SUPPLEMENT D TO "BAYESIAN COMPLEMENTARY CLUSTERING, MCMC AND ANGLO-SAXON PLACENAMES": MODEL EXTENSIONS AND VARIATIONS

By Giacomo Zanella*

University of Warwick

This supplementary material contains extensions of the null-hypothesis of Section 2.3.1 of Zanella (2014) and of the Bayesian complementary clustering model defined in Section 3 of Zanella (2014).

1. Null-hypothesis using Strauss point processes. In Section 2.3.1 of Zanella (2014) we defined the following null hypothesis for the distribution of the marked point process $\mathbf{x}$ under consideration: each point pattern $\mathbf{x}^{(j)}$ is an inhomogeneous Poisson point process with intensity function $\lambda_j(\cdot)$. Here $\mathbf{x}^{(j)}$ denotes the type j subpattern of points. In order to make such a null hypothesis more realistic we could introduce some repulsion among points of the same type. In fact it is reasonable to expect settlements with the same placename not to be too close to each other. This could be modeled by assuming that each point pattern $\mathbf{x}_j$ is distributed according to an inhomogeneous Strauss process, and $\mathbf{x}^{(j)}$ is independent from $\mathbf{x}^{(i)}$ for i different from j. A Strauss point process $\mathbf{x}^{(j)} = \{x_1^{(j)}, \dots, x_{n(\mathbf{x}_j)}^{(j)}\}$ has probability density

$$f(\mathbf{x}^{(j)}) = \alpha\, \gamma^{s(\mathbf{x}^{(j)})} \prod_{i=1}^{n(\mathbf{x}^{(j)})} \lambda_j(x_i^{(j)})\,,$$

with respect to the distribution of a unitary homogeneous Poisson point process. Here α is a normalizing constant, γ is a inhibition parameter between 0 and 1, $s(\mathbf{x}^{(j)})$ is the number of (unordered) couples of points in $\mathbf{x}^{(j)}$ closer than some distance $R > 0$ apart, and $\lambda_j(\cdot)$ is the intensity function. See Stoyan, Kendall and Mecke (1987) for more rigorous definitions of the Strauss process and other Gibbs-type point processes.

We then perform the same approximate Monte Carlo test of Section 2.3.1 of Zanella (2014), replacing the inhomogeneous Poisson point process model with the Strauss one (the estimated intensities $\hat{\lambda}_j(\cdot)$ are obtained through Gaussian kernel smoothing as describe there). In order to perform such a test we need to choose the values of the inhibition parameter γ and the maximal

*Supported by EPSRC through a PhD position under the CRiSM grant EP/D002060/1

inhibition distance R determining the distribution of the Strauss process. We considered γ equal to 0.1, 0.5 and 0.9 (corresponding to strong, medium and mild interaction). Given the historical context we considered values of the inhibition distance R equal to 5, 10 and 20 km. We tried all the 9 resulting combinations of γ and R. The results did not change significantly from the ones obtained in Section 2.3.1 of Zanella (2014) using the inhomogeneous Poisson point process model. Figure 1 shows the result obtained using $\gamma = 0.1$ and $R = 20$ (the strongest interaction among the ones we considered). Note that the 95% envelopes with such a null hypothesis are very similar to the ones obtained in Section 2.3.1 of Zanella (2014).

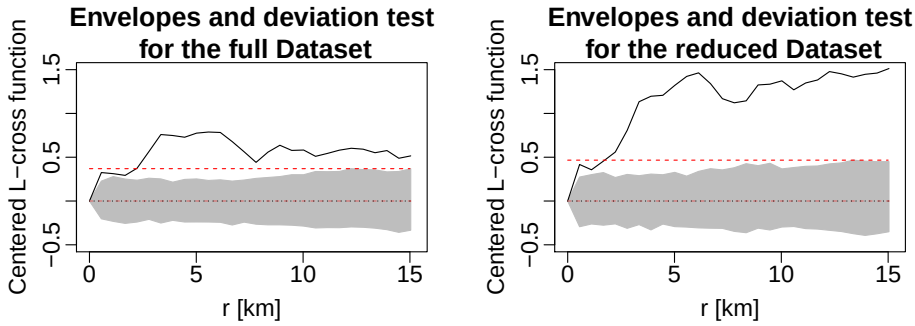

FIG 1. *Testing the null hypothesis of Section 1, based on Strauss point processes, with the procedure described in Section 2.3.1 of Zanella (2014). Black solid lines represent the centred L-function for the observed pattern. The* 95% *envelopes (gray areas) are obtained using 99 simulated patterns and 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* $\alpha = 0.05$.

