Spatial Interpolation using Multiple Regression

Orlando Ohashi
LIAAD - INESC TEC / DCC
Faculdade de Ciências - Universidade do Porto
Porto, Portugal
ohashijr@dcc.fc.up.pt

Luís Torgo
LIAAD - INESC TEC / DCC
Faculdade de Ciências - Universidade do Porto
Porto, Portugal
ltorgo@dcc.fc.up.pt

***Abstract*—Many real world data mining applications involve analyzing geo-referenced data. Frequently, this type of data sets are incomplete in the sense that not all geographical coordinates have measured values of the variable(s) of interest. This incompleteness may be caused by poor data collection, measurement errors, costs management and many other factors. These missing values may cause several difficulties in many applications. Spatial imputation/interpolation methods try to fill in these unknown values in geo-referenced data sets. In this paper we propose a new spatial imputation method based on machine learning algorithms and a series of data pre-processing steps. The key distinguishing factor of this method is allowing the use of data from faraway regions, contrary to the state of the art on spatial data mining. Images (e.g. from a satellite or video surveillance cameras) may also suffer from this incompleteness where some pixels are missing, which again may be caused by many factors. An image can be seen as a spatial data set in a Cartesian coordinates system, where each pixel (location) registers some value (e.g. degree of gray on a black and white image). Being able to recover the original image from a partial or incomplete version of the reality is a key application in many domains (e.g. surveillance, security, etc.). In this paper we evaluate our general methodology for spatial interpolation on this type of problems. Namely, we check the ability of our method to fill in unknown pixels on several images. We compare it to state of the art methods and provide strong experimental evidence of the advantages of our proposal.**

***Keywords*-spatial prediction; data pre-processing;**

I. Introduction

In spatial data analysis the data is frequently obtained from measurements of real systems, e.g. wind speed, oil resources analysis, water quality assessment, satellite images, pictures and/or paintings repair, etc. This process of data collection is not fully controllable and it is prone to failures. These failures can lead to missing values on the collected data sets, which in turn may have a serious impact on the posterior analysis. Other constraints, e.g. financial and human resources, may even increase the amount of missing data. In this context, it is of key importance to have methods that help in trying to fill in these gaps, which is confirmed by the amount of literature and methods available for spatial interpolation (see [1] for an overview).

The main idea behind any approach to spatial imputation is the assumption that the value at any location has some form of dependence on the values on neighboring locations. This is supported by the first law of the geography that says that "everything is related to everything else, but near things are more related than distant things" [2]. Our work is also based on this assumption. However, the fundamental difference of our proposal when compared to the state of the art is the fact that *we also allow the use of data from faraway regions provided these neighborhoods have similar spatial dynamics to the target location* for which we want to fill in a value. The fact that two neighborhoods are distant from each other does not preclude them from having similar spatial behavior. Ignoring these similarities and data seems a waste. The main contribution of our work is to allow for the use of this extra data from distant apart regions. This means that our methods will tend to use much more data than existing methods to estimate the unknown values. The hypothesis driving our work is that this extra information will lead to gains in terms of the precision of the imputation.

We have tested and compared our proposal against a series of alternative state of the art methods on a particular task - filling in the missing pixels on pictures. Still, the approach is by no means restricted to this particular application and can actually be seen as a general approach to spatial imputation and even to the problem of formalizing prediction tasks in the context of spatial data. Our experiments show that our approach significantly outperforms the most common techniques used for spatial interpolation (IDW and Kriging according to [1]). In the research area of image processing other approaches exist to this problem. One of the most used approaches is the Inpaint technique [3], [4]. We have also compared our approach against the Inpaint method and the results show a clear advantage of our technique.

II. Spatial Interpolation - an overview of the state of the art

Forecasting the missing values in spatial data sets is not a new problem and it is usually known as spatial imputation or interpolation. Spatial interpolation methods address the problem of estimating unknown values of a variable of interest, Z, on certain geographical locations, based on a spatial data set $\mathcal{Z} = \{z_1, \cdots, z_n\}$, where z_i is the value of the variable Z at location i.

