Heredity (2005) 95, 7–15

www.nature.com/hdy

REVIEW

Estimating dispersal from short distance spatial autocorrelation

BK Epperson
Michigan State University, East Lansing, MI 48824, USA

A series of theoretical studies has formed a strong connection between spatial statistics observed in populations and summary measures of the amount of dispersal. Synthesized, these developments allow dispersal to be indirectly estimated from standing spatial patterns of genetic variation under a range of conditions broad enough to be likely met in most populations of either plants or animals. The spatial correlations at the shortest distances are particularly robust to range of conditions and have disproportionately high statistical power. This review integrates theoretical results in a way that maximizes robustness and flexibility in the use of short distance autocorrelation to estimate Wright's neighborhood size, or the total variance in dispersal distances. Empirical guidelines are developed that are meant to be as practical and broad as possible. The guidelines focus on Moran's *I*-statistics for diploid genotypes converted to allele frequencies, but are also extended to or compared with several other approaches.
Heredity (2005) **95,** 7–15. doi:10.1038/sj.hdy.6800680
Published online 4 May 2005

Keywords: dispersal; genetic structure; isolation-by-distance; Moran's *I* statistic; neighborhood size; spatial autocorrelation

In 1943, Sewall Wright established the fundamental relationship between the total amount of dispersal within a large continuous population and theoretical inbreeding coefficients relative to an initial population. Wright's approach was hierarchical, and he described the spatial structure in terms of how the inbreeding coefficient within a given block size (number of 'neighborhoods') changes as block size increases. He showed that under certain conditions, structure depended only on a function of the variance of dispersal distances for both sexes and the density. He termed this function the neighborhood size, and recognized that there was not likely to be either random mating or lack of structure within a neighborhood (Wright, 1946). Malécot (1948) reformulated theoretical models in terms of probabilities of identity by descent, and developed results largely in terms of pairs of genes each sampled randomly from each of two diploid individuals. Malécot's models had the advantages. In principle, there was no loss of spatial resolution, because even adjacent individuals could be specified, and there was no need to choose arbitrary block sizes. Recent spatial statistical literature has shown that there are no advantages to hierarchical spatial statistics unless a system is operating hierarchically (eg Hooper and Hewings, 1981), and few dispersal processes are even partially hierarchical, still fewer fully so.

In 1955, Malécot suggested an empirical estimator of spatial correlation of gene frequency (which is a function of the probability of identity by descent) at distance y, $r(y)$:

$$r(y) = \frac{\sum_x (q_x - q)(q_{x+y} - q)}{\sum_x (q_x - q)^2} \quad (1)$$

where q is the mean gene frequency. If the summation in the numerator is taken over all pairs of individuals in a distance class, then this estimator is the Moran (1950) I statistic, in the case where diploid genotypes are converted to allele frequencies. Here homozygotes for an allele are assigned the value 1.0, heterozygotes for that allele 0.5 and all other genotypes 0.0 (Heywood, 1991). This is a very common statistic in recent empirical studies. However, there remained a number of theoretical problems (eg the 'pain in the torus' problem of Felsenstein (1975)) that limited application of Malécot's theoretical models. Moreover, some of these problems are particularly acute for the shortest distances in populations existing in two spatial dimensions. Approximations for model theoretical ('*a priori*') values for probabilities of identity by descent for moderate distances can be obtained using Bessel functions. However, the *a priori* values generally cannot be observed in real populations, although surrogates can be close approximations in many cases (see Vekemans and Hardy, 2004). In addition, much of the predicted theoretic values require known, nonzero rates of mutation or some other 'recall coefficient,' and equilibrium is assumed. As will be discussed later, Moran's I statistic is interconvertible with the spatial covariance (Cockerham, 1969), also termed the 'conditional kinship' (Hardy and Vekemans, 1999), if the fixation index is specified. As this article has its focus on spatial structure *per se*, the methodology is laid out for Moran's I-statistic, which appears to depend only on dispersal and not directly on the rate of self-

Correspondence: BK Epperson, Department of Forestry, Michigan State University, 126 Natural Resources Building, East Lansing, MI 48824, USA. E-mail: Epperson@msu.edu
Received 4 November 2004; accepted 28 February 2005; published online 4 May 2005

fertilization (Epperson, 1990). Although recently some have claimed that conditional kinship bears a closer relationship than does Moran's *I* (eg Loiselle *et al*, 1995) to the theoretical models of Malécot and others, the above description of Equation (1) shows this to be untrue. Moran's *I* has the advantage in that analyses may be based on any of per-allele values, per-locus averages, or averages over all alleles and loci, unlike truly multilocus measures (eg Smouse and Peakall, 1999). Both per-allele (Cliff and Ord, 1981) and averaged (Epperson, 2004) Moran's *I*-statistics have known distributions under the null hypothesis that the spatial distribution of genotypes is random. Values of Moran's *I* under isolation by distance processes are very well characterized through a series of simulation studies.

While the mathematical theory of Malécot and others was crucial for laying the foundation for the theory of isolation by distance, another important approach uses Monte Carlo simulations that explicitly generate spatial distributions of genotypes. As will be discussed, the simulation-predicted values of spatial autocorrelation statistics essentially do not require the equilibrium assumption, and they can be directly compared to empirical data as well as loosely compared to mathematical models. The variances (possibly containing statistical as well as stochastic 'noise,' neither of which are available from the mathematical theory), and thus levels of uncertainty, of predicted values can also be obtained. In addition, populations with relatively high dispersal have little structure and low autocorrelations, hence errors of approximations (which would be involved if mathematical models were used) can become critical. Nonetheless, by averaging over alleles and loci, even low autocorrelations can be accurately measured using spatial autocorrelation statistics. For example, using about 40 alleles it was possible to detect a statistically significant difference in autocorrelations between two populations of eastern white pine (an outcrossing and wind-pollinated species), one an old growth site, the other logged, even though the values were very small, 0.022 and 0.004, respectively (Marquardt and Epperson, 2004). A long series of simulation studies have now been conducted, and in total they represent a massive number of simulations (eg Sokal and Wartenberg, 1983; Epperson, 1990, 1995, 2003a; Doligez *et al*, 1998).

