

A technique for constructing monotonic regression splines to enable non-linear transformation of GIS rasters

Manion, G.

*Landscape modeling & Decision Support Section,
NSW Department of Environment and Climate Change
Email: glenn.manion@environment.nsw.gov.au*

Abstract:

The inherent nonlinearity commonly encountered in large scale ecological data sets often presents problems for linear regression models generated with data from highly diverse biological groups, such as plants and invertebrates. This can be especially frustrating when working in regions exhibiting high rates of spatial turnover in biological composition.

Novel techniques such as generalised dissimilarity modelling accommodate this nonlinearity with the application of piecewise polynomials or splines to address the curvilinear relationship between increasing ecological distance and, observed compositional dissimilarity between sites. But to make spatial predictions using raster data, one needs to be able to use the modelled coefficients to transform environmental surfaces.

The transformation of a Digital Elevation Model (DEM) raster surface to represent compositional dissimilarities between vascular plant assemblages in Northern NSW, Australia is used as a simple case study.

This paper presents a flexible technique that has many applications, not only ecological, to enable non-linear transformations of raster data.

Keywords: *I-Splines, Generalised Dissimilarity Modelling (GDM).*

Manion G., A technique using monotonic regression splines to enable non-linear transformation of GIS rasters.

1. INTRODUCTION

An increasing number of research techniques in ecology deal with abstract spaces in which the objects of the research are located. The spaces used include habitat spaces with environmental gradients as axes and compositional spaces with gradients of changing species compositions of communities as axes (Gauch 1973).

Two types of non-linearity are commonly found within these spaces. The first is the curvilinear relationship between increasing ecological distance (the distance between two sites on a given environmental gradient), and observed compositional dissimilarity between sites (the number of species shared at both sites). The second is the variation in the rate of compositional turnover (the change in species composition), at different positions along environmental gradients (Ferrier et al 2007).

Most measures of compositional dissimilarity, including the Bray-Curtis index (Bray et al 1957), are constrained between 0 and 1. As the environmental and spatial distance increases between sites, the compositional dissimilarity increases also. This turnover in species composition progresses continuously along the ecological gradient until no species are shared. At this point, the dissimilarity value is 1 and beyond this point the value remains at 1, no matter how far the ecological separation increases.

Data sets within a study that exhibit relatively low rates of compositional turnover, will show lower rates of curvilinear tendencies and can be treated as approximately linear for analysis purposes. However, this approximation is less tenable for data sets exhibiting higher levels of beta diversity, in which case a sizable proportion of sites may share little or no species with one another (Ferrier et al 2007).

The variation in the rate of compositional turnover along environmental gradients at different positions can be illustrated by these two examples. Figure 1a shows a hypothetical environmental gradient for native vegetation where the x-axis refers to mean annual temperature and the y-axis refers to compositional dissimilarity. At the lower end of the gradient there is a greater turnover in species composition relative to the upper end of the continuum the beyond a certain mean annual temperature, there is a greater number of species in common. Conversely figure 1b shows a hypothetical environmental gradient where the x-axis refers to mean annual rainfall and the y-axis refers to compositional dissimilarity. At the lower end of the rainfall gradient there is less composition dissimilarity that at the higher end. If one looked at the compositional dissimilarity over both of the above gradients as analogous to the species turnover occurring along an ecotone from arid woodland up to a tropical rainforest then this non-linearity becomes readily apparent.

2. A DESCRIPTION OF A NOVEL APPLICATION BASED ON THE USE OF I-SPLINES

A recently developed application, Generalised Dissimilarity Modeller (GDM), is a statistical technique for analysing and predicting patterns of turnover in community composition (beta diversity) across large regions. The approach is an extension of matrix regression designed to accommodate the two different types of non-linearity encountered in large-scale data sets discussed previously (Ferrier et al 2007).

Fitting a GDM model to biological and environmental data from a set of survey sites involves the following steps:

- 1 Calculate compositional dissimilarity, d between all possible pairs of survey sites (e.g. using the Bray-Curtis index).
- 2 For each environmental variable, x_p , derive a set of m_p I-spline basis functions and calculate the value of each survey site against each of these functions, $I_{pk}(x_p)$.
- 3 For each of the I-spline basis functions generated in Step 2, calculate the absolute difference in value between sites i and j , $|I_{pk}(x_{pi}) - I_{pk}(x_{pj})|$, for all possible pairs of sites, and save these distances as ΔI_p .

