Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology & Evolution doi: 10.1111/j.2041-210X.2010.00048.x

Python based GIS tools for landscape genetics: visualising genetic relatedness and measuring landscape connectivity

Thomas R. Etherington*†

The Food and Environment Research Agency, Sand Hutton, York YO41 1LZ, UK

Summary

1. Landscape genetics is an area of research that can help to understand many spatial ecological processes, but requires significant interdisciplinary collaboration. Use of geographic information system (GIS) software is essential, but requires a degree of customisation that is often beyond the non-specialist.

2. To help address this, a series of Python script based GIS tools have been developed for use in landscape genetics studies.

3. The scripts convert files, visualise genetic relatedness, and measure landscape connectivity using least-cost path analysis. The scripts are housed in an ArcToolbox that is freely available along with the underlying Python code.

4. The Python scripts allow researchers to use more current software, provide the option of further development by the user community, and reduce the amount of time that would be spent developing common solutions.

Key-words: ArcGIS, geographic information system, least-cost modelling, Python

Landscape genetics is a rapidly growing area of research (Holderegger & Wagner 2006) that contributes to a better understanding of many spatial ecological processes by combining knowledge from population genetics, landscape ecology, and spatial analysis to conduct 'research that explicitly quantifies the effects of landscape composition, configuration and matrix quality on gene flow and spatial genetic variation' (Storfer *et al.* 2007).

Although landscape genetics requires interdisciplinary collaboration (Holderegger & Wagner 2006), such collaboration remains a major challenge for the future (Balkenhol *et al.* 2009). For example, landscape connectivity, as a species-specific measure of how a landscape facilitates or impedes movement between two locations on that landscape (Tischendorf & Fahrig 2000), is of interest in landscape genetics. For a sample of genetic data, pair-wise relatedness values can be compared against pair-wise landscape connectivity values to assess if relatedness can be explained by landscape connectivity. This ability to quantify the connectivity of a landscape for a given species as represented in the genetics of a population is of great interest to ecologists concerned with a wide range of issues such as habitat fragmentation, invasive species, or wildlife diseases.

Landscape connectivity can be measured using geographic information system (GIS) software, but for direct application in landscape genetics work, a degree of customisation is required that is beyond the non-GIS specialist. As seen with the PATHMATRIX GIS extension for ArcView 3.x (Ray 2005), such customisation is possible and popular, but other solutions are required to take advantage of other and newer GIS software.

With this in mind I have used the open source programming language Python (version 2·5) to produce thirteen scripts (Table 1) for use in landscape genetics. These scripts have a variety of software requirements, with the GIS functionality provided by ArcGIS software (version 9·3). I chose to use ArcGIS as although a commercial product, the system is popular and widespread, and allows for the scripts to be housed within an ArcToolbox so that scripts can be run as tools through simple dialogue boxes. This makes the scripts accessible to the non-GIS specialist, and allows them to be readily linked with other ArcGIS tools to automate larger workflows. However, as Python is an open source language, and the script code is accessible, other GIS specialists can not only review, modify, or develop the scripts for their own purposes, but could even

*Correspondence author. E-mail: teth001@aucklanduni.ac.nz
†Present address: School of Environment, The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
Correspondence site: http://www.respond2articles.com/MEE/

Table 1. The Python script tools developed with their function, and the type and level GIS software required

Tool	Function	GIS software required		
		ArcGIS	Spatial analyst	ArcInfo
Genetics text file conversions				
Convert raw lists to matrix	Converts a raw text file of relatedness values in pair-wise column format into a text file matrix of relatedness values.			
Convert raw matrix to matrix	Converts a raw text file of relatedness values in full matrix or half matrix format into a text file matrix of relatedness values.			
KINGROUP to matrix MSA to matrix SPAGeDi to matrix	These tools convert outputs from the KINGROUP (Konovalov, Manning, & Henshaw 2004), MSA (Dieringer & Schlötterer 2003), and SPAGeDi (Hardy & Vekemans 2002) genetic analyses programs to make a text file matrix of relatedness values.			
Genetics visualisations				
Kinship links	Produces a shapefile of polylines that can be used to visualise genetic relatedness between pair-wise combinations of points.	✔		
Landscape connectivity				
Distance matrix	Produces a matrix of Euclidean distances between each pair-wise combination of points.	✔		
Cost-distance	Produces a matrix of cost-distances between each pair-wise combination of points.	✔	✔	
Least-cost paths	Produces a matrix of cost-distances and least-cost path lengths between each pair-wise combination of points. A polyline shapefile of the least-cost paths is also produced.	✔	✔	
Landscape separation				
Line barrier matrix Zone barrier matrix	These tools produce a matrix showing which pair-wise combinations of points are separated by identified barrier features. Separation is defined either using a straight-line intersection with barriers, or by creating landscape zones.	✔ ✔		✔
Matrix conversion				
Log transform matrix	Log_e transforms values in a matrix, which can be useful in some contexts (Rousset 1997).			
Matrix to pairs	Coverts data in a matrix format to a pair-wise column format.			

substitute the ArcGIS functions for equivalent functions from another GIS product.