This review focuses on the spatial autocorrelations between individuals separated by *short* distances, more specifically and throughout, Moran's *I* for converted genotypes and distance class one, which may contain pairs of individuals that are adjacent or otherwise near neighbors (either in the population or in the sample). There are several reasons for doing so. First, the use of the entire spatial structure (or as it is characterized by the *I*-correlogram) is complicated by the fact that estimates for different distances (distance classes) are not independent, but have correlations that depend on the spatial distribution itself. Thus, for example, the *Q* statistic of Oden (1984), which tests entire correlograms for statistical significance, is conservative. Remarkably, Oden (1984) found that tests of significance for the first distance class and the *Q*-test have about the same statistical power under the process of isolation by distance. The goal here is to outline robust methods for estimating dispersal, which methods usually require averaging over multiple genetic markers. By using values separately for each distance class, in our case the first distance class, it becomes possible to obtain averages that have known distributions. Here, a critical advance is the recent characterization of another form of correlations, those between *I*-statistics (for a given distance class) for different alleles of a locus (Epperson, 2004). Second, simulations show that values of correlations stabilize fastest for short distances, and values for all distance classes remain stable for very long periods. Thus, the number of generations for which a population has existed need not be known. It is only required that a population is say 10–30 or more generations old, in order that the short distance autocorrelation can be confidently used to estimate dispersal.

Standard relationship

Figure 1 illustrates the basic relationship between Moran's *I* and physical distance. Distance is on the scale where 1.0 is the distance between (rook's move) nearest neighbors on a lattice. Generally, the correlations drop smoothly with increasing distance and often become negative. In this example, distance class one contains all pairs of individuals separated by distances up to 1.5; thus, it contains both rook's move (4) and bishop's move (4) nearest neighbors on the lattice. All other distance classes D are such that $D-0.5<\text{distance}\leq D+0.5$.

The relationship between Moran's *I* for distance class one and dispersal is illustrated in Figure 2. The results (averages for one allele at a single locus for 100 space-time simulations of a population of 10 000 individuals – Epperson, 1990) are shown for a very wide range of dispersal models listed in Table 1. All results shown were for a locus with two equally frequent alleles, for sake of consistency and simplicity. Various other models, having differing arrays of alleles numbers and frequencies, generally give nearly identical results, as discussed elsewhere in this article. Models involve dispersal of males and females, which may have the same or differing amounts of dispersal. For females, N_f is the number of nearest neighbors of a location from which the maternal parent of an offspring at that location is randomly (with uniform probability) chosen, and N_m is

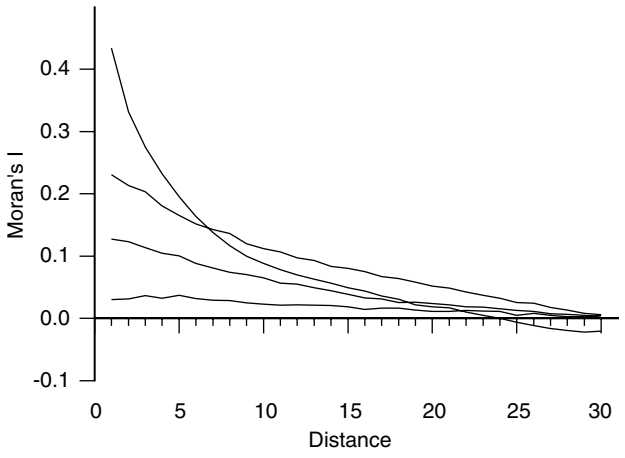

Figure 1 Decrease of Moran's *I* with distance, typified by four examples. Curves of 100 simulation averages are from top to bottom: set 1 ($N_e=4.2$); set 4 ($N_e=25.1$); set 6 ($N_e=50.2$); and set 9 ($N_e=115.2$), respectively.

the analogous parameter for males. In some simulations, females self-fertilize ($N_f = 1$) with rate $1/N_m$. From N_m and N_f sex-specific axial variances of dispersal distances can be calculated, and Wright's neighborhood size can be calculated from these variances (Table 1). In plant species, dispersal is through pollen and seed, but we can use Crawford's (1984) formula $\sigma_t^2 = \sigma_s^2 + \sigma_p^2/2$ together with Wright's (1943) formula $N_a = 4\sigma_t^2$, where σ_p^2 and σ_s^2 are the corresponding axial variances of dispersal distance and N_a is the neighborhood area. Neighborhood size $N_e = N_a$ because the density is 1.0. Clearly, Moran's I for the first distance class (nearest 8 neighbors, including both 'rooks' and 'bishops' on the lattice) decreases smoothly with N_e; indeed, there is a near one-to-one transformation. Disparate dispersal and mating system models that have similar values of N_e produce similar Moran's I statistics. For example, one model with $N_f = N_m = 121$ ($N_e = 125.7$) (which could represent a population of an animal species having both sexes dispersing fairly long distances), and another with $N_f = 1$, $N_m = 225$ ($N_e = 115.2$) (which could represent in effect a wind pollinated plant with near zero seed dispersal) have I values of 0.05 and 0.04, respectively. Such results fit with Malécot's (1948) finding that spatial structure depends primarily on the variance of dispersal distance, and not much on the shape of the dispersal function on distance.

The standard deviations of values among the 100 simulations for distance class one, for each set are also shown in Table 1. They are generally very small, in the range of 0.02–0.04, for low to moderate dispersal levels, and similar in value to those for other distance classes (results not shown). Estimated standard errors on the set means are 10 times $(\sqrt{100})$ smaller, indicating that the predicted values for each set are extremely precise.

