Simultaneous Reconstruction of Spatial Frequency Fields and Field Sample Locations

Ross Anthony Haines

St John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2016

Abstract

Classically, spatial smoothing methods such as kriging estimate smooth interpolating fields for features measured at well-located points. In this thesis, we make a simultaneous reconstruction of interpolating spatial fields and measurement locations.

We give models, and sample-based Bayesian inference, for estimating locations of dialect samples on a map of England. The method exploits dialect-based spellings to locate these samples.

The data are feature vectors extracted from written dialect samples. Just a fraction of the feature vectors ('anchors') have an associated spatial location. When coupled to a prior for the smoothly varying feature field, and the anchor texts, the unlocated feature vectors are jointly informative of their own location and the feature fields.

The dataset is large, but sparse, since a given word has a large number of variant spellings which may appear in just a few documents. We report an analysis including Bayesian model fitting and validation on a large and representative subset of the data.

The thesis has two main aims - to provide statistical tools for the linguists who collected the data, and to meet the computational and inferential challenge of simultaneously locating large numbers of feature vectors. The results presented in this thesis show that we have largely succeeded in meeting these challenges.

Acknowledgements

My doctoral studies at the University of Oxford have, at times, been immensely frustrating, and at others, extremely rewarding. Undoubtedly, they have been the most challenging experience of my life. It would be remiss to overlook the contributions made by friends and family towards helping me reach the finish line. Unfortunately, there is not nearly enough space here to acknowledge everyone, so to those not explicitly named, thank you for your help.

First and foremost, I would like to extend my gratitude towards the Woolf Fisher Trust for their generous support, which enabled this wonderful opportunity to undertake the research presented in this thesis. It really has been a life-changing experience, and opening the scholarship award letter five-and-a-half years ago still remains one of my happiest memories.

Writing this thesis would not have been possible without the guidance and steadfast support of my supervisor, Professor Geoff Nicholls. Thank you especially for your constant generosity of time, especially whilst juggling the Departmental reins. Thanks also to Professor Michael Benskin for providing the dataset used throughout this project, and for the invaluable advice imparted in our discussions.

Thanks to all the friends I've made during my studies for making my time in Oxford so enjoyable, and particularly, to Dr Tom Hills, Dr Madura Jayatunga, Dr Ian Ashpole, Dr Struan Murray, Dr Gaëlle Coullon, Dr Matt Lewis, Dr Richard Lau, Dr Ed Greening, Dr Adrastos Omissi, Dr

Leigh Paton, John O'Rourke, and Zack Goldman for making College life so entertaining. Special thanks to Dr Liv Faull, who was a source of support throughout (as well as an incredible birthday cake), and to my partners in crime Dr Ant Hibbs and Tobias Wieczorek.

Oxford also offered great opportunities outside of the academic sphere, and I was fortunate to have many wonderful sporting experiences with some great people. Thanks to everyone from SJCCC, OUSRC, OUCC Authentics and Blues, and Oxford MCCU for all the fun memories, and especially to Ian Madden and Tara McSweeney for their constant hospitality and friendship, and the world's best backyard cricket ground.

Thanks to my wonderful family (Mum, Dad, Andrew & Anna-Louise, and my grandparents) for their constant support, encouragement and love throughout my studies. Thanks for helping me through the tougher times - you all really did go the extra mile for me. Words cannot express my appreciation.

For helping to guide me to the completion of this journey, and for her wonderful laugh, I am extremely grateful towards Dr Jan Cosgrave. I really am incredibly lucky to have you by my side, and cannot wait for more laughter and more adventures with you.

Contents

CHAPTER 1

INTRODUCTION

Spatial statistics presents substantial methodological and computational challenges. Progress on full likelihood-based inference has been limited by the complexity of spatial models and our consequent inability to deal with intractable likelihoods. However, improvements in computer hardware and the development of new efficient Monte Carlo algorithms has brought previously intractable problems within reach. The analysis presented in this thesis is in this class.

A common problem in spatial statistics is to reconstruct a spatial field from field values observed at scattered locations. In some settings, like in Cressie & Kornack (2003), there is some variability in the specified locations at which field observations are made.

We consider a related problem, in which the locations where field values are observed are of two kinds. At 'anchor' points, the location for the associated field measurement is perfectly observed. At other non-anchor ('floating') points, there is no direct location information at all. The desire is to reconstruct the locations of

these floating points. The spatial field is missing data of secondary importance.

Such a problem arises for a dataset collected from mediaeval manuscripts. This dataset is presented in *A Linguistic Atlas of Late Mediaeval English* (McIntosh et al., 1986), which is henceforth referred to as 'the Atlas'. During the late mediaeval period, the spoken form of the English language exhibited considerable regional variation (McIntosh, 1987). The Atlas showcases this variation.

The Atlas was the result of the examination of over a thousand manuscripts originating from across England and Wales between ca. 1325-1450. Some could be linked to specific places on non-linguistic grounds (McIntosh, 1987; Laing, 1991), however, these anchor texts comprise only a small percentage of the manuscripts. The origins of the large remaining body of manuscripts were estimated by McIntosh et al. (1986) using a non-statistical method of elimination. This method cannot provide a rigorous quantitative measure of uncertainty in the estimated locations, nor can goodness-of-fit be assessed. One of the two main aims of this project is to provide reliable statistical tools for these researchers.

We develop models and a Bayesian inference procedure to locate the manuscripts from the Atlas, using spatially interpolated maps of dialect. Our models allow us to consider the validity of the placings found using the non-statistical technique employed by McIntosh et al. (1986). Measures of precision were quoted when this technique was used to locate new manuscripts. Our models also allow us to check if these

precision measures are reasonable.

Very roughly, our approach is to use the anchor texts to inform "dialect fields". The non-anchor texts can then be placed at locations where their dialect patterns are likely. This is a simplification, because the non-anchor texts themselves feed back to inform the field. An advantage of our approach is that our modelling makes full use of all information in the data, and leads to quantitative measures of uncertainty for the estimated locations. The technique of McIntosh et al. does not give such rigorous measures.

To some extent, we may choose this set of anchor texts which inform the "dialect fields", depending on how we frame the applied problem. There are two frameworks we consider:

1. In the first, we locate all of the non-anchor texts. Here, the anchor texts are the 'true' anchors, i.e. those which could be linked to specific places on non-linguistic grounds.

2. In the second, we trust the provided Atlas locations for the non-anchor texts, and locate a new text(s). Therefore, all of the texts in the Atlas are anchors here: both those located on non-linguistic grounds and those located using McIntosh et al.'s non-statistical technique.

We consider these two frameworks as they let us consider two separate issues. The first allows us to calculate location estimates to compare to those provided in

the Atlas, and thus check for consistency. On the other hand, the second framework allows us to check the claimed precision in the Atlas associated with locating new manuscripts.

Whilst the second framework presents a simpler problem than the first, it is still a difficult problem due to the unobserved "dialect fields", which are missing data of high dimension. The first framework suffers all the problems of the second, and more, as the target distribution is more diffuse, and we have less informative data available. Further, there are likely to be multiple "modes" of reconstruction, and even "phases" characterised by different parameter values and different levels of spatial order.

In both cases, modelling the Atlas data is a challenging problem, with the magnitude of the data and the sparseness of non-zero observations contributing to the complexity. We face challenges in carrying out Monte Carlo inference for a system with complex spatial correlation structure. It simply takes many Markov Chain Monte Carlo (MCMC) updates to bring this system to equilibrium. Meeting the computational and inferential challenge of simultaneously locating large numbers of manuscripts is the other of the two principal statistical aims of this project.

1.1 Layout

In this thesis, we develop a model and Bayesian inference procedure to locate the Atlas manuscripts on a map based on their dialect. We apply parallel sample-based Bayesian inference to new spatial models for the Atlas data. We consider analyses

under both frameworks of the applied problem described above.

We begin in Section 1.2 with a review of the relevant literature. In particular, we discuss modelling approaches taken in similar research, and explain how our methods differ, and why.

The raw data we have been provided by collaborating scientists is introduced in Section 1.3. A registration of the data in presence/absence form is given in Section 1.3.1. This section also gives some exploratory data analysis, and defines the terms of the key scientific questions. A secondary dataset, defining hierarchical structure in the primary dataset, is then described in Section 1.3.2. We introduce the new data and explain how the datasets may be brought together to form a coarsened dataset.

Our ℓ_1-Dirichlet zero-inflated model for the linguistic data (primary or coarsened) is described in Chapter 2, with the likelihood presented in Section 2.1.1. Prior distributions for the parameters of the model are given in Section 2.1.2. The joint posterior distribution is given in Section 2.1.3, alongside conditional posterior distributions for each of the parameters.

We then extend our model to the ℓ_1-Dirichlet outlier model, which features explicit modelling of outlier manuscripts and spellings. This model is outlined in Section 2.2. An adjustment to the construction and smoothing of the dialect fields is then considered in Section 2.3, leading to our ℓ_2-logistic models.

Chapter 3 contains the Markov Chain Monte Carlo (MCMC) algorithms used

to fit our models to the data. These include reversible updates for all parameters. The algorithm for the core model is provided in Section 3.1. The algorithm for the extended model allowing for outliers is then detailed in Section 3.2. Finally, the algorithm for the model with the new dialect field construction is provided in Section 3.3.

With synthetic data, we then compare theoretical distributions for each parameter to those found using the samplers from each algorithm. These quick checks, provided in Section 3.4, are not intended as rigorous proof that the algorithms sample from the correct distributions, rather to provide some peace-of-mind.

Chapter 4 explores alterations to the algorithms from Chapter 3 to utilise more advanced MCMC methods. Firstly, Section 4.1 considers parallel computation as a means to improve the MCMC algorithm, specifically through improving the mixing of the dialect field parameters. We then consider a Metropolis adjusted Langevin algorithm (MALA) based approach in Section 4.2.

In Chapter 5, we assess the performance of the overall inference (Bayesian framework, models and samplers) as a method for locating manuscripts. Analyses in Section 5.2 consider the estimation of the locations of origin of multiple manuscripts, with all others fixed in place. We then extend the scope of analysis in Section 5.3, using the full (coarsened) linguistic data to locate all of the non-anchor texts.

We compare our model-based methods for locating manuscripts to a variety of

other statistical methodologies in Chapter 6. We first consider location estimates derived from k-Nearest Neighbour algorithms in Section 6.1. The use of classification trees to estimate a manuscript's location of origin within the lattice is then explored in Section 6.2. Finally, in Section 6.3, we explore multi-dimensional scaling methods for location estimation.

The thesis concludes in Chapter 7 with a summary of results and the conclusions we draw from them. Some avenues for future work are considered, based on obstacles identified in Chapter 5.

Appendix A provides additional sampler checks akin to those in Section 3.4, and Appendix B contains example convergence diagnostics results for analyses presented in Chapter 5. Finally, Appendix C provides a derivation of the gradient used in our MALA-based methods described in Section 4.2.

1.2 Literature Review

In this section, we refer to pertinent work in MCMC and Bayesian inference that we draw upon. We then very briefly consider existing location estimation work where the location information is known but with mild associated uncertainty. Finally, we review in more detail work considering similar location estimation problems to ours, whereby at least some locations are completely unknown.

1.2.1 Markov Chain Monte Carlo and Bayesian Inference

In this thesis, we develop a Bayesian inference procedure to estimate the locations of origin of Atlas manuscripts. Gelman et al. (2014) describes Bayesian inference as 'the process of fitting a probability model to a set of data, and summarising the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations.' This probability model for the data, $\pi(y \mid \theta)$, and the prior distributions $\pi(\theta)$ (capturing our knowledge about the parameters θ of the model, without reference to the data) are combined using Bayes' theorem to give the posterior distribution

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\int \pi(y \mid \theta)\pi(\theta)\, d\theta}, \tag{1.1}$$

upon which our inference is based. For further details, see (for example) Gelman et al. (2014) and Robert (2007).

Evaluating the integral in Equation 1.1 has, in the past, been the primary obstacle to the implementation of Bayesian inference. Since the mid-twentieth century, the development of Markov Chain Monte Carlo (MCMC) methods for drawing samples from the posterior distribution has brought previously intractable problems within reach (like the problem considered in this thesis).

MCMC sampling methods involve constructing a Markov chain whose stationary distribution is the posterior $\pi(\theta \mid y)$. Standard MCMC simulation methods which we use in this thesis are the Gibbs sampler and the Metropolis-Hastings algorithm.

The Gibbs sampler was introduced by Geman & Geman (1984) for image processing. This method allows us to sample from the joint posterior distribution $\pi(\theta_1, \ldots, \theta_d \,|\, y)$ using the full conditional distributions $\pi(\theta_i \,|\, y, \theta_{-i})$ for each parameter θ_i, $i \in \{1, \ldots, d\}$, where θ_{-i} denotes all the other parameters. At each iteration of the algorithm, the θ_i are updated sequentially using these conditional distributions.

In the Metropolis-Hastings algorithm (Metropolis et al. (1953), Hastings (1970)), we sample a proposed value θ' from the proposal distribution $q(\theta' \,|\, \theta^{t-1})$ at each iteration t of the algorithm, and an acceptance/rejection rule to determine θ^t. With probability

$$\alpha(\theta', \theta^{t-1}) = \min\left(1, \frac{\pi(\theta' \,|\, y)/q(\theta' \,|\, \theta^{t-1})}{\pi(\theta^{t-1} \,|\, y)/q(\theta^{t-1} \,|\, \theta')}\right), \tag{1.2}$$

the proposal is accepted ($\theta^t = \theta'$); otherwise $\theta^t = \theta^{t-1}$.

For the majority of analyses in this thesis, we adopt random-walk Metropolis-within-Gibbs sampling. This standard sampling method is a hybrid of the two methods described above, using Metropolis-Hastings updates (with normal proposal distributions centred on the current value) at each step of the Gibbs sampler.

We also include analyses using a more advanced gradient-descent MCMC sampling algorithm, based on work presented by Roberts & Tweedie (1996) and Roberts & Rosenthal (1998). The Metropolis Adjusted Langevin Algorithm (MALA) constructs proposals based upon the Langevin diffusion, such that

$$q(\theta' \,|\, \theta^{t-1}) = \theta^{t-1} + \sigma\epsilon_t + \frac{\sigma^2}{2}\nabla \log\left(\pi(\theta^{t-1} \,|\, y)\right), \tag{1.3}$$

where ϵ_t are independent and identically distributed standard normal variates, and $\nabla \log\left(\pi(\theta^{t-1} \mid y)\right)$ is the gradient of the log-posterior distribution.

1.2.2 Location Estimation with Mild Uncertainty

In some settings where one wishes to estimate spatial fields, there is some uncertainty in the specified locations at which field observations are made. Such work differs from ours, because the missing locations are tightly constrained, and not of primary interest in the inference, unlike in our application. For a review of such work, we refer the reader to Cressie & Kornack (2003) for more detail.

1.2.3 Location Estimation with Completely Unknown Locations

We found relatively few applications of Bayesian or likelihood-based inference for completely unknown location information. In this section, we review the modelling approaches used in such applications. We describe where these approaches are similar to our model-based methods, as well as how they differ from ours. Finally, we give an overview of other peripherally related work.

A problem analogous to ours is considered in the series of papers by Wasser et al. (2004, 2007, 2008, 2009, 2015). The problem encompasses Bayesian location estimation for samples of poached elephant ivory, based on genetic data displaying spatial variation. Like the linguistic data from the Atlas, this genetic data divides into anchor and non-anchor points. The scale of the problem in this work is quite

different to ours, as they have far more anchor points than us (1350 versus 300), and far fewer non-anchor points (28 versus 1211).

In a similar fashion to our "dialect-field" approach, Wasser et al. estimate an underlying field of parameters (frequencies $\theta_{j\ell k}$ of allele j at locus ℓ for sampling location k) which are smoothed to relate the parameters from neighbouring areas. There are far fewer of these parameters to estimate than in our work, with only 28 sampling locations k, and two alleles j at each of the 16 loci ℓ.

A parameter transformation is applied to these frequencies $\theta_{j\ell k}$, such that

$$\tilde{\theta}_{j\ell k}(\theta) = \frac{\exp(\theta_{j\ell k})}{\sum_{j'} \exp(\theta_{j'\ell k})}. \tag{1.4}$$

We explore a similar parameter transformation when developing our ℓ_2-logistic models in Section 2.3, as this approach allows us to remove the summation constraint present in the ℓ_1-Dirichlet models we develop in Section 2.1. We find, however, that the ℓ_1-based models are better.

Spatial smoothing is performed by Wasser et al. on $\theta_{j\ell k}$ via a joint normal distribution with a common mean, and the transformed parameters $\tilde{\theta}_{j\ell k}(\theta)$ are used in modelling to estimate the origins of the samples from unknown locations. These locations are estimated in two-dimensional coordinate space, unlike our estimated locations, which are cell-references within a lattice overlaid onto the geographic space.

Wasser et al. (2004) considered estimating a single location at a time, whereas in Wasser et al. (2007) and subsequent work, the authors locate multiple samples

simultaneously. Unlike our work, however, this estimation is not done jointly with the estimation of the underlying parameter field. Thus, the multiple samples do not feed information back into the frequency fields. It is worth nothing that there is information about the frequency field parameters even in the ivory samples of unknown origin, but without joint location estimation with the frequency field, minimal advantage can be taken of this information.

The model developed by Wasser et al. has been used directly in a variety of work with different data, generally for similar applications within the field of conservation biology; for example, Pope et al. (2007), Ghobrial et al. (2010), Mondol et al. (2014) and Puckett & Eggert (2016).

Other work has also considered exploiting underlying parameter fields as a means towards location estimation. Yang et al. (2012) and Yang et al. (2014) model allele frequencies f_j as logistic functions of locations x, with slope parameters (a_j, b_j):

$$f_j(x) = \frac{1}{\exp(-a_j^T x - b_j) + 1}. \tag{1.5}$$

As noted by the authors, these functions are not able to successfully model cases where the spatial distribution of the frequencies are complex, for example when there is multimodality. As this often proves to be the case with the linguistic data from the Atlas, such an approach would have limited value with our data.

Like the linguistic data in the Atlas, the genetic data used for modelling in Yang et al. (2012) divides into anchor and non-anchor points. The authors consider two

different modelling frameworks for their problem, similar to the ones we described earlier on page 3. Like in Wasser et al. (2004), and unlike our work, the frequency fields are estimated separately from the coordinate locations of unknown origin.

Ranola et al. (2014) consider the same problem (with the same genetic data) as Yang et al. (2012). Unlike Yang et al. (2012), the authors divide the region of interest into pixels, and estimate allele frequencies for each. This approach more closely mirrors that taken in this thesis, whereby we overlay a rectangular lattice onto the map of England. The authors smooth the frequency field by relating values in neighbouring pixels with an ℓ_2-penalty. Unlike in this thesis, the authors use a neighbourhood structure including pixels sharing corners (with lower weights).

Bradburd et al. (2016) also estimate allele frequency surfaces as a means to location estimation. The authors use a Bayesian inference procedure, and spatial smoothing is performed on the allele frequencies such that the covariance between frequencies decays exponentially with a power of distance. Despite these similarities to our work, the problem considered by Bradburd et al. differs, as the authors seek to infer locations on a population-level. Further, the 'locations' inferred are not geographical; rather, the distance between the 'geogenetical' locations of two samples is sought to be proportional to their genetic differentiation.

Amos & Manica (2006) and Elhaik et al. (2014) consider a similar problem to Yang et al. (2012) and Ranola et al. (2014), but adopt a more straightforward analysis

approach, using a series of regressions to find the location of best fit for the points with unknown origins.

Motivated by Wasser et al. (2004), Giorgi & Diggle (2014) also consider inference for multiple missing locations when a spatially varying phenomenon has been measured at these missing locations, as well as at known ones. The modelling approach taken utilises a mixture of normal distributions. The scale of the applications presented differs from that which we face, given the authors estimate only one or two completely unknown locations, using data from between 100 and 300 anchor points. Unlike the approach taken in this thesis, the authors do not include the observations with unknown locations in the likelihood for parameter estimation, though they do note this as an interesting avenue for further exploration.

Eisenstein et al. (2010) consider a similar modelling problem to ours, using linguistic data derived from Twitter[1]. Like the Atlas data, this dataset is divided into anchor points with known origins (created by Twitter users from GPS-enabled devices), and non-anchor points with unknown origins. The authors also consider prediction of the location of these unlabelled non-anchor documents; however, the linguistic data available differs from the Atlas data, leading to a different approach. Latent 'topics' (similar to our 'items') are estimated, with the probability of topic usage varied spatially, and words assigned probabilities to belong to the regional-version of each topic. Spatial clustering of topics is assumed, and the spatial distribution of each topic is

[1] http://www.twitter.com

modelled as bivariate normal.

The modelling of latent topics in Eisenstein et al. (2010) follows a similar approach to Blei et al. (2003), whose 'Latent Dirichlet Allocation' (LDA) hierarchical model clusters co-occurring words into topics. This model assumes that a document is a 'bag-of-words', in which the the order of words can be neglected. The Atlas dataset does not specify the order that the 'forms' (spellings) of the 'items' (words) are used in, so such an assumption is implicit in our work.

Inferring the geographical origins of textual or image-based content posted on social media platforms (like Twitter, Facebook[2] and Flickr[3]) is an area of research that has gathered much attention in the literature since the work of Eisenstein et al. (2010). Some of this work directly draws from the methods detailed in Blei et al. (2003) and Eisenstein et al. (2010).

Examples of textual content location are Wing & Baldridge (2011), Eisenstein et al. (2011), Kinsella et al. (2011), Hong et al. (2012), Ikawa et al. (2012), Roller et al. (2012), Schulz et al. (2013), Intagorn & Lerman (2014) and Priedhorsky et al. (2014). Image location problems are considered by Hays & Efros (2008), Serdyukov et al. (2009), Crandall et al. (2009), Chen & Grauman (2011), Laere et al. (2012), Hauff & Houben (2012), Laere et al. (2013) and OHare & Murdock (2013).

Similarly to the Atlas linguistic data, the datasets used in these sets of work divide

[2] http://www.facebook.com
[3] http://www.flickr.com

into geotagged content with known origins, and other content with unknown origins to be estimated. The papers considering image location often exploit textual image tags, so many of those listed above use linguistic data for location estimation. Like in Eisenstein et al. (2010), however, this linguistic data differs from the Atlas data, leading to different location estimation approaches to ours.

These approaches vary in complexity. Simple methods are used in Hays & Efros (2008), Wing & Baldridge (2011), Ikawa et al. (2012) and Roller et al. (2012), for example, with new documents assigned the location of the 'closest' anchor (based on some distance-metric). We explore related methods in Chapter 6 to benchmark the performance of our more complex model-based methods.

More complex model-based approaches are taken in other papers in this set, like Hong et al. (2012) and Priedhorsky et al. (2014). Often, these are based around a mixture of normal distributions, like in Eisenstein et al. (2010). Some, like Serdyukov et al. (2009), use spatial smoothing to relate features to neighbours. Hong et al. (2012) use zero-mean Laplace prior distributions over certain parts of their model, which are similar to the prior distribution used for the dialect-fields in our ℓ_1-Dirichlet models. None of the work, though, takes a particularly close route to solving the applied problem as ours. Interestingly, Schulz et al. (2013) do take a similar approach to the 'fit-technique' used in the Atlas by McIntosh et al. (1986), with potential regions of document origin successively overlaid to give the final result.

Another popular area of work relating to social media and location estimation is that of inferring the home locations of the social media users. Approaches are typically content-based, using the data provided in the user's social media profile and the content they share; or network-based, using the locations of other users in the user's social network. Cheng et al. (2010), Hecht et al. (2011), Chandra et al. (2011), Dalvi et al. (2012), Chang et al. (2012), Ahmed et al. (2013), and Han et al. (2014) take a content-based approach, whereas Backstrom et al. (2010), Abrol & Khan (2010), Davis et al. (2011), Sadilek et al. (2012), Jurgens (2013), McGee et al. (2013), and Rout et al. (2013) take a network-based approach to the problem. Li et al. (2012) takes a both content- and network-based approach.

The data used in these problems again divides into anchor points with known origins (other users in the network), and points (users) with unknown origins to be estimated. As before, the complexity of the approach taken varies from basic methods (e.g. McGee et al. (2013) train a decision tree to separate users into groups likely to live in proximal locations), to more complex methods (e.g. Ahmed et al. (2013) use a hierarchical Bayesian model based on normal and Dirichlet distributions), but are universally quite different to our spatial-smoothing modelling approach.

Feature Selection

The problem of feature selection, whereby a subset of location-indicative data is selected for use in location estimation, is considered in work such as Yang & Pedersen

(1997), Bo et al. (2012), Laere et al. (2014) and Ranola et al. (2014). Some of this work is linguistic in nature; for example, Han et al. (2014) consider a variety of methods to select 'location-indicative words', including using the compactness of the geographical spread of a word.

The authors' motivations for pursuing such work are often applicable to our problem. For example, Ranola et al. (2014) believe the majority of their vast genetic dataset is uninformative, and use a form of feature selection to reduce this dataset (twentyfold) before modelling. This eases the considerable computational burden of analysing their data.

We take a different approach in this thesis, but for similar reasons. We utilise a secondary dataset to coarsen the Atlas data in Section 1.3.2, with similar words merged together. Our approach is not dissimilar in nature to stemming algorithm work (begun by Lovins (1968)), in which words with the same root/stem are reduced to a common form (e.g. 'divers', 'diving' and 'dived' are reduced to 'div'). However, we do not use an algorithm to infer which words should be merged; rather, this information has been provided by the authors of the Atlas.

We discuss feature selection further in Chapter 7.

Location Estimation using Principal Components Analysis

A body of work peripherally related to ours concerns the use of Principal Components Analysis (PCA) for location inference. Work in this area includes Lao et al.

(2008), Novembre et al. (2008), Tian et al. (2008), Chen et al. (2009), Price et al. (2009), Xing et al. (2009), Bryc et al. (2010), Engelhardt & Stephens (2010), Xing et al. (2010), Liu et al. (2013) and Petkova et al. (2015). Although this work considers the problem of estimating locations, this location information is often known, and PCA is applied to attempt to recover the known spatial structure. Furthermore, PCA is not based on an explicit probabilistic model for the spatial structure, and it is not clear how to relate the estimated structure to geographical coordinates (or regions). We briefly consider this avenue in Section 6.3, but were unable to reproduce anything resembling the spatial configuration provided in the Atlas.

Assignment Methods & Bayesian Clustering

A popular area of research regards the use of assignment methods and Bayesian clustering to assign individuals to a population based on feature vectors. In some of this work, the populations considered have distinct regions of origin, meaning that the population assignment can be thought of as a coarse location estimation problem, and distantly related to our work. The seminal work in this area is Pritchard et al. (2000), in which a Bayesian model is developed to assign individuals to a set of populations (based on genetic data). More recently, Gopalan et al. (2015) developed an algorithm to fit this model to massive genetic datasets using variational inference.

For an overview of the vast body of work around population assignment and Bayesian clustering, we refer the reader to Guillot et al. (2009).

1.3 Data

The manuscripts examined in the Atlas are divided into 'literary texts' and 'documents'. McIntosh et al. (1986) describe 'documents' as covering "*legal instruments, administrative writings, and personal letters: the type of material that is calendared by historians, likely to be of known date and local origins*", whereas the 'literary texts' include "*imaginative and discursive writings, regardless of their quality as literature... less obviously, perhaps, they number such things as translations of the Bible, glosses, paraphrases of the Psalms, medical recipes, charms*".

The authorship of these manuscripts is generally unclear, as "*Middle English scribes are self-effacing and for the most part anonymous*" (McIntosh et al., 1986). In the twelfth- and thirteenth-centuries, when writing in English was uncommon, writing was essentially a clerical accomplishment, and lay literacy was uncommon (McIntosh et al., 1986). Moving forward, "*the more English manuscripts... were produced, the more likely it is that copying texts became the habitual occupation of a large number of scribes*" (McIntosh et al., 1986).

For further detail regarding the manuscripts contained within the Atlas, we refer the reader to the Introduction of Volume 1 of the Atlas (McIntosh et al., 1986)[4].

[4]This can be accessed electronically via http://www.lel.ed.ac.uk/ihd/elalme/intros/atlas_gen_intro.html

To collect the raw Atlas data from these manuscripts, McIntosh et al. (1986) first compiled a list of items (words and grammatical concepts). Each item comes in a variety of forms (spellings). Different forms of a word are equivalent in their function and/or meaning but may vary significantly in appearance (McIntosh, 1987). For example, the singular noun 'brother' is an item, with distinct forms including 'BROTHER', 'BROTHer', 'BROyer', 'BROyEr', 'BROTHYR', and 'BRUTHIR'.

Note that in the data provided, manuscript symbols (including extra symbols and accented letters) have been mapped into the modern English alphabet using lower and upper case letters. Thus, 'E' and 'e' (for example) represent distinct symbols in the manuscript, rather than upper and lower case letters.

For each manuscript, the dialect of the scribe who wrote it was then characterised in the following way:

(i) For each item on the list, the forms present in the manuscript were recorded.

(ii) For each present form, an indicative measure of frequency was also recorded. The most commonly used form(s) were assigned indicative frequency 3. Forms used between a third and two-thirds as often were assigned the value 2, and forms occurring less than a third as often were assigned the value 1 (McIntosh, 1987). The indicative frequency is an ordinal categorical variable.

(iii) If historical data allowed for the dialect sample to be located geographically without reference to the forms themselves, then this location was recorded for

the manuscript. This location of origin was recorded in the form of a six-digit Ordnance Survey (OS) grid reference, where the first three digits were the easting measurement and the next three digits the northing measurement of the reference.

Once all the manuscripts had been examined, an iterative non-statistical method of location-elimination called the 'fit-technique' was applied to the manuscripts whose location of origin could not be determined from a historical source:

(iv) For each form of each item, the geographical regions where the form appeared in the anchor texts were identified.

(v) A manuscript with unknown location of origin was selected.

(vi) The relevant geographical regions from Step (iv) - i.e. those for the forms appearing in this manuscript - were overlaid on a map. The most likely location of origin for the manuscript was taken as the area of the map with greatest overlap.

(vii) Treating this manuscript now as an anchor text, Steps (iv)-(vi) were repeated for the next manuscript of unknown origin. This process was then repeated until all manuscripts were placed.

Figure 1.1 provides an illustration of this process. Clockwise from top-left, for each of three imagined forms, the regions where these forms are found in the anchor profiles

are superimposed on the map (in red, blue and purple). In the bottom left, the area with the greatest overlap (i.e. most forms co-occurring) is identified in yellow as the region from which the profile is most likely to originate. Linguistic expertise was used to select items for use in location, and to resolve conflicts.

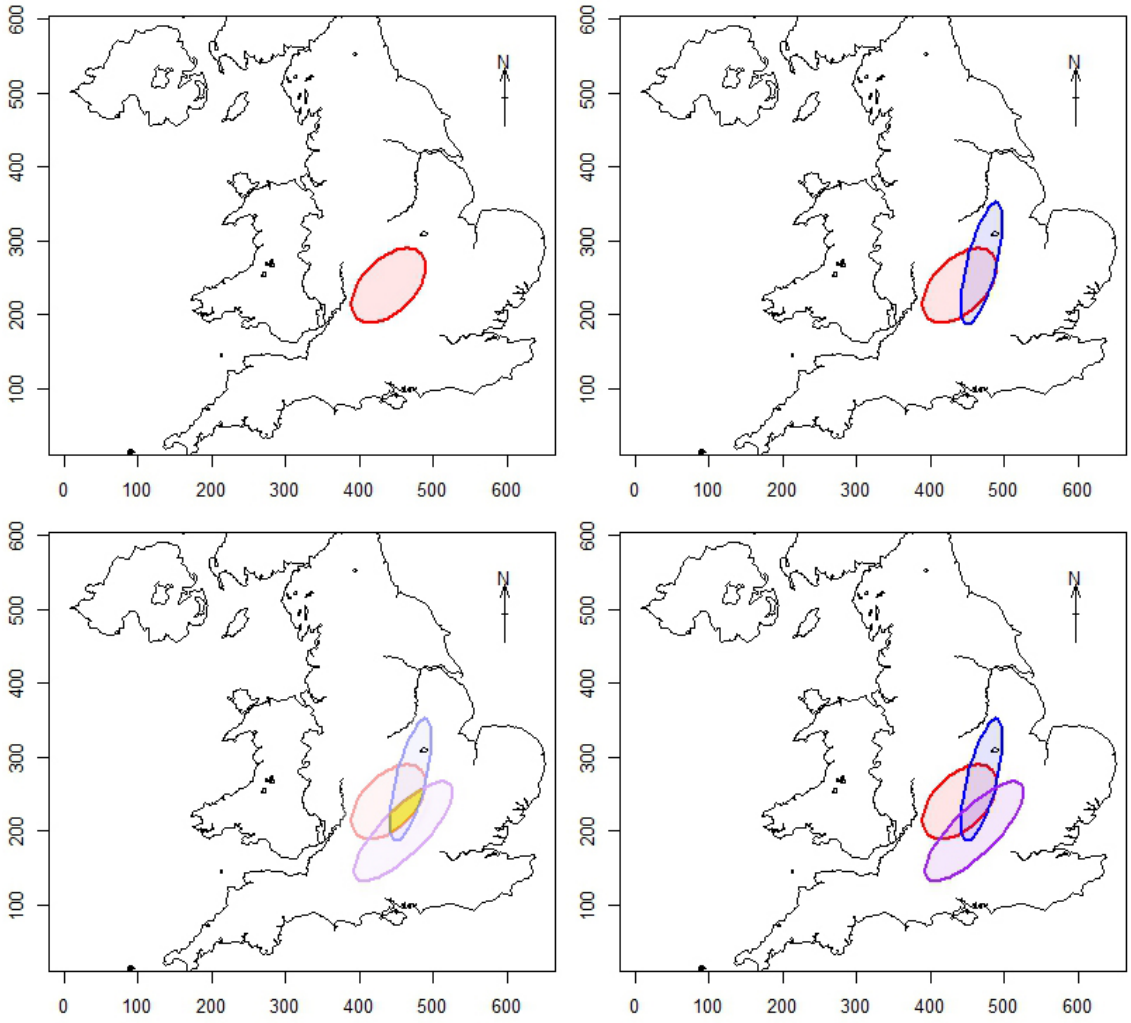


Figure 1.1: Clockwise from top left: an illustration of steps of the 'fit-technique' (the non-statistical method of location-elimination used by McIntosh et al. (1986) to determine the unknown location of origin of a manuscript). Co-occurrence of forms (represented by the yellow region) was used to locate the manuscripts.

1.3.1 Primary Data

The data resulting from this collection process were published in the Atlas in the form of linguistic profiles (LPs). Table 1.1 displays an example linguistic profile. The collection process yielded a set of $P = 1511$ linguistic profiles.

All linguistic profiles were published with OS locations to three significant figures. The majority of these locations were computed by the fit-technique. The Atlas itself does not distinguish anchors and floating (non-anchor) texts. We have this dataset directly from the scientists who examined the manuscripts.

The anchor texts are a collective of manuscripts whose origins are explicitly available, or readily deducible. As McIntosh et al. (1986) describes, these have diverse origins, ranging from "*personal correspondence, the records of manors and municipal-*

LP # 1	OS 478 332
Items	**Forms**
brother (n)	BRODER 3, BROTHER 2
soul (n)	SOULE 3, SAULE 2
then	THEN 3, THAN 1
upon	APON 3, APPON 2, VPPON 1
(205 more items)	⋮
young (aj)	YONG 3

Table 1.1: An example of raw Atlas linguistic data collected from a manuscript. Each row features a word (e.g. the noun 'brother'), along with the spellings ('broder' and 'brother') present in the manuscript and the indicative measure of the frequency of their use.

ities, the records of courts, secular or ecclesiastical (though the latter are commonly in Latin), and legal instruments - depositions and indentures, conveyances and arbitrations". Table 1.2 provides some examples of linguistic profiles derived from such manuscripts.

Let A represent the set of these profiles p with known origins; i.e. the anchor profiles. Let $\bar{A}$ represent the set of all profiles p with unknown origins; i.e. the floating profiles.

LP	OS	County	Description
48	558 306	Norfolk	Register of Crabhouse Nunnery.
50	225 484	Isle of Man	Parchment roll containing inquests of 1428 held at Peel and Castle Rushen, before Henry of Dyrom, lieutenant of Man.
53	435 457	Yorkshire, West Riding	Latin and Middle English verse and prose, relating to the life and cult of St Robert of Knaresborough. Comparisons with the language of local documents indicate that it is unlikely to derive from other than the district of Knaresborough.
64	547 258	Cambridgeshire	Ordinances of the Gild of St Clement, Cambridge, 1431.
81	357 339	Shropshire	Letter from the son of Jon Hullemore to a person unknown, concerning his father's disputed title to Eddysley (Edgeley, one mile southeast of Whitchurch).
82	365 352	Cheshire	Indenture made and written at Wych Maubanke (Wich Malbank, i.e. Nantwich), 15 April, 6 Henry V [1418].

Table 1.2: Descriptions of a subset of anchor profiles, provided by McIntosh et al. (2013).

As mentioned earlier in this chapter, to some extent we may choose A, depending on the framing of the problem: are we trying to locate all of the profiles whose origins were determined using the fit-technique; or just a small set of new profiles, trusting the fit-technique locations? These correspond to problems (1) and (2) stated on Page 3. For each analysis presented in this thesis, we clearly define which of the two frameworks we are working under, and consequently the set of profiles in A and $\bar{A}$. If we try to locate all of the texts whose origins were determined using the fit-technique, then we have $|A| = 300$ and $|\bar{A}| = 1211$.

For $p \in \{1, \ldots, P\}$, let Λ_p denote the true spatial location for profile p, and λ_p denote the Atlas location coordinate vector for profile p. If $p \in A$, then $\Lambda_p = \lambda_p$. If $p \in \bar{A}$, then Λ_p is unknown.

Figure 1.2 shows the spread of the $P = 1511$ linguistic profiles. The top plot displays the known locations of origin Λ_p for the 300 anchors $p \in A$ in red. The bottom plot displays the Atlas location estimates λ_p (derived from the fit-technique) for the 1211 floating profiles $p \in \bar{A}$ in blue. There are only a few anchor profiles whose scribes learned to write in southern England, southern Scotland and western Wales, with the majority from the Midlands and the North. The fit-technique estimates for the origins of the floating profiles are spread more evenly across England, as well as covering similar regions of Scotland and Wales.

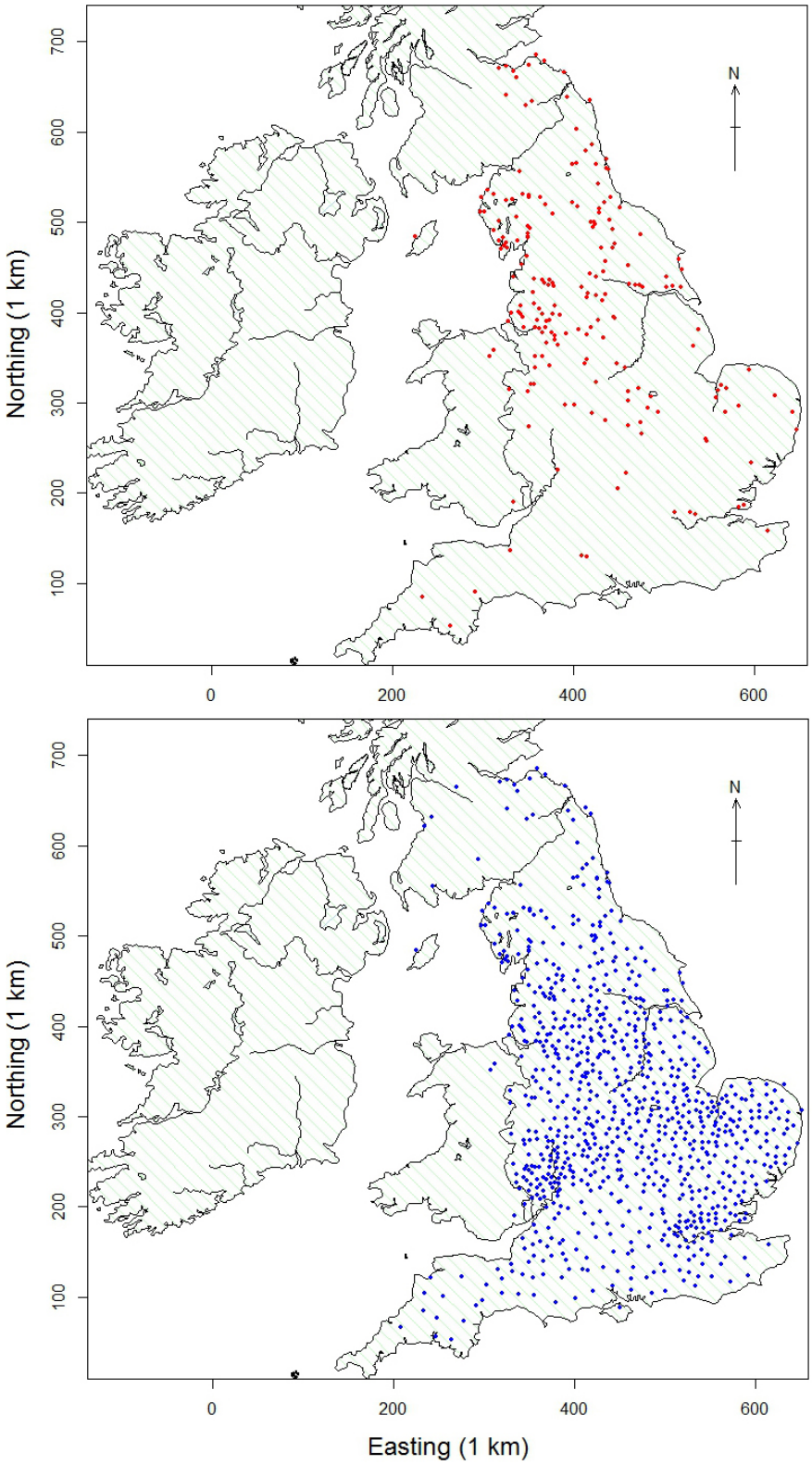


Figure 1.2: The known locations of origin Λ_p for the anchor profiles $p \in A$ (top plot, displayed in red), and the Atlas location estimates λ_p derived from the fit-technique for floating profiles $p \in \bar{A}$ (bottom plot, displayed in blue). Some profiles sit just outside the coastal boundaries, suggesting slight plotting mis-alignment between the profile locations and displayed map.

McIntosh et al. (1986) describes that "*the Atlas tells us, in essence, where the scribe of a manuscript learned to write; the question of where he actually worked and produced the manuscript is a matter of extrapolation and assumption.*" Thus, strictly speaking, when we refer to 'locating a manuscript' in this thesis, we mean inferring the location where the scribe learned to write. The spatial distribution of these locations will not be uniform, and will relate to the spatial clustering of population. This clustering will differ relative to a sample of documents of this size gathered in the present time, due to the population spread in mediaeval England differing to that of today.

When estimating the dialect fields, we fix the anchor profiles to their known locations, and use a uniform prior distribution for the remaining profiles. This prior distribution will be overdispersed relative to the truth, and future work could seek to refine this prior by incorporating additional historical information (for example, census records). However, our prior is a reasonable starting point, given that the prior is over a lattice of 10x10 km cells (many of which, when considering the area covered by towns, are likely to contain places where scribes could have learned to write).

The survival of documents from the period covered by the Atlas is likely to have spatial dependence based on where they were stored. Indeed, the uneven spatial survival of anchor texts is exhibited in Figure 1.2. When estimating the locations of origin of the non-anchor profiles, we do not use the spatial distribution of the anchors directly; rather, we use inferred dialect fields. These dialect fields are informed by

the anchors, but also by the non-anchors. If the anchors have been thinned spatially, and their locations are clustered, it means we will be able to best reconstruct the locations of the non-anchors near these clusters, and the variance of our estimates will increase elsewhere.

In this thesis, we model over a sub-region of the map, within which the anchors are more densely and more homogeneously spread. This region was chosen for precisely this reason.

Across the $P = 1511$ profiles, $I = 623$ items were used. When McIntosh et al. (1986) collected the raw data, they recorded the occurrence of 280 separate items, however, many of these had sub-categories (called sub-items), which we treat as separate 'items'.

Let item $i \in \{1, \ldots, I\}$ have F_i forms across all profiles. For example, the noun 'brother' is an item, with 56 forms used across the $P = 1511$ profiles. These forms include 'BRODER', 'BROTHER', 'BROYYR' and 'BRUTHIR'. The total number of forms F_i for items $i \in \{1, \ldots, I\}$ varies between a minimum of 1 and a maximum of over 500.