Manion G., A technique using monotonic regression splines to enable non-linear transformation of GIS rasters.

4 Use maximum likelihood estimation to fit coefficients, a_{pk} , to the I-spline basis functions. This can be achieved using the iteratively re-weighted least squares (IRLS) algorithm (McCullagh & Nelder, 1989), with compositional dissimilarity as the response and the series of derived ΔI_{pk} variables as predictors. The only non-standard requirement of this step is that all fitted coefficients must be non-negative to ensure monotonicity in $f_p(x_p)$. This is readily achieved by replacing the least-squares regression normally employed within the IRLS algorithm with a non-negative least-squares regression.

(Ferrier et al 2007).

The model fitting process for GDM uses a response vector derived from the Bray-Curtis distances between site data provided for input. The predictor data for the regression matrix is derived from converting the raw environmental distances between each site pair for all the predictors used as inputs to the GDM. The environmental differences are converted into a set of I-Splines with three or more knots so that the regression matrix will accommodate the curvilinear relationship described in figures 1a and 1b. By applying more than three knots for each predictor a more complex curvilinear relationship can be modelled. The risk however with using too many knots is that the model can be easily 'over-fitted'.

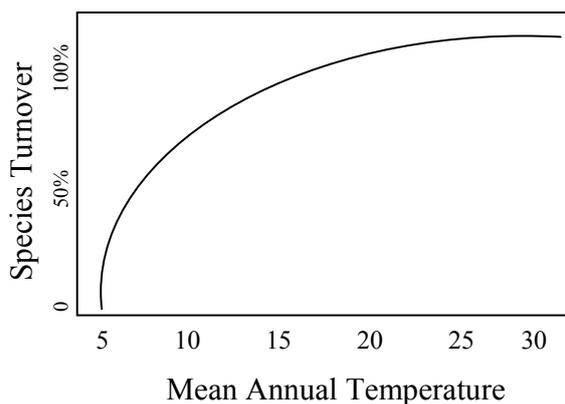


Figure 1a.

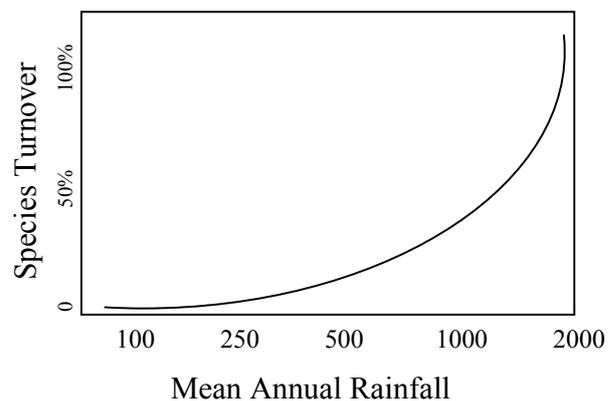


Figure 1b.

The results from a GDM are a set of coefficients that are used to transform the raw predictor grids so that by treating these grids as a stack of matrices, an abstract GDM space is created that will allow simple and useful metrics such as manhattan distance to be used to describe distance in terms of compositional dissimilarity.

For more general non-linear transformations the criteria of non-negativity can be dropped but this characteristic is critical for environmental data modelling in that it needs to follow the ecological principle that these further two sites are apart along an environmental gradient then the less compositional similarity will be exhibited.

3. RASTER SURFACES IN GIS

In geographic information systems, rasters are used to represent continuous layers, such as elevation, slope and aspect, abiotic values such as temperature and rainfall variants, and so on. Rasters are also used for the storage of aerial photo information and remotely sensed information such as multi and hyper spectral data acquired from satellite data.

A raster spatial data model defines space as an array of equally sized cells arranged in rows and columns, essentially, a matrix. Each cell contains an attribute value and location coordinates. Unlike a vector structure, which stores coordinates explicitly, raster coordinates are contained in the ordering of the matrix.

Manion G., A technique using monotonic regression splines to enable non-linear transformation of GIS rasters.

Most industry standard geographic information systems offer map calculators to enable calculation of rasters from other rasters and to apply a suite of mathematical and logical operators. They also have more functional analysis such as those provided by distance, surface, extraction, reclassification and hydrological functions.

