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

Guillot, Gilles; Jónsson, Hákon; Hinge, Antoine; Manchih, Nabil; Orlando, Ludovic

Published in:
Bioinformatics

Link to article, DOI:
10.1093/bioinformatics/btv703

Publication date:
2016

Document Version
Early version, also known as pre-print

Supplementary material for *Accurate continuous geographic assignment from low- to high-density SNP data.*

Gilles Guillot[*], Hákon Jónsson[†], Antoine Hinge[*], Nabil Manchih[*], Ludovic Orlando[†]

July 14, 2015

[*]Department of Applied Mathematics and Informatics, Technical University of Denmark, 2800, Lyngby, Denmark. gigu@dtu.dk

[†]Centre for Geogenetics, Museum of Natural History and University of Copenhagen, Øster Voldgade 5-7, 1350 København K, Denmark

Method

Statistical model

We consider datasets consisting of a set of allelic counts at bi-allelic loci for a set of reference populations of known geographic locations. Additionally, genotypes for orthologous loci are available for individuals of unknown geographic origin. Our method is tailored to geoposition the latter individuals given the set of geo-referenced genetic data (hereafter referred to as training data). We denote by f_{sl} the frequency of a reference allele at locus l at geographic location s. We assume that the number of reference alleles is binomial $B(n_{sl}, f_{sl})$ with statistical independence across loci. This amounts to assuming that inviduals located around location s form a population at Hardy-Weinberg equilibrium with linkage equilibrium across markers. Our model has therefore the same likelihood function as described by Pritchard et al. (2000). We assume that spatial variation of allele frequencies can be described by a non-parametric surface in two dimensions. Following Wasser et al. (2004), we model the spatial variation of $(f_{sl})_s$ by a set of spatially auto-correlated random variables with Gaussian distribution (a random field) denoted by y_{sl}. We assume that f_{sl} and y_{sl} relate through a logistic function $f_{sl} = 1/[1 + \exp -(a_l + y_{sl})]$ where a_l is a locus-specific intercept. We model the spatial auto-covariance of allele frequencies by imposing a parametric form to $\mathrm{Cov}[y_{sl}, y_{s'l}]$.

We should stress that our method is designed to perform continuous assignment. Therefore, we cannot only rely on a covariance matrix, but need instead a covariance function, which models covariance variation in the continuous space. We assume that $\mathrm{Cov}[y_{sl}, y_{s'l}] = C(|s - s'|) = C(h)$ for some function C, implying that the spatial auto-covariance only depends on the geographical distance $h = |s - s'|$. As commonly assumed in spatial statistics and for reasons that will appear later, we consider that C belongs to the Matérn family i.e. $C(h) = \sigma^2(\kappa h)^\nu 2^{1-\nu}\Gamma^{-1}(\nu)K_\nu(\kappa h)$ where K_ν is the modified Bessel function of the second kind of order $\nu > 0$, $\kappa > 0$ is a scaling parameter and σ^2 is the marginal variance. This model can be defined either in a flat geographical domain, using straight-line distances (2D) or on the sphere using great circle distances (a sub-model referred to below as 3D model) which is more appropriate when analyzing worldwide datasets. The Matérn family of covariance

function is broad and flexible, it includes for example the widely used exponential covariance function $\sigma^2 \exp(-\kappa h)$ as a particular case (Gelfand et al., 2010; Porcu et al., 2010). Under our model, the covariance between allele frequencies at geographical locations s and s' decays with the geographical distance $|s - s'|$ and therefore models the form of population structure known as isolation-by-distance (Guillot et al., 2009; Guillot and Orlando, 2015). However, its main advantage is computational, as explained in the next section.

