Genet. Res., Camb. (1999), **73**, *pp.* 251–261. *With* 2 *figures.* Printed in the United Kingdom © 1999 Cambridge University Press

Measures of spatial structure in samples of genotypes for multiallelic loci

B. K. EPPERSON*, Z. HUANG AND T.-Q. LI
Department of Forestry, Michigan State University, East Lansing, MI 48824, USA

(*Received 29 December 1997 and in revised form 15 October 1998 and 7 January 1999*)

Summary

Various spatial autocorrelation statistics have been widely used both in theoretical population genetics and to study the spatial distribution of diploid genotypes in many plant and animal populations. However, previous simulation studies have considered only diallelic loci. In this paper, we use a large number of space–time simulations to characterize for the first time the parametric and statistical values of Moran's *I*-statistics for converted individual genotypes as well as for join-count statistics. A wide range of levels of dispersal and numbers of alleles and allele frequencies are modelled and the results reveal the different general effects of each of these factors on these statistics. We also examine the range of appropriate sampling designs and sizes for which predicted values can be interpolated for specific sampling schemes for any given population genetic field survey. Numbers of alleles and allele frequencies each affect some statistics but not others. The results indicate generally low standard deviations. The results also develop precise and efficient methods of estimating gene dispersal, based on the various autocorrelation measures of standing spatial patterns of genetic variation within populations. The results also extend these methods to loci with multiple alleles, typical of those studied through modern molecular methods.

1. Introduction

Most populations have substantial spatial structure: limits to the distances that individuals (or propagules) disperse result in consanguineous matings by proximity, and consequently the build-up of genetic isolation by distance within populations. Isolation by distance *among* discrete populations has been studied analytically both in terms of consanguinity coefficients (e.g. Malécot, 1950) and in terms of inbreeding coefficients (e.g. Wright, 1943). However, even in discrete systems, with two-dimensional space, analytical methods lead to closed forms only as the distances become large, whereas results for short distances are expressed in terms of Bessel functions (Malécot, 1950). The situation for models for spatial distributions *within* populations (which are not equivalent to models of discrete populations) occupying two spatial dimensions is yet more complex, even in terms of probabilities of identity by descent between two genes or for covariances of gene frequencies, where for example diffusion approximations fail (Felsenstein, 1975). Moreover, to quantify structure more fully in terms of the distribution of *diploid* genotypes, neither measure, separately or in combination, is sufficient for complete characterization. Gillois (1966) has demonstrated that a large number of additional measures are required, making the situation further intractable analytically. Thus there is no analytical theory for the spatial distributions of diploid genotypes within populations.

Instead, stochastic space–time simulations can be employed. Sokal and colleagues (Sokal & Wartenberg, 1983; Sokal *et al.*, 1989) have examined such simulations extensively, using Moran's *I*-statistics calculated for gene frequencies among subpopulational quadrats of 25 individuals each. It was recently shown that a popular experimental measure, based on individual genotypes converted into gene frequencies (Heywood, 1991), has very different properties and values (Epperson, 1995*a*). This is an extreme example of a more general critical effect of the size of quadrats on the values of Moran's *I* (Epperson & Li, 1996). This popular method converts the diploid genotype at each location into the values 0, 0·5 or 1·0 according to the numbers (none, one, two) of a particular allele

* Corresponding author.

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

that were carried in the genotype, and then for each allele Moran's *I*-statistics are calculated for the numerical values. Sets of other spatial autocorrelation statistics known as join-count statistics provide more detailed measures of spatial distributions because they are expressed in terms of all the different combinations or pairs of diploid genotypes (in this respect they are analogous to the measures of Gillois). In this paper we characterize the statistical properties of Moran's *I*-statistics for converted genotypes and join-count statistics for multiallelic loci under isolation-by-distance processes.

Spatial autocorrelation statistics have become widely used for characterizing the actual genetic structure of populations of various species. Most experimental studies that survey the genetic structure within a population produce maps of locations of sampled individuals and their diploid genotypes at various loci, determined through various molecular or biochemical assays. Starting with such maps of genotypes, in order to minimize subsequent loss of information it is possible to calculate either join-count statistics or Moran's *I*-statistics for genotypes converted to numeric values, rather than combine data into arbitrary quadrat areas, which results in a loss of statistical power corresponding approximately to decreasing sample size by a multiple equalling the number of individuals per quadrat (Epperson & Li, 1996). Spatial autocorrelation studies based on individual genotypes have been highly successful and these studies continue to increase in number (reviewed in, for example, Epperson, 1993; Peakall & Beattie, 1995; Shapcott, 1995; Real & McElhany, 1996; Leonardi *et al*., 1996). Moreover, spatial statistics have been successful even in populations where structure is very weak, whereas, in the same populations, analyses of F_{st} often failed to detect significant spatial structure (e.g. Bacilier *et al*., 1994).

Parametric values of join-count statistics and Moran's *I*-statistics for individual genotypes have been examined (Epperson, 1995*a*, *b*), and recently we studied the statistical values of join-count statistics under appropriate sampling schemes for diallelic loci (Epperson & Li, 1997). The results showed that there were low standard deviations, and that join-count statistics could be used to estimate dispersal based solely on standing spatial distributions in actual populations. These statistics constitute the most detailed characterization of spatial structure to date. They take on a role complementary to summary measures in experimental studies. Individual genotype measures of spatial autocorrelations and summary measures are not mutually exclusive choices, since both have relative advantages and disadvantages. Summary measures have advantages in that fewer statistics need be reported and they require smaller corrections for multiple tests, and this becomes increasingly important as the numbers of alleles and loci are increased (e.g. Bertorelle & Barbujani, 1995; Michalakis & Excoffier, 1996). However, summary measures of pairwise genetic similarity (or distance) suffer from the fact that they are not directly related to *a priori* kinship coefficients, and what is measured is relatedness always relative to the existing population. This has been an issue since the initial attempts by Morton and colleagues (Morton, 1973). Moreover, they lose the advantage for standard spatial autocorrelation statistics, which can be distribution-free. On the other hand, spatial autocorrelation statistics based on pairs of individual genotypes have advantages in that less information is lost, and more detailed expectations can be obtained under different models and different kinds of structures may be detected. Moreover, under the randomization hypothesis spatial autocorrelation methods require no assumptions about the underlying generative process. It is worth noting that the oft-cited issue of applicability of measures to modern molecular data, such as DNA sequence data, is largely irrelevant in the context of genetic structure within populations. Ewens (1974) has shown that there is almost no additional information in the infinite-sites mutation model (e.g. DNA sequence data) compared with infinite-alleles models (allele frequency data) when the product, '$4Nu$', of four times the mutation rate and the population size, is less than 1·0. Most reported populations have values an order of magnitude smaller or yet less (Ewens, 1974).

