A Bayesian algorithm to identify zones of shared evolution in space

Peter Ranacher[a]

[a]*Department of Geography, University of Zurich*

Abstract

TBD

Keywords: TBD

1. Introduction

2. Related Work

3. Methodology

Human traits, such as language, material culture, religion, mythology or genes diffuse in geographical space either through expansion or contact between societies. *Sbayes* analyses the spatial distribution of human traits and identifies zones of shared evolution. The algorithm proposes zones – geographical areas where traits might have diffused – (Section 3.1) and evaluates these against evidence from the data (Section 3.2). Thus, *Sbayes* computes a posterior distribution of zones where human traits are likely to have diffused (Section 3.3).

3.1. Proposing zones of shared evolution

Sbayes connects the spatial locations of all societies in the data by a Delaunay triangulation. The resulting auxiliary graph creates an adjacency between the societies making it easy to propose spatial zones in an efficient and fast way. *Sbayes* proposes an initial random zone - a connected subgraph in the auxiliary graph consisting of five societies (Figure 1a). Starting from the initial zone, the algorithm generates new random candidate zones by applying one of the following three steps:

- Grow step: A random neighbor is added to the zone (Figure 1b). Neighbors are societies adjacent to the current zone.
- Shrink step: A random node is removed from the zone (Figure 1c). The resulting candidate zone is not necessarily a connected subgraph of the auxiliary graph.
- Swap step: A shrink step followed by a grow step (Figure 1d).

3.2. The likelihood of a zone

Sbayes evaluates the likelihood of a zone in terms of evidence for shared evolution (feature likelihood) and spatial proximity (geo-likelihood). We implement two models :

- The *Generative model (GM)* assesses the likelihood of a zone to result from a common generative process.
- The *Particularity model (PM)* assesses the likelihood of a zone to be *distinctive*.

Feature likelihood

Sbayes requires traits to be binary: a trait is either *present* and can be found in a society, or it is *absent* and cannot be found. We assume that only presence is indicative of a possible shared evolution.

In *GM* ... (Nico).

In *PM*, *Sbayes* evaluates how exceptional the shared evolution in the zone is, taking into account both global probabilities and the local presence of a trait. Let us assume a zone of size n, with the trait f present in k sites. The feature likelihood equals the inverse of the probability of finding a zone which is at least as *distinctive* in terms of evidence for shared evolution (Figure 2, red part of the histogram). Hence, in *PM* the feature likelihood is defined as

$$\mathcal{L}_f^{PM} = \frac{1}{Pr(K \geq k)}, \tag{1}$$

where $Pr(K \geq k)$ is the probability that the trait is present in k societies in a random sample of size n, given the global probability of presence p. In a sense, *PM* emulates the implicit reasoning of a domain expert who attempts to determine the particularity of a sample. If f is hardly present there is

