

**Original citation:**

Zanella, Giacomo. (2015) Random partition models and complementary clustering of Anglo-Saxon place-names. *The Annals of Applied Statistics*, 9 (4). pp. 1792-1822.

**Permanent WRAP url:**

<http://wrap.warwick.ac.uk/76100>

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions. Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher statement:**

Link to final version : <http://dx.doi.org/10.1214/15-AOAS884>

**A note on versions:**

The version presented in WRAP is the published version or, version of record, and may be cited as it appears here. For more information, please contact the WRAP Team at: [publications@warwick.ac.uk](mailto:publications@warwick.ac.uk)



<http://wrap.warwick.ac.uk/>

## RANDOM PARTITION MODELS AND COMPLEMENTARY CLUSTERING OF ANGLO-SAXON PLACE-NAMES

BY GIACOMO ZANELLA<sup>1</sup>

*University of Warwick*

Common cluster models for multi-type point processes model the aggregation of points of the same type. In complete contrast, in the study of Anglo-Saxon settlements it is hypothesized that administrative clusters involving complementary names tend to appear. We investigate the evidence for such a hypothesis by developing a Bayesian Random Partition Model based on clusters formed by points of different types (complementary clustering).

As a result, we obtain an intractable posterior distribution on the space of matchings contained in a  $k$ -partite hypergraph. We apply the Metropolis–Hastings (MH) algorithm to sample from this posterior. We consider the problem of choosing an efficient MH proposal distribution and we obtain consistent mixing improvements compared to the choices found in the literature. Simulated Tempering techniques can be used to overcome multimodality and a multiple proposal scheme is developed to allow for parallel programming. Finally, we discuss results arising from the careful use of convergence diagnostic techniques.

This allows us to study a data set including locations and place-names of 1316 Anglo-Saxon settlements dated approximately around 750–850 AD. Without strong prior knowledge, the model allows for explicit estimation of the number of clusters, the average intra-cluster dispersion and the level of interaction among place-names. The results support the hypothesis of organization of settlements into administrative clusters based on complementary names.

### 1. Introduction.

1.1. *The historical problem.* The starting point of this work is a data set supplied by Professor John Blair of Queen’s College, Oxford. The data set consists of the locations and place-names of 1316 Anglo-Saxon settlements dated approximately around 750–850 AD [data set fully available in [Zanella \(2015a\)](#)]. In the data set there are 20 different kinds of place-names in total. Place-names form an important source of information regarding the Anglo-Saxon civilization and are intensively studied by the historical community [see, e.g., [Gelling and Cole \(2000\)](#) and [Jones and Semple \(2012\)](#)].

---

Received September 2014; revised June 2015.

<sup>1</sup>Supported by EPSRC through a PhD position under the CRiSM grant EP/D002060/1.

*Key words and phrases.* Random partition models, complementary clustering, data association problems, Metropolis–Hastings algorithm, efficient proposal distribution, K-cross function, kernel smoothing, bandwidth, Anglo-Saxon place-names locations.

In particular, the place-names included in this data set are often described as *functional* place-names, as they were probably used to indicate specific functions or features of their corresponding settlements. For example *Burton* is thought to label fortified settlements having a military role, *Charlton* the settlements of the peasants and *Drayton* the settlements dedicated to portage.

Moreover, historians expect the settlements in this data set (especially those having one of the place-names underlined in Table 1) to have been formed approximately at the same time and in the same context (specifically, royal administration in the period c. 750–850). This suggests that there could be some coherence in the distribution of such place-names. In particular, Professor Blair’s hypothesis is that those settlements were not independent units, but rather that they were organized into administrative clusters (or districts) where place-names were used to indicate the role of each settlement within the district. According to this hypothesis, such clusters would tend to involve a variety of complementary place-names in each of them. For example, Figure 1 indicates a plausible administrative cluster made of four settlements, with, for example, a settlement dedicated to military functions (*Burton*) and one dedicated to agriculture (*Carlton*).

The objective of our statistical approach to the study of settlements names and geographical locations is to address the following questions: Is there statistical support for Blair’s hypothesis? What is the typical distance between settlements in the same cluster? How many settlements are clustered together and how many are singletons? Which place-names tend to cluster together? Can we provide a list of those clusters which are more strongly supported by the analysis?

Our intention is to provide a useful contribution to historical research on this topic based on a quantitative approach, bearing in mind the scarcity of textual evidences regarding the Anglo-Saxon period. Since there is a lot of uncertainty and controversy regarding the meaning of place-names, even the apparently obvious ones, we should try to be fairly neutral from the historical point of view, avoiding strong assumptions on the functions of place-names and relationships among them. This will help our statistical analysis to be a genuine contribution to the ongoing historical debate on this topic.

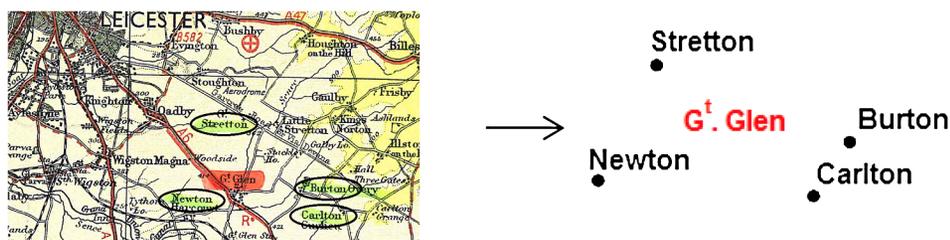


FIG. 1. A cluster of four Anglo-Saxon settlements (highlighted in green and circled) in the region of Great Glen (highlighted in red).

We note that there has already been statistical work related to Anglo-Saxon place-names. In particular, Keith Briggs did various works on this topic (see <http://keithbriggs.info/place-names.html> for a full list). Nevertheless, both the historical questions considered and the statistical methodologies used are substantially different from ours.

*1.2. Modeling approach.* By considering the place-names as marks attached to points, we model our data as the realization of a  $k$ -type point process (also called  $k$ -variate point process), where  $k$  is the number of different place-names available [see Baddeley (2010)]. We can view our problem as a clustering problem based on aggregations of points of different types. In fact, we seek a *complementary clustering*: each cluster may contain at most one settlement for each place-name. This simplifying requirement is motivated by the assumption that each place-name represents a different administrative function (role) within the cluster.

Our intention is to perform explicit inferences on the partition of settlements into clusters. As with hierarchical models, it would be desirable to analyze the data set all at once, so as not to lose statistical power, and also to provide inferences at the single cluster level to facilitate visualization and historical interpretation of the results of the analysis.

We employ Random Partition Models (RPMs), often used in the Bayesian Non-parametric literature [e.g., Lau and Green (2007)], as they permit natural inferences on the cluster partition and they have enough flexibility to allow specification of a useful model for complementary clustering.

Standard approaches for point process cluster modeling, like the Log-Gaussian Cox Processes [see Lawson and Denison (2010), Chapter 3] or the Neyman–Scott model [e.g., Loizeaux and McKeague (2001)], are not appropriate here, as such models usually provide inferences on the cluster centers or on the point process intensity, while we seek explicit inferences on the cluster partition. Moreover, standard cluster methods for marked point process consider the marks as an additional dimension and search for aggregations of points with similar marks. In complete contrast, we seek aggregations of points of different types.

Diggle, Eglén and Troy (2006) seek evidence for repulsion among points of different types in a bivariate spatial distribution of amacrine cells. They use a pairwise interaction model, which has theoretical limitations which prevent its use for clustering. While this approach could be extended to our case by using area-interaction point processes, which can model clustering [Baddeley and van Lieshout (1995)], it would not provide us with explicit estimates of the cluster partition and it would not easily allow complementary clustering specification (at most one point of each type in each cluster).

Multitarget tracking involves the Data Association problem, that is to group together measurements recorded at different time intervals to create objects tracks [e.g., Oh, Russell and Sastry (2009)]. This problem is similar to the problem of

performing complementary clustering of a  $k$ -type point process. In Data Association problems, however, the interest is to find the best association, while we are interested in assessing the strength of clustering and the level of interaction between different place-names, and in quantifying the uncertainty of our estimates. In fact, the modeling aspects we have to be careful about are different from the ones of Data Association problems, while the computational challenges are similar (see Sections 3.6 and 4).

**1.3. Organization of the paper.** In Section 2 we perform preliminary analysis of the data set, testing whether there is a significant clustering interaction between points of different types by using common Spatial Statistics tools such as K-cross functions. In Section 3 we define a RPM for complementary clustering and discuss appropriate prior distributions for the cluster partition [see also Section 3.1 of Zanella (2015b)]. The resulting model leads to an intractable posterior distribution. We express such a posterior in terms of matchings contained in hypergraphs. We thus link the problems of sampling from the posterior and finding the posterior mode to the more classical problems of Data Association and Optimal Assignment. In Section 4 we design a Metropolis–Hastings algorithm to obtain approximate samples from the posterior. We carefully consider the problem of choosing an efficient proposal distribution, we explore the use of Simulated Tempering to overcome multimodality, and we develop a multiple proposal scheme to allow for parallel computation. In Section 5 we analyze the Anglo-Saxon place-name location data with our RPM, using the algorithm of Section 4. The results support the hypothesis of settlements being organized into administrative clusters and give explicit inferences of various quantities of historical interest. Finally, in Section 6 we discuss future directions of research. Supplementary material includes extensive calculations, additional tables and plots, the settlements data set and R codes to perform the data analysis.

**2. Preliminary analysis of the Anglo-Saxon settlements data set.** We describe the Anglo-Saxon settlements data set supplied by Professor John Blair and the data cleaning operations that we carried out. We then perform preliminary analysis on the resulting point pattern using Spatial Statistic tools.

**2.1. Format of the data set.** The data set [available in Zanella (2015a)] is made of 20 different groups, each of which contains the list of settlements having one of the 20 place-names (see Table 1). The historians involved in the project expect the clustering behavior to involve, in particular, 13 of those place-names, indicated in Table 1. We refer to the settlements relative to those 13 place-names as the *reduced data set*, and to all the settlements recorded as the *full data set*. We will perform the analysis on both data sets.

For each settlement the following variables are given: County, place, Parish or Township, grid ref, date of first evidence (see Table 2).

TABLE 1

*Number of settlements in the Anglo-Saxon place-names location data set supplied by Professor Blair. The historians expect the clustering behavior mainly to involve 13 of those place-names (underlined and emboldened in this table). Settlements with less precise locations (third column) are settlements whose location is given with 1 km accuracy, rather than 100 m, or having a more uncertain location (see Section 2.1). The term “couples” (last two columns) refers to multiple records of the same settlements (see Section 2.2 for discussion). The “total number” column refers to the count after merging the couples classified by historians*

| Place-names                     | Total number | # of settlements with less precise location | # of couples (as classified by historians) | # of couples (as classified by proximity) |
|---------------------------------|--------------|---|--|---|
| <b><u>Aston/Easton</u></b>      | 90           | 0   | 1  | 8   |
| Bolton                          | 17           | 1   | 1  | 0   |
| Burh-Stall                      | 29           | 2   | 1  | 0   |
| <b><u>Burton</u></b>            | 108          | 2   | 1  | 7   |
| Centres                         | 46           | 0   | 0  | 0   |
| <b><u>Charlton/Charlcot</u></b> | 98           | 3   | 7  | 1   |
| Chesterton                      | 9            | 0   | 0  | 0   |
| Clæg                            | 84           | 13  | 0  | 5   |
| <b><u>Draycot/Drayton</u></b>   | 55           | 1   | 0  | 2   |
| <b><u>Eaton</u></b>             | 33           | 1   | 1  | 5   |
| <b><u>Kingston</u></b>          | 71           | 1   | 1  | 1   |
| <b><u>Knighton</u></b>          | 26           | 1   | 0  | 0   |
| Newbold                         | 34           | 3   | 1  | 0   |
| <b><u>Newton</u></b>            | 191          | 5   | 4  | 5   |
| <b><u>Norton</u></b>            | 74           | 1   | 8  | 1   |
| <b><u>Stratton</u></b>          | 37           | 0   | 5  | 0   |
| <b><u>Sutton</u></b>            | 101          | 2   | 4  | 5   |
| Tot                             | 77           | 17  | 1  | 1   |
| <b><u>Walton/Walcot</u></b>     | 51           | 4   | 1  | 0   |
| <b><u>Weston</u></b>            | 85           | 3   | 3  | 2   |
| Total                           | 1316         | 60  | 40   | 43  |

TABLE 2

*Data available regarding the first 6 settlements with the name Burton. The acronym DB stands for Domesday Book, compiled in 1086*

| County | Place   | Parish or Township     | Grid ref. | Date of first evidence |
|--------|---------|------------------------|-----------|------------------------|
| BRK    | Bourton | Bourton                | SU 230870 | c. 1200                |
| BUC    | Bierton | Bierton with Broughton | SP 836152 | DB                     |
| BUC    | Bourton | Buckingham             | SP 710333 | DB                     |
| CHE    | Burton  | Burton (T)             | SJ 509639 | DB                     |
| CHE    | Burton  | Burton (T)             | SJ 317743 | 1152                   |
| CHE    | Buerton | Buerton (T)            | SJ 682433 | DB                     |

The locations are expressed through the Ordnance Survey (OS) National Grid reference system. A set of OS National Grid coordinates, like *SU 230870*, identify a  $100\text{ m} \times 100\text{ m}$  square on a grid covering Great Britain. Some locations have just 2 letters and 4 digits (e.g., *SU 2387*) and they identify a  $1\text{ km} \times 1\text{ km}$  square, and some have a letter *c* in front of them (e.g., *c. SU 2387*) to indicate that the location is less accurate (see Table 1 for amounts).