Many different approaches have been applied to solve this problem. Existing approaches are usually motivated by the first law of the geography [2] that prescribes that nearby points should have strongly correlated values. Li [1] classifies the approaches in three main classes: non-geostatistical interpolators, geostatistical interpolators and combined procedures that integrate approaches from the two former classes.

Non-geostatistical interpolators are based on the distance between the neighbors. The simplest method is the Distance Interpolator (DI) that consists on the use of the average value of the spatial neighbors as an approximation to the value at the missing location, where the neighborhood of point o is defined as $\mathcal{N}_o^\beta = \{z_i \in \mathcal{Z} : d(o,i) < \beta\}$.

The Inverse Distance Weighted Interpolation (IDW) [5] is a simple improvement of the DI method. It is based on the assumption that the values that are farther apart within the neighborhood of a point should contribute less to the average calculation. In this context, this method approximates the value at an unknown location as the weighted average of the known neighborhood values, with weights inversely proportional to the distance from the target location.

The second class of existing methods are geostatistical interpolators that have origin in the work of Krige [6]. Kriging is a generic name for a family of generalized spatial interpolation models. According to Mitas [7] kriging assumes that the spatial distribution of a geographical region can be modeled by the realization of a random function, using a statistical technique to analyze the data. Kriging uses the same basic principal behind the inverse distance weighting technique - it approximates the unknown value at a location by interpolating the values at known locations given more importance to the closer neighbors. However, the way the weights are calculated is different as kriging uses the covariation between known data at various spatial locations [6]. There are several variants of kriging, most of which differ on the way these weights are approximated. Frequently used variants include ordinary kriging and co-kriging. In this paper we have only considered ordinary kriging because co-kriging requires an auxiliary variable (covariable) [5], which was not available in the domain considered in this paper.

III. Our Proposal

Spatial interpolation aims at filling in the values of a variable of interest at geographical locations for which they are unknown. This problem is usually solved by assuming that the unknown values can be filled in by using the information of the known values in their vicinity. It is possible to look at this task as a prediction problem where the target variable is the variable of interest at a certain geographical location and the predictors are the values of this variable within the respective neighborhood. We have taken this approach, by mapping the spatial interpolation problem into a multiple regression problem.

Other authors have addressed the use of regression tools with spatial data (e.g. [8]). Still, to the best of our knowledge all these works constrained the use of data to make predictions for a certain location to the neighboring data (e.g. through kernels [8]). In our approach we do not impose this constraint. We let the regression methods decide which observations should be used for a certain prediction. Depending on the used tools this may lead to data from different locations being used to make the forecasts. This will happen if the used tools find these data to be similar in terms of the predictor variables. With the goal of helping the models to find neighborhoods with similar spatial dynamics we propose the use of a series of *spatial indicators* as predictors.

Summarizing, our proposal for the spatial interpolation problem consists of two key ideas:

- Mapping the spatial interpolation problem into a multiple regression task;
- Propose a series of spatial indicators to better describe the spatial dynamics of the variable of interest.

The first idea has two main advantages: (i) allowing the use of the large number of sophisticated function approximators that are available; and (ii) allowing the use of data from faraway neighborhoods if the models find them similar to the region being interpolated in terms of the predictor variables. Regards the second idea, we have considered three classes of properties to describe the spatial dynamics between the variable values in a neighborhood: i) properties describing the typical value of the target variable; ii) properties describing the variability of the variable; and iii) properties describing the tendency (in spatial terms) of the variable. Among these, the third class is the one that differentiates more our work from the information used in standard approaches to spatial interpolation. Still, we should remark that standard approaches use these indicators for directly forecasting the unknown values, while we are using them as predictors in a regression model, thus allowing for the discovery of possible interactions between the properties.

