Using spatial randomisations to improve the utility of Geographically Weighted Regression model results

[1]Laffan, S.W. and [2]S.A. Bickford

[1]University of New South Wales, [2]CSIRO Plant Industry, E-Mail: Shawn.Laffan@unsw.edu.au

Keywords: *Geographically Weighted Regression; Climate surfaces; GIS; species richness, randomisations.*

EXTENDED ABSTRACT

Geographically Weighted Regression (GWR) and its variants are analysis methods that can cope with the multi-scale, spatially non-stationary relationships common to spatial data. They achieve this by using geographical sub-samples of the data for which one expects the complexity of any relationships to be simpler than over the whole study area.

The current implementation of GWR does not allow for locally varying analysis scales, but these can be inferred using a post-processing step to combine a set of GWR models, for example choosing the model with the best local r^2 at each analysis location. With this information, one can assess the spatial scale at which a set of variables are related, and how this varies through space. Such an understanding is extremely useful when trying to understand the processes operating across a landscape, and would improve the utility of GWR type analyses.

However, the above approach has two fundamental limitations. First, the chance of high goodness of fit statistics increases as the sample size decreases. Second, goodness of fit statistics normally assume that errors in the sample data are independent. Such independence is unlikely with spatial data, and therefore the goodness of fit statistics are likely to be over-estimates. One therefore cannot use local goodness of fit statistics alone to choose the best local GWR model.

One approach that can be applied to counter this effect is to use spatial randomisations to augment the goodness of fit statistics. We describe such an approach using species richness data of ferns across the Australian continent, exploring the effect of two randomisation models. Species richness is the number of unique species occurring in a region, where we used 50 km by 50 km cells.

Using GWR, the species richness surface is correlated with two climatic parameters, the mean water relations and standard deviation of annual rainfall within each analysis cell. These were correlated one at a time using Gaussian spatial weights with bandwidths at 100, 200, 300, 400, 500, 600 and 800 km.

The first randomisation used a model of complete spatial randomness. This is the same randomisation as used in the GWR software, but adapted to produce spatially local assessments. In this model, we randomly allocated species richness values across the landscape without regard to any spatial structure in the original data. However, such a random distribution is unlikely in reality, and ignoring spatial structures represents a test that is easy to satisfy when comparing the randomised data against the original data.

The second randomisation allocated the data to the landscape on a species by species basis, where each species range occurs in a circular pattern around a randomly selected cell and the total number of species records in the data set is kept exactly the same as in the original. This is a very conservative model of the spatial structure.

Each GWR model was compared against 1000 randomisations for each random model, and the best analysis scale for each GWR model location was taken as that with the highest local r^2 of those that were better than the randomised model 95% of the time.

There is almost no difference for the results between the two randomisation approaches. This is because even the circular model is randomly located, and so the resulting species richness surface is much less structured than the original surface. More variable results are to be expected if using indices of species level distributions such as endemism or genetic diversity, as opposed to simple species counts as used here. In these cases replicating the species richness surface can be used as an additional control on the randomisation.

However, despite the lack of difference between the two approaches, this spatially local approach does provide greater confidence in GWR model results than when using a standard global randomisation.

1. INTRODUCTION

Geographically Weighted Regression (GWR) (Fotheringham et al., 2002) and its variants are analysis methods that can cope with the multi-scale, spatially non-stationary relationships common to spatial data. They achieve this by using geographical sub-samples of the data in a moving window approach, where one expects the complexity of any relationships to be simpler within the geographical sub-sample than over the whole study area. It should be noted that these analyses do not assess the degree of autocorrelation or spatial dependence. Rather, they can operate in the presence of these effects.

The current implementation of GWR does not allow for locally varying analysis scales, but these can be inferred using a post-processing step to combine a set of GWR models, for example by choosing the model with the best local r^2 at each analysis location. With this information, one can assess the spatial scale at which a set of variables are related, and how this varies through space. Such an understanding is extremely useful when trying to understand multi-scalar processes operating across a landscape, and would improve the utility of GWR type analyses.

However, this approach has two fundamental limitations. First, the chance of achieving a high goodness of fit statistics increases as the sample size decreases. Second, goodness of fit statistics usually assume that the errors in the sample data are independent, causing the errors to cancel out. Such independence is unlikely with spatial data, and therefore the goodness of fit statistics are likely to be over-estimates. One approach that can be applied to counter these effects is the use of spatial randomisations to augment the goodness of fit statistics when assessing model results. We describe such an approach using species richness data of ferns across the Australian continent, exploring the effect of two randomisation models on the results.