Let $d_{p,i,f}$ denote the indicative frequency with which form f of item i was used in profile p, where $d_{p,i,f} \in \{\text{NA}, 0, 1, 2, 3\}$. $d_{p,i,f} = 0$ if form f of item i was not used in profile p, but at least one other form of item i was used in the profile. $d_{p,i,f} = \text{NA}$ if item i was not used in profile p.

Examination of the data collection process and a conversation with the scientists who gathered the data revealed that there was no consistent standard for the recording of the indicative frequencies $d_{p,i,f}$. We were advised by these scientists that although these indicative frequencies were consistent within certain subsets of the data, they were inconsistent between subsets.

One reliable feature of these data is the ordering of the indicative frequencies within each profile. Future work could use this ordering to help inform the 'dialect fields', since an indicative frequency for one form being greater than the frequency for another ($d_{p,i,f_1} > d_{p,i,f_2}$) implies that the probability of using that form should be higher than the other.

The only other uniformly reliable information in the data is the presence or absence of forms. The data were therefore reduced to indicator variables $y_{p,i,f}$ as follows:

$$y_{p,i,f} = \begin{cases} 0 & \text{if } d_{p,i,f} \in \{\text{NA}, 0\} \\ 1 & \text{if } d_{p,i,f} \in \{1, 2, 3\} \end{cases} \tag{1.6}$$

The coding of NA as a zero is appropriate here. The item has been used zero times in the profile. We come back to this point shortly.

Let $y_{p,i}$ denote the vector $(y_{p,i,1}, \ldots, y_{p,i,F_i})^T$, let $y_p = (y_{p,1}, \ldots, y_{p,I})$, and $y = (y_1, \ldots, y_P)$. For $1 \leq a < b \leq I$, let $y_{p,a:b} = (y_{p,a}, \ldots, y_{p,b})$.

Figure 1.3 provides a graphical representation of the data, displaying how form usage varies between different profiles by region. The data vectors ($y_{407,V}$, $y_{1351,V}$, $y_{5140,V}$, $y_{5371,V}$) are depicted, for the set of items $V = \{19, 43, 57, 587\}$. We can see that

similar forms of the items $i \in V$ are used in $p \in \{407, 1351\}$ and in $p \in \{5140, 5371\}$, but the forms used in these two pairs of profiles differ.

This representation also illustrates the issue of sparsity in the data. The data vector y contains over forty million entries, but 99% of these are zeroes. The majority of these zeroes (54%) arise from item non-usage. We might think of item usage as informing of the "topic" of the document. For example, a given document may not use the item "brother" at any point in the text, and so we do not know what form or forms the author prefers when using this item. As mentioned above, all the form-use indicators in the raw data y are zero for such items within these profiles, since the item has been used zero times. However, they can and should still be understood as "NA", and that is how we will treat them in our models in Chapter 2.

1.3.2 Secondary Data

Some forms of an item appear more visually similar than others. For example, the forms "BROTHER" and "BROTHERE" of the item "brother" differ only slightly in appearance when compared to the forms "BROYYR", "BRUTHIR" and "BRODRE". In this section, we will describe a coarsening of the data in which similar forms are merged, leading to a dramatic reduction in the number of forms per item. This merging operation, as illustrated later in Table 1.5, was done under the supervision of the scientists involved, and exploited their expertise.

It seems credible that a scribe knowing a certain form was more likely to know

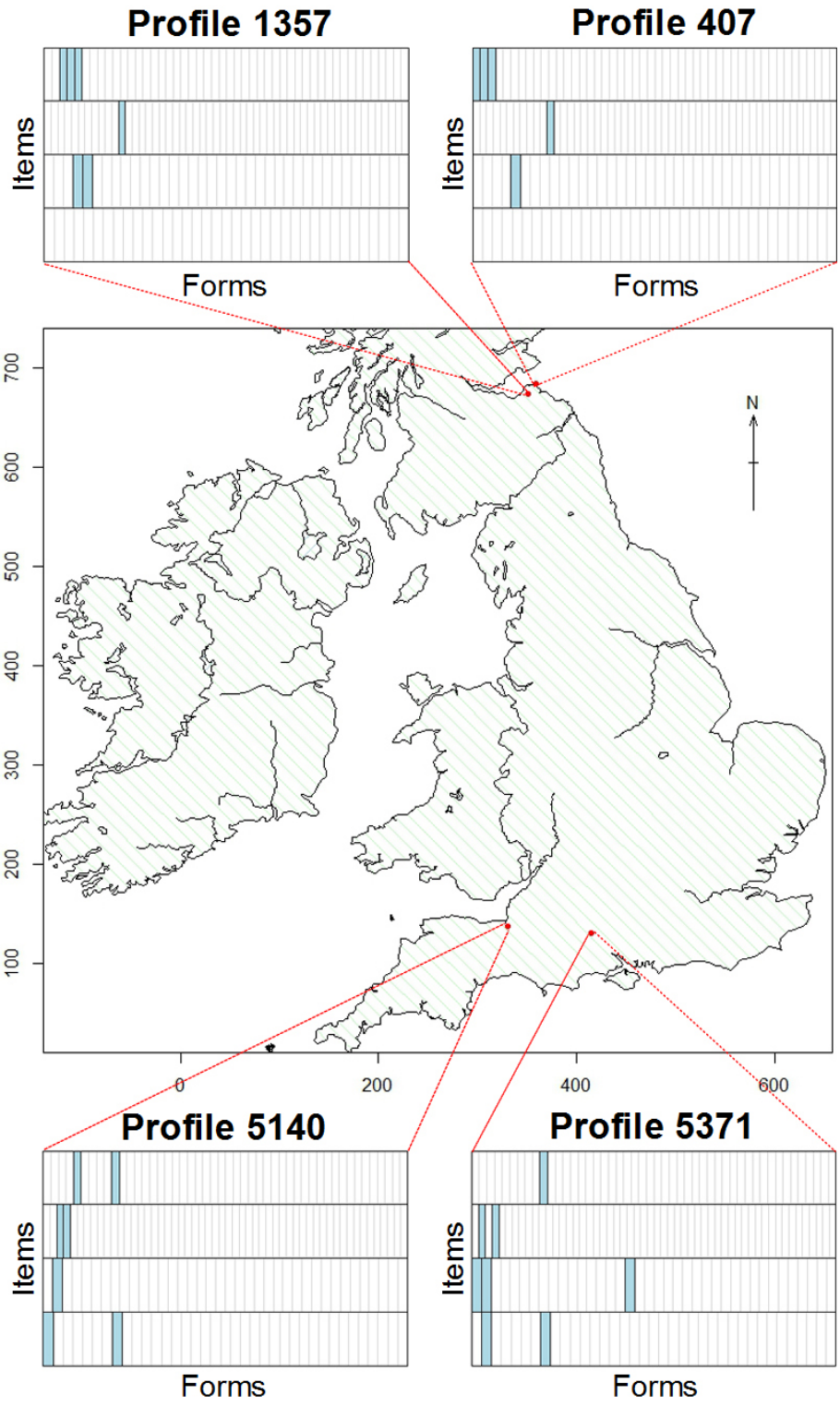


Figure 1.3: A graphical representation of a subset of data from the anchor profiles $p \in \{407, 1351, 5140, 5371\}$. For each vector, if $y_{p,i,f} = 1$ for form f of item $i \in \{19, 43, 57, 587\}$, then a corresponding box is coloured blue. The known locations of the four profiles are shown in red on the map.

and use another form with a similar spelling rather than one that differs more philologically. From this observation arises the question of which linguistic features of a profile are informative of its location. Statistically, finding a means to identify these features is very desirable. Such knowledge would allow us to coarsen the data to the most informative features, reducing both the sparseness and the magnitude of the data without necessarily a great loss of information informing location. This in turn would ease the computational and inferential burden of modelling the data.

We could take a statistical approach to discovering these features; for example, clustering data based on the correlations $r^{(i)}_{f,f'}$ between each pair of forms f and f' of the same item i in profile p. We observe (but do not report here) that such an approach leads to clusters that generally contain forms of similar appearance.

It is clear though that the most efficient means to this end is to ask the scientists directly. Correspondingly we obtained a secondary dataset from the authors of the Atlas, detailing hierarchical structure in the primary data.

For each of 510 items, the secondary data provides the possible spellings of its forms, segment by segment. An example of secondary data is provided in Table 1.3 for the item "whether". Recall, the upper and lower case letters presented in this table represent distinct symbols in the original manuscripts, rather than the same letter in capital and lower-case form.

By referring to these data in Table 1.3, we can decompose each form of "whether"

into its constituent segments. The form "QWEyER", for example, is "QW:E:y:E:R" (where ':' separates each segment). Similarly, the form "QWEyE" is "QW:E:y:E:@", where "@" denotes an empty segment.

For each item, the potential number of forms that could be constructed using the secondary data is enormous. For example, there are 34560 possible combinations of the segments from Table 1.3 for the item "whether". In reality, we only observe a small percentage of these in the data, but this still leaves us with a large number of forms which we want to reduce.

For some items in the secondary dataset, the scientists provided supplementary information to use for coarsening. An example of this information is detailed in Table 1.4 for the item "against", which has forms including "AGAYNST", "AYEYNSTE", "OGAINIS", "AzENSSE" and "TOGEYNus".

In the supplementary information, equivalent spellings within each segment which did not imply different sound values were marked with an "=" sign. The scientists advised us to ignore some segments entirely, because they did not signify unique features. These segments are marked "**EXCLUDE**".

Segment	Possible Spellings
1	HW, QU, QUH, QV, QW, Qu, QWH, W, WH, Y
2	A, E, EI, ER, EY, I, O, Y
3	@, D, DD, T, TH, Y, YT, YTH, d, y, z, zH
4	@, E, EI, I, O, Ou, U, Y, YH, e, i, u
5	@, R, r

Table 1.3: Secondary data for the item "whether", which has forms including "WHEITHIR", "WETHER", "QWEyER", "QWEyE", "WHEDERe" and "HWADER".

Segment	Possible Spellings
1	@, A, E, I=j=J, O, ON, TO
2	@, -, -A, -I **EXCLUDE**
3	G, Y=Z=z, y
4	A, AI=AIY=AY=AYI, E=EE, EI=EII=EY=EYI, I=Y, IEI
5	@, N=NN=Nn=n=nN, ND, NT
6	@, E, I, U, Y, e, u **EXCLUDE**
7	@, C, S=SS=s, ST, y, z
8	@, E, e **EXCLUDE**

Table 1.4: Secondary data for the item "against", which has forms including "AGAYNST", "AYEYNSTE", "OGAINIS", "AzENSSE" and "TOGEYNus". Supplementary information about segments which are equivalent and segments which can be ignored is also provided.

We have secondary data, like that shown in Table 1.3, for 510 of the 623 items. Supplementary information, like that shown in Table 1.4, is provided for 73 items. Let $\mathbb{H}$ denote the set of these 73 items. Using this supplementary information, the secondary data can be combined with the primary data to produce coarsened data. For each item $i \in \mathbb{H}$, we map its forms $f \in \{1, \ldots, F_i\}$ to the coarsened forms $f' \in \{1, \ldots, F_i'\}$ as follows.

Consider the form "AYEYNSTE" of the item "against". Using the secondary data in Table 1.4, we see that this form is created using the segments "A:@:Y:EY:N:@:ST:E". The second, sixth and eighth segments were marked by the scientists to be ignored, leaving us with "A:Y:EY:N:ST". Replacing equivalent spellings with the left-most spelling from the supplementary data gives the coarsened form, "AYEINST".

Table 1.5 shows this process for nine further forms f of the item. These nine forms reduce to three common coarsened forms f', illustrating the data reduction through this coarsening. Overall, the 344 forms of "against" reduce to 69 coarsened forms.

After completing this process for all forms $f \in \{1, \ldots, F_i\}$ of items $i \in \mathbb{H}$, we then coarsen the primary data as follows:

- Suppose forms $f \in \{1, \ldots, F_i\}$ of item $i \in \mathbb{H}$ reduce to coarsened forms $f_1', \ldots, f_n'$, where $n < F_i$.
- Let $F_1', \ldots, F_n'$ be the number of original forms that reduce to each coarsened form, where each $F' \geq 1$.

Form	Segment by Segment	Coarsened Form
A-GEYNST	"A, -, G, EY, N, @, ST, @"	AGEINST
AGEYNYSTE	"A, @, G, EY, N, Y, ST, E"	AGEINST
AGEYNEST	"A, @, G, EY, N, E, ST, @"	AGEINST
⋮	⋮	⋮
AzAYNESTE	"A, @, z, AY, N, E, ST, E"	AYAINST
AzAINST	"A, @, z, AI, N, @, ST, @"	AYAINST
AzAYNEST	"A, @, z, AY, N, E, ST, @"	AYAINST
⋮	⋮	⋮
TO-AzENST	"TO, -A, z, E, N, @, ST, @"	TOYENST
TOzENST	"TO, @, z, E, N, @, ST, @"	TOYENST
TOzENSTE	"TO, @, z, E, N, @, ST, E"	TOYENST

Table 1.5: Example coarsening of the primary data. The supplementary information from Table 1.4 is used to reduce each form $f \in \{1, \ldots, F_i\}$ of the item "against" to its coarsened form f'.

- For each coarsened form $f' \in \{f'_1, \ldots, f'_n\}$,

$$y'_{p,i,f'} = \begin{cases} 0 & \text{if } \sum_{f \in \{f_1, \ldots, f_{F'}\}} y_{p,i,f} = 0 \\ 1 & \text{if } \sum_{f \in \{f_1, \ldots, f_{F'}\}} y_{p,i,f} \geq 1 \end{cases} . \tag{1.7}$$

Note that we denote the coarsened data $y'_{p,i,f'}$, as compared to the primary data $y_{p,i,f}$ from which it is derived. We use the two interchangeably later in this report, but do clearly mention in the text which dataset we are working with.

There are 6012 forms across the 73 items $i \in \mathbb{H}$, with a median of 60 forms per item. After the reduction process, there are 1182 coarsened forms across these items, with a median of 11 forms per item. Figure 1.4 illustrates this reduction. There is much greater variation in the number of forms per item with the primary data, and

the coarsening has clearly offered a massive reduction.

We believe that the differences between the spellings from the finer (original) scale to the coarsened scale are largely noise, and thus uninformative of location. That is, for all practical purposes, the coarsened data are sufficient for the inference. We explore this supposition in further detail in Chapters 5 and 6. It is infeasible to estimate "dialect fields" with all the primary data vectors.

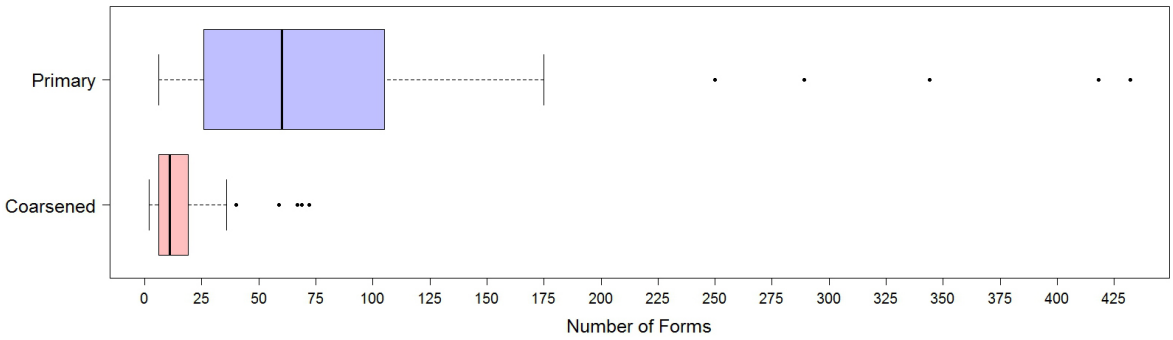


Figure 1.4: Box-and-whiskers plot showing how the number of forms of the 73 items $i \in \mathbb{H}$ (those items for which both primary and coarsened data are available) varies between the primary data (in blue) and the coarsened data (in red).

chapter 2

Model and Inference

In this chapter, we detail our models for the linguistic data introduced in Chapter 1. We first describe a generative model for the process realising a data vector y_p, with the likelihood presented in Section 2.1.1. Prior distributions for the parameters of the model are given in Section 2.1.2, and the joint posterior distribution is given in Section 2.1.3, alongside conditional posterior distributions for each of the parameters.

We then extend our model, with explicit modelling of outlier manuscripts and spellings outlined in Section 2.2. Finally, an alteration to the construction and smoothing of the dialect fields is considered in Section 2.3, and the two models arising from this alteration are described.

From earlier, we have a set of profiles $p \in 1, \ldots, P$ divided into two sets: the anchor profiles A and the floating profiles $\bar{A}$. Each profile p has a true spatial position Λ_p, and a spatial position λ_p provided by the Atlas. If $p \in A$, then $\Lambda_p = \lambda_p$, otherwise Λ_p is unknown.

We work in a lattice of cells $x \in \{1, \ldots, C\}$ laid over the geographic map. Let $J : \mathbb{R}^2 \to \{1, 2, \ldots, C\}$ be the function mapping from coordinate space to the lattice. Let $\xi_p = J(\Lambda_p)$ be the true cell-location within the lattice for profile $p \in A$, and x_p denote the unknown true cell-location for profile $p \in \bar{A}$.

Finally, let $\eta_{x,i,f}$ be the probability that item i takes form f in any manuscript originating from within cell x.

2.1 ℓ_1-Dirichlet Zero-Inflated Model

2.1.1 Likelihood

In this section, we describe a generative model for the process realising a data vector y_p. Under our model, the unobserved count data $z_{p,i,f}$ (the number of times form f of item i was used in profile p) are generated as follows:

(i) Consider a manuscript $p \in \{1, \ldots, P\}$. Suppose item $i \in \{1, \ldots I\}$ appears $z_{p,i} \sim \text{Poisson}(\mu_i)$ times in a manuscript, where μ_i is the mean number of times item i is used in a manuscript. The parameter μ_i is assumed constant across manuscripts and independent of location (this assumption is discussed further below).

(ii) If item i appears $z_{p,i} > 0$ times in manuscript p, we assume forms $f \in \{1, \ldots, F_i\}$ appear

$$(z_{p,i,1}, \ldots, z_{p,i,F_i}) \sim \text{Multinomial}([\eta_{x_p,i,1}, \ldots, \eta_{x_p,i,F_i}], z_{p,i}) \tag{2.1}$$

times in the manuscript, independently of all other $z_{p,i,f}$. If item i appears $z_{p,i} = 0$ times in manuscript p, it follows that $z_{p,i,f} = 0$ for all forms $f \in \{1, \ldots, F_i\}$.

(iii) Repeat (i)-(ii) to generate count data for all other manuscripts $p \in \{1, \ldots, P\}$.

The observation model for the count data is therefore

$$\begin{aligned} \mathrm{p}(z_{p,i,f_1}, \ldots, z_{p,i,F_i} \,|\, \mu_i, \eta_{x_p,i}, x_p) &= \mathrm{p}(z_{p,i,f_1}, \ldots, z_{p,i,F_i} \,|\, z_{p,i}, \eta_{x_p,i}) \times \mathrm{p}(z_{p,i} \,|\, \mu_i) \\ &= \prod_{f=1}^{F_i} \left\{ \frac{e^{-\mu_i \eta_{x_p,i,f}} (\mu_i \eta_{x_p,i,f})^{z_{p,i,f}}}{z_{p,i,f}!} \right\}, \end{aligned} \tag{2.2}$$

where $\eta_{x_p,i} = (\eta_{x_p,i,f_1}, \ldots, \eta_{x_p,i,F_i})^T$.

We do not observe the count data, but the observed presence/absence data $y_{p,i,f}$ are derived from $z_{p,i,f}$ as follows:

$$y_{p,i,f} = \begin{cases} 0 & \text{if } z_{p,i,f} = 0 \\ 1 & \text{if } z_{p,i,f} > 0 \end{cases}. \tag{2.3}$$

From the above process, the probability to realise data y is thus

$$\mathrm{p}(y_{p,i,f} \,|\, \eta_{x_p,i,f}, \mu_i, x_p) = \begin{cases} 1 - e^{-\mu_i \eta_{x_p,i,f}} & \text{if } y_{p,i,f} = 1 \\ e^{-\mu_i \eta_{x_p,i,f}} & \text{if } y_{p,i,f} = 0 \end{cases}. \tag{2.4}$$

Given the assumed independence of form usage, the likelihood for the data y is thus

$$\mathrm{p}(y|\eta, \mu, x) \propto \prod_{p=1}^{P} \prod_{i \in I_p} \prod_{f=1}^{F_i} \beta_{p,i,f}, \tag{2.5}$$

where $\beta_{p,i,f} = (1 - e^{-\mu_i \eta_{x_p,i,f}})^{y_{p,i,f}} (e^{-\mu_i \eta_{x_p,i,f}})^{1-y_{p,i,f}}$, and $I_p = \{i : \sum_f y_{p,i,f} > 0\}$ is the set of items i that were used at least once in profile p.

Towards the end of Section 1.3.1, it was noted that 99% of observations $y_{p,i,f} = 0$. Of these zero observations, a slight majority (54%) arise from item non-usage, i.e. $\sum_f y_{p,i,f} = 0$. Such cases are excluded from the likelihood above.

The remainder of the zeroes in the data arise when item i is used in profile p at least once, but form f of the item is not. To account for the prevalence of these zeroes in the data, we introduce zero-inflation parameters ϕ_i, such that

$$\mathrm{p}(y_{p,i,f} = 0 \,|\, \eta_{x_p,i,f}, \mu_i, x_p, \phi_i) = \phi_i + (1 - \phi_i)e^{-\mu_i \eta_{x_p,i,f}}. \tag{2.6}$$

ϕ_i is the extra probability of form f of item i not being used in profile p. This probability is constant across all forms $f \in \{1, \ldots, F_i\}$ of item i, as well as across all profiles $p \in \{1, \ldots, P\}$.

From (2.6), it follows that

$$\mathrm{p}(y_{p,i,f} = 1 \,|\, \eta_{x_p,i,f}, \mu_i, x_p, \phi_i) = (1 - \phi_i)\left\{1 - e^{-\mu_i \eta_{x_p,i,f}}\right\}, \tag{2.7}$$

and so the likelihood for the data y is now

$$\mathrm{p}(y|\Theta) \propto \prod_{p=1}^{P} \prod_{i \in I_p} \prod_{f=1}^{F_i} \left\{ \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\beta_{p,i,f} \right\}, \tag{2.8}$$

where $\Theta = \{\eta, \mu, x, \phi\}$.

As discussed in Section 1.3.1, there was no consistent standard for the recording of the indicative frequencies $d_{p,i,f}$, and we therefore reduced to presence/absence data $y_{p,i,f}$. We could, however, somehow use the indicative frequencies to impose an order on the η-fields, given that $d_{p,i,f_1} > d_{p,i,f_2}$ implies that $\eta_{x_p,i,f_1} > \eta_{x_p,i,f_2}$. Exploiting this information is a non-trivial problem and not one we have explored in depth, but is one that could be explored in future work.

We make the simplifying assumption the item usage rate, μ_i, is constant across

manuscripts and independent of location. It would be possible to relax this assumption and to allow μ_i to vary spatially, or perhaps by the class of document (e.g. legal documents, personal correspondence, etc). We do not pursue this avenue because the words from the Atlas are (for the most part) very basic elements of speech, or core vocabulary.

The manuscripts examined in the Atlas cover the period ca. 1325-1450. However, this range does not apply uniformly across the country: in the south, the majority of the manuscripts fall within a narrower period ca. 1325-1425; and in the Midlands and the north, within the period ca. 1350-1450. We later model using data from a subregion of the Midlands. Allowing form usage parameters η to vary temporally, as well as spatially, may be considered desirable even given the restricted time-period covered in the dataset we model with. However, the data available for the manuscripts' temporal origins varies in detail (for some, to the year they were written; for others, to the century). Incorporating this information in the model would therefore be difficult. We therefore assume in this thesis form usage to be constant across time. This removes a layer of complexity, and reduces the number of parameters to estimate.

2.1.2 Prior Distributions

We use a Bayesian framework for parameter estimation, so we now specify prior distributions for the parameters $\Theta = \{\eta, \mu, x, \phi\}$. η is a set of $C \times I \times \sum_i F_i$ form usage probabilities; μ is a set of I item usage rate parameters; x is a set of P profile-

location parameters; and ϕ is a set of I zero-inflation probabilities.

2.1.2.1 Profile Locations

For each floating profile $p \in \bar{A}$, we use a uniform prior distribution over the lattice's cells $x \in \{1, \ldots, C\}$ for the profile's unknown cell-location of origin x_p:

$$\pi_x(x_p = x) = \frac{1}{C}. \tag{2.9}$$

We fix the anchor profiles $p \in A$ to their true cell-location $\xi_p = J(\Lambda_p)$.

As mentioned in Chapter 1, McIntosh et al. (1986) describes how "*the Atlas tells us, in essence, where the scribe of a manuscript learned to write; the question of where he actually worked and produced the manuscript is a matter of extrapolation and assumption.*" Thus, not all locations on the map are actually equally likely locations of origin for manuscripts. We discussed how future work could seek additional historical data to incorporate into this prior distribution, and thus allow for additional weight to be given to places with higher literacy rates, or higher populations. Such data could be, for example, census records. However, the prior specified here gives a good starting point.

2.1.2.2 Item Usage Rates

We assign each of the rate parameters μ_i a gamma prior distribution with hyper-parameters (a_i, b_i):

$$\pi_\mu(\mu) \propto \prod_{i=1}^{I} \mu_i^{a_i - 1} e^{-b_i \mu_i}. \tag{2.10}$$

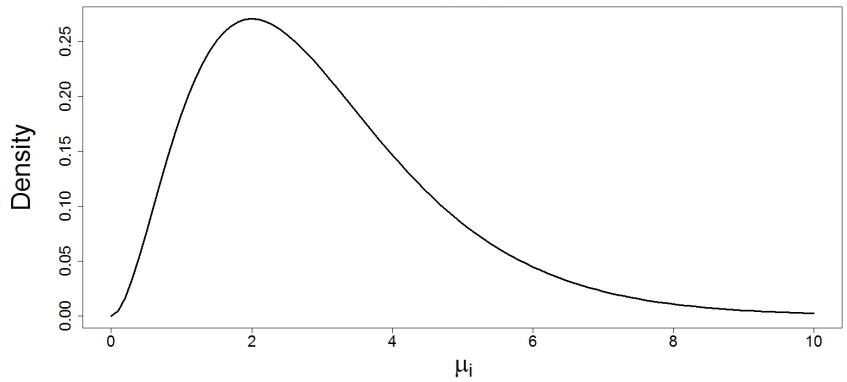


Figure 2.1: Gamma prior distribution chosen for the item usage rates μ_i, for $i \in \{1, \ldots, I\}$.

Following prior elicitation from the scientists, these hyper-parameters were set to $a_i = 3$ and $b_i = 1$ for all $i \in \{1, \ldots, I\}$, inducing a distribution for each μ_i as displayed in Figure 2.1, with $E(\mu_i) = 3$ and $V(\mu_i) = 3$. The hyper-parameters could be varied for different items i if we had reason to believe certain items would be used more than others, however, without such data, we kept these constant across items.

2.1.2.3 Form Usage Probabilities

We set the prior distribution for the form usage probabilities to

$$\pi_\eta(\eta) \propto \prod_{i=1}^{I} \prod_{f=1}^{F_i} \prod_{x=1}^{C} (\eta_{x,i,f})^{\chi-1} \exp\left(-\frac{\theta_i}{2} \sum_{x' \in \mathcal{N}(x)} |\eta_{x,i,f} - \eta_{x',i,f}| \right), \qquad (2.11)$$

where $\mathcal{N}(x)$ denotes the set of cells neighbouring cell x. Cells are neighbours if they share a border. On a rectangular lattice with free boundary conditions and nearest-neighbour interactions, corner cells have two neighbours, other cells on the boundary of the lattice have three, and interior cells have four.

This prior distribution defines a Markov random field (MRF). The η parameters

are independent between items i, from the product structure of the distribution. Within items, the random vectors $\eta_{x,i} = (\eta_{x,i,1}, \eta_{x,i,2}, \ldots, \eta_{x,i,F_i})$ satisfy the Markov property within neighbourhoods: that is, they are conditionally independent of $\eta_{x',i}$ for $x' \notin \{x\} \cup N(x)$ (i.e. at locations x' outside their neighbourhood) given the η-values of their neighbours (that is, given $\eta_{x^*,i}$ for $x^* \in N(x)$).

This distribution, which we name a ℓ_1-Dirichlet field, is essentially a Dirichlet distribution for form probabilities in each cell, with form probabilities smoothed across cells. The probabilities $\eta_{x,i,f}$ from neighbouring cells are correlated through the exponential penalty term. The ℓ_1 penalty $|\eta_{x,i,f} - \eta_{x',i,f}|$ allows sharp discontinuities, modelling population movement, rather than diffusion of forms as a quadratic penalty might suggest.

The magnitude of the applied problem means that there are a very large number of η-field parameters to estimate. The ℓ_1-penalty is chosen because it offers a form of dimensionality reduction, bringing the scale of the problem to a more manageable level. Given this, we expect the choice of ℓ_1-penalty to be very advantageous, but do also explore a model based on a quadratic penalty later in the chapter.

The level of smoothing is controlled by the parameters θ_i. The number of forms F_i of item i varies greatly between different items, and since $\sum_{f=1}^{F_i} \eta_{x,i,f} = 1$, the average size of $\eta_{x,i,f}$ decreases as F_i increases. This means that $-\frac{\theta_i}{2}|\eta_{x,i,f} - \eta_{x',i,f}|$ cannot be measured on the same scale for different items without allowing θ_i to depend on F_i.

Thus, we set $\theta_i = \theta F_i$, where θ is the overall smoothing parameter. When $\theta = 0$, no smoothing is performed. Increased smoothing is performed as θ increases.

Our sensitivity analysis suggests that an appropriate level of smoothing can be obtained when $1 \leq \theta \leq 2$, but over-smoothing occurs when $\theta > 2$. These choices are supported by a goodness of fit analysis, in which we leave out Atlas information about profile locations, and compute Bayes factors to see if the Atlas position is rejected. This can be thought of as a kind of Bayesian cross-validation. We tried a range of values for θ and considered their impact. The choices given here are supported by good performance in these analyses, whereas other values tried gave obviously worse results. We find that $\theta = 1.75$ gives a good level of smoothing, so use this value throughout this report.

This analysis tells us the scale of variation in η that has an impact is around $\frac{1}{F_i}$. A separate issue is that η needs to be able to vary across the spatial region in such a way that a form is able to dominate in one area, and be absent in another. We believe that around 100 km is a sensible scale for a form to vary in such a fashion. With the lattice we use (with 10 km by 10 km cells), this means variation as big as $\frac{1}{10}$ might be allowed. For the coarsened data, this is essentially the same scale of variation as $\frac{1}{F_i}$.

In our sensitivity analysis (described above), we considered both of these different ways of thinking about the scale of variation. Each of the different methods gave us

a starting point for θ, and the sensitivity analysis results confirmed that variation on a scale of $\frac{1}{F_i}$ was sensible. Nevertheless, this is an area that could be explored further in future work.

The Dirichlet term $(\eta_{x,i,f})^{\chi-1}$ allows us to incorporate prior knowledge about the relative usage rates of the forms f of an item i. When $\chi = 1$, the prior reduces to the exponential smoothing term. If $\chi = 1$ and $\theta = 0$, the prior further reduces to a uniform distribution. $\chi > 1$ represents belief that the forms are used at very similar rates, whereas $\chi < 1$ signifies belief in higher usage rates for some forms than others, which is contextually more plausible. Sensitivity analysis suggests if χ is too small, our estimates of the location of origin of manuscripts are overly confident, but decreasing χ from 1 by a small amount has minimal impact. For simplicity, $\chi = 1$ is used throughout this report.

2.1.2.4 Zero-Inflation Probabilities

We assign each of the zero-inflation probabilities ϕ_i a uniform prior distribution:

$$\pi_\phi(\phi) \propto 1. \tag{2.12}$$

2.1.3 Joint Posterior Distribution

Combining the likelihood presented in Equation (2.8) with the prior distributions presented in Equations (2.9), (2.10), (2.11), and (2.12) gives us the joint posterior

distribution from which we wish to sample:

$$\pi(\Theta \,|\, y) \propto \left(\prod_{p=1}^{P} \prod_{i \in I_p} \prod_{f=1}^{F_i} \left\{ \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\beta_{p,i,f} \right\} \right) \left(\prod_{i=1}^{I} \mu_i^{a_i - 1} e^{-b_i \mu_i} \right) \times \left(\prod_{i=1}^{I} \prod_{f=1}^{F_i} \prod_{x=1}^{C} (\eta_{x,i,f})^{\chi - 1} \exp\left(-\frac{\theta_i}{2} \sum_{x' \in \mathcal{N}(x)} |\eta_{x,i,f} - \eta_{x',i,f}| \right) \right), \quad (2.13)$$

(recall, $\beta_{p,i,f} = (1 - e^{-\mu_i \eta_{x_p,i,f}})^{y_{p,i,f}} (e^{-\mu_i \eta_{x_p,i,f}})^{1 - y_{p,i,f}}$).

2.1.3.1 Conditional Posterior Distributions

We derive conditional posterior distributions for each of the parameters from the joint posterior distribution specified in Equation (2.13). We later use these in the MCMC algorithm outlined in Section 3.1, since they directly give the form we arrive at after all possible cancellation.

The conditional posterior distribution for x_p is

$$\pi_x(x_p \,|\, y, \Theta \backslash \{x_p\}) \propto \prod_{i \in I_p} \prod_{f=1}^{F_i} \left\{ \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\beta_{p,i,f} \right\}, \quad (2.14)$$

where $\Theta \backslash \{x_p\}$ is the set of values for all parameters apart from x_p, and the RHS depends on x_p through $\beta_{p,i,f} = \beta(\mu_i, \eta_{x_p,i,f})$ - see page 41.

The conditional posterior distribution for μ_i is

$$\pi_\mu(\mu_i \,|\, y, \Theta \backslash \{\mu_i\}) \propto \left(\prod_{p \in P_i} \prod_{f=1}^{F_i} \left\{ \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\beta_{p,i,f} \right\} \right) \mu_i^{a_i - 1} e^{-b_i \mu_i}, \quad (2.15)$$

with additional μ_i-dependence on the RHS through $\beta_{p,i,f} = \beta(\mu_i, \eta_{x_p,i,f})$, and where $P_i = \{p : \sum_f y_{p,i,f} > 0\}$ is the set of profiles p in which item i was used at least once.

The conditional posterior distribution for $\eta_{x,i} = (\eta_{x,i,1}, \ldots, \eta_{x,i,F_i})$, the vector of all probabilities of form usage for item i in a given cell x, is:

$$\pi_\eta(\eta_{x,i} \,|\, y, \Theta \backslash \{\eta_{x,i}\}) \propto \left(\prod_{p \in P_{x,i}} \prod_{f=1}^{F_i} \left\{ \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\beta_{p,i,f} \right\} \right) \times \prod_{f=1}^{F_i} (\eta_{x,i,f})^{\chi - 1} \exp\left(-\theta_i \sum_{x' \in \mathcal{N}(x)} |\eta_{x,i,f} - \eta_{x',i,f}| \right), \quad (2.16)$$

where $P_{x,i} = \{p : \sum_f y_{p,i,f} > 0, x_p = x\}$ is the set of profiles p located in cell x in which item i was used at least once. Each pair of neighbours is only counted once in Equation (2.16), compared to twice in Equation (2.11), thus the factor of $\frac{1}{2}$ in the penalty term is removed.

The conditional posterior distribution for ϕ_i is:

$$\pi_\phi(\phi_i \,|\, y, \Theta \backslash \{\phi_i\}) \propto \left(\prod_{p \in P_i} \prod_{f=1}^{F_i} \left\{ \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\beta_{p,i,f} \right\} \right). \quad (2.17)$$

2.2 ℓ_1-Dirichlet Outlier Model

Some of the manuscripts in the Atlas are themselves copies from source texts. When the manuscript's scribe 'translated' from the dialect of a source text into their own dialect, sometimes they would copy spellings from the source text, rather than using one of the spellings they would normally use. Such copied forms are outliers, which do not reflect the written dialect of the scribe.

For example, suppose a source text contained only the form 'BROTHERE' of the item 'BROTHER', and a given scribe normally instead used the form 'BRODRE'.

When this scribe copied this source text, the scribe copied the form 'BROTHERE' directly from the source text, but translated the forms of all other items into forms he would normally use. In such a case, the form 'BROTHERE' would be an outlier in the resulting manuscript.

In this section, we alter the likelihood from the ℓ_1-Dirichlet zero-inflated model (presented in Equation (2.8)) to explicitly model this possibility, and describe the resulting extended model, which we call the ℓ_1-Dirichlet outlier model.

The reader is reminded that we have a lattice of cells $x \in \{1, \ldots, C\}$ laid over the geographic map; that $\eta_{x,i,f}$ represents the probability that item i takes form f in any manuscript originating from within cell x; that μ_i is the mean number of times item i is used in a manuscript; and that ϕ_i represents the (additional) zero-inflation probability for form f of item i not being used in profile p.

2.2.1 Likelihood

Let $\bar{\eta}_{i,f}$ be the average probability of using form $f \in \{1, \ldots, F_i\}$ of item i across the lattice, such that $\bar{\eta}_{i,f} = \frac{1}{C} \sum_{x=1}^{C} \eta_{x,i,f}$. Notice that $\sum_{f=1}^{F_i} \bar{\eta}_{i,f} = 1$. Let $\psi_{p,i}$ be the probability that item $i \in \{1, \ldots, I\}$ in profile $p \in \{1, \ldots, P\}$ contains aberrant forms (i.e. the item is an outlier).

We alter the ℓ_1-Dirichlet zero-inflated model so that if item i in profile p contains aberrant forms, we replace the local probabilities $\eta_{x,i,f}$ with the global probabilities $\bar{\eta}_{i,f}$ in the likelihood. This is done because $\bar{\eta}_{i,f}$ is the probability of using form f of

item i in profile p if we don't know the profile's location:

$$\begin{aligned}\pi(y_{p,i,f} = 1 \,|\, i) &= \sum_{x=1}^{C} \pi(y_{p,i,f} = 1, x_p = x) \\ &= \sum_{x=1}^{C} \pi(y_{p,i,f} = 1 \,|\, x_p = x)\, \pi(x_p = x) \\ &= \bar{\eta}_{i,f}. \end{aligned} \tag{2.18}$$

With $\Theta = \{\eta, \mu, x, \phi, \psi\}$, the likelihood for data y now becomes

$$\mathrm{p}(y \,|\, \Theta, O) \propto \prod_{p=1}^{P} \prod_{i \in I_p} \prod_{f=1}^{F_i} \Big\{ \phi_i(1 - y_{p,i,f}) + (1 - \phi_i) \Big[(1 - O_{p,i}) \beta_{p,i,f} + O_{p,i} \bar{\beta}_{p,i,f} \Big] \Big\}, \tag{2.19}$$

where $\bar{\beta}_{p,i,f} = (1 - e^{-\mu_i \bar{\eta}_{i,f}})^{y_{p,i,f}} (e^{-\mu_i \bar{\eta}_{i,f}})^{1 - y_{p,i,f}}$, and $O_{p,i} \sim \text{Bernoulli}(\psi_{p,i})$ is the outlier indicator variable for item i in profile p. The ℓ_1-Dirichlet zero-inflated model is nested within this ℓ_1-Dirichlet outlier model, since it can be obtained if all $\psi_{p,i} = 0$.

We assume the product form $\psi_{p,i} = \psi_p^{(P)} \psi_i^{(I)}$ in order to keep the number of new parameters manageable. We can think of $\psi_i^{(I)}$ as the probability for an event that allows the form-choice of item i to be aberrant, and $\psi_p^{(P)}$ as the probability for an event that allows the form-choice by the scribe of profile p to be aberrant. Item i in profile p is thus an outlier if both of these events occur.

The outlier probability parameters ψ are of greater interest than the outlier indicator variables O, so we marginalise the indicators out of the likelihood:

$$\begin{aligned}
\mathrm{p}(y \mid \Theta) &= \prod_{p=1}^{P} \prod_{i \in I_p} \left\{ \sum_{O_{p,i}=0}^{1} \mathrm{p}(y_{p,i} | \eta_{x_p,i,f}, x_p, \mu_i, \phi_i, \psi_{p,i}, O_{p,i}) \times \mathrm{p}(O_{p,i} | \psi_{p,i}) \right\} \\
&= \prod_{p=1}^{P} \prod_{i \in I_p} \left\{ \sum_{O_{p,i}=0}^{1} \left[\prod_{f=1}^{F_i} \left\{ \phi_i(1 - y_{p,i,f}) + (1 - \phi_i) \Big[(1 - O_{p,i}) \beta_{p,i,f} + O_{p,i} \bar{\beta}_{p,i,f} \Big] \right\} \right] \right. \\
&\qquad \left. \times \mathrm{p}(O_{p,i} | \psi_{p,i}) \right\},
\end{aligned}$$

and thus,

$$\begin{aligned}
\mathrm{p}(y \mid \Theta) \propto \prod_{p=1}^{P} \prod_{i \in I_p} \Bigg\{ & (1 - \psi_p^{(P)} \psi_i^{(I)}) \left(\prod_{f=1}^{F_i} \phi_i(1 - y_{p,i,f}) + (1 - \phi_i) \beta_{p,i,f} \right) \\
& + \psi_p^{(P)} \psi_i^{(I)} \left(\prod_{f=1}^{F_i} \phi_i(1 - y_{p,i,f}) + (1 - \phi_i) \bar{\beta}_{p,i,f} \right) \Bigg\}. \qquad (2.20)
\end{aligned}$$

2.2.2 Prior Distributions

The model parameters are now $\Theta = (\eta, \mu, x, \phi, \psi)$, with ψ being a set of $I + P$ probabilities, and all other parameters as before. We continue to use the prior models specified in Equations (2.9), (2.10), (2.11), (2.12) for x, μ, η, and ϕ respectively.

We use beta prior distributions for the probabilities $\psi_i^{(I)}$ and $\psi_p^{(P)}$, with hyperparameters $(\sigma_1, \sigma_2)^T$ and $(\rho_1, \rho_2)^T$:

$$\pi_\psi(\psi) = \left\{ \prod_{p=1}^{P} (\psi_p^{(P)})^{\rho_1 - 1} (1 - \psi_p^{(P)})^{\rho_2 - 1} \right\} \left\{ \prod_{i=1}^{I} (\psi_i^{(I)})^{\sigma_1 - 1} (1 - \psi_i^{(I)})^{\sigma_2 - 1} \right\}. \qquad (2.21)$$

We set $(\sigma_1, \sigma_2) = (2, 3)$ for each item $i \in \{1, \ldots, I\}$, and $(\rho_1, \rho_2) = (2, 3)$ for each profile $p \in \{1, \ldots, P\}$, as we believed these choices to give weight to sensible values for

both the component probabilities $\psi_i^{(I)}$ and $\psi_p^{(P)}$ and the overall outlier probabilities $\psi_{p,i}$, as well as being relatively diffuse. The priors are displayed in Figure 2.2.

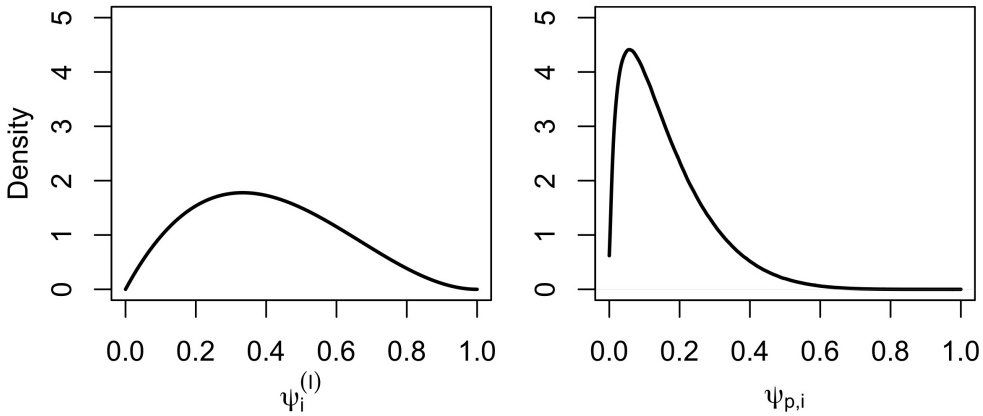


Figure 2.2: Left: beta prior distribution chosen for $\psi_i^{(I)}$, for $i \in \{1, \ldots, I\}$ (this distribution was also used as a prior for the $\psi_p^{(P)}$, for $p \in \{1, \ldots, P\}$). Right: the resulting prior distribution induced for $\psi_{p,i}$, for $i \in \{1, \ldots, I\}$ and $p \in \{1, \ldots, P\}$.