The typical value of the target variable within a neighborhood can be captured by both the Distance Interpolator (DI) and the Inverse Distance Weighted Interpolation (IDW), the difference being that the latter weights the contribution of the neighbors by the distance to the target. In this context, we will use these values as predictors in our models. To simplify our notation we will use $\overline{z}(\mathcal{N}_o^\beta)$ for the standard averages ($= DI$), and $\widetilde{z}(\mathcal{N}_o^\beta)$ for the weighted averages ($= IDW$).

To capture the notion of spread of the values within a certain vicinity we have used the standard deviation calculated with the values in this neighborhood,

$$\sigma_z(\mathcal{N}_o^\beta) = \sqrt{\frac{1}{|\mathcal{N}_o^\beta|}\sum_{z_i \in \mathcal{N}_o^\beta}(z_i - \overline{z}(\mathcal{N}_o^\beta))^2} \quad (1)$$

In financial forecasting it is common to describe the tendency of a price time series by means of a ratio between two moving averages calculated using two different time spans. If the value of the moving average with shorter time span surpasses the longer moving average we know that the time series is on an upwards tendency, while the opposite indicates a downwards direction. We have imported this idea into the spatial dimension. The ratio between two averages calculated on two spatial neighborhoods with different sizes around the target location provides us with information on how the target variable values evolve in the vicinity of this location. If the shorter average is above the longer, then we know that values are increasing as we approach the target location, while the opposite occurs if the shorter average is smaller. This ratio can be defined as follows,

$$\overline{Z}_o^{\beta_1,\beta_2} = \frac{\overline{z}(\mathcal{N}_o^{\beta_1})}{\overline{z}(\mathcal{N}_o^{\beta_2})} \quad (2)$$

where β_1 and β_2 are two neighborhood sizes ($\beta_1 < \beta_2$) and $\overline{z}()$ is the average of a set of points in the neighborhood of o.

A variation of this indicator can be easily obtained by using weighted averages of the values within the spatial neighborhood,

$$\widetilde{Z}_o^{\beta_1,\beta_2} = \frac{\widetilde{z}(\mathcal{N}_o^{\beta_1})}{\widetilde{z}(\mathcal{N}_o^{\beta_2})} \quad (3)$$

where $\tilde{z}()$ is the weighted average of a set of points in the neighborhood of o.

Having defined a series of spatial indicators, we can proceed to map the spatial interpolation problem into a multiple regression task. The target variable of this task is the value of the variable Z at a geographical location. As predictors we propose to use several variants of the spatial indicators we have described above. Namely, we will estimate the value of Z at a target location o as a function of the following predictors,

$$\begin{aligned}\hat{z}_o = f(&\overline{z}(\mathcal{N}_o^{k_1}), \overline{z}(\mathcal{N}_o^{k_2}), \overline{z}(\mathcal{N}_o^{k_3}), \overline{Z}_o^{k_1,k_2}, \overline{Z}_o^{k_2,k_3},\\ &\widetilde{z}(\mathcal{N}_o^{k_1}), \widetilde{z}(\mathcal{N}_o^{k_2}), \widetilde{z}(\mathcal{N}_o^{k_3}), \widetilde{Z}_o^{k_1,k_2}, \widetilde{Z}_o^{k_2,k_3},\\ &\sigma_z(\mathcal{N}_o^{k_1}), \sigma_z(\mathcal{N}_o^{k_2}), \sigma_z(\mathcal{N}_o^{k_3}))\end{aligned} \quad (4)$$

where $f()$ is the unknown regression function we are trying to model using a set of training data $\mathcal{Z}$, and k_1, k_2 and k_3 (with $k_1 < k_2 < k_3$) are 3 spatial neighborhood sizes. In the experiments described in this paper we have used the values $10, 30$ and 50, respectively, for these spatial neighborhood sizes.

It is important to remark that several other indicators/predictors could have been used. The same can be said regards the sizes of the spatial neighborhoods. Which predictors to use is a well-studied problem on predictive data mining. Several established methods exist to search and select the best predictors for a given data set and learning algorithm. It is not the goal of this paper to address this well-studied subject. Our contribution is the idea of mapping the problem of spatial interpolation into a multiple regression task and also to provide some new predictors that capture the spatial dynamics on a certain vicinity.

