A new approach to analysing spatial data using sparse grids

Shawn W. Laffan[a], Howard Silcock[b], Ole Nielsen[b] and Markus Hegland[b]

[a] Centre for Remote Sensing and GIS, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, Australia. Shawn.Laffan@unsw.edu.au [b] Mathematical Sciences Institute, Australian National University, Canberra, Australia.

Abstract: We describe in this paper a new data mining approach for the analysis of spatial data for environmental modelling. The sparse grids analysis system models the functional relationship between a set of predictor variables and a response variable by using a combination of easily computable functions defined on grids of varying mesh sizes in attribute space. The approach circumvents the so-called "curse of dimensionality" by using, instead of a costly high-dimensional grid a with a fine mesh size in every dimension, a collection of grids that are coarse along some dimensions but fine along others. Adaptive sparse grid regression and classification methods select combinations of grids that suit a particular data set. One advantage of the sparse grids approach from an environmental analysis perspective is that it uses machine learning approaches, and so can deal with correlated data, as are common in environmental problems. One advantage of the sparse grids approach from an environmental analysis perspective is that it uses machine learning approaches, and so can deal with correlated data, as is commonly the case with geographic data. They also require fewer degrees of freedom than do full grid models, allowing them to be applied to more datasets. The parameters defining the adaptive sparse grids can be used to interpret relationships in terms of scale and resolution. For example, the distribution of mesh points used in the set of lattices describes the complexity of the relationships present. It can be used to understand if the system is responding to fine scale variations (many mesh points used) or to gross patterns (few mesh points used). This is valuable information for environmental modelling.

Keywords: *environmental modelling, predictive modelling, sparse grids, geographic information systems*

1. INTRODUCTION

We introduce in this paper a new analysis tool for environmental data known as sparse grids. Sparse grids were originally developed for the solution of partial differential equations (Zenger, 1991), and later adapted to data mining (Garcke et al., 2001). They have great applicability to the analysis and understanding of environmental data and processes.

This paper represents a work in progress with initial results. More detailed descriptions of the method and results will be presented in a later publication.

1.1. Approaches to environmental correlation

The process of finding relationships between a response variable and some set of predictor variables is known as environmental correlation. It can be used for a variety of purposes, including classification, predictive mapping, or simply to better understand the relationships in a system.

There are many different approaches to environmental correlation, including Classification and Regression Trees, Artificial Neural Networks, Generalised Linear Models, Generalised Additive Models, and Multivariate Adaptive Regression Splines (see Hastie et al. 2001).

All these approaches effectively address the curse of dimensionality and use few degrees of freedom, both important considerations for analysing geographic data (Gahegan, 2003). They differ in how well they can approximate the data and, most importantly for our purposes, in the way they can be used to extract information about the underlying system relationships.

The sparse grid predictive model is additive, where all the components are piecewise multilinear functions. It generalises linear and additive models (in the sense of Hastie and Tibshirani, 1986) and can be interpreted as a computationally very efficient variant of a multivariate regression spline. Sparse grid function evaluation is very fast, and effective parallel algorithms are used to fit the functions to very large data sets using high performance computers. The function evaluation uses a machine learning approach, which allows it to effectively deal with correlated data. Geographic

datasets are commonly correlated, and so this is an advantage over parametric statistical approaches.

2. SPARSE GRIDS

The sparse grids system uses a sum of piecewise multilinear functions to represent relationships between a subject variable and a set of predictor variables. In the simplest case, each of these component functions depends on only one variable, and is piecewise linear. The number of grid points of the components is the degrees of freedom used by that component and characterises the complexity of the model and its capability to approximate fine fluctuations (see Figure 1).

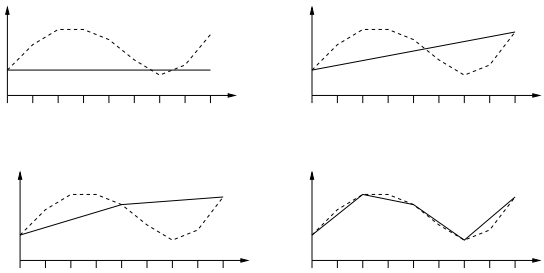

Figure 1. Modelling a function (dotted line) using piecewise linear functions (solid lines) with various degrees of freedom.

A piecewise linear function can be represented as a combination of components which all have an intuitive meaning like the height, slope, curvature, and other fluctuations including noise (see Figure 2). It is this interpretation which is inherited by the sparse grids and provides useful insights into the underlying system in attribute space.