*2.2. Data cleaning and data assumptions.* Our analysis is concerned with place-names (variable “place”) and geographical locations (variable “Grid reference”). We convert the data to a *k*-type point process form as described below. Such a data cleaning process entails historical assumptions on the data set and, thus, we have been guided by the judgment of the subject-specific historians involved in this project in doing so.

*Place-names:* we express the variable “place” as a categorical variable with *k* possible values (i.e., *k* types). By doing so we ignore minor variations in place-names. For example, we consider the settlements of Table 2 as having place-name *Burton*: their actual recorded place-names vary among *Burton*, *Bourton*, *Bierton*, *Buerton*.

Four groups (out of 20) are made up of two subgroups each with similar place-names: *Aston–Easton*, *Charlton–Charlcot*, *Drayton–Draycot* and *Walton–Walcot*. We consider such subgroups to be the same, for example, *Charlton* and *Charlcot* are treated as the same place-name.

*Locations:* we convert OS National Grid coordinates to two-dimensional Euclidean coordinates and each settlement is assumed to be located at the center of the corresponding OS National Grid square.

*“Multiple” records:* it is sometimes indicated in the original data set that some couples (or triples) of settlements, with same place-name and very close locations, have to be considered as multiple records of the same settlement. We replaced such couples (or triples) of settlements with one settlement located at their midpoint. Moreover, there are some other pairs of records having very close locations and the same place-name (see Table 1 for amounts). It is primarily a matter of historical interpretation whether these couples have to be considered as single settlements. We performed the analysis under both hypotheses (keeping them separated and merging them) without seeing significant changes in the results. The analysis presented here is made with those settlements merged together (3 km is the threshold distance below which we identify two records of settlements with the same place-name).

*Observation region  $W$ :* a point processes realization consists of points locations and of the region  $W$  where the points have been observed. Indeed, both the K-cross function analysis of Section 2.3 and the Bayesian analysis of Section 3 will use information about  $W$ . In our case we define  $W$  as Great Britain [coastline obtained from the *mapdata R* package [Becker, Wilks and Brownrigg \(2013\)](#)] cropping the

region where the point process intensity  $g$  falls below a certain threshold, approximately at the borders between England–Scotland and England–Wales. We also added a small buffer zone of 3 km around the region to include the few points that were falling outside the region (e.g., because the coastline has moved or because the location was inaccurate). See Figure 12 for a plot of the region.

**2.3.  $K$ -cross function analysis.** Second moment functions are a useful tool to investigate interpoint interaction [e.g., Chiu et al. (2013)]. In particular, given a multitype point pattern, bivariate (or cross-type)  $K$ -functions provide good summary functions of the interaction across points of different types. The bivariate  $K$ -function  $K_{ij}(r)$  is the expected number of points of type  $j$  closer than  $r$  to a typical point of type  $i$ , divided by the intensity  $\lambda_j$  of the type  $j$  subpattern of points  $\mathbf{x}_j$  [e.g., Baddeley (2010), Section 6]. For testing and displaying purposes we define a single summary function, a multitype  $K$ -function  $K_{\text{cross}}(r)$ , as the weighted average of  $K_{ij}(r)$  for  $i \neq j$ , where the weights are the product of the intensities  $\lambda_i \lambda_j$ .

Classical  $K$ -functions, however, rely strongly on the assumption that the point pattern is stationary, which is not appropriate for our data set. Therefore, we use the inhomogeneous version of the  $K$ -functions, where the contribution coming from each couple of points is reweighted to take into account for spatial inhomogeneity [Baddeley, Møller and Waagepetersen (2000)]. Standard estimates of the inhomogeneous bivariate  $K$ -functions  $\hat{K}_{ij}$  are obtained using the *spatstat* R package [Baddeley and Turner (2005)].

**2.3.1. Null hypothesis testing.** In order to test whether the interaction shown by  $K$ -functions is significant or not, we need to define a null hypothesis (representing no interaction among place-names). Section 8 of Baddeley (2010) describes three classical null hypotheses for multivariate point processes: random labeling (given the locations the point types are i.i.d.), Complete Spatial Randomness and Independence (CSRI, the locations arise from a uniform Poisson point process and the point types are i.i.d.) and independence of components (points of different types are independent). The random labeling and the CSRI hypotheses are unrealistic assumptions for our data set because our point pattern is clearly not stationary and the distribution of place-names is not spatially homogeneous (some place-names are more concentrated in the South, some in the North and so on). The independence of components hypothesis is realistic, but, in order to test it, stationarity of the points pattern is usually assumed. Instead, we define the following no-interaction null hypothesis: each subpattern of points  $\mathbf{x}_j$  is an inhomogeneous Poisson point process [with intensity function  $\lambda_j(\cdot)$  potentially varying over  $j$ ]. Note that a more realistic null hypothesis would also include repulsion among points of the same type. In Section 1 of Zanella (2015b) we implement such a null hypothesis using Strauss point-processes. The results are very similar to the ones presented here and require additional tuning of various parameters.

Given the null hypothesis, we perform the following approximate Monte Carlo test. First, we estimate the intensities  $\lambda_j(\cdot)$  with  $\hat{\lambda}_j(\cdot)$  [see Figures 4 and 5 of Zanella (2015c)] obtained through standard Gaussian kernel smoothing with bandwidth chosen according to the cross-validation method [e.g., Diggle (2003), pages 115–118], and edge correction performed according to Diggle (1985). Second, we sample 99 independent multivariate inhomogeneous Poisson point patterns according to  $\{\hat{\lambda}_j(\cdot)\}_{j=1}^k$ . Finally, we use those samples to plot simulation envelopes and to perform a deviation test with significance  $\alpha = 0.05$  using as a summary function a centered version of the  $L$ -function  $\hat{L}_{\text{cross}}(r) = \sqrt{\frac{\hat{K}_{\text{cross}}(r)}{\pi}}$  for  $r \in (0, r_{\text{max}})$ , with  $r_{\text{max}} = 15$  km. The deviation test [Grabarnik, Myllymäki and Stoyan (2011)] summarizes the summary function with a single value  $D = \max_{r \in (0, r_{\text{max}})} \hat{L}_{\text{cross}}(r) - \mathbb{E}[\hat{L}_{\text{cross}}(r)]$  and compares it to the ones obtained from the 99 simulated samples.

The null hypothesis is rejected for both the full and the reduced data set (see Figure 2). For the reduced data set this provides evidence of a stronger clustering effect. The R code used to perform this test and produce Figure 2 is given in Zanella (2015a). Application of the same deviation test on the bivariate  $L$ -functions  $\hat{L}_{ij}(r)$  provides an indication of which couples of place-names exhibit significant interaction [see Figure 6 of Zanella (2015c)].

The preliminary analysis we just presented indicates a clustering interaction between points of different types. Nevertheless,  $K$ -functions do not provide explicit estimates and quantification of uncertainty for the parameters of interest (including the cluster partition itself). In the next section we develop a more advanced model in order to provide more informative answers to the questions of historical interest. We regard  $K$ -functions as a useful exploratory tool and the fact that they indicate interaction is a motivation to pursue further statistical analysis.

We note that Dr. Stuart Brookes from UCL has already used second moment functions to do some preliminary analysis on the Anglo-Saxon settlements data set presented here (personal communication by Professor John Blair).

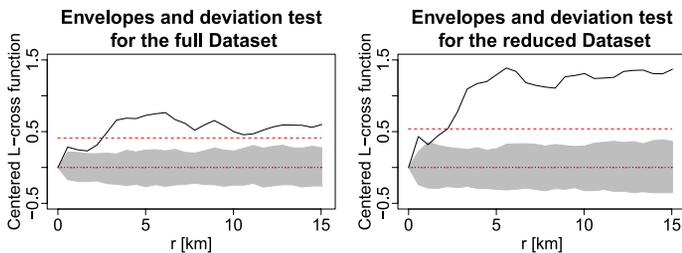


FIG. 2. Black solid lines represent  $\hat{L}_{\text{cross}}(r) - \mathbb{E}[\hat{L}_{\text{cross}}(r)]$  for the observed pattern, the 95% envelopes (gray areas) are obtained using 99 simulated patterns and the red dashed lines indicate the upper deviations. Deviation test: if the (black) solid line rises above the (red) dashed line, then the interaction can be considered significant at significance level  $\alpha = 0.05$ . The values of  $\mathbb{E}[\hat{L}_{\text{cross}}(r)]$  are estimated using independently simulated point patterns generated according to the null hypothesis.

### 3. A Bayesian complementary clustering model.

3.1. *Random Partition Models.* We present Random Partition Models (RPMs) in the specific context of planar  $k$ -type point processes. For more general and detailed discussions see Lau and Green (2007) and Müller and Quintana (2010). Let  $\rho$  be a partition of an ordered set of marked points  $\mathbf{x} = ((x_1, m_1), \dots, (x_{n(\mathbf{x})}, m_{n(\mathbf{x})}))$ , with each  $(x_i, m_i)$  belonging to  $\mathbb{R}^2 \times \{1, \dots, k\}$ . Thus,  $\rho$  can be represented as an *unordered* collection  $\{C_1, \dots, C_{N(\rho)}\}$  of disjoint nontrivial subsets of the indices  $\{1, \dots, n(\mathbf{x})\}$  whose union is the whole set  $\{1, \dots, n(\mathbf{x})\}$ . RPMs are used to draw inferences on the partition  $\rho$  given the observed points  $\mathbf{x}$ . Given  $C_j = \{i_1^{(j)}, \dots, i_{s_j}^{(j)}\}$ , we define  $\mathbf{x}_{C_j} = \{(x_{i_1^{(j)}}, m_{i_1^{(j)}}), \dots, (x_{i_{s_j}^{(j)}}, m_{i_{s_j}^{(j)}})\}$ , for  $j$  running from 1 to  $N(\rho)$ . We call  $\mathbf{x}_{C_j}$  cluster and  $s_j$  the size of the cluster. Given the partition  $\rho$ , we suppose that locations in each cluster  $\mathbf{x}_{C_j}$  are generated independently of locations in other clusters, according to a probability density function  $h_{(s_j, \sigma)}(\cdot)$  depending on  $s_j$  and on a global intra-cluster dispersion parameter  $\sigma$ . Thus, the probability density function of  $\mathbf{x}$  conditional on  $\rho$  and  $\sigma$  is  $\prod_{j=1}^{N(\rho)} h_{(s_j, \sigma)}(\mathbf{x}_{C_j})$ .

