A COMPARISON OF DATA MINING AND SPATIAL TECHNIQUES: AN APPLICATION TO PROPERTY DATA.

T.J. Hine
University of Pretoria
and
S.M. Millard
University of Pretoria
and
F.H.J. Kanfer
University of Pretoria

Key words: Spatial Statistics, Data-Mining, Stacking, Property.

Summary: The improvement of data management and data capturing techniques has led to the availability of large amounts of data for analysis. This is especially true in the field of spatial data, where data is indexed by location. Traditionally, spatially correlated data has been analysed using methods that rely on the spatial component of the data. This article will compare the results of using traditional spatial methods such as Kriging and geographically weighted regression against the use of other statistical data mining methods, given the large amount of data available. Using a dataset containing property values for the Tshwane Metropolitan area, different spatial and statistical models will be applied for predictive purposes in order to determine which model represents the data most accurately. Finally, these methods will be combined using stacking, to determine whether the combination of models has better predictive abilities than the single models.

1. Introduction

The improvement of data capturing and storing techniques has led to an increase in the amount of information available for analysis. This improvement has also been seen in the field of spatial data, leading to high dimensional datasets with many observations that are indexed by location. In this paper, methods that have been used in the analysis of large amounts of data will be compared to spatial statistical methods in which the spatial aspect of an observation is taken as the most important variable. This paper will also consider combining the two approaches using the method of stacking, which results in a model that performs relatively well in all regions.

2. Dataset

The data that will be used in this application is property valuation data of the Tshwane Metropolitan. The dataset has approximately 400 000 observations (after cleaning the dataset) and has variables regarding the value of the property, size of the property, the land rights, land use, whether the owner is private and whether the property can be charged rates. The dependent variable is property value. In order to account for the high variability in the scale, the dependent variable was transformed into the log of rand per area (calculated by dividing the property value by the size of the property). The models will be used to predict property value, and will be compared on the basis of how well they predict the property value.

3. Data Mining Models

Three statistical models commonly used in data mining were considered, Classification and Regression Trees (CART), Artificial Neural Networks (ANN) and Multivariate Adaptive Regression Splines (MARS).

3.1. Artificial Neural Networks

An artificial neural network is comprised of many interconnected neurons. These neurons simultaneously receive inputs and evaluate their outputs. In the first "training" phase the weights are determined using a subset of the data, and in the second "recognition" phase these weights are tested on another subset. The network is evaluated based on the sum of squared error between the output and true property value. ANN are most commonly used in applications involving speech or image recognition, classification and prediction problems. This method is also used to model high dimensional time series data (Gardner and Dorling, 1998). In this case, due to the high number of observations, a sample of size 30 000 from the dataset was used to build the network. The network had 30 hidden layers, in order to account for the number of observations and variables. When this model was used to classify the entire dataset, an R^2 of 0.7264 was achieved.

3.2. Classification and Regression Trees

Classification and regression trees (CART) refers to a method of classification commonly used for high dimension data. This method is particularly relevant in situations where the data is of mixed type (both qualitative and quantitative) or does not conform to usual assumptions. CART is a popular method in practice due to its lack of prior assumptions regarding the data. This method differs from cluster analysis in that it allows for a quantitative response variable (Breiman, Friedman, Olshen and Stone, 1984). The aim of CART is to classify the data into J mutually exclusive classes, contained in the set C in terms of a dependent variable. The classification is based on a set of predictor variables and is done in such a way that within group variance $\sum_{classes} N_c V_c$ is minimised, where N_c is the number of observations in class c and V_c is the variance of class c. The process begins with all observations in a single group and then splits this group into classes in a stepwise fashion, based on the predictor variables, in order to minimise this variance (Shalizi, 2009). The classification and regression tree identified the binary variable separating development types (sectional schemes against single properties) as the most important discriminator. The remainder of the discriminators were location based. This model resulted in an R^2 of 0.7627.

