Discovering Homogeneous Regions in Spatial Data through Competition

Slobodan Vucetic SVUCETIC@EECS.WSU.EDU
Zoran Obradovic ZORAN@ EECS.WSU.EDU
School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-2752, USA

Abstract

Here, a supervised machine learning algorithm for the analysis of heterogeneous spatial data is proposed. It is based on partitioning the data set into more homogeneous regions by competition of regression models (linear or nonlinear). The algorithm starts from learning a global model, and adds new models into the competition until each model becomes specialized for one of the regions. The competition convergence is proven theoretically. Also, the influence of filtering the competing models residuals for improving convergence speed and accuracy is discussed. A number of experiments on artificial and real-life spatial data are performed to validate some aspects of the algorithm and to illustrate its potential applications. The obtained results provide strong evidence that homogeneous regions can be identified with high accuracy by using the proposed approach even when their observed feature spaces highly overlap.

1. Introduction

Distinctive properties of spatial data require specific data analysis and modeling techniques (Cressie, 1993). Spatial data can often be heterogeneous, such that in distant areas the dependencies between observed features are different. If it is possible to partition heterogeneous data sets into homogeneous subsets, a local model could be learned separately on each of them. This would result in better overall prediction accuracy as compared to constructing a single model on all data since a global model is likely to be biased on most homogeneous subsets. Thus, in this study a data set is said to be *heterogeneous* if its partitioning can improve prediction accuracy. Given multivariate spatial data consisting of real-valued features and a real-valued target, the learning tasks addressed in this study are: (1) determining if the data set is heterogeneous; (2) discovering an appropriate partitioning into more homogeneous subsets; and (3) constructing an ensemble of regression models specialized to the identified subsets.

If all features causing heterogeneity were observed, a mixture of experts approach (Jacobs et al., 1991) is likely to be superior to using a single model. When unobserved or very noisy spatial features are the cause for the heterogeneity, the observed feature spaces of homogeneous subsets can highly overlap, leading to a biased global model or biased mixture of experts. Our goal is to allow more accurate predictions in such situations.

An assumption of data independence valid for most of standard machine learning data sets is often unrealistic for spatial variables, whose dependence is strongly tied to a location, where observations spatially close to each other are more likely to be similar than observations widely separated in space. As a consequence, errors of spatial prediction models are also spatially correlated (Cressie, 1993). The method proposed here incorporates knowledge of spatial correlation for more accurate partitioning of heterogeneous spatial data sets.

The new partitioning algorithm is based on three important mechanisms: (a) competition among learning models for spatial data points, (b) averaging errors of each competing model over neighboring data points, and (c) an incremental introduction of additional models into the competition when needed.

In Section 2, the consequences of applying a global versus an ensemble of local models for heterogeneous spatial data are formally discussed. The initial algorithm for spatial data partitioning is proposed and analyzed in Section 3. An extension of the algorithm through error filtering is suggested in Section 4. The experimental results on artificial and real life data are discussed in Section 5.

2. Learning on Heterogeneous Spatial Data

Let us assume that a spatial data set is a union of K disjoint homogeneous regions R_i, i = 1, …,K, where the number of regions, their size and location are unknown. The data generating process for R_i can be represented as

$$y_s = E_i[y_s|\mathbf{x}_s] + \varepsilon_s, \quad s \in R_i,$$

where $\boldsymbol{x}_s$ is a vector of observed features and ε_s is the error term representing unexplained variance. Subscript s is used to indicate the spatial position of each data point. Domain D_i of observed features $\boldsymbol{x}_s$ corresponding to region R_i generally overlaps with the domains of other homogeneous regions. This means that the same input vector $\boldsymbol{x}_s$ can, in different regions, result in quite different outputs, y_s.

Without prior knowledge of regions, learning a global regression model on the whole data set would ideally result in learning $E^*[y_s|\boldsymbol{x}_s]$ defined as

$$E^*[y_s|\boldsymbol{x}_s] = \arg\min_{f(\boldsymbol{x})} E_D[(y_s - f(\boldsymbol{x}_s))^2] \ ,$$

where $D = D_1 \cup D_2 \cup \ldots \cup D_K$. Since whole data is a mixture of homogeneous regions, $E^*[y_s|\boldsymbol{x}_s]$ can be expressed as

$$E^*[y_s|\boldsymbol{x}_s] = \sum_{i=1}^{K} r_i \ p_i(\boldsymbol{x}_s) E_i[y_s|\boldsymbol{x}_s] / \sum_{i=1}^{K} r_i \ p_i(\boldsymbol{x}_s) \ ,$$

where r_i is the fraction of data points from region R_i used for regression, and $p_i(\boldsymbol{x}_s)$ is the probability density of input $\boldsymbol{x}_s \in D_i$. As could be seen, the higher the r_i the closer $E^*[y_s|\boldsymbol{x}_s]$ is to $E_i[y_s|\boldsymbol{x}_s]$.

