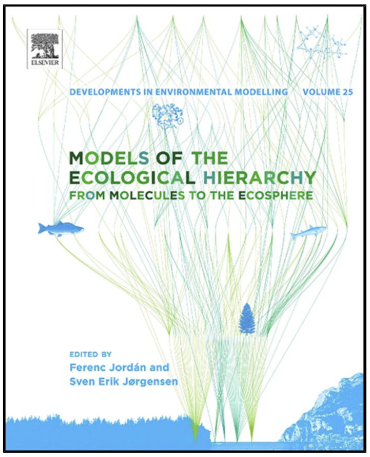

From Stojanova, D., Debeljak, M., Ceci, M., Appice, A., Malerba, D., Džeroski, S., 2012. Dealing with Spatial Autocorrelation in Gene Flow Modeling. In: Jordán, F., Jørgensen, S.E. (Eds), Models of the Ecological Hierarchy: From Molecules to the Ecosphere. Elsevier B.V., pp. 35–49.
ISBN: 9780444593962

3 Dealing with Spatial Autocorrelation in Gene Flow Modeling

Daniela Stojanova*, Marko Debeljak*, Michelangelo Ceci†, Annalisa Appice†, Donato Malerba†, Sašo Džeroski*,‡

*DEPARTMENT OF KNOWLEDGE TECHNOLOGIES, JOŽEF STEFAN INSTITUTE, JAMOVA CESTA, LJUBLIJANA, SLOVENIA; JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL, JAMOVA CESTA, LJUBLJANA, SLOVENIA, †DIPARTIMENTO DI INFORMATICA, UNIVERSITÀ DEGLI STUDI DI BARI "ALDO MORO," VIA ORABONA 4, BARI, ITALY, ‡CENTRE OF EXCELLENCE FOR INTEGRATED APPROACHES IN CHEMISTRY AND BIOLOGY OF PROTEINS

3.1 Introduction

Spatial data can identify the geographic locations of different kinds of ecological and environmental data. The main characteristic of spatial data is that they are geo-referenced, that is, they are usually presented with coordinates and a topology which can be visualized on a map. Spatial data are common in ecological studies. For instance, studies on population dynamics examine changes of the size and the structure of populations over space and time. Moreover, several studies on habitat modeling look for the suitability of different ecosystem environments (e.g., aquatic, arable, and forest ecosystems) for various organisms (flora and fauna) (Debeljak and Džeroski, 2011).

The main problem with spatial data is that measured at locations relatively close to each other tend to have more similar values than data measured at locations further apart (Tobler's first law of geography (Tobler, 1970)). For example, species richness at a given site is likely to be similar to that of a site nearby, but very different from that of sites far away. This is mainly due to the fact that the environment is more uniform within a shorter radius. This phenomenon is referred to as spatial autocorrelation. More specifically, the *positive* spatial autocorrelation refers to a map pattern according to which similar values of a geographic feature tend to cluster closely on the map, whereas *negative* spatial autocorrelation indicates a map pattern according to which similar values scatter throughout the map. When no statistically significant spatial autocorrelation exists, the pattern of spatial distribution is considered to be random. The inappropriate treatment of spatial autocorrelation could obfuscate important insights and the observed patterns may even be inverted when spatial autocorrelation is ignored (Kühn, 2007).

The motivation to take spatial autocorrelation into account in ecological and environmental data is manifold, but the following four factors are of particular importance (Legendre, 1993; Legendre et al., 2002):

Models of the Ecological Hierarchy. DOI: http://dx.doi.org/10.1016/B978-0-444-59396-2.00003-1
ISSN 0167-8892,

1. Biological processes of speciation, extinction, dispersal, or species interactions are typically distance-related.
2. Nonlinear relationships may exist between species and environments, but these relationships may be incorrectly modeled as linear.
3. Classical statistical modeling may fail in the identification of the relationships between different kinds of data without taking into account their spatial arrangement (Besag, 1974).
4. The spatial resolution of data should be taken into account: Coarser grains lead to spatial smoothing of data.

Unfortunately, classical statistical models are based on the assumption that the values of observations in each sample are independent of each other. This is in contrast with spatial autocorrelation, which clearly indicates a violation of this assumption. As observed by LeSage and Pace (LeSage and Pace, 2001), "anyone seriously interested in mining sample data which exhibits spatial dependence should consider a spatial model," since this can take into account different forms of spatial autocorrelation. They showed how the inclusion of autocorrelation of the dependent variable in a predictive task provides an improvement in fitting, as well as a dramatic difference in the significance and impact of explanatory variables included in the predictive model. Even if we are not interested in analyzing the spatial structure within a data sample (i.e., the relationship between the response variable and spatial variables such as longitude and latitude), we should still care about spatial autocorrelation, as it can result in severe problems, that is, can stretch the statistics beyond the basic assumptions and, worse, still overlook important factors affecting the functioning of the ecosystems (Legendre, 1993; Legendre et al., 2002; Betts et al., 2009).

Applications of autocovariate regression, spatial eigenvector mapping, generalized least squares, conditional/simultaneous autoregressive models, and generalized estimating equations to spatial regression tasks are presented by Dormann et al. (Dormann, 2007). All these methods account for spatial autocorrelation in the analysis of species distribution data by considering both species presence/absence (binary response) and species abundance data (Poisson or normally distributed response). The suggested way to detect spatial autocorrelation effects in regression is by measuring spatial autocorrelation of the residuals of the linear model.

