Hidden spatial clusters – and how to find them

Peter Ranacher[1,2,3][0000−0002−8680−4063]
and Nico Neureiter[1,2,3][0000−0002−3719−2259]

[1] University Research Priority Program (URPP) Language and Space, University of Zurich, Zurich, Switzerland
[2] Department of Geography, University of Zurich, Zurich, Switzerland
[3] Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich, Switzerland
peter.ranacher@geo.uzh.ch

Abstract. Spatial clustering finds groups of neighbouring objects with similar attributes, revealing patterns of spatial interaction and influence. However, not all similarities in spatial data are due to areal effects. Confounders can mask similarities and hide the spatial signal in the data. We see this, for example, in cultural evolution where language similarities due to shared ancestry mask similarities due to contact and interaction. In this article, we present `sBayes` a Bayesian mixture model for spatial clustering in the presence of confounders. `sBayes` learns which similarities in a set of spatial point objects are explained by confounding effects and assigns objects to clusters based on the remaining similarities in the data. We introduce the algorithm to a geographic audience on the example of a fictional mobility analysis. We discuss how `sBayes` can be applied to ecology, health, and economy problems, revealing hidden geographic structures and patterns.

Keywords: spatial clustering · confounders · Bayesian modelling

DOI: https://doi.org/10.25436/E2G01J

1 Introduction

In the URPP Language and Space at the University of Zurich, geographers and linguists study language evolution in space. One of our research problems is particularly challenging: language contact. When speakers of different languages interact, they likely exchange properties and their languages become more similar. Spatial clustering methods promise to recover these traces, revealing geographic contact areas and past human interaction. However, contact areas are notoriously difficult to find. Only a few and usually weak language similarities come from contact, while more and stronger ones result from either universal preference or common ancestry. English, German, French and Italian, for example, belong to the Indo-European language family. They have all inherited similar properties from their common ancestor, which overshadow potential contact signals.

Standard clustering algorithms group objects based on their attribute similarity, their spatial proximity, or both (4). However, in the presence of confounders, clustering might yield undesired results. In the case of language contact, clustering likely returns the language families in the data, missing out on the weaker contact signal. The contact areas are hidden spatial clusters whose similarity is masked by the stronger similarities of shared ancestry.

We developed `sBayes`, a Bayesian algorithm for spatial clustering in the presence of confounders (6). The algorithm learns which similarities in a set of data result from confounders and which come from areal effects. Initially, we designed `sBayes` to find areas of cultural contact, but we believe that it can be applied to a broader range of spatial clustering problems, revealing hidden spatial patterns. In this paper, we present a generalized version of the `sBayes` algorithm to a geographic audience on the example of a fictional mobility analysis.

2 Hidden spatial clusters

Among many other factors, we can imagine that the place of residence influences individual mobility behaviour. In areas with easy access to affordable, safe, and regular bus and train services, citizens might be inclined to use public transport. At the same time, they might opt for private cars in places where public transport is poorly developed. We could imagine exploring the role of residence on individual mobility in a survey similar to that in Table 1. We assume that answers in the survey are fixed-choice such that each question can be encoded as a categorical variable, with each answer being one of the applicable categories. For example, the question ”Do you own a car?” is encoded as a binary variable with categories ”yes” and ”no”. The question ”How often do you use public transport?” is encoded as a multinomial variable with categories ”daily”, ”weekly”, ”monthly”, ”yearly”, ”never”. Each participant in the sample has a vector of variables describing their mobility behaviour. To simplify the subsequent visualizations, we map the participant to a colour gradient reflecting the main variation in mobility behaviour. In Figure 1, yellow indicates a participant with an overall preference for public transport, e.g. participant A, while dark purple indicates a preference for private motorized transport, e.g. participant B.

Question	**Part. A**	**Part. B**	···
Do you own a car?	no	yes	
What is your preferred means of transport?	train	car	
How often do you use public transport?	daily	never	
Do you have a half-fare travel card?	yes	no	
Do you do your daily shopping by car?	no	yes	
⋮			

Table 1. Excerpt of a fictional mobility survey on public and private transport use.