Mean squared error (MSE) of the global regression model $E^*[y_s|\boldsymbol{x}_s]$ on region R_i can be expressed as

$$mse_i = E_{D_i}[(y - E_i(y|\boldsymbol{x}))^2] + E_{D_i}[(E^*(y|\boldsymbol{x}) - E_i(y|\boldsymbol{x}))^2] ,$$

where the first term corresponds to an unavoidable error (obtained by a local model specialized for region R_i) and the second term corresponds to the bias of the global model on data from region R_i. If locations of homogeneous regions were known, the second term from the previous equation would be canceled by learning a local model on each homogeneous region.

Let us now discuss the connection between heterogeneous spatial data, bias and complexity of global models. For spatial data, observed features and target are likely to be spatially correlated, and so the error e_s of a global model will likely be spatially correlated. In spatial statistics, e_s is often decomposed into two components: a slowly varying bias component w_s, and a spatially uncorrelated component u_s, such that $e_s = w_s + u_s$. The bias w_s can be caused by: (i) the misspecification of an estimation model, e.g. applying a linear model on spatial data with nonlinear relationships; or (ii) an application of a global model on heterogeneous spatial data (mse_i equation). In both cases it should be possible to suppress w_s, either by: (a) introducing more complex learning algorithms; or (b) partitioning the data into several homogeneous regions, and learning a separate local model on each of them. For the rest of the paper, we assume that the employed models are of sufficient complexity such that bias component is a consequence of data heterogeneity. Therefore, discovering more homogeneous regions in spatial data is equivalent to bias reduction.

Observe that examining the error of a global model can provide an indication of possible heterogeneities in spatial data. The level of bias can be effectively estimated from an error correlogram used to measure spatial continuity of error. A *correlogram* of a spatial variable is a plot of the correlation coefficient computed as a function of the separation distance between spatial data (Cressie, 1993). The correlogram is characterized by its type, range and nugget effect as shown in Figure 1. The *correlogram type* describes the way spatial dependency changes over distance, and is named by its approximate functional form (e.g., exponential, spherical, Gaussian). The *range* corresponds to a distance where spatial dependence vanishes, e.g. where the value of the correlogram drops somewhere between 0.01-0.1. The *nugget effect*, *Nugget*, equals the size of the spike of the correlogram at distance zero, and corresponds to the nonspatial component of a spatial variable. After calculating the error correlogram of a global model, the value (1–*Nugget*) represents the percent of a total error variance contained in the bias component w_s. The larger (1–*Nugget*) is, the larger the potential benefits of spatial data partitioning can be.

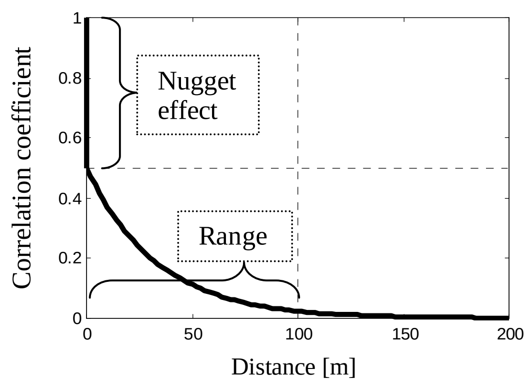


Figure 1. An exponential correlogram with range 100 m and nugget effect 0.5

3. Spatial Partitioning Algorithm

The proposed spatial partitioning algorithm relies on the competition among regression models for each part of a given data set. By $M(\boldsymbol{x}, \beta)$ we denote the regression model with input variables $\boldsymbol{x}$ and the set of parameters β. A whole data set (heterogeneous region R containing N data points) is denoted as $S_0 = \{(\boldsymbol{x}_s, y_s), s \in R\}$, where y_s is the target variable. If a given region is partitioned into L disjoint subsets R_i, $i=1, \ldots L$, the corresponding data subsets are denoted as S_i, where $S_i = \{(\boldsymbol{x}_s, y_s), s \in R_i\}$, and the whole partitioning as $\mathbb{S}_L = \{S_1, \ldots S_L\}$. By $M_i(\boldsymbol{x}, \beta)$ we denote a local regression model built on subset R_i.

The algorithm (formally described in Table 1) starts by learning a global regression model on the whole data set. Next, the region is split spatially into two disjoint subsets of equal size and separate regression models are trained for each of them. The two models then compete for each data point such that a given point is assigned to the model achieving smaller error. This competition procedure iterates until a stable partitioning $\mathbb{S}_2$ is obtained. The competition procedure is described in Table 2 and it contains an optional error filtering step which will be ignored in this section, but will be discussed in the next section. Subsequently, the prediction error obtained by such a partition is compared to the prediction error of the global model. If the improvement is not sufficiently large, the algorithm is stopped and it is concluded that the spatial data set is homogeneous. If using two regression models is better than one global model, the algorithm proceeds by partitioning data into three subregions in an attempt to further decrease the total error.