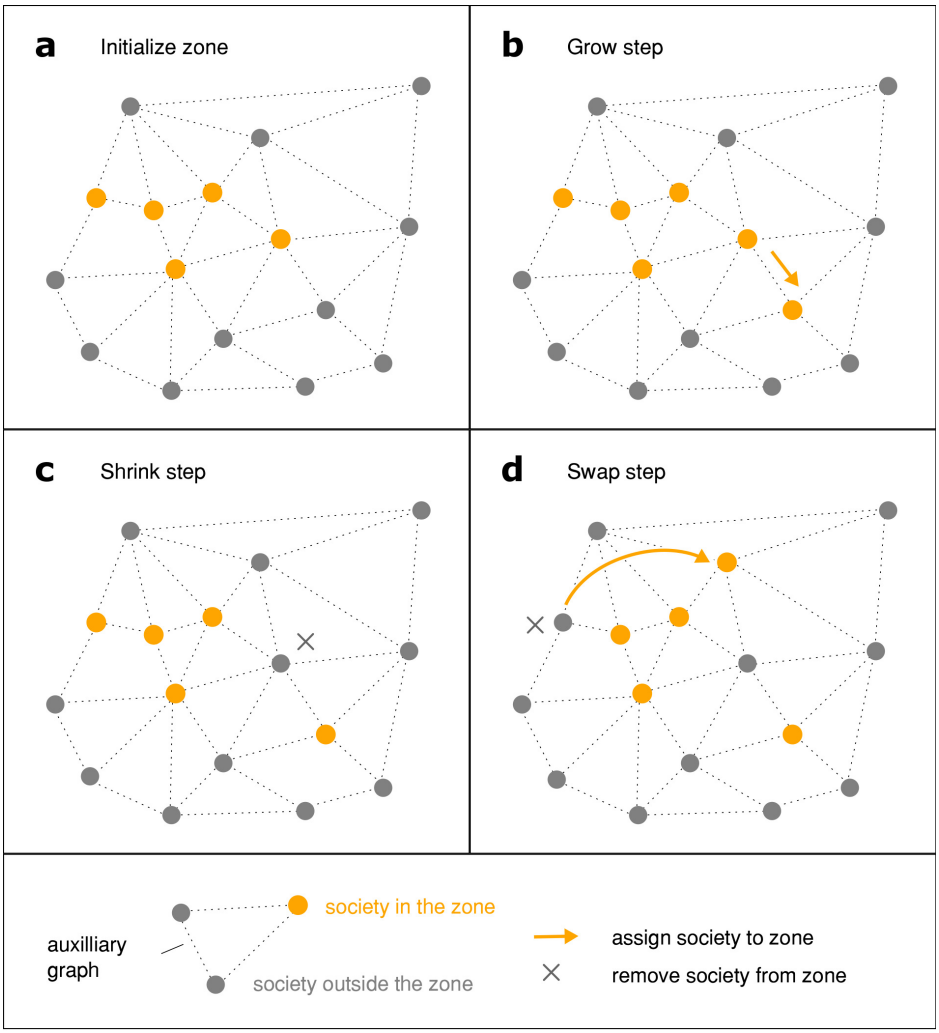

Figure 1: Proposing contact zones

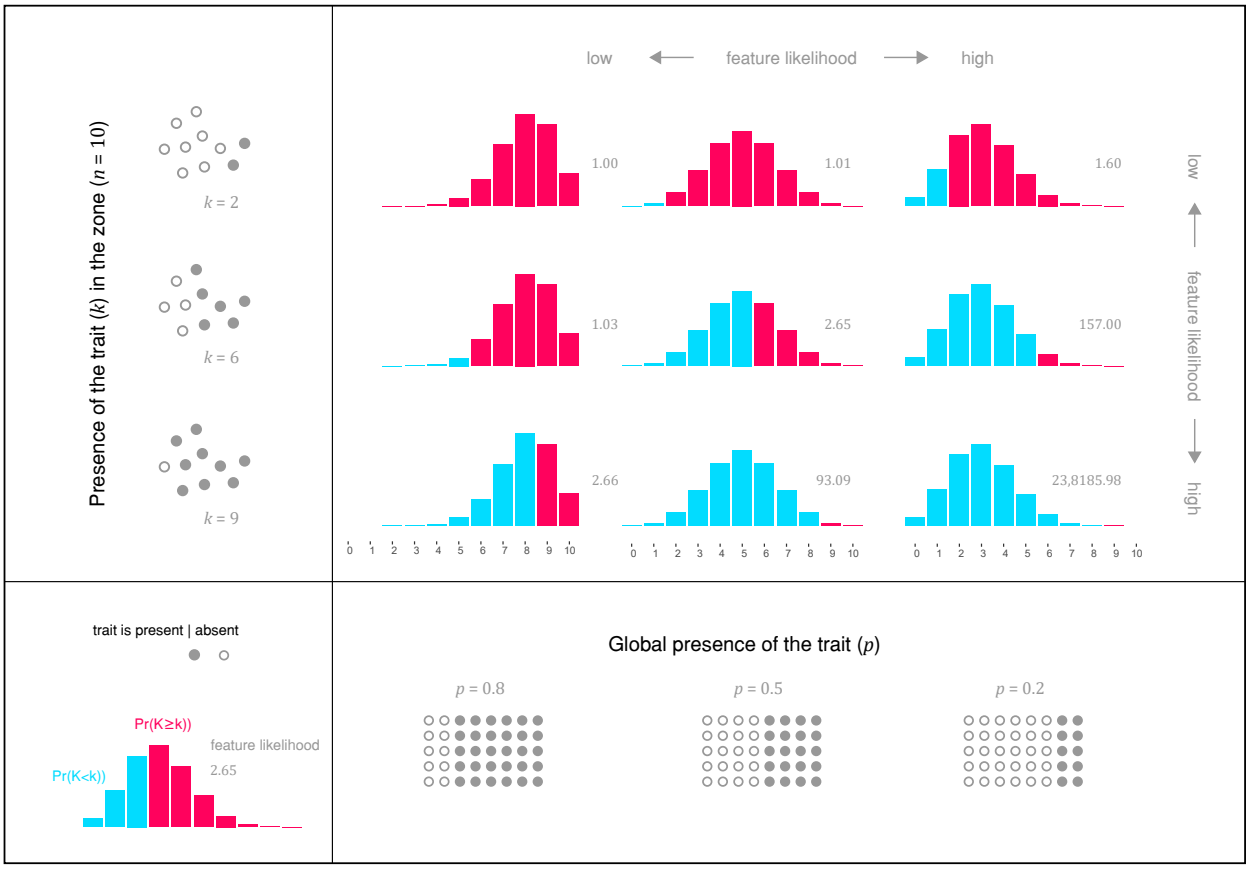

Figure 2: The feature likelihood in the *PM*

little to no evidence for shared evolution (Figure 2, $k = 2$). If f is present the evidence is weaker for traits which are globally present (Figure 2, $k = 9$ and $p = 0.8$) and stronger for traits which are globally absent (Figure 2, $k = 9$ and $p = 0.2$). Simply put, it is not exceptional to find a common trait in a zone, whereas it is exceptional to find a rare one.

Usually, data will consist of several traits f_1, f_2, ... f_m. Each trait represents one piece of evidence for shared evolution. The overall feature likelihood is simply the product of the feature likelihood of all m independent traits:

$$\mathcal{L}_F = \mathcal{L}_{f_1} \cdot ... \cdot \mathcal{L}_{f_m} \tag{2}$$

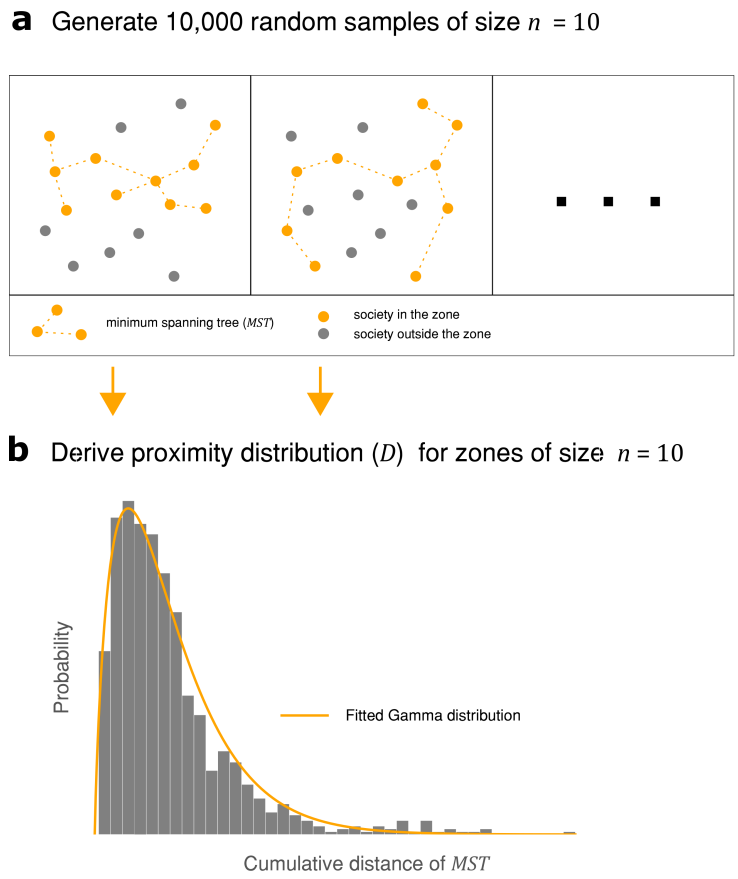

Figure 3: The proximity distribution for zones of size n=10

Geo-likelihood

The geo-likelihood quantifies the spatial proximity of the societies in a zone.

In *GM*, ... (Nico)

In *PM*, *Sbayes* finds the minimum spanning tree (*MST*) of the zone and evaluates it against the proximity distribution (D). D is derived empirically from the data for each sample size n: First, 10,000 random zones of size n are generated. For each zone the *MST* is found and its cumulative distance is computed (Figure 3, a). Finally a gamma distribution is fitted to the resulting sample, which yields D (Figure 3, b). To compute the geo-likelihood *Sbayes* finds the *MST* of the societies in the zone, retrieves its cumulative distance d, and evaluates it against the proximity distribution for zones of

size n. Thus, in PM the geo-likelihood is defined as

$$\mathcal{L}_{geo}^{PM} = Pr(D \leq d). \tag{3}$$

The MST connects all societies in a zone minimizing the distance between these. It reflects the minimum effort necessary for the societies in the zone to diffuse or to have contact. The geo-likelihood shows how great this effort is, taking into account the global proximity in the data. The geo-likelihood is high for close societies (Figure 4, a and b) and low for distant ones (Figure 4, c and d). Since the geo-likelihood derives from a cumulative distance measure it is robust to single outliers.

3.3. Markov chain Monte Carlo Sampling

Sbayes computes the posterior probability for shared evolution in a zone, according to Bayes' theorem:

$$Pr(\text{zone}|\text{data}) = \frac{Pr(\text{data}|\text{zone}) \cdot Pr(\text{zone})}{Pr(\text{data})}. \tag{4}$$

In equation 4, $Pr(\text{zone})$ is the prior probability for shared evolution in a zone. We assume that $Pr(\text{zone})$ is uniform, i.e. a priori, shared evolution is equally probable in each zone. $Pr(\text{data}|\text{zone})$ is the combined feature and geo-likelihood:

$$Pr(\text{data}|\text{zone}) = \mathcal{L} = {\mathcal{L}_F}^{w_F} \cdot {\mathcal{L}_{geo}}^{w_{geo}}, \tag{5}$$

where w_F and w_{geo} are weights applied to either likelihood. $Pr(\text{data})$ is the marginal likelihood of the model and acts as a normalizing constant of the posterior.

There is no analytical solution for equation 4. *Sbayes* performs Markov-chain-Monte Carlo (MCMC) sampling to calculate a numerical approximation. The MCMC first generates a random initial zone z and evaluates the likelihood $\mathcal{L}(z)$. Then, a candidate zone z' is proposed, by either shrinking, growing or swapping, and the candidate likelihood $\mathcal{L}(z')$ is computed. The MCMC evaluates the acceptance probability

$$\alpha = \frac{\mathcal{L}' \cdot Q(z|z')}{\mathcal{L} \cdot Q(z'|z)}, \tag{6}$$

where $Q(z|z')$ is the proposal probability to move from the current zone z to the candidate zone z' and $Q(z'|z)$ is the corresponding back probability