Other methods include spatial filtering, kriging, and geographically weighted regression (GWR). **Spatial filtering** actually transforms a variable exhibiting spatial autocorrelation into a variable that is free of the spatial dependence. The transformation is obtained by building two new synthetic variables from the original georeferenced variable. The former is a spatial filter variable capturing the latent spatial dependency that otherwise would remain in the response residuals. The latter is a nonspatial variable that is free of spatial dependence (Griffith, 2002). **Kriging** (Cressie, 1991) is a spatial statistics technique which exploits spatial autocorrelation and determines a local model of the spatial phenomenon. It applies an optimal

linear interpolation method to estimate the response variable at each location across the landscape. The response variable is decomposed into a structural component, which represents a mean or a constant trend, a random (spatially correlated) component, and a random noise component, which expresses measurement errors or variations inherent in the response variable. Another standard way to take into account spatial autocorrelation in spatial statistics is the **geographically weighted regression (GWR)** (Fotheringham et al., 2002). GWR extends the traditional multiple linear regression framework so that all parameters are estimated within a local context. In this way, GWR takes advantage of positive autocorrelation between neighboring sites in space and provides valuable information on the nature of the processes being investigated.

In this chapter, we focus on the issues raised by spatial autocorrelation when we are interested in learning a predictive spatial regression model from a sample of geo-referenced data. In the spatial regression setting, the sample data are distributed on some domain $\boldsymbol{X}$, are geo-referenced in the space $U \times V$ (e.g., latitude $\times$ longitude), and labeled according to an unknown function g with range Y. The domain $\boldsymbol{X}$ is spanned by m independent (or explanatory) random variables X_i (both continuous and categorical), while Y is a subset of $\Re$, that is, the dependent (or response) variable y is continuous. A learning algorithm receives a training sample $E = \{(u, v, \boldsymbol{x}, y) \in | U \times V \times \boldsymbol{X} \times Y | \; y = g(u, v, \boldsymbol{x})\}$ and returns a function f close to g on the domain $\boldsymbol{X}$, where closeness is often measured by means of the expected square error over a testing sample or a part of E withheld for learning.

This chapter focuses on handling spatial autocorrelation when predicting gene flow from genetically modified (GM) to non-GM maize fields under real multifield crop management practices at a local scale. The gene flow measured at neighboring locations tends to have more similar values than the ones measured at locations further apart. Because this is a distance-related problem and nonlinear relationships may exist between the factors affecting gene flow environments, classical statistical modeling may fail in the identification of the true relationships between the different kinds of data if it does not take into account their spatial arrangement (Besag, 1974). Therefore, in this chapter, we present a different approach to spatial regression which accounts for spatial autocorrelation in tasks such as the one of predicting gene flow.

The proposed method, spatial predictive clustering system (SCLUS) (Stojanova et al., 2011), is an extension of the system CLUS (Blockeel et al., 1998) that builds Predictive Clustering Trees (PCTs). The SCLUS method can be used to build spatially aware regression trees by considering both local and global, spatial autocorrelation and can deal with the "ecological fallacy" problem (Robinson, 1950) according to which individual subregions do not have the same data distribution of the entire region. We also prove the usefulness of the proposed method in predicting the gene flow (outcrossing rate) from GM to non-GM maize fields under real multifield crop management practices at a local scale.

3.2 Modeling Method: Spatially Aware Predicting Clustering Trees

Decision tree learning is among the most popular machine learning techniques used for ecological modeling. Decision trees can be used to predict the value of one or several target (dependent) variables. They are hierarchical structures, where each internal node contains a test on an attribute, each branch corresponds to an outcome of the test, and each leaf node gives a prediction for the value of the class variable(s). Depending on whether we are dealing with a classification (discrete target) or a regression problem (continuous target), the decision tree is called a classification or a regression tree, respectively.

Predictive Clustering Trees (PCTs) (Blockeel et al., 1998) are decision trees that combine elements from both prediction and clustering. As in clustering, clusters of observations that are similar to each other are identified, but a predictive model is associated to each cluster. The predictive model can be viewed as a hierarchy of clusters complemented by a symbolic description of the clusters—the top-node corresponds to one cluster containing all data, which is recursively partitioned into smaller clusters while moving down the tree. The construction of PCTs is not very different from that of standard decision tree learning—at each internal node, a test has to be selected according to a given evaluation function. The main difference between standard decision trees and PCTs is that PCTs select the best test by maximizing the (intercluster) variance reduction with respect to an arbitrary variance function Var (.), defined as $\Delta(E,P) = Var(E) - \sum_{E_k \in P} \frac{|E_k|}{|E|} Var(E_k)$, where E represents the observations in the node and P defines the partition $\{E_1, E_2\}$ of E. When dealing with single outcrossing target Y, Var (.) denotes the standard variance function known from statistics $Var(E) = Var(Y)$. When dealing with several outcrossing targets, $Y_1, \ldots Y_t$, $Var(E) = \sum_{i=1}^{T} (Y_i)$. The PCT framework allows different definitions of appropriate variance functions for different types of data and can thus handle complex structured data as targets. PCTs can work with complex structured data by introducing adequate variance measures, PCTs are a good candidate to appropriately deal with spatial data where the complexity comes from the need of taking autocorrelation into account. Following this idea, SCLUS extends CLUS in order to build spatially aware PCTs. In fact, the tree structure of the obtained models permits us to naturally deal with the ecological fallacy: the entire region is hierarchically partitioned into subregions such that the data distribution in each subregion is as uniform as possible across the landscape. Additionally, the tree structure permits us to model the presence of autocorrelation in the sample data at different levels of the tree: global modeling is possible at the top-levels of the tree, while local modeling is possible at the bottom levels of the tree. This ability of supporting a hierarchical modeling of spatial autocorrelation is the main difference of SCLUS with respect to the classical spatial statistical approaches (GWR and Kriging), which can only

deal with autocorrelation locally and result in many different local models that make separate predictions at each site.

