GenAlEx Tutorial 3

Spatial Genetic Analysis

Based on material provided at the national graduate workshop *Genetic Analysis for Populations Studies* offered by Rod Peakall and Peter Smouse at the Australian National University, Canberra, Australia, July 2009.

Table of Contents

About this Tutorial Module **3**
Goals of the Tutorial **3**
Principal Coordinate Analysis (PCA) **4**
Ex 3.1 Steps for performing PCA **4**
Q 3.1 Questions 4
Mantel Tests for Matrix Correspondence **5**
Ex 3.2 Hand Calculation of Mantel Tests **5**
Box 3.1 Mantel Test Formulae 8
Ex 3.3 Mantel Tests for Isolation-by-Distance **8**
Q 3.3 Questions 9
Ex 3.4 Statistical Testing for Mantel **10**
Q 3.4 Questions 11
Box 3.2 Calculation of Geographic Distance in GenAlEx 11
Ex 3.5 Advanced Mantel (Optional) **12**
Global Spatial Autocorrelation **13**
Ex 3.6 Hand Calculation of Spatial Autocorrelation **13**
Q 3.6. Questions 16
Box 3.3 Autocorrelation Formulae 18
Ex 3.7 Spatial Autocorrelation Analysis in GenAlEx **18**
Q 3.7. Questions 20
Box 3.4 Understanding Distance Classes 20
Statistical Testing for Spatial Autocorrelation **20**
Ex 3.8 Statistical Tests for Autocorrelation (Optional) **21**
Q 3.8 Questions 22
Ex 3.9 Spatial Analysis of Multiple Loci **22**
Q 3.9 Questions 23
Ex 3.10 Alternative Outputs to Correlograms **23**
Q 3.10 Questions 25
Box 3.5 Correlograms and the meaning of the X-intercept 25
Local 2D Spatial Autocorrelation Analyses **26**
Ex 3.11 Understanding 2D Spatial Analyses **26**
Q 3.11 Questions 27
Statistical Testing for 2D LSA **28**
Box 3.6 2D LSA - How many neighbours? 28
References and Further Reading **29**
Appendix 3.1 – Advanced Spatial Analysis Options **30**

About this Tutorial Module

This GenAlEx tutorial is the third in a series of modules. It is strongly recommended that you complete the first and second module before attempting this third one since important background to both the underlying statistics and the use of GenAlEx is provided in the earlier modules. Furthermore, the first module includes a glossary of terms that are used throughout all of the modules.

This document is based on material provided at a five-day national graduate workshop entitled *Genetic Analysis for Populations Studies* offered by Rod Peakall (Australian National University) and Peter Smouse (Rutgers University, USA) at the Australian National University in July 2009.

Rod Peakall and Peter Smouse, Aug 2012

Professor Rod Peakall

Evolution, Ecology and Genetics
Research School of Biology
The Australian National University
Canberra ACT 0200 Australia
Email: rod.peakall@anu.edu.au

Professor Peter Smouse

Department of Ecology, Evolution and Natural Resources. School of Environmental and Biological Sciences, Rutgers University
New Brunswick NJ 08901-8551 USA
Email: smouse@AESOP.Rutgers.edu

Goals of the Tutorial

1. To introduce Principal Coordinate Analysis (PCoA) as a method for visualizing variation among elements in a matrix.
2. To demonstrate by way of hand calculations how to perform Mantel Tests for Matrix Correspondence and Spatial Autocorrelation Analysis.
3. To learn how to use the software GenAlEx to perform PCoA and conduct Mantel Tests and Spatial genetic analysis.
4. To explore the biological interpretation of the statistics described for some real data sets.

Principal Coordinate Analysis (PCoA)

Matrices such as the pairwise *PhiPT* matrix shown below for 10 populations of bush rats can be difficult to read and interpret. Ideally what we need is a way of visualizing the patterns of genetic relationship contained in such a matrix. Principal Coordinate Analysis (PCoA) provides such a tool. PCoA represents a procedure by which the essential patterns can be revealed, without actually altering the data itself. In this respect PCoA is unlike Tree building methods, such as UPGMA and neighbour joining, where the algorithms always assume a hierarchical genetic structure. While this is a reasonable assumption at higher taxonomic levels, it is not necessarily true at the population level. Another disadvantage of Tree building methods for population analysis is that often the number of samples analysed is vast, making the generation of Trees unwieldy, and the interpretation difficult. On the other hand, Tree building methods do offer some procedures for statistical testing of the clusters, such as bootstrap tests, that are not available with PCoA.

Pairwise population *PhiPT* values among 10 bush rat populations

CAM5	CAMM	335	U3	T3	B5	1875	1875A	MC	MD	
0.000										**CAM5**
0.068	0.000									**CAMM**
0.052	0.089	0.000								**335**
0.076	0.071	0.059	0.000							**U3**
0.058	0.057	0.073	0.013	0.000						**T3**
0.122	0.128	0.128	0.070	0.072	0.000					**B5**
0.112	0.125	0.128	0.079	0.068	0.093	0.000				**1875**
0.097	0.152	0.119	0.090	0.096	0.118	0.048	0.000			**1875A**
0.125	0.120	0.150	0.100	0.116	0.128	0.087	0.097	0.000		**MC**
0.138	0.161	0.154	0.083	0.105	0.119	0.084	0.053	0.077	0.000	**MD**

In brief, PCoA is a multivariate technique that allows one to find and plot the major patterns within a multivariate data set (e.g., multiple loci and multiple samples). The mathematics is complex, but in essence PCoA is a process by which the major axes of variation are located within a multidimensional data set. Each successive axis explains proportionately less of the total variation, such that when there are distinct groups, the first 2 or 3 axes will typically reveal most of the separation among them.

Ex 3.1 Steps for performing PCoA

In this course we will leave the complex mathematics behind PCoA to GenAlEx. In which case, all that is needed is an appropriately formatted distance matrix. Here we will use PCoA to visualize the genetic relationships among the 10 populations of bush rats.

Step 1. Open the workbook *Ex 3.1 PCA Rats.xls* and locate then activate the *PhiPTP* worksheet.

Step 2. Choose ***PCoA->Analysis***, then accept the default options of *TriDistance Matrix*, *Covariance-Standardized* and *Data Labels.*

Step 3. Inspect the outcomes of the PCoA analysis and answer question 1.

Step 4. Compare the PCoA plot with the map of the population locations in the worksheet *Pop Coords* and answer questions 2 & 3.

Q 3.1 Questions

1. Summarise the outcomes of your PCoA analysis in words. How well does this PCoA plot represent the original data?

2. Does the PCoA plot suggest genetic structure in bush rats? How would you test for this pattern?

3. Do you predict 'Isolation by Distance' for this data set? How would you test this hypothesis?

Mantel Tests for Matrix Correspondence

Over a larger geographic scale it is unlikely that the null hypothesis (*H0*) of no genetic difference among populations will hold. Instead, we frequently expect to find a positive relationship between geographic and genetic distance – the so called 'Isolation-by-Distance' hypothesis (*H1*). The Mantel test for Matrix Correspondence is an ideal statistical tool for testing this hypothesis.

The Mantel test allows tests for a statistical relationship between the elements of any two distance matrices with matching entries (Mantel 1967). While it is easy to plot a graph of the relationship between elements from any two matrices, we cannot use the *P*-values of standard regression analysis, because the N $(N - 1)/2$ elements of a triangular matrix cannot be independent. Consequently, we need another way to test the significance of two matrices, and the Mantel test provides such an option. This method yields a correlation coefficient for the two data matrices, with a range from –1 to +1, with a test for a significant relationship by random permutation.

In the case of Mantel Test the null (*H0*) and alternative hypotheses (*H1*) are:
H0=No relationship exists between the elements of the *X* and *Y* matrices (Rxy=0)
H1=There is a relationship between the elements of the *X* and *Y* matrices (Rxy>0). This relationship can be either negative or positive.

If the null hypothesis is true, it follows that if we shuffle (randomise) the *Y* data while keeping the *X* data unchanged, and recalculate the Mantel statistic Rxy we should obtain an Rxy close to zero (as expected by chance under the null hypothesis). Because of sampling effects, the results will of course vary from shuffle to shuffle. On the other hand, if we perform multiple shuffles (say 100 or 1000 times) we can obtain a good estimate of the value we would expect if the null hypothesis was true. If our observed value is greater than the permuted values 95% or more of the time, we declare the results significant at the 5% level. That is, if there is significant relationship between the two data sets, the observed correlation will be more extreme (closer to +1 or –1) than the values generated by random permutation, at least 5% of the time.

When testing for isolation by distance don't neglect to also test the relationship between Log(Geographic distance) and genetic distance. If using F_{ST} you should also test linearized F_{ST} (which can be calculated with GenAlEx via AMOVA).

Ex 3.2 Hand Calculation of Mantel Tests

The mathematics of Mantel Tests for Matrix Correspondence is straightforward and can be performed by hand for small data sets. Here we will work through the calculations, step by step, although for ease some of the early steps have been completed for you. Our hypothetical data consist of X and Y matrices for which the elements are intentionally highly correlated.

X Matrix

1	**2**	**3**	**4**	**5**	**6**	
0						**1**
1	0					**2**
4	1	0				**3**
3	1	1	0			**4**
3	1	1	1	0		**5**
4	3	4	3	1	0	**6**

Y Matrix

1	**2**	**3**	**4**	**5**	**6**	
0						**1**
2	0					**2**
5	2	0				**3**
3	2	2	0			**4**
4	2	2	2	0		**5**
6	4	5	4	2	0	**6**

Step 1. Convert distance matrices to table (optional).

Step 2. Calculate the mean of the *X* matrix elements,($\overline{X}$).

Step 3. Calculate the mean of the *Y* matrix elements,($\overline{Y}$).

Step 4. For each element *x* in the *X* matrix, calculate (x- $\overline{X}$) and (x- $\overline{X}$)^2.

Step 5. For each element *y* in the *Y* matrix, calculate (*y*- $\overline{Y}$) and (y- $\overline{Y}$)^2.

Step 6. For each element, calculate the cross product (*x*- $\overline{X}$)(*y*- $\overline{Y}$).

Lower triangular matrices as table excluding diagonal for ease of calculations

Sample i	Sample j	X Matrix	Y Matrix	$(x-\overline{X})$	$(y-\overline{Y})$	$(x-\overline{X})^2$	$(y-\overline{Y})^2$	$(x-\overline{X})(y-\overline{Y})$
1	2	1	2	-1.133	-1.133	1.284	1.284	1.284
1	3	4	5	1.867	1.867	3.484	3.484	3.484
2	3	1	2	-1.133	-1.133	1.284	1.284	1.284
1	4	3	3	0.867	-0.133	0.751	0.018	-0.116
2	4	1	2	-1.133	-1.133	1.284	1.284	1.284
3	4	1	2	-1.133	-1.133	1.284	1.284	1.284
1	5	3	4	0.867	0.867	0.751	0.751	0.751
2	5	1	2	-1.133	-1.133	1.284	1.284	1.284
3	5	1	2	-1.133	-1.133	1.284	1.284	1.284
4	5	1	2	-1.133	-1.133	1.284	1.284	1.284
1	6	4	6	1.867	2.867	3.484	8.218	5.351
2	6	3	4	0.867	0.867	0.751	0.751	0.751
3	6	4	5	1.867	1.867	3.484	3.484	3.484
4	6	3	4	0.867	0.867	0.751	0.751	0.751
5	6	1	2	-1.133	-1.133	1.284	1.284	1.284
Mean		2.133	3.133			***SSx***	***SSy***	***SPxy***
Sum								

Based on the values provided in the table for you, complete steps 7 to 10. Enter your answers in the tables above and below below.

