Supplementary material for

Accurate continuous geographic assignment
from low- to high-density SNP data.

Gilles Guillot, Hakon Jonsson, Antoine Hinge, Nabil Manchih, Ludovic Orlando

September 17, 2015
Method

Statistical model

In the following, we assume that a set of allelic counts at bi-allelic loci is available for a set of reference populations of known geographic locations. Genotypes for the same loci are also known for a set of individual of unknown geographic origin. Our method undertakes to predict the most likely geographic origin of these individuals, given the set of geo-referenced panel (hereafter referred to as training data).

Let $f_{sl}$ be the frequency of a reference allele at locus $l$ and geographic location $s$. Let assume that the number of reference alleles is binomial $B(n_{sl}, f_{sl})$ and that loci are statistically independent. Individuals located around location $s$ are, thus, implicitly assumed to belong to a population at Hardy-Weinberg equilibrium, with linkage equilibrium across loci, which provides the same likelihood function for XX PLEASE ADD XX as the one described by [Pritchard et al.] (2000).

Let now assume that spatial variation of allele frequencies, $f_{sl}$s, can be described by a non-parametric surface in two dimensions, which can be modelled following [Wasser et al.] (2004) as a set of spatially auto-correlated random variables with Gaussian distribution (i.e. a random field) denoted $y_{sl}$. Assuming that $f_{sl}$ and $y_{sl}$ are logistically-related, we have $f_{sl} = 1 / [1 + \exp (- (a_l + y_{sl}))]$ where $a_l$ is a locus-specific intercept.

Finally, we model the spatial auto-covariance of allele frequencies by imposing a parametric form to Cov[$y_{sl}, y_{sl'}$]. By assuming that the spatial auto-covariance only depends on the geographic distance $h = |s - s'|$ (i.e. $C(|s - s'|) = \text{Cov}[y_{sl}, y_{sl'}]$), and decays with the geographical distance, we model the form of population structure known as isolation-by-distance [Guillot et al., 2009; Guillot and Orlando, 2015]. We also do not restrict assignment to discrete geographic categories (typically, the geographic locations represented in the training data) but open for assignments in the continuous space.

Estimation within the INLA-GMRF-SPDE framework

A key feature of our model is that it can be handled within the theoretical and computational framework developed by [Rue et al., 2009] and [Lindgren et al., 2011]. The former develops
a framework for Bayesian inference in a broad class of models enjoying a latent Gaussian
structure. The latter bridges a gap between Markov random fields (MRF) and Gaussian
random fields (GRF) theory and makes it possible to combine the flexibility of Gaussian
random fields for modelling and the computational efficiency of Markov random fields for
inference. The approach of [Lindgren et al., 2011] is based on the observation that a Gaussian
random field $y(s)$ with a Matérn covariance function is the solution of the stochastic partial
differential equation (SPDE). Solving numerically this SPDE with finite element techniques
and a smart choice of basis functions makes it possible to use Markov properties. This
framework can be embedded in the INLA method of [Rue et al., 2009], which makes use of the
Markovian structure of the model during computation. The INLA and SPDE approximate
inference methods are implemented in the R-INLA package [Rue et al., 2014], the accuracy of
the INLA method in spatial statistics being widely validated [Lindgren et al., 2011; Simpson
et al., 2012; Martins et al., 2013]. See also Guillot et al. (2013) for the use of a related model
in genomics.

**Practical implementation of INLA-GMRF-SPDE**

We now describe specific steps for casting the problem of continuous geographic assignment
in the INLA-GMRF-SPDE framework. The location of samples with unknown geographic
location is estimated following three steps.

In the first step, we estimate the parameters of the GMRF-SPDE model from the set of geo-
referenced genetic data. There are three parameters ($\sigma, \kappa, \nu$). However, in line with [Lindgren
et al., 2011] and to minimize the computational burden, we set $\nu = 1$. We stress that the
inferential difficulties reported under Markov Random field models by Sørbye and Rue (2014)
bear on Intrinsic Markov Random fields (IMRF). The SPDE-GMRF model considered here
differs sharply from the IMRF model and is not subject to this issue. The estimated pa-
rameters ($\sigma, \kappa$) of the GMRF-SPDE model summarize information on the magnitude and
the spatial scale of variation of allele frequencies. Estimating the ($\sigma, \kappa$) parameters involves
processing the whole dataset jointly, which can be typically computed for datasets consist-
ing of... typically $\sim 500$ individuals and $\sim 1,000$ loci. For larger datasets, we devised a less
computationally intense approach, which proceeds on a limited subset of randomly selected
loci. The parameters estimated on this subset are then used in downstream analyses.

