Learning and Transferring Geographically Weighted Regression Trees across Time

Annalisa Appice, Michelangelo Ceci, Donato Malerba, and Antonietta Lanza

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

Abstract. The Geographically Weighted Regression (GWR) is a method of spatial statistical analysis which allows the exploration of geographical differences in the linear effect of one or more predictor variables upon a response variable. The parameters of this linear regression model are locally determined for every point of the space by processing a sample of distance decay weighted neighboring observations. While this use of locally linear regression has proved appealing in the area of spatial econometrics, it also presents some limitations. First, the form of the GWR regression surface is globally defined over the whole sample space, although the parameters of the surface are locally estimated for every space point. Second, the GWR estimation is founded on the assumption that all predictor variables are equally relevant in the regression surface, without dealing with spatially localized collinearity problems. Third, time dependence among observations taken at consecutive time points is not considered as information-bearing for future predictions. In this paper, a tree-structured approach is adapted to recover the functional form of a GWR model only at the local level. A stepwise approach is employed to determine the local form of each GWR model by selecting only the most promising predictors. Parameters of these predictors are estimated at every point of the local area. Finally, a time-space transfer technique is tailored to capitalize on the time dimension of GWR trees learned in the past and to adapt them towards the present. Experiments confirm that the tree-based construction of GWR models improves both the local estimation of parameters of GWR and the global estimation of parameters performed by classical model trees. Furthermore, the effectiveness of the time-space transfer technique is investigated.

1 Introduction

A main assumption underpinning geographic thinking is spatial non-stationarity, according to which a phenomenon varies across a landscape. In a regression task, where the predictor variables and the response variable are collected at several locations across the landscape, the major consequence of spatial non-stationarity is that the relationship between the predictor variables and the response variable is location-dependent. The consequence of this spatial variability is that a spatial analyst is discouraged from employing any conventional regression-based

M. Atzmueller et al. (Eds.): MSM/MUSE 2011, LNAI 7472, pp. 97–117, 2012.

model which assumes the independence of observations from the spatial location. Indeed, LeSage and Pace [19] have shown that the application of conventional regression models leads to wrong conclusions in spatial analysis and generates spatially autocorrelated residuals. One of the best known approaches to spatial regression is GWR (Geographical Weighted Regression) [2], a spatial statistics technique which addresses some challenges posed by spatial non-stationarity. In particular, GWR maps a local model as opposed to the global linear model conventionally defined in statistics, in order to fit the relationship between predictor variables and a response variable. In fact, unlike the conventional regression equation which defines single parameter estimates, GWR generates a *parametric* linear equation, where parameter estimates vary from location to location across the landscape. Each set of parameters is estimated on the basis of distance-weighted neighboring observations. The choice of a neighborhood is influenced by the observation that the positive spatial autocorrelation of a variable is common to many geographical applications [14]. In particular, the positive spatial autocorrelation of the response variable occurs when the response values taken at pairs of locations a certain distance apart (neighborhood) are be more similar than expected for randomly associated pairs of observations [18].

The focus of our attention on GWR is motivated by a number of recent publications which have demonstrated that this local spatial model is appealing in areas of spatial econometrics, including climatology [6], social segregation [20], industrialisation [15] as well as environmental and urban planning [28]. Despite of this, there are still research issues which are not faced by GWR. In the following we introduce a novel local algorithm which aims to solve them.

The main issue of GWR is that it outputs a single parametric equation, which represents the linear combination of all the predictor variables in the task, and considers the coefficients of the linear combination as parameters for the local estimation. This means that GWR assumes that predictor variables are all equally relevant for the response everywhere across the landscape, although it admits spatially varying parameters. Consequently, GWR does not deal with the spatially localized phenomenon of *collinearity*. In general, collinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly linearly correlated. In this case, the coefficient estimates may change erratically in response to small changes in the model or the data, thus decreasing the predictive accuracy. In conventional regression, the problem of collinearity is addressed by identifying the subset of the relevant predictor variables and outputting the linear combination of only the variables in this subset [11]. Based on this idea, we argue that a solution to the spatial collinearity in GWR is to determine a parametric regression surface, which linearly combines a subset of the predictor variables. As we expect that variables in the subset may vary in space, we define a new spatially local regression algorithm, called GWRT (*G*eographically *W*eighted *R*egression *T*rees learner), which integrates a spatially local regression model learner with a tree-based learner. The tree-based learner recursively segments the landscape along the spatial dimensions (e.g. latitude and longitude), according to a measure of the positive spatial

autocorrelation over the response values. In practice, the leaves of an induced tree represent a segmentation of the landscape into non-overlapping areal units which spatially reference response values positively autocorrelated within the corresponding leaf. The high concentration of autocorrelated response values falling in a leaf motivates the search for a parametric surface equation to be associated to the leaf. The leaf surface reasonably combines only a subset of relevant predictor variables and the parameters of this surface are locally estimated across the leaf. In particular, at each leaf, the predictor variables and the local parameters are learned by adapting the forward stepwise approach [11] defined for a global aspatial models to local spatial model learning.

Another important issue of both GWR and GWRT is that they do not capitalize on the *time dependence* among observations repeatedly collected across the same landscape at consecutive time points. This issue cannot be neglected due to the ubiquity of sensor network applications which continuously feed an unbounded amount of georeferenced and timestamped data (for instance, the temperature is periodically measured by weather stations across the Earth's surface). We face this issue by tailoring a transfer learning technique which adapts predictions obtained by geographically weighted regression trees learned by GWRT in the recent past towards the present. The research problem we consider focuses on applying knowledge from one set of past instances of a task to improve the performance of learning the same task in the present [26]. In our case the transferred model will reflect the spatial non-stationarity of phenomenon at present and also the time dependence among consecutive observations of the same phenomenon. For the transfer process, we sample few training key data in the present which are regularly distributed across the landscape. For each key observation, a transfer observation is computed with one predictor variable for each GWRT tree to be transferred from the past. The transferred model is a (spatially piecewise) regression model learned from these transfer data.

Therefore, the innovative contributions of this work with respect to original formulation of GWR are highlighted as follows. We propose a tree-based learner which allows the segmentation of the landscape in non-overlapping areal units that group positively autocorrelated response values. We do not assume any global form of the geographically weighted regression model, but we allow the variation across the landscape of the subset of predictive variables included the model. We design a stepwise technique to determine a geographically weighted regression model, where only the most promising predictive variables are selected. We define a transfer technique which allows us to use geographically weighted regression trees previously learned on past source domain data in order to improve the accuracy of prediction over the target domain data. We empirically prove that geographically weighted regression trees allow us a more accurate prediction of unknown response values spread across the landscape than three competitive methods: the traditional spatial statistic predictor GWR, the inductive aspatial model tree learner M5' [30] and the transductive spatial regression learner SpReCo [4] . Finally, we evaluate the viability of the transfer technique in a real application.