Step 7. Calculate the Sum of Squares for the *X* matrix (*SSx*) as the sum of the $(x-\overline{X})^2$ values.

Step 8. Calculate the Sum of Squares for the *Y* matrix (*SSy*) as the sum of the $(y-\overline{Y})^2$ values.

Step 9. Calculate the Sum of Cross Products (*SPxy*) as the sum of the $(x-\overline{X})(y-\overline{Y})$ values.

Step 10. Calculate *Rxy* as $SPxy/(SSx*SSy)^{0.5}$.

Mantel table for the highly correlated hypothetical data

SSx	***SSy***	***SPxy***	***Rxy***

Step 11. Check your answers using GenAlEx. The workbook containing this data set is called *Ex 3.2 Mantel by Hand.xls.*

Box 3.1 Mantel Test Formulae

$$SSx = \sum_{i \neq j}^{N} \left(x_{ij} - \bar{x}\right)^2$$

$$SSy = \sum_{i \neq j}^{N} \left(y_{ij} - \bar{y}\right)^2$$

$$SPxy = \sum_{i \neq j}^{N} \left(x_{ij} - \bar{x}\right)\left(y_{ij} - \bar{y}\right)$$

$$R_{xy} = \frac{SPxy}{\sqrt{[SSxSSy]}}$$

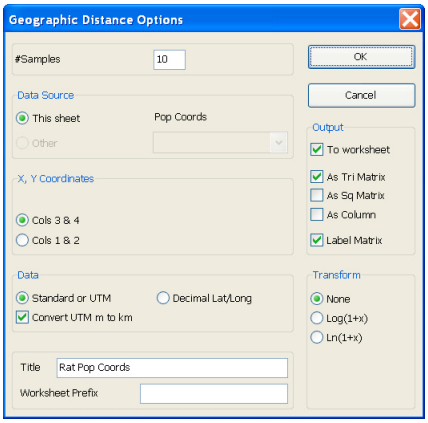
Where *SPxy* is the sum of cross products of corresponding elements of the ***X*** and ***Y*** Matrices; *SSx* is the sum of products of ***X*** matrix elements and *SSy* that of ***Y*** matrix elements (Smouse et al. 1986. Smouse and Long 1992).

Ex 3.3 Mantel Tests for Isolation-by-Distance

In Ex 3.1 we used PCoA to visualise the genetic relationships among 10 populations of bush rats. The findings indicated there may be regional genetic structure (north vs south) and the possibility of isolation by distance. Armed with the Mantel Test procedure, we can now test that hypothesis.

Step 1. Return to Excel and reopen *Ex 3.1 PCA Rats.xls, Save* the workbook as *Ex 3.3 Mantel Rats.xls* and inspect the worksheets.

Step 2. First we need to compute the pairwise Geographic distance matrix (our X matrix). Activate the *Pop Coords* worksheet that contains the geographic coordinates in the universal map grid system (UTM). These units are in metres. Choose ***Distance->Geographic.*** At the Geographic Distance Options dialog box select *Cols 3 & 4*, *Standard*, and *Convert UTM to km*. Check *TriMatrix* and *To worksheet*.



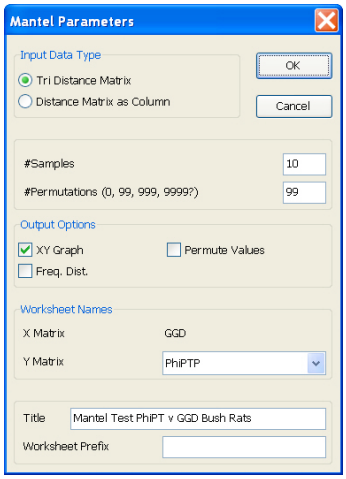
Tip: You can calculate log(geographic distance) by selecting the appropriate transform option in this dialog box.

Step 3. Ordinarily at this step you would need to compute the pairwise genetic distance matrix (our Y matrix). In this case we will use the pairwise *PhiPT* matrix already calculated in the worksheet *PhiPTP*.

While not essential, it is convenient to move the X and Y matrices for Mantel tests to sheet positions 1 and 2, before proceeding. This will simplify the selection of worksheets when prompted by GenAlEx.

Step 4 Activate the worksheet [GGD] containing the geographic distance matrix. This will become the *X* matrix, then choose ***Mantel->Paired***. At the Mantel Parameters dialog box,

choose the worksheet [PhiPTP] as the *Y* matrix from the drop down worksheet list and select *XY graph*.



Step 5. Inspect the outcome of your analysis in the worksheet [MT]. Complete the table and then answer the questions below.

Mantel table for a test of isolation by distance in bush rats

	SSx	***SSy***	***SPxy***	***Rxy***
GGD v PhiPT				

Step 6. If time permits you may wish to try other Mantel Tests such as Log(1+GGD) vs *PhiPT* and Log(1+GGD) vs *Linearized Fst*. Extra rows are provided above to record your results.

Step 7. If you completed step 6 you might also try using Mantel Tests to evaluate the relationship between *PhiPT* and *Fst*.

Q 3.3 Questions

1. Summarise the outcomes of your Mantel test in words. Include the value of *Rxy* in your summary, and comment on its magnitude.

2. What factors might explain the variance about the predicted linear relationship?

3. Why did we use the geographic distance as the *X* matrix? (Which is the independent axis?). What happens if you reverse the matrices. Do your answers change?

Ex 3.4 Statistical Testing for Mantel

This exercise will revise the basics of random permutation and demonstrate how this tool for statistical tests can be applied to Mantel. We have already noted that within AMOVA the shuffling of sample data can be used to generate a distribution of *PhiPT* values typical of those expected under the null hypothesis of no genetic difference among the populations. Can we use the same trick for Mantel? How can we shuffle data in a distance matrix? Do we need to shuffle both *X* and *Y* matrices?

It turns out that for Mantel statistical testing we can randomly permute the *Y* matrix (each permute shuffles the rows and columns of the matrix) calculating an *Rxy* value each permutation. Next we compare the frequency distribution of the permuted *Rxy* values with the *Rxy* for the original data.

Step 1. Return to *Ex 3.3 Mantel Rats.xls* and activate the [PhiPTP] worksheet. Next use ***Rand Data->Shuffle Tri*** to create a new worksheet called [ShuffleTri]. Inspect the new data matrix before proceeding.

Step 2. Re-activate the [GGD] worksheet containing the geographic distance matrix (*X* matrix), then choose ***Mantel->Paired***. At the Mantel Parameters dialog box, select the worksheet [ShuffleTri] as the *Y* matrix from the drop down worksheet list and *XY graph*. For now you can set the number of permutations to zero.

Step 3. Repeat steps 1 and 2, four more times, recording the *Rxy* values obtained in the table below. Calculate the mean *Rxy* value for the five shuffles. Answer question 1 before proceeding.

Run	***Rxy***
Original Data	0.707
ShuffleTri 1	
ShuffleTri 2	
ShuffleTri 3	
ShuffleTri 4	
ShuffleTri 5	
Mean Rxy of Shuffles	

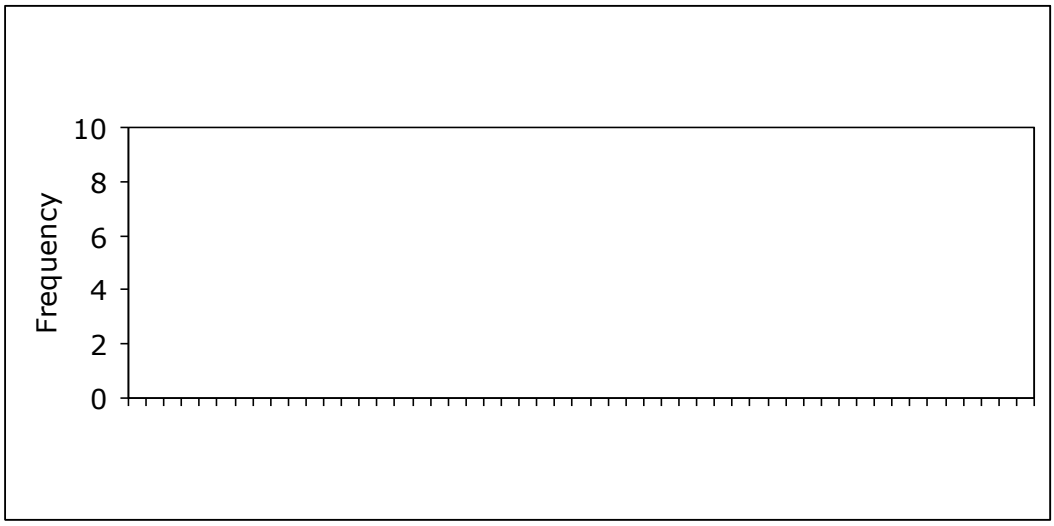
Step 4. Activate the original X matrix worksheet [GGD] and repeat the Mantel analysis. At the Mantel Parameters dialog box set the number of permutations to 99, and select *XY Graph, Freq. Dist.* and *Permute Values.* Choose *PhiPTP* as the *Y* matrix.

Step 5. Inspect the results of your analysis. The outcomes are presented in the 3 worksheets: [MT], [FDMT] and [PVMT].

Step 6. In the worksheet [PVMT] a sorted list of the *Rxy* values for the 99 permutations and the observed value is shown. Scroll down the list and locate the position of the observed value (*Rxy*=0.707). Calculate the Probability as: *Number of Random Values => Observed Value (Including Observed Value)/Number of Permutations + 1*. (Where '=>' means 'greater than or equal to')

Step 7. Compare your calculated probability with the outcome in GenAlEx (listed in the Mantel summary table in the worksheet [MT]).

Step 8. In the worksheet [FDMT], a graph showing the 'Frequency Distribution of Permuted Rxy versus Observed Rxy' is shown. Draw a schematic of this graph below, include a title and axis labels. Next answer question 2.



Q 3.4 Questions

1. How does the mean *Rxy* value across the shuffles compare with the *Rxy* for the original data?

2. Is the hypothesis of 'Isolation-by-Distance' supported in this study of bush rats? What does this finding tell us about the biology of bush rats? Given the scale of this study (<25 km), did you expect this result?

Box 3.2 Calculation of Geographic Distance in GenAlEx

GenAlEx calculates geographic distance among standard coordinates by the standard formula:

$$D = \sqrt{(xi - xj)^2 + (yi - yj)^2}$$

Where, *xi and yi* are the coordinates for the *i*-th sample and *xj* and *yj* are the coordinates for the *j*-th sample. In this case the distance units will depend on the units in which the coordinates were entered. If you use UTM coordinates from a GPS (or read directly from a map) the distance units will be in m's, unless you choose the GenAlEx option to convert m's to km's

GenAlEx also offers the option to calculate distance from Lat/Long coordinates in decimal degrees. Be sure to use negative and positive values where appropriate if your coordinates span across zero degrees latitude (the equator) or across zero degrees longitude (the Greenwich line). GenAlEx uses a modification of the Haversine Formula developed by R.W. Sinnott, "Virtues of the Haversine", Sky and Telescope, vol. 68, no. 2, 1984, p. 159 following computer code published in the online document by Bob Chamberlain from JPL, NASA (As of 8-9-05 available from http://www.usenet-replayer.com/faq/comp.infosystems.gis.html). GenAlEx calculations match those calculated by the web site http://www.movable-type.co.uk/scripts/LatLong.html and closely approximate the output of Garmin GPS software (which likely uses a more complex formula). Distances calculated via Lat/Long coordinators are returned in km's.