We assign independent prior distributions to  $\rho$  and  $\sigma$ . With a slight abuse of notation, we denote them by  $\pi(\rho)$  and  $\pi(\sigma)$ , respectively. We require  $\pi(\rho)$  to be exchangeable with respect to the point indices  $\{1, \dots, n(\mathbf{x})\}$  to reflect the fact that point labels are purely arbitrary and have no specific meaning. We obtain the following expression for the posterior density function:

$$\pi(\rho, \sigma | \mathbf{x}) \propto \pi(\rho)\pi(\sigma) \prod_{j=1}^{N(\rho)} h_{(s_j, \sigma)}(\mathbf{x}_{C_j}).$$

3.2. *Likelihood function.* Given  $\rho$  and  $\sigma$ , each cluster  $\mathbf{x}_{C_j}$  is constructed as follows. First, an unobserved center point  $z_j$  is sampled from the observation region  $W \subseteq \mathbb{R}^2$  with probability density function  $g(\cdot)$ . Then the observed points  $x_{i_1^{(j)}}, \dots, x_{i_{s_j}^{(j)}}$  are given by

$$(3.1) \quad x_{i_l^{(j)}} = z_j + y_{i_l^{(j)}}, \quad l = 1, \dots, s_j,$$

where  $y_{i_l^{(j)}}$  is defined as  $w_{i_l^{(j)}} - s_j^{-1} \sum_{l=1}^{s_j} w_{i_l^{(j)}}$  with  $w_{i_1^{(j)}}, \dots, w_{i_{s_j}^{(j)}}$  being independent bivariate  $N(0, \frac{\sigma^2}{\pi} \mathbb{I}_2)$  random vectors, where  $\mathbb{I}_2$  is the  $2 \times 2$  identity matrix. The variance parametrization  $\frac{\sigma^2}{\pi}$  is chosen so that  $\sigma$  equals the expected distance between two points in the same cluster, independently of the value of  $s_j$ . In fact, if  $x_1$  and  $x_2$  belong to the same cluster, it holds

$$\mathbb{E}[\sqrt{(x_1 - x_2)^\top (x_1 - x_2)}] = \mathbb{E}[\sqrt{(w_1 - w_2)^\top (w_1 - w_2)}] = \sqrt{\frac{\pi}{2}} \sqrt{\frac{2\sigma^2}{\pi}} = \sigma,$$

where  $a^\top a = \sum_{i=1}^2 a_i^2$  for  $a$  in  $\mathbb{R}^2$ , and we used the fact that the Euclidean norm of a two-dimensional  $N(0, \eta^2 \mathbb{I}_2)$  random vector follows the Rayleigh distribution and its mean equals  $\sqrt{\frac{\pi}{2}} \eta$  for  $\eta \geq 0$ .

Finally, the marks  $m_{i_1^{(j)}}, \dots, m_{i_{s_j}^{(j)}}$  are sampled uniformly from the set  $\{\{m_1, \dots, m_{s_j}\} \subseteq \{1, \dots, k\} \mid m_{l_1} \neq m_{l_2} \text{ for } l_1 \neq l_2\}$ .

The resulting likelihood function is

$$(3.2) \quad h_{(s_j, \sigma)}(\mathbf{x}_{C_j}) = \frac{g(\bar{\mathbf{x}}_{C_j}) \prod_{l_1, l_2 \in C_j: l_1 \neq l_2} \mathbb{1}(m_{l_1} \neq m_{l_2})}{\binom{k}{s_j} s_j (2\sigma^2)^{s_j-1}} \exp\left(-\frac{\pi \delta_{C_j}^2}{2\sigma^2}\right)$$

[Section 1 of Zanella (2015d) gives calculations], where  $\bar{\mathbf{x}}_{C_j}$  is the Euclidean barycenter of  $\mathbf{x}_{C_j}$  and  $\delta_{C_j}^2 = \sum_{i \in C_j} (x_i - \bar{\mathbf{x}}_{C_j})^\top (x_i - \bar{\mathbf{x}}_{C_j})$ .

Here we treat  $g(\cdot)$  as a known function. For the purposes of data analysis we will replace  $g$  with an estimate using Gaussian kernel smoothing [see, e.g., Figure 4 of Zanella (2015c)] with bandwidth chosen according to the cross-validation method Diggle [(2003), pages 115–118] and edge correction performed according to Diggle (1985). Note that this replacement commits us to the use of a data-driven prior.

REMARK 1. Given the heterogeneity in the number of settlements across different place-names, the assumption of the marks being sampled uniformly seems not to be very realistic. In Zanella (2015b) we propose an empirical Bayes approach to include nonuniformity of marks in the model while keeping the computation feasible and we present inferences under that assumption. Here we retain the uniform marks assumption for simplicity and because the two approaches produce similar inferences. Moreover, the inferences with the uniform marks assumption are more conservative [see Zanella (2015b)] and therefore preferable in this context.

REMARK 2. This model does not constrain  $x_{i_l^{(j)}} = z_j + y_{i_l^{(j)}}$  to lie in the observation region  $W$ . To make the model more realistic, one could condition the distribution of  $y_{i_l^{(j)}}$  in (3.1) on  $z_j + y_{i_l^{(j)}} \in W$  (which would be an additional form of edge correction). Nevertheless, in our application the density function  $g$  is not concentrated on the borders and the values of  $\sigma$  are small (below 10 kilometers) compared to the size of  $W$ . Therefore, most correction terms would be negligible. Moreover, computing a correction term for each center point  $z_j$  would result in a consistent additional computational burden for each step of the Markov chain Monte Carlo (MCMC) algorithm in Section 4. Therefore, we avoid such correction terms here. Note that, since such correction terms would increase the probability of points being clustered, this approximation has a conservative effect.

3.3. *Prior distribution on  $\sigma$ .* History and context suggest some considerations regarding the expected intra-cluster dispersion (in particular,  $\sigma$  between 3 and 10 km). For example, a basic consideration is that settlements of the same cluster needed to be at no more than a few hours walking distance in order for the inhabitants of the settlements to interact administratively and politically. Nevertheless, we prefer not to impose strong prior information on  $\sigma$ , as this gives us the opportunity to see whether our study of geographical location is in accordance with available contextual information. We use a flat uniform prior for  $\sigma$ , as, for example, it is recommended in Gelman (2006), Section 7.1:

$$\sigma \sim \text{Unif}(0, \sigma_{\max}).$$

We set  $\sigma_{\max} = 50$  km. Given the historical context, such an upper bound for  $\sigma$  constitutes a safe and conservative assumption. We tested other values of  $\sigma_{\max}$ , namely, 20 and 100 km, and the inferences presented in Section 5 were not sensible to such changes, which is in accordance with Gelman (2006), Section 2.2.

3.4. *Prior distribution on  $\rho$ .* We need to model a partition made up of many small clusters. In fact, each cluster can contain at most  $k$  points (one for each color), and the historians expect most of the original clusters to have had fewer than 6 settlements. Common RPMs usually result in clusters with many data points each, and therefore do not seem to be appropriate to our case (see, e.g., Remark 3). We now define a prior distribution  $\pi(\rho)$  designed for situations where each cluster can have at most  $k$  points, with  $k$  small compared to the number of points  $n$ .

3.4.1. *Poisson model for  $\pi(\rho)$ .* The number of clusters  $N(\rho)$  follows a Poisson distribution with mean  $\lambda$  and each cluster size  $s_j$  is sampled from  $\{1, \dots, k\}$  according to a probability distribution  $\mathbf{p}^{(c)} = (p_1^{(c)}, \dots, p_k^{(c)})$ . Note that in such a model the (unobserved) point process of centers  $\{z_1, \dots, z_{N(\rho)}\}$  is a Poisson point process with intensity measure  $\lambda g(\cdot)$  and the number of observed points need not equal  $n$ . Conditioning on observing  $n$  points, the induced prior distribution on  $\rho$  is  $\pi(\rho|\lambda, \mathbf{p}^{(c)}) \propto \prod_{j=1}^{N(\rho)} \lambda p_{s_j}^{(c)}$ . We assign the following conjugate priors to  $\lambda$  and  $\mathbf{p}^{(c)}$ :

$$\lambda \sim \text{Gamma}(k_\lambda, \theta_\lambda), \quad \mathbf{p}^{(c)} = (p_1^{(c)}, \dots, p_k^{(c)}) \sim \text{Dir}(\alpha_1^{(c)}, \dots, \alpha_k^{(c)}).$$

Combinations of the following choices of hyperparameters did not change the posterior significantly:  $k_\lambda = 100, 300, 600$ ;  $\theta_\lambda = 0.5, 1, 3$  and  $(\alpha_1^{(c)}, \dots, \alpha_k^{(c)}) = (1/k, \dots, 1/k), (1, \dots, 1)$  and  $(1, 1/(k-1), \dots, 1/(k-1))$ . In the data analysis of Section 5 we set  $k_\lambda = 300, \theta_\lambda = 1$  and  $(\alpha_1^{(c)}, \dots, \alpha_k^{(c)}) = (1/k, \dots, 1/k)$ .

REMARK 3. In the RPMs literature it is common to assign a Dirichlet Process (DP) prior to  $\rho$ , which is  $\pi(\rho|\theta) \propto \prod_{j=1}^{N(\rho)} \theta(s_j - 1)!$ . The parameter  $\theta$  is often called a concentration parameter and can be either fixed or random. A DP prior

(conditioning on having no cluster with more than  $k$  points) would be equivalent to the Poisson model with fixed  $\mathbf{p}^{(c)}$  given by  $p_l^{(c)} = \frac{(l-1)!}{\sum_{l=1}^k (l-1)!}$ , for  $l = 1, \dots, k$ . Such a choice would enforce most clusters to have almost  $k$  points, and thus is not appropriate to this context where we expect most clusters to be smaller.

REMARK 4. In Zanella (2015b) we describe an alternative model for  $\pi(\rho)$  based on the Dirichlet-Multinomial distribution rather than the Poisson one. Although the inferences we obtain from the two models are almost equivalent, the Poisson model is preferable because its posterior distribution factorizes over clusters and thus allows for cheaper computation.

3.5. *Model parameters and Posterior Distribution.* The model presented above results in the following unknown elements:

$$(\rho, \sigma, \mathbf{p}^{(c)}, \lambda) \in \mathcal{P}_n \times \mathbb{R}_+ \times [0, 1]^k \times \mathbb{R}_+,$$

where  $\mathcal{P}_n$  is the set of all partitions of  $\{1, \dots, n\}$ . Figure 3 provides a graphical representation of the underlying conditional independence structure. Given the prior and likelihood distributions described in Sections 3.2, 3.3 and 3.4, we obtain the following conditional posterior distributions:

$$(3.3) \quad \pi(\rho | \mathbf{x}, \sigma, \mathbf{p}^{(c)}, \lambda) \propto \prod_{j=1}^{N(\rho)} \left( \frac{g(\bar{x}_{C_j}) \lambda p_{s_j}^{(c)}}{c_{s_j} \sigma^{2(s_j-1)}} \exp\left(-\frac{\pi \delta_{C_j}^2}{2\sigma^2}\right) \prod_{i,l \in C_j, i \neq l} \mathbb{1}(m_i \neq m_l) \right),$$

$$(3.4) \quad \pi(\sigma | \mathbf{x}, \rho, \mathbf{p}^{(c)}, \lambda) \propto \frac{\mathbb{1}_{(0, \sigma_{\max})}(\sigma)}{\sigma^{2(n-N(\rho))}} \exp\left(-\frac{\pi \sum_{j=1}^{N(\rho)} \delta_{C_j}^2}{2\sigma^2}\right),$$

$$(3.5) \quad \mathbf{p}^{(c)} | \mathbf{x}, \rho, \sigma, \lambda \sim \text{Dir}(\alpha_1^{(c)} + N_1(\rho), \dots, \alpha_k^{(c)} + N_k(\rho)),$$

$$(3.6) \quad \lambda | \mathbf{x}, \rho, \sigma, \mathbf{p}^{(c)} \sim \text{Gamma}(k_\lambda + N(\rho), \theta_\lambda / (\theta_\lambda + 1)),$$

where  $c_s = \binom{k}{s_j} s_j 2^{s_j-1}$  and  $\mathbb{1}_{(0, \sigma_{\max})}(\cdot)$  is the indicator function of  $(0, \sigma_{\max})$ .

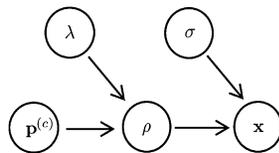


FIG. 3. Conditional independence structure of the random elements involved in the Poisson model.

3.6. *The posterior distribution of the partition  $\rho$ .* The posterior distribution  $\pi(\rho|\mathbf{x}, \sigma, \mathbf{p}^{(c)}, \lambda)$  in (3.3) is intractable, meaning that we cannot obtain exact inferences from it and even performing approximate inferences is challenging. In fact, the posterior sample space  $\mathcal{P}_n$  is too large (of order between  $n!$  and  $n^n$ ) to perform brute force optimization or integration, and the complementary clustering condition makes it not easy to move in the state space. To make these statements more precise, we describe  $\pi(\rho|\mathbf{x}, \sigma, \mathbf{p}^{(c)}, \lambda)$  in terms of hypergraphs and then we consider complexity theory results regarding its intractability. For simplicity, we will denote  $\pi(\rho|\mathbf{x}, \sigma, \mathbf{p}^{(c)}, \lambda)$  by  $\hat{\pi}(\rho)$ .