Table 1. Spatial partitioning algorithm

- ❑ Learn a single regression model on global data set S_0 and calculate its MSE, mse_1.
- ❑ Split S_0 into $\mathbb{S}_2^0 = \{S_{1,2}^0, S_{2,2}^0\}$, where $S_{1,2}^0$ and $S_{2,2}^0$ are disjoint and of equal size.
- ❑ Modify $\mathbb{S}_2^0$ through the *competition procedure* (Fig. 3) and calculate error mse_2 achieved by the obtained partitioning $\mathbb{S}_2 = \{S_{1,2}, S_{2,2}\}$.
- ❑ Terminate the algorithm if the ratio mse_2/mse_1 is larger than α. Spatial data is homogeneous.
- ❑ $L=2$.

repeat
    for $i = 1$ to L
- Split $S_{i,L}$ into two disjoint subsets of the same sizes, $S_{i,L}'$ and $S_{i,L}''$, to obtain the initial partitioning $\mathbb{S}_{L+1,i}^0 = \{\mathbb{S}_L/S_{i,L}, S_{i,L}', S_{i,L}''\}$.
- Modify $\mathbb{S}_{L+1,i}^0$ through the *competition procedure* (Fig. 3) and calculate error $mse_{L+1,i}$ achieved by the obtained partitioning $\mathbb{S}_{L+1,i}$.

    end
- ❑ Out of L partitions, $\mathbb{S}_{L+1,i}$, $i = 1, \ldots L$, choose the partitioning achieving the smallest error, $mse_{L+1,i}$, to represent the new partitioning, $\mathbb{S}_{L+1} = \{S_{1,L+1}, \ldots S_{L+1,L+1}\}$, and save its error as mse_{L+1}.
- ❑ $L \leftarrow L + 1$

until the ratio mse_L/mse_{L-1} is smaller than $\alpha, \alpha \in (0,1)$
Output: $L-1$ 'homogeneous regions' defined by partitioning $\mathbb{S}_{L-1}$ and their corresponding models.

The third model is added to the competition in a tree-like manner as described in Table 1. Starting from a partitioning into $S_{1,2}$ and $S_{2,2}$, subregion $S_{1,2}$ is split spatially into two equal size disjoint subsets, while $S_{2,2}$ is left intact. The competition of three models starting from such a partitioning is performed and after converging to a stable partitioning $\mathbb{S}_{3,1}$, the error is calculated. Also, the same steps are repeated by splitting S_2 spatially into two subsets and leaving S_1 intact to obtain $\mathbb{S}_{3,2}$. The better of the two partitions (denoted as $\mathbb{S}_3$) is compared to $\mathbb{S}_2$. The procedure continues by adding new models into the competition until it is concluded that further partitioning does not improve predictability of the spatial data.

Table 2. The competition procedure

Start from a partitioning into L subsets $\mathbb{S}_L^0$ and set $n=0$.
repeat
- ❑ Learn L models on L subsets of $\mathbb{S}_L^n$, S_i, $i=1, \ldots L$.
  - ❑ Optionally, filter errors or absolute errors of all models on the whole data set with MAF.
- ❑ If $M_i(\mathbf{x}, \beta)$ gives the smallest filtered error on $(\mathbf{x}_s, y_s)$, assign this point to subset S_i. Form a new partitioning, $\mathbb{S}_L^{n+1}$, by reassigning all data in this way.
- ❑ $n \leftarrow n+1$

until less than $p\%$ of data points change subsets.

3.1 A Technique for Comparison of Partitionings

When using complex nonlinear models (e.g. neural networks) on spatial domains special attention should be given to training and testing procedures. For spatial data, input features, target and residuals are likely to be spatially correlated. Therefore, a random split of spatial data into a training and a test subset would lead to overly optimistic generalization estimates. This is a particularly serious problem for complex nonlinear models able to memorize training data. To avoid this problem and to allow proper performance estimation of the partitioning algorithm, testing points should be as spatially distant as possible from data points used for training and validation. One solution can be to divide the spatial region into squares of size S×S meters and assign a fraction of the squares exclusively for testing (Vucetic et al., 1999).

This choice of test set is incorporated into the algorithm in order to provide correct termination signaling. Before starting the algorithm, a fraction of squares of size S×S is designated for testing. However, the whole data set is used in the competition procedure to provide as much data as possible for the partitioning algorithm. The other reason for using the whole data set for partitioning is that big spatial holes in the data caused by removing a number of squares from a training data set would deteriorate the proper delineation between homogeneous subregions.

After competing models have converged to their corresponding regions, all data points are labeled according to the region to which they belong. Then, new regression models are learned on each subregion, excluding the squares designated for testing. Each of the new models is then tested on the test points of the corresponding label, and the achieved MSE is reported as a measure of goodness for the obtained partitioning.

If using less powerful linear regression models, the choice of a test set is not so critical as there is no danger of overfitting, and so standard statistical performance measures can be used without a need for testing on S×S squares.

3.2 Analysis for Two Homogeneous Regions

In this section we provide insight into the performance of the competition procedure by analyzing its behavior for the heterogeneous data with two (K=2) homogeneous regions R_1 and R_2. It is assumed that errors of ideal models for regions R_1 and R_2 are spatially uncorrelated and normally distributed with mean zero and variances σ_1^2 and σ_2^2, respectively.

