Towards Using Geovisual Analytics to Interpret the Output of Geographically Weighted Discriminant Analysis

P. Foley[1], U. Demšar[2]

[1]National Centre for Geocomputation, NUI Maynooth, Co. Kildare, Ireland
Telephone: ++353 01 708 6731
Fax: ++353 01 708 6456
Email: peter.f.foley@nuim.ie

[2]National Centre for Geocomputation, NUI Maynooth, Co. Kildare, Ireland
Telephone: ++353 01 708 6178
Fax: ++353 01 708 6456
Email: urska.demsar@nuim.ie

1. Introduction

Geographically Weighted Methods are statistical techniques developed to model spatially varying (non-stationary) processes (Fotheringham et al. 2002). Their outputs are spatial datasets which are highly dimensional, complex and large. Interpreting these datasets is a significant challenge. One way to help with this is to use Geovisual Analytics methods.

Geovisual Analytics is the sub-discipline of Visual Analytics that deals with data with a spatial and possibly temporal extent. Visual Analytics combines "automated analysis techniques with interactive visualisations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets" (Keim et al. 2010).

This research aims to use Geovisual Analytics methods to transform the information contained in the output of a specific Geographically Weighed method: Geographically Weighted Discriminant Analysis (GWDA) into new knowledge about the underlying spatial process. This has been done before for other Geographically Weighted methods (Demšar et al. 2008a,b; 2010), but here we extend the principle to GWDA. This abstract describes progress to date and outlines a plan for the remainder of the research.

1.1 Geographically Weighted Discriminant Analysis

Discriminant Analysis is a supervised classification technique used to assign objects in a dataset to distinct classes. Training data are used to estimate the class means, covariance matrices and prior probabilities in attribute space and this information is used to calibrate classification functions of the attributes. Objects are assigned to the class with the maximum classification score. Linear Discriminant Analysis (LDA) outputs include; classification functions that are linear combinations of the attributes, the assigned class and the posterior probabilities which represent the probability that an object belongs to a particular class.

GWDA (Brunsdon et al. 2007) models spatial non-stationarity in the relationship between class membership and the attributes by allowing the parameters of the classification functions to vary spatially. GWDA outputs include; spatially varying classification functions, the assigned class and the posterior probabilities. The GWDA classification functions require analysis to understand the causes of spatial non-

stationarity but this is complex. Not only is there a cognitive difficulty comparing the values of multiple parameters for a single variable (Brunsdon et al. 2007) but in addition, the values of the classification functions are not absolute (Klecka 1980) which means that parameter values cannot be compared directly.

2. Combining Linear Discriminant Analysis with Geovisual Analytics

In this abstract we present the use of tools from the GeoViz Toolkit (Hardisty and Robinson 2010) to interpret the output of LDA and GWDA. Later, we will develop new visualizations specifically suited to exploring the output of GWDA.

2.1 Implementation of Discriminant Analysis in the GeoViz Toolkit

The GeoViz Toolkit is a free and open-source collection of visual and computational tools for exploring geographical datasets. These tools can be used in tandem so that multiple dynamically linked visualizations of the data are possible. Since one of the goals of Geovisual Analytics is to integrate visualization methods and spatial analysis techniques (Hardisty and Robinson 2010), we implemented LDA and GWDA in the GeoViz Toolkit as a first step.

2.2 Data

A data requirement for GWDA to work is that the classes are relatively evenly mixed spatially. In addition, the relationship between the classes and the attributes should vary spatially. We use a simulated dataset to ensure that both of these conditions are met. An advantage of this approach is that the non-stationary spatial patterns are already known so we are able to test the ability of different visualizations to detect them.

We used an existing well-known non-spatial Iris dataset and spatialised it to meet the requirements for GWDA (fig. 1). This dataset was first used by Fisher (1936) and comprises 150 Iris plants of three different species: 50 Iris Setosa, 50 Iris Versicolor and 50 Iris Virginica. Each plant has four associated measurements: sepal length, sepal width, petal length and petal width. To spatialise these data, we assigned the plants to cells on a rectangular grid with 10 rows and 15 columns using the following rules:

1. To ensure an even spatial mix of species, we reserved a random selection of 50 grid cells for each species.
2. To incorporate spatial non-stationarity, plants of each species were assigned to the reserved set of grid cells in a manner that created local patterns. Plants with the shortest petal length were assigned to cells in the bottom left corner of the grid and plants with the longest petal length were assigned to cells in the upper right corner. The ordering is equivalent to the height of an oblique plane over the study area such that the height at the bottom left corner is minimized and the height at the upper right corner is maximized.