The paper is organized as follows. In the next Section we revise related work on regression in spatial statistics and spatial data mining as well as related work on inductive transfer learning. In Section 3, we illustrate the problem setting and introduce some preliminary concepts. In Section 4, we present the geographically weighted regression tree induction algorithm. In Section 5, we present the inductive transfer of this kind of tree-based models learned in the recent past to the present. In Section 6 we describe experiments we have performed with several benchmark spatial data collections. Finally, we draw some conclusions and outline some future work.

2 Background and Related Work

In order to clarify the background of this work, in this Section we illustrate related research on regression in both spatial data analysis and transfer learning.

2.1 Spatial Regression

Several definitions of the regression task have been formulated in spatial data analysis over the years. The formulation we consider in this work is the traditional one, where a set of attribute-value observations for the predictor variables and the response variable are referenced at point locations across the landscape. So far, several techniques have been defined to perform this task, both in spatial statistics and spatial data mining. A brief survey of these techniques (e.g. k-NN, geographically weighted regression, kriging) is reported in [27].

In particular, the k-Nearest Neighbor (k-NN) algorithm [23] is a machine learning technique which appears to be a natural choice for dealing with the regression task in spatial domains. Each test observation is labeled with the (weighted) mean of the response values of neighboring observations in the training set. A distance measure is computed to determine neighbors. As spatial coordinates can be used to determine the Euclidean distance between two positions, k-NN predicts the response value at one position by taking into account the observations which fall in the neighborhood. Thus, k-NN takes into account a form of positive autocorrelation over the response attribute only.

GWR [2] is a spatial statistic technique which extends the regression framework defined in conventional statistics by rewriting a globally defined model as a locally estimated model. The global regression model is a linear combination of predictor variables, defined as: $y = \alpha + \sum_{k=1}^{n} \beta_k x_k + \epsilon$ with intercept α and parameters β_k globally estimated for the entire landscape, by means of the least square regression method [11]. Then GWR rewrites this equation in terms of a *parametric* linear combination of predictor variables, where the parametric coefficients (intercept and parameters) are locally estimated at each location across the landscape. Formally, the parametric model at location i is in the form:

$$y(u_i, v_i) = \alpha(u_i, v_i) + \sum_{k=1}^{n} \beta_k(u_i, v_i) x_k(u_i, v_i) + \epsilon_i, \quad (1)$$

where (u_i, v_i) represents the coordinate location of i, $\alpha(u_i, v_i)$ is the intercept at location i, $\beta_k(u_i, v_i)$ is the parameter estimate at the location i for the predictor variable x_k, $x_k(u_i, v_i)$ is the value of the k-th variable for location i, and ϵ_i is the error term. Intercept and parameter estimates are based on the assumption that observations near one another have a greater influence on each other. The weight assigned to each observation is computed on the basis of a distance decay function centered on the observation i. This decay function is modified by a bandwidth setting, that is, at which distance the weight rapidly approaches zero. The bandwidth is chosen by minimizing the Akaike Information Criteria (AIC) score [7]. The choice of the weighting scheme is a relevant step in the GWR procedure and, at this purpose, several different weighting functions are defined in the literature [2]. The more common weighting functions are Gaussian and the bi-square kernels.

Kriging [5] is a spatial statistic technique which exploits positive autocorrelation and determines a local model of the spatial phenomenon. It applies an optimal linear interpolation method to estimate unknown response values $y(u_i, v_i)$ at each location i across the landscape. $y(u_i, v_i)$ is decomposed into a structural component, which represents a mean or constant trend, a random but spatially correlated component and a random noise, which expresses measurement errors or variations inherent to the attribute of interest.

A different approach is reported in [22], where the authors present a relational regression method (Mrs-SMOTI) that builds a regression model tightly integrated with a spatial database. The method considers the geometrical representation and relative positioning of the spatial objects of different types to decide the split condition for the tree induction (e.g., towns crossed by a river and towns not crossed by any river). For the splitting decision the heuristic based on the error reduction is used. The regression problem addressed in this paper is clearly different from the task faced by Mrs-SMOTI, as we assume data which are produced at a time by a sensor network, i.e. measurements of one or more variable taken from sensors which are georeferenced through the latitude-longitude position of the measuring sensor. In any case, we have found appealing the idea of partitioning data and learning the model in a stepwise fashion at each partition to solve the problem of linear collinearity also in spatial domains. We extend this idea by proposing a segmentation of the landscape in areal units according to Boolean tests on spatial dimensions and not on predictor variables as traditional model trees do. The segmentation is tailored to identify boundaries of areal units across the landscape which group (positively) autocorrelated response values. Finally, the regression model associated to each leaf is built stepwise, but it is also synthesized to be a locally estimated regression model.

Finally, SpReCo [4] is a data mining method which addresses the spatial regression problem in a transductive learning setting and takes into account the autocorrelation of spatial data by resorting to a co-training algorithmic solution. Traditional model tree based regressors are learned from two different views of the same data. One view is defined only on original predictor variables. The other view, which accounts for the possible spatial autocorrelation, is based on

aggregate variables, whose values are derived by aggregating measurements of the predictor variables in the neighborhood of each considered spatial location. According to the co-training paradigm, the model learned from a view is used to predict unlabeled data for the other during the learning process. However, only some unlabeled data are considered, namely the most reliable. The final prediction of unlabeled observations is the weighted average of the regression estimates generated by both learners. According to the transductive formulation of a regression problem, SpReCo inputs both labeled and unlabeled georeferenced data and outputs a prediction of the unlabeled ones, but no regression model is produced to predict data which are not available during the learning phase.

2.2 Transfer Learning

The major assumption in many data mining techniques is that the training and future data must be in the same feature space and have the same distribution. However, in many real world applications, this assumption may not hold. For example, in time-dependent spatial applications, such as applications of sensor network analysis, we may have to define a regression task for the domain of interest in the present (target domain), but have sufficient training data for this task available only in the past (source domain). In these cases, past data may follow a different data distribution with respect to present data and knowledge transfer would greatly improve the performance of learning, by avoiding much expensive labeling effort. In recent years, transfer learning has emerged as a new learning framework to address this kind of problem.

A survey focusing on categorizing and reviewing the current progress in transfer learning is reported in [26]. This survey revises several transfer learning techniques which are defined for different data mining tasks, including regression. Independently of the task, these techniques are classified with respect to the learning setting in which they operate. In particular, the learning setting, and consequently the transfer technique, may be inductive, transductive or semi-supervised. In the inductive setting, the target task is different from the source task, although it does not matter whether the source and target domains are the same or not. Some labeled data in the target domain are required to transfer an objective predictive model for use in the target domain. In the transductive setting, the source and target tasks are the same, while the source and target domains are different. No labeled data in the target domain are available, while labeled data in the source domain are available. Finally, in the unsupervised transfer learning setting, similar to the inductive transfer learning setting, the target task is different from but related to the source task. There are no labeled data available in either source and target domains in training. According to this categorization, the transfer learning problem we address in this paper stops halfway between the transductive transfer setting and the inductive transfer setting. As in the transductive setting, we have a unique task which admits several timestamped domains. In particular, we observe that the source timestamped domains share the same feature vector which varies across the landscape, but this spatial data distribution may drift in time [12]. On the other hand, as in the

inductive transfer setting, we assume the existence of some labeled observations in the present domain (the target one), which are used to transfer the predictive models learned in the past source domains to the present ones. This transfer learning problem is related to that of transferring a knowledge from WiFi localization models across time periods and space to perform WiFi localization tasks [31,25]. Additionally, this transfer with a single task is also connected to domain adaptation which has already been investigated with similar assumptions for the knowledge transfer in text classification [9].