3.3. Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines, or MARS, refers to a regression method that is used as a solution to nonlinearity (Izenman, 2008). It is often applied in situations when there are many variables for a large number of observations (Izenman, 2008). The model involves recursive partitioning, separating the entire dataset into smaller segments to which regression models (splines) can be fitted. The model itself is a weighted sum of basis functions, which in turn are comprised of the products between the splines. The basis functions to be included are determined in a stepwise fashion, in which terms are added to the model in order to reduce the sum of squares. This results in a highly flexible model. It can also be considered as an expansion of the piecewise linear regression method (Leathwick, Rowe, Richardson, Elith and Hastie, 2005).

The MARS model used in this case was limited to second degree interaction (that is, it limited the basis function to be comprised of the product of no more than 2 splines). Again, the binary variable indicating development type was the most significant variable, followed by a binary variable indicating whether the property had residential rights, and finally the location of the property. The MARS model performed the best of the statistical methods, resulting in an R^2 of 0.821.

4. Spatial Models

There were two spatial methods considered in this study, namely geographically weighted regression (GWR) and Kriging.

4.1. Geographically Weighted Regression

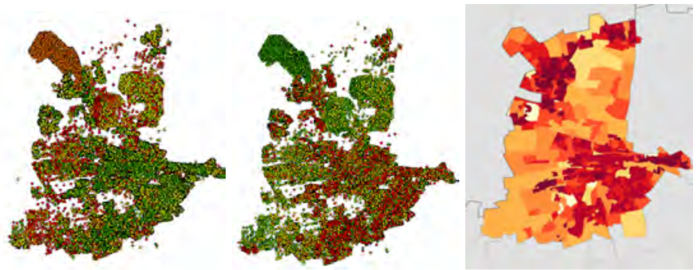
As a special form of weighted regression, geographically weighted regression is commonly used to deal with serial correlation that arises when the assumption of stationarity in the data does not hold. The stationarity assumption implies that if the explanatory variables are the same for any observations, the response will also be the same. Models in which the stationarity assumption is valid are often termed global models, as the model can be applied over the entire dataset. If these global models are fitted to datasets in which stationarity is not present (that is, in datasets where the relationships between observations cannot be expected to be homogenous over the entire dataset), the result is serial correlation – as a result of the spatial dependence between observations (Charlton and Fotheringham, 2009).

Geographically weighted regression is a specific form of weighted regression, in which the weighting matrix is based on the distance of the observations from one another. The weights are generally Gaussian based and the parameters may be estimated using the formula:

$$\hat{\mathbf{B}}(\mathbf{u}) = (\mathbf{X}^{\mathrm{T}}\mathbf{W}(\mathbf{u})\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}(\mathbf{u})\mathbf{Y}$$

where $\mathbf{W}(\mathbf{u})$ represents the weights matrix at location $\mathbf{u}$. Thus parameters are estimated for each specific location $\mathbf{u}$.

This flexible model allows the relationships between variables to change based on location – thus allowing for a spatial relationship in which observations that are close to each other are more



(a) ANN Errors (b) Kriging Errors (c) Population Density

influential than observations which are far apart (Charlton and Fotheringham, 2009). When all observations are included in the geographically weighted regression model, the model becomes very computationally expensive, due to the construction of the weights matrix. When a subset of the data is used, the model performs relatively well with an R^2 of 0.77.

4.2. Kriging

Kriging is one of the most popular methods for analysing spatially correlated data. This method is appropriate for Geostatistical data or discrete point data as it essentially considers the impact of an observation on surrounding observations. It is also a form of weighted regression, which uses the spatial covariance as weights. Thus it essentially weights observations based on their relationship to neighboring observations. This method could also be considered as a data driven method, since the weights are based on observed values (Bohling, 2005).

In order to determine which Kriging model should be used, the semi-variogram (which measures he correlation between 2 points as a function of distance) was drawn and examined. The data was most appropriately modeled by universal Kriging (that is, a second order trend is removed and Kriging is performed on the residuals) using an exponential model of the form $\gamma(h)$=$c(1-e^{\frac{-h}{a}})$, where h is the the distance between 2 points, a is the maximum range of influence from a point and c is the value of the semi-variogram function at distance a (Oliver and Webster, 1990) . The Kriging model performed the best of all models, with an R^2 of 0.8.

