Theor Appl Genet (2008) 117:435–447
DOI 10.1007/s00122-008-0788-1

ORIGINAL PAPER

Non-parametric smoothing of multivariate genetic distances in the analysis of spatial population structure at fine scale

C. Bruno · R. Macchiavelli · M. Balzarini

Received: 26 December 2007 / Accepted: 2 May 2008 / Published online: 1 July 2008
© Springer-Verlag 2008

Abstract Species dispersal studies provide valuable information in biological research. Restricted dispersal may give rise to a non-random distribution of genotypes in space. Detection of spatial genetic structure may therefore provide valuable insight into dispersal. Spatial structure has been treated via autocorrelation analysis with several univariate statistics for which results could dependent on sampling designs. New geostatistical approaches (variogram-based analysis) have been proposed to overcome this problem. However, modelling parametric variograms could be difficult in practice. We introduce a non-parametric variogram-based method for autocorrelation analysis between DNA samples that have been genotyped by means of multilocus-multiallele molecular markers. The method addresses two important aspects of fine-scale spatial genetic analyses: the identification of a non-random distribution of genotypes in space, and the estimation of the magnitude of any non-random structure. The method uses a plot of the squared Euclidean genetic distances *vs.* spatial distances between pairs of DNA-samples as empirical variogram. The underlying spatial trend in the plot is fitted by a non-parametric smoothing (LOESS, Local Regression). Finally, the predicted LOESS values are explained by segmented regressions (SR) to obtain classical spatial values such as the extent of autocorrelation. For illustration we use multivariate and single-locus genetic distances calculated from a microsatellite data set for which autocorrelation was previously reported. The LOESS/SR method produced a good fit providing similar value of published autocorrelation for this data. The fit by LOESS/SR was simpler to obtain than the parametric analysis since initial parameter values are not required during the trend estimation process. The LOESS/SR method offers a new alternative for spatial analysis.

Keywords Microsattellite markers · Variograms · Correlograms · Smoothing

Communicated by J. Yu.

C. Bruno (✉) · M. Balzarini
Biometry Unit, College of Agriculture,
Universidad Nacional de Córdoba,
Av. Valparaíso s/n. Ciudad Universitaria,
CC 509 (5000) Córdoba, Argentina
e-mail: cebruno@agro.uncor.edu

R. Macchiavelli
Department of Agronomy and Soils,
University of Puerto Rico, P·O.Box 9030,
Mayagüez 00681-9030, Puerto Rico

Introduction

Population genetic structure is a subject of highly refined statistical analyses. The dispersion of individuals of a given species is an evolutionary process contributing to spatial structures even in a single population. Understanding the spatial population structure is crucial for species management and conservation (Manel et al. 2003). Populations can modify their genetic constitutions by processes such as genetic drift, mutation, gene flow, and natural selection (Hedrick 2005). While gene flow and genetic drift act over the whole genome, natural selection generally operates on individual loci. If natural selection favours different alleles in different populations, spatial differentiation is expected. Even in a single population local spatial patterns are expected under restricted gene flow. Therefore, restricted dispersal may give rise to non-random distribution of genotypes in space.

Non-random patterns of spatial distribution at short distances produce positive spatial autocorrelations, i.e., nearby observations tend to be more similar than the distant ones. By definition, the spatial autocorrelation is the correlation of a variable with itself through space (Lembo 2007). If nearby or neighbouring areas yield observations that are more alike, a positive spatial correlation arise. Such type of correlation measures the extent to which the occurrence of an event in a location affects the occurrence of the same event in a nearby location (Wagner et al. 2005). Detection of spatial genetic structure by autocorrelation analysis may therefore provide valuable insight into dispersal.

A number of quantitative methods and models that use genetic correlated data to address the study of population structure are available. Some rely on genetic variability measures using statistics like F_{st} (Cockerham 1969) to test correlation of population genetic diversity with geographic distance by means of permutation tests (Hardy and Vekemans 2002). These studies usually involve predefined populations that are large distances apart and hypothesise overall absence of spatial structure. The designs are conducted under the premise of random mating and restricted gene flow between populations.

Other methods are focused in the isolation-by-distance effects that may arise within continuous populations under restricted gene flow. Isolation-by-distance effects may lead to spatial process (Sokal and Neal 1978; Sokal and Wartenberg 1983). These studies commonly imply genetic data from a sample of individuals at short spatial distances or collected from plots embedded within a larger population. Usually a goal is to cluster together individuals who are genetically similar to analyse how these clusters relate to spatial distances (Pritchard et al. 2000). Indices such as Moran's *I* (Moran 1950), Geary's *C* (Geary 1954), Ripley's *K* (Ripley 1977) and joint-count statistics (Epperson 2003) account for autocorrelation and have been widely used to quantify short distance spatial structures. However, the common practice of assessing the extent of spatial genetic structure from these indexes, for example the distance at which a Moran's *I* correlogram reaches zero, could be misleading since the estimation strongly depend on the sampling design (Vekerman and Hardy 2004).

Contrary to theoretical expectations, the genetic spatial structure at a short scale is rarely consistent across loci, so the analysis of each allele separately may not be sensitive enough to detect spatial structures (Heywood 1991). It is, therefore, recommended to use multivariate approaches based on the analysis of multiple loci simultaneously (Smouse and Peakall 1999). Therefore, hypervariable markers such as microsatellites (Litt and Luty 1989) are useful for these types of studies. Microsatellite (SSR) markers are common since they increased significantly the number of alleles and loci to assess differences between individuals (Epperson 2000; Coltman et al. 2003; Peakall et al. 2003; Marquardt and Epperson 2004). Polymerase chain reaction (PCR) amplification of microsatellites shows in many species that they are highly polymorphic, somatically stable and inherited in a codominant Mendelian manner. Smouse and Peakall (1999) defined a multivariate genetic distance between a pair of individuals for several multiallelic codominant loci. The unweighted distances as well as other weighting schemes for allele-specific contributions at the locus level (Paetkau et al. 1995) provide a multiple loci Euclidean dissimilarity between molecular profiles. Smouse and Peakall (1999) introduce an approach to autocorrelation analysis from SSR data which is based on the expected multivariate genetic distances of DNA samples that are 'h steps apart' (lags) in the space. They defined the squared genetic distance between a pair of genotypes as one half of the Euclidean distance between vectors containing allele frequencies as elements. The multilocus distance is obtained adding single-locus distances across loci and they are used to calculate empirical correlograms that shows correlation coefficients between pairs of observations at each one of several lags (Smouse and Peakall 1999). The method is implementing in GenAlEx 6 software (Peakall and Smouse 2006). A characteristic of these empirical correlograms is that the magnitude of autocorrelation and the extent of non-random structure may depend on the choice of lag size (Hardy et al. 2003; Peakall et al. 2003).