In the second step, we compute estimated geographic maps of allele frequencies for each locus using the parameters previously estimated.

In the third step, we assign samples of unknown origin by maximizing the likelihood that a sample comes from a specific location over the study area (in practice, the nodes of a grid which can be easily chosen to be fine enough to avoid any discretization issue). More specifically we maximise the likelihood \( p(\text{genotypes}|\text{allele freq.}, \text{locations}) \) with respect to the geographic locations, assuming allele frequencies are perfectly estimated. The method therefore provides not only a point estimate of the unknown geographic origin but also a likelihood surface, representing a map summarizing assignment uncertainties, as illustrated in figure I. See (Rue et al., 2009; Lindgren et al., 2011; Simpson et al., 2012; Martins et al., 2013) for details on the INLA method and its implementation with random fields models.

The main competitors of SPASIBA are the SCAT program of Wasser et al. (2004) and the SPA program of Yang et al. (2012). We therefore compare our method to the latter.

Our model is very similar to that of Wasser et al. (2004). As running SCAT on a single dataset of more than 1,000 loci typically requires weeks of computations, we did not carry out full comparison of SPASIBA and SCAT. The comparison was, therefore, limited to SPASIBA and SPA. Furthermore, our focus is on medium-density SNP datasets which are becoming increasingly more common in the field of ecology and we therefore do not compare to recent methods that require high-density SNP data (Drineas et al., 2010; Baran et al., 2013; Rañola et al., 2014; Yang et al., 2014). We also stress that our method is tailored to perform continuous geographic assignment, therefore we do not compare it to methods designed to assign individuals to a set of known populations, such as GENECLASS (Piry et al., 2004).
Figure I: Map of SPASIBA likelihood scores and assignment error (green arrow) recovered for one individual. Data were simulated under the model underlying the SPASIBA program (50 diploid individuals with known origin, 200 SNP markers). We used SPASIBA to assign the most likely geographic origin of a given individual. The red dot indicates the true geographic position of the individuals, while the green triangle corresponds to the position inferred by SPASIBA. Typically, an individual located in an area of low spatial sampling density (left panel) is assigned with larger errors than an individual located in an area of high spatial sampling density or close to an individual of the training sample (right panel). The map relative to a specific individual can be checked for the existence of several local maxima. The global maxima inferred for each and every individual can be compared to help identify which individuals are assigned with low or large confidence.
Results

Model validation on simulated data

We validated our method on datasets simulated under various spatially explicit models, in line with the validation strategy used earlier by Novembre et al. (2008) and Bradburd et al. (2013). A set of individuals is randomly selected and removed from the dataset. Remaining individuals are used to train the algorithm (training dataset) while individuals initially removed from the dataset are used as testing data for which we predict their spatial origin using genotype information only. The accuracy of each method is assessed using the average geographic distance obtained between predicted and known geographic positions.

We first simulated datasets under the model underlying the SPA program (Yang et al., 2012), in which variation of allele frequencies is given by a logistic function in two dimensions characterized by an origin, a slope and a direction. We considered a training set consisting of 100 diploid individuals and evaluated accuracy in assignment for 200 individuals. The locations of individuals were sampled from a uniform distribution on the unit square, the direction of the cline was sampled uniformly on $[-\pi, \pi]$ and the slope was sampled uniformly on $[1, 10]$. This type of simulation can be seen as the best-case scenario for the SPA method.

We then simulated data under the geostatistical random field model underlying the SPA-SIBA program. The data simulated here display far more variability than those generated under the SPA model. We considered a training set consisting of 100 diploid individuals and evaluated accuracy in assignment for 200 individuals. The marginal variance of the random field was set to one and the scale parameter to $10/3$ on a unit square domain.

Lastly, we used the MS program (Hudson, 2002) to simulate data under a two-dimensional stepping stone model. This approach was selected because it explicitly accounts for demographic and mutational processes and therefore provides spatial genetic structure. Importantly, it does not rely on any of the assumptions underlying the SPA and the SPASIBA program. Data were simulated for haploid individuals on a 20x20 grid with training and testing sets of size 380 and 20 individuals respectively. In all cases the mutation and migration were controlled by setting mutation rate $4N\mu = 1$ and the migration rate $4Nm = 0.4$. Simulations were performed for a number of loci varying from 20 to 5,000. Results reported
for each condition are obtained as averages over five independent datasets. Results for the three types of simulations are summarized on figure II.