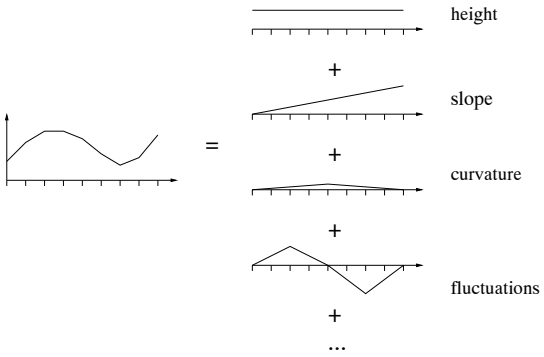

Figure 2. The combination of intuitive grid functions.

In the full grids case, one could represent a relationship using a single system of regularly spaced grid points. However, it can be seen that most of the degrees of freedom used do not contribute anything to the approximation provided by the model. In the sparse grids case, one can represent a similar amount of complexity by using a series of partial grids, the combination of which will give the solution (Figures 3 and 4). To take a very simple example, a full grid might use a 5 by 5 lattice of grid points, requiring 25 grid points in total (Figure 3). A sparse grid system might instead use two grids, each with five grid points along one data axis and two grid points along the other (denoted $V_{1,3}$ and $V_{3,1}$). The functions derived from these grids are added together, but to avoid 'double counting' one then needs to subtract the function associated with the intersection grid—in this case, the grid with two grid points along each data axis ($V_{1,1}$) (Figure 4). In this simple system a total of only 16 grid points are used, although there is some redundancy because the constant term is repeated in each model. The additive system used by sparse grids also makes it easy for interaction terms to be included where parts of the system are represented in more than one model.

Figure 3. The number of grid points used by the sparse grid system (right) is much less than that used for the full grid system (left).

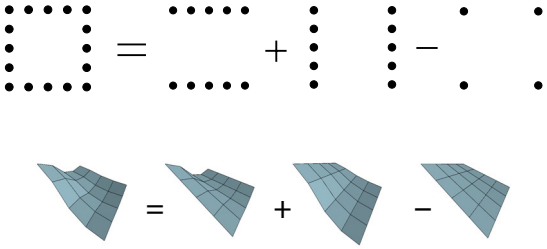

Figure 4. The combination of sparse grids approximates more complex functions in attribute space, plan view and three dimensional view.

The computational advantage of sparse grids becomes apparent when one moves to higher dimensional data spaces, such as are commonly used for environmental correlation analyses. Full grids with d variables and m grid points per variable have a total of m^d grid points whereas classical sparse grids have only $m(\log_2 m)^{d-1}$ grid points, and newer sparse grids have even fewer. The actual number of grid points used depends on the data, and in many cases the number of grid points is of the order dm. For example, in the case when the data can be approximated by a linear function (two grid points) with eight variables, the full grid requires $2^8 = 512$ data

points, where the sparse grid might use a combination of eight linear one dimensional functions with a total of $2 \times 8 = 16$ degrees of freedom. This is a significant reduction. Furthermore, the number of degrees of freedom of the sparse grid is a more realistic estimator of the number of degrees of freedom inherent in the underlying data.

As with most analysis systems, it is possible to use sparse grids as a spatial interpolator by using spatial coordinates as the predictor variables. However, it is likely that other systems specifically designed for this purpose would be more effective. Sparse grids are most useful when dealing with high dimensional datasets.

The interpretation of the system of grids provides a means of understanding the relationships within the system studied. One advantage of sparse grids over other systems is that the number of mesh points used by the system to represent a relationship can be used to gain some understanding of the scale of the relationships. This is done through visualisation of the results for each sparse grid in the system. Such visualisation can be done through mapping the relationships modelled by each sparse grid, as has been done for Artificial Neural Networks by Laffan (1998), and by plotting error matrices for the individual sparse grids (Figure 6). The error matrix approach can also be used as a means of pruning grids that contribute little to the overall solution from the system, possibly in a step-wise manner, and forms the basis of a means to understand the scales represented in the system.

3. AN APPLICATION

We applied the sparse grids system to an extensive geochemical dataset from Weipa, Far North Queensland, Australia. The objective of the test application is to find correlations between subsurface regolith properties and some set of features that are easily measured from the surface. The purpose of this is to better understand how applicable the mapping of regolith properties is. Previous work using this dataset is presented in Laffan (2001, 2002) and Laffan and Lees (submitted).

The dataset consists of a set of 57,642 drill cores collected between 1955 and 1980, 54,757 of which intersect bauxite. These are sampled on a magnetic north aligned grid at spacings from 38 m to 308 m (1000, 500, 250 and 125 feet) using an infilling sample design. Each drill core contains data for percentage abundance of oxides of aluminium, iron, silica and titanium, and for the depth to the base of the bauxite layer, and the depth of the overburden (topsoil or A-horizon).

There is also a set of eight surface measurable features consisting of: a DEM and derived attributes of slope and flow accumulation (fD8 algorithm, Freeman, 1991); Landsat Thematic Mapper bands two, four and seven, captured 16 June 1988; and the Euclidean distances from swamps (melon holes) and from drainage lines (defined as cells with flow accumulation greater than 200,000 m^2).

