ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2015)
178, *Part* 2, *pp.* 465–480

Partially supervised spatiotemporal clustering for burglary crime series identification

Brian J. Reich

North Carolina State University, Raleigh, USA

and Michael D. Porter

University of Alabama, Tuscaloosa, USA

[Received March 2013. Final revision April 2014]

Summary. Statistical clustering of criminal events can be used by crime analysts to create lists of potential suspects for an unsolved crime, to identify groups of crimes that may have been committed by the same individuals or group of individuals, for offender profiling and for predicting future events. We propose a Bayesian model-based clustering approach for criminal events. Our approach is semisupervised, because the offender is known for a subset of the events, and utilizes spatiotemporal crime locations as well as crime features describing the offender's *modus operandi*. The hierarchical model naturally handles complex features that are often seen in crime data, including missing data, interval-censored event times and a mix of discrete and continuous variables. In addition, our Bayesian model produces posterior clustering probabilities which allow analysts to act on model output only as warranted. We illustrate the approach by using a large data set of burglaries in 2009–2010 in Baltimore County, Maryland.

Keywords: Bayesian hierarchical model; Crime linkage; Markov chain Monte Carlo methods; Model-based clustering

1. Introduction

Advanced statistical methods are increasingly being used to aid in criminal investigations. One objective is to cluster crime events that share a common offender or group of co-offenders. As an investigative tool, the results from clustering will allow crime analysts to operate more efficiently and effectively by jointly investigating crimes that are likely to share a common offender instead of investigating each crime individually (Grubin *et al*., 2001; Woodhams *et al*., 2007). Clustering can also be useful for discovering previously unknown serial offenders or identifying additional crimes that are part of a series. In addition, practical interest is often focused on identifying who is responsible for a set of crimes (suspect prioritization). In this case, a set of crimes with unknown offender(s) is compared with the crimes that have been committed by a given set of suspects (perhaps from past offending records). This calls for a type of semisupervised clustering, where the known series of past offences are labelled (i.e. associated with their own cluster), providing a direct way to incorporate suspect information. New crimes can then be examined to determine how closely they are related to existing clusters corresponding to the crimes that have been committed by known individuals or groups. This can potentially provide information on the type of offender (offender profiling) and inform on future behaviour (next event prediction).

Address for correspondence: Brian J. Reich, Department of Statistics, North Carolina State University, 4264 SAS Hall, Box 8302, Raleigh, NC 27695, USA.
E-mail: bjreich@ncsu.edu

0964–1998/15/178465

One of the primary uses for criminal cluster analysis is in crime linkage. Crime linkage often refers to several connected but slightly different tasks, all of which may be employed in the course of a police investigation. First, crime linkage can refer to the process of linking crimes to other crimes (that share common offender(s)). We refer to *pairwise case linkage* as the processes of determining whether a given pair of crimes share the same offender or group of co-offenders. This is essentially a binary classification problem where each crime pair is considered independently. Extending beyond crime pairs, *crime series identification*, or series linkage, is the process of identifying the set of crimes that share a common offender or group of co-offenders. Crime linkage may also refer to the process of linking crimes to offenders. We refer to *suspect identification* as the process of identifying the offender or group of co-offenders that is responsible for a crime or set of crimes by comparing the unsolved crime(s) with the past offenders' crime series.

The majority of the research that is related to linkage analysis has focused on (pairwise) behavioural case linkage. Also known as comparative case analysis (Bennell and Canter, 2002), behavioural case linkage attempts to determine whether the same offender(s) committed two crimes on the basis of *modus operandi* evidence rather than physical or forensic evidence (e.g. DNA or fingerprint). The *modus operandi* behaviour that is exhibited by an offender can include, *inter alia*, aspects of their site selection (where they choose to commit an offence), the timing, methods for carrying out the crime and what they did during the commission of the crime. Practically, the premise of behavioural case linkage is that offenders behave consistently across their crime series and distinctively among the other criminals in the region (Canter, 2004; Woodhams *et al.*, 2007). As such, the main emphasis in case linkage research has been in determining how well the assumptions regarding consistency and distinctiveness hold. This has resulted in numerous studies detailing the performance of several case linkage methods across a variety of types of crime (Brown and Hagen, 2003; Bennell and Jones, 2005; Goodwill and Alison, 2006; Lin and Brown, 2006; Cocx and Kosters, 2006; Woodhams and Toye, 2007; Tonkin *et al.*, 2008, 2011, 2012; Bennell *et al.*, 2009; Markson *et al.*, 2010; Woodhams and Labuschagne, 2012).

Expanding consideration beyond crime pairs, crime series identification seeks to find all of the crimes corresponding to the same offender(s). This can be performed simultaneously for all crimes through clustering (Adderley and Musgrove, 2001; Ma *et al.*, 2010) for deriving groups of similar crimes for further investigation or even offender profiles. Alternatively, the focus can be on identifying all the crimes sharing a cluster with a particular crime or set of crimes. This could be useful for investigative or interrogative purposes by providing a list of additional crimes that a suspect may have committed. For example, Adderley and Musgrove (2003) and Adderley (2004) used data mining methods to identify all the unsolved crimes that were similar to the crimes perpetrated by a known criminal group and known single offender respectively.

Another practical use of crime linkage is in suspect identification. This task attempts to link crimes to offenders by comparing a particular crime with the crimes that are known to be perpetrated by a set of past offenders. As the focus is often on *suspect prioritization*, or developing a ranked list of suspects, several classification methods have been employed (Yokota and Watanabe, 2002; Santtila *et al.*, 2004, 2005, 2008; Ewart *et al.*, 2005; Snook *et al.*, 2006; Canter and Hammond, 2007; Salo *et al.*, 2013). By restricting the suspects to a known set of offenders, these approaches are essentially conditioning on the crime in question belonging to one of the known series.

This paper presents a Bayesian model-based clustering methodology that fuses spatial and temporal information with features of the crimes (e.g. the method of entry), crime scenes (e.g. the type of property) and offenders (e.g. their crime history) to identify crime clusters. Instead of an unsupervised approach (where no crimes are attributed to an offender), our partially supervised model can incorporate identifying information about the known offenders of crimes (i.e. the labels) when that information is available. Including this additional source of data will help to identify the true crime series accurately. Model-based clustering (Fraley and Raftery, 1998) is an attractive choice for crime events because it is straightforward to include both solved and unsolved crimes, it directly models the distribution of the crime features across the crimes of an offender and thus it allows for prior expert knowledge to be included in the clustering algorithm. Our hierarchical Bayesian model also naturally handles missing data and interval censoring (the time of an event is known only to be in an interval), which are very common features of crime data, and provides an assessment of uncertainty for all model parameters, including the relative influence of each feature (space, time, method of entry, etc.) in the model. Finally, our approach provides the posterior probability that each pair of crimes are linked. This allows crime analysts to act on the evidence only as appropriate. For example, our model provides a list of the 10 most likely criminals associated with an unsolved crime but, if no criminal is associated with probability higher than 0.05, then crime analysts may choose not to take action on the basis of this information.

2. Baltimore burglary data

To help to motivate the development of our model, this section describes some aspects of the crime data that we subsequently use for analysis. The study region is Baltimore County, Maryland, and a surrounding buffer region. Baltimore County has a land area of 598.30 miles^2 (1549.5 km^2), a 2010 population of 805029, providing a population density of 1345.5 people mile^{-2} (519.5 km^{-2}), an average of 561.0 housing units mile^{-2} (216.6 km^{-2}) and an average of 2.47 people per household (US Census Bureau, 2011).