2.2.3 Posterior Distribution

Combining the observation model presented in Equation (2.20) with the prior distributions in Equations (2.9), (2.10), (2.11), (2.12) and (2.21) gives us the joint posterior distribution:

$$\begin{aligned}\pi(\Theta \,|\, y) \propto \left(\prod_{i=1}^{I} \mu_i^{a_i-1} e^{-b_i\mu_i}\right) \left(\prod_{p=1}^{P} \prod_{i \in I_p} \left\{(1-\psi_p^{(P)}\psi_i^{(I)}) \left[\prod_{f=1}^{F_i} \phi_i(1-y_{p,i,f}) + (1-\phi_i)\beta_{p,i,f}\right]\right.\right. \\ \left.\left. + \psi_p^{(P)}\psi_i^{(I)} \left[\prod_{f=1}^{F_i} \phi_i(1-y_{p,i,f}) + (1-\phi_i)\bar{\beta}_{p,i,f}\right]\right\}\right) \left(\prod_{p=1}^{P} (\psi_p^{(P)})^{\rho_1-1}(1-\psi_p^{(P)})^{\rho_2-1}\right) \\ \times \left(\prod_{i=1}^{I} (\psi_i^{(I)})^{\sigma_1-1}(1-\psi_i^{(I)})^{\sigma_2-1}\right) \left(\prod_{i=1}^{I} \prod_{f=1}^{F_i} \prod_{x=1}^{C} \eta_{x,i,f}^{\chi-1} e^{-\frac{\theta_i}{2}\sum_{x' \in \mathcal{N}(x)} |\eta_{x,i,f} - \eta_{x',i,f}|}\right).\end{aligned} \tag{2.22}$$

As mentioned earlier, we obtain the ℓ_1-Dirichlet zero-inflated model if we set $\psi^{(P)} = 0$ and $\psi^{(I)} = 0$ for all $p \in \{1, \ldots, P\}$ and $i \in \{1, \ldots, I\}$.

2.2.3.1 Conditional Posterior Distributions

From Equation (2.22), we can derive conditional posterior distributions for each of the parameters, which we later use in the MCMC algorithm outlined in Section 3.2. We omit the conditional posterior distributions for x_p, μ_i, ϕ_i and $\eta_{x,i}$ here as they are straightforward variations of expressions provided earlier in this chapter.

The conditional posterior distribution for $\psi_i^{(I)}$ is:

$$\pi_\psi(\psi_i^{(I)} \mid y, \Theta \backslash \{\psi_i^{(I)}\}) \propto \left(\prod_{p \in P_i} \left\{ (1 - \psi_p^{(P)}\psi_i^{(I)}) \left[\prod_{f=1}^{F_i} \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\beta_{p,i,f} \right] + \psi_p^{(P)}\psi_i^{(I)} \left[\prod_{f=1}^{F_i} \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\bar{\beta}_{p,i,f} \right] \right\} \right) \times (\psi_i^{(I)})^{\sigma_1 - 1}(1 - \psi_i^{(I)})^{\sigma_2 - 1}. \tag{2.23}$$

The conditional posterior distribution for $\psi_p^{(P)}$ is:

$$\pi_\psi(\psi_p^{(P)} \mid y, \Theta \backslash \{\psi_p^{(P)}\}) \propto \left(\prod_{i \in I_p} \left\{ (1 - \psi_p^{(P)}\psi_i^{(I)}) \left[\prod_{f=1}^{F_i} \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\beta_{p,i,f} \right] + \psi_p^{(P)}\psi_i^{(I)} \left[\prod_{f=1}^{F_i} \phi_i(1 - y_{p,i,f}) + (1 - \phi_i)\bar{\beta}_{p,i,f} \right] \right\} \right) \times (\psi_p^{(P)})^{\rho_1 - 1}(1 - \psi_p^{(P)})^{\rho_2 - 1}. \tag{2.24}$$

2.3 ℓ_2-Logistic Models

In this section, we propose alterations to the ℓ_1-Dirichlet zero-inflated model (presented in Section 2.1) and to the ℓ_1-Dirichlet outlier model (presented in Section 2.2). We call these altered models the ℓ_2-logistic zero-inflated and the ℓ_2-logistic outlier models respectively.

The alteration to the models consists of modelling a set of parameters γ (from which the probability field parameters η are derived), rather than modelling the η parameters directly. This approach is similar to that taken in Wasser et al. (2004).

We assign a prior to the γ-fields, rather than the η-fields. The γ-fields are not restricted to lie within the simplex as the η-fields are, so this approach removes a constraint which is slightly awkward for MCMC. In particular, it makes efficient MALA updates more straightforward (or even possible).

The reader is once again reminded that we have a lattice of cells $x \in \{1, \ldots, C\}$ laid over the geographic map; that $\eta_{x,i,f}$ represents the probability that item i takes form f in any manuscript originating from within cell x; that μ_i is the mean number of times item i is used in a manuscript; and that ϕ_i represents the (additional) zero-inflation probability for form f of item i not being used in profile p.

The η-fields are derived from the γ-fields using

$$\eta_{x,i,f} = \frac{e^{\gamma_{x,i,f}}}{\sum_{f'=1}^{F_i} e^{\gamma_{x,i,f'}}}, \tag{2.25}$$

where $\gamma_{x,i,1} = 0$ for each item i and cell x. The inverse relationship is defined by

$$\gamma_{x,i,f} = \log\left(\frac{\eta_{x,i,f}}{\eta_{x,i,1}}\right). \tag{2.26}$$

2.3.1 Prior Distribution for γ

For the ℓ_1-Dirichlet zero-inflated and outlier models, we modelled the η-fields using a Dirichlet-ℓ_1 field as specified in Equation (2.11). This prior distribution was essentially

a Dirichlet distribution for the form-usage probabilities η in each cell, smoothed across cells.

For the ℓ_2-logistic models, we set the prior distribution for the γ parameters to

$$\pi_\gamma(\gamma) \quad \propto \quad \prod_{i=1}^{I}\prod_{x=1}^{C}\left[e^{-\frac{1}{2}(\tilde{\gamma}_{x,i}-\lambda_i)^T\Sigma_i^{-1}(\tilde{\gamma}_{x,i}-\lambda_i)}\prod_{f=1}^{F_i}e^{-\frac{1}{2\sigma_i^2}\sum_{x'\in\mathcal{N}(x)}(\gamma_{x,i,f}-\gamma_{x',i,f})^2}\right], \tag{2.27}$$

where $\tilde{\gamma}_{x,i} = (\gamma_{x,i,2}, \ldots, \gamma_{x,i,F_i})$, λ_i is a vector of length $F_i - 1$, and Σ_i is a $(F_i - 1) \times (F_i - 1)$ matrix.

This prior distribution, which we name a ℓ_2-MVN field, is essentially a multivariate-normal distribution for the γ-fields in each cell, smoothed across neighbouring cells through the exponential penalty term. We use a quadratic penalty $(\gamma_{x,i,f} - \gamma_{x',i,f})^2$, modelling diffusion of forms, rather than an ℓ_1 penalty as before.

The MVN term $e^{-\frac{1}{2}(\tilde{\gamma}_{x,i}-\lambda_i)^T\Sigma_i^{-1}(\tilde{\gamma}_{x,i}-\lambda_i)}$ acts similarly to the Dirichlet term in the Dirichlet-ℓ_1 prior distribution, allowing us to incorporate prior knowledge about the relative usage rates of the forms f of an item i in a single cell. Based on Aitchison & Shen (1980), this MVN(λ_i, Σ_i) distribution on the $\tilde{\gamma}_{x,i}$ scale induces a logistic-normal distribution LN(λ_i, Σ_i) on the $\eta_{x,i}$ scale.

It is possible to choose λ_i and Σ_i so this distribution approximates a Dirichlet distribution. We found no issues with our earlier choice of within-cell model, so we make our ℓ_2-logistic model mimic that part of the ℓ_1-Dirichlet models. We really wish to change the spatial correlation imposed in the prior, but not anything else. This

change in the spatial correlation gives us a higher-level sensitivity analysis - does this choice make a difference in the results obtained? Additionally, and more importantly, this choice leads to an unconstrained spatial field on which it may be easier to do MCMC inference.

Aitchison & Shen (1980) give the appropriate choice of MVN parameters λ_i and Σ_i to mimic the Dirichlet distribution with parameters $(\chi_1, \chi_2, \ldots, \chi_{F_i})$, by minimising the Kullback-Liebler divergence between the two distributions.

For $f \in \{2, \ldots, F_i\}$, the choices are:

$$\begin{aligned} \lambda_{i,f-1} &= \delta(\chi_f) - \delta(\chi_1), \\ (\Sigma_i)_{f-1,f-1} &= \delta'(\chi_f) + \delta'(\chi_1), \\ (\Sigma_i)_{f-1,f'} &= \delta'(\chi_1) \text{ when } f - 1 \neq f', \end{aligned} \tag{2.28}$$

where δ and δ' are the digamma and trigamma functions respectively.

When earlier choosing the prior for the within-cell correlation of $\eta_{x,i,f}$, we set the Dirichlet parameter $\chi = 1$ for all forms f of each item $i \in \{1, \ldots, I\}$ (see Section 2.1.2.3). Thus, here we set $\chi_f = 1$ for $f \in \{1, \ldots, F_i\}$, and use Equation (2.28) to accordingly set λ_i and Σ_i. The resulting parameters, for $f \in \{2, \ldots, F_i\}$, are:

$$\begin{aligned} \lambda_{i,f-1} &= 0, \\ (\Sigma_i)_{f-1,f-1} &= 3.290, \\ (\Sigma_i)_{f-1,f'} &= 1.645 \text{ when } f - 1 \neq f'. \end{aligned} \tag{2.29}$$

We found, using a simple simulation check, that these choices did indeed give a good approximation to the Dirichlet$(1, \ldots, 1)$ distribution. The Dirichlet-approximated prior distribution for a form usage probability η of a five-form item is shown in Figure 2.3, alongside the corresponding Dirichlet$(1, \ldots, 1)$ prior used in the ℓ_1-Dirichlet models. We note these closely coincide.

Earlier we noted how the number of forms F_i of item i varies greatly between different items, meaning the average size of $\eta_{x,i,f}$, and hence the scale for variation of η between cells, decreases as F_i increases. Thus, we ensured the level of smoothing in the Dirichlet-ℓ_1 field depended on F_i.

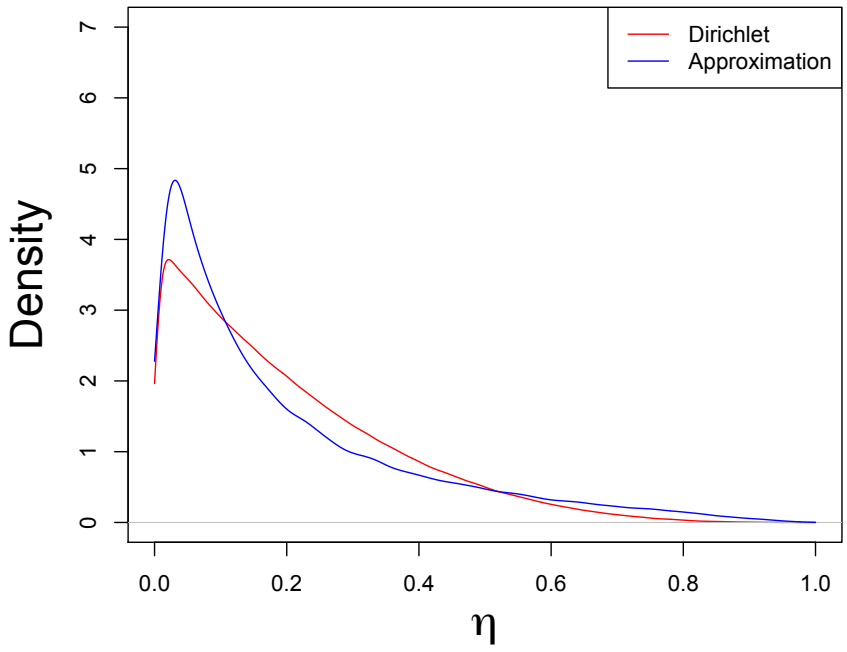


Figure 2.3: The Dirichlet within-cell prior distribution for the usage probability η of a form f of an item i with $F_i = 5$ forms is shown in red. The prior distribution induced on η by our MVN prior on γ is shown in blue, with the MVN parameters set as in Equation 2.28 to approximate the Dirichlet distribution in red.

Here, the level of smoothing is controlled by the parameters σ_i^2. Differences $\Delta\eta = |\eta_{x',i,f} - \eta_{x,i,f}|$ much smaller than $\frac{1}{F_i}$ are 'small', so we wish to choose σ_i^2 such that we smooth differences of order $\frac{1}{F_i}$, unless the data overwhelms the prior. This will set the scale of variation of γ.

With $\Delta\gamma = |\gamma x', i, f - \gamma_{x,i,f}|$, we set σ_i^2 such that $|\Delta\gamma| > \sigma_i$ approximately one-third of the time; thus, we want to choose σ_i^2 such that $\Delta\eta = 1/F_i$ gives $\Delta\gamma = \sqrt{2}\sigma_i$ (thinking of $\gamma_{x,i,f}$ and $\gamma_{x',i,f}$ as zero-mean normal random variables). Therefore, consider two γ vectors in neighbouring cells x and x' that differ only for one form f^*, i.e. $\gamma_{x,i,f} = \gamma_{x',i,f}$ for $f \in \{1, \ldots, F_i\} \backslash f^*$. Let $\gamma_{x',i,f} = 0$ for $f \in \{1, \ldots, F_i\}$, and let $\gamma_{x,i,f} = 0$ for $f \in \{1, \ldots, F_i\} \backslash f^*$, so that γ_{x,i,f^*} is the only non-zero entry. So, if

$$\eta_{x,i,f^*} - \eta_{x',i,f^*} = \frac{1}{F_i},$$

then

$$\left(\frac{e^{\gamma_{x,i,f^*}}}{\sum_{f'=1}^{F_i} e^{\gamma_{x,i,f'}}} - \frac{e^{\gamma_{x',i,f^*}}}{\sum_{f'=1}^{F_i} e^{\gamma_{x',i,f'}}} \right) = \frac{1}{F_i}.$$

Substituting in the assumed γ-values, we find

$$\gamma_{x,i,f^*} = \log \left\{ \frac{2(F_i - 1)}{F_i - 2} \right\}$$

$$\implies \gamma_{x,i,f^*} - \gamma_{x',i,f^*} \simeq \log 2$$

$$\implies \sigma_i^2 \simeq \{\log 2\}^2/2 \tag{2.30}$$

Thus, with $\theta > 0$, we set

$$\sigma_i^2 = \frac{\{\log 2\}^2}{2\theta}. \tag{2.31}$$

We want $(\gamma_{x,i,f^*} - \gamma_{x',i,f^*})^2/\sigma^2$ to be $O(1)$ when $\eta_{x,i,f^*} - \eta_{x',i,f^*}$ is $O(1/F_i)$, so it is natural to choose $\theta = 1$. Sensitivity analysis suggests that an appropriate level of smoothing can be obtained when $1 \leq \theta \leq 2$, but over-smoothing occurs when $\theta > 2$. Having found that $\theta = 1$ does indeed give a good level of smoothing, we use this value for analyses throughout this report.

The smoothing (as with the ℓ_1-Dirichlet models) is partly set by the number of forms, but also by the spatial scale. It is clear that if the cells of the lattice were made larger, less smoothing for γ would be required. We arrive at the same conclusions relating to this issue as we did earlier with the ℓ_1-Dirichlet models, for the same reasons.

2.3.2 Posterior Distributions

2.3.2.1 ℓ_2-Logistic Zero-Inflated Model

Combining the observation model presented in Equation (2.8) with the prior distributions in Equations (2.9), (2.10), (2.12), and (2.27) gives us the joint posterior distribution:

$$\pi(\Theta \,|\, y) \propto \left(\prod_{i=1}^{I} \mu_i^{a_i - 1} e^{-b_i \mu_i} \right) \left(\prod_{p=1}^{P} \prod_{i \in I_p} \prod_{f=1}^{F_i} \phi_i (1 - y_{p,i,f}) + (1 - \phi_i) \beta_{p,i,f} \right) \times \left(\prod_{i=1}^{I} \prod_{x=1}^{C} \left[e^{-\frac{1}{2}(\tilde{\gamma}_{x,i} - \lambda_i)^T \Sigma_i^{-1} (\tilde{\gamma}_{x,i} - \lambda_i)} \prod_{f=1}^{F_i} e^{-\frac{1}{2\sigma_i^2} \sum_{x' \in \mathcal{N}(x)} (\gamma_{x,i,f} - \gamma_{x',i,f})^2} \right] \right). \tag{2.32}$$

2.3.2.2 ℓ_2-Logistic Outlier Model

Combining the observation model presented in Equation (2.20) with the prior distributions in Equations (2.9), (2.10), (2.12), (2.21) and (2.27) gives us the joint posterior distribution:

$$\begin{aligned}\pi(\Theta \,|\, y) \propto & \left(\prod_{i=1}^{I} \mu_i^{a_i-1} e^{-b_i\mu_i}\right)\left(\prod_{p=1}^{P}\prod_{i\in I_p}\left\{(1-\psi_p^{(P)}\psi_i^{(I)})\left[\prod_{f=1}^{F_i}\phi_i(1-y_{p,i,f})+(1-\phi_i)\beta_{p,i,f}\right]\right.\right. \\ & \left.\left. + \psi_p^{(P)}\psi_i^{(I)}\left[\prod_{f=1}^{F_i}\phi_i(1-y_{p,i,f})+(1-\phi_i)\bar{\beta}_{p,i,f}\right]\right\}\right)\left(\prod_{p=1}^{P}(\psi_p^{(P)})^{\rho_1-1}(1-\psi_p^{(P)})^{\rho_2-1}\right) \\ & \times\left(\prod_{i=1}^{I}\prod_{x=1}^{C}\left[e^{-\frac{1}{2}(\tilde{\gamma}_{x,i}-\lambda_i)^T\Sigma_i^{-1}(\tilde{\gamma}_{x,i}-\lambda_i)}\prod_{f=1}^{F_i}e^{-\frac{1}{2\sigma_i^2}\sum_{x'\in\mathcal{N}(x)}(\gamma_{x,i,f}-\gamma_{x',i,f})^2}\right]\right) \\ & \times\left(\prod_{i=1}^{I}(\psi_i^{(I)})^{\sigma_1-1}(1-\psi_i^{(I)})^{\sigma_2-1}\right). \end{aligned} \tag{2.33}$$

As with the ℓ_1-Dirichlet models, the ℓ_2-logistic zero-inflation model is nested within this ℓ_2-logistic outlier model, since it can be obtained if all $\psi^{(I)} = 0$ and $\psi^{(P)} = 0$.

CHAPTER 3

MCMC Methods

3.1 Monte-Carlo Sample-Based Inference for ℓ_1-Dirichlet Zero-Inflated Model

In this section, we outline the Markov Chain Monte Carlo (MCMC) algorithm used to fit the model described in Section 2.1 to either the primary data described in Section 1.3.1, or the coarsened data described in Section 1.3.2. We used random-walk Metropolis-within-Gibbs updates for the model parameters. Where possible, the calculations described were performed on the log-scale for numerical stability.

Let $\Theta^{(0)} = (x^{(0)}, \eta^{(0)}, \mu^{(0)}, \phi^{(0)})$ be initial values for the parameters Θ of the model. $x^{(0)}$, $\mu^{(0)}$ and $\phi^{(0)}$ are drawn from the prior distributions specified in Equations (2.9), (2.10) and (2.12) respectively. In each cell $x = 1, \ldots, C$, we set

$$\eta_{x,i}^{(0)} \sim \text{Dirichlet}(\chi), \tag{3.1}$$

where χ was a vector of ones, with length F_i.

Suppose $\Theta^{(t)} = (x, \eta, \mu, \phi)$. Then $\Theta^{(t+1)}$ is found as follows:

1. **Update for form usage probability fields η.**

   For each item $i \in \{1, \ldots, I\}$, and each location $x \in \{1, \ldots, C\}$:

   (i) Randomly choose non-equal $f_1, f_2 \in \{1, 2, \ldots, F_i\}$. Without loss of generality, say $\eta_{x,i,f_1} < \eta_{x,i,f_2}$.

   (ii) Draw $\delta \sim U(-\eta_{x,i,f_1}, \eta_{x,i,f_2})$.

   (iii) Set:

   $$\eta'_{x,i,f} = \begin{cases} \eta_{x,i,f} & \text{if } f \notin \{f_1, f_2\} \\ \eta_{x,i,f} + \delta & \text{if } f = f_1 \\ \eta_{x,i,f} - \delta & \text{if } f = f_2 \end{cases} . \tag{3.2}$$

   (iv) By construction (illustrated in Figure 3.1), the proposal distribution q is symmetric, and therefore $q(\eta'_{x,i}|\eta_{x,i}) = q(\eta_{x,i}|\eta'_{x,i})$. Thus, draw $u \sim U(0, 1)$, and accept the proposed $\eta'_{x,i}$ if $u \leq \alpha_\eta(\eta_{x,i}, \eta_{x',i})$, where

   $$\alpha_\eta(\eta_{x,i}, \eta'_{x,i}) = \min\left(1, \frac{\pi_\eta(\eta'_{x,i}|y, \Theta \backslash \{\eta'_{x,i}\})}{\pi_\eta(\eta_{x,i}|y, \Theta \backslash \{\eta_{x,i}\})}\right). \tag{3.3}$$

   We give this acceptance probability (and the ones which follow) in terms of conditional distributions (rather than the full joint posterior) because it is faster to calculate using just non-cancelling terms.

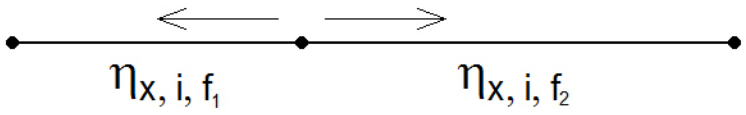


Figure 3.1: The proposal distribution for the form usage probabilities, $q(\cdot|\eta_{x,i})$, can be viewed as an exchange of length between η_{x,i,f_1} and η_{x,i,f_2}, where the 'cut' between the two usage probabilities is moved along the number line.

2. **Update for item usage rates μ.**

   For each item $i \in \{1, \ldots, I\}$:

   (i) Draw $\epsilon \sim N(0, W_\mu)$. We find from trial runs that $W_\mu = 0.75$ is sensible.

   (ii) Set $\mu_i' = \mu_i + \epsilon$ (and $\mu_j' = \mu_j$ for $j \neq i$).

   (iii) The proposal distribution $q(\cdot|\mu_i) = N(\mu_i, W_\mu)$ is symmetric, so $q(\mu_i'|\mu_i) = q(\mu_i|\mu_i')$. Thus, draw $u \sim U(0,1)$, and accept the proposed μ_i' if $u \leq \alpha_\mu(\mu_i, \mu_i')$, where

$$\alpha_\mu(\mu_i, \mu_i') = \min\left(1, \frac{\pi_\mu(\mu_i'|y, \Theta \backslash \{\mu_i'\})}{\pi_\mu(\mu_i|y, \Theta \backslash \{\mu_i\})}\right). \tag{3.4}$$

3. **Update for zero-inflation probabilities ϕ.**

   For each item $i \in \{1, \ldots, I\}$:

   (i) Draw $\epsilon \sim N(0, W_\phi)$. We find from trial runs that $W_\phi = 0.125$ is sensible.

   (ii) Set $\phi_i' = \phi_i + \epsilon$.

   (iii) The proposal distribution $q(\cdot|\phi_i) = N(\phi_i, W_\phi)$ is symmetric, so $q(\phi_i'|\phi_i) = q(\phi_i|\phi_i')$. Thus, draw $u \sim U(0,1)$, and accept the proposed ϕ_i' if $u \leq \alpha_\phi(\phi_i, \phi_i')$, where

$$\alpha_\phi(\phi_i, \phi_i') = \min\left(1, \frac{\pi_\phi(\phi_i'|y, \Theta \backslash \{\phi_i'\})}{\pi_\phi(\phi_i|y, \Theta \backslash \{\phi_i\})}\right). \tag{3.5}$$

4. **Update for profile locations x.**

   For each profile $p \notin A$:

   (i) For each cell $x \in 1, \ldots, C$, evaluate $\pi_x(x \mid y, \Theta \backslash \{x_p\})$ (as defined in Equation (2.14)).

   (ii) π_x is only defined up to a normalising constant, so we normalise these quantities by calculating for each cell $x \in \{1, \ldots, C\}$

$$\frac{\pi_x(x \mid y, \Theta \backslash \{x_p\})}{\sum_{x=1}^{C} \pi_x(x \mid y, \Theta \backslash \{x_p\})}, \tag{3.6}$$

   which gives the conditional probability for profile p to be in cell x. A new location x_p is then sampled straightforwardly using these probabilities.

3.2 Extending to the ℓ_1-Dirichlet Outlier Model

If we use the ℓ_1-Dirichlet outlier model described in Section 2.2, rather than the ℓ_1-Dirichlet zero-inflated model of Section 2.1.1, we use the slightly altered MCMC algorithm outlined below.

Let $\Theta^{(0)} = (x^{(0)}, \eta^{(0)}, \mu^{(0)}, \phi^{(0)}, \psi^{(0)})$ be the initial values for the parameters Θ of the model. $\psi^{(0)}$ are drawn from the prior distribution specified by Equation (2.21), and all other initial values are set using the same methods as earlier.

Suppose $\Theta^{(t)} = (x, \eta, \mu, \phi, \psi)$. Then $\Theta^{(t+1)}$ is found as follows:

1. **Update for form usage probability fields η.**

2. **Update for item usage rates μ.**

3. **Update for zero-inflation probabilities ϕ.**

   Update parameters η, μ and ϕ as in Section 3.1, though with conditional distributions derived from Equation (2.22) instead of Equation (2.13).

4. **Update for outlier probabilities $\psi_i^{(I)}$.**

   For each item $i \in \{1, \ldots, I\}$:

   (i) Draw $\epsilon \sim N(0, W_{\psi^{(I)}})$. We find from trial runs that $W_{\psi^{(I)}} = 0.0625$ is sensible.

   (ii) Set $\psi_i'^{(I)} = \psi_i^{(I)} + \epsilon$.

   (iii) The proposal distribution $q(\cdot|\psi_i^{(I)}) = N(\psi_i^{(I)}, W_{\psi^{(I)}})$ is symmetric, thus, draw $u \sim U(0,1)$, and accept the proposed $\psi_i'^{(I)}$ if $u \leq \alpha_{\psi^{(I)}}(\psi_i^{(I)}, \psi_i'^{(I)})$, where

$$\alpha_{\psi^{(I)}}(\psi_i^{(I)}, \psi_i'^{(I)}) = \min\left(1, \frac{\pi_\psi(\psi_i'^{(I)}|y, \Theta\backslash\{\psi_i'^{(I)}\})}{\pi_\psi(\psi_i^{(I)}|y, \Theta\backslash\{\psi_i^{(I)}\})}\right), \tag{3.7}$$

   and $\pi_\psi(\psi_i^{(I)}|y, \Theta\backslash\{\psi_i^{(I)}\})$ is given in Equation (2.23).

5. **Update for outlier probabilities $\psi_p^{(P)}$.**

   For each profile $p \in \{1, \ldots, P\}$:

   (i) Draw $\epsilon \sim N(0, W_{\psi^{(P)}})$. Though $W_{\psi^{(P)}}$ need not be the same as $W_{\psi^{(I)}}$, we similarly find from trial runs that $W_{\psi^{(P)}} = 0.0625$ is sensible.

   (ii) Set $\psi_p'^{(P)} = \psi_p^{(P)} + \epsilon$.

   (iii) The proposal distribution $q(\cdot|\psi_p^{(P)}) = N(\psi_p^{(P)}, W_{\psi^{(P)}})$ is symmetric, thus, draw $u \sim U(0,1)$, and accept the proposed $\psi_p'^{(P)}$ if $u \leq \alpha_{\psi^{(P)}}(\psi_p^{(P)}, \psi_p'^{(P)})$, where

$$\alpha_{\psi^{(P)}}(\psi_p^{(P)}, \psi_p'^{(P)}) = \min\left(1, \frac{\pi_\psi(\psi_p'^{(P)}|y, \Theta\backslash\{\psi_p'^{(P)}\})}{\pi_\psi(\psi_p^{(P)}|y, \Theta\backslash\{\psi_p^{(P)}\})}\right). \tag{3.8}$$

   and $\pi_\psi(\psi_i^{(P)}|y, \Theta\backslash\{\psi_i^{(P)}\})$ is given in Equation (2.24).

6. **Update for profile locations x.**

   Update parameters x in the same fashion as with the ℓ_1-Dirichlet zero-inflated model, though with the conditional distribution for x derived from Equation (2.22) instead of Equation (2.13).

3.3 Monte-Carlo Sample-Based Inference for ℓ_2-Logistic Models

3.3.1 ℓ_2-Logistic Outlier Model

In this section, we briefly outline the MCMC algorithm used with the ℓ_2-logistic outlier model described in Section 2.3.

Let $\Theta^{(0)} = (x^{(0)}, \gamma^{(0)}, \mu^{(0)}, \phi^{(0)}, \psi^{(0)})$ be initial values for the parameters Θ of the model. $x^{(0)}$, $\mu^{(0)}$, $\phi^{(0)}$ and $\psi^{(0)}$ are set using the same methods as earlier. In each cell $x = 1, \ldots, C$, for item $i = 1, \ldots, I$, we draw $\gamma_{x,i}^{(0)} = (\gamma_{x,i,2}^{(0)}, \ldots, \gamma_{x,i,F_i}^{(0)})$ from the multivariate normal distribution:

$$\gamma_{x,i}^{(0)} \sim \text{MVN}(\mathbf{0}, \Sigma_i). \tag{3.9}$$

Suppose $\Theta^{(t)} = (x, \gamma, \mu, \phi, \psi)$. Then $\Theta^{(t+1)}$ is found as follows:

1. **Update for form usage parameter fields γ.**

   For each item $i \in \{1, \ldots, I\}$, and each location $x \in \{1, \ldots, C\}$:

   (i) Set $\gamma_{x,i}^{'} \sim \text{MVN}(\gamma_{x,i}, (W_\gamma)_i \Sigma_i)$. We find that $W_\gamma)_i = \frac{1}{64}$ is sensible.

   (ii) By construction, the proposal distribution q is symmetric, and therefore $q(\gamma'_{x,i}|\gamma_{x,i}) = q(\gamma_{x,i}|\gamma'_{x,i})$. Thus, draw $u \sim U(0,1)$, and accept the proposed $\gamma'_{x,i}$ if $u \leq \alpha_\gamma(\gamma_{x,i}, \gamma_{x',i})$, where

$$\alpha_\gamma(\gamma_{x,i}, \gamma'_{x,i}) = \min\left(1, \frac{\pi_\gamma(\gamma'_{x,i}|y, \Theta \backslash \{\gamma'_{x,i}\})}{\pi_\gamma(\gamma_{x,i}|y, \Theta \backslash \{\gamma_{x,i}\})}\right). \tag{3.10}$$

2. **Update for item usage rates μ.**

3. **Update for zero-inflation probabilities ϕ.**

4. **Update for outlier probabilities $\psi_i^{(I)}$.**

5. **Update for outlier probabilities $\psi_p^{(P)}$.**

6. **Update for profile locations x.**

   After converting from the γ-scale to the η-scale using Equation (2.26), update

parameters μ, ϕ, $\psi^{(P)}$, $\psi^{(I)}$ and x_p using the same methods as with the ℓ_1-Dirichlet outlier model (see Section 3.2).

3.3.2 ℓ_2-Logistic Zero-Inflated Model

The MCMC algorithm used to fit the ℓ_2-logistic zero-inflated model is the same as the one used to fit the ℓ_2-logistic outlier model, except that we skip steps 4 and 5, and set $\psi^{(P)} = 0$ and $\psi^{(I)} = 0$ for all $p \in \{1, \ldots, P\}$ and $i \in \{1, \ldots, I\}$ respectively.

3.4 Checking the MCMC Samplers

With synthetic data, we now compare theoretical distributions for the parameters of the ℓ_1-Dirichlet zero-inflated model to those found using the samplers from the algorithm presented in Section 3.1. These quick checks are not intended as rigorous proof that the algorithms sample from the correct distributions, rather, to provide some peace-of-mind. We find excellent agreement between estimated and theoretical distributions.

We work within a small 5×5 lattice L with cells $x \in \{1, \ldots, 25\}$, with a small number of items ($I = 8$) and forms ($F_i = (5, 5, 4, 4, 3, 3, 3, 3)$), and with a small number of profiles ($N = 50$).

Similar checks for the ℓ_1-Dirichlet outlier and the ℓ_2-logistic models are presented in Appendix A.

3.4.1 ℓ_1-Dirichlet Zero-Inflated Model

In this section, we check each of the sampling methods from the algorithm described in Section 3.1 for the ℓ_1-Dirichlet zero-inflated model. This model, as specified in Section 2.1, has parameters $\Theta = (x, \mu, \eta, \phi)$. We choose the following values for these parameters:

- **Locations** x_p such that there are two profiles in each cell $x \in \{1, \ldots, 25\}$.
- **Form usage probabilities** η as displayed in Figures 3.2 and 3.3, with two copies of each generated such that $\eta_1 = \eta_2$, $\eta_3 = \eta_4$, $\eta_5 = \eta_6$ and $\eta_7 = \eta_8$.
- **Item usage rates** $\mu = (3, 5, 3, 5, 3, 5, 3, 5)$, so that one copy of each η-field is used more frequently than the other.
- **Zero-inflation probabilities** $\phi = (0.05, 0.1, 0.05, 0.1, 0.05, 0.1, 0.05, 0.1)$ such that the more frequently used items have a higher probability of extra zeroes.

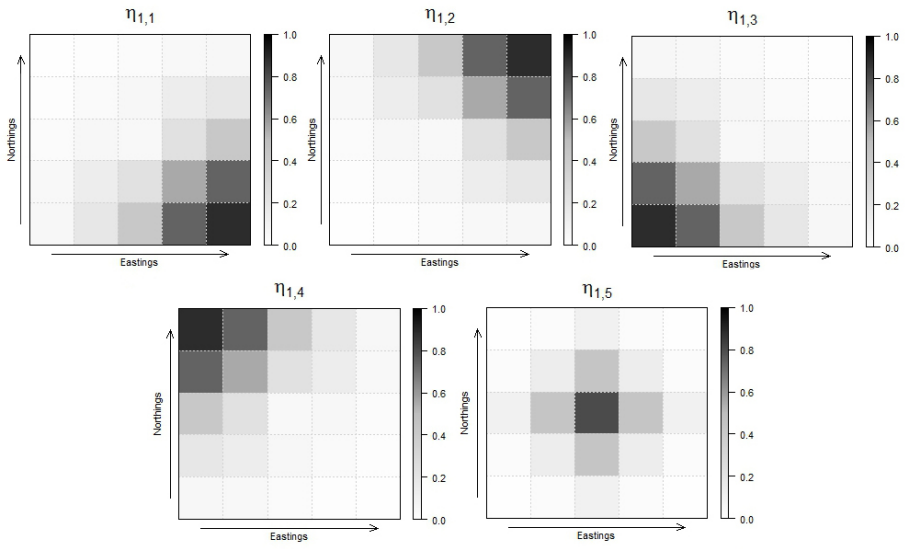


Figure 3.2: Assigned values of η_1, and η_2 (since $\eta_2 = \eta_1$), used throughout Section 3.4.1.

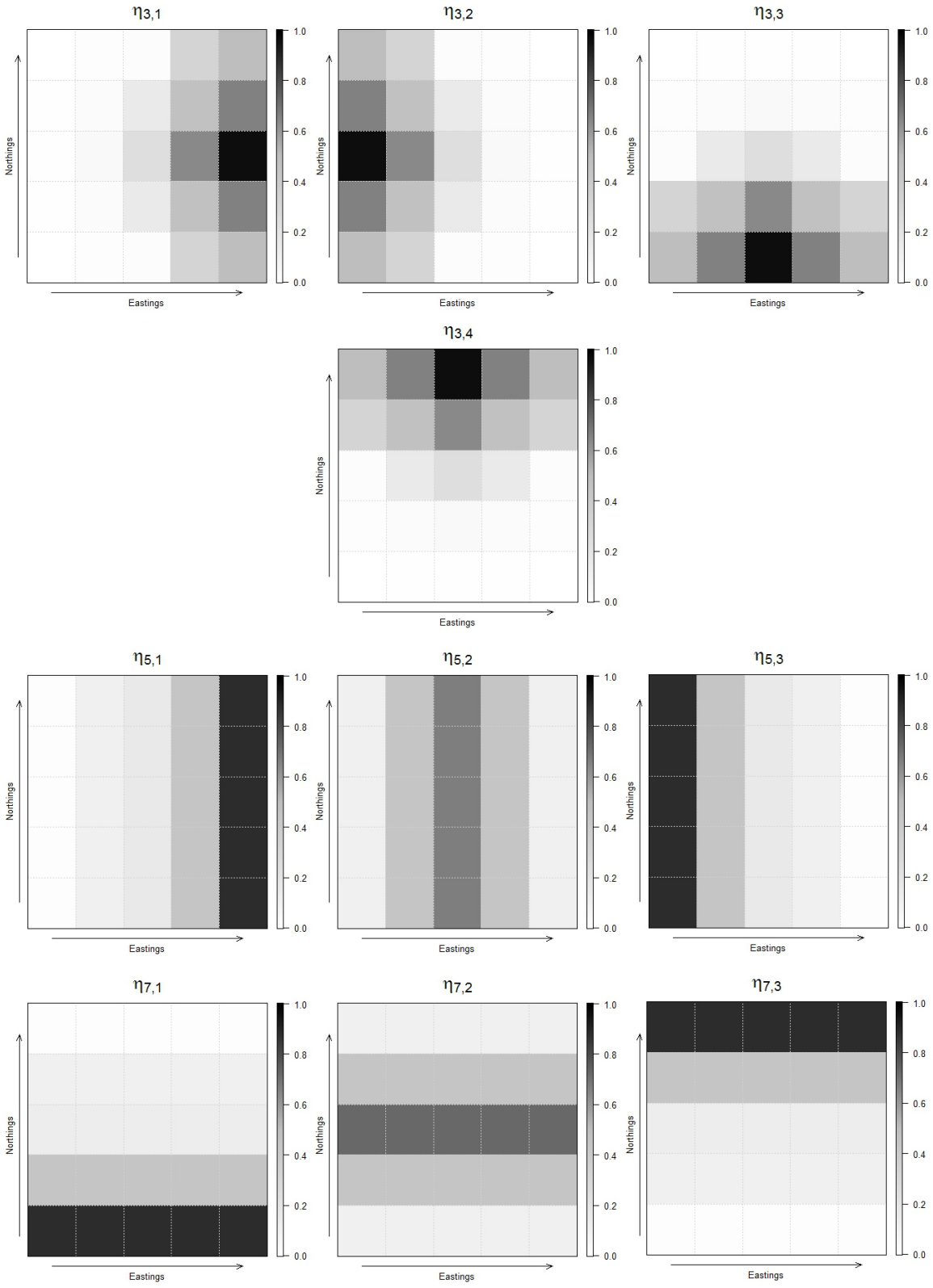


Figure 3.3: The top four plots give the assigned values of η_3 and η_4 (since $\eta_4 = \eta_3$); the middle three give the assigned values of η_5 and η_6 (since $\eta_6 = \eta_5$); and the bottom three give the assigned values of η_7 and η_8 (since $\eta_8 = \eta_7$) used throughout Section 3.4.1.

These parameter values are assigned, not simulated. Using them, we simulate data y for the profiles $p \in \{1, \ldots, 50\}$ using the generative process for this model.

3.4.1.1 Form Usage Probabilities

To test the η-field sampler from the MCMC algorithm for the ℓ_1-Dirichlet zero-inflated model, we use the sampler to estimate the posterior distributions for eight randomly selected $\eta_{x,i}$ vectors (one for each item) with all other parameters (x, μ, ϕ and all other $\eta_{x,i}$) fixed to their assigned values. We then compare the estimated posterior distributions to the true posterior distributions for these parameters (we obtain the normalising constant for these distributions using quadrature).

Figure 3.4 shows the exact posterior distributions for one of the $\eta_{x,i}$ vectors ($\eta_{11,5}$) alongside summaries of the MCMC samples obtained from our algorithm, where the numbers of MCMC samples taken is varied. We can see that the estimated posterior converges (slowly) to the exact posterior distribution.

Further evidence for convergence to the desired posterior distribution is provided in Figure 3.5. This shows the estimated marginal distributions for the components of $\eta_{11,5}$. The exact marginal posterior distributions are overlaid, and we note that the estimated marginal posterior distributions for all η parameters converge to the exact distributions. Similar figures can be obtained for the other η-fields, but are not included here for brevity.

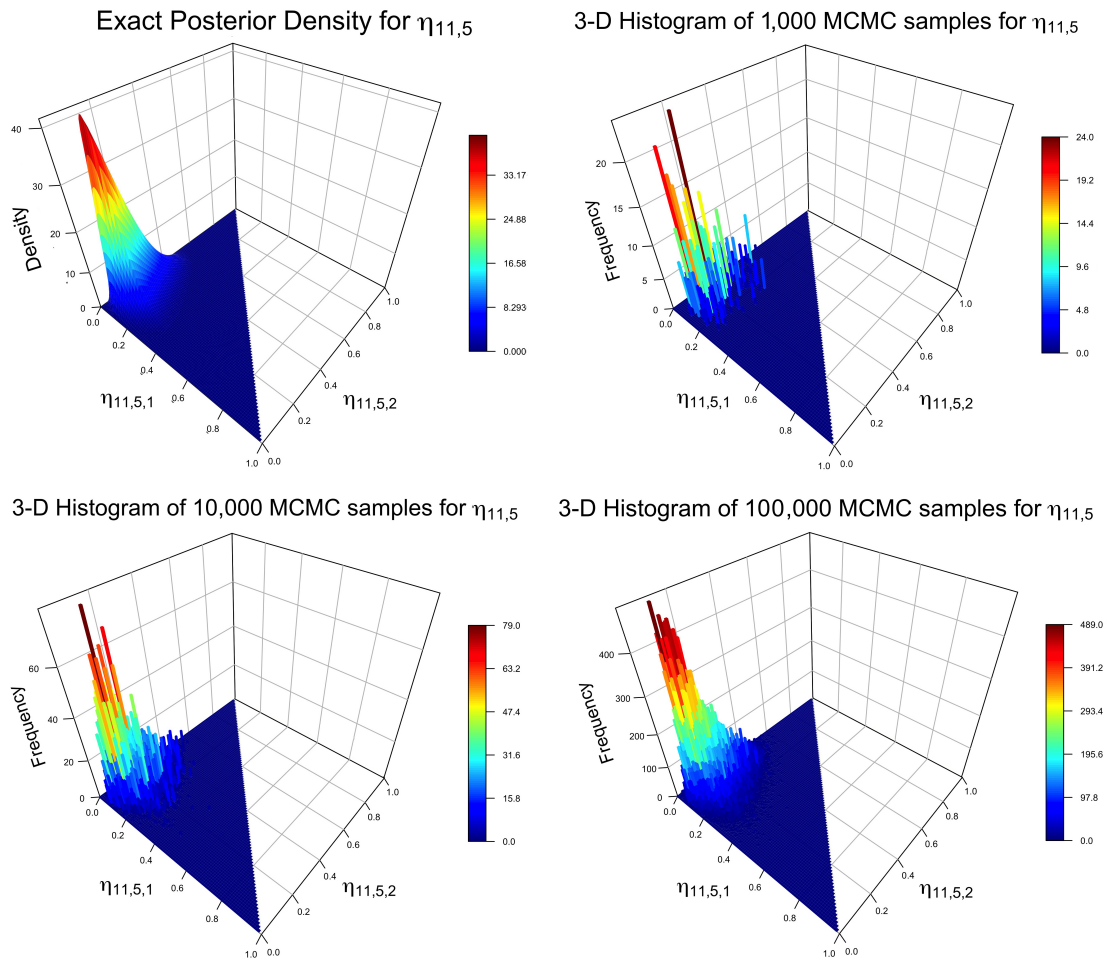


Figure 3.4: The exact posterior distribution for the form usage probabilities $\eta_{11,5}$ (for forms of item $i = 5$ in cell $x = 11$), shown in the top left, as compared to 3-D histograms for $\eta_{11,5}$ after 1000 (top right), 10000 (bottom left) and 100000 (bottom right) MCMC samples.