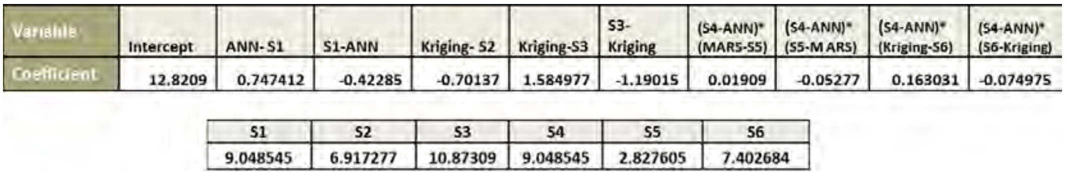
5. Analysis Results

The fact that the Kriging model resulted in the highest R^2, when Kriging uses only the location of the observation in the model and does not consider any other variables, is testament to the importance of spatial correlation in this case. Even in the statistical models, the location variable was found to be one of the most significant variables. Upon examining the spatial representation of the error terms in each model, a pattern is evident. The error maps for the CART and MARS method were not included, however they also performed better in regions with higher population density. The map for GWR is not shown since the method did not use the entire dataset and thus is not comparable.

The figures show the difference between the regions that are well predicted using spatial methods and those using data mining methods. Specifically, spatial methods (Kriging- shown in 1(b)) perform well in regions with lower population density (1(c)), that is rural and farming regions, where the neighbouring properties are all relatively similar. Conversely, the data mining models perform best in regions of high disparity (ANN error model shown in 1(a)), with many different types of properties in one area. Since the data mining models make use of variables other than the spatial location, they are more successful in these areas.

6. Combinining the Models

In order to create a model that could model both types of regions relatively well, stacking was considered as a method to combine spatial and statistical models (Breiman, 1996). The method of stacking involves 2 levels, namely level 0 and level 1. In level 0, a training dataset is used to fit the methods in question to the data. Predictions are calculated for each of the methods and these predictions are then used as inputs into a new model. In this case the methods of Kriging, MARS and ANN will be combined. There were different areas in which each of these methods excelled (although MARS and ANN both performed well in regions with high disparities, there were some differences in between the models within those regions), and so the combination of the methods should provide a model with relatively good predictions for each region. These predictions will then be combined into a new MARS model. CART predictions were not included since they performed well in the same regions as MARS and ANN and GWR was not included since the model was not built on the entire dataset. Although one of the major concerns of stacking is the presence of multicollinearity (since all predicted values are trying to predict the same output) this may not be as much of a problem in this analysis, since there are specific regions that are well predicted by one model but not by the other model. Thus the observations with good predictions for one model will differ from the observations with good predictions for another model. In order to test for multicollinearity, the variance inflation factors (VIF) were calculated for all inputs. As a rule of thumb, a VIF of greater than 10 is an indicator of a high degree of multicollinearity (Gujarati and Porter, 2009) . However, VIF values of the inputs were not high enough to indicate multicollinearity. This may be due to the fact that all of these methods complement each other, and do not perform well in the same regions. The final stacked model is shown in Table 1.