The Baltimore County Police Department provided data on $n = 11\,524$ burglaries reported in 2009–2010. Each crime record includes a geocoded location, the time interval in which the crime could have occurred (the exact timing of the crime cannot be ascertained in many cases, in which case victims indicate the likely interval of time during which the crime could have occurred), the type of property, point of entry, method of entry and an anonymized offender identifier if it is available.

Fig. 1 shows the spatial distribution of the crime events. Baltimore County participates in the ‘Regional crime analysis program’ which facilitates the sharing of crime data across jurisdictional boundaries. As such, some crime events (0.7%) were located outside Baltimore County but, as they were suspected to be part of a series occurring inside the borders of the county, they were included in the data set and subsequently used in our analysis. Unfortunately, we do not have the information on the criteria that they used to select these cases. Only 47, or 0.4%, of the spatial co-ordinates were missing.

Because burglaries often occur when the victim is away from their property (Catalano, 2010), uncertainty is prevalent in the event timing. Thus, the crime records include the temporal interval in which the crime occurred (i.e. the event times are interval censored). Table 1 shows the empirical distribution for the time interval length. Note that only 20% of crimes have an exact time recorded and 48% have an interval shorter than 6 h. Because there is so much uncertainty, using only one point (e.g. earliest time, midpoint or latest time) to represent the event time could result in bias (Ratcliffe, 2002). In Section 3, we describe how uncertainty in the event times are incorporated in our model. This interval-censoring-based approach is slightly different from aoristic analysis (Ratcliffe, 2000, 2002). An aoristic analysis deals only with missing times and typically assumes that missing times are uniformly distributed; our analysis handles missing

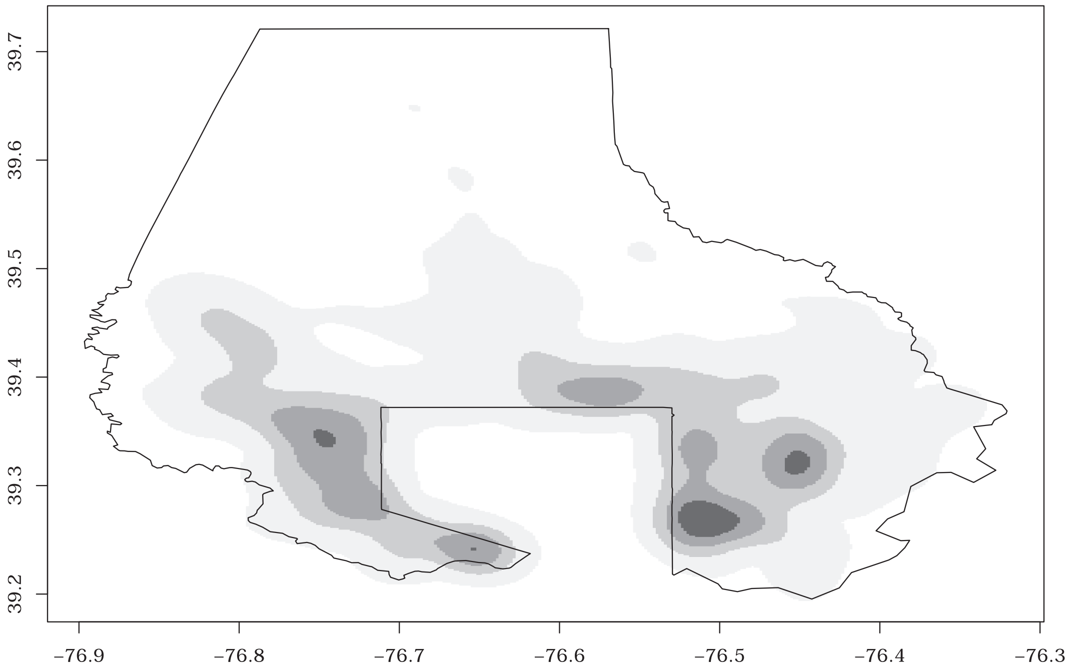


Fig. 1. Spatial density for burglaries in Baltimore County, Maryland, in 2009–2010

Table 1. Uncertainty in event timing: distribution for the length of the time window in which a crime could have occurred

	Results for the following lengths of time interval in which a crime could have occurred:						
	Exact	*⩽1 h*	*⩽6 h*	*⩽12 h*	*⩽24 h*	*⩽48 h*	*⩽120 h*
Number	2337	3701	5559	7758	9514	10082	10870
%	20	32	48	67	83	87	94

times as part of a larger analysis and thus uses more information to deal with missingness. Using such an interval censoring approach, we estimate the crime rate over the analysis period in Fig. 2 and the time-of-day distribution by the day of the week in Fig. 3.

Fig. 3 shows that there are clear differences between the weekday and weekend crime patterns. On weekdays, there is a spike in burglaries between the late morning and early afternoon. This spike is missing on the weekends which experience a rise in crime late at night. Because these patterns can possibly be attributed to different types of offenders (Coupe and Blake, 2006), we created a new variable indicating the time of day (day or night) and day of the week (weekday or weekend) to use in the model. We define daytime as 6 a.m.–8 p.m. and weekends as Friday after 8 p.m. until Sunday at midnight.

Besides the spatiotemporal information, each crime record also contains three categorical variables. The *type of property* took values ‘other’, 27%, ‘single home’, 24%, ‘apt/condo’, 13%, ‘yard’, 13%, ‘rowhome/townhouse’, 12%, and ‘shed/garage’, 11%. The *point of entry* took values

