Spatial targeting of infectious disease control: identifying multiple, unknown sources

Robert Verity[1*], Mark D. Stevenson[1*], D. Kim Rossmo[2], Richard A. Nichols[1], Steven C. Le Comber[1†]

[1]School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom; and [2]Center for Geospatial Intelligence and Investigation, School of Criminal Justice, Texas State University, 601 University Drive, San Marcos, TX 78666

[*]These authors contributed equally to this paper,

[†]Correspondence author. E-mail: s.c.lecomber@qmul.ac.uk

Summary

1. Geographic profiling (GP) was originally developed as an analytical tool in criminology, where it uses the spatial locations of linked crimes (for example murder, rape or arson) to identify areas that are most likely to include the offender's residence. The technique has been extremely successful in this field, and is now widely used by police forces and investigative agencies around the world. More recently, the same method has been applied to biological data, notably in spatial epidemiology, where it uses the locations of disease cases to identify infection sources: the identification of these sources is critical to control efforts of diseases such as malaria, since targeted intervention is more efficient and cost effective than untargeted intervention.

2. Here we solve the problem of identifying multiple sources, even when the number of sources is unknown – a requirement for many biological studies. We present a new, rigorous mathematical and computational method, and show why previous Bayesian methods were often outperformed by the empirically-developed Criminal Geographic Targeting (CGT) algorithm used in criminology.

3. We use simulations and real-world examples to compare our model to both the CGT algorithm and to an existing Bayesian model. We demonstrate that our method combines the advantages of both previous methods, particularly in cases featuring large data sets and multiple sources.

4. Our approach provides an increase in search efficiency over other methods and is likely to lead to improved targeting of interventions and more efficient use of resources. We suggest that the Dirichlet process mixture (DPM) model provides a useful and practical tool for conservation biologists and epidemiologists that can be used to inform management decisions and public health policy.

Keywords

Bayesian statistics, criminology, Dirichlet process mixture, epidemiology, geographic profiling

Abbreviations

GP, geographic profiling; DPM, Dirichlet process mixture; MCMC, Markov Chain Monte Carlo

Introduction

In many areas of biology (for example invasion biology and epidemiology), models describing the ways in which animals, plants or pathogens spread outwards from a central source are of considerable importance. Such models are routinely used to generate risk maps in epidemiology, or to predict the effect of global climate change on the spread of invasive species (Kolar & Lodge 2001). Surprisingly, very few models exist which run backwards in time, using current spatial patterns to identify sources of infections or biological invasions, despite the fact that the identification of these sources can be used to target control efforts, dramatically improving the efficiency of interventions. Recently, geographic profiling (GP) – a technique originally developed in criminology to help prioritise large lists of suspects in cases of serial crime (Rossmo, 2000) – has been successfully applied to biological data, providing a way of doing exactly this (Le Comber & Stevenson 2012).

Investigations of serial crime typically involve too many, rather than too few, suspects; for example, the investigation into the Yorkshire Ripper murders in the UK between 1975 and 1980 generated 268,000 names (Doney 1990). In criminology, GP techniques use spatial data concerning the locations of connected crime sites to create a surface of search priority that is overlaid on a map of the study area to produce a geoprofile, which in turn allows the police to prioritise investigations by systematically checking suspects associated with locations in descending order of the height on the geoprofile (Rossmo 2000). There are a number of different geographic profiling software programs available, including Rigel (Miller 2003), developed by Environmental Criminology Research Inc. (ECRI), CrimeStat (Levine 1996), funded by the U.S. National Institute of Justice, and Dragnet (Canter 2000), developed at the University of Liverpool. Other authors (for example Snook et al. (2002, 2005)) have made a case for the use of human judges. Of different programmes available, the most widely used is the criminal geographic targeting (CGT) algorithm of Rossmo (Rossmo 1993), which forms the basis of Rigel (Miller 2003), in which information from multiple crime sites is combined by means of summing over independent distributions. The CGT is used by organisations including the Royal Canadian Mounted Police, the Bureau of Alcohol, Tobacco, Firearms and Explosives, the Los Angeles Police Department, the National Crime Agency in the UK and the United States Marine Corps and has also been used to identify source populations during biological invasions and sources of infection during disease outbreaks (Le Comber et al. 2006; Raine et al. 2009; Le Comber et al. 2011; Stevenson et al. 2012).

The development of geographic profiling has – understandably – been driven by the need for practical solutions to the problems encountered by law enforcement agencies. O'Leary (O'Leary 2009; O'Leary 2010; O'Leary 2012) placed GP in a Bayesian framework, mathematically formalising the problem. However, the model put forward by O'Leary makes the simplifying assumption that all observed data points originate from a single source, and hence performs extremely badly in cases where there are actually multiple sources (see Methods and Results). Thus, despite the mathematical appeal of O'Leary's approach, the CGT algorithm continues to be widely used as a result of its proven track record (Rossmo 2000).

Here, we present a well-defined mathematical approach that unifies existing methods in a single framework. Crucially, our method explicitly deals with the issue of multiple sources – a situation typical of biological data sets, but less common in criminology. Under these circumstances, our model outperforms both the CGT algorithm and a simple Bayesian model based on the work of O'Leary (O’Leary 2010). Further, we develop a computational approach using Markov Chain Monte Carlo (MCMC) methods that extends the technique to large data problems. Finally, we demonstrate the effectiveness of our model using a real-life example of malaria cases in Egypt.

Specifically, we assert that (1) one of the reasons for the CGT algorithm's improved performance relative to the simple Bayesian model lies in its ability to deal with multiple sources; and hence by constructing a Bayesian model that incorporates the ability of the CGT algorithm to deal with multiple sources while maintaining the mathematical rigour of the simple Bayesian model, we can outperform both of the existing methods; (2) this method can be extended to large data problems using MCMC; (3) this method can be used to provide practical solutions to real-life problems, such as those found in epidemiology.