To consider autocorrelation, SCLUS groups examples that show high-positive autocorrelation. This means that examples in each cluster not only have similar response values, but also they fall in the same spatial neighborhood (very close to each other in space). The SCLUS method uses this rationale while building the spatially aware predictive models.

At present, SCLUS does not learn models which perform spatial splits (using the coordinates as attributes). The motivation behind this choice is that of avoiding to lose generality in the induced models. In this way, we can learn more general models that can be easily applied in the same domain, but in different spatial contexts (Ester et al., 1997). Therefore, the spatial arrangement is only used in the split evaluation. In particular, SCLUS uses the level of spatial autocorrelation among examples in the dataset. Spatial autocorrelation is measured by using standard spatial statistics, such as Global Moran's I and Global Geary's C. Equation (1) defines the Global Moran's I as

$$I = \frac{N \sum_i \sum_j w_{ij}(Y_i - \overline{Y})(Y_j - \overline{Y})}{\sum_i \sum_j w_{ij} \times \sum_j (Y_j - \overline{Y})^2} \tag{1}$$

where N is the number of geo-referenced examples indexed by i and j; Y_i and Y_j are the values of the variable Y for the examples o_i and o_j, respectively; Y is the variable of interest (response variable in SCLUS); $\overline{Y}$ is the overall mean of Y; and $W = [w_{ij}]_{i=1,\ldots,N \;\; j=1,\ldots,N} \sum_{ij} w_{ij}$ is the matrix of spatial weights. Values of I generally range from -1 (negative autocorrelation) to $+1$ (positive autocorrelation) and 0 indicates a random distribution of the data.

Equation (2) defines the Global Geary's C as

$$C = \frac{(N-1) \sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{2 \sum_i \sum_j w_{ij} \times \sum_j (Y_j - \overline{Y})^2} \tag{2}$$

Values of C typically range from 0 (positive autocorrelation) to 2 (negative autocorrelation) and 1 indicates a random distribution of the data.

Both indexes (Moran's I and Geary's C) use a spatial weight matrix $W = [w_{ij}]_{i=1,\ldots,N \;\; j=1,\ldots,N}$ that reflects the intensity of the spatial relationship between observations in a neighborhood. Each weight w_{ij} is inversely proportional to the distance between the examples o_i and o_j. More specifically, if the examples are far away from each other, that is, the distance between o_i and o_j is larger than a predefined threshold *b (bandwidth)*, w_{ij} tends to zero. In addition, the elements are normalized so that for each i, $\sum_j w_{ij} = 1$. In this work, the bandwidth b is represented as a percentage of the maximum spatial distance between two examples in the dataset.

While both statistics reflect the spatial dependence of values, they do not provide identical information. C emphasizes the differences in values between pairs of observations, while I emphasizes the covariance between the pairs. This means that Moran's I values are smoother, whereas Geary's C is more sensitive to differences in small neighborhoods.

In order to evaluate the possible splits and find the best split for each internal node of the tree structure, one of these spatial statistics is linearly combined with the variance reduction. A user-defined parameter α regulates the relative influence of both parts of the split evaluation heuristic as reported in Eqn. (3)

$$h(E,P) = \alpha \times \Delta(E,P) + (1-\alpha) \times S(E,P) \tag{3}$$

where $\alpha \in [0,1]$; $\Delta(E,P)$ is the variance reduction associated to the split P, while $S(E,P) = \sum_{E_k \in P} \frac{|E_k|}{|E|} S(E_k)$ is the autocorrelation measurement (with $S(E_k)$ computed as either I or C) associated to the split P. Both $\Delta(E,P)$ and $S(E, P)$ are scaled to the same interval [0, 2] in order to be combined.

Using the split evaluation heuristic described above, the top-down induction of spatially aware PCTs proceeds recursively by taking as input a set of training examples E and partitioning the descriptive space until the stopping criterion is satisfied. In the construction of the tree, at each internal node, the split that maximizes $h(E,P)$ is chosen. As stated before, splits are derived only from the nonspatial variables. Possible tests are of the form $X \leq \beta$ for continuous variables, and

$X \in \{x_{i1}, x_{i2}, \ldots, x_{ie}\}$ (where $\{x_{i1}, x_{i2}, \ldots, x_{ie}\}$ is a subset of the domain of X) for discrete variables. Finally, as in standard decision trees, the tree construction stops when the number of examples in a leaf is less than $\sqrt{N}$, where N is the number of examples in the whole datasets. This value is considered to be a good locality threshold that does not permit to lose too much in accuracy (Gora and Wojna, 2002). As in the original CLUS method, when a leaf is constructed, the prediction associated to it is the mean of the response values of training examples clustered in the leaf.