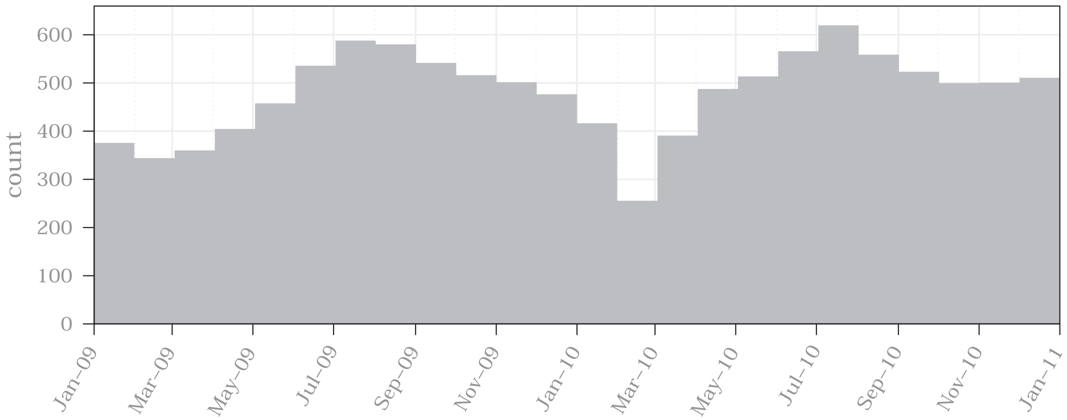
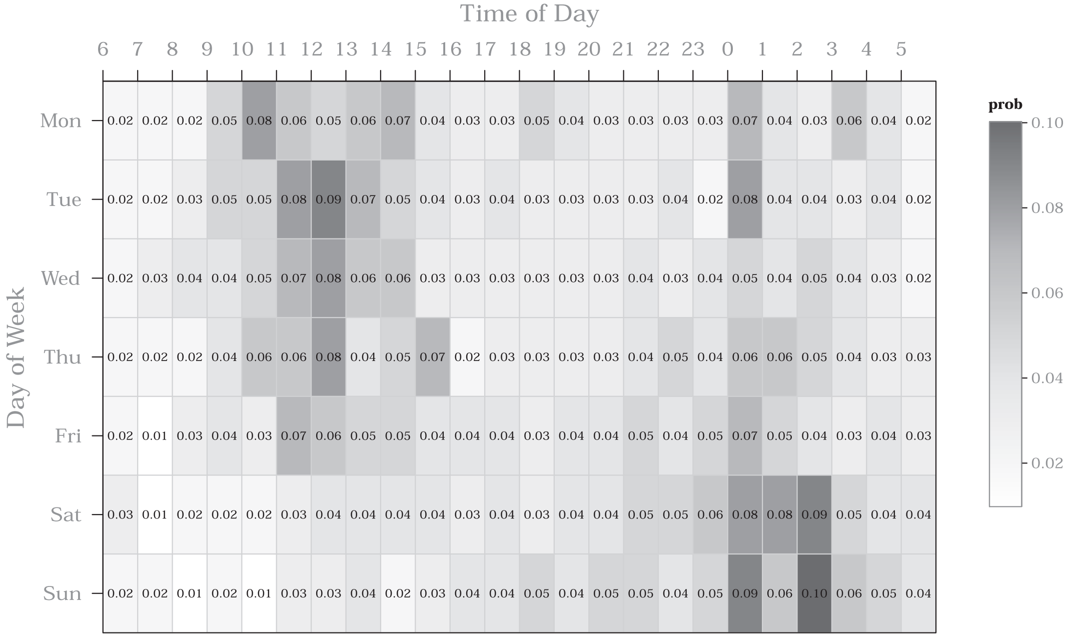


Fig. 2. Temporal histogram of burglaries in Baltimore County, Maryland (bin width 30.4 days): the large drop in February 2010 is probably the result of a series of record breaking snowstorms



	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	0	1	2	3	4	5
Mon	0.02	0.02	0.02	0.05	0.08	0.06	0.05	0.06	0.07	0.04	0.03	0.03	0.05	0.04	0.03	0.03	0.03	0.03	0.07	0.04	0.03	0.06	0.04	0.02
Tue	0.02	0.02	0.03	0.05	0.05	0.08	0.09	0.07	0.05	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.04	0.02	0.08	0.04	0.04	0.03	0.04	0.02
Wed	0.02	0.03	0.04	0.04	0.05	0.07	0.08	0.06	0.06	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.04	0.03	0.04	0.05	0.04	0.05	0.04	0.03	0.02
Thu	0.02	0.02	0.02	0.04	0.06	0.06	0.08	0.04	0.05	0.07	0.02	0.03	0.03	0.03	0.03	0.04	0.05	0.04	0.06	0.06	0.05	0.04	0.03	0.03
Fri	0.02	0.01	0.03	0.04	0.03	0.07	0.06	0.05	0.05	0.04	0.04	0.04	0.03	0.04	0.04	0.05	0.04	0.05	0.07	0.05	0.04	0.03	0.04	0.03
Sat	0.03	0.01	0.02	0.02	0.02	0.03	0.04	0.04	0.04	0.04	0.03	0.04	0.03	0.04	0.04	0.05	0.05	0.06	0.08	0.08	0.09	0.05	0.04	0.04
Sun	0.02	0.02	0.01	0.02	0.01	0.03	0.03	0.04	0.02	0.03	0.04	0.04	0.05	0.04	0.05	0.05	0.04	0.05	0.09	0.06	0.10	0.06	0.05	0.04

Fig. 3. Time-of-day distribution for burglaries in Baltimore County, Maryland (2009–2010), conditioned on the day of the week (i.e. each row sums to 1)

‘door’, 45%, ‘window’, 21%, ‘none’ (the event occurred outside the home), 8%, ‘other’, 6%, and ‘missing’, 19%. The *method of entry* took values ‘no force’, 28%, ‘other’, 20%, ‘forced’, 16%, ‘pried’, 10%, ‘broke glass’, 9%, and ‘missing’, 17%. To reduce the number of category levels, for each of the three factors we recategorized the levels with the smallest counts as other. Lumping rare (unconventional) events together into an other category is certainly not necessary, but it reduces the uncertainty that comes from estimating many small probabilities. In our method, two crimes with the same category, including other as the response for the same feature (say

Table 2. Number of offenders for each cluster of crimes known to be committed by the same individual or group of individuals

	Results for the following numbers of offenders per group:			
	1	*2*	*3*	*>3*
Number	1130	303	112	31
%	72	19	7	2

Table 3. Number of crimes for each cluster of crimes known to be committed by the same individual or group of individuals

	Results for the following numbers of crimes per group:					
	1	*2*	*3*	*4*	*5*	*>5*
Number	1329	136	46	22	12	31
%	84	9	3	1	1	2

point of entry) is taken as evidence of similarity. However, an other for point of entry and an other for method of entry would not be used as evidence of similarity.

Of the $n = 11524$ burglaries, 2264 are considered solved (i.e. an arrest was made and an offender identifier has been assigned to the crime). The solved crimes were perpetrated by a total of 2032 different offenders (assuming that each unique offender identifier corresponds to one individual). However, several crimes had multiple offenders. For use in our clustering model, the solved crimes must be preassigned to groups. We choose to define a group as a unique set of offenders. Thus all the crimes that offender A and offender B committed together are grouped together. However, if some crimes were committed by offender A only or if some crimes were committed with the help of offender C, then they would be in separate groups. Although this leads to some group overlap (i.e. an offender can belong to multiple groups), it prevents the grouping together of long series of crimes from multiple offenders. Approaches other than declaring all unique combinations of offenders as separate groups would be very complex. One could imagine that each individual offender has certain preferences, and a group of offenders takes some combination of its individuals' preferences. But specifying a model for this would require strong assumptions. In total, this gave 1576 groups. The distributions of group size (number of offenders) and crimes per group are given in Tables 2 and 3. The majority of groups are comprised of a single offender (72%) and 16% of groups have some overlap with another group. Also, most groups are known to have committed only a single crime (84%).

3. Model-based clustering for burglary data

Denote the spatial location and time for crime $i = 1, \ldots, n$ as $\mathbf{s}_i$ and T_i respectively. The time is often not known exactly. In these cases, T_i is interval censored, so we know only that $T_i \in [L_i, U_i]$. In addition to spatiotemporal information, crimes are associated with features

$\mathbf{X}_i=(X_{i1},\ldots,X_{ip})$ that describe the circumstances of the crime in terms of the type of property, point of entry, method of entry and time of day. For our data all these features are categorical so $X_{ij}\in\{1,\ldots,N_j\}$, where N_j is the number of levels for feature j.