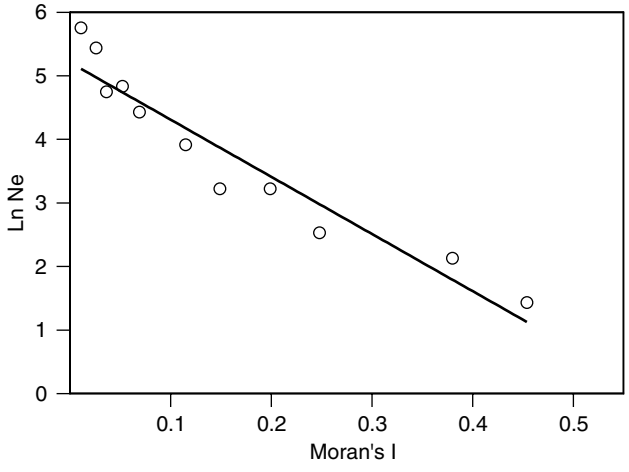

Figure 2 Natural logarithm of N_e plotted against Moran's I for distance class one. The line was fitted by least-squares, omitting set 12 ($N_e = 632.4$).

It appears that the relationship of finest scale spatial autocorrelation to dispersal is essentially exponential, since the plot of $\ln N_e$ on I is nearly linear. The line fitted using least squares ($R^2 = 0.92$) in Figure 2 omits the model with $N_e = 632.4$, but when that model is included the slope (−0.091) and intercept (0.509) of the fitted line ($R^2 = 0.88$) are scarcely changed. The mean value of I for the model with $N_e = 632.4$ was very small (0.008) and somewhat of an outlier on the logarithm scale. Stochastic variation relative to the predicted value should be great for models with extremely high dispersal (eg Table 1). Linear regression using $\ln N_e$ as the independent variable (with the $N_e = 632.4$ model omitted) yielded a slope of −0.102 and an intercept of 0.544, that is, $I = 0.544 - 0.102 \ln N_e$. Thus a good point estimate of N_e, appropriate across a very wide range of dispersal, is

$$\hat{N}_e = \exp[(0.544 - \hat{I})/0.102] \quad (2)$$

It should be recognized that as dispersal becomes very large, this estimate should generally become less precise and possibly less accurate. At the other extreme, this formula suggests that the maximum amount of spatial autocorrelation that can be created by isolation by distance is approximately 0.55. However, it should be noted that the models may not fully address all possible roles of self-fertilization that may be associated with extremely small neighborhood sizes, that is, approaching 1.0. Another method for estimation is to use tables of numeric values in Epperson *et al* (1999) and interpolate between values. A recent review of values of I-statistics for distance class one observed in empirical studies showed similar correspondence with the ranges of neighborhood sizes that were either estimated or could be presumed from dispersal biology (Table 7.1 of Epperson, 2003a).

Factors for consideration of application to an experimental system

In considering whether or not this indirect estimation method should be applied to a specific system, we will assume that the population is at least fairly large and somewhat continuous. The most important other considerations fall into two categories. The first is how many generations the population has existed. The second is a

Table 1 Dispersal parameters in the various sets of simulations, set means (I) and standard deviations (SD)

	Simulation dispersal model[a]											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
N_f	1	9	1	25	1	49	81	121	1	225	1	625
N_m	9	9	25	25	49	49	81	121	225	225	625	625
N_e	4.2	8.4	12.6	25.1	25.1	50.2	83.7	125.7	115.2	230.4	316.2	632.4
I	0.454	0.380	0.248	0.199	0.149	0.115	0.069	0.052	0.036	0.026	0.010	0.008
SD	0.024	0.032	0.034	0.037	0.026	0.033	0.023	0.024	0.015	0.012	0.008	0.005

[a]N_f and N_m are the numbers of nearest female and male individuals from which parents of an offspring are randomly chosen, and N_e is Wright's neighborhood size. Adapted from Epperson and Li (1997) and Epperson *et al* (1999).

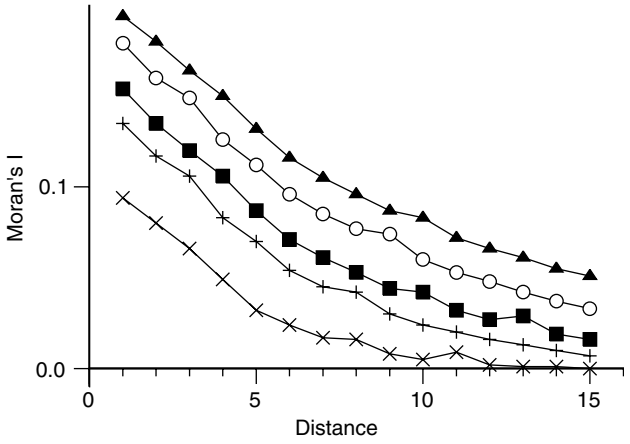

Figure 3 Increase of Moran's I with time in generations, starting from a spatially random initial population: (×) 10; (+) 30; (■) 50; (○) 100; and (▲) 200. Curves show 100 simulation averages of set four ($N_e = 25.1$).

set of demographic factors associated with population growth rates and spatial or temporal variation in density. Again, the goal here is not to review all demographic intricacies, rather to define a broad range of conditions under which the relationship expressed in Equation (2) is not substantially altered.

It has been shown repeatedly that isolation by distance processes in large but finite populations produce spatial genetic correlations that become 'quasi-stationary' (Sokal and Wartenberg, 1983). This quasi-stationarity obtains within 50–100 generations and persists for very long periods (eg Sokal and Wartenberg, 1983; Epperson, 1990). Figure 3 illustrates the temporal changes in spatial autocorrelations, when beginning from a random distribution. The trends are essentially the same for all dispersal models (eg Epperson, 1990), and in the case shown the plotted values are averages for 100 simulations of a single locus with five alleles in a population having moderate dispersal ($N_f = N_m = 25$, $N_e = 25.1$). By about 10–20 generations, the increases in spatial autocorrelation over time have markedly slowed. Thereafter, the percentage increases are much lower for short distance classes than for relatively long distances. Hence, the percentage error that would occur relative to values under quasi-stationarity are much larger for large distance classes. Moreover, the statistical errors in estimating (from empirical data) I-statistics for large distance classes under many circumstances would be very large relative to the predicted values. In contrast, for using distance class one only, the percentage error may be tolerable, if the population is more than 30–50 generations. For example, if a population like the model above was only 30 generations old, the predicted value (0.136) corresponds to a neighborhood size of 54.6, using Equation (2). Given that N_e is a function of squared distances (hence estimates of N_e based on *direct* measures of dispersal distance are likewise liable), the fact that the estimate is about twice that of the quasi-stationarity predicted value (25.1) is not surprising. Other courses of action are: (1) if the age of the population is known, then more refined (non-quasi-stationary) values could be used; or (2) if the age is not known, which is more often the case, then use the quasi-stationary value and consider the estimate of N_e to be an upper bound.