As alternative to solve some of above problems, variograms of different genetic diversity measures have been proposed (Piazza and Menozzi 1983; Monestiez and Goulard 1997; Wagner 2003, 2004). Commonly, the term variogram refers to a plot of the semivariance against spatial distances. However, several variance measures have been estimated from pairwise comparisons to build variograms that provide an estimate of genetic diversity as a function of spatial distances (Wagner et al. 2005). Supported by the fact that under stationarity (constant mean and variance for the spatial process) the autocorrelation depends only on the spatial distance between sampling units, the empirical semivariance can be expressed as half of the squared Euclidean distance between molecular profiles (Schabenberger and Gotway 2005). Therefore, variogram-based analysis could be implemented from genetic distances. The geostatistical approach avoids the need to select groups of pairs of observations for specified lags, so since it models a continuous function on the whole spatial distance domain. Except for distances between observations smaller than by the nugget effect (which can be estimated), spatial distances are regarded as continuous variables avoiding impacts of different neighbourhood size of discrete correlograms (Double et al. 2005). In geostatitics of genetic distances, the samples which are spatially separated by a spatial distance greater than a threshold value could be regarded as not genetically correlated. Such threshold is estimated by the range of the variogram. A high value for the range suggests that the spatial structure extends to longer distances. Thus, the existence of a range different from zero suggests a non-random distribution of genotypes in space.

In the classical geostatistical approach, the common practice is to fit parametric variogram models, which are non-lineal functions of the semivariances. Numerical estimation procedures which demand initial values for the variogram parameters are required. In this paper, we propose to work under a geostatistical approach, but modelling the relationship underlying the plot of genetic distances (instead semivariance) and spatial distances between pairs of multilocus-multiallele marker profiles by means of a non-parametric variogram. The proposal is based on the smoothing of the spatial trend in the squared Euclidean genetic distances using local regression technique (LOESS) (Cleveland et al. 1988; Cleveland and Grosse 1991). The predicted LOESS values are then modelled by segmented regressions (SR) (Jennrich and Moore 1975) to obtain classical semivariogram parameters such as the range (extent of autocorrelation). The approach is supported by the Euclidean property of the multivariate genetic distances that allows visualising such plot as a semivariogram. An important advantage with regard to parametric variogram analysis is that the new method does not require initial parameter values.

We illustrate the application of the new geostatistical method using a microsatellite dataset from the native Australian bush rat for which strong spatial genetic analysis has previously been reported (Peakall et al. 2003). Full background on bush rats is provided in a series of paper by Peakall et al. that have explored the genetic and ecological impacts of habitat fragmentation on the Australian bush rat, *Rattus fuscipes* (Lindenmayer and Peakall 2000; Peakall et al. 2003, 2006; Peakall et al. 2006). Peakall et al. (2003) reported finding of strong fine-scale spatial genetic structure in bush rats leading to the prediction that restricted per generational dispersal was the basis for this pattern. Subsequent mark-recapture studies employing genetic tagging confirmed highly restricted movements of animals consistent with this prediction. An experimental perturbation study revealed patterns of population recovery consistent with the emerging evidence for restricted gene flow in bush rats. We use this previous knowledge to compare the results obtained from the parametric and non-parametric approaches of variogram analyses.

Models and methods

Geostatistical approach

The autocorrelation in a random spatial process can be analysed via variograms (Cressie 1993), which model mean or variance of observation differences as function of a continuous domain of spatial distances. If $Z(s)$ is an attribute Z observed at the spatial location $s = [x, y]'$, the term spatial autocorrelation refers to the correlation between $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$, i.e., the correlation between the same attribute at two different locations $\mathbf{s}_i$ and $\mathbf{s}_j$. Second-order (or weak) stationarity of a random process implies that the observation mean is constant, $E[Z(\mathbf{s})] = \mu$, and the covariance between attributes at different locations is only a function of their spatial separation (the lag-vector), $\mathrm{Cov}[Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})] = \mathrm{C}(\mathbf{h})$. Thus, the semivariogram function of a stationary random process, denoted by $\gamma(\mathbf{s}_i - \mathbf{s}_j)$, only depend on the spatial distance between observations, and usually is expressed as a semivariance,

$$\gamma(\mathbf{s}_i - \mathbf{s}_j) = \frac{1}{2}\left\{\mathrm{Var}\ \left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]\right\} \tag{1}$$

where $\mathbf{s}_j = \mathbf{s}_i + \mathbf{h}$ and $\mathbf{h}$ is the spatial distance o lag between sample point $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$. For a second order stationarity process, the semivariogram can be also expressed as the expected value of the squared distance between observations,

$$\gamma(\mathbf{s}_i - \mathbf{s}_j) = \frac{1}{2}\mathrm{E}\left[\left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right)^2\right] \tag{2}$$

Therefore, an empirical estimator of a semivariogram, due to Matheron (1962, 1963), can be obtained as one half of the average squared distances between pairs of observations which are $\mathbf{h}$ steps apart,

$$\hat{\gamma}(\mathbf{s}_i - \mathbf{s}_j) = \frac{1}{2\left|N(\mathbf{s}_i - \mathbf{s}_j)\right|}\sum_{N(\mathbf{s}_i - \mathbf{s}_j)}\left[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right]^2 \tag{3}$$

where $N(\mathbf{s}_i - \mathbf{s}_j)$ represents the set of location pairs with coordinate difference $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ and $\left|N(\mathbf{s}_i - \mathbf{s}_j)\right|$ is the number of distinct pairs in this set (Schabenberger and Gotway 2005). The name semivariogram is used both for the function $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ as well as the graph of $\gamma(\mathbf{h})$ against $\mathbf{h}$. When working with covariances, $\mathrm{C}(\mathbf{s}_i - \mathbf{s}_j) = \mathrm{Cov}\left[\mathrm{Z}(\mathbf{s}_i), \mathrm{Z}(\mathbf{s}_j)\right]$, the graph of $\mathrm{C}(\mathbf{h})$. against $\mathbf{h}$ is referred to as the covariogram, and $\mathrm{C}(\mathbf{0})$ represent a variance component. Similarly, a graph of the correlation $\mathrm{R}(\mathbf{h})$ (standardised covariance) against $\mathbf{h}$ is termed the correlogram. If equal variance is assumed (stationarity), the correlation can be expressed as:

$$\mathrm{Corr}[Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})] = 1 - \frac{\gamma(\mathbf{h})}{\mathrm{Var}[Z(\mathbf{s})]} \tag{4}$$

When the covariance function is monotonous decreasing, the semivariogram is increasing and three parameters are commonly used to describe the spatial trend, the sill, the range and the nugget. The sill is the upper asymptote value of the semivariogram and represents the variance under autocorrelation. The range is the spatial distance at which the semivariogram reaches the sill. From a practical point of view, the observations are regarded as uncorrelated if they are spatially separated at a distance larger than the range. When the semivariogram approaches the sill only asymptotically, i.e., an exponential semivariogram, the practical range is defined as the distance at which the semivariogram achieves 95% of the sill. A third parameter of the semivariogram is the nugget effect, θ_0, which relates to variability in a microscale, it represents a variance component that is not spatially structured and is visualised as a discontinuity at the origin. If a semivariogram has nugget θ_0 and sill $\mathrm{C}(\mathbf{0})$, the difference $\mathrm{C}(\mathbf{0}) - \theta_0$ is called the partial sill, θ_η. The practical range is then defined as the distance at which the semivariogram has achieved $\theta_0 + 0.95\theta_\eta$. In the no-nugget model the population variance is directly represented by $\mathrm{Var}[\mathrm{Z}(\mathbf{s}_i)] = \mathrm{C}(\mathbf{0})$ (Fig. 1). The faster the semivariogram rises from the origin to the sill, the faster the spatial structure declines. Commonly used theoretical semivariogram models are expressed as exponential, spherical and Gaussian functions; all of them are continuous function of **h** (Fig. 2). The empirical correlogram can be computed from a semivariogram as:

$$\text{correlogram} = 1 - \frac{\text{predictedsemivariogram}}{\text{estimated population variance}} \tag{5}$$