For data simulated under the logistic curve underlying the SPA program, our method performed similarly to or better than the SPA method, as long as a large number of loci was considered (superior to 1,000). For smaller datasets, SPASIBA achieved a better accuracy than SPA, with, for example an average error twice smaller for 20 loci (Fig. II top panel).

For data simulated under the geostatistical model underlying the SPASIBA program, the assignment errors are typically larger than those observed for data simulated under the SPA model, which reflects the greater spatial complexity in the genetic variation simulated. In such conditions, the SPASIBA method outperforms the SPA method regardless of the number of loci analyzed (Fig. II middle panel).

In our attempts to implement the SPA program on the stepping-stone data, we faced numerous cases where the assignment error appears of several orders of magnitude larger than the size of the geographic domain considered. This phenomenon becomes increasingly important with increasing numbers of loci (Tab. I). Even when such problematic datasets were discarded from the analysis, the assignment error of the SPA method is larger (up to about 10-fold over the range of loci considered) than that of SPASIBA (Fig. II bottom panel).

As SPASIBA provided great performance in simulated settings, we next applied our method to three real datasets, selected to represent a range of possible biological situations.
Figure II: Assignment error on simulated data. We simulated spatially explicit genetic datasets using three methods (Top: SPA, Middle: SPASIBA, Bottom: MS). In the bottom plot, the curve for the SPA method corresponds to the subset of data where SPA did not fail, see text for detail. The average distance between true and inferred locations is used as a measure of the accuracy of each method.
<table>
<thead>
<tr>
<th>Nb loci</th>
<th>Index sim</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>200</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0</td>
<td>0</td>
<td>13</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td>8</td>
<td>10</td>
<td>20</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>10000</td>
<td></td>
<td>7</td>
<td>11</td>
<td>9</td>
<td>13</td>
<td>11</td>
</tr>
</tbody>
</table>

Table I: Summary about problematic runs with the SPA program on data simulated under a stepping stone model: number of individuals with outlier estimated coordinates. These are defined conventionally as those larger than $10^{64}$.
Florida scrub jays

We consider here a dataset consisting of 1,311 Florida scrub jay birds (*Aphelocoma coerulescens*), which are known for their short dispersal distances (Woolfenden and Fitzpatrick, 1984, 1996; Fitzpatrick et al., 1999). For example, Coulon et al. (2010) reported dispersal distances of the order of 1.3-4.2 km (depending on sex and habitat). This species is therefore expected to show strong geographic population structure, which should facilitate geospatial assignment. The species was sampled extensively over Florida and genotyped for a limited number of SNP markers (for a total of 41). This allowed us to explore how the method performs with types of datasets that are classical for ecological surveys and population monitoring.

The population density and the spatial sampling strategy are both characterized by the absence of clusters, which are known to be problematic for traditional population-based assignment methods (Manel et al., 2005). We investigated the assignment accuracy of our method by splitting the dataset into a random training set of 1,000 individuals, the 311 remaining individuals being used as a testing set. Running the SPA program on the same training and testing dataset returned non-sensical results with a large proportion of individuals assigned at locations farther than several thousands of kilometers away from Florida. For SPASIBA outputs, we computed the distance between the predicted origin and the sampling location and used this as a genuine measure of the assignment error. This distance has a median of 26.4 km, a 75% quantile of 76.6 km and a maximum of 274.5 km. The distribution of the distance between predicted origin and sampling location is displayed on figure [III]. This, together with the short dispersal distances of Florida scrub jays, suggests that even if a dispersal event occurred for individuals of our testing set, at the scale of Florida, our method is able to detect their birthplace with relatively high accuracy. This is particularly striking as only 41 SNPs were considered and those had not been pre-selected for the purpose of making assignment, not even for their ability to a priori reflect population structure.
Figure III: SPASIBA geospatial assignments of Florida scrub jays with the SPASIBA method. Arrows originate from the true sampling site and point towards the estimated origin which provide a measure of assignment errors. They are displayed for different quantiles: Top left, 0-median; top right, median-$q_{0.75}$; bottom left, $q_{0.75}-q_{0.9}$. The full distribution of assignment errors is indicated for the 311 individuals of the testing set in the bottom right panel.
**Arabidopsis thaliana in Europe**