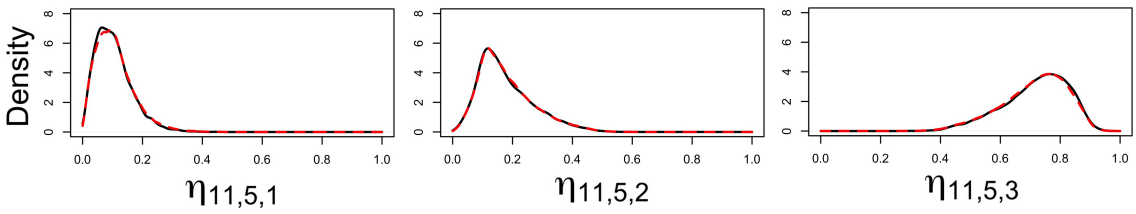


Figure 3.5: Marginal posterior distributions for form usage probabilities $\eta_{11,5,1}$ (left), $\eta_{11,5,2}$ (middle) and $\eta_{11,5,3}$ (right) based on 100000 MCMC samples. The exact marginal posterior distributions are overlaid in red.

3.4.1.2 Item Usage Rates

To test the sampler for μ from the MCMC algorithm for the ℓ_1-Dirichlet zero-inflated model, we use it to estimate the posterior distributions for μ_i for $i \in \{1, \ldots, 8\}$, given our simulated data y, and with η, ϕ and x_p fixed to their assigned values.

We compare the estimated posterior distributions to the exact posterior distributions for μ_i, which can be computed analytically in this case. To derive the analytic expression for the posterior distribution for μ_i, the normalising constant Z_{μ_i} for Equation (2.15) must be computed.

We find this using a quadrature approach, with the hyper-parameters of the prior specified by Equation (2.10) set to $a_i = 3$ and $b_i = 1$ for all items $i \in \{1, \ldots, I\}$.

The estimated posterior distributions for selected μ_i are shown in Figure 3.6. It can be seen that the estimated posteriors converge to the exact posteriors (overlaid in red). The posterior distributions for other μ_i display similar patterns.

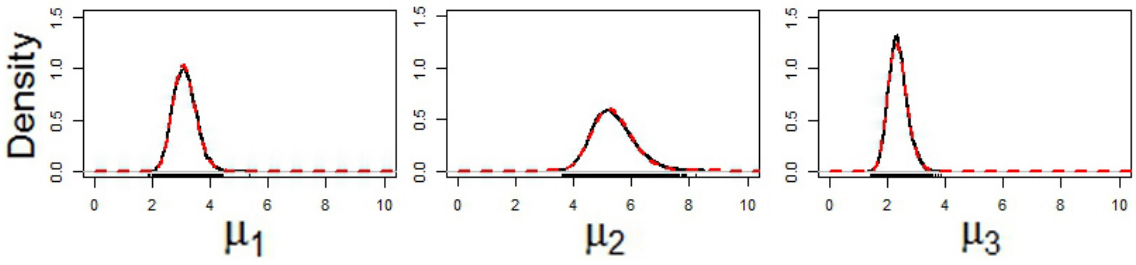


Figure 3.6: Posterior distributions for item usage rates μ_1, μ_2, and μ_3, based on 10000 MCMC samples. The exact posterior distribution is overlaid in red for each.

3.4.1.3 Zero-Inflation Probabilities

To test the sampler for ϕ from the MCMC algorithm for the ℓ_1-Dirichlet zero-inflated model, we use it to estimate the posterior distributions for ϕ_i for $i \in \{1, \ldots, 8\}$, given our simulated data y, and with η, μ and x_p fixed to their assigned values.

We compare the estimated posterior distributions to the exact posterior distributions for ϕ_i, which can be computed analytically in this case. To derive the analytic expression for the posterior distribution for ϕ_i, the normalising constant Z_{ϕ_i} for Equation (2.17) must be computed. Again, we find this using a quadrature approach.

The estimated posterior distributions for selected ϕ_i are shown in Figure 3.7. It can be seen that the estimated posteriors converge to the exact posterior (overlaid in red). The posterior distributions for other ϕ_i display similar patterns.

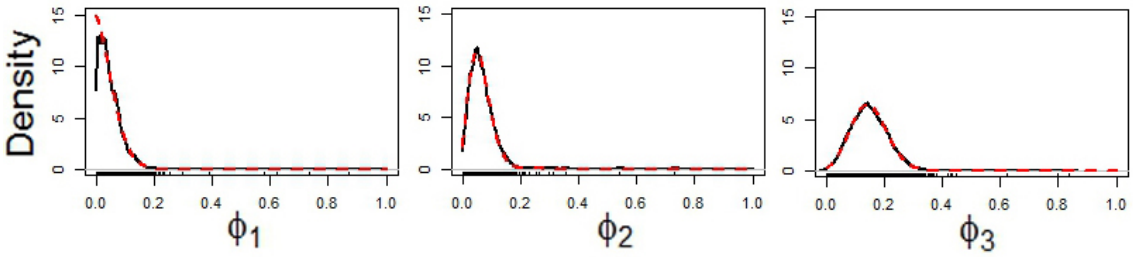


Figure 3.7: Posterior distributions for zero-inflation probabilities ϕ_1, ϕ_2, and ϕ_3, based on 10000 MCMC samples. The exact posterior distribution is overlaid in red for each.

3.4.1.4 Location

The location sampler of the MCMC algorithm for the ℓ_1-Dirichlet zero-inflated model already samples locations for the profiles directly from the posterior distribution π_x. As a double-check though, we calculate the posterior probabilities for profile p to locate in each cell x, with η, μ and ϕ fixed to their 'true' values, using both our sampler code and independent code. We find complete agreement between the two.

CHAPTER 4

MCMC Improvements

A limitation of the MCMC algorithms presented in Sections 3.1 and 3.2 is that the η-field parameters mix much slower than the other parameters of our model, due to serial correlation. For each item $i \in \{1, \ldots, I\}$, and each location $x \in \{1, \ldots, C\}$, the proposed $\eta'_{x,i}$ differs from $\eta_{x,i}$ in only two forms. If the number of forms F_i of item i is large, then it takes a long time before updates have been proposed for every form $f \in \{1, \ldots, F_i\}$ of the item (let alone accepted).

What exactly does a 'long time' translate to? If we have F_i forms and choose two forms at a time at random, then the expected number of samples $\mathbb{E}(T)$ before we have chosen them all can be derived by applying a result from combinatorics (Stadje, 1990). This application gives us that

$$\mathbb{E}(T) = \left(\sum_{f=1}^{F_i-2} \binom{F_i}{f} \left\{ \frac{\binom{F_i}{2}}{\binom{F_i}{2} - \binom{F_i-f}{2}} \right\} (-1)^{f-1} \right) + F_i(-1)^{F_i-2} + (-1)^{F_i-1}. \quad (4.1)$$

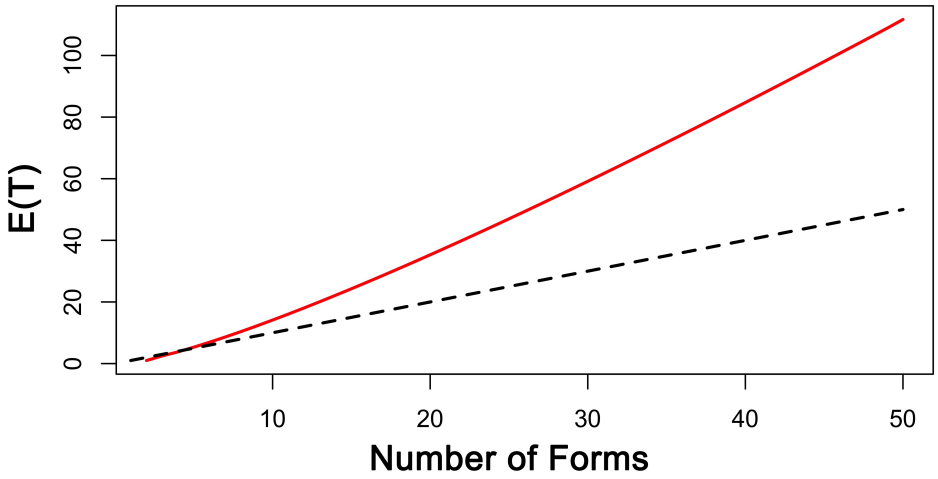


Figure 4.1: The expected number $\mathbb{E}(T)$ of MCMC algorithm iterations needed to propose updates for every form usage probability $\eta_{x,i,f}$ of item i, for forms $f \in \{1, \ldots, F_i\}$, is plotted in red against the number of forms F_i of the item. The dashed line shows the line $E(T) = F_i$ as a point of reference for the increase in $\mathbb{E}(T)$ with increase in F_i.

$\boldsymbol{F_i}$	5	10	20	30	40
$\mathbb{E}(\boldsymbol{T})$	5.3	14.1	35.3	59.2	84.7

Table 4.1: The expected number $\mathbb{E}(T)$ of iterations of the MCMC algorithm needed to propose updates for every form usage probability $\eta_{x,i,f}$ of item i, for forms $f \in \{1, \ldots, F_i\}$, as the total number of forms F_i varies.

Figure 4.1 shows $\mathbb{E}(T)$ with varying number of forms F_i, and a summary of some of these values is provided in Table 4.1. For the specified number of forms F_i, these $\mathbb{E}(T)$ tell us how many iterations of the MCMC algorithms from Sections 3.1 and 3.2 we would expect to require before we had *proposed* updates for all forms of an item (which has that number of forms). Obviously, many more iterations of the algorithm would be required before at least one update had been *accepted* for all the forms.

4.1 Parallel MCMC Updates for η

The typical run-time of the algorithms from Sections 3.1 and 3.2 with our linguistic data is in the order of weeks, depending on how much data are used. Although a solution to compensate for poor mixing within the η-fields would be to simply perform longer model runs, taking this option in this scenario is costly.

In this section, we consider another solution, proposing an alteration to our MCMC algorithms. This alteration consists of updating the η-fields N_η times for every sweep through the other parameters. By exploiting the conditional independence of the η-field parameters, these updates are easily performed in parallel, and are thus more computationally efficient than those from the previous algorithm.

Conditional on the η-fields, the locations of the profiles are independent, and we could also update these in parallel. However, calculating Equation (3.6) is not so time consuming, and they mix perfectly in one sample in the conditional distribution (conditioned on η) in question, since they are independently Gibbs sampled.

We therefore replace step 1 in the algorithms presented for the ℓ_1-Dirichlet models (in Sections 3.1 and 3.2) with the following. The only difference is the addition of (iii) - to repeat the step multiple (N_η) times.

1. **Altered update for form usage probability fields η.**

   For each item $i \in \{1, \ldots, I\}$, and each location $x \in \{1, \ldots, C\}$:

(i) Propose $\eta'_{x,i}$ as in the algorithm presented in Section 3.1.

(ii) Draw $u \sim U(0,1)$, and accept the proposed $\eta'_{x,i}$ if $u \leq \alpha_\eta(\eta_{x,i}, \eta_{x',i})$, with $\alpha_\eta(\eta_{x,i}, \eta'_{x,i})$ as defined in Equation (3.3) and using $\pi_\eta(\eta_{x,i}|y, \Theta \backslash \{\eta_{x,i}\})$ as derived from the relevant posterior distribution: Equation (2.13) for the ℓ_1-Dirichlet zero-inflated model, or Equation (2.22) for the ℓ_1-Dirichlet outlier model.

(iii) Repeat (i)-(ii) N_η times.

The $\eta_{x,i}$ vectors are conditionally independent, so we are able to update the η-fields for different items on different processor cores. The intent is to select N_η in an 'optimal' fashion, i.e. such that the η-fields returned after the N_η updates are a relatively independent sample from the target distribution, and we strike the right balance between the computational time spent on the cores and the reduction in the η-dependence that we gain.

We therefore must consider how many processor cores[1] (N_c) to use, as well as how many updates N_η to propose for each cell $x \in \{1, \ldots, C\}$ on each core. These considerations are made in turn in Sections 4.1.1 and 4.1.2 below.

4.1.1 How Many Cores?

Figure 4.2 depicts one iteration of our parallel η-sampler. We have slave processors S_n for $n \in \{1, \ldots, N_c\}$, where the number of cores is no greater than the number of

[1] In the case of CPUs with hyperthreading, by core, we refer to virtual cores.

items ($N_c \leq I$). We also have the master processor, M, which controls these slaves. Let $\eta_i = (\eta_{1,i}, \eta_{2,i}, \ldots, \eta_{C,i})$.

In our parallel MCMC algorithm, the master processor M passes each slave processor S_n one of the η_i, as well as other parameter values and data ($\mathbb{D}$) necessary for the calculations in the algorithm described above on page 81. After N_η updates are proposed on the slave for each $\eta_{x,i}$ vector for $x \in \{1, \ldots, C\}$, the new field η_i' is returned to the master M.

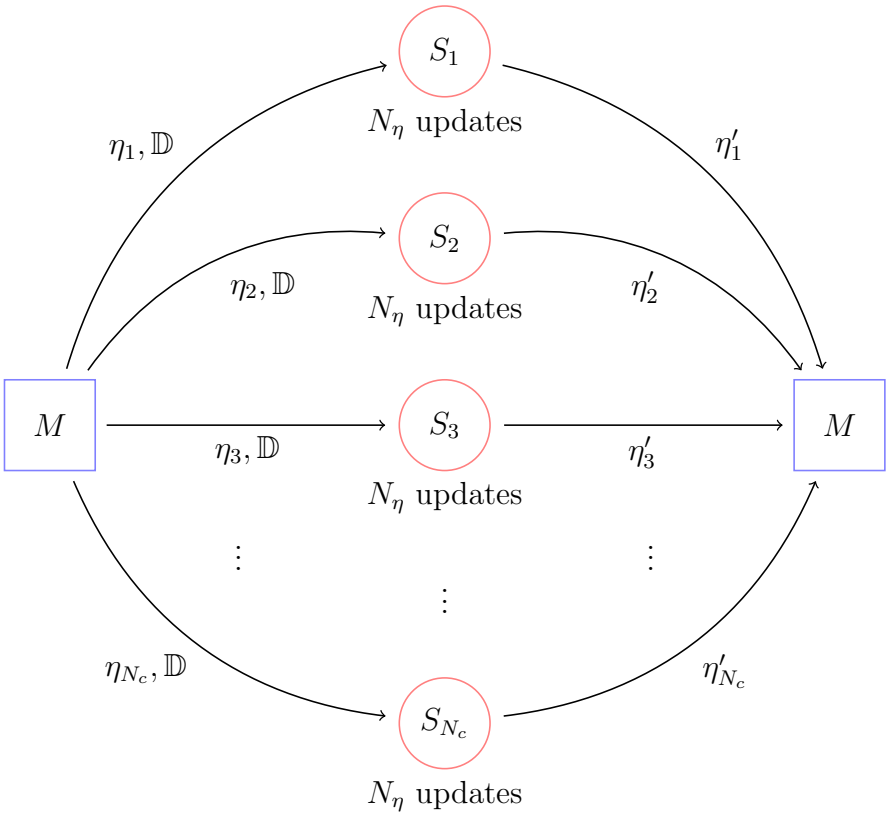


Figure 4.2: A schematic diagram of the parallel η-sampler. The master processor M passes each slave S_n all the form usage probability vectors $\eta_{x,i}$ for a given item i, plus all data $\mathbb{D}$ needed by the η sampler. After N_η updates have been proposed for each $\eta_{x,i}$ on the slave, the resulting vectors $\eta_{x,i}'$ are returned to the master.

If the number of slave processors $S_n = I$, this iteration of η-field sampling is completed when all the slaves return these new fields η_i' to the master. If $S_n < I$, the process depicted is repeated for the remaining items, with the η-fields for these items passed sequentially to the slaves as they complete their earlier tasks.

Intuitively, we should employ as many (slave) processor cores S_n as possible, up to a maximum of $S_n = I$ (i.e. equal to the number of items I). Figure 4.3 shows the results of exploratory analyses, confirming that this is indeed the case.

In these analyses, we measured the time taken to complete 500 iterations of the parallel MCMC algorithm for varying numbers of cores N_c, when $N_\eta \in \{1, 10, 20, 30\}$. In each case, we used the ℓ_1-Dirichlet zero-inflated model (presented in Section 2.1) and an eight-item subset of the coarsened linguistic data (described in Section 5.2.1).

Irrespective of the number of cores N_c used, the increase in time to complete 500 iterations of the parallel algorithm is linear with respect to the number of updates N_η proposed for each η vector at each iteration.

We note that performing the updates in parallel (coloured lines in Figure 4.3) is much faster than not (the black line), and that the runtime decreases as we increase N_c towards I.

The notable exception to both of these observations is when $N_\eta = 1$, i.e. we only proposed one update for each η vector per iteration of the algorithm. Here, the runtime increases with increasing N_c, reflecting the increased time spent communicat-

ing with the slave processors. With larger N_η, however, this communication time is negligible compared to the gains made by performing many η-updates simultaneously.

4.1.2 How Many Parallel η Updates?

The exploratory analyses presented above confirmed that using as many cores as possible, up to a maximum of $N_c = I$, leads to the smallest expenditure of computational time to run our model (assuming that $N_\eta > 1$). Here, we consider the choice of N_η with further exploratory analyses. As mentioned earlier in the chapter, we aim to choose N_η such that the η-fields returned after the N_η proposed updates are a relatively independent sample from the target distribution, but also such that we strike

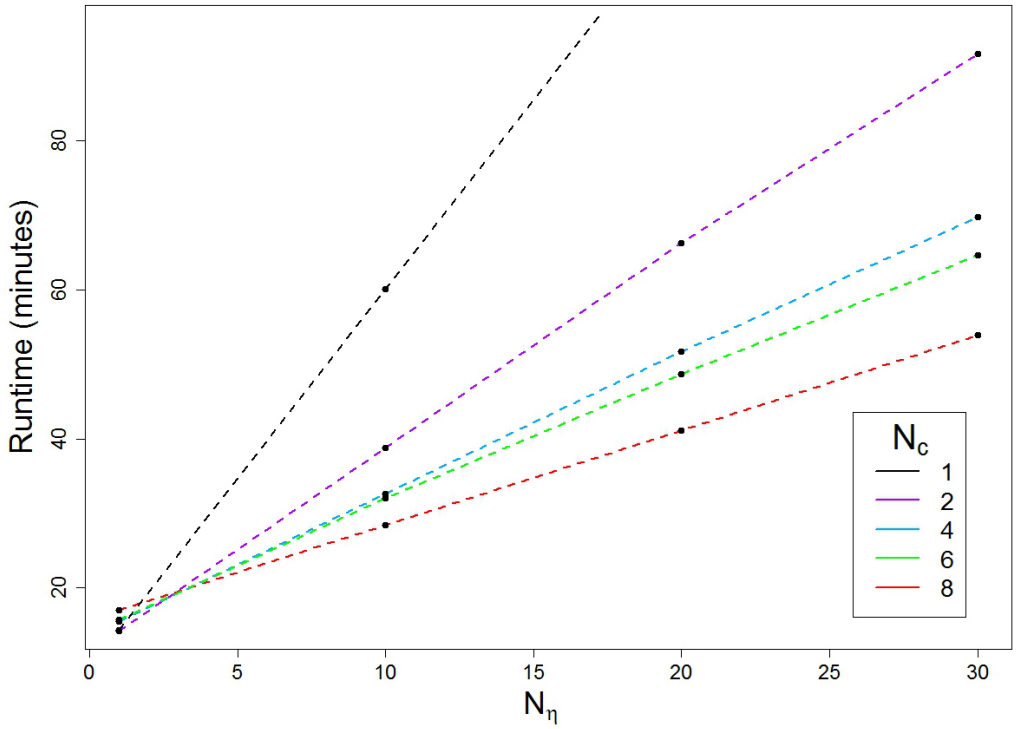


Figure 4.3: Time taken to complete 500 iterations of the parallel MCMC algorithm for the ℓ_1-Dirichlet model, with varying number N_η of parallel updates proposed for each η-vector per iteration of the algorithm. Observations were taken using varying numbers of cores N_c to fit the model.

the right balance between the computational time taken for extra updates and the reduction in the η-correlation obtained by performing them.

We fit the ℓ_1-Dirichlet zero-inflated model (presented in Section 2.1) to an eight-item subset of the coarsened linguistic data (described later in Section 5.2.1), using $N_c = 8$ cores. Our analyses suggest an optimal range for N_η of between 10-20 proposed updates for each η vector per iteration of the overall algorithm.

For $N_\eta \in \{1, 2, \ldots, 14, 15, 20, 30, 40, 50\}$, we measured the time taken to complete the first 500 iterations of the MCMC algorithm. We also computed the effective sample size (ESS) for every $\eta_{x,i,f}$ parameter over this period, and took the overall median of these effective sample sizes. This quantity, which we denote $\overline{\text{ESS}}(\eta)$, is a measure of the overall η-dependence in our samples.

The left panel of Figure 4.4 shows $\overline{\text{ESS}}(\eta)$ for varying N_η. We observe that we are able to substantially reduce the η-dependence by increasing N_η. The right panel then shows $\overline{\text{ESS}}(\eta)$ per minute for varying N_η. The optimal rate achieved around $10 \leq N_\eta \leq 15$ gives us approximately eight effective samples (on average) for each $\eta_{x,i,f}$ every 10 minutes.

Figure 4.5 shows the log-likelihood over the first 500 iterations of the MCMC algorithm (corresponding to part of the burn-in phase) for a subset of the N_η values used. We note that the log-likelihoods converge to a narrow equilibrium range of values if we run the algorithm for longer, and that this equilibrium is reached fastest

with large N_η. We conclude that large gains in convergence rate (as opposed to mixing rate) are made for N_η up to about 10-20, and based on this and our earlier observation, the optimal choice for N_η seems to be about 10-20.

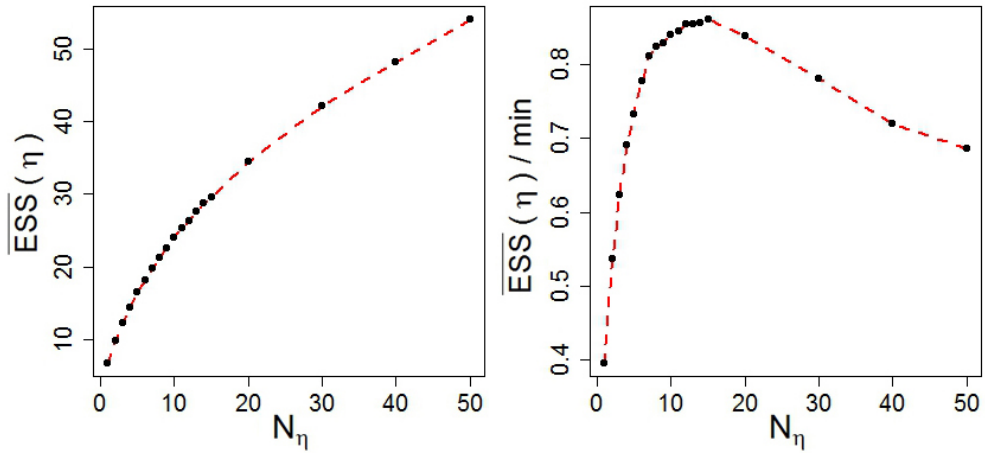


Figure 4.4: Left: Overall median effective sample size ($\overline{\text{ESS}}$) for the form usage probabilities η, against N_η (the number of parallel η-updates per iteration of the MCMC algorithm). Right: Overall median effective sample size for η per minute against N_η.

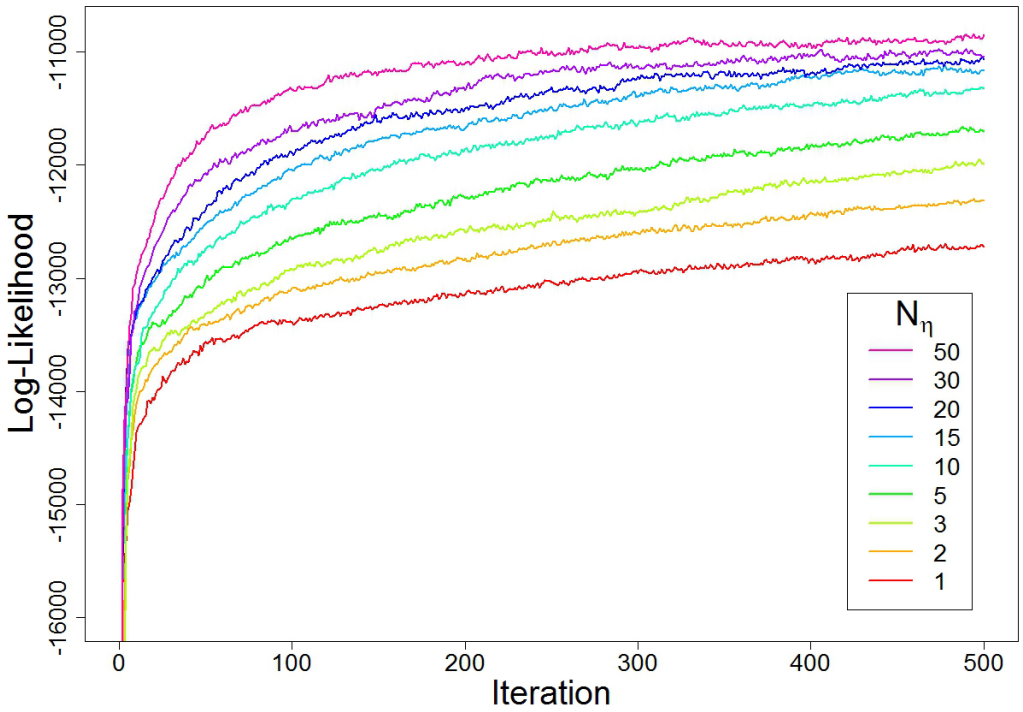


Figure 4.5: Log-likelihood for varying N_η (the number of parallel η updates per iteration of the MCMC algorithm) from the first 500 iterations of an MCMC run using the ℓ_1-Dirichlet zero-inflation model. All log-likelihood traces converge to a narrow equilibrium range of values if we run the algorithm for longer.

4.2 Metropolis Adjusted Langevin Algorithm

The problem of poor mixing for the dialect-field parameters is not restricted solely to the ℓ_1-Dirichlet zero-inflated and outlier models. We find similar issues with the γ-field parameters from the ℓ_2-logistic models when using the algorithms from Section 3.3. Particularly, the mixing of the γ-fields is quite sensitive to the choice of tuning parameters $(W_\gamma)_i$. If we propose new $\gamma_{x,i}$ vectors for the Metropolis-Hastings updates directly from a $\text{MVN}(\gamma_{x,i}, \Sigma_i)$ distribution, i.e. set $(W_\gamma)_i = 1$ for each item $i \in \{1, \ldots, I\}$, then the mixing is extremely poor.

One option to improve the mixing of the γ-field parameters revolves around careful selection of tuning parameters $(W_\gamma)_i < 1$ as well as proposing updates to these γ-fields in parallel, in a similar fashion to that outlined earlier in this chapter for the η-fields.

Another option which we explore in this section is an adjusted MCMC algorithm for use with the ℓ_2-logistic models. We adapt our earlier algorithm to utilise the Metropolis Adjusted Langevin Algorithm (MALA) (Roberts & Tweedie, 1996, Roberts & Rosenthal, 1998) to update the γ fields.

Under MALA, the Langevin diffusion is incorporated into the algorithm, which means that proposals for $\gamma_{x,i}$ are based upon the gradient of the log-posterior distribution, $\nabla \ln \pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})$. A derivation of this gradient for the ℓ_2-logistic outlier model is provided in Appendix C. The gradient for the ℓ_2-logistic zero-inflated

model can be obtained using this result by setting $\psi^{(P)} = 0$ and $\psi^{(I)} = 0$ for all $p \in \{1, \ldots, P\}$ and $i \in \{1, \ldots, I\}$, since it is nested within the ℓ_2-logistic outlier model.

We do not consider MALA for the ℓ_1-Dirichlet models given difficulty in obtaining the gradient function due to the model constraints (namely, the positivity and summation constraints on the η-parameters). The ℓ_2-logistic models do not feature such constraints, and this extra tractability is one of their attractions, if in other respects they are adequate models.

The altered MALA-based algorithm for the ℓ_2-logistic models only differs from that presented in Section 3.3 in the γ-update step, which we replace with the following:

1. **Altered update for form usage parameter fields γ.**

   For each item $i \in \{1, \ldots, I\}$, and each location $x \in \{1, \ldots, C\}$:

   (i) Compute $\nabla \ln \pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})$. This gradient is specified by Equation (C.28) in Appendix C for the ℓ_2-logistic outlier model. The gradient for the ℓ_2-logistic zero-inflated model can be obtained from Equation (C.28) by setting the outlier probability parameters ψ to zero.

   (ii) Set

   $$\gamma'_{x,i} = \gamma_{x,i} + \frac{W_\gamma^2}{2} \nabla \ln \pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\}) + \epsilon_i, \tag{4.2}$$

   where $\epsilon_i = (\epsilon_{i,1}, \ldots, \epsilon_{i,F_i})$, and $\epsilon_{i,f} \sim N(0, W_\gamma^2)$. We find from trial runs that $W_\gamma = 0.2$ is sensible.

(iii) Draw $u \sim U(0,1)$, and accept the proposed $\gamma'_{x,i}$ if $u \leq \alpha_\gamma(\gamma_{x,i}, \gamma_{x',i})$, where

$$\alpha_\gamma(\gamma_{x,i}, \gamma'_{x,i}) = \min\left(1, \frac{\pi_\gamma(\gamma'_{x,i} \mid y, \Theta \backslash \{\gamma'_{x,i}\})}{\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})} \frac{q(\gamma_{x,i} \mid \gamma'_{x,i})}{q(\gamma'_{x,i} \mid \gamma_{x,i})}\right), \tag{4.3}$$

with

$$\begin{aligned} q(\gamma'_{x,i} \mid \gamma_{x,i}) &= \prod_{f=2}^{F_i} \frac{1}{W_\gamma \sqrt{2\pi}} \exp\left(-\frac{\epsilon_{i,f}^2}{2W_\gamma^2}\right), \\ q(\gamma_{x,i} \mid \gamma'_{x,i}) &= \prod_{f=2}^{F_i} \frac{1}{W_\gamma \sqrt{2\pi}} \exp\left(-\frac{(\epsilon'_{i,f})^2}{2W_\gamma^2}\right), \end{aligned} \tag{4.4}$$

and

$$\epsilon'_{i,f} = \gamma_{x,i} - \left(\gamma'_{x,i} + \frac{W_\gamma^2}{2} \nabla \ln \pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})\Big|_{\gamma_{x,i}=\gamma'_{x,i}}\right). \tag{4.5}$$

Given that a MALA-based update for the γ-fields is computationally more costly than our old method, we performed exploratory analyses to assess whether the reduction in γ-dependence obtained offsets this additional computational time. Further, the MALA updates for the γ-fields for each item can be performed in parallel, so we also considered in these analyses whether the reduction in dependence in the γ-fields gained by multiple MALA updates per iteration of the algorithm was worth the additional computational time.

We find that despite the additional computational time required, our MALA-based updates offer a large improvement on our old methods. Further, we find that performing these updates in parallel and choosing N_γ (the number of proposed updates to each $\gamma_{x,i}$ vector per iteration of the algorithm) of around 3-5 offers the optimal rate of reduction in γ-dependence.

For our analyses, we used the ℓ_2-logistic zero-inflated model (presented in Section 2.3) and a eight-item subset of the coarsened linguistic data (described later in Section 5.2.1). Results are presented for 500 iterations of both our MALA-based and original MCMC algorithms for this model, with $N_c = 8$.

For $N_\eta \in \{1, 2, \ldots, 9, 10, 15, 20\}$, we once again measured the time taken to complete 500 iterations of the MCMC algorithm, as well as $\overline{\text{ESS}}(\eta)$ (the median effective sample size for the η parameters), which is a measure of the overall η-dependence in our samples (and, thus, the γ-dependence).

Figure 4.6 shows the additional computational time required to perform γ-updates using the MALA-based algorithm (blue line) instead of the old algorithm (red line) increases with N_γ.

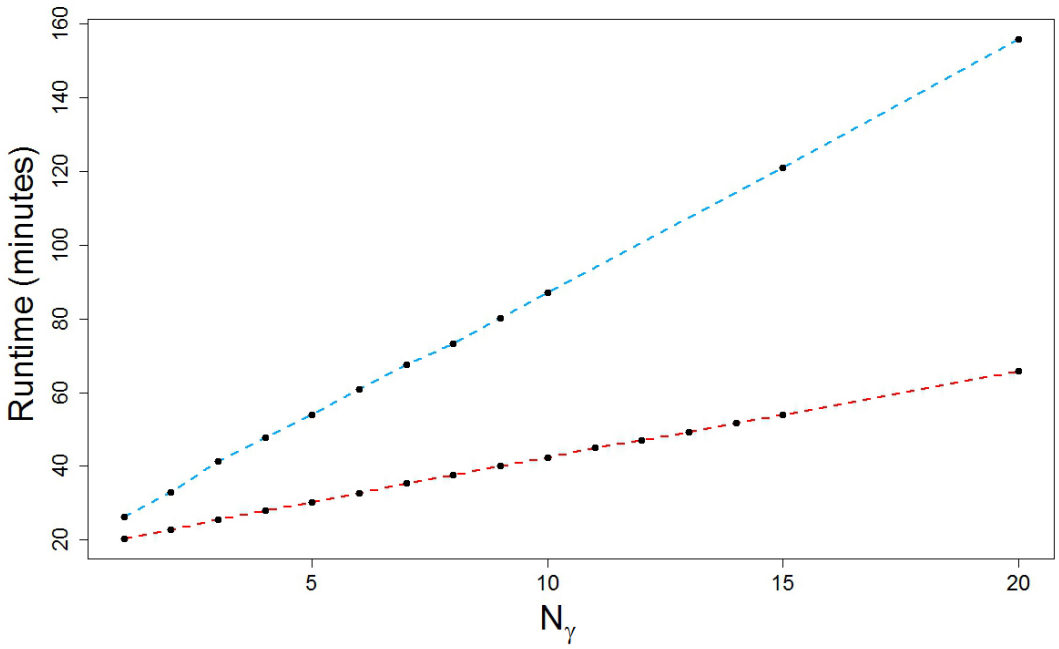


Figure 4.6: Time taken to complete 500 iterations of the MALA (blue line) and non-MALA (red line) MCMC algorithms for the ℓ_2-logistic zero-inflated model, against the number N_γ of parallel γ updates per iteration of the algorithm.

The left panel of Figure 4.7 shows that the MALA-based updates (in blue) do lead to substantially lower γ-dependence, and thus higher $\overline{\text{ESS}}(\eta)$. Furthermore, as displayed in the right panel of Figure 4.7, this reduction in dependence outweighs the additional computational time spent to achieve it. The optimal rate achieved around $3 \leq N_\gamma \leq 5$ gives us approximately one effective sample (on average) for each $\eta_{x,i,f}$ every minute.

Finally, Figure 4.8 displays the log-likelihood for a subset of N_γ values, for both the MALA-based and original γ-field samplers. The log-likelihoods from MALA-based sampling converge together within 500 iterations for all values of N_γ, and extremely quickly when $N_\gamma \geq 5$. With the original algorithm for the ℓ_2-logistic zero-inflated model, the log-likelihood only reaches this equilibrium range within this window when N_γ is large, though the log-likelihoods do all converge to a narrow equilibrium range when run for longer.

4.2.1 Checking the MALA-Based Sampler

In Appendix A.3, we perform similar checks to those provided in Section 3.4, but for the MALA-based γ-field sampler described above for the ℓ_2-logistic outlier model. These checks are compared to similar checks provided in Appendix A.2 for the original γ-field sampler from the algorithm in Section 3.3. We find no cause for concern.

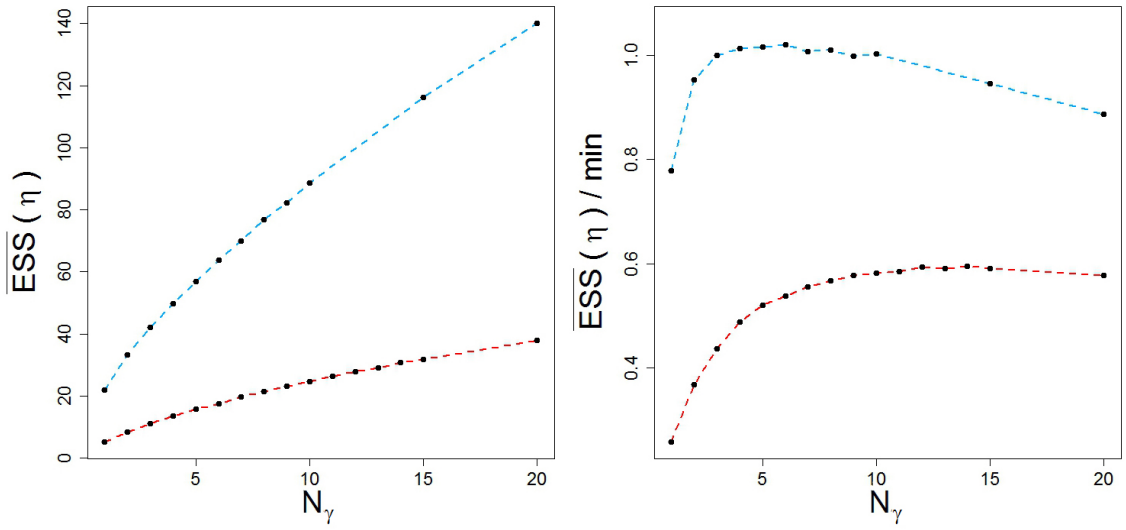


Figure 4.7: Left: Overall median effective sample size ($\overline{\text{ESS}}$) for the form usage probabilities η, against N_γ (the number of parallel γ updates per iteration of the MCMC algorithm). Right: Overall median effective sample size per minute against N_γ. MALA-based updates are shown in blue, and non-MALA updates in red.

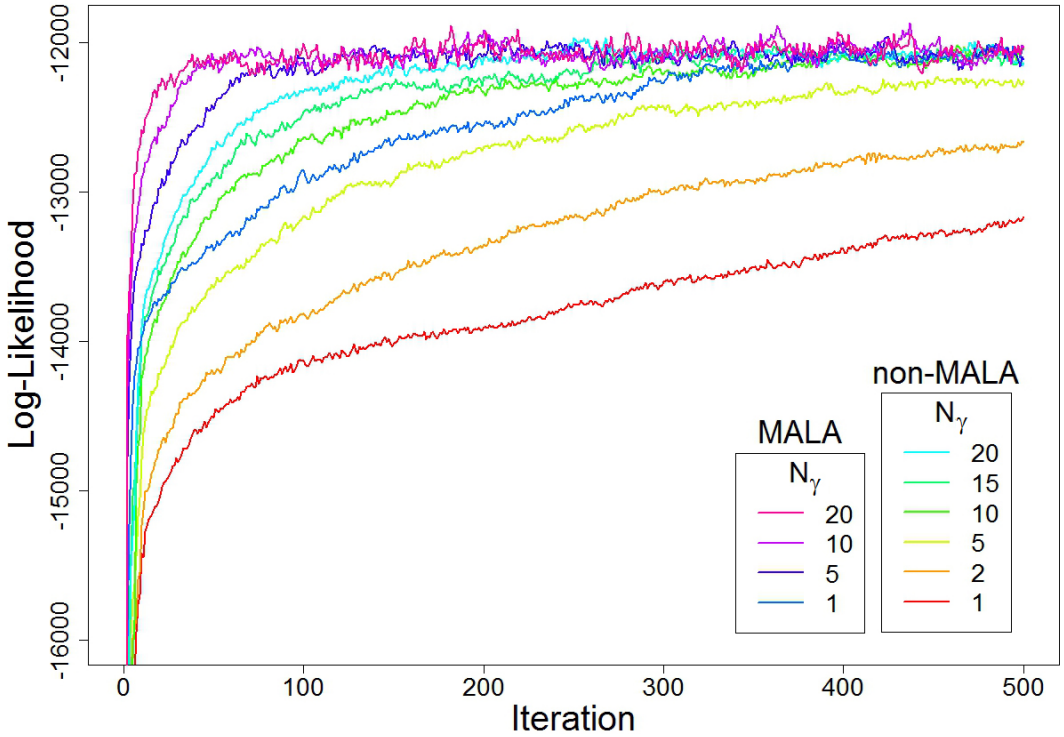


Figure 4.8: Log-likelihood for varying N_γ (the number of parallel γ updates per iteration of the MCMC algorithm) from the first 500 iterations of an MCMC run using the ℓ_2-logistic outlier model. All log-likelihood traces converge to a narrow equilibrium range of values if we run the algorithm for longer.

CHAPTER 5

Locating Linguistic Profiles

In this chapter, we address a range of pertinent questions. Are the linguistic data informative of location? To what extent does data coarsening cost us information? Can we fit our models to manageable sized datasets? How do these models perform when applied to the two applied problems outlined in Chapter 1? And finally, does our model predict locations of origin for the manuscripts that agree with, or conflict with, the Atlas' predicted locations via the 'fit-technique'?

To consider the questions above, we present a series of analyses. As described in Chapters 1 and 2, there are two ways we can frame the applied problem, depending on what we fix as anchor texts to inform the "dialect fields". We have a choice of whether we locate all of the non-anchor texts, or instead we trust the provided Atlas locations for the non-anchor texts and just locate new texts. We present analyses of both kinds.

As the latter framework presents the less difficult modelling challenge, we begin with analyses of this variety. We compare the models specified in Chapter 2, each

fit to a representative subset of the data described in Sections 1.3.1 and 1.3.2. Since these analyses are intended to explore the predictive capabilities of the models, a subset of the data is sufficient. We find that our ℓ_1-Dirichlet zero-inflated model is able to successfully locate the Atlas profiles. Even though the vast number of form-usage probabilities are only measured through the impact of sparse Bernoulli fields, with the summation constraint, information on one form informs others as well.

We further conclude that the ℓ_1-Dirichlet zero-inflated model performs just as well as the more complex ℓ_1-Dirichlet outlier model on the problems tested; similarly, that the models with ℓ_1 smoothing prior distributions perform better than those with ℓ_2 smoothing; and finally, that little information is lost by the data coarsening operation.

We then present an analysis under the other framework, whereby we locate all of the non-anchor texts. As this is a more difficult task, we extend the scope of analysis to use the full coarsened linguistic data. Though an advantage was not observed when using the ℓ_1-Dirichlet outlier model to model the smaller data subsets, given the larger and noisier dataset we scale to, we would expect this model to be useful for handling the big missing data (i.e. the massive parameterisation - see Equation 5.1 and Table 5.2). However, given the additional computational burden of using this model (as well as the much larger linguistic dataset), we proceed with the ℓ_1-Dirichlet zero-inflated model.

The results we obtain under this framework are of a poorer quality, with the non-anchor texts pulled towards the largest cluster of anchors. Equivalent findings are obtained even if we vary the level of smoothing, if we switch to using a different model or if we use the primary (uncoarsened) data. We then present results explaining how this issue is related to the spatial distribution of the anchors.

Throughout all these analyses, we jointly estimate the unknown locations of origin for linguistic profiles. Unlike previous research, such as Wasser et al. (2007), this 'joint' estimation is joint in the sense that we are jointly estimating these locations and the underlying probability fields η (or γ), so that the profiles with unknown origins feed information back into the estimation of fields. The last analyses presented in this chapter consider the impact this has, to explain how the results obtained when locating all non-anchor profiles can be impacted by this feedback.

An overview of these analyses is provided in Table 5.1, alongside further details of the scope of these analyses in Table 5.2. As is illustrated by Table 5.2, the applied problems considered pose a very sizeable statistical challenge, with many parameters to estimate, and data issues of sparsity and zero-inflation (outlined in Chapter 1). For the zero-inflated models, the number of parameters to estimate is

$$|\Theta| = \left(C \times I \times \sum_{i=1}^{I} F_i \right) + 2I + |x_p|, \tag{5.1}$$

where $|x_p|$ is the number of floating profiles, C is the number of cells in the lattice, I is the number of items, and F_i is the number of forms of item i. An additional $I + P$

parameters are estimated with the outlier models, due to the addition of the outlier probabilities ψ.

For each of the analyses presented in this chapter, three separate Markov chains were generated using the relevant algorithm from Chapter 3 or Chapter 4. Each chain was initialised at different starting values $\Theta^{(0)}$ for the model parameters Θ. We check that these chains give consistent results.