Some of the tools perform useful but basic functions such as file conversion from popular genetic analysis program output files to the core file format used by the scripts, which is a text file containing a tab-delimited matrix of values. Additional scripts will convert other output file formats with some prior manual manipulation, transform data values within matrices, and convert data in matrix format to a pair-wise listing format. As these tools are straightforward to use, and are summarised in Table 1, I will concentrate on highlighting the more novel or complicated scripts, giving examples of where the approaches have been used in previous landscape genetics studies so that users can examine detailed applications elsewhere.

While working on landscape genetics for the first time I was surprised to find that there was no way to visualise differences in pair-wise genetic relatedness. As with any kind of spatially explicit analysis, the ability to visualise your data in the context of its landscape is very useful when data is being explored and hypotheses developed. To rectify this, the *Kinship Links* script will take a series of points and a text file matrix of kinship values, and will produce a polyline shapefile of links between each pair-wise combination of points. These lines can then be symbolised based on the kinship value to try and illustrate where the stronger and weaker genetic links are in reference to other landscape data (Fig. 1a). A threshold level for relatedness can also be specified in order to produce links between pairs of highly related or unrelated individuals. This function has proven to be useful in helping to identify potential landscape barriers to wildlife (Frantz *et al*. 2010).

The landscape connectivity tools use least-cost modelling (Adriaensen *et al*. 2003), a method based upon a friction surface, which is a raster GIS map for which each cell describes the permeability of different parts of the landscape being studied to the species of interest. This friction surface can be used to derive the most efficient route, called a least-cost pathway

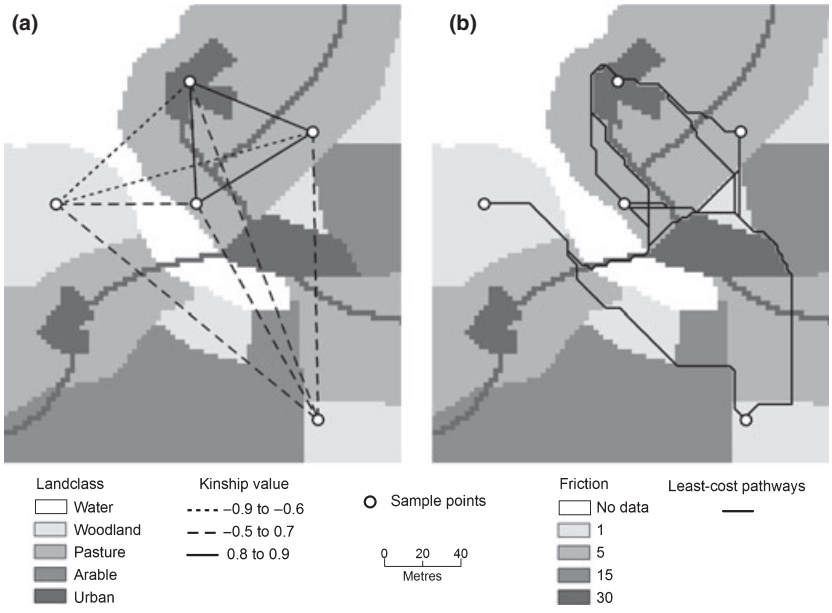

Fig. 1. The test data supplied with the ArcToolbox can be used with a) the *Kinship links* tool to create lines between sample points which can be symbolised based on the strength of each pair-wise kinship value. This allows for the spatial visualisation of pair-wise relatedness to try and identify genetic patterns in relation to the landscape. From interpretation of the kinship links it is possible to b) assign sensible friction values to the landscape and use the *Least-cost path* tool to generate a least-cost pathway (LCP) between each pair-wise combination of sample points to see if variation in LCP is correlated to variation in genetic relatedness.