However the computation of grids with parameterised splines is not something found in any of the class libraries associated with GIS map calculators.

4. WHAT ARE SPLINES?

The archetypical spline is the polynomial spline, a piecewise polynomial function joined at knots. The original mechanical spline was a flexible piece of wood curved to the desired shape and tacked down at selected knots (Yandel 1993).

The curved section between the knots can be usually approximated by a cubic equation, with segment joined seamlessly by appropriate side conditions on the slope and values to ensure continuity. This is known as a interpolating cubic spline or I-Spline. The continuity characteristics of the spline and the number of independent parameters which define it depend on a knot sequence t which partitions this interval into a number of subintervals.

Because of the well observed monotonically increasing nature of compositional dissimilarity over environmental gradients, the focus of this application is the use of I-Splines as quadratic polynomial functions joined end to end such that an interval $[L,U]$ is subdivided by a mesh Δ consisting of points $L = \xi_1 < \dots < \xi_q = U$. Within any subinterval $[\xi_j, \xi_{j+1})$ the function is a polynomial P_j of specified degree $k - 1$ or order k (Ramsey 1998).

In the case of I-Splines used for GDM, $k = 2$, when the spline I_i will be piecewise quadratic, that is, zero for $x \leq t_i$ and unity for $x \Rightarrow t_{i+2}$. For $x > t_i$ and $x < t_{i+2}$ we have the following direct expressions:

$$I_{i-1}(x/2, t) = 1 - \frac{(t_{i+1} - x)^2}{(t_{i+1} - t_i) * (t_{i+1} - t_{i-1})}$$

$$I_i(x/2, t) = \frac{(x - t_i)^2}{(t_{i+1} - t_i) * (t_{i+2} - t_i)}$$

The useful preliminary knot placement is to position the single interior knot at the median when the most basic form of three splines are used, or two interior knots at the terciles (the 33% and 66% positions), when four knots are used. When running a GDM with a number of predictors one should be careful about exceeding these knot counts as over-fitting the model may result.

5. USING SPLINES TO TRANSFORM RASTER SURFACES

The transformation of a raster using I-Splines requires a non-negative coefficient for each knot position. The parameters required are therefore

1: number of knots

2: for each knot, a data position along the continuum spanned by the raster data. The first spline should be at the minimum value and the last at the maximum value. The interior knots can be located anywhere between the minimum and maximum.

3: for each knot, a transformation coefficient greater than or equal to zero.

Manion G., A technique using monotonic regression splines to enable non-linear transformation of GIS rasters.

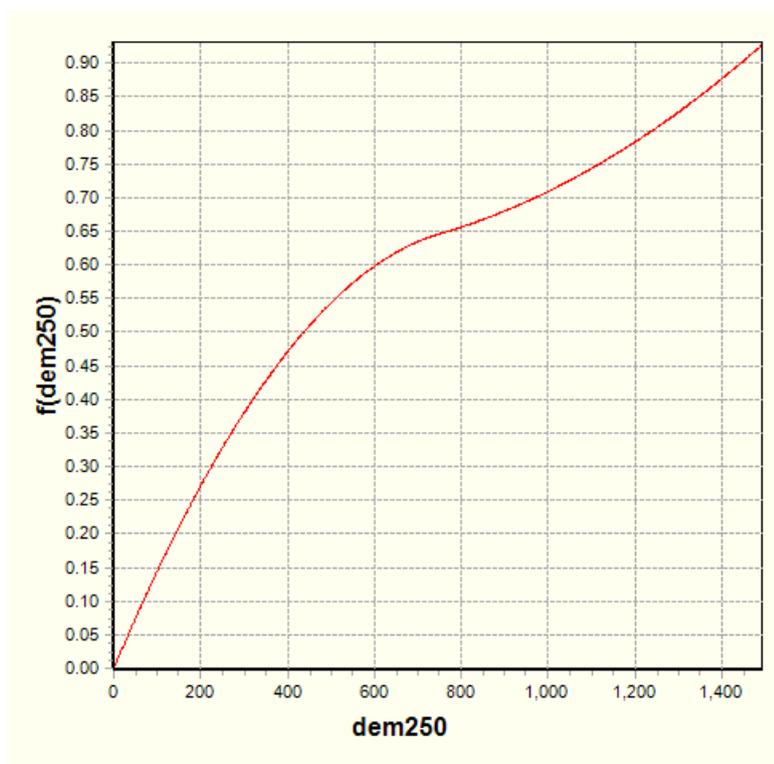
Computing I-Splines is done using three quantile positions at a time. These three positions are progressively moved along the data continuum so that all sub-intervals are addressed. Given that the piecewise polynomials are treated as linear terms, the final value attributed to each raster cell will be the sum of these derived spline calculations.