Although the existence of this somehow related research is documented in the literature, to the best of our knowledge our work is the first attempt to tailor a transfer technique of a purely spatially local regression model to a framework of spatio-temporal data analysis.

3 Problem Setting and Preliminary Concepts

In this Section, we formulate a definition for the regression relationship between the predictor variables and the response variable observed in a geographically distributed environment. This relationship is defined according to a field-based [29] modeling of the variables which allows us to fit ubiquity of data across the landscape. The inductive regression task is formulated to learn a definition of the regression relationship in a geographically distributed training sample. We propose to address this task by learning a piecewise definition of a space-varying (parametric) regression function which met the requirements of spatial non-stationarity posed by this task without suffering of collinearity problems. Finally, the transfer learning task is formulated to allow us to transfer regression models on a landscape across time.

3.1 Spatial Regression Definition

Formally, a *spatial regression relationship*, denoted as $\tau(U, V, Y, X_1, X_2, \ldots, X_m)$, defines the (possibly unknown) space-varying relationship between a response numeric variable Y and m predictor numeric variables X_j (with $j = 1, \ldots, m$). This relationship varies across a 2D landscape $U \times V$ (e.g. Latitude $\times$ Longitude) due to the phenomenon of spatial non-stationarity. In this formulation and according to the *field-based model*, the variation of both the response variable and the predictor variables across the landscape is mathematically defined by means of one response function $y(\cdot, \cdot)$ and m distinct predictor functions $x_j(\cdot, \cdot)$ which are respectively:

$$y \colon \mathbb{U} \times \mathbb{V} \mapsto \mathbb{Y} \qquad x_j \colon \mathbb{U} \times \mathbb{V} \mapsto \mathbb{X}_j \text{ (with } j = 1, \ldots, m), \tag{2}$$

where $\mathbb{U} \times \mathbb{V} \subseteq \mathbb{R} \times \mathbb{R}$ is the range of the Cartesian product $U \times V$; $\mathbb{Y}$ is the numeric range of response function $y(\cdot, \cdot)$ (variable Y); $\mathbb{X}_j$ is the numeric range of the predictor function $x_j(\cdot, \cdot)$ (variable X_j).

An extensional definition D of the relationship τ comprises any set of observations which are simultaneously collected across the landscape according

to both the response function ($y(\cdot,\cdot)$) and the predictor functions ($x_j(\cdot,\cdot)$ with $j = 1,\ldots,m$). The observation i of this set is the data tuple defined as follows:

$$[i,\ u_i, v_i,\ x_1(u_i, v_i),\ x_2(u_i, v_i),\ \ldots,\ x_m(u_i, v_i),\ y(u_i, v_i)], \tag{3}$$

where i is the primary key of the data tuple one-to-one associated to the point location with coordinates (u_i, v_i). $x_j(u_i, v_i)$ is the value measured for the predictor variable X_j at the location (u_i, v_i) across the landscape, while $y(u_i, v_i)$ is the (possibly unknown) value measured for the response variable Y at (u_i, v_i). The response value $y(u_j, v_j)$ may be unknown, in this case the tuple i is unlabeled.

3.2 Spatial Regression Inductive Task

The *inductive regression task* associated to τ can be formulated as follows. Given a training data set $T \subset D$ which consists of a sample of n randomly tuples taken from D and labeled with the known values for the response variable. The goal is to learn a space-varying functional representation $f\colon \mathbb{U} \times \mathbb{V} \mapsto \mathbb{R}$ of the relationship τ such that f can be used to predict unlabeled responses at any location across the landscape. Our proposal to address this task consists of a new learner which receives training data T as input and outputs a *piecewise* definition for the space-varying function f which is defined as a geographically weighted regression tree. This tree recursively partitions the landscape surface along the spatial dimensions U and V and associates the areal unit at each leaf with a parametric (space-varying) linear combination of an opportunely chosen subset of the predictor variables. The parameters of this equation are locally estimated at each training location which falls in the leaf.

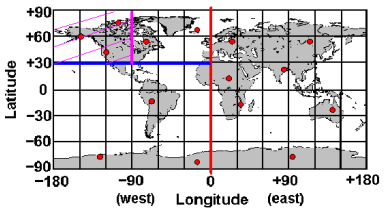
3.3 Geographically Weighted Regression Tree

Formally a *geographically weighted regression tree* f is defined as a binary tree $f = (N, E)$ where:

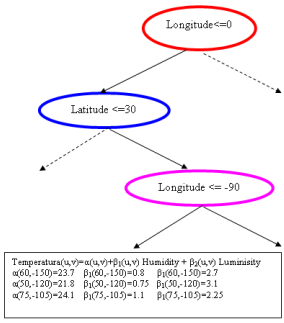
1. each node $n \in N$ is either an internal node or a leaf node ($N = N_I \cup N_L$)
   (a) an internal node $n \in N_I$ identifies a rectangular surface $s(n)$ over the landscape $U \times V$. The root identifies the entire landscape $U \times V$;
   (b) a leaf node $n \in N_L$ is associated with a parametric multiple linear regression function, that, for each location i falling in $s(n)$, allows the prediction of y_i according to predictor values and coefficients of the linear combination as they are locally estimated at the location i;
2. each edge $(n_i, n_j) \in E$ is a splitting edge labeled with a Boolean test over U or V which allows the identification of the rectangular surface $s(n_j) \subset s(n_i)$.

An example of a geographically weighted regression tree is reported in Figure 1.

Once the geographically weighted regression tree f is learned, it can be used to predict the response for any unlabeled observation $i' \in D$. During classification, the leaf of f which spatially contains i' is identified. The parametric function associated to this leaf is then applied to predict the unknown response value of i' by taking into account the $(u_{i'}, v_{i'})$ localization of i'.



(a) Sensors possibly installed across the Earth to obtain geographically distributed training observations of the Temperature (response), Humidity and Luminosity (predictors).



(b) A geographically weighted regression tree.

Fig. 1. An example of geographically weighted regression trees with spatial splits and space-varying parametric functions at the leaves (b) learned from spatial data (a).

