
EXTENDING GEOGRAPHIC PROFILING LIKELIHOODS
TO INCLUDE A RANGE OF DATA TYPES
GENERATED IN ECOLOGY AND EPIDEMIOLOGY

MICHAEL STEVENS (MSci)

SUPERVISED BY:

DR. STEVEN LE COMBER,

-

PROF. RICHARD NICHOLS, DR. ROBERT VERITY

AND DR. HANNAH FRY

THESIS SUBMITTED TO QUEEN MARY UNIVERSITY OF LONDON FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY APRIL 17TH, 2021.

-

FUNDED BY THE NATIONAL ENVIRONMENTAL RESEARCH COUNCIL AND
QUEEN MARY UNIVERSITY OF LONDON.

To Steve.

Statement of Originality

I, Michael Stevens, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below. I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: Michael Stevens

Date: April 17th, 2021

Abstract

This thesis revolves around the development of geographic profiling, a spatial model originally developed in criminology to identify areas that likely contain a suspect's home or workplace, based upon where they committed their crimes. Geographic profiling is still used to this day in investigations of serial crime but has recently found a cornucopia of applications in ecology and epidemiology. For example, the sightings of an invading species, responsible for the decline of local wildlife, can be used to target their nesting sites for efficient removal from an environment. Similarly, households testing positive for an infectious disease can be used to target breeding sites of vectors responsible for the disease's transmission. Despite countless applications, geographic profiling models are limited to considering only a single type of spatial data; a set of points on a map. The work I have conducted addresses this issue by specifying a set of geographic profiling models that can deal with multiple kinds of spatial data. In ecology, field experiments surveying alien species often record the location of an encounter and count the number of individuals present. I make the first development accordingly, by describing a model that produces interventions based on spatial count data. I then describe the first instance that geographic profiling is applied to simulated and real-world count data, comparing the conclusions of this new model to that of an existing model that only considers the location recorded but not the number of individuals counted. The next development focusses on applications in epidemiology. When testing members of a household for an infectious disease, the test location and prevalence rate; the proportion of individuals testing positive, are recorded. Hence, I build and test a geographic profiling model that draws conclusions via spatial prevalence data. In the final part of the thesis, I return to analysing the type of data common to geographic profiling, a set of points on a map. I equip users with the flexibility to specify varying assumptions about the process generating these spatial points, ultimately leading to a model for better describing real data. To conclude, I summarise the impact of each new development in this thesis and take a step back, establishing where geographic profiling fits within a universe of spatial models.

Acknowledgements

I owe much gratitude to my supervisor Richard Nichols. You stepped into this role without hesitation and in this short time, you have taught me so much. Your ability to always find some new way of approaching a problem is truly admirable, and I intend to take this mindset with me, long after my work here has concluded. To Bob Verity, whom it has been a genuine pleasure to know and work with on this project. You were under no obligation to provide support, but you did so anyway. I will come away from this PhD looking fondly back upon our meetings that prompted some of the most engaging discussions I have had over these four years. To Sally Faulkner, you've been my rock throughout this entire process. From start to finish, you have inspired confidence within me that has pulled me through the challenges I have faced. Thank you for helping me enjoy this journey as much as you did. To my second supervisor, Hannah Fry, thank you for the pleasant meetings and for keeping me on my toes. Getting to bounce ideas around with you was always a unique and valuable experience.

There are a number of collaborators that I wish to thank for their help with this thesis. To John Beier and André Wilke for providing the data for my third chapter, and whose patience and support were greatly appreciated as the work of chapters two and three developed from our first correspondence in June 2017 to the acceptance of a manuscript in December 2020. I would also like to thank Ross Boyce for providing the data set in chapter four that provoked many interesting questions that I discuss in this thesis. This research would not have been possible without the aid of Queen Mary's high performance computing cluster, Apocrita and the support of The Natural Environmental Research Council.

To the Croft-Street crew: Judith Ament, Natalie Bakker, Rory Walshe, Thomas Baird and Benjamin Taylor. Thank you guys so much for the best company I could possibly have asked for. The countless boardgame sessions, parties, movie nights and cooking lessons will be looked back upon fondly. To Daniel Nicholson, thanks for being such a good mate. This whole experience

would have been dreadful without you. I must also thank the other members of the London NERC DTP. I feel so privileged to have known you all over these past years. You guys have provided some of the most spectacular days and nights, but you've also pushed me to be a better person and a better scientist.

To those staff, students, office mates, collaborators and friends at Queen Mary, thank you for being the warm and welcoming community that you are. I could always rely on you to be approachable, for a friendly chat or a technical discussion. To those individuals at the Centre for Advanced Spatial Analysis at UCL, thank you for being such a delightful bunch of people to be around. Navigating this process from the point of view of two different departments has kept these four years fresh.

To my parents, my brother and my family. I am extremely grateful for your persistent patience and the support that you have provided me, both during the PhD and beyond it. To my close friends: James Boyle, Michael Joyce, Timothy May, Charlotte Parker and Timothy Simmonds. Your company and willingness to be dragged along to random gigs are always greatly appreciated, I feel very lucky to be able to call you friends.

Finally, to my supervisor and friend, Steve Le Comber. It is difficult to express the extent of my gratitude towards everything you did for me during this PhD. I owe you so much for your relentless enthusiasm and yes-man attitude that pulled me through a lot of the hardships. Whenever I met a problem, I would wander over to your office, talk it through and I would always leave with a new found sense of motivation to attack the seemingly impossible problem. I sincerely wish that I could share this accomplishment with you, and that we might chat about where the loose ends of the past 4 years were tied up. I am truly grateful for the time that I knew you. You were more than a supervisor, you were my friend and for that Steve, I once again thank you.

Contents

List of Figures	8
List of Tables	12
Publications and Presentations	13
1 Introduction	15
1.1 A brief history of geographic profiling	16
1.2 The criminal geographic targeting algorithm	17
1.2.1 Applications in ecology and epidemiology	19
1.2.2 Disadvantages of Rossmo’s CGT algorithm	21
1.3 A Bayesian approach	21
1.4 Mixture models	27
1.5 Applications in ecology and epidemiology, revisited	30
1.6 Markov Chain Monte Carlo methods	30
1.7 Platforms for building a geographic profiling model	36
1.8 Conclusions and thesis objectives	37
2 Building a Poisson Geographic Profiling Model	39
2.1 Introduction	40
2.2 Model derivation	42
2.3 An MCMC algorithm for the new model	45
2.3.1 Metropolis-Hastings steps	46
2.3.2 Gibbs sampling step	46
2.4 Volume under surface approximation	49
2.4.1 Deriving an approximation	49
2.4.2 Deriving an upper bound on the error	50

2.5	Integrating out the expected event size	51
2.6	Discussion and conclusions	53
3	Applying a Poisson Geographic Profiling Model	56
3.1	Introduction	57
3.2	Simple exploratory examples	57
3.3	Poor mixing and Metropolis-Hastings coupling	59
3.4	Power analysis	63
3.4.1	Methods	63
3.4.2	Results	65
3.5	Bromeliad analysis	69
3.5.1	Over dispersion in ecological count data	69
3.5.2	Model settings	72
3.5.3	Results	73
3.6	Discussion and conclusions	74
4	Building a Prevalence Geographic Profiling Model	78
4.1	Introduction	79
4.2	Model derivation	81
4.3	Data and model settings	83
4.4	Results	86
4.5	Discussion and conclusions	89
5	A Gaussian Finite Mixture Model Beyond the Normal	92
5.1	Introduction	93
5.2	Methods	95
5.2.1	The Laplace and Cauchy distributions	95
5.2.2	Model derivation	97
5.2.3	Proposal distributions	98
5.2.4	Data and model settings	100
5.3	Results	102
5.4	Discussion and conclusions	105
6	Conclusion	108
6.1	On the developments and findings of this thesis	109
6.2	Points for further development	111

6.3	Contextualising geographic profiling	114
6.4	Linking geographic profiling to other models	116
6.5	Conclusion	118
Bibliography		119
A Additional Work		135
A.1	Error in dispersal literature	135
A.2	Non-conjugate Dirichlet process mixture model	137
B R Package: <i>silverblaze</i>		139
B.1	Landing page	139
B.2	Basic tutorials	140
B.3	Advanced tutorials	141
B.4	Function list	143

List of Figures

1.1	Rossmo’s choice of dispersal kernel, a combination of buffer zone and distance decay, used to denote the score of a crime given its distance from an anchor point (at zero).	18
1.2	The jeopardy surface produced in Rossmo et al. (2014).	19
1.3	The normalised likelihood, prior and posterior distributions for a set of simulated data. The toy anchor point (cross) and crimes locations (dots) are plotted along the x-axis.	24
1.4	Jeopardy and posterior surfaces produced by the CGT algorithm and O’Leary’s Bayesian model respectively when analysing a set of crimes (dots) distributed around two anchor points (crosses)	25
1.5	Jeopardy and posterior surfaces produced by the CGT, single-source Bayesian and DPM models for the same data in Figure 1.4	29
1.6	A map illustrating the diversity and locations of studies within the geographic profiling literature.	31
2.1	The process of simulating a set of count data under the Poisson finite mixture model.	44
3.1	Geographic profiles produced by the Poisson and DPM models when analysing count data (a and c) and point-pattern data (b and d), respectively. Panels a) and b) show the top 20% of the geographic profiles, whilst c) and d) show the top 75%. Geographic profiles associated with the Poisson model were produced via MCMC sampling (algorithm 1) within the <i>Silverblaze</i> R package whilst the profiles associated with the DPM model were produced via the sampling protocol from the <i>RgeoProfile</i> package (Verity et al. 2014, Faulkner et al. 2016).	58

3.2	An example of manually simulated count data under the Poisson finite mixture model. Red sentinel sites denote those that observe one or more events and grey sites are those containing no observations. Each source location is marked with a cross. This map was created using the <i>CartoDB positron</i> layer via the R package <i>leaflet</i> (Cheng et al. 2019).	60
3.3	Results from five different MCMC chains running the Poisson finite mixture model. The output consists of 95% credible intervals for log-likelihoods, dispersal σ_{c_i} and expected event size λ . True values for the dispersal and expected event size are marked with red horizontal lines.	61
3.4	Results from running the Poisson model on the same data, ten times, five with (triangles) and five without (circles) Metropolis-Hastings coupling turned on. The output consists of 95% credible intervals for the log-likelihood, dispersal σ_{c_i} and expected event size λ	63
3.5	Tile plots illustrating the error incurred under different K models when selecting K via the deviance information criterion. From left to right the panels show results for a) every sample size, b) a small sample size (25), c) a medium sample size (100) and d) a large sample size (400).	66
3.6	Estimates for σ_{c_i} across priors and differing expectations on the number of events, λ .	66
3.7	Estimates for λ given each of the true rates: a) 100, b) 1,000 and c) 10,000 across each prior (tight and wide).	67
3.8	A map of the trap surveillance data and bromeliad breeding sites from Wilke et al. (2018, 2019). Traps are coloured depending on the number of mosquitoes caught at each site. Bromeliad patches containing <i>Ae. aegypti</i> larvae are marked with a cross. Map created using the <i>ESRI ocean</i> layer via QGIS.org (2021).	69
3.9	A frequency plot illustrating the number of mosquitoes caught in each trap.	70
3.10	The geographic profiles in Miami-Dade County Florida, created by a) the negative binomial model via the 2017 mosquito count data under informative priors ($K = 14$) and b) the DPM model via repeat point-pattern data ($K = 91$). Locations of bromeliad breeding sites are marked with a cross. Given the proximity between positive and empty traps, some positive traps are only visible in panel b).	74

3.11	The deviance information criterion for each negative binomial model searching for K source locations under different priors. Of these models, the most suitable, indicating the best number of source locations describing the data, was greatly affected by the combination of priors on parameters. For each prior combination, the best model is marked with a diamond (corresponding to minimum DIC). Full DIC values are displayed in panel a) with a zoomed version in panel b).	75
4.1	The prevalence data set in Uganda. This map was created using the <i>OpenStreetMap.de</i> layer via QGIS.org (2021).	84
4.2	Two metrics for judging the number of mixture components K that best describe the prevalence data: a) the conventional deviance information criterion, where each DIC value is coloured based on the proportion of samples allocating data to that given K model and b) the alternative allocation sampling, where the realised number of sources with the maximum proportion of allocation are marked with a cross. . .	87
4.3	The geographic profiles and risk maps produced for each model type metric: the DIC (a and d), realised sources (b and e) and the combined DIC (c and f).	88
5.1	Probability densities of the univariate Cauchy, Laplace and normal distributions. .	97
5.2	Point-pattern data concerning positive tests of malaria in Cairo (black dots). Water bodies containing the species responsible for the transmission of malaria are marked with a cross. This map was created using the <i>Voyager (no labels) retina</i> layer via QGIS.org (2021).	101
5.3	DIC values (a) and hit score percentages (b) for each dispersal kernel model. The hit scores in b) are associated with the K model determined via the lowest DIC value in a). The dashed line in b) indicates those hit scores above and below the value obtained from a random search (50%)	103
5.4	Geographic profiles created when fitting a finite mixture of bivariate a) normal, b) Laplace or c) Cauchy distributions. Data points associated with households testing positive for malaria are coloured based on the source they are allocated to. Known bodies of water harbouring breeding mosquitoes are marked with a cross.	104
5.5	Source dependent weight and scale estimates for each dispersal kernel.	105
A.1	A simulated set of point-pattern data (a) transformed into histograms that (b) correct and (c) do not correct for the area of each bin.	136

A.2	The search and filtering process for identifying suitable literature. The number of studies in each category is shown in brackets.	136
A.3	An illustration of a simple non-conjugate Dirichlet process mixture model fitting source locations in Cartesian space. Contours represent the non-conjugate prior (orange) and posterior (purple) densities, where data points and source locations are black and red dots respectively.	138

List of Tables

1.1	A list of available software, in addition to their cost and type, for implementing different geographic profiling models.	36
2.1	Examples of spatial point-pattern and count data common to criminology, ecology, epidemiology.	40
2.2	Terminology used in geographic profiling (GP) and species distribution models (SDM) alongside joint terms used in this thesis	41
2.3	A list of parameters and their meanings adopted throughout this chapter.	43
3.1	Gini coefficients for the extensive set of simulated data. Each cell represents the average coefficient across one hundred replicates.	68
3.2	AIC values for those models fitting a linear and quadratic relationship between the count mean and variance at pseudo-sources dictated via each grid resolution.	71
4.1	An example of spatial count and prevalence data sets.	80
4.2	Parameter estimates for σ and λ under different models.	87
5.1	An example of spatial point-pattern data analysed in this chapter.	95
6.1	The different spatial data types analysed within this thesis. From left to right: point-pattern, count and prevalence data.	109

Publications and Presentations

Publications

- Stevens & Faulkner (2018): Stevens, M. C. A. & Faulkner, S. C. (2018), 'Geographic Profiling: murder, maths, malaria and mammals', *Chalkdust magazine* pp. 18-25.
- Faulkner et al. (2018): Faulkner, S. C., Stevens, M. C. A., Romañach, S. S., Lindsey, P. A. & Le Comber, S. C. (2018), 'A spatial approach to combatting wildlife crime', *Conservation Biology* pp. 1-24.
- Aygin et al. (2019): Aygin, D. T., Cox, L. A., Faulkner, S. C., Stevens, M. C. A., Verity, R. & Le Comber, S. C. (2019), 'Double cross: geographic profiling of V-2 impact sites', *Journal of Spatial Science* pp. 1-12.
- Heald et al. (2019): Heald, O. J., Fraticelli, C., Cox, S. E., Stevens, M. C. A., Faulkner, S. C., Blackburn, T. M. & Le Comber, S. C. (2019), 'Understanding the origins of the ring-necked parakeet in the UK', *Journal of Zoology* pp. 1-11.
- Stevens, Ray, Faulkner & Le Comber (2020): Stevens, M. C. A., Ray, G., Faulkner, S. C. & Le Comber, S. C. (2020), 'Investigating Sherlock Holmes: Using Geographic Profiling to analyse the novels of Arthur Conan Doyle', *The Professional Geographer* pp. 1-9.
- Stevens, Chen, Stringer, Clemmow & Lewis (2020): Stevens, M. C. A., Chen, Y., Stringer, A., Clemmow, C. & Lewis, A. (2020), Key factors driving obesity in the UK, in 'Proceedings of the 28th conference for Geographic Information Systems Research in the UK (GISRUK)'.
- Stevens et al. (2021): Stevens, M. C. A., Faulkner, S. C., Wilke, A. B. B., Beier, J. C., Vasquez, C., Petrie, W. D., Fry, H., Nichols, R. A. Verity, R. & Le Comber, S. C. (2021) 'Spatially clustered count data provide more efficient search strategies in invasion biology and disease control', *Ecological Applications*.

Presentations

- Poster presentation: 'The impact of "absent" crime on a geographic profile' - NERC student conference 2017: Frontiers in natural environmental research.
- Poster presentation: 'Can we utilise "absences" in data to better locate the source of an invasive species using geographic profiling?' - The British Ecological Society's annual meeting 2017: Ecology across borders.
- Oral presentation: 'Incorporating absences into geographic profiling' - NERC student conference 2018: A changing planet.
- Poster presentation: 'Exploring alternative likelihoods in geographic profiling with the R package *silverblaze*' - Bayesian Computation 2020.

Chapter 1

Introduction

Abstract:

Geographic profiling was established in criminology to aid in investigations of serial crime such as murder, rape and arson. The finite resources of such operations are often diluted by the overwhelming number of potential suspects that require investigating. Geographic profiling offers a solution to this overload of data by providing a manner to prioritise a list of suspects using the spatial locations of an offender's crimes. The fundamental objective of the model is to identify spatial regions that are most likely to contain an offender's base of operation, such as their home or workplace. Almost two decades following the first geographic profiling model, it was shown that such methods could be applied to problems in ecology and epidemiology, targeting nests of invasive species or origins of infectious diseases. Several methodologies have been employed to resolve the objective of geographic profiling; beginning with a simplistic model that scores different regions based on their distance from a set of crimes, to rigorous statistical models that derive explicit estimates for each base of operation. In this introduction, I will provide a detailed description of the methods and applications linked to the past 30 years of geographic profiling. I will showcase the model's many successes, however I also highlight the ample room for further development, motivating the contents of this thesis.

1.1 A brief history of geographic profiling

In criminology, investigations into serial crime often generate a number of suspects far too large to systematically interview, DNA test or survey. In the late 1970s, the investigation into Peter Sutcliffe, also known as the Yorkshire ripper, generated over 268,000 suspects and flagged 5.4 million vehicle registrations (Doney 1990). The next biggest manhunt in UK history resulted in the arrest of Clive Barwell, an individual responsible for five serial rapes in Leeds, Leicester and Nottingham between 1982 and 1995 (Rossmo 2005*b*). A partial fingerprint was recovered from one of the crime scenes but existing technology did not allow for automatic fingerprint checks of partial prints. As a potential plan of action, those involved with the investigation were to manually cross check the partial print with that of over two million individuals.

In 1987, Kim Rossmo identified this persistent problem of large suspect lists in serial investigations and began development on a spatial model to solve this problem (Rossmo 1987). Using the theory of criminal psychology (Brantingham & Brantingham 1981, 1984), Rossmo produced a model that offered a way to prioritise these large lists of suspects into something more manageable (Rossmo 1993). In his thesis, Rossmo formally defined the criminal geographic targeting (CGT) algorithm to solve the following problem. Given a set of locations, all linked with the same investigation of serial crime, can a criminal's base of operation (referred to in criminology as an anchor point) be efficiently identified (Rossmo 2000)? This model and subsequent developments by Rossmo and other authors are referred to as *geographic profiling*.

Rossmo's method did indeed prove to find offenders efficiently. Equipped with geographic profiling, the investigation into Clive Barwell's crimes identified his home and his mother's home by searching less than three percent of the total area he was suspected to be living in. In 2006, a study showed that the CGT algorithm could be applied to problems in ecology (Le Comber et al. 2006). In this case, the study did not focus on finding the home of an offender, given their crimes, but rather, used sightings of a particular species in order to infer its nesting site(s). A separate study demonstrated that the CGT algorithm could be applied to cases in epidemiology, using households testing positive for an infection to find water bodies harbouring vectors responsible for the disease's transmission (Le Comber et al. 2011).

Despite its many successes, Rossmo's method would come under heavy criticism over the two decades following its invention for being unable to explicitly quantify model accuracy. Around

2009, Mike O’Leary would recognise that the underlying process generating a series of data could be parametrised by a probability distribution and in this manner built a Bayesian model such that the uncertainty in estimating an anchor point, for example, could be stated (O’Leary 2009). Finally, Verity et al. (2014) unified the concepts Rossmo and O’Leary brought to geographic profiling and described a Dirichlet process mixture (DPM) model that produced significantly better results over the CGT algorithm and O’Leary’s Bayesian model.

In this chapter, I will walk the reader through these pivotal moments of methodological development, focussing in on each new model’s process for reaching a solution to the original problem of geographic profiling. At each milestone, I will pause to elaborate on the many applications that each model provides whilst highlighting the gaps in each model motivating the incremental developments over the past 30 years, ultimately leading to the basis with which this thesis will be built.

1.2 The criminal geographic targeting algorithm

In his book, Rossmo laid the foundations of geographic profiling by describing the criminal geographic targeting algorithm in four main stages (Rossmo 2000). Firstly, generate the domain over which anchor points will be searched for. This domain is created by choosing a search area based on the crime locations and discretizing this space into a grid of cells. Next, specify the measure of distance used to describe how an offender is travelling. Crimes occurring within city streets prompted the so-called Manhattan distance metric

$$d(a, b) = |x_a - x_b| + |y_a - y_b|, a = (x_a, y_a), b = (x_b, y_b), \quad (1.1)$$

where $d(\cdot)$ denotes the chosen metric that describes the distance between points a and b . This distance metric is suitable for describing individuals that navigate through cities, but for those that can travel in straight lines, the Euclidean Distance is adopted,

$$d(a, b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}. \quad (1.2)$$

Rossmo goes on to encode offender behaviour by describing where an offender will commit a crime, based on this distance d from an anchor point. A criminal will not commit crimes close to home out of fear of being recognised, justifying the need for a buffer zone around an anchor point. Too far from home and the crime becomes too costly to commit, thus the need for a distance decay function (Brantingham & Brantingham 1981). A one dimensional version of this configuration can

be seen in Figure 1.1. Formally, the score for each cell (\hat{x}) within the domain, given a set of crime

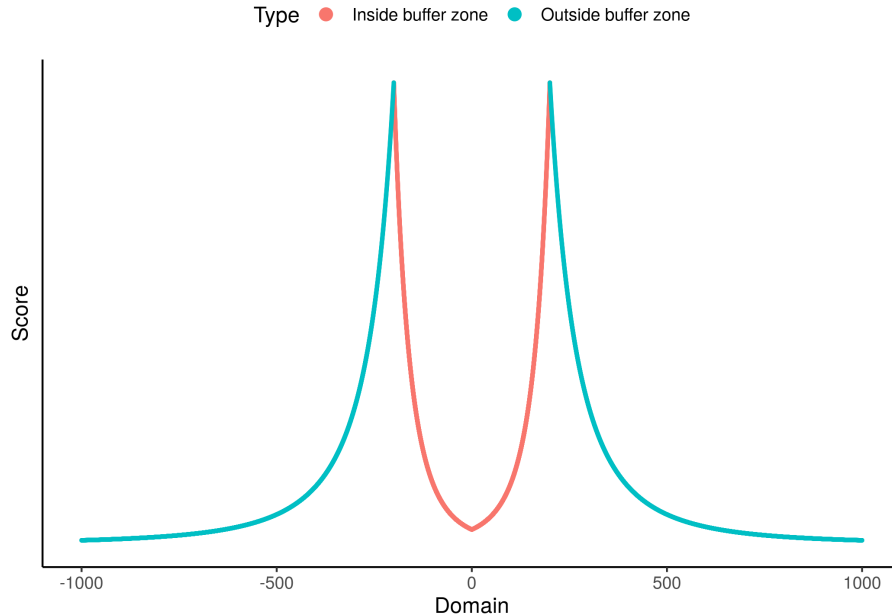


Figure 1.1: Rossmo’s choice of dispersal kernel, a combination of buffer zone and distance decay, used to denote the score of a crime given its distance from an anchor point (at zero).

locations, is obtained via the function:

$$f(\hat{x}) = k \sum_{i=1}^C \left[\left(\frac{\phi}{d(c_i, \hat{x})^f} \right) + \left(\frac{(1 - \phi)(B)^{g-f}}{(2B - d(c_i, \hat{x}))^g} \right) \right], \quad (1.3)$$

where a cell with a higher score denotes an area more likely to contain an anchor point than one with a low score. Here, B is the distance defining the buffer zone and d is the distance between each cell of the domain (\hat{x}) and crime i . The value ϕ , set to either zero or one, indicates if a distance is inside or outside the buffer zone:

$$\phi = \begin{cases} 0 & \text{if } d \leq B, \\ 1 & \text{if } d > B. \end{cases} \quad (1.4)$$

C is the total number of crimes and g , f , and k are all pre-determined constants fitted from the data. Essentially this procedure is akin to kernel density estimation, where the kernel in question is specified as in Equation 1.3 to encode offender behaviour and g and f control the kernel’s bandwidth. Once scores have been calculated for each cell in the domain, the resulting three-dimensional surface is lain over the grid to obtain a *jeopardy surface*. An example can be seen in Figure 1.2.

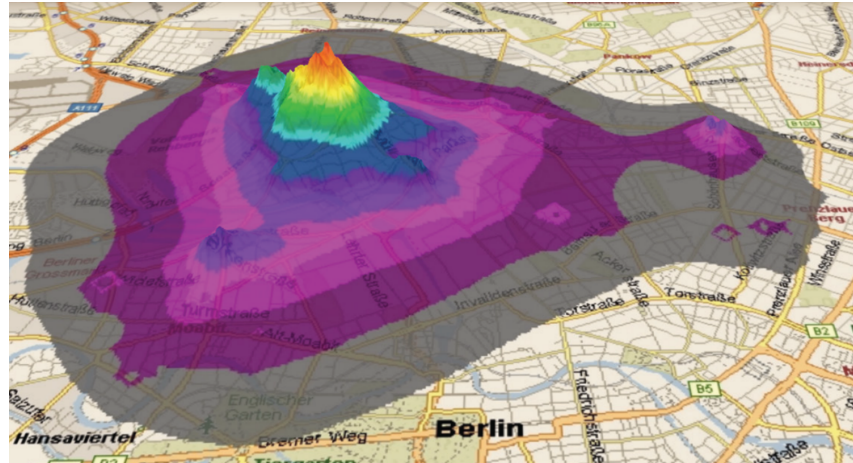


Figure 1.2: The jeopardy surface produced in Rossmo et al. (2014).

A list of suspects are then investigated, where priority is given to those that reside at the highest parts of the jeopardy surface. A metric used throughout geographic profiling literature to determine the efficiency of a search strategy is through a *hit score percentage*. This value is calculated by dividing the area searched before an anchor point is found by the total search area in question. It follows that the lower the hit score percentage, the more efficiently its respective anchor point was found. If an area were to be searched at random, an anchor point would be expected to be found, on average, by searching 50% of the total area (Rossmo 2000).

To this day, geographic profiling is still being used in criminology (Rossmo & Harries 2011, Rossmo 2012, Hipp & Williams 2020). Recent studies have implemented geographic profiling to target various forms of criminal activity such as cybercrime, card skimming and spear phishing (Mburu & Helbich 2015, Butkovic et al. 2018). The CGT algorithm has also been used to find anchor points when targeting criminal gangs in cordon and search operations (Huddleston et al. 2013).

1.2.1 Applications in ecology and epidemiology

Le Comber et al. (2006) revolutionised the field of geographic profiling by demonstrating its applicability to problems in environmental science, focussing specifically on applications in ecology and epidemiology. In ecology, data were no longer associated with locations of crime but instead with sightings of an invasive species or an elusive animal in order to identify its nesting sight. Similarly, in epidemiology the data being analysed consisted of households testing positive for an infectious disease in order to infer sites responsible for early transmission. Le Comber et al. (2006) applied the CGT algorithm to two pipistrelle bats (*Pipistrellus pipistrellus* and *P. pygmaeus*) in the north east of Scotland. Not only did the model successfully identify each species' roosting sites

more efficiently than a random search, but it was also able to distinguish between their foraging patterns.

Since this 2006 study, there have been many other instances in which the CGT algorithm has been applied to ecological problems. For example, geographic profiling was used to study white shark (*Carcharodon carcharias*) predation on Cape fur seals (*Arctocephalus pusillus pusillus*) (Martin et al. 2009). Rossmo imposed a buffer zone on criminal behaviour, but in biology, this may not be suitable since, synonymously to criminology, this would mean an animal purposefully avoids foraging close to a central location. However, this study concluded that a buffer zone is present, since a shark engaging in an attack too close to shore risks its prey escaping back onto land. Raine et al. (2009) and Suzuki-ohno et al. (2010) investigated experimental studies for bumble-bee foraging using geographic profiling. In these studies, the CGT algorithm was able to distinguish between different simulated foraging patterns in the species *Bombus terrestris* and demonstrated the capabilities of the model when applied to laboratory experiments.

Multiple studies have also dealt with applying geographic profiling to invasion biology. Stevenson et al. (2012), for example, used the CGT algorithm to identify central points of interest for 53 different invasive species in Great Britain. The results of this analysis found that the CGT algorithm identified points of origin more efficiently than a search strategy obtained via simple metrics of spatial tendency such as the spatial mean and median. In addition to comparing results to these simple metrics, the CGT was compared to another kernel density estimator (Worton 1989) when searching for points of origin of the invasive giant hogweed, *Heracleum mantegazzium*. Hit score percentages were again found to be significantly better using the CGT algorithm.

Le Comber et al. (2011) was the first instance that geographic profiling entered the field of epidemiology. This study investigated two sets of epidemiological data: John Snow's classic case of cholera from the Broad street pump in Soho, London and cases of malaria from the mosquito vector *Anopheles gambiae* in Cairo, Egypt. Similarly to Stevenson et al. (2012), the algorithm was compared to other simple metrics of spatial tendency and proved to find sources of infection more efficiently in the majority of cases.

In a short letter, Le Comber & Stevenson (2012) described the growing presence of geographic profiling across multiple disciplines, citing much of the literature above. This letter demonstrated that geographic profiling also has the potential to be applied to novel problems. For example, this

letter revisited the infamous case of Jack the ripper, applying the CGT algorithm to the five body dump sites associated with the investigation. Another study revisited the historical case of Otto and Elise Hampel. This couple were prosecuted for dropping anti-Nazi propaganda across Berlin in 1941. Given these locations, the CGT algorithm efficiently identified the home and work places of the couple, in addition to the homes of their relatives (Rossmo et al. 2014).

1.2.2 Disadvantages of Rossmo’s CGT algorithm

Despite an extensive list of successful applications spanning multiple disciplines, the accuracy of geographic profiling to identify an offender’s anchor point(s) has been heavily criticised. In many studies, individuals have been asked to predict the location of an offender’s anchor point given a map of crimes. Participants were then introduced to some simple geographic profiling concepts and asked to make predictions based on new maps. Snook et al. (2004) described how participants with these simple heuristics were able to predict anchor points as accurately as geographic profiling algorithms. As pointed out by Rossmo however, participants were asked to make predictions based on small sample sizes ($n = 3$) for this study (Rossmo 2005a). Further studies have questioned the accuracy of the model compared to individuals with simple heuristics (Paulsen 2006a, Harries & LeBeau 2007). Studies have also concluded geographic profiling algorithms are more efficient than a human equipped with a heuristic (Bennell et al. 2009).

Rossmo repeatedly remarks that geographic profiling *is not an X marks the spot* kind of model. Rather, it is a tool to be utilised by criminal investigations to prioritise questioning a list of suspects (Rossmo 2000). As described above, the methodology has come under fire for only being able to measure accuracy via the hit score metric. What if the process of crimes generated from an anchor point were governed by some underlying probability distribution? Then, it would be possible to treat an anchor point as an unknown parameter, estimate it and state the uncertainty associated with it. The jeopardy surface produced by the CGT algorithm is not probabilistic since the underlying decay functions that are summed to produce a jeopardy surface generate scores in place of probabilities. Given a set of crime sites, geographic profiling tries to infer the location of the criminal’s anchor point(s), so it seems intuitive that a Bayesian approach is suitable.

1.3 A Bayesian approach

In 2009, Mike O’Leary produced a solution to the problem of measuring model accuracy in geographic profiling. In his model, an anchor point is treated as an unknown parameter that should

be estimated under a Bayesian framework (O’Leary 2009, 2010a). Bayesian analysis originated in the 18th century, where the Reverend Thomas Bayes derived a *posterior* distribution for the location of a ball on a table based upon the proximity of this to other balls being dropped onto the same table (Bayes & Price 1763). In the context of geographic profiling, O’Leary derived the posterior distribution on an anchor point, based upon the locations of a set of crimes. Bayes’ theorem provides the probability that an offender resides at the anchor point μ , given a crime at x ,

$$\Pr(\mu|x) = \frac{\Pr(x|\mu) \cdot \Pr(\mu)}{\Pr(x)} \propto c \cdot \Pr(x|\mu) \cdot \Pr(\mu), \quad (1.5)$$

where c is some normalising constant. O’Leary’s model works under the assumptions that crimes occur independently of each other, and, originate from a single anchor point. The model tries to calculate the posterior distribution on the anchor point $\Pr(\mu|x)$ which can be expressed as a combination of the probability of observing the data (the likelihood), and the prior belief on where this anchor point could be (the prior). In order to compute this posterior, the model requires a user to specify said likelihood ($\Pr(x|\mu)$) and prior ($\Pr(\mu)$).

A suitable likelihood is one that encodes the process of committing crimes and captures the underlying behaviour of the criminal in question. As stated in Rossmo (2000), a criminal is unlikely to commit crimes close to home out of fear of being recognised whilst avoiding locations too costly to reach, hence the kernel in Figure 1.1. O’Leary (2009) argued that a Gaussian distribution is a reasonable choice to capture offender behaviour and, as such, a normal distribution has been adopted in his example below. Then, the probability of observing a crime at x given μ is governed by the normal density:

$$\Pr(x|\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (1.6)$$

Equation 1.6 is then extended to describe the likelihood of observing a set of crimes $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$

$$\Pr(\mathbf{x}|\mu) = \frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{i=1}^n e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}, \quad (1.7)$$

assuming each criminal offence occurs independently of every other. The choice to adopt a Gaussian kernel may seem counter-intuitive as Rossmo’s kernel in Figure 1.1 dictates the mostly likely place for crime to occur is at a specific distance from an anchor point (the border of the buffer zone). In contrast, O’Leary’s Gaussian kernel places the most likely location for crime on top of an offender’s anchor point. Ultimately, the decision of the most suitable kernel should be influenced by the phenomenon being studied, but other kernels such as the negative exponential or log-normal

distributions could be adopted (O’Leary 2009). This section follows a Gaussian distribution to demonstrate the architecture of O’Leary’s model but a full discussion on suitable kernels appears in chapter 5.

Next, a user is required to specify a suitable prior on μ . With no inclination to where the culprit resides, an uninformative prior could be chosen. Suppose it is known that the offender’s residence lies within some interval ($\mu \in [a, b]$), then a prior to reflect this knowledge could be uniform on this interval and zero everywhere else. A more informative prior could come in the form of a normal distribution centred on a simple spatial metric, such as the spatial mean or median of the crimes:

$$\Pr(\mu) = \frac{1}{\sigma_p \sqrt{2\pi}} e^{-\frac{(\mu_p - \mu)^2}{2\sigma_p^2}}, \quad (1.8)$$

where μ_p is the prior mean and the extent to which the prior informs the model can be controlled via the standard deviation σ_p . Combining the likelihood and normal prior leads to the posterior:

$$\Pr(\mu|\mathbf{x}) \propto \frac{1}{\sigma_p \sqrt{2\pi}} e^{-\frac{(\mu_p - \mu)^2}{2\sigma_p^2}} \cdot \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}. \quad (1.9)$$

Figure 1.3 shows an example of O’Leary’s model in action. It demonstrates that his method is simply an implementation of the classic Bayesian approach to estimating a parameter by combining information from the data, encoded by the likelihood (dashed red line), with prior beliefs (dotted blue line), to obtain the posterior of interest (solid green).

The normal density was chosen to describe the process of generating the data where its standard deviation σ was assumed to be known. O’Leary points out that this parameter translates to the distance willing to be travelled by an individual (O’Leary 2010a). This so called *mobility* parameter reflects the ability of an individual to navigate space. One would expect for example, a much smaller value of σ for a delinquent without a driver’s license compared to an adult. If a suitable value for σ cannot be chosen then it is possible to integrate it out:

$$\Pr(\mu|\mathbf{x}) = \int_0^\infty \Pr(\mathbf{x}|\mu, \sigma) \Pr(\mu) \Pr(\sigma) d\sigma, \quad (1.10)$$

resulting in the posterior on μ based entirely on the data collected and the prior on μ . Here, $\Pr(\sigma)$ denotes the prior on σ .

In addition to assuming a normal distribution on crime locations, O’Leary works through a few

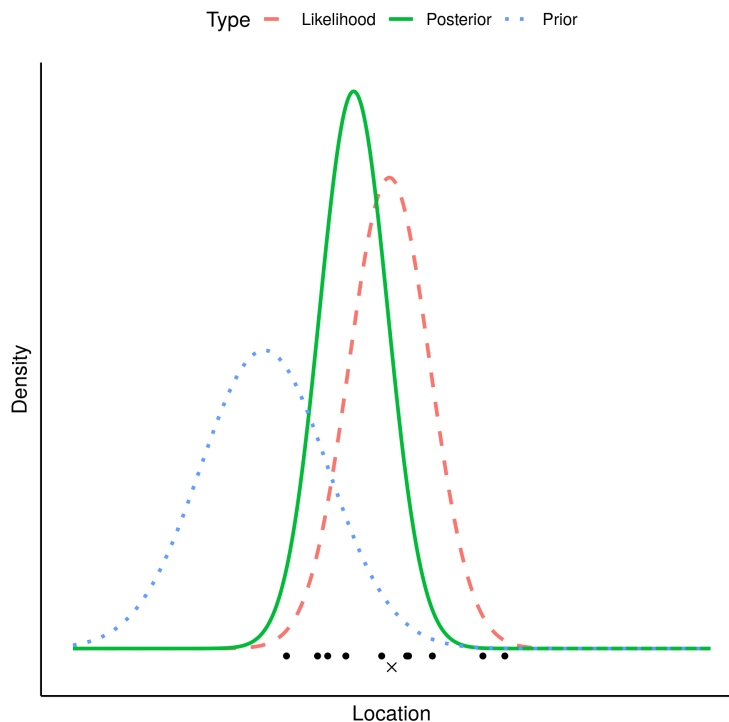


Figure 1.3: The normalised likelihood, prior and posterior distributions for a set of simulated data. The toy anchor point (cross) and crimes locations (dots) are plotted along the x-axis.

more examples to demonstrate other potential forms for the likelihood and priors. In one instance, O’Leary expands upon the likelihood to account for offender choice in target selection. A function $G(x)$ is introduced, that takes the form:

$$G(x) = \begin{cases} 1 & \text{if } x \in J, \\ 0 & \text{if } x \notin J, \end{cases} \quad (1.11)$$

to indicate regions where crimes cannot occur, such as bodies of water or where they will not be detected, such as outside jurisdictional boundaries. G could also be considered as part of the prior on an anchor point, since μ is expected to be inside jurisdictional boundaries, similarly to x .

The model described in Mohler & Short (2012) expands on the topic of spatial heterogeneities in geographic profiling. This study built an agent based model for geographic profiling that estimates anchor point density based on criminal behaviour. Similarly to existing models, an offender’s decision to commit a crime at a certain location is dependent on the distance from their anchor point, however, this model also incorporates various other spatial features such as barriers (e.g. water bodies), directional preference (e.g. incline) and site attractiveness (e.g. housing density).