For each group of quantiles, $[q1, q2, q3]$, in order to ensure full continuity across the data continuum, the first spline calculation will use the $[q1, q1, q2]$ values for the knot positions, The second spline calculation will use the $[q1, q2, q3]$ positions and the last spline calculation will use the $[q2, q2, q3]$ knot positions.

If the transformation coefficients were derived from data that did not fully span the raster that is being transformed, then the result will be values that are either zero for values less than the minimum quantile or $1 * \text{coefficient}$ for values that exceed the maximum quantile. This may not be an optimal result for the transformed grid. A strategy that could be applied to the transformed grid would be to derive the slope of the first or final 10% of the spline and extrapolate the minimum or maximum spline value for raster values that are outside the quantile range.

6. AN EXAMPLE OF A RASTER PRE AND POST TRANSFORMATION USING THE DECC GENERALISED DISSIMILARITY MODELLER (GDM) APPLICATION

The following examples of raster transformation were produced with the DECC Generalised Dissimilarity Modeller described in section 2. This is a simple model for vascular plant distribution in northern NSW, Australia. It used a 250 metre grid-cell resolution digital elevation model (DEM) as a single predictor.



The X-Axis describes the original elevation values in metres that are to be transformed.

The Y-Axis describes the resulting transformed elevation values in units of Bray-Curtis Dissimilarity.

The DEM predictor was modelled with three splines positioned at the minimum, midpoint and maximum elevation values for the DEM.

The GDM produced the following three coefficients to apply to the original DEM data to best fit the model.

Percentile=0%
Value: 0.000000
Coefficient: 0.575507

Percentile=50%
Value: 786.000000
Coefficient: 0.140679

Percentile=100%
Value: 1572.000000
Coefficient: 0.213676

Figure 2. A screen shot from the GDM package showing the transformation that is to take place on the DEM to best fit the model. This plot describes the predicted Bray-Curtis dissimilarities for vascular plants along an elevation gradient defined by the DEM.

Manion G., A technique using monotonic regression splines to enable non-linear transformation of GIS rasters.

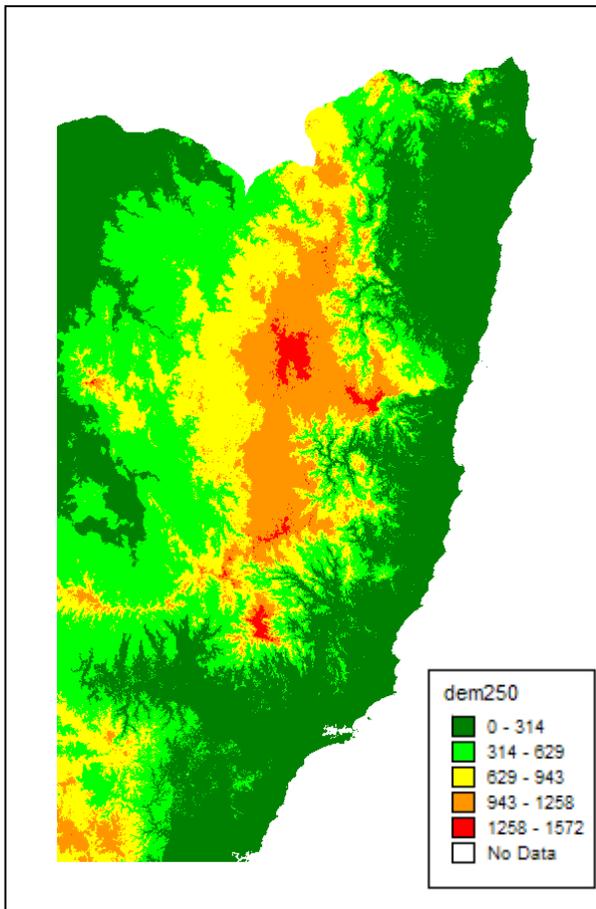


Figure 3a.