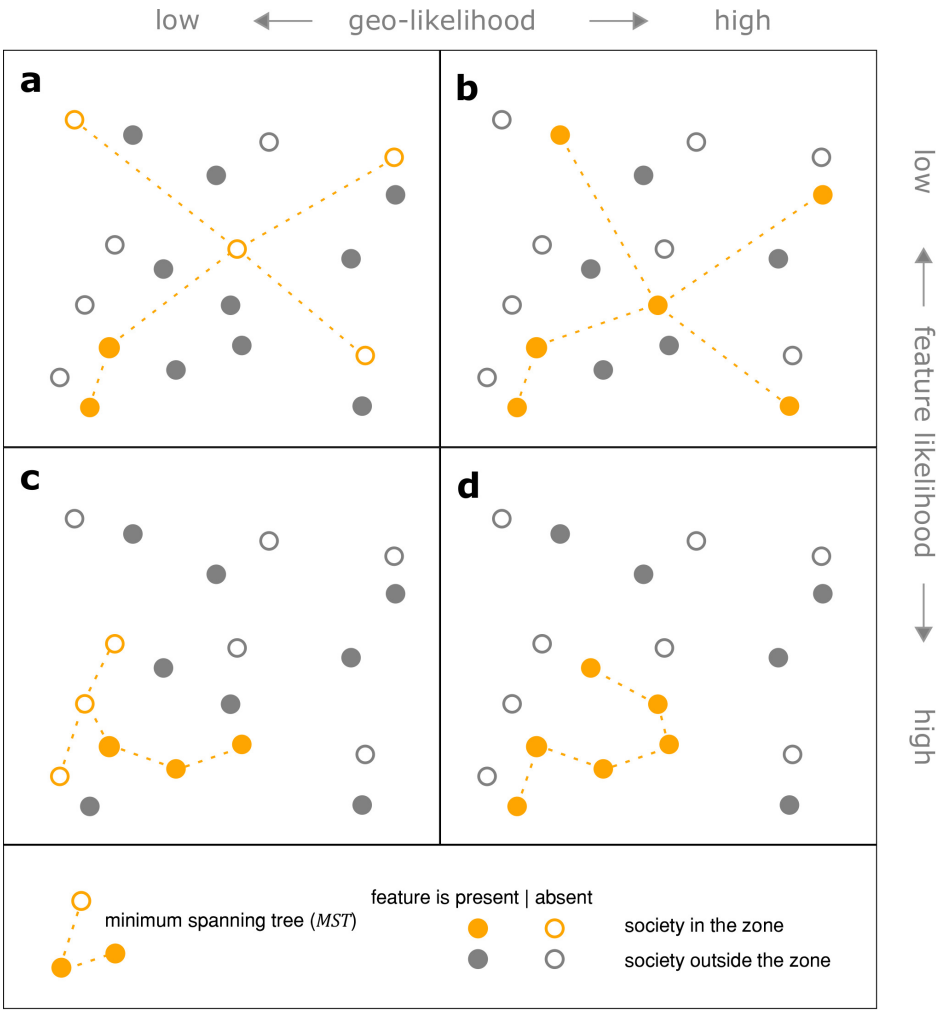

Figure 4: Feature and Geo-likelihood

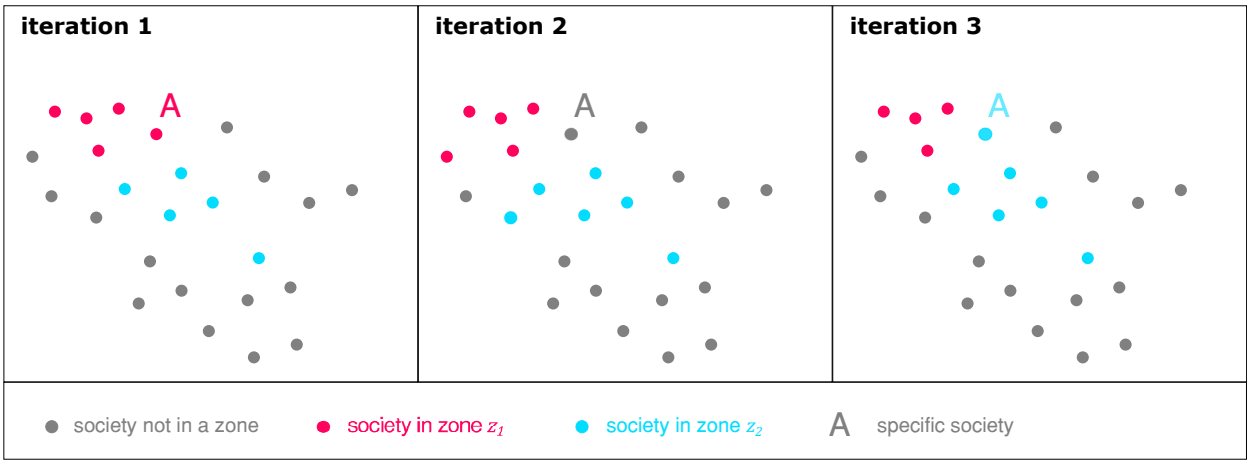

Figure 5: Parallel analysis

to move from z' to z. A move to the candidate zone is accepted with probability min $(1, \alpha)$ and rejected otherwise. The MCMC repeatedly proposes new candidate zones, evaluates the acceptance probability and accepts or rejects a move. The resulting Markov chain gives an approximation of the posterior probability for zones of shared evolution in space. For numerical convenience the MCMC uses the log-likelihood rather than the likelihood *(Missing: explain briefly why this does not matter).*

3.4. Parallel analysis

The diffusion of traits in space is a complex process in which expansion and contact might create multiple and possibly overlapping areas of shared evolution. Hence, it is not realistic to observe a posterior probability with a single mode resulting from one zone of shared evolution. *Sbayes* uses parallel zones to explore multiple modes in the data and to identify complex and possibly intertwined patterns of shared evolution.

In the parallel analysis, the MCMC initializes l individual Markov chains, to evaluate l parallel zones. Each chain explores the data, i.e. proposes and evaluates zones independently. However, a society can only belong to a single zone at a time. In Figure at iteration 1, the society A belongs to z_1. Only after A is removed from z_1 (iteration 2) it can be assigned to z_2 (iteration 3). The number of parallel zones is fixed.

Missing: a method to estimate an appropriate number of parallel zones (l)

- Compute the marginal likelihood of each chain