Ex 3.5 Advanced Mantel (Optional)

The Mantel test allows tests for a statistical relationship between the elements of any two distance matrices with matching entries. It is therefore widely applicable. For example, beyond testing for 'Isolation-by-Distance' we might also wish to compare the relationship between different loci or relationship between different genetic markers. Under the assumption of independent and neutral loci, we would expect to find general concordance between different genetic markers. We can test this expectation with Mantel. Examples of comparisons among markers are provided by Peakall et al. (1995) and Maguire et al. (2000).

A subset of SSR and AFLP data obtained by Maguire et al. (2000) in their study of a mangrove tree is provided in the workbook *Ex 3.5 Mangrove.xls*. If time permits, you may wish to run a Mantel test to compare the patterns of genetic variation revealed by the two kinds of markers.

Another invaluable application of Mantel is to compare different statistics. For example, AMOVA allows estimates of *PhiPT*, F_{ST} and R_{ST}. One question you may wish to address is: What is the relationship among *PhiPT* and F_{ST} for the bush rat data introduced in Ex. 3.1? Alternatively, you might ask how strongly correlated is *PhiPT* with Nei Genetic Distance *D* or the Shannon Mutual Information Index. You should now be sufficiently skilled in the use of GenAlEx to tackle these questions without instructions.

Global Spatial Autocorrelation

Our first consideration of spatial genetic pattern was in Ex 3.2 where we explored the relationship between geographic distance and genetic distance among bush rat populations using Mantel tests. The geographic scale was in km for that study. It is now time to move to a smaller geographic scale and consider the patterns of individual genotypes in space. Spatial autocorrelation is a well established tool in biology for exploring the relationships between ecological or genetic variables and geographic location. 'Global' spatial autocorrelation attempts to describe the genetic structure across an entire study site. In contrast, 'local' analyses use a subset of the available data, investigating the presence of clusters or patches of genetic autocorrelation. Such clusters may or may not occur in the absence of significant global autocorrelation.

In this course we will restrict our attention to the multivariate spatial autocorrelation methods developed by Smouse and Peakall (1999) and extended by Peakall et al. (2003), Double et al. (2005) and Smouse et al. (2008). These methods are exclusively implemented in GenAlEx. These methods employ a multivariate approach to simultaneously assess the spatial signal generated by multiple genetic loci. Unlike classical spatial autocorrelation analysis, usually executed one allele at a time, the procedure is intrinsically multivariate, avoiding the need for allele-by-allele, locus-by-locus analysis (although such analyses can always be conducted, if desired). By combining alleles and loci, we strengthen the spatial signal by reducing stochastic (allele- to-allele and locus-to-locus) noise. The autocorrelation coefficient generated (r) is a proper correlation coefficient, bounded by [–1, +1] and is closely related to Moran's-I. The autocorrelation coefficient r provides a measure of the genetic similarity between pairs of individuals whose geographic separation falls within the specified distance class. The results are often summarised by a correlogram.

A key feature of the autocorrelation method of Smouse and Peakall (1999) is that the starting point for analysis is a pair of genetic and geographic distance matrices. Since genetic distance matrices can be generated for any kind of genetic marker (haploid, binary or codominant) or DNA sequence, the method is generic. Furthermore, as a distance-based method, with appropriate distance metrics it can also be extended to non-genetic data.

In the exercises that follow we will focus on 'global spatial' analysis before concluding with a brief exercise on 'local spatial' analysis. While we illustrate the analyses with genetic data, the methods described are generic and therefore applicable to ecological data (see Andrew et al. 2007 for an ecological example).

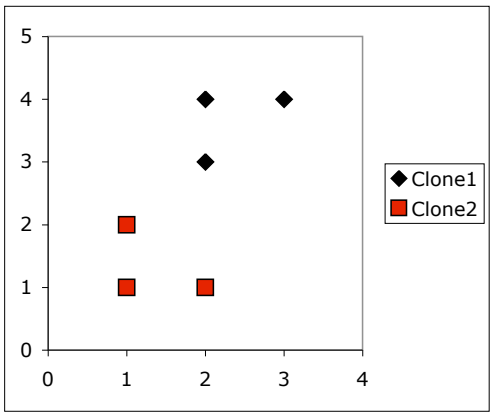
Ex 3.6 Hand Calculation of Spatial Autocorrelation

To illustrate the steps in a spatial autocorrelation analysis, let's begin with a hypothetical data set containing 6 samples scored at a single locus, representing two genetic clones spatially arranged as shown below. For ease the first 5 steps have been completed for you. However, before you begin complete questions 1 to 3.

Hypothetical data

Sample	Clone	Locus1			X	Y
1	1	1	1		3	4
2	1	1	1		2	4
3	1	1	1		2	3
4	2	1	2		1	1
5	2	1	2		1	2
6	2	1	2		2	1

Map of the clones in m



Step 1. Calculate a pairwise genetic distance matrix.

The codominant genotypic distance matrix

1	**2**	**3**	**4**	**5**	**6**	
0						**1**
0	0					**2**
0	0	0				**3**
1	1	1	0			**4**
1	1	1	0	0		**5**
1	1	1	0	0	0	**6**

Step 2. Calculate a pairwise linear geographic distance matrix.

The geographic distance matrix

1	**2**	**3**	**4**	**5**	**6**	
0.000						**1**
1.000	0.000					**2**
1.414	1.000	0.000				**3**
3.606	3.162	2.236	0.000			**4**
2.828	2.236	1.414	1.000	0.000		**5**
3.162	3.000	2.000	1.000	1.414	0.000	**6**

Step 3. Convert the genetic distance matrix to a square covariance matrix. The mathematics is a little tricky, so we will let GenAlEx do the work.

The square covariance matrix

0.25	0.25	0.25	-0.25	-0.25	-0.25
0.25	0.25	0.25	-0.25	-0.25	-0.25
0.25	0.25	0.25	-0.25	-0.25	-0.25
-0.25	-0.25	-0.25	0.25	0.25	0.25
-0.25	-0.25	-0.25	0.25	0.25	0.25
-0.25	-0.25	-0.25	0.25	0.25	0.25

Step 4. For ease of calculation, output in a single table the covariance, genetic distance and X matrices (for the lower triangular matrix only).

Covariance, geographic and X matrices as a table with the relevant *XijCij*, *XijCii* and *XijCjj* values for the valid contrasts at the first two distance classes

				X_1	X_2	Distance<=1			Distance<=2		
i	**j**	**Cov**	**Geog**	**<=1**	**<=2**	**XijCij**	**XijCii**	**XijCjj**	**XijCij**	**XijCii**	**XijCjj**
1	1	0.25	0	0	0						
1	2	0.25	1	1	0	0.25	0.25	0.25			
2	2	0.25	0	0	0						
1	3	0.25	1.4142	0	1				0.25	0.25	0.25
2	3	0.25	1	1	0	0.25	0.25	0.25			
3	3	0.25	0	0	0						
1	4	-0.25	3.6056	0	0						
2	4	-0.25	3.1623	0	0						
3	4	-0.25	2.2361	0	0						
4	4	0.25	0	0	0						
1	5	-0.25	2.8284	0	0						
2	5	-0.25	2.2361	0	0						
3	5	-0.25	1.4142	0	1				-0.25	0.25	0.25
4	5	0.25	1	1	0	0.25	0.25	0.25			
5	5	0.25	0	0	0						
1	6	-0.25	3.1623	0	0						
2	6	-0.25	3	0	0						
3	6	-0.25	2	0	1				-0.25	0.25	0.25
4	6	0.25	1	1	0	0.25	0.25	0.25			
5	6	0.25	1.4142	0	1				0.25	0.25	0.25
6	6	0.25	0	0	0						
Sum						**1**	**1**	**1**	**0**	**1**	**1**
SumXijCii											
SumXijCjj											
SumXijCij											

Step 5. For each pairwise comparison with a geographic distance less than or equal to 1 (=first distance class), determine the values of *XijCij*, *XijCii* and *XijCjj* from the table. You can easily find the valid pairwise contrasts by reference to the X_1 matrix. The valid comparisons for the first distance class have a value of 1, all other comparisons have the value zero.

For any valid pairwise contrast this means looking up 3 values in the table. For example, the first valid pair is sample 1 and 2. The XijCij value is the covariance matrix value for this contrast – i.e. the element in the matrix with i = 1 and j = 2. XijCii is the covariance matrix value of the diagonal element with i = 1 and j=1. XijCjj is the covariance matrix values of the diagonal element with i=2 and j=2. Here the answers are provided in the table above.

Using the data in the table above, complete steps 6 to 8.

Step 6. Sum the respective *XijCij*, *XijCii* and *XijCjj* values across the data set to give *SumXijCij*, *SumXijCii* and *SumXijCjj.* Enter these values at the bottom of the table.

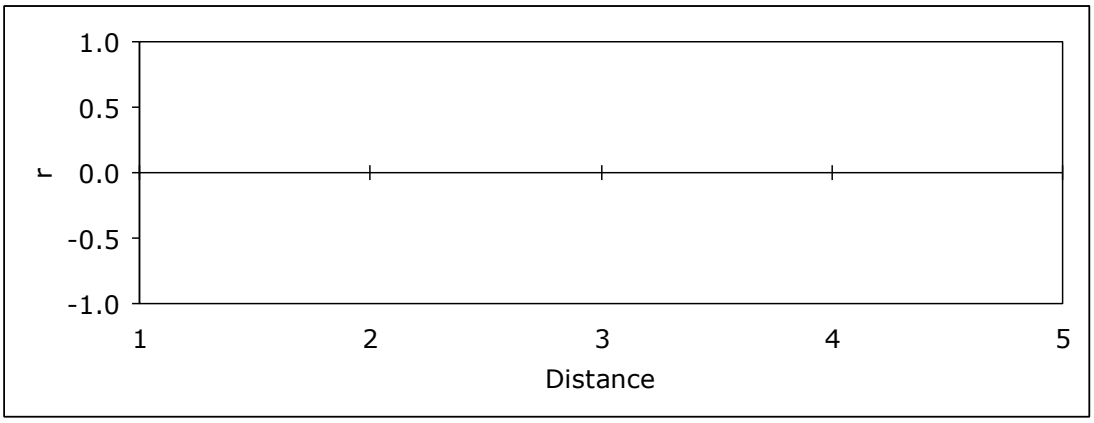
Step 7. Now calculate *r* as 2**SumXijCij*/(*SumXijCii* + *SumXijCjj*). Enter your results in the table below.

Step 8. Repeat steps 6 & 7 for the second distance class <= 2 (X_2 matrix). Add your results to the table below. Results for third and fourth distance class are provided for you.

Outcome of autocorrelation analysis

n	4	4	4	3
Distance	1	2	3	4
r			-1.000	-1.000

Step 9. Plot your outcomes as a correlogram below.