Spatial structure in terms of Moran's *I*-statistics for converted diploid genotypes – statistics that have become popular measures for field studies – have not been characterized, and neither have any of the statistics for loci with multiple alleles, which often is the case in field studies, all the more so when modern molecular markers are analysed. In the present paper we conduct a massive simulation study, with a number of space–time simulations an order of magnitude greater, and a number of samples two orders of magnitude greater than in previous studies. We use sets of simulations designed to determine the general effects of the number of alleles per locus and the allele frequencies on the parametric and statistical values of Moran's statistics for converted genotypes and join-count statistics. It is to be expected that spatial distributions must depend on the number and frequency of alleles, but the effects on spatial statistics are not obvious. Moreover, we characterize for a broad range of conditions the statistical values of join-count statistics as measures of genetic structure for multiallelic loci within populations. Because spatial statistics depend on the actual spatial distribution (which in turn varies widely with the distances of dispersal) and the spatial orientation and size of samples from that population, we use a wide range of

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

(five different) dispersal levels and six different arrays of allele frequencies. Join-count statistics for the diallelic case were studied earlier (Epperson & Li, 1997). In the present paper we also study Moran's *I*-statistic for converted genotypes for the diallelic case and both types of statistics for multiple allele cases. By considering a large number of combinations we can examine the joint effects of dispersal, number of alleles and allele frequencies, so that we can examine not only the marginal effects but also any potential interactions. For the multiallele cases we conducted highly replicated sets of simulations (100 for each set, or 3000 space–time simulations in total), and a wide range of sampling schemes (7 different sampling schemes for each simulation, or 21000 samples in total), and similarly in addition 1300 simulations for the various diallelic cases. Naturally, it is intractable to consider all possible sampling schemes, and numbers and frequencies of alleles, but our results can be widely interpolated and extrapolated to specific cases. These massive results essentially complete analyses of autocorrelation statistics as parametric and statistical values under isolation by distance models and complete the framework for their applications for selectively neutral loci in field studies.

2. Methods

(i) *Simulations*

Each simulated population consisted of 10000 individuals with diploid genotypes, located on a 100×100 square lattice. Each simulation was initialized with a random distribution of diploid genotypes at an arbitrary locus, in Hardy–Weinberg proportions. We simulated a variety of sets with different numbers of alleles (three, four or five). For each multiplicity of alleles (and for each dispersal model, discussed below), a set of simulations was conducted where each simulation had equal initial frequencies of all alleles. Allele frequencies changed very little during the course of a simulation. In addition, three types of sets of simulations where the frequencies of the alleles were unequal were conducted (Table 1). Together, the sets allowed contrasts that separate effects of allele frequency *per se* from possible effects of number of alleles. These contrasts capture the paramount features of the effects of numbers of alleles. We did not consider cases with more than five alleles, in part because this forces low allele frequencies. It has already been established that when allele frequencies decrease below about 0·2 to 0·1, the degree of autocorrelation and scale of patches both decrease even in a diallelic case (Epperson, 1995*b*). Other details of the FORTRAN simulation program, which uses Monte Carlo methods to simulate stochastic generations of life cycles, were described previously (Epperson, 1990). All models were pure isolation by distance, with no selection processes (Epperson, 1990). We ran 100 simulations for each set, on a Sun Sparcstation 20.

Table 1. *Models of unequal allele frequencies for simulation sets with multiple alleles*

Allele				
1	2	3	4	5
0·1	0·3	0·6	—	—
0·1	0·1	0·3	0·5	—
0·1	0·1	0·1	0·1	0·6

Table 2. *Dispersal parameters in the various sets of simulations with multiple alleles*

	Simulation dispersal model[a]				
	1	2	3	4	5
N_f	1	25	49	1	625
N_m	9	25	49	225	625
N_e	4·2	25·1	50·2	115·2	632·4

[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 neighbourhood size.

Table 3. *The sample sizes in terms of sampled individuals for various sampling schemes from the simulated populations with multiple alleles*

	Area[a]		
Porosity[b]	10000	5000	2500
1	10000	—	—
1/4	—	1250	625
1/9	—	544	272
1/25	—	200	100

[a] The total number of individuals in the population that the entire sample lattice covered.
[b] The proportion of total population individuals that were sampled.

Sets with more than two alleles also varied according to five dispersal models, which together represent a very wide range of dispersal levels (Table 2). The five dispersal models were a representative set of a more continuous set of dispersal models simulated in earlier studies for the simpler two-allele cases (Epperson & Li, 1997). Either or both the female and male parents of an offspring were chosen at random (using two Uniform (0, 1) pseudorandom numbers to choose the two coordinates for each parent) generally from one of the nearest N_f and N_m (respectively) neighbours including self. Thus, each individual within the group

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

Table 4. *Mean values of Moran's I-statistics for individual genotypes (standard deviation in parentheses) for different dispersal models and sampling schemes. Also shown are the rates of rejection of the null hypothesis*