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 appproximate inference methods are implemented in the R-INLA package (Rue et al., 2014). 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 from unknown geographical origin 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 (σ, κ, ν). 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 parameters (σ, κ) of the GMRF-SPDE model summarize information on the magnitude and the spatial scale of variation of allele frequencies. This step involves processing the whole dataset jointly and can be computed for datasets consisting of typically ~500 individuals and ~1,000 loci. For larger datasets, we devised a strategy limiting computational demands and running times by picking a random subset of loci and performing inference of σ and κ on this subset. 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). In the latter step, we maximimise the likelihood $p(\text{genotypes}|\text{allele freq., locations})$ with respect to the geographical locations, assuming allele frequencies are perfectly estimated. The method provides therefore not only a point estimate of the unknown geographic origin but also a map informative about uncertainty in assignment and multiple putative origins, 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. The accuracy of the INLA method in spatial statistics being widely validated (Lindgren et al., 2011; Simpson et al., 2012; Martins et al., 2013). Additionally, 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. Therefore, we 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,

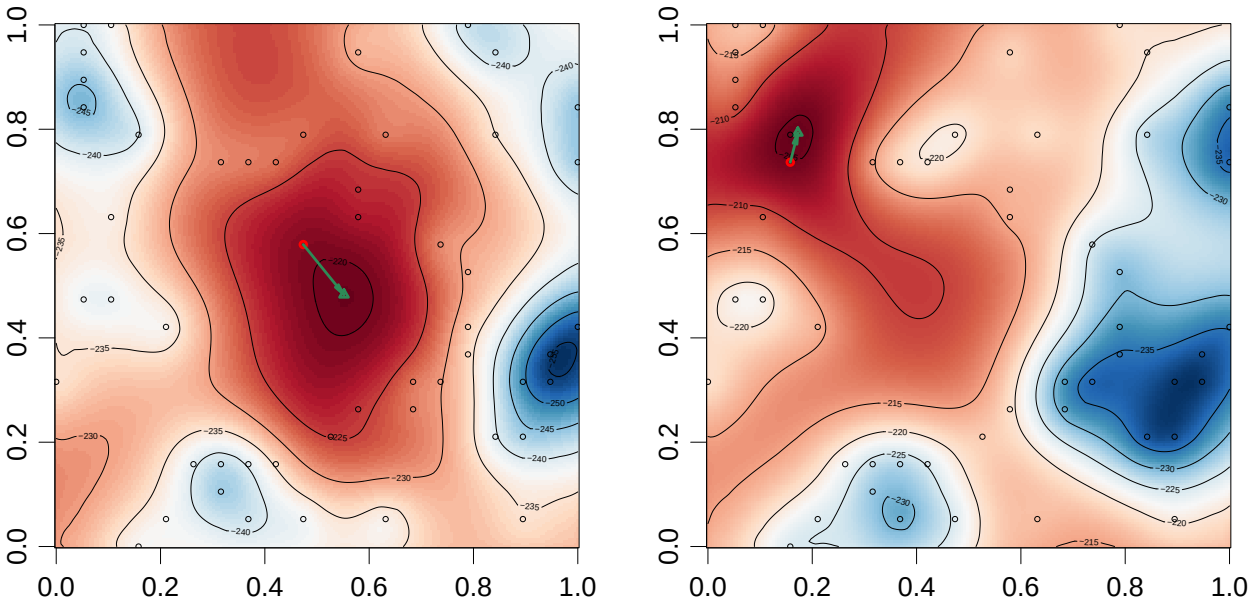


Figure I: Map of SPASIBA likelihood scores and assignment error (green arrow) recovered for one individual. Data were simulated under model underlying the SPASIBA program (50 diploid individuals with known origin, 200 SNP markers). We used SPASIBA to assign the most likely geographic origin of a given 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 a 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 various global maxima corrponding to the various individuals can be compared and help identify which individuals are assigned with low and large confidence.

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).

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 geographical distance obtained between predicted and known geographical 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 SPASIBA 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 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 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 discarding such problematic datasets 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.