Geographic Profiling Models

The traditional (CGT) and Bayesian approaches to geographic profiling differ in both their construction and implementation. In the following sections we specify each in common terms.

CGT algorithm

The traditional method begins by considering a distance-decay function around each individual data point. The height of the surface is a measure of how confident we are that the source location lies at this point. The decay function can take a number of forms, but in criminological applications it is typical to use a two-part distribution that increases to a maximum at a distance *B* from the data point, and then declines beyond this:

$$f(d) = \begin{cases} \frac{1}{d^h}, & \text{if } d > B \\ \frac{kB^{g-h}}{(2B-d)^g}, & \text{if } d \leq B \end{cases} \qquad [1]$$

where *d* is the distance (either Euclidian or Manhattan) from the observation. This distribution was originally proposed by Rossmo (2000), but here we have used the notation of O'Leary (O'Leary 2009; O'Leary 2010) (correcting for a mistake in the direction of the inequalities). In this paper we use the Euclidean distance throughout. Although this decay function is often referred to as a probability distribution, this is not technically true as there is no requirement for the surface to integrate to unity (nor, in criminology, any need for it to do so, since the analysis is used to produce ranked scores rather than probabilities). Thus, in the traditional method the decay function is better described as a surface of search priority, subject to the more general constraint that points high up on the surface represent areas of high priority. This measure of priority is modelled as an additive quantity, meaning that the information from several observations can be combined by summing together the independent surfaces. The end result of this process of summation is a single surface that represents our integrated knowledge of the source location, which is referred to as a jeopardy surface (Rossmo, 2000).

The search efficiency of the model can be calculated using the hit score percentage; the proportion of the area that we must search before the true source location is found. The smaller the hit score percentage, the more accurate the geoprofile, with a hit score percentage of 50% representing what we would expect from a non-prioritised random or uniform search (see Rossmo 2000).

Simple Bayesian model

We compare the CGT algorithm against a simple Bayesian model based on the initial approach described by O'Leary (O'Leary 2010; O'Leary 2012), and ignoring subsequent extensions relating to the choice of priors. This approach differs from the CGT in that distributions are defined and manipulated according to the laws of probability. The starting point is to write down the probability of the data, given the known location of the source. This is achieved through the use of a probability distribution, which we will refer to as the migration profile, in which the probability of finding an observation at any point in the domain is expressed relative to the location of the source. Assuming independence between observations, the probability of the sample is simply the product over the probabilities of the individual data points (in fact, Rossmo (1995) considered a similar formulation in which the CGT algorithm is applied in log space). By placing a suitable prior on the source location and applying Bayes' rule it is possible to derive the posterior distribution of the source location, given the observations.

Unsurprisingly, the choice of method makes a big difference to the results. While the CGT algorithm tends to create a patchy distribution of peaks and troughs, entertaining the possibility of a number of different source locations, the simple Bayesian method tends to place the majority of the posterior probability mass around the spatial mean of the data points (at least for many choices of prior and likelihood, including those considered here). Another important difference between the methods is in the rate of convergence. In the Bayesian approach the variance of the posterior distribution tends to decrease rapidly as more data is added, whereas in the CGT method the variance of the geoprofile can never be less than the variance of the decay function. Generally, when there is in fact a single source location the Bayesian method is predicted to outperform the traditional method. However, if there is the potential for multiple source locations then the Bayesian method is predicted to converge quickly on the wrong answer, while the traditional method will still perform well. In this study, we test this prediction using a variety of simulations (see Results 1 and 2, below).

The Dirichlet process mixture model

Our primary objective is to address the issue of multiple sources within a well-defined Bayesian framework. The tool that allows us to do this is the Dirichlet Process Mixture (DPM) model, which has a strong mathematical foundation (Ferguson 1983; Green & Richardson 2001) and is finding increasing application within biology (e.g. Huelsenbeck et al. 2006; Huelsenbeck & Andolfatto 2007; Dorazio et al. 2008). Unlike many clustering approaches, DPM models do not require the user to specify the number of clusters beforehand, making them extremely useful in situations where there is no strong prior information about the exact number of clusters. In place of a fixed number of clusters, the DPM model describes the process of cluster formation using a single ‘concentration parameter’, α. Specifically, if we have already seen n observations, of which n_A came from cluster A, then the (prior) probability of the next observation also belonging to cluster A is given by $n_A/(n + \alpha)$. It follows that, no matter how many observations we have seen, there is always a positive probability $\alpha/(n + \alpha)$ of the next observation originating from a previously undiscovered cluster. While we may not believe there to be a truly unlimited number of clusters, by allowing for the possibility of an expanding number of clusters we can ensure that our model is always appropriate for the quantity of data at hand. Obviously the choice of the concentration parameter α has a strong influence on the model. Although an appropriate value of α could be fitted from training data, here we chose instead to integrate over our uncertainty by placing a diffuse hyper-prior over α (of the form $h(\alpha)=1/(1+\alpha)^2$, see Appendix 2 for details). Where stronger prior information is available, the model can easily be adapted to include this.

The second part of the DPM model is the calculation of the posterior distribution of source locations, conditional on a particular partition of the data into clusters. This part is mathematically very similar to the simple Bayesian model, with the only difference being that a different posterior distribution is produced for each cluster. The likelihood of all observations in the same cluster is equal to the product of the migration profile over each of the observations. By incorporating an appropriate prior on the source location and applying Bayes’ rule we arrive at the posterior distribution of the source location from which this particular subset of observations derived. Carrying out this step for each cluster independently we obtain a set of posterior distributions – one for each of the (potentially) multiple source locations.

Finally, in order to obtain an analytical solution to the DPM model described above we would be required to sum over all possible partitions of the n data points into up to n clusters, weighted by the posterior probability of the partition in each case. The number of such partitions is given by the n^{th} Bell number (B_n) which becomes prohibitively large for values as low as n=10 (B_{10}=115,975). Thus, for any reasonably sized data set we must turn to MCMC methods for a practical solution. Fortunately, a detailed exposition of MCMC algorithms for DPM models is provided by Neal (2000), and we need only to adapt these algorithms to our specific application. A more detailed description of the DPM model, including expressions relating to posterior inference under the analytical and MCMC forms of the solution, is provided in Appendices 1 to 3.