In summary, our proposal for the spatial imputation problem using a spatial data set $\mathcal{Z}$ consists on: (i) use these data to build a new multiple regression data set where the target variable is the value of Z on a location and the predictor variables are calculated using the values in the vicinity of this location (an example are the variables mentioned in Equation 4 but others could be used); (ii) use this new data set to build a regression model with some existing algorithm; and (iii) apply this model to locations where the value of the target variable is unknown.

IV. A Concrete Application - image inpainting

This section describes a real world application of our proposed methodology. The application consists on filling in missing pixels on a image. An image can be seen as a spatial data set, given that each pixel has a different location in a system of Cartesian coordinates. At each location one or more values may be measured. In our problem it is a single degree of gray (a value in the interval $[0, 255]$) that is measured. In the research area of image processing this type of problems are referred to as "image inpainting" [3], [4]. The term "inpainting" has its origin in the manual task of restoring damaged paintings and/or photos by professional restorers [3]. Digital inpainting is a relatively new research area with the goal of developing tools that automatically restore damaged images. Examples of damages include: noise (missing pixels caused by some equipment failure), unwanted objects (persons, cars, red-eye, etc.), logos, stamps, scratches (old pictures), etc.

Since the target application of this work is the repairing of images, we have also compared our approach against one successful implementation of an image inpainting algorithm based on the "exemplar based approach" [3], [4]. In our experiments we have used an open source implementation[1] of this inpainting algorithm.

Figure 1 shows the two original images that were selected to evaluate and compare our proposal. The first picture (Figure 1a) is a dog face and the second picture (Figure 1b) is the Coliseum of Rome. Based on these images we will generate several data sets with an increasing number of the original pixels removed.

[1]Publicly available at http://sourceforge.net/projects/imageinpainting

(a) Dog face (260x222) (b) Coliseum (320x240)

Figure 1: Original pictures.

V. Experimental Evaluation

The main goal of our experiments is to check the validity of our proposal for spatial interpolation. We have carried out an extensive set of experiments under different conditions. All data, code and extra results not shown due to space restrictions are provided in a web page[2] to ensure that our work is replicable.

A. Experimental methodology

We have considered several setups in terms of the amount of missing pixels. Namely, we have created 9 different data sets from each original image (Figure 1) with an increasing number of pixels being randomly removed[3]: 10%, 20%, $\cdots$, 90%. Moreover, to ensure the statistical significance of the results we have repeated this random selection 10 times for each of the 9 settings. This means that we have compared the models on 180 different data sets generated from the two original images.

For each data set the models were given the available pixel data and asked to forecast the value of the target variable (degree of gray) at the missing pixel locations. The predictions were compared against the true values (Figure 1) using the mean absolute error,

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{z}_i - z_i| \quad (5)$$

where, n is the number of missing pixels, $\hat{z}_i$ is the level of gray predicted by the models, and z_i is the real value according to the pictures in Figure 1.

B. Models

Our methodology is based on the use of a regression algorithm to obtain the models that are then used to carry out the spatial imputation of unknown values. In order to fully test our ideas we have selected a diverse range of modeling approaches to confirm its validity independently of the technique used to forecast. We have used the following regression algorithms:

[2]http://www.dcc.fc.up.pt/~ltorgo/ICDM12

[3]We should remark that the values at these locations were actually removed, i.e. set as unknown, and not set as white pixels as the graphical representations of the data sets we will see later, may suggest.

Regression Trees (RT) - a regression tree based on the R package `rpart`. In our experiments we have used the function `rpartXse` provided in package `DMwR` [9] and have tried 4 different variants by using the parameter `se` that controls the level of pruning with values: 0, 0.5, 1 and 1.5.