For solved crimes, we have the identification number of the cluster or criminal (or group of criminals) G_i; for unsolved crimes G_i is missing. We assume that the cluster labels $G_i\in\{1,\ldots,M\}$, where M is the upper bound on the number of criminal groups responsible for the n crimes. Denote the prior probability of crime i being attributed to cluster g as $\mathrm{Prob}(G_i=g)=\pi_g$. The probability π_g determines the expected proportion of the crimes in the data set that are committed by cluster g. We model these probabilities as $(\pi_1,\ldots,\pi_M)\sim\mathrm{Dirichlet}(D_0,\ldots,D_0)$. $D_0>0$ controls the variation in the number of crimes in each cluster. If D_0 is large, all clusters account for roughly the same proportion of crimes and, if D_0 is small, many clusters are empty and some have a large number of crimes. An assumption underlying this model is that the likelihood of solving each crime is the same, i.e. that the cluster labels are missing completely at random. This assumption would be violated, for example, if certain criminals are simply better at evading detection than others. Though it is difficult to determine whether this assumption holds in practice, our test set validation suggests robustness to this assumption for the burglary analysis in Section 4.

We model the remaining variables conditioned on the cluster index. The spatial and temporal distribution of the events for each cluster are assumed to be Gaussian:

$$\begin{aligned}\mathbf{s}_i|G_i=g&\sim N(\boldsymbol{\mu}_g,\sigma_s^2 I_2),\\ T_i|G_i=g&\sim N(\theta_g,\sigma_t^2),\end{aligned}\tag{1}$$

where all spatial locations are projected so that $\|\mathbf{s}_i-\boldsymbol{\mu}_g\|$ is given in kilometres. Censoring of the times T_i is handled by using latent variable methods as described in Appendix A. We denote the probability of a crime being level $k\in\{1,\ldots,N_j\}$ of feature j for cluster g as $\mathrm{Prob}(X_{ij}=k|G_i=g)=P_{kjg}$. Conditioned on the cluster indictor G_i, the variables $\mathbf{s}_i$, T_i and X_{ij} are mutually independent.

The spatiotemporal variability across clusters is determined by the cluster means $\boldsymbol{\mu}_g$ and θ_g. The cluster means are modelled as

$$\begin{aligned}\boldsymbol{\mu}_g&\overset{\mathrm{IID}}{\sim}N(\bar{\boldsymbol{\mu}},\tau_s^2 I_2),\\ \theta_g&\overset{\mathrm{IID}}{\sim}N(\bar{\theta},\tau_t^2).\end{aligned}$$

In this model the marginal (over cluster label) distribution of the spatiotemporal data $(\mathbf{s}_i,T_i)$ follows a flexible mixture of normal distributions model. The variances σ_s^2 and τ_s^2 (and similarly σ_t^2 and τ_t^2) control the proportion of variability attributed to variability within (σ_s^2) and across (τ_s^2) cluster centres. If $\tau_s^2=0$ and all clusters have the same mean, then spatial information does not help to distinguish between the clusters and therefore is not useful for cluster identification. Therefore, we inspect the posterior of $r_s=\mathrm{var}(\mathbf{s}_i|G_i)/\mathrm{var}(\mathbf{s}_i)=\sigma_s^2/(\sigma_s^2+\tau_s^2)$ to quantify the contribution of the spatial (and similarly temporal) data.

The variation in the probabilities for the categorical covariates across clusters is also modelled hierarchically. Let $\mathbf{P}_{jg}=(P_{1jg},\ldots,P_{N_jjg})$ be the probabilities for cluster g for the N_j levels of feature j. We assume

$$\mathbf{P}_{jg}\sim\mathrm{Dirichlet}(D_j\bar{P}_{1j},\ldots,D_j\bar{P}_{N_jj}),\tag{2}$$

where $\bar{P}_{1j},\ldots,\bar{P}_{N_jj}$ are the means of the probabilities across clusters, $\Sigma_{k=1}^{N_j}\bar{P}_{kj}=1$, and $D_j>0$ controls the variability of the probabilities across clusters, with $\mathrm{var}(P_{kjg})=\bar{P}_{kj}(1-\bar{P}_{kj})/(D_j+1)$.

If D_j is large, then the probabilities for covariate j are similar for all clusters and thus covariate j does not help with clustering. In contrast, if D_j is small, then $\mathbf{P}_{jg}$ varies considerably across clusters, and covariate j helps to separate observations into clusters.

For priors, we use uninformative gamma priors $\sigma_s^{-2}, \sigma_t^{-2}, \tau_s^{-2}, \tau_t^{-2}, D_j \sim^{\mathrm{IID}}$ gamma(a, b) with $a = b = 0.1$. The overall means $\bar{\mu}$ and $\bar{\theta}$ have uniform priors over their domains. The spatial domain was taken to be the rectangle defined by the range of observed latitudes and longitudes. We experimented with uninformative priors for the mean probabilities $\bar{P}_{kj}$, but this led to poor convergence. Therefore, we elected to fix these hyperparameters at the sample proportions $\bar{P}_{kj} = \Sigma_{i=1}^{n} I(X_{ij} = k)/n$.

4. Analysis of Baltimore burglaries

In this section, we analyse the Baltimore burglary data that were described in Section 2. We begin by comparing several versions of the model in Section 4.1 to illustrate the relative importance of various model features. We then summarize the parameters in the final model in Section 4.2. To illustrate how this analysis could be used in practice, we perform crime series identification and crime classification in Sections 4.3 and 4.4 respectively.

4.1. Model comparisons

We compare clustering models by using test set validation. We evaluate models' performance to replicate classification of crimes that are known to be from the same cluster. Observations with known cluster index were randomly selected with probability 0.2 to be assigned to the test set, and the remaining observations with known labels and all observations with missing labels were assigned to the training set. After this initial randomization, observations in the test set with no observations in the training set from the same cluster were reassigned to the training

Table 4. Summary of test set clustering performance†

Model	*P cor*	*P rank 1*	*P rank 5*	*P rank 10*	*P rank 25*
(a) All observations					
Spatial	0.043	0.108	0.450	0.526	0.638
Spatiotemporal	0.130	0.101	0.584	0.654	0.723
Full	0.226	0.087	0.599	0.664	0.749
(b) 1 crime in the training set					
Spatial	0.009	0.019	0.068	0.117	0.374
Spatiotemporal	0.028	0.068	0.366	0.532	0.672
Full	0.045	0.091	0.374	0.494	0.630
(c) Two 10 crimes in the training set					
Spatial	0.039	0.139	0.601	0.715	0.783
Spatiotemporal	0.140	0.136	0.689	0.727	0.766
Full	0.250	0.109	0.684	0.727	0.791
(d) More than 10 crimes in the training set					
Spatial	0.102	0.167	0.661	0.688	0.694
Spatiotemporal	0.253	0.070	0.661	0.667	0.699
Full	0.433	0.032	0.731	0.763	0.785

†'*P* cor' is the average posterior probability that observations are assigned to the correct cluster, and '*P* rank *R*' is the proportion of the observations where the correct cluster is one of the *R* clusters with highest posterior probability.