Sampling scheme			Wright's neighbourhood sizes												
Porosity[a]	Area[b]	Size	4·2	8·4	12·6	25·1	25·1	50·2	83·7	125·7	115·2	230·4	316·2	632·4	Random
1	1	10000	0·45	0·38	0·25	0·20	0·15	0·12	0·07	0·05	0·04	0·03	0·01	0·01	−0·000
			(0·02)	(0·03)	(0·03)	(0·04)	(0·03)	(0·03)	(0·02)	(0·02)	(0·02)	(0·01)	(0·01)	(0·01)	(0·006)
			0·96	1·00	1·00	1·00	1·00	1·00	1·00	0·99	0·96	0·96	0·49	0·32	0·09
4	1	2500	0·33	0·31	0·22	0·18	0·14	0·11	0·07	0·05	0·04	0·03	0·01	0·01	0·000
			(0·03)	(0·03)	(0·04)	(0·04)	(0·03)	(0·04)	(0·03)	(0·03)	(0·02)	(0·02)	(0·01)	(0·01)	(0·010)
			1·00	1·00	1·00	1·00	1·00	1·00	1·00	0·91	0·86	0·66	0·17	0·13	0·01
9	1	1089	0·26	0·25	0·18	0·16	0·13	0·10	0·06	0·05	0·03	0·02	0·01	0·01	−0·002
			(0·03)	(0·04)	(0·04)	(0·05)	(0·03)	(0·04)	(0·03)	(0·03)	(0·02)	(0·02)	(0·02)	(0·02)	(0·017)
			1·00	1·00	1·00	1·00	1·00	0·98	0·83	0·71	0·49	0·34	0·08	0·11	0·07
25	1	400	0·16	0·17	0·13	0·12	0·09	0·08	0·05	0·04	0·03	0·02	0·01	0·01	−0·001
			(0·04)	(0·04)	(0·05)	(0·05)	(0·04)	(0·05)	(0·04)	(0·04)	(0·03)	(0·03)	(0·03)	(0·03)	(0·027)
			1·00	1·00	0·97	0·93	0·85	0·74	0·47	0·37	0·25	0·17	0·08	0·07	0·07
4	0·5	1250	0·33	0·30	0·22	0·17	0·14	0·10	0·06	0·04	0·03	0·02	0·01	0·01	−0·001
			(0·04)	(0·05)	(0·05)	(0·05)	(0·04)	(0·04)	(0·03)	(0·03)	(0·02)	(0·02)	(0·02)	(0·01)	(0·013)
			1·00	1·00	1·00	1·00	1·00	0·99	0·90	0·71	0·50	0·37	0·15	0·06	0·05
9	0·5	544	0·26	0·24	0·17	0·15	0·12	0·09	0·06	0·04	0·03	0·02	0·01	0·00	−0·004
			(0·05)	(0·06)	(0·06)	(0·06)	(0·05)	(0·04)	(0·04)	(0·03)	(0·03)	(0·03)	(0·02)	(0·02)	(0·024)
			1·00	1·00	1·00	1·00	0·98	0·87	0·57	0·41	0·31	0·17	0·05	0·09	0·08
25	0·5	200	0·16	0·16	0·12	0·11	0·09	0·07	0·04	0·03	0·02	0·02	0·00	0·01	−0·000
			(0·06)	(0·06)	(0·07)	(0·07)	(0·06)	(0·06)	(0·05)	(0·04)	(0·04)	(0·04)	(0·04)	(0·04)	(0·041)
			0·95	0·93	0·74	0·73	0·64	0·51	0·24	0·17	0·12	0·12	0·07	0·05	0·11
4	0·25	625	0·33	0·29	0·20	0·15	0·12	0·08	0·05	0·03	0·02	0·02	0·01	0·00	−0·004
			(0·06)	(0·06)	(0·06)	(0·07)	(0·05)	(0·04)	(0·03)	(0·03)	(0·03)	(0·02)	(0·02)	(0·02)	(0·019)
			1·00	1·00	1·00	1·00	0·95	0·81	0·63	0·41	0·19	0·14	0·03	0·05	0·03
9	0·25	272	0·25	0·22	0·16	0·13	0·11	0·07	0·04	0·03	0·01	0·01	0·00	0·00	−0·007
			(0·07)	(0·08)	(0·07)	(0·07)	(0·06)	(0·05)	(0·04)	(0·04)	(0·04)	(0·03)	(0·04)	(0·04)	(0·033)
			1·00	1·00	0·96	0·83	0·82	0·56	0·34	0·16	0·10	0·05	0·08	0·09	0·06
25	0·25	100	0·15	0·14	0·10	0·09	0·07	0·06	0·02	0·02	0·01	0·01	−0·01	0·00	−0·001
			(0·07)	(0·09)	(0·09)	(0·09)	(0·07)	(0·08)	(0·06)	(0·05)	(0·06)	(0·06)	(0·06)	(0·05)	(0·058)
			0·70	0·65	0·57	0·42	0·28	0·29	0·14	0·08	0·08	0·08	0·09	0·06	0·06

[a] The porosity of the sample design.
[b] The proportion of the total population actually sampled under a given sample design.

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

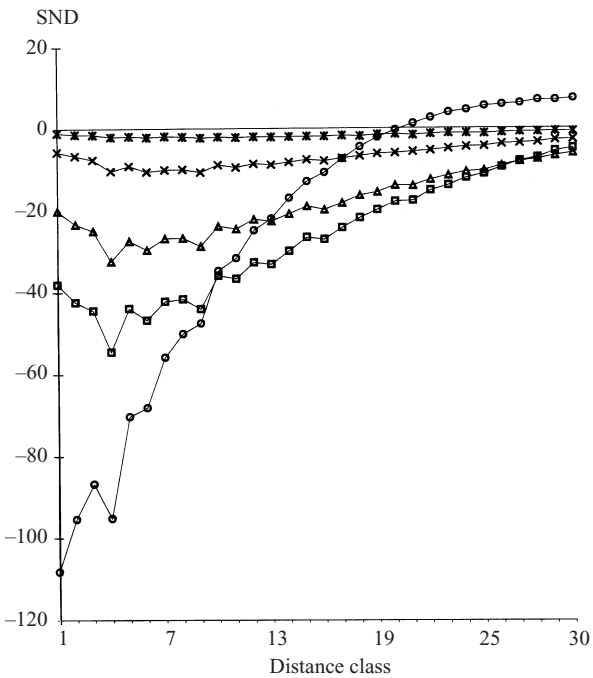

Fig. 1. Correlograms of SNDs for simulations with different levels of dispersal for the case of three alleles with unequal frequencies, for the total number of joins between unlike genotypes, for dispersal models: 1, circles ($N_e = 4{\cdot}2$); 2, squares ($N_e = 25{\cdot}1$); 3, triangles ($N_e = 50{\cdot}2$); 4, crosses ($N_e = 115{\cdot}2$); 5, asterisks ($N_e = 632{\cdot}4$). SNDs are calculated based on full sampling.

of size N_f and N_m had an equal chance of being the female or male parent, respectively. This may be considered unrealistic for many species, in which the probability of dispersal decays with distance. However, it is justified by the fact that the form of the dispersal curve has very little effect on spatial structure; rather it is the standardized neighbourhood size that matters (e.g. Rohlf & Schnell, 1971). In total, 3000 ($6 \times 5 \times 100$) space–time simulations were run for loci with multiple loci. A total of 1300 simulations were run for diallelic loci with a more continuous range of dispersal parameters (see Table 4).