The method, as described above, is implemented in a system that offers the opportunity to the user to define the parameters and explore their influence on the final results. The user can choose which of the spatial statistics (Global Moran's I or Global Geary's C) to include in the model and the relative influence α. The user can also define the neighborhood (bandwidth b) in which autocorrelation is considered and set the level of spatial autocorrelation to be considered in split selection when building the model.

3.3 Results and Discussion: Modeling Gene Flow from GM to Non-GM Fields

In this section, we describe the use of SCLUS to model gene flow from GM to non-GM fields under real muftifiled crop management practices at a local scale. We first provide a description of the data and then give the SCLUS parameter settings and the evaluation

measures used for the evaluation of the effectiveness of the predictions provided by SCLUS. Then, we compare the performance of the SCLUS method to competitive regression and spatial modeling methods. Finally, we present the map of the predicted outcrossing rates constructed by using the obtained models, as a final visual product of this analysis.

3.3.1 Data Description

The area under study encompasses 400 ha of the Foixa region in Spain, a region with intensive production of GM and non-GM maize. The data are provided for a period of three successive years, from 2004 to 2006. A detailed description of the field setup, data used, and methods of sampling is given by Messeguer et al. (2006). The definition of the variables used in the modeling process is based on expert knowledge and previous studies (Messeguer et al., 2006; Debeljak et al., 2011) about the gene flow modeling problem.

The explanatory variables are defined by taking into account the multifield effects from the neighboring fields at a local scale. This includes the consideration of a *secureDistance*[1], that is, a distance threshold above which there is no (or negligibly small) influence from the surrounding fields and a *floweringDelay*, which expresses the difference in number of days between the flowering of the GM and the non-GM field. We use prior knowledge of previous studies to set these thresholds. The threshold for the flowering delay between the GM and non-GM fields is set to 10 days. This means that if there are a non-GM and a GM field in the same neighborhood and one of them flowers ten or more days later than the other field, the non-GM field cannot be contaminated by the GM field. The secure distance around each sampling point is set to 150 m. According to domain expertise, GM fields or parts of fields that are further away than 150 m from a conventional field cannot contaminate it significantly, even if they have an overlap in the flowering periods.

Both variables were identified by a domain expert as crucial for the analysis and further on serve as the basis for the definition of the other independent variables:

- Average (minimum) distance and average (minimum) "weighted edge" from a sampling point to surrounding GM fields
- Number of GM fields surrounding a sampling point within the *secureDistance* of 150 m considering only fields that have *floweringDelay* less or equal to 10 days
- Total GM area surrounding a sampling point within the *secureDistance* of 150 m considering only fields that have *floweringDelay* less or equal to 10 days
- Relative non-GM area defined as the proportion of the non-GM area in which the sampling point is located and the total GM area surrounding the sampling point considering both the *secureDistance* of 150 m and the *floweringDelay* less or equal to 10 days.

An image of the area under study is given in Fig. 3.1 whereas an image illustrating the definition of the independent variables used in this study is given in Fig. 3.2.

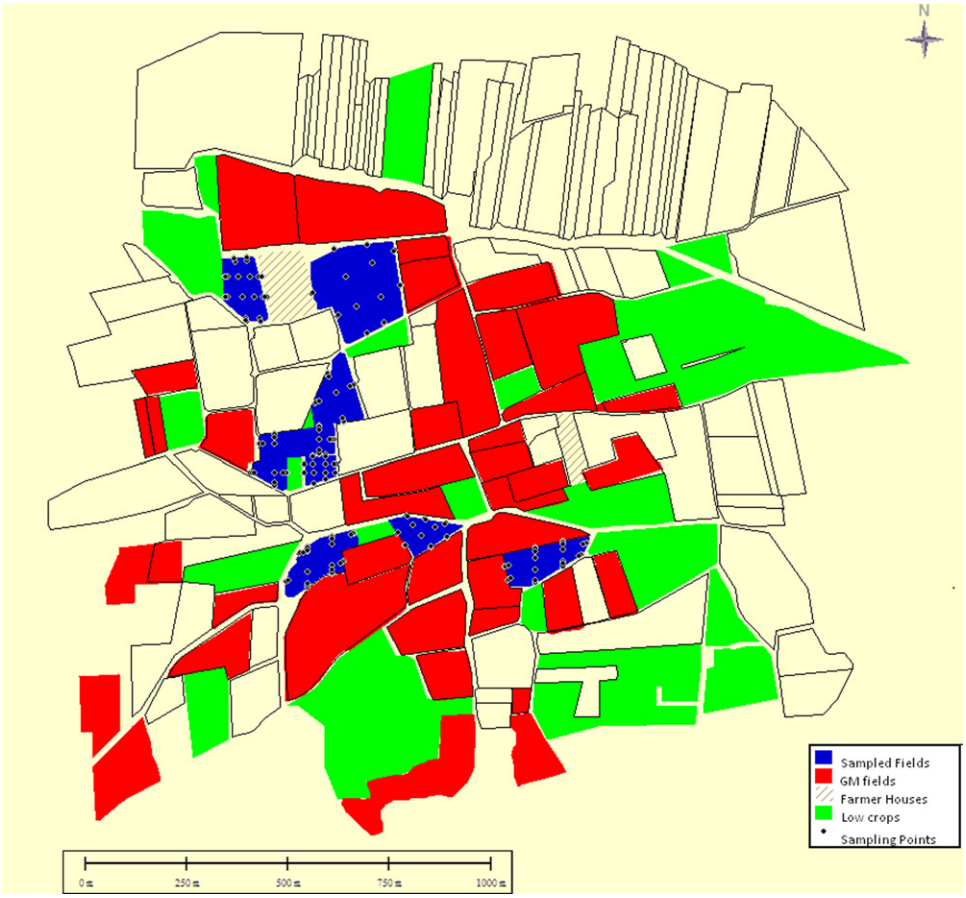

FIGURE 3.1 An image of the area under study, the Foixa region in Spain. The legend shows the different crops grown in 2004. For color version of this figure, the reader is referred to the online version of this book.