For two models (L=2), determining the regression function with minimum error on a data point $(\boldsymbol{x}_s, y_s)$ is equivalent to determining the sign of $\Delta e_s^2 = e_{s,2}^2 - e_{s,1}^2$, where $\{e_{s,1}\}$ and $\{e_{s,2}\}$ are the errors on the whole data set of the two competing regression models, $M_1(\boldsymbol{x}, \beta)$ and $M_2(\boldsymbol{x}, \beta)$, built on disjoint subsets S_1 and S_2, respectively. A positive sign of Δe_s^2 means assigning a data point to model M_1, and a negative sign to model M_2. The following proposition gives the expression for a conditional distribution of Δe_s^2 on the two regions, R_1 and R_2. Without loss of generality, it is assumed that $mse_{1,1} < mse_{2,1}$ and $mse_{2,2} < mse_{1,2}$, where $mse_{i,j}$ is MSE of $M_i(\boldsymbol{x}, \beta)$ on region R_j and it can be expressed as

$$mse_{i,j} = E_{D_j}[(y - E_j(y|\,\boldsymbol{x}))^2] + E_{D_j}[(M_i(\boldsymbol{x}, \beta) - E_j(y|\,\boldsymbol{x}))^2],$$

with D_j denoting the domain for region R_j, i, j =1, 2.

Proposition 1. $\{\Delta e_s^2 \,|\, \boldsymbol{x}_s, s\in R_1\} \sim \mathcal{N}(A(\boldsymbol{x}_s)+\delta_1,\ 4A(\boldsymbol{x}_s)\sigma_1^2)$, where δ_1 is a small positive number, and $A(\boldsymbol{x}_s) = (M_1(\boldsymbol{x}_s, \beta) - M_2(\boldsymbol{x}_s, \beta))^2$. Similarly, $\{\Delta e_s^2 \,|\, \boldsymbol{x}_s, s\in R_2\} \sim \mathcal{N}(-A(\boldsymbol{x}_s)-\delta_2,\ 4A(\boldsymbol{x}_s)\sigma_2^2)$, where δ_2 is a small positive number.

Proof. Omitted for lack of space.

The following proposition proves the convergence of the competition procedure, where $P_{i,j}(n)$ denotes the fraction of data points from region R_j in subset S_i at iteration n of the competition procedure.

Proposition 2. If $P_{j,j}(n_0+1) > P_{j,j}(n_0)$ for some n_0, then $P_{j,j}(n)$ is a strictly increasing function for $n>n_0$.

Proof sketch. $E_{D_j}[(M_j(\boldsymbol{x}, \beta) - E_j(y|\,\boldsymbol{x}))^2]$ decreases from n_0^{th} to $(n_0+1)^{th}$ iteration, which in turn leads to an increase of $A(\boldsymbol{x}_s)$. Consequently, the probabilities $P\{\Delta e_s^2 > 0 \,|\, \boldsymbol{x}_s, s\in R_1\}$ and $P\{\Delta e_s^2 < 0 \,|\, \boldsymbol{x}_s, s\in R_2\}$ increase, leading to an increase of $P_{j,j}$ for the following iterations.

However, $P_{j,j}(n)$ does not reach 1 in the limit (when $n\to\infty$), since $P\{\Delta e_s^2 > 0 \,|\, \boldsymbol{x}_s, s\in R_1\} < 1$. The bound on $P_{j,j}(n)$ for $n\to\infty$ depends on the variances σ_1^2 and σ_2^2, and the difference between the underlying regression functions of regions R_1 and R_2, as shown by the following proposition.

Proposition 3. When $n\to\infty$, the probability of correctly assigning data points is upper bounded as

$$P_{j,j}(n) < \int_{D_j} P\{Z(\mathbf{x}_s) > 0\} p_j(\mathbf{x}_s) d\mathbf{x}_s \text{ , where}$$

$$Z(\boldsymbol{x}_s) \sim \mathcal{N}((E_1[y_s|\boldsymbol{x}_s] - E_2[y_s|\boldsymbol{x}_s])^2,\ 4\sigma_j^2(E_1[y_s|\boldsymbol{x}_s] - E_2[y_s|\boldsymbol{x}_s])^2).$$

Proof sketch. Assuming a perfect separation at some iteration, $M_1(\boldsymbol{x}, \beta) = E_1[y|\boldsymbol{x}]$ and $M_2(\boldsymbol{x}, \beta) = E_2[y|\boldsymbol{x}]$, the probability of a correct classification, $P_{j,j}$, at the next iteration depends on the distribution given by Prop. 1.

Therefore, a perfect separation of homogeneous regions should not be expected. However, as will be explained in Section 4, and experimentally tested in section 5.1.1, spatially filtering errors of competing models can improve the data partitioning by increasing $P_{j,j}(n)$.