Variogram estimation

As showed in Fig. 2 the commonly used parametric models are nonlinear in the parameters, so they are estimated by means of iterative numerical procedures which demand initial parameter values (Schabenberger and Gotway 2005). In this paper, we propose to replace the classical parametric fitting of theoretical models by smoothing procedures such LOESS (local regression) (Cleveland et al. 1988; Cleveland and Grosse 1991) to avoid the specification of initial parameter values. LOESS assumes that, for $i = 1$ to N, the ith measurement y_i of a response variable y and the corresponding measurement x_i of its predictor are related as $y_i = g(x_i) + \varepsilon_i$ where g is a regression function and ε_i is a random error. The idea of local regression is that near $x = x_0$, the regression function $g(x)$ can be locally approximated by the value of come function (Loader 2004). Such a local approximation is obtained by fitting a regression surface to the data points (x_i, y_i) within a chosen neighbourhood of the point x_0. In the LOESS method, weighted least squares are used to fit linear or quadratic regression surface at each neighbourhood centre (x_0). Data points in a given neighbourhood are weighted by a decreasing function of their distance from the centre of the neighbourhood (smooth). The radius of each neighbourhood is chosen so that the neighbourhood contains a specified percentage of data points (local points). The fraction of the data in each local neighbourhood controls the smoothness of the estimated function and is called the smoothing parameter. There are several alternatives that can be used to select the smoothing parameter. One strategy is to use several values for the smoothing parameters and then statistics, such the Akaike Information Criterion (Akaike 1973), that allow statistical fitting comparisons. Since classical selectors, such AIC, tend to undersmooth and tend to be non-robust in the sense that small variations of the input data can change the choice of the smoothing parameter value significantly, Hurvich et al. (1998) obtained two bias corrected AIC criteria, named as $\mathrm{AIC_c}$ and $\mathrm{AIC_{c1}}$. Another strategy is to use generalised cross-validation (Craven and Wahba 1979) or residual plots. When using these criteria for smoothing parameter value selection (smaller values are better) more than one smoothing function could result appropriated (similar criteria values). The final strategy is to choose the largest smoothing parameter that yields no clearly discernible fitting. Cohen (1999) provided a SAS macro for automatically select the smoothing parameter.

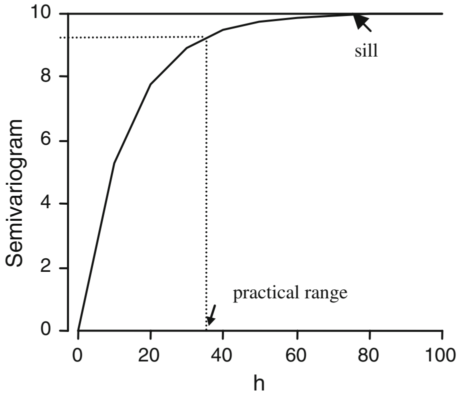
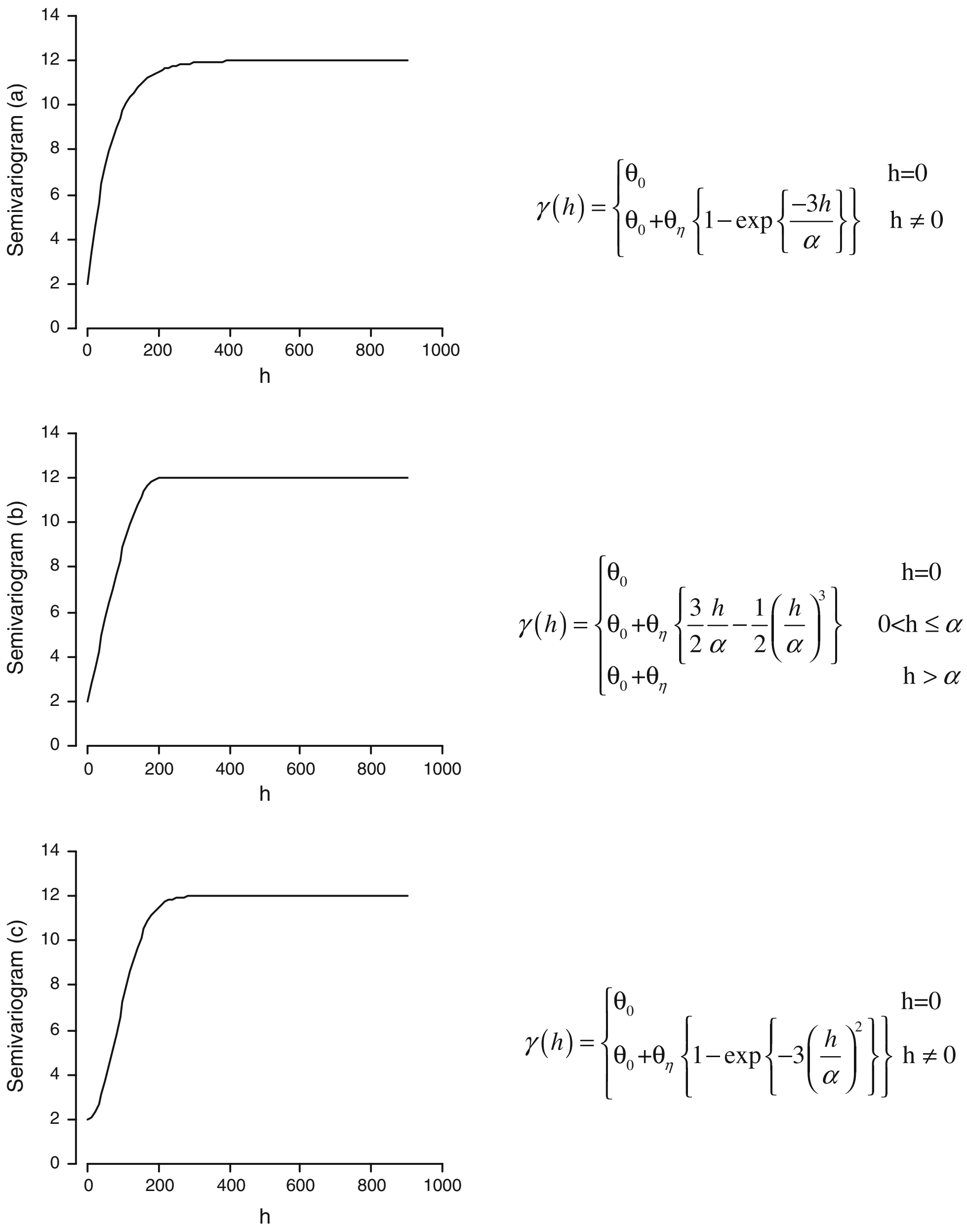