(ii) *Statistical characterization of populations*

As in previous studies, the spatial distributions of genotypes during the period of the quasi-stationary phase were characterized by computing the statistics at generation 200 for each simulation run. Characterizations of a single generation anywhere in the range from about 50 to several thousand are adequate, because during this period the simulated populations exist in a quasi-stationary phase (e.g. Epperson, 1990), and it is more meaningful to replicate over entire simulations rather than over generations. Each simulated population for loci with more than two alleles was sampled in seven different ways (Table 3), which have been shown to cover the range of possible optimal strategies (Epperson & Li, 1997). Briefly, based on simple arguments, the scale of sampling should be constructed to cover an area containing four to nine patches (especially critical for low to moderate dispersal) to avoid, for example, sampling single patches or areas between two patches (and otherwise improve *spatial* stationarity at large scales), yet include 10–20 individual sample points within each patch. Thus, for example, sampling every individual over an area is not efficient, because it requires sampling very large numbers of individuals (unless dispersal is large); instead a sampling grid should be used. A larger set of sampling schemes was conducted for the otherwise simpler diallelic cases (Epperson & Li, 1997), and the ones chosen for the present study avoided redundancy, which is unnecessary since we may extrapolate by adjusting sample sizes (see below). The sampling schemes varied according to the total number of individuals in the population area sampled and the proportion (for sake of brevity we term this 'porosity') of the total number that were actually sampled from the population of individuals covered by a sample area. Thus a sample lattice was superimposed onto a simulated population surface of genotypes. Note that porosity affects the spatial scale of sampling as well as the total size of the sample (Table 3). For porosity equal to 1 all individuals in the sample area were sampled. We omitted stochastic location sampling, since this has virtually no effect (Epperson & Li, 1996, 1997).

We calculated Moran's *I*-statistic for the individual genotypes. For each allele, the genotype at each location was converted into the values 0, 0·5 or 1·0 according to the numbers (none, one, two) of that allele that were carried in the genotype. For each allele Moran's *I*-statistics were calculated for the numerical values.

For each of the seven samples for each of the multiallelic simulations, pairs or 'joins' were classified according to the two genotypes and with respect to distances classes, D, in multiples of lattice units. To reduce over 1 gigabyte of output to that which is most useful, we focus on statistics for the joins between identical homozygotes, among the many types of joins in a multiallelic setting, and the total number of unlike joins. For the former, the number of joins between two identical homozygotes for allele j for distance class D, $n_{jj}(D)$, was compared with the expected number, $u_{jj}(D)$, and standard deviation, $\mathrm{SD}_{jj}(D)$, under the null hypothesis, H_0, of sampling pairs without replacement [$u_{jj}(D) = W(D)\, n_j(n_j-1)/2n(n-1)$; where n_j is the number of individuals with genotype j, n is the total sample size, and $W(D)$ is the total number of all joins for distance class D], and without regard to their locations. The difference was standardized to form a test statistic,

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

Table 5. *Mean SND values for total unlike joins for various alleles for different levels of dispersal* (*Wright's neighbourhood size*, N_e) *and with different types of sampling* (*standard deviations in parentheses*)

Dispersal Model (N_e)	Allele model[a]	Porosity[b] Area[c]:	Sampling strategy 1 1	4 0·5	4 0·25	9 0·5	9 0·25	25 0·5	25 0·25
4·2	3e		−108·2	−25·9	−17·4	−12·1	−7·9	−4·6	−3·0
			(6·3)	(3·4)	(2·9)	(2·6)	(2·2)	(1·6)	(1·5)
4·2	3u		−95·5	−23·0	−15·4	−10·9	−7·1	−4·0	−2·6
4·2	4e		−133·3	−31·2	−20·9	−14·7	−9·7	−5·6	−3·8
4·2	4u		−110·6	−26·1	−17·9	−12·5	−8·5	−4·7	−3·2
4·2	5e		−152·1	−34·7	−23·4	−16·1	−10·5	−6·0	−4·0
4·2	5u		−123·3	−29·1	−19·0	−13·7	−8·7	−5·0	−3·0
25·1	3e		−38·0	−11·4	−6·9	−6·5	−3·7	−3·2	−1·8
			(5·4)	(3·0)	(2·4)	(2·3)	(1·8)	(1·7)	(1·5)
25·1	3u		−35·0	−10·2	−6·8	−5·7	−3·6	−2·4	−1·5
25·1	4e		−44·7	−13·0	−8·2	−7·3	−4·5	−3·3	−1·9
25·1	4u		−41·5	−12·5	−8·0	−7·1	−4·5	−3·3	−2·1
25·1	5e		−50·7	−14·9	−9·2	−8·4	−5·2	−4·0	−2·6
25·1	5u		−48·5	−14·1	−8·8	−8·0	−4·8	−3·6	−2·0
50·2	3e		−20·0	−6·1	−03·7	−3·5	−1·9	−1·7	−1·1
			(3·9)	(1·9)	(1·4)	(1·7)	(1·3)	(1·2)	(1·1)
50·2	3u		−19·5	−5·7	−3·5	−3·6	−2·2	−1·9	−1·1
50·2	4e		−23·2	−6·6	−3·8	−3·7	−2·0	−2·0	−1·2
50·2	4u		−23·2	−6·7	−4·1	−3·8	−2·4	−1·8	−1·0
50·2	5e		−26·1	−7·7	−4·7	−4·5	−2·7	−2·2	−1·4
50·2	5u		−27·3	−8·2	−4·7	−4·8	−2·7	−2·5	−1·3
230·4	3e		−5·6	−1·6	−0·8	−1·0	−0·4	−0·6	−0·1
			(3·9)	(1·9)	(1·4)	(1·7)	(1·3)	(1·2)	(1·1)
230·4	3u		−5·7	−1·6	−1·0	−0·9	−0·6	−0·5	−0·3
230·4	4e		−6·7	−2·0	−1·0	−1·3	−0·7	−0·6	−0·5
230·4	4u		−7·0	−1·9	−1·1	−1·3	−0·7	−0·7	−0·3
230·4	5e		−7·6	−2·2	−1·3	−1·5	−0·7	−9·8	−0·4
230·4	5u		−8·7	−2·5	−1·3	−1·6	−0·7	−0·9	−0·5
637·4	3e		−1·1	−0·6	−0·4	−0·2	0·0	−0·1	−0·0
			(0·9)	(1·0)	(1·0)	(0·9)	(1·0)	(1·1)	(1·0)
637·4	3u		−1·4	−0·3	−0·2	−0·1	−0·1	0·0	0·1
637·4	4e		−1·2	−0·4	−0·1	−0·4	−0·3	−0·2	−0·2
637·4	4u		−1·2	−0·4	−0·2	−0·2	−0·0	−0·1	−0·0
627·4	5e		−1·5	−0·2	−0·0	−0·1	−0·0	−0·2	0·1
637·4	5u		−1·7	−0·7	−0·2	−0·4	−0·1	−0·2	0·0