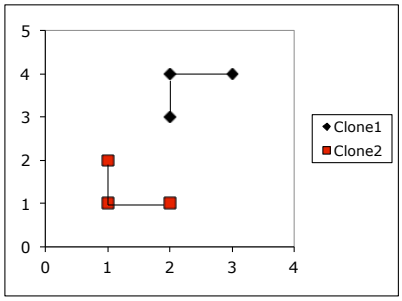
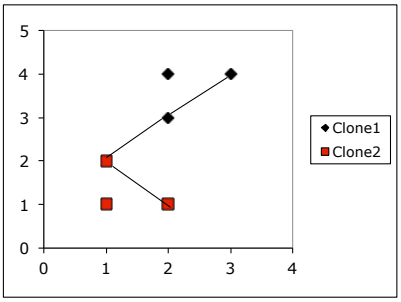
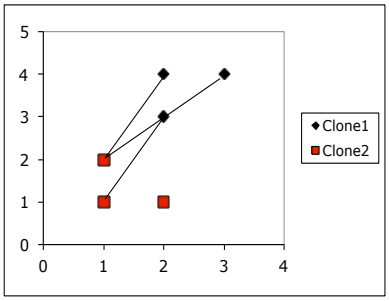
Q 3.6. Questions

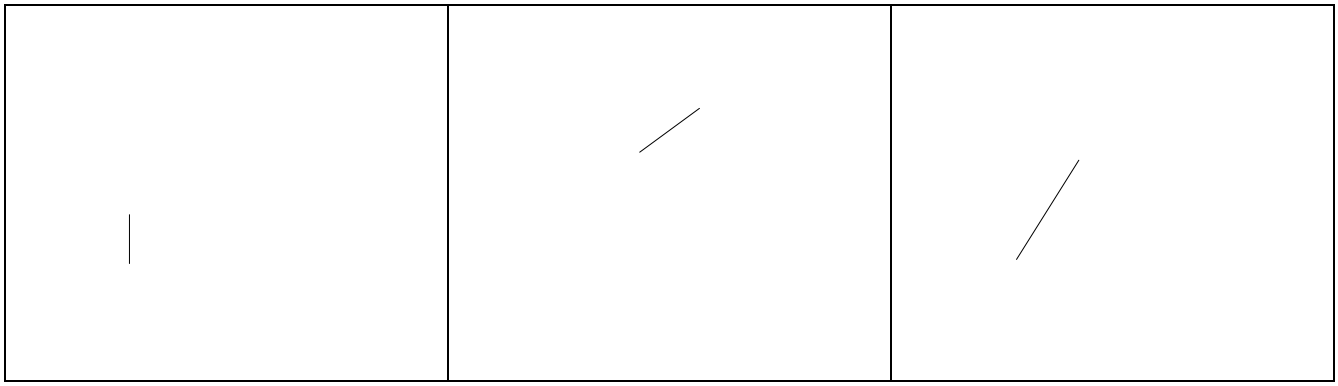
1. Study the map of the clones shown below and consider the connections between individuals that are 1 m or less apart. This is the first distance class in the analysis. What value do you predict for *r*?

2. What value do you predict for *r* for those individuals that are >1 to 2 m apart?

3. What value do you predict for *r* for those individuals that are >2 to 3 m apart?

4. Compare your predictions with the answers you obtained above. Does it all make sense?

Connections 1 m or less apart	Connections >1 to 2 m apart	Connections >2 to 3 m apart
		



Box 3.3 Autocorrelation Formulae

In general terms a correlation coefficient is calculated as:

$$r = \frac{Cov(x,y)}{\sqrt{[Var(x)Var(y)]}}$$

$$\text{Where, } Cov(x,y) = \frac{\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)} \text{ and } Var(x) = \frac{\sum_{i}^{n}(x_i - \bar{x})^2}{(n-1)}.$$

For the spatial autocorrelation coefficient defined by Smouse and Peakall (1999) and computed from the lower triangular matrix the following (analogous) formula applies:

$$r = \frac{2\sum x_{ij}c_{ij}}{\sum x_{ij}c_{ii} + \sum x_{ij}c_{jj}}$$

Covariance matrix				
c_{11}	c_{21}	c_{31}	…	c_{N1}
c_{12}	c_{22}	c_{32}	…	c_{N2}
c_{13}	c_{23}	c_{33}	…	c_{N3}
…			…	
c_{1N}	c_{2N}	c_{3N}	…	c_{NN}

X matrix				
x_{11}	x_{21}	x_{31}	…	x_{N1}
x_{12}	x_{22}	x_{32}	…	x_{N2}
x_{13}	x_{23}	x_{33}	…	
…			…	
x_{1N}	x_{2N}	x_{3N}	…	x_{NN}

Where c_{ij}, c_{ii} and c_{jj} are the respective elements of the covariance matrix, and x_{ij} the respective element of the *X* matrix. Each distance class is represented by its own *X* matrix, with elements having the value 1 when that specific pairwise comparison falls within the distance class, all other elements have the value 0. The covariance matrix is derived from the genetic distance matrix (following Smouse and Peakall 1999) by the formula:

$$C_{ij} = \frac{1}{2}\left[-d_{ij}^{\,2} + \frac{1}{N}\left(\sum_{i=1}^{N} d_{ij}^{\,2} + \sum_{j=1}^{N} d_{ij}^{\,2}\right) - \frac{1}{N^2}\left(\sum_{i\neq j}^{N} d_{ij}^{\,2}\right)\right]$$

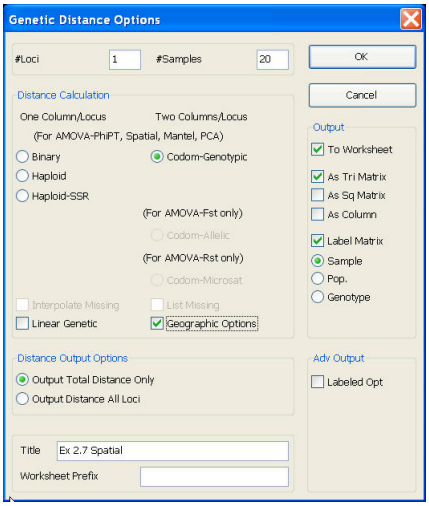
Ex 3.7 Spatial Autocorrelation Analysis in GenAlEx

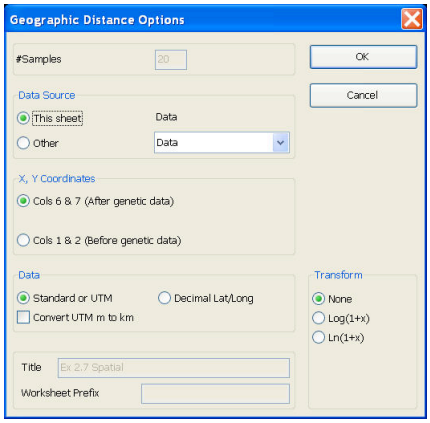
It is time to learn the steps for performing spatial analysis in GenAlEx. To illustrate these steps we will start with a hypothetical plant data set similar to that used in Ex 3.6. In this case, there are 3 distinct clones within a 10 x 10 m plot.

Step 1. Open *Ex 3.7 Spatial.xls*. The *Data* worksheet shows the genotypes for a single codominant locus with 2 alleles. X and Y coordinates are provided for each sample. A plot of the clones is also shown.

Step 2. For spatial analysis we need pairwise genetic and geographic matrices as inputs. Both these distance matrices can be computed in one easy step. Activate the *Data* worksheet then choose ***Distance->Genetic.*** At the Genetic Distance Options dialog box select *Codom-Genotypic*, *Output Total Distance Only* and *Geographic Options*. From the Geographic Distance Options dialog box select *Cols 6 & 7 (after data)* and *Standard*.

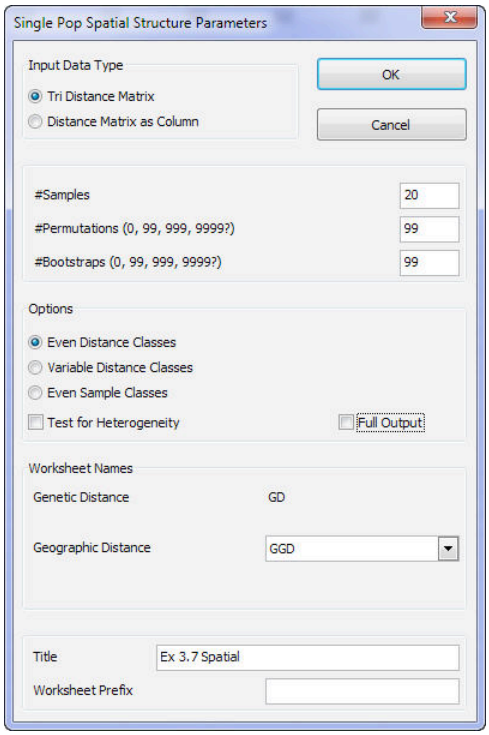
Note that GenAlEx places the geographic (*GGD*) and genetic distance (*GD*) matrices in front of the data worksheet. If not already, it is always recommended that you move these matrices to worksheet positions 1 and 2, respectively.

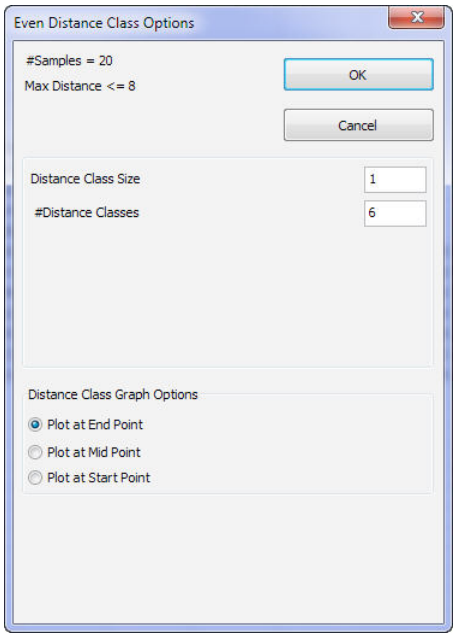




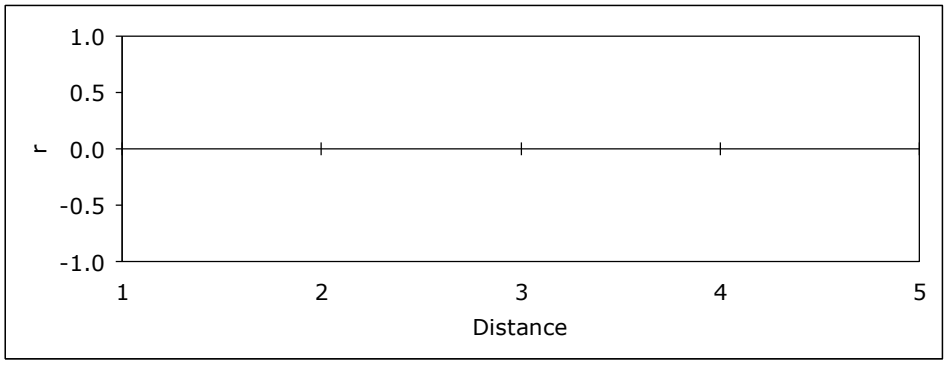
Step 3. Activate the genetic distance worksheet [GD]. Choose ***Spatial->Single Pop***. At the Single Pop Spatial Structure Parameters Options dialog box choose worksheet [GGD] from the drop down worksheet list and select *Even Distance Classes*. For now you can set the number of bootstraps and permutations to zero.

Step 4. At the Even Distance Class Options dialog box set: *Distance Class Size* = 1, and *#Distance Classes* = 6.





Step 5. Inspect the results of your analysis. The outcome is presented in the worksheet [R]. Draw a schematic of the correlogram below and answer questions 1 & 2.



Q 3.7. Questions

1. Summarise the outcomes of your spatial analysis in words.

2. What do you conclude about the magnitude of local spatial genetic structure?

Box 3.4 Understanding Distance Classes

The concept of 'Distance Classes' as used in Spatial autocorrelation analysis, may at first be a little confusing. Each distance class is bounded by an upper and lower value as specified either implicitly or explicitly by the user. Within GenAlEx the default *Even Distance Classes* option takes the inputs *Distance Class Size* and *#Distance Classes* and from these determines the upper and lower value for each distance class. For example, if Distance Class Size =1 and #Distance Classes = 5, the first distance class has the lower bound of 0 and an upper bound of 1. All pairwise comparisons with a geographic distance less than or equal to 1 belong in the first distance class. The second distance class includes all those comparisons with a geographic distance greater than 1 and less than or equal to 2, etc. With only 5 distance classes, all pairwise comparisons of geographic distance greater than 5 are ignored in the spatial autocorrelation analysis.