2. METHODS

Species occurrence data were obtained from geo-referenced specimen records from Australian herbaria. The determinations of species records from taxonomically confusing groups were checked and removed if there was uncertainty in their likely veracity. All duplicate records, naturalised species, hybrids and specimens from cultivated plants were removed. Obvious geo-locational errors resulting from incorrect data entry were identified by comparison with other published range maps (Orchard 1998) and by expert checking. The final dataset consisted of 37,071 geo-referenced records of 418 species in 105 genera. All of the data locations were projected into a Lambert's conic conformal projection with two standard parallels at 18ºS and 36ºS, centred on a meridian at 134ºE. This projection has minimal spatial distortion across large areas like the Australian continent.

Species richness was calculated as the number of unique species occurring in a region, where we used a square cell 50 km on a side to reduce the effects of biased and non-random sampling common to museum and herbarium data (Figure 1)(Crisp et al., 2001; Bickford et al., 2004).

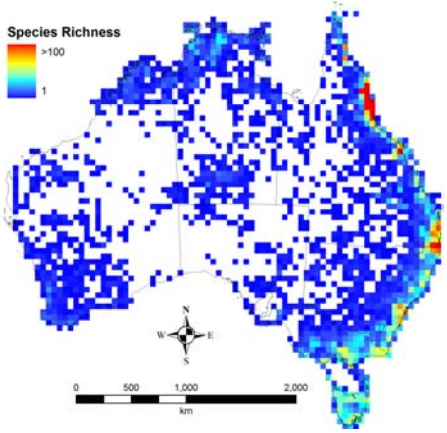


Figure 1. The distribution of fern species richness across Australia at a 50 km by 50 km cell size.

Using GWR, the species richness surface was correlated with the mean water relations within a cell and with the standard deviation of annual rainfall within each species richness cell. The standard deviation was used to provide a proxy for the heterogeneity within the cell. The relationship with other variables is described elsewhere (Bickford and Laffan, in prep.). We used a Gaussian spatial weights scheme for the GWR weightings, with seven bandwidths at 100, 200, 300, 400, 500, 600 and 800 km. The models were fitted one variable at a time to avoid the multicollinearity effects to which GWR is subject (Wheeler & Tiefelsdorf 2005). The randomisations and collation of model results were implemented in PERL.

The two randomisation approaches have been previously used to assess plant endemism patterns (Laffan & Crisp 2003). The first uses a model of complete spatial randomness (Figure 2). This is the same randomisation as used in the GWR software (Fotheringham et al., 2002). In this model we randomly allocated species richness values across the landscape without regard to any spatial structure in the original data. However,

such a random distribution is unlikely in reality. Ignoring the spatial structure in the data represents a test that is easy to satisfy when comparing the randomised data against spatially structured data.

The second randomisation allocates the data to the landscape on a species by species basis, where each species range occurs in an approximately circular pattern around a cell randomly selected from the original distribution (Figure 2). The randomly generated species richness values were the number of unique species randomly allocated to each cell, where the total number of species records is kept exactly the same as in the original data set. This is a very conservative model of the spatial structure at the species level.

Each GWR model was compared against 1000 randomisations for each random model, and the best analysis scale for each GWR model location was taken as that with the highest local r^2 of those that were better than the randomised model 95% of the time.

Random values were generated using the Mersenne Twister Pseudo Random Number Generator (Matsumoto & Nishimura 1998) because it has a period length of $2^{19,937}$ ($\sim 4.3*10^{6002}$) and passes all of the current tests of randomness essential for reliable analyses (Van Neil & Laffan 2003; McCullough & Wilson 2005).