Note that, although we have little hope of solving the problem in its general form (see Section 3.6.2), Monte Carlo methods, for example, can still give satisfactory results in specific applications.

3.6.1. *Formulation of the model in terms of hypergraphs.* Hypergraphs are the generalization of graphs where each hyperedge can contain more than two vertices [Berge (1973)]. In particular, the complete  $k$ -partite hypergraph induced by  $k$  sets  $V_1, \dots, V_k$  is defined as  $G = (V, E)$ , where  $V = V_1 \cup \dots \cup V_k$  and  $E = \{e \subseteq V : |e \cap V_l| \leq 1 \forall l, |e| \geq 2\}$ . See Figure 4(a). A partition  $\rho \in \mathcal{P}_n$  of  $n$  points into clusters is admissible for our model if and only if no cluster of  $\rho$  contains two points of the same type. Therefore, a set of points is an admissible cluster if and only if the hyperedge connecting them belongs to the complete  $k$ -partite hypergraph induced by the  $k$  set of points corresponding to the  $k$  types. Every admissible partition  $\rho$  can then be interpreted as a partial matching (i.e., hypergraph with at most one hyperedge containing each point) contained in  $G$  as follows: each cluster with at least two points corresponds to a hyperedge and each unlinked point is a cluster by itself [see Figure 4(b)]. Moreover, we can define a weight  $w(e)$  for each hyperedge  $e = \{x_1, \dots, x_s\}$  in  $E$ ,

$$(3.7) \quad w(e) = \frac{(c_1)^s \lambda p_s^{(c)} g(\bar{x}) \sigma^{-2(s-1)}}{c_s (\lambda p_1^{(c)})^s g(x_1) \dots g(x_s)} \exp\left(-\frac{\pi \sum_{i=1}^s (x_i - \bar{x})^2}{2\sigma^2}\right),$$

in such a way that  $\hat{\pi}(\rho)$  is proportional to the weight of the matching  $\rho$ , defined as  $\prod_{e \in \rho} w(e)$ . In (3.7)  $\bar{x}$  denotes the barycenter of  $x_1, \dots, x_s$ .

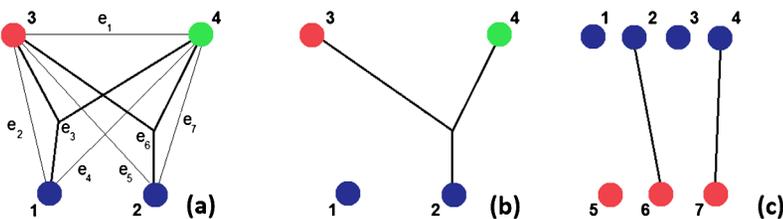


FIG. 4. (a): Complete 3-partite hypergraph induced by the sets  $V_1 = \{1, 2\}$ ,  $V_2 = \{3\}$  and  $V_3 = \{4\}$  corresponding to the colors blue, red and green. (b)–(c): Partial matching corresponding to  $\rho = \{\{1\}, \{2, 3, 4\}\}$  and  $\rho = \{\{1\}, \{2, 6\}, \{3\}, \{4, 7\}, \{5\}\}$ , respectively.

In the remainder of the paper we will treat  $\rho$  indifferently as a partition or as a matching, as the two formulations are equivalent. Note that in the two-color case  $\rho$  reduces to a matching in a bipartite graph; see Figure 4(c).

3.6.2. *Complexity theory results for  $\hat{\pi}(\rho)$ .* Given the hypergraph formulation of Section 3.6.1, we can appeal to complexity theory results to obtain rigorous statements on the intractability of  $\hat{\pi}(\rho)$ . In particular, we consider the following tasks: (a) finding the normalizing constant of  $\hat{\pi}(\rho)$ , (b) finding the mode  $\rho_{\max} = \arg \max_{\rho \in \mathcal{P}_n} \hat{\pi}(\rho)$  and (c) sampling from  $\hat{\pi}(\rho)$ . In this section we briefly summarize the complexity of such tasks. Zanella (2015e) provides a more detailed analysis. Note that the two-color case ( $k = 2$ ) and the multicolor case ( $k \geq 3$ ) present substantially different complexity issues.

(a) The normalizing constant of  $\hat{\pi}(\rho)$  is the sum of the weights of all the matchings  $\rho$  contained in  $G$ , that is, the total weight of  $G$ . The problem of computing the total weight of a  $k$ -partite hypergraph is an  $\#P$ -hard counting problem [Valiant (1979)], even for  $k = 2$ . The  $\#P$ -hard complexity class for counting problems is analogous to the  $NP$ -hard complexity class for decision problems [see Valiant (1979) or Jerrum (2003) for definitions of these terms].

(b) Finding the posterior mode  $\rho_{\max} = \arg \max_{\rho} \hat{\pi}(\rho)$  can be reduced to a  $k$ -dimensional optimal assignment problem [see Zanella (2015e)]. For  $k = 2$  this problem is efficiently solvable, for example, in  $O(n^3)$  steps with the Hungarian Algorithm [Kuhn (1955)]. In contrast, for  $k \geq 3$  this is an  $NP$ -hard optimization problem. Even more, unless  $P = NP$ , there is no deterministic polynomial-time approximation algorithm for a general cost function (i.e., the problem is not in  $APX$ ). Heuristics algorithms exist, but no constant of approximation is provided [see Zanella (2015e)]. Therefore, while heuristics might still work in particular cases, the literature does not appear to provide a generic bounded-complexity method to obtain or approximate  $\rho_{\max}$ .

(c) For  $k = 2$ ,  $\hat{\pi}(\rho)$  can be interpreted as a monomer-dimer system [see Zanella (2015e)]. Jerrum and Sinclair (1996) describe a polynomial-time MCMC algorithm to draw approximate samples from  $\hat{\pi}(\rho)$ . Unfortunately, the polynomial bound they provide on the number of MCMC steps needed is not practically feasible [more details in Zanella (2015e)]. More recent results [Karpinski, Rucinski and Szymanska (2012)] suggest that the techniques used by Jerrum and Sinclair (1996) cannot be extended to  $k \geq 3$ , and they prove a negative result for  $k \geq 6$  [see Zanella (2015e)].

Theoretical results like the ones above do not rule out, for example, the possibility of obtaining approximate samples in specific situations, but do exclude the possibility of finding a scheme that does so (in polynomial time) for arbitrary instances of a certain class of distributions. Since the problem we consider is by no mean arbitrary, it is feasible that special methods may produce good approximate samples. In Section 4 we propose an MCMC algorithm for the two-color case and

one for the  $k$ -color case. As a consequence of the results presented in this section, it is clear that additional care is needed when empirically studying MCMC mixing properties.

**4. Description of proposed MCMC algorithm.** We use the Metropolis-within-Gibbs algorithm to sample from  $\pi(\rho, \sigma, \mathbf{p}^{(c)}, \lambda | \mathbf{x})$  given in (3.3)–(3.6). Direct sampling from  $\pi(\mathbf{p}^{(c)} | \rho, \sigma, \lambda, \mathbf{x})$  and  $\pi(\lambda | \rho, \sigma, \mathbf{p}^{(c)}, \mathbf{x})$  is straightforward and, given  $(\rho, \mathbf{p}^{(c)}, \lambda, \mathbf{x})$ , few steps of the Metropolis–Hastings algorithm are sufficient for the distribution of  $\sigma$  to be close to its stationary distribution  $\pi(\sigma | \rho, \mathbf{p}^{(c)}, \lambda, \mathbf{x})$ . In contrast, sampling from  $\pi(\rho | \mathbf{x}, \sigma, \mathbf{p}^{(c)}, \lambda)$ , which for simplicity we will denote by  $\hat{\pi}(\rho)$ , is challenging (see Section 3.6.2). To do this, we use the Metropolis–Hastings (MH) algorithm. We consider ways of improving the efficiency and of assessing the convergence of MH algorithms in this framework.

4.1. *2-color case.* We commence by considering the two-color case because there is more known theory than in the general case and because the combinatorial structure of the sample space is simpler. We view  $\rho$  as a matching in a bipartite graph with  $n_1$  red points and  $n_2$  blue points (see Section 3.6.1). We denote the edge connecting the  $i$ th red point and the  $j$ th blue point by the ordered couple  $(i, j) \in \{1, \dots, n_1\} \times \{1, \dots, n_2\}$ .

The proposal  $Q^{2D}(\rho_{\text{old}}, \rho_{\text{new}})$  for  $\rho$  is defined in two steps. First, we select an edge  $(i, j)$  according to some probability distribution  $q_{\rho_{\text{old}}}(i, j)$  on  $\{1, \dots, n_1\} \times \{1, \dots, n_2\}$ . Then, having defined  $i'$  as the index such that  $(i', j) \in \rho_{\text{old}}$ , if such an  $i'$  exists, and similarly  $j'$  as the index such that  $(i, j') \in \rho_{\text{old}}$ , if such a  $j'$  exists, we propose a new state  $\rho_{\text{new}} = \rho_{\text{old}} \circ (i, j)$  defined as

$$(4.1) \quad \left\{ \begin{array}{ll} \rho_{\text{old}} + (i, j), & \text{if neither } i' \text{ nor } j' \text{ exists,} & \text{(Addition)} \\ \rho_{\text{old}} - (i, j), & \text{if } (i, j) \in \rho_{\text{old}}, & \text{(Deletion)} \\ \rho_{\text{old}} - (i, j') + (i, j), & \text{if } j' \text{ exists and } i' \text{ does not exist,} & \text{(Switch)} \\ \rho_{\text{old}} - (i', j) + (i, j), & \text{if } i' \text{ exists and } j' \text{ does not exist,} & \text{(Switch)} \\ \rho_{\text{old}} - (i', j) - (i, j') + (i, j) + (i', j'), & \text{if } i' \text{ and } j' \text{ exist and } (i, j) \notin \rho_{\text{old}}, & \text{(Double-Switch)} \end{array} \right.$$

where  $\rho - (i, j)$  and  $\rho + (i, j)$  denote the matchings obtained from  $\rho$  by respectively removing or adding the edge  $(i, j)$ . Display (4.1) defines the set of allowed moves starting from  $\rho_{\text{old}}$  and it induces a neighboring structure on the space of matchings as follows:  $\rho_{\text{new}}$  is a neighbor of  $\rho_{\text{old}}$  if  $\rho_{\text{new}} = \rho_{\text{old}} \circ (i, j)$  for some  $(i, j)$ . Jerrum and Sinclair (1996) and Oh, Russell and Sastry (2009) consider similar but slightly smaller sets of allowed moves, given by the addition and deletion

moves and addition, deletion and switch moves, respectively. It is plausible that increasing the set of allowed moves improves the mixing of the MH Markov chain.

Display (4.1) does not identify uniquely the proposal  $Q^{2D}(\rho_{\text{old}}, \rho_{\text{new}})$  because we still need to choose  $q_{\rho_{\text{old}}}(\cdot, \cdot)$ . Different choices of  $q_{\rho_{\text{old}}}(\cdot, \cdot)$  will affect the mixing properties of the MH algorithm. Previous works [e.g., [Jerrum and Sinclair \(1996\)](#) and [Oh, Russell and Sastry \(2009\)](#)] chose  $q_{\rho_{\text{old}}}(i, j)$  to be a uniform measure over the edges  $(i, j) \in E$ . A naive implementation of such choice leads to poor mixing because most proposed matchings  $\rho_{\text{new}}$  are improbable and therefore are typically rejected (in our experiments usually less than 1% of the proposed moves were accepted). Some authors overcome this problem using a truncation approximation of the posterior: they force edge weights below a certain threshold  $\delta$  to be zero, and then choose

$$(P1) \quad q_{\rho_{\text{old}}}(i, j) \propto \mathbb{1}_{\{w_{ij} > \delta\}},$$

where  $w_{ij}$  is the weight of the edge  $(i, j)$  defined in (3.7) and  $\mathbb{1}$  denotes the indicator function. See, for example, the measurement validation step in [Oh, Russell and Sastry \(2009\)](#).

In the following we propose a choice of  $q_{\rho_{\text{old}}}$  that achieves a better mixing than (P1) and does so without requiring to target an approximation of the posterior.