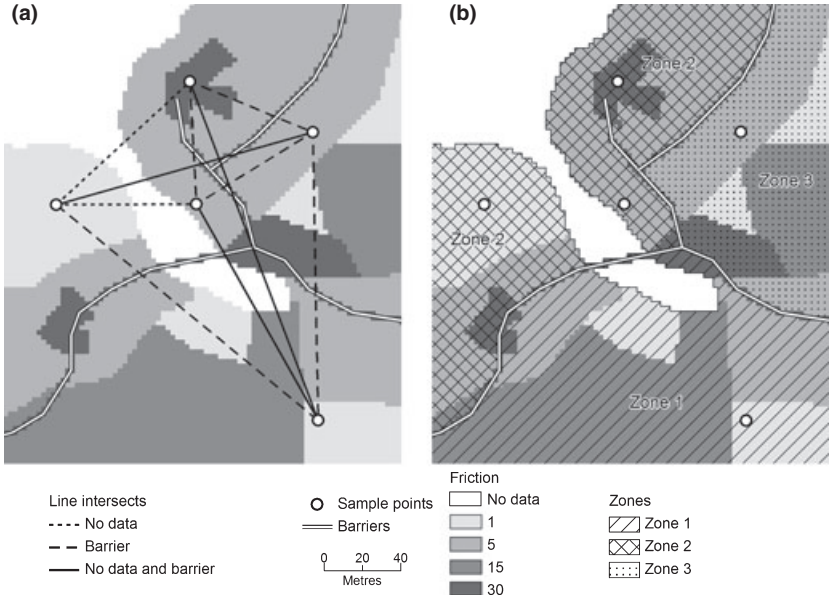

Fig. 2. Two tools can be used to identify samples separated by barriers. The *Line barrier matrix* tool classifies separation if a) a straight line between the samples intersects friction surface no data cells and or linear barriers. The *Zone barrier matrix* tool classifies separation if b) samples are in different zones that are created by partitioning the friction surface based on no data cells and linear barriers. As can be seen with Zone 2, which is only one zone because of a thin strip of split cells, careful attention needs to be paid to the zones created to ensure they are ecologically meaningful.

(LCP), which balances distance and friction between genetic sample locations (Fig. 1b). Use of LCP modelling has become a core method of landscape genetics studies, with genetic relatedness being compared to connectivity measured in terms of the LCP cost-distance value, which is a combination of the distance that would be travelled and the cost of the landscape friction traversed (Stevens *et al*. 2006; Walker, Novaro, & Branch 2007), the LCP length (Broquet *et al*. 2006; Wang *et al*. 2008), or both (Lada *et al*. 2008). These measures of connectivity can be calculated using two tools. The *Cost-distance Matrix* tool produces a text file matrix of pair-wise cost-distance values, while the *Least-cost Paths* tool, produces a polyline shapefile of pair-wise LCPs as well as text file matrices of both LCP cost-distance and LCP length. These are computationally demanding tools, for which the number of samples is most important in determining processing times. For example, when processing 10 samples with the *Cost-distance Matrix* tool on a standard desktop computer, increasing the friction surface size from 1×10^4 cells to 1×10^6 cells triples the processing times, but total time remains on a scale of tens of seconds, whereas increasing the number of samples from 10 to 50 will shift processing times from a scale of seconds to minutes. The *Least-cost Paths* tool responds to changes in number of cells similarly, but the same increase in sample size shifts processing times from a scale of minutes to hours. Progress is reported during processing so total-processing times can be easily gauged.

I would advise that it is worthwhile generating LCP polylines, at least initially, even if only for a subset of samples, as it allows the LCPs to be visualised in the context of the landscape to check for any obvious flaws in the friction surface. For instance once the LCPs are plotted, it may become apparent that landscape features that are supposed to act as barriers are not doing so. This could be because the value used in the

friction surface is too low, or because there are small gaps through which an LCP can pass. These kinds of flaws would not be evident if just a text matrix of cost-distance values were created with the *Cost-distance Matrix* tool.

In addition to measures of connectivity, the presence of potential barriers between pair-wise combinations can form the basis of an analysis to determine if a landscape feature is a barrier (Frantz *et al*. 2010), and can also be incorporated into analyses of landscape connectivity to control for the effect of barriers on gene flow (Epps *et al*. 2007). There are two tools that produce a matrix file that identifies pairs separated by identified barriers within the extent of a friction surface. The *Line Barrier Matrix* draws a straight line between all pair-wise combinations of samples and classifies a pair as separated if the straight line intersects either a no data cell in the friction surface or barrier feature (Fig. 2a). This separation based on a straight line can be misleading. For instance a pair of samples could be on the same side of a river but a straight line between the two could intersect a meander. Therefore, the *Zone Barrier Matrix* tool provides an alternative as it partitions the friction surface up into separate zones, and classifies a pair as separated if the samples are within different zones (Fig. 2b). The friction surface is partitioned into zones based on linear barriers and no data cells within the friction surface. Both tools produce shapefiles to help visualise and interpret these results.