O’Leary offered the first statistically rigorous model for geographic profiling, however, this model, and similar Bayesian models of the time, made the assumption that only a single anchor point was responsible for generating a series of crimes. This assumption is not always valid of course. The beginning of this chapter described operation Lynx, in which Clive Barwell was arrested for committing crimes around two distinct locations: his home and mother’s home. Should the investigation have implemented O’Leary’s method, search strategies would have spuriously focussed on the spatial mean of his crimes, potentially avoiding the two known anchor points.

Figure 1.4 illustrates the pitfalls in O’Leary’s model when analysing a set of data generated from two anchor points. This figure clearly shows that the simple Bayesian model focusses search priority near the mean of these data. The CGT algorithm is superior in this case, as it demonstrates it can produce a multi-modal surface, with peaks near true anchor points.

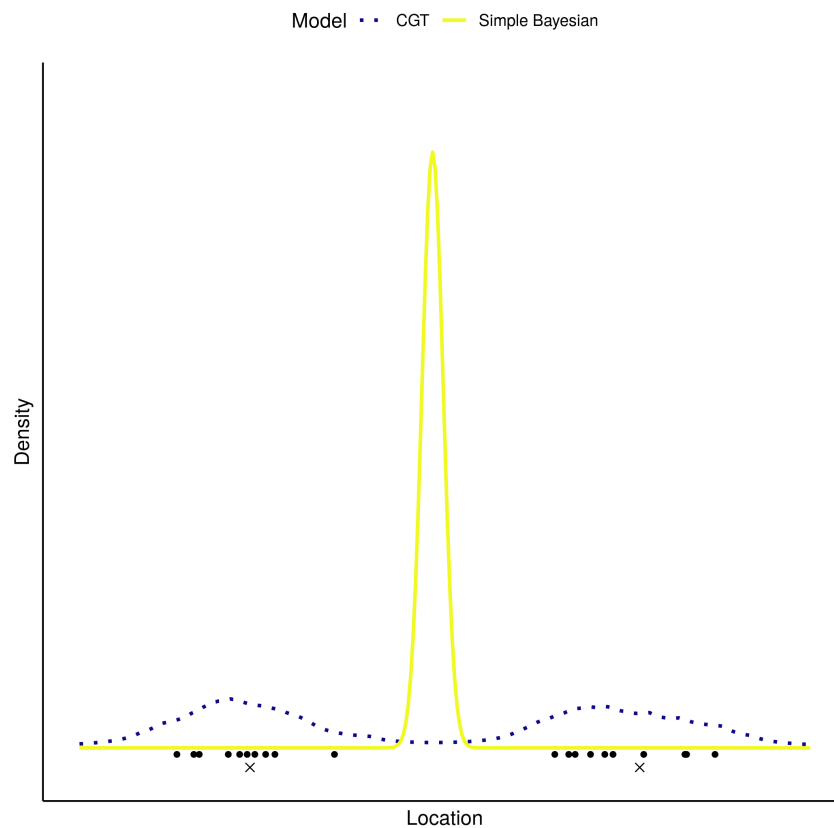


Figure 1.4: Jeopardy and posterior surfaces produced by the CGT algorithm and O’Leary’s Bayesian model respectively when analysing a set of crimes (dots) distributed around two anchor points (crosses)

Rossmo's algorithm has been used to analyse data sets with multiple anchor points (Le Comber et al. 2006, 2011, Rossmo et al. 2014) and in essence, treats each data point as if it originates from its own unique anchor point by summing over multiple distance-decay functions. As previously mentioned (subsection 1.2.2), Rossmo's kernel density estimator is non-parametric and thus its accuracy is hard to measure.

Similarly, the model in Mohler & Short (2012) entertains the idea of multiple anchor points but accomplishes this artificially by estimating a single anchor point for a time series and then summing over every time series to obtain a multi-peaked surface. O'Leary appeared to take the first step in considering a model for multiple anchor points, however his work focussed on combining a weighted average of different models each assuming the same anchor point (O'Leary 2010*b*).

The field of geographic profiling needed some way to combine the advantages of the multi-modal CGT algorithm with the pure mathematical approach of the simple Bayesian model in order to identify multiple anchor points. Five years on from O'Leary (2009) and a solution to the complex problem of estimating an unknown number of anchor points came from the Dirichlet process mixture model in Verity et al. (2014). This powerful class of model allows users to estimate parameters when the number of parameters itself is also unknown. The DPM model can also be referred to as an infinite mixture model, where the number of anchor points is assumed infinite but only a finite number are realised by the data.

Before walking through the DPM model's methodology, I will introduce the reader to finite mixture models as a first solution to estimating multiple anchor points. I will then expand on this, moving onto the DPM model and its applications in geographic profiling. It is worth also briefly discussing the terminology used in geographic profiling. This method is applied to problems in criminology, ecology and epidemiology, hence it is often common for introductory literature, such as this chapter, to confuse the terms used in different disciplines. For example a study aiming to identify the nesting site of an invasive species may refer to a sighting location as a crime. To avoid confusion, I will speak generally of an anchor point, nesting site of an invasive or the origin of an infectious disease as a *source location*, or *source* for short. Similarly, I will refer to the location of a crime, sighting of an organism or an individual testing positive for a disease as *data*. In both cases, I will stray from these generalised terms but only when referring to a particular case study that requires a specific word.

1.4 Mixture models

Expanding on these Bayesian models to accommodate multiple source locations requires a mixture model that introduces the allocation of each data point x_i to a finite set of K source locations (Aitkin 2001). The probability that data point x_i is allocated to source $c_i \in 1 : K$ is denoted as ω_k , for $k \in 1 : K$. Here, ω_k is often referred to as a mixture component's *weight* and the set of weights $\boldsymbol{\omega}$ follows the condition $\sum_{k=1}^K \omega_k = 1$. Note that c_i and k are inter-changeable since both index over the total number of mixture components K , however, c_i provides more information, stating the cluster that data point i originates from. The complete data set \boldsymbol{x} can be thought of as generated under the following model

$$x_i \sim f(\theta_{c_i}), \quad (1.12)$$

$$\theta_{c_i} \sim \mathcal{F}, \quad (1.13)$$

$$c_i \sim \text{Multinomial}(\omega_1, \omega_2, \dots, \omega_k), \quad (1.14)$$

$$\omega_k \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_k). \quad (1.15)$$

Here, the data point, x_i is drawn from density $f(\theta_{c_i})$, where $\theta_{c_i} = (\mu_{c_i}, \sigma_{c_i})$ are drawn from a suitable prior \mathcal{F} . As O'Leary (2010a) remarks, f most likely takes the form of some Gaussian distribution and setting f to the normal density and K to one, leads to the example in section 1.3. Data are allocated to sources via a multinomial distribution dependent on mixture weights and a suitable prior for mixture weights is often chosen to be Dirichlet as its realisations meet the condition that the ω_k sum to one (Aitkin 2001). Such a prior is adopted here, with the hyper parameters α_k controlling the strength of each mixture component's weight.

Adapting O'Leary's likelihood in Equation 1.6 to account for multiple sources leads to the probability of observing x_i

$$\Pr(x_i|\boldsymbol{\theta}) = \sum_{c_i=1}^K \omega_{c_i} \cdot f(\theta_{c_i}), \quad (1.16)$$

which leads to the likelihood of observing the entire data set

$$\Pr(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^n \sum_{c_i=1}^K \omega_{c_i} \cdot f(\theta_{c_i}). \quad (1.17)$$

Note that the likelihood of observing the data is now conditional on the allocation of data to different mixture components.

The finite mixture model described in this section provides a solution to the multiple source problem. However, a fundamental weakness of this model is the requirement to pre-specify the number of mixture components K , i.e. the number of source locations. Although the explicit locations of each source are the desired parameters from a geographic profiling model, a practitioner may also be interested in estimating K . To deal with this, Verity et al. (2014) developed a Dirichlet process mixture model for geographic profiling to not only infer source locations but to also estimate their numbers. Dirichlet process mixture models are particularly useful tools when clustering data given that a user does not need to specify the number of clusters prior to running the model and instead assumes an infinite number of source locations, with which only a finite number are responsible for generating the observed data (Gershman & Blei 2012).

The formulation of the Dirichlet process mixture model can be obtained by slightly altering the model described in Equation 1.13 to Equation 1.15. Data x_i are still drawn from a density parameterised by θ_{c_i} , but allocations c_i are now governed by the Chinese restaurant process (CRP). A description of the CRP can be found in Gershman & Blei (2012), but briefly, this process concerns the problem of allocating customers to an infinite number of tables in a Chinese restaurant. If a table is populated by many individuals, a new customer is more likely to choose this table to sit at. There is however, always some non-zero probability that a new customer will choose to sit at a fresh, unoccupied table. This process is analogous for geographic profiling. Data are allocated to source locations where there is always a non-zero probability that a data point is allocated to a new, unseen source location. Full details of the DPM model can be found in the appendices of Verity et al. (2014).

Figure 1.5 illustrates the output of the three main geographic profiling models: the CGT algorithm, the simple Bayesian model and the DPM model, when analysing the same data shown in Figure 1.4. The DPM model (purple dashed line) clearly demonstrates it is able to accurately return the source locations more efficiently than the CGT algorithm whilst also operating under a rigorous statistical methodology, such as O’Leary’s method.

When implementing a mixture model, a problem arises when component specific parameter estimates are obtained. Specifically, the likelihood of these models remains invariant when assigning labels

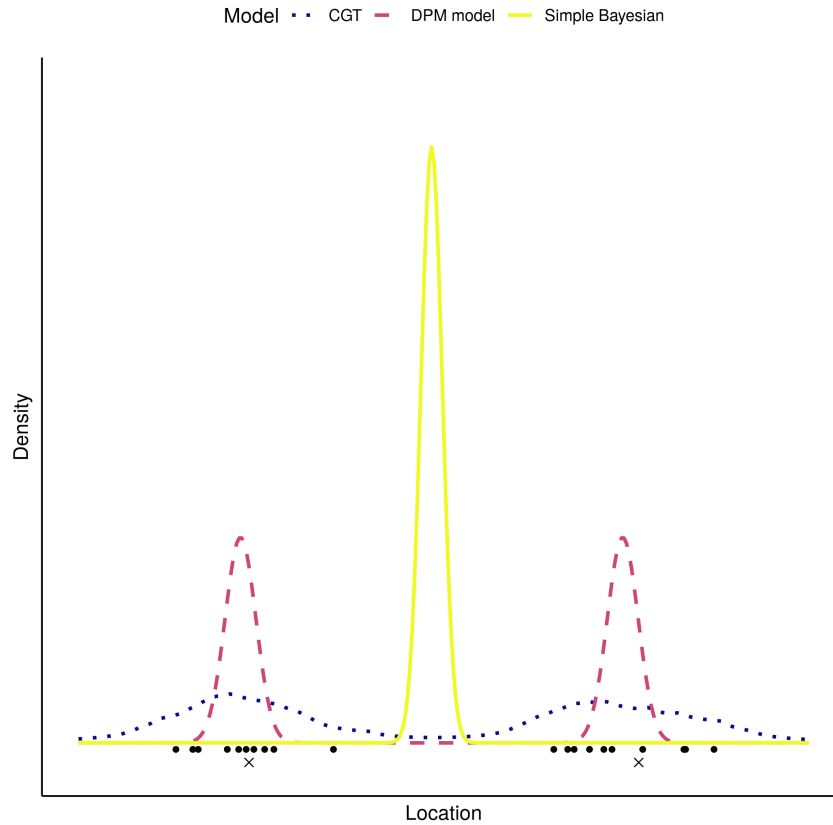


Figure 1.5: Jeopardy and posterior surfaces produced by the CGT, single-source Bayesian and DPM models for the same data in Figure 1.4

to each mixture component. For example, a two-component mixture model may label one source as *source A* and the other as *source B*, but it is entirely possible to swap these labels without effecting the likelihood. When stating parameter estimates of interest (the location of *source A* for example), reporting the posterior mean of the marginal distribution on a component specific parameter will lead to erroneous results, given that this distribution will be multi-modal (for example, see Figure 1.5). To resolve this so-called *label switching* problem, the DPM model employs an algorithm from Stephens (2000b) that aims to assign labels to components by minimising a cost function. For the DPM model, the cost function is the Kullback-Leibler divergence, where the distribution on current component labels is compared to the distribution based upon previous inferences.

Geographic profiling has already seen many applications spanning ecology, epidemiology and criminology. The extension to a Dirichlet process mixture model has been no exception to this. In criminology for example, it was demonstrated that a DPM model could be used in cases of counter-terrorism by considering instances of improvised explosive devices in Ireland to infer locations of

IED factories (Tench 2018).

1.5 Applications in ecology and epidemiology, revisited

In ecology, the DPM model has successfully located sleeper trees of elusive, nocturnal tarsiers (*Tarsius tarsier*) (Faulkner et al. 2015). The model also played a role in identifying the origin of several invasive species: from American mink (*Neovison vison*) in Scotland (Faulkner et al. 2016) to ring-necked parakeets (*Psittacula krameri*) in London (Heald et al. 2019), Siberian chipmunks (*Eutamias sibiricus*) in Italy (Cerri et al. 2020) and African red-headed agamas (*Agama picticauda*) in Florida (Gray 2020). Ecological applications also include addressing human-wildlife conflict, such as interactions between humans and tigers in Sumatra (Struebig et al. 2018) and endangered species and poachers in Zimbabwe (Faulkner et al. 2018).

Beyond ecology, the DPM model has also been applied to cases in epidemiology. The model was used to propose a method for badger culling to eradicate bovine tuberculosis (Smith, Downs, Mitchell, Hayward, Fry & Le Comber 2015). The proposed culling process however, was not shown to efficiently identify infected badger setts. In addition to developing the DPM model, Verity et al. (2014) revisited the malarial vector data in Cairo from Le Comber et al. (2011). The DPM identified sources of malaria more efficiently than Rossmo’s CGT model and O’Leary’s single source Bayesian model.

Finally, the DPM model has contributed several novel applications to the field of geographic profiling. The first of which saw the inference of the identity of the notorious artist Banksy by analysing the locations of his graffiti (Hauge et al. 2016). Aygin et al. (2019) saw the DPM model applied to the locations of V2 rocket landing sites to infer information about a deception from the British government against Germany in World War II. In Stevens, Ray, Faulkner & Le Comber (2020), the model was applied to locations where Sherlock Holmes stories were set in order to infer information about the author, Sir Arthur Conan Doyle.

1.6 Markov Chain Monte Carlo methods

Although finite and infinite mixture models can both estimate multiple source locations, applying these models proves a somewhat troublesome task since a relevant posterior distribution would integrate over all possible allocations of data points to source locations. To put this problem into

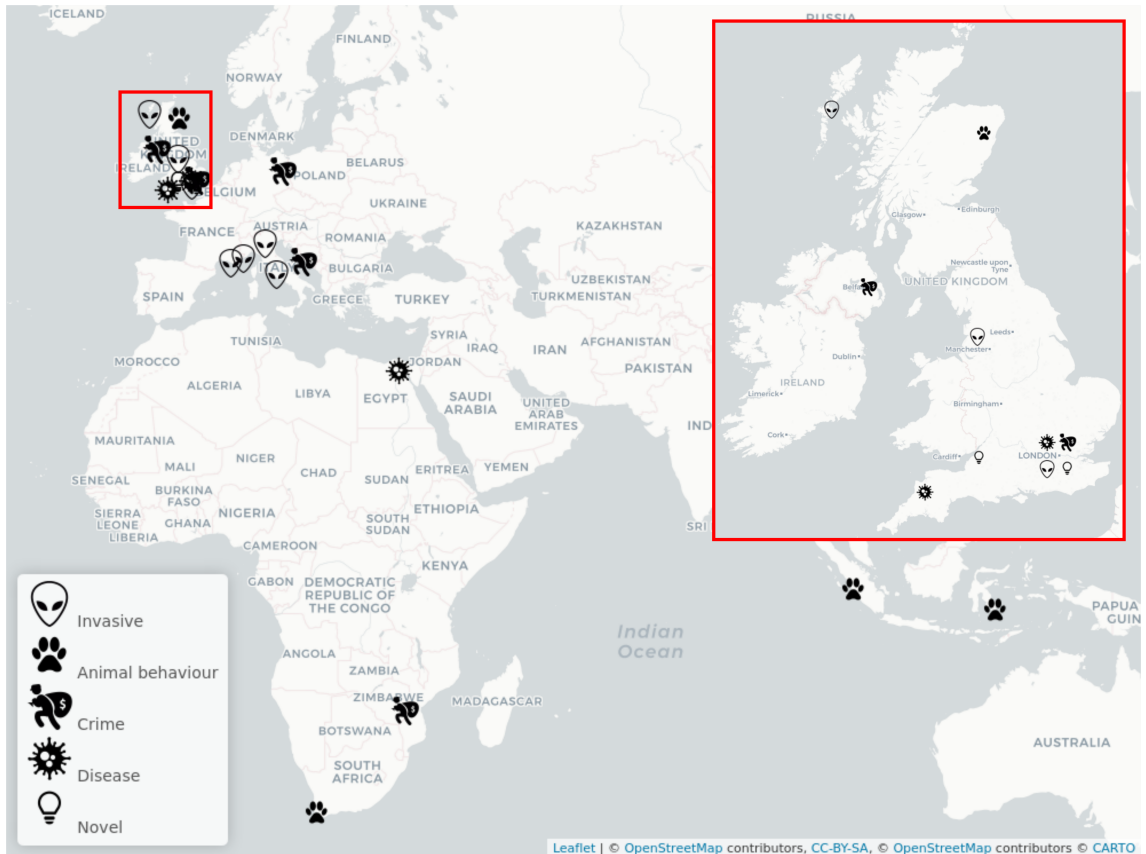


Figure 1.6: A map illustrating the diversity and locations of studies within the geographic profiling literature.

perspective, there are 115,975 ways to partition ten data points among ten clusters. Doubling this, the number of ways to partition twenty data points across twenty clusters increases by eight orders of magnitude to 51,724,158,235,372 partitions. Dealing with an intractable posterior distribution is a common problem in Bayesian modelling and Markov Chain Monte Carlo (MCMC) methods offer a solution (Gelman et al. 2004).

It was realised that information about a posterior distribution can be gathered by sampling from said distribution many, many times (Metropolis & Ulam 1949). An MCMC algorithm aims to do exactly this; collect a reasonably sized set of samples from the posterior distribution in order to draw conclusions about the parameters that govern it. Samples from the posterior of interest are collected by constructing a Markov chain, a sequence of values that is built up over time where, at any instance, the value of a Markov chain at a certain iteration is dependent on the previous iteration's value.

A Markov chain must satisfy a few conditions for its values to be considered samples from the

posterior of interest. Firstly, a Markov chain is required to be *ergodic*, or, as it's referred to in some texts, *irreducible*. An ergodic Markov chain is such that any state can be reached from any other i.e. the transition probability from the current state of a Markov chain to any other is always non-zero (Gelman et al. 2004). If the current state of a Markov chain can return to the same state without the need of taking the same route and the same number of steps then the Markov chain is referred to as *aperiodic*. If a Markov chain holds these ergodic and aperiodic properties then there exists a unique *stationary* distribution for which the Markov chain asymptotically behaves as the size of the chain approaches infinity (Albert 2009). This *stationary* property ensures that, regardless of the transitions that occur from one state to another, the frequency with which a particular state is observed within a Markov chain is proportional to the posterior of interest.

As previously mentioned, the next state of a Markov chain is chosen via some probability distribution parameterised by the current state. This process of randomly sampling over and over is what earns a Markov chain the title of Monte Carlo method. With all MCMC methods, it may take some time for a Markov chain to reach a point where it meets the previously described conditions and thus it is customary to discard the first X iterations of a Markov chain. This interval is known as the *burn-in* period of a MCMC algorithm. After the burn-in phase, a Markov chain is believed to be sampling from the posterior of interest and is thus referred to as the *sampling* period. There are a variety of MCMC algorithms that a practitioner can adopt to sample from a desired posterior. The first covered here is the *Gibbs sampler*, one of the first MCMC algorithms to be implemented for geographic profiling.

Suppose the desired full conditional posterior distribution is governed by the parameter set $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3, \dots, \theta_n\}$. Deriving the marginal distribution for any of these parameters requires an integral over $n - 1$ dimensions. In general, a Gibbs sampler breaks up the complex problem of sampling from the posterior distribution by sampling from the conditional distribution on each unique parameter of interest (by fixing the remaining $n - 1$ parameters) (Geman & Geman 1984). The Markov chain for each θ_i is first initialised by setting $\boldsymbol{\theta} = \{\theta_1^0, \theta_2^0, \theta_3^0, \dots, \theta_n^0\}$, where each θ_i^0 is drawn from a suitably dispersed prior on θ_i . Denote the conditional distribution for θ_1 as

$[\theta_1|\mathbf{x}, \theta_2, \theta_3, \dots, \theta_n]$, then the set of conditional distributions

$$\begin{aligned} &[\theta_1|\mathbf{x}, \theta_2, \theta_3, \theta_4 \dots \theta_n], \\ &[\theta_2|\mathbf{x}, \theta_1, \theta_3, \theta_4 \dots \theta_n], \\ &[\theta_3|\mathbf{x}, \theta_1, \theta_2, \theta_4 \dots \theta_n], \\ &\dots \\ &[\theta_n|\mathbf{x}, \theta_1, \theta_2, \theta_3 \dots \theta_{n-1}] \end{aligned}$$

are used to update each θ_i at each iteration of the MCMC algorithm (Albert 2009).

Verity et al. (2014), appendix 3, describes a Gibbs sampling algorithm for a DPM model from Neal (2000) to estimate the explicit location of each source as well as the allocation of data to spatial clusters (i.e. the number of source locations) (Green & Richardson 2001). The Gibbs sampler in question uses the following steps. Data are allocated to source locations based on the CRP and the initialised Markov chain. The location of each source is then estimated following O’Leary’s framework. Based on these updated source locations, the data are then re-allocated to each source. During this re-allocation step, there is always some non-zero probability that a data point will be allocated to a never seen before source location. This process is repeated many times and allows for the birth and death of sources at each MCMC iteration, providing a way of allowing the data to be described via differing numbers of sources.

Following on from the work of Verity et al. (2014), Faulkner et al. (2016) provide a development to the DPM model that allows a user to estimate the dispersal parameter σ via an additional Gibbs sampling step in the MCMC algorithm. The derivation for the conditional distribution for σ can be found in Faulkner et al. (2016). For mathematical convenience, the mixture components of Verity et al. (2014) and Faulkner et al. (2016)’s DPM models are made up of bivariate normal distributions, hence, the dispersal parameter σ is the standard deviation of these bivariate normals.

The DPM model is an incredibly useful tool, with a myriad of applications spanning a wide number of disciplines (see Figure 1.6). It is cost effective; requiring only a set of longitudinal and latitudinal points from a user. However, its reliance on a Gibbs sampling algorithm restricts the input from a user when building a model. Gibbs sampling relies on conjugacy between prior and likelihood, i.e. the prior distributions chosen must match that of each parameter’s respective

conditional likelihood. This condition leaves very little room for a user to customise priors when running an analysis. For example, the prior on source locations for every study implementing the DPM model from Verity et al. (2014) uses a bivariate normal distribution centred on the spatial mean of the data. The spatial mean may be an unsuitable location for a source, hence a suitable prior should reflect this (Mohler & Short 2012).

In place of Gibbs sampling, a suitable MCMC algorithm that bypasses the requirement of conjugacy between likelihood and prior is the Metropolis-Hastings algorithm. This method provides a user with a way to customise and specify their model with a prior that could take many forms. The Metropolis-Hastings algorithm relies on a system where one proposes a new value for a particular parameter based on the current value and updates said parameter based on the ratio between the likelihood of each parameter set (Metropolis et al. 1953, Hastings 1970, Dunson & Johndrow 2019).

Once again, the Markov chain is initialised by setting $\boldsymbol{\theta} = \{\theta_1^0, \theta_2^0, \theta_3^0, \dots, \theta_n^0\}$. Then for each parameter, a candidate value of θ_i , denoted as θ'_i , is proposed via some *proposal distribution* $g_{\theta_i}(\cdot)$ as the next value in the Markov chain for θ_i . The proposed θ'_i is either accepted, taking its position as the new value in the Markov chain, or it is rejected, and the original θ_i is retained. The probability of accepting the proposed θ'_i is given by:

$$Acc(\theta'_i|\theta_i) = \min\left(1, \frac{L(\mathbf{x}|\boldsymbol{\theta}') \cdot p(\theta'_i)}{L(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\theta_i)} \cdot \frac{g_{\theta_i}(\theta_i|\theta'_i)}{g_{\theta_i}(\theta'_i|\theta_i)}\right), \quad (1.18)$$

where $L(\cdot)$ refers to the likelihood of the data given $\boldsymbol{\theta}$ and the prior on θ_i is denoted by $p(\theta_i)$. Note that this prior can take on many different forms lending an element of flexibility to a user over the Gibbs sampler.

The Metropolis-Hastings algorithm is essentially trying to find peaks of posterior density and Equation 1.18 does so by calculating the posterior density of observing the proposed θ'_i and compares it to the current density given θ_i . If the posterior density is higher for the proposed value, it will almost certainly be accepted as the new value in the Markov chain. If the density is lower, then the new value is accepted dependent on how much lower the posterior density is compared to the current value. In the case of a symmetric proposal distribution ($g_{\theta_i}(\theta'_i|\theta_i) = g_{\theta_i}(\theta_i|\theta'_i)$), Equation 1.18 simplifies to comparing the ratio between posteriors when proposing a new value.

Though the Metropolis-Hastings algorithm described lends flexibility to a user when building

a model, it also comes with its share of pitfalls. For example, candidate values obtained via a proposal distribution may not always be accepted. In this case, the current parameter value is stored leading to a repeat in the Markov chain. This consistent repetition leads to high auto-correlation within the Markov chain and ultimately, samples from the posterior of interest cannot be assumed as independent. This can be resolved by running the MCMC algorithm for longer, but large auto-correlation can lead to an inefficient algorithm. In contrast, the Gibbs sampler bypasses this problem since new candidate values are immediately accepted when drawn from each parameter's conditional posterior distribution.

Another common challenge of using a Metropolis-Hastings algorithm is the necessity of specifying a proposal distribution. A poorly chosen proposal can lead to a poorly explored posterior distribution (often referred to as poor *mixing*). Furthermore, once a proposal has been chosen, one must also consider the scale of the distribution. Take a Gaussian distribution for example. If a new value θ'_i is chosen by drawing from a Gaussian centred on the current θ_i , the variance of said distribution will determine how near or far proposed values fall from the centre. Too larger a step size and the algorithm may consistently move past peaks of posterior density, too small and the algorithm may never explore the posterior surface fully, getting stuck at local maxima. A simple solution to this problem is to allow the proposal distribution's variance to adapt over time thereby optimally adapting the proposal distribution to efficiently explore the posterior (Haario et al. 1999, Garthwaite et al. 2016).

Beyond the Metropolis-Hastings and Gibbs sampling algorithms there are a collection of gradient based MCMC methods for posterior sampling. For example, the Metropolis-adjusted Langevin algorithm (Roberts & Tweedie 1996) proposes a candidate value via the combination of the distribution centred on the current state of the Markov chain and its gradient on the logged target density. Another alternative is the Hamiltonian Monte Carlo method (Neal 2011). This process aims to sample from the density of interest by considering the dynamics of a frictionless body moving over a surface (the negative logged target density). A candidate for the new state of the Markov chain is produced in two steps. Firstly, propose a new momentum value for the frictionless body, and allow it to move in response to this momentum under the laws of Hamiltonian dynamics. After some time interval, the body is frozen and its current location and momentum are accepted or rejected as the new states of the Markov chain via a Metropolis step. Dunson & Johndrow (2019) offer an excellent review summarising the algorithms described in this section to celebrate the developments made since the original Hastings (1970) paper, fifty years on.

The primary algorithm that will be adopted in this thesis is the Metropolis-Hastings algorithm for its ease of implementation, focussing on the flexibility it brings to geographic profiling by allowing a user to estimate source locations under a custom prior. At each step, I will justify the choice of proposal distribution and associated measures put into place to ensure healthy MCMC mixing.

1.7 Platforms for building a geographic profiling model

Each of the major models described in this chapter have multiple platforms, both open and commercial software, to solve the objectives of geographic profiling. Many of these platforms come with user interfaces for clean implementation (Rossmo 2000, Levine & Block 2011, Canter et al. 2013). Others require knowledge of programming languages such as R (Verity et al. 2014, R Core Team 2019) and Python (Van Rossum & Drake Jr 1995, Santosuosso & Papini 2018).

An extensive list of available software can be found in Table 1.1 below. Of the existing software, the majority implement Rossmo’s CGT algorithm with various options allowing a user to choose different kernels compared to that in Figure 1.1. Some also make use of O’Leary’s single-source Bayesian model (Levine & Block 2011, Canter et al. 2013).

Name	Type	Model	Cost?	Reference
Rigel	User interface	CGT	Commercial/paid	Rossmo (2000)
CrimeStat	User interface	CGT/Bayesian	Free	Levine & Block (2011)
Dragnet	User interface	CGT/Bayesian	Free	Canter et al. (2013)
Rgeoprofile	R package	DPM	Open software	Verity et al. (2014)
NA	Python script	CGT	Open software	Santosuosso & Papini (2018)
GeoCrime	User interface	CGT	Free	Butkovic et al. (2018)
rgeoprofile	R package	CGT	Open software	Spaulding & Morris (2020)
Silverblaze	R package	Finite mixture	Open software	Stevens et al. (2021)

Table 1.1: A list of available software, in addition to their cost and type, for implementing different geographic profiling models.

The RgeoProfile software from Verity et al. (2014) is the only platform to implement a Dirichlet process mixture model for geographic profiling, despite evidence that it finds source locations more efficiently than the CGT or simple Bayesian model. Paulsen (2006b) details a comparison between geographic profiling software, covering Rigel, Crimestat and Dragnet. Like many reviews in criminology (Paulsen 2006a, Harries & LeBeau 2007), Paulsen (2006b) question the validity of

geographic profiling since measures of model accuracy consist of simple summary statistics such as the hit score.

1.8 Conclusions and thesis objectives

In this chapter, I have covered the pivotal developments of geographic profiling, from its origin as a kernel density estimator to a fully fledged non-parametric Bayesian model. Despite many fundamental changes over the past 30 years, the main objective of geographic profiling has remained the same. Given a set of locations, attributed to crimes, sightings of an invasive species or positive tests of an infectious disease, estimate the spatial source locations(s) responsible for generating these data.

Each model has and continues to demonstrate a wide range of applications across multiple disciplines. Of these models, the Dirichlet process mixture model has the most to offer by providing a solution to the problem of estimating multiple source locations even when prior knowledge on their numbers is unknown. Inferences made by the DPM model rely on a specific configuration where data and prior must follow a set form and thus deny a user the possibility of analysing other data types. Although the DPM model has proven applicable to problems across multiple disciplines, the developments it made to the field of geographic profiling are one of many steps that could have been taken.

In this PhD, I will produce a toolbox that opens up user flexibility to implement a variety of geographic profiling models that estimate desired parameters based on new types of data and varying prior beliefs. The remaining contents of the thesis will be set out as follows:

- **Chapter 2:** I build a Poisson finite mixture model that utilises a Metropolis-Hastings within Gibbs sampler to estimate parameters via spatial count data; a common data type in ecology where locations and associated counts of a particular species are recorded. I specify the Poisson finite mixture model such that the spatial prior on source locations is discretized, compared to the continuous domain of the existing model. This flexibility allows a user to choose a spatial prior of their choice.
- **Chapter 3:** I apply the Poisson finite mixture model built in chapter two, testing its ability to estimate parameters when making inference via simulated count data. I will also develop the Poisson finite mixture model, equipping it to deal with over-dispersed count data. I then test this development by inferring breeding sites of mosquitoes given a set of over-dispersed surveillance data in Miami-Dade county, Florida.

- **Chapter 4:** I build a Binomial finite mixture model that estimates desired parameters based on a set of spatial prevalence data; a data type common to epidemiology where the location, number of individuals tested and those testing positive for a disease are recorded. This model is then applied to malarial survey data in Kasese, Uganda.
- **Chapter 5:** I return to Gaussian mixture models and the data set analysed in Le Comber et al. (2011) and Verity et al. (2014), comparing the results inferred from mixture components made up of bivariate normal, Laplace or Cauchy distributions.
- **Chapter 6:** I conclude and reflect upon the implications of the work in this thesis on the field of geographic profiling. I will also describe the avenues future developments may take in building upon this work. Finally, I take a step back, to specify where geographic profiling fits into a universe of models with similar objectives across the three cornerstone disciplines.

Each new model in this thesis will be built in the coding language *R*, and extensive documentation including tutorials for model implementation will be available at <https://github.com/Michael-Stevens-27/silverblaze>.

Chapter 2

Building a Poisson Geographic Profiling Model

Abstract:

Historically, geographic profiling studies have consisted of analysing a specific kind of spatial data. A key challenge with this kind of data is that locations associated with no observations may or may not have been sampled, hence it is difficult to distinguish between evidence of absence and an absence of evidence. This problem can be addressed by considering counts at known locations; that is, a known set of locations are surveyed and observations are counted, with the possibility of counts being zero. In this chapter, I build a geographic profiling model that solves this issue by deriving the probability of observing a set of locations with associated counts. Utilising this scheme, a user can formally specify various prior distributions describing the locations of the sources of the observations. The model also estimates a variable describing the total number of individuals present in a landscape. Finally, the model built in this chapter allows for sources to generate observations at independent scales. The new model broadens the flexibility available to a user and widens the pool of applications in an already multi-disciplinary field.

The contents of this chapter are published in Stevens et al. (2021).

2.1 Introduction

In the introduction of this thesis, advantages and disadvantages of the three main geographic profiling models were described. The criminal geographic targeting algorithm (Rossmo 2000), the simple Bayesian model (O’Leary 2009, 2010a) and the more advanced Dirichlet process mixture (DPM) model (Verity et al. 2014) have all successfully been used to infer parameters across various disciplines. These models however, were designed to analyse point-pattern data, a finite collection of longitudinal/latitudinal points each associated with a single event, when making their inferences. This data consisted of, for example, a unique set of crime locations (Rossmo et al. 2014) or recorded sightings of invasive species (Stevenson et al. 2012). By contrast, these algorithms were not designed to analyse count data, a finite set of locations with an associated count corresponding to the number of individuals observed at each location.

On one occasion, the DPM model was applied to data of this kind. Faulkner et al. (2016) used the DPM model to search for nesting sites of invasive American mink (*Neovison vison*) using expensive trapping data vs cheap citizen science sighting data. The trap data considered in this study were forced into the form of point-pattern data by considering only those locations of successful traps. Any locations yielding no captures were discarded.

Longitude	Latitude	Longitude	Latitude	Counts
-0.0404	51.5239	-0.0404	51.5239	2
-0.0404	51.5239	-0.1335	51.5245	0
-0.0892	51.5238	-0.0892	51.5238	3
-0.0892	51.5238	-0.0757	51.5050	0
-0.0892	51.5238	-0.0512	51.5070	0
...

Table 2.1: Examples of spatial point-pattern and count data common to criminology, ecology, epidemiology.

By considering count data, an important distinction can be made between evidence of absence and absence of evidence. In an ecological context, this might relate to areas where traps were set but failed to catch any animals and areas where no traps were set; in criminology, between areas where crimes could have been committed but were not and areas where no information was recorded (such as outside a jurisdictional boundary); and in epidemiology, between areas where people were tested and found negative, and areas where no-one was tested.

In ecology, there are existing models that use such count data to infer parameters of biological interest. For example, spatially explicit capture recapture models aim to estimate the underlying

population density in a study area given the locations of a set of traps with associated counts (Borchers & Efford 2008, Chandler & Royle 2013). These models even go so far as to estimate an individual’s *activity centre*, a latent variable synonymous to *source location* or *anchor point* used throughout geographic profiling literature. These models, however, assume each individual from a species is associated with its own unique activity centre of which are estimated from the data. The DPM model does not make this assumption and is built to deal with the complex problem of partitioning individuals into spatial clusters of which each cluster is governed by a single source location (Verity et al. 2014).

Discipline	Terms						Examples
Criminology	GP	Criminal	Crime	No crime	Anchor point	Potential crime site	Rossmo et al. (2014)
Epidemiology	GP	Disease host	Positive result	Negative test result	Source of outbreak	Patient postcode	Verity et al. (2014)
Ecology	GP	Invasive species	Capture	Empty trap	Nesting location	Trap	Faulkner et al. (2016)
	SDM	Animal	Observed individual	No observation	Activity centre	Single level, multilevel, proximity	Chandler & Royle (2013)
Joint Terms		Event	Encounter	No encounter	Source location	Sentinel site	

Table 2.2: Terminology used in geographic profiling (GP) and species distribution models (SDM) alongside joint terms used in this thesis

This chapter addresses the gap in existing models by developing a fully Bayesian geographic profiling model for analysing spatial count data. This model consists of calculating the likelihood of a particular number of crimes (or captures, or positive tests) at a given location where crucially, this number can be zero. In addition to including count data in the model’s likelihood, this chapter will lead to, for the first time, an estimation of the expected population size over a specified search area and time period. This parameter is of consistent interest across disciplines, from criminology - in estimating the number of prostitutes or migrating fugitives (Rossmo & Routledge 1990), to ecology - in estimating the population size of many avian species (Royle 2004).

Once the architecture of the model has been described, the reader will be taken through

the MCMC algorithm that fits the parameters of interest step by step. The validity of the model's assumptions will then be explored, focussing specifically on an approximation to an integral utilised in the model's likelihood. Additionally, an alternative model in which the unknown population size is treated as a nuisance parameter will be described.

2.2 Model derivation

The Poisson finite mixture model begins by assuming K sources, with locations $\boldsymbol{\mu}_k = (\mu_x, \mu_y)$ for $k \in 1 : K$ drawn from some suitable prior distribution, \mathcal{F} . Here the model followed O'Leary (2010a) in assuming that \mathcal{F} is defined over a two-dimensional grid of cells, allowing the prior probability mass to be defined separately for each cell (for example, zero probability over water bodies is often desired). Next, the number of events - both encountered and un-encountered - in the study area is said to be governed by some expectation, λt , where λ is the expected number of events over the search area per unit time and t is the time interval with which data were collected. From this expectation, the total number of events, N , in the study area is drawn from a Poisson distribution with the associated mean. Explicitly, an event is the existence of an invasive species, a host of a disease or a criminal in the search area.

Each event originates from a single source with equal probability $\frac{1}{K}$, and the source from which event i originates can be written as $c_i \in 1 : K$. So event i is denoted to originate from source c_i , but may also be referred to as originating from source k .

The spatial location of event i , denoted \mathbf{x}_i , is drawn from a dispersal distribution centred on its source. Here this distribution is assumed to be a bivariate normal distribution with mean $\boldsymbol{\mu}_{c_i}$, variance $\sigma_{c_i}^2$ and zero correlation between dimensions. Choosing a bivariate normal keeps this model consistent with previous geographic profiling studies that recognise the probability of encountering an event is defined over two-dimensional space as opposed to spatial capture recapture models that consider a univariate half-normal distribution between source and capture location (Efford 2004).