Fig. 1 Semivariogram. Spatial process with positive covariance function

The LOESS-based method

We propose to fit a semivariogram from eq. (3). Instead of semivariances, our input data are the multivariate observation squared distances. The data $Z(\mathbf{s})$ are microsatellite allele frequencies in the sampled genotype at location $\mathbf{s}$. The response variable y, that will be fitted by LOESS, is one-half of the Euclidean genetic distance between genotypes, which is expressed as proposed by Smouse and Peakall (1999). The squared genetic distance will be plotted as function of x = spatial distance between $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$. After the predicted trend of the squared Euclidean genetic versus spatial distances is obtained by LOESS the classical parameters describing the underlying population structure are estimated by means of segmented regressions



$$\gamma(h)=\begin{cases}\theta_0 & h=0\\ \theta_0+\theta_\eta\left\{1-\exp\left\{\frac{-3h}{\alpha}\right\}\right\} & h\neq 0\end{cases}$$

$$\gamma(h)=\begin{cases}\theta_0 & h=0\\ \theta_0+\theta_\eta\left\{\frac{3}{2}\frac{h}{\alpha}-\frac{1}{2}\left(\frac{h}{\alpha}\right)^3\right\} & 0<h\leq\alpha\\ \theta_0+\theta_\eta & h>\alpha\end{cases}$$

$$\gamma(h)=\begin{cases}\theta_0 & h=0\\ \theta_0+\theta_\eta\left\{1-\exp\left\{-3\left(\frac{h}{\alpha}\right)^2\right\}\right\} & h\neq 0\end{cases}$$

Fig. 2 Exponential (**a**), spherical (**b**) and Gaussian (**c**) semivariogram theoretical models. Nugget effect $\theta_0 = 2$, partial sill $\theta_\eta = 10$ and range $\alpha = 200$ m

(SR) (Jennrich and Moore 1975) of the predicted LOESS values as a continuous function of the spatial distance. The proposed LOESS/SR method is non-parametric with regard to the estimation of spatial trend, the correlograms are derived from Eq. (5) and no initial parameter values are required.

Discrete correlograms

Smouse and Peakall (1999) introduce a procedure to obtain correlograms from direct estimation of correlation coefficient for a discrete series of lag (**h**) previously selected. An Euclidean metric is used to compute a correlation coefficient for the series of chosen lags which reflects the covariance pairs of individuals that are ‘h steps apart’. The correlation coefficient has a zero value when there is not autocorrelation. Two procedure parameters, the distance class size and the number of distance classes, should be defined prior to calculate the correlation coefficients that will yield the correlogram. These parameters determine the series of lags. The first distance class for which the coefficient correlation will be calculated includes all distances in the interval between zero and the fixed distance class. The spatial analysis considers all samples that are represented by a distance greater than the previous distance class, and lesser or equal than the upper distance class (Flanagan 2006). The sets of pseudocorrelations are then sorted, and a $(1 - \alpha)\%$ confidence interval is constructed from the (1 − α/2)th value and the (α/2)th value, respectively (Smouse and Peakall 1999). A significance test at each lag is obtained by comparing the observed correlation coefficient with those obtained from a large number of spatial permutations of the same sampled individuals.

Applications to data

Data and statistical approaches

The non-parametric and parametric variogram approaches to model trends in genetic distance versus spatial distance plots are illustrated by using a dataset involving 38 individual DNA samples of *Rattus fuscipes* generated by Peakall et al. (2003). This dataset contains just one of the eight populations in the original dataset published by Peakall et al. (2003). The file involves four microsatellite loci called C2, E5, CR and PB used to evaluate ecological and genetic impacts of habitat fragmentation on the spread of Australian bush rat. Lindenmayer and Peakall (2000) provide the background to the microsatellite loci used to genotype georeferenced samples in such study. The samples were taken at random from 1 km. transects within a single population. A thorough understanding of the sampling and biological context underlying this dataset can be obtained from Peakall et al. (2003) and (2006). The dataset used here is named as BushRat SinglePop.xls, and constitutes an example file in GenAlEx 6 software (Peakall and Smouse 2006). The dataset was previously analysed by Peakall et al. (2003), reporting a positive spatial correlation up to 200 m. This result was used here to compare the findings obtained from the non-parametric and parametric variogram-based analyses.

The relationship between squared multivariate genetic and spatial distances was modelled using LOESS as implemented in SAS Proc Loess (SAS Institute 2004, version 9.1). By means of information criteria derived from alternative fittings with different smoothing parameters (Table 1), we selected a smoothing parameter value equals to 0.6. Selecting the optimal smoothing function in this way will have little impact on the outcomes of the spatial analysis as shows the last column in Table 1.

Fig. 3 Non-parametric (**a**) and parametric exponential (**b**), spherical (**c**) and Gaussian (**d**) fitting of squared Euclidean genetic distances versus spatial distances and derived correlograms ▶

Two connected polynomial functions or segmented regressions (Jennrich and Moore 1975), one of zero-slope, were fitted over the distance predicted by LOESS. These linear approximations to the smoothed function allowed the estimation of the “sill” (genetic variability under spatial structure), and the “range” (distance beyond which observations are spatially uncorrelated). An indicator variable, functionally dependent on the range, was used to fit the segmented regressions (Appendix A). Three geo-statistical parametric models (exponential, spherical and Gaussian) were also fitted using SAS Proc Nlin (SAS Institute, 2004. version 9.1) (Appendix B). Both, non-parametric and parametric variograms were applied to the Euclidean measure of multivariate genetic distances as well as to single-locus distances. The LOESS fit was compared with those from parametric models using residual mean squares. The dataset was also analysed using GenAlEx software (Peakall and Smouse 2006) with distance classes of 10, 35, 50 and 200 m which implied discrete correlograms base on 10, 35, 50 and 200 lags, respectively.