2. Dropping the uniform marks assumption. In Section 3 of Zanella (2014) we defined the Bayesian random partition model we used to analyze the Anglo-Saxon settlements dataset. In particular, when defining the likelihood function (Section 3.2 of Zanella, 2014) we assume that, given the number of points s in a cluster $\mathbf{x}_C$, the marks $m_1, \ldots, m_s$ of such points are sampled uniformly from the set

$$\mathcal{M}_s = \left\{ \{m_1, \ldots, m_s\} \subseteq \{1, \ldots, k\} \mid m_{l_1} \neq m_{l_2} \text{ for } l_1 \neq l_2 \right\}. \tag{2.1}$$

Since the cardinality of $\mathcal{M}_S$ is $\binom{k}{s}$, this leads to the term

$$\frac{1}{\binom{k}{s}} \prod_{l_1 \neq l_2} \mathbb{1}(m_{l_1} \neq m_{l_2}) \tag{2.2}$$

in the likelihood function $h_{(s,\sigma)}(\mathbf{x}_C)$ given in (3.2) of Zanella (2014).

Nevertheless, as already mentioned in Remark 1 of Zanella (2014), the assumption of the marks being sampled uniformly seems unrealistic because of the heterogeneity in the number of settlements across different placenames (see Table 2 of Zanella, 2014). In this section we develop a model where the marks within each cluster are sampled non uniformly.

Suppose we have a probability vector $\mathbf{p}^{(m)} = (p_1^{(m)}, \dots, p_k^{(m)})$ on the set of possible marks $\{1, \dots, k\}$, with $p_i > 0$ for any i and $\sum_{i=1}^k p_i^{(m)} = 1$. Then, given the number of points s in a cluster $\mathbf{x}_C$, the marks $m_1, \dots, m_s$ are independently sampled from $\{1, \dots, k\}$ according to $\mathbf{p}^{(m)}$, conditioning on all the marks being different among themselves. Therefore the probability of a certain configuration $m_1, \dots, m_s$ is

$$\frac{p_{m_1}^{(m)} \cdots p_{m_s}^{(m)}}{Z_s} \prod_{l_1 \neq l_2} \mathbb{1}(m_{l_1} \neq m_{l_2}), \tag{2.3}$$

where Z_s is a normalizing constant defined as

$$Z_s = Z_s(\mathbf{p}^{(m)}) = \sum_{\{a_1, \dots, a_s\} \in \mathcal{M}_s} p_{a_1}^{(m)} \cdots p_{a_s}^{(m)}. \tag{2.4}$$

Note that if the probability vector $\mathbf{p}^{(m)}$ is uniform then (2.3) equals (2.2). Replacing (2.2) with (2.3) in the likelihood function (3.2) of Zanella (2014) we obtain the new likelihood function

$$h_{(s,\sigma)}(\mathbf{x}_C) = \frac{p_{m_1}^{(m)} \cdots p_{m_s}^{(m)} \, g(\overline{x}_C) \prod_{l_1 \neq l_2} \mathbb{1}(m_{l_1} \neq m_{l_2})}{Z_s \, s \, (2\sigma^2)^{s-1}} \exp\left(-\frac{\pi \delta_C^2}{2\sigma^2}\right), \tag{2.5}$$

where, as in Section 3.2 of Zanella (2014), $\overline{x}_C$ is the Euclidean barycenter of $\mathbf{x}_C$ and $\delta_C^2 = \sum_{i \in C} (x_i - \overline{x}_C)^\top (x_i - \overline{x}_C)$.