Support Vector Machines (SVM) - an implementation of SVMs available in the R package `e1071`. We have used 6 variants of the parameters *cost* and *gamma*. For the parameter *cost* we used the values: 1, 10, 100 and for the parameter *gamma* the values: 0.1 and 0.5.

Random Forest (RF) - an implementation of random forests available in the R package `randomForest`. We have used 3 variants of the parameter *ntree* with the values: 500, 1000 and 1500.

These multiple regression algorithms were applied to data sets obtained using the setup of Equation 4 (cf. Section III).

Regards the competitive approaches for spatial imputation we have selected a series of techniques that are a good representation of the state of the art on this area:

Distance Interpolator (DI) - a simple baseline method that uses the mean value of a circular neighborhood region. We have considered 3 neighborhood sizes: 10, 30 and 50.

Inverse Distance Weighted Interpolator (IDW) - a variation of the previous method that uses the weighted average value within the neighborhood region as the approximation for the unknown location. The weights are inversely proportional to the distance. We have considered the same neighborhood sizes as in DI.

Ordinary Kriging (OK) - we have used an implementation of this method available on the R package `automap`. The implementation in this package automatically selects the best parameters for the kriging method, including the neighborhood size and the function used in the calculation of the semivariograms (it considers spherical, exponential, Gaussian and two variants of the Matern family). To limit the search space, in our experiments we have set the maximum neighborhood size to 90.

All the used tools are freely available in the R software environment [10], ensuring easy replication of our work.

C. Results

Figure 2 summarizes the results obtained by the alternative models on the two pictures using the experimental settings described before. Each bar represents the estimated MAE value averaged over the 10 repetitions of the best model variant on the data sets DS_{30} and DS_{70}. These are the original pictures with 30% and 70% of the pixels removed, respectively. The first two bars present the results of the distance interpolator (DI) approach, using the smallest neighborhood size. The second group of bars show the results of the IDW technique using the same spatial neighborhood size. Then we have the best variant of regression trees (`se` = 0), the best SVM (`cost` = 100 and

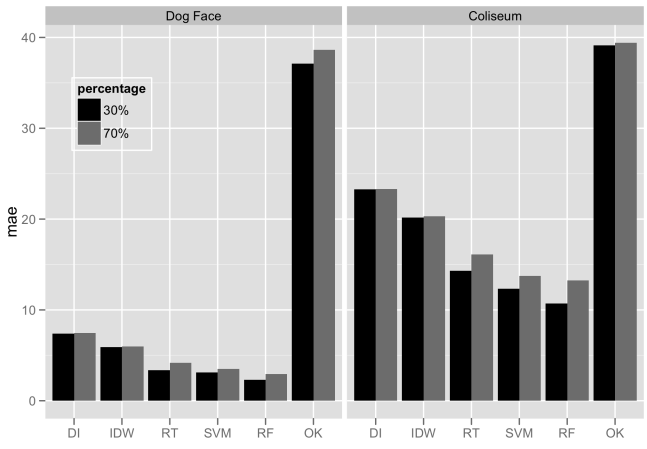


Figure 2: Estimated MAE of the different approaches.

gamma = 0.1) and random forest (ntree = 1000). The last two bars show the results of the ordinary kriging approach (OK), whose parameters are automatically tuned by the used software package.

The results of Figure 2 (together with the ones available at the accompanying web page) show an overwhelming advantage of our approaches when compared to these state of the art methods. In particular, both the SVM and RF variants achieve remarkably good scores, although even with the simple RT approach the results are superior. These experiments provide clear evidence of the advantage of: (i) using more sophisticated function approximators; (ii) using more elaborated information concerning the spatial dynamics through spatial indicators; and (iii) allowing the use of data from distance points in space provided the regression models find this useful in terms of accuracy. Another noticeable observation is the surprisingly bad scores obtained by the used ordinary kriging method, which was unable to beat even the simple DI variants. This may indicate that the automatic tuning provided by the used software package may not be adequate for all situations and that these particular problems could require a more careful hand-tuning of the kriging parameters. Our approach, however, did not require any tuning at all, and it may even be the case that with different variants of our spatial indicators, for instance through the use of some feature selection algorithm, the performance could be further improved.