In Chapter 4, we proposed methods to compensate for the otherwise high autocorrelation between successive states of each of the dialect-field parameters. As a result,

Section	**Data**	**Model**	**Framework**	**Purpose**
5.2.2	Primary & coarsened	ℓ_1-D ZI	New	- Data informative of location? - ℓ_1-D ZI model any good? - Impact of coarsening?
5.2.3	Primary & coarsened	ℓ_1-D O	New	- Data informative of location? - ℓ_1-D outlier model better than ZI? - Impact of coarsening?
5.2.4	Coarsened	ℓ_2-L ZI	New	- How does ℓ_2-based model compare to earlier ℓ_1-based models?
5.3	Coarsened	ℓ_1-D ZI	All	- Consider problem fit-technique was applied to.
5.3.1	Coarsened	ℓ_1-D ZI	All	- Impact of anchor spread?
5.3.2	Coarsened	ℓ_1-D ZI	All	- Impact of estimating locations and fields separately?

Table 5.1: For each analysis presented in this chapter, we list which type of data and which model was used, as well as the framework of the applied problem we worked within (locating *new* profiles, or locating *all* profiles whose locations were determined using the fit-technique). We also briefly describe the questions the analysis sought to answer.

Section	Data	Items	Forms	Anchors	Floating	Parameters (Θ)	$\lvert\Theta\rvert$
5.2.2	P	8	873	244	123	η, μ, ϕ, x_p	314,419
	C	8	128	244	123	η, μ, ϕ, x_p	46,219
5.2.3	P	8	873	244	123	$\eta, \mu, \phi, \psi^{(I)}, \psi^{(P)}, x_p$	314,794
	C	8	128	244	123	$\eta, \mu, \phi, \psi^{(I)}, \psi^{(P)}, x_p$	46,594
5.2.4	C	8	128	244	123	γ, μ, ϕ, x_p	43,339
5.3	C	71	741	120	247	η, μ, ϕ, x_p	425,909
5.3.1	C	71	741	120	247	η, μ, ϕ, x_p	425,909
5.3.2	C	60	298	120	0	η, μ, ϕ	107,296

Table 5.2: For each analysis presented in this chapter, we list which data was used (primary, or coarsened), alongside the number of items used, and the corresponding total number of forms of these items. We also specify the number of profiles fixed in place as anchors, and the number of profiles whose locations we estimate. Finally, we list the model parameters to be estimated, as well as the total number of these.

when using the ℓ_1-Dirichlet zero-inflated or outlier models, we proposed MCMC updates for the η-fields in parallel, with 10 updates proposed for each $\eta_{x,i}$ vector for every one update proposed for each of the remaining parameters. When using the ℓ_2-logistic models, we used our MALA-based algorithm, and proposed five updates for each $\gamma_{x,i}$ vector for every one update proposed for each of the remaining parameters.

We performed $M = 50000$ iterations of our algorithm for each of the three chains in each of our analyses. We discarded the initial $B = 1000$ iterations of each chain as a burn-in period. This burn-in may appear relatively short, but given the above choices, it corresponds to 10000 updates proposed for each of the η parameters (and 5000 for the γ parameters, when using ℓ_2-logistic models). We found this burn-in sufficiently

long, with little change in posterior means and variances for the parameters when dropping larger initial segments.

With primary data, these choices gave typical effective sample sizes for each chain of 350 for the form usage probabilities η, 200 for the item usage rate parameters μ, 150 for the zero-inflation probabilities ϕ, 600 for ψ_i, and 1000 for ψ_p. Typical posterior variances were 0.00007 for η, 0.2 for μ, 0.007 for ϕ, 0.004 for ψ_i, and 0.04 for ψ_p.

With coarsened data, these choices gave typical effective sample sizes for each chain of 1500 for η parameters, 1500 for μ, 1000 for ϕ, 3000 for ψ_i, and 1000 for ψ_p. Typical posterior variances were 0.0003 for η, 0.01 for μ, 0.001 for ϕ, 0.005 for ψ_i, and 0.04 for ψ_p.

To assess if there was evidence for a lack of convergence to the target distributions of interest, diagnostics were calculated using the methods of Gelman & Rubin (1992), Heidelberger & Welch (1983) and Geweke (1992). Traceplots and autocorrelation plots for the parameters were also examined. We did not find any cause for concern.

For a summary of these convergence diagnostics and the `coda` package in R we used to implement them, see Robert & Casella (2010). For an example of convergence diagnostic results for some randomly selected parameters, refer to Appendix B.

5.1 Modelling Region

Throughout this chapter, we restrict attention to a large rectangular sub-region of the country, which has good coverage of profiles, a relatively high proportion of which are anchors. This region, which we label $\mathbb{W}$, is displayed in Figure 5.1.

367 linguistic profiles are believed to lie within $\mathbb{W}$, based on their Atlas location estimates λ_p. This equates to approximately one-quarter of the profiles from the Atlas. 120 of these profiles are derived from manuscripts with known origins. The locations of origin of the other 247 profiles were estimated using the fit technique. These two sets of profiles are displayed in red and blue respectively in the lower half of Figure 5.1.

Let $A_{\mathbb{W}}$ and $\bar{A}_{\mathbb{W}}$ denote our sets of anchor and floating (non-anchor) profiles within $\mathbb{W}$ respectively. These sets are determined by which of the two frameworks described earlier we work within.

If we seek to locate all of the non-anchor texts, then $|A_{\mathbb{W}}| = 120$ and $|\bar{A}_{\mathbb{W}}| = 247$. If instead we seek to locate a set of new profiles, then profiles whose location were estimated using the 'fit-technique' will be fixed in place to their Atlas locations; and thus treated as anchors in the same way as profiles derived from manuscripts with known origins. If $N_{(F)}$ of these profiles are treated as anchors, then $|A_{\mathbb{W}}| = 120 + N_{(F)}$ and $|\bar{A}_{\mathbb{W}}| = 247 - N_{(F)}$.

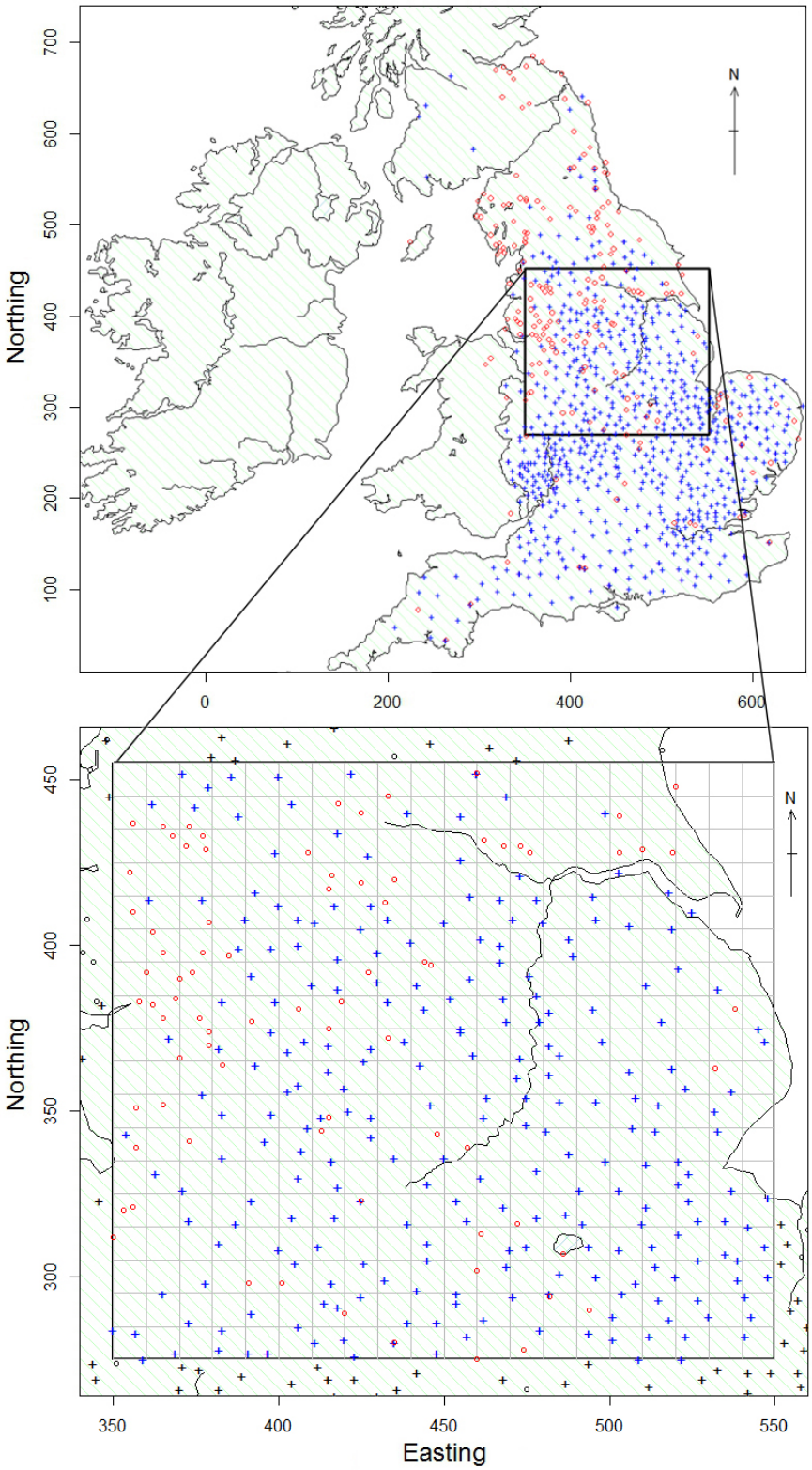


Figure 5.1: The location of the sub-region $\mathbb{W}$ of England used throughout the chapter (top plot). $\mathbb{W}$ is divided into a rectangular lattice of $C = 360$ square cells, each 100 km^2 (bottom plot). The Atlas locations for anchor profiles (blue crosses) and floating profiles (red circles) are shown in both plots.

5.2 Locating a Subset of Non-Anchor Profiles

To assess the performance of our models and Bayesian inference procedure for locating the Atlas manuscripts, we present here exploratory analyses using models detailed in Chapter 2: the ℓ_1-Dirichlet zero-inflated model from Section 2.1; the ℓ_1-Dirichlet outlier model from Section 2.2; and the ℓ_2-logistic zero-inflated model from Section 2.3.

We work under the framework of locating new profiles. As such, we randomly select a set of profiles Q whose locations were derived from the fit-technique, remove their cell-locations $J(\lambda_q)$ from the data and then estimate them. All other profiles are fixed to their Atlas locations. Thus, some profiles whose locations were derived from the fit-technique are fixed in place as if they were anchors.

This process leads to a Bayesian validation method of predictive testing of held-out data, with profiles $q \in Q$ treated as the 'new' profiles. Through this methodology, we can compare derived quantifications of uncertainty in the location estimates with those provided in the Atlas.

After we estimate the location x_q of each profile q within the lattice, we check to see if it coincides with the cell $J(\lambda_q)$ containing the Atlas location estimate. We do this by visual inspection of plots of marginal posterior probabilities for location (considering whether $\pi_x(x_q = J(\lambda_q) \,|\, y, \Theta \backslash \{x_q\})$ is relatively large), as well as assessing whether the computed Bayes factor supports the hypothesis that the profile belongs in (λ_q).

We estimate the unknown locations of origin for 123 linguistic profiles $q \in Q$, which equates to half of the profiles in $\mathbb{W}$ whose locations were derived from the fit-technique. We then compare the results obtained using the various models.

5.2.1 Item Subset

Since these exploratory analyses are intended only to explore the predictive capabilities of the models, a subset of the data is sufficient. We therefore restrict attention to eight of the available items, displayed in Table 5.3.

These items were randomly chosen from those for which both primary and coarsened data were available. We label the eight-item data subset arising for the primary data $\mathbb{D}$, and the subset for the coarsened data $\mathbb{S}$.

Though the items were chosen randomly, we hope that the forms of these items display contrasting spatial coverage across $\mathbb{W}$, as we believe this characteristic to aid in estimating the location of origin of profiles. For data subset $\mathbb{D}$, this behaviour is illustrated in Figure 5.2 by some forms of the item 'vps13' (a collection of third person singular verbs in the present tense).

Profiles located in the far south are unlikely to contain the form '-ES', profiles in the north being unlikely to contain the forms '-ETH' or '-TH', and profiles from the west being unlikely to contain the form '-US'. A similar plot is obtained for the corresponding coarsened forms of the item.

Item Label	which	against	each	much	there	vps13	then	vpp
$\boldsymbol{F_i}$	251	165	160	87	65	65	48	32
$\boldsymbol{F_i'}$	24	34	19	14	6	17	4	10

Table 5.3: The eight items used for exploratory analyses. The item labels describe the function/meaning of the item's collection of forms, are presented alongside the number of primary and coarsened data forms (F_i and F_i' respectively) for each item.

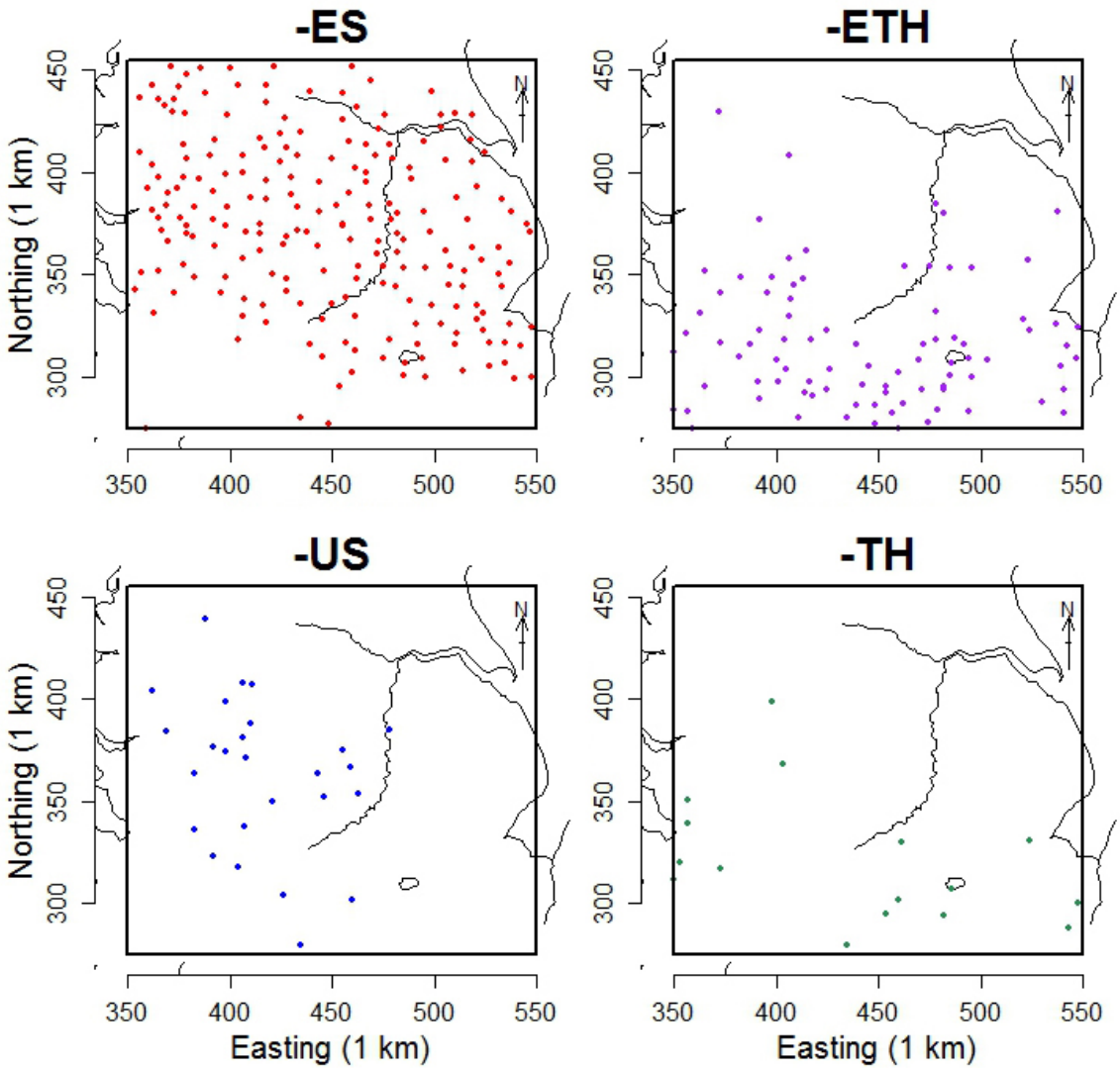


Figure 5.2: Dialect maps for four of the forms of the item 'vps13', which is a collection of third person singular verbs in the present tense. Each map shows the scatter for one form, with the dots corresponding to the Atlas-provided locations of profiles containing the mapped form.

5.2.2 ℓ_1-Dirichlet Zero-Inflated Model

We first test the placement of the 123 profiles q using the ℓ_1-Dirichlet zero-inflated model outlined in Section 2.1. This model has parameters $\Theta = (\mu, \phi, \eta, x)$. We fit the model to the eight-item data subsets $\mathbb{S}$ and $\mathbb{D}$ described above.

Estimated posterior distributions for μ_i and ϕ_i are displayed in Figure 5.3 for each data subset for four of the eight items. The separate MCMC chains gave closely overlaying density plots, with most items apparently appearing between one and four times in the manuscripts. The linguists found these estimates quite reasonable. The estimates for ϕ_i from the coarsened data were smaller than those from the primary data, as would be expected, given the reduction in the prevalence of zeroes in the data following the coarsening operation.

With each data subset, we estimated the mean posterior usage probability in each of the cells x of the lattice for form f of item i using

$$\hat{\pi}_\eta(\eta_{x,i,f} \mid y) = \frac{\sum_{t=B+1}^{M} \eta_{x,i,f}^{(t)}}{M - B}. \tag{5.2}$$

For data subset $\mathbb{D}$, Figure 5.4 displays these values $\hat{\pi}_\eta$ for the four forms f of item 'vps13' plotted earlier in Figure 5.2. Cells with higher $\hat{\pi}_\eta$ are represented by darker shades, and the Atlas locations of profiles which display the given form are overlaid. Encouragingly, the interpolated probability to choose a given form is highest where the Atlas believes it is commonly seen.

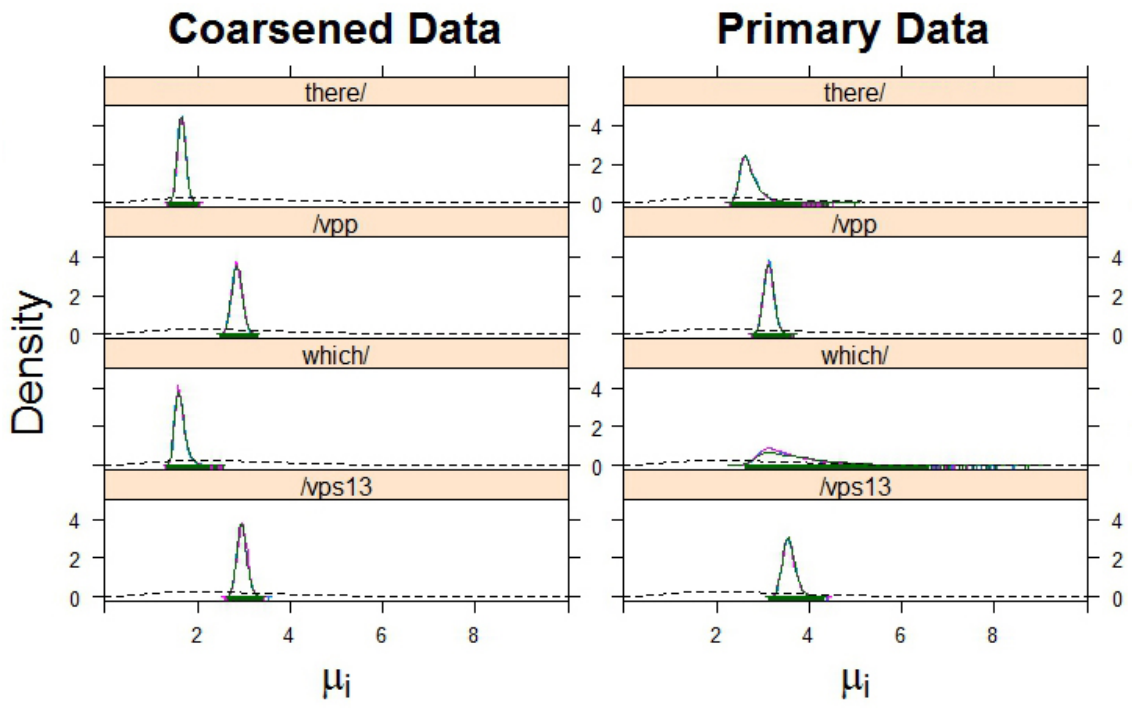
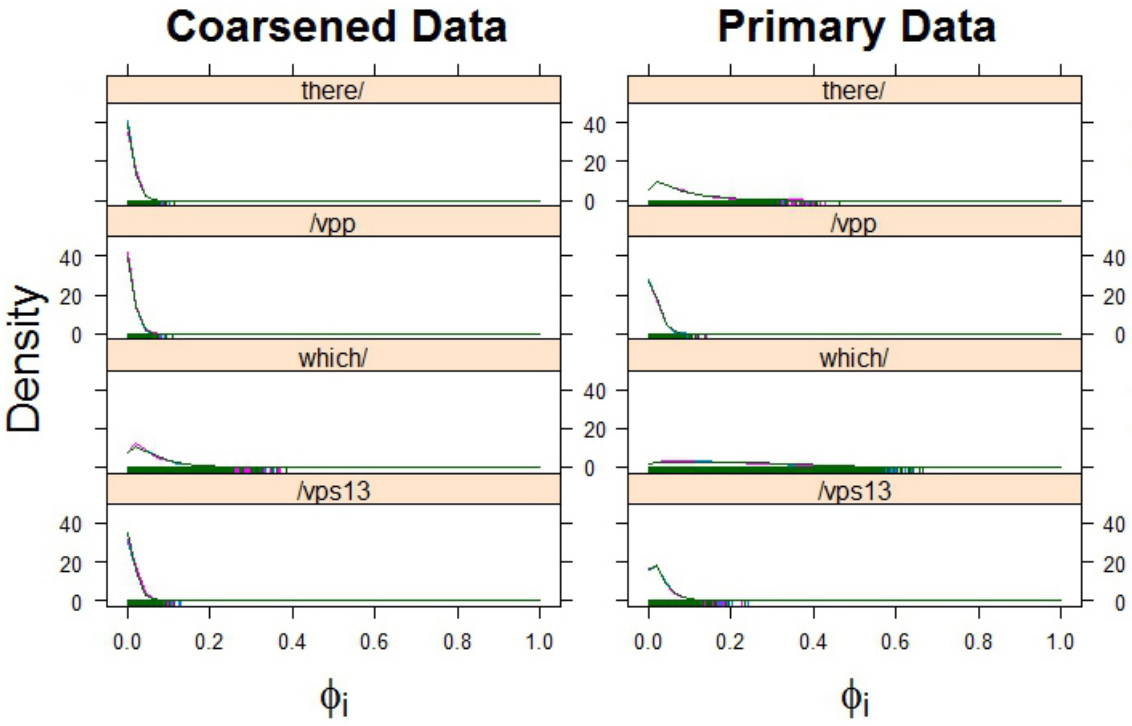


Figure 5.3: Estimated posterior distributions for mean item usage rates μ_i (above) and zero inflation probabilities ϕ_i (below) from the ℓ_1-Dirichlet zero-inflated model. Results are shown for eight-item subsets of the coarsened (left) and primary data (right). Three density plots are shown for each item i, with one for each parallel MCMC chain. The Gamma prior distributions for μ_i are overlaid with a dotted black line (ϕ_i had uniform priors).



The reader will note an edge-effect on the south border of the modelling region $\mathbb{W}$ for forms '-US' and '-TH', caused by inhomogeneities in smoothing due to the free boundary condition. We discuss this further in Chapter 7.

We can see that the primary forms are providing a useful feature-based segmentation of the region (i.e. they are distinct). Of course, these are just four primary forms of one of the eight items; however, these are representative of the patterns seen across other forms and items.

For data subset $\mathbb{S}$, Figure 5.5 then displays the values of $\hat{\pi}_\eta$ for the coarsened forms '-ES', '-ED', '-US' and '-D' of the same item. These are the corresponding coarsened forms for the primary data forms displayed in Figure 5.4. The interpolated form-usage probability is again generally highest around the Atlas locations. These regions of high posterior probability are generally larger than those found with $\mathbb{D}$.

With each data subset, we estimate the expected marginal posterior probability $\hat{\pi}_x(x_q = x \,|\, y)$ for each profile $q \in Q$ to originate from each of the lattice's cells $x \in \{1, \ldots, C\}$, calculating

$$\hat{\pi}_x(x_q = x \,|\, y) = \frac{\sum_t \mathbb{I}(x_q^{(t)} = x)}{M}. \tag{5.3}$$

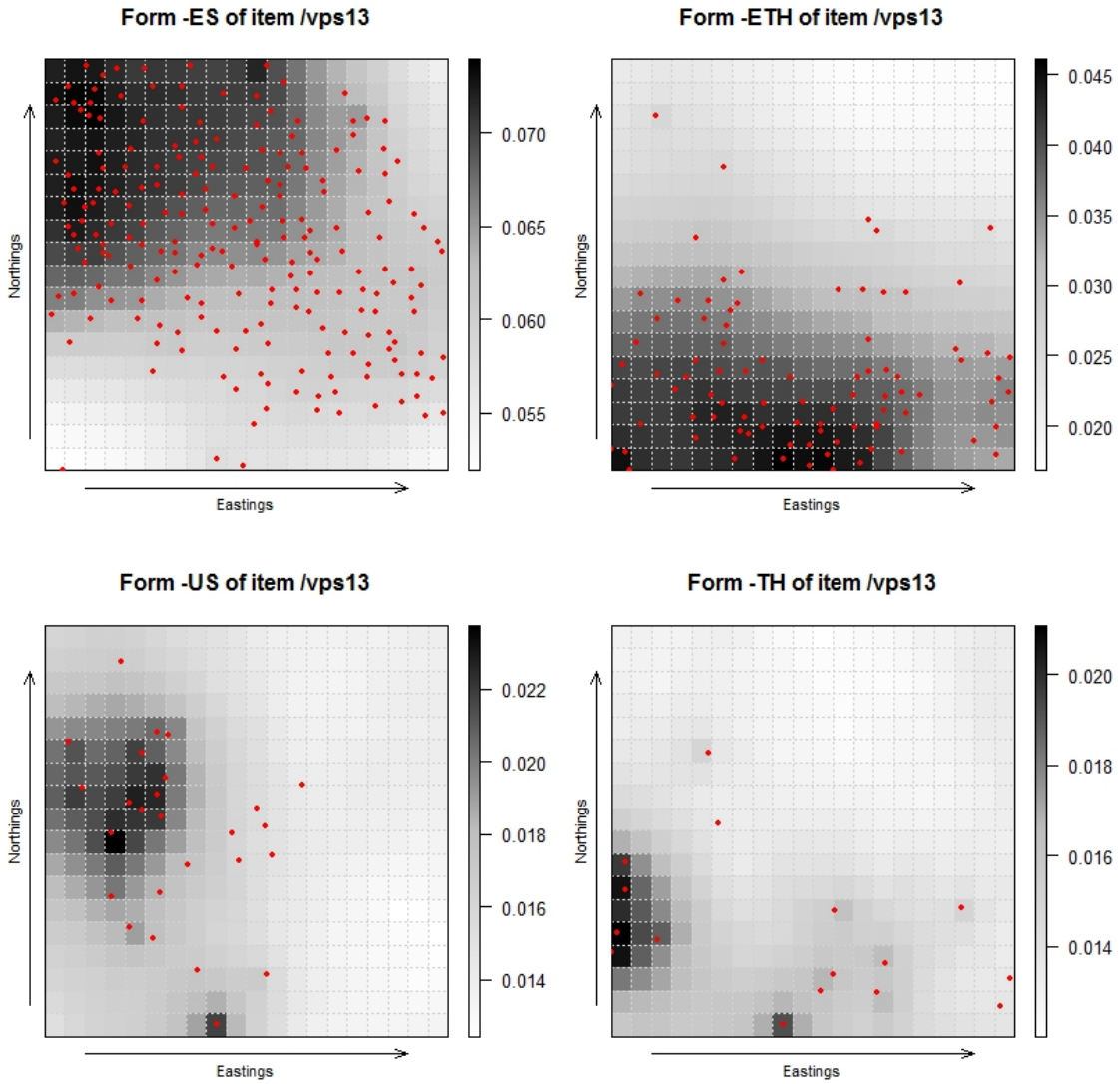


Figure 5.4: Mean posterior form usage probabilities $\hat{\pi}_\eta(\eta_{x,i,f} \mid y)$ across the lattice for four primary data forms of the item 'vps13', estimated using the ℓ_1-Dirichlet zero-inflated model (fit to the eight-item primary data subset $\mathbb{D}$). The colour bar beside each plot gives the scale of the $\hat{\pi}_\eta$ values. The red dots indicate the locations provided in the Atlas for the manuscripts that contained these forms.

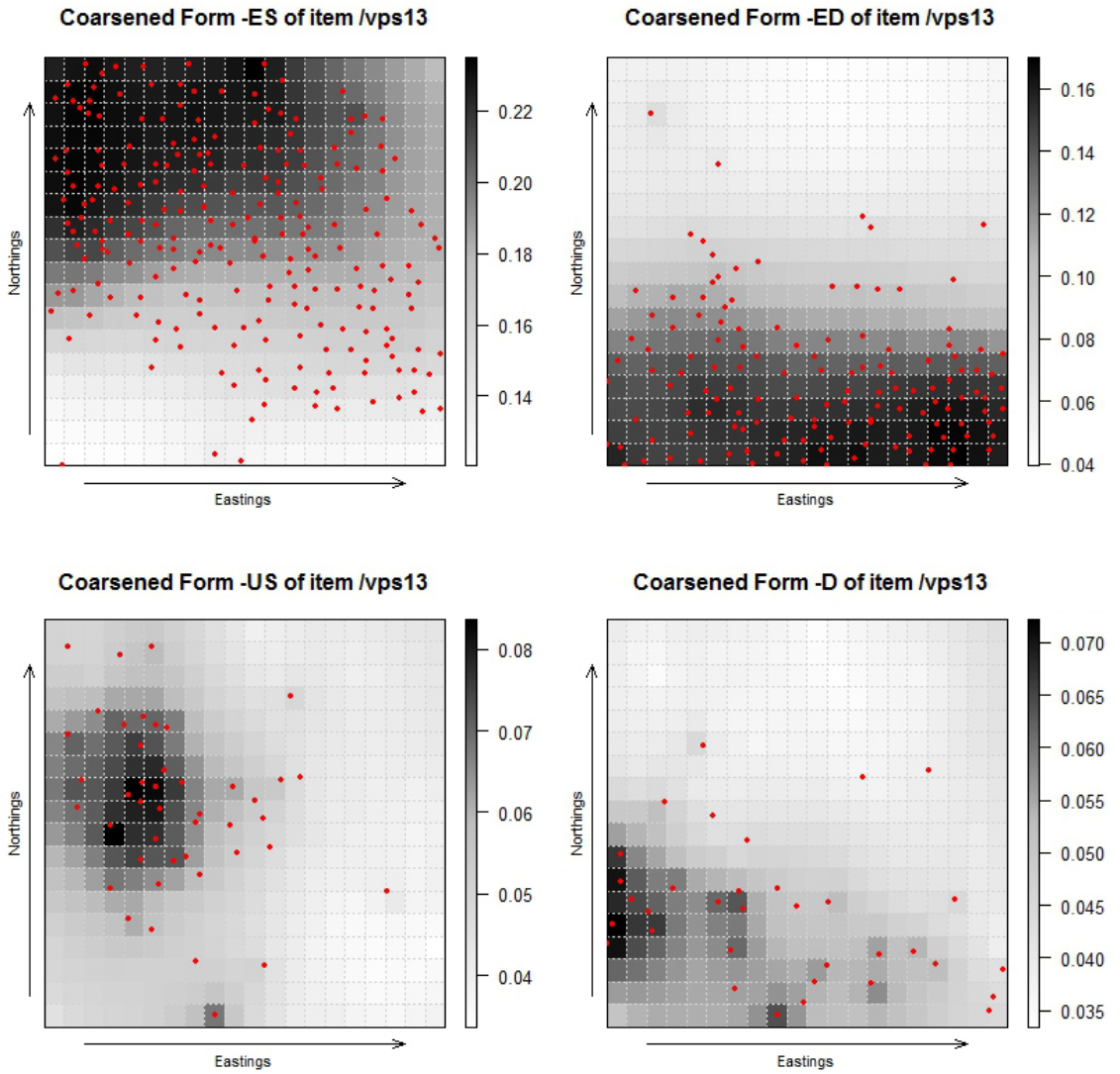


Figure 5.5: Mean posterior form usage probabilities $\hat{\pi}_\eta(\eta_{x,i,f} \mid y)$ across the lattice for the four coarsened data forms of the item 'vps13' equivalent to the primary forms shown in Figure 5.4. The $\hat{\pi}_\eta$ were estimated using the ℓ_1-Dirichlet zero-inflated model (fit to the eight-item coarsened data subset $\mathbb{S}$). The colour bar beside each plot gives the scale of the $\hat{\pi}_\eta$ values. The red dots indicate the locations provided in the Atlas for the manuscripts that contained these forms.

Figure 5.6 displays these estimated probabilities for each data subset for six randomly selected profiles $q \in Q$, with darker shades representing more probable cells than lighter shades. With both datasets, each profile has either one or two regions of high estimated posterior probability, with a relatively smooth decline in probability away from these regions. There is generally reasonable agreement between the area of highest estimated mean posterior probability and the Atlas cell-location $J(\lambda_q)$. Although the Atlas does not record uncertainty in individual locations, these $J(\lambda_q)$ can be thought of as having some (unknown) confidence region around them. As such, comparing our area of highest estimated mean posterior probability to the $J(\lambda_q)$ (red crosses) is conservative.

These estimated marginal posterior location probabilities for the two different datasets are generally similar in appearance. The areas of high estimated posterior probability for the location of origin of these six profiles are generally of a similar size for inference from primary and coarsened data. This is interesting, given they are based on coarsened data, and one might expect more diffuse posterior distributions given we have 'less' information.

However, we may actually gain by coarsening the data, beyond the obvious that coarsening leads to far fewer fields and far more rapid MCMC convergence. The information added when we coarsen is the group structure of the forms. As an example, if we were to split a form so that every instance of it was recorded as a distinct form, then we have no replication and are clearly worse off than with the original grouped

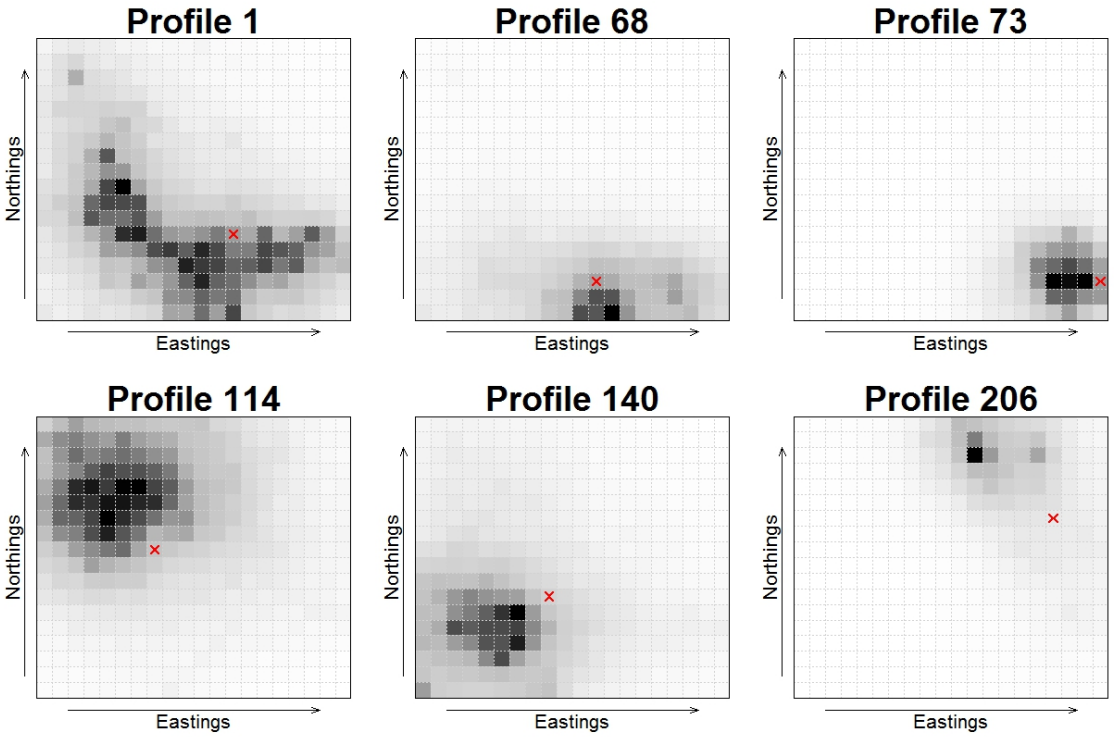
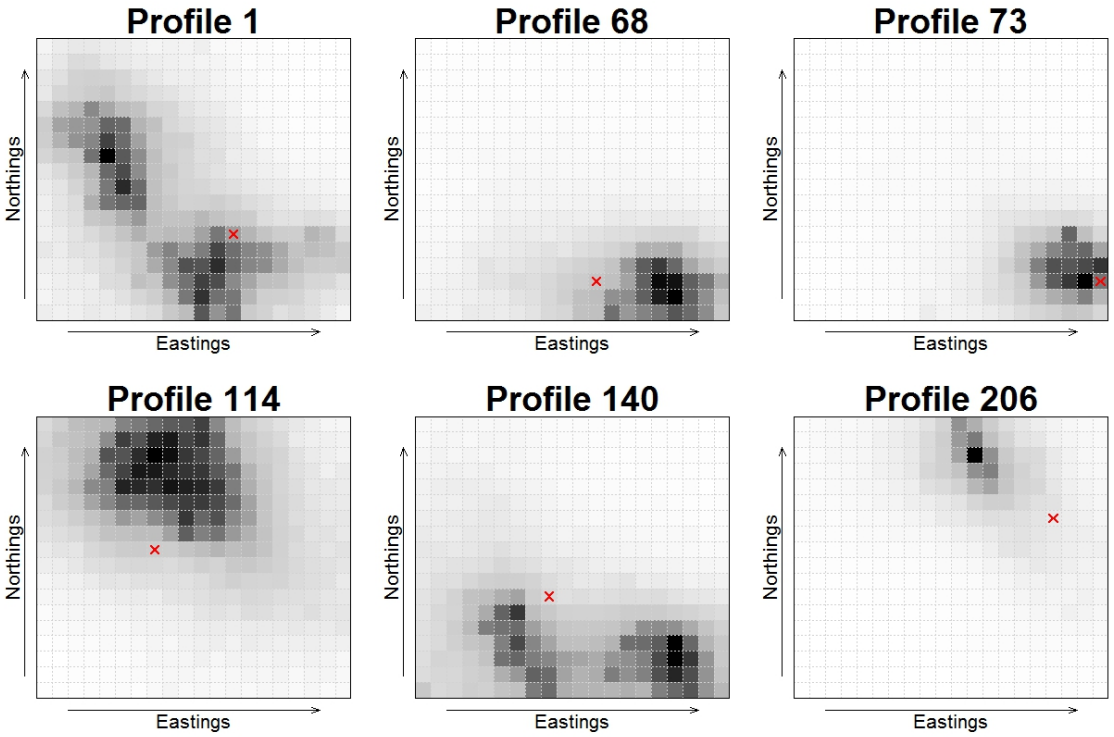


Figure 5.6: Estimated expected marginal posterior location probabilities $\hat{\pi}_x(x_q = x \mid y)$ (for cells $x = 1, \ldots, C$) are shown for six randomly selected floating profiles, out of the 123 whose locations we estimated using the ℓ_1-Dirichlet zero-inflated model. Darker cells have higher $\hat{\pi}_x$ values. The locations provided in the Atlas for these profiles are marked in red. Results with the eight-item coarsened data subset $\mathbb{S}$ are displayed in the top plot, and those with the eight-item primary data subset $\mathbb{D}$ below.



data. A fully informed model would have both fine scale and coarsened hierarchy in the model, as when we coarsen the data, we effectively add some information about the grouping of forms. We discuss this concept further in Chapter 7, but in any case, it seems as though our coarsened data is still informative of location.

With each data subset, for each profile $q \in Q$, we formally test the disagreement between the Atlas location and the estimated marginal posterior probabilities. We compute the Bayes factor $B_{0,1}$ measuring the evidence against the hypothesis H_0 : $x_q = J(\lambda_q)$, when compared to the alternative $H_1 : x_q \in \{1, \ldots, C\}$. Therefore, we measure the evidence against profile q being located in the cell-location predicted by the Atlas, when compared to the non-informative model prior that it is located anywhere on the lattice. Informally, we are asking if our model is so bad it would cause us to reject the Atlas "truth".

The two priors (representing H_0 and H_1) are nested, so the Bayes factor for this test is an instance of a Savage-Dickey ratio (see Verdinelli & Wasserman (1998)).

We evaluate

$$\begin{aligned} B_{0,1} &= \frac{\mathrm{p}(y \,|\, H_0)}{\mathrm{p}(y \,|\, H_1)} \\ &= \frac{\pi_x(x_q = J(\lambda_q) \,|\, x_{-q}, y)}{\pi_x(x_q = J(\lambda_q))} \frac{\pi_x(x_q \in \{1, \ldots, C\})}{\pi_x(x_q \in \{1, \ldots, C\} \,|\, x_{-q}, y)} \\ &= \frac{\pi_x(x_q = J(\lambda_q) \,|\, x_{-q}, y)}{\pi_x(x_q = J(\lambda_q))}. \end{aligned} \tag{5.4}$$

The Bayes factor is thus the posterior probability for profile q to be in the Atlas cell-location $J(\lambda_q)$, divided by the prior probability $\pi_x(x_q = J(\lambda_q))$ for profile q to be

in $J(\lambda_q)$. Note that because x_q has a uniform prior, then $\pi_x(x_q = J(\lambda_q)) = 1/C$.

Kass & Raftery (1995) describe a guideline of strong evidence against H_0 when $-2\log(B_{0,1}) > 6$, and very strong evidence when $-2\log(B_{0,1}) > 10$. If $B_{0,1} > 1$, then our data support H_0.

Table 5.4 displays the Bayes Factors $B_{0,1}$ for the six profiles $q \in Q$ displayed in Figure 5.6. The Bayes Factors for these profiles are similar in each case though generally higher with the coarsened data.

	$\mathbb{S}$		$\mathbb{D}$	
Profile	$\boldsymbol{B_{0,1}}$	$\boldsymbol{-2\log(B_{0,1})}$	$\boldsymbol{B_{0,1}}$	$\boldsymbol{-2\log(B_{0,1})}$
1	1.89	-1.27	1.93	-1.32
68	4.80	-3.14	3.16	-2.30
73	8.72	-4.33	7.44	-4.01
114	1.60	-0.94	1.20	-0.36
140	1.73	-1.09	1.16	-0.30
206	2.00	-1.39	1.73	-1.10

Table 5.4: Bayes factors $B_{0,1}$ (and log-Bayes factors) for the six profiles displayed in Figure 5.6, following modelling with the ℓ_1-Dirichlet zero-inflated model. These measure the evidence against the fit-technique locations for these profiles. Results are shown for both the eight-item primary data subset $\mathbb{D}$ and the eight-item coarsened data subset $\mathbb{S}$.

Of course, these are only six of the 123 profiles which we have estimated locations for. Figure 5.7 displays the (log-transformed) Bayes Factors for all these profiles $q \in Q$ obtained when modelling each data subset. It can be observed that we would not reject the Atlas location on any occasion, with either data subset, using $-2\log(B_{0,1}) >$

6 as a threshold. A very slight majority of profiles have larger Bayes Factors (i.e. smaller log-Bayes Factors) with the coarsened data subset, but these differences are often small.

The plot is split into quadrants: evidence against the fit-technique location from both datasets (top-right), evidence supporting the fit-technique location from both datasets (bottom-left), and disagreement between datasets (top-left and bottom-right). It can be seen that there are hardly any profiles for which there is disagreement, and that their log-Bayes Factors are universally small.