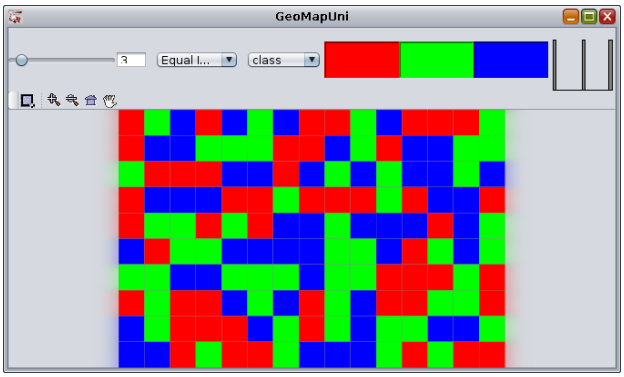

Figure 1. The spatialised Iris dataset showing Iris Setosa cells as red, Iris Versicolor cells as green and Iris Virginica cells as blue.

2.3 Experiment: Visualising LDA Results

Using our implementation of LDA, we classified the simulated spatial dataset using all four Iris measurements as predictor variables and used tools from the GeoViz Toolkit to visualise the output.

The confusion matrix for the classification is shown in table 1. The classification accuracy is 98% and only 3 out of 150 plants were misclassified.

	Iris Setosa	Iris Versicolor	Iris Virginica	Class Total
Iris Setosa	50	0	0	50
Iris Versicolor	0	48	2	50
Iris Virginica	0	1	49	50
LDA Total	50	49	51	150

Table 1. Confusion Matrix from an LDA classification of the spatialized Iris Dataset.

The following tools were found to be useful in visualizing the output of LDA:

1. GeoMapUni is a classified univariate choropleth map. It shows the spatial distribution of a single variable, in our case the 3 species of Iris (fig. 1).

2. GeoMap is a classified bivariate choropleth map. It shows the spatial distribution of two variables with a bivariate colour scheme. We used a bivariate map with a complementary colour scheme (Eyton 1984) to visualize the spatial distribution of the misclassified plants (fig. 2).

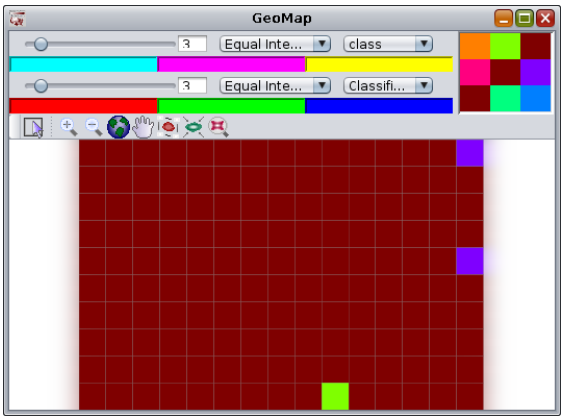

Figure 2. Location of misclassified Iris plants. Dark red cells contain correctly classified plants. The two purple cells contain Iris Versicolor plants misclassified as Iris Virginica and the single green cell contains an Iris Virginica plant misclassified Iris Versicolor.

3. ParallelPlot is a Parallel Coordinates Plot (PCP) to visualize a dataset in attribute space. We used a PCP to visualize the relationship between the three species of Iris and the predictor variables (fig. 3) and to visualize the relationship between the predictor variables, the posterior probabilities of the classification and the species of Iris for the 3 misclassified plants (fig. 4).