Since $\mathbf{p}^{(m)}$ is unknown, the standard bayesian approach would be to define a prior distribution on $\mathbf{p}^{(m)}$ and to consider the joint posterior distribution of $\mathbf{p}^{(m)}$ and the other unknown quantities. In order to explore such a posterior distribution, one should add to the MCMC algorithm of Section 4 of Zanella (2014) a Metropolis-Hastings step updating $\mathbf{p}^{(m)}$. This step would require the evaluation of the normalizing constants $Z_1(\mathbf{p}^{(m)})$ up to $Z_k(\mathbf{p}^{(m)})$ defined in (2.4) for the proposed value of $\mathbf{p}^{(m)}$. Note that the evaluation of $Z_s(\mathbf{p}^{(m)})$ is costly because its definition involves a summation over all the elements of $\mathcal{M}_s$. Expressing $Z_s(\mathbf{p}^{(m)})$ as the permanent of an appropriate $k \times k$ matrix, we could use Ryser's algorithm (Ryser, 1963), whose complexity is of order $O(2^k k)$. This allows us to evaluate $Z_1(\mathbf{p}^{(m)})$ up to $Z_k(\mathbf{p}^{(m)})$ but the cost is

too high to perform such evaluation at each MCMC step (the step updating $\mathbf{p}^{(m)}$ would dominate the others in terms of computational cost, making the algorithm too expensive).

In order to circumvent this problem we replace $\mathbf{p}^{(m)}$ with a plug-in estimator, in an empirical Bayes fashion. In such a way, the posterior distribution will not account for the uncertainty over $\mathbf{p}^{(m)}$. Nevertheless this will allow us to understand what is the impact of using a non-uniform $\mathbf{p}^{(m)}$ over the estimates of the quantities of interests (e.g. σ and $\mathbf{p}^{(c)}$) in a computationally feasible way. A natural estimator for the probability of the i-th mark, $p_i^{(m)}$, is the number of points with such a mark divided by the total number of points, $\frac{n_i(\mathbf{x})}{n(\mathbf{x})}$. We performed posterior inferences setting $p_i^{(m)} = \frac{n_i(\mathbf{x})}{n(\mathbf{x})}$ for i running from 1 to k and replacing the likelihood (3.2) of Zanella (2014) with the non-uniform version in (2.5). The results are in accordance with each other, although there are some differences (see Figure 2). In particular the ones obtained with the uniform marks assumptions are more conservative, meaning that they produce less clustering. The results presented in Zanella (2014) use the uniform marks assumption for simplicity and because it produces more conservative results.

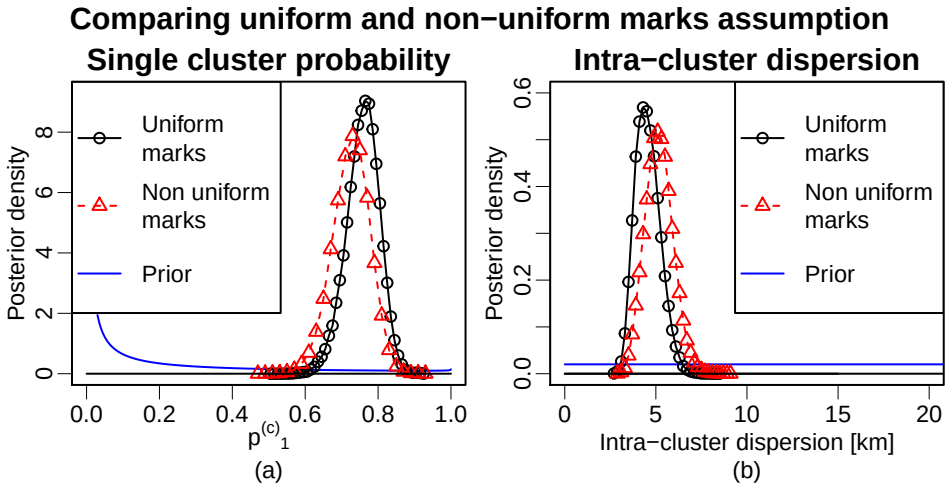

FIG 2. *Comparison of the posterior distributions of (a) $p_1^{(c)}$ and (b) σ, obtained with and without the assumption of the marks being sampled uniformly (see Section 2 for details).*

Finally we tested the sensitivity of the results to the choice of the plug-in estimator $p_i^{(m)} = \frac{n_i(\mathbf{x})}{n(\mathbf{x})}$. In particular we sampled perturbed values $(\tilde{n}_1, \dots, \tilde{n}_k)$ according to a multinomial distribution $\text{Mult}(n(\mathbf{x}), \mathbf{p}^{(m)})$, with $p_i^{(m)} = \frac{n_i(\mathbf{x})}{n(\mathbf{x})}$,

and used the perturbed values $\tilde{p}_i^{(m)} = \frac{\tilde{n}_i}{n(\mathbf{x})}$ as plug-in estimator. The results with and without the perturbation were extremely similar.