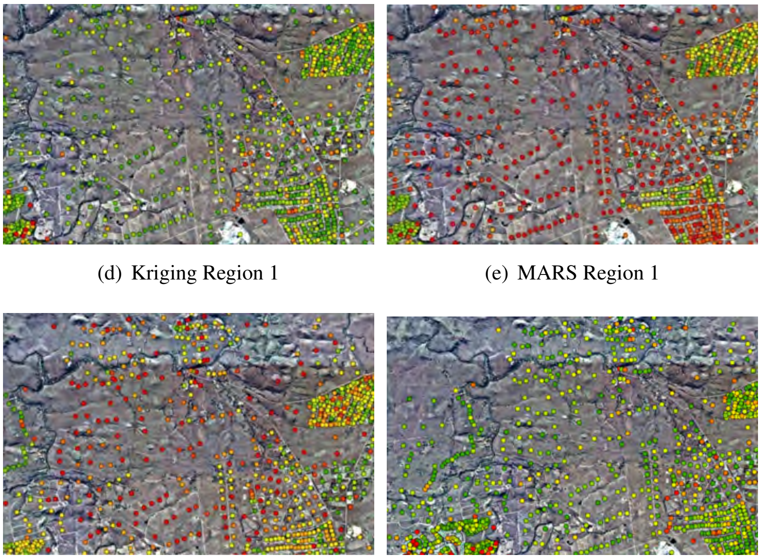


Variable	Intercept	ANN- S1	S1-ANN	Kriging- S2	Kriging-S3	S3-Kriging	(S4-ANN)*(MARS-S5)	(S4-ANN)*(S5-MARS)	(S4-ANN)*(Kriging-S6)	(S4-ANN)*(S6-Kriging)
Coefficient	12.8209	0.747412	-0.42285	-0.70137	1.584977	-1.19015	0.01909	-0.05277	0.163031	-0.074975

S1	S2	S3	S4	S5	S6
9.048545	6.917277	10.87309	9.048545	2.827605	7.402684

Table 1

The interpretation of the coefficients is difficult (which is one of the disadvantages of the stacking method) since the input variables are in fact outputs of other models. MARS appears to be the least important of the predictors and features only once in an interaction term. The results of the analysis are promising. When the entire dataset is used in the analysis, the R^2 is 0.90122. Under cross validation, the model results in an R^2 of 0.899. Thus even the testing set provides a higher R^2 than the best performing single model (Kriging, with an R^2 of 0.88). Although the results of the analysis are very positive, more research into the method of stacking is required. The cross validation results shows that the model can be considered as a more "general" model for the region of Tshwane, which is a good result. Analysis has also been done on the residuals of the model, and it was found that there is no serial correlation and they are uncorrelated with the input values. Aerial photographs may be used to compare the results of the MARS model, ANN mode, Kriging model and the stacked model. The errors for each model have been plotted over the aerial photograph for various regions. The same legend has been used for all models. Again, a green dot indicates the prediction error is low and a red dot indicates that the prediction error is high.