In order to better understand what the methods are doing in terms of approximating the original image, we have selected one of the ten repetitions for the two image variations DS_{30} and DS_{70}, and represented graphically both the original data and the approximations provided by the competing methods. These results are show on Figure 3. The first row of graphs shows the two original data sets (Dog Face and Coliseum) with 30% and 70% pixels removed, respectively. The remaining rows show the approximations

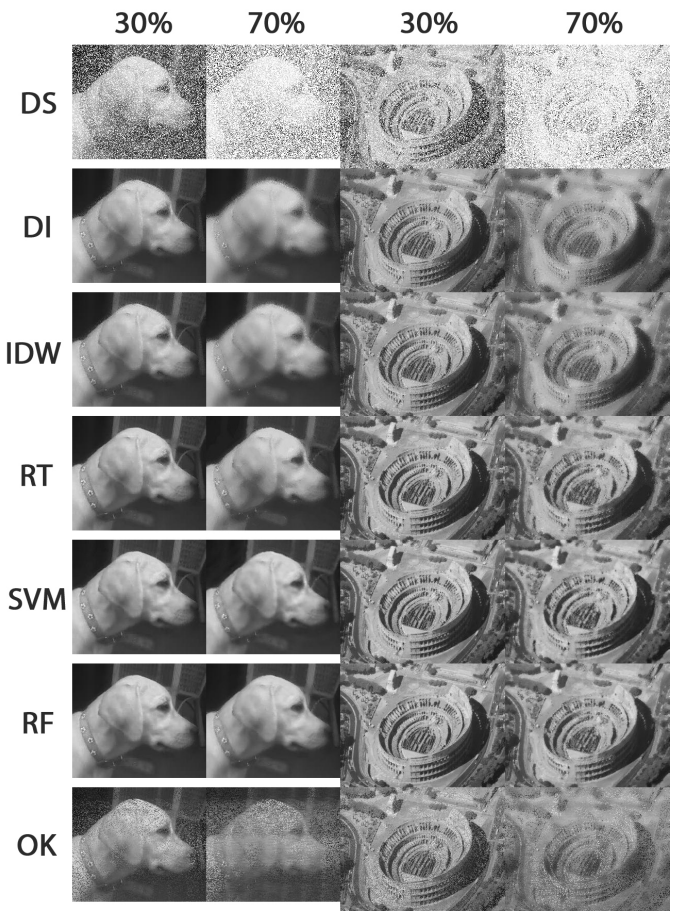
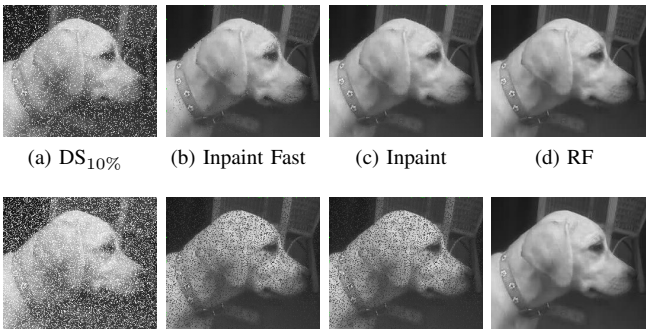


Figure 3: Estimated MAE of the different approaches for the Figure 1.

provided by each of the alternative approaches. The figure shows the best variants of each alternative, namely : i) DI and IDW with the neighborhood size of 10; ii) regression trees with $se = 0$; iii) the SVM model with $cost = 100$ and $gamma = 0.1$; and iv) and the random forest (RF) model with $ntree = 1000$. These graphs illustrate the remarkable job that our approaches are able to achieve in terms of recovering the original image, even at very high levels of noise. The quality of the pixel imputation even with 70% of the pixels removed is impressive.