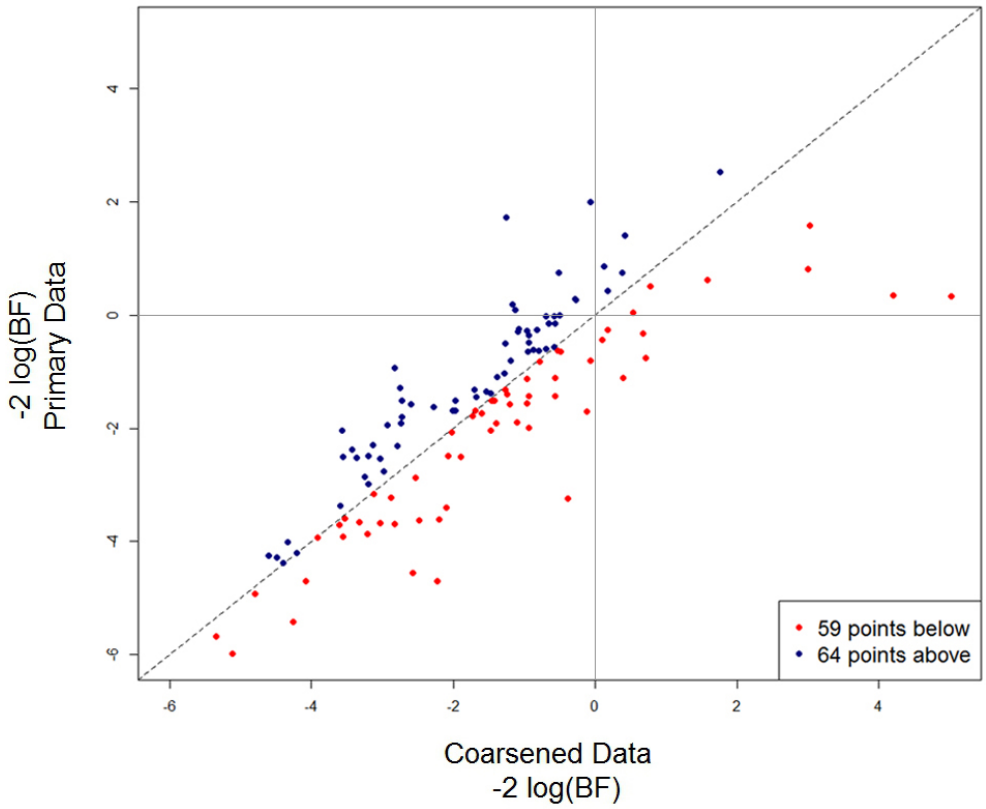


Figure 5.7: Bayes factors (on the log-scale) for all 123 floating profiles following modelling with the ℓ_1-Dirichlet zero-inflated model. The Bayes factors obtained from modelling the eight-item primary data subset $\mathbb{D}$ are compared to those obtained from modelling the eight-item coarsened data subset $\mathbb{S}$ (the diagonal dotted-line represents no difference), with 59 profiles having larger log-Bayes Factors with $\mathbb{S}$, and 64 with $\mathbb{D}$.

Although we would not reject the Atlas location on any occasion, we have evidence against the Atlas locations for 17 profiles with the coarsened data $\mathbb{S}$, and 19 with the primary data $\mathbb{D}$. Of these, the evidence is considered 'positive' (using the guidelines in Kass & Raftery (1995)) for four profiles with $\mathbb{S}$ and one with $\mathbb{D}$, and 'not worth more than a bare mention' for the rest.

Hence, we find no evidence for model misspecification in this goodness-of-fit test on reserved data. The reader may object that the posterior location distributions are so diffuse there was never any chance of the reserved data conflicting with the model anyway. In order to test this, we consider the typical magnitude of Bayes Factors if the Atlas locations (marked with a red 'X' in the location posterior plots) were completely random.

For each profile $q \in Q$, we compute the set of Bayes factors $B_{0,1}$ with $H_0 : x_q = x$ for every $x \in \{1, \ldots, C\}$, and consider how many of the 360 cells x would lead to evidence against H_0 (i.e. a Bayes factor $B_{0,1} < 1$). We find a median of 247 such cells per profile $q \in Q$ when using the primary data subset $\mathbb{D}$, ranging between [166, 319] cells. When using the secondary data subset S, we find a median of 243 cells per profile, ranging between [179, 310] cells.

Finally, for each profile $q \in Q$, we consider in how many cells x we would reject H_0, based on strong evidence against H_0 (i.e. $-2\log(B_{0,1}) > 6$). We obtain a median of 32 such cells x with data $\mathbb{D}$, ranging between [0, 235] cells. With the data $\mathbb{S}$, this

median is 60 cells with a range of [0, 248] cells.

A straightforward calculation shows that, if the Atlas locations for the non-anchor profiles were uncorrelated with our posterior maps (i.e. if they were random), with the primary data $\mathbb{D}$ we would expect evidence against the Atlas locations ($B_{0,1} < 1$) for approximately 84 (i.e. 123*247/360) of our profiles $q \in Q$. Further, we would expect strong evidence against the Atlas locations ($-2\log(B_{0,1}) > 6$) with approximately 11 (i.e. 123*32/360) of the profiles $q \in Q$. In fact we find 19 and 0 respectively, so much higher levels of conflict between data and model are possible for our goodness-of-fit test, but are not seen.

Similar conclusions are drawn with the coarsened data $\mathbb{S}$. We conclude that the objection that our posterior distributions are too diffuse to reject anything does not apply. We further conclude that the ℓ_1-Dirichlet zero-inflated model is able to successfully jointly estimate the unknown location of origin of a large number of randomly selected profiles, with all the other floating profiles fixed to the positions provided in the Atlas. Additionally, the linguistic data still appears to be informative of location after our coarsening operation.

5.2.3 ℓ_1-Dirichlet Outlier Model

We now repeat the modelling exercise of the previous section, using instead the ℓ_1-Dirichlet outlier model described in Section 2.2. Evidence is provided in this section that this outlier model adds little, with the results presented being very similar to

those obtained earlier with the ℓ_1-Dirichlet zero-inflated model.

The ℓ_1-Dirichlet outlier model has parameters $\Theta = (\mu, \phi, \eta, x, \psi)$. We fit the model, in turn, to the data subsets $\mathbb{S}$ and $\mathbb{D}$ described earlier in the chapter, and compare the results to those obtained using the ℓ_1-Dirichlet zero-inflated model. We do not provide plots of the estimated posterior distributions for μ_i and ϕ_i here, because they mirror those provided in Figure 5.3 so closely that they are indistinguishable.

The estimated posterior distributions for the outlier-probabilities $\psi_i^{(I)}$, for a handful of items, are displayed in Figure 5.8 for each data subset. With both datasets, the highest posterior weight is generally for values between 0 and 0.1. Figure 5.8 also shows the estimated posterior distributions for the outlier-probabilities $\psi_p^{(P)}$ for a few profiles $q \in Q$. We can see that these are identical to the Beta$(2, 3)$ prior distributions imposed on the parameters. It is not surprising that we can estimate $\psi^{(I)}$ better than $\psi^{(P)}$, given that for item i, each profile p is a test of $\psi_i^{(I)}$ versus $1 - \psi_i^{(I)}$. Since there are many profiles and only a few items, this means that $\psi^{(I)}$ are well informed. Conversely, there are far fewer tests to inform $\psi^{(P)}$.

As with the item usage rates μ_i and the zero-inflation probabilities ϕ_i, the estimated mean posterior form usage probabilities $\hat{\pi}_\eta$ from the ℓ_1-Dirichlet outlier model are very similar to those from the ℓ_1-Dirichlet zero-inflated model. Plots equivalent to Figures 5.4 and 5.5 are so similar that we do not include them here (the only noticeable difference is that the area of high estimated mean posterior probability for

the (primary data) form '-ES' extends further south-east).

Figure 5.9 displays the estimated expected marginal posterior probabilities $\hat{\pi}_x(x_q = x \mid y)$ for the same six profiles $q \in Q$ as Figure 5.6. With both datasets, these posterior probabilities are nearly identical to those obtained from the ℓ_1-Dirichlet zero-inflated model. The Bayes Factors $B_{0,1}$ for these six profiles are consequently very similar to those obtained earlier.

Indeed, this similarity extends beyond these six profiles. The equivalent plot to Figure 5.7 for the ℓ_1-Dirichlet outlier model displays a very similar pattern, and hence is not included here. Instead, in Figure 5.10, we compare the Bayes Factors (on the log-scale) from the ℓ_1-Dirichlet outlier model to those from the ℓ_1-Dirichlet zero-inflated model. These Bayes Factors are for the coarsened data subset $\mathbb{S}$, but a similar pattern is observed for the primary data subset $\mathbb{D}$. We can see that there is very little difference between the Bayes Factors derived in each case, with this few items.

As earlier, we would not reject the Atlas location on any occasion, with either data subset, when using the ℓ_1-Dirichlet outlier model. We have evidence against the Atlas locations for 14 profiles with the coarsened data $\mathbb{S}$, and 25 with the primary data $\mathbb{D}$. Of these, the evidence is considered 'not worth more than a bare mention' (using the guidelines in Kass & Raftery (1995)) for nearly all of them.

As noted throughout this section, the results obtained here are very similar to

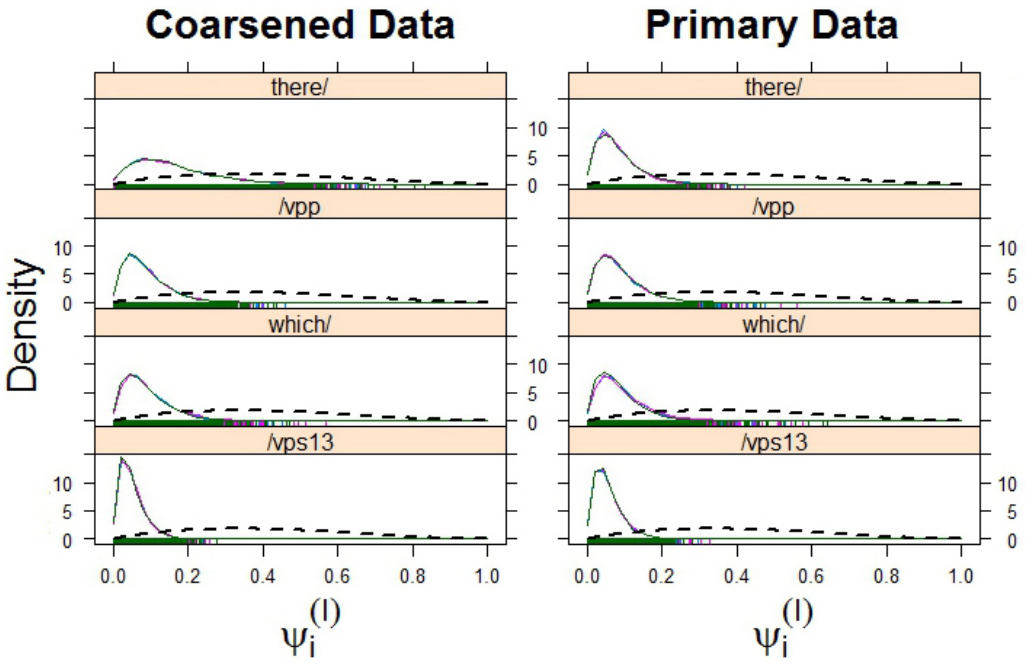
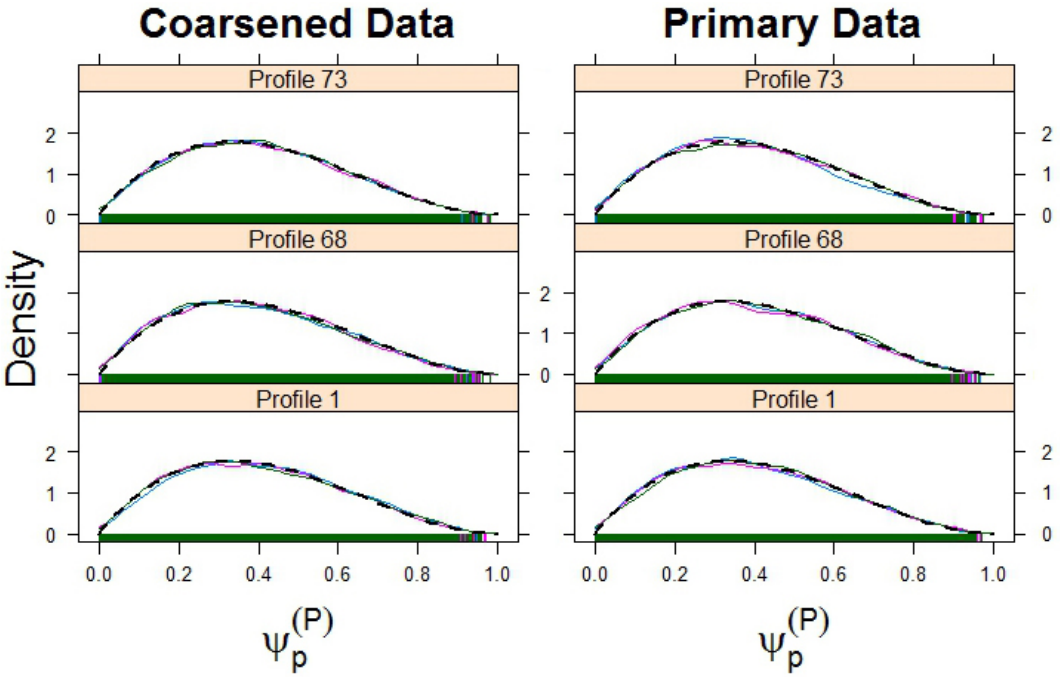


Figure 5.8: Estimated posterior distributions for outlier probabilities $\psi_i^{(I)}$ (above) and outlier probabilities $\psi_p^{(P)}$ (below) from the ℓ_1-Dirichlet outlier model. Results are shown for eight-item subsets of the coarsened (left) and primary data (right). Three density plots are shown for each parameter, with one for each parallel MCMC chain. The Beta prior distribution for each parameter is overlaid with a dotted black line.



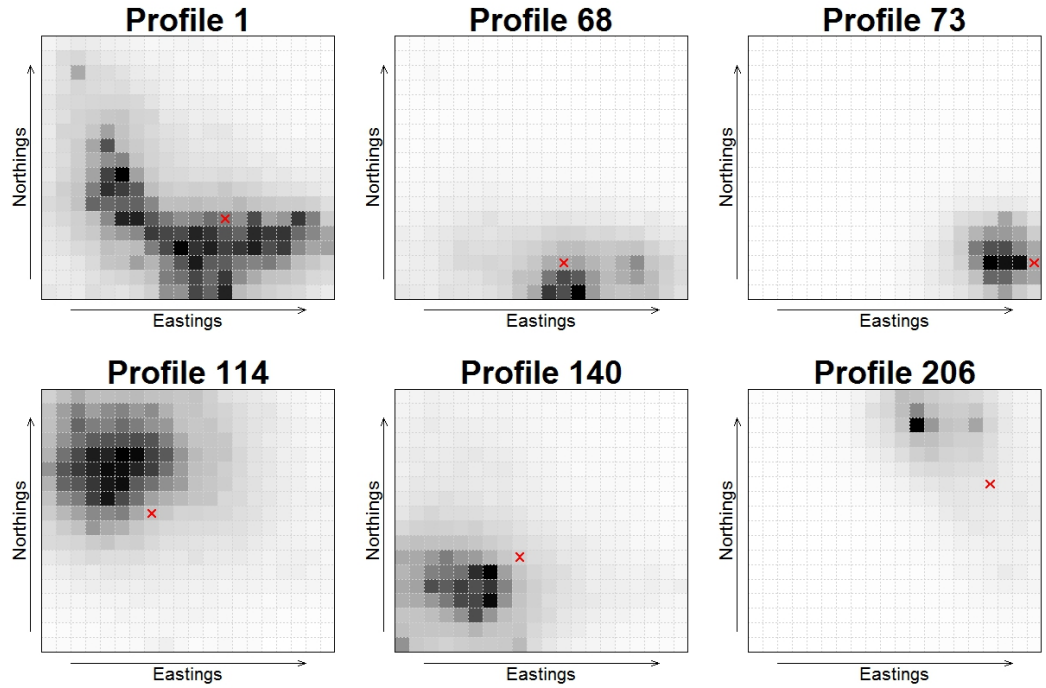
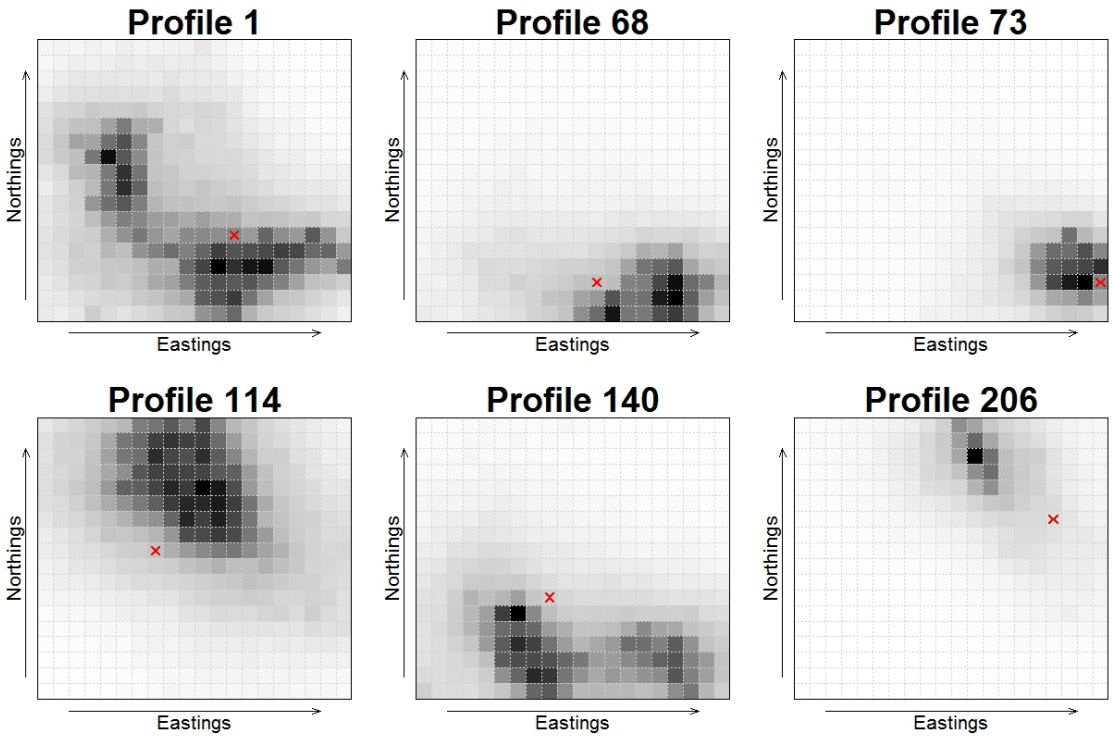


Figure 5.9: Estimated expected marginal posterior location probabilities $\hat{\pi}_x(x_q = x \,|\, y)$ (for cells $x = 1, \ldots, C$) are shown for the same six of 123 floating profiles as in Figure 5.6. Locations were estimated using the ℓ_1-Dirichlet outlier model. Darker cells have higher $\hat{\pi}_x$ values. The locations provided in the Atlas for these profiles are marked in red. Results with the eight-item coarsened data subset $\mathbb{S}$ are displayed in the top plot, and those with the eight-item primary data subset $\mathbb{D}$ below.



those in Section 5.2.2. Given this, as well as that the ℓ_1-Dirichlet outlier model is computationally much more expensive to simulate than its zero-inflated equivalent, we set the outlier probabilities to zero (i.e. use the zero-inflated model) in future ℓ_1-Dirichlet modelling. A counter argument to this decision would be the belief that the ℓ_1-Dirichlet outlier model will be important when modelling with large numbers of items (rather than just eight). We do not further investigate this hypothesis in this thesis, but additional investigation would be an interesting avenue to pursue in future work.

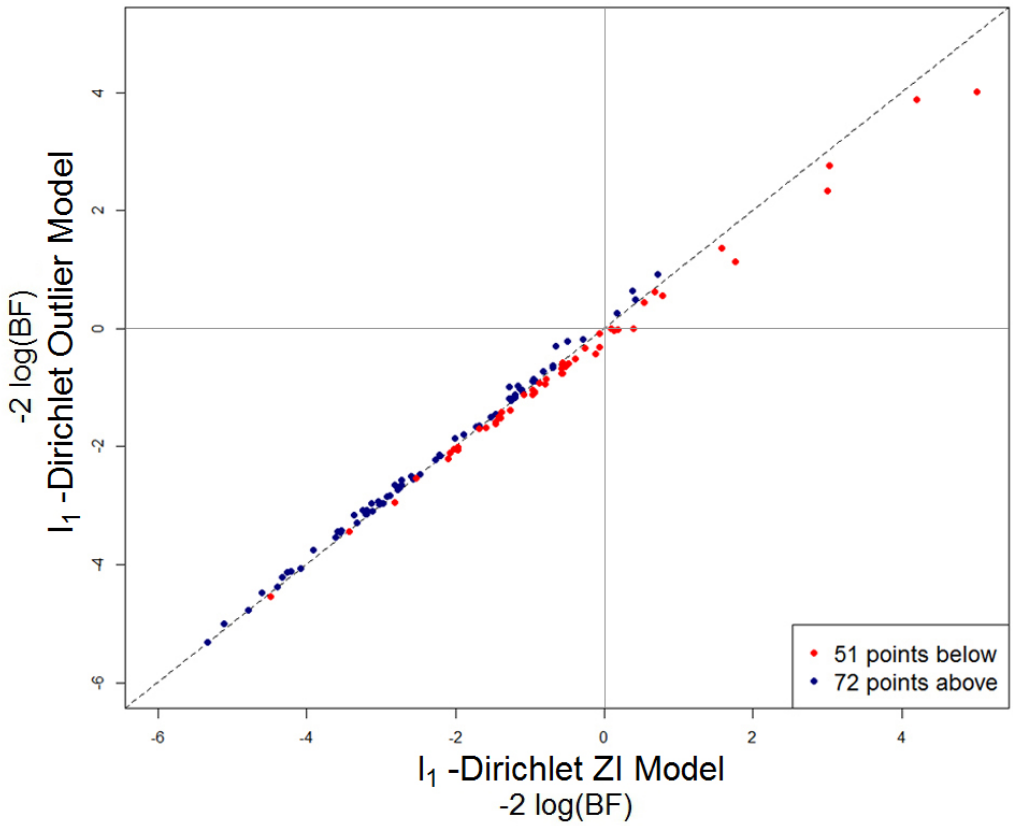


Figure 5.10: Comparison of Bayes Factors (on the log-scale) for all 123 floating profiles whose locations were estimated earlier using the ℓ_1-Dirichlet zero-inflated and ℓ_1-Dirichlet outlier models. Results are shown for the eight-item coarsened data subset $\mathbb{S}$.

5.2.4 ℓ_2-Logistic Models

In the previous two sections, we established that the ℓ_1-Dirichlet zero-inflated and outlier models give near identical results on the common problem considered. We now repeat the same modelling exercise one final time, using this time ℓ_2-logistic zero-inflated model described in Section 2.3. This model has parameters $\Theta = (\mu, \phi, \gamma, x)$.

Unlike with the ℓ_1-Dirichlet zero-inflated and outlier models, we fit this model only to the coarsened data subset $\mathbb{S}$ described earlier in the chapter, and compare the results to those previously obtained from the other models. The reasons for not fitting to the primary data subset are two-fold. For one, our results to this point have shown very little impact of the choice of data on the results obtained; and furthermore, MALA-based MCMC updates for the primary data are very time-consuming due to the vast increase in number of forms per item.

Once again, we find that the estimated posterior distributions for μ_i and ϕ_i very closely mirror those obtained with the other models (as displayed in Figure 5.3). We therefore do not provide these plots here.

The point of difference for the ℓ_2-logistic zero-inflated model to the ℓ_1-Dirichlet zero-inflated model is in how we model the spatial structure in the dialect fields. It is unsurprising then that mean posterior form usage probabilities $\hat{\pi}_\eta(\eta_{x,i,f} \mid y)$ obtained from this model look very different to those from the ℓ_1-Dirichlet zero-inflated model. These differences can be observed by comparing Figure 5.5 (for the ℓ_1-Dirichlet zero-

inflated model) to Figure 5.11 (for the ℓ_2-logistic zero-inflated model). Of course, we model in γ-space with the ℓ_2-logistic model, but the $\hat{\pi}_\eta(\eta_{x,i,f} \mid y)$ follow in a straightforward fashion by transforming into η-space using Equation (2.25).

We observe from Figure 5.11 that the regions of high posterior usage probability for each of the displayed forms are much smaller than those in Figure 5.5. Further, switching from an ℓ_1-norm to ℓ_2 has resulted in much sharper posterior distributions, with some multimodality in the regions of high posterior usage probability.

This multimodality and decrease in smoothness of the posterior form usage probability propagates through to the posterior distributions for the locations of our floating profiles, as evidenced in Figure 5.12. Similarly, as with the form usage probabilities, the areas of high posterior probability for the locations are smaller than those obtained with the earlier models (see Figures 5.6 and 5.9).

Despite this, we would still not reject any of the Atlas locations based on the Bayes Factors. However, there is evidence against the Atlas positions for the majority (68) of the floating profiles $q \in Q$.

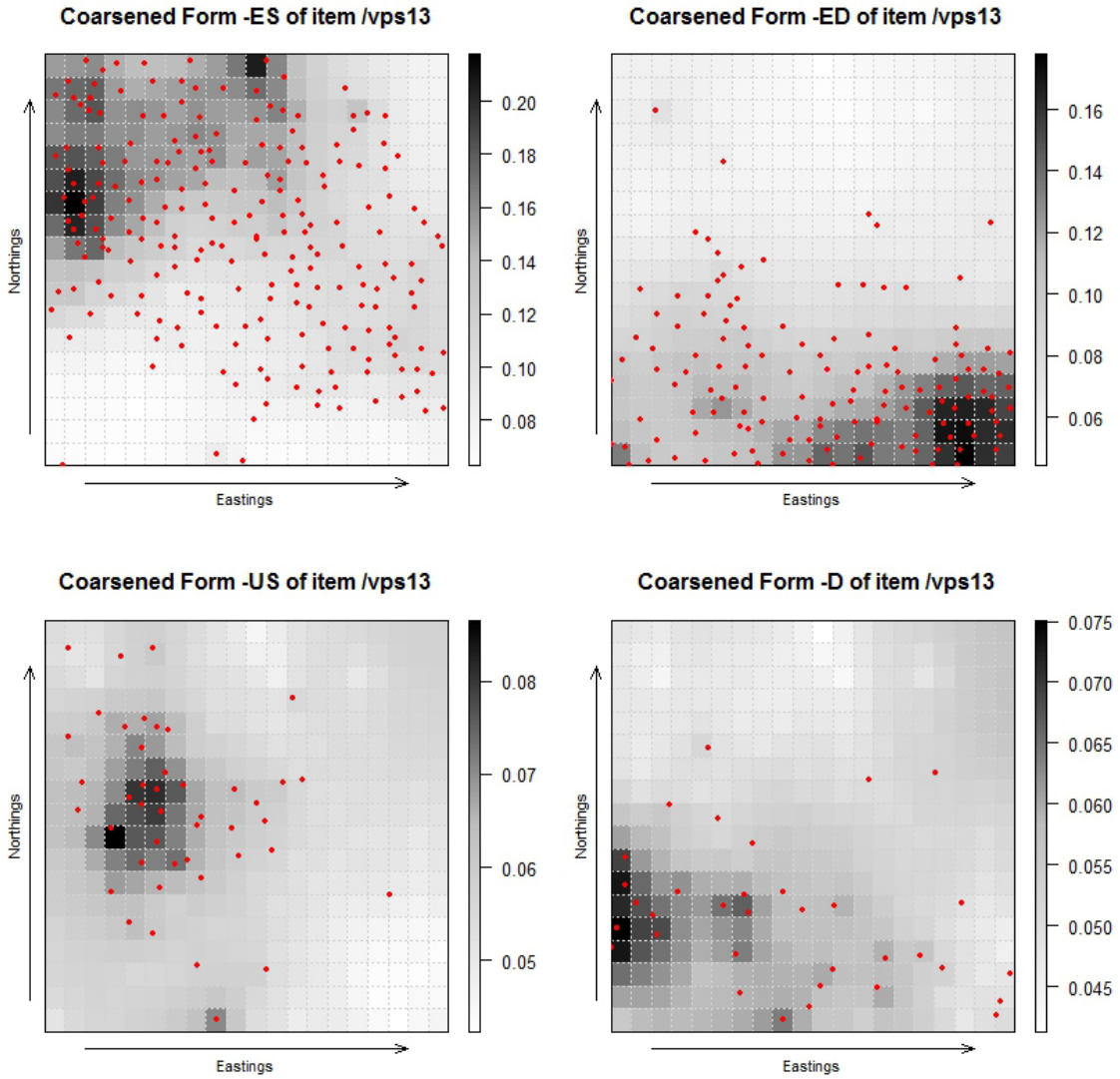


Figure 5.11: Estimated mean posterior form usage probabilities $\hat{\pi}_\eta(\eta_{x,i,f} \mid y)$ across the lattice for the four coarsened forms of the item 'vps13' shown in Figure 5.5. The $\hat{\pi}_\eta$ were estimated using the ℓ_2-logistic zero-inflated model (fit to the eight-item coarsened data subset $\mathbb{S}$). The colour bar beside each plot gives the scale of the $\hat{\pi}_\eta$ values. The red dots indicate the locations provided in the Atlas for the manuscripts that contained these forms.

Table 5.5 shows a comparison of the Bayes Factors from the ℓ_2-logistic zero-inflated model and the ℓ_1-Dirichlet zero-inflated model for the six profiles displayed in Figures 5.6 and 5.11. Most of these are bigger with the ℓ_1-Dirichlet zero-inflated model, and this trend extends beyond these six profiles. Figure 5.13 shows indeed that the estimated locations of the vast majority of the floating profiles are closer to their Atlas locations when using the ℓ_1-Dirichlet zero-inflated model.

The results presented in this section thus far have been from the ℓ_2-logistic zero-inflated model rather than the ℓ_2-logistic outlier model. We find results based on the latter model to be very similar to those from the former, so do not present them in full here. Briefly, though, this similarity can be observed from Figure 5.14.

	ℓ_2-logistic ZI model		**ℓ_1-Dirichlet ZI model**	
Profile	$\boldsymbol{B_{0,1}}$	$\boldsymbol{-2\log(B_{0,1})}$	$\boldsymbol{B_{0,1}}$	$\boldsymbol{-2\log(B_{0,1})}$
1	0.48	1.48	1.89	-1.27
68	1.67	-1.03	4.80	-3.14
73	13.94	-5.27	8.72	-4.33
114	0.60	1.01	1.60	-0.94
140	0.53	1.28	1.73	-1.09
206	0.29	2.49	2.00	-1.39

Table 5.5: Bayes factors $B_{0,1}$ (and log-Bayes factors) for the six profiles displayed in Figures 5.6 and 5.12, following modelling with the ℓ_1-Dirichlet and ℓ_2-logistic zero-inflated models. These Bayes Factors measure the evidence against the fit-technique locations for these profiles.

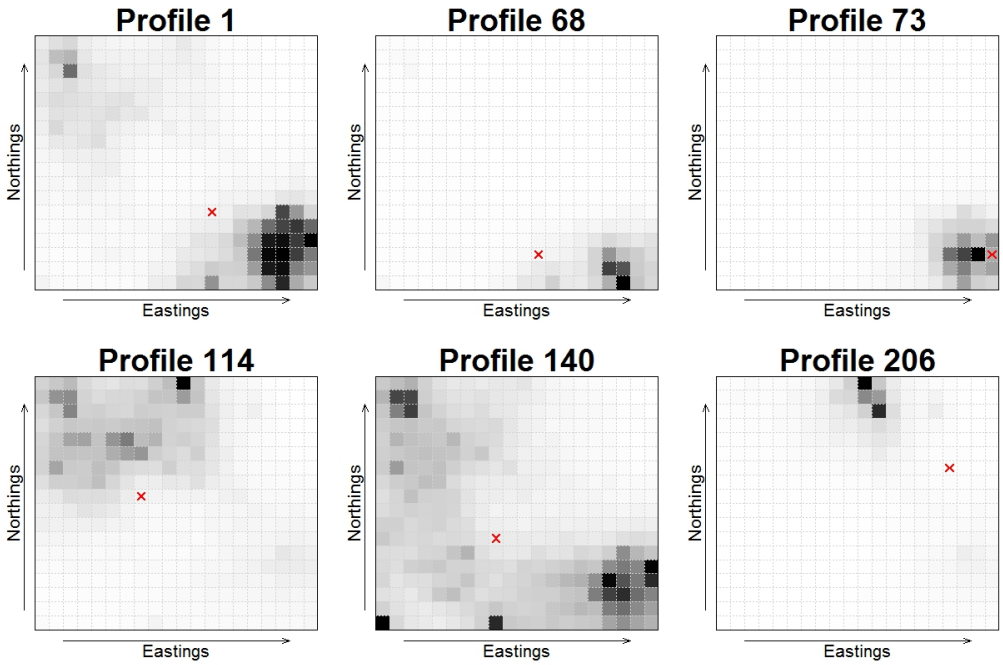


Figure 5.12: Estimated expected marginal posterior location probabilities $\hat{\pi}_x(x_q = x \mid y)$ (for cells $x = 1, \ldots, C$) are shown for the same six of 123 floating profiles as in Figures 5.6 and 5.9. Locations were estimated using the ℓ_2-logistic zero-inflated model. Darker cells have higher $\hat{\pi}_x$ values. The locations provided in the Atlas for these profiles are marked in red. Results are shown for the eight-item coarsened data subset $\mathbb{S}$.

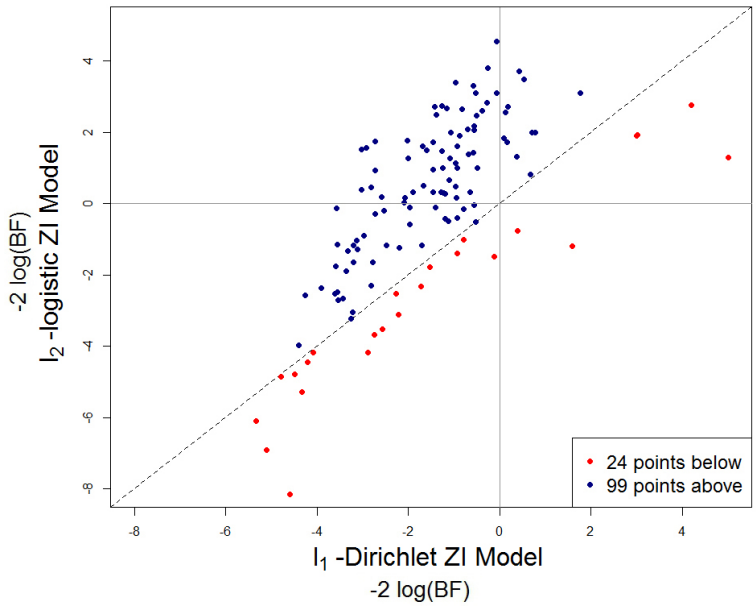


Figure 5.13: Comparison of Bayes Factors (on the log-scale) for all 123 floating profiles from the ℓ_1-Dirichlet zero-inflated and ℓ_2-logistic zero-inflated models. Bayes Factors are shown for model results based on the coarsened data subset $\mathbb{S}$.

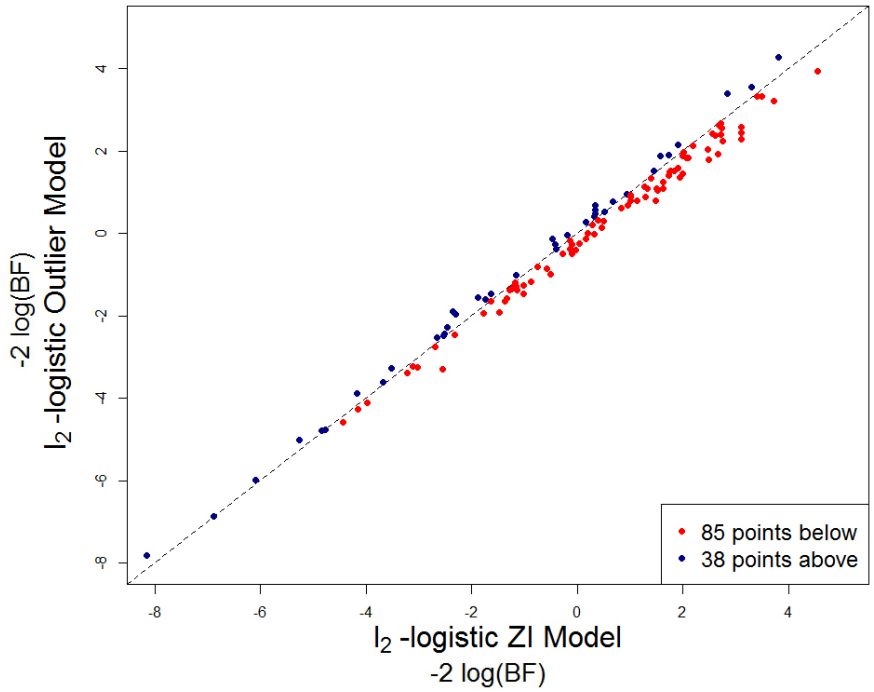


Figure 5.14: Comparison of Bayes Factors (on the log-scale) for all 123 floating profiles from the ℓ_2-logistic zero-inflated and ℓ_2-logistic outlier models. Bayes Factors are shown for model results based on the coarsened data subset $\mathbb{S}$.

5.3 Locating All Non-Anchor Profiles

We remind the reader that there are two ways in which we can frame the applied problem, depending on the profiles fixed in place to inform the "dialect fields". We either locate all of the non-anchor texts, or instead we trust the provided Atlas locations for the non-anchor texts and just locate new texts. The analyses presented throughout this chapter, to this point, have been of the latter variety. We now consider analysis under the former, more difficult, setting.

Our earlier explorations of the capabilities of our various models established several things. First, our coarsened data gave very similar results to the original (primary)

binary data. Second, the simplest of our models (the ℓ_1-Dirichlet zero-inflated model) gave very similar results to the more complex ℓ_1-Dirichlet outlier model, and better results than the equivalent model with ℓ_2-penalties on the dialect fields.

To some extent, these results give justification to our decision to proceed with the coarsened data and the ℓ_1-Dirichlet zero-inflated model, though we did expect the advantage of using the ℓ_1-Dirichlet outlier model to be evident with larger datasets than it was tested with. Really, these choices happen to be computationally convenient ones. Given the small number of anchors (120) relative to floating profiles (247), we are now tackling a far more difficult problem, and thus need to extend beyond the subsets of the linguistic data used so far. Utilising the simplest model and the reduced data helps to offset the increase in computational processing required.

We use the entire coarsened dataset, which consists of 741 forms across 71 items (two of the 73 items with coarsened data available are not used in any of the 367 profiles) observed at least once within $\mathbb{W}$. With this data, we estimate the unknown locations of origin for the 247 linguistic profiles whose locations were derived from the fit-technique. We then compare these estimates with the fit-technique locations.

The results we obtain are poor. Figure 5.15 shows how the areas of high posterior probability for the locations of origin of the floating profiles are drawn towards the north-west corner of $\mathbb{W}$. Referring back to Figure 5.1, it can be observed that this is where the majority of the 120 anchor profiles in $\mathbb{W}$ are concentrated. Figure 5.16

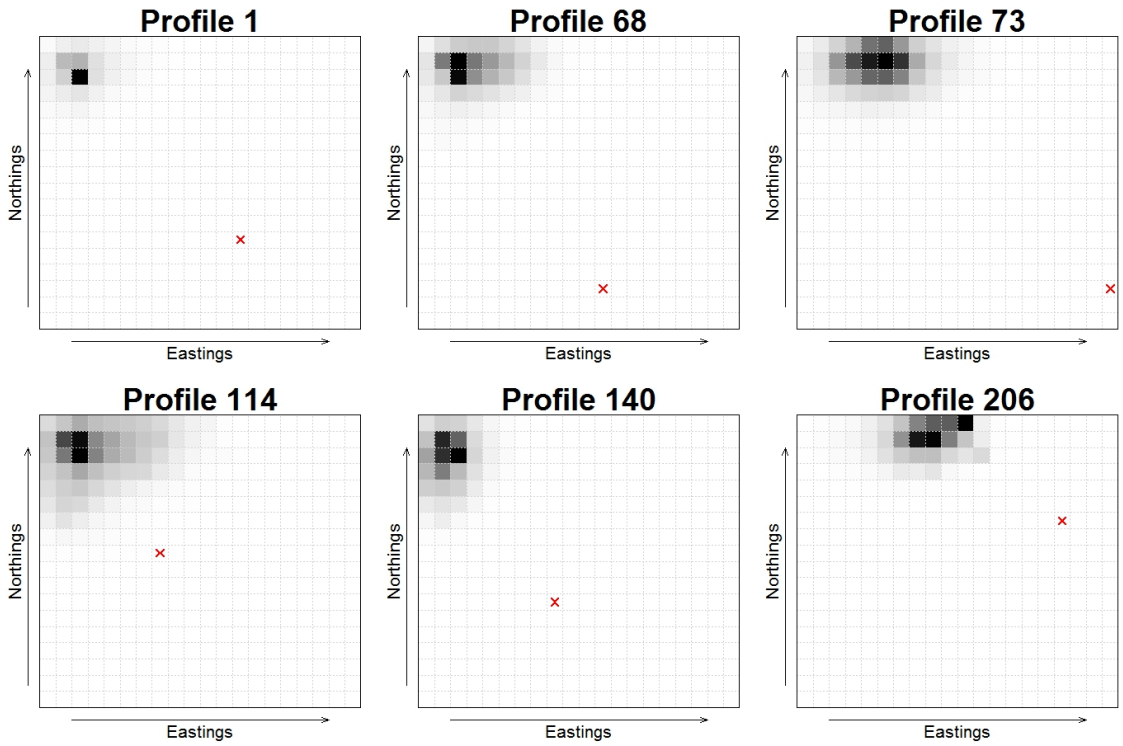


Figure 5.15: Estimated expected marginal posterior location probabilities $\hat{\pi}_x(x_q = x \mid y)$ (for cells $x = 1, \ldots, C$) are shown for the same six floating profiles as in Figures 5.6, 5.9 and 5.12. This time 247 locations were estimated using the ℓ_1-Dirichlet zero-inflated model. Darker cells have higher $\hat{\pi}_x$ values. The locations provided in the Atlas for these profiles are marked in red. Results are shown for the entire coarsened dataset.

displays the relationship between the anchor locations and the η-fields.

There are two related features associated with this modelling failure. One of these is that the model tends to fail with the anchors so unevenly spread. We show in Section 5.3.1, with the same linguistic data, our model returns meaningful location estimates if the anchors are more evenly spread.

The other feature associated with the poor results obtained relates to the interaction between the anchor and floating profiles. We estimate the locations of the

floating profiles jointly with the form usage probability fields (i.e. η). This allows the floating profiles to contribute back into the dialect fields. However, it appears that this feedback from the floating profiles is overwhelming the anchors.

Take the posterior distribution for the coarsened form '-ED', for example. The area of highest posterior probability surrounds the one north-western anchor profile (which coincides with where the majority of the floating profiles are now located), rather than the southern area with multiple anchors (which is also where the Atlas believed this form to be most widely used).

To further demonstrate the impact this issue has, we present an analysis in Section 5.3.2 whereby the η-fields are estimated using only the anchor profiles, and show how the inference changes. We do not fully understand this phenomenon, and this is certainly an interesting area to be explored in future work.

5.3.1 Modelling with a Geographically Even Spread of Anchor Profiles

In this section, we present an analysis to show the impact the clustering of the anchors has on our floating profile location estimates. To do this, we investigated how the inference would change with a more even geographical spread of anchors across $\mathbb{W}$.

The anchor profiles from the Atlas were those which could be located based on the non-linguistic grounds. Here, we ignore this information, and instead select profiles to be 'anchors' ourselves using the following methodology. We split the modelling

region $\mathbb{W}$ into 12 equally sized sub-regions, each measuring 50 x 60 km. From each sub-region, we randomly chose 10 profiles whose Atlas locations fell within the sub-region as 'anchors'. This results in the same number of anchors (120) as there actually are, but these 'anchors' are distributed spatially far more evenly across $\mathbb{W}$.