Statistical Testing for Spatial Autocorrelation

For spatial autocorrelation the null (*H0*) and alternative hypotheses (*H1*) are:
H0=A random distribution of genotypes in space (r=0)
H1=A non random distribution of genotypes in space (r<>0)

In order to distinguish among these hypotheses, GenAlEx offers statistical testing for spatial autocorrelation based on two methods: (*i*) random permutation, similar to that used for AMOVA and Mantel, and (*ii*) bootstrap estimates of r.

Random permutation allows us to generate a distribution of permuted (r_p) values under the assumption of no spatial structure, by the random shuffling of all individuals among the geographic locations. From 1000 such random permutations, the values of the 25th and 975th ranked r_p values are taken to define the upper and lower bounds of the 95% confidence interval. If the calculated r-value falls outside this confidence belt, then significant spatial genetic structure is inferred. This is the classic two-tailed test. When one's interest is in the detection of positive autocorrelation, as predicted under restricted dispersal, GenAlEx also computes a one-tailed probability. In this case the individual r_p values are compared with the observed r-value, to estimate the probability of randomly achieving a value greater than or equal to the observed r. If this probability is less than 0.05, the alternative hypothesis of positive spatial genetic structure is accepted.

Bootstrap estimates allow us to place a confidence interval around the observed estimate of r by drawing (with replacement) from within the set of pairwise comparisons for a specific distance class. For each of 1,000 bootstrap trials, the bootstrap autocorrelation coefficient (r_{bs}) is calculated for each distance class. The 25th and 975th ranked r_{bs} are then taken to define the 95% confidence interval. When the bootstrap confidence interval does not straddle r = 0, significant spatial genetic structure is inferred.

Note that while providing an alternative statistical test, this bootstrap test is less powerful than permutational tests, since the number of samples per distance class is much smaller than the $n(n-1)/2$ comparisons used during permutation. Thus, for small sample sizes, bootstrap errors tend to be larger than the permutational errors. The bootstrap test is conservative, favouring the null

hypothesis to a greater extent than does the permutational test. Despite this limitation, the calculation of bootstrap errors enable a graphical test of statistical significance among different *r* values, using the respective 95% confidence intervals.

More recently Smouse et al. (2008) have developed procedures for testing the overall significance of correlograms and for testing for heterogeneity among correlograms. These options are implemented in GenAlEx 6.2 onwards.

Ex 3.8 Statistical Tests for Autocorrelation (Optional)

In this exercise we will explore the application of random permutation for spatial autocorrelation. Much of this should be revision from earlier exercises, therefore only brief instructions will be provided. If in doubt, refer to previous exercises.

Step 1. Return to *Ex 3.7 Spatial.xls.* Use ***Rand Data->Shuffle Tri*** to create a shuffled genetic distance matrix.

Step 2. Re-run a spatial analysis on the new shuffled data using the same options and parameters as in Ex 3.7. Record the *r* values across the 4 distance classes in the table below.

Step 3. Repeat steps 1 and 2, four more times, recording the *r* values obtained in the table below. Calculate the mean *r* value for the five shuffles across the 4 distance classes. Answer question 1 before proceeding.

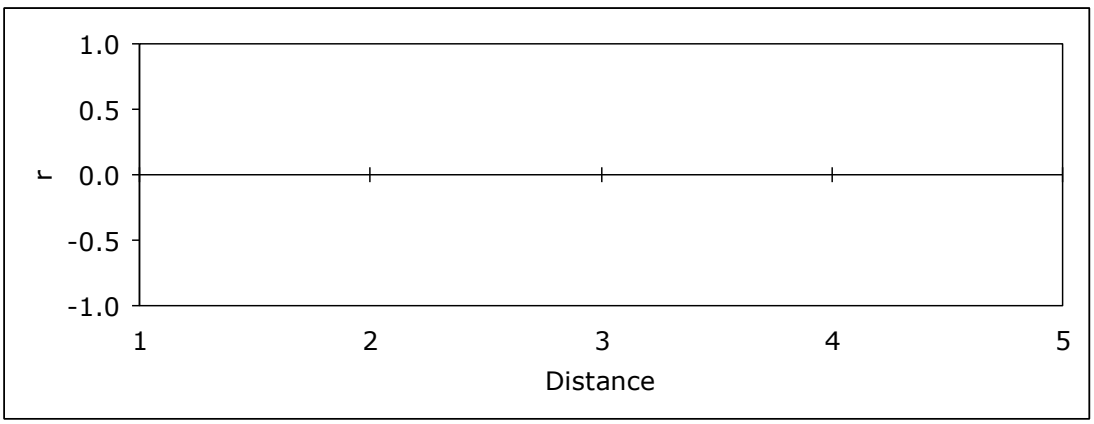
Autocorrelation analysis for shuffled genetic distance matrices

Distance	**1**	**2**	**3**	**4**
Original r	1.000			
ShuffleTri1 r				
ShuffleTri2 r				
ShuffleTri3 r				
ShuffleTri4 r				
ShuffleTri5 r				
Mean r of Shuffles				

Step 4. Re-run the original spatial analysis with the number of permutations set to 99. For now, set the number of bootstraps to zero.

Step 5. Inspect the results of your analysis and plot the outcome as a correlogram below. Answer questions 2 onwards.

Step 6. (Optional) Re-run the original spatial analysis with the number of bootstraps set to 99. Plot the error bars below.



Q 3.8 Questions

1. How does the mean *r* value across the shuffles compare with the *r* for the original data?

2. Summarise the outcomes of your spatial analysis in words. Include the value of *r* and comment on the outcome of statistical tests.

3. Briefly describe the process of permutational testing during spatial autocorrelation analysis and compare with the process for Mantel.

4. How does the bootstrap test differ from permutation? What advantages and limitations does this test offer for spatial analysis?

Ex 3.9 Spatial Analysis of Multiple Loci

The 'global' spatial autocorrelation method implemented in GenAlEx employs a multivariate approach to simultaneously assess the spatial signal generated by multiple genetic loci. However, it can be of interest to determine if any overall signal of spatial autocorrelation is consistent across loci, especially for coding loci where the assumption of neutrality may not hold.

In this exercise, the data set is drawn from a study of allozyme variation at 5 loci in the orchid *Caladenia tentaculata* (see Peakall and Beattie 1995, Smouse and Peakall 1999). We will examine the patterns of spatial genetic autocorrelation for each of the loci independently, and in combination. While this can always be done, locus by locus, we will let GenAlEx assist in the automation of the process.

Step 1. The first step is to generate the pairwise genetic and geographic matrices. Open the workbook *Ex 3.9 Caladenia Spatial.xls*, activate the *Data* worksheet then choose ***Distance->Genetic.*** At the Genetic Distance Options dialog box select *Codom-Genotypic*, *Output Distance All Loci* and *Geographic Options*. From the Geographic Distance Options dialog box select *Cols 14 & 15 (after data)* and *Standard*.

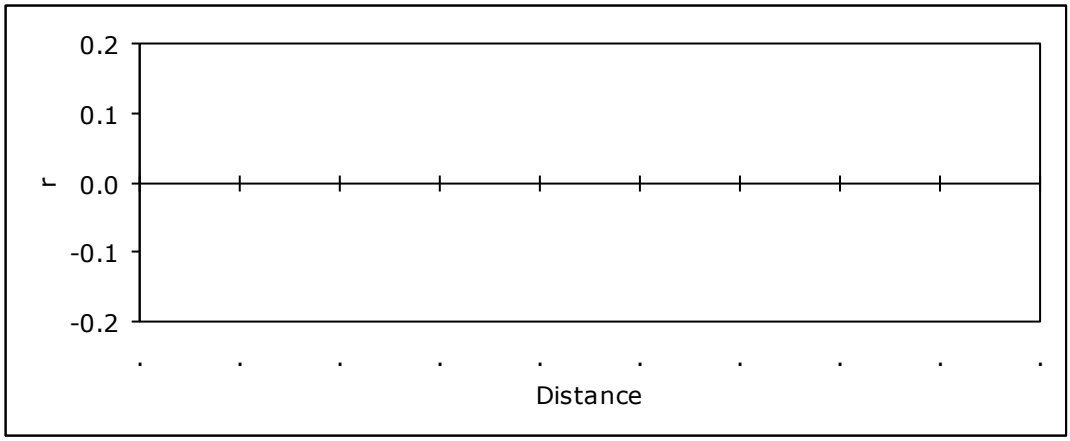
Note that GenAlEx places the geographic (*GGD*) and genetic distance (*GD*) worksheets in front of the data worksheet. For the Multiple Loci analysis, the genetic distance worksheets for loci 1 to n must be in the matching worksheet positions 1 to n. The geographic distance matrix must be placed after the final genetic distance worksheet.

Step 2. Manually move the worksheet *GGD* from worksheet position 1 to worksheet position 6 (right after *GOT3 GD*).

Step 3. Choose ***Spatial->Multiple Loci***, then enter the number of loci as 5. At the Multiple Loci Spatial Structure Parameters dialog box, choose *GGD* from the drop down worksheet list and select *Even Distance Classes*. Set the number of bootstraps and permutations to 99.

Step 4. At the Even Distance Class Options dialog box set: *Distance Class Size* = 1, and *#Distance Classes* = 10.

Step 5. Inspect the results of your analysis. The outcome is presented in the worksheet *RML*. Draw a schematic of the correlogram below for the loci MR and PGM only. Answer questions 1 and 2.



Q 3.9 Questions

1. Summarise the outcomes of your spatial analysis in words. Did you detect a consistent pattern of spatial genetic structure across the loci?

2. List some hypotheses that might explain your findings? What additional analyses might it be useful to conduct?

Ex 3.10 Alternative Outputs to Correlograms

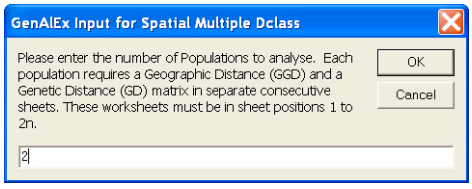
So far we have used the traditional correlogram as a descriptor of the patterns of spatial genetic autocorrelation. Correlograms plot the autocorrelation coefficient *r* as a function of distance class. GenAlEx further superimposes the 95% confidence interval about the null hypothesis of no spatial genetic structure (random) on this graph, as determined by permutation. However, there are some limitations of correlograms as indicators of spatial genetic structure (see also Box 3.5). Here we consider an alternative output option uniquely provided in GenAlEx. The ***Spatial->Multiple Dclass*** option performs autocorrelation analysis for multiple distance class sizes and multiple populations, in a way that is designed to explore the interplay between sample size and distance class size. Unlike a standard analysis, multiple analyses are performed, with automatically increasing distance size classes.

For this exercise we will return to a bush rat data set.

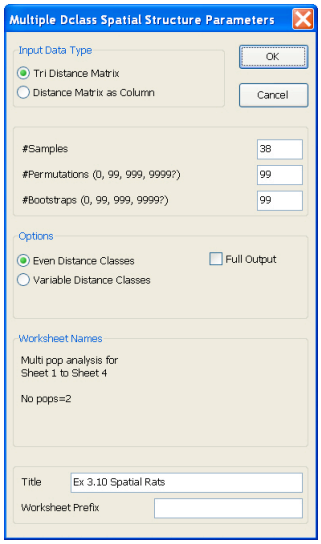
Step 1. Open *Ex 3.10 Spatial Rats.xls*. Inspect the workbook that contains genetic results for two populations *MD* and *U3*. Geographic and genetic distance matrices have also been calculated. For this analysis (and the Multiple Pops analysis) the distance matrices must be ordered from 1 to 2n with the Geographic distance matrix (*GGD*) for population 1, followed by genetic distance (*GD*) matrix for population 1, *GGD* for population 2, *GD* for population 2, etc.