As mentioned before, the problem we are addressing is named image inpainting within the image processing research area. We have compared our best variant (RF $ntree = 1000$) to one of the most common methods in image inpainting (see Section IV). We were not able to compare these two techniques on the 9 data sets with increasing percentage of removed pixels from Figure 1a, because the used inpaiting software crashed on data sets with too many unknown pixels. In this context, we were only able to collect results for the $DS_{10\%}$ and $DS_{20\%}$ data sets.



(a) $DS_{10\%}$ (b) Inpaint Fast (c) Inpaint (d) RF

(e) $DS_{20\%}$ (f) Inpaint Fast (g) Inpaint (h) RF

Figure 4: Random Forest vs Inpaint Technique

Figure 4 shows the results of this comparison. We show: i) the original data sets; ii) the approximations provided by two variants of the inpainting algorithm[4] - the fast implementation (Figures 4b and 4f) and the standard implementation (Figures 4c and 4g); and iii) the results of the random forest in Figures 4d and 4h. Although the inpainting algorithm is able to achieve similar results on the data set with a lower level of unknowns (particularly in the standard implementation), in the data set with 20% of removed pixels we already see a marked advantage of our approach.

VI. Conclusions

This paper describes a novel approach to the problem of spatial interpolation. Our general methodology is based on the idea of transforming this problem into a multiple regression task and then applying standard algorithms to a data set that is constructed from the original spatial data using a series of spatial indicators designed to better describe the spatial dynamics of the variable of interest. The key distinctive feature of this methodology is the data that is used to obtain the approximations of the unknown values of the variable of interest. Existing state of the art methods use only values within a certain neighborhood of the target location for which we want an estimate. Our proposal is based on the assumption that other distant vicinities may be used provided they show a similar spatial correlation pattern. The decision to use this extra data is left to the the optimization process of the regression models. With the goal of improving the discovery of similar neighborhoods we have also introduced the notion of spatial indicators. These are features constructed from the original data that try to provide useful information on the spatial correlation dynamics within a neighborhood. Their goal is to help the models in uncovering similarities among different regions.

Although the described methodology is a general spatial imputation method, in this paper we have tested it on a particular task with strong impact in several application domains: image repairing. We have tested and compared our method under different setups in terms of missing information on the given images. On all setups we have observed a strong advantage of our approach that has achieved impressive results in terms of recovering an image even at high levels of noise. These initial results are very encouraging and provide strong empirical evidence towards the advantages of our approach to spatial imputation.

[4]To apply the inpaint software to our data set variants we needed to convert the missing pixels to the RGB green color.

Acknowledgment

This work is partially funded by the ERDF - European Regional Development Fund through the COMPETE Programme, by the Portuguese Funds through the FCT within project FCOMP - 01-0124-FEDER-022701, and by a FCT PhD grant (SFRH/BD/61795/2009) to Orlando Ohashi.

References

[1] J. Li and A. Heap, "A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors," *Ecological Informatics*, 2010.

[2] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, 1970.

[3] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *27th Conf. on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 417–424.

[4] A. Agrawal, P. Goyal, and S. Diwakar, "Fast and enhanced algorithm for exemplar based image inpainting," in *Image and Video Technology (PSIVT), 2010 Fourth Pacific-Rim Symposium on*. IEEE, 2010, pp. 325–330.

[5] E. Isaaks and R. Srivastava, *Applied geostatistics*. Oxford University Press New York, 1989, vol. 2.

[6] D. Krige, *A statistical approach to some mine valuation and allied problems on the Witwatersrand*. Univ. of the Witwatersrand, 1951.

[7] L. Mitas and H. Mitasova, "Spatial interpolation," *Geographical Information Systems: Principles, Techniques, Management and Applications, Wiley*, vol. 481, 1999.

[8] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically weighted regression-modelling spatial non-stationarity," *Journal of the Royal Statistical Society. Series D*, pp. 431–443, 1998.

[9] L. Torgo, *Data Mining with R, learning with case studies*. CRC Press, 2010.

[10] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2010.