3. Alternative model for the prior distribution of ρ. In Section 3.4.1 of Zanella (2014) we define a model, namely the Poisson model, for the prior distribution of the partition ρ. As mentioned in Remark 4 of Zanella (2014), we also consider another model for $\pi(\rho)$ based on the Dirichlet-Multinomial distribution. We now define such a model and we provide the results of the analysis of the Anglo-Saxon settlements dataset obtained using the Dirichlet-Multinomial model. The results are almost equivalent to the ones obtained with the Poisson model. In Zanella (2014) we preferred to use the Poisson model because its posterior distribution factorizes over the clusters and this simplifies drastically the computations needed at each MCMC step.

3.1. *Dirichlet-Multinomial Model for $\pi(\rho)$.* For l running from 1 to k, we define $N_l(\rho)$ as the number of clusters of ρ having size l and $Y_l(\rho) = l \cdot N_l(\rho)$ so that $Y_l(\rho)$ is the total number of points in all the clusters of size l. Note that $\sum_{l=1}^k Y_l(\rho) = n(\mathbf{x})$, where $n(\mathbf{x})$ is the number of points in the k-type point pattern $\mathbf{x}$. In this model the random vector $\mathbf{Y}(\rho) = (Y_1(\rho), \dots, Y_k(\rho))$ follows a Dirichlet-Multinomial distribution conditioned on Y_l being a multiple of l (for l running from 1 to k)

(3.1)
$$\Pr(Y_1 = y_1, \dots, Y_k = y_k) \propto \begin{cases} \frac{n!}{y_1! \cdots y_k!} p_1^{y_1} \cdots p_k^{y_k} & \text{if } \sum_{l=1}^k y_l = n \text{ and} \\ & \quad y_l \text{ is a multiple of } l, \\ 0 & \text{otherwise.} \end{cases}$$

We assume that the parameter vector $\mathbf{p} = (p_1, \dots, p_k)$ is unknown with prior distribution $\mathrm{Dir}(\alpha_1, \dots, \alpha_k)$. The resulting prior distribution of ρ given $\mathbf{p}$, recalling that we want such distribution to be exchangeable, is

$$\pi(\rho \mid \mathbf{p}) \quad \propto \quad \frac{1}{\eta(\rho)} \frac{n(\mathbf{x})!}{Y_1(\rho)! \cdots Y_k(\rho)!} p_1^{Y_1(\rho)} \cdots p_k^{Y_k(\rho)}, \tag{3.2}$$

where $\eta(\rho) = \#\{\tilde{\rho} \mid \mathbf{Y}(\rho) = \mathbf{Y}(\tilde{\rho})\} = n! \left(\prod_{l=1}^k (l!)^{\frac{Y_l}{l}} (Y_l/l)! \right)^{-1}$.

Remark 1. *This model can be seen as a Dirichlet-Multinomial mixture of k classes having Y_1, Y_2, up to Y_k points corresponding to singletons, couples, up to k-tuples. We are therefore converting the problem of finding an unknown number (between $\frac{n}{k}$ and n) of small clusters into the problem of finding k big clusters, with k fixed and relatively small (20 in our case).*

REMARK 2. *Note that p_l represents the probability of a point being in a cluster of size l. Since we conditioned Y_l on being a multiple of l, though, this is just an approximation. Nevertheless for $n(\boldsymbol{x})$ big (e.g. $n(\boldsymbol{x}) \geq 10$) the approximation error is negligible.*

3.2. *Model parameters and Posterior Distribution.* The Dirichlet-Multinomial model results in the following unknown elements

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

where $\mathcal{P}_n$ is the set of all partitions of $\{1, \ldots, n\}$. Figure 3 provides graphical representations of the underlying conditional independence structure. Given the prior distribution described above and the likelihood distribution