4. Spatial Partitioning with Error Filtering

One of the most important mechanisms for efficient discovery of homogeneous subregions is competition of regression models with assigning a data point to the model achieving the smallest error of all competing models. If there are K homogeneous subregions R_i and a specialized model M_i is learned on each, it is reasonable to assume that models M_j, i≠j, would have larger bias on R_i then the model M_i. However, for large unexplained variance in spatial data, the nonspatial error component is also large. Therefore, the probability of M_j, achieving the smaller error then M_i on a point from R_i can be significant, causing a number of points to be misclassified to a wrong homogeneous region. In practice, the competing models are being learned on a mixture of points from the whole data set, and a misclassification due to unexplained variance is even higher (see Proposition 1).

For a better separation between homogeneous regions, it is desirable to decrease the effect of nonspatial error component u_s while preserving bias w_s of all competing models. From spatial statistics, it is well known (Cressie, 1993) that extracting w_s from u_s is an unidentifiable problem and therefore some assumptions should be made to solve this problem approximately. Since u_s is a spatially uncorrelated component, it can be assumed to correspond to a nugget effect of the error correlogram. Bias w_s then corresponds to the part of the correlogram at distances larger than zero. For improving the spatial partitioning algorithm, we propose an efficient heuristic for extraction of the bias component based on a spatial moving average filtering (MAF) of the prediction error. MAF of the proper size should cancel out u_s, while preserving w_s as much as possible. Observe that for too small size of MAF a large potion of u_s would be preserved. Also, for too large MAF, w_s would be filtered out together with u_s.

The choice of proper MAF size should be closely related to the error correlogram. We propose selecting MAF size based on an analysis of the range and nugget effect of the error correlogram of a global model. MAF size should not exceed the range of the error correlogram (to prevent filtering out the bias component) and should decrease with the nugget effect. Based on this analysis, instead of the competition procedure described in Section 3, its modification with filtering the errors of competing models is used (an optional step in Table 2). Two types of moving average filtering are examined in the experimental section: (Type 1) MAF of errors of competing models, and (Type 2) MAF of absolute errors of competing models.

5. Experimental Results

Spatial data used in the experiments were: (1) artificially generated sets designed for fully controlled examination of some aspects of the partitioning algorithm, and (2) a real-life precision agriculture data set aimed at demonstrating a possible application scenario.

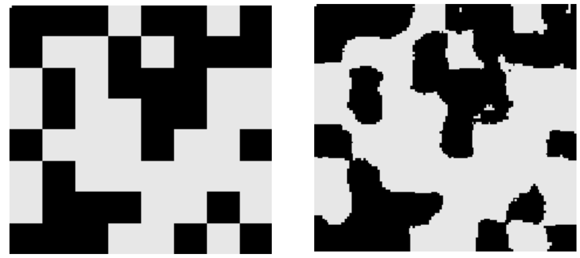
5.1 Experiments on Artificially Generated Spatial Data Sets

We generated two types of artificial heterogeneous spatial data such that observed feature spaces of homogeneous regions highly overlap. For both data sets, it was assumed that 5 input spatial features were observed, while a spatial feature that was an indicator of existing homogeneous regions was hidden. It was assumed that both spatial data sets originated from a squared region of size 1280×1280 meters, with data points gathered along a square grid with 10 meters distance (totaling 128×128 data points). Five layers with Gaussian distribution, mean zero, variance one, and the spatial characteristics shown in the Table 3 were generated as input features for both data sets using the method of the moving averages (Oliver, 1995).

Table 3. Description of generated layers

Layer	Type of correlogram	Range [m]
X_1	Exponential	100
X_2	Exponential	150
X_3	Spherical	150
X_4	Spherical	200
X_5	Gaussian	100

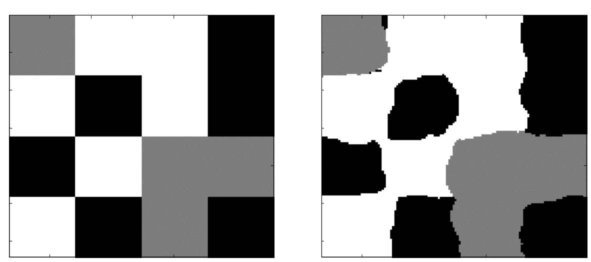
Data set DS_1 with two linear homogeneous regions. The squared region was first partitioned into 64 equal squares of size 16×16 data points. The value of the hidden feature was 1 and 2 in randomly selected halves of the squares (representing homogeneous regions R_1 and R_2) as shown in Figure 2.a. The overlap of normalized histograms of R_1 and R_2 in observed features X_1-X_5 was 95%, 98%, 95%, 96% and 93%, respectively. In both regions, the target was generated to be a linear function of the observed input features, such that in both regions it had mean zero and variance one. Finally, Gaussian noise with variance one was also added to the target variable.