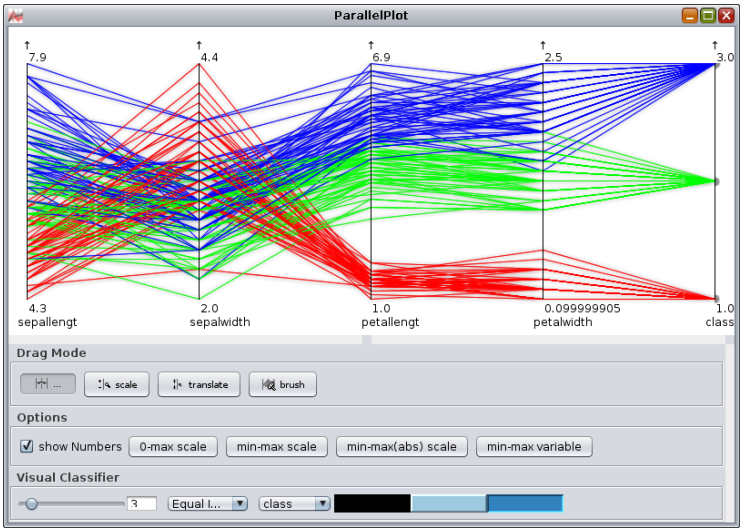

Figure 3. Relationship of the 4 Iris measurements to the species of Iris. Iris Setosa plants are in red, Iris Versicolor plants are in green and Iris Virginica plants are in blue.

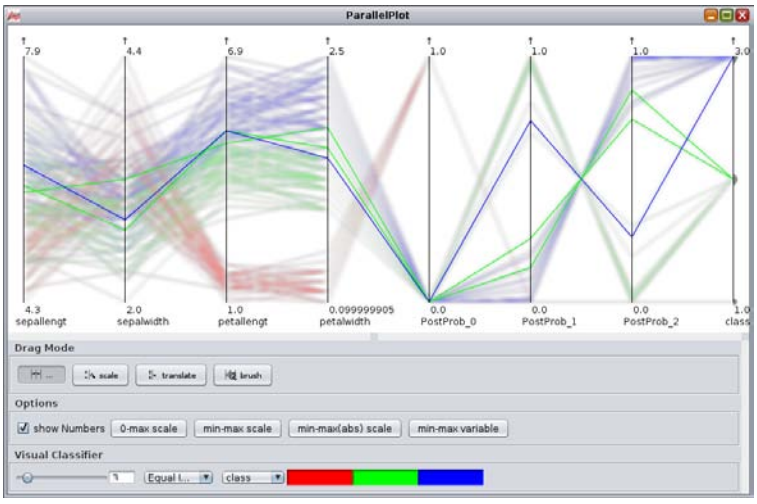

Figure 4. Highlights the relationship of the 3 misclassified plants to the 4 Iris measurements and the LDA posterior probabilities. Iris Setosa plants are in red, Iris Versicolor plants are in green and Iris Virginica plants are in blue.

4. StarPlotMap. This tool shows the spatial distribution of more than two variables using Star Plot icons. We used a Star Plot map to visualize the spatial distribution of the LDA posterior probabilities (fig. 5).

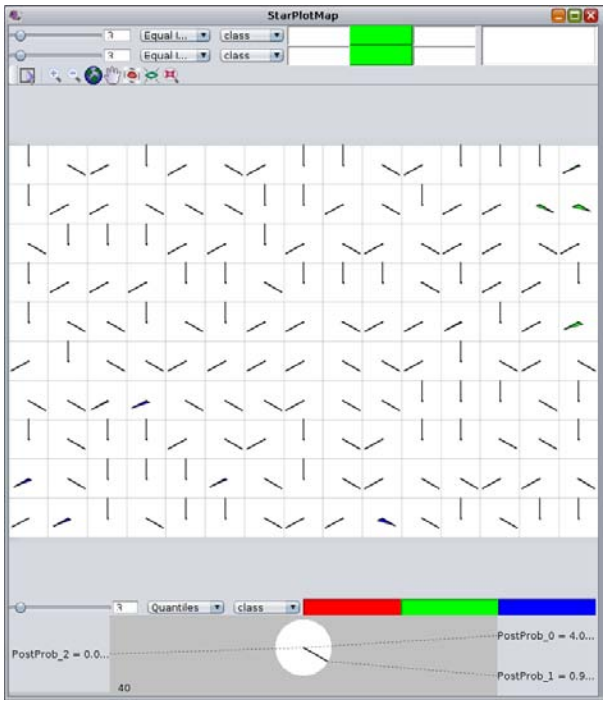

Figure 5. Spatial distribution of the posterior probabilities. The lengths of the rays are proportional to the posterior probabilities for each of the 3 species: rays pointing north for Iris Setosa, rays pointing south-east for Iris Versicolor and rays pointing south-west for Iris Virginica.