It is important to emphasise that the DPM model can be adapted to use any migration profile that satisfies the laws of probability (i.e. integrates to unity). The essence of the DPM model lies in the way that information is combined between clusters, and not in the specific details of the migration profile used. This can be seen in the logic of our study, which has four parts. (i) First, when comparing directly the CGT, simple Bayesian, and DPM models, we use the distribution from the CGT (described in equation [1]) as our migration profile in all three approaches. This ensures that the only difference between methods lies in the way that information is being combined, and not in any other assumptions relating to migration. (ii) Next, we validate the MCMC version of our proposed solution using this same migration profile, thereby ensuring that our MCMC results are directly comparable with our analytical results. (iii) From this, we move on to consider simulated data generated from a distribution more typical of those assumed in biology – the normal distribution – and explicitly compare the full form of the DPM model with the CGT under this assumption. (iv) Finally, we examine a real-world data set – an outbreak of malaria in Cairo – using all three models.

Methods(i) Comparing the simple Bayesian, CGT and DPM models

As mentioned above, our first task is to compare the simple Bayesian, CGT and DPM models purely in terms of the way that information is combined in each case, and controlling for any differences between models, such as the migration profile. We simulated 6, 7, 8 or 9 data points from the distribution given in equation [1] (B=0.5, f=4, g=4), emanating from either 1, 2 or 3 sources, truncated them to fit the available grid. For the purposes of simulation we split the domain into a 100*100 grid, and replicated each combination of the number of data points and sources 1000 times. Sources were chosen to fall within the central 50*50 cells in a random, uniform manner. For each simulated data set we then used each of the three methods described above to search for the 'unknown' source locations, with search efficiency being measured in terms of the hit score percentage. The same distribution (distribution [1] with B=0.5, f=4, g=4) was used as the search distribution in each of the three methods. By designing simulations in this way we can capture an idealised situation in which all three methods make the same assumptions about the true dispersal distribution, and furthermore these assumptions are exactly correct (thereby removing another possible source of model error).

(ii) MCMC validation

For the reasons described previously, the analytical form of the DPM model can deal with only small data sets, and for larger data sets an MCMC implementation of the solution is required. For each of the 12000 simulations described above (1000 replicates of each combination of 1, 2 and 3 sources and 6, 7, 8 or 9 data points), we also used an MCMC implementation of the model, and calculated the correlation between the surface produced by the analytical form of the model and the MCMC form (see Appendix 3 for details of the MCMC algorithm). We also repeated the comparison of the DPM model with the CGT for larger data sets (1, 2 and 5 source locations; 20, 40, 60, 80 and 100 spread points), using just the MCMC implementation of the model.

When running the MCMC, multiple chains were run simultaneously, with convergence being assessed using the Gelman-Rubin (GR) diagnostic statistic (Gelman et al. 2003) evaluated on the concentration parameter α (using a value of GR=1.1 as a threshold for convergence). After the burn-in period, samples were obtained until the largest standard error of any point on the estimated surface was less than 0.01. Samples were not thinned, as it has previously been shown that this does not increase statistical power in situations such as this (Link & Eaton 2012).

(iii) Further comparison of the CGT and DPM models

The migration profile used above (distribution [1]) was designed for criminological applications. In some cases, including many biological applications, it may be more appropriate to assume alternative migration profiles. Here, we assume a bivariate normal migration profile, centred on the unknown source location(s), and with variance σ^2. In some cases, there will be biological data on dispersal patterns that can be used to inform the choice of σ; for example, studies have shown that most malaria transmission occurs close to the larval breeding sites – usually between a few hundred meters and a kilometer– and rarely exceeds 2-3 km (Carter et al. 2000).

We are also required, as part of the DPM model, to choose a prior on the source location(s). For the sake of simplicity we use an empirical Bayes approach, assuming a bivariate normal prior, centred on the spatial mean of the observed data, and with variance τ^2, where τ was set to the maximum distance in either latitude or longitude between the crime sites. τ equals one standard deviation of the normal prior; hence, we expect our source to lie within this distance of the centre around two-thirds of the time, and the model allows for sources well outside the area bounding the crimes. Hence, there is a diffuse, non-informative prior over and beyond the normal search area.

We simulated 6, 7, 8 or 9 data points from a bivariate normal distribution with standard deviation sigma = 1 and emanating from either 1, 2 or 3 sources. For the purposes of simulation we split the domain into a 100*100 grid, and replicated each combination of the number of data points and sources 1000 times. For each simulated data set we then used the two best performing methods described above (CGT and DPM) to search for the ‘unknown’ source locations, with search efficiency being measured in terms of the hit score percentage. The CGT uses the distribution describe in equation [1] with parameters fitted from the data as described by Rossmo (2000), while the DPM uses the spatial mean to fit phi, with sigma fixed at 1.

(iv) Case study

We tested the performance of our model in a real world example by using the MCMC implementation of the DPM model to reanalyse data from Le Comber et al. (2011). In this study, spatial data relating to 139 recorded *Plasmodium vivax* malaria cases were collected, and buffer zones of 2 km were created around the locations of these malaria cases and merged to form a polygon of 296.5 km^2 (Hassan 2006). All accessible aquatic habitats within this study area (surface/cryptic; temporary/semipermanent/permanent) were located and characterised between April and September 2005. These included water tanks, water pools created through pipelines or drainage system breakage, seepage from slum housing, natural springs, pools and ditches filled with ground water. Water sources included in this analysis were identified as bodies of water harbouring at least one mosquito larva over the study period (n= 59). A total of 11 mosquito species were identified, including the malaria vectors *An. sergentii* and *An. pharoensis*, as well as other, non-vector, species. Of these 59 sites, seven tested positive for one or both of the malaria vectors *An. sergentii* and *An. pharoensis* (*An. sergentii* is well established as the most dangerous malaria vector in Egypt (Said et al. 1986)).