All datasets use a 30 m cell size, and the drill core dataset reduces to 14,833 locations after raster conversion and exclusion of locations identified in the Landsat dataset as having been mined (regrowth or mine floor), or as cleared for mining.

A sparse grids model was trained using two thirds of the silica data, with an accuracy of 48% correct within a tolerance of 1% silica abundance. The testing set returned an accuracy of 26% for the same error tolerance. The testing accuracy is almost identical to that obtained using an ANN on the same data (25%, Laffan, 2001), with some key differences (Figure 5). The ANN predictions were similar for both the training and testing data, where the sparse grids training accuracies are higher than the testing accuracies. The sparse grids error distribution is also less skewed than that for the ANN, and is aligned to the axis of correct prediction. However, the errors are more dispersed than for the ANN.

We attribute the low predictive accuracy to the effect of spatial non-stationarity in the relationships. Previous analyses using Geographically Weighted Regression (GWR, Laffan, 2001), which fitted local regression models to samples within 300 m of each location, returned accuracies of 67% for a 1% error tolerance. However, further investigation showed that only 15% of sample locations had any relationship that could not also be explained using a local constant model (mean of the 300 m radius sample; Laffan, 2001). For comparison, only 6% of locations were better predicted by the ANN than by its related constant model (the global mean), and 7% for these sparse grids predictions.

Visualisations of the error matrices for individual sparse grids are shown in Figure 6. None of the variables analysed appear have a strong relationship with percentage silica abundance. Again, we attributed this to spatial non-stationarity in the relationships.

CONCLUSION

The sparse grids system is a means of analysing environmental datasets with high dimensionality. The initial results are comparable to those used for previous work on this dataset.

Given that previous results showed a better response for the geographically local analyses, an avenue currently being pursued is to extend the sparse grids model to use such a weighting scheme.

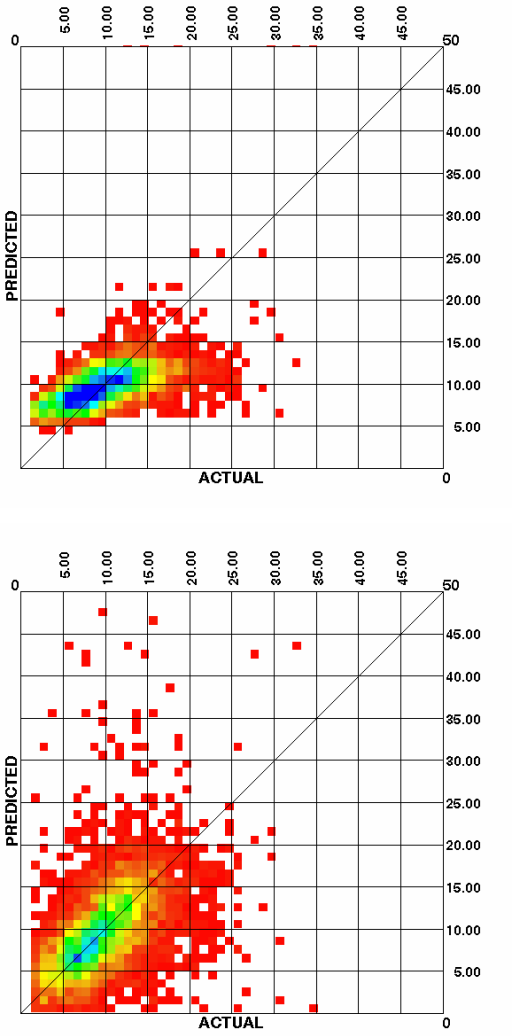

Figure 5. Distribution of errors for (a) ANN and (b) sparse grids predictions (negative predictions not shown). Colours are scaled between 1 and 100, where red is low and blue is high.

4. REFERENCES

Freeman, T.G., Calculating catchment area with divergent flow based on a regular grid, *Computers and Geosciences*, 17, 413-422, 1991.

Gahegan, M., Is inductive machine learning just another wild goose chase (or might it lay the golden egg)?, *International Journal of Geographical Information Science*, 17, 69-92, 2003.

Garcke, J., Griebel, M. and Thess, M., Data mining with sparse grids, *Computing*, 67, 225-253, 2001.

Hastie, T.J. and Tibshirani, R., Generalized additive models (with discussion), *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 1, 297-318, 1986.

Hastie, T.J., Tibshirani, R. and Friedman, J., *The elements of statistical learning: Data mining, inference and prediction*, Springer, 2001.

Laffan, S.W., Visualising neural network training in geographic space, *Third International Conference on GeoComputation*, University of Bristol, UK, http://www.geocomputation.org/1998/, 1998.