Imagine three different age groups in the sample, young, working-age, and elderly participants (Figure 1). Each age group has a different mobility behaviour.

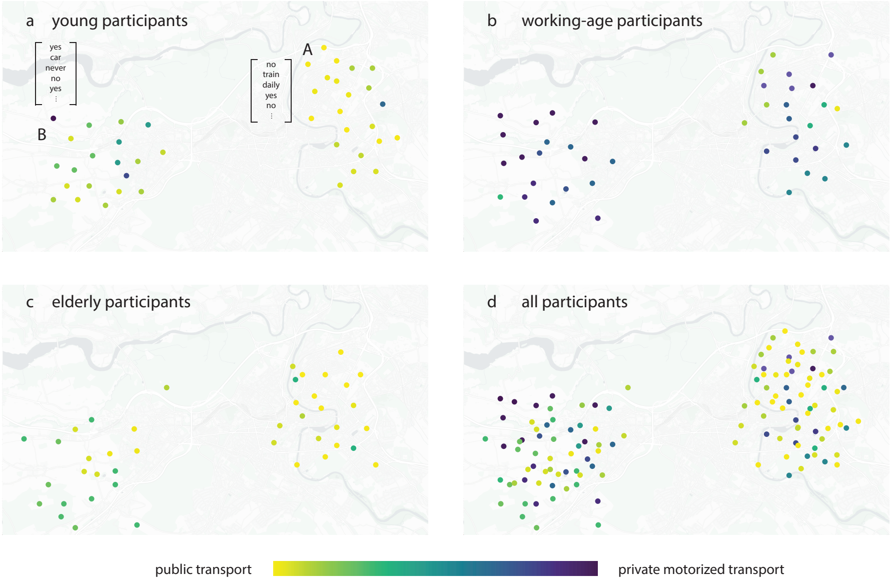

Fig. 1. The spatial clustering of similar mobility behaviour in the three age groups young (a), working-age (b) and elderly (c), is masked by the tendencies in the different groups. Hence, no clear clustering is visible in the joint data (d). Accounting for the confounding effect of age can uncover the hidden spatial clusters.

The young (a) and elderly participants (b) prefer public transport, while the working-age participants (c) lean towards motorized private transport. In each age group, participants with similar mobility behaviour also tend to cluster in space. In the Western region, participants across age groups prefer motorized private transport, while in the East, they prefer public transport. However, in the entire sample (d) the spatial signal disappears, because the confounder, age, masks the clustering. We call these areas with masked similarities *hidden spatial clusters.*

3 sBayes

sBayes is a Bayesian mixture model for finding hidden spatial clusters. For an introduction to Bayesian data analysis and modelling, see the textbooks by Gelman et al. (3) and McElreath (5). Let us assume a set of spatial point objects O with categorical features F. Each feature $f \in F$ takes one of N_f mutually exclusive states:

$$\mathcal{S}_f = \{s_1, ..., s_{N_f}\}, \tag{1}$$

In our example, the spatial objects are participants and their locations of residence. The features capture the individual mobility behaviour. The feature

$f_{\text{preferred}}$ could indicate the preferred means of transport with states

$$\mathcal{S}_{\text{preferred}} = \{\text{car}, \text{train}, \text{bus}, \text{tram}\}.$$

sBayes aims to identify the relevant effects to predict why feature f of object o has state s. Specifically, sBayes proposes an areal effect and one or several discrete confounding effects. In the mobility example, the confounder is age demographics. Participants belong to an age group $a(o)$, which influences their mobility behaviour. For each confounder and the areal effect, sBayes models a likelihood function. In the example, P_{age} is the likelihood that s is preferred because of age, P_{areal} is the likelihood that s is preferred in the areal cluster $Z(o)$.