(a) (b)

Figure 2. (a) Assignment of two homogeneous regions in DS_1, and (b) the partitioning obtained after a single iteration of the competition procedure

Data sets $DS_{2,1}$ and $DS_{2,2}$ with three highly nonlinear homogeneous regions. The squared region was first partitioned into 16 equal squares of size 32×32 data points. The hidden feature values were 1, 2 and 3 in randomly selected thirds of the squares (representing homogeneous regions R_1, R_2 and R_3) as shown in Figure 3.a. An averaged pairwise overlap of histograms of R_1, R_2 and R_3 in observed features X_1-X_5 was 94%, 92%, 89%, 88% and 90%, respectively. One of three highly nonlinear functions was assigned to each homogeneous region and each was of the form $y=\prod_{i=1}^{5} f_i(x_i)$, where $f_i(x)$'s were highly nonlinear functions (Pokrajac et al., in review). Also, in all three homogeneous regions the target means were set to zero and target variances to one, thus forming $DS_{2,1}$. Data set $DS_{2,2}$ with statistically identical properties was generated for the purposes of testing the generalization capabilities of the spatial partitioning algorithm. A new set of 5 input features was generated according to Table 3, and it was assumed that $DS_{2,2}$ was obtained from the same spatial region after some time period, such that the locations of homogeneous regions and their nonlinear relationships did not change (hidden features remained the same). Finally, Gaussian noise with variance 1 was added to the target variable in both $DS_{2,1}$ and $DS_{2,2}$.



(a) (b)

Figure 3. (a) Assignment of three homogeneous regions in $DS_{2,1}$ and $DS_{2,2}$, and (b) the resulting partitioning

For both types of generated data sets observed input features could not help in discriminating between different homogeneous regions, making the problem of spatial partitioning extremely difficult. Using the first data set (DS_1), convergence speed of the competition procedure and influence of error filtering were tested. With the second data set $DS_{2,1}$, neural networks and linear models were used as regression models for the algorithm, and generalization capabilities of the obtained partitionings were examined on $DS_{2,2}$.

5.1.1 Testing Convergence and Accuracy on DS_1

Assume that two linear models compete for DS_1 data, such that at iteration n of the competition procedure model $M_1(n)$ is learned on set $S_1(n)$ which is a mixture of $P_{1,1}(n)$ data from R_1 and $1–P_{2,2}(n)$ of the data from R_2, while model $M_2(n)$ is learned on set $S_2(n)$ which is a mixture of $1–P_{1,1}(n)$ from R_1 and $P_{2,2}(n)$ from R_2. Assuming $P_{1,1}(n)$ and $P_{2,2}(n)$ are larger than 0.5, models $M_1(n)$ and $M_2(n)$ can be considered as experts for regions R_1 and R_2, respectively. We were interested in the speed of convergence of the competition procedure to more homogeneous regions, and in how well the obtained partitioning resembles the actual locations of homogeneous regions. Therefore, we measured the fraction of correctly classified points at iteration n of the competition defined as:

$$P(n) = \{|R_1 \cap S_1(n)| + |R_2 \cap S_2(n)|\} / |R_1 \cup R_2| \;,$$

and we plot $P(n+1)$ vs. $P(n)$.

Using DS_1 (Figure 2.a) experiments were performed over the whole range of $P(n)$, such that for various fractions $P(n)$ we constructed equal size subsets $S_1(n)$ and $S_2(n)$ with $P_{1,1}(n) = P_{2,2}(n) = P(n)$, and learned linear models $M_1(n)$ and $M_2(n)$ on them, respectively. Based on the competition between $M_1(n)$ and $M_2(n)$, new sets $S_1(n+1)$ and $S_2(n+1)$ were constructed and the corresponding $P(n+1)$ was calculated.

Plots of $P(n+1)$ vs. $P(n)$ for the competition procedure (a) without error filtering; (b) with error filtering and MAF size of 70×70 meters, and (c) with absolute error filtering and MAF size of 70×70 meters are shown in Figure 4. It could be seen that an absolute error filtering resulted in the most accurate discovery of homogeneous regions. Also, the shape of the plot reveals that the competition procedures converge to a stable partitioning very fast even for an initial partitioning with P(1) value near 0.5.

Starting from an initial partitioning into 2 equal size subsets with a fraction of only $P(1)$=0.55, the fraction of correctly classified points after one iteration of the competition procedure increased to $P(2)$=0.93 in the case of absolute error filtering. The result of this partitioning is shown in Figure 2.b.

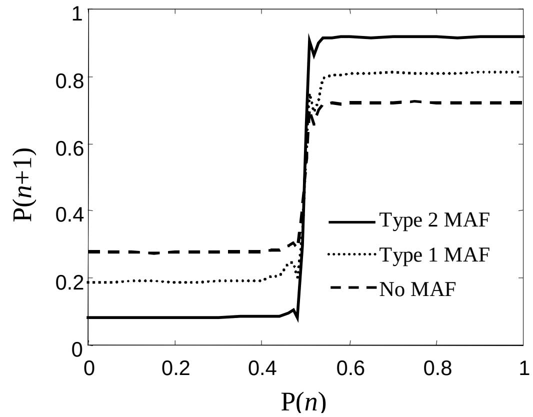


Figure 4. P(n+1) vs. P(n) for three types of error filtering on DS_1

To test the importance of choosing an appropriate MAF size, the values of $P(n+1)$ for $P(n)$=0.55, and different sizes of MAF for error filtering and for absolute error filtering are shown in Figure 5. The optimum MAF size was 70×70 meters, but it should be noted that for a large range of MAF sizes, $P(n+1)$ was significantly better than using the competition algorithm without error filtering (with $P(n+1)$=0.72). Also, the absolute error filtering consistently outperformed simple error filtering for any MAF size.