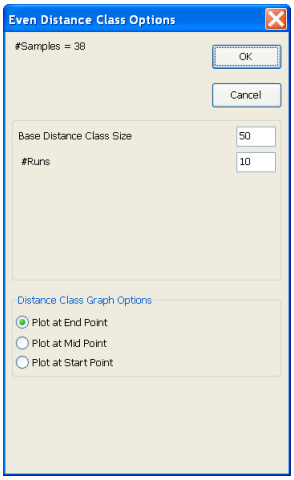
Note that GenAlEx does not check that your sheets are in the correct order so be sure to double-check before analysis. It is also essential to ensure that no other worksheets interrupt the data worksheets.

Step 2. Choose ***Spatial->MultipleDclass*** then specify the number of populations as 2.



Step 3 At the Multiple Dclass Spatial Structure Parameters dialog box select *Even Distance Classes* and set the number of bootstraps and permutations to 99. At the Even Distance Class Options dialog box set: *Base Distance Class Size* = 50, *#Runs* = 10.

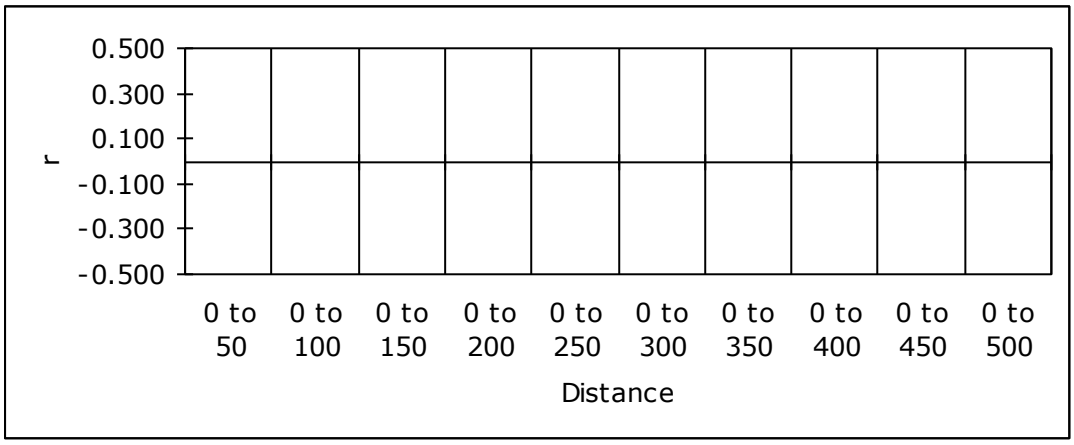




For each run, GenAlEx will calculate the Distance Class Size as *Run no.* * *Base Distance Class Size*. For example, if you set the base size to 50 for 3 runs, the routine will calculate the combined *r* values across populations, based on all those pairwise comparisons with a geographic distance of 0 to 50m. Next *r* will be calculated for pairwise comparisons with a geographic distance of 0 to 100m. Finally, *r* will be calculated for all those pairwise comparisons with a geographic distance of 0 to 150m.

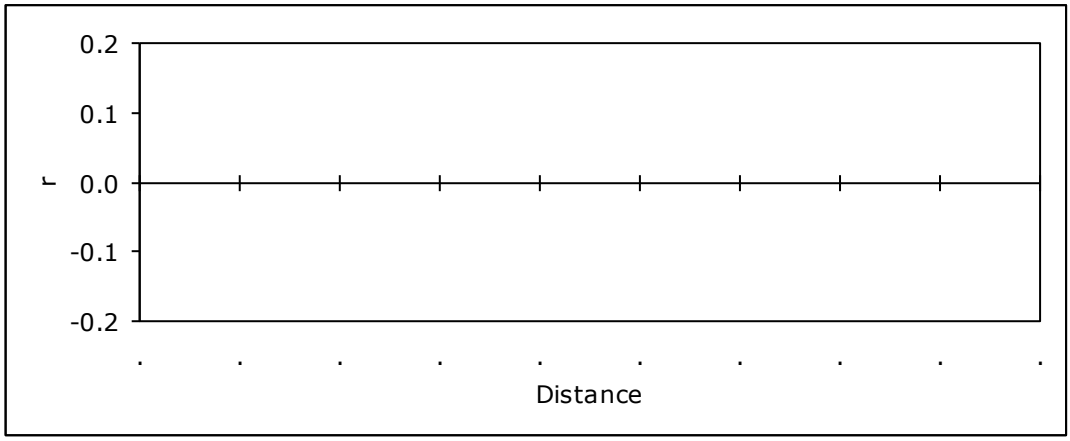
The *r* calculated for each distance class size is actually the *r* of the first distance class had you run a spatial autocorrelation analysis at that distance class size with either of the single or multiple population options. You can test this by comparing *r* values in the table below the autocorrelogram.

Step 4. Inspect the results of your analysis. The outcome is presented in the worksheet *MDC*. Draw a schematic of the graph below and answer question 1.



Step 5. For comparative purposes generate a correlogram for the data using ***Spatial->Multiple Pops*** with the same settings as above.

Step 6. The worksheets *RMP* and *RC* show correlograms for each population, and combined populations, respectively. Draw a schematic of the combined correlogram below and answer questions 2 and 3.



Q 3.10 Questions

1. Summarise the outcomes of your Multiple Dclass spatial analysis in words. What is the extent of genetic structure?

2. Summarise the outcomes of your Multiple Pops spatial analysis in words. What is the extent of genetic structure detected by this analysis.

3. Are the conclusions drawn from the two respective analyses, congruent? If not, why not?

Box 3.5 Correlograms and the meaning of the X-intercept

Correlograms plot the autocorrelation coefficient *r* as a function of distance class. If positive spatial genetic structure is found, the first x-intercept has often been interpreted in the literature as an estimate of the extent of non-random (positive) genetic structure or 'genetic patch size'. However, a single correlogram may not reveal the true extent of non random genetic structure, since the capacity to detect spatial genetic structure in any given analysis reflects the interplay between the true (but unknown) extent of genetic structure, the distance class sizes chosen, and the associated number of samples per distance class. Sampling at intervals greater than the scale of genetic structure will fail to detect the genetic structure at all, while sampling at intervals well below the scale of genetic structure may be associated with unnecessarily small sample sets, and therefore limited statistical power. Furthermore, as demonstrated by Peakall et al. (2003), changing the distance class size can lead to a very different value for the x-intercept. Thus, caution is required in the interpretation of the meaning of the x-intercept for any given correlogram.

Since there is no *a priori* way to predict the extent of genetic structure from a single autocorrelation analysis, GenAlEx also allows the calculation of *r* (along with associated errors about *r* and the null hypothesis), for increasing distance class sizes, ranging from the minimum distance between samples to the maximum distance of sampling. When significant positive structure is present, the estimated value of *r* will decrease with increasing size of the distance class. The distance class size at which the estimate of *r* is no longer significant provides an approximation of the extent of detectable positive spatial genetic structure (Peakall et al. 2003).

Local 2D Spatial Autocorrelation Analyses

Double et al. (2005) extended the spatial analysis procedures of Smouse and Peakall (1999) to enable investigation of the local patterns of spatial genetic autocorrelation across a two-dimensional landscape. The approach taken is analogous to the procedures and statistics introduced by earlier workers for the analysis of local spatial autocorrelation. We use the acronym 2D LSA for '2 Dimensional Local Spatial Autocorrelation' to refer to this approach (Double et al. 2005).

All local autocorrelation methods employ a sampling strategy that focuses on a subset of points surrounding a pivotal data point. Our approach defines a 'local' subset as an individual and its *n* nearest neighbours. For each subset, we estimate *lr* the local autocorrelation, according to the method of Smouse and Peakall (1999), based on the *n* pairwise comparisons between the pivotal individual and its *n* nearest neighbours. In a complete analysis of *s* samples, we estimate the local autocorrelation *lr* for each of *s* local data subsets. This 2D LSA method only differs from standard autocorrelation analysis in the way individuals are selected for calculating *lr*. In essence, our local subset is merely a specially constructed 'distance class'.

Ex 3.11 Understanding 2D Spatial Analyses

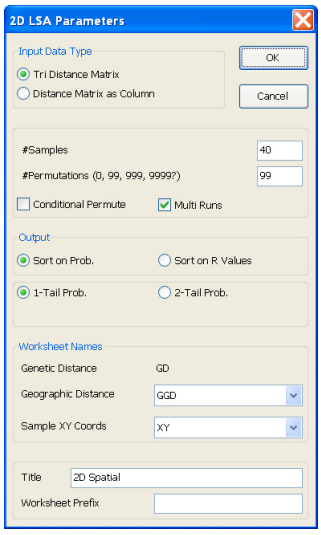
In this exercise we will explore 2D LSA using a hypothetical plant data set that is an expansion of Ex 3.7. In this case, the 10 x10 m plot contains 3 distinct clones plus 20 additional unique genotypes (non-clonal, see figure below).

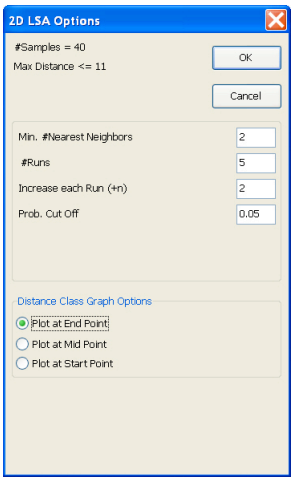
Step 1. Open *Ex 3.11 Spatial 2D.xls* and inspect the *Data* worksheet.

Step 2. For a 2D LSA analysis, pairwise genetic and geographic distance matrices are required, as well as a list of the XY coordinators. For ease of GenAlEx analysis, these should be in worksheet positions 1 to 3, respectively. These worksheets have been prepared for you.

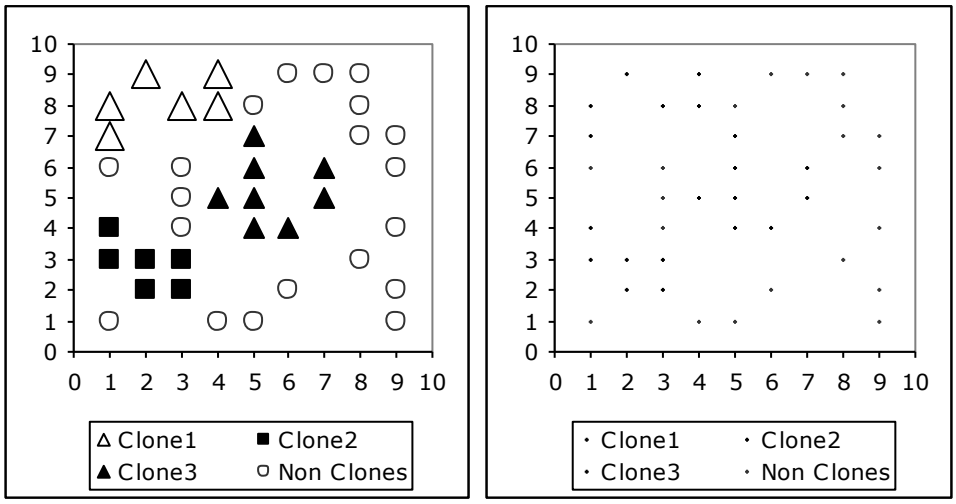
Step 3. Activate the worksheet *GD* and then choose ***Spatial->2D LSA***. At the 2D LSA Parameters dialog box, check that *GGD* and *XY* have been selected as *Geographic data* and *Sample XY coordinates* respectively, select *Multi Runs* and set the number of permutations to 99.