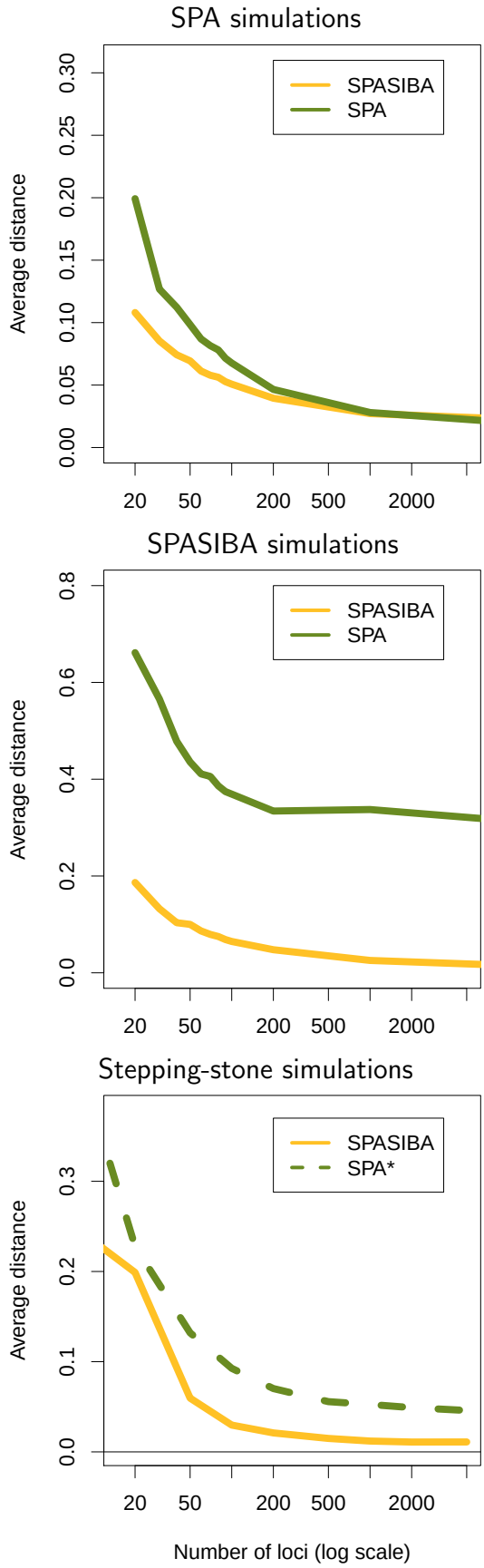


Figure II: Assignment error on simulated data. We simulated spatially explicit genetic datasets using three methods (Top: SPA, Middle: SPASIBA, Bottom: MS). In the bottom plot, the curve for the SPA method corresponds to the subset of data where SPA did not fail, see text for detail.

Nb loci \Index sim	1	2	3	4	5
10	0	0	2	0	0
20	0	0	0	0	0
50	0	0	0	0	0
100	0	0	0	0	0
200	0	0	0	0	0
500	0	0	13	11	0
1000	0	1	0	6	13
2000	8	0	3	0	9
5000	8	10	20	12	9
10000	7	11	9	13	11

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 cœrulescens*), 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 geographical 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 occured 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.