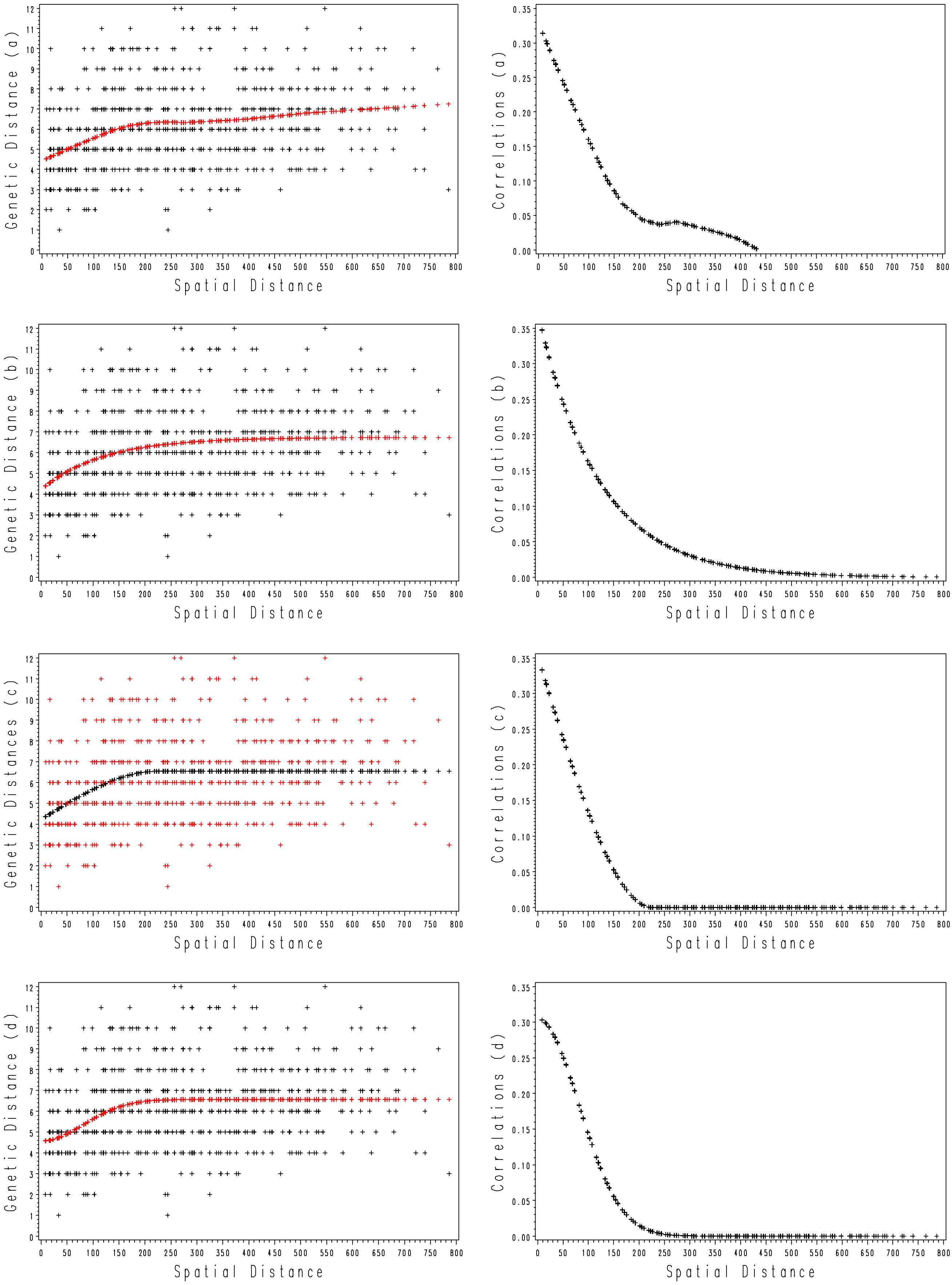
Results

The following results were obtained from the multivariate genetic distances. Figure 3 shows the fitted trends in the genetic distance versus spatial distance plot and the derived correlograms for the parametric and the non-

Table 1 Statistical criteria for smoothing parameters value selection in the example dataset

Smoothing parameter	Local points	Fitting criteria			Predicted range
		GCV	AIC_C	AIC_{C1}	
0.1	69	0.00554	2.3542	1641.08975	167
0.2	139	0.00546	2.3392	1630.50268	169
0.3	209	0.00544	2.3362	1628.31969	175
0.4	278	0.00544	2.3364	1628.48877	180
0.5	348	0.00545	2.3367	1628.70751	186
0.6	418	0.00545	2.3372	1629.01259	213
0.7	487	0.00547	2.3412	1631.80800	269
0.8	557	0.0055	2.3459	1635.06294	324
0.9	627	0.00552	2.3511	1638.72101	409

GCV Generalised cross-validation; AIC_c and AIC_{c1}: bias corrected Akaike information criterion


Genetic Distance (a)
Correlations (a)
Genetic Distance (b)
Correlations (b)
Genetic Distances (c)
Correlations (c)
Genetic Distance (d)
Correlations (d)
Spatial Distance
0 1 2 3 4 5 6 7 8 9 10 11 12
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750 800

parametric approaches to variogram-based analysis. In Table 2, we present the estimated parameters (sill, range and nugget effect) under each approach. The plot of squared Euclidean genetic distance between DNA samples $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$ against $\mathbf{h}$ (sample spatial distance) fitted by LOESS suggested an empirical range of 213 m. This result is similar to those obtained by the Gaussian and Spherical models (range 199 and 225 m, respectively) because the error is similar into both ways. These findings show that individuals might be genetically related up to distances of about 200 m. Even though very similar fittings were obtained for all approaches, the residual mean square (Table 3) shows a relative better fit whit the LOESS/SR method. The estimate of the sill by the segmented regressions was 6.6, and the estimate of the nugget effect 4.5. These values allow inference about the population variance under autocorrelation. In agreement with the findings of Peakall et al. (2003) the new approach for autocorrelation analysis also indicates that proximate individuals are more related than average. If individual relationship were spatial distance independent (samples random distributed in space), a constant function at the average genetic distance will fit the trend. It is of further interest that the extent of spatial genetic structure detected in this analysis (199–359 m) under the different models is similar to that estimated by Peakall et al. (2003) (200 m for a distance class size of 50 m, 400 m for a distance class of 200 m, see also Fig. 4).

In Table 4, we show the number of observation pairs available at each lag ($\mathbf{h}$) for distance classes of 35 and 50 m with the average and the variance of the Euclidean genetic distances at each lag. Those mean values are defining the semivariograms trends captured by the discrete correlogram and the variogram based-analyses. However, in the variogram approaches the selection of a distance class size is not a requirement.

Table 2 Estimates of the nugget, sill and range parameters and non-parametric functions modelling multivariate squared Euclidean genetic distances versus spatial distances

Fit	Parameter	Estimate
Exponential semivariogram	Nugget	4.2
	Sill	6.8
	Range	359.0
Spherical semivariogram	Nugget	4.2
	Sill	6.5
	Range	225.0
Gaussian semivariogram	Nugget	4.6
	Sill	6.6
	Range	199.0
LOESS/SR semivariogram	Nugget	4.5
	Sill	6.6
	Range	213.0

Table 3 Goodness of fit of squared Euclidean genetic distances versus spatial distances from parametric and non-parametric geostatistical approaches

Fit	RMS
Exponential semivariogram model	3.7967
Spherical semivariogram model	3.7942
Gaussian semivariogram model	3.7866
LOESS	3.7784

RMS Residual mean square

In Fig. 4, the correlograms obtained from GenAlEx 6 software (Peakall and Smouse 2006) for several sizes and number of distance class are shown. Significant autocorrelations are suggested for all values of distance classes since all curves cross the *x*-axis and the verticals bars (which represent 95% confidence intervals) that not touch the zero line (null correlation coefficient) for at last one lag. However, the spatial distance with significant correlation coefficient is highly dependent on the selected distance classes. The standard errors of the correlation coefficient estimates are also dependent on the distance class size since the number of data points changes (Table 5). The stability of the estimates throughout hundred runs of the procedure was high, which was reflected by a low coefficient of variation of the obtained results (Table 5).

In the locus-by-locus analyses, the LOESS/SR method as well as the Gaussian parametric semivariogram produced the best fits of the squared genetic distance versus spatial distance plot for most loci. Table 6 shows the estimates of the nugget effect, sill and range of the best fit among the parametric semivariogram models and the LOESS/SR method for the single-locus analyses. The spatial structure is not consistent across loci, as is reflected by the high variability of semivariogram range. The estimated range from a plot obtained from multivariate genetic distances is not the average single-locus range. However, for the E5 and CR loci the function relating genetic distance with spatial distance showed similar shape than the function modelling multivariate distances. The E5 and CR were those loci with higher effective allele number and higher expected heterozygosis (Table 7).