Step 4. When prompted by the 2D LSA Options dialog box set: *#Nearest Neighbours* = 2, and *#Runs* = 5, *Increase each Run* = 2, *Prob. Cut-off* = 0.05. With these settings, GenAlEx will perform 2D LSA analysis for 5 runs. Each run the number of nearest neighbours will be increased by 2 (2, 4, 6, 8, 10, 12).





Step 5. Inspect the results of your analysis. The outcomes are presented in a series of worksheets *R2D1* to *R2D5*. Draw a schematic of the outcome for one run in the blank graph on the right, below. Answers Questions 1 and 2.



Q 3.11 Questions

1. Summarise the outcomes of your 2D LSA analysis in words. Do your findings make sense?

2. How many nearest neighbours were optimal for this analysis? Explain your answer.

Statistical Testing for 2D LSA

While calculation of the local autocorrelation statistic, *lr,* is straightforward, in common with all other local spatial autocorrelation methods, testing for statistical significance poses some complications. Permutational procedures offer a partial (but incomplete) solution. As for global autocorrelation analyses, permutation can reveal the extremeness of observed *lr* values, relative to the permuted lr_p values, under the null hypothesis of no spatial structure. One of two permutational approaches may be used: (*i*) standard permutation, in which all samples are shuffled over all locations in the data set, and (*ii*) conditional permutation, in which the relevant sample at the pivotal location of the subset in question is held fixed, while the remaining samples are permuted over all other locations in the data set. GenAlEx 6 offers both types of permutational procedures. In practice, there appears to be little difference in the outcomes between the two alternative permutational procedures. As for global autocorrelation procedures, GenAlEx determines the upper and lower bounds of the 95% confidence interval by ranking the permuted local spatial autocorrelation lr_p values. In addition, a one-tailed probability for positive local spatial autocorrelation is estimated as the fraction of times a value of lr_p greater than or equal to *lr* is observed for each local sub-sample.

It is important to note that these permutational procedures only go part way toward providing an assessment of statistical significance in a local spatial analysis. One complicating factor is that the neighbouring estimates of *lr* will be partially correlated, because they share some overlapping samples. This is not the case for global autocorrelation. An additional complication is that multiple comparisons are involved, requiring a Bonferroni-type adjustment, though a strict Bonferroni adjustment is probably far too conservative. For example, in Double et al. (1999) our data set involved 74 dominant fairy wren males. In this case, the Bonferroni adjustment for a one-tailed test of α at 0.05 is 0.05/74 = 0.00067, making the possibility of detecting a significant value remote. Furthermore, the prospects of detecting statistical significance become more and more remote with increasing sample size.

Pending resolution of these statistical testing issues, it is important to recognize that the 2D LSA method is best viewed as a tool for data exploration.

Box 3.6 2D LSA - How many neighbours?

In 2D LSA analysis, GenAlEx allows user control of the number of nearest neighbours and also offers a multiple analysis function that automatically outputs analyses for increasing numbers of neighbours. It is recommended that the user take full advantage of this multiple analysis function and look for consistency of pattern across a range of different numbers of nearest neighbours, when interpreting the pattern.

To fully exploit the power of genetic autocorrelation analyses, we require highly variable markers, appropriate sampling and detailed ecological data. Without such information, the biological meaning of the spatial genetic patterns can easily be misinterpreted. Indeed, irrespective of the statistical procedure, a greater comprehension of the biological significance and meaning of spatial autocorrelation results is best achieved when both ecological and genetic data are combined (Double et al. 2005).

References and Further Reading

Note that for a more extensive literature on these topics, please see the Appendix 1 provided with GenAlEx: Freely available from the Australian National University, Canberra, Australia. http://biology.anu.edu.au/GenAlEx/

Austerlitz F and Smouse PE (2001) Two-generation analysis of pollen flow across a landscape. II. Relation between Phi(ft), pollen dispersal and interfemale distance. *Genetics*, **157**, 851-857.

Austerlitz F and Smouse PE (2001) Two-generation analysis of pollen flow across a landscape. III. Impact of adult population structure. *Genetical Research*, **78**, 271-280.

Austerlitz F and Smouse PE (2002) Two-generation analysis of pollen flow across a landscape. IV. Estimating the dispersal parameter. *Genetics*, **161**, 355-363.

Double MC, Peakall R, Beck NR and Cockburn A (2005) Dispersal, philopatry and infidelity: dissecting local genetic structure in superb fairy-wrens (*Malurus cyaneus*). *Evolution*, **59**, 625-635.

Excoffier L, Smouse PE and Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitocondrial DNA restriction sites. *Genetics*, **131**, 479-491.

Huff DR, Peakall R and Smouse PE (1993) RAPD variation within and among natural populations of outcrossing buffalograss *Buchloë dactyloides* (Nutt) Engelm. *Theoretical and Applied Genetics*, **86**, 927-934.

Maguire TL, Peakall R and Saenger P (2002) Comparative analysis of genetic diversity in the mangrove species *Avicennia marina* (Forsk.) Vierh. (Avicenniaceae) detected by AFLPs and SSRs. *Theoretical and Applied Genetics*, **104**, 388-398.

Mantel N (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209-220.

Michalakis Y and Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics*, **142**, 1061-1064.

Orloci L (1978) *Multivariate analysis in vegetation research*. The Hague: Dr W. Junk B. V.

Peakall R and Beattie AJ (1995) Does ant dispersal of seeds in *Sclerolaena diacantha* (Chenopodiaceae) generate local spatial genetic structure. *Heredity*, **75**, 351-361.

Peakall R and Beattie AJ (1996) Ecological and genetic consequences of pollination by sexual deception in the orchid *Caladenia tentactulata*. *Evolution*, **50**, 2207-2220.

Peakall R, Ruibal M and Lindenmayer DB (2003) Spatial autocorrelation analysis offers new insights into gene flow in the Australian bush rat, *Rattus fuscipes*. *Evolution*, **57**, 1182-1195.

Peakall R, Smouse PE and Huff DR (1995) Evolutionary implications of allozyme and RAPD Variation in diploid populations of dioecious buffalograss *Buchloë dactyloides*. *Molecular Ecology*, **4**, 135-147.

Peakall R and Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288-295.

Peakall, R. and Smouse P.E. (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* In press. First published online July 20, 2012 doi:10.1093/bioinformatics/bts460. Advanced print Epub available here

Smouse PE, Dyer RJ, Westfall RD and Sork VL (2001) Two-generation analysis of pollen flow across a landscape. I. Male gamete heterogeneity among females. *Evolution*, **55**, 260-271.

Smouse PE and Long JC (1992) Matrix correlation analysis in anthropology and genetics. *American Journal of Physical Anthropology*, **35**, 187-213.

Smouse PE, Long JC and Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology*, **35**, 627-632.

Smouse PE and Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, **82**, 561-573.

Smouse PE, Peakall R and Gonzales E (2008) A heterogeneity test for fine-scale genetic structure. *Molecular Ecology*, **17**, 3389-3400.

Sork VL and Smouse PE (2006) Genetic analysis of landscape connectivity in tree populations. *Landscape Ecology*, **21**, 821-836.

Sokal RR, Oden NL and Thomson BA (1998) Local spatial autocorrelation in biological variables. *Biological Journal of the Linnean Society*, **65**, 41-62.

Vekemans X and Hardy OJ (2004) New insights from fine-scale spatial genetic structure Analysis in plant populations. *Molecular Ecology*, **13**, 921-935.

Appendix 3.1 – Advanced Spatial Analysis Options

In 2008, GenAlEx 6.2 was offered to provide users with access to new spatial genetic analysis procedures that extended application of the multiallelic, multilocus spatial autocorrelation analysis methods of Smouse and Peakall (1999), Peakall et al. (2003), Double et al. (2005). Collectively, these methods provide valuable insights into fine-scale genetic processes across a wide range of animals and plants. The additional procedures were developed and described in the following publications:

Smouse, P. E., Peakall, R., and Gonzales, E. (2008) A heterogeneity test for fine-scale genetic structure. *Molecular Ecology* 17, 3389-3400.

Beck, N., Peakall, R., and Heinsohn, R. (2008) Social constraint and an absence of sex-biased dispersal drive fine-scale genetic structure in white-winged choughs. *Molecular Ecology* 17, 4346-4358.

Gonzales E, Hamrick J, Smouse P, Trapnell D, Peakall R (2010) The impact of landscape disturbance on spatial genetic structure in the Guanacaste tree, Enterolobium cyclocarpum (Fabaceae). Journal of Heredity 101, 133-143.

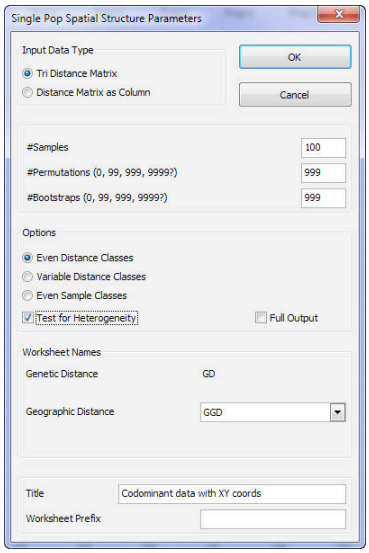
In 2012, with the release of GenAlEx 6.5 (Peakall and Smouse 2012), the heterogeneity tests of Smouse et al. (2008) were fully integrated into the standard GenAlEx Spatial options menus. Banks and Peakall (2012) also confirmed the statistical power and performance of this heterogeneity test by spatially explicit computer simulations.

Banks SC, Peakall R (2012) Genetic spatial autocorrelation can readily detect sex-biased dispersal. Molecular Ecology 21, 2092-2105.

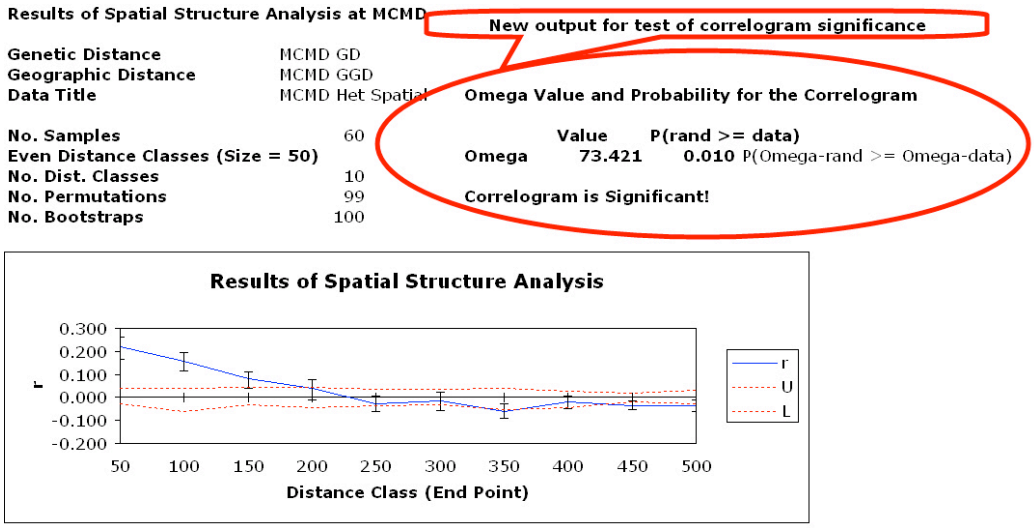
This appendix provides a brief overview of these options. Readers are strongly advised to complete the Spatial Tutorial Module fully before attempting to use these Advanced Spatial options. This appendix assumes that readers are experienced GenAlEx users and familiar with the above publications.