- Based on the marginal likelihood define a statistical procedure to estimate l (AIC, BIC, scree plot, ...)

4. Test and evaluation

In this section we test *Sbayes* on simulated data and evaluate its performance.

Missing: a structured approach to test the algorithm on simulated data

- How do we simulate zones?

  Draw zones by hand? Make an algorithm propose zones? Curdin? Nico?

- Single mode

  We simulate shared evolution in space and hope that the algorithm finds it. We can vary:

  - the approach (GM or PM)
  - the size of the simulated zones (i.e. number of societies)
  - their shape (i.e. elongated, compact, tree-like, ...)
  - the intensity in terms of how many features are indicative of a shared evolution
  - the intensity in terms of how many societies in the simulated zones are indicative of a shared evolution

- Multi-mode (parallel analysis)

  - all of the above
  - the number of simulated zones
  - the degree of how much the zones are intertwined

- Define a test statistics to evaluate the performance of the algorithm (Curdin?), either quantitative or qualitative

5. Empirical Analysis

Missing: Rik and Manuel, should we test the algorithm on the Amazon data, on the Asian data or on both? - It is strongest if we do both - Rik

This depends a bit on the consecutive papers that we plan to publish and the story that we might or might not want to anticipate.

6. Discussion

- Explain how the marginal likelihood of a model can be used for spatial inference.
- Compare the algorithm to DBSCAN, Structure, TESS, the geo-model in BayesTraits ...
- Explain the difference between GM and PM.
- Explain how the results of the algorithm depend on the given set of societies and the given spatial extent.