Unlike the DPM model, not every event is assumed to have been encountered. Instead the model assumes that there are m sentinel sites, denoted \mathbf{s}_j for $j \in 1 : m$, within the study area and that events are only encountered if they fall within a distance ρ from one of these sites. A sentinel site could take on many forms as shown in Table 2.2. Biologically, these sites could refer

Parameters	Description
K	The true number of source locations
$\boldsymbol{\mu}_{c_i}$	The spatial location of source c_i in $1 : K$
\mathcal{F}	The prior on source locations
N	The number of events in a search area
λt	The rate of events in a search area in time t
(ζ, η)	The shape and rate of the gamma prior on λ
$c_i = k$	The source that event i originates from
\mathbf{x}_i	The spatial location of event i
σ_{c_i} (km)	The standard deviation of the dispersal distribution centred on $\boldsymbol{\mu}_{c_i}$
(γ, δ)	The mean and variance of the log normal distribution on σ_{c_i}
m	The number of sentinel sites
ρ (km)	The sentinel site radius
\mathbf{s}_j	The spatial location of sentinel site j
n_j	The number of events encountered at site j
θ_j	The height of sentinel site j on the mixture of bivariate normals
$B_\rho(\mathbf{s}_j)$	The ball of radius ρ centred on sentinel site \mathbf{s}_j
$f_{BN}(\cdot)$	The density on the bivariate normal f
$g_\mu(\cdot), g_\sigma(\cdot)$	Proposal distributions for $\boldsymbol{\mu}$ and σ
$\epsilon_\mu, \epsilon_\sigma$	Standard deviations of the proposal distributions on $\boldsymbol{\mu}$ and σ

Table 2.3: A list of parameters and their meanings adopted throughout this chapter.

to multiple methods of observing an organism, such as camera traps, hair snares or bio-acoustic sensors (Royle et al. 2018). In this study, sentinel sites can encounter any non-negative integer of events, akin to multi-catch traps in ecology (Borchers 2012), leading to a set of count data. The model is made fully Bayesian by placing suitable priors on the remaining unknown quantities of interest. The complete model can be written:

Likelihood:

$$c_i \sim \text{Categorical}\left(\frac{1}{K}\right), \quad i \in 1 : N, \quad (2.1)$$

$$\mathbf{x}_i \sim \text{Normal}(\boldsymbol{\mu}_{c_i}, \mathbf{I}_2 \sigma_{c_i}^2), \quad i \in 1 : N, \quad (2.2)$$

$$n_j = \#\{\mathbf{x}_i : d_E(\mathbf{x}_i, \mathbf{s}_j) < \rho\}, \quad j \in 1 : m. \quad (2.3)$$

Priors:

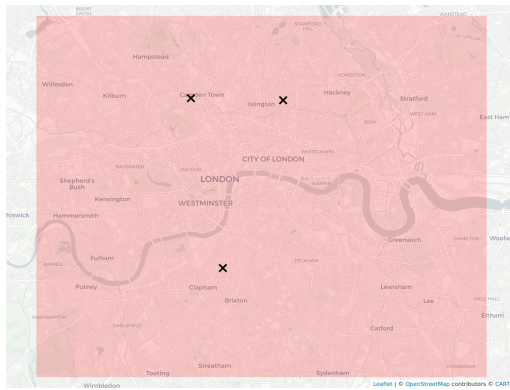
$$\boldsymbol{\mu}_k \sim \mathcal{F}, \quad k \in 1 : K, \quad (2.4)$$

$$\sigma_k \sim \text{Log-Normal}(\gamma, \delta), \quad k \in 1 : K, \quad (2.5)$$

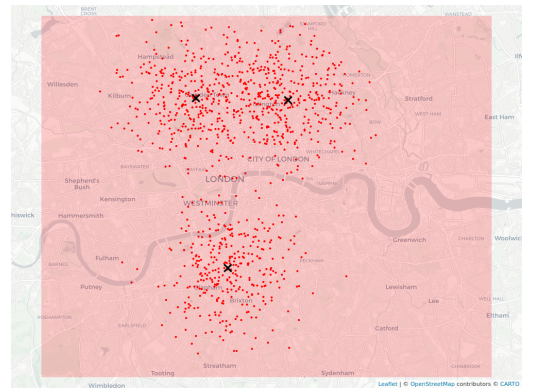
$$N \sim \text{Poisson}(\lambda t), \quad (2.6)$$

$$\lambda \sim \text{Gamma}(\zeta, \eta), \quad (2.7)$$

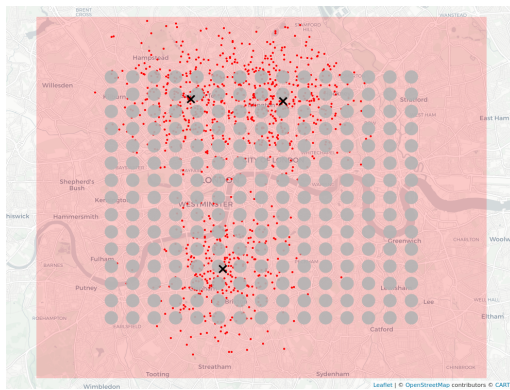
where \mathbf{I}_2 is the two-dimensional identity matrix, and $d_E(\mathbf{x}_i, \mathbf{s}_j)$ is the Euclidean distance between event location \mathbf{x}_i and sentinel site \mathbf{s}_j . When performing its inference, the model only has access to the final counts n_j at each of the m sentinel sites, and not the raw data \mathbf{x}_i for $i \in 1 : N$. The procedure of generating this count data is illustrated in Figure 2.1.



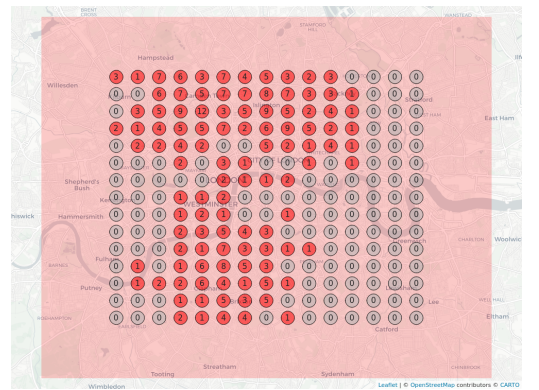
(a) Source locations drawn from \mathcal{F}



(b) Event locations (\mathbf{x}_i)



(c) A uniform array of sentinel sites



(d) Count data observed

Figure 2.1: The process of simulating a set of count data under the Poisson finite mixture model.

Now, the probability of the observed n_j (the likelihood) given the parameters $\{\boldsymbol{\mu}_{c_i}\}, \{\sigma_{c_i}\}$ and λ needs to be calculated. The probability that an event is observed is equal to the probability that

it falls within a distance ρ of a sentinel site, which can be obtained by integrating the dispersal distribution over the ball $B_\rho(\mathbf{s}_j)$ of radius ρ centred on \mathbf{s}_j . In general, this integral will not have a simple analytical solution, but under certain conditions can be approximated by a cylinder centred on \mathbf{s}_j with radius ρ and height equal to the dispersal distribution at the central point:

$$\Pr(d_E(\mathbf{x}_i, \mathbf{s}_j) < \rho | \boldsymbol{\mu}, \sigma, c_i) = \int_{B_\rho(\mathbf{s}_j)} f_{BN}(x, y | \boldsymbol{\mu}_{c_i}, \mathbf{I}_2 \sigma_{c_i}^2) dx dy, \quad (2.8)$$

$$\approx \pi \rho^2 f_{BN}(\mathbf{s}_j | \boldsymbol{\mu}_{c_i}, \mathbf{I}_2 \sigma_{c_i}^2), \quad (2.9)$$

where $f_{BN}(\mathbf{s}_j | \boldsymbol{\mu}_{c_i}, \mathbf{I}_2 \sigma_{c_i}^2)$ is the density of the bivariate normal distribution at sentinel site j with mean $\boldsymbol{\mu}_{c_i}$ and covariance matrix $\mathbf{I}_2 \sigma_{c_i}^2$. The validity of this approximation is explored in detail in section 2.4. The total probability of being detected by sentinel site j can be obtained by averaging Equation 2.9 over all sources, leading to the following expression defined as θ_j for convenience:

$$\theta_j \equiv \Pr(d_E(\mathbf{x}_i, \mathbf{s}_j) < \rho | \boldsymbol{\mu}, \sigma, c_i) = \frac{\pi \rho^2}{K} \sum_{c_i=1}^K f_{BN}(\mathbf{s}_j | \boldsymbol{\mu}_{c_i}, \mathbf{I}_2 \sigma_{c_i}^2). \quad (2.10)$$

Given that a Poisson distribution with rate λt is applied to the total number of events N and given that every event has the same independent probability of being detected given by Equation 2.10, it follows that the probability of detecting n_j events at sentinel site j is Poisson distributed with rate $\lambda t \theta_j$. The likelihood is obtained by multiplying this Poisson probability over all sentinel sites:

$$\Pr(\mathbf{n} | \lambda t, \boldsymbol{\theta}) = \prod_{j=1}^m \frac{(\lambda t \theta_j)^{n_j} e^{-\lambda t \theta_j}}{n_j!}. \quad (2.11)$$

Here, the unit of time t is assumed to be the interval in which the data was collected and thus it is set to one. The likelihood in Equation 2.11, combined with the priors in Equation 2.5 - Equation 2.7, is used to estimate the unknown parameters $\{\boldsymbol{\mu}_{c_i}\}$, $\{\sigma_{c_i}\}$ (for $c_i \in 1 : K$) and λ . These parameters are fitted via MCMC methods using a combination of Metropolis-Hastings and Gibbs sampling, details of which will now be described in full.

2.3 An MCMC algorithm for the new model

The Poisson finite mixture model, as described in section 2.2, is expected to estimate the set of parameters: $\{\boldsymbol{\mu}_{c_i}\}$ for $c_i \in 1 : K$, $\{\sigma_{c_i}\}$ for either $c_i \in 1 : K$ or c_i equal to one and finally the expected event size λ . This section describes the process of estimating these desired parameters through either a Metropolis-Hastings ($\boldsymbol{\mu}_{c_i}$ and σ_{c_i}) or Gibbs sampling step (λ).

2.3.1 Metropolis-Hastings steps

Each $\boldsymbol{\mu}_{c_i}$ for $c_i \in 1 : K$ is updated via a separate Metropolis-Hastings step. First, a proposed value $\boldsymbol{\mu}'_{c_i}$ is drawn from a bivariate normal proposal distribution centred on $\boldsymbol{\mu}_{c_i}$, denoted $g_{\mu}(\boldsymbol{\mu}'_{c_i} | \boldsymbol{\mu}_{c_i}, \epsilon_{\mu})$, where ϵ_{μ} is the proposal standard deviation. New values of θ_j , denoted θ'_j , are then calculated using Equation 2.10 for $j \in 1 : m$, and the likelihood is recalculated from these θ'_j values using Equation 2.11. The likelihood is then combined with the prior on source locations, \mathcal{F} , to produce an acceptance probability through the standard Metropolis-Hastings expression (Hastings 1970):

$$Acc(\boldsymbol{\mu}'_{c_i}, \boldsymbol{\mu}_{c_i}) = \min \left(1, \frac{L(\boldsymbol{\mu}'_{c_i}) \mathcal{F}(\boldsymbol{\mu}'_{c_i}) g_{\mu}(\boldsymbol{\mu}_{c_i} | \boldsymbol{\mu}'_{c_i}, \epsilon_{\mu})}{L(\boldsymbol{\mu}_{c_i}) \mathcal{F}(\boldsymbol{\mu}_{c_i}) g_{\mu}(\boldsymbol{\mu}'_{c_i} | \boldsymbol{\mu}_{c_i}, \epsilon_{\mu})} \right), \quad (2.12)$$

where $L(\cdot)$ is shorthand for the likelihood in Equation 2.10. The proposal standard deviation ϵ_{μ} adapts itself at every iteration of the MCMC burn in via the Robbins-Monro process (Garthwaite et al. 2016) to hit a target acceptance rate of 23% for a multivariate proposal distribution. Essentially if a proposed move is rejected, ϵ_{μ} will shrink to ensure the next proposed step is further from areas where moves are being rejected. Similarly, if a proposed move is accepted, ϵ_{μ} will grow, to ensure the MCMC reaches areas of interest quickly. An analogous method is used to update each σ_{c_i} ; this time proposing from a univariate normal distribution reflected about zero, and using a log-normal(γ, δ) prior. Again, the proposal standard deviation is chosen automatically via the Robbins-Monro method, but this time to reach a target acceptance rate of 44% in the case of a univariate proposal distribution (Garthwaite et al. 2016).

2.3.2 Gibbs sampling step

The expected number of events within the total search area, λ , is updated at each iteration of the MCMC algorithm via a Gibbs sampling step. Deriving a Gibbs sampling step starts with re-writing Equation 2.11 in terms of λ by dropping any terms that do not pertain directly to λ and setting t equal to one.

$$\Pr(\mathbf{n} | \lambda, \boldsymbol{\theta}) \propto \prod_{j=1}^m \lambda^{n_j} e^{-\lambda \theta_j}, \quad (2.13)$$

$$= (\lambda)^{\dot{n}} e^{-\lambda \dot{\theta}} \quad (2.14)$$

where $\dot{n} = \sum_{j=1}^m n_j$ and $\dot{\theta} = \sum_{j=1}^m \theta_j$. Notice this expression is conjugate to the Gamma density, and hence multiplying Equation 2.14 by a Gamma(ζ, η) prior on λ and normalising, the following

conditional posterior distribution is obtained:

$$\Pr(\lambda|\mathbf{n}, \boldsymbol{\theta}) = \frac{(\eta + \hat{\theta})^{\eta + \hat{n}}}{\Gamma(\zeta + \hat{n})} (\lambda)^{\zeta + \hat{n} - 1} e^{-\lambda(\eta + \hat{\theta})} \quad (2.15)$$

which is equal to a $\text{Gamma}(\zeta + \hat{n}, \eta + \hat{\theta})$ distribution. Combining this Gibbs sampling step with the Metropolis-Hastings step that was just defined results in the MCMC algorithm shown in algorithm 1.

Data: A set of longitudinal and latitudinal sentinel sites with associated counts

Result: The set $\{\boldsymbol{\mu}_{c_i}\}, \{\sigma_{c_i}\}$ and λ , for $c_i \in 1 : K$

Initialise: draw $\{\boldsymbol{\mu}_{c_i}\}, \{\sigma_{c_i}\}, \lambda$ from priors, set *burn in, sampling, con check* iterations;

for *iteration* in $1 : (\text{burn in} + \text{sampling})$ **do**

for c_i in $1 : K$ **do**

Propose new $\boldsymbol{\mu}'_{c_i}$ from $g_{\mu}(\cdot)$;

Calculate M-H ratio, R_{μ} , via Equation 2.12, generate random $u \in [0, 1]$;

if $u < R_{\mu}$ **then**

Accept $\boldsymbol{\mu}'_{c_i}$ as the new $\boldsymbol{\mu}_{c_i}$;

if *iteration* < *burn in* **then**

 Increase prop s.d. ϵ_{μ} for $\boldsymbol{\mu}_{c_i}$ via Robbins-Monro;

else

Reject $\boldsymbol{\mu}'_{c_i}$ and retain $\boldsymbol{\mu}_{c_i}$;

if *iteration* < *burn in* **then**

 Decrease prop s.d. ϵ_{μ} for $\boldsymbol{\mu}_{c_i}$ via Robbins-Monro;

for c_i in $1 : K$ **do**

Propose new σ'_{c_i} from $g_{\sigma}(\cdot)$;

Calculate M-H ratio, R_{σ} , via Equation 2.12, generate random $u \in [0, 1]$;

if $u < R_{\sigma}$ **then**

Accept σ'_{c_i} as the new σ_{c_i} ;

if *iteration* < *burn in* **then**

 Increase prop s.d. ϵ_{σ} for σ_{c_i} via Robbins-Monro;

else

Reject σ'_{c_i} and retain σ_{c_i} ;

if *iteration* < *burn in* **then**

 Decrease prop s.d. ϵ_{σ} for σ_{c_i} via Robbins-Monro;

Update θ given new $\{\boldsymbol{\mu}_{c_i}\}, \{\sigma_{c_i}\}$, draw new λ via Equation 2.15;

Store $\{\boldsymbol{\mu}_{c_i}\}, \{\sigma_{c_i}\}$ and λ for this iteration ;

if *iteration* < *burn in* AND *iteration* $\equiv 0 \pmod{(\text{con check})}$ **then**

 Check for convergence via Geweke's metric (Cowles & Carlin 1996);

if convergence is met **then**

 set *iteration* = *burn in* + 1;

Algorithm 1: The MCMC algorithm associated with the Poisson model.

The Poisson finite mixture model and the procedure for fitting the set of unknown parameters via MCMC methods has now been described in detail. Before implementing and testing the model's ability to fit these parameters, I will explore the model in more detail. A list of all the parameters described in this chapter can be found in Table 2.3 .

2.4 Volume under surface approximation

Computing the Poisson finite mixture model's likelihood in Equation 2.11 requires the evaluation of a complicated integral; the bivariate normal distribution integrated over the ball of radius ρ centred on sentinel site \mathbf{s}_j . This calculation cannot be obtained analytically and so instead I make use of the approximation described in Equation 2.10, where the bivariate normal density at the sentinel site multiplied by $\pi\rho^2$ is used. Here, I will describe the origin of this approximation and go on to evaluate its validity by deriving an upper bound for the error associated with it.

2.4.1 Deriving an approximation

To derive the approximation stated in Equation 2.10, the integral

$$\iint_{B_\rho(\mathbf{s}_j)} f_{BN}(x, y | \boldsymbol{\mu}_{c_i}, \mathbf{I}_2 \sigma_{c_i}^2) dx dy$$

is shifted into polar co-ordinates. For notational simplicity the mean of the bivariate normal $\boldsymbol{\mu}_{c_i}$ is written as (μ_x, μ_y) and the variance of the bivariate normal $\sigma_{c_i}^2$ is written as σ^2 . Shifting the integral in Equation 2.8 to polar co-ordinates ($x = r \cos \Theta$ and $y = r \sin \Theta$) yields:

$$\int_0^{2\pi} \int_0^\rho \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}((r \cos \Theta - \mu_x)^2 + (r \sin \Theta - \mu_y)^2)} r dr d\Theta, \quad (2.16)$$

and expanding on the exponential term leads to

$$\frac{1}{2\pi\sigma^2} \int_0^{2\pi} \int_0^\rho r e^{-\frac{1}{2\sigma^2}(r^2 + (\mu_x^2 + \mu_y^2) - 2r(\cos \Theta \mu_x + \sin \Theta \mu_y))} dr d\Theta. \quad (2.17)$$

Taking the Taylor series expansion about the location of the sentinel site (for simplicity it is set to the origin ($x = 0, y = 0$)) leads to the integral in the form:

$$\frac{1}{2\pi\sigma^2} e^{-\frac{(\mu_x^2 + \mu_y^2)}{2\sigma^2}} \int_0^{2\pi} \int_0^\rho r \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{-r^2}{2\sigma^2}\right)^i \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{r \cos \Theta \mu_x}{\sigma^2}\right)^j \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{r \sin \Theta \mu_y}{\sigma^2}\right)^k dr d\Theta, \quad (2.18)$$

As an approximation to this expression, i , j and k are set to zero, greatly simplifying by reducing each summation to one:

$$\frac{1}{2\pi\sigma^2} e^{-\frac{(\mu_x^2 + \mu_y^2)}{2\sigma^2}} \int_0^{2\pi} \int_0^\rho r dr d\Theta = \frac{1}{2\pi\sigma^2} e^{-\frac{(\mu_x^2 + \mu_y^2)}{2\sigma^2}} \int_0^{2\pi} \frac{\rho^2}{2} d\Theta \quad (2.19)$$

$$\frac{1}{2\pi\sigma^2} e^{-\frac{(\mu_x^2 + \mu_y^2)}{2\sigma^2}} \left[\Theta \frac{\rho^2}{2} \right]_0^{2\pi} = \frac{1}{2\pi\sigma^2} e^{-\frac{(\mu_x^2 + \mu_y^2)}{2\sigma^2}} \cdot \pi\rho^2 \quad (2.20)$$

$$= f_{BN}(0, 0) \cdot \pi\rho^2. \quad (2.21)$$

This result shows that a suitable approximation to the volume under the bivariate normal bound by a circle of radius ρ centred at the origin is simply the density on the bivariate normal at the origin multiplied by the area of the circle of radius ρ . This result can be generalised to the case of any sentinel site location by translating the co-ordinate system such that the sentinel site is now on the origin. A sentinel site's density located at (s_x, s_y) on a bivariate normal with mean (μ_x, μ_y) is equivalent to a sentinel site located at the origin on a bivariate normal with mean $(\mu_x - s_x, \mu_y - s_y)$, hence the approximation in Equation 2.10 holds for any site location. The work of Gilliland (1962) should be explored for a more accurate approximation, as opposed to simply reducing the summations in Equation 2.18 to one. When deriving their approximation, Gilliland also places less restrictions on the behaviour of the bivariate normal, for example that the standard deviations in each dimension are not necessarily equal.

2.4.2 Deriving an upper bound on the error

How trustworthy is this alternative to the integral in Equation 2.10? To describe the error in this approximation consider first the gradient of f_{BN} (in polar co-ordinates and centred at the origin)

$$\nabla f_{BN} = \mathbf{e}_r \frac{\partial}{\partial r} f_{BN} = -\frac{r}{\sigma^2} f_{BN} \mathbf{e}_r, \quad (2.22)$$

where \mathbf{e}_r is the unit vector in the radial direction. The maximum of this gradient is then calculated by setting $\nabla|\nabla f_{BN}|$ to zero and solving for r . In this case $r = \sigma$. Hence the maximum of ∇f_{BN} is

$$\max|\nabla f_{BN}| = |\nabla f_{BN}(r = \sigma)| = \frac{1}{2\pi\sqrt{e}\sigma^3}. \quad (2.23)$$

Note that for any $\mathbf{y} = (x, y)$ or \mathbf{s}_j that the gradient

$$\frac{|f_{BN}(\mathbf{y}) - f_{BN}(\mathbf{s}_j)|}{|\mathbf{y} - \mathbf{s}_j|} \leq \frac{1}{2\pi\sqrt{e}\sigma^3}. \quad (2.24)$$

Hence the error in the integral

$$\left| \iint_{B_\rho(\mathbf{s}_j)} f_{BN}(\mathbf{y}) dxdy - \pi\rho^2 f_{BN}(\mathbf{s}_j) \right| = \left| \iint_{B_\rho(\mathbf{s}_j)} f_{BN}(\mathbf{y}) - f_{BN}(\mathbf{s}_j) dxdy \right| \quad (2.25)$$

$$\leq \frac{1}{2\pi\sqrt{e}\sigma^3} \left| \iint_{B_\rho(\mathbf{s}_j)} |\mathbf{y} - \mathbf{s}_j| dxdy \right| = \frac{1}{2\pi\sqrt{e}\sigma^3} \left| \int_0^{2\pi} \int_0^\rho r \cdot r dr d\theta \right| \quad (2.26)$$

$$= \frac{1}{3\sqrt{e}} \left(\frac{\rho}{\sigma} \right)^3. \quad (2.27)$$

So, the maximum error in the approximation is governed by the ratio of site radius ρ and dispersal parameter σ . If ρ is fixed, then as σ increases the shape of the bivariate normal becomes flatter with respect to ρ and the approximation moves closer to the true value. Conversely if σ remains fixed and ρ increases, then the approximation will be invalid. This scenario is not expected within ecological surveys though, as little information is to be gained from a sentinel site whose radius is much larger than the dispersal of a particular species (Sun et al. 2014).

2.5 Integrating out the expected event size

The Poisson finite mixture model demonstrates the first time a geographic profiling model has estimated the expected event size λ in the search area. Although there is value to estimating this parameter, the original objective of geographic profiling is to provide a method of searching a landscape for one or more source locations (Rossmo 2000). It is possible then, to consider λ as a nuisance parameter that can be integrated out of the likelihood. Recall the likelihood of observing a set of count data (sentinel site locations and associated counts) is:

$$\Pr(\mathbf{n}|\lambda t, \boldsymbol{\theta}) = \prod_{j=1}^m \frac{(\lambda t \theta_j)^{n_j} e^{-\lambda t \theta_j}}{n_j!}. \quad (2.28)$$

Again, t is assumed to be the interval of time with which the data were collected and thus is set to one. λ is to be integrated out of the likelihood in the hope of improving MCMC mixing by reducing the number of parameters the model must estimate. Before λ is integrated out of the model, the product in Equation 2.28 is expanded and simplified on:

$$\prod_{j=1}^m \frac{(\lambda \theta_j)^{n_j} e^{-\lambda \theta_j}}{n_j!} = \frac{(e^{-\lambda \sum_j \theta_j}) \prod_j^m (\lambda \theta_j)^{n_j}}{\prod_j^m n_j!} \quad (2.29)$$

$$= \lambda^{\dot{n}} e^{-\lambda \dot{\theta}} \cdot \frac{\prod_j^m \theta_j^{n_j}}{\prod_j^m n_j!} \quad (2.30)$$

$$= \lambda^{\dot{n}} e^{-\lambda \dot{\theta}} \cdot c, \quad (2.31)$$

where $\dot{n} = \sum_j^m n_j$, $\dot{\theta} = \sum_j^m \theta_j$ and $c = \prod_j^m \frac{\theta_j^{n_j}}{n_j!}$. The prior on λ is given by a gamma distribution with shape and rate α and β respectively. Multiplying the likelihood with said prior yields:

$$(\lambda^{\dot{n}} e^{-\lambda \dot{\theta}} \cdot c) \cdot \left(\frac{1}{\beta^\alpha \Gamma(\alpha)} \lambda^{\alpha-1} e^{-\frac{\lambda}{\beta}} \right) \quad (2.32)$$

Now in order to derive the posterior in terms of $\boldsymbol{\mu}$ and σ the following integral must be evaluated:

$$\frac{c}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty \lambda^{\dot{n}+\alpha-1} e^{-\lambda(\dot{\theta}+\frac{1}{\beta})} d\lambda. \quad (2.33)$$

This expression looks somewhat like the Gamma function, so recall that:

$$\Gamma(a+1) = \int_0^\infty x^a e^{-x} dx. \quad (2.34)$$

Substituting in $x = by$, for some constant $b > 0$, leads to

$$\Gamma(a+1) = \int_0^\infty (by)^a e^{-by} b dy = b \cdot b^a \int_0^\infty y^a e^{-by} dy, \quad (2.35)$$

which produces the identity:

$$\int_0^\infty y^a e^{-by} dy = \frac{\Gamma(a+1)}{b^{a+1}}. \quad (2.36)$$

Going back to the integral in Equation 2.33, replacing y , a and b with $y = \lambda$, $a = \dot{n} + \alpha - 1$ and $b = \dot{\theta} + \frac{1}{\beta}$, yields

$$\frac{c}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty \lambda^{\dot{n}+\alpha-1} e^{-\lambda(\dot{\theta}+\frac{1}{\beta})} d\lambda = \frac{c}{\beta^\alpha \Gamma(\alpha)} \frac{\Gamma(\dot{n} + \alpha)}{(\dot{\theta} + \frac{1}{\beta})^{\dot{n}+\alpha}}. \quad (2.37)$$

Hence the integrated posterior probability of observing a set of count data is

$$\Pr(\mathbf{n}|\boldsymbol{\theta}) = \frac{c}{\beta^\alpha \Gamma(\alpha)} \frac{\Gamma(\dot{n} + \alpha)}{(\dot{\theta} + \frac{1}{\beta})^{\dot{n}+\alpha}}. \quad (2.38)$$

It is possible to integrate out λ from the likelihood in Equation 2.11 such that inferences on parameters of consistent interest in geographic profiling from count data can still be made whilst reducing the potential computational burden of an extra parameter.

2.6 Discussion and conclusions

This chapter sees the construction of a Poisson geographic profiling model that can distinguish between an absence of evidence and evidence of absence by taking as input count data into the model's likelihood of which can consist of locations associated with no encounters.

The new Poisson finite mixture model estimates a variety of desired parameters as well as additional information yet to be fitted by any geographic profiling model. It allows for the estimation of source locations and dispersal distances (as in the DPM model). Additionally, it can estimate dispersal distances independent to each source as well as the expected number of events (i.e. broadly equivalent to the underlying population size of the species of interest).

This new model offers the first instance of the potential inclusion of temporal variability in relation to the sentinel site radius. A key change in the Poisson finite mixture model is that each sentinel site is associated with an expected number of events. This expectation is dependent on: the site radius ρ , the expected number of events in the search area λ , the time period the sentinel site is left open t (the time spent susceptible to events) and the site's spatial location with respect to a set of source locations. Intuitively, the longer a sentinel site is open for, the more events that are expected to be encountered. As suggested in many studies (Rossmo 2000, Raine et al. 2009, Santosuosso & Papini 2018), a more accurate geographic profiling model is one that considers temporal variability in the data to draw its inferences.

In the same way, the sentinel site radius ρ is also kept constant throughout analyses to ensure the approximation in Equation 2.10 is not erroneous. The effect of the site radius is like that of time: the larger the radius the more events expected at each site. Another way of considering the site radius is the distance with which a baited trap with different substances is sniffed out by a species (Brockerhoff et al. 2006). An assumption of the model is that an event was encountered by a single sentinel site only; dependent upon whether an event fell within a site's radius. Should an event be encountered by two sites, for example by proximity detectors (Borchers 2012), then it must have been observed at two distinct points in time. This motivates further need for temporal variability within geographic profiling models.

A sentinel site that encounters at least one event is indicative of the presence of, for example, an invasive species. The opposite however is not necessarily true for a site that encounters nothing.

If a sentinel site yields no encounters, then either an event is not present in that area or, it is, but the sentinel site failed to observe it. An assumption of the model is that if an event fell within a sentinel site's radius then it was immediately detected by that site. Detection probabilities are not always one and future studies may investigate relaxing this condition. Furthermore, the observation model could be adapted so that encounters are not governed by a site radius, such as in Chandler & Royle (2013). Capture probability could be dictated by a bivariate normal distribution around a sentinel site. Then the spatial component of the expected number of events at each site would be a product of two bivariate normals, one governing the distance from sources and one from a sentinel site.

The Poisson finite mixture model's likelihood has been built such that each source location can fit its own σ value; the standard deviation of the bivariate normal distribution associated with that specific mixture component. The ability to fit a single σ value across all mixture components was first built into the DPM model in Faulkner et al. (2016). The ability to fit independent σ values allows geographic profiling to capture more realistic scenarios in criminology. Offender behaviour is restricted by other commitments during a normal day, such as the requirement to return to work after a lunch break. As such a σ value associated with an anchor point centred on an offender's workplace may be considerably smaller than that of their home (Brantingham & Brantingham 1981, 1984).

When estimating source locations from the data the MCMC now explores a discretized space as opposed to a continuous one adopted in the DPM model (Verity et al. 2014). A discretised domain allows for the spatial prior on source locations to be defined in various ways. For example O'Leary (2009) described a spatial prior for a criminal's anchor point(s) via a set of ones and zeros, corresponding to inside or outside jurisdictional boundaries respectively. In Faulkner et al. (2018), the DPM model of geographic profiling was used to investigate individuals associated with poaching within the Savé valley conservancy, Zimbabwe. Given poachers could not reside within the conservancy, the final posterior surface was manipulated using a shapefile to notably reduce the probability in that area. This manipulation however was post-hoc to the analysis instead of being implemented within the Bayesian framework of the model. Given the DPM model adopts a Gibbs sampling algorithm, the prior on source locations must be conjugate to the form of the likelihood and has historically been set to a bivariate normal centred on the spatial mean of the data. The introduction of a discretised domain in the Poisson finite mixture model allows the user far more flexibility in building their priors on source locations.

Count data is common in ecology, for example in spatially explicit capture-recapture models and site occupancy models (MacKenzie et al. 2002, Royle et al. 2011, Kéry et al. 2011, Chandler & Royle 2013). The primary purpose of these models is to estimate abundance, rather than, as here, the location of sources. Spatially explicit capture-recapture models do treat these source locations (known as *activity centres*) as a latent variable but make differing assumptions about their numbers. Instead of assuming each encountered event is associated with a unique source, geographic profiling aims to partition the count data into clusters and finds the source location associated with each cluster. The main aim of this chapter was to build a model that estimates source locations using count data and so the architecture of the Poisson finite mixture model was built from the point of view of historical geographic profiling models that consistently focus on this objective. Geographic heterogeneities have been accounted for in previous geographic profiling models such as Kent (2009) and Mohler & Short (2012). This is however, the first time such information is accounted for in a geographic profiling model's likelihood whilst also estimating multiple sources locations.

This Poisson model is the first instance of a geographic profiling model to make its inferences using count data. The Poisson finite mixture model introduces the ability to fit independent dispersal distances across clusters in addition to the expected population size across the search area. Finally, the Poisson finite mixture model allows for added flexibility in the choice of spatial prior introduced by a user. Although the model boasts various developments to the field of geographic profiling, the model's ability to fit these parameters has yet to be tested. Model behaviour under different circumstances must also be explored. This point leads to the work of the next chapter, in which the model will be tested rigorously through extensive simulated data and explore a case study consisting of mosquito surveillance data in Miami-Dade County, Florida.

Chapter 3

Applying a Poisson Geographic Profiling Model

Abstract:

Building on the foundations of the previous chapter, the Poisson geographic profiling model is tested and developed further here in many ways. Simple exploratory examples comparing the new and existing models are provided as a visual aid to demonstrate that each model leads to different search strategies when attempting to identify source locations. Against a complex data set, the new model struggles to accurately estimate parameters and as such, a procedure to improve the quality of the algorithm estimating these parameters is implemented. The new model then demonstrates its ability to accurately estimate parameters of biological interest. Evidence in favour of this conclusion is provided via a Bayesian analogue of a power analysis. Model performance is also measured via real-world data in which the model infers breeding locations of mosquitoes in bromeliads in Miami-Dade County, Florida. In this case, the model undergoes further development by being equipped to process count data when the number of zero counts (i.e. the number of empty sentinel sites) is much greater than expected. At its core, the new model consistently demonstrates its ability to produce efficient search strategies when targeting invasive species or infectious diseases. By drawing its conclusions from count data, in place of point-pattern data, these interventions are shown to not only prove efficient, but are also seen to be unique to any pre-existing geographic profiling model.

The contents of this chapter are published in Stevens et al. (2021).

3.1 Introduction

In the previous chapter, I demonstrated how a Poisson finite mixture model can infer parameters of interest, common to geographic profiling, whilst also offering flexibility to a user when building their model. I am however, yet to rigorously test the ability of the Poisson model to return said parameters. Here I will test the ability of the Poisson model in a variety of ways.

Firstly, I will explore some simple, manually simulated data sets and produce a side by side comparison of the search strategies generated via the Poisson model, analysing a set of count data, and the Dirichlet process mixture (DPM) model analysing a set of repeat point-pattern data.

I will then push the model to its limits, analysing a more complex simulated data set, which will illustrate how the Poisson model's MCMC algorithm struggles to mix properly in certain cases. Consequently, I will introduce and implement a procedure to resolve this issue known as Metropolis-Hastings coupling. Once the Metropolis-Hastings coupling procedure has been implemented I will perform a Bayesian analogue of a power analysis on an extensive range of data sets simulated under the Poisson model.

I will then explore a real-world dataset in which the Poisson and DPM models attempt to infer the locations of breeding sites of the mosquito *Aedes aegypti* in ornamental bromeliads using trap surveillance data in Miami-Dade County, Florida (Wilke et al. 2018, 2019). *Ae. aegypti* is one of the primary transmitters of Zika virus across the globe (Hayes 2009, Hennessey et al. 2016), hence efficient search strategies are vital to find breeding locations and intervene accordingly. I will provide evidence to support that this mosquito surveillance data exhibits over-dispersion, a common attribute of ecological count data. I will describe the changes required for the model to consider over-dispersed count data in its inference process. Finally, the following output from each model is compared: estimates for sources locations, number of sources and dispersal patterns associated with the mosquito surveillance data.

3.2 Simple exploratory examples

Before I examine the ability of the Poisson model in too much detail I wish to evaluate its behaviour when analysing some extremely basic data sets. By exploring this simplistic data, the behaviour of the new model can be observed compared to what would normally be concluded from the DPM model. The two manually generated sets of count data can be seen in Figure 3.1. The

first consists of five sentinel sites. Two of these sentinel sites contain three observed events and are situated at either end of the spatial domain. The remaining three sites contain no observations and lie in between the two other sites. The second simple data set consists of ten sentinel sites. Five of these sites contain two observed events and are clustered together in the eastern part of the spatial domain. The remaining five contain no observations and are clustered in the north-western section of the domain. For these examples, the Poisson model inferred parameters based on the full data set whereas the DPM model discards those locations with no observations and counts become repeat point-pattern data (such as in Table 2.1).

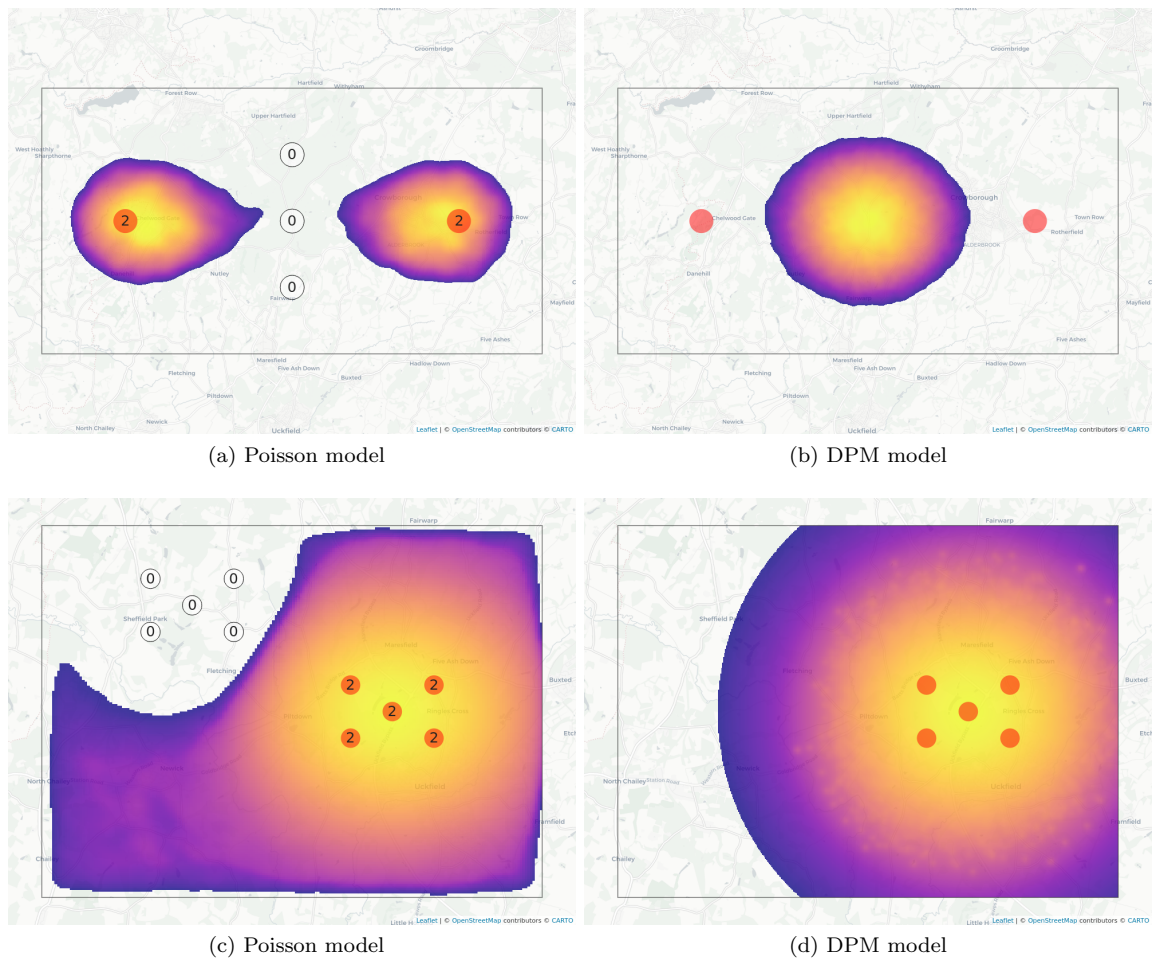


Figure 3.1: Geographic profiles produced by the Poisson and DPM models when analysing count data (a and c) and point-pattern data (b and d), respectively. Panels a) and b) show the top 20% of the geographic profiles, whilst c) and d) show the top 75%. Geographic profiles associated with the Poisson model were produced via MCMC sampling (algorithm 1) within the *Silverblaze* R package whilst the profiles associated with the DPM model were produced via the sampling protocol from the *RgeoProfile* package (Verity et al. 2014, Faulkner et al. 2016).