sBayes then models each feature as coming from a distribution that is a weighted mixture of the areal effect and the confounders. The unknown weights — in the example w_{areal} and w_{age} — quantify the contribution of each effect. For a single participant o who belongs to the demographic age group $a(o)$ and cluster $Z(o)$, the following mixture likelihood gives the probability of feature f being in state s:

$$\begin{aligned} P(X_{o,f} = s|\mathcal{Z}, w, \beta, \gamma) = w_{\text{age},f} \cdot P_{\text{age}}(X_{o,f} = s|\beta_{f,a(o)}) \\ + w_{\text{areal},f} \cdot P_{\text{areal}}(X_{o,f} = s|\gamma_{f,Z(o)}) \end{aligned} \tag{2}$$

The mixture components — P_{age} and P_{areal} — are categorical distributions parameterised by probability vectors $\beta_{f,a(o)}$ and $\gamma_{f,Z(o)}$. That is, the probability of observing state s in feature f is $\beta_{f,s}$ if it is the result of age demographics and $\gamma_{f,Z(o),s}$ if it is the result of an areal effect in $Z(o)$. While the assignment to age groups is fixed – each participant belongs to one demographic age group – the assignment to areal clusters is inferred from the data. sBayes allows for multiple clusters $\mathcal{Z} = \{Z_1, ..., Z_K\}$, each with their own set of areal probability vectors. A detailed explanation of all mixture components together with examples can be found in the Supporting Information of the original publication.

The mixture model combines the weighted likelihood for age demographics and areal effects across all objects. The model has parameters $\Theta = \{\mathcal{Z}, \beta, \gamma, w\}$, which are evaluated against the data D – in the example this is the mobility behaviour of all participants. The likelihood of the whole model is the joint probability of the observed feature values $D_{o,f}$ over $o \in O$ and features $f \in F$, given Θ:

$$P(D|\Theta) = \prod_{o \in O} \prod_{f \in F} P(X_{l,o} = D_{l,o}|\Theta) \tag{3}$$

Since sBayes is a Bayesian model, each of its parameter needs a prior distribution. sBayes uses Dirichlet priors for the mixture weights and the probability vectors of the categorical distributions, and purpose-built geo-priors for the assignment of objects to clusters. The original paper gives a detailed explanation of each of the priors $P(\Theta)$ and in the discussion, we will explore the geo-prior in more detail. The posterior of the model is proportional to the likelihood times the prior:

$$P(\Theta|D) \propto P(D|\Theta) \cdot P(\Theta). \tag{4}$$

sBayes uses a Markov chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution. For each new sample, the algorithm either (a) randomly assigns objects to one of *k* clusters (b) updates the probability vectors for the areal effect, e.g. setting a strong preference for public transport in one of the clusters, (c) updates the probability vectors for the confounding effects, e.g. setting a high probability for owning a car in the working-age population or (d) alters the weights for either effect. After each update, the algorithm evaluates the likelihood. A sample has a high likelihood, if

- the estimates for the areal and the confounding effect fit the data,
- the entropy for the areal effect is low and the participants in a cluster are similar across many features,
- the areal effect differs from the confounding effect, e.g. age demographics do not explain the similarity in the cluster.

The algorithm accepts a move to a new sample with Metropolis-Hastings probability. More details on the sampling procedure can be found in the original publication, together with two case studies and detailed simulation studies (6). The sBayes algorithm is available on GitHub (github.com/derpetermann/sbayes), both with fixed confounders for finding contact areas in cultural data (branch master), and customizable discrete confounders for finding hidden spatial clusters in general (branch geo_sbayes).

4 Discussion

This paper presented a generalized version of sBayes, an algorithm to find hidden spatial clusters in categorical multivariable geographic data. sBayes is an interpretable machine learning model. In our idealized example, the feature variation reduces to a single dimension, with petrolheads and train aficionados on either end of the spectrum. In actual data, we likely find variation along several axes, in which case the labelling of objects and the interpretation of clusters is less straightforward. Besides clusters, sBayes returns weights and feature distributions. The weights indicate how important each feature was to delineate the cluster, the feature distribution captures the intra-cluster propensity, allowing for interpretation and labelling. The Bayesian mixture model yields a posterior distribution, which reflects the robustness of clustering. For a strong and concentrated spatial signal, the posterior distribution is narrow, such that the clusters in each sample contain the same objects. For a diffuse spatial signal, the assignment of objects to clusters varies across posterior samples, reflecting the uncertainty in the data.