We further explore the performance of our method using a large genetic dataset of *Arabidopsis thaliana*, which represents an extensively studied model organism. We consider here a subset of the data from Horton et al. (2012), consisting of the 1,007 samples located in Eurasia with longitude between 20°W and 100°E. We perform assignment on random training sets of eight hundreds specimens at random subsets of $L = 100$ then $L = 1,000$ loci. Geospatial assignment was performed in each case using the remaining 207 samples. As our dataset encompasses a large geographical scale, we investigate both the 2D and 3D versions of these programs. In many runs of SPA in the 3D option, the output was non-sensical, showing samples assigned to geographic regions located well outside the margins of the sampling area. We therefore limited our exploration of the 3D option to $L = 100$. We found that SPASIBA was more accurate than SPA for all cases considered, and predicted the geographic position of a large number of specimens to be extremely close to their known positions (Fig. IV). More specifically, three quarters of the samples were assigned within 375 and 93 kilometers of their exact geographic origin, when using 100 loci and 1,000 loci respectively.
Figure IV: Assignment errors estimated using datasets of 100 SNPs and 1,000 SNPs on *A. thaliana* data. Eight hundreds specimens were used as a training dataset and geospatial assignment was performed using the remaining 207 samples. Assignment errors are indicated in increasing order. On the vertical axis, the assignment error is expressed as a fraction of the distance separating the two most remote points in our sampling window (i.e. 7,500 km).
Lastly, we explore the performance of our SPASIBA method in a case where extensive genetic information is available for a large number of individuals. More specifically, we consider here the subset of the Population Reference Sample (POPRES [Nelson et al., 2008], used by [Novembre et al., 2008]) which consists of 1,385 individuals with grandparents of similar ancestry. We use genotypes at 197,146 loci (after pruning tightly linked loci). In this dataset, the exact geographic origin of individuals is unknown and each individual is conventionally geo-referenced to the centre of its reported country of origin (except for a few countries for which another location was considered as more reflective of the origins of these individuals). This implies that the uncertainty in the known geographic origin of samples varies with the size of the country of origin, ranging from around 80 km in Macedonia up to thousands of kilometres in Russia.

To assess the accuracy of methods on this dataset, we proceeded in two different ways to compute predicted maps of allele frequencies. In a first assessment, we used the whole dataset to compute these maps and estimated origins of each individual using these maps. This is likely to produce unrealistically low estimates of assignment errors. To assess the accuracy of the two methods in a more realistic setting, and following a strategy taken by [Wasser et al., 2004], we removed all individuals of a given country at a time from the dataset, then computed predicted maps of allele frequency with a training set of geo-referenced genotypes only consisting of individuals from all other countries (which we refer below to as 'leave-one-population-out') and estimated origins of remaining individuals from these maps. The detail of estimated origins is displayed in figure V.

In the approach using the whole dataset to obtain allele frequencies maps, the median distance of the estimated origins to the centre of the country is 72.8 km for SPASIBA (187 km for SPA) and the bias defined as the mean distance, per country, between inferred individual geographic origins and the country center) is 7.9 km for SPASIBA (21.8 km for SPA). Therefore, under this validation scheme, both methods show great accuracy, albeit SPASIBA consistently shows slightly better performance than SPA. Under the leave-one-population-out strategy, these statistics are respectively 696 km and 45.8 km for SPASIBA (543 km
and 75km for SPA). This suggests that the accuracy of both methods is extremely reduced when the training dataset does not include a population from the same genetic background as the test individuals. Importantly, while SPA appears to perform better than SPASIBA in this setting, the assignment errors of SPASIBA appear to be homogeneously distributed geographically in contrast to those of SPA, which all appear to converge to the center of the study domain.

Miscellaneous remarks

Our INLA framework enabled us to perform geospatial inference within reasonable running times, scaling linearly with the number of loci considered. Typically, analyses of the Florida Scrub-jay dataset were performed within 10 minutes using a single 3GHz-CPU. In our experience, SPASIBA appears to be an ideal method for the analysis of reduced-representation sequencing data that become increasingly available in ecology. However, for a larger number of loci, SPASIBA is best carried on a computer cluster where the predictive maps of allele frequencies can be computed in parallel. Implementing this strategy on the POPRES data on a 80-CPU cluster, allowed us to carry out the analysis in 24-48 hours.