An example of effects of young population age was observed in contrasts between populations of sumac, *Rhus tricocarpha* and *R. javanica*, both woody perennials, dioecious, and having substantial frequencies of clonal reproduction. Despite the fact that dispersal is quite limited, owing to insect pollination and seed dispersal by gravity and birds, the autocorrelations are small. In the two study populations of *R. tricocarpha*, the younger population (12 years old) had a much lower value of I (0.02) than that (0.17) for the other population (26 years old) (Chung *et al*, 1999). The same pattern was observed among nonclonal individuals in the two study populations of *R. javanica*, one apparently younger ($I = -0.02$) than the other ($I = 0.08$) (Chung *et al*, 2000). For example, for *R. javanica*, if Equation (2) were used, N_e would be estimated to be effectively infinite for the younger population. For both younger populations of *Rhus* spp., the inferred neighborhood sizes appear to be too high, given the dispersal mechanisms.

The lattice models used to generate the predicted values assume in effect that density is spatially uniform and does not change over time. A recent rapid increase in density could alter the relationship between spatial autocorrelation and dispersal. Such appears to be the case in populations of the perennial *Silene dioica* in their successional stages on uplifted emerging islands off the coast of Sweden (Ingvarsson and Giles, 1999; discussed in Epperson, 2003a). On one island studied in detail, there were three distinct subpopulations. Spatial genetic structure steadily increases for the older and denser subpopulations. The amount of autocorrelation was low in the young subpopulation compared to what would be predicted based on simulations and the fact that *S. dioica* is pollinated by bumblebees and has limited seed dispersal. Moreover, limited seed dispersal resulted in demographic patches that were essentially matrilineal and produced a nonstandard form of spatial genetic structure. In this case, join count statistics (Cliff and Ord, 1981) provided a more flexible and somewhat more powerful method for characterizing structure.

Relatively little is known about what happens to spatial autocorrelation when a population rapidly decreases in density. However, in some respects this may be similar to what occurs across life stages in populations that have strong intraspecific competition (although it should be noted that N_e would also be decreasing in many cases). It has been observed in many tree species that relatively large amounts of autocorrelation are found in seedlings, less in juveniles and still less in reproductive adults and older age classes (Epperson and Alvarez-Buylla, 1997; Chung *et al*, 2003).

Spatial variations in density also need to be considered. Epperson and Li (1997) examined a very mild form of nonuniform density, and found there to be no measurable effect. Clumping is a complex issue, possibly involving many factors, to name a few inter- and intraspecific competition, spatial distributions of appropriate habitat, and limited seed dispersal in plants and territoriality in animals. Doligez *et al* (1998) found that clumping increased Moran's I. However, the form of clumping was rather extreme, the model allowed no intraspecific competition, and the effect on I was still

fairly mild. In most continuous populations, clumping should not be a major problem in using Equation (2). Similar analyses by Lee and Hastings (personal communication) have also found little effect of clumping. Recently, Barton *et al* (2002) examined the effects of clumping on identity in state for a general density-dependent survival model (of Bolker and Pacala, 1997) that uses two parameters (ρ and α) to represent the effects of local density on mortality. They gave an example for specific values of ρ and α, and found virtually no effect on spatial genetic structure. This approach could be used to determine more precisely the effects of clumping. However, it appears that violations of the uniform density of simulations usually are not serious enough to cause difficulties.

Choice of markers

There appear to be five major factors that may drive choice of genetic markers: (1) neutrality; (2) numbers of alleles; (3) allele frequencies; (4) how alleles are 'binned' and the extent to which scored identity in state corresponds to identity by descent; and (5) possible direct effects of mutations. Markers must be neutral because even nonspatial selection can markedly alter spatial autocorrelations (Epperson, 1990). Simulations have shown that the number of alleles does not change the predicted values of spatial autocorrelation, except to the degree that high numbers may force very low allele frequencies (Epperson *et al*, 1999). Markers with greater numbers of alleles generally are preferred, because the spatial autocorrelations of each allele are nearly independent (see below), hence greater statistical power is achieved for averaged values. Allele frequency does not substantially change Moran's I in the large simulation samples (10 000 individuals) (Epperson *et al*, 1999), unless it is reduced to about 0.02 – then the I-statistic is reduced by about 15% (Epperson, 2003a). However, there may be additional issues for smaller samples (Epperson *et al*, 1999).

In typical-sized samples, say of one to a few hundred diploid individuals, a low frequency allele could be present only once. If an allele is present in only one individual (in one or possibly two copies, for example, if the rate of selfing is high), it is a foregone conclusion that the I-statistic will be negative, and possibly large. There are two opposing ways of viewing such an occurrence. One is that it is meaningful – the fact that no other copies are found in the vicinity should tilt the conclusion toward spatial structure being weak. The other is that the finding is meaningless with respect to spatial structure, and it could be that the allele is the result of a recent but else-rare, long-distance migration event or perhaps even a recent mutation. To err on the side of caution and avoid negative bias, such alleles should generally be omitted from analyses. When an allele occurs in more than one individual, there are no hard and fast rules. For example, if an allele occurs in two individuals, and those two are neighbors, it probably is meaningful. On the other hand, if dispersal is fairly high, and not highly restricted, it is unlikely that the two would be adjacent. Such an *a priori* view can introduce a positive bias, but it is probably very minor. Various criteria have been used, but the following guideline can be broadly applied: if dispersal is believed to be very low, then use all alleles that are carried in more than one individual; if not then omit any allele that is present in less than five copies in the sample. If the allele frequency is less than 2%, then some adjustment should be made.