Table 5. Summary of parameters in the final fit

Parameter	*Mean*	*90% interval*
Within-cluster spatial standard deviation (km), σ_s	1.072	(1.053, 1.093)
Between-cluster spatial standard deviation (km), τ_s	5.759	(5.624, 5.901)
Spatial variance ratio $r_s = \sigma_s^2/(\sigma_s^2 + \tau_s^2)$	0.034	(0.032, 0.036)
Within-cluster temporal standard deviation (days), σ_t	66.466	(64.224, 68.725)
Between-cluster temporal standard deviation (days), τ_t	191.061	(184.566, 197.910)
Temporal variance ratio, $r_t = \sigma_t^2/(\sigma_t^2 + \tau_t^2)$	0.175	(0.165, 0.185)
D_1, property type	4.058	(3.878, 4.232)
D_2, point of entry	6.931	(6.448, 7.455)
D_3, method of entry	5.039	(4.718, 5.359)
D_4, time period	21.919	(18.419, 25.760)
Cluster-specific proportions of property type, P_{k1g}		
Other	0.275	(0.024, 0.657)
Residential yard	0.125	(0.001, 0.453)
Single home	0.236	(0.014, 0.611)
Town house	0.126	(0.001, 0.443)
Apartment	0.131	(0.001, 0.448)
Garage	0.106	(0.000, 0.399)
Cluster-specific proportions of point of entry, P_{k2g}		
Door	0.556	(0.260, 0.837)
Other	0.081	(0.001, 0.290)
Window	0.262	(0.051, 0.549)
None	0.102	(0.002, 0.325)
Cluster-specific proportions of method of entry, P_{k3g}		
Other	0.246	(0.026, 0.587)
Broke glass	0.106	(0.001, 0.377)
Pried	0.123	(0.001, 0.402)
Forced physically	0.190	(0.010, 0.506)
No force required	0.334	(0.065, 0.682)
Cluster-specific proportions of time period, P_{k4g}		
Weekday–day	0.554	(0.383, 0.721)
Weekday–night	0.223	(0.095, 0.382)
Weekend–day	0.101	(0.022, 0.222)
Weekend–night	0.121	(0.031, 0.251)

set. This was repeated independently five times and the results below are averaged over the five test set data sets.

For each training set, we apply three special cases of the clustering model that was described in Section 3. The first model ('Spatial') uses only the spatial co-ordinates $\mathbf{s}_i$ and discards temporal information T_i and other features $\mathbf{X}_i$. The second model ('Spatiotemporal') includes both space and time but ignores the other features. The full model ('Full') uses all available information as described in Section 3.

For each model and test set observation, let p_{ik} be the posterior probability that observation i is assigned to cluster k and denote G_i as the true cluster for observation i. We compare models by using the average of p_{iG_i}, i.e. the posterior mean probability of the correct cluster ('P cor'), and the probability that p_{iG_i} is ranked in the top R of $\{p_{i1}, \ldots, p_{iM}\}$ ('P rank R'). By comparing results across multiple R, we are essentially computing the receiver operating characteristic statistics that were recommended by Bennell *et al.* (2009). These statistics are averaged over the test set observations and the five random splits of the data. They are presented separately on the basis of the number of observations in the training data with cluster index equal to G_i.

To select the maximum number of clusters M, we fit the spatiotemporal model with $M = 2000$,

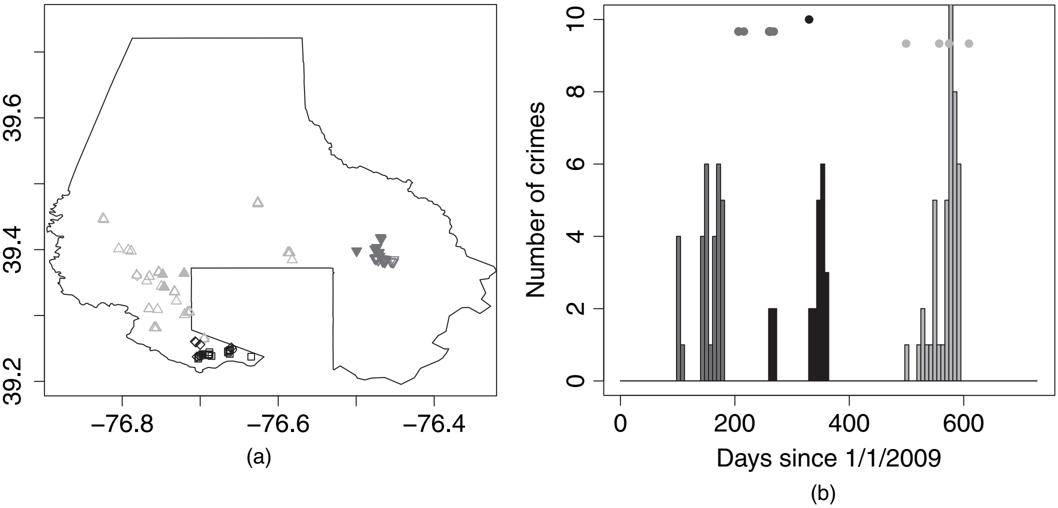


Fig. 4. (a) Spatial locations (○, other; □, single home; ◇, town house; △, apartment; ∇, garage) and (b) times of crimes from the largest three clusters, as well as unsolved crimes with posterior probability at least 0.8 of being from one of the three largest clusters: the spatial locations of solved crimes are indicated by empty symbols and unsolved crimes by filled symbols; the times (midpoints of plausible intervals) of solved crimes are shown as histograms and the times of unsolved crimes are represented by points

4000, 6000. We found that $M=4000$ and $M=6000$ were comparable; for example, the probability of the correct cluster ('P cor') averaged over all test set observations was 0.119, 0.130 and 0.132 for $M=2000$, $M=4000$ and $M=6000$ respectively. Therefore, we used $M=4000$ for the remaining analysis. Table 4 presents the model comparison results. Over all observations, the posterior probability of the correct cluster increases from 0.043 for the spatial model, to 0.130 for the spatiotemporal model and to 0.236 for the full model. Therefore, accounting for both the event time and other features substantially improves cluster detection.

Although the overall probability of the correct cluster is fairly low, the full model includes the correct cluster in the top 5 59.9% of the time and the top 25 74.9% of the time. Not surprisingly, the probability of the correct cluster increases with the number of observations from the cluster in the training set, as the characteristics of these criminals are estimated more precisely with a large sample. The probability of the correct cluster is very low for crimes with only a single observation in the training set (0.045 for the full model), but even for this challenging case the probability that the true cluster is ranked in the top 25 is 0.630, which may provide investigators useful information. Because of the way that we defined groups to be the unique combination of co-offenders, several clusters will contain the same subset of offenders. Thus, these results underestimate the performance of the model to identify at least one of the offenders who were involved in the crime.