[a] u denotes unequal allele frequency, e denotes equal.
[b] The porosity of the sample design.
[c] The proportion of the total population actually sampled under a given sample design.

$\mathrm{SND}_{jj}(D) = (n_{jj}(D) - u_{jj}(D))/\mathrm{SD}_{jj}(D)$, which has an asymptotic standard normal distribution under H_0, which is appropriate for the large samples typically used in genetic studies (Cliff & Ord, 1981). These statistics are completely free of assumptions about the underlying distributions, which would not be the case if they were based on sampling with replacement, in which case the mean must be known and constant at all locations, as was proved by Cliff & Ord (1981). Positive values greater than 1·96 indicate statistically significant excesses of a type of join at a distance, and negative values indicate deficits. In addition, the total number of joins between unlike genotypes was computed, and a SND test statistic for H_0 was formed for each distance class.

3. Results

(i) *Moran's I-statistics for individual genotypes*

Changes in the number of alleles and the allele frequencies did not cause statistically significant differences in Moran's *I*-statistics (nor their standard deviations and rates of rejection of the null hypothesis) for individual genotypes. Thus we do not display all the values, which can be obtained on request from the authors. Instead, we focus on the 1300 simulations (13000 samples) for the two-allele case, by examining the statistical values of Moran's *I*-statistics for individual genotypes (Table 4). In almost all samples the results are determined by the sample porosity or

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

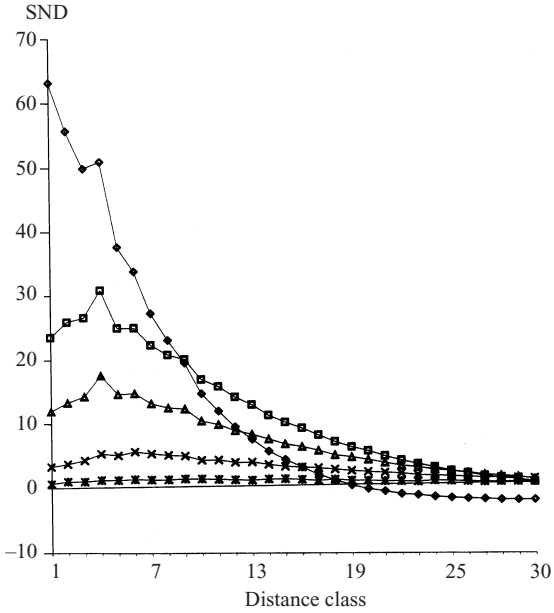

Fig. 2. Correlograms of SNDs for simulations with different levels of dispersal for the case of three alleles with unequal frequencies, for joins between like homozygotes, for dispersal models: 1, diamonds (N_e = 4·2); 2, squares (N_e = 25·1); 3, triangles (N_e = 50·2); 4, crosses (N_e = 115·2); 5, asterisks (N_e = 632·4). SNDs are calculated based on full sampling.

distance class but *not* by the sample size. Only a small decrease is observed for the smallest sample (n = 100). A simple relationship is also seen in the effects of sampling on the standard deviations. For an otherwise identical scheme (i.e. same sample porosity), when sample size (area) is decreased the standard deviation is increased approximately by the square root of the ratio. Statistical power, i.e. the observed rates of rejection of the null hypothesis, is given in Table 4. In all cases the values are close to those found by treating the set means as a normal deviate with the observed standard deviation and calculating the resulting probability of I having values less than or equal to zero. For all cases, Moran's I-statistics for individual genotypes have smaller statistical power than do SNDs for the total number of unlike joins, and the differences become large as the number of alleles increases. The I-statistics generally have essentially equivalent power to SNDs for like-homozygotes (see below), although slightly greater for alleles with low frequencies.

Increasing dispersal has a much stronger effect on the mean values. Dispersal has relatively little effect on the standard deviations. Thus dispersal regime may be inferred from observed values in samples, with a good degree of certainty.