A cline structure may cause a linear change of semivariances with distance (Wagner et al. 2005), so linear fit may result better than nonlinear variograms for same loci. The extent of the genetic dissimilarities is higher in the multivariate approach than in the single-locus one as is reflected by the sill estimates of Table 6 with respect to those in Table 2.

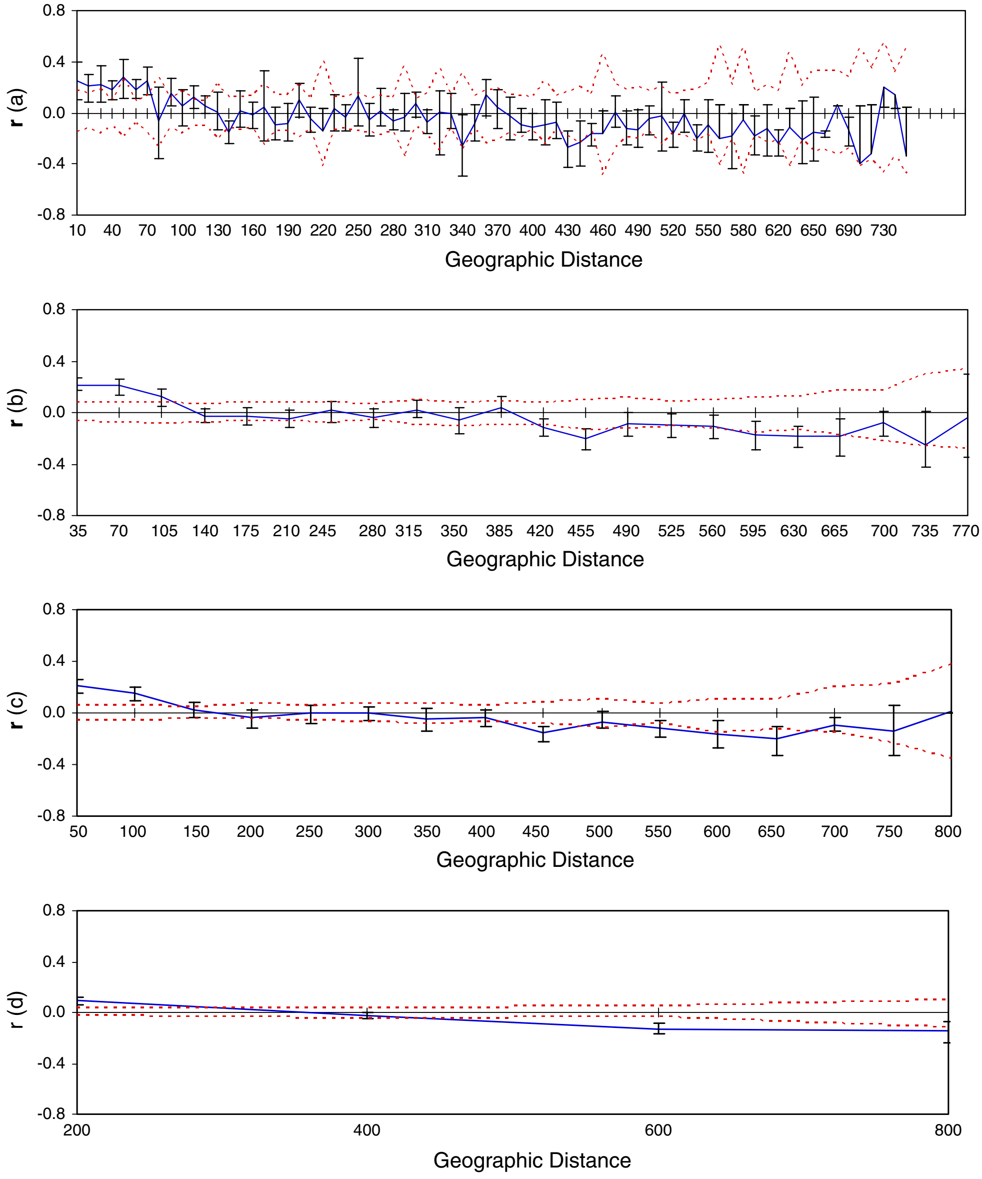


Fig. 4 Correlograms obtained from GenAlEx 6 software for 10 (**a**), 35 (**b**), 50 (**c**) and 200 m (**d**) as distance class sizes. *Vertical bars* show standard errors of the correlation coefficient ***r*** at each distance class. *Dotted lines* show lower and upper 95% confidence limits for the expected correlation

Table 4 Number of pairs of observations available at each lag (**h**), and the corresponding averages and variances of genetic distance values

35 h step apart	Data pairs	Average d_{ij}^2	Variance d_{ij}^2	50 h step apart	Data pairs	Average d_{ij}^2	Variance d_{ij}^2
35	54	4.80	3.07	50	68	4.85	2.90
70	110	5.02	2.68	100	167	5.56	3.68
105	57	6.32	4.08	150	62	6.40	4.08
140	53	6.32	3.95	200	65	6.25	3.78
175	44	6.50	3.47	250	65	6.42	5.25
210	44	6.14	4.35	300	47	6.40	5.72

d_{ij}^2 : squared Euclidean genetic distance between $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$

Discussion

This paper introduces a new approach for autocorrelation analysis from genetic data at fine scale. The method is supported by a well-established geostatistical approach, i.e., the variogram-based analysis, with modifications that attempt to reduce some of the underlying assumptions to facilitate variogram fittings. Instead to work with semi-variances, the method uses a plot of the squared Euclidean genetic distances *vs*. spatial distances between pairs of

Table 5 Expected values of correlogram x-intercepts ("range") for distance class 10, 35, 50 and 200 m from discrete correlograms

Distance class size (m)	Average x-intercept (m)	Coefficient of variation x-intercept (%)	Number of observations at each lag
10	79	0.61	10
35	134	0.59	32
50	170	1.53	44
200	363	0.73	176

Coefficient of variation for the estimated x-intercept from 100 runs of 1,000 permutation cycles for each distance class

Table 7 Diversity indices, number of alleles (N_a), effective number of alleles (N_e), information index (I), the observed (H_o) and expected heterozygosity (H_e) per locus and at the whole population

Locus	N	N_a	N_e	I	H_o	H_e
C2	38	6.000	3.903	1.551	0.816	0.744
E5	38	9.000	6.763	2.034	0.921	0.852
CR	38	9.000	5.378	1.833	0.763	0.814
PB	38	7.000	3.446	1.429	0.763	0.710
Population mean	38	7.750	4.873	1.712	0.816	0.780

DNA-samples. The use of this plot as empirical variogram is supported by the fact that for any stationary spatial process the variances can be expresses as Euclidean distances (Schabenberger and Gotway 2005) and the calculated genetic distances are metrics (Smouse and Peakall 1999). Distances are commonly easier to understand the variances, then the plot is easier to understand than the classical semivariogram used in geostatistics.

