Some background on the program *Geneland*

Alexa R. Warwick and Peter Beerli

***Abstract*—A short overview of the theory used in the program *Geneland*. This program uses a Bayesian model implemented in a Markov chain Monte Carlo scheme and colored Poisson-Voronoi tessellation to detect and locate genetic discontinuities in a spatially explicit framework.**

***Index Terms*—Voronoi tessellation, Spatial dependence, Bayesian inference, Markov chain Monte Carlo, Landscape genetics, Population structure**

I. Background

Although methods are available to investigate genetic divergence, they often do not consider the spatial location for each sample. To address this deficiency, the program *Geneland* [Guillot et al., 2005] uses multi-locus, geo-referenced genotype data to investigate the spatial modeling of genetic discontinuities. Specifically, the method (1) estimates the number of populations within the geographical area of interest, (2) maps borders between populations, (3) assigns individuals to populations, and (4) detects possible migrants. Furthermore, it can be used to analyze phenotypic and genotypic data under a consistent framework and address how well divergence in neutral loci predicts phenotypic trait divergence [see example in Guillot et al., 2012].

II. Overall Model

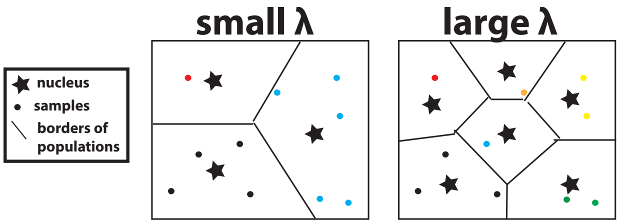
Geneland is based on a statistical model, not an explicit evolutionary model. The basic model assumes Hardy-Weinberg equilibrium within populations and linkage equilibrium between loci within populations. It also assumes that clusters are genotypic/phenotypic homogeneous. These similarities within clusters are the result of shared history, which is inferred from the allele frequencies (or means and variances of phenotypic traits, but we focus on genetic data in this overview).

III. Voronoi Tessellation

Within a geographical region of interest (Δ), each sampled individual (total = n diploid individuals) has a two-dimensional spatial location $t_i = \{x, y\}$ (vector of coordinates, $t = (t_{ij})_{i=1...n}$) and some genetic marker data (vector of all genotypes, $z = (z_{ij})_{i=1...n}$), where L indicates the genotype for each locus ($l = 1, ..., L$).

Across the region of interest, we consider K different populations present, and each population occupies a subdomain Δ1... ΔK . By splitting up space into these subdomains, the program co-estimates the value of K and allele frequencies. It assumes that each subdomain can be approximated by a union of convex polygons. Because we do not know where these non-overlapping polygons should be placed, the model uses a random number of points (m) that are scattered across the landscape under a uniform distribution. Each of these random points ($u_1, ..., u_m$) is then the nucleus for the non-overlapping convex polygon. Sampling points that are located within a particular polygon belong to that particular nucleus. If we assume that each polygon (A_i) contains individuals from a single population we can then label the polygons with a value between 1 and K.

The process of generating polygons via a random number and placement of points (nuclei) is Poisson-Voronoi tessellation (Poisson process generates the random number of nuclei u_i). For any point, x, in the geographical region of interest (Δ), $c(x)$ is the population of its closest nucleus. As a result, the domain is colored by population and the model is therefore often called colored Voronoi tiling. If we assume all populations have equal probability a priori, then each tile in Δ has a $1/K$ probability of belonging to a certain population. Because the amount of spatial dependence is contingent on the fragmentation of the subdomains, the model includes a parameter λ for the rate of the Poisson process which controls the number of nuclei and therefore the number of polygons. Low values of λ have fewer fragments and greater spatial dependence; high values have more fragments and weaker spatial dependence (see Figure 1).



At very high λ, with many points m, each tile would only have a single sampled individual and is simply a non-spatial cluster model [independent identically distributed mixture model like the program Structure; Pritchard et al., 2000].