4.2. Summary of the final model

The posteriors of the full model parameters from the fit to the complete data set are summarized in Table 5. These results show the relative importance of each of the clustering features. The posterior mean of the within-cluster spatial standard deviation is $\hat{\sigma}_s^2=1.072$ km. This is fairly tight spatial clustering relative to the size of the spatial domain; the posterior mean ratio of within-cluster variance to total variance is $\hat{r}_s=0.034$. There is also evidence of temporal clustering, even in this fairly narrow time window. The posterior mean of the within-cluster temporal

Table 6. Data relevant to the crime classification analysis†

	Probability	*Property type*	*Point of entry*	*Method of entry*	*Time*
Crime 1		*Other*	*Window*	*Broke glass*	*Weekday–day*
Rank 1	0.091	Other	Door	Broke glass	Weekend–day
Rank 2	0.036	Other	Missing	Missing	Missing
Rank 3	0.031	Other	None	No force	Weekend–day
Rank 4	0.028	Apartment	Window	Pried	Weekday–night
Rank 5	0.025	Other	Door	No force	Weekday–day
Crime 2		*Single home*	*Door*	*Physically forced*	*Weekday–day*
Rank 1	0.127	Single home	Door	Pried	Weekday–day
		Single home	Window	Broke glass	Weekday–day
		Single home	Door	Other	Weekday–day
Rank 2	0.126	Garage	Missing	No force	Missing
		Single home	Door	Broke glass	Weekday–day
		Single home	Window	Other	Weekday–day
		Town house	Window	Broke glass	Weekday–day
Rank 3	0.101	Single home	Other	Physically forced	Missing
		Single home	Door	Pried	Weekday–day
Rank 4	0.062	Single home	Window	Broke glass	Weekday–day
		Apartment	Door	Pried	Weekday–day
Rank 5	0.055	Apartment	Door	Physically forced	Weekday–day
		Apartment	Door	Physically forced	Weekday–day
		Apartment	Door	Physically forced	Weekday–day
Crime 3		*Yard, residential*	*Missing*	*No force*	*Weekend–day*
Rank 1	0.672	Yard, residential	Missing	No force	Weekday–day
		Yard, residential	None	No force	Weekday–day
Rank 2	0.133	Town house	Missing	Missing	Weekend–night
		Yard, residential	None	No force	Weekday–night
Rank 3	0.006	Yard, residential	Missing	No force	Weekday–day
Rank 4	0.005	Yard, residential	Missing	No force	Weekday–night
Rank 5	0.005	Town house	Window	No force	Missing
		Garage	Door	Other	Missing

†For each unsolved crime, the data for the jth most likely cluster of solved crimes is denoted 'Rank j'.

standard deviation and the ratio of within-cluster variance to total variance are $\hat{\sigma}_t^2 = 66$ days and $\hat{r}_t = 0.175$ respectively. Therefore, a typical criminal commits 95% of their crimes within a 264-day period.

The posterior means of the parameters that control variability of the distributions of event features across clusters, D_j, range from 4.058 for property type to 21.919 for the time period. To illustrate the induced variability in the feature probabilities across clusters, Table 5 gives the 90% interval of P_{kjg}, which represents the proportion of crimes with category k for feature j for an arbitrary criminal g. For example, the proportion of crimes that are committed at apartments varies from 0.001 for some criminals to 0.448 for others. These probabilities show considerable variability across criminals, which supports the result that including these features in the analysis improves clustering accuracy.

4.3. Crime series identification

To illustrate a typical crime series identification analysis, we analyse the three largest clusters of solved crimes, and the unsolved crimes associated with these clusters with posterior probability at least 0.8. Fig. 4 plots the spatial and temporal locations of these clusters and the crimes

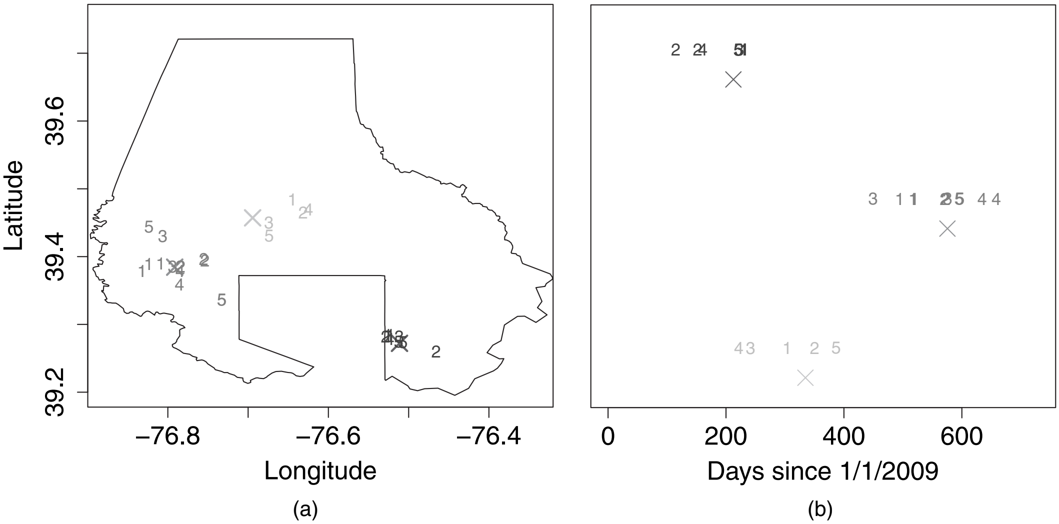


Fig. 5. (a) Spatial (×, crime 1; ×, crime 2; ×, crime 3) and (b) temporal locations for the crime classification example: times are given as the midpoints of the plausible interval; the location or time of the three unsolved crimes is indicated by 'x'; for each unsolved crime the location or time of the events in the jth most likely cluster of solved crimes is marked by 1–5

attributed to these clusters. The second and third criminals' spatial locations (medium and dark grey) form tight clusters, and the event times for tight clusters for all three criminals. The circumstances of the crimes also follow a distinctive pattern. Of the 43 crimes in the first cluster, all 42 were committed in an apartment, 39 had door as the point of entry and 29 had pried as the method of entry (11 others had forced physically). The four unsolved crimes that were associated with cluster 1 all fall in the centre of the spatial cluster, occur in the same time window and were committed at an apartment with door as the point of entry and two had pried as the method of entry (one had forced physically; one was missing). The second cluster (medium grey) also has a characteristic pattern. All the 31 crimes occurred in a garage with door as the point of entry and other as the method of entry. All unsolved events associated with this cluster share these features. The final cluster (dark grey) is associated with only one unsolved crime.

4.4. Crime classification

To illustrate crime classification, we select three unsolved crimes on the basis of their probability of being linked with a solved crime, i.e. the posterior probability that their cluster G_i includes at least one solved crime. We pick three crimes with probability 0.40, 0.98 and 0.99 to represent three likely scenarios. The data for these crimes and their likely clusters are given in Table 6 and Fig. 5.

The first unsolved crime is associated with no cluster of solved crimes with probability higher than 0.1. This crime is difficult to classify; its property type (other) and time ('weekday–day') are the most likely categories for these features. Therefore, a crime classification analysis is not likely to be useful for this case. However, this illustrates the importance of assigning uncertainty estimates to the clustering output, as this allows investigators to act on the results only when appropriate.

The second unsolved crime is linked with three clusters of solved crimes with probability between 0.101 and 0.127. As with the unsolved crime, these clusters of solved crimes are predominately committed in western Baltimore County, in single homes and on weekday days. The clusters that are linked with the third unsolved crime form a tighter spatiotemporal cluster (dark grey in Fig. 5) than for those of the other two crimes. This crime is associated with one cluster with probability 0.672. Both crimes in this cluster occur in a residential yard and with no force required. Another cluster of solved crimes shares these features but is ranked only the third most likely, perhaps because it has only a single event, whereas the most likely cluster has two crimes and thus there is less uncertainty about its features.

5. Discussion

In this paper, we have shown how model-based clustering analysis can aid investigators in crime series identification and crime classification. Our hierarchical Bayesian model naturally handles missing data and interval censoring, and provides uncertainty estimates for all associations. For the Baltimore burglary case-study, we demonstrated that this method can identify a short list of criminals which includes the true criminal with reasonably high probability.

