Heredity (2008) 100, 231–232

www.nature.com/hdy

NEWS AND COMMENTARY

Dispersal estimation

Demystifying Moran's *I*

F Rousset

Heredity (2008) **100,** 231–232; doi:10.1038/sj.hdy.6801065;
published online 26 September 2007

One of the most enduring ideas in spatial population genetics is that the 'neighbourhood size' determines the extent and geographic pattern of genetic differentiation. In particular, there is a widespread belief that the effects of dispersal on genetic structure are essentially determined by a straightforward parameter: the mean-squared parent offspring distance σ^2, from which the neighbourhood size can be calculated as $\mathrm{Nb} = 4\pi D\sigma^2$ (for two-dimensional habitats), where D is the population density. If this principle was sound, it would be highly convenient. For example, to determine how Nb affected a particular parameter (characterizing the observed variation in genetic data), it would be sufficient to derive the relationship between Nb and the parameter for some pattern of dispersal that is readily simulated. This relationship could then be applied to more complex patterns of dispersal with the same Nb. In one of the most recent incarnations of this logic, Epperson (2005, 2007) considered a method of estimation of neighbourhood size from values of Moran's I, a long-used descriptor of the spatial structure in genetic data.

However, there is a fundamental problem with this logic. The genetic consequences of dispersal do not depend solely on the variance of the dispersal distribution, σ^2. The shape of the dispersal distribution (for example, how it deviates from a normal distribution) does affect the magnitude of genetic differentiation from place to place (Rousset, 1997). This principle seems intuitively reasonable, so why should the converse view be widely established? It turns out that the neighbourhood size is actually a robust predictor of the increase of differentiation with distance (robust in the sense that the increase does not depend strongly on the shape of the dispersal distribution). Consequently, Nb can be estimated from this rate of increase, but is less directly related to the overall magnitude of differentiation. Here, I will recall how Moran's I relates to the increase of differentiation with distance, from which I will explain Epperson's simulation results and clarify their implications.

In a diploid population, Moran's I can be viewed as an estimator of $(Q_C - \bar{Q})/[(1+Q_w)/2 - \bar{Q}]$ (Hardy and Vekemans, 1999), where the Q's are probabilities of identity in state, Q_C for pairs of individuals within some distance class C, $\bar{Q}$ for all pairs in the sample and Q_w for gene diversity within individuals. As such I shares the desirable general properties of measures, such as this, which are defined as ratios of differences of probabilities of identity. In particular, when observations are taken at a small spatial scale, they have a fast approach to equilibrium and are robust with respect to details of mutation processes (Rousset, 2002). Further, the occurrence of $(1+Q_w)/2$ in the denominator makes I independent of the effects of selfing, except through effects on dispersal itself (Hardy and Vekemans, 1999).

In a two-dimensional plant population with selfing, the identity Q_r of genes at Euclidian distance r from each other obeys

$$a_r = \frac{Q_0 - Q_r}{(1+Q_w)/2 - Q_0} \approx \frac{\ln(r)}{2D\pi\sigma^2} + \text{constant} \qquad (1)$$

(Rousset, 2004, pp 131–132), where the constant does not depend on distance but is a complex function of the shape of the dispersal distribution (Rousset, 1997). Epperson (2005) argues that the shape has little effect, citing for example 'Malécot's (1948) finding that spatial structure depends (…) not much on the shape of the dispersal function'. But there is no such finding in Malécot's (1948) work. Indeed he gave general expressions in terms of the full dispersal distribution, and used the example for bivariate Gaussian dispersal.

Any method for estimating neighbourhood size will depend on the specific assumptions of the dispersal model on which it is based, or on relationships that can only be approximate when a wide range of dispersal distributions is considered. Insofar as the above approximation is valid for most distances, the average I for individuals within distance class C would be

$$I_C \approx \frac{(1+Q_w)/2 - Q_0}{(1+Q_w)/2 - \bar{Q}} \times \frac{\langle \ln(r)\rangle - \langle \ln(r)\rangle_C}{2D\pi\sigma^2} \qquad (2)$$

where $\langle\cdot\rangle$ and $\langle\cdot\rangle_C$ are the average values within the total sample and within distance class C, respectively.

There would be some problems assessing this relationship in simulations and then making use of it to interpret real data. For example, the value of $\langle \ln(r)\rangle - \langle \ln(r)\rangle_C$ was fixed in the simulations below, but more generally its variation should be taken into account to derive an estimator of neighbourhood size from I (Vekemans and Hardy, 2004). The constant term from Equation (1) still somewhat affects I_C since it affects the first ratio. In many cases, this ratio is $1/2 < . < 1$ and may be neglected as a first approximation. Otherwise, one can correct for it—as discussed by Vekemans and Hardy (2004) and Rousset (2004, p 132).