3. Next Step: Visualising the GWDA Results

The next steps are to identify the most useful tools from the GeoViz Toolkit to visualize the output of GWDA and finally, to develop new visualizations specifically for exploration of the GWDA output. These should provide additional insight into the output of GWDA and facilitate the interpretation of GWDA results. As this is work in progress, in this section we present some preliminary results from the GWDA classification of the same dataset.

Using our implementation of GWDA, we classified the simulated spatial dataset using all four Iris measurements as predictor variables. The confusion matrix for the classification is shown in Table 2. The classification accuracy is 100% and the variance in the GWDA posterior probabilities is reduced compared to the LDA posterior probabilities (fig. 6). The high classification accuracy for LDA and GWDA make it difficult to attribute the improved results to genuine spatial non-stationarity. Therefore these results should only be considered as preliminary and this experiment should be repeated for another, less ideal dataset. For example, since classification with GWDA performs so well, most of the posterior probability values are either 0 or 1 which accounts for considerable overprinting in the PCP (fig. 6). Therefore, this PCP should only be used in conjunction with other interactively connected visualisations to identify patterns. Note also the contrast between the variance of the posterior probability values in LDA (fig. 4) versus the almost binary separation in GWDA (fig. 6).

	Iris Setosa	Iris Versicolor	Iris Virginica	Class Total
Iris Setosa	50	0	0	50
Iris Versicolor	0	50	0	50
Iris Virginica	0	0	50	50
LDA Total	50	50	50	150

Table 2. Confusion Matrix from a GWDA classification of the spatialized Iris Dataset.

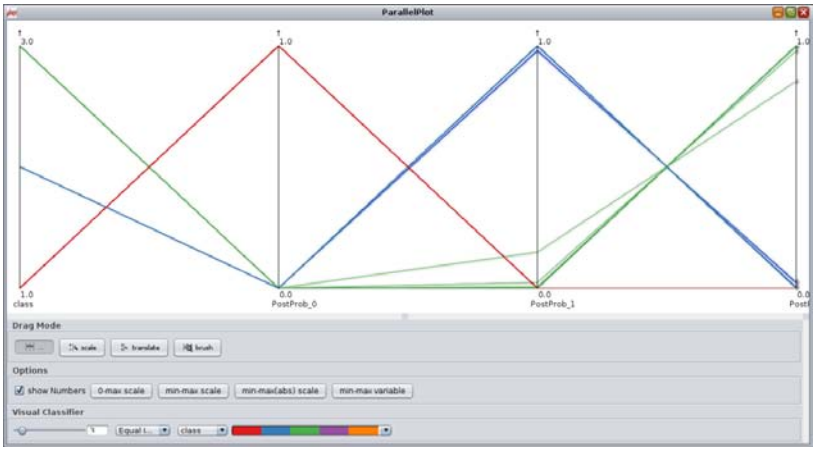

Figure 6. Relationship of species to the GWDA posterior probabilities. Iris Setosa plants are in red, Iris Versicolor plants are in blue and Iris Virginica plants are in green. This PCP shows all 150 plants, but as species are well separated by the posterior probabilities (i.e. they are either 0 or 1), there is a large amount of overprinting present in this PCP.

4. Conclusions

We have demonstrated that specific tools from the GeoViz Toolkit are useful in revealing spatial and non-spatial patterns in the output of LDA. For the remainder of the research we plan to develop new visualisations to provide insight into GWDA.

The tools for visualizing the output of LDA, described in section 2.3 can be used in exactly the same way with the GWDA output. However, the GWDA output presents additional challenges:

1. We need a method to visualize the spatially varying relationship between class membership and the predictor variables and this will require a new technique. A starting point could be to map the variation in posterior probabilities for a fixed set of predictor variables (Brunsdon et al. 2007). This could be improved by allowing the user to vary the predictor variables on the fly. We plan to visualize the posterior probabilities using a Treemap approach (Johnson and Shneiderman 1991). This should improve on the existing visualizations (StarPlot Map and PCP) which suffer from overprinting. We also plan to visualize the confusion matrix using a Mosaic Plot (Hartigan and Kleiner 1981).
2. For this particular dataset, the difference between the classification accuracy for LDA and GWDA is small. For a less ideal dataset, mapping the difference between the LDA and GWDA posterior probabilities would highlight cells where the confidence in the classification has been enhanced or reduced. To decrease the predictive accuracy of the four Iris measurements we have "confused" the dataset by perturbing them slightly (fig. 7). The contrast between the LDA and GWDA classification accuracies (~87% and ~91% respectively) has now increased.