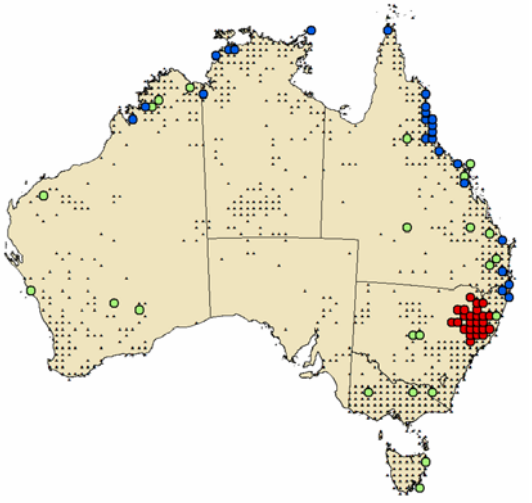
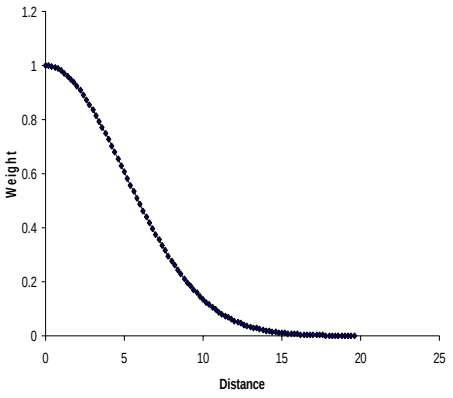
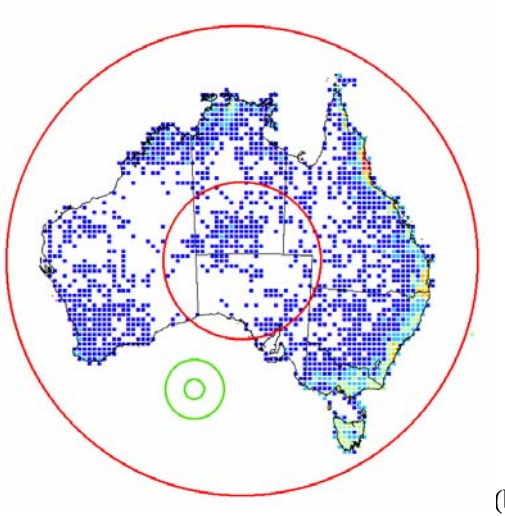


Figure 2. The two randomisations generate differing spatial structures. The blue circles represent the observed distribution of *Acrostichum aureum*, the green circles represent its distribution for one iteration of the complete spatial randomness model, while the red circles are for one iteration of the circular random model. The triangles represent the distribution of all species records after aggregation to the 50 km grid cells to which the species may be allocated.



(a)



(b)

Figure 3. The Gaussian sample weights decay to a zero asymptote at approximately three times the bandwidth (a). The circles (b) indicate the relative locations of the weights in relation to Australia, with the inner circle being the bandwidth and the outer circle being the zero asymptote. Red circles are a bandwidth of 800 km, green circles are 100 km.

3. RESULTS AND DISCUSSION

There is little difference between the cell based and the circular randomisations when considered at each individual spatial scale (Figure 4, Figure 5). The models are also better than random for all locations analysed with bandwidths of 400 km and above. Not surprisingly, those locations where the GWR models are worse than the randomisation tend to be those where the local model has a low r^2 value. These are also often locations which have low species richness values, such as near the edges of the Western Australian deserts. One exception to this is Tasmania, which has a high species richness of ferns, but a low correlation with the two variables used.

The lack of difference between the two randomisations is likely due to the use of species richness data as the variable that was randomised. Even with the circular model, the randomly generated species richness surfaces will be far less structured than the original, and therefore less likely to provide a better correlation with the climate surfaces than the original species richness data. While the results indicate little difference between the two randomisations for this data set, one would expect greater differences when using measures such as species endemism and genetic diversity (e.g. Bickford et al., 2004), and for which the species richness values are used as an additional constraint on the randomisation (see Laffan & Crisp 2003). In this case the species are randomly located across the landscape, but the spatial distribution of species richness must be replicated to some degree. One must also consider that there is a strong relationship between ferns and water related variables like those used here. This may not be the case with other biotic systems that do not depend so strongly on water, for example sclerophyll vegetation.

The aggregate results (Figure 6, circular model) indicate that the correlations are generally positive and strong (t>2) for mean water relations within a cell and with the standard deviation of annual rainfall in most parts of Australia, except for South West Western Australia and for central Australia and parts of the Northern Territory for the Annual Rainfall SD. The scale of the best relationship is also spatially clustered, and the smaller scales correspond to areas of higher species richness (compare with Figure 1). The unusual pattern near the Victorian/New South Wales border could be due to the inclusion of values from Tasmania in the larger analysis bandwidths. Other issues of multiple levels of relationships and the influence of other variables are considered in Bickford and Laffan (in prep.).