Laffan, S.W., *Inferring the spatial distribution of regolith properties using surface measurable features*, Unpublished PhD thesis, ANU, 2001.

Laffan, S.W., Using process models to improve spatial analysis, *International Journal of Geographical Information Science*, 16, 245-257, 2002.

Laffan, S.W. and Lees, B.G., Predicting regolith properties using environmental correlation: a comparison of spatially global and spatially local approaches, submitted to *Geoderma*.

Zenger, C., Sparse Grids. In Hackbusch, W. (ed), *Parallel Algorithms for Partial Differential Equations*, Proceedings of the Sixth GAMM-Seminar, Kiel, 1990, Notes on Num. Fluid Mech., Vieweg, volume 31, pp 241-251, 1991.

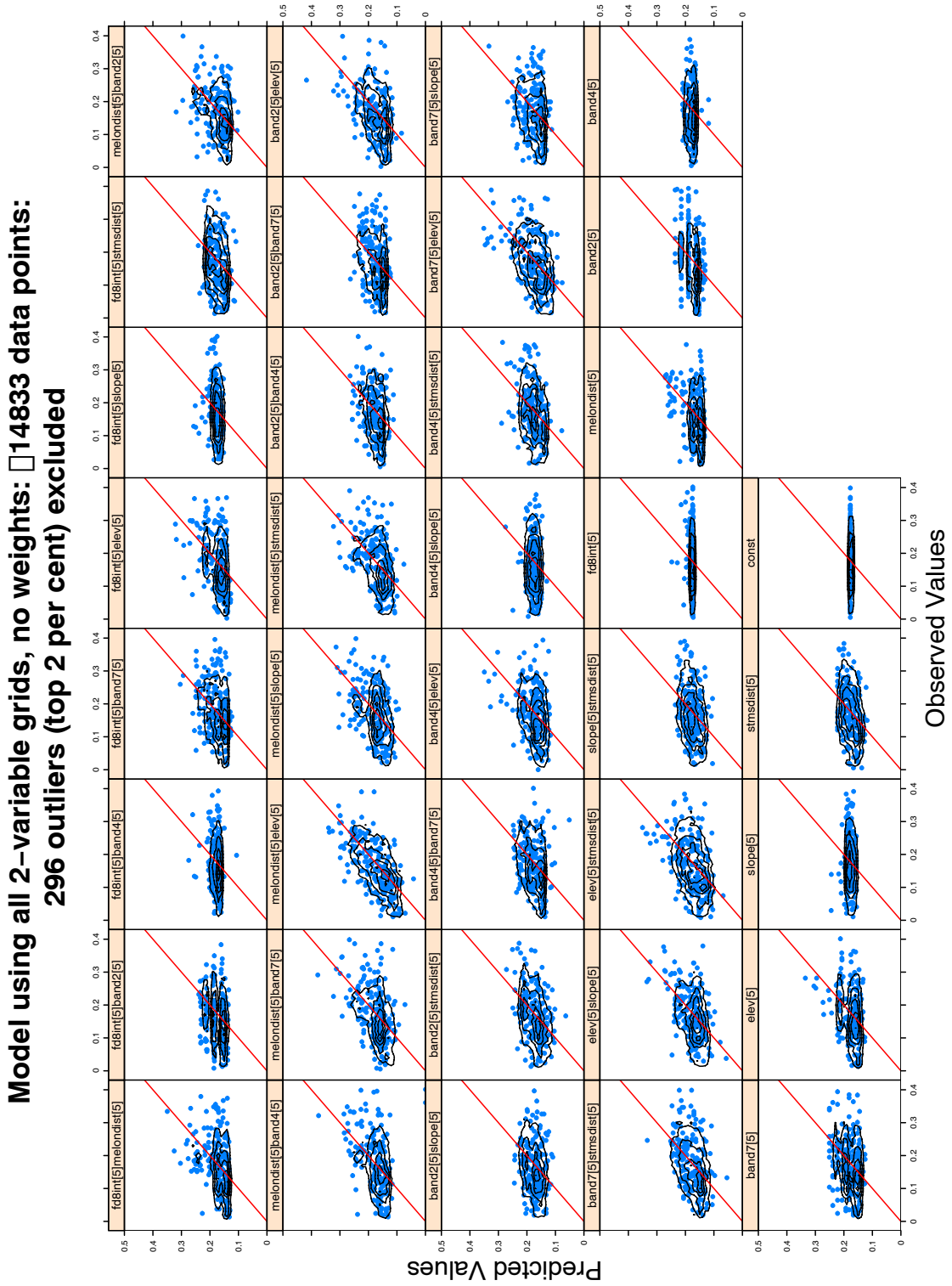

Figure 6. Scaled error plots for individual grids in the sparse grids system, including combination grids and the constant model. Contours relate to the density of data points.