(ii) *Join-count statistics*

Fig. 1 shows the SNDs for a typical case of the full sampling situation (n = 10000, porosity = 1) for the total number of unlike joins. As dispersal increases the deficits become reduced. In the full sampling case, for all but the highest amounts of dispersal (N_e = 632·4) the statistics have extremely high statistical power, rejecting the null hypothesis nearly 100% of the time (results not shown). Table 5 displays the SNDs for the total number of unlike joins for distance class 1 for the various sampling schemes. For most of the sample schemes that are appropriate under field conditions (e.g. porosity not equal to 1) the statistical power of the SND test statistics also remains quite high. To reduce these complicated tables we do not show the statistical power in terms of the observed rates of rejection of the null hypothesis. However, as for Moran's statistics, these rates were very close to the values found by treating the set means as a normal deviate with the observed standard deviation and calculating the resulting probability of SND having values crossing over zero. Thus the expected rejection rates can be inferred from the tables. For example, mean values approximately 1 standard deviation from zero have a rate of rejection near one-third. Again, allele number and frequencies had negligible effects on standard deviations; thus only those for the three alleles with equal frequencies are displayed in Table 5. For each sample scheme (except porosity 25, quarter population), the power was usually greater than 50% and in many cases at or near 100%, for low to moderate levels of dispersal ($N_e < 50$). The values of the standard deviations are virtually identical to those for the corresponding two-allele cases (Epperson & Li, 1997), as are the statistical powers or rates of rejection of the null hypothesis. In general, the effect of reducing sample size is to decrease a mean SND (and its standard deviation) by the square root of the ratio of sample sizes. For example, in Table 5 it is evident that, for a given porosity, the values of SNDs for the half-population (an area containing 5000 individuals) are close to $\sqrt{2}$ times smaller than those for the total area. When the area is restricted to one-quarter the SNDs are reduced by approximately one-half, but slightly less apparently owing to stochastic fluctuations (spatial non-stationarity over small fixed areas). Naturally, the SNDs and their statistical power are reduced as dispersal becomes quite large. Yet it is remarkable that a moderate-size sample is likely to detect structure even when $N_e > 200$, despite the classical claim that populations with this degree of dispersal should behave as randomly mating populations (Wright, 1943).

The absolute values of the SNDs for the total number of unlike joins consistently increase as the number of alleles increases. In contrast, the SNDs for the total number of unlike joins are essentially independent of the allele frequencies. For almost every dispersal model and every allele multiplicity,

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

Table 6. *Mean SND values for joins between identical homozygotes, for various alleles, for different levels of dispersal (Wright's neighbourhood size, N_e) and with different types of sampling (standard deviations in parentheses)*

Dispersal model (N_e)	Allele model[a]	Allele	Porosity[b] Area[c]:	Sampling strategy 1 1	4 0·5	5 0·25	9 0·5	9 0·25	25 0·5	25 0·25
4·2	3e	1		63·2	15·2	9·9	7·2	4·6	2·7	1·7
				(5·7)	(3·0)	(2·9)	(2·4)	(2·3)	(1·7)	(1·5)
	3u	1		51·3	12·0	7·5	5·7	3·6	1·7	1·0
		2		62·0	14·7	9·6	7·0	4·7	2·5	1·6
		3		72·3	16·9	11·2	7·9	5·1	2·7	1·7
	4e	1		59·2	13·7	9·2	6·5	4·5	2·2	1·5
	4u	1		49·6	11·4	7·8	5·4	3·8	1·8	1·0
		3		61·2	14·0	9·9	6·5	4·5	2·4	1·8
		4		69·6	16·8	11·3	7·9	5·3	3·0	1·8
	5e	1		57·1	13·5	9·2	5·9	3·8	2·5	1·7
	5u	1		51·1	11·0	7·1	5·3	3·4	1·7	0·9
		5		73·0	17·6	11·5	8·1	5·1	2·8	1·6
25·1	3e	1		23·5	6·5	3·8	4·0	2·3	1·6	0·9
				(5·2)	(2·6)	(2·5)	(2·1)	(1·9)	(1·4)	(1·2)
	3u	1		12·9	3·3	2·3	2·9	1·2	0·7	0·3
		2		20·9	5·9	3·9	3·2	2·0	1·3	0·7
		3		29·0	8·1	5·2	4·2	2·7	1·9	1·1
	4e	1		20·4	5·8	3·9	2·9	1·7	1·3	0·9
	4u	1		13·7	3·5	2·0	2·0	1·3	1·0	0·7
		3		21·8	6·4	4·0	3·3	2·0	1·4	0·8
		4		27·3	8·1	5·2	4·4	2·7	2·0	1·3
	5e	1		18·4	5·5	3·0	2·6	1·6	1·5	0·7
	5u	1		12·5	2·8	1·5	1·8	1·0	0·8	0·2
		5		30·0	8·6	5·2	4·7	2·8	2·1	1·1
50·2	3e	1		12·0	3·8	2·2	2·2	1·1	1·1	0·5
				(3·3)	(2·1)	(1·6)	(1·7)	(1·4)	(1·5)	(1·2)
	3u	1		6·2	1·7	1·0	1·1	0·7	0·5	0·1
		2		11·6	3·2	2·0	1·9	1·3	1·0	0·3
		3		16·4	4·5	2·7	2·7	1·6	1·2	0·7
	4e	1		9·8	2·7	1·4	1·4	0·6	0·6	0·2
	4u	1		5·9	1·8	0·8	0·8	0·3	0·4	0·2
		3		12·1	3·1	1·9	1·7	1·0	0·6	0·4
		4		16·0	4·4	2·7	2·5	1·5	1·1	0·6
	5e	1		9·2	2·9	1·5	1·5	0·7	0·7	0·3
	5u	1		5·5	1·4	0·8	0·7	0·4	0·2	0·1
		5		17·2	5·1	3·0	2·8	1·6	1·3	0·6
230·4	3e	1		3·3	1·1	0·8	0·7	0·2	0·1	0·0
				(3·3)	(2·1)	(1·6)	(1·7)	(1·4)	(1·5)	(1·1)
	3u	1		1·3	0·2	0·1	0·2	0·1	0·2	0·0
		2		3·0	0·8	0·4	0·5	0·2	0·2	0·2
		3		5·1	1·2	0·7	0·8	0·5	0·5	0·3
	4e	1		3·0	0·6	0·4	0·6	0·4	0·3	0·1
	4u	1		1·4	0·1	−0·2	0·4	0·1	0·3	0·0
		3		3·3	0·7	0·2	0·5	0·3	0·4	0·1
		4		5·2	1·3	0·7	1·0	0·5	0·4	0·2
	5e	1		2·4	0·6	0·3	0·6	0·4	0·4	0·1
	5u	1		1·5	0·6	0·4	0·4	0·2	0·1	−0·2
		5		5·3	1·5	0·8	0·9	0·4	0·4	0·1
637·4	3e	1		0·7	0·5	0·3	0·1	−0·1	−0·1	−0·0
				(1·2)	(1·1)	(1·0)	(1·1)	(0·9)	(1·0)	(0·1)
	3u	1		0·3	0·0	−0·0	−0·1	−0·2	−0·0	−0·0
		2		0·7	0·1	0·1	0·1	0·0	−0·1	−0·1
		3		1·3	0·2	0·1	0·2	0·1	−0·1	−0·1
	4e	1		0·4	0·1	−0·0	0·1	0·1	0·0	0·0
	4u	1		0·2	0·0	0·1	0·1	−0·2	−0·0	−0·0
		3		0·8	−0·1	−0·1	0·1	0·0	0·1	−0·1
		4		0·9	0·3	0·2	0·2	−0·0	0·1	0·0
	5e	1		0·6	0·0	−0·1	−0·2	−0·3	0·0	−0·2
	5u	1		0·2	0·1	0·3	−0·0	0·1	0·1	−0·1
		5		0·9	0·5	0·2	0·2	−0·0	0·2	−0·0