The dataset contains observations of the outcrossing rate, the response variable, measured 420 samples points during a period of three successive years, from 2004 to 2006. The sampling points are located in 15 (total) non-GM maize fields. The outcrossing comes from the surrounding 100 (total) GM maize fields, within a secure distance of 150 m radius around a sampling point, considering only fields that have an overlap in the flowering periods, that is, flowering delay of less or equal to 10 days. The explanatory variables include the number of GM fields, the size of the surrounding GM fields, the ratio of the size of the surrounding GM fields and the size of conventional field, the average distance between conventional and GM fields, as well as the coordinates of the sampling points. Later are the two (x, y) spatial coordinates that are not directly included into the models.

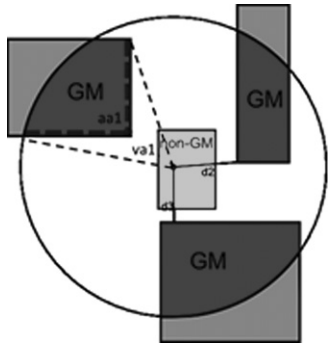

FIGURE 3.2 The influence area (circle) defined by the influence distance from a sample point. The parts of the GM fields that are outside the circle are treated as having no or insignificantly small influence on the conventional field. The (minimal) distances from sampling point to the neighboring GM fields within the influence area are marked as d_2, and d_3, visual angle from sampling point to GM field is marked as va1 (dashed lines), and active edges are marked as aa1 (double dashed lines). For color version of this figure, the reader is referred to the online version of this book.

The response (target) variable is the gene flow (outcrossing rate) that comes from the surrounding GM fields, measured at a sampling point that is located within a non-GM field. It is a nonnegative continuous variable; it has higher values at sampling points located very close to the GM fields and lower values at sampling points located further away from the GM fields.

3.3.2 Data Analysis Setup: Algorithms, Parameters, Evaluation Measures

We apply SCLUS to the data described above. The performance of the SCLUS methods depends on the choice of bandwidth b, the measure of spatial autocorrelation, and the user-defined parameter α. The bandwidth is represented as a percentage of the maximum distance (secure distance) between the sampling points in the dataset. Here, we use 1, 5, 10, and 20% of the maximum distance between the sampling points in the dataset. The spatial statistics are the Global Moran's I and Global Geary's C. The distance between the sampling points is calculated as the Euclidean distance between the points, using their spatial coordinates. Equal importance is given to variance reduction and spatial statistic by setting the user-defined parameter α to 0.5.

We compare the predictive performance of SCLUS with that of the original CLUS method. In addition, SCLUS is compared to other competitive regression methods, which include Linear Regression, M5′ Regression Trees (RT), M5′ Regression Rules (all three implemented in the WEKA system), and GWR. Only GWR and SCLUS take into account spatial autocorrelation in the data.

The predictive performance of the regression models on unseen cases is estimated using the standard 10-fold cross-validation method and evaluated according to the measure of average relative root mean squared error (AvgRRMSE). AvgRRMSE is defined by Eqn. (4) as the average of RMSE of the model predictions normalized with the

RMSE of the default model, that is, the model that always predicts (for regression) the average value of the target:

$$\text{AvgRRMSE} = \frac{1}{10}\sum_{\text{fold}_i \in E}\left(\sqrt{\frac{\sum_{o_j \in \text{fold}_i}\left(f_{E-\text{fold}_i}(x_j) - y_j\right)^2}{\sum_{o_j \in \text{fold}_i}\left(\overline{y}_{\text{fold}_i} - y_j\right)^2}}\right) \tag{4}$$

where $\{\text{fold}_1, \ldots, \text{fold}_{10}\}$ is a cross-validation partition of the sample data E, 10 is the number of folds, $f_{E-\text{fold}_i}(x_j)$ is the value predicted for the j-th test case by the model built from$E - \text{fold}_i$, y_j is the observed response value, and $\overline{y}_{\text{fold}_i}$ is the average of the actual response variable on the test set. The normalization by the denominator removes the influence of the range of values of the response. AvgRRMSE typically varies in the interval [0, 1].

3.3.3. Results: Predictive Performance

Table 3.1 presents the predictive performance results obtained by using the described method SCLUS and the competitive methods in terms of the RRMSE measure. SCLUS results are given for both spatial statistics and different bandwidth values b.

From the results reported in Table 3.1 we can observe that the selection of the bandwidth may influence the accuracy of the learned regression models. In our case, the best results are obtained when the bandwidth is greater than 5%. This indicates that the autocorrelation is more evident in larger neighborhoods than in very narrow ones.