4. CONCLUSIONS

While the results for these variables indicate that the spatial scales with the higher local r^2 values were always better than the random models, and that there was little difference between the two randomisations, this should not be expected for all data sets.

The randomisation approach described here allows a more rigorous assessment of what scale should be chosen when aggregating a multi-scaled group of GWR model surfaces into a single surface and enables a significant improvement over a fixed bandwidth GWR analysis. Randomisations also need to be developed that can represent "realistic" dispersal patterns of species across a landscape.

5. REFERENCES

Bickford, S. A., Laffan, S. W., De Kok, R. & Orthia, L. (2004). Spatial analysis of taxonomic and genetic patterns and their potential for understanding evolutionary histories. *Journal of Biogeography*, 31, 1715-1733.

Crisp, M. D., Laffan, S., Linder, P. & Monro, A. (2001). Endemism in the Australian flora. *Journal of Biogeography*, 28, 183-198.

Fotheringham, A. S., Brunsdon, C. & Charlton, M. (2002). *Geographically Weighted Regression, the analysis of spatially varying relationships* (New York: Wiley).

Laffan, S. W. & Crisp, M. D. (2003). Assessing endemism at multiple spatial scales, with an example from the Australian vascular flora. *Journal of Biogeography*, 30, 511-520.

Matsumoto, M. & Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modelling and Computer Simulation*, 8, 3-30.

Mccullough, B. D. & Wilson, B. (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics and Data Analysis*, 49, 1244-1252.

Orchard, A. (1998). *Flora of Australia Volume 48: Ferns, Gymnosperms and Allied Groups* (Melbourne: ABRS/CSIRO).

Van Neil, K. & Laffan, S. W. (2003). Gambling with randomness: The use of Pseudo-Random Number Generators in GIS. *International Journal of Geographical Information Science*, 17, 49-68.

Wheeler, D. & Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7, 161-187.

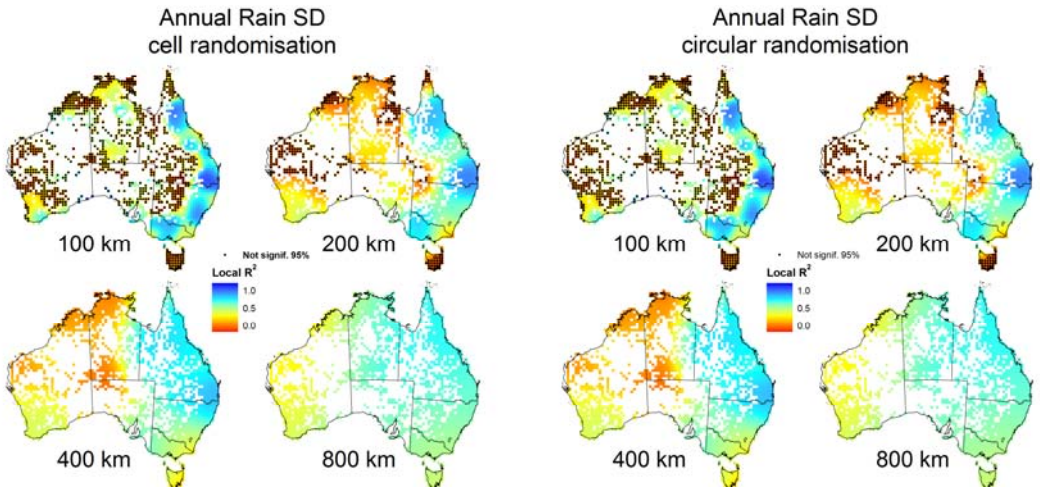


Figure 4. Local r^2 values for the standard deviation of annual rainfall at four scales using the cell based (left) and circular (right) randomisations. The black dots indicate where the model is better than random for less than 95% of the 1000 randomisations.