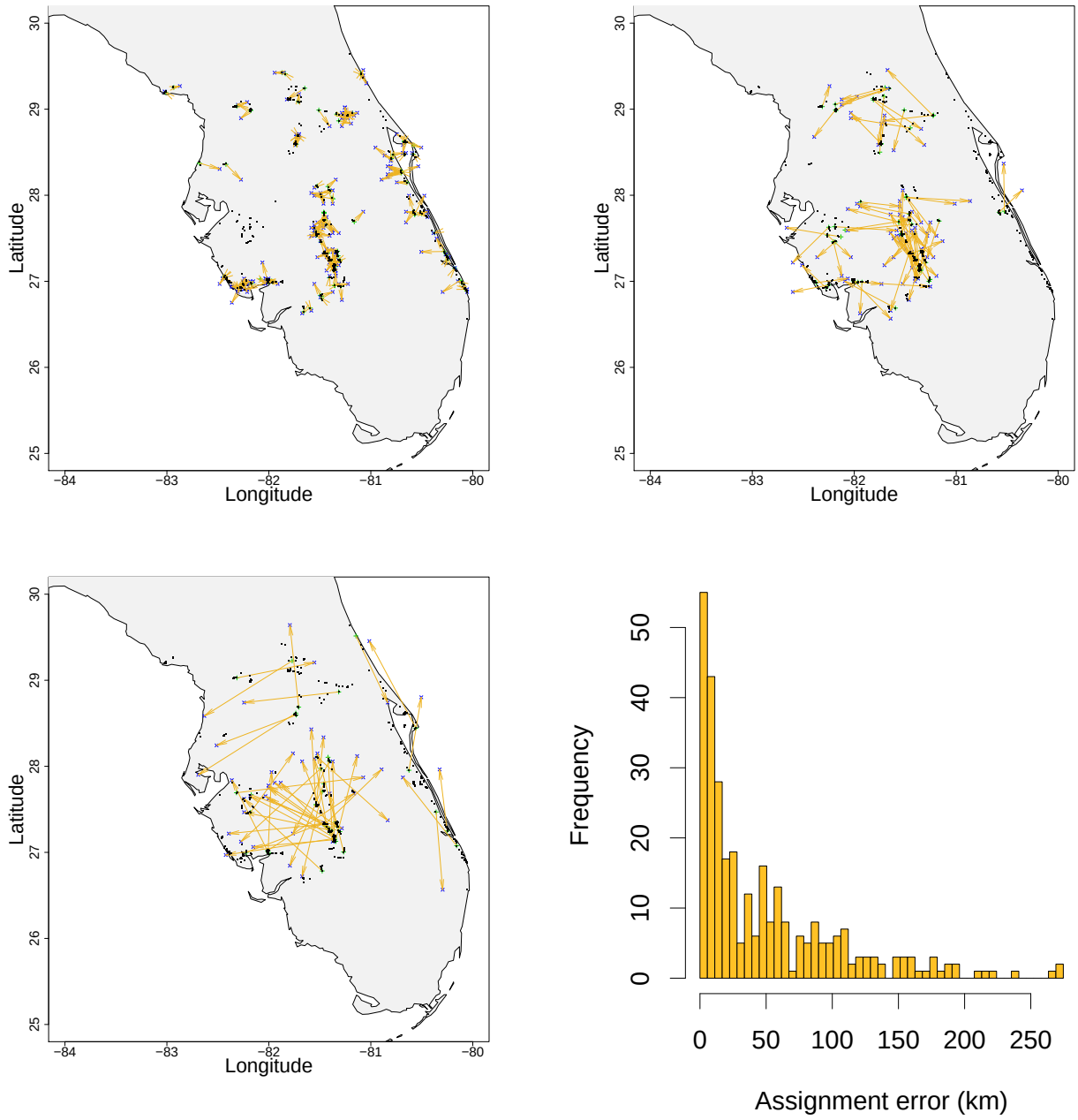
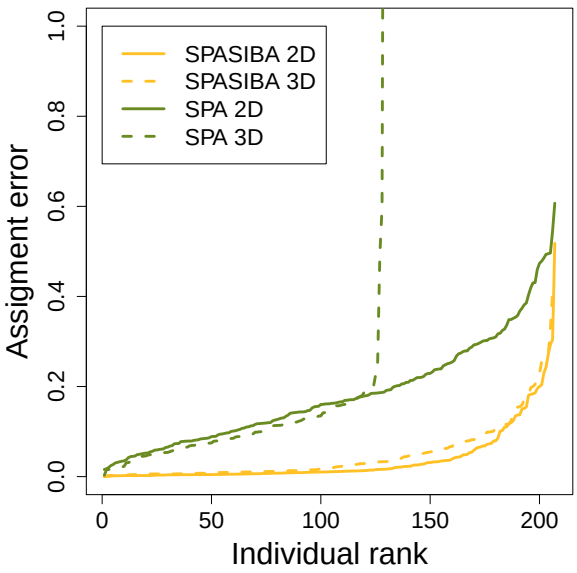