First note that, especially when  $\hat{\pi}(\rho)$  has a factorization in terms of edge weights, it is straightforward to evaluate  $\hat{\pi}$  up to a multiplicative constant on the set of neighbors of  $\rho_{\text{old}}$  defined in (4.1). For example, for the addition move,  $\frac{\hat{\pi}(\rho_{\text{old}} \circ (i, j))}{\hat{\pi}(\rho_{\text{old}})} = w_{ij}$ . Thus, one may be tempted to propose proportionally to  $\hat{\pi}$  restricted on the set of allowed moves as follows:

$$(P2) \quad q_{\rho_{\text{old}}}(i, j) \propto \hat{\pi}(\rho_{\text{new}}) \quad \text{where } \rho_{\text{new}} = \rho_{\text{old}} \circ (i, j).$$

Such a choice, however, does not take into account the fact that the normalizing constants of  $q_{\rho_{\text{old}}}(\cdot, \cdot)$  and  $q_{\rho_{\text{new}}}(\cdot, \cdot)$  differ for  $\rho_{\text{old}} \neq \rho_{\text{new}}$ . As a consequence, for example, detailed balance conditions,  $\frac{Q^{2D}(\rho_{\text{old}}, \rho_{\text{new}})}{Q^{2D}(\rho_{\text{new}}, \rho_{\text{old}})} = \frac{\hat{\pi}(\rho_{\text{new}})}{\hat{\pi}(\rho_{\text{old}})}$ , are not satisfied, not even approximately. A better choice for  $q_{\rho_{\text{old}}}(\cdot, \cdot)$  is

$$(P3) \quad q_{\rho_{\text{old}}}(i, j) \propto \frac{\hat{\pi}(\rho_{\text{new}})}{\hat{\pi}(\rho_{\text{old}}) + \hat{\pi}(\rho_{\text{new}})} \quad \text{where } \rho_{\text{new}} = \rho_{\text{old}} \circ (i, j).$$

Our experiments show that the latter choice leads to a significant improvement in the mixing of the MH Markov chain compared to (P1) and (P2) (see Section 4.1.2). The main reason is that the MH algorithm induced by such a proposal has a very high acceptance rate (usually above 99%) without changing the set of allowed moves. It can be shown that, under some regularity assumption on the weights, the proposal given by (P3) satisfies the detailed balance condition in the asymptotic regime (i.e., when the number of points tends to infinity), and this helps to explain why the acceptance rate is so high. Similarly, one could also derive Peskun ordering arguments in the asymptotic regime. We omit those theoretical results

here in favor of demonstrating the mixing improvement given by (P3) using the convergence diagnostic techniques in Section 4.1.2.

There is a trade-off between the complexity of the proposal and the mixing obtained (a complex proposal increases the cost of each step, while a poor proposal increases the number of MCMC steps needed). We seek a compromise with good mixing properties, like (P3), while still requiring little computation, like (P1). In Section 2 of Zanella (2015d) we derive the following proposal distribution to try to obtain such a goal:

$$(P4) \quad q_{\rho_{old}}(i, j) \propto \begin{cases} q^{(add)}(i, j), & \text{if } (i, j) \notin \rho_{old}, \\ q^{(rem)}(i, j), & \text{if } (i, j) \in \rho_{old}, \end{cases}$$

where  $q^{(rem)}(i, j) = w_{ij}^{-1/2}$  and

$$q^{(add)}(i, j) = \sqrt{w_{ij}} \left( 1 - \sum_{j' \neq j} \frac{w_{ij'} - \sqrt{w_{ij'}}}{1 + \sum_{s \neq i} w_{sj'} + \sum_l w_{il}} \right) \times \left( 1 - \sum_{i' \neq i} \frac{w_{i'j} - \sqrt{w_{i'j}}}{1 + \sum_{s \neq j} w_{i's} + \sum_l w_{lj}} \right).$$

Note that  $q^{(rem)}(i, j)$  and  $q^{(add)}(i, j)$  do not depend on  $\rho$  and can be precomputed at the beginning of the MCMC run. See Section 4.1.2 for discussion of performance.

4.1.1. *Scaling the proposal with a multiple proposal scheme.* When using the MH algorithm on continuous sample spaces, one can usually tune the variance of its proposal distribution to improve the efficiency of its algorithm [see, e.g., Roberts, Gelman and Gilks (1997)]. Given the very high acceptance rate obtained proposing according to (P3), it is natural to consider the possibility of scaling our proposal in order to obtain longer-scale moves. The scaling problem for MH algorithms in discrete contexts has been considered, for example, in Roberts (1998). In that case the sample space was  $\{0, 1\}^N$ , the vertices of the  $N$ -dimensional hypercube, and the scaling parameter, say  $l$ , was a positive integer representing the number of randomly chosen bits to be flipped at any given proposal.

Unfortunately, because of the nature of our sample space, it is not so straightforward to scale the proposal distribution  $Q^{2D}(\rho_{old}, \rho_{new})$ . One possibility is to scale by choosing  $l$  edges,  $\{(i_h, j_h)\}_{h=1}^l$ , and performing  $l$  moves defined in (4.1), proposing  $\rho_{new} = \rho_{old} \circ (i_1, j_1) \circ \dots \circ (i_l, j_l)$ . However, the  $l$  moves corresponding to  $\{(i_h, j_h)\}_{h=1}^l$  cannot be performed independently: consider, for example, the case where  $i_1 = i_2$ . We would then have to perform  $l$  moves sequentially, at a computational cost being roughly  $l$  times the one of a single move. Therefore, scaling the proposal in such a way does not seem to be effective.

Instead, if the  $l$  moves could be performed independently, it would be possible to implement a multiple proposal scheme using parallel computation, thus leading

to a significant computational gain. This can be obtained by considering an approximation of our model, where points at a distance greater or equal than some  $r_{\max}$  have probability 0 of being in the same cluster. The latter procedure is equivalent to the truncation procedure cited in Section 4.1 and can be viewed as coming from the use of truncated Gaussian distributions to model points distribution within clusters; see (3.1). Using this truncated model and dividing the observed region into a grid, we defined a multiple proposal scheme where the  $l$  moves are proposed and accepted/rejected simultaneously and independently. Therefore, at each MH step, such  $l$  moves can be performed in an embarrassingly parallel fashion, meaning that they can be performed without the need for any communication between them. In Zanella (2015f) we give more details on the implementation and we show that in practice the mixing of the resulting MH algorithm improves by a factor roughly equal to  $l$  itself (note that the maximum value of  $l$  is bounded above, in a way that depends on  $r_{\max}$  and the size of the observation region  $W$ ). A parallel-computing implementation of this algorithm would offer significant speedups [we anticipate speedups by a factor around 8 for our data set; see Zanella (2015f)]. Such speedups would increase with the size of the data set and window, making this proposal scheme especially relevant for applications to very large data sets. In Zanella (2015f) this scheme is presented and tested for fixed  $\sigma$ . In case  $\sigma$  is varying, either one requires an upper bound on  $\sigma$  or one needs different square grids for different values of  $\sigma$ .

4.1.2. *Convergence diagnostics.* We used various convergence diagnostic techniques in order to assess the reliability of our algorithm, to indicate the number of iterations needed and to compare the efficiency of the four proposals (P1)–(P4) of Section 4.1. We demonstrate such techniques on the posterior  $\pi(\rho|\sigma, \mathbf{p}^{(c)}, \lambda, \mathbf{x})$  with  $k = 2$ ,  $\sigma = 0.3$ ,  $p_1^{(c)} = p_2^{(c)} = 0.5$ ,  $\lambda = 50$  and the center intensity  $g(\cdot)$  being the uniform measure over  $W = [0, 10] \times [0, 10]$ . Here  $\mathbf{x}$  is a synthetic sample of 44 red and 47 blue points generated according to the model just defined; see Figure 6(a). We set the threshold  $\delta$  of (P1) to 0.001. The R code used to produce the results presented in this section is available in Zanella (2015a).

We first performed some qualitative output analysis by looking at summary plots of the MCMC samples of the partition [such as the one in Figure 6(a)]. Such plots can be helpful to spot when mixing has not yet occurred (see Section 4.1.3).

Second, we considered different real-valued summary statistics of the chain state (typically the number of different edges from some fixed reference matching). We plotted time series (see Figure 5) and empirical distributions of such real-valued functions for different runs of the MCMC starting from different configurations. We estimated the autocorrelation functions [see Figure 6(b)], the Integrated Autocorrelation Time (IAT) and the Effective Sample Size (ESS) of such real-valued time series using the R package *coda* [see Plummer et al. (2005)] in order to compare different versions of the algorithm (see Table 3).

Traceplots of a summary statistic for different proposals

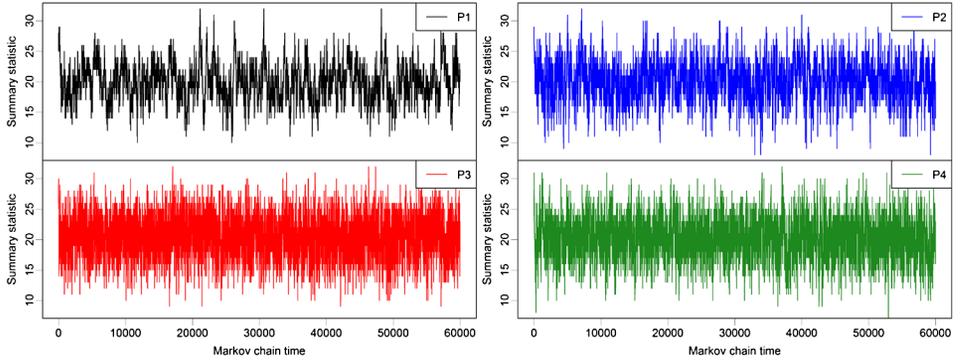


FIG. 5. Traceplots of the number of differences from a reference matching.

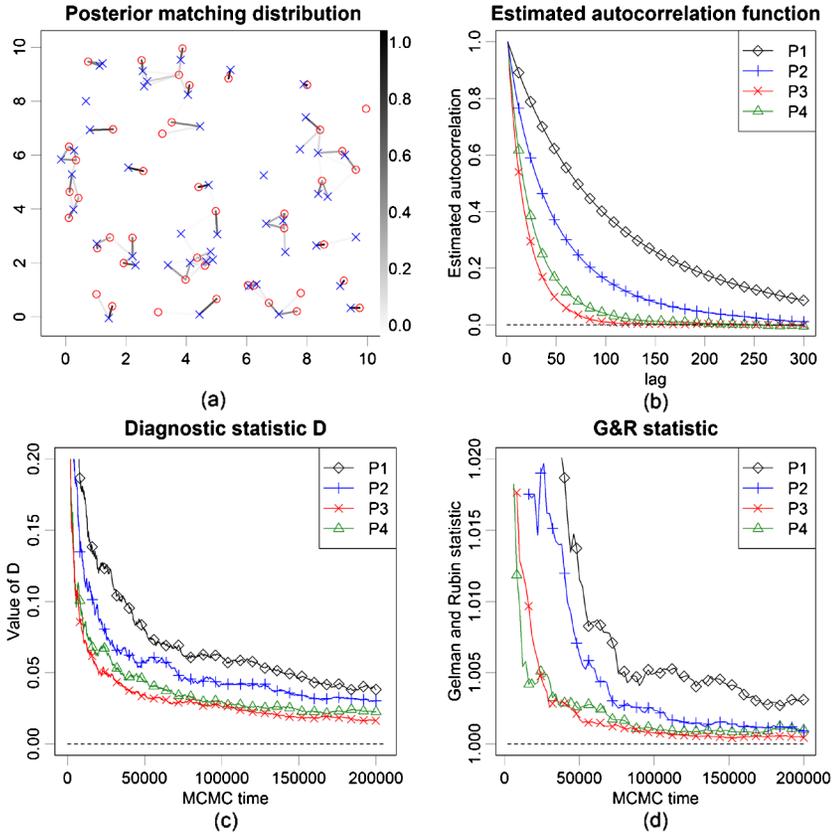


FIG. 6. Four convergence diagnostic techniques described in Section 4.1.2.