We have focused our analysis on burglaries, but our model could also be applied to other types of crime or even for crime series that include multiple types of crime. Our model will accommodate any crimes that have spatial, temporal and categorical attributes. The success of our approach with other types of crime, locations and time periods would need to be evaluated. Also, additional information could be included in our analysis to improve predictive performance. For example, considering more categories of the crime features type of property, point of entry, or method of entry or further spatial information such as distance to a major highway could improve clustering, especially for larger data sets. Because police data are not collected for research, they may suffer from accuracy and coding reliability issues. Improving the quality of these data (e.g. reducing time intervals, geocoding improvements or checking coding reliability) has the potential to reduce uncertainty and to improve results.

Another approach that is likely to produce similar results is clustering via a Dirichlet process mixture prior (Ghosh and Ramamoorthi, 2003). This has the advantage of avoiding an upper bound on the number of clusters. In our analysis, we found that results were not sensitive to the number of clusters, and so we elected to use the simpler finite mixture model. However, spatiotemporal clustering using non-parametric Bayesian methods is an area of future research.

Several methodological issues remain unresolved. For example, we must assume that all crimes are equally likely to be solved. In reality, it may be that larger clusters are more likely to be solved than small clusters or vice versa. One possibility would be to model the probability that a crime is solved by using the known crime features such as spatiotemporal location and mode of entry. However, it is not clear that this would have any effect on the clustering results. Another issue is that we have assumed that observations are independent conditioned on the cluster label. In practice it is likely that there is dependence between features within a cluster. Again, it is not clear what effect this would have on clustering results. Finally, it is not obvious how to deal with observations near the border. We have elected to include in the analysis solved crimes outside the study region that are linked to crimes inside the study region, but to exclude unsolved crimes outside the study region. In our study this represented only 0.7% of the solved crimes, but for other studies this issue may require careful consideration.

Acknowledgements

The authors express gratitude to the Baltimore County Police Department for the crime data. This work was supported by the National Institute of Justice through grant 2010-DE-BX-K255.

Appendix A: Computational details

An advantage of the proposed method for spatiotemporal clustering is that the Markov chain Monte Carlo code is fairly straightforward. We use Metropolis-within-Gibbs sampling. Most of the model parameters have conjugate full conditionals, which are given below, and thus are updated by using Gibbs sampling. For observation i, T_i is unknown for interval-censored observations, X_{ij} is missing for some observations, and G_i is unknown for unsolved crimes. These parameters are treated as latent variables in the Markov chain Monte Carlo sampler and are updated from the full conditionals:

$$T_i|\text{rest} \sim \text{TN}_{(L_i,U_i)}(\theta_{G_i}, \sigma_t^2),$$

$$P(X_{ij}=k|\text{rest}) = P_{kjG_i},$$

$$P(G_i=g|\text{rest}) \sim \frac{\pi_g\, \phi(\mathbf{s}_i|\boldsymbol{\mu}_g, \sigma_s^2 I_2)\, \phi(T_i|\theta_g, \sigma_t^2) \prod_{j=1}^{p} P_{X_{ij}jg}}{\sum_{h=1}^{M} \pi_h\, \phi(\mathbf{s}_i|\boldsymbol{\mu}_h, \sigma_s^2 I_2)\, \phi(T_i|\theta_h, \sigma_t^2) \prod_{j=1}^{p} P_{X_{ij}jh}}$$

where 'TN' denotes the truncated normal density and $\phi(y|m,s^2)$ is the Gaussian density with variate y, mean m and variance s^2. The cluster-specific parameters have full conditionals

$$\boldsymbol{\mu}_g|\text{rest} \sim N\left(\frac{\sigma_s^{-2}}{\sigma_s^{-2}|\mathcal{G}_g| + \tau_s^{-2}} \sum_{i\in\mathcal{G}_g} \mathbf{s}_i + \frac{\tau_s^{-2}}{\sigma_s^{-2}|\mathcal{G}_g| + \tau_s^{-2}} \bar{\boldsymbol{\mu}}, \frac{1}{\sigma_s^{-2}|\mathcal{G}_g| + \tau_s^{-2}} I_2 \right),$$

$$\theta_g|\text{rest} \sim N\left(\frac{\sigma_t^{-2} \sum_{i\in\mathcal{G}_g} T_i + \tau_s^{-2}\bar{\theta}}{\sigma_t^{-2}|\mathcal{G}_g| + \tau_t^{-2}}, \frac{1}{\sigma_t^{-2}|\mathcal{G}_g| + \tau_t^{-2}} \right),$$

$$\mathbf{P}_{jg}|\text{rest} \sim \text{Dirichlet}\left\{ D_j \bar{P}_{1j} + \sum_{i\in\mathcal{G}_g} I(X_{ij}=1), \ldots, D_j \bar{P}_{N_j j} + \sum_{i\in\mathcal{G}_g} I(X_{ij}=N_j) \right\}$$

where $\mathcal{G}_g = \{i|G_i=g\}$ and $|\mathcal{G}_g|$ is the number of elements in $\mathcal{G}_g$.

The full conditionals for hyperparameters with conjugate full conditionals are

$$\sigma_s^{-2}|\text{rest} \sim \text{gamma}\left\{ n+a, \sum_{i=1}^{n} \frac{(\mathbf{s}_i - \boldsymbol{\mu}_{G_i})^{\mathrm{T}}(\mathbf{s}_i - \boldsymbol{\mu}_{G_i})}{2} + b \right\},$$

$$\sigma_t^{-2}|\text{rest} \sim \text{gamma}\left\{ \frac{n}{2} + a, \sum_{i=1}^{n} \frac{(T_i - \theta_{G_i})^2}{2} + b \right\},$$

$$\tau_s^{-2}|\text{rest} \sim \text{gamma}\left\{ M+a, \sum_{m=1}^{M} \frac{(\boldsymbol{\mu}_m - \bar{\boldsymbol{\mu}})^{\mathrm{T}}(\boldsymbol{\mu}_m - \bar{\boldsymbol{\mu}})}{2} + b \right\},$$

$$\tau_t^{-2}|\text{rest} \sim \text{gamma}\left\{ \frac{M}{2} + a, \sum_{m=1}^{M} \frac{(\theta_m - \bar{\theta})^2}{2} + b \right\},$$

$$\bar{\boldsymbol{\mu}}|\text{rest} \sim \text{TN}_{\mathcal{D}}\left(\frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\mu}_m, \frac{\tau_s^2}{M} I \right),$$

$$\bar{\theta}|\text{rest} \sim \text{TN}_{(0,T^*)}\left(\sum_{m=1}^{M} \frac{\theta_m}{M}, \frac{\tau_t^2}{M} \right).$$

The only parameters with non-conjugate full conditionals are the D_j. These parameters are updated by using Metropolis sampling with log-normal candidate distributions, tuned to have acceptance probability 0.4. The full conditional distributions that are required for Metropolis updates are

$$f(D_0|\text{rest}) \propto \text{Dirichlet}(\boldsymbol{\pi}|D_0, \ldots, D_0)\, \text{gamma}(D_0|a,b),$$

$$f(D_j|\text{rest}) \propto \prod_{m=1}^{M} \text{Dirichlet}(\mathbf{P}_{jm}|D_j \bar{P}_{1j}, \ldots, D_j \bar{P}_{N_j j})\, \text{gamma}(D_j|a,b)$$

where Dirichlet and gamma are the Dirichlet and gamma densities respectively. We generate 25 000 samples from this model and discard the first 5000 as burn-in. Convergence is monitored by using trace plots of several representative parameters.