The comparison of the two statistics reveals that there is no statistically significant difference between the results. However, the obtained results confirm the distinguishing characteristics of both statistics described in Section 3.3. The errors of the SCLUS models using Geary's C are more sensitive to the bandwidth, whereas the errors of the SCLUS

Table 3.1 AvgRRMSE of the Regression Models Obtained by Using Different Methods (SCLUS, CLUS, GWR, Linear Regression, M5′ Trees, and M5′ Rules). Only GWR and SCLUS Consider Spatial Autocorrelation. The Performance of the SCLUS Models Depends on the Choice of Bandwidth, α, and the Measure of Spatial Autocorrelation (Moran's I and Geary's C) Used to Learn the Models

	1%	5%	10%	20%	CLUS[1]				
SCLUS Bandwidth		**$\alpha = 0.5$**			**$\alpha = 1$**	**GWR**	**Linear Regression**	**M5′ Trees**	**M5′ Rules**
Moran I	0.902	0.872	0.872	0.872	0.904	1.509	0.944	0.976	0.948
Geary C	0.891	0.873	0.878	0.872					

[1]Note that the secure distance depends on the design of the study area and it is set by a domain expert as a theoretical limit or an empirically estimated value obtained from previous studies. The bandwidth b, on the other hand, regulates the level of autocorrelation included into the models and depends on the data characteristics. The maximum value of the bandwidth b corresponds to the secure distance.

models using Moran's *I* are less sensitive to the bandwidth (therefore, the influence of the bandwidth is smoother).

In Table 3.1, we also report the accuracy obtained by CLUS, Linear Regression, M5' Regression Trees (RT), M5′ Regression Rules, and GWR. Besides SCLUS, only GWR (from the above listed methods) considers spatial autocorrelation. The other methods ignore this phenomenon.

The results show that SCLUS outperforms all methods. This is explained by the fact that all the competitors, except GWR, completely ignore autocorrelation when learning the regression models. GWR builds local models that consider only the local effects of spatial autocorrelation, that is, effects that are limited to only the closest neighborhood.

Beside the differences in accuracy, the obtained models may have different structure. For illustration, in Fig. 3.3a and b, we show the regression trees obtained by using CLUS and SCLUS (with Global Moran I and $b = 20\%$), respectively. The trees have different splits at the root and different size (number of internal nodes, number of leaves). It is noteworthy that the spatially aware PCT (Fig. 3.3b) is smaller and more compact than the PCT learned by CLUS. Moreover, according to domain expertise, the spatially aware PCT is more interpretable and understandable than the corresponding ordinary PCT. This is also due to the fact that the spatially aware PCT is balanced.

A more detailed analysis of the training examples falling in the leaves of both the ordinary PCT and the spatially aware PCT reveals that the leaves of both trees cluster examples with similar response values (this is due to the variance reduction). But training examples falling in the leaves of the spatially aware PCT are also close in space. This

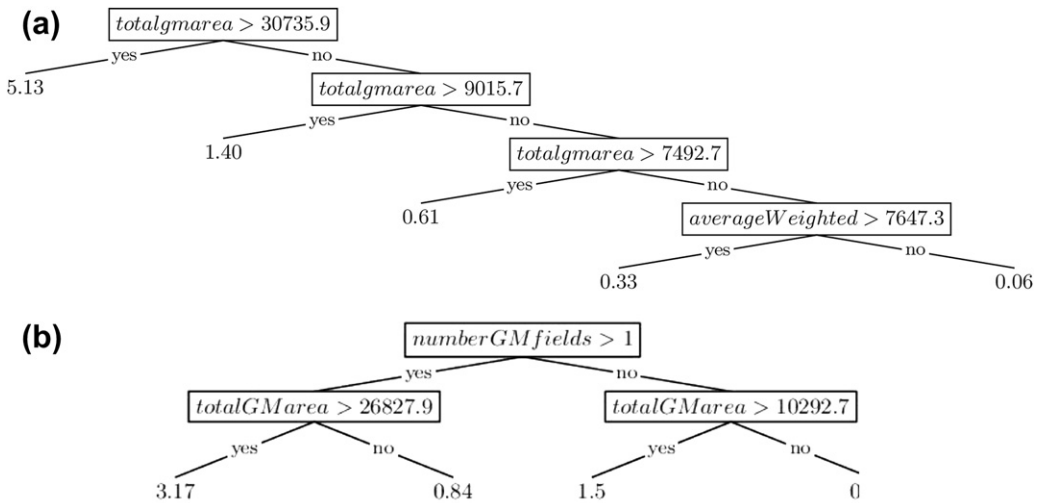

FIGURE 3.3 Two PCTs learned on one of the training fold of Foixa data: (a) The ordinary PCT learned by CLUS and (b) the spatially aware PCT learned by SCLUS.

guarantees spatially smoothed predictions. This means that the predictions that are close to each other in space tend to have similar values. When plotted on a geographical map they form a nice, smooth continuous surface and there are not sharp edges and discontinuities.

3.3.4. Results: Maps of the Predicted Outcrossing Rates

The predictions for testing examples obtained by models learned with GWR, SCLUS, and CLUS (the results are obtained by using onefold as a testing set and the remaining folds as a training set) are visualized in two geographical maps of the outcrossing rates (see Fig. 3.4a–c, respectively). The maps represent the FOIXA study area, as shown in Fig. 3.1.