3.4 Transfer of Spatial Regression Models across Time

By adding the time dimension to the spatial regression task formulation, data sets are collected across the same landscape, but at distinct time points. As the manual labeling of large data sets can be very costly, it is reasonable that after an initial extensive activity of labeling, regression model(s) learned from the past would be used to predict unknown label of data at the present time point. As data distribution may drift in time, any learned model should also take this drift into account. The challenge of the transfer learning is that of allowing us to avoid that a regression model is learned again form scratch; indeed the learning operation will require to label a large amount of data to guarantee a quite accurate training. To take under control the labeling cost, only few data are labeled at the present time point and they are used to transfer regression models learned in the past across the time and adapt them to the present data at best. The hypothesis we investigate in this paper is that in presence of a set of scarcely and sparsely labeled data, we can gain more accuracy by labeling the unlabeled data with regression models transferred from the past than by using a new regression model learned from a small training set. With this aim we formulate a transfer learning task. Given, the definition of a spatial regression task $\tau(Y, X_1, X_2, \ldots, X_m, U, V)$; a series of w geographically weighted regression trees f_j, each one learned from a training source domain D_j on collected on $\mathbb{U} \times \mathbb{V}$ for the task τ at the past time t_j; and observations in the *key target set* $K \subset D$ (with responses) timestamped with the present time point. The transfer learner induces a target predictive function $f_{f_1,f_2,\ldots,f_w,K} : \mathbb{U} \times \mathbb{V} \mapsto \mathbb{Y}$ by

using $f_1, f_2, \ldots f_w$ and the response values of the observations in K which allows us to predict the unlabeled observations of any *testing target set* T taken across $U \times V$ at the same time point of K.

4 Geographically Weighted Regression Tree Induction

Based on the classical Top-Down Induction of Decision Tree framework, GWRT recursively partitions the landscape in non-overlapping areal units and finds a parametric piecewise prediction model that fits training data in these areal units. Details of partitioning phase and regression model construction phase are discussed in this section. We also explain how geographically weighted regression trees can predict unknown response values across the landscape.

4.1 Splitting phase

The partitioning phase is only based on the spatial dimensions of the data. The choice of the best split is based on the well known Global Moran autocorrelation measure [18], computed on the response variable Y and defined as follows:

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{N}(y_i - \overline{y})^2}, \tag{4}$$

where y_i is the value of the response variable at the observation i, $\overline{y}$ is the mean of the response variable, w_{ij} is the spatial distance based weight between observations i and j and N is the number of observations.

The Gaussian kernel is an obvious choice to compute the weights:

$$w_{ij} = \begin{cases} e^{(-0.5 d_{ij}^2/h^2)} & \text{if } d_{ij} \leq h \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where h is the bandwidth and d_{ij} is the Euclidean spatial distance between observations i and j. The basic idea is that observations that show high positive spatial autocorrelation in the response variable should be kept in the same areal unit. Therefore, for a candidate node t, the following measure is computed:

$$I_t = \frac{(I_L N_L + I_R N_R)}{N_L + N_R} \tag{6}$$

where I_L (I_R) represents the Global Moran autocorrelation measure computed on the left (right) child of t and N_L (N_R) is the number of training observations falling in the left (right) child of t. The higher I_t, the better the split.

The candidate splits are in the form $u_i \leq \gamma_u$ or $v_i \leq \gamma_v$, where (u_i, v_i) is the spatial position of observation i. Candidate γ_u and γ_v values are determined by finding $n_{bins} - 1$ candidate equal-frequency cut points for each spatial dimension.

Our motivation of an autocorrelation measure as a splitting heuristic is that we look for a segmentation of the landscape in regions of highly correlated data which may lead to accurate GWR regression models [2].

The stopping criterion to label a node as a leaf requires the number of training observations in each node to be less than a minimum threshold. This threshold is set to the square root of the total number of training observations, which is considered a good locality threshold that does not allow too much loss in accuracy ever for rule-based classifiers [13].

4.2 Model Construction Phase

For each leaf a parametric linear regression function is associated to the areal unit associated to the leaf. Parameters of this function re estimated at each training location falling in such areal unit. After the tree is completely built, it defines a piecewice regression function f in this form $f(u,v) = \sum_{i=1}^{l} I((u,v) \in D_i) \times f_i(u,v)$, where l is the number of leaves and $D_1, D_2, \ldots, D_l$ represent the segmentation of the landscape due to the spatial partition defined by the tree; $f_i(\cdot,\cdot)$ is the parametric linear function learned for the areal unit D_i; and $I(\cdot)$ is an indicator function returning 1 if its argument is true and 0 otherwise.

Each parametric linear regression function is a parametric linear combination of a subset of the predictor variables. The variables are selected according to a forward selection strategy. Thus, the function is built with a stepwise process which starts with no variable in the function and tries out the variables one by one, including the best variable if it is "statistically significant". For each variable included in the model, parameters of the output combination are locally estimated across the landscape covered by the leaf areal unit.

To explain the stepwise construction of a parametric regression function we illustrate an example. Let us consider the case we build a function of the response variable Y with two predictor variables X_1 and X_2 and estimate the space-varying parameters of this function at the location (u_i, v_i). Our proposal is to equivalently build the parametric function:

$$\hat{y}(u_i v_i) = \alpha(u_i, v_i) + \beta(u_i, v_i)\ x_1(u_i, v_i) + \gamma(u_i, v_i)\ x_2(u_i, v_i), \tag{7}$$

through a sequence of parametric straight-line regressions. At this aim, we start by regressing Y on X_1 and building the parametric straight line

$$\hat{y}(u_i, v_i) = \alpha_1(u_i, v_i) + \beta_1(u_i, v_i)\ x_1(u_i, v_i). \tag{8}$$

This equation does not predict Y exactly. By adding the variable X_2, the prediction might improve. However, instead of starting from scratch and building a new function with both X_1 and X_2, we follow the stepwise procedure. First we build the parametric linear model for X_2 if X_1 is given, that is, $\hat{x}_2(u_i, v_i) = \alpha_2(u_i, v_i) + \beta_2(u_i, v_i) x_1(u_i, v_i)$. Then we compute the parametric residuals on

both the predictor variable X_2 and the response variable Y, that is:

$$x'_2(u_i, v_i) = x_2(u_i, v_i) - (\alpha_2(u_i, v_i) + \beta_2(u_i, v_i)x_1(u_i, v_i)) \quad (9)$$
$$y'(u_i, v_i) = y(u_i, v_i) - (\alpha_1(u_i, v_i) + \beta_1(u_i, v_i)x_1(u_i, v_i)). \quad (10)$$

Finally, we determine a parametric straight-line regression between parametric residuals Y' on X'_2, that is,

$$\hat{y}'(u_i, v_i) = \alpha_3(u_i, v_i) + \beta_3(u_i, v_i)x'_2(u_i, v_i). \quad (11)$$

By substituting Equations 9-10, we reformulate Equation 11 as follows:

$$y(u_i, v_i) - (\alpha_1(u_i, v_i) + \beta_1(u_i, v_i)x_1(u_i, v_i)) = \alpha_3(u_i, v_i) + \quad (12)$$
$$+ \beta_3(u_i, v_i)(x_2(u_i, v_i) - (\alpha_2(u_i, v_i) + \beta_2(u_i, v_i)x_1(u_i, v_i))).$$