As other variogram-based approaches the proposed method allows addressing two important aspects of fine-scale spatial genetic analyses: the identification of a non-random distribution of genotypes in space, and the estimation of the magnitude of any non-random structure. Both aspects are related with the x point at which the variogram reaches or approaches asymptotic values. If there is not a spatial structure a constant fit is expected since genetic distance is not related with spatial distance. If the variogram represents a stationary process as indicated by the presence of a sill, the range can be estimated. In such case, the underlying trend in the genetic distances against the spatial sample separation can be fitted by two simple polynomials, one with positive slope and other with zero slope, the magnitude of the autocorrelation is estimated by the x point at which this change exist. The direct statistical comparison of ranges among variograms is correct since it is estimated under the spatial process, and no permutational methods are required for assessing statistical significance of the extent of autocorrelation.

The analogous parameter in the classical variogram approach is the semivariogram range. Since the expected shape of spatial genetic structure is usually nonlinear in its parameter, the classical variogram fittings uses iterative numerical procedures which demand initial values (predefinition) for the parameters including the range. With the use of LOESS (Cleveland et al. 1988; Cleveland and Grosse 1991), which is a smoothing technique free of need to specify initial parameter values we successfully fitted the trend embedded in the distance-based variogram. In such way, the LOESS/SR method avoids typical assumptions on parameter values of classical variogram analysis.

LOESS fits the trend of interest by running several local linear weighted regressions. In the LOESS/SR method one still has to select a radius for a local neighbourhood in the smoothing process, but this can be done by means of well established information criteria. The smoothing parameter is related to the empirical fit and not to the underlying spatial process since the x points in a distance neighbour are not necessary neighbours in the space. The LOESS procedure does not work with fixed windows of data; it uses the idea of moving windows or overlapped subsets of distance data. Therefore, the method minimises potential problems associated with the predefinition of distance classes, a point made by other authors for discrete correlograms (Smouse and Peakall 1999; Hardy et al. 2003; Peakall et al. 2003). A range of smoothing parameters could produce good fittings of the data. However, when the smoothing parameters are identified by likelihood based criteria and conservatively selected, the effect if any different smoothing parameters

Table 6 Estimates of nugget, sill and range parameters of functions modelling single-locus squared Euclidean genetic distances versus spatial distances

Locus	Function[a]	Nugget estimate	Sill estimate	Range estimate
C2	Gaussian semivariogram	1.2	3.6	1,954
	LOESS/SR	1.1	2.1	928
CR	Spherical semivariogram	1.5	1.8	162
	LOESS/SR	1.5	1.8	107
E5	Gaussian semivariogram	1.2	1.7	130
	LOESS/SR	1.2	1.7	115
PB	Gaussian semivariogram	0.8	2.1	612
	LOESS/SR	0.7	2.0	551

[a] At each locus, results from the best of the parametrical fitting and the non-parametrical method

will have on the biological results will be smoothed. There exist automatically routines and statistical criteria that could help researches with an objective selection regarding smoothing parameters (Cohen 1999). We run the analyses under several smoothing parameters (0.01–1) and values from 0.3 to 0.6 produce good fittings according the corrected Akaike information criteria (smaller is better) and the GVC index. The biological inferences drawn from any of these neighbourhood sizes do not show significant differences. Most available statistical software provides facilities for easy implementation of LOESS.

The LOESS/SR procedure is similar to fit a tent (linear–linear) variogram model (Schabenberger and Gotway 2005) but with higher robustness to outliers because of the smoothing process carried out in the first procedure step. Since it is expected high variability of multivariate distances, the application of a smoothing procedure could favour the rescue of signal over noise in the variogram. The LOESS/SR procedure is more “robust” to the unbalanced in the number of data points at different distance classes compared to the parametric models since it is a characteristic of the smoothing procedure.

The LOESS/SR method is a new approach for variogram-based analysis that would simplify dispersal studies by autocorrelation analysis. However, it is important to remember that estimating biological parameters from empirical variogram analysis would be only valid if the scale of the study is appropriate, and the assumption of stationarity should be checked. In the parametric variogram approach, the assumption of stationarity may be statistically evaluated by comparing homoscedastics versus heteroscedastics spatial fits of the genetic distances versus spatial distance relationship using SAS PROC Mixed or other software with similar facilities. Further research on testing this assumption in the LOESS/SR procedure will be useful. However, it is logical to expect lesser impact of heterogeneity of variance (lack of stationarity) with smoothing techniques than classical model fitting. Nevertheless, simplicity of smoothing techniques as applied in variogram-based analyses invites researchers to shorten distances between theory and practice of geostatistics in genetics.

Acknowledgments We thank Professor Rod Peakall from Australian National University for very thorough and critical revisions of the manuscript. His review was very valuable for a better understanding of the biological aspects and scientific issues of this paper. We fully acknowledge Professor Rod Peakall and other authors of the research projects involving Australian Bush rat genetic studies for allowing us using the BushRat SinglePop.xls dataset which is an example file of GenAlEx 6 (Peakall and Smouse 2006) for re-analysis with other procedures. We also thank C. Gardenal a recognised argentinean geneticist (CONICET-UNC) and Dr. L. Beaver from the University of Puerto Rico for their rich comments. This work was supported by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, and by an exchange fellowship (Red de Macrouniversidades de América Latina y el Caribe, Universia). CB is a fellow of CONICET, RM is a professor of the University of Puerto Rico and MB is a professor of the University of Córdoba and researcher of CONICET.

Appendixes

Appendix A

SAS code to fit non-parametric semivariograms and correlograms form genetic squared Euclidean distance (xgen) and sample spatial (xg) distance.

```
/*---------------------------------- Non parametric fit: LOESS ------------------------------------------*/
proc loess ;
ods output OutputStatistics= Modelstats;
model xgen=xg /  residual  clm alpha=0.05 smooth=0.6 details(outputStatistics ModelSummary) ;
/*--------------------------------- Plot of LOESS semivariogram ---------------------------------------*/
proc gplot data= Modelstats;
  by SmoothingParameter;
    plot (depvar pred)*xg    / overlay  ;
/*---------------------------------- -------------Estimation of range --------------------------------------*/
proc nlin data=Modelstats;
parms nugget=4 sill=4.5 range=150;
b=(sill-nugget)/range;
 if xg<range    then    model pred=nugget+b*xg ;
 else  model pred=nugget+b*range ;
  end;
```