[a] u denotes unequal allele frequency, e denotes equal.
[b] The porosity of the sample design.
[c] The proportion of the total population actually sampled under a given sample design.

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

values for all unequal allele frequencies are very close to those for equal allele frequencies (Table 5). The only substantial exception is the case of full sampling in the model with very low dispersal, $N_e = 4{\cdot}2$, for which SNDs for equal allele frequencies are substantially more negative than for unequal allele frequencies.

The SNDs for joins between identical homozygotes are shown in Fig. 2 and Table 6. The values for alleles with frequencies 0·3, 0·5 or 0·6 are virtually identical among models with different numbers of alleles, including the two-allele case (Epperson, 1995*b*; Epperson & Li, 1997). The values for low-frequency alleles can be substantially smaller, indicating a strong effect when allele frequencies are reduced to about 0·1. Reductions of the same magnitude were observed in the two-allele case. The values are nearly identical regardless of the number of alleles. In sum, these results indicate that changes in the number of alleles affect the SNDs for the total number of unlike joins only when N_e is low to moderate and do not affect SNDs for like homozygotes. Skewing of allele frequencies has no effect unless a frequency is reduced to about 0·1 or less, under which condition, the SNDs for the corresponding joins between like homozygotes are substantially decreased, whereas the SNDs for total number of unlike joins are unaffected. These results are consistent with the idea that low-frequency effects are primarily reflected in similar patch structures and sizes of patches, only fewer patches of homozygotes for alleles in low frequency. Only when the allele frequency is about 0·1 or lower is the patch size reduced. Again, allele number had negligible effects on standard deviations; however, allele frequency did have a small effect similar to the diallelic case (Epperson & Li, 1997). Thus only those for the three alleles with equal frequencies are displayed in Table 6, and other values are available on request from the authors.

It should be noted that the standard deviations reported in the Tables 4, 5 and 6 combine statistical variation with stochastic variation, except in the case of full sampling which is subject only to stochastic variation. Thus the standard deviations are appropriate for corresponding sampling schemes in real populations. They may be interpreted as variances for test statistics for the null hypothesis that the spatial distribution of genotypes is that expected for a selectively neutral locus subject to the given level of dispersal, for samples of genotypes for a single locus.

4. Discussion

Our results demonstrate that both Moran's *I*-statistics for converted diploid genotypes and SND-statistics have high statistical power, and low stochastic and statistical variation, even when sample sizes are quite small. Thus, autocorrelation statistics can be used in experimental studies in a number of distinct fashions, for loci with two to many alleles.

(i) *Multiallelism and population structure*

The patch structures observed for two-allele cases are maintained in all the multiallele models studied. Thus the dominant features of the spatial distributions are maintained despite the fact that the distributions themselves cannot be identical for loci with different allele configurations. The stochasticity that drives the formation of patches dominates over the initial presence of multiple alleles within local areas in the founding populations. Thus, it appears that some emergent characterizations of population structure based on two-allele models are robust to the existence of as many as five alleles and possibly more. The major effect of increasing numbers of alleles is an indirect one in tending to skew allele frequencies away from 0·5. The results indicate that if allele frequency is reduced to about 0·2, then the SNDs remain unaffected, and thus the number of patches of concentration of that allele are reduced but the sizes of patches are unchanged. If frequency is reduced further to about 0·1 or less, the SNDs for joins between identical homozygotes do become reduced, and have smaller *X*-intercepts, indicating reductions in patch sizes. The same effects of allele frequency were observed in the two-allele case (Epperson, 1995*b*).

The SNDs for the total number of unlike joins do not change with allele frequencies (except where dispersal is extremely low, e.g. Wright's neighbourhood size, N_e, less than *c*. 5), as is also true in the two-allele case. They do take moderately greater negative values as the number of alleles is increased. The fact that the SNDs for joins between homozygotes are unchanged suggest this is apparently a purely statistical phenomenon. The SNDs are test statistics whose values may depend strongly on the numbers of the various genotypes as well as their spatial distributions.

Moran's *I*-statistics for genotypes converted into allele frequencies were not affected by either the number of alleles or the population allele frequencies, in our extensive studies of two-, three-, four- and five-allele models.

(ii) *Values in samples*

The most powerful test statistic, in all the sampling schemes for all studied multiallelic models, is that for the total number of unlike joins. In the two-allele cases, the absolute magnitude of means and standard deviations for the total number of unlike joins are almost identical to those for the joins between identical homozygotes, and the power (not shown) is nearly

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

identical to that for Moran's *I*-statistics for converted genotypes. The observed increase in the SND for the total number of unlike joins in the various sampling schemes, resulting from increasing numbers of alleles, is connected to the behaviour of the parametric values discussed above.