A dispersal distance of sigma = 0.018, roughly corresponding to 1km, was used in the DPM model in correspondence with values in the literature (e.g. Carter et al. 2000) and a value of tau = 0.328 was fitted from the observed data (see above).

The model is written in R (R core team 2012) and integrates with Google Maps via the R package RgoogleMaps (Loecher 2012). The model used in this paper is available from the authors on request as an R package called 'Rgeoprofile'.

Results

(i) Comparing the simple Bayesian, CGT and DPM models

Starting with the first set of simulations (1000 replicates of each combination of 1, 2 and 3 sources and 6, 7, 8 or 9 data points), we used a fully factorial ANOVA to test the effect on the hit score percentage (or average hit score percentage when the number of sources was > 1) of model type, number of sources and number of spread points. Three model types were examined; the analytical form of the DPM model, the classical CGT algorithm and the simple Bayesian model.

Model type, number of points and number of sources all significantly affected the relative performance of the three models (ANOVA: model type: $F_{2,35964}$=4787.05,p< 2e-16; sources: $F_{2, 35964}$=13099.30,p<2e-16; points: $F_{3, 35964}$=106.23, p<2e-16). All interactions were highly significant, with the *F* value for model type*sources interaction having the largest effect size ($F_{4, 35964}$=2840.12, p<2e-16); none of the other *F* values exceeded 52. Tukey post-hoc tests at α=0.05 showed that (1) the CGT significantly outperformed the simple Bayesian model, by an average of 1.81% (95% CI: 1.75-1.86%); (2) the DPM model showed a statistically significant improvement over both the CGT algorithm, albeit only by 0.3% (95% CI: 0.25-0.36%) and the simple Bayesian model, again by about 2% (95% CI: 2.1-2.2%). Across all 12,000 runs, the DPM model performed better than the CGT in 68.2% of trials, and as well or better in 74.9%, and better than the simple Bayesian model in 64.6% of trials, and as well or better in 91.5%.

However, although the DPM model outperformed the simple Bayesian model overall, the simple Bayesian model had a small advantage when there was a single source (Figure 1).

(ii) MCMC validation

For the same simulated data sets described above we calculated the correlation between the surface produced by the analytical form of the DPM model and the MCMC form. The two surfaces tended to extremely highly correlated (*r* (mean ±sd) = 0.9998 ± 0.0010), demonstrating that the MCMC algorithm does indeed find the same – or at least extremely similar – posterior distributions as the analytical form of the model.

For the second set of simulations (1000 replicates of each combination of 1, 2 and 5 sources and 20, 40, 60, 80 or 100 data points) we performed the same analysis as in Results part 1, with extremely similar results (ANOVA: model type: $F_{1,29992}$=167.7, p<2e-16; sources: $F_{2, 29992}$=10603.1, p<2e-16; points: $F_{4, 29992}$=1986.2, p<2e-16; model type*sources: $F_{2, 29992}$=463.5, p<2e-16; model type*points: $\mathrm{F}_{4, 29992}$=17.4, p<2e-16; sources*points: $F_{8, 29992}$=2916.7, p<2e-16; model type*sources*points: $F_{8, 29992}$=0.9, p=0.87). Tukey post-hoc tests at α =0.05 showed that the DPM model outperformed the CGT algorithm in a statistically significant way; again, this improvement was most marked when the number of sources was > 1 (Figure 2).

(iii) Further comparison of the CGT and DPM models

In the next set of simulations, in which a normal migration profile was assumed, we used ANOVA to test the effect on the hit score percentage (or average hit score percentage when the number of sources was > 1) of model type, number of sources and number of spread points. The two best performing model types from previous simulations were examined; the CGT and the DPM.

The best performing ANOVA was selected by AIC to include a single significant interaction term. Model type, number of points and number of sources all significantly affected the relative performance of the two models (ANOVA: model type: F_{19991}=3693.6,p< 2e-16; sources: $\mathrm{F}_{2, 19991}$=2038,p<2e-16; points: $\mathrm{F}_{3, 19991}$=39.1, p<2e-16). Model type*sources interaction was also significant ($F_{4, 19991}$=222.1, p<2e-16). Tukey post-hoc tests at α=0.05

showed that the DPM model showed a statistically significant improvement over the CGT algorithm with an effect size of 4.1% (95% CI: 3.9-4.2%). The MCMC implementation of the DPM outperforms the CGT 67.1% of the time, and performs as well or better 67.2% of the time. In our simulations this equates to searching on average 410 fewer cells (95% CI: 394-421) before finding all of the sources.

(iv) Case study

The median hit score percentages for the seven vector breeding sites identified in Hassan (2006) were 0.34% for the DPM model, compared to 0.43% for the CGT and 1.2% for the simple Bayesian model. Note that the hit scores reported here differ from those in Le Comber et al. (2011), although the surface produced is the same in both cases. The difference arises because the DPM model uses RgoogleMaps (Loecher 2012), and thus the exact dimensions of the search area (which affects the hit score) are set by the available zoom levels in the Google Maps data. To allow direct comparison, we used the same search area for the CGT and the DPM mode.

For five of the seven sites, hit score percentages for the DPM were less than half a per cent. An additional output of our model is that it can provide a barplot of the posterior probability of the number of realised sources (Figure 3). In this case our model indicated the highest probability for seven sources, with a likely range of 6-10. Interestingly, some of these correspond to areas where no vector species were found by Hassan (2006) (Figure 4). One possibility, of course, is that these are false-positive results. Alternatively, it is possible that some sources were missed in the original survey, especially given the often considerable difficulty of locating small, transient breeding populations of mosquitoes (Carter et al. 2000) and since searches were carried out in a single year (2005), whereas the malaria cases spanned four (2001-2004) (Hassan 2006; Le Comber et al. 2011).