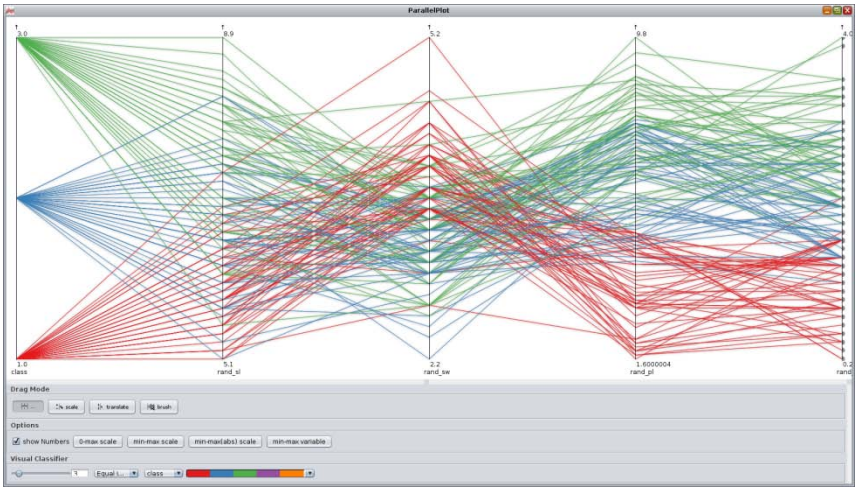

Figure 7. Relationship of the 4 randomized Iris measurements to the species of Iris. Iris Setosa plants are in red, Iris Versicolor plants are in blue and Iris Virginica plants are in green.

3. Identification of outliers in the classes is possible since the Mahalanobis Distance squared from each object to the class means follows a chi-squared distribution with *m* degrees of freedom where *m* is equal to the number of predictor variables (Manly 2005). We are investigating possible visualisations for outlier detection based on this.

5. Acknowledgements

We thank Frank Hardisty at the Pennsylvania State University for technical advice and assistance on extending the GeoViz Toolkit. The authors are supported by a Research Frontiers Grant (09/RFP/CMS2250) awarded to Urška Demšar by Science Foundation under the National Development Plan.

6. References

Brunsdon C, Fotheringham A S and Charlton M, 2007, Geographically Weighted Discriminant Analysis. *Geographical Analysis*, 39(4):376-396.

Demšar U and Fotheringham A S, 2010, Geographically Weighted Principal Components Analysis and the Curse of Dimensionality. *Journal of Visual Languages and Computing* (Under Review).

Demšar U, Fotheringham A S and Charlton M, 2008a, Combining Geovisual Analytics with Spatial Statistics: the Example of Geographically Weighted Regression. *The Cartographic Journal*, 45(3):182-192.

Demšar U, Fotheringham A S and Charlton M, 2008b, Exploring the spatio-temporal dynamics of geographical processes with geographically weighted regression and geovisual analytics. *Information Visualization*, 7(3-4):181-197.

Eyton J, 1984, Complementary-color, two variable maps. *Annals of the Association of American Geographers*, 74(3):477-490.

Fisher R A, 1936, The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7:179-188.

Fotheringham A S, Brunson C and Charlton M, 2002, *Geographically Weighted Regression – the analysis spatially varying relationships*. John Wiley & Sons, Chichester, England.

Hardisty F and Robinson A C, 2010, The GeoViz Toolkit: Using component-oriented coordination methods for geographic visualization and analysis. *International Journal of Geographical Information Science (Forthcoming).*

Hartigan JA, and Kleiner B, 1981, Mosaics for Contingency Tables. *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, Springer, New York

Johnson, B and Shneiderman, B, 1991, Treemaps: a space-filling approach to the visualization of hierarchical information structures. *Proceedings of the 2nd International IEEE Visualization Conference*.

Keim D, Kohlhammer J, Ellis G and Mansmann F, eds, 2010 *Mastering the Information Age - Solving Problems with Visual Analytics*, http://www.vismaster.eu/wp-content/plugins/cimy-counter/cc_redirect.php?cc=full_book&fn=http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf

Klecka W R, 1980, *Discriminant Analysis*. Quantitative Applications in the Social Sciences, Sage Publications, Beverly Hills, California.

Manly B F J, 2005, Multivariate Statistical Methods: A Primer. Chapman & Hall/CRC, Boca Raton, Florida, third edition.