Heterogeneity Tests

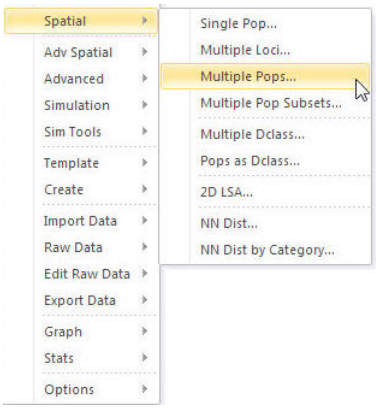
As noted above, GenAlEx 6.5 allows one to implement the nonparametric heterogeneity tests of Smouse et al. (2008).

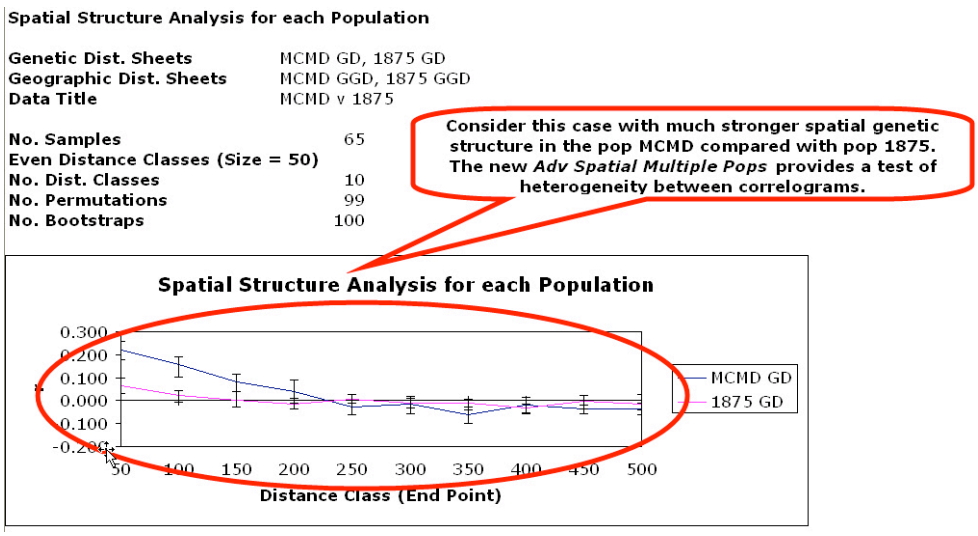


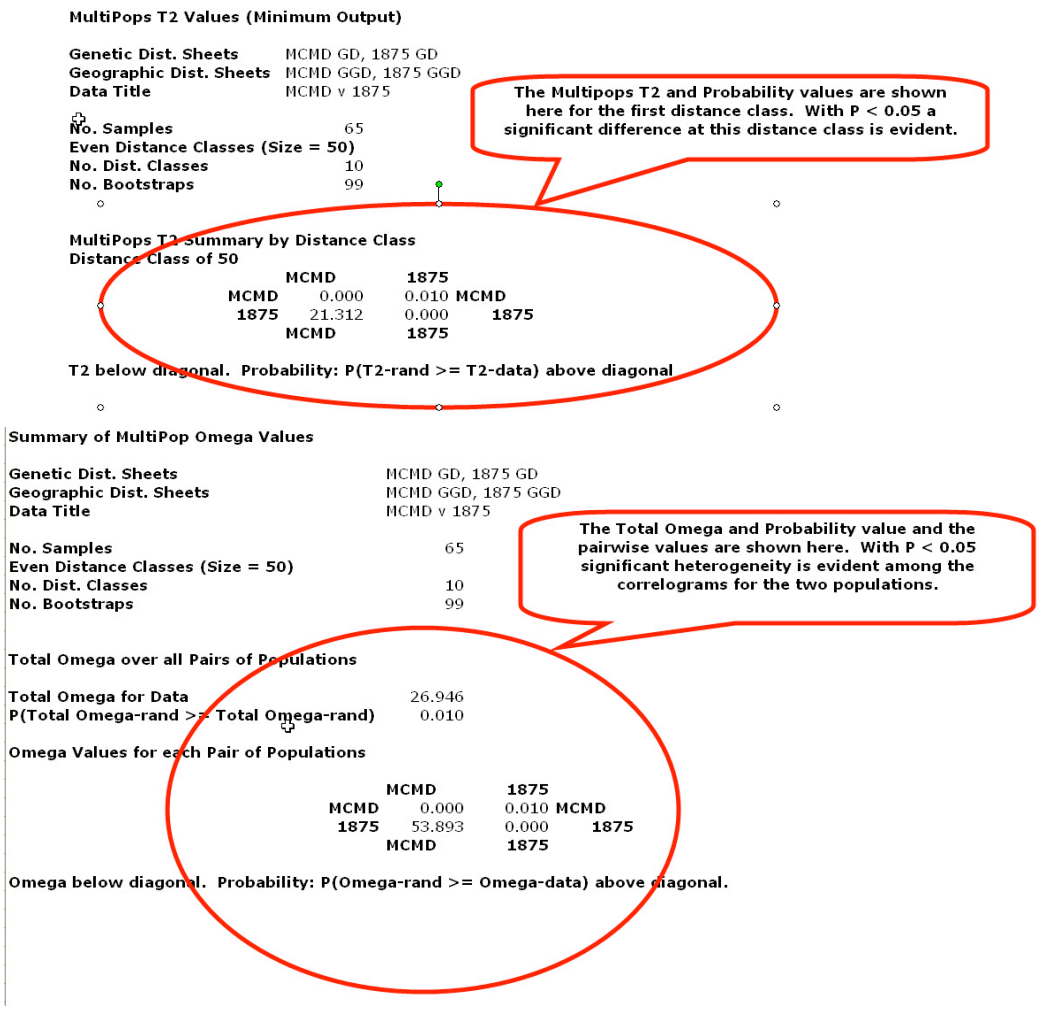
The *Test for Heterogeneity* is enabled at the Spatial Structure Parameters dialog box. In the case of a *Single Pop* spatial analysis, the Omega value and probability are provided above the spatial correlogram graph. In this case, the heterogeneity test provide a test of correlogram significance.



More often, applications of the heterogeneity tests will be of more interest when you have suitable data from two or more populations. This test is also useful for assessing difference in spatial genetic structure patterns between sexes (here sexes are treated as different populations for the purpose of the analysis). See Banks and Peakall (2012) for a comprehensive overview of how spatial autocorrelation analysis can be applied to detect sex-biased dispersal. To proceed, set up your data as per the instructions for the ***Multiple Pops*** option as outlined in the GenAlEx 6.5 guide.





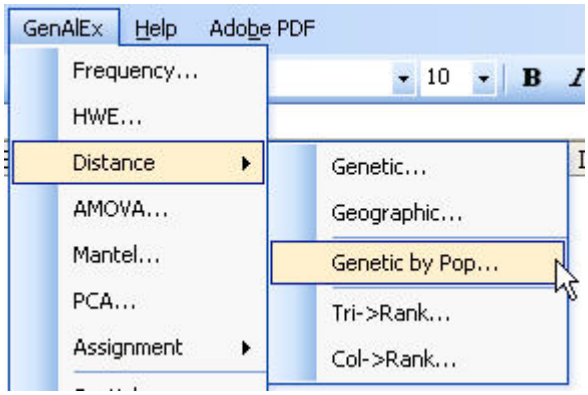


How does the heterogeneity test work?

To illustrate how this test works, assume we are testing for sex-biased dispersal. After standard spatial autocorrelation analysis of males and females as separate populations, GenAlEx computes the nonparametric heterogeneity tests of Smouse et al. (2008), as follows: First, the pooled within-population autocorrelation (across both sexes) is estimated, representing the base autocorrelation levels under the null hypothesis of no difference between the sexes. Next, the distribution of random departure from this average is tested by bootstrap resampling. The bootstrapping is achieved by randomly drawing paired samples from across the two populations, but maintaining the original samples sizes within each distance class. Next a squared paired-sample t-test statistic 'T2' (t_2) for each distance class is computed to evaluate the upper tail probability that the observed T2 value is larger than expected under the null hypothesis. In the final step, the two correlograms are compared, drawing on the P values for the T2 statistic at each distance class, across both sexes (populations), to compute the correlogram wide 'Omega'. Finally, the probability that observed Omega is larger than expected under the null hypothesis of homogeneous correlograms is determined. The null hypothesis for this test predicts homogeneity between the spatial correlograms of the two sexes, while the alternative hypothesis predicts heterogeneity.

The Genetic by Pop options

A new option ***Genetic by Pop*** for calculating the average pairwise genetic distance among populations as described in Beck et al. (2008) is available via the ***Distance*** menu.

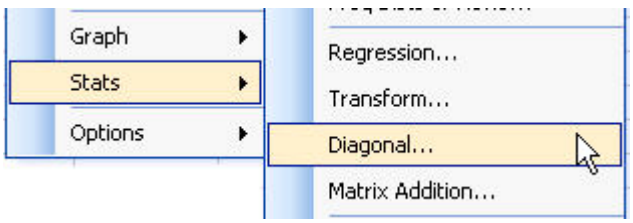


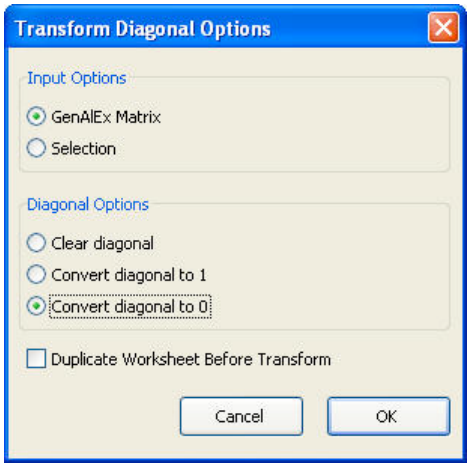
This option offers the standard genetic distance options dialog box for computing pairwise individual–by-individual genetic distances. After choosing this option, GenAlEx produces a new worksheet 'PopGD' in addition to the standard output options for the chosen genetic distance metric. In this worksheet the arithmetic average of the individual-by-individual pairwise genetic distances is provided by population. Note that in general it is expected that this matrix will be highly correlated with Φ_{PT}, F_{ST} and *Nei* Genetic distance matrices. However, we offer this output because the scale will be the same as the chosen individual-by-individual distance potentially enabling a more meaningful comparison between individual and population level spatial analyses.

	A	B	C	D	E	F	G	H	I
1	1	5							
2	Random cc	D	Pairwise Population Matrix of Mean Pop Codom Genotypic Genetic Distance						
3	Pop1	Pop2	Pop3	Pop4	Pop5				
4	3.511	3.920	3.350	3.470	3.500	Pop1			
5	3.920	4.156	3.520	3.870	3.800	Pop2			
6	3.350	3.520	3.156	3.300	3.300	Pop3			
7	3.470	3.870	3.300	3.644	3.500	Pop4			
8	3.500	3.800	3.300	3.500	3.622	Pop5			
9									

With the following additional steps this average population matrix can be used in standard spatial autocorrelation, 2D LSA and Mantel test routines:

1. Conversion of matrix diagonal to zero. This can be achieved via the *Stats->Transform Diagonal* options. This step is required because presently the matrix provided includes the average within population distance along the diagonal.





2. Creation of a pairwise population-by-population geographic distance matrix to match the average population genetic distance matrix. The easiest way to do this is to generate a list of XY coordinates for each population in GenAlEx format and use the *Distance->Geographic* option to create the matrix. (Note that for 2D LSA options a separate XY coordinates worksheet is required in any case).

	A	B	C	D
1	1	5	1	5
2	Pop Coords km			
3	Sample	Pop	X	Y
4	1	Pop1	70	18
5	2	Pop2	82	55
6	3	Pop3	23	62
7	4	Pop4	89	38
8	5	Pop5	16	53

Note that these steps may be provided automatically in future versions of GenAlEx