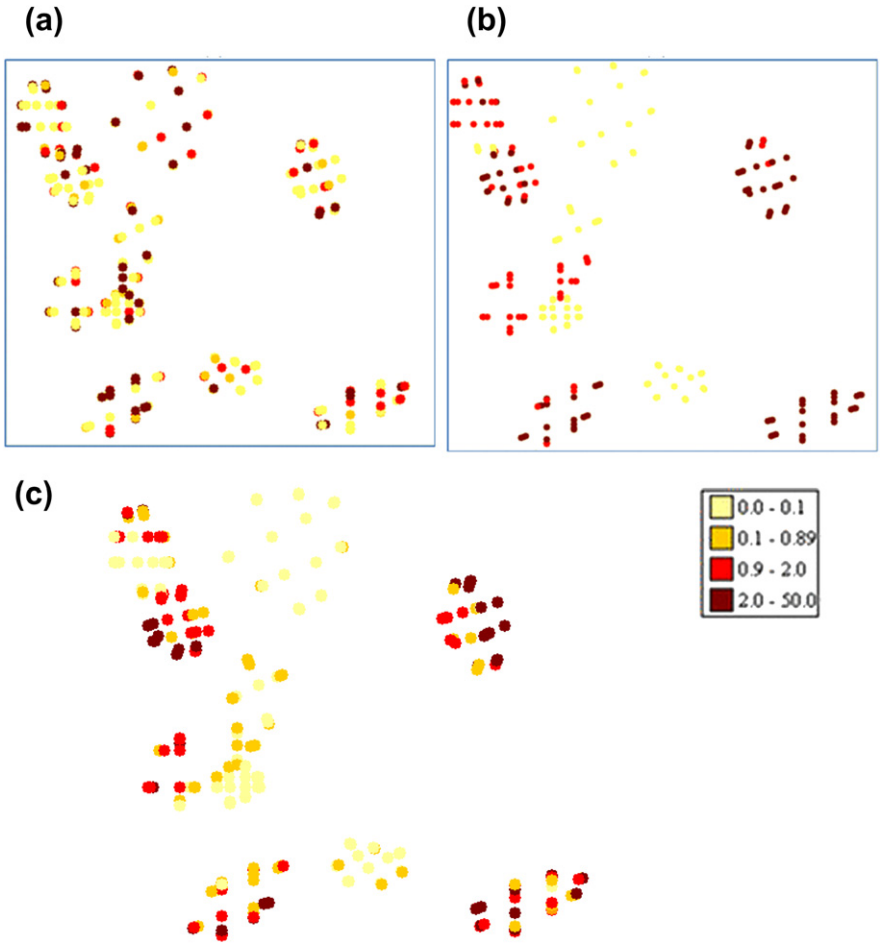

FIGURE 3.4 Map of the predicted outcrossing rate for sampling selected of points in the FOIXA region, generated by using (a) GWR, (b) SCLUS ($b = 20\%$, Global Moran's I, $\alpha = 0.5$), and (c) CLUS. For color version of this figure, the reader is referred to the online version of this book.

However, here we present only the sampling points and disregard the GM and non-GM fields, in order to get a more detailed view of the predicted outcrossing rates at the sampling points.

The map obtained by using the GWR model gives "salt and pepper" results (see Fig. 3.4a). This means that the predictions that are close to each other in space tend to have very different values (very high outcrossing rates come predicted very close to very low ones). When plotted on a geographical map, they do not form a nice smooth continuous surface, but there is sharpness and discontinuity so that the map looks like a mixture of "salt and pepper." This is due to the fact that GWR builds local models at each point, which are independent of the models built in the neighboring points. On the other hand, the GWR models exploit the spatial dimension by using the positive autocorrelation between neighboring points in space and in this way accommodate stationary autocorrelation.

In contrast, the map of predictions obtained by using the model learned by SCLUS (see Fig. 3.4b) shows that the predictions are smoother. This means that the predictions that are close to each other in space tend to have similar values (very high/low outcrossing rates are very close to each other) and when plotted on a geographical map they form a nice smooth continuous surface without sharp edges and discontinuities. Hence, SCLUS models are more accurate in terms of the obtained errors (Table 3.1) and sensible than competitors' models because they are more interpretable (Fig. 3.3b) than all other models. In addition, SCLUS predictions are also spatially smoothed. In practice, these characteristics make SCLUS models more useful than (both spatial and a-spatial) competitors because the geographical mass obtained by using these models is more realistic and easier to interpret.

The map obtained by using the model learned by CLUS (see Fig. 3.4c) is smother that the map obtained by using the GWR model, but we observe that it is a bit sharper and with some discontinues that the one obtained by using the model learned by SCLUS. This is supported by the fact that the accuracy of the predictions obtained by using CLUS is better than the one of the GWR model and worse that accuracy of the model learned by SCLUS.

3.4 Conclusion

In this chapter, we described spatial autocorrelation phenomenon and focused on the task of predicting the outcrossing rate from GM to non-GM maize fields. We demonstrated the use of spatially aware PCTs learned by SCLUS, which explicitly take into account the spatial arrangement of the data and the positive autocorrelation coming from this spatial arrangement. In particular, spatially aware PCTs allow us to learn a predictive model and hierarchically cluster examples at the same time.