This equation can be equivalently written as:

$$\hat{y}(u_i, v_i) = (\alpha_3(u_i, v_i) + \alpha_1(u_i, v_i) - \alpha_2(u_i, v_i)\beta_3(u_i, v_i)) + (\beta_1(u_i, v_i) - \quad (13)$$
$$- \beta_2(u_i, v_i)\beta_3(u_i, v_i))x_1(u_i, v_i) +$$
$$+ \beta_3(u_i, v_i)x_2(u_i, v_i).$$

It can be proved that the parametric function reported in this Equation coincides with the geographically weighted model built with Y, $X1$ and $X2$ (in Equation 7) since:

$$\alpha(u_i, v_i) = \alpha_3(u_i, v_i) + \alpha_1(u_i, v_i) - \alpha_2(u_i, v_i)\beta_3(u_i, v_i), \quad (14)$$
$$\beta(u_i, v_i) = \beta_1(u_i, v_i) - \beta_2(u_i, v_i)\beta_3(u_i, v_i) \quad (15)$$
$$\gamma(u_i, v_i) = \beta_3(u_i, v_i). \quad (16)$$

By considering the stepwise procedure illustrated before, two issues remain to be discussed: how parametric intercept and slope of a straight line regression (e.g. $\hat{y}(u_i, v_i) = \alpha(u_i, v_i) + \beta(u_i, v_i)\ x_j(u_i, v_i)$) are locally estimated across the landscape and how predictor variables to be added to the function are chosen.

The parametric slope and intercept are defined on the basis of the *weighted* least squares regression method [11]. This method is adapted to fit the geographically distributed arrangement of the data. In particular, for each training location which contributes to the computation of the straight-line function, the parametric slope $\beta(u_i, v_i)$ is defined as follows:

$$\beta(u_i, v_i) = (L^T \mathbf{W}_i L)^{-1} L^T \mathbf{W}_i Z, \quad (17)$$

where L represents the vector of the values of X_j on the training observations, Z is the vector of Y values on the same observations and $\mathbf{W}_i$ is a diagonal matrix defined for the training locations (u_i, v_i)as follows: $\mathbf{W}_i = \begin{pmatrix} w_{i1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & w_{iN} \end{pmatrix}$,

where w_{ij} is computed according to Equation 5. Finally, the parametric intercept $\alpha(u_i, v_i)$ is computed according to the function:

$$\alpha(u_i, v_i) = \frac{1}{N}\sum_i z_i - \beta(u_i, v_i) \times \frac{1}{N}\sum_i l_i. \tag{18}$$

where z_i and l_i are members of Z and L, respectively.

The choice of the best predictor variable to be included in the model at each step is based on the maximization of the Akaike information criterion (AIC) measure. The AIC is a measure of the relative goodness of fitting of a statistical model. First proposed in [3], AIC is based on the concept of information entropy, and offers a relative measure of the information lost when a given model is used to describe reality. It can be said to describe the trade-off between the bias and the variance in model construction, or, loosely speaking, between the accuracy and the complexity of the model. In this work we use the corrected AIC (AIC_c) [16] that has proved to give good performance even for small datasets [7]:

$$AIC_c = 2N\ ln(\hat{\sigma}) + N\ ln(2\pi) + N\left(\frac{N+p}{N-2-p}\right), \tag{19}$$

where N is the number of training data falling in the leaf, $\hat{\sigma}$ is the standard deviation of training residuals for the response variable and p is the number of parameters (number of variables included in the model – degrees of freedom of a χ^2 test). AIC_c is used to compare regression models; however, it does not provide a test of a model in the usual sense of testing a null hypothesis; i.e. AIC can tell nothing about how well a model fits the data in an absolute sense. This means that if all the candidate models fit poorly, AIC will not give any warning. To overcome this problem, once the best variable to be added to the function is identified, the new function is evaluated according to the partial F-test. This test allows the evaluation of the statistical contribution of a new predictor variable to the model [11]. If the contribution is not statistically significant, the previous model is kept and no further variable is added to the regression function.

4.3 Prediction

Once a geographically weighted regression tree T is learned, it can be used for prediction purposes. Let o be any georeferenced observation with unknown response value, then the leaf of T spatially containing o is identified. If a training parameter estimation exists in this leaf computed for the spatial coordinates (u_o, v_o), then these estimates are used to predict $y(u_o, v_o)$ according to the local function associated to the leaf; otherwise, the k closest training parameter estimations falling in the same leaf are identified. Closeness relation is computed by the Euclidean distance. These estimated neighbor parameters are used to obtain k predictions of the response value, then a weighted combination of these responses is output. Weights are defined according to the Gaussian schema.

5 Geographically Weighted Regression Tree Transfer

Let τ be a spatial regression relationship, $f_{q-w}, f_{q-w+1}, \ldots$, and f_q be a series of w[1] functions (in our study, geographically weighted regression trees) learned for the task at the source time points $t_{q-w}, t_{q-w+1}, \ldots$, and t_q and D_{q+1} be the target domain of the task which refers to the time point t_{q+1}. The unknown response values of data collected in the time point t_{q+1} are predictable by the unknown target predictive function f_{q+1}. Then the goal is to obtain a definition of f_{q+1} by transferring the source trees towards the set of labeled key observations in the target time t_{q+1}. We denote with K_{q+1} the set of labeled target keys in t_{q+1} and T_{q+1} the set of unlabeled non-key observations.

In our proposal, a new dataset, denoted as K' is computed. It contains one tuple for each key observation in K_{q+1}. Attributes of K' represent the responses predicted by $f_{q-w}, f_{q-w+1}, \ldots$, and f_q for each key observation in K, the spatial dimension coordinates U and V and the true response value collected for the key observation. K' is now employed as training data set for a new regression task $\tau'(Y, Y_{q-w}, Y_{q-w+1}, \ldots, Y_q, U, V)$, which is formulated with the aim of learning the the target predictive function f_{q+1}.

We investigate two alternative solutions to learn f_{q+1} by a transfer learning process. The former solution employs the classical stepwise least squared regression (LSR) method [11] which uses accuracy to determine the prominent variables (responses of the past trees) for the transfer and outputs a global multivariate linear combination of these variables. The latter solution adapts the idea of a piece-wise (tree-based) form for the multivariate regression model and bases the partitioning on Boolean splits on the spatial attributes. Also in this case, the stepwise least squared regression method is used to learn the multivariate function associated to each leaf. An example of piece-wise multivariate regression model (model tree) is reported in Figure 2. Again, leaves are associated with multivariate regression models learned stepwise.

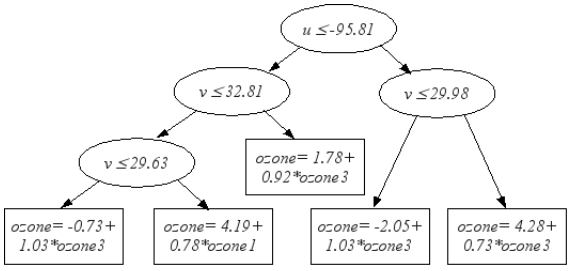


Fig. 2. An example of model tree used during transfer learning. *ozone*1 and *ozone*3 are predictor variables whose values are estimated according to f_1 and f_3, respectively.