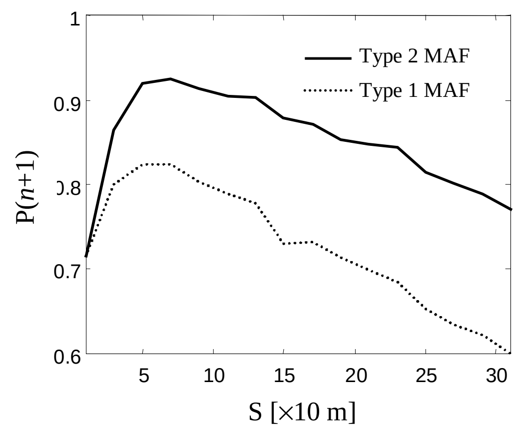


Figure 5. P(n+1) vs. size of MAF (expressed as S×S) for two types of error filtering on DS_1

For different types of error filtering and 4 different levels of noise with variances σ^2={0, 0.25, 1, 4} in DS_1, and starting from $P(1)$=0.55, $P(2)$ values were shown in Table 4. In each experiment, the best achieved $P(2)$ is reported together with the corresponding MAF size. As seen, both filtering of the error and filtering of the absolute error (Type 1 and Type 2 MAF) largely outperform the algorithm without error filtering (No MAF) even if small noise level was present in the data. Also, the filtering of absolute errors consistently outperformed the simple filtering of errors, and is therefore accepted as an error filtering procedure to be used with the algorithm in the remaining experiments. In addition, it should be noted that the size of the optimum filter increased with the level of the noise in the data. For comparison, the correlograms of errors of a global model had range at about 100 meters for all levels of noise.

Table 4. Influence of error filtering on the competition

σ^2	No MAF	Type 1 MAF		Type 2 MAF	
	P(2)	P(2)	Size [m]	P(2)	Size [m]
0	0.92	0.92	1	0.97	30
1/4	0.78	0.85	50	0.96	50
1	0.68	0.80	70	0.93	70
4	0.60	0.73	110	0.88	110

5.1.2 Performance on Highly Nonlinear $DS_{2,1}$ and DS_2 Data

A global neural network with 5 hidden nodes was first trained on the whole $DS_{2,1}$. The error correlogram of the global model is calculated to select the size S of 80 meters for error filtering. Neural networks with 5 hidden nodes were also competing in the spatial partitioning algorithm, with the stopping criterion α from Table 1 set to 0.95. 30% of squares of size 50×50 meters were randomly selected to be used as the testing data set for comparing the obtained partitionings.

Table 5. Intermediate results of partitioning $DS_{2,1}$

Number of regions	1	2	3	4
R^2	0.55	0.73	0.89	0.90

The intermediate results of applying the partitioning algorithm to $DS_{2,1}$ starting from the global model and ending with four regions are shown in Table 5. Performance was measured by the coefficient of determination R^2 computed as $1–MSE/\sigma^2$, where σ^2 was the target variance, and which is used as a measure of predictive capabilities of different regression model (R^2 of 1 and 0 corresponds to the perfect and a mean prediction, respectively). The small difference between R^2 corresponding to partitioning into three and four regions indicated that three homogeneous regions existed (shown in Figure 3.b). Here, 94% of all data points were correctly classified to the appropriate homogeneous regions. This suggests that the proposed algorithm can be successfully applied on highly complex heterogeneous data sets.

Table 6. Performance comparison on $DS_{2,2}$

Model	R^2
Global	0.54
Ideal partitioning	0.95
Partitioning algorithm	0.90

To further test the algorithm, the obtained partitioning and regression models from $DS_{2,1}$ were used to predict targets in $DS_{2,2}$. In Table 6 we show the R^2 of (a) a global regression model learned on the whole $DS_{2,1}$; (b) ideal 3 models assuming the positions of the 3 homogeneous subregions were known; and (c) the 3 partitions and 3 corresponding models obtained by the partitioning algorithm. The obtained R^2 results provide additional evidence for validity of the proposed algorithm.

To observe the influence of model complexity on partitioning, the proposed algorithm has been applied to the same data with linear models instead of neural networks. Due to highly nonlinear functions used for the target generation, linear models were not able to identify regions well. As expected, the competition of linear models resulted in partitioning the data set into more regions (seven). R^2 on training data increased from 0.32 for a global model to 0.68 for an ensemble of 7 local linear models. Although the improvement seems significant, the generalization properties as measured on the second data set $DS_{2,2}$ were poor and even worse than the global linear model (R^2 of 0.33 for a global model as compared to 0.06 using the obtained 7 subregions and their linear models). These results provide additional evidence that a regression model choice can be very important for the performance of the partitioning algorithm.