The results provide precise predictions for analyses of survey data for selectively neutral loci in natural populations with a given dispersal level. First, we may make any adjustments for the numbers and frequencies as outlined just above. Then we note that unless sample sizes are very small then Moran's *I*-statistics for converted genotypes are virtually unchanged from the parametric values, but we must interpolate the values of the SNDs in experiments by using the square root of the ratio of experiment sample sizes relative to the closest one found in the tables. The standard deviations of both types of statistics must be interpolated also by the square root of the ratio in sample sizes. Then by the approximation of treating the adjusted SND as a normal variate with the adjusted standard deviation, we may use the probability of such a variate being less than or equal to zero (in the case of negative means, we use instead greater than or equal to zero) to obtain interpolated predicted values of the probabilities of observing a significant value fitted to specific experimental studies (see also Epperson & Li, 1997).

The total number of unlike joins has the additional advantage that a single statistic may be constructed (for each distance class and for each locus) and thus smaller Bonferroni-type corrections for multiple tests are required. These corrections are very conservative, yet must be employed in sets of multiple join-count statistics or sets of multiple Moran's *I*-statistics for converted genotypes for multiallele loci, because the relationships among different elements are unknown.

A large number of simulations and sample schemes were examined in this study. The sample schemes are among the optimal ones examined in an earlier analysis of the two-allele case, which included a substantially wider range of sampling schemes (Epperson & Li, 1997). These schemes also incorporate some additional important general considerations for sampling strategies (Epperson, 1993). The scale of the sampling lattice should cover an area that is expected to contain at least four to nine patch areas, in order to avoid statistical fluctuations that might occur by inadvertently sampling only within one patch or the area between two patches. In addition, because genetic samples typically contain only a few hundred to several hundred sampled individuals, it follows that the percentages of sampled individuals over the sample area should be in the range of about 1/4 or 1/9 to 1/25, respectively. Complete census is very inefficient.

The standard errors are relatively small and for estimates based on k loci they will be $\sqrt{k}$ times smaller. Multiple-locus studies are an important strategy used in experimental studies. Using modern molecular methods a large number of multiallelic genetic loci may easily be assayed within natural populations. The present results for multiallelic loci follow those in our study for the diallelic case (Epperson & Li, 1997). The present results may be used to estimate the level of dispersal, in a way that is more often feasible and more efficient than those based on direct observations of movements. This can be done by finding predicted values, by adjusting the values in the tables for experimental sample size and porosity (for the latter, approximate population density may be easily estimated and the scale of the sampling is generally known), for various levels of dispersal, and matching these with the values observed in the experiment. An example, using data from *Ipomoea purpurea*, was presented in an earlier paper (Epperson & Li, 1997). Even in the case where there is only a single locus, with sample sizes of a few hundred for low-to moderate-dispersal cases, or several hundred to a thousand for high-dispersal cases, the statistical power is high.

This work was supported in part by National Institutes of Health grant GM48453 and McIntire-Stennis project number 1774 to B.K.E.

References

Bacilier, R., Labbe, T. & Kremer, A. (1994). Intraspecific genetic structure in a mixed population of *Quercus petraea* (Matt.) Lelbl and *Q. robur* L. *Heredity* **73**, 130–141.

Bertorelle, G. & Barbujani, G. (1995). Analysis of DNA diversity by spatial autocorrelation. *Genetics* **140**, 811–819.

Cliff, A. D. & Ord, J. K. (1981). *Spatial Processes*. London: Pion.

Epperson, B. K. (1990). Spatial autocorrelation of genotypes under directional selection. *Genetics* **124**, 757–771.

Epperson, B K. (1993). Recent advances in correlation studies of spatial patterns of genetic variations. *Evolutionary Biology* **27**, 95–155.

Epperson, B. K. (1995*a*). Fine-scale spatial structure: correlations for individual genotypes differ from those for local genotypes. *Evolution* **49**, 1022–1026.

Epperson, B. K. (1995*b*). Spatial distribution of genotypes under isolation by distance. *Genetics* **140**, 1431–1440.

Epperson, B. K. & Li, T.-Q. (1996). Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proceedings of the National Academy of Sciences of the USA* **93**, 10528–10532.

Epperson, B. K. & Li, T.-Q. (1997). Gene dispersal and spatial genetic structure. *Evolution* **51**, 672–681.

Ewens, W. J. (1974). A note on the sampling theory for infinite alleles and infinite sites models. *Theoretical Population Biology* **6**, 143–148.

Felsenstein, J. (1975). A pain in the torus: some difficulties with models of isolation by distance. *American Naturalist* **109**, 359–368.

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press

Gillois, M. (1966). Le concept d'identité et son importance en génétique. *Annales de Génétique* **9**, 58–65.

Heywood, J. S. (1991). Spatial analysis of genetic variation in plant populations. *Annual Review of Ecology and Systematics* **22**, 335–355.

Leonardi, S., Raddi, S. & Borghetti, M. (1996). Spatial autocorrelation of allozyme traits in a Norway spruce (*Picea abies*) population. *Canadian Journal of Forest Research* **26**, 63–71.

Malécot, G. (1950). Quelques schemas probabilistes sur la variabilité des populations naturelles. *Annales de l'Université de Lyon*, *Section A* **13**, 37–60.

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

Morton, N. E. (1973). Kinship bioassay. In *Genetic Structure of Populations* (ed. N. E. Morton), pp. 158–163. Honolulu: University of Hawaii Press.

Peakall, R. & Beattie, A. J. (1995). Does ant dispersal of seeds in *Sclerolaena diacantha* (Chenopodiaceae) generate local spatial genetic structure? *Heredity* **75**, 351–361.

Real, L. A. & McElhany, P. (1996). Spatial pattern and process in plant–pathogen interactions. *Ecology* **77**, 1011–1025.

Rohlf, F. J. & Schnell, G. D. (1971). An investigation of the isolation by distance model. *American Naturalist* **105**, 295–324.

Shapcott, A. (1995). The spatial genetic structure in natural populations of the Australian temperate rain forest tree *Atherosperma moschatum* (Labill) (Monimiaceae). *Heredity* **74**, 28–38.

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

Sokal, R. R., Jacquez, G. M. & Wooten, M. C. (1989). Spatial autocorrelation analysis of migration and selection. *Genetics* **121**, 845–855.

Wright, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.

https://doi.org/10.1017/S001667239900378X Published online by Cambridge University Press