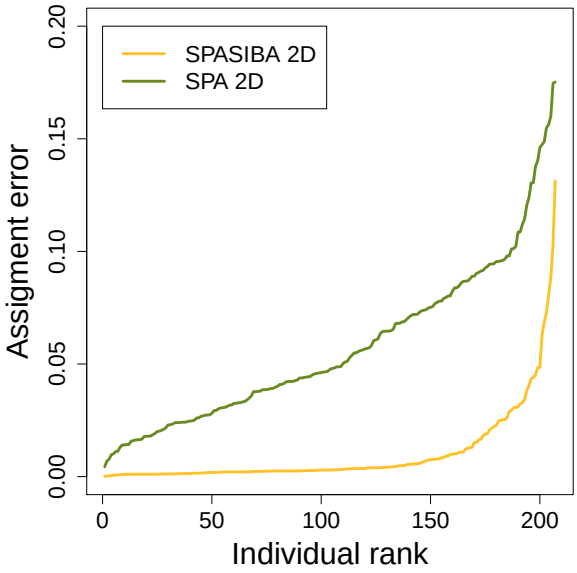
Figure III: SPASIBA geo-spatial assignments of Florida scrub jays with the SPASIBA method. Arrows originate from the true sampling site and point towards the estimated origin and provide a measure of assignment errors. They are displayed for different quantiles: Top left, 0-median; top right, median-$q_{0.75}$; 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. Because the data are sampled at large 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 at the antipodes 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 kilometers of their exact geographic origin, when using 100 loci and within 93 kilometers when using 1,000 loci.



L=100 loci



L=1,000 loci

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 between two most remote points of the geographical sampling window (7,500 km).

Geographic assignment of Europeans

Lastly, we explore the performance of our 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 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. Therefore, 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 country at a time from the dataset, then computed predicted maps of allele frequency with a training set of geo-referenced genotypes consisting of individuals from all other countries only (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 of the per-country average estimated location to 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

The statistical model underlying our method is largely reminiscent of the SCAT program (Wasser et al., 2004, 2007). However, taking advantage of INLA instead of MCMC allowed us to significantly reduce computing times typically by several orders of magnitudes. Additionally, our approach is free from MCMC convergence issues that can increase considerably the computation burden. In the Florida Scrub-jay dataset (1,311 individuals, 41 SNPs), SPASIBA achieved a full analysis in about ten minutes using a single 3GHz-CPU. SCAT required about a week of computation, while SPA provided results within a few seconds. These computing times scale linearly with the number of loci. With such running times and the accuracy levels demonstrated above, SPASIBA appears well tailored for the routine analysis of SNP datasets for non-model species consisting of a few tens of thousands of loci. In particular, it 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