The algorithm underlying SPA and SPASIBA are essentially deterministic, while SCAT is stochastic. Defining a computing time for an MCMC-based like SCAT is impossible as computations are usually carried out over a number of iterations, larger than what is assumed to be necessary, and it is checked a posteriori and over several independent runs that the MCMC algorithm did not experience any convergence issue.

In SPA, all computations are locus-specific, therefore the computing time scales linearly with the number of loci. In SPASIBA, the computing time for the inference of the parameters of the random field scales non-linearly with the size of the data matrix (whose dimension is given by the product of the number of geographic sampling sites and the number of loci). The task of computing predicted allele frequency maps scales linearly with the number of loci.

In the tasks above, deterministic algorithms seek to optimize one criterion until a condition is fulfilled. For the reasons described above, we are reluctant to provide exact computing
Figure V: Predicted geographic origins of Europeans. We used the POPRES data and evaluated the assignment error of SPA and SPASIBA using the whole dataset approach (top panels, using the whole dataset), or a leave-on-population-out approach (bottom panels, leave-one-pop-out).
times for the various methods discussed here. However, in our computations we observed that computations with SPA are in the order of hundred times faster than those with SPA-SIBA, which are themselves in the order of hundred times faster than those with SCAT. We note however that SCAT is the only program that handles micro-satellite data.

Current limitations of the SPASIBA method

A potential advantage of SCAT over our SPASIBA method is the computer implementation that allows SCAT to restrict geographic assignments to a set of polygonal areas. Implementing this feature in SPASIBA would be straightforward and could increase accuracy in assignment when the spatial sampling window includes areas known to be non-suitable habitats. We note however that in the Florida scrub jay case, SPASIBA assigned only a handful of individuals a few kilometers away from the landmass (Fig. [II]), even though the assignment was not restricted to any specific area of the rectangular domain encompassing Florida.

Lesser accuracy of the SPA method

The SPA method is based on the assumption that allele frequencies vary logistically on the plan or the sphere, displaying essentially a nearly linear behavior in a central region and no variation elsewhere with frequencies fixed to 0 or 1. This may be a reasonable approximation for the data used earlier to assess the SPA method, namely human data in Europe and at the synoptic scale. At smaller scales, spatial patterns of genetic variation also likely reflect the processes of local genetic drift, migration and relatedness, which presumably features more spatial complexity. Additionally, the logistic model underlying SPA has the property of being invariant under shifts orthogonal to the main axis of variation. We believe that a combination of these factors explain the lesser accuracy observed for SPA and also its propensity to numerical instabilities, as observed here with the Arabidopsis thaliana dataset (especially under the 3D option), the Florida scrub jay dataset and MS simulations.
Limitations of current continuous assignment methods

The interpolation of allele frequencies between reference populations assumes a model of isolation-by-distance, however in reality, many biological populations display restricted gene flow due to a range of barriers that disrupt this relationship. These includes habitat variation and physical dispersal barriers (Wang and Bradburd, 2014). This is not handled by any of the continuous assignment methods and may affect the accuracy obtained.

Related to the point above, current continuous assignment methods assume marker neutrality. While this is likely to be true for smaller microsatellite and SNP panels selected at random, genome-wide SNP panels, such as those produced by whole-genome or reduced-representation sequencing are likely to include loci under selection where the change in allele frequency may be completely disconnected from geographic distance. A recent study by Nielsen et al. (2012) suggests that such loci are highly informative for geographic assignment. However, the latter study is not based on an isolation-by-distance model and how the information gained from the use of highly informative loci will be offset by the use of a model that does not fit these loci, remains to be assessed.

Re-appraisal of assignment results on the POPRES dataset

The POPRES population reference sample has become an invaluable resource in many areas of human genetics, including pharmacogenetics and population genetics (Nelson et al., 2008). Here, we were able to bring the assignment error down to 72.8km but we caution that this figure only represents a lower bound for assignment errors. We note, however, that removing all individuals from a country from the training data (the leave-one-population-out approach) resulted in substantially larger assignment errors (696 km and 543 km for SPASIBA and SPA, respectively). Additionally, SPASIBA was characterized by relatively isotropic errors while SPA systematically biased predicted geo-spatial assignments towards the centre of the study area. Our leave-one-population-out approach revealed that none of the two methods is robust to uneven population sampling in the training dataset and are particularly inefficient at estimating the country of origin of an individual whose true country of origin is not represented in the training dataset. It opens avenues for novel
statistical approaches reducing the impact of uneven training sets on spatial assignments.
References