References

Adderley, R. (2004) The use of data mining techniques in operational crime fighting. In *Intelligence and Security Informatics* (eds H. Chen, R. Moore, D. Zeng and J. Leavitt), pp. 418–425. Berlin: Springer.

Adderley, R. and Musgrove, P. (2001) Data mining case study: modeling the behavior of offenders who commit serious sexual assaults. In *Proc. 7th Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining Int. Conf. Knowledge Discovery and Data Mining*, pp. 215–220. New York: Association for Computing Machinery.

Adderley, R. and Musgrove, P. (2003) Modus operandi modelling of group offending: a data-mining case study. *Int. J. Pol. Sci. Mangmnt*, **5**, 265–276.

Bennell, C. and Canter, D. (2002) Linking commercial burglaries by modus operandi: tests using regression and ROC analysis. *Sci. Just.*, **42**, 153–164.

Bennell, C. and Jones, N. (2005) Between a ROC and a hard place: a method for linking serial burglaries by modus operandi. *J. Invest. Psychol. Offend. Profilng*, **2**, 23–41.

Bennell, C., Jones, N. and Melnyk, T. (2009) Addressing problems with traditional crime linking methods using receiver operating characteristic analysis. *Leg. Crimin. Psychol.*, **14**, 293–310.

Brown, D. and Hagen, S. (2003) Data association methods with applications to law enforcement. *Decsn Supprt Syst.*, **34**, 369–378.

Canter, D. (2004) Offender profiling and investigative psychology. *J. Invest. Psychol. Offend. Profilng*, **1**, 1–15.

Canter, D. and Hammond, L. (2007) Prioritizing burglars: comparing the effectiveness of geographical profiling methods. *Pol. Pract. Res.*, **8**, 371–384.

Catalano, S. (2010) Victimization during household burglary. US Department of Justice Office of Justice Programs, Bureau of Justice Statistics, Washington DC. (Available from `http://www.bjs.gov/index.cfm?ty=pbdetail&iid=2172`.)

Cocx, T. and Kosters, W. (2006) A distance measure for determining similarity between criminal investigations. In *Advances in Data Mining: Applications in Medicine, Web Mining, Marketing, Image and Signal Mining* (ed. P. Perner), pp. 511–525. Berlin: Springer.

Coupe, T. and Blake, L. (2006) Daylight and darkness targeting strategies and the risks of being seen at residential burglaries. *Criminology*, **44**, 431–464.

Ewart, B., Oatley, G. and Burn, K. (2005) Matching crimes using burglars' modus operandi: a test of three models. *Int. J. Pol. Sci. Mangmnt*, **7**, 160–174.

Fraley, C. and Raftery, A. E. (1998) How many clusters?; which clustering method?; answers via model-based cluster analysis. *Comput. J.*, **41**, 578–588.

Ghosh, J. K. and Ramamoorthi, R. V. (2003) *Bayesian Nonparametrics*. New York: Springer.

Goodwill, A. and Alison, L. (2006) The development of a filter model for prioritizing suspects in burglary offences. *Psychol. Crime Law*, **12**, 395–416.

Grubin, D., Kelly, P. and Brunsdon, C. (2001) Linking serious sexual assaults through behaviour. *Research Studies 215*. Home Office, London.

Lin, S. and Brown, D. (2006) An outlier-based data association method for linking criminal incidents. *Decsn Supprt Syst.*, **41**, 604–615.

Ma, L., Chen, Y. and Huang, H. (2010) Ak-modes: a weighted clustering algorithm for finding similar case subsets. In *Proc. Int. Conf. Intelligent Systems and Knowledge Engineering*, pp. 218–223. New York: Institute of Electrical and Electronics Engineers.

Markson, L., Woodhams, J. and Bond, J. (2010) Linking serial residential burglary: comparing the utility of modus operandi behaviours, geographical proximity, and temporal proximity. *J. Invest. Psychol. Offend. Profilng*, **7**, 91–107.

Ratcliffe, J. (2000) Aoristic analysis: the spatial interpretation of unspecific temporal events. *Int. J. Geog. Inform. Sci.*, **14**, 669–679.

Ratcliffe, J. (2002) Aoristic signatures and the spatio-temporal analysis of high volume crime patterns. *J. Quant. Crimin.*, **18**, 23–43.

Salo, B., Sirén, J., Corander, J., Zappalà, A., Bosco, D., Mokros, A. and Santtila, P. (2013) Using Bayes theorem in behavioural crime linking of serial homicide. *Leg. Crimin. Psychol.*, **18**, 356–370.

Santtila, P., Fritzon, K. and Tamelander, A. (2004) Linking arson incidents on the basis of crime scene behavior. *J. Pol. Crim. Psychol.*, **19**, 1–16.

Santtila, P., Junkkila, J. and Sandnabba, N. (2005) Behavioural linking of stranger rapes. *J. Invest. Psychol. Offend. Profilng*, **2**, 87–103.

Santtila, P., Pakkanen, T., Zappala, A., Bosco, D., Valkama, M. and Mokros, A. (2008) Behavioural crime linking in serial homicide. *Psychol. Crime Law*, **14**, 245–265.

Snook, B., Wright, M., House, J. and Alison, L. (2006) Searching for a needle in a needle stack: combining criminal careers and journey-to-crime research for criminal suspect prioritization. *Pol. Pract. Res.*, **7**, 217–230.

Tonkin, M., Grant, T. and Bond, J. (2008) To link or not to link: a test of the case linkage principles using serial car theft data. *J. Invest. Psychol. Offend. Profilng*, **5**, 59–77.

Tonkin, M., Woodhams, J., Bull, R., Bond, J. and Palmer, E. (2011) Linking different types of crime using geographical and temporal proximity. *Crimin. Just. Behav.*, **38**, 1069–1088.

Tonkin, M., Woodhams, J., Bull, R., Bond, J. and Santtila, P. (2012) A comparison of logistic regression and classification tree analysis for behavioural case linkage. *J. Invest. Psychol. Offend. Profilng*, **9**, 235–258.

US Census Bureau (2011) State and county quickfacts. US Census Bureau, Washington DC. (Available from `http://quickfacts.census.gov/qfd/states/24/24005.html`.)

Woodhams, J., Hollin, C. and Bull, R. (2007) The psychology of linking crimes: a review of the evidence. *Leg. Crimin. Psychol.*, **12**, 233–249.

Woodhams, J. and Labuschagne, G. (2012) A test of case linkage principles with solved and unsolved serial rapes. *J. Pol. Crimin. Psychol.*, **27**, 85–98.

Woodhams, J. and Toye, K. (2007) An empirical test of the assumptions of case linkage and offender profiling with serial commercial robberies. *Psychol. Publ. Poly Law*, **13**, 59–85.

Yokota, K. and Watanabe, S. (2002) Computer-based retrieval of suspects using similarity of modus operandi. *Int. J. Pol. Sci. Mangmnt*, **4**, 5–15.