TABLE 3

*Performances of the four proposals of Section 4.1 on configuration in Figure 6(a) averaged over 5 independent runs for each proposal. GR denotes the multivariate Gelman and Rubin statistic (potential scale reduction factor). The running time indicated in brackets is evaluated using R software on a desktop computer with Intel i7 processor*

|    | Mean acc. rate | Estimated IAT | ESS for 10 <sup>4</sup> steps (for 1 sec) | Steps (sec) to D < 0.05 | Steps (sec) to GR < 0.005 |
|----|----------------|---------------|---|-------------------------|---------------------------|
| P1 | 17%            | 206           | 262 [270]                                 | 1.4e05 [7.3]            | 7.6e04 [13.5]             |
| P2 | 41%            | 108           | 544 [40]                                  | 7.1e04 [84.6]           | 6.2e04 [97]               |
| P3 | 97%            | 40            | 1358 [99]                                 | 2.0e04 [32.7]           | 2.4e04 [27.3]             |
| P4 | 68%            | 55            | 1038 [747]                                | 3.4e04 [2.2]            | 1.6e04 [4.8]              |

Third, we used some standard convergence diagnostic techniques [see Brooks and Roberts (1998) and Cowles and Carlin (1996) for an overview of the techniques available]. In particular, we used the multivariate version of the Gelman and Rubin diagnostic [see Gelman and Rubin (1992) and Brooks and Gelman (1998)]. Figure 6(d) shows the results obtained by using a 10-dimensional summary statistic of  $\rho$ . In this context univariate summary statistics are not sufficiently informative and, therefore, misleading results can be obtained if these are used as the sole basis for convergence diagnostics.

Finally, we compared two independent runs of the algorithm (with different starting states) by looking at estimates of the association probabilities  $p_{ij} = \Pr((i, j) \in \rho)$  with  $\rho \sim \hat{\pi}$ . We consider the following measure of proximity:

$$(4.2) \quad D = \sup_{(i,j) \in E} |\hat{p}_{ij}^{(1)} - \hat{p}_{ij}^{(2)}|,$$

where  $\hat{p}_{ij}^{(1)}$  and  $\hat{p}_{ij}^{(2)}$  denote the proportion of time that  $(i, j)$  was present in the two MCMC runs. As starting states we considered the empty matching (each point is a cluster), the posterior mode (obtained with the Hungarian algorithm) and matchings obtained as the output of the MCMC itself. Since equation (4.2) considers each link individually, we expect the resulting convergence diagnostic indicator  $D$  to be more severe than the ones obtained from one or few summary statistics. Results are shown in Figure 6(d).

None of the convergence methods just presented indicate convergence issues except in the complete matching case (when the parameter  $p_1^{(c)}$  is equal or very close to 0), that is considered in the next subsection.

All convergence diagnostic techniques agree in indicating that proposal (P3) gives the best mixing; however, in terms of real computation time, the most efficient proposal is (P4). Note that such performances depend on the measure being targeted and, when running time is considered, on the computer implementation of such proposals. For the case considered in this section, proposal (P4) gives a 3–4

times speedup over the commonly used choice (P1). Depending on the configuration, such a speedup may vary. According to our experiments, for “flatter” distributions (e.g., increasing  $\sigma$  to 1 and  $p_1^{(c)}$  to 0.9, while keeping the other parameters unchanged) the speedup almost disappears, while for “rougher” distributions (e.g., decreasing both  $\sigma$  and  $p_1^{(c)}$  to 0.1, while keeping the other parameters unchanged) the speedup increases and (P4) can be up to 10 times faster than (P1).

4.1.3. *Multimodality and simulated tempering.* In the complete matching case the posterior distribution of  $\rho$  presents a strongly multimodal behavior. Cycle-like configurations like the one in Figure 7(a) are local maxima for  $\hat{\pi}(\rho)$ . In fact, in order to reach a higher probability configuration (i.e., shorter links) from such a “cycle” configuration, with the set of allowed moves defined by (4.1), the chain needs to pass through lower probability configurations (i.e., longer links). If we consider extreme cycle-like configurations [e.g., Figure 7(b)], then the MCMC will typically get stuck in such local maxima. In order to overcome this potential multimodality problem, we implemented a simulated tempered version of our MCMC algorithm; see, for example, [Geyer and Thompson \(1995\)](#) or [Marinari and Parisi \(1992\)](#) for references. This technique manages to overcome local maxima for the complete matching case even when extreme cycle-like configurations are present [as in Figure 7(b)]. Nevertheless, our specific application does not present a complete matching case and, therefore, we have a milder multimodality and the MCMC algorithm exhibits sufficient mixing without the use of Simulated Tempering. Therefore, Simulated Tempering is not used for the real data analysis, as convergence diagnostic tools do not show suspicious behavior.

We note that [Dellaert et al. \(2003\)](#) deal with multimodality in a similar posterior space (made of perfect matchings in a bipartite graph) arising from the Structure from Motion problem. In order to allow the MH algorithm to overcome local maxima like the one in Figure 7(b), they allow the MH proposal to include “long” moves that they call “chain flipping.”

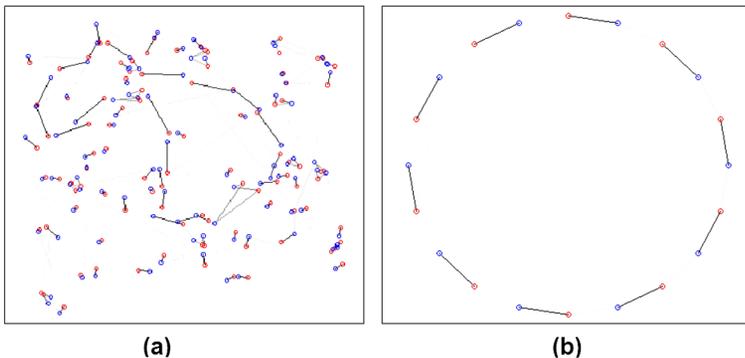


FIG. 7. Configurations corresponding to local maxima of  $\pi(\rho|\mathbf{x})$  for (a) a synthetic sample and (b) an artificially designed points configuration.

4.2. *k-color case.* We now define an MCMC algorithm that targets  $\hat{\pi}(\rho)$  when  $k \geq 3$ . This case is harder than the two-dimensional one because it involves clusters with different dimensions and not just pairwise interaction.

4.2.1. *Description of proposed Gibbs projection MCMC algorithm.* We define the transition kernel  $P$  of our MCMC algorithm as a mixture of  $\binom{k}{\lfloor k/2 \rfloor}$  MH transition kernels, each of which corresponds to a group  $A$  of  $\lfloor k/2 \rfloor$  colors

$$(4.3) \quad P(\rho_{old}, \rho_{new}) = \binom{k}{\lfloor k/2 \rfloor}^{-1} \sum_{A \subset \{1, \dots, k\}, |A| = \lfloor k/2 \rfloor} P^{(A)}(\rho_{old}, \rho_{new}),$$

where  $\lfloor k/2 \rfloor$  denotes the integer part of  $k/2$  and  $\binom{k}{\lfloor k/2 \rfloor}$  denotes a binomial coefficient. Here  $P(\cdot, \cdot)$  selects a set of colors  $A$ , “projects” the  $k$ -color configuration to a 2-colors configuration where the new two colors correspond to  $A$  and  $A^c = \{1, \dots, k\} \setminus A$  and then acts on the two-colors configuration. More precisely, the action of  $P^{(A)}$  is the following (see Figure 8):

1. reduce the  $k$ -color configuration  $(\mathbf{x}, \rho_{old})$  to a two-color one  $(\mathbf{x}^{2D}, \rho_{old}^{2D})$  by replacing the points having colors in  $A$  and  $A^c$ , respectively, with their cluster centroids. We denote by  $d_i$  the number of points merged together into the  $i$ th point  $x_i^{2D}$ ,
2. obtain  $\rho_{new}^{2D}$  from  $(\mathbf{x}^{2D}, \rho_{old}^{2D})$  with one or more MH moves using the proposal  $Q^{2D}$  of Section 4.1 on a target measure  $\hat{\pi}^{2D}$  being the two-dimensional version of  $\hat{\pi}$  [modified to take account of the multiplicity of the points  $d_i$ ; see Section 3 of Zanella (2015d)],
3. obtain the  $k$ -color configuration  $(\mathbf{x}, \rho_{new})$  from  $(\mathbf{x}^{2D}, \rho_{new}^{2D})$  by the inverse operation of Step 1 (note that here one needs to know what  $A$  is).

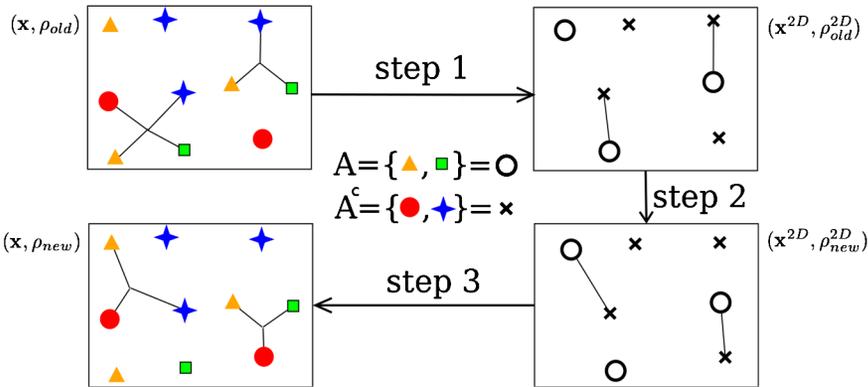


FIG. 8. The action of a transition kernel  $P^{(A)}$  for a given  $A$ .

In order for this algorithm to be correct,  $\hat{\pi}^{2D}$  must be proportional to  $\hat{\pi}$  on the collection of possible moves of  $P^{(A)}$ , so that  $P^{(A)}$  satisfies detailed balance conditions with respect to  $\hat{\pi}$ . This follows from basic properties of the Gaussian density function and is proven in Section 3 of Zanella (2015d). Note that, when  $k$  is even,  $P^{(A)}$  is the same transition kernel as  $P^{(A^c)}$ . This is not an issue and it is indeed equivalent to never using  $P^{(A^c)}$  and using  $P^{(A)}$  twice more often.

By merging colors together we allow proposals which move many points at the same time from one cluster to another. Therefore, the induced set of allowed moves is broader than, for example, the one of a scheme that moves one point at a time. Oh, Russell and Sastry (2009) consider also, for example, “birth” moves proposing to create a cluster from three or more single points in one step. Such moves are likely to be useful to speed up mixing in applications where there appear clusters with many points.

The mixture proposal in (4.3) allows us to reuse the two-color algorithm and, in particular, the approximation given in (P4). In fact,  $\hat{\pi}^{2D}$  involves only pairwise interaction among points, meaning that  $\hat{\pi}^{2D}(\rho^{2D}) \propto \prod_{(i,j) \in \rho^{2D}} w_{ij}^{2D}$  for some weights  $w_{ij}^{2D}$  depending on  $\mathbf{x}^{2D}$  [see Remark 1 of Zanella (2015d)]. Therefore, given  $(\mathbf{x}^{2D}, \rho_{\text{old}}^{2D})$ , it is possible to perform informed MH moves in the two-color matching space in a computationally efficient way using the approximation given in (P4) (see Table 3 for performances with two colors).

It would be desirable to design informed proposals like (P3) or (P4) directly in the  $k$ -color space, without the need of projecting on two-color subspaces. However, it would not be easy to do so in a computationally efficient way. In fact, given the high-dimensionality of the space of matchings contained in a complete  $k$ -partite hypergraph, the set of neighboring states  $\rho_{\text{new}}$  of the current state  $\rho_{\text{old}}$  would be extremely large. Therefore, it would be very expensive to use a scheme like (P3) in this context. Moreover, since  $\hat{\pi}(\rho)$  involves interactions between three or more points, it would be difficult to design an approximation like (P4) that can be evaluated efficiently.

Note that the mixture proposal in (4.3) first chooses uniformly at random a lower-dimensional subspace and then performs informed proposals in such a space. Therefore, such a scheme is a compromise between a “fully uninformed” proposal (which would choose uniformly at random some neighbor of  $\rho_{\text{old}}$  and thus mix poorly) and a “fully informed” proposal (which, in order to make informed proposals in the  $k$ -color space, would be computationally expensive).

Since the  $k$ -color sample space is more complicated than the two-color one, additional care and longer MCMC runs are needed. We implemented analogous convergence diagnostic techniques to the ones in Section 4.1.2. As might be expected, the number of MCMC steps needed to reach stationarity and to obtain mixing is much higher than in the two-color case (see end of Section 5). Nevertheless, our experiments suggest that, as in the two-color case, the MCMC manages to mix properly unless we are in a case close to complete matching (see Section 4.1.3).

