* number of migrants (364 EDs) ...
    - **split** on *EDs'* area (120 EDs)
        - **leaf** on *EDs'* area (22 EDs)
        - **regression** on *EDs'* area (98 EDs) ...

Fig. 3. Top-level description of a portion of the model mined by Mrs-SMOTI on the entire dataset at BK_2 level with no spatially lagged response attributes

The number of regression nodes and leaves are indicators of the complexity of the induced regression models. In this case, results show that the model induced by Mrs-SMOTI is much simpler than the model induced by SMOTI in both settings independently from data transformation. The relative simplicity of the spatial regression models mined by Mrs-SMOTI makes them easily to be interpreted. In particular, the tree structure can be easily navigated in order to distinguish among global and local effects of explanatory attributes. For instance, in Fig. 3 it is shown the top-level description of the spatial regression model mined by Mrs-SMOTI on the entire dataset at BK_2 level with no spatially lagged response attributes. Mrs-SMOTI captures the global effect of the area of EDs over Stockport covered by the 458 EDs having "number of migrants $\leq$ 47". The effect of this regression is shared by all nodes in the corresponding sub-tree.

Finally, the comparison of Mrs-SMOTI with M5', does not show any clear difference in terms of MSE. Anyway, M5' presents two important disadvantages with respect to Mrs-SMOTI. First, M5' cannot capture spatial global and local effects. Second, mined model trees cannot be interpreted by humans because of the complexity of the models (there is an increase of one order of magnitude in the number of leaves from Mrs-SMOTI to M5')

6 Conclusions

In this paper we have presented a spatial regression method Mrs-SMOTI that is able to capture both spatially global and local effects of explanatory attributes. The method extends the stepwise construction of model trees performed by its predecessor SMOTI in two directions. First, by taking advantage from a tight-integration with a spatial database in order to mine both spatial relationships and spatial attributes which are implicit in spatial data. Indeed, this implicit information is often responsible for the spatial variation over data and it is extremely useful in regression modelling. Second, the search strategy is modified in order to mine models that capture the implicit relational structure of spatial data. This means that spatial relationships (intra-layer and inter-layer) make possible to consider explanatory attributes that influence the response attribute but do not necessarily come from a single layer. In particular, intra-layer relationships make available spatially lagged response attributes in addition to spatially lagged explanatory attributes.

Experiments on real-world spatial data show the advantages of the proposed method with respect to SMOTI. As future work, we intend to extend the method in order to mine both geometrical (e.g., distance) and directional (e.g., north of) relationships in addition to topological relationships.

Acknowledgment

This work is partial fulfillment of the research objective of ATENEO-2005 project “Gestione dell’informazione non strutturata: modelli, metodi e architetture”.

References

1. A. Appice, M. Ceci, A. Lanza, F. A. Lisi, and D.Malerba. Discovery of spatial association rules in georeferenced census data: A relational mining approach. *Intelligent Data Analysis*, 7(6):541–566, 2003.
2. A. Appice, M. Ceci, and D. Malerba. Mining model trees: A multi-relational approach. In T. Horvath and A. Yamamoto, editors, *Proceedings of ILP 2003*, volume 2835 of *LNAI*, pages 4–21. Springer-V., 2003.
3. N. R. Draper and H. Smith. *Applied regression analysis.* John Wiley & Sons, 1982.
4. S. Džeroski and N. Lavrač. *Relational Data Mining.* Springer-V., 2001.
5. R. Haining. *Spatial data analysis in the social and environmental sciences.* Cambridge University Press, 1990.
6. W. Klosgen and M. May. Spatial subgroup mining integrated in an object-relational spatial database. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of PKDD 2002*, volume 2431 of *LNAI*, pages 275–286. Springer-V., 2002.
7. J. Knobbe, M. Haas, and A. Siebes. Propositionalisation and aggregates. In L. D. Raedt and A. Siebes, editors, *Proceedings of PKDD 2001*, volume 2168 of *LNAI*, pages 277–288. Springer-V., 2001.
8. K. Koperski. *Progressive Refinement Approach to Spatial Data Mining.* PhD thesis, Computing Science, Simon Fraser University, British Columbia, Canada, 1999.
9. G. Kuper, L. Libkin, and L. Paredaens. *Constraint databases.* Springer-V., 2001.
10. D. Malerba, F. Esposito, M. Ceci, and A. Appice. Top down induction of model trees with regression and splitting nodes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):612–625, 2004.
11. D. Malerba, F. Esposito, A. Lanza, F. A. Lisi, and A. Appice. Empowering a gis with inductive learning capabilities: The case of ingens. *Journal of Computers, Environment and Urban Systems, Elsevier Science*, 27:265–281, 2003.
12. M. May. Spatial knowledge discovery: The spin! system. In K. Fullerton, editor, *Proceedings of the EC-GIS Workshop*, 2000.
13. M. Orkin and R. Drogin. *Vital Statistics.* McGraw Hill, New York, USA, 1990.
14. H. Samet. *Applications of spatial data structures.* Addison-Wesley longman, 1990.
15. S. Shekhar, P. R. Schrater, R. Vatsavai, W. Wu, and S. Chawla. Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia*, 4(2):174–188, 2002.
16. L. Torgo. *Inductive Learning of Tree-based Regression Models.* PhD thesis, Department of Computer Science, University of Porto, Porto, Portugal, 1999.
17. Y. Wang and I. Witten. Inducing model trees for continuous classes. In M. Van Someren and G. Widmer, editors, *Proceedings of ECML 1997*, pages 128–137, 1997.
18. S. Weisberg. *Applied regression analysis.* Wiley, New York, USA, 1985.