Figure 3.1 illustrates the results of these simple analyses. The colour scheme in this map indicates

areas likely to contain source locations with yellow, whilst those areas less likely to contain sources are highlighted in purple. The first case (Figure 3.1a and b) demonstrates the alternative results obtained when using count data via the Poisson model, including those locations associated with no encounters vs point-pattern data, via the DPM model ignoring locations with no encounters. These figures make it evident that sentinel sites with no encounters draw search priority away from common practices such as looking near the spatial mean of observed data, a method that is only effective when searching for a single source location (Stevenson et al. 2012, Verity et al. 2014)

In the second case (Figure 3.1c and d), the Poisson model shifted search priority away from areas with no encounters, compared to the DPM model, that gave equal search priority to areas where no encounters occurred and those areas with no information at all. The data in this section were produced manually by hand. I will now describe the process for simulating under the Poisson model in section 2.2.

3.3 Poor mixing and Metropolis-Hastings coupling

Here, I analysed a set of data explicitly simulated under the model described in section 2.2. For this data set, the number of sources, K , was set to three and the value for σ_{c_i} was set to 1.5 km for every source ($\sigma_1 = \sigma_2 = \sigma_3 = 1.5$). The sentinel radius, ρ was set to 0.3 km (resulting in an absolute error less than or equal to approximately 0.002 for the integration of the bivariate normal (section 2.4)). The number of events N was Poisson distributed with rate 1,000 and the number of sentinel sites was set to 225. The sampling strategy, i.e. the configuration of sentinel sites, was set as a uniform grid (15 x 15). The spatial prior for sources was also set to uniform whose extent was governed by the sentinel sites. As a result of this uniform prior, source locations were distributed randomly over the spatial extent. The mean and standard deviation for the univariate log-normal prior on σ_{c_i} were set to 1.5 and 5 respectively. The mean and standard deviation for the prior on λ were set to 1,000 and 500 respectively. The model was also set to estimate a single σ_{c_i} value, shared across sources.

When running the model, the initial parameter states for the MCMC algorithm were drawn from the priors described in the previous paragraph. Additionally, the number of mixture components was set to three, meaning the model knew the number of sources *a priori* to running the model. To evaluate the model's ability to infer parameters, the model was run five separate times on the same data. For each run, the MCMC algorithm ran for $5 \cdot 10^4$ burn-in iterations and $1 \cdot 10^4$ sampling

iterations. Convergence of the MCMC chains during burn-in were determined by Geweke’s metric (Geweke 1991). This metric derives a Z score by comparing the mean log-likelihood of the first 10% and last 50% of burn-in iterations (Cowles & Carlin 1996). If the samples produced are drawn from the stationary distribution of MCMC the MCMC chain then this Z score will follow a standard normal distribution. To determine MCMC convergence, Geweke’s metric was employed every $1 \cdot 10^4$ burn-in iterations. Figure 3.2 illustrates the particular data set to be analysed (full simulation process visible in Figure 2.1).

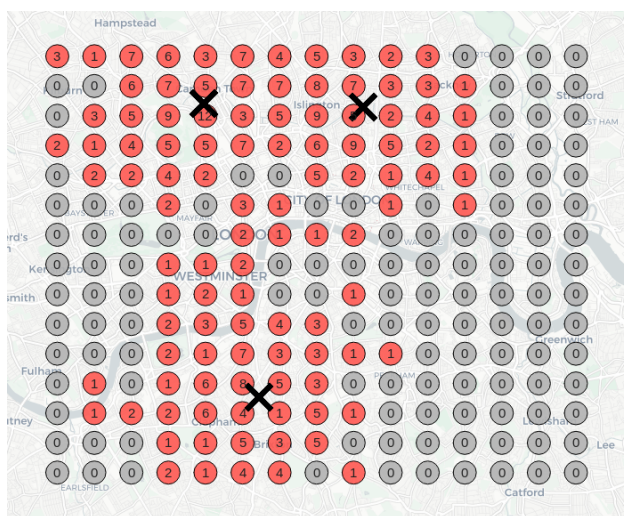


Figure 3.2: An example of manually simulated count data under the Poisson finite mixture model. Red sentinel sites denote those that observe one or more events and grey sites are those containing no observations. Each source location is marked with a cross. This map was created using the *CartoDB positron* layer via the R package *leaflet* (Cheng et al. 2019).

Figure 3.3 shows the results of this analysis, stating 95% credible intervals for the model’s fit (the log-likelihood) and estimates for σ_{c_i} and λ . Each MCMC chain converged within the fifty thousand burn-in iterations. This example illustrates how analysing the same data on five separate occasions can lead to different, but more importantly incorrect, conclusions about parameter estimates. The model also struggles to agree upon certain parameters despite it being generously offered the number of mixture components prior to running the MCMC algorithm, a luxury that is rarely available to problems in geographic profiling.

Additionally, this analysis reveals that Geweke’s metric for checking single-chain MCMC convergence is unreliable for problems in geographic profiling. Clearly the five parallel chains failed to converge to the same answer. A potential solution to this problem could be to bin the MCMC samples for each chain together and instead rely on the Gelman-Rubin diagnostic for

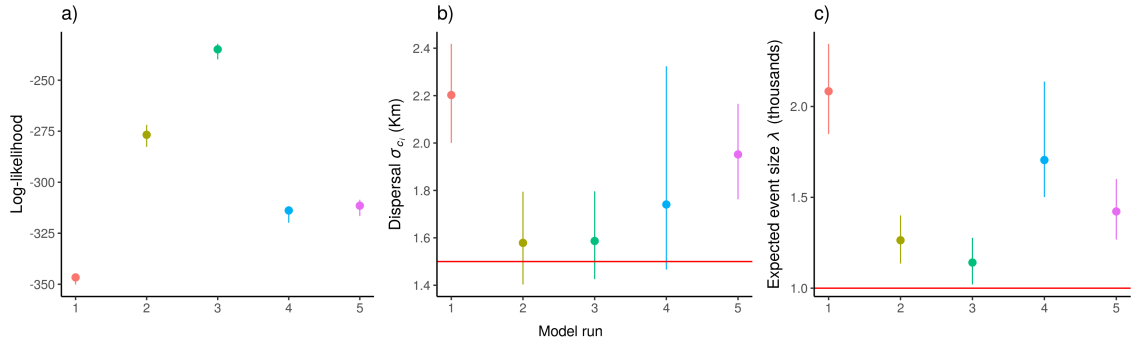


Figure 3.3: Results from five different MCMC chains running the Poisson finite mixture model. The output consists of 95% credible intervals for log-likelihoods, dispersal σ_{c_i} and expected event size λ . True values for the dispersal and expected event size are marked with red horizontal lines.

parallel chain convergence (Gelman et al. 2004, Verity et al. 2014). Given the Poisson model utilises a Metropolis-Hastings algorithm I will instead allow for communication between multiple MCMC chains via the Metropolis-Hastings coupling algorithm.

Metropolis-Hastings coupling allows for multiple MCMC chains to communicate with one another in the following way. Firstly, parameter values are updated within chains following a similar procedure to algorithm 1. The MCMC algorithm then proposes to swap current parameter values between chains given a similar acceptance probability. Instead of each chain exploring the posterior distribution however, a *heat* is applied to each chain such that the hotter a chain, the more Metropolis-Hastings steps are accepted within said chain. If there are h parallel chains, then each chain is associated with a heat β_i , where

$$\beta_i = \begin{cases} \frac{1}{i^P} & \text{for } i \in \{1, 2, \dots, h-1\}, \\ 0 & i = h. \end{cases} \quad (3.1)$$

Here P is referred to as the *thermodynamic power* (Atchadé et al. 2011). Each chain is heated by raising the likelihood to the power of β_i where the coldest chain explores the posterior of interest (when β_i is equal to one) and the hottest explores the model's priors (when β_i is equal to zero).

Within chain i the probability of accepting a new proposed parameter value is altered as follows. For a Metropolis-Hastings step such as Equation 2.12, the acceptance probability becomes:

$$Acc(\boldsymbol{\mu}'_{c_i}, \boldsymbol{\mu}_{c_i}) = \min \left(1, \frac{L(\boldsymbol{\mu}'_{c_i})^{\beta_i} \mathcal{F}(\boldsymbol{\mu}'_{c_i}) g_{\mu}(\boldsymbol{\mu}_{c_i} | \boldsymbol{\mu}'_{c_i}, \epsilon_{\mu})}{L(\boldsymbol{\mu}_{c_i})^{\beta_i} \mathcal{F}(\boldsymbol{\mu}_{c_i}) g_{\mu}(\boldsymbol{\mu}'_{c_i} | \boldsymbol{\mu}_{c_i}, \epsilon_{\mu})} \right), \quad (3.2)$$

An analogous method is then used to update each σ_{c_i} . Note also that for chain i , the conditional likelihood for λ in Equation 2.14 changes to:

$$L(\lambda)^{\beta_i} = (\lambda)^{\dot{n}\beta_i} e^{-\lambda\dot{\theta}_i^{\beta_i}}, \quad (3.3)$$

leading, in the same manner, to the altered conditional posterior distribution for λ

$$\Pr(\lambda|\mathbf{n}, \boldsymbol{\theta}) = \frac{(\eta + \dot{\theta})^{\eta + \dot{n}\beta_i}}{\Gamma(\zeta + \dot{n})} (\lambda)^{\zeta + \dot{n}\beta_i - 1} e^{-\lambda(\eta + \dot{\theta}\beta_i)}. \quad (3.4)$$

Once the MCMC algorithm has updated $\{\boldsymbol{\mu}_{c_i}\}$, $\{\sigma_{c_i}\}$ and λ a swap between chains is proposed. Using a bubble sort approach, swaps between chains i and $i + 1$, for i up to $h - 1$, are proposed using an acceptance ratio reliant on the relative likelihoods between chains. The probability of accepting a swap between the current parameter values of chain i and $i + 1$ is:

$$Acc(L_i, L_{i+1}) = \min \left(1, \frac{L_i^{\beta_{i+1}} L_{i+1}^{\beta_i}}{L_i^{\beta_i} L_{i+1}^{\beta_{i+1}}} \right). \quad (3.5)$$

I will now revisit the simulated data from earlier where algorithm 1 has been altered in light of Equation 3.2, Equation 3.4 and Equation 3.5. Once again, the model ran five times, each with the same random number generator seed as before, now utilising the Metropolis-Hastings coupling process described above. The choice of β_i values can be manually chosen in the *silverblaze* package. Additionally, values can be optimised such that transition probabilities between heated chains reach a certain threshold. The protocol for optimising these heats involves running a pre-specified number of chains, then the MCMC algorithm is executed and checks transition probabilities for the burn-in phase only. Any transition probabilities below a certain threshold results in the birth a new MCMC chain with an intermediate heat based on the two chains producing the probability below the threshold. Following this protocol, 30 chains were used when running the MCMC algorithm with the Metropolis-Hastings coupling process to ensure a minimum transition probability of 50%.

Results for these five chains with Metropolis-Hastings coupling turned on, alongside the original five, with coupling turned off, can be seen in Figure 3.4. This figure visibly demonstrates that the MCMC algorithm utilising the Metropolis-Hastings coupling procedure returns consistently sensible parameter estimates. Now that MCMC mixing is no longer an issue for the Poisson model, I will describe an extensive set of simulations that were used to test the model via a Bayesian analogue of a power analysis.

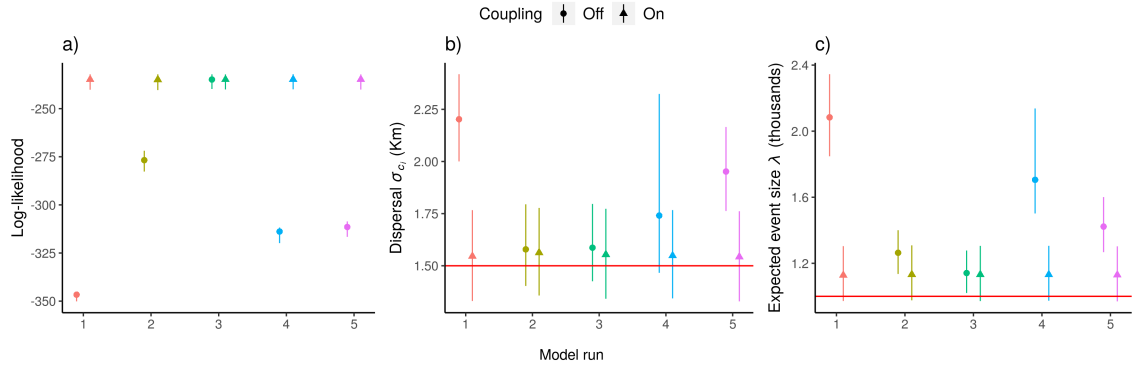


Figure 3.4: Results from running the Poisson model on the same data, ten times, five with (triangles) and five without (circles) Metropolis-Hastings coupling turned on. The output consists of 95% credible intervals for the log-likelihood, dispersal σ_{c_i} and expected event size λ .

3.4 Power analysis

3.4.1 Methods

I performed a Bayesian analogue of a traditional power analysis by simulating data from the Poisson model described in Equation 2.11 and explored the ability of the model to infer the true parameter values. For the validation of the Poisson model I explored a parameter space similar to Verity et al. (2014).

Source locations were generated uniformly at random from a longitudinal and latitudinal extent of -0.2 to 0.0 and 51.45 to 51.55 respectively. The spatial prior \mathcal{F} was defined over a 100x100 grid whose extent matched the same values as the source locations plus a 25% margin at each limit (-0.25 to -0.05 and 51.425 to 51.575). The specification of this search area led to a spatial coverage of 345.53 km². The number of sources K ranged from one to five, the true value of σ_{c_i} was set to 1.5 km and the number of events N was Poisson distributed with rates 100, 1,000 and 10,000. For the power analysis, note that each source shared the same σ_{c_i} (i.e. $\sigma_1 = \sigma_2 \dots = \sigma_k$). This model type was chosen for simplicity given the study focussed on the models ability to estimate source locations in place of independent dispersal. Finally, the number of sentinel sites was set to 25, 100 or 400 and they were distributed over space either uniformly at random or as a grid.

To determine the correct number of source locations, the Poisson model ran seven times, in each case searching from one up to seven sources to allow for cases where the model overestimates K . The most suitable value of K was then chosen via the deviance information

criterion (DIC) from Spiegelhalter et al. (2014) given by:

$$\text{DIC} = -2 \cdot \text{E}[\log(\text{Pr}(\mathbf{n}|\boldsymbol{\theta}, \lambda))] + c \cdot \text{Var}[\log(\text{Pr}(\mathbf{n}|\boldsymbol{\theta}, \lambda))]. \quad (3.6)$$

The DIC is a metric for model comparison, used here to determine the best of a set of models, where each model went in search of a different number of source locations. The DIC was calculated similarly to other model comparison metrics such as Akaike’s information criterion (AIC) and the Bayesian information criterion. The DIC consists of two terms: a model fit term, equivalent to -2 multiplied by the mean of the log-likelihood of the data, and a penalty term for model complexity, equivalent to the variance of the log-likelihood of the data multiplied by some positive constant. This constant c controls how harshly models are penalised for their complexity and throughout this and remaining chapters c was set to four (see Spiegelhalter et al. (2014) for a discussion on choosing c). Parameter estimates for $\{\boldsymbol{\mu}_{c_i}\}$, $\{\sigma_{c_i}\}$ and λ were then pulled for the value of K chosen by the DIC.

These settings lead to 360 parameter combinations, each of which were repeated one hundred times and results were averaged. The log-normal prior on the dispersal σ_{c_i} was set as either tight (standard deviation of one) or wide (standard deviation of 100) around the true value (1.5 km). The gamma prior on λ was set such that the mean was equal to the true rates (100, 1,000 or 10,000) and the standard deviation was either the true rate or a tenth of the true rate. The burn-in and sampling period for the MCMC chains were set to $5 \cdot 10^4$ iterations. Geweke’s metric was once again employed to determine if the MCMC chains reached a stationary distribution (Cowles & Carlin 1996). This was tested at multiples of $1 \cdot 10^4$ iterations during burn-in. Once again, the initial states for each parameter within the MCMC algorithm were drawn from the model’s priors. Following the protocol described in the previous section for optimising the heats of each MCMC chain (β_i), simulated data sets ran between five and 184 rungs to ensure an acceptance rate greater than 50%. Depending on the sample size of the data set (n), the number of sources searched for by the model (K) and the number of heated rungs (h), the execution time of the model was anywhere between a few minutes and a couple of days.

The success of the Poisson model was measured in a similar way to many other geographic profiling models, via a source’s hit score (Rossmo 2000). This metric is defined as the area searched before finding a source divided by the total search area, indicating that the lower a source’s hit score, the more accurately it was identified by the model. The search strategy itself,

consists of starting at the location with the highest value on the geographic profile and working downwards. Given the number of sources produced for each data set varied from one to five, results for each simulation produced between one and five hit scores. As such, the Gini coefficient was used to summarise these hit scores into a single statistic.

The Gini coefficient is a metric originally developed in economics to describe the imbalance in wealth distribution of a country across its population (Dorfman 1979). It describes what proportion of an economy’s income is received by the lowest income receivers. In the context of geographic profiling, it was used to describe the proportion of source locations discovered over area searched via a geographic profile. A coefficient of one corresponded to a perfect search strategy where all sources were found by searching the smallest area possible. Additionally, a Gini coefficient of zero corresponded to a linear relationship between sources found and area searched ($x\%$ of sources found by searching $x\%$ of the search area). Finally, a Gini coefficient of minus one corresponded to the most inefficient search strategy (all sources found simultaneously once the entire area had been searched).

The DPM and Poisson models were both developed in the R programming language (R Core Team 2019) and implemented in the *Rgeoprofile* (github.com/bobverity/Rgeoprofile) and *silverblaze* (github.com/Michael-Stevens-27/silverblaze) packages respectively.

3.4.2 Results

The results of the Bayesian power analysis, in the form of average Gini coefficient, can be seen in Table 3.1. There was a consistent decrease in power as the number of sources increased but an increase in power given more sampling locations. Of the 360 parameter combinations, 278 (77%) reached a Gini coefficient of 0.9 or higher. Table 3.1 also shows that a uniform site configuration yielded a higher Gini coefficient more often than a random layout (134 of 180 cases). Additionally, tight priors on σ_{c_i} and λ in place of wide priors yielded higher Gini coefficients in 94 and 116 of 180 cases respectively.

The new model was also tested on its ability to return the true number of source locations K , the true dispersal σ_{c_i} and finally, the expected number of events λ . The new model correctly fitted the true value of K in 57% of cases, it fitted within 1 of the true value in 76% of cases and within 2 in 88%. The error in estimated K compared to the true K for the full data along with each sample size can be seen in Figure 3.5. The Poisson model was more likely to underestimate

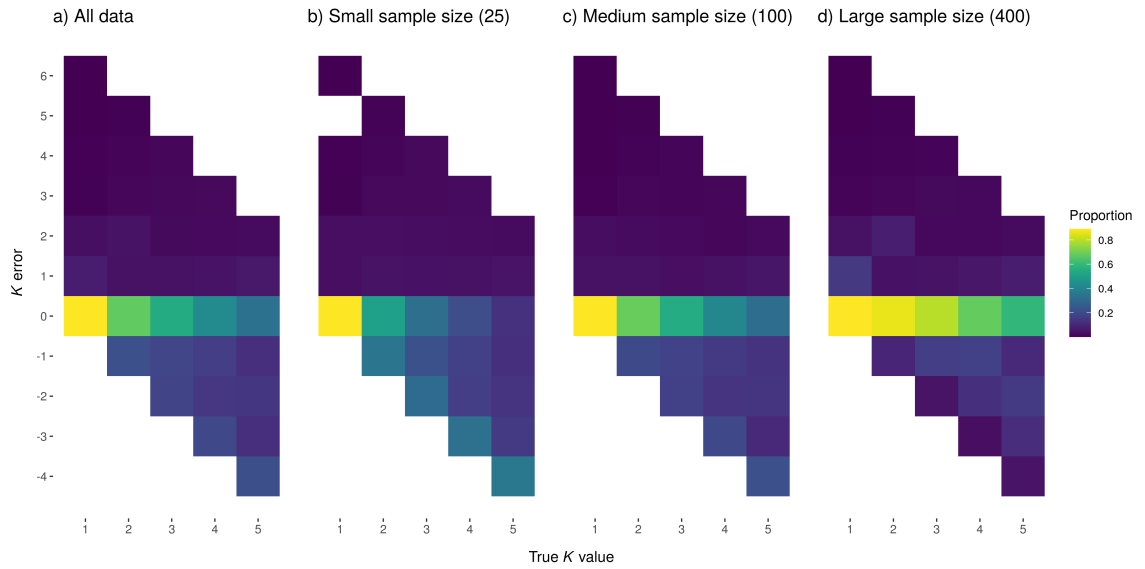


Figure 3.5: Tile plots illustrating the error incurred under different K models when selecting K via the deviance information criterion. From left to right the panels show results for a) every sample size, b) a small sample size (25), c) a medium sample size (100) and d) a large sample size (400).

the value of K (approximately a third of the time) than it was to overestimate it (a tenth of the time). Additionally, larger sample sizes lead to a more accurate estimate of K .

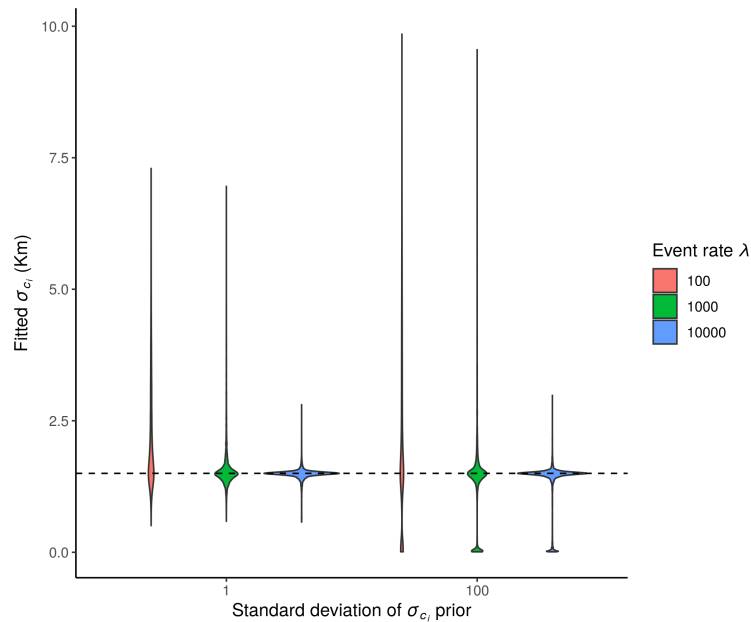


Figure 3.6: Estimates for σ_{c_i} across priors and differing expectations on the number of events, λ .

The true value of σ_{c_i} was set to 1.5 km. The model's average estimate for σ_{c_i} was 1.68 km (standard deviation of 0.94). Model estimates for σ_{c_i} , split by prior type, can be seen in Figure 3.6. True values of λ were set to 100, 1,000 and 10,000. The model's average estimates for λ were 118,

1,094 and 10,501 (with standard deviations of 34, 274 and 1,856 respectively). Model estimates for λ , split by prior type, can be seen in Figure 3.7.

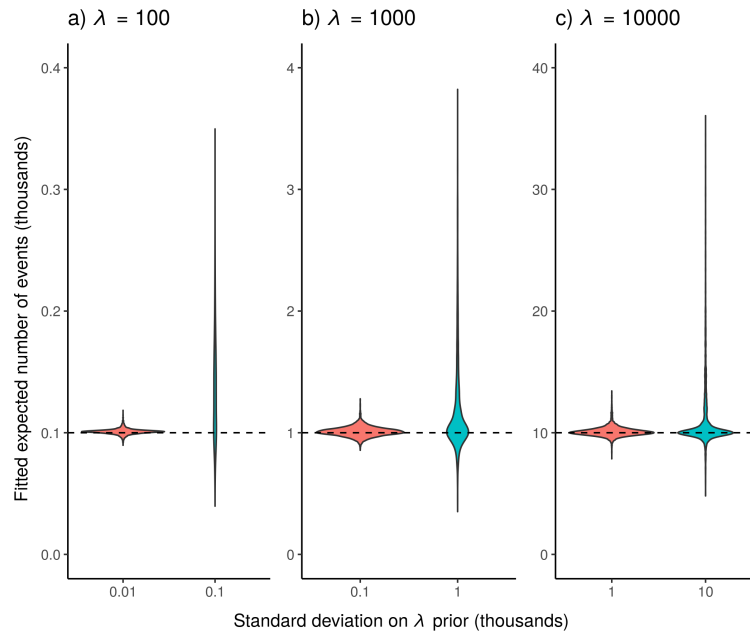


Figure 3.7: Estimates for λ given each of the true rates: a) 100, b) 1,000 and c) 10,000 across each prior (tight and wide).

	K sources	1				2				3				4				5			
	λ prior	Tight		Wide		Tight		Wide		Tight		Wide		Tight		Wide		Tight		Wide	
	σ_{c_i} prior	T	W	T	W	T	W	T	W	T	W	T	W	T	W	T	W	T	W	T	W
λ	m sites	Uniform																			
100	25	0.994	0.992	0.992	0.993	0.892	0.874	0.863	0.877	0.798	0.751	0.771	0.769	0.740	0.721	0.683	0.690	0.673	0.662	0.655	0.618
100	100	0.997	0.996	0.996	0.995	0.948	0.951	0.943	0.941	0.878	0.875	0.838	0.818	0.789	0.791	0.763	0.729	0.730	0.684	0.712	0.683
100	400	0.999	0.999	0.999	0.999	0.993	0.992	0.988	0.986	0.970	0.969	0.955	0.949	0.931	0.925	0.899	0.903	0.876	0.865	0.857	0.816
1000	25	0.998	0.998	0.998	0.998	0.985	0.982	0.981	0.980	0.943	0.929	0.929	0.936	0.872	0.866	0.874	0.876	0.823	0.795	0.814	0.795
1000	100	0.999	0.999	0.999	0.999	0.996	0.997	0.996	0.996	0.989	0.988	0.982	0.984	0.973	0.971	0.968	0.954	0.946	0.935	0.934	0.918
1000	400	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.997	0.997	0.996	0.996	0.993	0.993	0.992	0.992	0.986	0.988	0.980	0.983
10000	25	0.999	0.999	0.999	0.999	0.998	0.998	0.997	0.996	0.995	0.993	0.994	0.989	0.985	0.985	0.984	0.977	0.969	0.963	0.954	0.955
10000	100	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.998	0.997	0.998	0.997	0.996	0.995	0.996	0.993	0.994
10000	400	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.998
		Random																			
100	25	0.994	0.994	0.991	0.992	0.904	0.895	0.899	0.868	0.822	0.807	0.773	0.783	0.715	0.750	0.702	0.700	0.689	0.677	0.660	0.664
100	100	0.996	0.994	0.995	0.994	0.950	0.942	0.935	0.928	0.853	0.855	0.816	0.829	0.775	0.768	0.747	0.755	0.721	0.722	0.693	0.692
100	400	0.997	0.996	0.998	0.997	0.984	0.990	0.984	0.984	0.959	0.946	0.943	0.939	0.904	0.907	0.894	0.872	0.847	0.845	0.845	0.810
1000	25	0.998	0.996	0.997	0.998	0.969	0.964	0.969	0.965	0.908	0.910	0.901	0.918	0.853	0.845	0.833	0.821	0.814	0.809	0.744	0.766
1000	100	0.999	0.999	0.999	0.999	0.994	0.995	0.995	0.993	0.986	0.987	0.980	0.981	0.968	0.958	0.949	0.957	0.941	0.931	0.906	0.902
1000	400	0.999	0.999	0.999	0.999	0.998	0.996	0.997	0.998	0.994	0.995	0.995	0.993	0.989	0.990	0.991	0.989	0.981	0.977	0.978	0.978
10000	25	0.999	0.998	0.999	0.999	0.997	0.995	0.988	0.989	0.984	0.982	0.982	0.976	0.966	0.965	0.961	0.950	0.938	0.937	0.926	0.912
10000	100	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.998	0.998	0.997	0.996	0.995	0.995	0.992	0.991	0.991	0.990
10000	400	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.999	0.998	0.998	0.998	0.999	0.998	0.996	0.998	0.997	0.996

Table 3.1: Gini coefficients for the extensive set of simulated data. Each cell represents the average coefficient across one hundred replicates.

3.5 Bromeliad analysis

3.5.1 Over dispersion in ecological count data

Trap surveillance data from Wilke et al. (2018, 2019) of the mosquito *Ae. aegypti* in Miami- Dade County, Florida were used to test the Poisson model's ability to find breeding sites in ornamental bromeliads. Data consisted of 124 traps with encounters per trap ranging from 0-1033. A total of 94 traps contained *Ae. aegypti* and 30 did not. The average distance between an empty trap and its nearest positive trap was 55 metres with a standard deviation of 77 metres. There were 51 ornamental bromeliad patches that were checked for immature stages of mosquitoes where 30 contained *Ae. aegypti* larvae and 21 did not. In this section, trap data recorded during 2017 were analysed to match the time period bromeliad patches were surveyed. A map of the trap surveillance data, alongside bromeliad patches, can be seen in Figure 3.8.

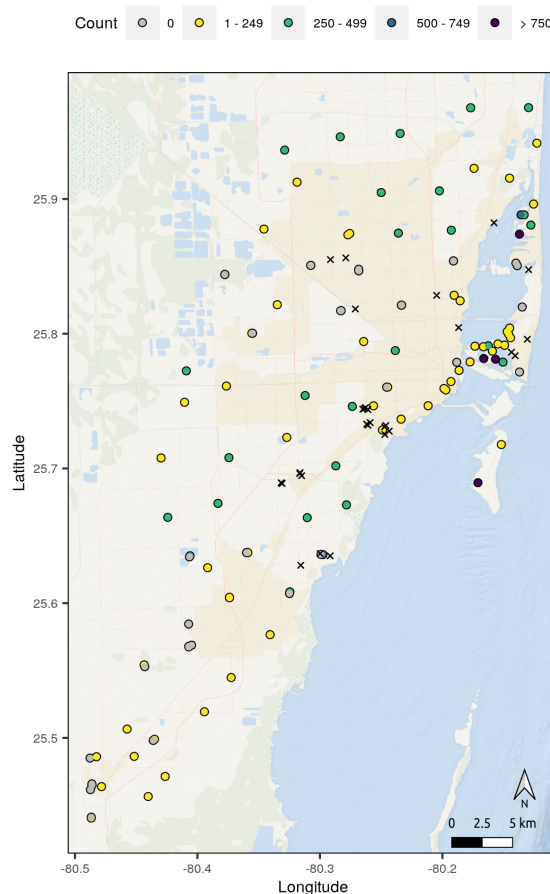


Figure 3.8: A map of the trap surveillance data and bromeliad breeding sites from Wilke et al. (2018, 2019). Traps are coloured depending on the number of mosquitoes caught at each site. Bromeliad patches containing *Ae. aegypti* larvae are marked with a cross. Map created using the *ESRI ocean layer* via QGIS.org (2021).

A fundamental feature of the Poisson model is its assumption of equal mean and variance across the counts exhibited in a data set. However, in ecology, it is often common for count data to stray from this assumed equal mean and variance and instead allow for count variation to be governed by a linear or quadratic function of the mean (Ver Hoef & Boveng 2007). This phenomenon is referred to as *over-dispersion*. Over-dispersion can be caused by a range of factors such as sampling, aggregation, environmental variability or a combination of the above (Lindén & Mäntyniemi 2011).

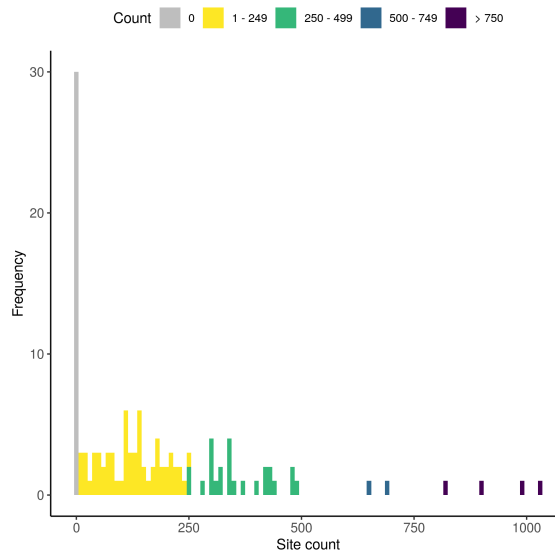


Figure 3.9: A frequency plot illustrating the number of mosquitoes caught in each trap.

Although the frequency plot in Figure 3.9 may indicate that the mosquito counts are indeed over-dispersed, this phenomenon can only be confirmed by knowing the mean and variance in counts for each source location. So to demonstrate that over-dispersion is present within the trap surveillance data, I spatially binned the counts into pseudo-sources by splitting up the spatial domain in the form of a grid at eleven different resolutions. In each case I calculated the count mean and variance within those pseudo-sources (for bins containing two or more counts) and compared Akaike’s information criterion for a linear and quadratic model connecting the count means and variances. Multiple models were compared to investigate which best described the relationship between the mean and variance in counts.

The grid resolutions, along with the AIC for each model, can be seen in Table 3.2. In each case the quadratic model performs as well or better than the linear model. To account for over-dispersed data, the Poisson model can be altered such that the probability of observing the data (the likelihood) is governed by a negative-binomial density in place of a Poisson density. The

Resolution	Model AIC	
	Linear	Quadratic
5 x 5	230.04	231.68
6 x 6	301.49	288.12
7 x 7	281.78	283.05
8 x 8	275.65	258.90
9 x 9	417.09	411.50
10 x 10	275.07	265.97
11 x 11	249.84	247.94
12 x 12	250.50	250.95
13 x 13	237.49	224.72
14 x 14	166.24	149.55
15 x 15	114.09	111.43

Table 3.2: AIC values for those models fitting a linear and quadratic relationship between the count mean and variance at pseudo-sources dictated via each grid resolution.

negative-binomial density normally takes the form

$$\Pr(X = x) = \binom{x+r-1}{x} p^r (1-p)^x, \quad (3.7)$$

where x refers to the number of failures in a set of Bernoulli trials, p is the probability of trial success and the Bernoulli trials themselves are terminated after r trials. In ecology, this expression is often re-parameterised such that the distribution is governed by a mean and a variance in the following way (Lindén & Mäntyniemi 2011)

$$\Pr(X = x) = \frac{\Gamma(r+x)}{x!\Gamma(r)} \left(\frac{r}{r+M}\right)^r \left(\frac{M}{r+M}\right)^x. \quad (3.8)$$

Here, $M = \frac{pr}{1-p}$ is the mean of the distribution and the variance is of the form $\sigma^2 = M + \frac{M^2}{r}$. By substituting in the Poisson mean derived after Equation 2.10, the likelihood from Equation 2.11 becomes

$$\Pr(\mathbf{n}|\lambda t, \boldsymbol{\theta}, r) = \prod_{j=1}^m \frac{\Gamma(r+n_j)}{n_j!\Gamma(r)} \left(\frac{r}{r+\lambda t\theta_j}\right)^r \left(\frac{\lambda t\theta_j}{r+\lambda t\theta_j}\right)^{n_j}. \quad (3.9)$$

So each n_j are now assumed to be drawn from a negative-binomial density with mean $M = \lambda t\theta_j$ and variance $\sigma_{c_i}^2 = \lambda t\theta_j + \alpha(\lambda t\theta_j)^2$. I have chosen to re-write the co-efficient of the second order term as $\alpha = \frac{1}{r}$ for ease when interpreting the variance in response to changes in r . Note that as α tends to zero, the variance neatly returns to $\lambda t\theta_j$, matching the assumption of the Poisson model.

In the previous chapter I demonstrated how the Poisson model can estimate an independent σ_{c_i} per source. It is entirely possible to alter the expectation $\lambda t\theta_j$ to also estimate an independent

expected number of events for each source. An assumption of the Poisson model is that each source is associated with the same expected number of events $\frac{\lambda}{K}$. This assumption is not necessarily valid for the mosquito surveillance data, hence I introduce an independent λ_{c_i} per source, where $\lambda = \sum_{c_i=1}^K \lambda_{c_i}$. The likelihoods in Equation 2.11 and Equation 3.9 can then be altered accordingly to accommodate independent λ_{c_i} . Recall that the expected number of events observed at sentinel site \mathbf{s}_j is given by $\lambda\theta_j t$ where

$$\lambda\theta_j t = \frac{\lambda t \pi \rho^2}{K} \sum_{c_i=1}^K f_{BN}(\mathbf{s}_j | \boldsymbol{\mu}_{c_i}, \mathbf{I}_2 \sigma_{c_i}^2).$$

By slightly altering this equation, it is possible to derive the expectation for each sentinel site, given an independent λ_{c_i}

$$\theta_j t = t \pi \rho^2 \sum_{c_i=1}^K \lambda_{c_i} \cdot f_{BN}(\mathbf{s}_j | \boldsymbol{\mu}_{c_i}, \mathbf{I}_2 \sigma_{c_i}^2). \quad (3.10)$$

Hence, the expected number of events has been merged into the source dependent part of the expectation for each sentinel site (θ_j). As a result of this change, the MCMC algorithm can no longer use the Gibbs sampling step that draws a new λ from the conditional posterior in Equation 2.15. In place of this, each λ_{c_i} is updated with its own Metropolis-Hastings step analogously to $\boldsymbol{\mu}_{c_i}$ and σ_{c_i} in Equation 2.12.

3.5.2 Model settings

When running the analysis on the mosquito surveillance data set, model priors were set as follows. For source locations, the DPM model used a bivariate normal centred on the mean of the positive surveillance locations with standard deviation equal to the maximum distance between the positive data and their mean (Verity et al. 2014). The final surface was then manipulated post-hoc to exclude the possibility of source locations in the sea using a shape file (*South Florida Region Shapefile, Miami-Dade County - Open Data Hub* 2018, Faulkner et al. 2018). The negative binomial model used the same shape file for its prior on source locations where each cell's probability mass was uniform on land and zero in the sea.

For the dispersal parameter σ_{c_i} , a diffuse prior was set for the DPM (mean of 2.5 and standard deviation of 10). The same hyper-parameters were used for the negative binomial model's prior on σ_{c_i} in addition to a tight prior (standard deviation of 1) to explore model behaviour under different priors. These priors conform to previous studies placing *Ae. Aegypti*

dispersal somewhere between zero and five km (Service & Place 1997, Gorrochotegui-Escalante et al. 2000). The sentinel radius ρ was set to 0.1 km. For the negative binomial model, tight and diffuse log-normal priors were set for the expected number of events λ_{c_i} (means of $1 \cdot 10^6$ and standard deviations of $1 \cdot 10^5$ and $1 \cdot 10^6$). Note that for this analysis the negative binomial model estimated a single σ_{c_i} (i.e. $\sigma_1 = \sigma_2 \cdots = \sigma_k$) shared across sources, whilst also estimating an independent λ_{c_i} for each source. The prior on α was also log-normal with mean 1 and standard deviation 100.