IV. Frequencies Model

For each new configuration of polygons, the allele frequencies need to be recalculated. If allele frequencies follow an independent Dirichlet distribution then it is called the D-model (or spatial D-model):

$$F_{kl} \sim \text{Dirichlet}(\alpha, ..., \alpha), k = 1, ..., K, l = 1, ..., L$$

where f denotes the vector containing the allele frequencies of each allele at locus l in population k. This model is biological realistic under the neutral theory of mutations, but it does not take into account that different allele frequencies of the populations tend to be alike. In contrast, the F-model [Falush et al., 2003] takes into account both the present population and a hypothetical ancestral population when determining allele frequencies. The F-model also includes a parameter for divergence of the present population (from the ancestral population) as the result of genetic drift (d_1 ... d_K), as shown here:

$$f_{kl} \sim \text{Dirichlet}(f_{Al1}\frac{1-d_k}{d_k}, f_{Al2}\frac{1-d_k}{d_k}, ..., f_{AlJ_l}\frac{1-d_k}{d_k}),$$
$$k = 1, ..., K, l = 1, ..., L \qquad (1)$$

for each allele $1, 2, ..., J_l$ in every locus.

V. Bayesian Inference

For the Bayesian model, we now have a set of observables (genotypes z; coordinates t) and non-observables, which random variables with a distribution, for which we need to definepriors. These unknown parameters are denoted by the following vector (θ):

$$\theta = (K, \lambda, m, u, c, d, f, f_A, s)$$

with

- Spatial parameters: λ, m, u, c
- Genetic parameters: f, f_A, d
- Phenotypic parameters: μ, σ, β (not shown in full model here)

Each of these spatial and genetic parameters and their prior distributions are defined here:

- K = number of population (sampled from a uniform distribution from K_{min}, K_{max})
- λ = rate of Poisson process generating m nuclei; since we do not really know λ we define a hyper prior drawn from a uniform (0, λ max)
- u = events/points/nuclei of the Poisson process
- c = color of the tiles (membership of the partitioned subdomains; sampled from a uniform distribution)
- f = current population frequency of an allele at a locus in a particular cluster
- f_A = ancestral population frequency of an allele at locus in a particular cluster
- d = drift constant
- s = true individual location (t_i = s_i + ϵ_i, using a normal distribution for the error ϵ)

To infer the unknown parameters θ a Markov chain Monte Carlo method is used. Guillot et al. [2005] calculates the likelihood of the observed values (t, z) as

$$\begin{aligned} P(t, z|\Theta) &= P(t|\Theta)P(z|t, \Theta) & (2)\\ &= P(t|\Theta)\prod_i^n \prod_l^L P(z_{il}|\Theta), & (3)\end{aligned}$$

$$P(z_{il}|\Theta) = P(z_{il} = (\alpha, \beta)|\Theta) \begin{cases} 2f_{kl\alpha}f_{kl\beta} & \text{if } \alpha \neq \beta \\ f_{kl\alpha}^2 & \text{if } \alpha \equiv \beta. \end{cases} \qquad (4)$$

where $f_{kl.}$ is the population allele frequency of the alleles α and β at locus l seen in the sampled individuals.

VI. Other Assumptions

The sampling of individuals across the region of interest is likely problematic when tightly clustered rather than dispersed. Also, the model assumes no repeat sampling (repeated individuals over multiple coordiantes because of non-stationarity over time). Each individual must belong to one K cluster. The current version of *Geneland* also does not make use of the heterozygote information.

Bibliography

Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, August 2003.

Gilles Guillot, Arnaud Estoup, Frédéric Mortier, and Jean François Cosson. A spatial statistical model for landscape genetics. *Genetics*, 170(3):1261–1280, 2005. doi: 10.1534/genetics.104.033803.

Gilles Guillot, Sabrina Renaud, Ronan Ledevin, Johan Michaux, and Julien Claude. A unifying model for the analysis of phenotypic, genetic and geographic data. *Systematic Biology*, doi:10.1093/sysbio/sys038, 2012.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

VII. Disclaimer

This text was written by Alexa R. Warwick and Peter Beerli, Florida State University for a course on practical population genetics inference, Fall 2015.