If we ignore the latter correction, $\ln(I_C)$ should be approximately linearly related to ln(Nb), with a slope of -1. A reanalysis of Epperson's (2005) simulations estimates the slope as -1.13 (Figure 1). In the original paper, ln(Nb)

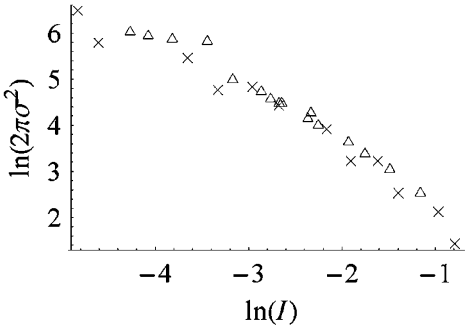

Figure 1 Moran's I revisited. ×: I values from Table 1 of Epperson (2005); △: I values from Table 4 of Epperson (2007).

was fitted to I, and it can be shown that $\ln(\mathrm{Nb}) = 1.11 - 1.13 \ln(I)$ is a better predictor of neighbourhood size, whether the quality of the predictor is assessed by mean, mean square or maximum relative error.

The original predictor performs poorly: the relative error ((estimated Nb)/(true Nb)) ranges from 70% (too low) to 91% (too high). The improved predictor reduces the root mean squared error from 0.86 to 0.39. Likewise, for more recent simulations (Epperson, 2007, Table 4), $\ln(\mathrm{Nb}) = 1.38 - 1.16 \ln(I)$, with root mean squared error reduced from 0.93 (for the old predictor) to 0.19. This reanalysis illustrates that, if one really wants to use the simulation results to estimate Nb, the relationship with $\ln(I)$ should be used rather than I. The remaining variation around the fits confirms that I is not exactly a function of neighbourhood size.

Should we then use Moran's I at all? There is no clear reason for doing so. The parametric bias of the originally proposed methods is generally larger under their simulation conditions than the realized bias reported in simulations to assess some alternative estimators (Rousset, 2000; Leblois *et al*., 2003, 2004; Watts *et al*., 2007). My aim here is not to advocate a particular alternative, but to make clear what is known or what is not. Further discussion should be based on the performance of rival estimators in head-to-head comparison in biologically relevant conditions.

It has been claimed that I performs well if used for 'only short distances' in comparison to an alternative method of Vekemans and Hardy (2004). But in such an approach, I would still be a function of Q and thus of all distances within the sample. An alternative estimator 'may have some disadvantages, including the fact that it assumes the decrease with distance is exponential' (actually, logarithmic). But I also depends on the increase of differentiation with distance, so the simulation-based method makes the implicit assumptions about this increase that are inherent to the simulation conditions. Although not fitted specifically to these conditions, the analytical results presented above suggest a better approximation to describe the simulation results, and form a better basis for understanding the properties of Moran's I.

Dr F Rousset is at Institut des Sciences de l'Evolution (UM2-CNRS), Université Montpellier 2, Place Eugène Bataillon, CC 065, Montpellier cedex 5 34095, France.

e-mail: Rousset@isem.univ-montp2.fr

Epperson BK (2005). Estimating dispersal from short distance spatial autocorrelation. *Heredity* **95**: 7–15.

Epperson BK (2007). Plant dispersal, neighborhood size and isolation by distance. *Mol Ecol* **16**: 3854–3865.

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.

Leblois R, Estoup A, Rousset F (2003). Influence of mutational and sampling factors on the estimation of demographic parameters in a 'continuous' population under isolation by distance. *Mol Biol Evol* **20**: 491–502.

Leblois R, Rousset F, Estoup A (2004). Influence of spatial and temporal heterogeneities on the estimation of demographic parameters in a continuous population using individual microsatellite data. *Genetics* **166**: 1081–1092.

Malécot G (1948). *Les mathématiques de l'hérédité*. Masson: Paris.

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

Rousset F (2000). Genetic differentiation between individuals. *J Evol Biol* **13**: 58–62.

Rousset F (2002). Inbreeding and relatedness coefficients: what do they measure? *Heredity* **88**: 371–380.

Rousset F (2004). *Genetic Structure and Selection in Subdivided Populations*. Princeton Univ. Press: Princeton, NJ.

Vekemans X, Hardy OJ (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol Ecol* **13**: 921–934.

Watts PC, Rousset F, Saccheri IJ, Leblois R, Kemp SJ, Thompson DJ (2007). Compatible genetic and ecological estimates of dispersal rates in insect (*Coenagrion mercuriale*: Odonata: Zygoptera) populations: analysis of 'neighbourhood size' using a more precise estimator. *Mol Ecol* **16**: 737–751.