Geographic clustering should explicitly consider spatial neighbourhood and contiguity to find clusters. In sBayes, the *geo-prior* addresses this issue. The geo-prior connects all point objects in a cluster with a linkage criterion and then evaluates spatial coherence. In the original publication, we used the minimum spanning tree (MST) as linkage criterion, and we let the geo-prior decrease exponentially with the average distance in the MST. In this case, the geo-prior regularizes, preferably reporting spatially compact clusters. However, the likelihood might still overwhelm the prior if the similarity in non-compact clusters is strong enough. The geo-prior is much more flexible and can accommodate different spatial contiguity and neighbourhood scenarios. For example, when the spatial influence is known not to exceed a given threshold, as is the case for noise or pollution, we might want to set the probability of distant points to occur together in a cluster to zero. In this case, the linkage criterion would connect all pair-wise objects. The geo-prior would evaluate the maximum distance against a distribution truncated at the threshold distance and exclude all spread-out clusters. There is no prior on the number of clusters, k, in the model. Instead, sBayes uses methods from model selection to find a suitable k that avoids overfitting but captures the variance in the data.

It is a standard practice in geographic regression analysis to visualize the residuals of a model on a map (1). Suppose the residuals are spatially autocorrelated, such that positive and negative residuals occur together in space. Autocorrelated residuals either point at a distance-related interaction between the objects or a misspecified model where a critical, spatially structured predictor is missing (2). Our approach differs from classical spatial residual analysis, but we can make a similar analogy. The mixture model assumes that all relevant non-spatial predictors of a phenomenon are available as confounders. It assigns objects to clusters based on the remaining similarities in the data. Consequently, the model either reveals spatial interaction or a spatially structured predictor. In the mobility example, the clusters could identify regions with different accessibility to public transport or point at a distance-based interaction in space: spatially close participants interact with each other and reinforce their views on mobility. Only context can tell which of the two the clusters reflect. In case the clustering does not show a spatial pattern, the algorithm has revealed a non-spatially structured confounding effect. In the mobility example, this could be the participants' profession or income.

So far, we have presented mobility behaviour as a potential new application for sBayes, but the idea of clustering in the presence of confounders is very general. Hence, many areas of application could benefit from such an approach. We will briefly discuss potential applications in ecology, public health and economics. The question at the heart of ecology is how we can explain the spatial distributions of community composition and biodiversity patterns on the planet. sBayes could find hidden spatial clusters in these distributions accounting for known confounding effects of climate, geology, soil or human influence. Based on health records and demographic data, sBayes could be used to detect clusters of high or low incidence of different diseases, accounting for known confounders of age or preexisting conditions of patients. The influence of different policies, investments and institutions on economic activity is a central question in economics. Using data of economic activity in a municipality (e.g. number of businesses, share of different industries, average revenue) to detect clusters of surprisingly low or high economic activity, given predefined confounders like certain policies and regulations in the municipality.

Bibliography

[1] Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V., Pebesma, E.J.: Applied spatial data analysis with R, vol. 747248717. Springer (2008)

[2] F. Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., et al.: Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography **30**(5), 609–628 (2007)

[3] Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian data analysis. Chapman and Hall/CRC (1995)

[4] Liu, Q., Deng, M., Shi, Y., Wang, J.: A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. Computers & Geosciences **46**, 296–309 (2012)

[5] McElreath, R.: Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC (2018)

[6] Ranacher, P., Neureiter, N., van Gijn, R., Sonnenhauser, B., Escher, A., Weibel, R., Muysken, P., Bickel, B.: Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact. Journal of The Royal Society Interface **18**(181), 20201031 (2021). https://doi.org/10.1098/rsif.2020.1031