To estimate the number of sources K , the negative binomial model was run 25 times, in each case searching for that specific number of sources, where the DIC was again utilized to pick the most suitable value of K to explain the data (Spiegelhalter et al. 2014). The DPM model used five sampling chains, each with a burn-in period of $5 \cdot 10^2$ iterations and a sampling period of $1 \cdot 10^4$ iterations. The negative binomial model ran for $5 \cdot 10^4$ burn-in and sampling iterations with convergence checked at each multiple of $1 \cdot 10^4$ iterations during burn-in (Cowles & Carlin 1996). Similarly to the power analysis, a Metropolis-Hastings coupling step was utilised to ensure healthy MCMC mixing, leading to between 30 and 45 rungs to ensure an acceptance rate greater than 50%.

3.5.3 Results

The mosquito surveillance data and bromeliad patches can be seen alongside the geographic profiles created by the negative binomial and DPM models in Figure 3.10. The DPM model determined 91 clusters best described the data. Within the negative binomial model, different combinations of parameter priors greatly affected model choice. A wider prior on the expected population size lead to a much lower number of clusters ($K = 2$). Under a more informative expected population prior, model choice for K yielded larger values than under an uninformative prior (K equal to 14 and 18 for informative and uninformative prior on the dispersal, respectively). See Figure 3.11 for more details. Hit score percentages for the DPM model ranged from 0.13% to 41.64% with an average of 11.12%. The negative binomial models hit scores percentages ranged from 1.64% to 66.41% with an average of 21.25%.

Under informative priors the negative binomial model returned a dispersal σ_{c_i} value between 1.41 and 7.03 km (95% credible interval) whereas under less informative priors estimates reached up to 22 km. Comparatively, the DPM model estimated σ_{c_i} between 9 and 10 metres. The total expected population density of *Ae. aegypti* was estimated between 3.64 to 28.28 million for 2017.

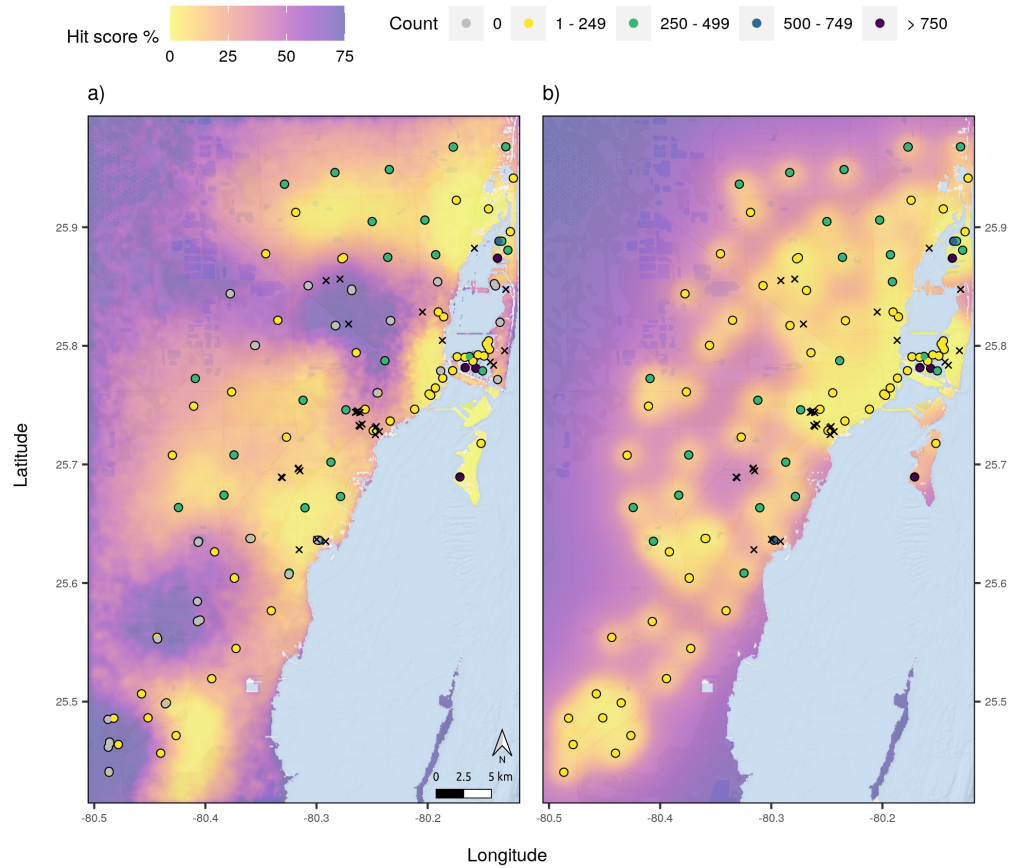


Figure 3.10: The geographic profiles in Miami-Dade County Florida, created by a) the negative binomial model via the 2017 mosquito count data under informative priors ($K = 14$) and b) the DPM model via repeat point-pattern data ($K = 91$). Locations of bromeliad breeding sites are marked with a cross. Given the proximity between positive and empty traps, some positive traps are only visible in panel b).

The over-dispersion parameter α was consistently estimated between 2.40 and 4.35.

3.6 Discussion and conclusions

Accounting for different information can lead to different search strategies. As illustrated by the simple example at the beginning of this chapter, sentinel sites with no encounters drew search priority away from common search practices such as looking near the spatial mean of observed data. Additionally, the new model drew search priority away from those areas containing no encounters to those with no information at all, compared to the DPM model, where these areas were treated equally. An assumption when using the DPM model is that perfect observations are made, meaning all events that occur will be seen. This assumption is valid in studies where the exact locations of events are recorded (Faulkner et al. 2015, Smith, Downs, Mitchell, Hayward, Fry & Le Comber 2015, Struebig et al. 2018) but is less suitable in those that adopt a sampling

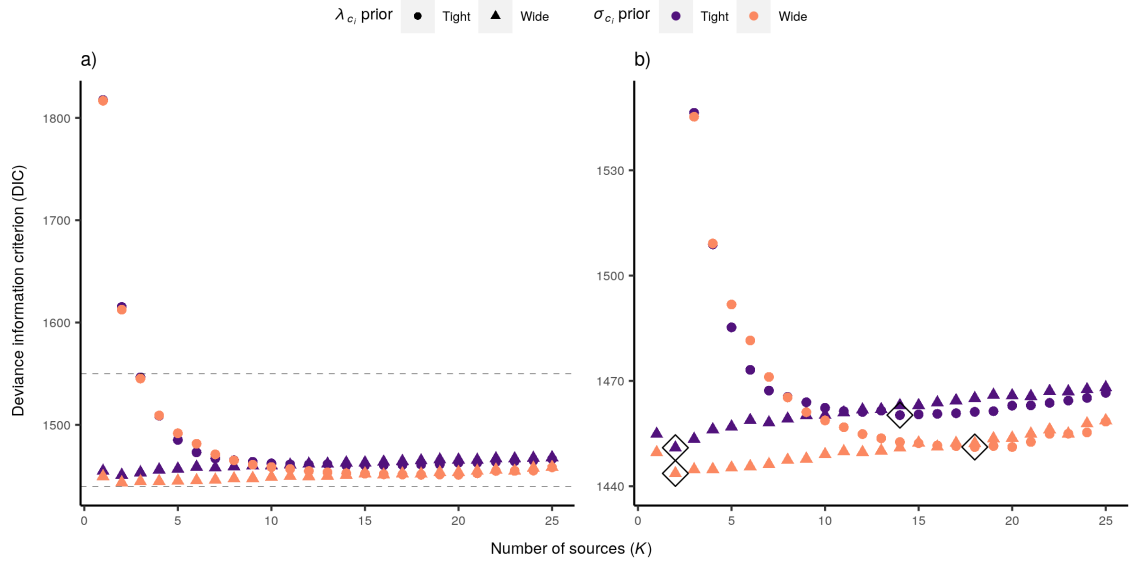


Figure 3.11: The deviance information criterion for each negative binomial model searching for K source locations under different priors. Of these models, the most suitable, indicating the best number of source locations describing the data, was greatly affected by the combination of priors on parameters. For each prior combination, the best model is marked with a diamond (corresponding to minimum DIC). Full DIC values are displayed in panel a) with a zoomed version in panel b).

strategy using sentinel sites (Faulkner et al. 2016).

I have shown via a power analysis and real-world case study that the new model can estimate a variety of parameters common to geographic profiling in addition to new ones. It accurately estimated source locations, dispersal σ_{c_i} and the number of source locations, K , in addition to the newly fitted expected number of events, λ_{c_i} , and over-dispersion parameter, α . Finally, it allows for cases where σ_{c_i} , and λ_{c_i} vary from source to source.

The new model was able to identify source locations efficiently, as was reflected in consistently high Gini coefficients across parameter combinations. Average Gini coefficients never fell below zero, the value associated with a random search strategy. In the case of the mosquito surveillance and bromeliad data set, the DPM model returned better hit scores than that of the negative binomial model. This was likely due to the negative binomial model's tendency to push search priority away from those areas containing empty sentinel sites, coupled with the fact that a subset of the bromeliad locations were identified near empty traps. In contrast, the DPM model placed search priority nearest the positive data. In this case, defaulting to a search strategy that looks near positive data has proven to be better than a strategy advising to steer clear of empty surveillance traps.

Estimating the number of sources in the new model was less straightforward than in the DPM model. A major strength of the latter is that it did not require the specification of the number of source locations in advance. In the new model, the algorithm ran many times and used the deviance information criterion to find the most appropriate number of source locations. This process produced accurate results for simulated data with large enough sample sizes but was shown to produce different results dependent on prior choice in the real-world case study. Here, combinations of diffuse and informative priors indicated suitable K values at 2, 14 and 18 (Figure 3.11). Although a K value of two corresponded to the lowest DIC value, estimates of σ_{c_i} in these cases were up to 22 km. For a K value of 14, estimates for σ_{c_i} were much more sensible. The DPM model estimated σ_{c_i} between nine and ten metres. In both cases, each model's estimate for σ_{c_i} heavily contradicted the prior beliefs built from the biological understanding of *Ae. Aegypti* dispersal. I therefore suggest careful consideration be taken when building priors and advice from field experts and collaborators is sought.

Furthermore, it would be naive to assume the number of sources fitted by either model or the known number of breeding sites reflects the true number of sources, of which could consist of any body of stagnant water (Ramasamy et al. 2011). Given the ground truth about the true number of sources is unknown, there is no way of evaluating the hit scores of these hypothetical locations. I therefore suggest that the number of source locations fitted by either model play the role of a lower bound on the true value of K . This suggestion is especially valid for the Poisson model that showed it was more likely to underestimate K with smaller sample sizes. I also suggest future work could focus on migrating the new model to a non-parametric framework, similarly to the DPM model, in place of estimating K by running the model multiple times.

As is visible in Figure 3.4, the Metropolis-Hastings coupling protocol is a necessity to ensure that the MCMC algorithm for the Poisson or negative-binomial models mix correctly. Additionally, section 3.3 illustrated that Geweke's metric for MCMC convergence is unreliable for multi-modal problems. The most one can assume from Geweke's metric is that the samples obtained via the MCMC algorithm are from a stationary distribution, but not necessarily the posterior of interest.

In addition to real data, it is entirely possible for the new model to utilise pseudo-absences in its inference process (Barbet-Massin et al. 2012). If un-sampled locations were to be replaced with pseudo-absences, I would expect the model to focus search priority entirely on locations with positive data. However, this behaviour could be obtained by employing a suitably informed

Bayesian prior on source locations, such as the Miami-Dade coastline shapefile that was used to ignore locations in the sea. Comparing the utility between a Bayesian prior and a set of pseudo-absences derived from a habitat suitability model was not tested in this chapter but could be explored in future work. Geographic heterogeneities have been considered in previous geographic profiling models (Mohler & Short 2012). This study however, is the first time a geographic profiling model utilises spatial heterogeneities whilst also being able to estimate multiple numbers of sources.

The analyses and results in this chapter have shown that a geographic profiling model that utilises count data can alter search strategies when intervening in cases of species invasion, outbreaks of infection or crime by making the distinction between evidence of absences in data and an absence of evidence. In doing so, search strategies produced move priority away from those locations containing absences to those containing no information at all; a substantial change over existing models that treat these areas with equal search priority. Additionally, the new model introduces the ability to estimate spatial dispersal and expected population size unique to each source location as well as the flexibility to a user to implement a spatial prior of their choosing. Different models should be used in differing circumstances dependent on the type of data to hand. The DPM model should be used when data are in point-pattern form (each location is associated with a single instance of crime, an invasive species or disease etc.) and the new model should be chosen when data consist of a list of sentinel site locations and associated counts (bioacoustics monitors, camera or pitfall traps etc.).

Chapter 4

Building a Prevalence Geographic Profiling Model

Abstract:

One of the many disciplines that geographic profiling has been applied to is epidemiology. Within this field, locations associated with a positive test for an infectious disease are used to target the spatial origin responsible for their transmission. Despite many successful applications, previous studies neglect to account for the number of individuals tested at each location in addition to those locations where zero positive tests were recorded. Previous chapters have already highlighted the importance of retaining such absence data. In this chapter I introduce another important innovation by incorporating the number of individuals tested at a location, in addition to those testing positive. Two locations may yield the same number of positive tests, but more resources should be focussed near the location with a higher percentage of positive tests. In this chapter I address this gap in geographic profiling by building a model that estimates hot spots of disease via a data type common to epidemiology. Additionally, I describe how the new model obtains an estimate for the probability of testing positive for a disease at a local scale. The new model is successfully applied to cases of malaria in Kasese, Uganda, leading to targeted interventions informing future surveys of the area. I also highlight the need to consider other environmental predictors of disease transmission when generating targeted interventions.

4.1 Introduction

Geographic profiling has successfully been used in many different disciplines, falling predominantly within ecology, epidemiology or criminology. Although the previous two chapters described the broader applications of a geographic profiling model based on spatial count data, the developments made focussed primarily within ecology; where sentinel sites could be considered as a baited, camera or proximity trap etc. Of the few geographic profiling studies that have been applied to problems in epidemiology, inferences were, once again, based upon a set of point-pattern data in place of other common data types in the field. Le Comber et al. (2011), for example, inferred breeding sites of the malarial vector *Anopheles sergentii* given the locations of household testing positive for the disease. This study also revisited John Snow’s classic epidemiological instance of the cholera outbreak in London. Another study applied geographic profiling to bovine tuberculosis by inferring badger setts given locations of infected cattle (Smith, Downs, Mitchell, Hayward, Fry & Le Comber 2015). When it comes to epidemiological studies, it should also be possible to make use of data comprising of counts at a set of locations, but for geographic profiling, this has yet to be done.

Within spatial epidemiology, there are many other data types with which inferences are made (McLafferty 2015). One such data type is a list of administrative boundaries/zones in which the number of positive tests for a particular disease are counted (Lawson & Clark 2002). Models then seek to estimate the disease’s prevalence, the probability of contracting said disease, within each zone (Staubach et al. 2002, Shaweno et al. 2017). Such models do possess a spatial component, in the form of a neighbouring effect, but understandably lack the complexity to estimate those parameters desired by geographic profiling, given the scale and aggregated form of the data.

A variety of models exist within epidemiology that focus on estimating parameters similar to geographic profiling. Diggle (1990) for example, described a method for estimating a *pre-specified point* where a particular phenomenon, such as cancer of the larynx, is expected to be observed more often within the vicinity of such a location. This idea resonates with the concepts of geographic profiling, where data are expected to be observed closer to a source location. Diggle’s approach only allowed for a single pre-specified point to contribute to incidences within a particular area, but Lawson (1995, 2000) built upon Diggle’s method and specified a procedure for estimating a set of *cluster centres* as well as their numbers. In these models, estimates are

conditioned on a set of locations each associated with a single instance of a disease but do not incorporate locations where individuals have been tested for a disease but yielded no positive results.

In this chapter I will focus on building a geographic profiling model that infers source locations and other parameters of interest via a set of prevalence data defined as follows. Within this chapter, *prevalence data* refers to a set of trial-site locations at which both the number of individuals testing positive for a disease and the total number tested have been recorded. An example of such prevalence data, compared to the spatial count data, modelled in previous chapters, is shown in Table 4.1.

Longitude	Latitude	Count	Longitude	Latitude	Tested	Positive
-0.0404	51.5239	1	-0.0404	51.5239	5	1
-0.1335	51.5245	1	-0.1335	51.5245	3	1
-0.0892	51.5238	2	-0.0892	51.5238	5	2
-0.0757	51.5050	0	-0.0757	51.5050	1	0
-0.0512	51.5070	3	-0.0512	51.5070	7	3
...

Table 4.1: An example of spatial count and prevalence data sets.

Although the aim of geographic profiling is to produce a search strategy for efficiently identifying source locations within a specified area, models in epidemiology consistently aim to estimate the true prevalence for a given spatial unit. That is, each pixel or administrative zone within a search area is associated with its own unique probability of contracting a disease (Besag et al. 1991). Hence, it is favourable for geographic profiling models to produce such risk maps since these will be familiar to epidemiologists and consequently, more useful when informing targeted interventions.

The structure of this chapter is as follows. I will begin by building a geographic profiling model that infers source locations, amongst other parameters of interest, via a set of prevalence data described above. Once the model is built, it will then be applied to a set of prevalence data that tested households in Kasese, Uganda for the malarial *Plasmodium falciparum*.

Many regions of Uganda experience devastating levels of malarial prevalence, in line with seasonal rainfall each year (Yeka et al. 2012). Whilst prevalence levels within Kasese were estimated to be over 60% around 2012 (Yeka et al. 2012), the specific area covered by the data analysed in this chapter recently recorded prevalence around 18% (Ugandan Ministry of Health

2015, Boyce et al. 2018). Early diagnostics combined with targeted interventions are vital for the eradication of diseases that cluster in space, such as malaria (Marsh 2010, The malERA Consultative Group on Diagnoses 2011, Bousema et al. 2012), indicating that geographic profiling can be an incredibly valuable tool for these kind of epidemiological problems.

In addition to the analysis, I will describe a few alternative processes for judging the most suitable number of source locations that best describe the data. Although the deviance information criterion (DIC) from Spiegelhalter et al. (2014) was useful for a quick comparison between models, previous chapters have demonstrated the metric is sensitive to different priors. Hence, I will describe and implement two alternatives and compare them to the output of models determined as most suitable via the DIC.

4.2 Model derivation

A binomial geographic profiling model, much like the Poisson model, starts by assuming the existence a set of source locations $\boldsymbol{\mu}_k$, for $k \in 1 : K$, distributed over space via some spatial prior \mathcal{F} on a two-dimensional grid of cells. Also distributed across space is a set of trial sites \boldsymbol{s}_j , for $j \in 1 : m$, at which N_j individuals are tested for a disease. Let $\phi_{i,j} \in \{0, 1\}$ denote the positive (one) or negative (zero) outcome when testing individual i at trial site j for the disease of interest. The total number of individuals testing positive at site j is then obtained by summing over the index i , denoted

$$n_j = \sum_{i=1}^{N_j} \phi_{i,j}. \quad (4.1)$$

The probability p_j of testing positive for the disease at trial site j is derived by considering the contribution from each source, broken into two parts. The first component comes from the distance between source and trial site. Working under the principles of geographic profiling, proximity to a source location should have an impact on the probability of testing positive for the disease. This idea is reflected by the summed heights of the trial site \boldsymbol{s}_j on the bivariate normals centred on each source location:

$$\sum_{k=1}^K f_{BN}(\boldsymbol{s}_j | \boldsymbol{\mu}_k, \mathbf{I}_2 \sigma_k^2). \quad (4.2)$$

The second component contributing to the probability of infection derives from the total number of infections associated with each source location λ_k , since the more infections attributed to a source, the higher the probability of infection within its vicinity. Pooling these contributions leads to the

expression:

$$\Phi_j = \sum_{k=1}^K \lambda_k \cdot f_{BN}(\mathbf{s}_j | \boldsymbol{\mu}_k, \mathbf{I}_2 \sigma_k^2). \quad (4.3)$$

The domain of Φ_j lies in the range of non-negative real numbers, however, p_j is required to lie in the interval between zero and one, given it is the probability of infection. Hence the following the transformation:

$$\Phi_j \mapsto p_j = \frac{\Phi_j}{1 + \Phi_j} \quad (4.4)$$

is used to ensure the mapping from $\Phi_j \in [0, \infty)$ becomes a value between zero and one. The full binomial model is as follows:

Likelihood:

$$\phi_{i,j} \sim \text{Bernoulli}(p_j), \quad i \in 1 : N_j, \quad j \in 1 : m, \quad (4.5)$$

$$n_j = \sum_{i=1}^{N_j} \phi_{i,j}, \quad j \in 1 : m. \quad (4.6)$$

Priors:

$$\boldsymbol{\mu}_k \sim \mathcal{F}, \quad k \in 1 : K, \quad (4.7)$$

$$\sigma_k \sim \text{Log-Normal}(\gamma, \delta), \quad k \in 1 : K, \quad (4.8)$$

$$\lambda_k \sim \text{Gamma}\left(\frac{\zeta}{K^2}, \frac{\eta}{K}\right) \quad k \in 1 : K. \quad (4.9)$$

The priors of this binomial finite mixture model are chosen in a similar manner to the models of the previous chapters. The prior on the expected number of infected individuals is slightly altered such that this number is governed by a global mean controlling the expected number of infections over the total area λ . This attribute follows from

$$\lambda_1 + \lambda_2 + \lambda_3 + \dots + \lambda_K = \lambda \sim \text{Gamma}\left(K \cdot \left(\frac{\zeta}{K^2}\right), \frac{\eta}{K}\right) \quad (4.10)$$

$$= \text{Gamma}\left(\frac{\zeta}{K}, \frac{\eta}{K}\right), \quad (4.11)$$

where $\frac{\zeta}{\eta}$ is the global mean and $\frac{\zeta}{\eta^2} \cdot K$ is the global variance that linearly increases with K to reflect that the uncertainty in the number of infections should increase with the number of source locations.

A binomial geographic profiling model requires the evaluation of a likelihood for observing the prevalence data set in Table 4.1. The probability of observing this data is governed by the binomial density:

$$\Pr(n_j|p_j) = \binom{N_j}{n_j} p_j^{n_j} (1 - p_j)^{N_j - n_j}, \quad (4.12)$$

hence the probability of observing n_j positive tests at sentinel site j is:

$$\Pr(n_j|p_j) = \binom{N_j}{n_j} p_j^{n_j} (1 - p_j)^{N_j - n_j} \quad (4.13)$$

$$= \frac{N_j!}{n_j!(N_j - n_j)!} \left(\frac{\Phi_j}{1 + \Phi_j} \right)^{n_j} \left(\frac{1}{1 + \Phi_j} \right)^{N_j - n_j}. \quad (4.14)$$

To obtain the full likelihood of observing the entire set of prevalence data \mathbf{n} , conditional on the sources and their attributes, the probability of observing each n_j are multiplied together (assuming independence between sites) to obtain

$$\Pr(\mathbf{n}|\mathbf{p}) = \prod_{j=1}^m \frac{N_j!}{n_j!(N_j - n_j)!} \left(\frac{\Phi_j}{1 + \Phi_j} \right)^{n_j} \left(\frac{1}{1 + \Phi_j} \right)^{N_j - n_j}. \quad (4.15)$$

4.3 Data and model settings

To test the new binomial finite mixture model for geographic profiling, I analysed a set of prevalence data across five villages in Kasese, Uganda. The data consisted of 485 individuals tested for malaria (the parasitic *Plasmodium falciparum* spread via *Anopheline* mosquitoes (Okello et al. 2006)) at 455 unique locations in January 2020. At each location one to five individuals were tested and of these 455 locations, 97% tested 1 individual. There were a total of 436 negative test results at 406 of the 455 locations and of the remaining 49 test locations, a single individual tested positive for malaria. The prevalence data analysed in this chapter can be seen in Figure 4.1.

When running the binomial model, priors were set as follows. The prior on source locations (\mathcal{F}) was set to uniform and its extent was governed by the data plus a 10% guard rail (Le Comber et al. 2006), covering an area of 61 km². The lower left quadrant of this uniform prior contained a large, un-sampled area, far from any positive data hence its prior probability mass was set to zero. The northern-most negative data point, isolated from the rest of the data (see Figure 4.1), was removed for this analysis given that a negative data point far away from the positive data is not expected to greatly effect estimates on source locations.

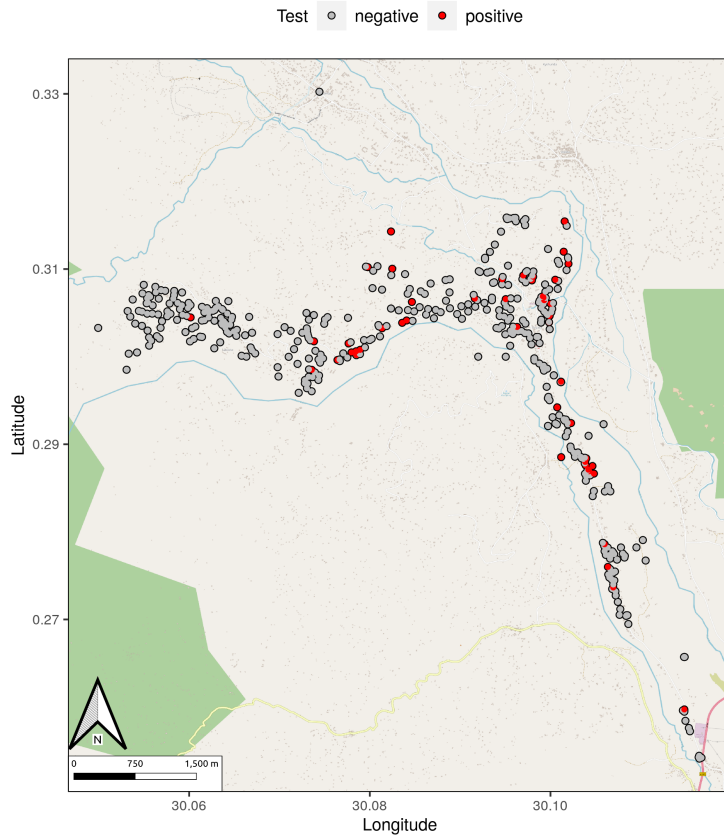


Figure 4.1: The prevalence data set in Uganda. This map was created using the *OpenStreetMap.de* layer via QGIS.org (2021).

Once again, a log-normal prior was utilised for the dispersal parameter σ_k . In this analysis, it was assumed that the dispersal parameter was the same for each source ($\sigma_1 = \sigma_2 = \dots = \sigma_k$). The log-normal prior was set with a mean of 0.5 (equivalent to 500 metres) and a standard deviation of 0.25. In the previous chapter, the prior on σ_k was set to reflect the tendencies of mosquito dispersal to lie anywhere between a couple of hundred metres and 5 km (Service & Place 1997, Carter et al. 2000, Gorrochotegui-Escalante et al. 2000, Le Comber et al. 2011, Verity et al. 2014). In this particular example, the prior on σ_k was restricted to the lower end of this interval as the data were situated in between bodies of water (the rivers Esha, Sebwe and Mubuku), where the maximum distance between a positive data point and one of these water bodies was approximately 500 metres.

To account for spatial heterogeneity in the expected number of infections, an independent λ_k value was estimated for each source. The prior on these values, shown in Equation 4.9, was set such that the global mean (the mean of the prior on λ) was 100 and the standard deviation was $10 \cdot \sqrt{K}$. This prior ensured that the mean and standard deviation of each mixture component's expected number of infections was $\frac{100}{K}$ and 10 respectively, whilst also imposing extra uncertainty

in the total expected number of infections when adding mixture components.

When estimating the number of source locations, K , the model went in search of up to fifteen sources. As shown in the previous chapter, the deviance information criterion can lead to various conclusions based upon different priors. Hence two alternatives were used in addition to the DIC to deal with the uncertainty in model selection. The first relies on an allocation matrix calculated at each MCMC iteration to obtain the probability that a positive test originated from a particular source location. This probability was calculated by considering the contribution to the infection rate from a particular source (Φ_j^i), and dividing it by the total contributions (Φ_j):

$$\frac{\Phi_j^i}{\Phi_j} = \frac{\lambda_i \cdot f_{BN}(\mathbf{s}_j | \boldsymbol{\mu}_i, \mathbf{I}_2 \sigma_i^2)}{\sum_{k=1}^K \lambda_k \cdot f_{BN}(\mathbf{s}_j | \boldsymbol{\mu}_k, \mathbf{I}_2 \sigma_k^2)}. \quad (4.16)$$

At each iteration of the MCMC algorithm, positive data were assigned to source locations by sampling from the allocation matrix. Then, the number of realised sources was obtained by counting up the number of sources that had positive data assigned to them. Posterior draws for any source not producing data were discarded, resulting in a density responsible for generating the data.

The second alternative to the vanilla DIC model comparison metric produced results by following a weighted model averaging protocol similarly to O'Leary (2010*b*). This procedure required comparing DIC values to the minimum to calculate a set of Δ_r values, where

$$\Delta_r = \text{DIC}_r - \{\text{DIC}_{min}\}, \quad (4.17)$$

and the weight for each model was then obtained via the expression:

$$w_k = \frac{e^{-0.5\Delta_r}}{\sum_{r=1}^K e^{-0.5\Delta_r}}. \quad (4.18)$$

Posterior densities for each parameter of interest were then produced by averaging over model estimates weighted by each w_k .

For each model type a geographic profile was produced in order to inform search strategies for intervening in cases of malaria. However, to conform to the conventions of other epidemiological models, a set of risk maps were also produced such that the prevalence for each cell of the spatial domain was stated. These prevalence values were calculated at each iteration of the MCMC algorithm via Equation 4.3 and Equation 4.4 given the values of $\{\boldsymbol{\mu}_k\}$, σ_k and $\{\lambda_k\}$. The final

prevalence value for each cell was then obtained by averaging over the values across sampling iterations.

MCMC parameters were set as follows. The burn-in and sampling phases of the MCMC were set to 5×10^4 iterations where convergence was checked during the burn-in phase for every 2.5×10^3 iterations. The Metropolis-Hastings coupling protocol was once again utilised to ensure healthy MCMC mixing. The number of heated rungs ranged from 36 to 45 ensuring proposed swaps between heated chains were accepted in at least 75% of cases. The model described in this chapter was implemented in R via the *silverblaze* package. Detailed tutorials for implementing the model are available at <https://michael-stevens-27.github.io/silverblaze/>.

4.4 Results

Results for the binomial geographic profiling model can be seen in Figure 4.2 and Figure 4.3. Figure 4.2 illustrates the two metrics that were used to determine the most suitable number of source locations, K ; via the DIC metric and via sampling from the allocation matrix. DIC values ranged from 319.00 to 340.80 with a mean of 324.4 and standard deviation 5.87. Following the protocol of the previous chapter, the DIC judged that six source locations best described the data. Following the alternative approach, the model allocated data to the number of sources it was searching for when the model was set to search for one to six sources (see Figure 4.2b). However, when searching for greater than six source locations, the model rarely allocated data to the specific number of sources it was searching for.

Geographic profiles produced via the DIC metric ($K = 6$), via the realised-source metric ($K = 15$) and by combining models ($K = 1 : 15$) can be seen in Figure 4.3a, b and c respectively. The DIC and combined-model profiles placed some search priority near areas containing large numbers of positive data whilst also avoiding those areas with high numbers of negative data. Additionally, these two profiles placed search priority around areas with no information. The realised profile also avoided areas with high numbers of negative data but focussed much of its prioritisation on those areas with positive data. The realised profile avoided areas with no information.

Risk maps for these models can be seen in Figure 4.3d, e and f. Prevalence values for each map were as follows. For the DIC risk map, values ranged from 0.34% to 77.85% with a mean of 28.11%, for the realised risk map, values ranged from 0.53% to 54.06% with a mean of

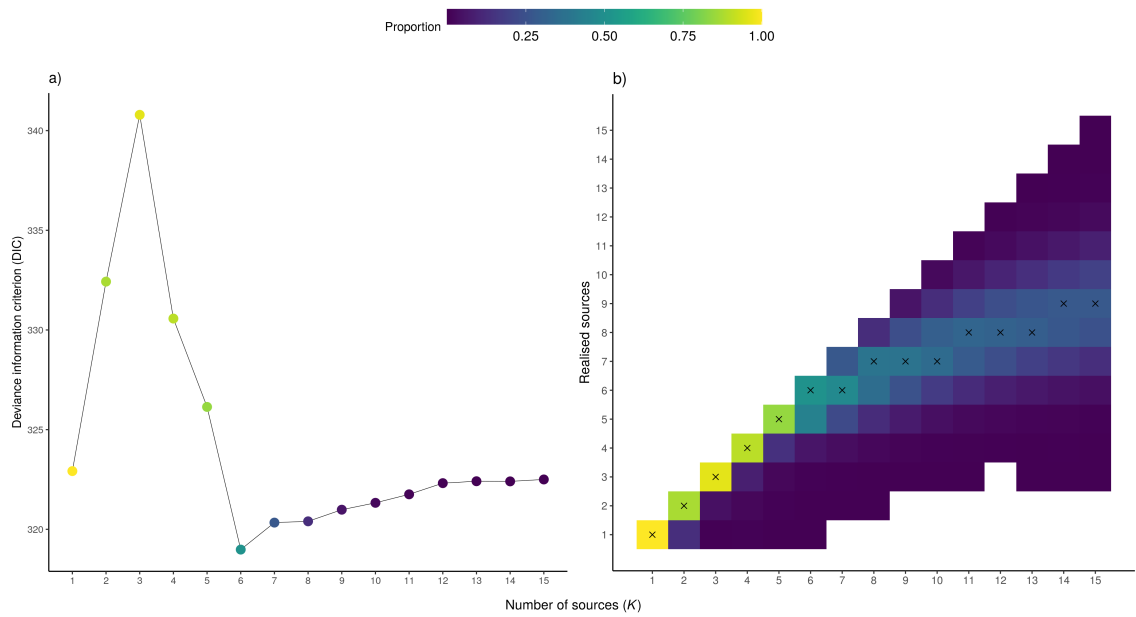


Figure 4.2: Two metrics for judging the number of mixture components K that best describe the prevalence data: a) the conventional deviance information criterion, where each DIC value is coloured based on the proportion of samples allocating data to that given K model and b) the alternative allocation sampling, where the realised number of sources with the maximum proportion of allocation are marked with a cross.

16.73% and for the combined risk map, values were between 1.33% to 64.71% with a mean of 25.45%. Once again, the DIC and combined-model risk maps determined the highest prevalence in areas near large numbers of positive data whilst also prompting risk in areas with no information. Similarly to the geoprofile, the risk map produced by considering only those source locations generating data prompted the highest prevalence near large areas of positive data.

Estimates for the spatial dispersal of infections, σ , and expected number infections for the total area, λ , can be seen in Table 4.2. Estimates for the dispersal parameter σ remained

Parameter		σ (km)		λ	
Model metric	K	mean	sd	mean	sd
DIC	6	0.51	0.18	76.68	21.75
Realised	15	0.46	0.20	30.51	19.42
Combined	1:15	0.46	0.19	63.06	24.27

Table 4.2: Parameter estimates for σ and λ under different models.

consistent across model types, however, the expected number of infected individuals across the search area (λ) in the case of the realised-source model were approximately half that of the DIC and combined models.

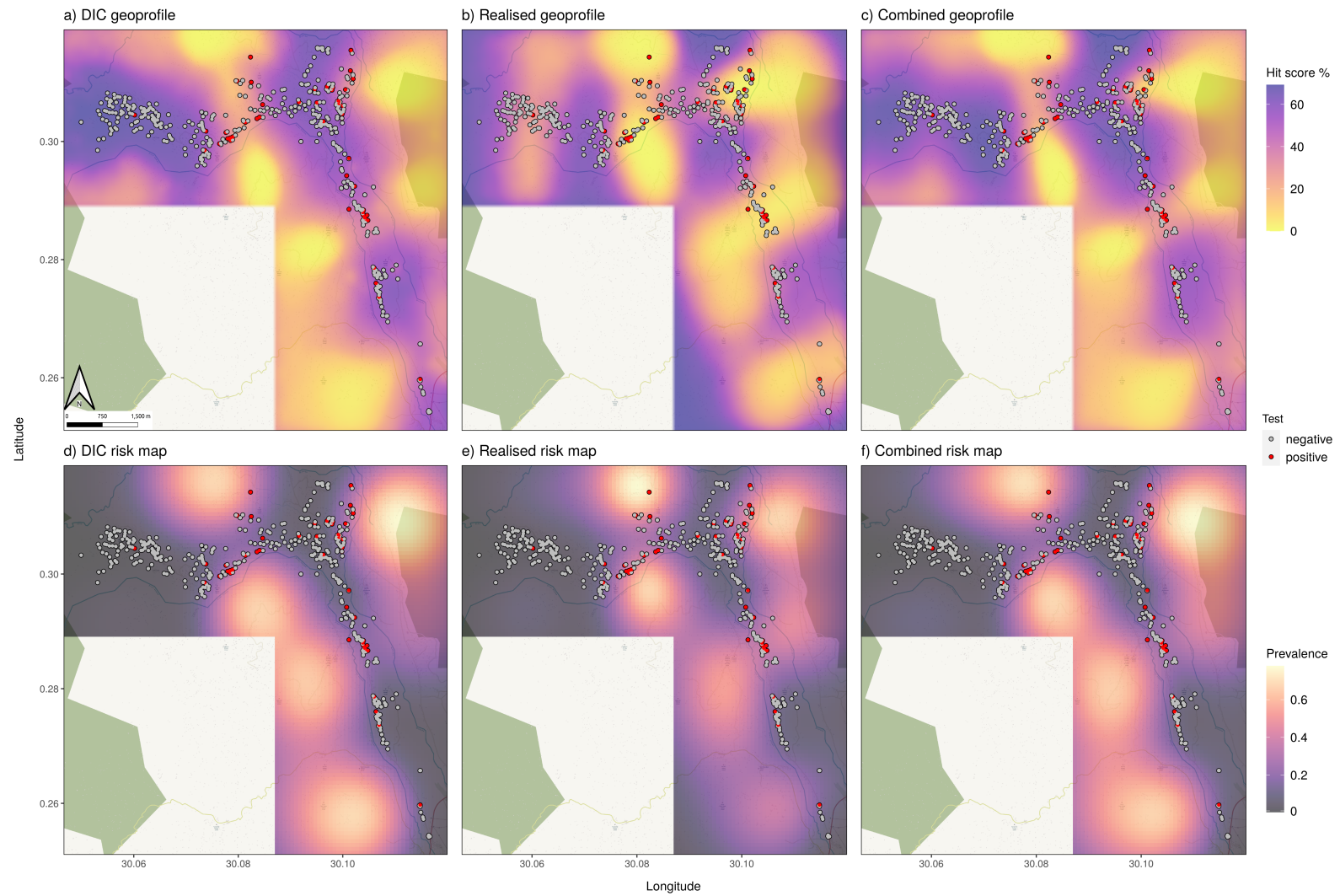


Figure 4.3: The geographic profiles and risk maps produced for each model type metric: the DIC (a and d), realised sources (b and e) and the combined DIC (c and f).

4.5 Discussion and conclusions

In this chapter, I have built and applied the first geographic profiling model to make inferences via prevalence data; a list of trial sites each recording the number of individuals tested for a disease in addition to the number testing positive. I demonstrated how the new binomial finite mixture model estimates those parameters common to geographic profiling, such as source locations, whilst also describing the process for obtaining a parameter common to epidemiology, a local prevalence.

In the analysis of the Kasese data set, model success could not be measured via the conventional hit score metric, since no source locations were available. However, model estimates for prevalence can be compared to existing records. Recent surveys placed malarial prevalence in the mid-western region of Uganda at around 18% (Ugandan Ministry of Health 2015). Although this value fell within the estimates of each model type, the intervals were broad, ranging from 0% to 77%. This result is likely due to the scale at which the model is operating whereas the existing estimates cover entire administrative zones. Further work is needed to verify the accuracy of the results from these three models.