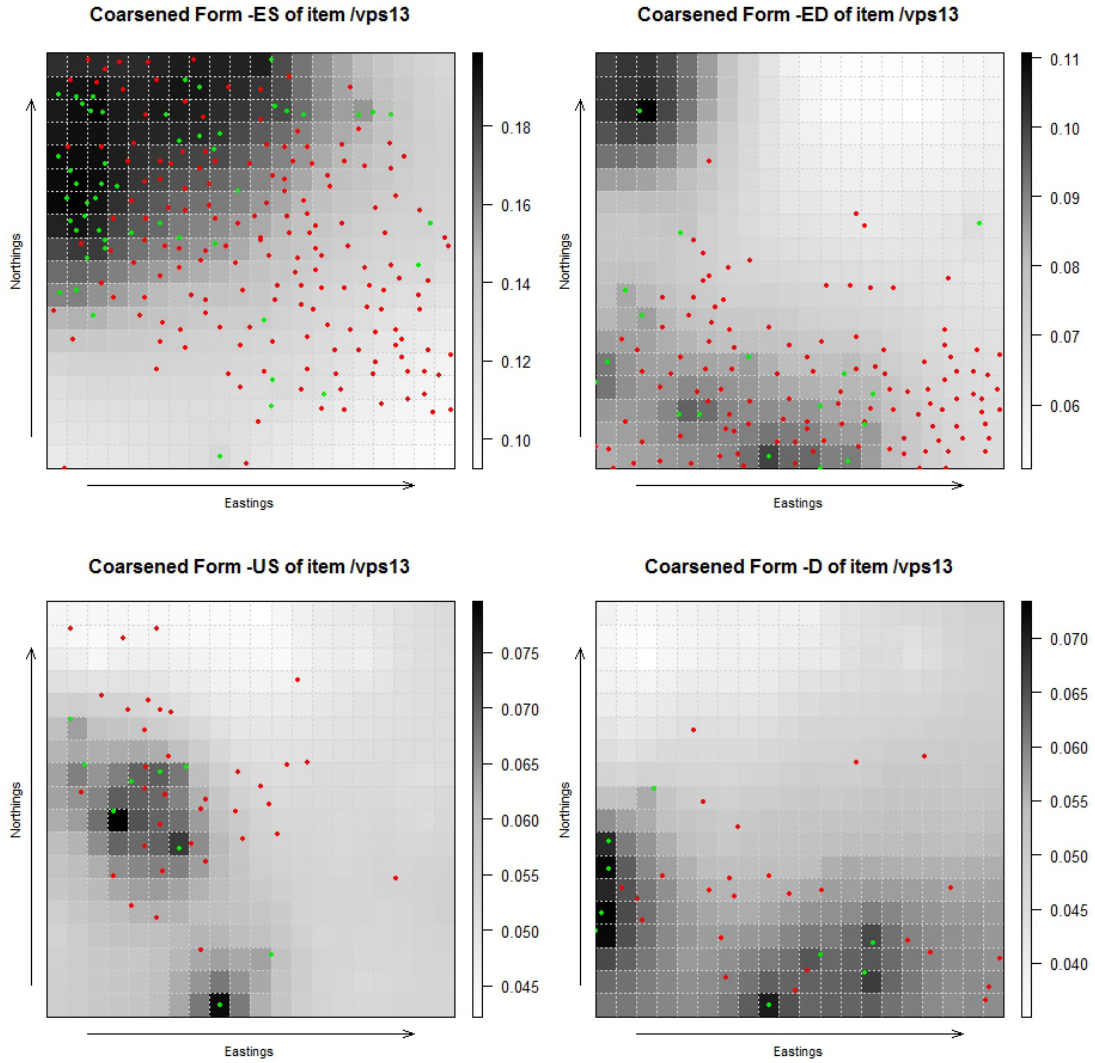


Figure 5.16: Estimated mean posterior form usage probabilities $\hat{\pi}_\eta(\eta_{x,i,f})$ across the lattice for four coarsened forms of the item 'vps13'. The $\hat{\pi}_\eta$ were estimated using the ℓ_1-Dirichlet zero-inflated model (fit to the full coarsened dataset). The colour bar gives the scale of $\hat{\pi}_\eta$ values. The green dots indicate the known locations of the Atlas profiles, whereas the red dots show the fit-technique locations for the floating profiles which contained these forms.

We repeat the modelling exercise of Section 5.3, using all of the coarsened data with the ℓ_1-Dirichlet zero-inflated model to estimate locations of origin for the other 247 profiles with Atlas location estimates within $\mathbb{W}$.

Figure 5.17 shows the estimated expected marginal posterior location probabilities $\hat{\pi}_x(x_q = x \mid y)$ for six of these profiles. Note that five of these are the same profiles as shown in the earlier equivalent plots (the other profile from earlier is now an 'anchor'). We observe that the areas of high posterior probability for the locations of origin of these profiles are now much closer to the Atlas estimates. More importantly, these estimates are no longer drawn to the north-west.

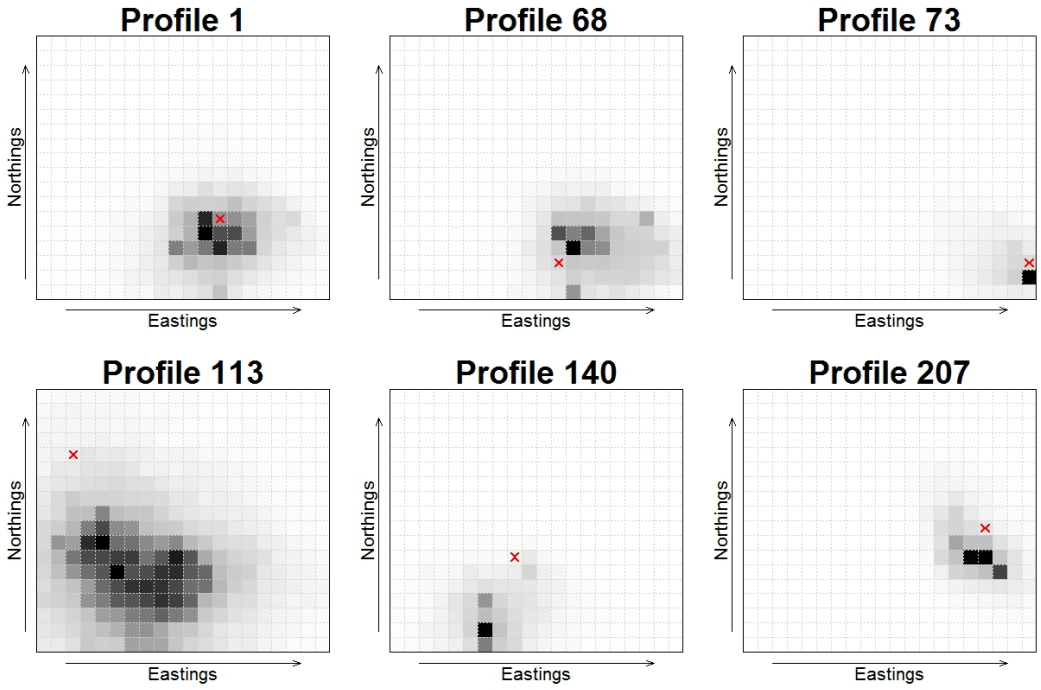


Figure 5.17: Estimated expected marginal posterior location probabilities $\hat{\pi}_x(x_q = x \mid y)$ (for cells $x = 1, \ldots, C$) are shown for six of 247 'floating' profiles, where anchors were selected artificially so they were evenly spread across the region. These locations were estimated using the ℓ_1-Dirichlet zero-inflated model. Darker cells have higher $\hat{\pi}_x$ values. The locations provided in the Atlas for these profiles are marked in red. Results are shown for the entire coarsened dataset.

5.3.2 Estimating Locations and Dialect Fields Separately

In Section 1.2, we described related work by Wasser et al. (2004, 2007, 2008, 2009, 2015), which considers a similar problem to ours, but which takes a modelling approach that differs from ours in several ways. The scale of the problem in this work is quite different to ours, having far more anchors (1350) and far fewer samples with unknown locations (28), as well as fewer 'items' (16). The underlying parameter fields are smoothed differently (not using ℓ_1 smoothing as we do with our much larger fields), and not estimated jointly with the unknown locations.

The intuition of our approach is that joint estimation exploits the information about the field parameters contained in the multiple samples of unknown origin. However, as mentioned earlier, it appears that these samples of unknown origin are overwhelming the information contained within the anchor profiles for some forms.

To investigate this further, we consider in this section an approach similar to that taken by Wasser et al. (2004). In this approach, the dialect fields and field sample locations are estimated separately, though still using our ℓ_1-Dirichlet prior on the dialect fields. This allows us to make inference about the dialect fields using only the anchor profiles, and to then use these fields to determine the locations for the floating profiles.

Thus, we remove the 247 floating profiles, and fit the ℓ_1-Dirichlet zero-inflated model to just the 120 anchor profiles (whose locations were determined using non-

linguistic means). Since these profiles are fixed in place, the model parameters to estimate are $\Theta = (\mu, \phi, \eta)$. Following modelling, locations for the floating profiles are estimated using Equation 2.14, with posterior means for μ, ϕ and η substituted in.

We again fit the ℓ_1-Dirichlet zero-inflated model to the full coarsened dataset. A large proportion (59%) of this data is not observed in the anchor profiles: that is, many forms are only used in the floating profiles. Therefore, we remove these forms, leaving us with with 298 forms of 60 items.

Figure 5.18 shows the resulting estimated mean posterior form usage probabilities $\hat{\pi}_\eta(\eta_{x,i,f})$ across the lattice for four coarsened forms of the item 'vps13'. These are noticeably different for the coarsened form '-ED' to those obtained when jointly modelling the dialect fields and unknown locations, as displayed in Figure 5.16. Encouragingly, the area of highest posterior usage probability for this form now coincides with the location of the cluster of southern anchor profiles displaying this form, rather than the solitary northern anchor which does.

Figure 5.19 shows the estimated expected posterior location probabilities for the same six floating profiles as Figure 5.15. We can see that whilst these still have high posterior weight in the north-west, the posterior distributions are now more diffuse, with some weight given to areas around the Atlas location for profile 68, for example. The apparent multimodality is unsurprising, representing the use of forms within the profile which lead to conflicting location estimates.

Figure 5.20 shows the posterior estimates for six more of the floating profiles. In the top panel, the results obtained from the earlier modelling with joint estimation of η and locations x_p are displayed. These exhibit the same north-western clustering

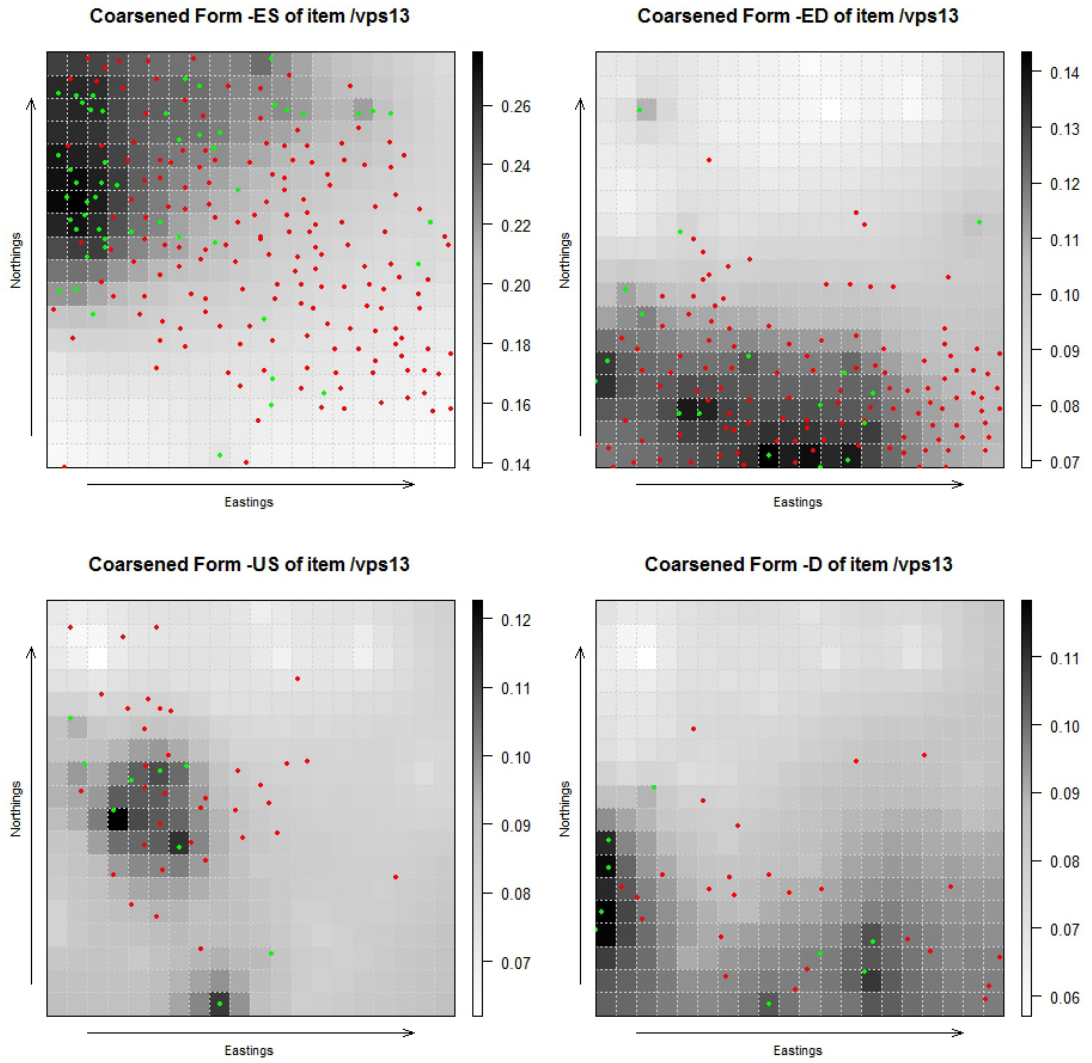


Figure 5.18: Estimated mean posterior form usage probabilities $\hat{\pi}_\eta(\eta_{x,i,f})$ across the lattice for four coarsened forms of the item 'vps13'. Like in Figure 5.16, $\hat{\pi}_\eta$ were estimated using the ℓ_1-Dirichlet zero-inflated model (fit to the full coarsened dataset), but this time η were not estimated jointly with the unknown locations, so these fields are based solely on the anchors. The colour bar gives the scale of $\hat{\pi}_\eta$ values. The green dots indicate the known locations of the Atlas profiles, whereas the red dots show the fit-technique locations for the floating profiles which contained these forms.

behaviour as the six previously shown. However, the results obtained from separate estimation of η and locations x_p (shown in the bottom panel) display marked improvement.

In Figure 5.21, we compare the Bayes Factors from the two modelling approaches. We can see that nearly all of the floating profiles have larger Bayes Factors when the locations are modelled separately from the η-fields. Whereas we would reject the Atlas locations for almost all of the floating profiles based on joint modelling, we observe a fewer number of rejections with separate η and location estimation.

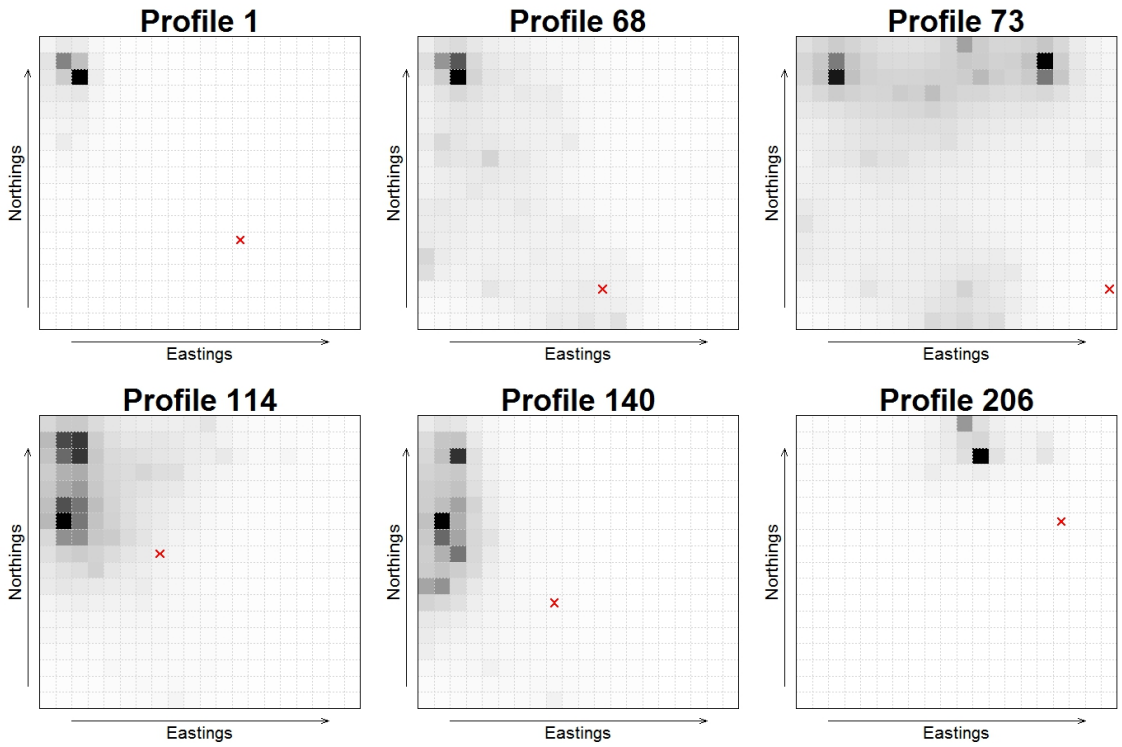


Figure 5.19: Estimated expected marginal posterior location probabilities $\hat{\pi}_x(x_q = x \mid y)$ (for cells $x = 1, \ldots, C$) are shown for the same six of 247 floating profiles as Figure 5.15. Again, these locations were estimated using the ℓ_1-Dirichlet zero-inflated model with all of the coarsened data, but this time the locations and form usage probabilities were not estimated jointly. Darker cells have higher $\hat{\pi}_x$ values. The locations provided in the Atlas for these profiles are marked in red.

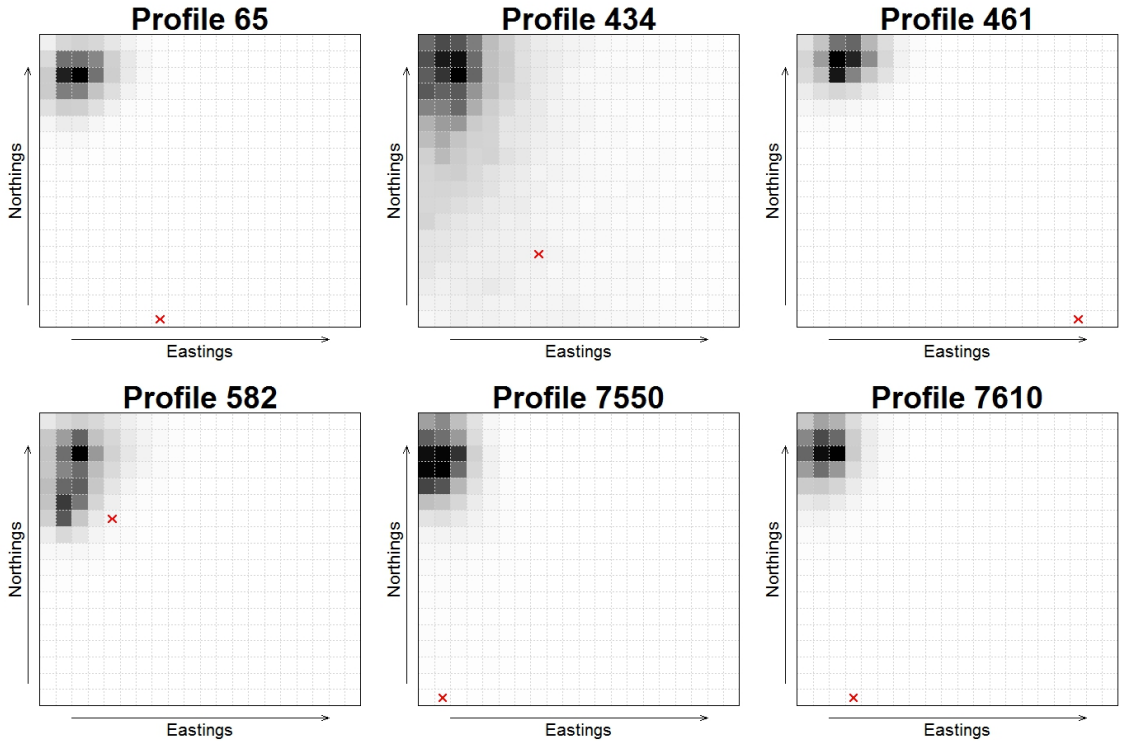
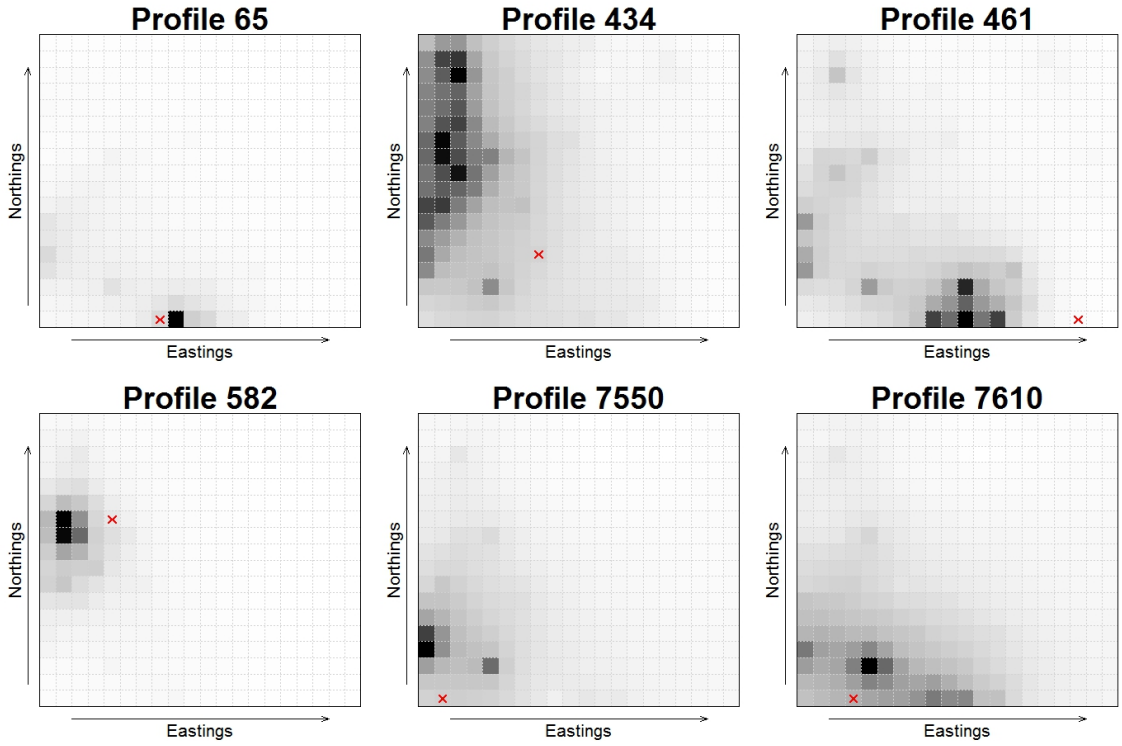


Figure 5.20: Estimated expected marginal posterior location probabilities $\hat{\pi}_x(x_q = x \mid y)$ (for cells $x = 1, \ldots, C$) are shown for six additional floating profiles. The locations were estimated using the ℓ_1-Dirichlet zero-inflated model with all of the coarsened data. Results in the top plot are obtained with the locations and form usage probabilities estimated jointly; the bottom plot show results with these estimated separately. Darker cells have higher $\hat{\pi}_x$ values. The locations provided in the Atlas for these profiles are marked in red.



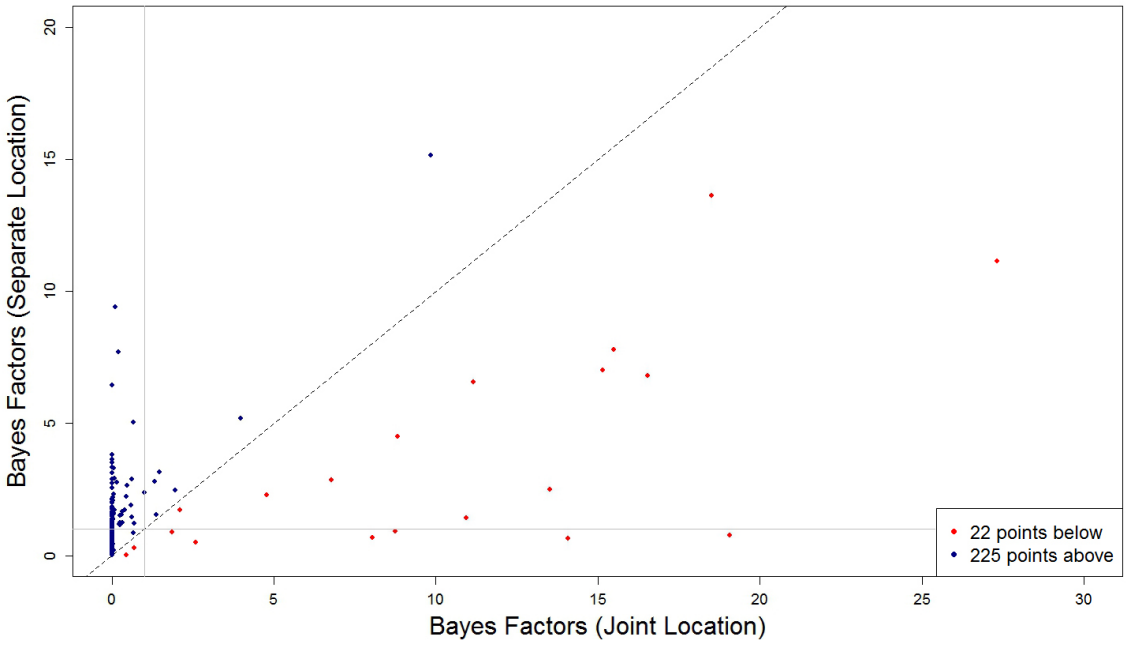


Figure 5.21: Bayes Factors for the 247 floating profiles from modelling with joint estimation of locations and η-fields, compared to those from modelling with separate location and η-field estimation. Bayes Factors are shown for model results based on the full coarsened dataset.

CHAPTER 6

Alternative Location Methods

In this chapter, we consider a variety of simple statistical methods for locating manuscripts. We do this to investigate whether a simple method exists which can successfully estimate the locations of origin of the manuscripts. If such a method does exist, we need not bother with the more complex and computationally expensive model-based methods presented throughout this report.

We present to the reader an overview of manuscript location using a selection of standard statistical methods, such as nearest-neighbour algorithms, classification trees, and multi-dimensional scaling. We do not attempt to fully resolve their potential, instead seeking to set a benchmark for our more complex model-based methods.

Simple statistical methods offer some advantages over our current model-based methods. As they are less computationally demanding than our model-based methods, they scale better, allowing us to more easily use a greater subset, if not all, of the linguistic data available. Further, with some of these methods, we are able to

easily estimate the location of origin of the manuscripts in coordinate-space, rather than the cell-location estimation our model-based methods currently offer.

Of course, such simple methods also have drawbacks. Like the 'fit-technique' used by McIntosh et al. (1986), they tend not to lead in a straightforward manner to quantitative measures of uncertainty for the estimated locations, which our model-based methods do offer. The provision of such measures is one of the primary concerns of this thesis; however, more simple methods can still provide a useful benchmark against which to assess the performance of our model-based methods. We therefore consider problems largely analogous to those tackled in Chapter 5, and compare the results obtained to those from our model-based methods. A summary outline of the analyses presented is provided in Table 6.1.

Though a nearest-neighbour approach shows some promise for providing a simple alternative to our model-based approach, we find the other methods totally inadequate. Such findings are unsurprising, given that earlier work by one of the Atlas authors (Benskin, 1988) found results from locating linguistic profiles using multi-dimensional scaling methods to give poor results. It was concluded that one of the reasons for this failure was that such methods fail to handle cases when forms have a multiply connected support. Our model-based methods do.

Section	Technique	Data	Items	Framework	Purpose
6.1.2	k-NN	P & C	8, 71	new	- k-NN any good? - More data = better results?
6.1.3	k-NN	P & C	8, 71	all	- Consider problem fit-technique was applied to.
6.1.4	k-NN	P & C	71	all	- Mimic fit-technique process. - Importance of sequence?
6.2	Classification trees	P & C	all	all	- Classification trees any good?
6.3	MDS	P & C	all	all	- MDS any good?

Table 6.1: For each analysis presented in this chapter, we list which statistical technique was used to analyse the data, as well as the type and quantity of data used. We also list the framework of the applied problem we worked within (locating *new* profiles, or locating *all* profiles whose locations were determined using the fit technique). Finally, we briefly describe the questions the analysis sought to answer.

We do not delve deeply into analysis using these simple methods. It is quite possible that we could improve the results obtained by investing more time into exploring variations of these methods. We do not proceed down this path, however; it would defeat the purpose of comparing our model-based methods to *simple* methods.

6.1 k-Nearest Neighbours

We begin this section with a brief explanation of location estimation using k-nearest neighbours, and provide the algorithm we use. We then perform a short series of simple analyses using this algorithm. We keep these analyses comparable to those

reported in Chapter 5, working within the same sub-region $\mathbb{W}$ of co-ordinate space.

We perform each analysis on both primary and coarsened data to assess the impact of our coarsening on the inference using k-nearest neighbours. As in Chapter 5, the results suggest this impact is minimal. Also, we initially perform analyses with the same eight-item data subsets described in Chapter 5, only later extending to larger datasets to see if our estimation improves with additional data. We find that this addition provides little improvement.

As in Chapter 5, we present analyses under both of the modelling frameworks defined on Page 3. We begin by locating a new set of profiles with the rest fixed to their Atlas estimates, and later present analyses where we locate all profiles whose origins were estimated to lie within $\mathbb{W}$ based on the fit-technique.

Finally, we present an additional analysis analogous to the fit-technique itself, whereby floating profiles are estimated sequentially and treated as anchors for the location of subsequent floating profiles. We do this to consider the impact of the sequence itself on the resulting location estimates, as well as the impact of sequentially fixing the floating profiles on location estimates. We find the results heavily dependent on the sequence chosen.

6.1.1 k-Nearest Neighbour Algorithm

To estimate the unknown true spatial location Λ_p for floating profiles $p \in \bar{A}$ using k-Nearest Neighbours (k-NN), we find the k profiles $\{p_1, p_2, \ldots, p_k\} \subset A$ to which p is

closest, based on some distance measure d. Our estimate for profile p's true spatial location Λ_p is then some combination of the locations of these k profiles $\{p_1, p_2, \ldots, p_k\}$. We use an exponentially-weighted average, such that

$$\hat{\Lambda}_p = \sum_{n=1}^{k} \left\{ \left(\frac{e^{-d(p,p_n)}}{\sum_{n=1}^{k} e^{-d(p,p_n)}} \right) \lambda_{p_n} \right\}, \tag{6.1}$$

where $d(p, p_n)$ is the distance between profiles p and p_n (recall, λ_p is the true coordinate location for anchor profile p). We use this because it is a robust measure, weighting our location estimates to be closest to the nearest-neighbours with smallest $d(p, p_n)$. We did also consider a simple average and an inverse-distance weighted average, but did not find the results to be particularly sensitive to this choice, so the results presented are based on this robust exponentially-weighted average.

Given the binary data y, we use the distance measure proposed by Jaccard (1901), where the distance d between any two profiles p_1 and p_2 is

$$d(p_1, p_2) = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}}, \tag{6.2}$$

where $M_{ab} = \sum_{i=1}^{I} \sum_{f=1}^{F_i} \{\mathbb{I}(y_{p_1,i,f} = a) \times \mathbb{I}(y_{p_2,i,f} = b)\}$. Thus, $d(p_1, p_2)$ is the number of disagreeing forms f in p_1 and p_2, divided by the number of forms used in either profile.

There are many other options for $d(p_1, p_2)$. One based on the simple matching coefficient used in Sokal & Michener (1958) is

$$d(p_1, p_2) = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11} + M_{00}}. \tag{6.3}$$

Another is the Euclidean distance

$$d(p_1, p_2) = \sqrt{M_{01} + M_{10}}. \tag{6.4}$$

We find that the location estimates $\hat{\Lambda}_p$ in the following analyses are not particularly sensitive to the choice of $d(p_1, p_2)$ from those listed above.

We use K-fold cross validation to choose the number k of nearest-neighbours to base our location estimates on. With $K = 10$, we:

1. Partition A into $K = 10$ equally-sized sets of anchors $A_1, A_2, \ldots, A_{10}$.

2. For each $k \in \{1, \ldots, k_{\max}\}$:

   - For each profile $p \in A_t$, for partition A_t with $t \in \{1, 2, \ldots, 10\}$:

     - Find the k nearest-neighbours to p in $\{A_1, A_2, \ldots, A_{10}\} \backslash A_t$.

     - Calculate $\hat{\Lambda}_p$ from Equation (6.1) using these k nearest-neighbours.

     - Calculate $\epsilon_{p,k,t} = \Lambda_p - \hat{\Lambda}_p$. The resulting vector $\epsilon_{p,k,t} = (\epsilon^{(e)}_{p,k,t}, \epsilon^{(n)}_{p,k,t})$ gives the error in the estimated location of origin for p in the eastings and northings directions respectively.

   - For each $t \in \{1, \ldots, 10\}$, calculate

$$\text{MSE}_{k,t} = \frac{\sum_{p=1}^{|A_t|} \sqrt{(\epsilon^{(e)}_{p,k,t})^2 + (\epsilon^{(n)}_{p,k,t})^2}}{|A_t|}, \tag{6.5}$$

     where $|A_t|$ is the number of profiles in A_t.

– Calculate the cross-validation error

$$\mathrm{CV}_k = \frac{\sum_{t=1}^{10} \mathrm{MSE}_{k,t}}{10}. \tag{6.6}$$

3. Repeat steps 1-2 N times, and choose the k with the smallest cross-validation error CV_k. Denote this k by k^*.

To measure the overall agreement between nearest-neighbour estimates $\hat{\Lambda}_p$ and Atlas location estimates λ_p in each of the following analyses, we calculate the average Euclidean distance between these estimates:

$$\bar{d} = \frac{\sum_{p \in \bar{A}} \sqrt{\left(\lambda_p^{(e)} - \hat{\Lambda}_p^{(e)}\right)^2 + \left(\lambda_p^{(n)} - \hat{\Lambda}_p^{(n)}\right)^2}}{|\bar{A}|}. \tag{6.7}$$

6.1.2 Locating 'New' Profiles

In this section, we present results from application of the k-Nearest Neighbours algorithm within the framework of locating new profiles. We estimate the unknown locations of origin of 123 linguistic profiles believed to lie within the sub-region $\mathbb{W}$. All other profiles within $\mathbb{W}$ are fixed to their Atlas locations. This means that, along with the 'true' anchors, 124 profiles whose locations were derived from the fit-technique are fixed in place as if they were anchors. This framing of the problem is the same as the one considered in Section 5.2.2.

We gather results using the eight-item data subsets $\mathbb{S}$ and $\mathbb{D}$ defined in Section 5.2.1, as well as larger datasets: namely, the entire coarsened dataset, the correspond-

ing items from the primary dataset, and then the entire primary dataset. We do this to consider the sensitivity of the k-NN location estimates to the dataset used.

Data	Items	Forms	k^*	$\bar{d}$
Coarsened	8	128	6	39.8 km
Coarsened	71	1182	2	41.1 km
Primary	8	873	6	43.2 km
Primary	71	2789	2	42.3 km
Primary	623	26908	2	41.6 km

Table 6.2: Average distance $\bar{d}$ between k-NN and Atlas location estimates for the floating profiles $p \in \bar{A}$, where k-NN results were based on different linguistic datasets. 244 profiles were fixed in place, and locations were estimated for 123 floating profiles.

Table 6.2 gives a summary of these results. We note from this table that the average distance between the Atlas and the nearest-neighbour location estimates are similar regardless of the data used, though slightly larger with primary data than with the equivalent coarsened data. Particularly, it does not appear that an increase in linguistic data has much of an impact on the results.

As a point of reference, were we to place the 123 profiles $p \in \bar{A}$ at random within $\mathbb{W}$, we would expect on average that $\bar{d} = 96.2$ km; and if we were to place each profile at the point in $\mathbb{W}$ furthest from its Atlas location, we would obtain $\bar{d} = 197.4$ km.

For the eight-item data subsets, Figure 6.1 shows the nearest-neighbour location estimates (in green) for six of the profiles $p \in \bar{A}$, overlaid onto the estimated marginal posterior probability plot displayed earlier in Figure 5.6. These posterior probabilities were obtained using the ℓ_1-Dirichlet zero-inflated model. The Atlas location estimates

for the six profiles are marked in red.

The nearest-neighbour location estimates for the profiles displayed generally coincide with the areas of high posterior probability from our model, with a few exceptions. There is some evidence, though, that the position of the nearest-neighbour estimates is sensitive to the choice of data. Indeed, with these data subsets, the average distance between the k-NN location estimates for each floating profile is 29.8 km - despite the fact that $\bar{d}$ differs by only 4.4 km.

6.1.3 Locating Ensembles of Profiles from 'True' Anchors

In this section, we present results from use of the k-Nearest Neighbours algorithm under the other framework presented on Page 3. In this framing of the problem, we estimate the unknown location of origin of all of the floating profiles. Only the 120 'true' anchor profiles (those whose origins were determined by non-linguistic means) are fixed in place. This means we are estimating locations for 247 profiles $p \in \bar{A}$. This framing of the problem is the same as the one considered in Section 5.3. We consider nearest-neighbour estimation with the same data as in the previous section.

Table 6.3 gives a summary of the results obtained. Again, results with the primary data are worse than with the equivalent coarsened data, and an increase in linguistic data does not appear to have much of an impact on the results, with the exception of moving from the eight-item primary data subset to a larger subset. As would be expected, the average accuracy in k-NN location estimates has decreased with the

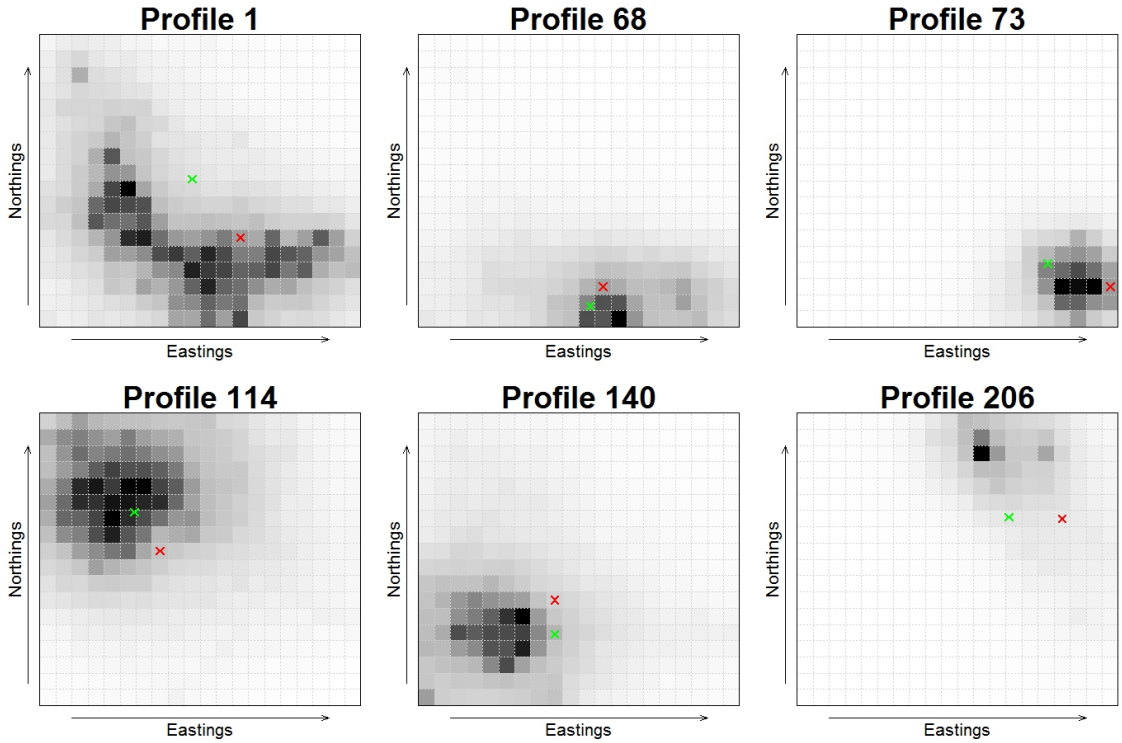
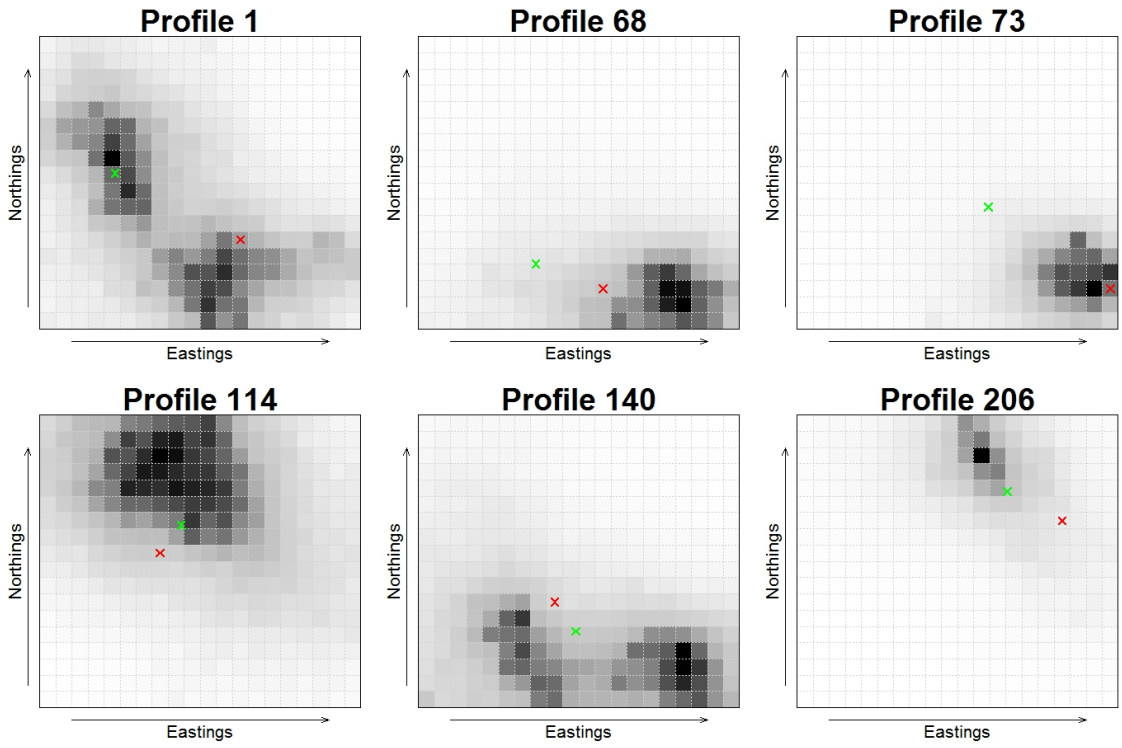


Figure 6.1: Estimated expected marginal posterior probabilities $\hat{\pi}_x$ for the locations of six profiles, obtained from the ℓ_1-Dirichlet zero-inflated model. Darker cells represent higher $\hat{\pi}_x$ values. The Atlas locations are marked in red and the nearest-neighbour estimates in green. Results using data subset $\mathbb{S}$ are displayed above and with data subset $\mathbb{D}$ below.



Data	Items	Forms	k^*	$\bar{d}$
Coarsened	8	128	3	57.2 km
Coarsened	71	1182	3	54.5 km
Primary	8	873	2	68.5 km
Primary	71	2789	2	58.0 km
Primary	623	26908	2	57.6 km

Table 6.3: Average distance $\bar{d}$ between k-NN and Atlas location estimates for the floating profiles $p \in \mathbb{W}$, from different linguistic data. The 120 true anchor profiles were fixed in place, and locations were estimated for all remaining 247 profiles.

decrease in the number of anchors.