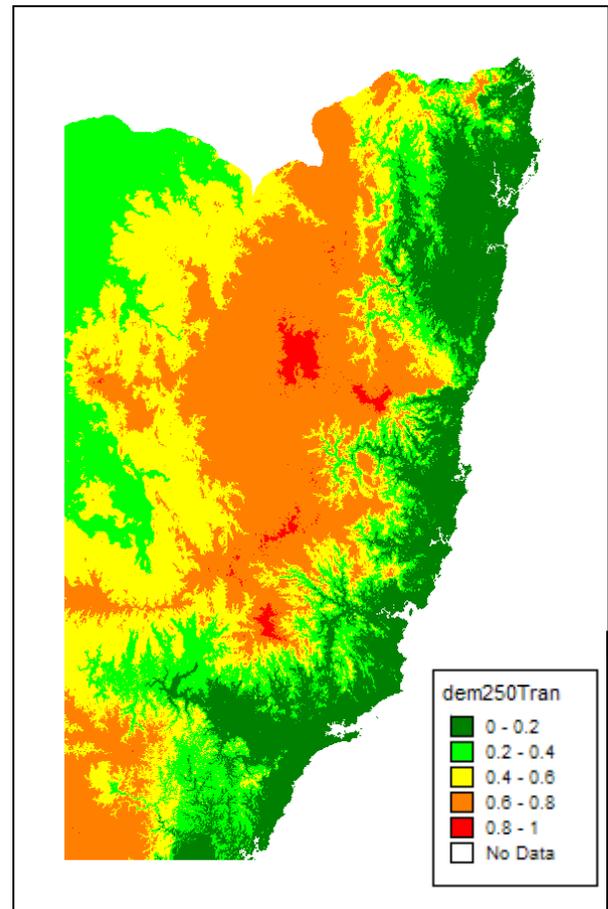


Figure 3b.

Figure 3a shows the original DEM raster and Figure 3b shows the same raster transformed by the coefficients described in the GDM.

7. CONCLUSIONS AND DISCUSSION

The example described in the paper for the non-linear transformation of rasters is a simple application of the use of I-Splines to GIS data. GDM uses matrix regression to derive the I-Spline coefficients for best fitting the model. It also constrains the coefficients to be non-negative to achieve a monotonically increasing relationship. In the GDM package, the raster surfaces produced by transforming the original predictor grids are used for species prediction and classification purposes. Using the transformed grids enables easy calculation of species assemblage dissimilarities. This technique allows for any number of predictors to be modeled collectively but for this example I have only included the single DEM predictor.

However, the technique is not limited to non-negative coefficients, as the same method can produce non-linear transformation of virtually any form. It is the actual coefficient values that define this.

The method can be adapted for using I-Splines to transform rasters to any shape by determining appropriate I-Spline numbers and position that will best allow derivation of meaningful coefficients. This subject is beyond the scope of this paper but points to the development of an interface that uses the distribution of data in the raster to be transformed in conjunction with methods to construct the appropriate transformation graphically and then derive coefficients or to read in a predefined set of coefficients.

In conclusion, I-Splines offer a great amount of scope to derive curvilinear transformations which can be directly applied to GIS rasters for the creation of abstract spaces.

Manion G., A technique using monotonic regression splines to enable non-linear transformation of GIS rasters.

REFERENCES

Bray J. R., Curtis J.T., (1957) An ordination of the upland forest of the southern Wisconsin. *Ecological Monographies*, **27**: 325-349.

Environmental Systems Research Institute Inc, (1996) *ArcView Spatial Analyst: Advanced Spatial Analysis Using Raster and Vector Data A*: 131-139.

Ferrier S., Manion G., Elith., Richardson K., (2007) Using generalised dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, **13**: 252-264.

Gauch, H. G., (1973) The relationship between sample similarity and ecological distance. *Ecology*, **54**: 618-622.

Hearn D., Baker M. P., (1986) *Computer Graphics: Prentice Hall*, **10**: 200-204.

Ramsay J. O., (1998) Monotone Regression Splines in Action. *Statistical Science* **3.4**: 425-461.

Winsberg S., De Soete G., (1997) Multidimensional scaling with constrained dimensions: CONSCAL. *British Journal of Mathematical and Statistical Psychology* **50**: 55-72.

Yandell B. S., Smoothing splines – a tutorial. *The statistician* **42**: 317-319.