This chapter provided some alternative metrics for model choice in place of the deviance information criterion. The first alternative, the realised source model, obtained the number of sources that best described the data by considering only those sources with which positive data were allocated. This process resonates with that of the Dirichlet process mixture model - where an infinite number of source locations are assumed to exist, but only a finite number are realised by the data. The geographic profile in Figure 4.3b illustrates that the realised model behaved as it should since areas of high priority were focussed around positive data. However, this process no longer placed priority in areas with no information, avoiding the distinction between those locations that generate no positive data and those that were not sampled. An alternative to model comparison consisted of combining each K model weighted based upon their respective DIC values. The geographic profile obtained by combining models (Figure 4.3c) produced a similar search strategy to the profile choosing a single K via the DIC (Figure 4.3a). Both methods placed high priority around positive data and areas with no information, whilst avoiding areas with negative data. It is entirely possible that this data set suffered from the same over-dispersion quality observed in the previous chapter and as such, the binomial model in this chapter could be altered to accommodate over-dispersed prevalence data.

Estimates for the spatial dispersal of infections σ were consistent across model types. This result was likely due to the informative prior on σ , as described in the methods sections, conforming to the beliefs of mosquito dispersal in Anopheline species that can reach up to distances of 5 km. However, the spatial structure of the landscape, i.e. proximity of positive data to water bodies (Ramasamy et al. 2011), meant that this estimate was reduced to the lower end of the prior interval used in previous chapters for mosquito dispersal. Estimates for the expected number of infected individuals in the search area, λ , were consistent for the DIC and combined DIC models. However, the mean of the posterior density for λ under the realised model was around half the estimates for the other two models. This result was likely due to the model failing to allocate positive data to source locations with larger λ_k values.

Ultimately then, which of the three processes should a user implement: the DIC metric, the realised source model or the combine DIC model? Via the realised source model, a user need only run the model once with a large value of K . Then the uncertainty in the most suitable value of K can be quantified in the manner shown in Figure 4.2b. Although this process may save computation by requiring only one execution of the MCMC algorithm, too large a K model may be cumbersome to evaluate. Hence a trade off exists between computation time and capturing the distribution in Figure 4.2b. Additionally, the profiles and risk maps indicate that interventions should be targeted nearer positive data, a more intuitive result compared to the other models. However, the realised model loses its ability to distinguish between those areas that were sampled and yielded no positive results and those with no information at all. For the combine DIC model, a user bypasses the tricky requirement of selecting a single, best model to describe the data, such as in chapter three, where the DIC was used to choose a single model, despite many models sharing similar values. Although this alternative method deals with the uncertainty in model choice, the results from the combine model were similar to that of the vanilla DIC method. To a user, it will be far more feasible to specify a range of K prior to running model, hence I recommend that the combined DIC method is chosen over the other two.

The new model considers only a hand full of predictors when estimating malarial prevalence. Though proximity to a breeding site is a key driver of malaria transmission, other environmental and biological factors such as altitude, temperature and precipitation also need to be considered when modelling outbreaks of an infectious disease (Lindsay & Martens 1998, Protopopoff et al. 2009). The area containing the Kasese data set is geographically diverse. It is spatially fragmented by multiple bodies of water running through steep valleys that reach heights of

2 km (Boyce et al. 2018, 2019). There are multiple avenues with which predictors such as altitude and proximity to water could be integrated into the model. The simplest route would be to implement a spatial prior favouring source locations within a certain distance of a river's flood catchment and or favouring those altitudes that are ideal for mosquito breeding. This chapter detailed a base line model for estimating breeding locations for vectors of malaria from prevalence data. Hence future work should focus on developing geographic profiling models for epidemiology by incorporating predictors beyond the conventional spatial proximity to a source.

Geographic profiling models share the implicit notion of an individual's fidelity to a central source location. As previously remarked, this location could be a criminal's home or work place or an invasive species' nesting site. Even in epidemiology, there are examples where an individual can be directly linked to a source of disease, such as John Snow's example of cholera caught from a particular water pump in London (Le Comber et al. 2011). For the Kasese example, there is an extra layer of complexity bypassing the direct fidelity between individual and source of disease. This complexity derives from the process that starts with a mosquito emerging from a water body and ends with the household of the individual testing positive for said disease. Future work could therefore focus on modelling this underlying process fully.

This chapter was successful in developing geographic profiling's place within epidemiology by making inferences via a common data type in the field. It also demonstrated its ability to produce targeted interventions when dealing with diseases that cluster in space such as malaria. Future work could focus on further validation of the model in cases where source locations are explicitly known to a user. Additionally, the model built in this chapter should be considered as the foundation for further work, where developments should start by considering other key predictors for disease prevalence.

Chapter 5

A Gaussian Finite Mixture Model Beyond the Normal

Abstract:

A fundamental principle of Rossmo's geographic profiling algorithm is that a serial criminal will not commit crimes far from home due to the cost incurred by travelling such large distances. This phenomenon was encoded into early geographic profiling models via Rossmo's distance decay function. In recent Bayesian models, this function takes the form of a normal distribution to conform to the laws of probability. This distribution was in many ways a convenient first choice; it was both easily implemented into new algorithms and was shown to capture spatial phenomenon across disciplines. There is, however, overwhelming evidence that other probability distributions would be more appropriate for many of the extensions of geographic profiling in biology. In ecology, studies show that it is common to observe an organism travelling large distances and as such, a distribution with a heavier tail than that of the normal should be utilised. In this chapter, I implement two alternative distributions for describing distances travelled by an organism when larger distances are more likely to occur. I compare each of these three distributions to judge which best describe a set of data previously analysed by earlier geographic profiling models and determine that a heavier tailed distribution is favourable over the normal. Generally I conclude that these alternative distributions allow for more reliable inference, in practical problems in which the profiling approach is used to make decisions for deploying limited resources for targeted interventions.

5.1 Introduction

One of the fundamental principles of Rossmo’s geographic profiling is that offenders do not want to commit crimes far from an anchor point (Rossmo 2000). This behaviour is due to the costly nature of travelling large distances. This common trope of criminal activity was encoded in the distance decay function (see Figure 1.1) used within Rossmo’s criminal geographic targeting (CGT) algorithm. When Rossmo’s method migrated to a Bayesian framework in O’Leary (2009), this distance-decay function needed to be replaced with a distribution following the laws of probability, such that it integrated to unity.

In his examples, O’Leary made use of a normal distribution when estimating a source location and similarly, a bivariate normal was adopted when the Dirichlet process mixture (DPM) model was first constructed in Verity et al. (2014). Subsequently, all studies that have implemented the DPM model for geographic profiling (Faulkner et al. 2015, Smith, Downs, Mitchell, Hayward, Fry & Le Comber 2015, Faulkner et al. 2016, Hauge et al. 2016, Faulkner et al. 2018, Struebig et al. 2018, Tench 2018, Aygin et al. 2019, Heald et al. 2019, Cerri et al. 2020, Gray 2020, Stevens, Ray, Faulkner & Le Comber 2020) have assumed that a bivariate normal distribution governs the generation of data around a source location. A desirable feature of the bivariate normal that made it an attractive choice was its self conjugacy, making it immediately suitable for use within a Gibbs sampling algorithm. A consequence of migrating to a Metropolis-Hastings algorithm, as described in previous chapters, allows for the condition of self conjugacy to be relaxed and thus any desired distribution can be adopted.

The topic of the most suitable distribution, referred to as a *dispersal kernel* from here on, that governs the distance travelled by, for example, a criminal to commit a crime or a species to forage for food is well documented across disciplines (Johnson 2014). When building his model, O’Leary specified many alternatives to the normal distribution when analysing a series of crimes (O’Leary 2009) and also pointed to platforms that can implement various dispersal kernels (Levine & Block 2011). Ultimately, the choice of kernel is dependent on a variety of factors, such as the type of crime being committed, the time of day and the value to an offender (Capone & Nichols 1975, Kent et al. 2006). Hence, it is desirable to have an approach capable of drawing from different distributions depending on the problem to hand.

Within ecology, a cornucopia of different dispersal distributions are used to describe the

distances travelled by a species from a central source location. Nathan et al. (2012) provided an extensive review, listing many kernels as well as their wide pool of applications in ecology. Mammals (Krkošek et al. 2007), insects (Chapman et al. 2007), fish (Coombs & Rodríguez 2007), birds (Van Houtan et al. 2007), seeds (Sagnard et al. 2007, Schurr et al. 2008) and pollen (Austerlitz et al. 2004) have all been demonstrated to exhibit different dispersal behaviours encoded by a variety of kernels. The most appropriate choice of dispersal kernel will also be dependent upon species-specific factors, such as their ability to disperse, parental influence, habitat quality, physical environment and landscape structure (Matthysen 2012).

In addition to listing the wide collection of dispersal kernels, Nathan et al. (2012) highlighted an important feature of dispersal behaviour within ecology. That is, there are a variety of studies pointing towards the deviation of dispersal distances from a Gaussian kernel, such as the normal distribution, to a kernel with a much heavier tail or higher *kurtosis*. Within ecology, it is common to find that data are best described by a heavy-tailed distribution to reflect that distances travelled by a species from a central source location often produce outliers; a collection of small distances travelled with the occasional larger step.

The probability of observing a dispersal distance is defined over a two-dimensional domain, however, many studies (such as Alonso et al. (1998) and Forero et al. (2002)), report this distribution of dispersal distances by transforming from a two-dimensional density to a one-dimensional histogram. This transformation often exhibits a geometrical mistake leading to erroneous inference about the true dispersal behaviour of an organism. This problem was touched upon in Stevenson (2015), but lacked a systematic literature search for those studies exhibiting the error in question. In appendix A.1, I outline a systematic approach and provide evidence that this error is still being made in recent ecological literature.

Despite the evidence indicating that heavy-tailed distributions are better suited when modelling dispersal distances within ecology, nearly all geographic profiling studies have chosen a bivariate normal distribution. Hence, in this chapter I will describe a finite mixture model for geographic profiling, that expands upon the model described in the introduction (Equation 1.13 - Equation 1.15), by implementing two alternative dispersal kernels alongside the conventional normal density. This chapter will take a step back from the count and prevalence data sets analysed in previous chapters to return to inferences based on point-pattern data, the form of which is shown in Table 5.1.

Longitude	Latitude
-0.0404	51.5239
-0.0892	51.5238
-0.0175	51.5223
-0.0190	51.5239
-0.0165	51.5110
...	...

Table 5.1: An example of spatial point-pattern data analysed in this chapter.

Once built, I will apply the finite mixture model to a well-known data set in the geographic profiling literature. Data consisted of households testing positive for malaria in Cairo, Egypt. This case study was originally analysed in Le Comber et al. (2011) by the criminal geographic targeting algorithm and went on to be compared to O’Leary’s model and the DPM model in Verity et al. (2014).

5.2 Methods

Here I will walk the reader through the two alternative dispersal distributions, the Laplace and Cauchy, that I am introducing for modelling point-pattern data for geographic profiling. I will then describe the model in full and briefly touch upon the suitable proposal distributions required for the MCMC algorithm.

5.2.1 The Laplace and Cauchy distributions

The Laplace distribution, the first alternative used in the model, has been fully described in Kotz et al. (2001), but briefly, it was originally defined as the first law of errors. That is, the frequency of errors or, in the context of this chapter, dispersal distances from a centre were described using an exponential distribution with a linear exponent. The second law, more commonly known as Gauss’ law, refers instead to a quadratic exponent, corresponding to the well-known normal distribution.

The univariate density of the Laplace distribution can be written as

$$f(x; \mu, \sigma) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}}, \quad (5.1)$$

where, similarly to the normal density, the distribution is governed by a central mean μ and scale parameter σ . The bivariate form of the Laplace distribution that was implemented in this chapter was adapted (assuming no correlation across dimensions) from Kotz et al. (2001) as follows

$$\begin{aligned} f(\mathbf{x}; \boldsymbol{\mu}, \sigma) &= \frac{1}{\pi\sigma^2} K_0 \left(\sqrt{2} \left[\frac{\sqrt{(x - \mu_x)^2 + (y - \mu_y)^2}}{\sigma} \right] \right) \\ &= \frac{1}{\pi\sigma^2} K_0 \left(\sqrt{2} \left[\frac{d}{\sigma} \right] \right), \end{aligned} \quad (5.2)$$

where d is the standard Euclidean distance and K_0 represents the modified Bessel function of the third kind with order zero.

The mean of the bivariate Laplace distribution is given by its centre $\boldsymbol{\mu} = (\mu_x, \mu_y)$ with variance equal to $2\sigma^2$, making it easily comparable with the bivariate normal. Additionally, the kurtosis or *heavy tailed-ness* of the bivariate Laplace is twice that of the normal, making it an appropriate candidate to use as an alternative kernel for encoding dispersal distances within ecology.

The other distribution chosen for use in place of the normal density is the bivariate Cauchy. This distribution is also popular within ecology for being able to explain anomalously large dispersal distances travelled by an organism (Paradis et al. 2002). The univariate form of the Cauchy distribution is given by

$$f(x; \mu, \sigma) = \frac{1}{\pi\sigma} \left(\frac{\sigma^2}{(x - \mu)^2 + \sigma^2} \right). \quad (5.3)$$

The Cauchy distribution's heavy tail can accommodate large dispersal distances, but does pose some disadvantages as its mean and variance are undefined. It is instead parameterised by its median μ and half median at half height σ , which will be referred to as its scale parameter. This definition is a bit of a mouth full but is simply referring to half the width of the distribution at half the maximum of the distribution. Considering the two dimensional version of the multivariate Cauchy in Lee et al. (2014) leads to the bivariate form:

$$f(\mathbf{x}; \boldsymbol{\mu}, \sigma) = \frac{1}{2\pi\sigma} \frac{\sigma^{\frac{3}{2}}}{[\sigma + (x - \mu_x)^2 + (y - \mu_y)^2]^{\frac{3}{2}}} = \frac{1}{2\pi} \frac{\sigma^{\frac{1}{2}}}{[\sigma + d^2]^{\frac{3}{2}}}, \quad (5.4)$$

where d is the standard Euclidean distance between its centre and a data point.

Figure 5.1 illustrates the differing forms of each distribution. Visibly, both the Cauchy and

the Laplace distributions have similar, if not larger densities around their centres and tails. Additionally, the density of each new kernel is lower around medium distances from the centre, compared to the normal.

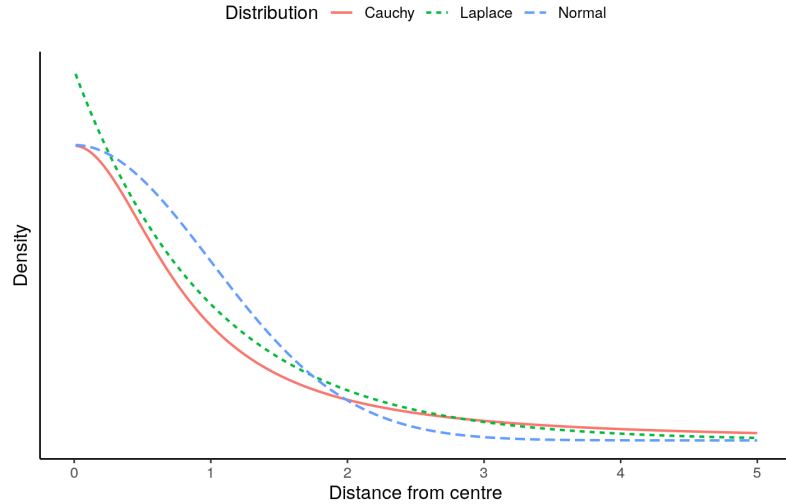


Figure 5.1: Probability densities of the univariate Cauchy, Laplace and normal distributions.

5.2.2 Model derivation

The model used in this chapter followed the same structure as the finite mixture model in the introduction (Equation 1.13 - Equation 1.15 from Aitkin (2001) and Kaiser et al. (2002)). Recall the model as:

Likelihood:

$$c_i \sim \text{Multinomial}(\omega_1, \omega_2, \dots, \omega_k), \quad i \in 1 : N,$$

$$\mathbf{x}_i \sim f(\theta_{c_i}), \quad i \in 1 : N,$$

Priors:

$$\theta_{c_i} \sim \mathcal{F},$$

$$\omega_{c_i} \sim \text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_k), \quad k \in 1 : K,$$

where $f(\cdot)$ represents the new dispersal kernels that will be implemented within the model. Once again, following the structure of Aitkin (2001), the probability density of observing a data point \mathbf{x}_i is given by the weighted contributions of each density centred on a particular source location

(Equation 1.16):

$$\sum_{c_i=1}^K \omega_{c_i} f(\mathbf{x}_i; \boldsymbol{\mu}_{c_i}, \sigma_{c_i}).$$

Expanding to a set of point-pattern data $\mathbf{n} = \{\mathbf{x}_1, \mathbf{x}_2 \cdots \mathbf{x}_n\}$, the probability of observing \mathbf{n} conditional on the sources and their attributes, is simply the product of probabilities of individual data points (assuming independence between them, from Equation 1.17).

$$L(\mathbf{n} | \{\boldsymbol{\mu}_{c_i}\}, \{\sigma_{c_i}\}, \{\omega_{c_i}\}) = \prod_{i=1}^n \sum_{c_i=1}^K \omega_{c_i} \cdot f(\mathbf{x}_i | \boldsymbol{\mu}_{c_i}, \sigma_{c_i}).$$

This model is implemented in the *silverblaze* package for estimating the set of unknown parameters $\{\boldsymbol{\mu}_{c_i}\}$, $\{\sigma_{c_i}\}$ and $\{\omega_{c_i}\}$. Before I describe the data analysed in this chapter, I need to mention some adjustments to the current proposal distributions used within the MCMC algorithm to cater to the developments described above.

5.2.3 Proposal distributions

It is well understood that the choice of proposal distribution is critical to ensure that an MCMC algorithm properly explores the target density of interest (Rosenthal 2011, Thawornwattana et al. 2018). In previous chapters, a bivariate normal proposal was used to produce candidates for new source locations but as a consequence of altering the model to fit a different kind of dispersal kernel, the MCMC algorithm will also require a shift to different proposal distributions.

Explicitly, the form of the proposal density for source locations within the MCMC algorithm was chosen to match the dispersal kernel being fitted to the data. Drawing candidate values from a bivariate normal distribution is straight forward since this density can be decomposed into two univariate proposals. The methods used for proposing from bivariate Laplace and Cauchy distributions were as follows. For the Laplace (Kotz et al. 2001, Daojing Wang et al. 2009), the candidate value for a new source $\boldsymbol{\mu}' = (\mu'_x, \mu'_y)$ location was obtained by calculating

$$\mu'_x = \mu_x + \sqrt{v} \cdot \tau_x \quad \text{and} \quad \mu'_y = \mu_y + \sqrt{v} \cdot \tau_y, \quad (5.5)$$

where v was drawn from an exponential distribution with rate one and τ_x and τ_y were both drawn from an independent univariate normal distribution centred at zero with standard deviation equal to the current proposal scale. Similarly, a new source location $\boldsymbol{\mu}'$ was proposed via the Cauchy

distribution by calculating

$$\mu'_x = \mu_x + \frac{\tau_x}{\sqrt{\chi}} \quad \text{and} \quad \mu'_y = \mu_y + \frac{\tau_x}{\sqrt{\chi}}, \quad (5.6)$$

where χ was drawn from a chi-squared distribution with one degree of freedom. Following the protocol of previous chapters, the proposal's scales were adapted using the Robbins-Monro process to optimise the mixing of the MCMC algorithm.

On the topic of optimally scaling proposal distributions, the protocol from Garthwaite et al. (2016) is straight forward for the three densities described above. The scale of the proposal increases and decreases depending on whether a candidate value was accepted or rejected. This chapter introduces a set of source weights $\{\omega_k\}$ that need estimating. Since a proposal for this set of weights must operate over the domain $[0, 1]$, such that their sum is also one, a natural choice is the Dirichlet distribution.

Explicitly, a new set of weights $\{\omega'_k\}$ can be drawn from a Dirichlet distribution given the current weights $\{\omega_k\}$, i.e.

$$\{\omega'_k\} \sim \text{Dirichlet}(\{\omega_k\}), \quad (5.7)$$

where the density of the Dirichlet distribution (Ng et al. 2011) is given by:

$$g(\{\omega'_k\}|\{\omega_k\}) = \Gamma\left(\sum_{k=1}^K \omega_k\right) \prod_{k=1}^K \frac{\omega_k^{\omega'_k-1}}{\Gamma(\omega_k)}. \quad (5.8)$$

Note that the mean and variance for each of the ω'_i given the set of existing concentration parameters, $\{\omega_i\}$, are

$$\text{E}[\omega'_i] = \frac{\omega_i}{\sum_{k=1}^K \omega_k}, \quad (5.9)$$

and

$$\text{Var}[\omega'_i] = \frac{\frac{\omega_i}{\sum_{k=1}^K \omega_k} \left(1 - \frac{\omega_i}{\sum_{k=1}^K \omega_k}\right)}{1 + \sum_{k=1}^K \omega_k}. \quad (5.10)$$

When proposing a new set of weights based upon the current set, there needs to be some method of controlling the variance on each weight's distribution whilst keeping the expectation fixed on the current value. A neat way to control the variance is by introducing a scaling factor $\epsilon > 0$ and multiplying it by the current set of weights. By introducing the distribution to $\epsilon > 0$, the

expectation and variance of each proposed weight becomes

$$\mathbb{E}[\omega'_i] = \frac{\epsilon\omega_i}{\sum_{k=1}^K \epsilon\omega_k} = \frac{\epsilon\omega_i}{\epsilon} = \omega_i, \quad (5.11)$$

and

$$\text{Var}[\omega'_i] = \frac{\frac{\epsilon\omega_i}{\sum_{k=1}^K \epsilon\omega_k} \left(1 - \frac{\epsilon\omega_i}{\sum_{k=1}^K \epsilon\omega_k}\right)}{1 + \sum_{k=1}^K \epsilon\omega_k} = \frac{\omega_i(1 - \omega_i)}{1 + \epsilon}. \quad (5.12)$$

So by setting the concentration parameters equal to the previous weights multiplied by a scaling factor, the mean remains unchanged whilst allowing for an increase or decrease in variance. Clearly, as ϵ increases, the variance on proposed values will decrease and the opposite will occur when decreasing ϵ . To conform to the methods of Garthwaite et al. (2016) in optimising the variance of the proposal distribution such that a certain acceptance rate is obtained, ϵ is simply replaced with $\frac{1}{\epsilon}$.

Recall that a proposed value for a parameter is accepted given the probability from Equation 1.18

$$\text{Acc}(\theta'_i|\theta_i) = \min\left(1, \frac{L(\mathbf{x}|\boldsymbol{\theta}') \cdot p(\theta'_i)}{L(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\theta_i)} \cdot \frac{g_{\theta_i}(\theta_i|\theta'_i)}{g_{\theta'_i}(\theta'_i|\theta_i)}\right).$$

So far, the proposal distribution $g(\cdot)$ has been set to a normal density. The symmetric property of this proposal conveniently leads to the Hastings ratio (the right most term in Equation 1.18) to cancel out. A Dirichlet density does not come with the luxury of this symmetric property and as such needs to be included when calculating the acceptance probability.

5.2.4 Data and model settings

To test the normal, Laplace and Cauchy finite mixture models, I revisited a well-known data set in geographic profiling. These data concern 139 locations associated with cases of *Plasmodium vivax* malaria. These data were collected between 2001 and 2004 and were analysed by the CGT algorithm (Le Comber et al. 2011) and the DPM model alongside the simple Bayesian model (Verity et al. 2014) in order to identify breeding sites of the vector responsible for the transmission of malaria. Eleven different mosquitoes species were found in 59 water bodies but only two were of biological interest for these analyses. *Anopheles sergentii* and *Anopheles pharoensis* are both regarded as the most dangerous vectors for transmission of malaria in Cairo (el Said et al. 1986) and of the 59 water bodies these two species were found at ten unique locations. The data analysed in this chapter can be seen in Figure 5.2, where the search area covered 963 km².

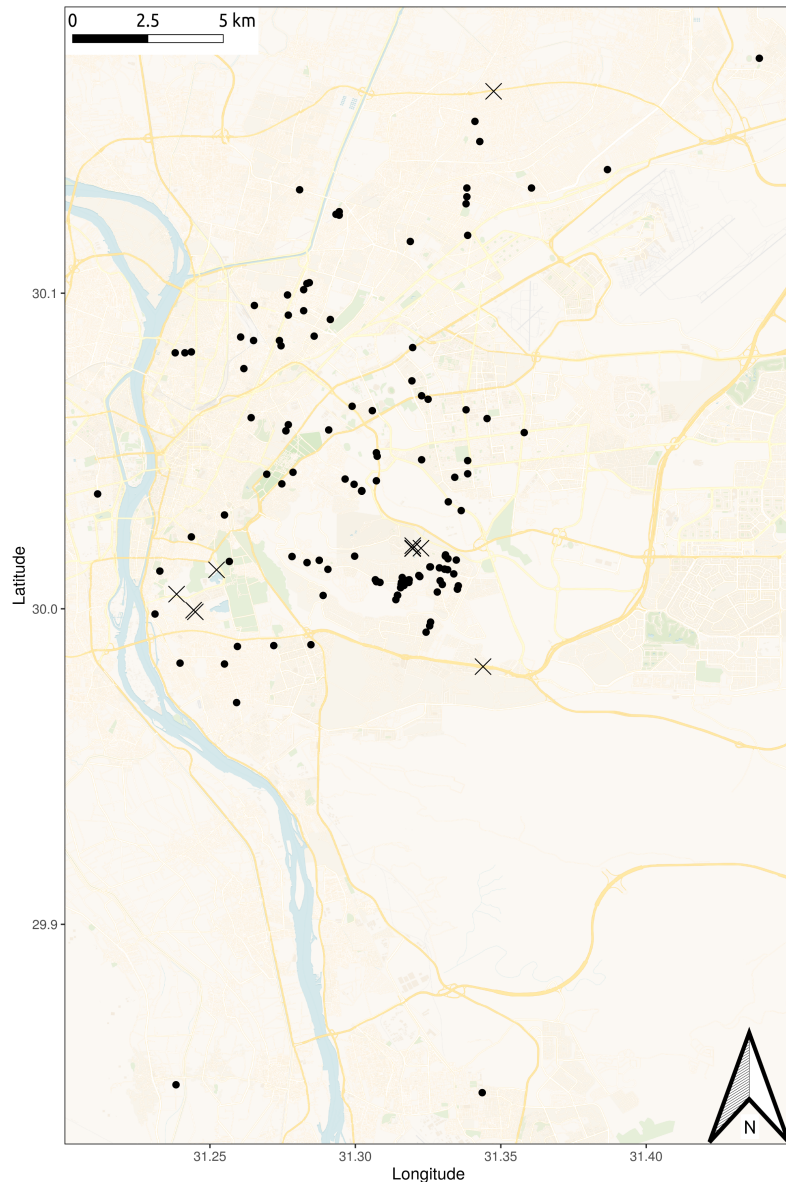


Figure 5.2: Point-pattern data concerning positive tests of malaria in Cairo (black dots). Water bodies containing the species responsible for the transmission of malaria are marked with a cross. This map was created using the *Voyager (no labels) retina* layer via QGIS.org (2021).

Model priors were set as follows. The prior on source locations (\mathcal{F}) was set in a similar manner to that of the DPM model in (Verity et al. 2014). The prior was a bivariate normal density with the mean centred on the spatial average of the data and a standard deviation equal to the maximum distance between the this mean and the data (≈ 22 km).

When estimating the scale parameter for each dispersal kernel (σ_k), each source was assumed to have a unique value. For the bivariate normal, the log-normal prior for the scale parameter was

set with mean of 1.5 km and a standard deviation of 2. The bivariate Laplace's variance was defined as $(2\sigma^2)$, so by equating this variance with that of the bivariate normal, the mean for the log-normal prior under the Laplace model was set to $\frac{1.5}{\sqrt{2}}$ km, again with a standard deviation of 2. When considering the bivariate Cauchy distribution, it was not possible to equate its variance with that of the bivariate normal, given that it is undefined. Hence, the mean on the log-normal prior for the bivariate Cauchy's scale parameter was set to a smaller value than that of the normal; a mean of 0.5 km, again with a standard deviation of 2.

The Dirichlet prior on source weights was set such that each concentration parameter, α_i , was equal to three. Setting this hyper-parameter at three ensured that the model erred in favour of equal weight for each source whilst avoiding the scenario where a single source location was responsible for the majority of the weights.

When running the model, the number of sources searched for (K) was between one and fifteen. Once again the deviance information criterion (DIC) was utilised to determine the most suitable number of source locations to best describe the data. Additionally, the DIC was used to determine which of the three dispersal kernels was the most suitable choice for best describing the data. Following the same protocol as the previous chapter, the number of realised sources were recorded at each MCMC iteration via the allocation probability derived in Equation 4.16.

MCMC parameters were set as follows. The number of burn-in and sampling iterations were set to 5×10^5 and 3×10^5 respectively, where MCMC convergence was checked during the burn-in stage at multiples of 2.5×10^3 . Similarly to previous chapters, Metropolis-Hastings coupling was implemented to ensure healthy MCMC mixing. The number of heated chains ranged from 36 to 92 with an average of 71 to ensure the acceptance probability between chains was at least 75%.

5.3 Results

Results from running the three finite mixture models on the Cairo data can be seen in Figure 5.3, Figure 5.4 and Figure 5.5. Figure 5.4 displays the three geographic profiles produced by each dispersal kernel. Via the deviance information criterion, the number of source locations K was determined to be six for the bivariate normal model and seven for the bivariate Laplace and Cauchy models. This result is illustrated in Figure 5.3a).

DIC values ranged from 1358.61 to 1626.79 with a mean of 1457.16 (normal), 1317.99 to 1561.40 with a mean of 1389.04 (Laplace) and 1351.74 to 1615.30 with a mean of 1413.00 (Cauchy). DIC values for the Laplace model were consistently lower than those of the normal whereas values for the Cauchy were similar to the normal for $K \leq 7$ but were consistently lower for $K > 7$. The choice of dispersal kernel significantly effected the DIC value (ANOVA: $F_{2,42} = 3.68, p = 0.03$).

Hit score percentages for each model can be seen in Figure 5.3b). For each dispersal kernel, hit score percentages ranged from 4.16% to 60.04% with a mean of 26.21% (normal), 4.76% to 27.07% with a mean of 16.04% (Laplace) and 4.66% to 24.62% with a mean of 14.01% (Cauchy). The type of kernel did not significantly effect the hit score (ANOVA: $F_{2,27} = 2.17, p = 0.13$). From these hit scores, Gini coefficients were also obtained. The Gini coefficient under the normal, Laplace and Cauchy models were 0.54, 0.71 and 0.74 respectively.

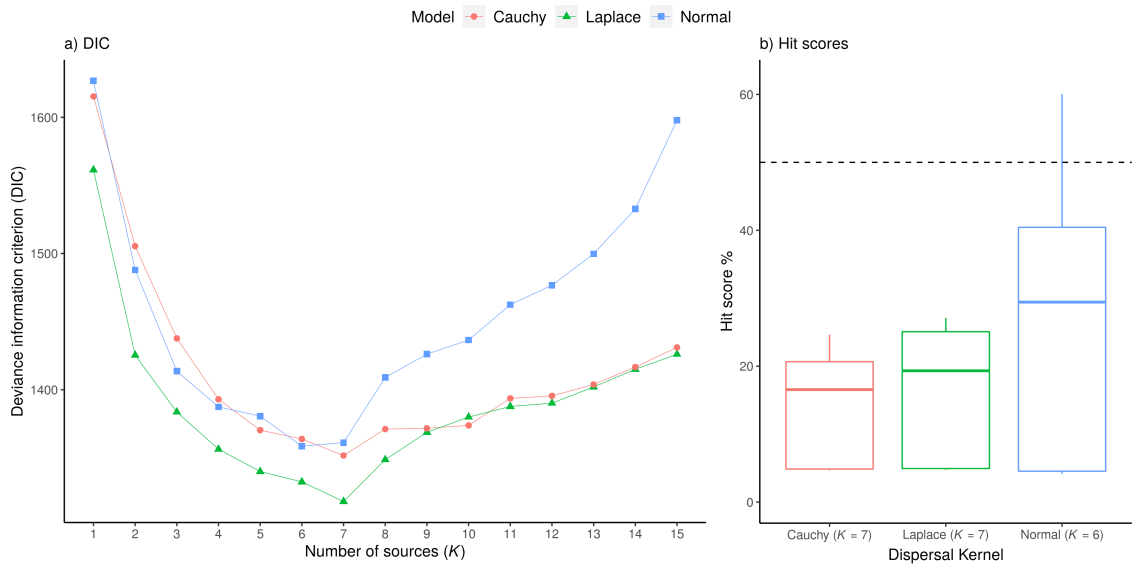


Figure 5.3: DIC values (a) and hit score percentages (b) for each dispersal kernel model. The hit scores in b) are associated with the K model determined via the lowest DIC value in a). The dashed line in b) indicates those hit scores above and below the value obtained from a random search (50%)

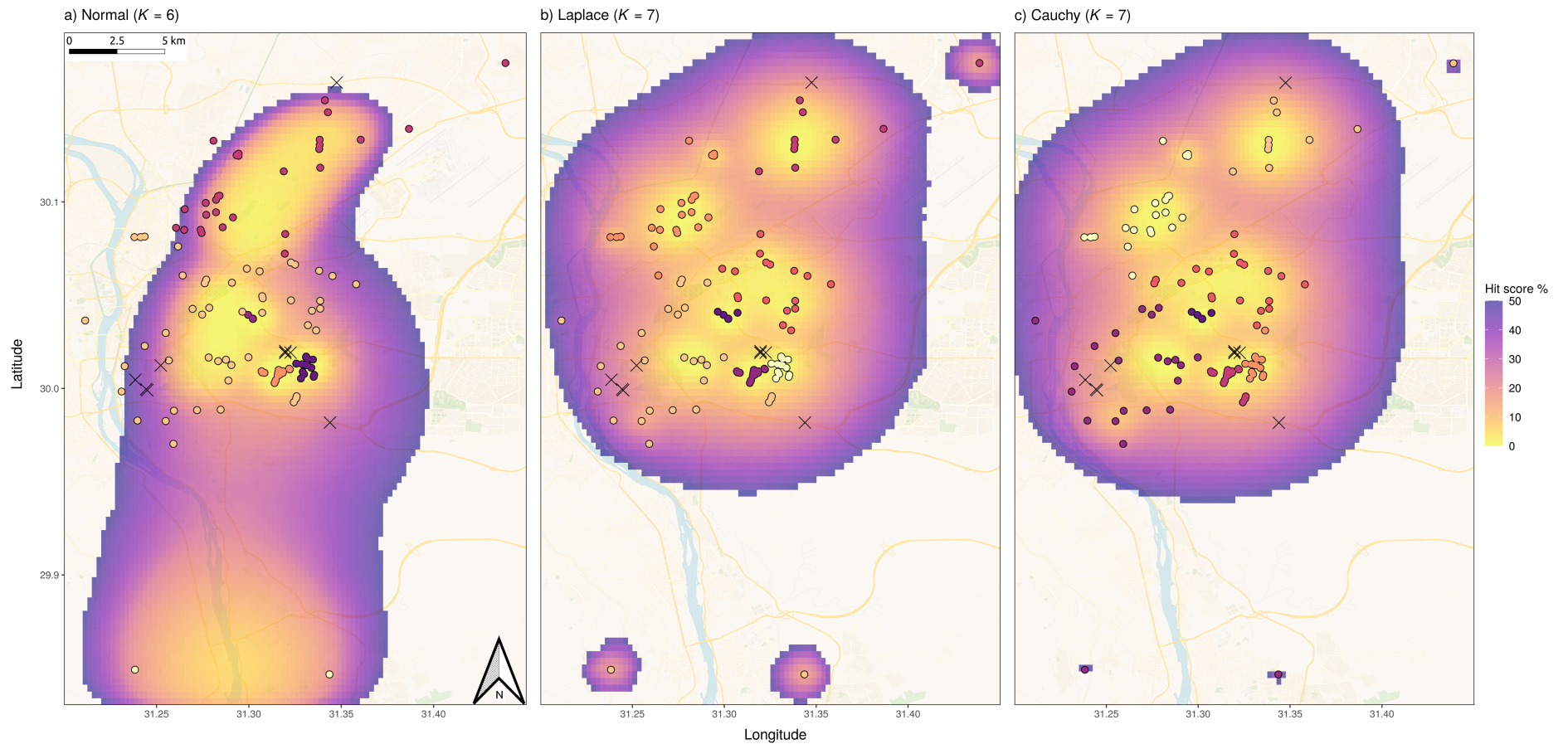


Figure 5.4: Geographic profiles created when fitting a finite mixture of bivariate a) normal, b) Laplace or c) Cauchy distributions. Data points associated with households testing positive for malaria are coloured based on the source they are allocated to. Known bodies of water harbouring breeding mosquitoes are marked with a cross.

Estimates for each model’s scale parameters and source weights can be seen in Figure 5.5. Each model exhibited three sources associated with small scale parameters (10 - 700 metres) with the remaining source locations producing much larger estimates (1.5 - 30km).

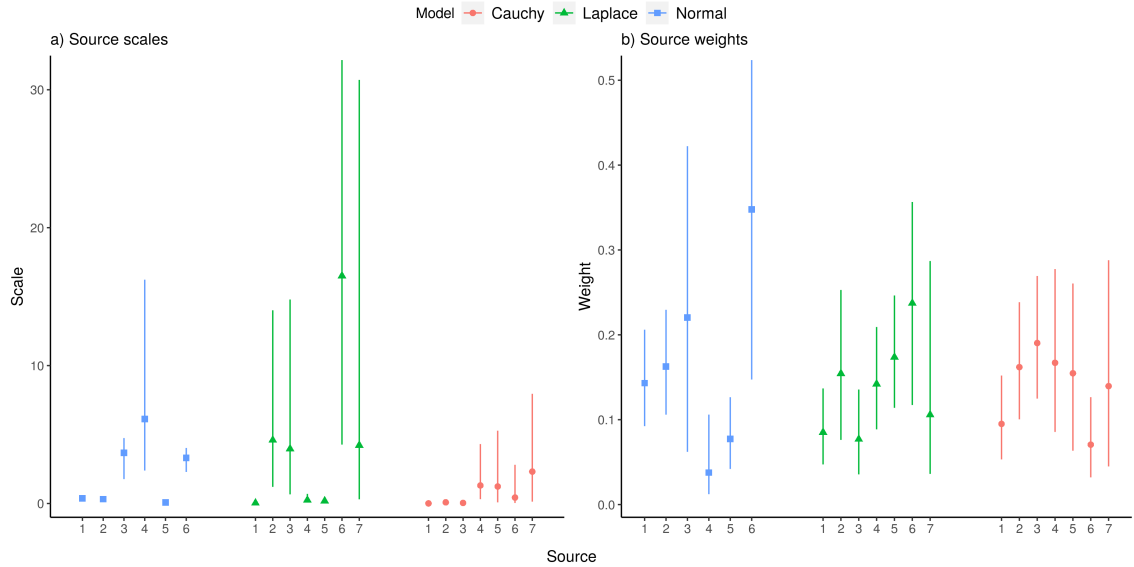


Figure 5.5: Source dependent weight and scale estimates for each dispersal kernel.

Following the protocol of the previous chapter (via Equation 4.16), each iteration of the MCMC algorithm stored the number of realised sources; the number of sources with which the observed data were generated. As seen in previous chapters, the model may go searching for K source locations but only assign the positive data to K' , such that $K' < K$. When analysing the Cairo data, every source location that the model estimated was responsible for generating at least one observed data point.