Regarding the possibility that scored identity in state does not correspond to identity by descent, which is what builds spatial structure (Malécot 1948; Barton and Wilson 1995), we may view this discrepancy as adding nearly spatially random or 'white' noise to the spatial signal of isolation by distance. The discrepancy should be greatest for allozymes, although there has not been empirical confirmation of this. Binning may be similar in effect to backmutation, in that it assigns identity in 'state' (bin) to genes that are not identical by descent, and therefore probably has little effect. The short time scale over which mating-by-proximity-caused coalescences (see below) occur is relevant to this issue, as well as critical to possible direct effects of mutations.

Recently, it has been found that mutation at very high rates can substantially reduce spatial autocorrelations (Epperson, 2005). Mutation has virtually no effect at normal rates, but when the rate is 10^{-3}, spatial autocorrelations at the smallest distances are decreased by an average of seven percent. The critical range, of 10^{-2} down to 10^{-3}, is spanned by microsatellite loci in animals (eg Bruford *et al*, 1992; Jarne and Lagoda, 1996) and plants (Udupa and Baum, 2001; Thuillet *et al*, 2002; Vigouroux *et al*, 2002), as well as some other hypervariable markers. For loci with more than a few alleles, mutation at a rate of 10^{-2} causes 30–40% reductions in values for the finest scale autocorrelation. The percent reductions vary little over wide ranges of dispersal level, numbers of alleles and mutation processes (Epperson, 2005). The reductions are primarily caused by the fact that (forward) mutation removes identity by descent between spatially proximal, genealogically related genes, in particular between genes having recent coalescences, which are the ones that are responsible for most of the spatial autocorrelation (Barton and Wilson, 1995).

Backmutation has very little effect on spatial autocorrelations (Epperson, 2005). First, for most mutation processes, the frequency of backmutations to a given allele is usually considerably lower than the forward mutation rate. However, even when the backmutation rate is relatively high, its effect is small and usually negligible. The reasons for the lack of effect center on the fact that any allele, spatially proximal or not, can backmutate with near-equal likelihood. In essence, the resulting identity in state (and contribution to spatial correlation) without identity by descent is nearly spatially random. As spatial autocorrelations are spatial correlations normalized by denominator of Equation (1), theoretically they could be affected. However, as noted, the effect is very small. It should be pointed out that the fundamental dependence of autocorrelation on recent coalescences and forward mutations means that the specifics of the mutation process are irrelevant. This is fortunate, because microsatellites can have fairly high ratios of backward/forward mutations, for example under the strictly stepwise mutation model (SMM). Thus, the reductions caused by forward mutation can be confidently applied to nearly any locus, including microsatellites.

There is some empirical evidence for mutation-reduced spatial autocorrelation. In the previously men-

tioned old growth population of eastern white pine, one, highly variable, microsatellite, Rps50 (14 alleles), had a much lower *I*-statistic (−0.010) for distance class one than the other loci (average $I = 0.025$, average number of alleles = 6.7), and the difference was statistically significant (Marquardt and Epperson, 2004). Although the mutation rates of these loci have not been studied, it is possible to estimate relative rates by using the well-known relationship (eg Ewens, 2004) between the effective number of alleles and the product of mutation rate and population size (both are unknown but the latter should be uniform over loci). By this method, it was estimated that Rps50 has a mutation rate that is approximately six- or seven-fold greater than the average for other loci. The same differences in spatial structure and estimated relative mutation rates between Rps50 and other loci was also observed in the white pine seedlings found in the same forest (Walter and Epperson, 2004).

This is an important lesson, particularly because the general and otherwise supported tendency is to choose microsatellites (or other markers) that have the highest numbers of alleles. While as noted earlier, a larger number of alleles generally increases statistical power per locus (or per genotyping-effort), numbers of alleles are sometimes so large that it seems likely that mutation is occurring at rates near 10^{-2}. For example, albeit at a larger spatial scale, Morand *et al* (2002) employed loci having as many as 56 alleles. The demonstrated effects of high mutation rates suggest that some caution is in order. Choice of most variable markers could affect the results. With regard to estimating dispersal, there are two remedies, either avoid such highly variable markers or measure the mutation rate directly and adjust the resulting values of *I* statistics.

Sample design

In most cases, the best design for sampling is simply to size an area such that it contains the desired number of individuals, or select a starting point and sample contiguous individuals radiating from it, using GPS or other methods to record locations. In such analyses, assuming that density is at least fairly uniform, the *I* value for the first distance class (with an upper bound configured along the guidelines discussed below) will fit those generated in the simulations (and exemplified in Equation (2)). However, there are cases where this method is inappropriate. If prior to study, it is believed that dispersal may be very limited, say with an N_e of 15 or less, for example, *Ipomoea purpurea* with its gravity dispersed seed and bumblebee pollination (Epperson and Clegg, 1986), the spatial distribution is constituted mainly by large patches (of several hundred individuals) of homozygous genotypes, with heterozygotes occurring primarily along patch boundaries (eg Epperson, 1990). If the sample size is usual, one to a few hundred, then for some loci the sample area may be fully contained in one patch (giving almost fixation and probably little autocorrelation), whereas for others it might be largely constituted by a patch boundary area (which might even create a cline; Epperson 2003a). Thus, high statistical variation among loci might result. An alternative in such cases is to sample at a larger spatial scale, that is, to design a sample that aims to collect only one out of every *X* individuals. If this route is chosen, then in most situations sampling on a lattice is best. A grid can be laid out and the individual closest to each intersection chosen (if the distance is measured, then density can also be estimated (Pielou, 1977)). Clearly, however, as exemplified in Figure 1, as *X* increases the spatial scale of sampling increases, and the amount of autocorrelation for the first distance class is decreased. The effects of the value of *X*, which has been termed 'porosity,' was studied in detail by Epperson *et al* (1999), and an approximate adjustment can be made for the relationship of *I* to N_e by spatial re-scaling in accordance with theoretical *I*-correlograms (using graphs like Figure 1).