Discussion

Overall the DPM model is an improvement on the existing methods. When the number of sources is greater than one it outperforms them (Results (i)), it does not require that the number of sources is known *a priori* and, in addition, it generates estimates of their number. Even in conditions specifically designed to maximise the performance of the CGT algorithm, the DPM model still obtains a small advantage, reflecting the way in which it appropriately combines information from observations, rather than taking a simple sum (as in the CGT) or product (as in the simple Bayesian model). The DPM model's analytical method cannot be extended to very large numbers of observations, but the approach can be implemented in an MCMC algorithm which accurately constructs the posterior distribution, as demonstrated in Results (ii).

With these facts established we move on to consider cases in which the DPM model may have a practical advantage over other approaches. The later set of simulations (Methods (iii) and Results (iii)) demonstrate that there are biologically plausible settings in which the use of the DPM model can result in an appreciable increase in search efficiency compared with other methods. Finally, and perhaps most encouragingly, we find that the DPM model leads to an increase in search efficiency when applied to a real-world data set describing malaria transmission in Cairo. The improvement over the CGT algorithm is small, but justifies further investigation of this model on a range of data sets.

In its construction, the DPM model forms a bridge between the seemingly disparate methodologies of the CGT and the simple Bayesian approach to geographic profiling. From a practical point of view the major difference between the two existing approaches lies in whether distributions should be summed (CGT) or multiplied (simple Bayesian). The DPM model works by splitting the data into groups, with each group corresponding to a different source location. The laws of probability then dictate that distributions should be multiplied within groups, but summed between groups. Thus, if all points are assigned to a single source we arrive back at the simple Bayesian model, while if all points are assigned to different sources we arrive at something more akin to the CGT algorithm. In this context, our concentration parameter α can be understood as a prior over the complete spectrum of models, which allows us to transition between a single-source model and a multiple-source model. When α is set to zero, the DPM model becomes mathematically equivalent to the simple Bayesian model; conversely, as α tends to infinity, we converge on the CGT algorithm. In the majority of cases – particularly those dealing with biological data – the most likely explanation for the data will often lie between these two extremes. For example, in the malaria analysis, the DPM model assigned the highest probability to seven sources from 139 disease case locations (Figure 3).

In our simulations, the DPM model outperformed both other approaches when there were multiple sources. In cases with a single source – a common scenario in criminology – the improvement over the CGT, although statistically significant, was minimal when the dispersal distribution was drawn from Equation [1] (when this assumption was relaxed, the improvement was more marked). The comparison between the DPM model and the simple Bayesian model shows that latter has a small advantage when there is a single source. However, when there is more than one source, the DPM shows a large improvement (this is perhaps unsurprising, since the simple Bayesian model assumes that there is a single source). In real-world applications of GP models it will often (perhaps even always) be the case that the true number of sources is unknown, therefore the principal advantage of the DPM model lies in its ability to rigorously handle the problem of multiple sources. In fact, since the difference between the simple Bayesian model and the DPM model is small when there is a single source, and the advantage offered by the DPM model when there are multiple sources is larger, we would argue that the DPM model is preferable in real-world applications of GP. In our simulations, the DPM model outperformed both other approaches in cases with multiple sources. In cases with a single source – a common scenario in criminology – the improvement over the CGT, although statistically significant, was minimal when the dispersal distribution was drawn from Equation [1] (when this assumption was relaxed, the improvement was more marked).

However, formulating the problem in a rigorous Bayesian framework also allows for a number of useful extensions. First, our model produces a true probability surface, allowing us to calculate the marginal probability of different numbers of sources, as in Figure 3. Second, we can produce a probability surface conditional on a particular number of sources, thereby allowing us to break the overall picture down into different scenarios (we can imagine a different search strategy, conditional on there being one source, two sources etc.). Third, the DPM model explicitly calculates the posterior probability under the model that a particular observation is derived from a particular source. This may be of interest in criminology, where crime linkage is an important problem (Rossmo 2000), and may also be useful in biological data sets, where the spatial linkage can be validated against other forms of information (for example genetic data).

So far, the DPM model is constructed with flexibility in mind, rather than statistical power. For particular cases it may be possible to increase the power of the model by incorporation of stronger prior information – for example, by inferring the concentration parameter from training data. Similarly, where empirical evidence has shown that non-normal dispersal profiles are appropriate (for example, Cauchy distributions in some bird species (Winkler et al.

2005; VanHoutan et al. 2007) or bivariate Student's t-distributions in seeds (Nathan & Muller-Landau 2000)), these can be used within the same general framework.

As well as producing a range of new outputs, the DPM model could also be extended to incorporate new inputs. For example, one useful possible extension of our approach is the utilisation of the outputs produced by niche models to generate priors in the DPM model. Niche modelling is a well-developed field that has recently been placed on a Bayesian footing (Elith & Leatherwick 2009), making its incorporation into the DPM model relatively straightforward. A Bayesian niche model produces a probabilistic estimate of the suitability of habitat for the organism being studied that can be used as a prior in the DPM model. Combining these two approaches would go some way towards producing a spatially explicit niche model approach, as called for by Peterson et al (2003).

In epidemiology and invasion biology, much more attention is paid to models that run forwards in time to generate risk maps or forecasts of future incidence than those that run backwards to locate sources. GP, on the other hand, is radically different, running backwards in time to use current locations to infer sources (Le Comber & Stevenson 2012). The DPM model structure described above also differs from many spatially explicit epidemiological models, such as the shot noise Cox process (Møller 2003), in assuming a distribution of point sources, rather than a smoothly varying hazard function over space. This feature also distinguishes the DPM approach from many existing methods that are routinely used to detect clusters in ecological and epidemiological data (see Pullan et al. 2012 for a review). The impact that these different modeling assumptions may have on our conclusions should be explored in further work. In fact, as O'Leary (O'Leary 2010; O'Leary 2012) has shown, a fully Bayesian implementation of GP can easily be extended to run forwards in time. Despite the difficulties faced by all predictive models, this could potentially be important in areas of biology including epidemiology, invasion biology and in conservation biology (e.g. planning reintroductions of animals or plants).