**5. Analysis of Anglo-Saxon settlements with the Bayesian model.** In this section we present the main results obtained by analyzing the Anglo-Saxon settlements data set with the Random Partition Model described in Section 3. The computation is done using the MCMC algorithm described in Section 4. The analysis gives support to the historians’ hypothesis that settlements are clustered according to complementary functional place-names, and it permits inference about ranges of values for relevant parameters.

Here the no-clustering null hypothesis corresponds to  $p_1^{(c)} = 1$  (see Section 3). As shown in Figure 9(a), such a hypothesis clearly lies outside the region where the posterior distribution is concentrated. As a sanity check we also fitted our model to synthetic samples generated according to the no-clustering null hypothesis of Section 2.3.1 (both with and without inhibition among points of the same type). As one would expect, in this case  $p_1^{(c)} = 1$  is included in the posterior support [see Figure 9(a) for an example].

Figure 10(a) shows the estimated posterior distribution of  $\sigma$  for the reduced data set, which is clearly peaked around 4–5 km.

The 95% Highest Posterior Density interval is (3.3, 5.9) km and the posterior mean is 4.6 km. Therefore, according to the fit given by our model, the clustering behavior consists of clusters with settlements having distance being approximately 5 km on average. It is satisfying to note that this value is in accordance with the value suggested by the historians involved in the project and coherent with the historical interpretation (see Section 3.3).

Figure 11(a) shows a box plot representation of the posterior distribution of  $(Y_1, \dots, Y_k)$ , where  $Y_l$  is the number of settlements in clusters of size  $l$  (i.e., with

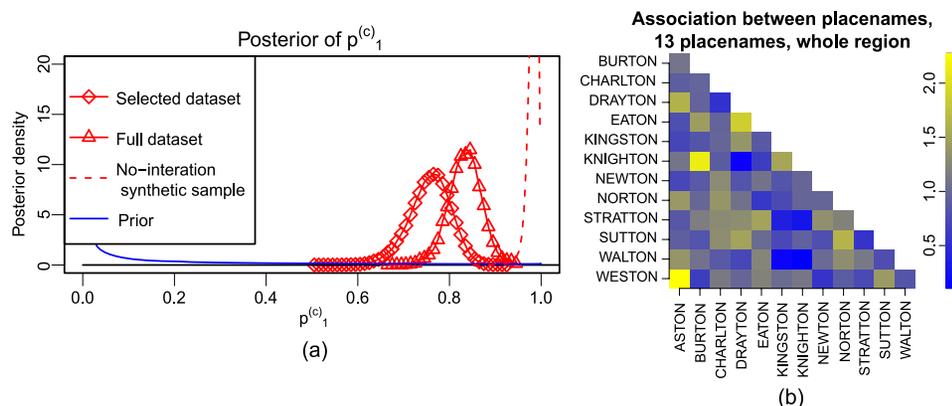


FIG. 9. (a) Estimated posterior distribution of  $p_1^{(c)}$  (see Section 3) for the reduced and full data set (13 and 20 place-names, resp.). The no-clustering hypothesis ( $p_1^{(c)} = 1$ ) lies outside the support of the posterior for the real data set. (b) Measure of association between place-names (see end of Section 5).

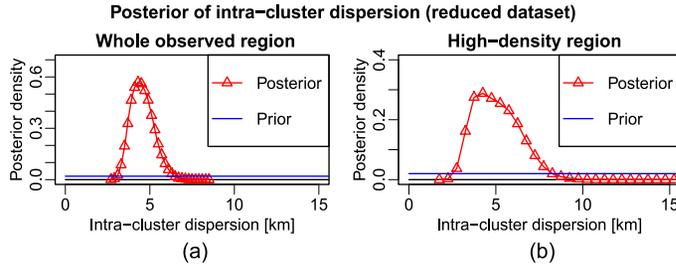


FIG. 10. (a)  $\pi(\sigma|\mathbf{x})$  for the reduced data set. (b)  $\pi(\sigma|\mathbf{x})$  considering only a high-density region (see Section 6).

$l$  settlements). Note that on average more than half of the settlements are not clustered (i.e., they belong to clusters of size 1). Moreover, most of the clustered settlements belong to clusters of size 2. Historians expected to see more clusters involving three or four settlements than what was reported by our model. Inspection shows that model-fitting, and the requirement to fit clusters in the low-density region (which mostly contain couples with a high posterior probability), forces all the clusters in the high-density region to be couples too. In fact, when the high-density region is analyzed separately (approximately 600 settlements), more triples appear and the posterior of  $\sigma$  includes also slightly bigger values; see Figures 10(b) and 11(b). This suggests that there might be a heterogeneity in the clustering behavior between high- and low-density regions which is not captured in the model when applied to the whole region. This indicates a possible direction for future work (see Section 6).

Figure 12 shows a graphical representation of the posterior distribution of the partition  $\rho$  for the reduced data set. This representation is of considerable use since it provides a visual understanding of how the model is fitting the data and enables comparison with contextual information.

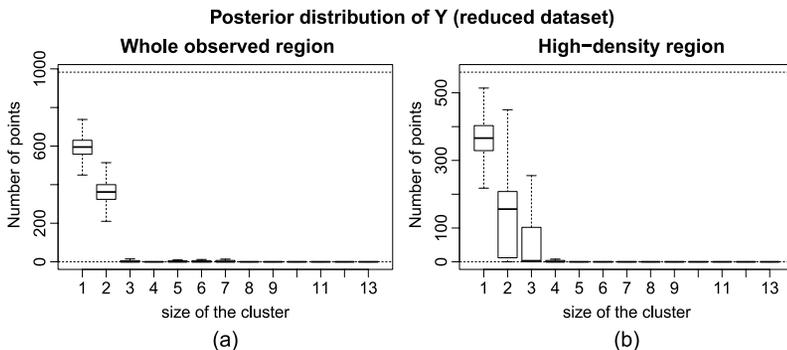


FIG. 11. (a) Posterior distribution of  $\mathbf{Y} = (Y_1, \dots, Y_k)$  for the reduced data set. (b) Same but considering only the settlements in a high density region (see Section 6).

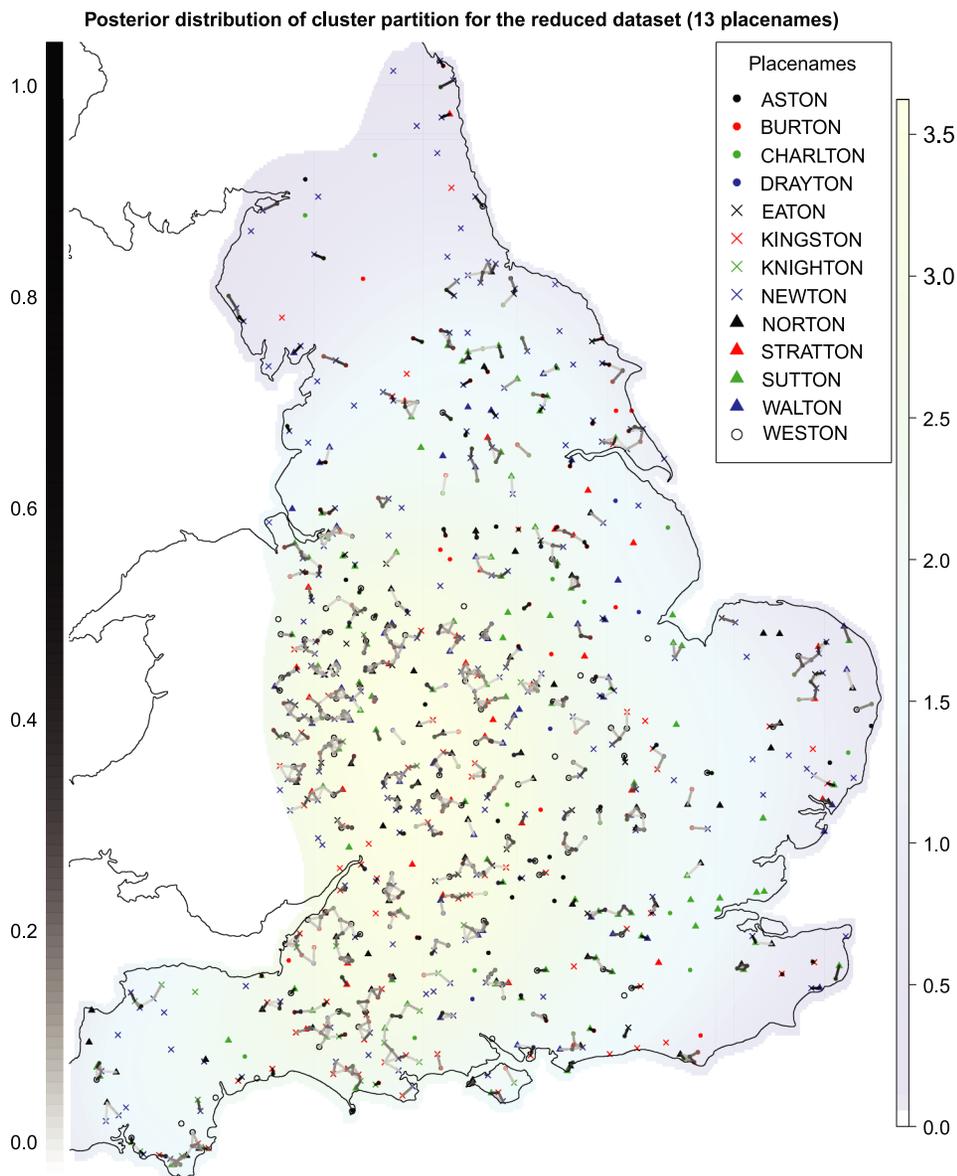


FIG. 12. Graphical representation of  $\pi(\rho|\mathbf{x})$ , where  $\mathbf{x}$  is the reduced data set (13 place-names) in the whole observed region. The intensity of gray corresponds to the estimated posterior probability of the cluster. The truncated kernel density estimation of  $g$  is plotted in the background, with values expressed in relative terms with respect to the uniform measure.

We perform sensitivity analysis on the values of the hyperparameters of  $\sigma$ ,  $\lambda$  and  $\mathbf{p}^{(c)}$  (see Section 3 for details on tested values), and the posterior distribution did not seem to be much sensitive to their specification. As a further sensitivity

analysis, in Zanella (2015b) we specify and implement an alternative model for the prior distribution of the partition  $\rho$ .

Figure 9(b) represents a measure of association between place-names. Given two place-names, say,  $a$  and  $b$ , the measure is defined as

$$(5.1) \quad \frac{\Pr[A|B]}{\Pr[A]} = \frac{\Pr[A \cap B]}{\Pr[A] \cdot \Pr[B]} = \frac{\Pr[B|A]}{\Pr[B]},$$

where  $A$  and  $B$  are the events of observing place-name  $a$  and  $b$ , respectively, in a cluster chosen uniformly at random from the clusters of  $\rho$ , with  $\rho$  distributed according to  $\pi(\rho|\mathbf{x})$ . In Figure 9(b) we plot the value of (5.1), estimated from the MCMC run, in relative terms with respect to a null hypothesis. In the null hypothesis we first choose a cluster from  $\rho$  as before and then, denoting the number of settlements in the cluster by  $s$ , we sample  $s$  place-names independently of each other with place-name probabilities proportional to their numerosity in the data set, conditioning on having pairwise different place-names. The expected values of interest under the null distribution have been estimated using standard Monte Carlo methods. High values in Figure 9(b) suggest positive interaction between place-names, while low values suggest negative interaction. Most of the positive associations suggested by Figure 9(b), such as *Knighton–Burton*, *Weston–Aston* or *Eaton–Drayton*, are coherent with the current historians' hypothesis. We note that, for a fixed  $\rho$ , the measure in (5.1) reduces to the *coefficient of association* used by ecologists to measure association between species [Dice (1945)]. Many different measures of association have been proposed in the ecological literature [see, e.g., Janson and Vegelius (1981)]. We chose (5.1) because it is symmetric, clearly interpretable and our experiments suggest that (5.1) is not much influenced by the numerosity of place-names  $a$  or  $b$ , unlike most measures proposed in Janson and Vegelius (1981).