The number of individuals, *n*, to be assayed generally should be in the range of one hundred to a few hundred (see Table 7.1 of Epperson, 2003a). In the experience of the author and collaborators, a sample size of 50 may be borderline. Naturally, the number of individuals could also be larger than a few hundred, and an appropriate choice also generally depends on the number of genetic traits assayed. The number of genetic traits can be discussed in terms of number of loci and alleles per locus. Given the near independence of *I*-statistics for different alleles (discussed below) of a locus, at least if there are more than a few alleles, it is the total number of alleles, *K*, rather than number of loci, *m*, that is critical. A perhaps somewhat finer guide would be to use $K-m$. Theoretical results indicate that the product of *K* and *n* usually dominates issues of statistical power and precision. To some degree, the power for tests of significance will depend on the amount of structure, generally being greater for populations with larger amounts. However, it appears that values of *Kn* in the range of two to several thousand are adequate for most purposes (Epperson and Li, 1996), especially where multilocus averages are used. For example, if $n = 100$ and $K = 20$, the sampling intensity is probably adequate. Such sizes are generally feasible.

Although not a sampling issue *per se*, another important consideration for analyses is the choice of the upper bound on the first distance class. Many of the software packages that calculate *I*-statistics have two standard settings. One makes the number of pairs equal across distance classes, and the other makes the range of distances uniform. Both have some desirable properties (Epperson, 2003a), but they are arbitrary with respect to a focus on distance class one. Experience suggests that the first (or any other) distance class should have at least a few hundred pairs of individuals. Most studies use 10 or so distance classes. For very large sample sizes, *n*, both standard settings will generally place far more pairs than needed into distance class one, which will decrease spatial resolution and make necessary spatial rescaling for estimating dispersal. Pairs of individuals that are separated by the shortest distances, and those predicted to have the highest correlations, will be averaged with pairs separated by considerably larger distances. An alternative that seems to work well, as long as *n* is ca 100 or greater, is based on the density of sampled individuals, *d* (m^{-2}). The inverse of *d* is a measure of the average amount of space taken up by an individual, and the square root of the inverse, $d^{-1/2}$, would be the distance between (rook's move) nearest neighbors if the individuals were located on a lattice, and analogously $\sqrt{2}d^{-1/2}$ would be the distance between 'bishop's move' nearest neighbors (Epperson, 2003a). It appears that the

latter value captures the majority of nearest neighbors in nonlattice samples, so long as the density is fairly uniform, and it will also include other near neighbors, but it in a sense maximizes retention of spatial resolution as well as provides sufficient numbers of pairs in distance class one. I statistics for distance class one, under the foregoing conditions, can be directly compared to theoretical values (without rescaling), that is, Equation 2 can be used directly.

Obtaining multiple-allele, multilocus averages and their variances

Conditions and methods for accounting for the correlations between I statistics for alleles of a locus, in order to obtain averages across alleles and then across loci, turn out to be deceptively simple. First, a fundamental problem had to be solved on how different pairs of types (in our case diploid genotypes) are correlated, in spatial distributions where one and only one type is mapped to each location (Epperson, 2003b). Then, it is recognized that Moran's I-statistic for diploid genotypes converted to allele frequencies is a function of the numbers of types of pairs of genotypes (Epperson, 1995). Thus, the covariance (and correlation) between the I-statistics for two alleles is a complex function of the counts of pairs of types and the products between these counts (Epperson, 2004). The expressions are huge, so much so that it is not feasible to write them down. For example, for a locus with five alleles, there are 15 possible genotypes, 120 types of pairs, and 14 400 products of pairs, that is 14 400 terms. However, calculation algorithms can be constructed for computer use.

The correlations were first found mathematically under the null hypothesis of a random distribution, as needed for tests of significance of averages. However and remarkably, isolation by distance does not substantially change the correlations, as was shown in simulations (Epperson, 2004). Hence, they can be used for constructing correct estimated standard errors for averages, apart from those used in significance tests. The main result is that the correlations can be large, but only between pairs of alleles where at least one of the alleles is common, that is, has a large frequency in the population.

The expressions for the correlations under the null hypothesis of sampling pairs without replacement involve not only many algebraic terms, but these terms contain such expressions as $n(n-1)(n-2)$. Hence, the finding discussed below is all the more surprising. First, it is necessary to give some definitions. Let I_i be Moran's I for genotypes converted to frequencies of allele A_i, for distance class one, and let $\mathrm{Cov}(I_i,I_j)$ be the covariance between I_i and I_j. Let I be the unweighted average of I_i over all k alleles. Then the variance of I is given by (eg Feller 1957)

$$\sigma^2(I) = \left(\frac{1}{k}\right)^2 \left[\sum_{i=1}^{k} \sigma^2(I_i) + 2\sum_{i,j} \mathrm{Cov}(I_i, I_j)\right] \qquad (3)$$

where the second summation is over all alleles i and j such that $i<j$, and $\sigma^2(I_i)$ is the variance of I_i. If the values of I_i were independent, then all covariances are zero and the second term is zero.

There are several possible courses of action. One course would be to compare the allele frequencies in a sample to those listed in Epperson (2004), find the closest cases to approximate the correlations, use observed variances to determine the covariances, and then insert those into Equation (3). It turns out there is a much simpler alternative.

The cases of three, four, and five alleles, wherein all alleles had equal frequencies (ie 0.33, 0.25, and 0.20) may be examined first. In these cases, when the ratio of the variance of the average ignoring the covariances and the true variance using Equation (3) was calculated, they took the precise (to the many digits calculated) values 0.66, 0.75 and 0.80, respectively. It should be noted the variance is underestimated by 20–33%, a substantial error. More interesting is the coincidence that the variances would be correct if the sum of variances were divided by $k(k-1)$ rather than by k^2. This result is astoundingly simple, given the complexity of the expressions for correlations among I statistics for different alleles. It is worth noting that, while the constraint that all alleles have equal frequencies makes them essentially interchangeable, this result is not simply a matter of degrees of freedom, although the spatial distributions of $k-1$ alleles do determine that of the remaining allele. Under a χ^2 degrees of freedom argument (and noting that each I_i is asymptotically normal (Cliff and Ord, 1981)), the variance of an average is obtained by dividing the sum of $\sigma^2(I_i)$ by $(k-1)^2$, whereas in this case the correct variance is obtained by dividing by $k(k-1)$. While examination of a few cases does not prove a generality, and algebraic proof is prohibited by the huge numbers of terms involved, one may conjecture that this finding, which can also be characterized by the (double) sum of the correlations equaling $k/(k-1)$, involves a 'universal constant' that is characteristic of spatial patterns.