The DPM model we present here is a general method that can be applied to data describing spread from common source. Evidence-based targeting of interventions is a crucial component in the fight against infectious disease, and targeted interventions are more efficient and more cost-effective than untargeted interventions; for example, malaria is strongly dependent on the location of vector breeding sites, and most transmission only occurs within short distances of these sites (Carter et al. 2000). Because of this clustering, untargeted intervention is highly inefficient. In the Cairo study, the DPM model identified five of the seven breeding sites in less than half a percent of the total search area, representing a dramatic improvement over a non-targeted search.

Although our implementation of the DPM model can deal with large data sets (>1000 data points), GP methods also work well with very small data sets (Rossmo 2000; Stevenson et al. 2012), allowing their use in the early stages of an outbreak or invasion, when control efforts are most likely to be successful. The DPM model provides a useful practical tool for conservation biologists and epidemiologists, offering improvements over other methods that are likely to lead to improved targeting of interventions, and more efficient use of resources.

Acknowledgements

We would like to thank Ali Hassan for permission to use the malaria data, and the Environmental Criminology Research Institute (ECRI) Canada for support and useful comments.

References

Canter, D., Toby C., Huntley, M., and Missen, C. (2000). Predicting Serial Killers' Home Base Using a Decision Support System. *Journal of Quantitative Criminology*, 16 , 457 – 478.

Carter. R., Mendis, K., Roberts, D. (2000) Spatial targeting of interventions against malaria. *Bulletin Of The World Health Organization*, 78, 1401-1411.

Doney, R.*The aftermath of the Yorkshire Ripper: the response of the United Kingdom Police Service*. In Egger, S. A. (1990) *Serial Murder: An Elusive Phenomenon,*Praeger.

Dorazio, R.M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H.L., Jordan, F. (2008) Modelling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics*, 64, 635-644.

el Said, S.,Beier, J. C., Kenway, M. A., Morsy, Z. S., Merdan, A. I. (1986) *Anopheles* population dynamics in two malaria endemic villages in Faiyum governorate, Egypt. *Journal of the American Mosquito Control Association*,2.

Elith, J., Leatherwick, J. R. (2009) Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40, 677-697.

Ferguson, T. S. (1983) Bayesian density estimation by mixtures of normal distributions.*Recent advances in statistics*, 287-303.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2003) *Bayesian data analysis*, Chapman and Hall/CRC, 2nd Ed.

Green, P. J., Richardson, S. (2001) Modelling heterogeneity with and without Dirichlet process. *Scandinavian Journal of Statistics*, 28, 355-375.

Hassan, A. N. (2006) WHO-TDR-SGS, Final report.

Huelsenbeck, J.P., Andolfatto, P. (2007) Inference of population structure under a dirichlet process model. *Genetics*, 175, 1787-1802.

Huelsenbeck, J.P., Jain, S., Frost, S. W. D., Kosakovsky Pond, S. L. (2006) A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proceedings of the National Academy of Sciences*, 103, 6263-6268.

Kolar, S.C., Lodge, D. M. (2001) Progress in invasion biology: predicting invaders. *Trends in Ecology and Evolution*, 16, 199-204.

Le Comber, S. C., Nicholls, B., Rossmo, D. K.,Racey, P. (2006) Geographic profiling and animal foraging. *Journal of theoretical biology*, 240, 223-240.

Le Comber, S. C., Rossmo, D. K., Hassan, A. N., Fuller, D. O., Beier, J. C. (2011) Geographic profiling as a novel spatial tool for targeting infectious disease control. *International journal of health geographics*,10-35.

Le Comber, S. C., Stevenson, M. D. (2012) From Jack the Ripper to epidemiology and ecology. *Trends In Ecology and Evolution*, 27, 307.

Levine, N. (1996). Spatial statistics and GIS: Software tools to quantify spatial patterns. *Journal of the American Planning Association*, 62, 381-392.

Link, W. A., Eaton, M. J. (2012) On thinning chains in MCMC. *Methods in Ecology and Evolution*, 3, 112-115.

Loecher, M. (2012) *RgoogleMaps: Overlays on Google map tiles in R*. R package version 1.2.0.2. Berlin School of Economics and Law. URL http://CRAN.R- project.org/package=RgoogleMaps

Miller, C. (2003) Geographic Profiling Serial Offenses with ECRI's Rigel. *Law Enforcement Technology,* 30, 130-135.

Møller, J. (2003) Shot noise Cox processes. *Advances in Applied Probability* 35, 614-640.

Nathan, R., Muller-Landau, H.C. (2000) Spatial patterns of seed dispersal, their determinants and consequences for recruitment. *Trends In Ecology and Evolution*, 15, 278–285.

Neal, R.M. (2000) Markov chain sampling methods for Dirichlet Process mixture models.*Journal of computational and graphical statistics*, 9, 249-265.

O'Leary, M. (2009) The mathematics of geographic profiling. *Journal of Investigative Psychology and Offender Profiling*,6, 253-265.

O'Leary, M. (2010) Implementing a Bayesian approach to criminal geographic profiling.*First International Conference on Computing for Geospatial Research and Application*, June 21-23 Washington, D.C.

O'Leary, M. (2012) *New Mathematical Approach to Geographic Profiling*, National Insitute of Justice, Washington, D.C.

Peterson, A. T. (2003) Predicting the geography of species' invasions via ecological niche modeling. *The Quarterly Review of Biology*, 78, 419–433.

Pullan, R. L., Sturrock, H. J. W., Magalhaes, R. J. S. et al. (2012) Spatial parasite ecology and epidemiology: a review of methods and applications. *Parasitology* 139, 1870-1887.

R Core Team (2012) *R: A language and environment for statistical computing.*R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/[accessed 20 November 2012]

Raine, N. E.,Rossmo, D. K., Le Comber, S.C. (2009) Geographic profiling applied to testing models of bumble-bee foraging. *Journal of the Royal Society*, 6, 307-319.