In order to obtain the results presented in this section, the MCMC algorithm of Section 4.2 was run for  $10^6$  steps, where at each step 200 moves of the two-color configuration  $(\mathbf{x}^{2D}, \rho^{2D})$  were proposed. We assessed convergence using the methods described in Section 4.1.2 [e.g., the value of  $D$  in (4.2) was approximately 0.02]. The time needed for such runs using a basic  $R$  implementation [available in Zanella (2015a)] on a desktop computer with an Intel *i-7* processor was approximately 40 hours.

**6. Discussion.** We have designed a Random Partition Model (RPM) that is able to capture the clustering behavior expected by the historians involved in the project. With no strong prior information, the model produces estimates that are meaningful for the historical context and in accordance with contextual information [e.g., see the posterior distribution of  $\sigma$  and the association between place-names in Figure 9(b)]. We also defined a flexible prior distribution for a clusters partition that is designed for a “small clusters” framework (where each cluster has

at most  $k$  points with  $k$  small). In doing so we developed a RPM to perform complementary clustering which is applicable to other contexts where one needs to find aggregations of elements of different types. For example, Professor Susan Holmes from Stanford University suggests that, in biological contexts, species living in the same geographical area assemble by dissimilarity as they fill different ecological niches, resulting in clusters of complementary species.

We carefully considered the computational aspects of this problem. After considering related problems in the complexity theory literature (see Section 3.6), we employed the Metropolis–Hastings (MH) algorithm. We proposed a choice of MH proposal distributions that, compared to the usual choices found in the literature, achieves a significantly better mixing by approximating detailed balance conditions (see Section 4.1). We developed a multiple proposal scheme to allow for parallel computation that could be relevant for applications to bigger data sets (see Section 4.1.1). Regarding the convergence diagnostic, we note that, when monitoring the convergence of the MCMC in the partition space, univariate summary statistics appear to be not sufficiently informative to be used as a basis for convergence diagnostics. Diagnostics based on multivariate summary statistics or on the matrix of the estimated association probabilities seem to give more robust results (see Section 4.1.2).

Although the proposed model manages to capture the pattern we were looking for, there is much room for improvement. For example, a direction for future work could be to extend the model in order to capture the heterogeneity in the clustering behavior between high- and low-density regions (see Section 5). One could try to do this by allowing the parameters  $\mathbf{p}^{(c)}$  and  $\sigma$  to vary over different regions, maybe as a function of the points density, while taking care not to over-parametrize the model (the amount of data is limited). An alternative approach would be to modify the metric we use to evaluate distances between settlements. For example, one could use a non-Euclidean distance, perhaps based on the inverse square root of the settlements density, in order to allow for larger clusters (meaning with points further apart) in less dense regions. One could also try to model the dispersion of settlements in the same cluster with a non-Gaussian distribution having heavier tails.

Another extension that could result in a better fit is to introduce spatial dependence of place-names probabilities. In fact, in our model, both under the assumption of uniform and nonuniform marks (see Remark 1), the probability of choosing a certain place-name does not depend on the location, while the data suggest that different place-names are more likely to be chosen in different regions.

The context suggests that we are observing a thinned version of the original settlements distribution. Nevertheless, it is not obvious how to incorporate missing data in this model without making further assumptions that do not seem realistic and are not supported by the historical information available (e.g., that in each cluster there is a settlement for each type).

An interesting direction for future work is to try to incorporate other sources of data in the model. For example, topographical information seems to be related to the settlements clustering (e.g., historians think that settlements named Burton are related to good viewpoints); it would be interesting to find an efficient way to incorporate them in the model.

**Acknowledgments.** Thanks to Professor Wilfrid Kendall for PhD supervision, and Professor John Blair for collaboration and arranging supply of data.

#### SUPPLEMENTARY MATERIAL

**Supplement A: Anglo-Saxon settlements data set and R codes** (DOI: [10.1214/15-AOAS884SUPPA](https://doi.org/10.1214/15-AOAS884SUPPA); .zip). Compressed folder containing the data set and the R codes used to perform the analysis carried out in the paper.

**Supplement B: Model extensions and variations** (DOI: [10.1214/15-AOAS884SUPPB](https://doi.org/10.1214/15-AOAS884SUPPB); .pdf). Extensions of the null-hypothesis of Section 2.3.1 and of the Bayesian complementary clustering model defined in Section 3.

**Supplement C: Additional plots** (DOI: [10.1214/15-AOAS884SUPPC](https://doi.org/10.1214/15-AOAS884SUPPC); .pdf). Additional plots relative to the Anglo-Saxon settlements data set and the corresponding preliminary analysis performed in Section 2.

**Supplement D: Additional calculations and derivations** (DOI: [10.1214/15-AOAS884SUPPD](https://doi.org/10.1214/15-AOAS884SUPPD); .pdf). Derivation of the expressions in (3.2) and (P3), and proof of the correctness of the algorithm in Section 4.2.

**Supplement E: Computational complexity of the model** (DOI: [10.1214/15-AOAS884SUPPE](https://doi.org/10.1214/15-AOAS884SUPPE); .pdf). Results and references from the complexity theory literature regarding the intractability of the model of Section 3.6.

**Supplement F: Multiple proposal scheme** (DOI: [10.1214/15-AOAS884SUPPF](https://doi.org/10.1214/15-AOAS884SUPPF); .pdf). Details on the multiple proposal scheme mentioned in Section 4.1.1.

#### REFERENCES

- BADDELEY, A. (2010). Multivariate and marked point processes. In *Handbook of Spatial Statistics*. 371–402. CRC Press, Boca Raton, FL. [MR2730956](https://doi.org/10.1201/9781420040200)
- BADDELEY, A. J., MØLLER, J. and WAAGEPETERSEN, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Stat. Neerl.* **54** 329–350. [MR1804002](https://doi.org/10.1002/9781118445397.ch25)
- BADDELEY, A. and TURNER, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *J. Stat. Softw.* **12** 1–42.
- BADDELEY, A. J. and VAN LIESHOUT, M. N. M. (1995). Area-interaction point processes. *Ann. Inst. Statist. Math.* **47** 601–619. [MR1370279](https://doi.org/10.1080/00140139508837679)
- BECKER, R. A., WILKS, A. R. and BROWNRIGG, R. (2013). Mapdata: Extra Map Databases. R package version 2.2-2.
- BERGE, C. (1973). *Graphs and Hypergraphs*. North-Holland, Amsterdam. [MR0357172](https://doi.org/10.1016/0011-8519(73)90001-0)

- BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. [MR1665662](#)
- BROOKS, S. P. and ROBERTS, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Stat. Comput.* **8** 319–335.
- CHIU, S. N., STOYAN, D., KENDALL, W. S. and MECKE, J. (2013). *Stochastic Geometry and Its Applications*, 3rd ed. Wiley, Chichester. [MR3236788](#)
- COWLES, M. K. and CARLIN, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Amer. Statist. Assoc.* **91** 883–904. [MR1395755](#)
- DELLAERT, F., SEITZ, S. M., THORPE, C. E. and THRUN, S. (2003). EM, MCMC, and chain flipping for structure from motion with unknown correspondence. *Mach. Learn.* **50** 45–71.
- DICE, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* **26** 297–302.
- DIGGLE, P. (1985). A kernel method for smoothing point process data. *J. Roy. Statist. Soc. Ser. C* **34** 138–147.
- DIGGLE, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Edward Arnold, London.
- DIGGLE, P. J., EGLEN, S. J. and TROY, J. B. (2006). Modelling the bivariate spatial distribution of amacrine cells. In *Case Studies in Spatial Point Process Modeling. Lecture Notes in Statist.* **185** 215–233. Springer, New York. [MR2232131](#)
- GELLING, M. and COLE, A. (2000). *The Landscape of Place-names*. Shaun Tyas, Stamford, Lincolnshire.
- GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.* **1** 515–534.
- GELMAN, A. and RUBIN, D. (1992). Inference from Iterative Simulation using Multiple Sequences. *Statist. Sci.* **4** 457–511.
- GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.
- GRABARNIK, P., MYLLYMÄKI, M. and STOYAN, D. (2011). Correct testing of mark independence for marked point patterns. *Ecol. Model.* **222** 3888–3894.
- JANSON, S. and VEGELIUS, J. (1981). Measures of ecological association. *Oecologia* **49** 371–376.
- JERRUM, M. (2003). *Counting, Sampling and Integrating: Algorithms and Complexity*. Birkhäuser, Basel. [MR1960003](#)
- JERRUM, M. and SINCLAIR, A. (1996). The Markov chain Monte Carlo method: An approach to approximate counting and integration. In *Approximation Algorithms for NP-hard Problems* (D. S. Hochbaum, ed.) 482–520. PWS Publishing, Boston, MA.
- JONES, R. and SEMPLE, S. (2012). *Sense of Place in Anglo-Saxon England*. Shaun Tyas, Donington.
- KARPINSKI, M., RUCINSKI, A. and SZYMANSKA, E. (2012). Approximate Counting of Matchings in Sparse Uniform Hypergraphs. Preprint. Available at [arXiv:1204.5335](#).
- KUHN, H. W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2** 83–97. [MR0075510](#)
- LAU, J. W. and GREEN, P. J. (2007). Bayesian model-based clustering procedures. *J. Comput. Graph. Statist.* **16** 526–558. [MR2351079](#)
- LAWSON, A. B. and DENISON, D. G. T. (2010). *Spatial Cluster Modelling*. CRC press, Boca Raton.
- LOIZEAUX, M. A. and MCKEAGUE, I. W. (2001). Perfect sampling for posterior landmark distributions with an application to the detection of disease clusters. In *Selected Proceedings of the Symposium on Inference for Stochastic Processes (Athens, GA, 2000)*. Institute of Mathematical Statistics Lecture Notes—Monograph Series **37** 321–331. IMS, Beachwood, OH. [MR2002518](#)
- MARINARI, E. and PARISI, G. (1992). Simulated tempering: A new Monte Carlo scheme. *EPL (Europhysics Letters)* **19** 451–458.
- MÜLLER, P. and QUINTANA, F. (2010). Random partition models with regression on covariates. *J. Statist. Plann. Inference* **140** 2801–2808. [MR2651966](#)

- OH, S., RUSSELL, S. and SASTRY, S. (2009). Markov chain Monte Carlo data association for multi-target tracking. *IEEE Trans. Automat. Control* **54** 481–497. [MR2191542](#)
- PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2005). Output analysis and diagnostics for MCMC. R Package Version 0.10-3, URL: <http://cran.rproject.org>.
- ROBERTS, G. O. (1998). Optimal Metropolis algorithms for product measures on the vertices of a hypercube. *Stoch. Stoch. Rep.* **62** 275–283. [MR1613256](#)
- ROBERTS, G. O., GELMAN, A. and GILKS, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120. [MR1428751](#)
- VALIANT, L. G. (1979). The complexity of enumeration and reliability problems. *SIAM J. Comput.* **8** 410–421. [MR0539258](#)
- ZANELLA, G. (2015a). Supplement to “Bayesian complementary clustering, MCMC and Anglo-Saxon place-names.” DOI:[10.1214/15-AOAS884SUPPA](https://doi.org/10.1214/15-AOAS884SUPPA).
- ZANELLA, G. (2015b). Supplement to “Bayesian complementary clustering, MCMC and Anglo-Saxon place-names.” DOI:[10.1214/15-AOAS884SUPPB](https://doi.org/10.1214/15-AOAS884SUPPB).
- ZANELLA, G. (2015c). Supplement to “Bayesian complementary clustering, MCMC and Anglo-Saxon place-names.” DOI:[10.1214/15-AOAS884SUPPC](https://doi.org/10.1214/15-AOAS884SUPPC).
- ZANELLA, G. (2015d). Supplement to “Bayesian complementary clustering, MCMC and Anglo-Saxon place-names.” DOI:[10.1214/15-AOAS884SUPPD](https://doi.org/10.1214/15-AOAS884SUPPD).
- ZANELLA, G. (2015e). Supplement to “Bayesian complementary clustering, MCMC and Anglo-Saxon place-names.” DOI:[10.1214/15-AOAS884SUPPE](https://doi.org/10.1214/15-AOAS884SUPPE).
- ZANELLA, G. (2015f). Supplement to “Bayesian complementary clustering, MCMC and Anglo-Saxon place-names.” DOI:[10.1214/15-AOAS884SUPPF](https://doi.org/10.1214/15-AOAS884SUPPF).

DEPARTMENT OF STATISTICS  
UNIVERSITY OF WARWICK  
COVENTRY, CV4 7AL  
UNITED KINGDOM  
E-MAIL: [g.zanella@warwick.ac.uk](mailto:g.zanella@warwick.ac.uk)