In addition, several other arrays of alleles, where frequencies were not uniform, were examined (Epperson, 2004). In every case, the ratio of biased to true variance was very close to but slightly smaller than that for the corresponding even-frequency cases, indicating that the error is slightly increased. However, since the differences are small, little bias would be created if instead the sum of variances were divided by $k(k-1)$ rather than k^2, as in the equal-frequency cases. In all cases where the number of alleles is small to moderate, the variance will be seriously underestimated by ignoring the correlations, unless a correction is made. Some markers may have as many as 25 or even 50 alleles, and here the bias would be negligible.

Estimated standard errors of single allele I-statistics are calculated and output by spatial statistics software packages such as PASSAGE (Rosenberg, 2002). Standard errors are usually quite uniform across alleles, in the experience of the author and collaborators. Once the single locus averages (and their standard errors) are obtained as outlined in this section, averages (and their standard errors) across loci can be calculated, in a straightforward manner, as long as the values for different loci can be treated as independent, which will be the case unless there is linkage disequilibrium.

Other measures

As was noted in the introduction, Moran's I is a genetic correlation, first suggested by Malécot (1955). Like the spatial covariance (Cockerham, 1969), R_i (of an allele A_i) also termed 'conditional kinship' (Hardy and Vekemans, 1999), Moran's I does not necessarily correspond to the kinship coefficients in the mathematical models of Malécot and others, although it can be a close approximation. Those models were developed in terms of *a priori* expectations (ie expected values given only the initial population and the dispersal and other parameters) for probabilities of identity by descent. Importantly, Moran's I does not depend on the fixation index, which is influenced by all forms of inbreeding, beyond just biparental inbreeding caused by spatial proximity. Existing simulations have found that at least low rates of selfing affect Moran's I only to the extent that male gamete dispersal distances are reduced, that is, through reductions in the neighborhood size. However, if the allele-specific maximum likelihood estimator of the fixation index (Brown, 1970), F_i, is known, then $R_i = I_i\ (1+F_i)/2$. The average (R) of values of R_i over alleles of a locus is a weighted average of the I statistics, with weights $(1+F_i)/2$. While the expected values of allele-specific inbreeding coefficients generally should be uniform, observed allele-specific fixation indices are not likely to be. However, it is worth noting that if they were, then $R = I\ (1+F)/2$, and the variance of R should be calculated by dividing the sum of variances of R_i by $k(k-1)$, as in the case of I. If not, then the correlations for pairs of alleles calculated for Moran's I (Epperson, 2004), together with an equation analogous to the Equation (3), but allowing for weights (Feller, 1957), can be used to find the variance of the average, R. Again, per-locus averages of R can be further averaged over loci, under the same conditions as for Moran's I. Thus, the distributions can be found for all estimates of spatial covariance that first find allele specific values and then average over alleles and loci. However, other multilocus methods, for example that of Smouse and Peakall (1999), which sum over alleles and loci first and then over pairs of individuals, cannot be obtained using the present method, unless there is no linkage disequilibrium.

Recently, Vekemans and Hardy (2004) developed an estimator based of the slope of the R-correlogram on the natural logarithm of distance. In principle, this estimator could have an advantage in that it uses more of the spatial structure than does Moran's I for distance class one. However, as was noted earlier, what little evidence there is suggests that the latter statistic has approximately the same statistical power as ones based on entire correlograms (Oden, 1984). The method based on slope may have some disadvantages, including the fact that it assumes the decrease with distance is exponential. In addition, use of only short distances is more robust to population age, and may also be more robust to differences in other demographic factors. Statistical noise can be substantial, especially for larger distance classes, and should be even greater for differences between distance classes (required for measuring slope). To date, very little is known about the statistical properties of slope-based measures, in contrast to Moran's I for distance class one.

Summary

The use of Moran's I-statistics for shortest distances to estimate dispersal is robust under a wide range of conditions. Normally, this will require averaging over alleles and loci. Either tables of values or an equation can be used. The method is valid as long as the population is fairly continuous, the population density is not changing rapidly, and the population has existed for more than 30 or so generations. In general, any set of neutral markers can be used, and low-frequency alleles can be dealt with. The total number of alleles (summed across m loci), K, times the number of sampled individuals, n, largely determines the statistical power of averages of I-statistics across alleles and loci. Adequate statistical power usually should be achieved when n exceeds 100 and nK exceeds a few thousand. Values for different alleles of a locus are mostly near-independent, so that loci with larger numbers of alleles provide greater efficiency per genotyping effort. However, some loci have so many alleles that their mutation rates may be near 10^{-2}, and this can cause substantial reductions in autocorrelations. Sample design is often optimized by the sampling of contiguous individuals, although there are exceptions. For normal sample sizes, the upper bound for the first distance class should be small yet contain nearest neighbor pairs and sufficient numbers of pairs. Simple methods are available for accounting for the correlations of I-statistics for different alleles of a locus, in order to estimate the variances of averages of I-statistics across alleles and loci. These methods can be extended to measures of spatial covariance or conditional kinship.

References

Barton NH, Depaulis F, Etheridge AM (2002). Neutral evolution in spatially continuous populations. *Theor Popul Biol* **61**: 31–48.

Barton NH, Wilson I (1995). Genealogies and geography. *Philos Trans Roy Soc London B, Biol Sci* **349**: 49–59.

Bolker B, Pacala SW (1997). Using moment equations to understand stochastically driven spatial pattern formation in ecological systems. *Theor Popul Biol* **52**: 179–197.

Brown AHD (1970). The estimation of Wright's fixation index from genotypic frequencies. *Genetica* **41**: 399–406.

Bruford MW, Hanotte O, Brookfield JFY, Burke T (1992). Multi- and single-locus fingerprinting. In: Hoelzel AR (ed) *Molecular Analysis of Populations: a Practical Approach*. IRL Press: Oxford, UK pp 225–269.