[1] w is the size of a backward time window with $w \geq 1$.

6 Experiments

The inductive learner GWRT and its transfer learner, called GWRTT, are implemented in a Java system which interfaces a MySQL DBMS. In the next subsections, we illustrate results obtained with benchmark spatial data sets and a real spatio-temporal data collection.

6.1 Geographically Weighted Regression Tree Induction

GWRT is evaluated on real data collections to seek answers to the following questions. (1) How does the spatial segmentation of the landscape in rectangular areal units of positively autocorrelated responses improve both the aspatial segmentation performed by the state-of-art model tree learner M5' and the co-training solution implemented by the transductive learner SpReCo? (2) How does the stepwise construction of a piecewise space-varying parametric linear function solve the collinearity problem and improve the accuracy of traditional GWR? (3) How does the boundary bandwidth h and the neighborhood size k affect accuracy of geographically weighted regression trees induced by GWRT? In the following, we describe the data sets, the experimental setting and we illustrate results obtained with these data in order to answer questions 1-3.

Datasets. GWRT has been evaluated on six spatial regression data collections whose description is reported in the following. *Forest Fires* (FF) [8] collects 512 observations of forest fires in the period January 2000 to December 2003 in the Montesinho Natural Park, Portugal. The predictor variables are: the Fine Fuel Moisture Code, the Duff Moisture Code, the Drought Code, the Initial Spread Index, the temperature in Celsius degrees, the relative humidity, the wind speed in km/h, and the outside rain in mm/m^2. The response variable is the burned area of the forest in ha (with 1ha/100 = 100 m^2). The spatial coordinates (U,V) refer to the centroid of the area under investigation on a park map. *USA Geographical Analysis Spatial Data* (GASD) [24] contains 3,107 observations on USA county votes cast in the 1980 presidential election. For each county the explanatory attributes are: the population of 18 years of age or older, the population with a 12th grade or higher education, the number of owner-occupied housing units, and the aggregate income. The response attribute is the total number of votes cast. For each county, the spatial coordinates (U,V) of its centroid are available. *North-West England* (NWE)(http://www.ais.fraunhofer.de/KD/SPIN/project.html) concerns the region of North West England, which is decomposed into 1011 censual wards. Both predictor and response variables available at ward level are taken from the 1998 Census. They are the percentage of mortality (response attribute) and measures of deprivation level in the ward, according to index scores such as, Jarman Underprivileged Area Score, Townsend Score, Carstairs Score and the Department of the Environment Index. Spatial coordinates (U,V) refer to the ward centroid. By removing observations including null values, only 979 observations are used in this experiment. *Sigmea-Real* [10] collects 817 observations of the

rate of herbicide resistance of two lines of plants (predictor variables), that is, the transgenic male-fertile (SMF) and the non-transgenic male-sterile (SMS) line of oilseed rape. Predictor variables are the cardinal direction and distance from the center of the donor field, the visual angle between the sampling plot and the donor field, and the shortest distance between the plot and the nearest edge of the donor field. Spatial coordinates (U,V) of the plant are available. *South California* (SC) [17] contains 8033 observations collected for the response variable, median house value, and the predictor variables, median income, housing median age, total rooms, total bedrooms, population, households in South California. Spatial coordinates represent the latitude and longitude of each observation.

Experimental Setting. The performed experiments aim at evaluating the effectiveness of the improvement of accuracy of the geographically weighted regression tree, with respect to the baseline model tree learned with the state of art model tree learner M5', the transductive learner SpReCo and the geographically weighted regression function computed by GWR. We used M5' as it is the state-of-art model tree learner which is considered as the baseline in almost all papers on model tree learning. At the best of our knowledge, no study reveals the existence of a model tree learner which definitely outperforms M5'. The implementation of M5' is publicly available in WEKA, while the implementation of GWR is publicly available in software R. M5' is run in two settings. The first setting adds the spatial dimensions to the set of predictor variables (sM5), the second setting filters out variables representing spatial dimensions (aM5). The empirical comparison between systems is based on the mean square error (MSE). To estimate the MSE, a 10-fold cross validation is performed and the average MSE (Avg.MSE) over the 10-folds is computed for each dataset. To test the significance of the difference in accuracy, we use the non-parametric Wilcoxon two-sample paired signed rank test.

Results. In Table 1, we compare the 10-fold average MSE of GWRT with the MSE of M5' (both sM5 and aM5 settings) and GWR . GWRT is run by varying h and k as we intend to draw empirical conclusions on the optimal tuning of these parameters. The results show that MSE comparison confirms that GWRT outperforms aspatial and spatial competitors, generally by a great margin. This result empirically proves the intuitions which lead us to synthesize a technique for the induction of geographically weighted regression trees. The spatial-based tree segmentation of the landscape aimed at the identification of rectangular areal units with positively autocorrelated responses improves the performance gained by the baseline M5' which partitions data (and not landscape) according to a Boolean test on the predictor variables. On the other hand, the spatial segmentation of landscape combined with linear models having a local estimate of parameters allows us to gain a more efficacious consideration (in terms of accuracy) of the spatial autocorrelation than SpReCo (which accounts the autocorrelation by a co-training procedure). The only exceptions are NWE and SMS. On the other hand, the stepwise computation of a geographically weighted regression model at each leaf is able to select the appropriate subset of

Table 1. 10-fold CV average MSE: GWRT vs M5', SpReCo and GWR. GWRT is run by varying both neighborhood size k and bandwidth h. M5' is run either by including the spatial dimensions (sM5) in the set of predictor variables or by filtering them out (aM5). GWR is run with the option for automatic bandwidth estimation. The best value of accuracy is in boldface for each dataset.

k	5	5	5	5	10	10	10	10	sM5	aM5	SpReCo	GWR
h	20%	30%	40%	50%	20%	30%	40%	50%				
FF	50.44	49.73	49.84	49.99	50.37	**49.64**	49.76	49.90	87.63	76.88	58.24	373.3
GASD	0.10	**0.09**	0.10	0.10	0.10	**0.09**	0.10	0.10	0.14	0.14	0.14	0.35
NWE	0.009	0.004	0.003	0.003	0.009	0.003	0.003	0.003	0.004	0.004	**0.0025**	0.004
SMS	11.11	5.82	4.19	4.58	18.33	5.48	4.42	4.56	3.98	4.73	**3.51**	5.22
SMF	3.66	2.24	**1.81**	1.92	3.67	2.37	1.90	1.92	2.40	1.98	1.91	1.98
SC	32.1e5	7.1e4	**5.3e4**	5.4e4	17.4e5	6.9e4	**5.3e4**	5.4e4	6.1e4	8.7e4	6.6e4	8.2e4

predictor variables at each leaf, thus solving the collinearity and definitely improving the baseline accuracy of traditional GWR. The statistical significance of the obtained differences is estimated in terms of the signed rank Wilcoxon test. The entries of Table 2 report the statistical significance of the differences between compared systems estimated with the signed rank Wilcoxon test. By insighting the statistical test results, we observe that there are three datasets, GASD, Forest Fires and South California, where GWRT statistically outperforms each competitor independently from the h and k setting (just with South California and h=20% the superiority of GWRT with respect to its competitor is not statistically obseravabe). The same primacy of GWRT is observable for the remaining three datasets, NWE, SigmeaMS and SigmeaMF, when we select higher values of h ($h >= 30\%$). In general, we observe that a choice of h between 30% and 40% leads to lower MSE in all datasets. GWRT seems to be less sensitive to the choice of k due to the weighting mechanism.