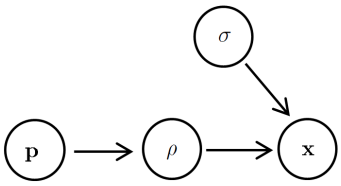

FIG 3. *Conditional independence of the random elements involved in the Dirichlet-Multinomial model.*

described in Section 3.2 of Zanella (2014) we obtain the following conditional posterior distributions for the Dirichlet-Multinomial model

$$\pi(\rho \mid \mathbf{x}, \sigma, \mathbf{p}) \ \propto \ \prod_{l=1}^{k} \frac{N_l!}{(lN_l)!} \cdot \cdot \prod_{j=1}^{N(\rho)} \left(\frac{g\left(\overline{x}_{C_j}\right) (p_{s_j})^{s_j}}{c_{s_j} \sigma^{2(s_j-1)}} \ \exp\left(-\frac{\pi \delta_{C_j}^2}{2\sigma^2}\right) \ \prod_{i,l \in C_j,\, i \neq l} \mathbb{1}(m_i \neq m_l) \right) , \tag{3.3}$$

$$\mathbf{p} \mid \mathbf{x}, \rho, \sigma \quad \sim \quad \mathrm{Dir}\left(\alpha_1 + Y_1(\rho), \ldots, \alpha_k + Y_k(\rho)\right) . \tag{3.4}$$

where $c_s = \binom{k}{s_j} s_j \, 2^{s_j - 1}$. Analogously to the Poisson model, the full conditional posterior distribution of σ, $\pi(\sigma|\mathbf{x}, \rho, \mathbf{p})$, depends only on σ, $\mathbf{x}$ and ρ and is given by (3.4) of Zanella (2014).

3.3. *Comparing the results obtained with the two models.* We used the MCMC algorithm described in Section 4 of Zanella (2014) to target the posterior distribution arising from the Multinomial-Dirichlet model when applied to the Anglo-Saxon settlements dataset. Figure 4 compares the results

obtained with the Dirichlet-Multinomial model to the ones obtained with the Poisson model, displayed in Section 5 of Zanella (2014). The posterior distributions obtained with the two models are very similar.

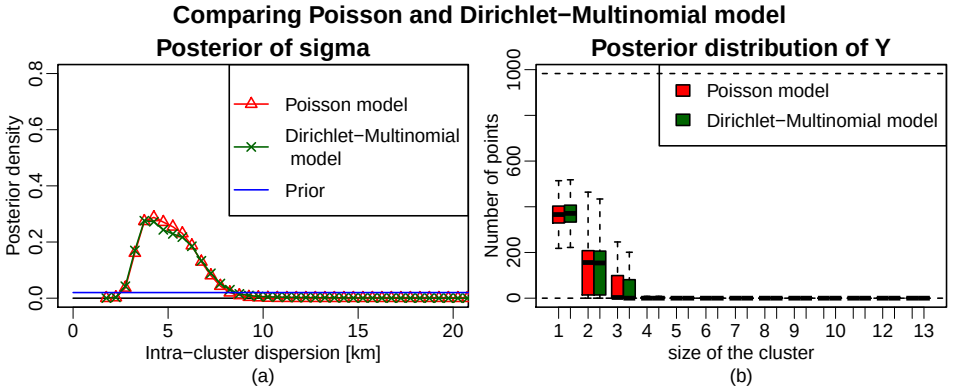

FIG 4. *Comparison of the the Dirichlet-Multinomial model (see Section 3.1) and the Poisson one (see Section 3.4.1 of Zanella, 2014). (a) Posterior distribution of σ and (b) posterior distribution of $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ (see Section 3.1 for a definition of $\boldsymbol{Y}$).*

References.

RYSER, H. J. (1963). *Combinatorial mathematics*. Washington: MAA.

STOYAN, D., KENDALL, W. and MECKE, J. (1987). *Stochastic geometry and its applications* **8**. Wiley Chichester.

ZANELLA, G. (2014). Bayesian Complementary Clustering, MCMC and Anglo-Saxon placenames. *arXiv preprint arXiv:1409.6994*.

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