These Python scripts are proving very useful in my own research, and so will hopefully be of interest to other researchers. They allow researchers to use more current software, provide the option of further development by the user community, and reduce the amount of time that would be spent developing common solutions. The Python scripts and ArcGIS Toolbox are freely available along with test data and further user guidance, and can be downloaded from a persistent website at http://purl.org/NET/python_land_gen/arcgis_toolbox.

Acknowledgements

A. Walker at Fera helped with code checking, and D. Lieske, C. Morrison, and J. Thomas at Mount Allison University provided user testing. A. Frantz of the University of Sheffield and N. Balkenhol of the University of Idaho advised on integration of genetics data. This work was partly supported by seedcorn funding from the Food and Environment Research Agency, an executive agency of the Department for Environment Food and Rural Affairs.

References

Adriaensen, F., Chardon, J.P., De Blust, G., Swinnen, E., Villalba, S., Gulinck, H. & Matthysen, E. (2003) The application of 'least-cost' modelling as a functional landscape model. *Landscape and Urban Planning*, **64**, 233–247.

Balkenhol, N., Gugerli, F., Cushman, S.A., Waits, L.P., Coulon, A., Arntzen, J.W., Holderegger, R. & Wagner, H.H. (2009) Identifying future research needs in landscape genetics: where to from here? *Landscape Ecology*, **24**, 455–463.

Broquet, T., Ray, N., Petit, E., Fryxell, J.M. & Burel, F. (2006) Genetic isolation by distance and landscape connectivity in the American marten (*Martes americana*). *Landscape Ecology*, **21**, 877–889.

Dieringer, D. & Schlötterer, C. (2003) Microsatellite Analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes*, **3**, 167–169.

Epps, C.W., Wehausen, J.D., Bleich, V.C., Torres, S.G. & Brashares, J.S. (2007) Optimizing dispersal and corridor models using landscape genetics. *Journal of Applied Ecology*, **44**, 714–724.

Frantz, A.C., Pope, L.C., Etherington, T.R., Wilson, G.J. & Burke, T. (2010) Using isolation-by-distance-based approaches to assess the barrier effect of linear landscape elements on badger (*Meles meles*) dispersal. *Molecular Ecology*, **19**, 1663–1674.

Hardy, O.J. & Vekemans, X. (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Holderegger, R. & Wagner, H.H. (2006) A brief guide to landscape genetics. *Landscape Ecology*, **21**, 793–796.

Konovalov, D.A., Manning, C. & Henshaw, M.T. (2004) KINGROUP: a program for pedigree relationship reconstruction and kin group assignments using genetic markers. *Molecular Ecology Notes*, **4**, 779–782.

Lada, H., Thomson, J.R., Mac Nally, R. & Taylor, A.C. (2008) Impacts of massive landscape change on a carnivorous marsupial in south-eastern Australia: inferences from landscape genetics analysis. *Journal of Applied Ecology*, **45**, 1732–1741.

Ray, N. (2005) PATHMATRIX: a geographical information system tool to compute effective distances among samples. *Molecular Ecology Notes*, **5**, 177–180.

Rousset, F. (1997) Genetic differentiation and estimation of gene flow from *F*-statistics under isolation by distance. *Genetics*, **145**, 1219–1228.

Stevens, V.M., Verkenne, C., Vandewoestijne, S., Wesselingh, R.A. & Baguette, M. (2006) Gene flow and functional connectivity in the natterjack toad. *Molecular Ecology*, **15**, 2333–2344.

Storfer, A., Murphy, M.A., Evans, J.S., Goldberg, C.S., Robinson, S., Spear, S.F., Dezzani, R., Delmelle, E., Vierling, L. & Waits, L.P. (2007) Putting the 'landscape' in landscape genetics. *Heredity*, **98**, 128–142.

Tischendorf, L. & Fahrig, L. (2000) On the usage and measurement of landscape connectivity. *Oikos*, **90**, 7–19.

Walker, R.S., Novaro, A.J. & Branch, L.C. (2007) Functional connectivity defined through cost-distance and genetic analyses: a case study for the rock-dwelling mountain vizcacha (*Lagidium viscacia*) in Patagonia, Argentina. *Landscape Ecology*, **22**, 1303–1314.

Wang, Y.H., Yang, K.C., Bridgman, C.L. & Lin, L.K. (2008) Habitat suitability modelling to correlate gene flow with landscape connectivity. *Landscape Ecology*, **23**, 989–1000.

Received 21 April 2010; accepted 02 June 2010
Handling Editor: Robert P. Freckleton