5.4 Discussion and conclusions

In this chapter, I have used heavy-tailed distributions to model dispersal, motivated by widespread evidence that these distributions are likely to be more appropriate to describe biologically realistic patterns. There was evidence for this expected improvement in inference for this dataset - since I could exploit information about the true sources generating the observed distribution of mosquitoes. The choice of Laplace and Cauchy distributions as alternatives to the normal were two of many potential dispersal kernels that could have been implemented in this chapter. Within criminology, it has been argued that a negative exponential or log-normal distribution could also be implemented to explain dispersal or *journey-to-crime* distances (O’Leary 2009). Additionally,

ecological studies fit a variety of other distributions such as Weibull, Wald or Gamma to explain dispersal distances (Nathan et al. 2012). Laplace and Cauchy distributions were appropriate candidates for comparison to the normal for multiple reasons; both exhibited heavier tails than that of the normal and both were conveniently parameterised with a centre and scale parameter allowing for quick comparisons between different models.

The DIC values seen in Figure 5.3a indicated that a Laplace distribution was the better fitting dispersal kernel of the three considered when analysing the Cairo data. In some cases the Cauchy provided a similar, if not better, fit to the data than the normal, however the consistent improvement from the Laplace distribution is evidence favouring the conclusion that geographic profiling models should not be relying entirely on a bivariate normal distribution when estimating source locations. Although the type of dispersal kernel did not significantly effect the hit score values, the Gini coefficients were shown to be higher for the Laplace and Cauchy models indicating that bodies of water containing vectors of malaria were found more efficiently than when searching using the normal model.

The number of source locations that best described the data was six under the normal model and seven under the Laplace and Cauchy. This result matched that of Verity et al. (2014), where the DPM model estimated between six and ten sources, where seven was the most likely. Curiously, the finite mixture model built in this chapter always allocated data to the number of sources it was searching for. There was no redundancy in a source location unlike the models of the previous chapters. The lack of redundancy was most likely due to the type of data analysed. The existence of negative data in the previous chapters lead the model to place higher density in areas containing no data compared to those containing negative data. This distinction cannot be made for point-pattern data, given that each location is attributed to the presence of, for example, a disease.

There is an overwhelming amount of evidence pointing to the deviation of dispersal distances from a normal to heavier-tailed distributions and in this chapter I have built a model that can fit these kinds of distributions to data. Previously, it has been hard to implement these types of distribution but I have shown that it can readily be done via an MCMC framework. These alternative kernels were implemented in *silverblaze* and the more computer literate user could easily adapt the package's code to implement their distribution of choice. My recommendation is that different study species will require different dispersal kernels and that future studies should

seek out the most suitable form for this distribution prior to running a geographic profiling model. Future work cannot simply rely on the parametrisation of a bivariate normal distribution to account for variation in dispersal distances.

Chapter 6

Conclusion

Abstract:

In this final chapter I bring together the ideas and developments of this thesis and conclude where geographic profiling fits among spatial models within many disciplines. I begin by reflecting upon the work I have conducted, describing the contributions each chapter offers to the field of geographic profiling as well as the developments shared across the entire thesis. Following this, I provide a list of potential next steps for the field of geographic profiling, as an immediate consequence of the work I have conducted, in addition to general gaps that still require attention. I also discuss geographic profiling from an abstract perspective, focussing on the methods at the core of the model, independently of the many fields it has been applied to and I conclude that its methodology is a subset of those found in the analysis of spatial point processes. Finally, I survey the wide range of contemporary biological models for spatial analysis, focussing on their objectives. I point out that as these models become more sophisticated, their aims and the data they analyse begin to align with one another.

6.1 On the developments and findings of this thesis

Chapter 1 of this thesis provided an in depth review that followed the history of geographic profiling. Chapter 1 covered the many pivotal developments as well as the numerous applications that geographic profiling has found along the way. Although this chapter did not provide any developments to the field of geographic profiling, I believe it covers what is needed to inform those individuals foreign to the subject with what they will need to understand this thesis. The jump from O’Leary’s simple Bayesian model (O’Leary 2009) to the complexities of an infinite mixture model (Verity et al. 2014) was a large one, and I hope that the introduction to finite mixture models in chapter 1 will help those individuals new to geographic profiling.

Long	Lat	Long	Lat	Count	Long	Lat	Tested	Positive
-0.040	51.523	-0.040	51.523	1	-0.040	51.523	5	1
-0.133	51.524	-0.133	51.524	1	-0.133	51.524	3	1
-0.089	51.523	-0.089	51.523	2	-0.089	51.523	5	2
-0.075	51.505	-0.075	51.505	0	-0.075	51.505	1	0
-0.051	51.507	-0.051	51.507	3	-0.051	51.507	7	3
...

Table 6.1: The different spatial data types analysed within this thesis. From left to right: point-pattern, count and prevalence data.

Chapter 2 introduced a new geographic profiling model that based its inferences upon spatial count data, defined as a set of locations at which the number of events (for example, the number of individuals caught in a pitfall trap) were counted. This chapter described the Poisson finite mixture model in full whilst also providing the details of an MCMC algorithm to estimate its desired parameters. This new model also offered the possibility of estimating the expected number of events within a search area; a parameter never before obtained by a geographic profiling model. With this information, targeted interventions can focus on the magnitude of resources required in addition to where they should be focussed. Although inferences based upon this expected event size offered information beneficial for informing interventions, it was not a direct parameter of interest for geographic profiling, hence a protocol for sampling from a posterior distribution independent of this parameter was also specified.

Chapter 3 focussed on testing the Poisson finite mixture model, and it did so in two different ways, quickly bringing points for further development to the surface. In applying the model, it became swiftly apparent that the MCMC algorithm described in the previous chapter suffered mixing issues even when the simulation and inference process were both made under the same

model. Hence, a Metropolis-Hastings coupling algorithm was implemented and shown to resolve the issue of poor mixing. Under this scheme, the model returned accurate parameter estimates when tested against an extensive set of simulations. When applied to a set of real spatial count data, I was able to alter the model to deal with over-dispersion, a phenomenon that yields a higher number of sites with a count of zero than what would be expected under the assumptions of the Poisson model. A measure of over-dispersion was built into the model and search strategies produced were compared with that of the Dirichlet process mixture (DPM) model in Verity et al. (2014). By considering the number of events observed at each location, the model was shown to derive new search strategies, unique to geographic profiling, that make the distinction between an area where nothing was observed and an area that was not sampled. Explicitly, this new model highlighted that search strategies should prioritise areas where no sampling took place over those where sampling did take place, but nothing was observed. This result is a clear improvement over the existing DPM model that treats these areas equally. Overall, the new model was shown to be preferable over the DPM model when analysing a set of spatial count data.

Chapter 4 moved away from the spatial count data common to ecology and instead built upon geographic profiling's place in epidemiology. This chapter described a model that estimated the same set of parameters as the previous chapters but via spatial prevalence data; a set of locations each associated with the number of individuals tested for an infectious disease in addition to those testing positive. Beyond the interests of geographic profiling, the methodology was also shown to derive a prevalence probability for each locality within the spatial domain of the model; a parameter of consistent interest in epidemiology but not obtained by current geographic profiling models.

Chapter 5 returned to analysing the data type most common to geographic profiling models. Point-pattern data, a set of locations each associated with a single event, were no longer assumed to be generated via a bivariate normal density around a source location. Instead, a user can specify heavy-tailed dispersal kernels, the Laplace and Cauchy, when fitting a geographic profiling model to their data. The new set of finite mixture models developed in this chapter provided evidence in favour of heavy tailed dispersal distributions over the more commonly used bivariate normal for problems in ecology.

In addition to the developments made in each chapter, this thesis offered a handful of other developments that are present throughout. Discretising the spatial domain over which the

prior on source locations was defined is an invaluable change compared to the continuous space over which the DPM model operates. Through this discrete domain, many biologically realistic priors on source locations can be chosen, opening geographic profiling up to solving new problems via habitat suitability matrices or distance metrics with directional biases. Each model was also equipped to estimate a set of parameters that were unique to each source location. These parameters consisted of the shape of each mixture component (the measure of dispersal distance) in addition to their absolute or relative weights (such as the expected number of events).

Finally, every model specified within this thesis can be implemented by a user through the R package *silverblaze*. For ease, each model comes with its own tutorial for loading data, specifying priors, running the MCMC algorithm and interpreting the results (see Appendix B). It is my hope that by providing these tutorials, geographic profiling becomes a more accessible topic to a broader range of practitioners, especially those less familiar with spatial mixture models. I will now walk through some points of further improvement to develop geographic profiling, both generally and as a direct result of the work in this thesis.

6.2 Points for further development

An assumption of geographic profiling models is that data being analysed exhibit some kind of spatial clustering behaviour. This assumption is implicit within the model since a set of focal points dictate the proximal locations of the data. However, studies implementing such models fail to actually determine if such spatial heterogeneity is present within the data prior to analysis. Many metrics and statistical techniques are available to quantify the spatial structure of a data set (Jacquez 2007, Fritz et al. 2013, Fortin et al. 2013, McLafferty 2015) and I recommend that such methods are utilised to clarify the applicability of geographic profiling to problems in spatial analysis.

The models in this thesis operate under a fine-scale, spatially discretised domain, allowing for the specification of a variety of priors on source locations. The extent of this flexibility is yet to be explored in full as the case studies in this thesis implemented simple spatial priors (jurisdictional boundaries and proximity to data). For example, habitat suitability matrices obtained via species distribution models (Kabir et al. 2017) could be easily adopted to favour those locations preferable to an organism. Similarly to the work of Mohler & Short (2012), the spatial metric assumed within these geographic profiling models (Euclidean distance) could be

manipulated to introduce a directional bias, indicating that the distance from A to B is not necessarily the same for B to A.

Each new innovation within this thesis has broadened the pool of problems that geographic profiling can be applied to. However, there are multiple predictors that are responsible for generating data analysed by geographic profiling that are yet to be included in any existing model. In criminology, the time of day and the existence of a victim and deterrent all drive an offender's decision to commit a crime (Brantingham & Brantingham 1981, 1984); in ecology, habitat type, predation risk, season and climate are but a few drivers of animal foraging and site fidelity behaviour (Avgar et al. 2013, Harris et al. 2020, Morrison et al. 2021) and in epidemiology, variations in altitude, temperature and precipitation can all drive the transmission of an infectious disease (Lindsay & Martens 1998, Protopopoff et al. 2009). Incorporating other predictors is a sensible next step for geographic profiling given the current set of variables fed into these models are few, drawing upon spatial proximity, dispersal tendency and population density.

Absence data has played a major role in this thesis. In the context of this work, absence data refer to those locations that were sampled, but yielded no observations. The ability to work with absence data leads to the possibility of introducing the models in this thesis to pseudo-absences (Barbet-Massin et al. 2012). Specifically, future work could focus on quantifying the utility of pseudo-absences vs an appropriately informed spatial prior. For example, mosquitoes will not be captured in the sea and this could be reflected either by a set of sampled locations yielding no captures (pseudo-absences) or a spatial prior whose probability mass over a large body of water, such as the sea, is zero.

A problem with current geographic profiling models is their persistent use of the hit score metric to measure model success. The hit score percentage and the Gini co-efficient (a value obtained directly from hit scores) are both useful summary statistics but discard a huge amount of information in their calculation. By ranking the final posterior surface on source locations to produce a geographic profile, the complex parameterised form of the model is lost. Given the Bayesian nature of the models in this thesis, and of geographic profiling in general, future studies should focus on reporting results via the posterior probability of a source location.

An obvious alternative to the hit score metric is to utilise protocols in Bayesian prediction. Specifically, a posterior predictive distribution would quantify the probability of observing new

data given the observed. Although a predictive distribution would govern the probability of observing a new crime, invasive sighting or instance of a disease, it may also be useful to quantify the probability of observing a new source location. Realising a currently unseen source location has been captured within existing geographic profiling models. In chapter 1, I described how the DPM model will always allocate some non-zero probability to realising a new source, dependent on the data. In later chapters, I show that making the distinction between areas that are sampled but yielded no observations and those that are not sampled at all leads to a higher posterior density on source locations in those areas with no information. Predicting the location of new data is common practice in criminology (Mohler et al. 2011), ecology (Dormann et al. 2018) and epidemiology (Smith, Le Comber, Fry, Bull, Leach & Hayward 2015), but few, if no, geographic profiling studies have discussed making these predictions based upon the posterior distribution obtained.

Some consideration is still needed as to how resources are explicitly allocated through targeted interventions guided by the output of different models. Although some studies do try to quantify the impact of geographic profiling when targeting resources (Struebig et al. 2018), most studies operate retrospectively, with the success of the model being measured via hit scores of known source locations. By directly contextualising the output of geographic profiling to targeted interventions, the explicit utility of such a model can be measured, allowing for ease of communication between a statistician and a practitioner directly applying geographic profiling to particular problem.

Each model developed in this thesis required a user to specify the number of mixture components (i.e. the number of source locations) that were describing the data, prior to running the MCMC algorithm. This shift from an infinite to a finite mixture model complicated matters by leaving a user to decide for themselves the most suitable value. Though a user could obtain the best of a set of models via the deviance information criterion, this metric lead to different conclusions based upon the prior information fed into each model. A well documented solution bypassing the need to pre-specify the number of mixture components is the so-called birth-death algorithm (Green 1995, Stephens 2000a, Cappé et al. 2003). For finite mixture models, this protocol introduces a birth-death step within the MCMC algorithm that proposes a switch between models, leading to an MCMC algorithm that explores alternate numbers of source locations whilst only needing to be executed once. Alternatively, future work could focus on placing the models in this thesis within a non-parametric framework, such as the DPM model (Neal 2000, Verity et al. 2014). The

bare bones of a non-conjugate DPM model for geographic profiling and corresponding MCMC algorithm have been explored in Appendix A.2.

When analysing spatial count and prevalence data, the models within this thesis have assumed perfect observational power. For count data, any individual present within a landscape was assumed to be observed if it fell within the spatial boundaries defined by a sentinel site's radius. For the prevalence data, every individual testing positive for malaria was assumed to have the disease and analogously those testing negative did not have it. These models are yet to capture the uncertainty arising from type I and type II errors hence future work could focus on developing the observational process that explains the data.

This thesis introduced the first geographic profiling model to allow for independent dispersal distances for each source location. This development was captured via the variance of the bivariate normal distribution that made up each spatial mixture component. The covariance matrix of the bivariate normal distribution, contained non-zero elements on the diagonal only, meaning that probability of observing data was radially symmetric around a source location. Faulkner et al. (2016) described the first instance of a geographic profiling model estimating a dispersal distance parameter and the push to independent values per source was made in chapter 2 (Stevens et al. 2021). A next sensible step is to relax this assumption that data are generated symmetrically around a source location and build a model whose covariance in dispersal distance across dimensions is either specified or estimated. There is evidence supporting that non-radially symmetric clusters capture the process of generating data within criminology and biology, and there are methods that exist that accommodate a multitude of cluster shapes (Lawson et al. 2007).

Despite the many developments that this thesis provides, there are still many avenues for future work and I hope that this section, along with the points raised in the discussions of Stevenson (2015) and Faulkner (2018), will guide the specification of new geographic profiling models.

6.3 Contextualising geographic profiling

Geographic profiling's history is rooted within its applications, but to better understand where its methodology fits into a universe of spatial models, it is worth taking a step back from these

applications in order to approach the subject from an abstract perspective. By viewing the methodology blind to the disciplines that it is applied to (criminology, ecology and epidemiology), it can be seen that the fundamental objectives of geographic profiling conform to those ideas of spatial point processes. The study of point processes is essentially to understand the structure underlying a finite collection of points distributed over some space (Cox & Isham 1980, Diggle 1990). This theoretical discipline resonates with the aims of geographic profiling; to obtain information about the underlying clustered structure of a set of point-pattern data. Generally, geographic profiling can be thought of as within the remit of point processes, but more specifically it fits under the category of spatially inhomogeneous Poisson point processes (Møller 2003, Kottas & Sansó 2007, Ji et al. 2009).

The definition of a spatial point process is always supplemented with some measure to count the number of points that are expected to be observed within a subset of the spatial domain (Daley & Vere-Jones 2007). In this case, the counting measure associated with a spatially inhomogeneous Poisson point process is broken into two parts. The first contribution comes from the expected number of events over the entire domain where as the second captures the spatial heterogeneity in events across a landscape via some probability density. Geographic profiling models prior to this thesis focus entirely on describing the latter of these two components. There are many ways to obtain this density of spatial heterogeneity, and some examples include: kernel density estimation (Rossmo 2000), finite or infinite mixture modelling (Verity et al. 2014, Micheas 2019, Stevens et al. 2021), covariates or marked data (Descombes & Zerubia 2002) and self-exciting point processes (Mohler et al. 2011, Price et al. 2016). In light of the developments within this thesis, geographic profiling now fits under more of the categories linked to analysing spatially inhomogeneous Poisson point processes. This development can be seen within the likelihood of the models that consider both the expected number of events within a domain and the spatial heterogeneity via a finite mixture model.

Concluding that geographic profiling is a method for modelling spatially inhomogeneous Poisson point processes is useful from a theoretical point of view, but implementing such a model, through the various platforms available (see Table 1.1) leads to the conclusion that geographic profiling also falls within the remit of geographic information systems (GIS) (Faulkner 2018). GIS are defined by software that can store, display and analyse a set of spatial data (Fotheringham & Wilson 2007) and cover a broad range of methods including those kinds of analyses for data that cluster in space. Jacquez (2007) splits clustered data into four categories: event-based data,

such as point locations or counts, population-based data, that captures information about a population from which events themselves originated, field-based data, such as covariates that change continuously over space and feature-based data, such as spatial boundaries or jurisdictions. Geographic profiling conforms to the definition of GIS and has found itself applied to all of the above. Hence I conclude that geographic profiling can be labelled as a geographic information system that models the spatial heterogeneity of a spatially inhomogeneous Poisson point process.

It is my hope that by contextualising the theoretical framework of geographic profiling, new developments for the model become clear and future work transcends the current application-driven narrative that most geographic profiling studies follow; that this is a method originally developed in criminology but can also be applied to problems in ecology and epidemiology.

6.4 Linking geographic profiling to other models

In addition to approaching geographic profiling from a theoretical perspective, I will now discuss its place amongst other spatial models by exploring the many fields in which it is directly applied. The term *geographic profiling* is well documented in criminology, from its very first use in Rossmo (1987) up until the present day (Hipp & Williams 2020). In 2006, geographic profiling was first applied to problems in ecology (Le Comber et al. 2006) and five years later, it was applied to problems in epidemiology (Le Comber et al. 2011). Since the first applications in biology, efforts have been made to reconcile the different terminologies across disciplines (Faulkner et al. 2015, Stevens et al. 2021), but less so for the underlying methodologies and as such, the aims and methods of contemporary biological models need to be considered to consolidate geographic profiling's place within these fields.

Early geographic profiling studies in ecology and epidemiology briefly mention alternative spatial methods. For example, it was shown that kernel density estimators for geographic profiling were more accurate than interventions based upon simple metrics of spatial tendency, such as the spatial mean or median of the data (Raine et al. 2009, Le Comber et al. 2011, Le Comber & Stevenson 2012). These studies touched on other spatial models: such as those in epidemiology for estimating putative sources of disease (Lawson 2006, Riley 2007) and heuristic methods that follow a similar approach to kernel density estimators (Buscema et al. 2009, Suzuki-ohno et al. 2010). Although other spatial methods were acknowledged, a detailed comparison between these models and geographic profiling was not present.

Only recently have studies made the connections between geographic profiling and other biological models clearer. In Stevenson et al. (2012) for example, it was suggested that trait-based risk analysis, that quantifies the probability of biological invasion at a location, could be utilised to inform geographic profiling models of potential source locations (Leung et al. 2002, Elith & Leathwick 2009). Stevenson et al. (2012) also noted that geographic profiling lacks the ability to draw conclusions based on an organism's preference of habitat and thus suggested integration with covariate driven ecological niche models (Peterson 2003, Leung et al. 2004, Thuiller et al. 2005). Verity et al. (2014) made a similar point; where integrating geographic profiling and niche modelling was proposed, but not implemented. Le Comber et al. (2011) and Verity et al. (2014) also identified a link between geographic profiling and the theory of point processes (Møller 2003, Henrys & Brown 2009). In particular, Verity et al. (2014) made the point that geographic profiling produces a distribution of source locations as opposed to methods for point processes that obtain a distribution governing the locations of the data.

The links between geographic profiling and the many other spatial methods in ecology and epidemiology form a complex network of similarities and the work within this thesis introduces many new connections in the following ways. Chapter 2 described the first instance of a geographic profiling model that can estimate the population density within an area. Estimating species abundance is a key objective of spatial capture recapture models in ecology (Kéry et al. 2011, Chandler & Royle 2013, Royle et al. 2018). Similarly to the model in chapter 2, these methods make their inferences via spatial count data and are capable of returning a set of *activity centres* synonymous with source locations. However, spatial capture recapture studies focus primarily on this abundance estimate where the process for obtaining activity centres is a possibility but not a necessity. In contrast to this, geographic profiling focusses on estimating source locations and although obtaining population estimates is possible within the model, it is not a necessity (section 2.5 removed this parameter from the model entirely).

New similarities between geographic profiling and other methods in biology can be seen in chapter 4. In this chapter, I described a protocol for obtaining the probability of contracting a disease within each pixel of the gridded spatial domain. This is an objective shared by many models in epidemiology where prevalence estimates can be obtained at a broad scale, such as administrative zones (Staubach et al. 2002, Anderson et al. 2014, Shaweno et al. 2017), or at a fine-scale, where the spatial domain is made of many small pixels (Besag et al. 1991). By

extending geographic profiling to produce prevalence estimates for each locality within the domain a direct comparison can be made between these two methods. In place of obtaining a prevalence rate, ecological models focus on estimating the probability that a species is present within each pixel of the domain. Methods such as maximum entropy (Elith et al. 2011), species distribution models and habitat suitability matrices (Hirzel et al. 2001, Austin 2007, Hengl et al. 2009, Kabir et al. 2017) all produce probabilities indicating a species presence within a search area. These ecological models produce such probabilities via site specific covariates, such as altitude or habitat type. A direct comparison then can be made between these ecological models, that are driven by site specific predictors, and geographic profiling, driven by the proximal relationship between data points. Ultimately, the methods of geographic profiling and niche models should be integrated as both methods stand to benefit from the different pools of information they draw upon.

6.5 Conclusion

In this concluding chapter, I have addressed the numerous advances to geographic profiling that have been made in this thesis. The work that I have conducted provides a broad range of new avenues for a user to explore spatial analyses via geographic profiling through new types of data, new assumptions about that data, various way of quantifying priors, estimating new information unseen to previous geographic profiling models all integrated into a single package with extensive documentation. Additionally, I have identified that at the core of geographic profiling is a geographic information system that focusses on modelling the spatial component of a spatially inhomogeneous Poisson point process. Whilst I do not doubt that geographic profiling will find more applications within its respective disciplines, it is my hope that geographic profiling graduates from its current status as a novel approach, usually applied retrospectively, to a tool for everyday use in the analysis of spatial data.

Bibliography

- Aitkin, M. (2001), ‘Likelihood and Bayesian analysis of mixtures’, *Statistical Modeling* **1**(4), 287–304.
- Albert, J. (2009), *Bayesian Computation with R*, 2nd edn, Springer, Berlin.
- Alonso, J. C., Martin, E., Alonso, J. A. & Morales, M. B. (1998), ‘Proximate and ultimate causes of natal dispersal in the great bustard *Otis tarda*’, *Behavioral Ecology* **9**(3), 243–252.
- Anderson, C., Lee, D. & Dean, N. (2014), ‘Identifying clusters in Bayesian disease mapping’, *Biostatistics* **15**(3), 457–469.
- Atchadé, Y. F., Roberts, G. O. & Rosenthal, J. S. (2011), ‘Towards optimal scaling of metropolis-coupled Markov chain Monte Carlo’, *Statistics and Computing* **21**(4), 555–568.
- Austerlitz, F., Dick, C. W., Dutech, C., Klein, E. K., Oddou-Muratorio, S., Smouse, P. E. & Sork, V. L. (2004), ‘Using genetic markers to estimate the pollen dispersal curve’, *Molecular Ecology* **13**(4), 937–954.
- Austin, M. (2007), ‘Species distribution models and ecological theory: A critical assessment and some possible new approaches’, *Ecological Modelling* **200**(1-2), 1–19.
- Avgar, T., Mosser, A., Brown, G. S. & Fryxell, J. M. (2013), ‘Environmental and individual drivers of animal movement patterns across a wide geographical gradient’, *Journal of Animal Ecology* **82**(1), 96–106.
- Aygin, D. T., Cox, L. A., Faulkner, S. C., Stevens, M. C. A., Verity, R. & Le Comber, S. C. (2019), ‘Double cross: geographic profiling of V-2 impact sites’, *Journal of Spatial Science* pp. 1–12.
- Barbet-Massin, M., Jiguet, F., Albert, C. H. & Thuiller, W. (2012), ‘Selecting pseudo-absences for species distribution models: How, where and how many?’, *Methods in Ecology and Evolution* **3**(2), 327–338.

- Bayes, T. & Price, N. (1763), 'LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S', *Philosophical Transactions of the Royal Society of London* **53**, 370–418.
- Bennell, C., Emeno, K., Snook, B. & Taylor, P. J. (2009), 'The Precision, Accuracy, and Efficiency of Geographic Profiling Predictions: A Simple Heuristic versus Mathematical Algorithms', *Crime Mapping: A Journal of Research and Practice* **1**(2), 65–84.
- Besag, J., York, J. & Mollié, A. (1991), 'A Bayesian image restoration with two applications in spatial statistics Ann Inst Statist Math 43: 159', *Find this article online* **43**(1), 1–20.
- Borchers, D. (2012), 'A non-technical overview of spatially explicit capture-recapture models', *Journal of Ornithology* **152**(SUPPL. 2), 435–444.
- Borchers, D. L. & Efford, M. G. (2008), 'Spatially explicit maximum likelihood methods for capture-recapture studies', *Biometrics* **64**(2), 377–385.
- Bousema, T., Griffin, J. T., Sauerwein, R. W., Smith, D. L., Churcher, T. S., Takken, W., Ghani, A., Drakeley, C. & Gosling, R. (2012), 'Hitting hotspots: Spatial targeting of malaria for control and elimination', *PLoS Medicine* **9**(1), 1–7.
- Boyce, R. M., Brazeau, N., Fulton, T., Hathaway, N., Matte, M., Ntaro, M., Mulogo, E. & Juliano, J. J. (2019), 'Prevalence of molecular markers of antimalarial drug resistance across altitudinal transmission zones in Highland Western Uganda', *American Journal of Tropical Medicine and Hygiene* **101**(4), 799–802.
- Boyce, R. M., Hathaway, N., Fulton, T., Reyes, R., Matte, M., Ntaro, M., Mulogo, E., Waltmann, A., Bailey, J. A., Siedner, M. J. & Juliano, J. J. (2018), 'Reuse of malaria rapid diagnostic tests for amplicon deep sequencing to estimate Plasmodium falciparum transmission intensity in western Uganda', *Scientific Reports* **8**(1), 1–10.
- Brantingham, P. J. & Brantingham, P. L. (1981), *Environmental Criminology*, Sage, Beverly Hills.
- Brantingham, P. L. & Brantingham, P. J. (1984), *Patterns in Crime*, Macmillan, New York.
- Brockerhoff, E. G., Jones, D. C., Kimberley, M. O., Suckling, D. M. & Donaldson, T. (2006), 'Nationwide survey for invasive wood-boring and bark beetles (Coleoptera) using traps baited with pheromones and kairomones', *Forest Ecology and Management* **228**(1-3), 234–240.

- Browne, L., Ottewell, K., Sork, V. L. & Karubian, J. (2018), ‘The relative contributions of seed and pollen dispersal to gene flow and genetic diversity in seedlings of a tropical palm’, *Molecular Ecology* **27**(15), 3159–3173.
- Buscema, M., Grossi, E., Breda, M. & Jefferson, T. (2009), ‘Outbreaks source: A new mathematical approach to identify their possible location’, *Physica A: Statistical Mechanics and its Applications* **388**(22), 4736–4762.
- Butkovic, A., Mrdovic, S., Uludag, S. & Tanovic, A. (2018), ‘Geographic Profiling for serial cybercrime investigation’, *Digital Investigation* (In Press), 1–7.
- Canter, D., Hammond, L., Youngs, D. & Juszczyk, P. (2013), ‘The Efficacy of Ideographic Models for Geographical Offender Profiling’, *Journal of Quantitative Criminology* **29**(3), 423–446.
- Capone, D. L. & Nichols, W. W. (1975), Crime and distance: An analysis of offender behavior in space, in ‘Proceedings of the Association of American Geographers’, Vol. 7, pp. 45–49.
- Cappé, O., Robert, C. P. & Rydén, T. (2003), ‘Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **65**(3), 679–700.
- Carter, R., Mendis, K. N. & Roberts, D. (2000), ‘Spatial targeting of interventions against malaria’, *Bulletin of the World Health Organization* **78**(12), 1401–1411.
- Cerri, J., Mori, E., Zozzoli, R., Gigliotti, A., Chirco, A. & Bertolino, S. (2020), ‘Managing invasive Siberian chipmunks *Eutamias sibiricus* in Italy: a matter of attitudes and risk of dispersal’, *Biological Invasions* **22**(2), 603–616.
- Chadœuf, J., Millon, A., Bourrioux, J. L., Printemps, T., van Hecke, B., Lecoustre, V. & Bretagnolle, V. (2018), ‘Modelling unbiased dispersal kernels over continuous space by accounting for spatial heterogeneity in marking and observation efforts’, *Methods in Ecology and Evolution* **9**(2), 331–339.
- Chandler, R. B. & Royle, A. J. (2013), ‘Spatially explicit models for inference about density in unmarked or partially marked populations’, *Annals of Applied Statistics* **7**(2), 936–954.
- Chapman, D. S., Dytham, C. & Oxford, G. S. (2007), ‘Modelling population redistribution in a leaf beetle: An evaluation of alternative dispersal functions’, *Journal of Animal Ecology* **76**(1), 36–44.
- Cheng, J., Karambelkar, B. & Xie, Y. (2019), *leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*.

- Coombs, M. F. & Rodríguez, M. A. (2007), ‘A field test of simple dispersal models as predictors of movement in a cohort of lake-dwelling brook charr’, *Journal of Animal Ecology* **76**(1), 45–57.
- Cowles, M. K. & Carlin, B. P. (1996), ‘Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review’, *Journal of the American Statistical Association* **91**(434), 883–904.
- Cox, D. R. & Isham, V. (1980), *Point processes*, Vol. 12, CRC Press.
- Daley, D. J. & Vere-Jones, D. (2007), *An introduction to the theory of point processes: volume II: general theory and structure*, Springer Science & Business Media.
- Daojing Wang, Chao Zhang & Xuemin Zhao (2009), ‘Multivariate Laplace Filter: A heavy-tailed model for target tracking’, pp. 1–4.
- Descombes, X. & Zerubia, J. (2002), ‘Marked point process in image analysis’, *IEEE Signal Processing Magazine* **19**(5), 77–84.
- Diggle, P. (1990), ‘A Point Process Modelling Approach to Raised Incidence of a Rare Phenomenon in the Vicinity of a Prespecified Point’, *Royal Statistical Society* **153**(3), 349–362.
- Doney, R. (1990), The aftermath of the Yorkshire Ripper: the response of the United Kingdom Police Service, in S. Egger, ed., ‘Serial murder: an elusive phenomenon’, Praeger, New York, pp. 95–112.
- Dorfman, R. (1979), ‘A Formula for the Gini Coefficient’, *The Review of Economics and Statistics* **61**(1), 146–149.
- Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Wood, S. N., Wüest, R. O. & Hartig, F. (2018), ‘Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference’, *Ecological Monographs* **88**(4), 485–504.
- Dunson, D. B. & Johndrow, J. E. (2019), ‘The Hastings algorithm at fifty’, *Biometrika* pp. 1–23.
- Efford, M. (2004), ‘Density estimation in live-trapping studies’, *Oikos* **106**(3), 598–610.
- el Said, S., Beier, J. C., Kenawy, 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**(2), 158–163.

- Elith, J. & Leathwick, J. R. (2009), ‘Species distribution models: Ecological explanation and prediction across space and time’, *Annual Review of Ecology, Evolution, and Systematics* **40**, 677–697.
- Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. & Yates, C. J. (2011), ‘A statistical explanation of MaxEnt for ecologists’, *Diversity and Distributions* **17**(1), 43–57.
- Faulkner, S. C. (2018), Integrating GIS approaches with geographic profiling as a novel conservation tool, PhD thesis, Queen Mary University of London.
- Faulkner, S. C., Stevens, M. C. A., Romañach, S. S., Lindsey, P. A. & Le Comber, S. C. (2018), ‘A spatial approach to combatting wildlife crime’, *Conservation Biology* pp. 1–24.
- Faulkner, S. C., Stevenson, M. D., Verity, R., Mustari, A. H., Semple, S., Tosh, D. G. & Le Comber, S. C. (2015), ‘Using geographic profiling to locate elusive nocturnal animals: A case study with spectral tarsiers’, *Journal of Zoology* **295**(4), 261–268.
- Faulkner, S. C., Verity, R., Roberts, D., Roy, S. S., Robertson, P. A., Stevenson, M. D. & Le Comber, S. C. (2016), ‘Using geographic profiling to compare the value of sightings vs trap data in a biological invasion’, *Diversity and Distributions* **23**(1), 104–112.
- Forero, M. G., Donázar, J. A. & Hiraldo, F. (2002), ‘Causes and fitness consequences of natal dispersal in a population of Black Kites’, *Ecology* **83**(3), 858–872.
- Fortin, M.-j., Dale, M. R. T. & Ver Hoef, J. M. (2013), ‘Spatial analysis in ecology’, *Encyclopedia of Environmetrics* .
- Fotheringham, A. S. & Wilson, J. P. (2007), Geographic Information Science: An Introduction, *in* ‘The Handbook of Geographic Information Science’, John Wiley & Sons, Ltd, pp. 1–7.
- Fritz, C. E., Schuurman, N., Robertson, C. & Lear, S. (2013), ‘A scoping review of spatial cluster analysis techniques for point-event data’, *Geospatial Health* **7**(2), 183–198.
- Garthwaite, P. H., Fan, Y. & Sisson, S. A. (2016), ‘Adaptive optimal scaling of MetropolisHastings algorithms using the RobbinsMonro process’, *Communications in Statistics - Theory and Methods* **45**(17), 5098–5111.
- Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004), ‘Bayesian data analysis’.
- Geman, S. & Geman, D. (1984), ‘Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**(6), 721–741.

- Gershman, S. J. & Blei, D. M. (2012), ‘A tutorial on Bayesian nonparametric models’, *Journal of Mathematical Psychology* **56**(1), 1–12.
- Geweke, J. (1991), Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments, Technical Report 148, Federal Reserve Bank of Minneapolis.
- Gilliland, D. C. (1962), ‘Integral of the Bivariate Normal Distribution Over an Offset Circle’, **57**(300), 758–768.
- Gorrochotegui-Escalante, N., De Lourdes Munoz, M., Fernandez-Salas, I., Beaty, B. J. & Black IV, W. C. (2000), ‘Genetic isolation by distance among *Aedes aegypti* populations along the northeastern coast of Mexico’, *American Journal of Tropical Medicine and Hygiene* **62**(2), 200–209.
- Gray, R. J. (2020), Exotic Hobos: Release, escape, and potential secondary dispersal of African red-headed agamas (*Agama picticauda* PETERS, 1877) through the Florida railway systems.
- Green, P. J. (1995), ‘Reversible jump Markov chain Monte Carlo computation Bayesian model determination’, *Biometrika* **82**(4), 711–732.
- Green, P. J. & Richardson, S. (2001), ‘Modelling Heterogeneity With and Without the Dirichlet Process’, *Scandinavian Journal of Statistics* **28**(2), 355–375.
- Haario, H., Saksman, E. & Tamminen, J. (1999), ‘Adaptive proposal distribution for random walk Metropolis algorithm’, *Computational Statistics* **14**(3), 375–395.
- Hall, L. A., Van Schmidt, N. D. & Beissinger, S. R. (2018), ‘Validating dispersal distances inferred from autoregressive occupancy models with genetic parentage assignments’, *Journal of Animal Ecology* **87**(3), 691–702.
- Harries, K. & LeBeau, J. (2007), ‘Issues in the Geographic Profiling of Crime: Review and Commentary’, *Police Practice and Research* **8**(4), 321–333.
- Harris, S. M., Descamps, S., Sneddon, L. U., Bertrand, P., Chastel, O. & Patrick, S. C. (2020), ‘Personality predicts foraging site fidelity and trip repeatability in a marine predator’, *Journal of Animal Ecology* **89**(1), 68–79.
- Hastings, W. K. (1970), ‘Monte Carlo Sampling Methods Using Markov Chains and Their Applications’, *Biometrika* **57**(1), 97–109.

- Hauge, M. V., Stevenson, M. D., Rossmo, D. K. & Le Comber, S. C. (2016), ‘Tagging Banksy: using geographic profiling to investigate a modern art mystery’, *Journal of Spatial Science* **61**(March), 1–6.
- Hayes, E. B. (2009), ‘Zika virus outside Africa’, *Emerging Infectious Diseases* **15**(9), 1347–1350.
- Heald, O. J., Fraticelli, C., Cox, S. E., Stevens, M. C. A., Faulkner, S. C., Blackburn, T. M. & Le Comber, S. C. (2019), ‘Understanding the origins of the ring-necked parakeet in the UK’, *Journal of Zoology* pp. 1–11.
- Hengl, T., Sierdsema, H., Radović, A. & Dilo, A. (2009), ‘Spatial prediction of species’ distributions from occurrence-only records: combining point pattern analysis, ENFA and regression-kriging’, *Ecological Modelling* **220**(24), 3499–3511.
- Hennessey, M., Fischer, M. & Staples, J. (2016), ‘Zika Virus Spreads to New Areas Region of the Americas, May 2015 January 2016’, *American Journal of Transplantation* **16**, 1031–1034.
- Henry, P. A. & Brown, P. E. (2009), ‘Inference for clustered inhomogeneous spatial point processes’, *Biometrics* **65**(2), 423–430.
- Hipp, J. R. & Williams, S. A. (2020), ‘Advances in Spatial Criminology: The Spatial Scale of Crime’, *Annual Review of Criminology* **3**, 75–95.
- Hirzel, A. H., Helfer, V. & Metral, F. (2001), ‘Assessing habitat-suitability models with a virtual species’, *Ecological Modelling* **145**, 111–121.
- Huddleston, S. H., Gerber, M. S. & Brown, D. E. (2013), Geographic Profiling of Criminal Groups for Military Cordon and Search, in A. M. Greenberg, W. G. Kennedy & N. D. Bos, eds, ‘Social Computing, Behavioral-Cultural Modeling and Prediction’, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 503–512.
- Jacquez, G. M. (2007), Spatial Cluster Analysis, in ‘The Handbook of Geographic Information Science’, John Wiley & Sons, Ltd, chapter 22, pp. 395–416.
- Ji, C., Merl, D., Kepler, T. B. & West, M. (2009), ‘Spatial mixture modelling for unobserved point processes: Examples in immunofluorescence histology’, *Bayesian Analysis* **4**(2), 297–316.
- Johnson, S. D. (2014), ‘How do offenders choose where to offend? Perspectives from animal foraging’, *Legal and Criminological Psychology* **19**(2), 193–210.
- Kabir, M., Hameed, S., Ali, H., Bosso, L., Din, J. U., Bischof, R., Redpath, S. & Nawaz, M. A. (2017), ‘Habitat suitability and movement corridors of Pakistan’, *PLoS ONE* **12**(11).