Appendix B

SAS code to fit parametric semivariograms and correlograms form genetic squared Euclidean distance (xgen) and sample spatial (xg) distance.

```
/*------------------------------- Parametric fit: Exponential semivariogram --------------------------*/
proc nlin       ;
parameters   sill=6.5       range=200           nugget=1.5       ;
bounds nugget > 0;
semivariogram = nugget + (sill-nugget)*(1-exp(-3*xg/range));
model xgen = semivariogram ;
output out=fitexp     predicted=pexp     residual=resexp     parms=sill range nugget;
/*------------------- Correlogram derived from Exponential semivariogram -----------------------------*/
data Correlogram ;
set fitexp;
corr=1-((nugget + (sill-nugget)*(1-exp(-3*xg/range)))/sill);
/*--------------------------- Parametric fit: Spherical semivariogram --------------------------------------*/
proc nlin  ;
parameters     sill=6.5       range=200              nugget=1.5;
bounds nugget > 0;
if xg <= range    then   do;
semivariogram = nugget+((sill-nugget)*(((3/2)*(xg/range))-(1/2*((xg/range)**3))));
model    xgen = semivariogram  ; end;
if xg > range          then       model xgen = nugget+(sill-nugget);
output out=fitSph            predicted=pSph   residual=resSph              parms=sill range nugget;
/*--------------------------- Correlogram derived from Spherical semivariogram -------------------------*/
data correlogram;
set fitSph;
if xg <= range    then         do;
Corr=1-((nugget+((sill-nugget)*(((3/2)*(xg/range))-(1/2*((xg/range)**3)))))/sill); end;
if xg > range        then     ;
Corr = 1-((nugget+((sill-nugget)*(((3/2)*(range/range))-(1/2*((range/range)**3)))))/sill);
/*------------------------------ Parametric fit: Gaussian semivariogram -----------------------------------*/
proc nlin ;
parameters      sill=6.5           range=200         nugget=1.5;
bounds nugget > 0;
semivariogram = nugget + ((sill-nugget)*(1-exp(-3*((xg/range)**2))));
model       xgen = semivariogram;
output out=fitG            predicted=pG      residual=resG             parms=sill range nugget;
/*--------------- Correlogram derived from Gaussian semivariogram -----------------------------------*/
data correlogram ;
set fitG;
corr=1-((nugget + ((sill-nugget)*(1-exp(-3*((xg/range)**2)))))/sill);
run;
```

References

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Proceedings of the second international symposium on information theory, pp 267–281

Cleveland W-S, Grosse E (1991) Computational methods for local regression. Stat Comput 1:47–62

Cleveland W-S, Devlin S-J, Grosse E (1988) Regression by local fitting. J Econometrics 37:87–114

Cockerham C-C (1969) Variance of gene frequencies. Evolution 23:72–84

Cohen R-A (1999) An introduction of PROC LOESS for local regression. Paper 273. SAS Institute Inc, Cary

Coltman D-W, Pilkington J-G, Pemberton J-M (2003) Fine-scale genetic structure in a free-living ungulate population. Mol Ecol 12(3):733–742

Craven P, Wahba G (1979) Smoothing noisy data with spline functions. Numer Math 31:377–403

Cressie N-A-C (1993) Statistics for spatial data, revised edn. Wiley, New York

Double M, Peakall R, Beck N, Cockburn A (2005) Dispersal, philopatry, and infidelity: dissecting local genetic structure in superb fairi-wrens (*Malurus cyaneus*). Evolution 59(3):625–635

Epperson B-K (2000) Spatial genetic structure and non-equilibrium demographics within plants populations. Plant Species Biol 15:269–279

Epperson B-K (2003) Geographical genetics. Princeton University Press, Princeton

Flanagan N (2006) A guide to GenAlEx6. Genetic analysis in Excel. http://www.anu.edu.au/BoZo/GenAlEx

Geary R-C (1954) The contiguity ratio and statistical mapping. Inc Stat 5:115–145

Hardy O-J, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol 2:618–620

Hardy O-J, Charbonnel N, Fréville H, Heuertz M (2003) Microsatellite allele size: test to assess their significance on genetic differentiation. Genetics 163:1467–1482

Hedrick P (2005) Genetic of populations, Third edn. Jones and Bartlett Publishers, Sudbury

Heywood J-S (1991) Spatial analysis of genetic variation in plant population. Ann Rev Ecol Syst 22:335–355

Hurvich C-M, Simonoff J-S, Tsai C-L (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J R Stat Soc B 60:271–293

Jennrich R-I, Moore R-H (1975) Maximum likelihood estimation by means of nonlinear least squares. In: American statistics proceedings of the statistical computing section, pp 57–65

Lembo A (2007) Course spatial modeling and analysis. Cornell University. http://www.css.cornell.edu/courses/620/css620.html. Cited spring 2007

Lindenmayer D, Peakall R (2000) The Tumult experiment-integrating demographic demographic and genetic studies to unravel fragmentation effects: a case study of the native Bush Rat. In: Young A, Clarke G (eds) Genetics, demography and the viability of fragmented population. Cambridge University Press, London, pp 173–201

Litt M, Luty J (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:398–401

Loader C (2004) Smoothing: local regression techniques. In: Gentle J, Wolfgang H, Yoichi M (eds) Handbook of computational statistics. Springer, Heidelberg

Manel S, Schwartz M, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol 18(4):189–197

Marquardt P, Epperson B-K (2004) Spatial and population genetic structure of microsatellite in white pine. Mol Ecol 13:3305–3315

Matheron G (1962) Traité de Geostatistique Appliquée, Tome I. Memoires du Bureau de Recherches Geologiques et Minières, N° 14. Editions Technip, Paris

Matheron G (1963) Principles of geostatistics. Econ Geol 58:1246–1266

Monestiez P, Goulard M (1997) Analysing spatial genetic structures by multivariate geostatistics: study of wild populations of perennial ryegrass (*Lolium perenne*). In: Baafi EY, Schofield NA (eds) Geostatistics Wollongong. Kluwer, Dordrecht, pp 1197–1208

Moran P-A-P (1950) Notes on continuous stochastic phenomena. Biometrika 37:17–23

Paetkau D, Calvert W, Stiling Y, Strobeek C (1995) Microsatellite analysis population structure in Canadian polar bears. Mol Ecol 4:347–354

Peakall R, Smouse P-E (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol 6:288–295

Peakall R, Rubial M, Lindenmayer D (2003) Spatial autocorrelation analysis offers new insights into gene flow in the Australian Bush rat, *Rattus fuscipes*. Evolution 57(5):1182–1195

Peakall R, Ebert D, Cunningham R, Lindenmayer D (2006) Mark-recapture by genetic tagging reveals restricted movements by bush rats (*Rattus fuscipes*) in a fragmented landscape. J Zool 268:207–216

Piazza A, Menozzi T (1983) Geographic variation in human gene frequencies. In: Felsenstein J (ed) Numerical taxonomy. Springer, Berlin, pp 444–450

Pritchard J, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Ripley B-D (1977) Modelling spatial patterns. J R Stat Soc B 39(2):172–212

SAS Institute Inc (2004) SAS STAT user's guide, version 9.1, Cary, pp 2997–3044

Schabenberger O, Gotway C (2005) Statistical methods for spatial data analysis. Chapman & Hall/CRC, London/New York, pp 42–105

Smouse P, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. Heredity 82:561–573

Sokal R-R, Neal O (1978) Spatial autocorrelation in biology, 2.Some implications and four applications of evolutionary interest. Biol J Linn Soc 10:229–249

Sokal R-R, Wartenberg D-E (1983) A test of spatial autocorrelation analysis using an isolation-by-distance model. Genetics 105:219–237

Vekerman X, Hardy J-O (2004) New insights from fine-scale spatial genetic structure analyses in plan populations. Mol Ecol 13:921–935

Wagner H-H (2003) Spatial covariance in plant communities: integrating ordination, geostatistics, and variance testing. Ecology 84:1045–1057

Wagner H-H (2004) Direct multiscale ordination with canonical correspondence analysis. Ecology 85:342–351

Wagner H, Holderegger R, Perth S, Gugerli F, Hoebee S, Scheidegger C (2005) Variogram analysis of the spatial genetic structure of continuous populations using multilocus microsatellite data. Genetics 169:1739–1752