5.2 Experiments on Real-Life Spatial Data

A database containing a grid of 8 topographic attributes and winter wheat yield from a 220 ha field located near Pullman, WA was used for experiments on real-life data. The goal was to predict wheat yield as a function of terrain attributes derived from USGS 30 m DEM data mapped to a 10 x 10 m grid. The terrain attributes were: Compound Topographic Index; Aspect East-West (0 to 180°); Aspect North-South (0 to 180°); Distance to Flow Paths > 232 m; Flow Direction; Slope; Profile Curvature; and Average Upslope Slope. All features except for flow direction were continuous and non-Gaussian. The crop yield values were collected with a combine mounted yield monitor and a global positioning system. There were 24,592 patterns in the entire data set.

From the previous results (Vucetic et al., 1999), it was known that topographical features alone can explain just a small part of crop yield variance. A global neural network with 5 hidden nodes achieved R^2 of only 0.11 on test data. The spatial partitioning algorithm was applied in order to examine if the field was heterogeneous. If homogeneous regions existed, each of them could be treated differently. Using the advances in precision agriculture, this would allow farmers to apply different practices on each of the homogeneous regions to obtain larger profits as compared to treating the whole field uniformly.

From the correlogram of residuals of a global neural network achieving the range at about 150 meters, and observing the large unexplained variance in data, the size of MAF was set to 150 meters. The spatial partitioning algorithm discovered 3 homogeneous regions shown at Figure 6, with the intermediate results of the partitioning shown in Table 7. As could be seen, a significant improvement in test R^2 has been achieved by the partitioning, indicating the existence of homogeneous regions within the field. A brief analysis of discovered regions (Table 8) shows that the largest homogeneous region corresponds to the high yielding part of the field, while the smaller two regions correspond to the low yielding part. The mean and standard deviation of yield in different regions were normalized versus statistics for the entire field with the global mean and the global standard deviation mapped to 0±1.

Figure 6. A 220 ha wheat field near Pullman, WA. Partitioning of the wheat field into 3 subsets by the algorithm

An additional analysis aimed at extracting a set of fuzzy classification rules is performed to explain the difference between identified homogeneous regions. The resolution of rules depends on an overlap between the observed feature spaces of discovered regions. A smaller feature overlap among identified regions would result in higher resolution rules that might be used for transfer of knowledge to remote geographical locations. However, even if feature overlap is too high for extraction of higher resolution fuzzy rules, knowledge on homogeneous regions limited to a specific geographical site can be useful for improving treatment in the subsequent years.

Table 7. Intermediate results of partitioning the wheat field

Number of regions	**1**	**2**	**3**	**4**
R^2	0.11	0.25	0.31	0.32

In our case, the three identified regions from Figure 6 had high overlap in all observed features. Aspect North-South (Aspect NS) and Slope has the smallest averaged pairwise overlap of histograms (73% and 81%, respectively). The corresponding qualitative fuzzy rules distinguishing between the regions are shown in Table 8. It can be concluded that the partitioning has significantly improved the prediction despite the large overlap in the observed feature space.

Table 8. Analysis of discovered regions

Region	**Size**	**Yield**	**Aspect NS**	**Slope**
1	61%	High 0.4±0.7	Average	Slightly lower
2	21%	Low −0.4±0.8	Slightly South	Slightly higher
3	18%	Low −0.8±0.9	Slightly North	Moderately higher

6. Conclusions

When learning on heterogeneous data with unobserved or noisy features, bias of a global model on homogeneous subsets is typically large but learning an ensemble of local models on appropriately partitioned data could decrease it. A method for identifying such a partitioning for heterogeneous spatial data through a competition of local models is proposed. Assuming local models of appropriate complexity, this method is likely to reduce the bias. The convergence of the algorithm was proven for a mixture of two homogeneous regions, while the experiments indicated that the convergence is very fast. In the presence of unexplained variance in spatial data, incorporating error filtering to the competition procedure can significantly improve the partitioning accuracy. In our experiments, the absolute error filtering consistently outperformed the alternatives. An analysis of the obtained partitioning can result in discovering potentially useful knowledge, as demonstrated on real-life precision agriculture data.

Acknowledgements

We thank Tim Fiez for providing real-life data used in this study and to anonymous reviewers for the many useful comments on the first version of this paper.

References

Cressie, N.A.C. (1993), *Statistics for Spatial Data,* New York , John Wiley & Sons, Inc.

Jacobs, R.A., Jordan, M.I., & Barto, A.G. (1991), Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks, *Cognitive Science, 15*, 219-250.

Oliver, D.S. (1995), Moving averages for Gaussian simulation in two and three dimensions, *Mathematical Geology, 27,* 939-960.

Pokrajac, D., Fiez, T. & Obradovic, Z. (2000), A tool for controlled knowledge discovery in spatial domains," *Proceedings of the International European Simulation Multiconference*, Gent, Belgium, in press.

Vucetic, S., Fiez, T., & Obradovic, Z. (1999), A data partitioning scheme for spatial regression, *Proceedings of the IEEE/INNS International Conference on Neural Networks*, No. 348, session 8.1A.