6.2 Geographically Weighted Regression Tree Transfer across Time

GWRTT is evaluated on a real spatio-temporal data collection in order to seek answers to the following questions. (1) How does the prediction function learned with the transfer technique vary in accuracy by tuning the percentage of key observations into the target domain and/or the time window size used to select source geographically weighted regression trees to be transfered across time? (2) When is the transfer learner better than the traditional inductive learner? In the following, we describe the data set, the experimental setting and we illustrate results obtained with these data in order to seek questions 1-2.

Dataset. We run experiments by considering data hourly collected by the Texas Commission On Environment Quality in the time period May 5-15, 2009. Data are obtained from 26 stations installed in Texas (http://www.tceq.state.tx.us/). Predictor variables are wind speed, temperature and solar radiation. The response variable is the ozone rate. Spatial dimensions are the latitude and longitude of the transmitting stations.

Table 2. The signed Wilcoxon test on the accuracy of systems: GWRT vs M5 (sM5 or aM5); GWRT vs SpReCo; GWRT vs GWR. The symbol "+" ("-") means that GWRT performs better (worse) than the competitor system. "+" ("–") denotes the statistically significant values ($p \leq 0.05$).

k																			
h	GWRT vs	5 20	5 30	5 40	5 50	10 20	10 30	10 40	10 50	GWRT vs	5 20	5 30	5 40	5 50	10 20	10 30	10 40	10 50
FF	sM5	+	+	+	+	+	+	+	+	aM5	+	+	+	+	+	+	+	+
	SpReCo	+	+	+	+	+	+	+	+	GWR	+	+	+	+	+	+	+	+
GASD	sM5	+	+	+	+	+	+	+	+	aM5	+	+	+	+	+	+	+	+
	SpReCo	+	+	+	+	+	+	+	+	GWR	+	+	+	+	+	+	+	+
NWE	sM5	–	+	+	+	–	+	+	+	aM5	–	+	+	+	–	+	+	+
	SpReCo	–	–	–	–	–	–	–	–	GWR	–	–	-	-	–	–	-	-
SMS	sM5	–	–	+	+	–	–	+	+	aM5	–	-	+	+	–	-	+	+
	SpReCo	–	-	-	-	–	-	-	-	GWR	–	+	+	+	–	+	+	+
SMF	sM5	–	+	+	+	+	=	+	+	aM5	–	-	+	+	–	-	+	+
	SpReCo	–	-	+	+	–	-	+	+	GWR	–	+	+	+	-	+	+	+
SC	sM5	-	+	+	+	-	+	+	+	aM5	=	+	+	+	=	+	+	+
	SpReCo	–	-	+	+	–	-	+	+	GWR	=	+	+	+	=	+	+	+

Experimental Settings. We define one transfer task for each day. The target domain for the transfer is the set of observations transmitted from the 26 stations at 24:00hrs of the corresponding day. The sources of the transfer are the Geographically Weighted Regression Trees learned on each source domain which is hourly collected in the corresponding day. We consider backward windows with size 1, 3, 6, 12, 18 and 24. We perform experiments by sampling 25%, 50% and 75% of the stations as key stations for the transfer.

Results. The MSE measured over the non-key observations of each target domain is averaged for the eleven days and it is illustrated in Table 3. The results of transfer are collected by varying the bandwidth h and the percentage of keys κ in the target domain. The number of neighbors k is set to 5 for all experiments. The target predictive function is induced by using either a model tree learner with Boolean test on spatial dimensions or the least square regression learner. The results of the transfer are compared with the baseline Geographically Weighted Regression Tree, induced on the key observations of the target domain only. These results suggest that, in this specific domain, LSR performs better than a model tree to approximate the target predictive function. Probably, this depends on the scarcity of key data. Additionally, we observe that the gain in accuracy due to the transfer is almost always appreciable when $\kappa = 50\%$. Concerning the bandwidth h and the window size w, we are not able to univocally identify the best setting for these parameters. This proves that further investigations are necessary in the direction of automatical parameter tuning.

Table 3. Avg MSE: eleven transfer tasks are daily defined for the Texas Ozone data collected at the 24:00 on May 5-15, 2009. For each task, sources domains are hourly collected from 0:00 to 23:00. In bold, the MSE where the transfer outperforms GWRT.

		Model Tree				LSR			
κ	w/h	20	30	40	50	20	30	40	50
25%	1	**7.58**	10.34	10.62	10.66	**4.64**	8.37	**6.17**	**6.54**
25%	3	11.78	12.31	32.60	15.84	**5.73**	**5.94**	12.87	7.91
25%	6	15.37	21.01	64.26	10.52	**6.30**	8.99	39.03	12.88
25%	12	14.55	12.38	50.30	10.79	**7.94**	**7.29**	11.55	22.46
25%	18	12.52	25.12	161.50	24.46	**7.79**	10.25	22.33	10.74
25%	24	16.33	11.29	129.89	24.37	13.77	11.34	27.30	16.73
25%	GWRT	8.03	8.08	7.85	7.59	8.03	8.08	7.85	7.59
50%	1	9.15	**7.31**	**7.31**	**6.27**	**5.71**	**5.93**	**5.97**	**6.59**
50%	3	10.99	**8.00**	10.09	**10.40**	**5.49**	**6.51**	**6.00**	**8.12**
50%	6	12.44	**10.76**	10.12	12.26	**6.31**	**7.03**	**6.36**	**8.51**
50%	12	11.78	**10.35**	10.60	12.73	8.76	**7.07**	**7.52**	**8.29**
50%	18	13.53	**11.27**	13.23	14.31	**7.22**	**7.89**	**7.94**	16.24
50%	24	17.28	**16.18**	12.96	12.94	**7.73**	**8.66**	**8.33**	**10.96**
50%	GWRT	8.33	20.24	8.67	11.10	8.33	20.24	8.67	11.10
75%	1	8.31	6.58	8.37	**6.86**	6.28	7.48	6.42	**7.28**
75%	3	10.43	7.92	11.45	15.58	**5.88**	8.31	**6.21**	**6.99**
75%	6	10.74	11.74	15.78	11.69	**5.37**	7.50	**6.09**	**7.83**
75%	12	8.69	11.67	10.57	16.29	**5.79**	7.57	**6.31**	**8.03**
75%	18	6.19	9.79	12.06	11.17	**5.19**	7.22	6.56	**7.20**
75%	24	6.67	15.90	12.15	10.92	**5.36**	8.09	6.53	**7.31**
75%	GWRT	6.08	6.08	6.42	11.45	6.08	6.08	6.42	11.45