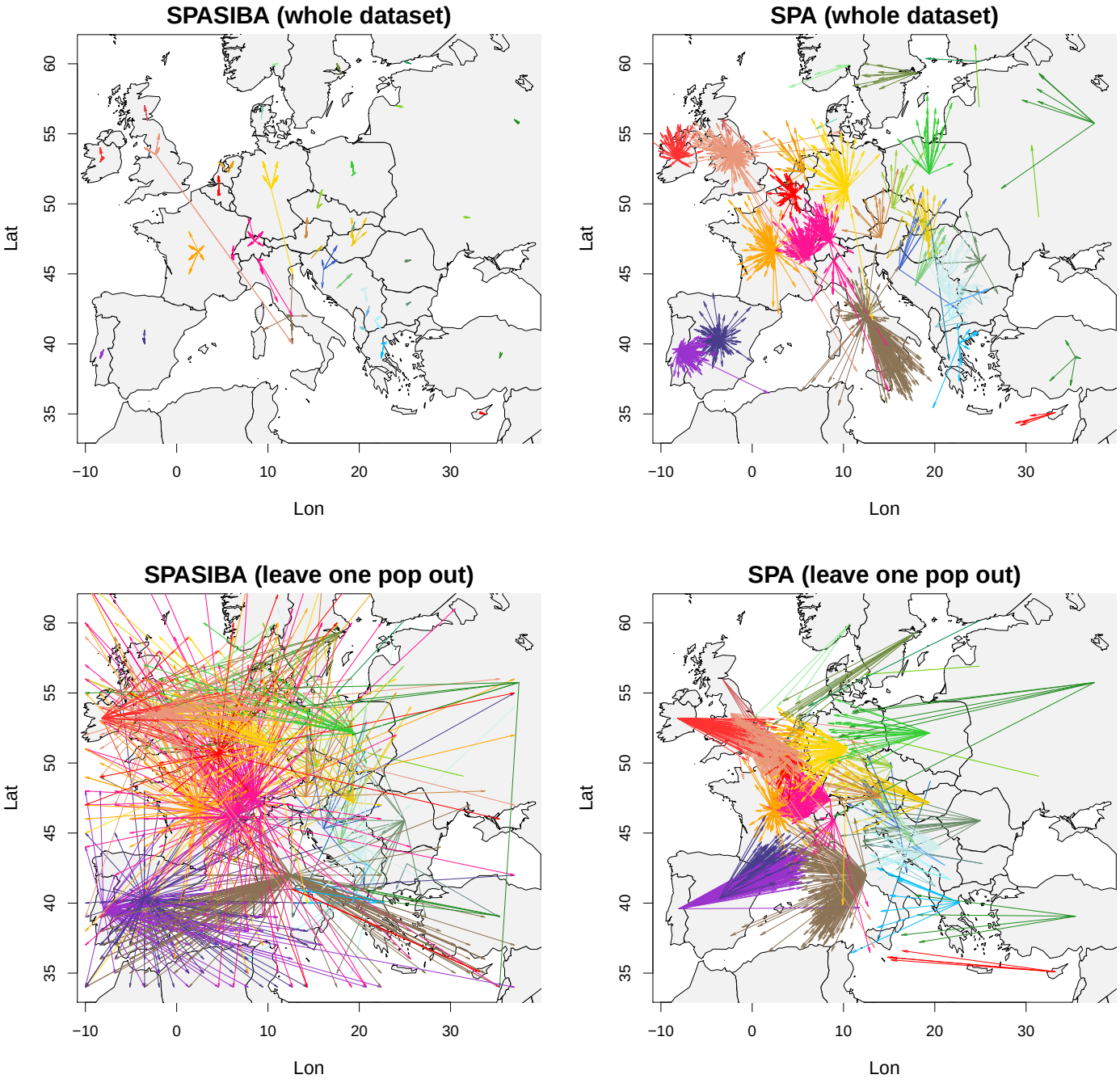


Figure V: Predicted geographic origins of Europeans. We used the POPRES data and evaluated the assignment error of SPA and SPASIBA using the whole dataset approach (top panels, using the whole dataset), or a leave-on-population-out approach (bottom panels, leave-one-pop-out).

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 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 SPASIBA, 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.

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. III), 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 alleles 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, has still 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

Y. Baran, I. Quintela, Á. Carracedo, B. Pasaniuc, and E. Halperin. Enhanced localization of genetic samples through linkage-disequilibrium correction. *The American Journal of Human Genetics*, 92(6):882–894, 2013.

G.S. Bradburd, P.L. Ralph, and G.M. Coop. Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, 67(11):3258–3273, 2013.

A. Coulon, J.W. Fitzpatrick, R. Bowman, and I. J. Lovette. Effects of habitat fragmentation on effective dispersal of Florida Scrub-Jays. *Conservation Biology*, 24(4):1080–1088, 2010.

Petros Drineas, Jamey Lewis, and Peristera Paschou. Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers. *PLoS One*, 5(8):e11892, 2010.

J. W. Fitzpatrick, G. E. Woolfenden, and Bowman R. Dispersal distance and its demographic consequences in the florida scrub-jay. In N. J. Adams and R. H. Slotow, editors, *22nd international ornithological congress*, pages 2465–2479, Johannesburg., 1999. BirdLife South Africa.

A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, editors. *Handbook of Spatial Statistics*. Handbooks of Modern Statistical Methods. Chapman & Hall/CRC, Boca Raton, 2010.

G. Guillot and L. Orlando. *Oxford Bibliographies in Evolutionary Biology*, chapter Population Structure. Oxford University Press, New York, 2015.

G. Guillot, R. Leblois, A. Coulon, and A. Frantz. Statistical methods in spatial genetics. *Molecular Ecology*, 18:4734–4756, 2009.