Taking spatial autocorrelation into account improves the predictive capability of the models. SCLUS clearly outperforms standard modeling techniques (CLUS, M5′) that do

not consider spatial autocorrelation and GWR that takes into account autocorrelation, but can only capture local (and not global) regularities because it builds local modes at each point and these models are independent from each other. The tree structure solves the ecological fallacy problem and considers, both globally and locally, the effect of autocorrelation.

The learned spatially aware PCTs adapt to local properties of the data, providing at the same time spatially smoothed predictions. In order to take autocorrelation into account, we use well known measures, such as Global Moran's I and Global Geary's C. The split evaluation heuristic that we use in the construction of PCTs is a weighted combination of variance reduction (related to predictive accuracy) and spatial autocorrelation of the response variable. We can also consider different sizes of neighborhoods (bandwidth) when calculating spatial autocorrelation. This allows SCLUS to offer a set of design options which can be adopted to the specific application at hand.

Further work would include the use of the presented approach to other regions and in other ecological modeling problems, since the phenomenon of spatial autocorrelation is ubiquitous. We would also like to extend the use of the presented approach to classification tasks as the current approach deals only with the regression task. In addition, an automated procedure for selecting the size of the neighborhood (bandwidth) is needed to replace the manual tuning of this parameter.

In summary, spatial data measured at locations relatively close to each other tends to have more similar values than data measured at locations further apart (Tobler's first law of geography (Tobler, 1970)). Inappropriate treatment of data with spatial dependencies could obfuscate important insights (e.g., when spatial autocorrelation is ignored). Taking spatial autocorrelation into account is thus crucial, but is only performed by few regression approaches. In our work, we have extended the very popular machine learning approach of decision trees in order to take into account spatial autocorrelation. This has led to the construction of more accurate and comprehensive models for predicting gene flow, which also produces spatially smoothed predictions.

Acknowledgments

This work is in partial fulfillment of the research objectives of the project "EMP3: Efficiency Monitoring of Photovoltaic Power Plants" funded by Fondazione Cassa di Risparmio di Puglia. Džeroski is supported by the Slovenian Research Agency (grants P2-0103, J2-0734, and J2-2285), the European Commission (grant HEALTH-F4-2008-223451), the Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins (operation no. OP13.1.1.2.02.0005), and the Jožef Stefan International Postgraduate School. Stojanova is supported by the grant J2-2285. The authors acknowledge the support of the Slovenian–French bilateral scientific collaboration program PROTEUS 2009–2010 and the Consorci Laboratori CSIC-IRTA de Genetica Molecular Vegetal, Departament de Genetica Vegetal, Barcelona, Spain, for providing the data.

References

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. Roy. Stat. Soc. B 36, 192–236.

Betts, M.G., Ganio, L.M., Huso, M.M.P., Som, N.A., Huettmann, F., Bowman, J., et al., 2009. Comment on "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review". Ecography 32, 374–378.

Blockeel, H., De Raedt, L., Ramon, J., 1998. Top-down induction of clustering trees. In: Proceedings of the Fifteenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, pp. 55–63.

Cressie, N., 1991. Statistics for Spatial Data. Wiley, New York.

Debeljak, M., Džeroski, S., 2011. Decision trees in ecological modeling. In: Modeling Complex Ecological Dynamics. Springer, pp. 197–209.

Debeljak, M., Trajanov, A., Stojanova, D., Leprince, F., Džeroski, S., 2011. Using relational decision trees to model flexible co-existence measures in a multi-field setting. In: Proc. 7th European Conf. on Ecological Modelling, p. 53.

Dormann, C.F., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30, 609–628.

Ester, M., Kriegel, H., Sander, J., 1997. Spatial data mining: a database approach. In: Proc. 5th Intl. Symp. on Spatial Databases. Springer, pp. 47–66.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley.

Gora, G., Wojna, A., 2002. RIONA: a classifier combining rule induction and k- NN method with automated selection of optimal neighbourhood. In: Proc. 13th European Conf. on Machine Learning. Springer, pp. 111–123.

Griffith, D., 2002. A spatial filtering specification for the auto-Poisson model. Stat. Probab. Lett. 58, 245–251.

Kühn, I., 2007. Incorporating spatial autocorrelation may invert observed patterns. Divers. Distrib. 13 (1), 66–69.

Legendre, P., 1993. Spatial autocorrelation—trouble or new paradigm. Ecology 74, 1659–1673.

Legendre, P., Dale, M.R.T., Fortin, M.-J., Gurevitch, J., Hohn, M., Myers, D., 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25, 601–615.

LeSage, J.H., Pace, K., 2001. Spatial dependence in data mining. In: Data Mining for Scientific and Engineering Applications. Kluwer, pp. 439–460.

Messeguer, J., Peñas, G., Ballester, J., 2006. Pollen-mediated gene flow in maize in real situations of coexistence. Plant Biotechnol. J. 4 (6), 633–645.

Robinson, W.S., 1950. Ecological correlations and the behavior of individuals. Am. Sociol. Rev. 15, 351–357.

Stojanova, D., Ceci, M., Appice, A., Malerba, D., Džeroski, S., 2011. Global and local spatial autocorrelation in predictive clustering trees. In: Proc. 15th Intl. Conf. on Discovery Science. Springer, pp. 307–322.

Tobler, W., 1970. A computer movie simulating urban growth in the Detroit region. Econ. Geogr. 46 (2), 234–240.