- Kaiser, M. S., Cressie, N. & Lee, J. (2002), ‘Spatial Mixture Models Based on Exponential Family Conditional Distributions’, *Statistica Sinica* **12**(2), 449–474.
- Kent, J. D. (2009), Essays on the Integration of Anisotropic Landscapes within Contemporary Geographic Profiling Models, PhD thesis.
- Kent, J., Leitner, M. & Curtis, A. (2006), ‘Evaluating the usefulness of functional distance measures when calibrating journey-to-crime distance decay functions’, *Computers, Environment and Urban Systems* **30**(2), 181–200.
- Kéry, M., Gardner, B., Stoeckle, T., Weber, D. & Royle, J. A. (2011), ‘Use of Spatial Capture-Recapture Modeling and DNA Data to Estimate Densities of Elusive Animals’, *Conservation Biology* **25**(2), 356–364.
- Kottas, A. & Sansó, B. (2007), ‘Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis’, *Journal of Statistical Planning and Inference* **137**(10), 3151–3163.
- Kotz, S., Kozubowski, T. J. & Podgorski, K. (2001), *The Laplace Distribution and Generalizations - A Revisit with New Applications*, Birkhauser, Boston.
- Krkošek, M., Lauzon-Guay, J. S. & Lewis, M. A. (2007), ‘Relating dispersal and range expansion of California sea otters’, *Theoretical Population Biology* **71**(4), 401–407.
- Lawson, A. (2006), Statistical methods in spatial epidemiology, Technical report, Wiley.
- Lawson, A. B. (1995), ‘MCMC methods for putative pollution source problems in environmental epidemiology’, *Statistics in Medicine* **14**(21-22), 2473–2485.
- Lawson, A. B. (2000), ‘Cluster modelling of disease incidence via RJMCMC methods: A comparative evaluation’, *Statistics in Medicine* **19**(17-18), 2361–2375.
- Lawson, A. B. & Clark, A. (2002), ‘Spatial mixture relative risk models applied to disease mapping’, *Statistics in Medicine* **21**(3), 359–370.
- Lawson, A. B., Simeon, S., Kulldorff, M., Biggeri, A. & Magnani, C. (2007), ‘Line and point cluster models for spatial health data’, *Computational Statistics and Data Analysis* **51**(12), 6027–6043.
- Le Comber, S. C., Nicholls, B., Rossmo, D. K. & Racey, P. A. (2006), ‘Geographic profiling and animal foraging’, *Journal of Theoretical Biology* **240**(2), 233–240.

- 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), 1–8.
- Le Comber, S. C. & Stevenson, M. D. (2012), 'From Jack the Ripper to epidemiology and ecology', *Trends in Ecology and Evolution* **27**(6), 307–308.
- Lee, H.-Y., Park, H.-J. & Kim, H.-M. (2014), 'A Clarification of the Cauchy Distribution', *Communications for Statistical Applications and Methods* **21**(2), 183–191.
- Leung, B., Drake, J. M. & Lodge, D. M. (2004), 'Predicting invasions: Propagule pressure and the gravity of allee effects', *Ecology* **85**(6), 1651–1660.
- Leung, B., Lodge, D. M., Finnoff, D., Shogren, J. F., Lewis, M. A. & Lamberti, G. (2002), 'An ounce of prevention or a pound of cure: Bioeconomic risk analysis of invasive species', *Proceedings of the Royal Society B: Biological Sciences* **269**(1508), 2407–2413.
- Levine, N. & Block, R. (2011), 'Bayesian Journey-to-Crime estimation: An improvement in Geographic profiling Methodology', *Professional Geographer* **63**(2), 213–229.
- Lindén, A. & Mäntyniemi, S. (2011), 'Using the negative binomial distribution to model overdispersion in ecological count data', *Ecology* **92**(7), 1414–1421.
- Lindsay, S. W. & Martens, W. J. (1998), 'Malaria in the African highlands: Past, present and future', *Bulletin of the World Health Organization* **76**(1), 33–45.
- MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, A. A. & Langtimm, C. A. (2002), 'Estimating site occupancy rates when detection probabilities are less than one', *Ecology* **83**(8), 2248–2255.
- Marsh, K. (2010), 'Research priorities for malaria elimination', *The Lancet* **376**(9753), 1626–1627.
- Martin, R. A., Rossmo, D. K. & Hammerschlag, N. (2009), 'Hunting patterns and geographic profiling of white shark predation', *Journal of Zoology* **279**(2), 111–118.
- Matthysen, E. (2012), Multicausality of dispersal: a review, in 'Dispersal Ecology and Evolution', chapter 1, pp. 3–18.
- Mburu, L. & Helbich, M. (2015), 'Evaluating the Accuracy and Effectiveness of Criminal Geographic Profiling Methods: The Case of Dandora, Kenya', *Professional Geographer* **67**(1), 110–120.

- Mclafferty, S. (2015), ‘Disease cluster detection methods: recent developments and public health implications’, *Annals of GIS* **21**(2), 127–133.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), ‘Equation of State Calculations by Fast Computing Machines’, *The Journal of Chemical Physics* **21**(6), 1087–1092.
- Metropolis, N. & Ulam, S. (1949), ‘The Monte Carlo Method’, *Journal of the American Statistical Association* **44**(247), 335–341.
- Micheas, A. C. (2019), ‘Cox Point Processes: Why One Realisation Is Not Enough’, *International Statistical Review* **87**(2), 306–325.
- Mohler, G. O. & Short, M. B. (2012), ‘Geographic Profiling from Kinetic Models of Criminal Behavior’, *SIAM Journal on Applied Mathematics* **72**(1), 163–180.
- Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. & Tita, G. E. (2011), ‘Self-exciting point process modeling of crime’, *Journal of the American Statistical Association* **106**(493), 100–108.
- Møller, J. (2003), ‘Shot Noise Cox Processes’, *Advances in Applied Probability* **35**(3), 614–640.
- Moraes, M. A., Kubota, T. Y., Rossini, B. C., Marino, C. L., Freitas, M. L., Moraes, M. L., Da Silva, A. M., Cambuim, J. & Sebbenn, A. M. (2018), ‘Long-distance pollen and seed dispersal and inbreeding depression in *Hymenaea stigonocarpa* (Fabaceae: Caesalpinioideae) in the Brazilian savannah’, *International Journal of Business Innovation and Research* **17**(3), 7800–7816.
- Morrison, T. A., Merkle, J. A., Hopcraft, J. G. C., Aikens, E. O., Beck, J. L., Boone, R. B., Courtemanch, A. B., Dwinnell, S. P., Fairbanks, W. S., Griffith, B., Middleton, A. D., Monteith, K. L., Oates, B., Riottelambert, L., Sawyer, H., Smith, K. T., Stabach, J. A., Taylor, K. L. & Kauffman, M. J. (2021), ‘Drivers of site fidelity in ungulates’, *Journal of Animal Ecology* (November 2020), 1–12.
- Morton, E. R., McGrady, M. J., Newton, I., Rollie, C. J., Smith, G. D., Mearns, R. & Oli, M. K. (2018), ‘Dispersal: a matter of scale’, *Ecology* **99**(4), 938–946.
- Nathan, R., Klein, E., Robledo-Arnuncio, J. J. & Revilla, E. (2012), Dispersal kernels: review, in ‘Dispersal Ecology and Evolution’, chapter 15, pp. 187–210.
- Neal, R. M. (2000), ‘Markov Chain Sampling Methods for Dirichlet Process Mixture Models’, *Journal of Computational and Graphical Statistics* **9**(2), 249–265.

- Neal, R. M. (2011), MCMC using hamiltonian dynamics, *in* ‘Handbook of Markov Chain Monte Carlo’, pp. 113–162.
- Ng, K. W., Tian, G.-L. & Tang, M.-L. (2011), *Dirichlet and related distributions: Theory, methods and applications*, John Wiley & Sons.
- Okello, P. E., Van Bortel, W., Byaruhanga, A. M., Correwyn, A., Roelants, P., Talisuna, A., D’Alessandro, U. & Coosemans, M. (2006), ‘Variation in malaria transmission intensity in seven sites throughout Uganda’, *American Journal of Tropical Medicine and Hygiene* **75**(2), 219–225.
- O’Leary, M. (2009), ‘The Mathematics of Geographic Profiling’, *Journal of Investigative Psychology and Offender Profiling* **6**, 253–265.
- O’Leary, M. (2010a), Implementing a Bayesian approach to criminal geographic profiling, *in* ‘1st International Conference and Exhibition on Computing for Geospatial Research & Application’.
- O’Leary, M. (2010b), ‘Multimodel inference and geographic profiling’, *Crime Mapping: A Journal of Research and Practice* **2**(1), 50–64.
- Paradis, E., Baillie, S. R. & Sutherland, W. J. (2002), ‘Modeling large-scale dispersal distances’, *Ecological Modelling* **151**(2-3), 279–292.
- Paulsen, D. (2006a), ‘Human versus machine: a comparison of the accuracy of geographic profiling methods’, *Journal of Investigative Psychology and Offender Profiling* **3**(2), 77–89.
- Paulsen, D. J. (2006b), ‘Connecting the dots: Assessing the accuracy of geographic profiling software’, *Policing* **29**(2), 306–334.
- Peterson, A. T. (2003), ‘Predicting the geography of species’ invasions via ecological niche modeling’, *Quarterly Review of Biology* **78**(4), 419–433.
- Price, S., Garner, T., Cunningham, A., Langton, T. & Nichols, R. A. (2016), Reconstructing the emergence of a lethal infectious disease of wildlife supports a key role for spread through translocations by humans, *in* ‘Royal Society B’, Vol. 283.
- Protopopoff, N., Van Bortel, W., Speybroeck, N., Van Geertruyden, J. P., Baza, D., D’Alessandro, U. & Coosemans, M. (2009), ‘Ranking malaria risk factors to guide malaria control efforts in African highlands’, *PLoS ONE* **4**(11), 1–10.
- QGIS.org (2021), ‘QGIS Geographic Information System’.
- URL:** <http://www.qgis.org>

- R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- 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**(32), 307–319.
- Ramasamy, R., Surendran, S. N., Jude, P. J., Dharshini, S. & Vinobaba, M. (2011), ‘Larval development of *Aedes aegypti* and *Aedes albopictus* in peri-urban brackish water and its implications for transmission of arboviral diseases’, *PLoS Neglected Tropical Diseases* **5**(11).
- Ramos, S. L. F., Dequigiovanni, G., Sebbenn, A. M., Lopes, M. T. G., de Macêdo, J. L. V., Veasey, E. A., Alves-Pereira, A., da Silva, P. P., Garcia, J. N. & Kageyama, P. Y. (2018), ‘Paternity analysis, pollen flow, and spatial genetic structure of a natural population of *Euterpe precatoria* in the Brazilian Amazon’, *Ecology and Evolution* **8**(22), 11143–11157.
- Riley, S. (2007), ‘Large-Scale Models of Infectious Disease’, *Science* **316**(June), 1298–1301.
- Roberts, G. O. & Tweedie, R. L. (1996), ‘Exponential convergence of Langevin distributions and their discrete approximations’, **2**(4), 341–363.
- Rosenthal, J. S. (2011), Optimal proposal distributions and adaptive MCMC, in ‘Handbook of Markov Chain Monte Carlo’, pp. 93–112.
- Rossmo, D. K. (1987), Fugitive migration patterns, PhD thesis, Arts and Social Sciences: Criminology.
- Rossmo, D. K. (1993), ‘A methodological model’, *American Journal of Criminal Justice* **17**(2), 1–21.
- Rossmo, D. K. (2000), *Geographic Profiling*, CRC Press, Boca Raton, Florida.
- Rossmo, D. K. (2005a), ‘Commentary: Geographic heuristics or shortcuts to failure?: Response to snook et al’, *Applied Cognitive Psychology* **19**(5), 651–654.
- Rossmo, D. K. (2005b), Geographic Profiling as Problem Solving for Serial Crime, in ‘Police Problem Solving’, pp. 121–131.
- Rossmo, D. K. (2012), ‘Recent Developments in Geographic Profiling’, *Policing* **6**(2), 144–150.
- Rossmo, D. K. & Harries, K. (2011), ‘The Geospatial Structure of Terrorist Cells’, *Justice Quarterly* **28**(2), 221–248.

- Rossmo, D. K., Lutermann, H., Stevenson, M. D. & Le Comber, S. C. (2014), 'Geographic Profiling in Nazi Berlin: fact and fiction', *Geospatial Intelligence Review* **12**(1), 44–57.
- Rossmo, D. K. & Routledge, R. (1990), 'Estimating the size of criminal populations', *Journal of Quantitative Criminology* **6**(3), 293–314.
- Royle, J. A. (2004), 'N-Mixture Models for Estimating Population Size from Spatially Replicated Counts', *Biometrics* **60**(March), 108–115.
- Royle, J. A., Fuller, A. K. & Sutherland, C. (2018), 'Unifying population and landscape ecology with spatial capturerecapture', *Ecography* **41**(3), 444–456.
- Royle, J. A., Magoun, A. J., Gardner, B., Valkenburg, P. & Lowell, R. E. (2011), 'Density estimation in a wolverine population using spatial capture-recapture models', *Journal of Wildlife Management* **75**(3), 604–611.
- Sagnard, F., Pichot, C., Dreyfus, P., Jordano, P. & Fady, B. (2007), 'Modelling seed dispersal to predict seedling recruitment: Recolonization dynamics in a plantation forest', *Ecological Modelling* **203**(3-4), 464–474.
- Santosuosso, U. & Papini, A. (2018), 'Geo-Profiling : beyond the Current Limits. A Preliminary Study of Mathematical Methods to Improve the Monitoring of Invasive Species', *Russian Journal of Ecology* **49**(May), 346–354.
- Schurr, F. M., Steinitz, O. & Nathan, R. (2008), 'Plant fecundity and seed dispersal in spatially heterogeneous environments: Models, mechanisms and estimation', *Journal of Ecology* **96**(4), 628–641.
- Service, M. W. & Place, P. (1997), 'Mosquito (Diptera : Culicidae) Dispersal The Long and Short of It', *Journal of Medical Entomology* **34**(6), 579–588.
- Shaweno, D., Trauer, J. M., Denholm, J. T. & McBryde, E. S. (2017), 'A novel Bayesian geospatial method for estimating tuberculosis incidence reveals many missed TB cases in Ethiopia', *BMC Infectious Diseases* **17**(1), 1–8.
- Sinclair, E. A., Ruiz-Montoya, L., Krauss, S. L., Anthony, J. M., Hovey, R. K., Lowe, R. J. & Kendrick, G. A. (2018), 'Seeds in motion: Genetic assignment and hydrodynamic models demonstrate concordant patterns of seagrass dispersal', *Molecular Ecology* **27**(24), 5019–5034.

- Smith, C. M., Downs, S. H., Mitchell, A., Hayward, A. C., Fry, H. & Le Comber, S. C. (2015), ‘Spatial targeting for bovine tuberculosis control: Can the locations of infected cattle be used to find infected badgers?’, *PLoS ONE* **10**(11), 1–14.
- Smith, C. M., Le Comber, S. C., Fry, H., Bull, M., Leach, S. & Hayward, A. C. (2015), ‘Spatial methods for infectious disease outbreak investigations: Systematic literature review’, *Eurosurveillance* **20**(39), 1–21.
- Snook, B., Taylor, P. J. & Bennell, C. (2004), ‘Geographic profiling: The fast, frugal, and accurate way’, *Applied Cognitive Psychology* **18**(1), 105–121.
- South Florida Region Shapefile, Miami-Dade County - Open Data Hub* (2018).
URL: <https://gis-mdc.opendata.arcgis.com/datasets/south-florida-region/>
- Spaulding, J. & Morris, K. (2020), *rgeoprofile: Geographic Profiling Methods for Serial Crime Analysis*.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Linde, A. V. D. (2014), ‘The deviance information criterion: 12 years on’, *Journal of the Royal Statistical Society: Series B* **76**(3), 485–493.
- Staubach, C., Schmid, V., Knorr-held, L. & Ziller, M. (2002), ‘A Bayesian model for spatial wildlife disease prevalence data’, **56**, 75–87.
- Stephens, M. (2000a), ‘Bayesian Analysis of Mixture Models with an Unknown Number of Components - An Alternative to Reversible Jump Methods’, *The Annals of Statistics* **28**(1), 40–74.
- Stephens, M. (2000b), ‘Dealing with label switching in mixture models’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **62**(4), 795–809.
- Stevens, M. C. A., Chen, Y., Stringer, A., Clemmow, C. & Lewis, A. (2020), Key factors driving obesity in the UK, in ‘Proc. of the 28th conference for Geographic Information Systems Research in the UK (GISRUK)’.
- Stevens, M. C. A. & Faulkner, S. C. (2018), ‘Geographic Profiling: murder, maths, malaria and mammals’, *Chalkdust magazine* pp. 18–25.
- Stevens, M. C. A., Faulkner, S. C., Wilke, A. B., Beier, J. C., Vasquez, C., Petrie, W. D., Fry, H., Nichols, R. A., Verity, R. & Le Comber, S. C. (2021), ‘Spatially clustered count data provide more efficient search strategies in invasion biology and disease control’, *Ecological Applications* pp. 1–11.

- Stevens, M. C. A., Ray, G., Faulkner, S. C. & Le Comber, S. C. (2020), 'Investigating Sherlock Holmes: Using Geographic Profiling to Analyze the Novels of Arthur Conan Doyle', *Professional Geographer* **72**(4), 566–574.
- Stevenson, M. D. (2015), Geographic Profiling in Biology, PhD thesis.
- 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* **35**(8), 704–715.
- Struebig, M. J., Linkie, M., Deere, N. J., Martyr, D. J., Millyanawati, B., Faulkner, S. C., Le Comber, S. C., Mangunjaya, F. M., Leader-Williams, N., McKay, J. E. & St. John, F. A. V. (2018), 'Addressing human-tiger conflict using socio-ecological information on tolerance and risk', *Nature Communications* **9**(1), 3455.
- Sun, C. C., Fuller, A. K. & Andrew Royle, J. (2014), 'Trap configuration and spacing influences parameter estimates in spatial capture-recapture models', *PLoS ONE* **9**(2).
- Suzuki-ohno, Y., Inoue, M. N. & Ohno, K. (2010), 'Applying geographic profiling used in the field of criminology for predicting the nest locations of bumble bees', *Journal of Theoretical Biology* **265**(2), 211–217.
- Tench, S. A. (2018), Space-Time Modelling of Terrorism and Counter-Terrorism, PhD thesis, University College London.
- Thawornwattana, Y., Dalquen, D. & Yang, Z. (2018), 'Designing simple and efficient Markov chain Monte Carlo proposal kernels', *Bayesian Analysis* **13**(4), 1033–1059.
- The malERA Consultative Group on Diagnoses (2011), 'A research agenda for malaria eradication: Diagnoses and diagnostics', *PLoS Medicine* **8**(1).
- Thuiller, W., Richardson, D. M., Pyssek, P., Midgley, G. F., Hughes, G. O. & Rouget, M. (2005), 'Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale', *Global Change Biology* **11**(12), 2234–2250.
- Ugandan Ministry of Health (2015), Malaria Indicator Survey 2014-15, Technical report, Uganda Bureau of Statistics (UBOS) and ICF International, Kampala, Uganda, and Rockville, Maryland.
- Van Houtan, K. S., Pimm, S. L., Halley, J. M., Bierregaard, R. O. & Lovejoy, T. E. (2007), 'Dispersal of Amazonian birds in continuous and fragmented forest', *Ecology Letters* **10**(3), 219–229.

- Van Rossum, G. & Drake Jr, F. L. (1995), *Python reference manual*, Center for Mathematics and Computer Science Amsterdam.
- Ver Hoef, J. M. & Boveng, P. L. (2007), ‘Quasi-Poisson vs. Negative Binomial Regression: How Should We Model Overdispersed Count Data?’, *Ecology* **88**(11), 2766–2772.
- Verity, R., Stevenson, M. D., Rossmo, D. K., Nichols, R. A. & Le Comber, S. C. (2014), ‘Spatial targeting of infectious disease control: Identifying multiple, unknown sources’, *Methods in Ecology and Evolution* **5**(7), 647–655.
- Wilke, A. B. B., Vasquez, C., Mauriello, P. J. & Beier, J. C. (2018), ‘Ornamental bromeliads of Miami-Dade County, Florida are important breeding sites for *Aedes aegypti* (Diptera: Culicidae)’, *Parasites & Vectors* **11**, 1–7.
- Wilke, A. B. B., Vasquez, C., Medina, J., Carvajal, A., Petrie, W. & Beier, J. C. (2019), ‘Community Composition and Year-round Abundance of Vector Species of Mosquitoes make Miami-Dade County, Florida a Receptive Gateway for Arbovirus entry to the United States’, *Scientific Reports* **9**(1).
- Worton, B. (1989), ‘Kernel Methods for Estimating the Utilization Distribution in Home-Range Studies’, *Ecology* **70**(1), 164–168.
- Yeka, A., Gasasira, A., Mpimbaza, A., Achan, J., Nankabirwa, J., Nsohya, S., Staedke, S. G., Donnelly, M. J., Wabwire-Mangen, F., Talisuna, A., Dorsey, G., Kamya, M. R. & Rosenthal, P. J. (2012), ‘Malaria in Uganda: Challenges to control on the long road to elimination. I. Epidemiology and current control efforts’, *Acta Tropica* **121**(3), 184–195.
- Yoshikawa, T., Masaki, T., Motooka, M., Hino, D. & Ueda, K. (2018), ‘Highly toxic seeds of the Japanese star anise *Illicium anisatum* are dispersed by a seed-caching bird and a rodent’, *Ecological Research* **33**(2), 495–504.

Appendix A

Additional Work

The contents of this appendix will walk through two pieces of additional work that I have carried out that did not make it into the main body of the thesis. The first part of this appendix will expand upon a point made in chapter five that discussed a common error in ecological literature when describing the dispersal distances travelled by different organisms from a central source location. The second part will briefly touch upon a further development for geographic profiling that unifies the non-conjugacy of the models built in this thesis with the non-parametric framework of the Dirichlet process mixture (DPM) model, bypassing the requirement of specifying a number of source locations (K) prior to running the model.

A.1 Error in dispersal literature

As previously mentioned in chapter five, there exists a common error in ecological literature attempting to describe the distribution of dispersal distances of fauna or flora travelling from a central source location. In many of these studies, point-pattern data sampled over a two dimensional space is often reduced down to one dimension by binning data that fall within concentric circles of larger radii around said source location. Although the interval length remains consistent across each circle, the area associated with each bin increases. As a consequence, more data is expected to be counted within bins farther away from a source. A visual representation of this problem can be seen in Figure A.1.

In his thesis, Stevenson (2015) highlighted this problem and walked through the process in which two dimensional data should be transformed to correct this error. In addition to specifying the solution, a few cherry-picked examples were critiqued, where data were extracted from two studies

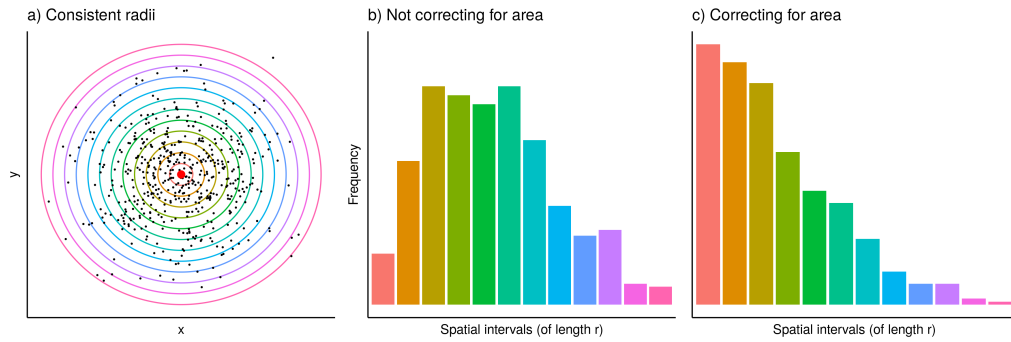


Figure A.1: A simulated set of point-pattern data (a) transformed into histograms that (b) correct and (c) do not correct for the area of each bin.

to reproduce the histograms making the mistake whilst also producing corrected histograms (Alonso et al. 1998, Forero et al. 2002). This work successfully demonstrated the error within the literature, however it focussed too closely on a handful of studies that took place many decades ago.

My contribution to resolving this problem is in the form of a systematic search of literature to identify if studies are still making the error described above. Web of science (<https://www.webofknowledge.com>) was searched for possible literature making this transformation error following the key terms and steps shown in Figure A.2. The advantage of following this procedure ensures that no particular study is being singled out, whilst also ensuring that only recent studies are considered.

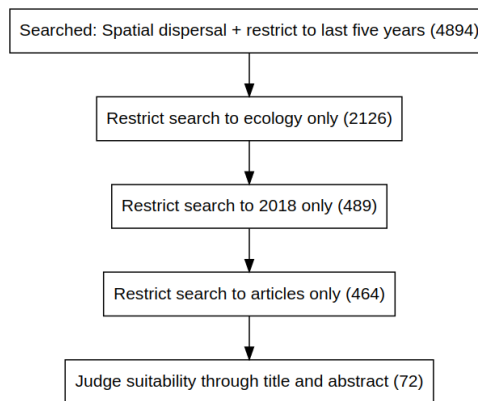


Figure A.2: The search and filtering process for identifying suitable literature. The number of studies in each category is shown in brackets.

Each of the 72 studies obtained via the protocol in Figure A.2 were read in full to identify if the error in obtaining dispersal distributions was present. Of these 72 studies, 8 papers (11%) reported point-pattern data in the form of a histogram that did not correct for bin size (Browne

et al. 2018, Chadoëuf et al. 2018, Hall et al. 2018, Moraes et al. 2018, Morton et al. 2018, Ramos et al. 2018, Sinclair et al. 2018, Yoshikawa et al. 2018).

These studies should be used as proof of concept that this error is still present within the literature and a thorough investigation should be carried out via different platforms and years to fully determine the extent with which this error is present but also, to what extent making the correct transformation impacts the results of the original study.

A.2 Non-conjugate Dirichlet process mixture model

Within this thesis, the procedure for determining how many source locations (K) best describe the data analysed was determined by running many models each with a different value of K and then making use of the deviance information criterion when choosing the best of this set of models. The finite mixture models built within this thesis present a problem shared by many other clustering algorithms; the necessity of choosing the number of clusters prior to running a model. As stated in Verity et al. (2014), a Dirichlet process mixture model is a powerful tool for clustering data as it does not require the specification of K prior to running the model. The Dirichlet process mixture model from Verity et al. (2014) can be obtained by re-writing Equation 1.13 - Equation 1.15 as:

$$x_i \sim f(\theta_{c_i}), \tag{A.1}$$

$$\theta_{c_i} \sim \mathcal{F}, \tag{A.2}$$

$$c_i \sim \text{CRP}(\alpha), \tag{A.3}$$

$$\alpha \sim \mathcal{H}. \tag{A.4}$$

In this instance, the allocation of data to source locations (c_i) no longer relies on a multi-nomial distribution, but is governed by the Chinese restaurant process (CRP) with a single concentration parameter α , with prior \mathcal{H} .

The MCMC algorithm built to implement the DPM model in Verity et al. (2014) relies on conjugacy between conditional likelihoods and their priors, indicating that said priors must be of a specific form. In collaboration with Bob Verity, I implemented a Metropolis-Hastings within Gibbs sampler for a non-conjugate DPM model that estimates source locations and allocations of data points to sources (i.e. K). The MCMC algorithm itself was built following algorithm seven from (Neal 2000). The major change that is made here compared to the existing DPM model's

MCMC algorithm is that source locations are now updated via a Metropolis-Hastings step as opposed to a Gibbs sampling step. An example of the model implemented with a non-conjugate prior on source locations can be seen in Figure A.3.

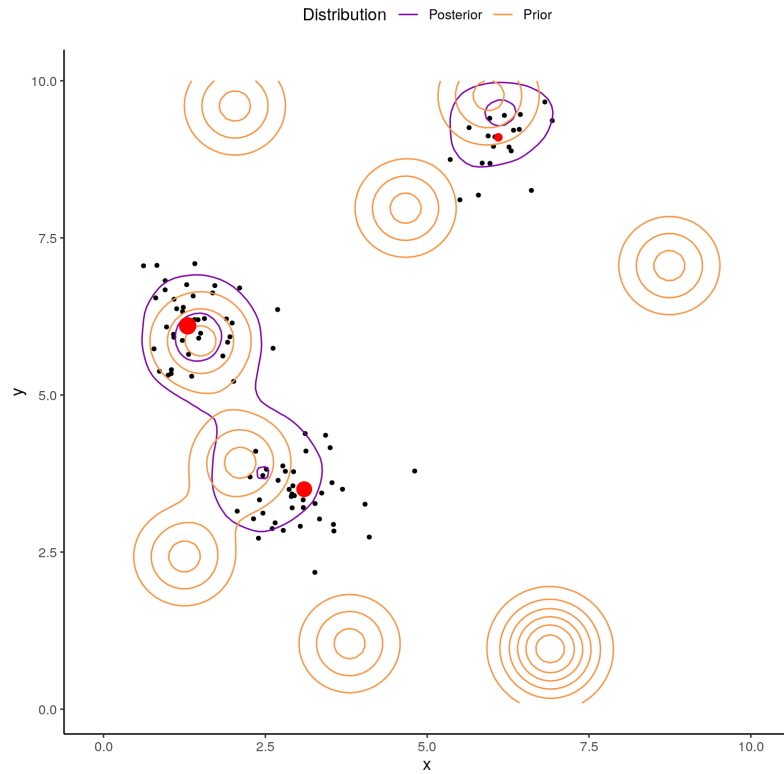


Figure A.3: An illustration of a simple non-conjugate Dirichlet process mixture model fitting source locations in Cartesian space. Contours represent the non-conjugate prior (orange) and posterior (purple) densities, where data points and source locations are black and red dots respectively.

Figure A.3 demonstrates that a non-parametric geographic profiling model employing a non-conjugate prior is possible and this section should guide future work that aims to unify the convenience of a Dirichlet process mixture model with the flexibility of the models associated with this thesis.

Appendix B

R Package: *silverblaze*

The *silverblaze* package for R was used to implement all of the models built within this thesis. Detailed tutorials describing how to build each model are available at <https://michael-stevens-27.github.io/silverblaze/>, but the landing page of each will be shown here.

B.1 Landing page

The landing page for *silverblaze* matches the repository's README.md file giving a brief description of the package's function in addition to useful links leading to the package license, a space for reporting bugs, links to source code, links to tutorials and the current status of any code quality checks.

silverblaze 1.0.0 Basic tutorials ▾ Advanced tutorials ▾ Functions

silverblaze

Version 1.0.0

Silverblaze offers a finite mixture model for Geographic Profiling. Through the package a user can

- infer source locations and dispersal patterns - common desirable parameters in geographic profiling
- make inferences via count, prevalence and point pattern data
- specify a spatial prior in source locations via shape files

Silverblaze is a toolbox for geographic profiling that utilises different MCMC algorithms (Metropolis-Hastings coupling and Gibbs sampling) to infer the above parameters. The package [RgeoProfile](#) uses a Dirchlet Process Mixture model (DPM) to solve the issue of multiple sources.

Links

Browse source code at <https://github.com/Michael-Stevens-27/silverblaze/>

Report a bug at <https://github.com/Michael-Stevens-27/silverblaze/issues>

License

[MIT + file LICENSE](#)

Developers

Bob Verity
Author, maintainer

Michael Stevens
Author

Dev status

build passing

build passing

B.2 Basic tutorials

The three main tutorials here cover the three main models built within this thesis. Each tutorial walks through the process of specifying and running a model that estimates parameters of interest, such as source locations, when analysing a set of spatial count, prevalence or point-pattern data.



A Poisson model for count data

Michael Stevens

2021-02-28

Source: vignettes/Poisson-basic-implementation.Rmd

Introduction

The following tutorial will introduce the user to the general structure required to run silverblaze when analysing count data. The protocol of silverblaze is to create one fundamental object, referred to in these tutorials as p , that describes the entire project. The structure of p consists of:

1. The data used to run the model
2. The parameter settings for the model
3. The output of running the model

Following this, the user will learn how to plot and diagnose the output of the model.

Contents

- Introduction
- Simulating data
- Creating a project
- Simple spatial prior
- Parameter sets
- Running the model
- Diagnosing the model
- Results



Binomial model and prevalence data

Michael Stevens

2021-02-28

Source: vignettes/Binomial-basic-implementation.Rmd

Introduction

The following tutorial will introduce the user to the general structure required to run silverblaze when analysing prevalence data. The protocol of silverblaze is to create one fundamental object, referred to in these tutorials as p , that describes the entire project. The structure of p consists of:

1. The data used to run the model
2. The parameter settings for the model
3. The output of running the model

Following this, the user will learn how to plot and diagnose the output of the model.

Contents

- Introduction
- Simulating data
- Creating a project
- Simple spatial prior
- Parameter sets
- Running the model
- Diagnosing the model
- Results



Point pattern data

Michael Stevens

2021-02-28

Source: vignettes/PointPattern-basic-implementation.Rmd

Introduction

The following tutorial will introduce the user to the general structure required to run silverblaze when analysing point pattern data. The protocol of silverblaze is to create one fundamental object, referred to in these tutorials as p , that describes the entire project. The structure of p consists of:

1. The data used to run the model
2. The parameter settings for the model
3. The output of running the model


Following this, the user will learn how to plot and diagnose the output of the model.

Contents

- Introduction
- Simulating data
- Creating a project
- Simple spatial prior
- Parameter sets
- Running the model
- Diagnosing the model
- Results

B.3 Advanced tutorials

The *silverblaze* package also offers a set of advanced tutorials for specifying and estimating parameters under more complex models. The first of which walks the user through a variety of ways to validate their models and MCMC output. The second focusses on the optimisation process to obtain heats for each chain via the Metropolis-Hastings coupling protocol. The models built in this thesis were defined over a two dimensional grid of cells and as such allow a user to specify the spatial prior of their choice on source locations and the next tutorial describes this process. The final two tutorials focus on estimating independent dispersal values and population sizes for each source location.

silverblaze 1.0.0 Basic tutorials ▾ Advanced tutorials ▾ Functions 

Model Choice and Validation

Michael Stevens

2021-02-28

Source: vignettes/model-validation.Rmd

Contents

- Introduction
- Model validity
- Model choice
- References

Introduction

If you have followed any of the basic tutorials for Silverblaze you would have already seen a few ways to validate the model that has been used. In this tutorial we will walk through an extensive list of ways that ensure the model is behaving as it should be. We are responsible for simulating the data in the basic tutorials, so a good way to know if the model is successful is if it returns the correct parameter values that were used to generate the data. Of course, this will not be the case in a real-world data set. So here we will be focussing on methods of model validity that are independent of parameter estimation. In addition to this, we shall also briefly discuss model choice. Silverblaze offers a finite mixture model for geographic profiling and we require a method for choosing the number of mixture components that best describe the data.

This tutorial will cover model validity:

- A healthy MCMC trace
- Geweke's metric for single chain MCMC convergence
- Auto-correlation
- Effective sample sizes
- Acceptance rates for proposed Metropolis-Hastings steps

and model choice:

- Deviance information criterion

silverblaze 1.0.0 Basic tutorials ▾ Advanced tutorials ▾ Functions 

Metropolis-Hastings Coupling

Michael Stevens

2021-02-28

Source: vignettes/MHCoupling.Rmd

Contents

- Introduction
- A difficult data set
- Poor mixing
- Metropolis-Hastings coupling
- Specifying a beta value for each heated chain
- Optimising beta values
- References

Introduction

Through the previous tutorials we have shown how parameters of interest associated with geographic profiling, such as source locations and dispersal patterns, can be inferred using count, prevalence and point-pattern data.

These examples were cherry picked such that the MCMC algorithm converged on the correct answer. This is not always the case though. In this tutorial we will show an example of a difficult data set that is a real challenge for the MCMC algorithm.

Complex Spatial Priors

Michael Stevens

2021-02-28

Source: vignettes/complex-spatial-priors.Rmd

Introduction

The geographic profiling models described in the previous tutorials all make use of a Metropolis-Hastings algorithm for updating source locations. As a result, we can impose any prior we like on source locations. So far, priors on source locations have assumed that the probability of observing a source at a specific location is either constant (inside the shapefile) or zero (outside the shapefile). When we update source locations via the Metropolis-Hastings algorithm any proposed value that is outside the shapefile is immediately rejected. This differs from the conventional geographic profiling model (see Verity et al. (2014) and Faulkner et al. (2018)) that implements the shapefile post-hoc. This tutorial will walk through different options for spatial priors in the silverblaze package. Let's start with a refresher: a uniform prior based on simulated data. The prior on source locations is defined over a grid of cells, hence we create a uniform prior via the `raster_grid()` function.

Contents

- Introduction
- Bi-variate normal prior
- Kernel density prior
- Using shapefiles
- A non-uniform spatial prior
- Converting a shapefile to a raster
- A non-uniform spatial prior in a GP project
- References

Variable Sigma

Michael Stevens

2021-02-28

Source: vignettes/Variable-sigma.Rmd

Motivation

A variable sigma model is available within the Silverblaze package but is yet to be discussed in great detail. The literature on geographic profiling has thus far assumed that the dispersal associated with an offender, an invasive species, or infectious disease is the same across source locations. Ratcliffe (2006) makes it clear that this is not always the case. An offender that commits crime around a work place will be restricted by time constraints to return to work within a certain time. This compares to more relaxed time constraints when committing crimes around a home. So we might expect a tight dispersal around a workplace but a wider dispersal around a home.

In this tutorial, we generate a couple of toy examples to investigate the following questions:

- Can the model successfully return different dispersal (sigma) values for different source locations?
- How well does the model perform when there is overlap in the dispersal ranges of sources with different sources?

Contents

- Motivation
- Simple two source example
- Overlapping sources and variable dispersal

Variable Expected Population

Michael Stevens

2021-02-28

Source: vignettes/Variable-population.Rmd

Motivation

An independent expected population model is available within the Silverblaze package but is yet to be investigated. In this tutorial we will generate a couple of toy examples to investigate the following questions:

- Can the model successfully return different population sizes for each source locations?
- How well does the model perform when there is overlap in the dispersal ranges of sources with different sources?

Contents

- Motivation
- Simple two source example

B.4 Function list

The remaining documentation for *silverblaze* lists the various R functions within the package, their arguments and examples for easy implementation.

silverblaze 1.0.0 Basic tutorials ▾ Advanced tutorials ▾ **Functions** ↻

Reference

Main

Main functions for running different models

sim_data()	Simulate data
bind_data()	Bind data to project
new_set()	Create new parameter set
delete_set()	Delete parameter set
optimise_beta()	Optimise beta values for MCMC runs
run_mcmc()	Run main MCMC
get_output()	Get specified output from project
get_hitscores()	Get hitscores
gini()	Calculate Gini coefficient

Contents

- Main
- Plotting
- Mapping
- Misc

silverblaze 1.0.0 Basic tutorials ▾ Advanced tutorials ▾ **Functions** ↻

Simulate data

Source: R/data.R

Simulate data under a specified model

```
sim_data(
  sentinel_lon,
  sentinel_lat,
  sentinel_radius = 0.1,
  K = 3,
  source_weights = NULL,
  source_lon_min = -0.2,
  source_lon_max = 0,
  source_lat_min = 51.45,
  source_lat_max = 51.55,
  source_lon = NULL,
  source_lat = NULL,
  sigma_model = "single",
  sigma_mean = 1,
  sigma_var = 0.1,
  expected_popsiz = 100,
  data_type = "counts",
  test_rate = 5,
  N = 150,
  dispersal_model = "normal"
)
```

Contents

- Arguments
- Examples

Arguments

sentinel_lon vector giving longitudes of sentinel sites.

sentinel_lat vector giving latitudes of sentinel sites.