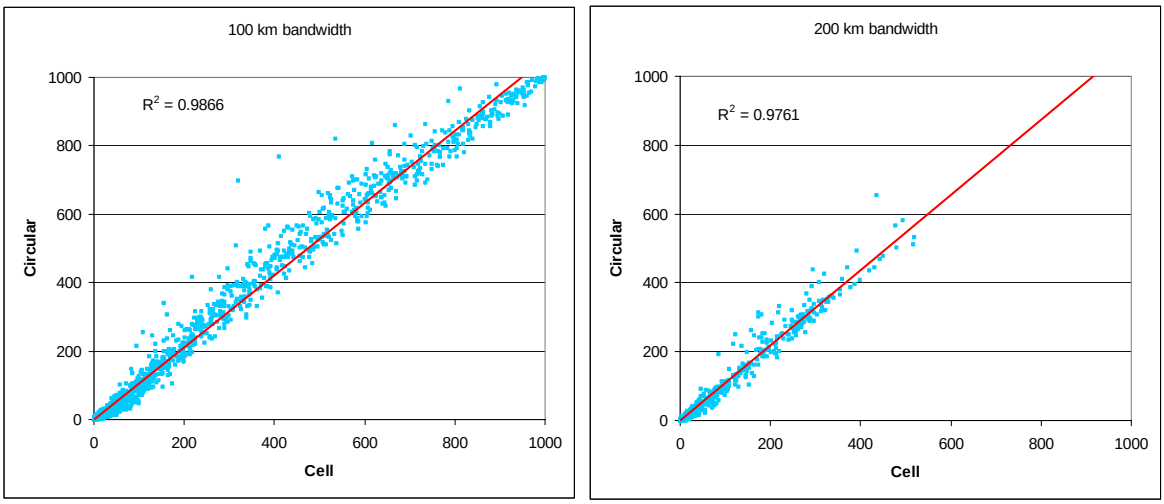


Figure 5. The randomisation surfaces are very similar between models (Annual Rainfall SD results). The y-axes represent the frequency the original model was better than the circular randomisation, while the x-axes are the frequency the original model was better than the cell based randomisation.

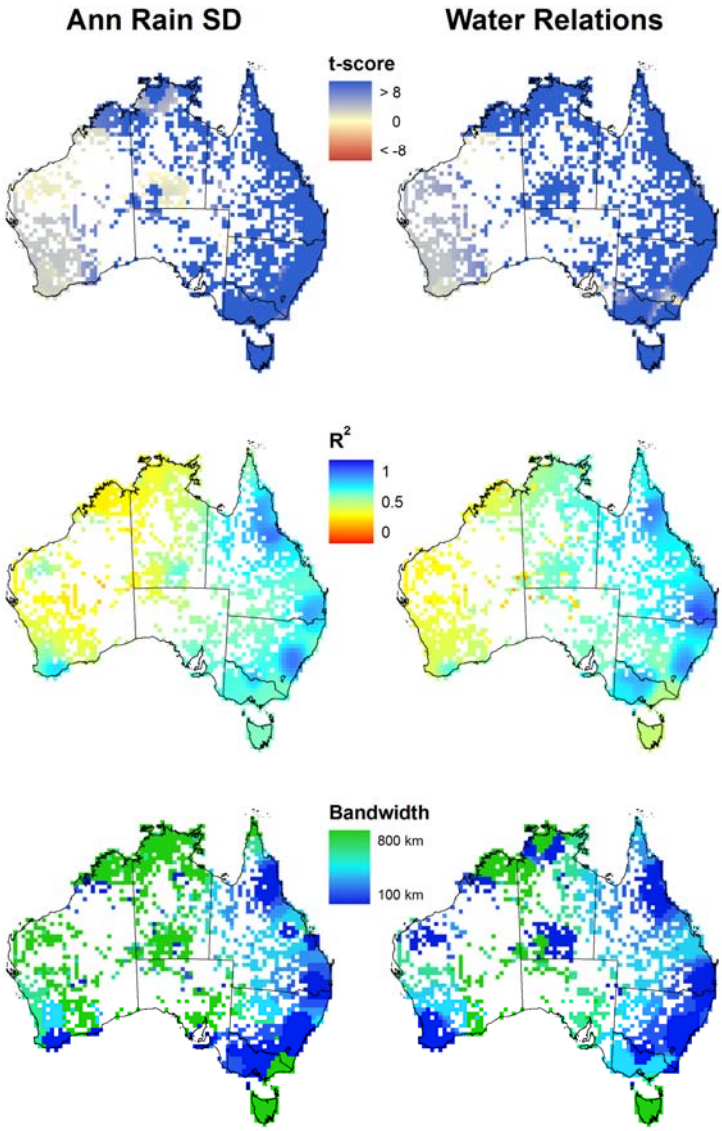


Figure 6. The spatial distributions of the combined model results are clustered (calculated using the circular randomisations).