Chung JM, Chung MG, Epperson BK (1999). Spatial genetic structure of allozyme polymorphisms within populations of *Rhus trichocarpha* (Anacardiaceae). *Silvae Genet* **48**: 223–227.

Chung MG, Chung JM, Chung MY, Epperson BK (2000). Spatial distribution of allozyme polymorphisms following clonal and sexual reproduction in populations of *Rhus javanica* (Anacardiaceae). *Heredity* **84**: 178–185.

Chung MY, Epperson BK, Chung MG (2003). Genetic structure of age classes in *Camellia japonica* (Theaceae). *Evolution* **57**: 62–73.

Cliff AD, Ord JK (1981). *Spatial Processes*. Pion: London.

Cockerham CC (1969). Variance of gene frequencies. *Evolution* **73**: 72–84.

Crawford TJ (1984). The estimation of neighborhood parameters for plant populations. *Heredity* **52**: 273–283.

Doligez A, Baril C, Joly HI (1998). Fine-scale spatial genetic structure with nonuniform distribution of individuals. *Genetics* **148**: 905–919.

Epperson BK (1990). Spatial autocorrelation of genotypes under directional selection. *Genetics* **124**: 757–771.
Epperson BK (1995). Fine-scale spatial structure: correlations for individual genotypes differ from those for local gene frequencies. *Evolution* **49**: 1022–1026.
Epperson BK (2003a). *Geographical Genetics*. Princeton University Press: Princeton, NJ.
Epperson BK (2003b). Covariances among join-count spatial autocorrelation measures. *Theor Popul Biol* **64**: 81–87.
Epperson BK (2004). Multilocus estimation of genetic structure within populations. *Theor Popul Biol* **65**: 227–237.
Epperson BK (2005). Mutation at high rates reduces spatial structure within populations. *Mol Ecol* **14**: 703–710.
Epperson BK, Alvarez-Buylla E (1997). Limited seed dispersal and genetic structure in life stages of *Cecropia obtusifolia*. *Evolution* **51**: 275–282.
Epperson BK, Clegg MT (1986). Spatial autocorrelation analysis of flower color polymorphisms within substructured populations of morning glory (*Ipomoea purpurea*). *Am Nat* **128**: 840–858.
Epperson BK, Huang Z, Li T-Q (1999). Spatial genetic structure of multiallelic loci. *Genet Res* **73**: 251–261.
Epperson BK, Li T-Q (1996). Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proc Natl Acad Sci USA* **93**: 10528–10532.
Epperson BK, Li T-Q (1997). Gene dispersal and spatial genetic structure. *Evolution* **51**: 672–681.
Ewens WJ (2004). *Mathematical Population Genetics, Vol. I. Theoretical Introduction*. Springer-Verlag: Berlin.
Feller W (1957). *An Introduction to Probability Theory and Its Applications*. Wiley: New York.
Felsenstein J (1975). A pain in the torus: some difficulties with models of isolation by distance. *Am Nat* **109**: 359–368.
Hardy OJ, Vekemans X (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83**: 145–154.
Heywood JS (1991). Spatial analysis of genetic variation in plant populations. *Annu Rev Ecol System* **22**: 335–355.
Hooper PM, Hewings GJD (1981). Some properties of space-time processes. *Geogr Anal* **13**: 203–223.
Ingvarsson PK, Giles BE (1999). Kin-structured colonization and small-scale genetic differentiation in *Silene dioica*. *Evolution* **53**: 605–611.
Jarne P, Lagoda PJL (1996). Microsatellites, from molecules to populations and back. *Trends Ecol Evol* **11**: 424–429.
Loiselle BA, Sork VL, Nason J, Graham C (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Bot* **82**: 1553–1564.
Malécot G (1948). *Les Mathématiques de l'Hérédité*. Masson: Paris.
Malécot G (1955). Remarks on the decrease of relationship with distance. Following paper by M Kimura. *Cold Spring Harb Symp Quant Biol* **20**: 52–53.
Marquardt PE, Epperson BK (2004). Spatial and population genetic structure of microsatellites in white pine. *Mol Ecol* **13**: 3305–3315.
Moran PAP (1950). Notes on continuous stochastic phenomena. *Biometrika* **37**: 17–23.
Morand ME, Brachet S, Rossignol P, Dufour J, Frascaria-Lacoste N (2002). A generalized heterozygote deficiency assessed with microsatellites in French common ash populations. *Mol Ecol* **11**: 377–385.
Oden NL (1984). Assessing the significance of a spatial correlogram. *Geogr Anal* **16**: 1–16.
Pielou EC (1977). *Mathematical Ecology*, 2nd edn. Wiley: New York.
Rosenberg MS (2002). *PASSAGE. Pattern Analysis, Spatial Statistics, and Geographic Exegesis. Version 1.0. Department of Biology*. Arizona State University: Tempe, AZ.
Smouse PE, Peakall R (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82**: 561–573.
Sokal RR, Wartenberg DE (1983). A test of spatial autocorrelation analysis using an isolation-by-distance model. *Genetics* **105**: 219–237.
Thuillet AC, Bru D, David JL, Roumet P, Santoni S, Sourdille P *et al* (2002). Direct estimation of mutation rate for ten microsatellite loci in durum wheat, *Triticum turgidum* (L.) Thell. ssp *durum* desf. *Mol Biol Evol* **19**: 122–125.
Udupa SM, Baum M (2001). High mutation rate and mutational bias at (TAA) (n) microsatellite loci in chickpea (*Cicer arietinum* L.). *Mol Genet Genomics* **265**: 1097–1103.
Vekemans X, Hardy OJ (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol Ecol* **13**: 921–935.
Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Smith JS *et al* (2002). Rate and pattern of mutation at microsatellite loci in maize. *Mol Biol Evol* **19**: 1251–1260.
Walter R, Epperson BK (2004). Microsatellite analysis of spatial structure among seedlings in populations of *Pinus strobus* (Pinaceae). *Am J Bot* **91**: 549–557.
Wright S (1943). Isolation by distance. *Genetics* **28**: 114–138.
Wright S (1946). Isolation by distance under diverse systems of mating. *Genetics* **31**: 39–59.