7 Conclusions

We present a novel spatial regression technique to tackle issues posed by spatial non-stationarity and positive spatial autocorrelation in geographically distributed data environments. To deal with spatial non-stationarity we decide to learn piecewise space-varying parametric linear regression functions with coefficients which are estimated to vary across the space. We combine local model learning with a tree structured segmentation approach that recovers the functional form of a spatial model only at the level of each areal segment of the landscape. A new stepwise technique is adopted to select the most promising predictors to be included in the model, while parameters are estimated at every point across the local area. The parameter estimation solves the problem of least square weighted regression and uses a positively autocorrelated neighborhood to determine a correct estimate of the weights. Finally, a transfer learning technique is defined to transfer spatial regression models learned in the past to the present time point. Experiments with several benchmark data collections confirm that the induction of our geographically weighted regression trees generally improves both the local estimation of parameters performed by GWR and the global

estimation of parameters performed by classical model tree learners like M5' as well as transductive solution for spatial regression problems as SpReCo. Furthermore, the transfer is proved to be effective in a real application. As future work, we plan to investigate techniques for automating tuning the bandwidth, the neighborhood size and the transfer window size. Additionally, we plan to use the defined transfer technique to frame this work in a streaming environment, where geographically distributed sensors continuously transmit observations across time.

Acknowledgment. This work fulfills the research objectives of both the project: "EMP3: Efficiency Monitoring of Photovoltaic Power Plants" funded by "Fondazione Cassa di Risparmio di Puglia," and the PRIN 2009 Project "Learning Techniques in Relational Domains and their Applications" funded by the Italian Ministry of University and Research (MIUR). Authors thank Lynn Rudd for her help in reading the manuscript.

References

1. Petrucci, C.S.A., Salvati, N.: The application of a spatial regression model to the analysis and mapping of poverty. Environmental and Natural Resources Series 7, 1–54 (2003)
2. Fotheringham, M.C.A.S., Brunsdon, C.: Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley (2002)
3. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6), 716–723 (1974)
4. Appice, A., Ceci, M., Malerba, D.: Transductive learning for spatial regression with co-training. In: Shin, S.Y., Ossowski, S., Schumacher, M., Palakal, M.J., Hung, C.-C. (eds.) Proceedings of the 2010 ACM Symposium on Applied Computing (SAC 2010), pp. 1065–1070. ACM (2010)
5. Bogorny, V., Valiati, J.F., da Silva Camargo, S., Engel, P.M., Kuijpers, B., Alvares, L.O.: Mining maximal generalized frequent geographic patterns with knowledge constraints. In: ICDM 2006, pp. 813–817. IEEE Computer Society (2006)
6. Brunsdon, C., McClatchey, J., Unwin, D.: Spatial variations in the average rainfall-altitude relationships in great britain: an approach using geographically weighted regression. International Journal of Climatology 21, 455–466 (2001)
7. Burnham, K., Anderson, D.: Model selection and multimodel inference: a practical information-theoretic approach. Springer (2002)
8. Cortez, P., Morais, A.: A data mining approach to predict forest fires using meteorological data. In: EPIA 2007, pp. 512–523. APPIA (2007)
9. Daumé III, H., Marcu, D.: Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research 26, 101–126 (2006)
10. Demšar, D., Debeljak, M., Lavigne, C., Džeroski, S.: Modelling pollen dispersal of genetically modified oilseed rape within the field. In: Annual Meeting of the Ecological Society of America, p. 152 (2005)
11. Draper, N.R., Smith, H.: Applied regression analysis. Wiley (1982)
12. Dries, A., Rückert, U.: Adaptive concept drift detection. Statistical Analysis and Data Mining 2(5-6), 311–327 (2009)

13. Góra, G., Wojna, A.: RIONA: A Classifier Combining Rule Induction and k-NN Method with Automated Selection of Optimal Neighbourhood. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) ECML 2002. LNCS (LNAI), vol. 2430, pp. 111–123. Springer, Heidelberg (2002)
14. Hordijk, L.: Spatial correlation in the disturbances of a linear interregional model. Regional and Urban Economics 4, 117–140 (1974)
15. Huang, Y., Leung, Y.: Analysing regional industrialisation in jiangsu province using geographically weighted regression. Journal of Geographical Systems 4, 233–249 (2002)
16. Hurvich, C.M., Tsai, C.-L.: Regression and time series model selection in small samples. Biometrika 76(2), 297–307 (1989)
17. Kelley, P., Barry, R.: Sparse spatial autoregressions. Statistics and Probability Letters 33, 291–297 (1997)
18. Legendre, P.: Spatial autocorrelation: Trouble or new paradigm? Ecology 74, 1659–1673 (1993)
19. LeSage, J., Pace, K.: Spatial dependence in data mining. In: Data Mining for Scientific and Engineering Applications, pp. 439–460. Kluwer Academic (2001)
20. Levers, C., Brückner, M., Lakes, T.: Social segregation in urban areas: an exploratory data analysis using geographically weighted regression analysis. In: 13th AGILE International Conference on Geographic Information Science 2010 (2010)
21. Longley, P., Tobon, A.: Spatial dependence and heterogeneity in patterns of hardship: an intra-urban analysis. Annals of the Association of American Geographers 94, 503–519 (2004)
22. Malerba, D., Ceci, M., Appice, A.: Mining Model Trees from Spatial Data. In: Jorge, A.M., Torgo, L., Brazdil, P.B., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS (LNAI), vol. 3721, pp. 169–180. Springer, Heidelberg (2005)
23. Mitchell, T.: Machine Learning. McGraw Hill (1997)
24. Pace, P., Barry, R.: Quick computation of regression with a spatially autoregressive dependent variable. Geographical Analysis 29(3), 232–247 (1997)
25. Pan, S.J., Shen, D., Yang, Q., Kwok, J.T.: Transferring localization models across space. In: AAAI, pp. 1383–1388. AAAI Press (2008)
26. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22(10), 1345–1359 (2010)
27. Rinzivillo, S., Turini, F., Bogorny, V., Körner, C., Kuijpers, B., May, M.: Knowledge discovery from geographical data. In: Mobility, Data Mining and Privacy, pp. 243–265. Springer (2008)
28. Shariff, N., Gairola, S., Talib, A.: Modelling urban land use change using geographically weighted regression and the implications for sustainable environmental planning. In: Proceeding of the 5th International Congress on Environmental Modelling and Software Modelling for Environment's Sake, iEMSs (2010)
29. Shekhar, S., Chawla, S.: Spatial databases: A tour. Prentice Hall (2003)
30. Wang, Y., Witten, I.: Inducing Model Trees for Continuous Classes. In: van Someren, M., Widmer, G. (eds.) ECML 1997. LNCS, vol. 1224, pp. 128–137. Springer, Heidelberg (1997)
31. Zheng, V.W., Xiang, E.W., Yang, Q., Shen, D.: Transferring localization models over time. In: AAAI, pp. 1421–1426. AAAI Press (2008)