G. Guillot, R. Vitalis, A. le Rouzic, and M. Gautier. Detection of correlation between genotypes and environmental variables. A fast computational approach for genomewide studies. *Spatial Statistics*, 8:145–155, 2013.

M. W. Horton, A. M. Hancock, Y. S. Huang, C. Toomajian, S. Atwell, A. Auton, N. W. Muliyati, A. Platt, F. G. Sperone, B. J. Vilhjálmsson, et al. Genome-wide patterns of genetic variation in worldwide arabidopsis thaliana accessions from the regmap panel. *Nature Genetics*, 44(2):212–216, 2012.

R.R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

F. Lindgren, H. Rue, and E. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, series B*, 73(4):423–498, 2011.

T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with INLA : New features. *Computational Statistics and Data Analysis*, 67:68–83, 2013.

M.R. Nelson, K. Bryc, K.S. King, A. Indap, A. R. Boyko, J. Novembre, L.P. Briley, Y. Maruyama, D.M. Waterworth, G. Waeber, et al. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3):347–358, 2008.

E.E. Nielsen, A. Cariani, E. Mac Aoidh, G. E. Maes, I. Milano, R. Ogden, M. Taylor, J. Hemmer-Hansen, M. Babbucci, L. Bargelloni, et al. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications*, 3:851, 2012.

J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Indap A. Auton, K.S. King, S. Bergman, M.R. Nelson, M. Stephens, and C.D. Bustamante. Genes mirror geography within Europe. *Nature*, 456:98–101, 2008.

A. Piry, S. Alapetite, J.M. Cornuet, D. Paetkau, L. Baudoin, and A. Estoup. Geneclass2: A software for genetic assignment and first-generation migrant detection. *Journal of Heredity*, 95(6):536–539, 2004.

E. Porcu, J.M. Montero, and M. Schlather, editors. *Advances and Challenges in Space-time Modelling of Natural Events*. Springer, Heidelberg Dordrecht London New York, 2010.

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

J.M. Rañola, D.H. Alexander, and K. Lange. Fast spatial ancestry via flexible allele frequency surfaces. *Bioinformatics*, 2014. URL `doi:10.1093/bioinformatics/btu418`.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, series B*, 71(2):1–35, 2009.

H. Rue, S. Martino, F. Lindgren, D. Simpson, A. Riebler, and E. Krainski. *INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximaxion*, 2014. http://www.r-inla.org/.

D. Simpson, F. Lindgren, and H. Rue. Think continuous : Markovian gaussian models in spatial statistic. *Spatial Statistics*, 1:16–29, 2012.

S. H. Sørbye and H. Rue. Scaling intrinsic gaussian markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51, 2014.

I J Wang and G S Bradburd. Isolation by environment. *Molecular ecology*, 23(23):5649–5662, 2014.

S.K. Wasser, A.M. Shedlock, K. Comstock, E.A. Ostrander, B. Mutayoba, and M. Stephens. Assigning African elephants DNA to geographic region of origin: applications to the ivory trade. *Proceedings of the National Academy of Sciences*, 101(41):14847–14852, 2004.

S.K. Wasser, C. Mailand, R. Booth, B. Mutayoba, E. Kisamo, and M. Stephens. Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proceedings of the National Academy of Sciences*, 104(10):4228–4233, 2007.

G. E. Woolfenden and J. W. Fitzpatrick. *The Florida Scrub Jay-demography of a cooperative-breeding bird*. Princeton University Press, 1984.

G. E. Woolfenden and J. W. Fitzpatrick. *Birds of North America*, chapter Florida Scrub-Jay *(Aphelocoma coerulescens)*. The Academy of Natural Sciences, Washington, D.C., and The American Ornithologists' Union, Philadelphia, Pennsylvania, 1996.

W.Y Yang, J. Novembre, E. Eskin, and E. Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics*, 44(6):725–731, 2012.

W.Y. Yang, A. Platt, C. W.K Chiang, E. Eskin, J. Novembre, and B. Pasaniuc. Spatial localization of recent ancestors for admixed individuals. *Genes, Genomes, Genetics*, 2014. doi:10.1534/g3.114.014274.