Rossmo, D.K , *Geographic profiling: Target patterns of serial murderers*, Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC, Canada.

Rossmo, D. K. (1993) A methodological model.*American Journal of Criminal Justice,* 17, 1-21.

Rossmo, D. K. (2000) *Geographic Profiling.* CRC Press 1st Ed, New york.

Snook, B., Canter, D., & Bennell, C. (2002) Predicting the home location of serial offenders: A preliminary comparison of the accuracy of human judges with a geographic profiling system. *Behavioral Sciences & the Law*, 20, 109–118.

Snook, B., Taylor, P. J., and Bennel, C. (2005) Shortcuts to Geographic Profiling success: A reply to Rossmo (2005). *Applied Cognitive Psychology*, 19, 655-661

Stevenson, M. D., Rossmo, D. K., Knell, R. J., Le Comber, S. C. (2012) Geographic profiling as a novel spatial tool for targeting the control of invasive species. *Ecography*, 10, 704-715.

Van Houtan, K.S.,Pimm, S. L., Halley, J. M., Bierregaard, R. O. Jr., Lovejoy, T. E. (2007) Dispersal of Amazonian birds in continuous and fragmented forest. *Ecology Letters*, 10, 219–229.

Winkler, D. W.,Wrege, P. W., Alllen, P. E., Kast, T. L., Senesac, P., Wasson, M. F., Sullivan, P. J. (2005) The natal dispersal of tree swallows in a continuous mainland environment. *Journal of Animal Ecology*, 74,1080–1090.

Figures

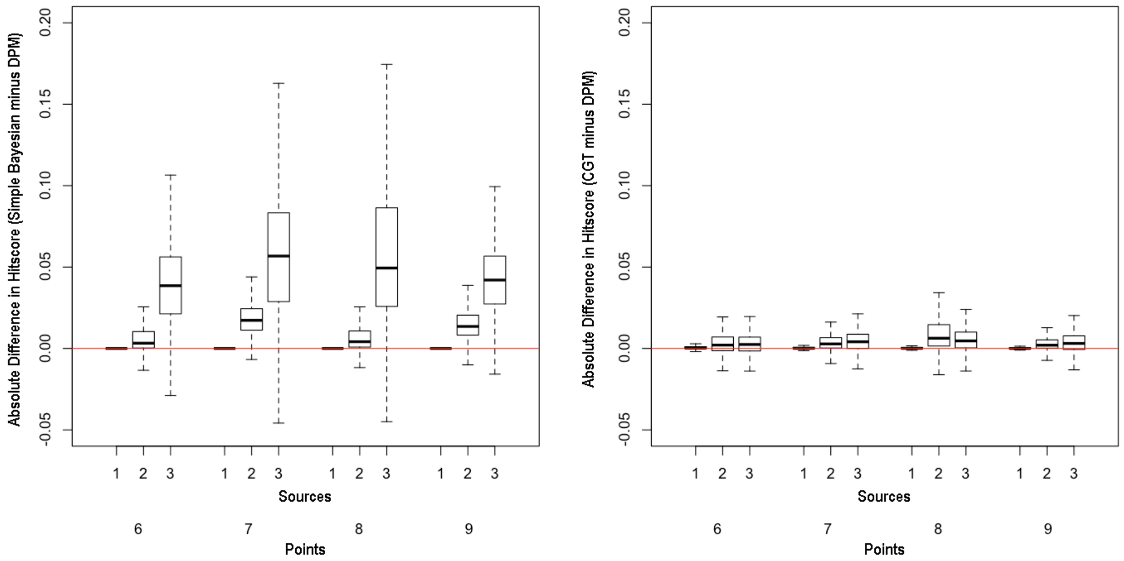

Figure 1 Comparison of the analytical form of the DPM model against (A) the simple Bayesian model, and (B) the CGT algorithm, expressed as the hit score percentage of the simple Bayesian model minus the hit score percentage of the DPM model, and the hit score percentage of the CGT algorithm minus the hit score percentage of the DPM model, respectively. Thus, points above the red line indicate cases in which the DPM model outperformed the other models. In both cases, the DPM model has a statistically significant advantage, although this is more pronounced for the comparison with the simple Bayesian model. In both comparisons, the relative performance of the DPM model improves as number of sources increases.

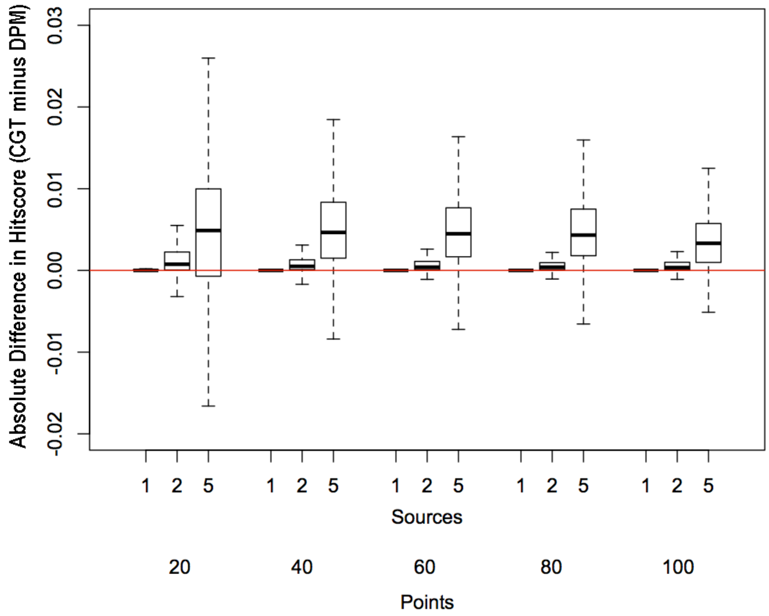

Figure 2 Comparison of the MCMC implementation of the DPM model against the CGT algorithm, expressed as the hit score percentage of the CGT algorithm minus the hit score percentage of the DPM model. Again, points above the red line indicate cases in which the DPM model outperformed the other model. The DPM model outperformed the CGT algorithm, especially as number of sources increases.