(d) Kriging Region 1

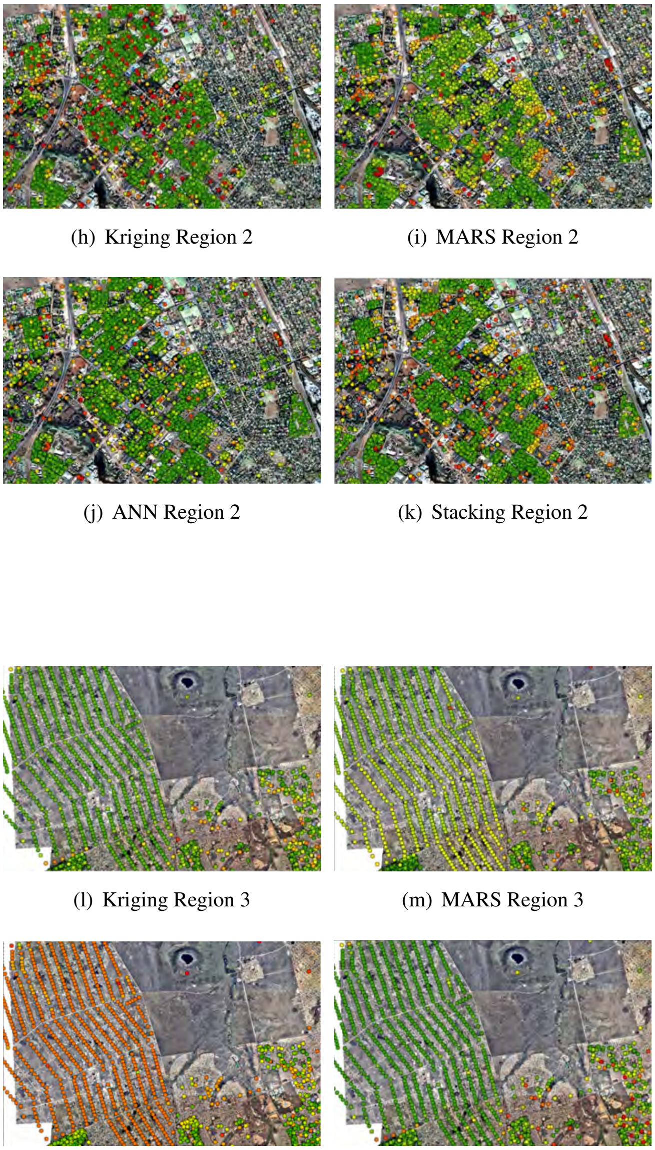
(e) MARS Region 1

(f) ANN Region 1

(g) Stacking Region 1

The first region has a relatively small amount of development. When the Kriging model is applied (a), it predicts the value of properties very well. However the MARS (b) and ANN (c) models predict the values very poorly. The stacked model (d) is able to perform as well as Kriging in this case.

The second region shows higher disparities between properties. Although the Kriging model (e) does not give very bad predictions, there is a set of properties within the neighbourhoods that has been poorly predicted. Comparatively, the MARS (f) and ANN (g) models manage to predict these areas well. The stacked model (h) predictive ability is comparable to that of MARS and ANN.



(h) Kriging Region 2

(i) MARS Region 2

(j) ANN Region 2

(k) Stacking Region 2

(l) Kriging Region 3

(m) MARS Region 3

(n) ANN Region 3

(o) Stacked Region 3

The third region shows an area of plots, or agricultural holdings. The Kriging model (i) predicts the uniform properties exceptionally well. The MARS model (j) does not predict the uniform properties to the same accuracy as the Kriging model, but manages to predict results with some degree of accuracy. The ANN model (k) offers the worst predictions for this area. Again the stacked model (l) is comparable to the better of the 3 models.

7. Conclusion

Spatial methods and data mining methods result in relatively good models for spatially indexed property value data. However, these models perform well in different regions based on the degree of disparity present between the properties. Spatial methods such as Kriging perform well in regions with low levels of disparity, where properties are similar whilst statistical methods such as ANN and MARS perform well in regions of high disparity. Using the method of stacking, these methods may be combined into a new model. This new model has the ability to predict both types of regions well, as it uses both spatial and data mining methods. The stacked model has an R^2 value of 0.901.

References

BOHLING, G. (2005). Kriging. Last accessed: 2013-03-10.
URL: `http://people.ku.edu/~gbohling/cpe940/Kriging.pdf`

BREIMAN, L. (1996). Stacked regressions. *Machine Learning*, **24**, 49 – 64.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R., AND STONE, C. (1984). *Classification and Regression Trees*. Wadsworth International Group.

CHARLTON, M. AND FOTHERINGHAM, A. (2009). Geographically weighted regression. *National Centre for Geocomputation[White Paper].*

GARDNER, S. AND DORLING, M. (1998). Artificial neural networks (the multilayer perceptron)-a review of applications in the atmospheric sciences. *Atmospheric Environment*, **32(14–15)**, 2627–2636.

GUJARATI, D. N. AND PORTER, D. C. (2009). *Basic Econometrics*. McGraw-Hill Irwin, New York.

IZENMAN, A. (2008). *Modern Multivariate Statistical Techniques*. Springer.

LEATHWICK, J., ROWE, J., RICHARDSON, J., ELITH, J., AND HASTIE, T. (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, **50**, 2034–2052.

OLIVER, M. AND WEBSTER, R. (1990). Kriging: A method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, **4** (3), 313–332.

SHALIZI, C. (2009). Classification and regression trees. Last accessed: 2013-03-10.
URL: `http://www.stat.cmu.edu/~cshalizi`