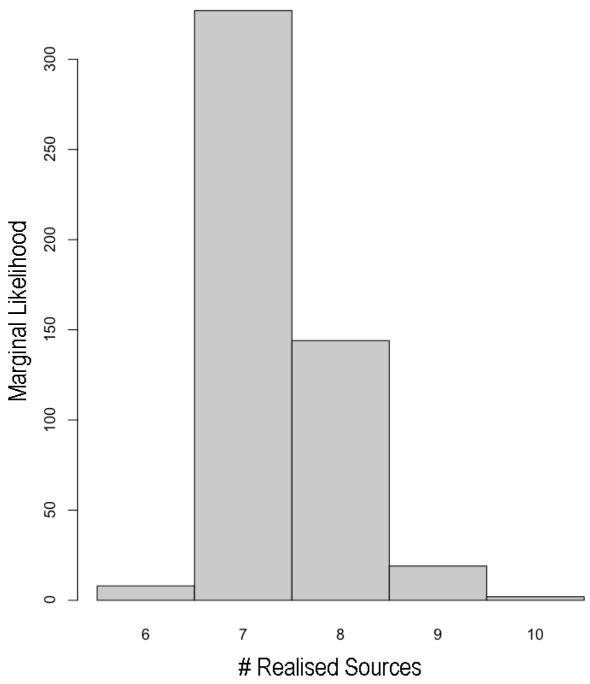

Figure 3 Marginal likelihood of different numbers of realised infection sources for the Cairo data. The DPM model estimates that there are 6-10 sources, and assigns the highest likelihood to seven sources.

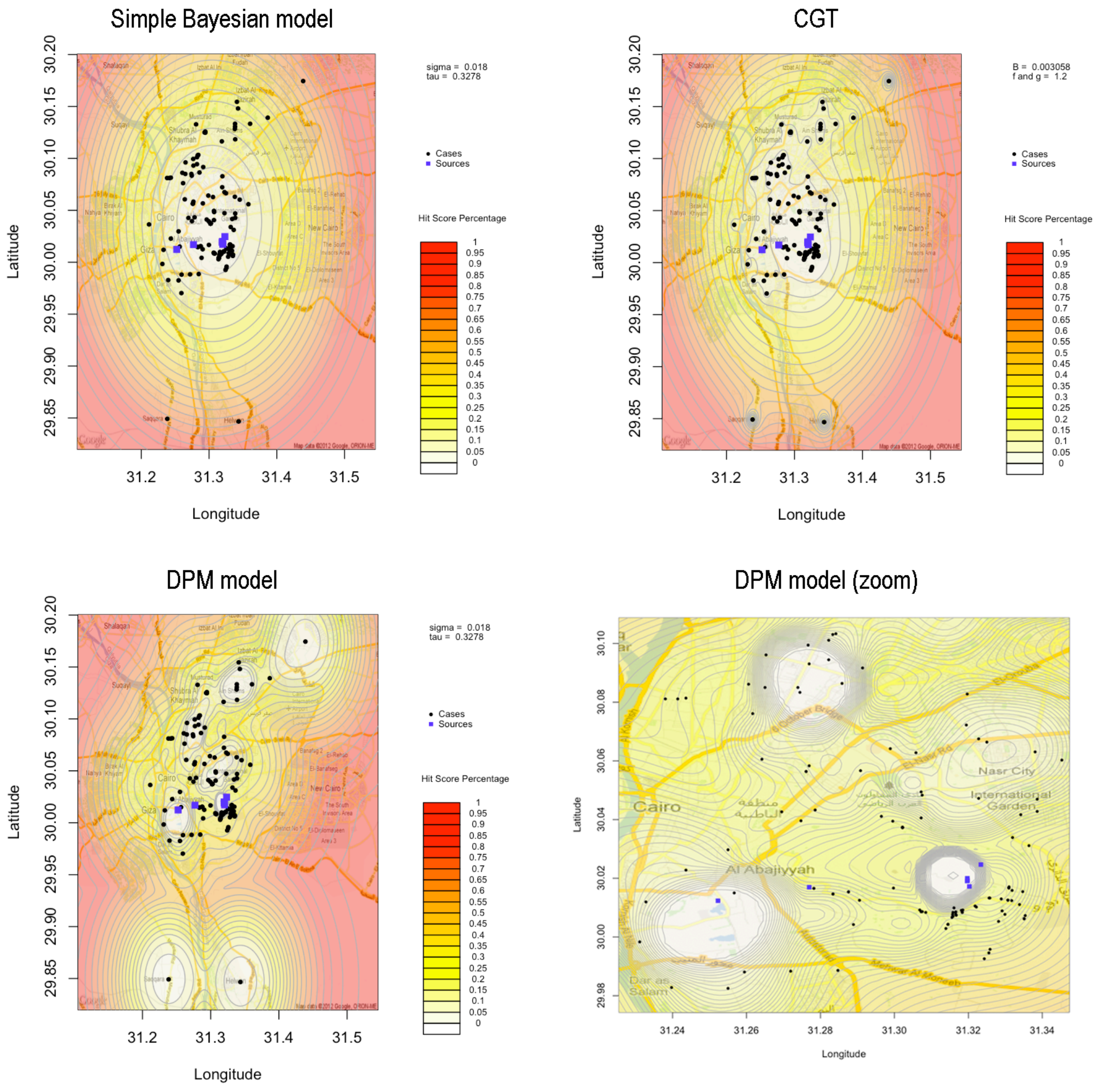

Figure 4 Geoprofile from 139 *Plasmodium vivax* cases in Cairo, Egypt, using (A) the simple Bayesian model; (B) the CGT algorithm; (C) the DPM model. (D) shows a close-up of the DPM surface. In all cases the observed data points are shown as black circles, while the empirically identified sources are shown as blue squares.