As a point of reference, were we to place the 247 profiles $p \in \bar{A}$ at random within $\mathbb{W}$, we would expect on average that $\bar{d} = 96.9$ km; and if we were to place each profile at the point in $\mathbb{W}$ furthest from its Atlas location, we would obtain $\bar{d} = 199.6$ km.

Figure 6.2 provides a comparison of the results obtained using k-NN location estimation to those obtained earlier using our model-based methods. The nearest-neighbour location estimates (in green) for six of the profiles $p \in \bar{A}$ are overlaid onto the estimated marginal posterior probability plot shown earlier in Figure 5.15. These posterior probabilities and k-NN estimates were obtained using the ℓ_1-Dirichlet zero-inflated model with the entire coarsened dataset.

As described in the previous chapter, the model-predicted areas of high posterior probability for the floating profiles are all drawn to the large cluster of anchors in the north-west corner of $\mathbb{W}$, rendering them essentially useless for most profiles. The nearest-neighbour location estimates for the six profiles displayed are universally closer to the Atlas location estimates (marked in red).

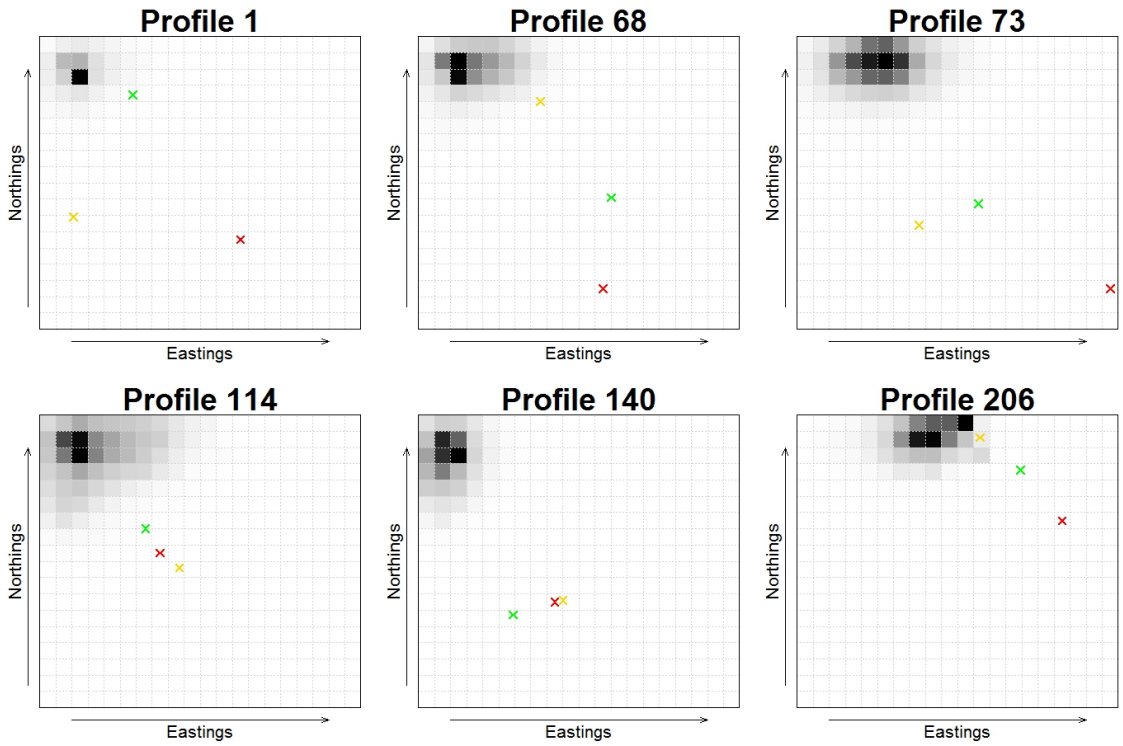


Figure 6.2: Estimated expected marginal posterior probabilities $\hat{\pi}_x$ for the locations of six profiles, obtained from the ℓ_1-Dirichlet zero-inflated model with only 'true' anchors fixed in place. Results are shown for (all of) the coarsened data. Darker cells represent higher $\hat{\pi}_x$ values. The Atlas locations are marked in red and the nearest-neighbour estimates in green. k-NN estimates obtained from the equivalent primary data are marked in gold.

However, by considering the k-NN location estimates obtained from the same model but with the dataset of corresponding primary data (marked in gold), we can see that even a k-NN approach has drawbacks with this framing of the problem. It is obvious that the results obtained with this approach are sensitive to the data used to derive them, and further, it is not clear how to get reliable quantitative measures of uncertainty in the estimates which capture this issue.

6.1.4 Locating Ensembles of Profiles using Sequential k-NN

The authors of the Atlas found their location estimates λ_p for floating profiles $p \in \bar{A}$ using the fit-technique: a sequential non-statistical method of elimination described in Section 1.3. As a final consideration of nearest-neighbour methods, we modify our k-NN algorithm to follow a similar process to the fit-technique.

We continue within the same framework as the previous section, whereby we estimate the unknown location of origin of all of the floating profiles. This time, though, we find our location estimates sequentially, with each profile p considered an anchor text after its location has been estimated.

We do this to assess the importance of the sequence itself on the resulting location estimates, as well as the impact of sequentially fixing the floating profiles on our estimates (as compared to fixing none in place, as in the previous analysis). Assessing the impact that the sequence has on the results may be a means to give insight into the impact it may have in the fit-technique. Indeed, we find that the results from sequential k-Nearest Neighbours are heavily dependent on the sequence used, and that there is little material advantage to locating profiles in this fashion.

We consider sequential location of the floating profiles in two fashions. In one, we choose the sequence in which we locate them randomly. We revisit this option later in the section. In the other, which we henceforth refer to as 'minimum-distance sequential k-NN', we locate the floating profiles $p \in \bar{A}_{\mathbb{W}}$ as follows:

1. For every $p \in \bar{A}_{\mathbb{W}}$, find the k nearest-neighbours to p in the anchors, based on our distance measure $d(p_1, p_2)$.

2. Find the $p \in \bar{A}_{\mathbb{W}}$ which minimises

$$\bar{d}_p = \frac{\sum_{a=1}^{k} d(p, p_a)}{k} \tag{6.8}$$

where $p_1, \ldots, p_k$ are the k nearest-neighbours in the relevant set of anchors (A or $A_{\mathbb{W}}$) to $p \in \bar{A}_{\mathbb{W}}$. Denoting this profile q_1, find $\hat{\Lambda}_{q_1}$ using Equation (6.1). We now treat q_1 as an anchor, letting $\bar{A}_{\mathbb{W}} \backslash \{q_1\}$ now be our set of floating profiles $\bar{A}_{\mathbb{W}}$, and $A_{\mathbb{W}} \cup \{q_1\}$ now be our set of anchor profiles $A_{\mathbb{W}}$. Repeat from Step (1) until all floating profiles $p \in \bar{A}_{\mathbb{W}}$ have been located.

In summary, we start by locating the profile closest to its k nearest neighbours, fix it in place, locate the profile next closest to its neighbours, fix it in place, and so on. Table 6.4 shows the results of applying this process. By comparison to Table 6.3, we can see that sequentially fixing the floating profiles in place in this fashion gives only a small boost in accuracy over fixing no profiles in place.

But what impact does the sequence in which we locate the floating profiles have? Given there are 247! possible sequences in which to locate the 247 profiles, it is not feasible to try them all. Instead, we perform k-NN location estimation for our floating profiles in 250 different random orderings, with the profiles fixed in place to their estimated location for our location estimation for subsequent profiles. The results for the best and worst sequence for each dataset are also displayed in Table

6.4. It is clear that the sequence chosen has a very large impact on the accuracy obtained, given the wide range of $\hat{d}$ obtained even when testing so few of the possible sequences,

Data	Items	Forms	k^*	$\bar{d}$	$\bar{d}_R$
Coarsened	8	128	3	51.1 km	(49.8, 61.3) km
Coarsened	71	1182	3	53.0 km	(47.9, 76.2) km
Primary	8	873	2	60.2 km	(56.6, 78.2) km
Primary	71	2789	2	56.4 km	(53.5, 85.2) km
Primary	623	26908	2	56.7 km	(52.3, 80.1) km

Table 6.4: Average distance $\bar{d}$ between k-NN and Atlas location estimates for all 247 floating profiles $p \in \mathbb{W}$, where k-NN results were based on different linguistic data, and location was done via minimum-distance sequential k-NN. $\bar{d}_R$ provides the range of $\bar{d}$ obtained when instead using sequential k-NN with random ordering.

6.2 Classification Trees

In this section, we consider estimating the cell-location of origin x_p (within a lattice over $\mathbb{W}$) of profiles $p \in \bar{A}_{\mathbb{W}}$ using classification trees. We use 10-fold cross-validation to determine the complexity parameter for cost-complexity pruning.

We consider classification problems within $\mathbb{W}$ with both the primary data $\mathbb{P}$ and coarsened data $\mathbb{C}$. We work with cells in a lattice over $\mathbb{W}$ as our classes. We find that even with very few classes, the classification trees perform dismally.

We initially work with just four classes corresponding to the four cells of a 2×2 lattice over $\mathbb{W}$. We train the tree on the anchors $A_{\mathbb{W}}$ using the data $\mathbb{C}$. The cross-validation error is smallest for a tree with 31 nodes.

Of the 1182 forms in the dataset, 27 are used in this tree's construction. The tree has a 96.67% classification success rate for the anchors, with 116 out of 120 profiles correctly classified to their true cell-locations $\xi_p = J(\Lambda_p)$.

Having trained the tree on the anchors, we then use it to predict cell-locations for the floating profiles $p \in \bar{A}_{\mathbb{W}}$. 131 of these 247 profiles (53%) are classified to the cell $J(\lambda_p)$ which contains their Atlas location estimate λ_p. Although we are uncertain of the exact level of accuracy in the Atlas estimates, given how few classes (cells) we are working with, we would expect far more of the profiles $p \in \bar{A}_{\mathbb{W}}$ to be classified to $J(\lambda_p)$.

We repeat the exercise with nine classes corresponding to cells in a 3×3 lattice. The tree with smallest cross-validation error has 4 nodes, and in training, classifies only 50% of the anchors correctly. Predictions for the floating profiles are terrible, with 48 of the 247 profiles (19.4%) classified to $J(\lambda_p)$.

The classification trees perform poorly with $\mathbb{C}$. There is little point reducing the size of the cells further.

We then consider the same methods applied to the primary data $\mathbb{P}$. Due to memory limitations, we are only able to run the classification routine with around 15,000 forms f, so we take random samples from the 26,908 available and compare the results obtained.

With four classes (cells of a 2×2 lattice), we find that the results vary depending

on which forms are included. The worst tree classified 50% of the anchors correctly in training, and then only 19.4% of the floating profiles $p \in \bar{A}_{\mathbb{W}}$ to $J(\lambda_p)$. The best tree classified all of the anchors correctly in training, and 50.2% of floating profiles $p \in \bar{A}_{\mathbb{W}}$ to $J(\lambda_p)$.

Typically though, $70 - 100\%$ of the anchors were classified successfully, and $30 - 50\%$ of the floating profiles $p \in \bar{A}_{\mathbb{W}}$ were classified to $J(\lambda_p)$. This is poor, given the tiny number of classes (cells), and given we performed better for the 2×2 lattice with the coarsened data, there is nothing to be gained from reducing the size of the cells and continuing with the primary data.

6.3 Multidimensional Scaling

Multidimensional scaling (MDS) refers to a set of methods which seek to capture underlying dimensions that explain similarities and dissimilarities between objects. From a matrix of dissimilarities between the objects, MDS algorithms produce a spatial configuration in N-dimensions. Each object is assigned coordinates in the N-dimensional space such that similar objects are placed close together, and dissimilar objects further apart. Thus, unlike our model-based methods, with MDS we do not get location estimates x_p on the geographical map, nor formal measures of uncertainty in the estimates. Rather, MDS offers a way to test the validity of a provided map.

We use MDS in two dimensions with our linguistic profiles, assessing whether profiles close to each other in the MDS representation have similar Atlas location

estimates. That is, can MDS recover the underlying spatial structure in the data?

We consider two problems. Using MDS with only the anchor profiles $p \in A_{\mathbb{W}}$, we first assess whether MDS can capture their known spatial structure. Finally, we use MDS with profiles p with Atlas location estimates $\lambda_p \in \mathbb{W}$, to see whether the MDS representation resembles the spatial structure provided in the Atlas for these profiles.

With each of these two problems, we use for comparability the subsets $\mathbb{D}$ and $\mathbb{S}$ of the primary and coarsened data used throughout Chapter 5, as well as in Section 6.1.2. As we did with the nearest-neighbour methods, we then extend beyond using only a subset of items, instead using all of the items available and assessing the impact on our inference.

We were unable to reproduce anything resembling the spatial configuration provided in the Atlas for even just the anchor profiles, regardless of the choice of MDS algorithm, distance measure, or data.

6.3.1 Recovering County Structure using MDS

To allow us to visually identify whether MDS captures underlying spatial structure, we label profiles by the counties they belong in, as in Figures 6.3 and 6.4. We look to see whether this cluster structure is present in the MDS representation: are profiles from the same county still close to each other, and are neighbouring counties close together?

We consider two different MDS methods: Sammon's mapping (Sammon, 1969), and classical multidimensional scaling (Torgerson, 1952). Sammon's mapping is a form of non-metric MDS, and iteratively finds the configuration which minimises the sum of squared differences between the input and output distances (the stress). Classical (metric) MDS produces an analytic solution, through an eigenvalue decomposition, which represents the similarities between objects in Euclidean space.

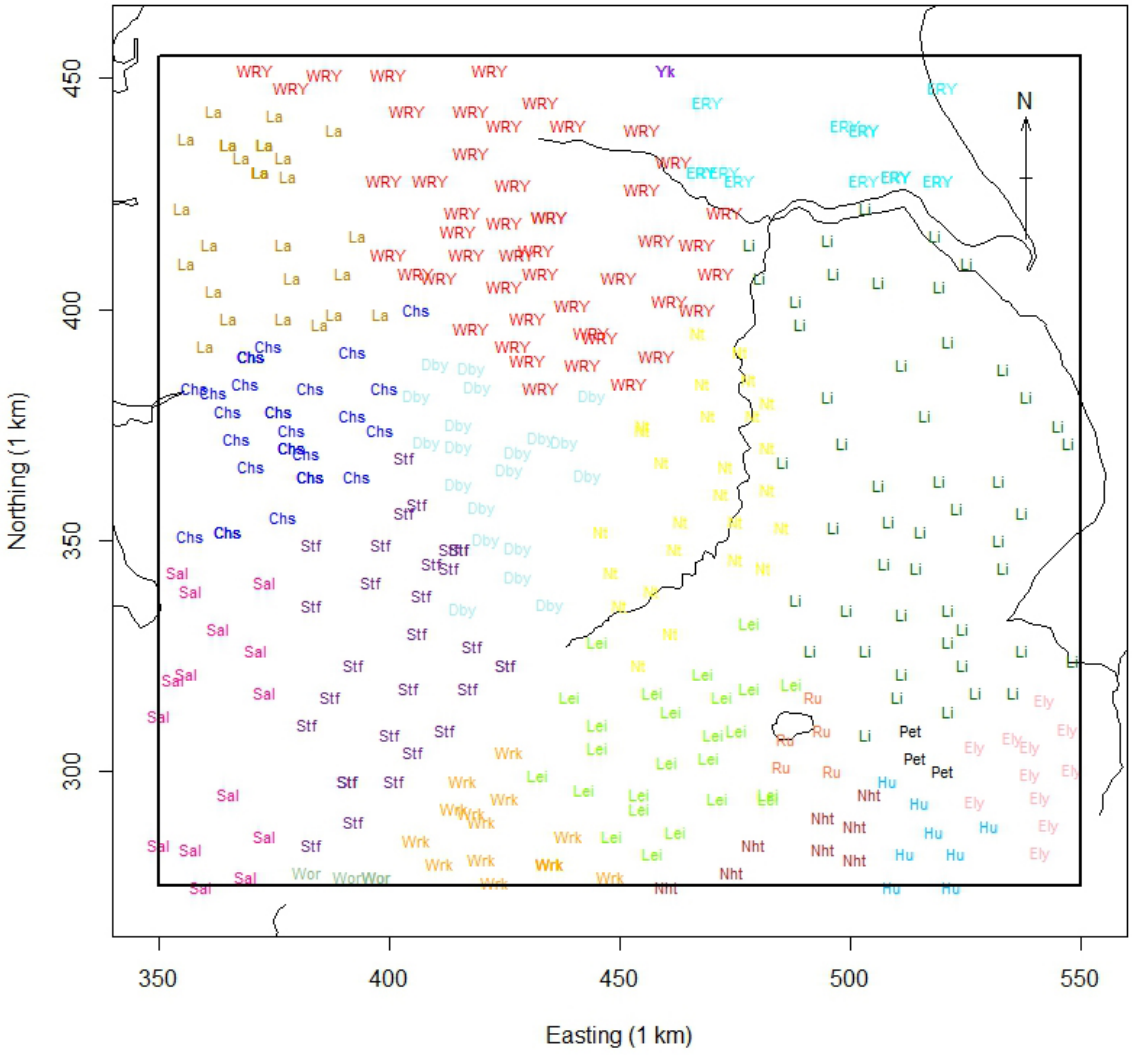


Figure 6.3: County information for both floating and anchor profiles. All Atlas location estimates λ_p are displayed, with each profile labelled by the county it belongs in. Profiles from the same county are plotted in the same colour.

With each algorithm, we consider Jaccard's distance as defined in Equation (6.2), and the Euclidean distance as defined in Equation (6.4), for use as our input distance measure $d(p_1, p_2)$ between profiles p_1 and p_2.

We found that regardless of the choice of MDS method, distance measure, or data, the MDS configuration did not come close to reflecting the true spatial configuration of the anchors $A_{\mathbb{W}}$. Figure 6.5 shows one of the 'best' configurations, obtained from non-metric MDS with all of the primary data.

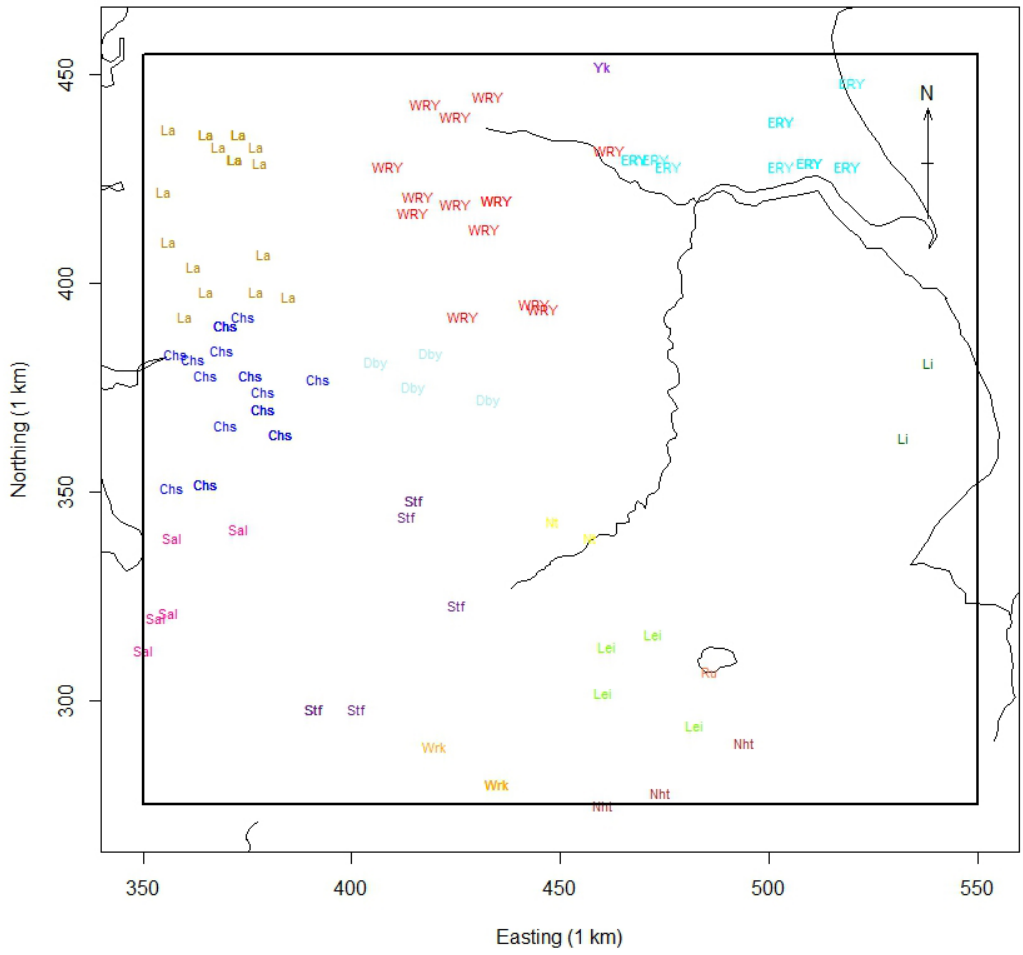


Figure 6.4: County information for anchor profiles. The true locations of origin Λ_p for $p \in A_{\mathbb{W}}$ are displayed, with each profile labelled by the county it belongs in. Profiles from the same county are plotted in the same colour (consistently with Figure 6.3).

We can see by comparison to Figure 6.4 that the MDS configuration is still a fairly poor reflection of the true structure, with a lack of clear clusters of profiles by county. Further, although the x-axis appears to capture east-west differences with some degree of success, the y-axis does not similarly capture north-south differences.

Similarly, we were unable to reproduce the spatial configuration provided in the Atlas for all the profiles p in $\mathbb{W}$, regardless of the choice of MDS method, distance measure, or data.

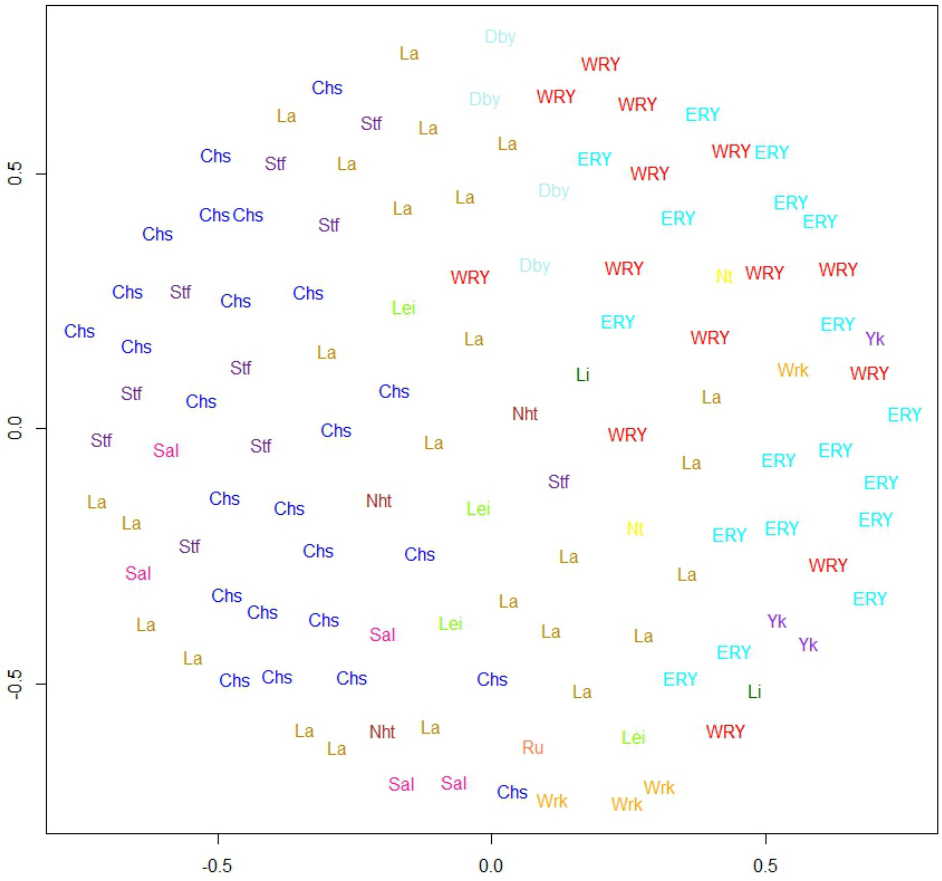


Figure 6.5: The two-dimensional configuration produced by Sammon's mapping for the anchor profiles $p \in A_{\mathbb{W}}$, using all of the primary data and Jaccard's distance to form the input distance matrix. Stress converged to 0.154.

CHAPTER 7

Conclusions

In this thesis, we made a simultaneous reconstruction of interpolating spatial fields and measurement locations, which sets this work apart from other approaches taken to similar problems, and gives a full quantification of uncertainty. The written dialect data we worked with is large, but very sparse, since a given word has a large number of variant spellings which may appear in just a few documents.

We aimed to meet the computational and inferential challenge of simultaneously locating large numbers of documents with this data. We also aimed to provide the linguists who collected the data reliable statistical tools under two modelling frameworks: to locate all of the non-anchor texts, or to locate a new text(s). The results presented in this thesis show that we have largely succeeded in meeting these challenges.

We begin this chapter by summarising areas highlighted throughout this thesis where changes to our methodology could be advantageous, and give suggestions of

how changes may be achieved. We then conclude this thesis by presenting our recommendations to potential users of our methods.

7.1 Further Work

7.1.1 Modelling Region Expansion

In Section 5.1, we described the rectangular sub-region $\mathbb{W}$ of England to which we restricted attention in this thesis. Based on the Atlas location estimates, this region contained 367 of the 1511 profiles, a relatively high proportion of which were anchors (120 of these 367 profiles).

A natural extension of the work presented in the thesis would be to expand the scope of analysis to a larger region of England, if not the whole country. As shown by Figure 1.2, the majority of the anchor profiles are located in the northern half of the country, so expansion to the entire country would present similar challenges to those faced in Section 5.3.

Of course, in other settings, such as that of Wasser et al. (2004), it is possible to engineer a better distribution of 'anchors' through good experimental design. The context of our application means that the number of anchor profiles is static and out of our control. However, expansion of the modelling region on a smaller scale than to the entire country, if chosen wisely, may overcome these challenges and improve the modelling results under the more difficult framework.

7.1.2 Further Investigation of the Outlier Model

In Chapter 5, we compared the results obtained when using the two ℓ_1-Dirichlet models to locate new profiles. Similar results were obtained for the two models on both of the eight-item data subsets of the primary and coarsened data. We therefore proceeded to use the ℓ_1-Dirichlet zero-inflation model when attempting to locate all of the floating profiles using the entire coarsened dataset. This enabled us to offset the additional computational challenge faced.

However, the expected strength of the ℓ_1-Dirichlet outlier model is for handling noisy and large subsets of the linguistic data. Exploring the potential benefits of this model over its zero-inflated counterpart in this setting would be worth further investigation in future work.

7.1.3 Spatial Smoothing Edge-Effects

In Section 5.2.2, we noted the apparent edge-effects present in the estimated posterior distribution for the form usage probabilities η. It was noticeable that the probabilities in some individual border cells of the lattice were inflated, due to inhomogeneities in smoothing due to the free boundary condition. Future work could address this issue by adjusting the coupling constants θ_i such that the spatial smoothing is stronger on the boundary.

7.1.4 Hierarchical Modelling to Exploit Form Structure

In Section 5.2.2, we discussed the impact our data coarsening operation had on the inference for the locations of the floating profiles. We noted that on one hand, one might expect the coarsened data to lead to more diffuse posterior distributions, given we have 'less' information. On the other hand, by coarsening the data, we may actually gain information by adding details about the group structure of the forms.

In this thesis, we viewed the situation as a choice between the coarsened and primary datasets. However, a fully informed model would feature both fine-scale (primary data) and form-hierarchy (coarsened data), and take advantage of the benefits of each. Future work could develop a hierarchical model with a generative process whereby the scribe of a document chooses a coarsened form first, and then chooses from the fine-scale forms which match the chosen coarsened form.

7.1.5 Exploiting Indicative Frequency Data

In this thesis, we reduced the indicative frequency data $d_{p,i,f}$ from McIntosh et al. (1986) (described in Section 1.3.1) to binary form $y_{p,i,f}$. The justification for this was the lack of a consistent standard across the entire sampling window for the recording of the indicative frequencies.

One avenue for future work could be to use these frequencies to impose an order on the form-usage probabilities, since an indicative frequency for one form being

greater than the frequency for another ($d_{p,i,f_1} > d_{p,i,f_2}$) implies that the form-usage probability should be higher ($\eta_{x_p,i,f_1} > \eta_{x_p,i,f_2}$).

Although the recording of the indicative frequencies was not done consistently across the entire sampling window, it was done consistently within sub-regions of this window. Another avenue to explore could therefore be to restrict attention to this sub-region, and model the indicative frequencies directly.

7.1.6 Returning to the Continuum

Estimation of the locations of origin of the linguistic profiles has in this report been restricted to membership of cells within the lattice, rather than to spatial coordinates, as were provided in the Atlas. This is for convenience of modelling the η-fields.

It would be relatively straightforward in future work to extend the model to allow the reconstructed location x_p to be a spatial location rather than a cell indicator. For example, one could use a local value for η interpolated to x_p by smoothing from the surrounding cells. Another option for future work would be to consider a hierarchical model, whereby we estimate η-fields for larger cells, and then split these into smaller ones at a second level of the model.

7.1.7 Feature Selection

In Section 1.2, we briefly reviewed related work which explores the concept of selecting location-informative features from a dataset. For example, Han et al. (2014) consider

a variety of methods to select 'location-indicative words', including the compactness of the geographical spread of a word.

In a similar direction, future work could develop our methods to automatically choose 'optimal' subsets of the linguistic data (i.e. the items/forms which are most informative of location). It is conceivable that removing forms, for example those that are very common or extremely rare, may aid in the location of the floating profiles.

7.2 Recommendations

7.2.1 Estimating the Unknown Location of New Samples

Of the two applied problems considered in thesis, the core problem of interest to the scientists was the estimation of the location of a new linguistic profile, and provision of a confidence region for this estimate. This was the application to which our methods were most successful. We also tested a range of standard statistical techniques for this purpose in Chapter 6, finding multidimensional scaling and decision-tree based approaches totally inadequate. Encouragingly, locations estimated using a k-Nearest Neighbours algorithm under this framework of the problem were in broad agreement with those obtained by our models. It is unclear how to get reliable quantification in the uncertainty of the k-NN estimates, meaning our model-based approach offers clear advantages over k-NN.

In Chapter 5, we showed that our methods were able to locate a large number

of new documents simultaneously with a very large number of dialect fields. Our ℓ_1-Dirichlet models used a spatial smoothing prior distribution featuring an ℓ_1-penalty term. Whilst not widely used in this context, this penalty performed well, as evidenced by the Goodness of Fit analysis, which was predictive in character and quantified using Bayes factors. It is likely that this penalty performs a dimensionality reduction and allows us to manage the very large number of form usage probability parameters to be estimated. Our ℓ_2-Dirichlet models allowed for more efficient MCMC updates, due to the absence of the parameter constraints present with the ℓ_1-Dirichlet models, but did not perform as well.

We would therefore recommend using our ℓ_1-Dirichlet zero-inflated model in future implementations of this type of applied problem.

7.2.2 Estimating Many Unknown Locations

Although locating new profiles is a difficult task, given the sparse data and massive parametrisation of the dialect fields, locating all of the non-anchors is a much more challenging problem. Contributing to this difficulty, we lack knowledge of the accuracy of the Atlas' fit-technique estimates to which we compare our results. Disagreement between the Atlas locations and our estimates can arise because our estimates are incorrect, but can also be because the Atlas locations are inaccurate, or are points within a location distribution which overlaps ours substantially.

The results of our modelling under this framework were less successful. In Chapter

5, we noted the quality of our results was related to the uneven geographical spread of the anchor profiles, along with the dialect-field feedback from the floating profiles. Better results were obtained using k-Nearest Neighbours, but for the moment, the best approach to this full problem is that taken by Wasser et al. (2004).

We thus recommend using our version of this methodology, which similarly conditions on dialect fields estimated from anchor profiles (cutting off feedback from non-anchors), but uses our ℓ_1-Dirichlet prior and is otherwise suitably modified and scaled for this problem. We would also recommend those wishing to develop our research to seek further understanding of failures of our full method with large problems.

appendix A

Further Checking of the MCMC Samplers

In Section 3.4, we performed peace-of-mind checks for the samplers from the algorithm for the ℓ_1-Dirichlet zero-inflated model. In this Appendix, we provide similar checks for the sampling methods from the algorithms for the ℓ_1-Dirichlet outlier and ℓ_2-logistic outlier models.

As before, using synthetic data, we compare theoretical distributions for the parameters of the models to those found using the samplers from the algorithms. These quick checks are not intended as rigorous proof that the algorithms sample from the correct distributions, rather, to provide some peace-of-mind. We find no cause for alarm with either algorithm. We do not provide these checks for the ℓ_2-logistic zero-inflated model in this thesis, but note that we obtain similarly unconcerning results by performing such checks.

A.1 ℓ_1-Dirichlet Outlier Model

The ℓ_1-Dirichlet outlier model, as specified in Section 2.2, has parameters $\Theta = (x, \mu, \eta, \phi, \psi)$. We set x, μ, η and ϕ as we did in Section 3.4 for the ℓ_1-Dirichlet zero-inflated model, and choose:

- **Outlier probabilities** $\psi_i^{(I)} = (0.1,\ 0.1,\ 0.1,\ 0.1,\ 0.5,\ 0.1,\ 0.1,\ 0.1)$, so that aberrant forms are more likely to be used in one randomly selected item than all the others.

- **Outlier probabilities** $\psi_p^{(P)} = 0.5$ for randomly selected profiles $p \in \{4, 5, 22, 26, 39\}$, and $\psi_p^{(P)} = 0.2$ for all other p. Thus, aberrant forms are more likely to be used in a small set of profiles than in all the others.

These parameter values are assigned, not simulated. Using them, we simulate data y for the profiles $p \in \{1, \ldots, 50\}$ using the generative process for this model.

A.1.1 Form Usage Probabilities

We test the η-field sampler from the MCMC algorithm for the outlier model in the same fashion as earlier with the ℓ_1-Dirichlet zero-inflated model, estimating the posterior distributions for the same eight randomly selected $\eta_{x,i}$ vectors, with all other parameters fixed to their assigned values.

Figure A.1 shows the exact posterior distribution for $\eta_{11,5}$, alongside summaries of the MCMC samples obtained from our algorithm, where the numbers of MCMC

samples taken is varied. We can see that the estimated posterior converges (slowly) to the exact posterior distribution.

Further evidence for convergence to the desired posterior distribution is provided in Figure A.2, which shows the estimated marginal distributions for the components of $\eta_{11,5}$. The exact marginal posterior distributions are overlaid, and we note that the estimated marginals for all η parameters match the exact distributions. Similar figures can be obtained for the other η-fields, but for brevity are not included here.

A.1.2 Item Usage Rates

We test the sampler for μ from the MCMC algorithm for the ℓ_1-Dirichlet outlier model in the same fashion as earlier with the ℓ_1-Dirichlet zero-inflated model, using it to estimate the posterior distributions for μ_i for $i \in \{1, \ldots, 8\}$ given our simulated data y, and with η, ϕ, ψ and x_p fixed to their assigned values.

We compare the estimated posterior distributions to the exact posterior distributions for μ_i, with the normalising constant Z_{μ_i} for Equation (2.15) again computed using a quadrature approach, and with the hyper-parameters of the prior specified in Equation (2.10) again set to $a_i = 3$ and $b_i = 1$ for all items $i \in \{1, \ldots, I\}$.

The estimated posterior distributions for selected μ_i are shown in Figure A.3. It can be seen that the estimated posterior matches the exact posterior (overlaid in red). The posterior distributions for other μ_i display similar patterns.

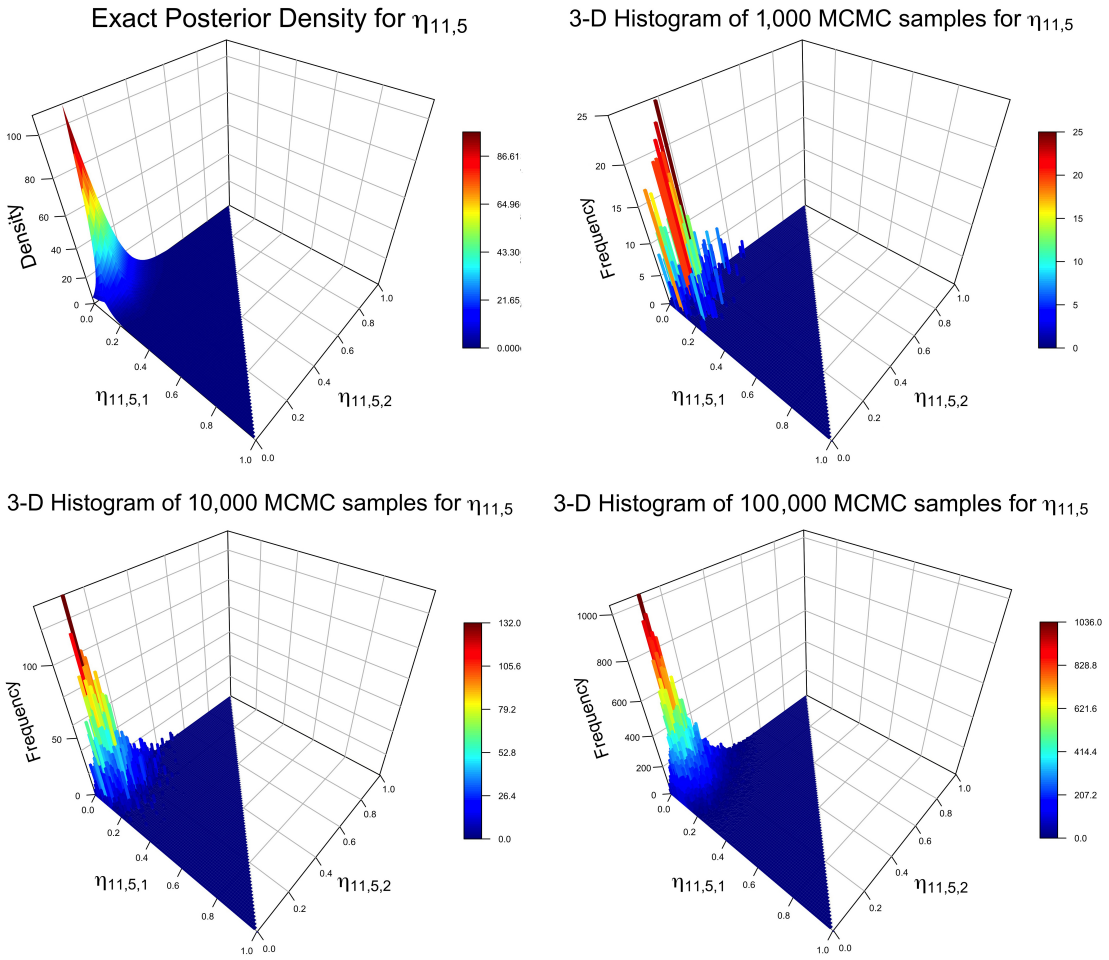


Figure A.1: The exact posterior distribution for $\eta_{11,5}$ (i.e. the form usage probability vector for item $i = 5$ in cell $x = 11$) is displayed in the upper left. 3-D histograms for $\eta_{11,5}$ are displayed after 1000 (top-right), 10000 (bottom left) and 100000 (bottom right) MCMC samples.

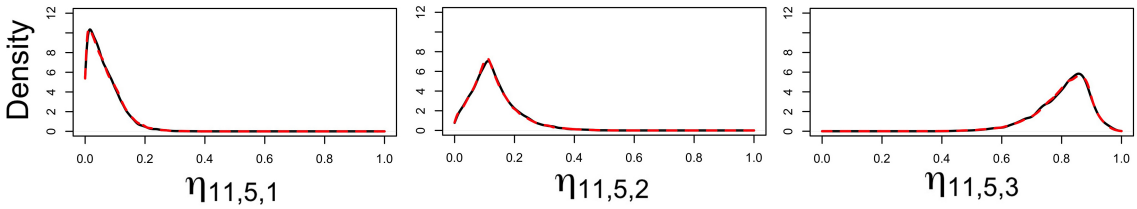


Figure A.2: Marginal posterior distributions for form usage probabilities $\eta_{11,5,1}$ (left), $\eta_{11,5,2}$ (middle) and $\eta_{11,5,3}$ (right) based on 100,000 MCMC samples. The exact marginal posterior distributions are overlaid in red.

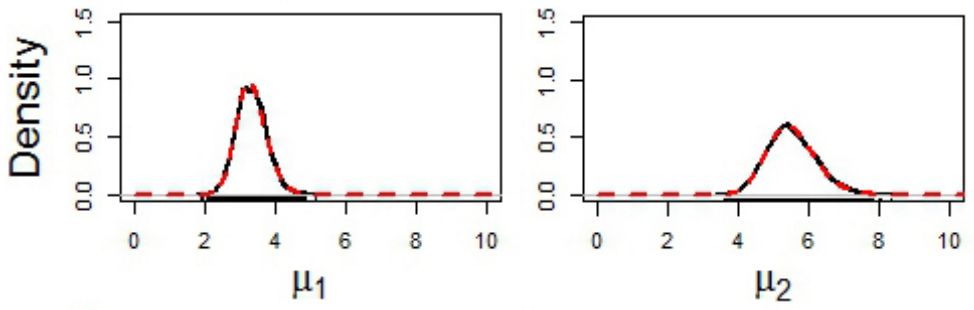


Figure A.3: Estimated posterior distributions for item usage rates μ_1 and μ_2, based on 10000 MCMC samples. The exact posterior distribution is overlaid in red for each.

A.1.3 Zero-Inflation Probabilities

We test the sampler for ϕ from the MCMC algorithm for the ℓ_1-Dirichlet outlier model in the same fashion as earlier with the ℓ_1-Dirichlet zero-inflated model, using it to estimate the posterior distributions for ϕ_i for $i \in \{1, \ldots, 8\}$, given our simulated data y, and with η, μ, ψ and x_p fixed to their assigned values.

We compare the estimated posterior distributions to the exact posterior distributions for ϕ_i, with the normalising constant Z_{ϕ_i} for Equation (2.17) again computed using a quadrature approach.

The estimated posterior distributions for selected ϕ_i are shown in Figure A.4. It can be seen that the estimated posterior matches the exact posterior (overlaid in red). The posterior distributions for other ϕ_i display similar patterns.

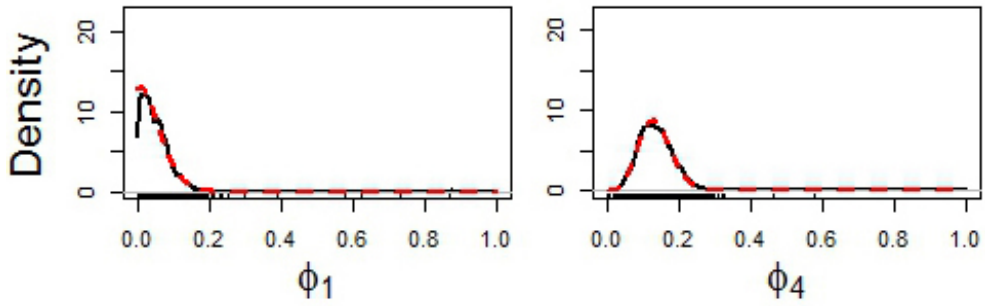


Figure A.4: Estimated posterior distributions for zero-inflation probabilities ϕ_1 and ϕ_4, based on 10000 MCMC samples. The exact posterior distribution is overlaid in red for each.

A.1.4 Outlier Probabilities

To test the sampler for $\psi_i^{(I)}$ from the MCMC algorithm for the ℓ_1-Dirichlet outlier model, we use it to estimate the posterior distributions for $\psi_i^{(I)}$ for $i \in \{1, \ldots, 8\}$, given our simulated data y, and with η, μ, ϕ, $\psi_p^{(P)}$ and x_p fixed to their assigned values.

We compare the estimated posterior distributions to the exact posterior distributions for $\psi_i^{(I)}$, which can be computed analytically in this case. To derive the analytic expression for the posterior distribution for $\psi_i^{(I)}$, the normalising constant $Z_{\psi_i^{(I)}}$ for Equation (2.23) must be computed. We find this using a quadrature approach, with the hyper-parameters of the prior given by Equation (2.21) set to $(\sigma_1, \sigma_2) = (2, 3)$.

The estimated posterior distributions for selected $\psi_i^{(I)}$ are shown in Figure A.5. It can be seen that the estimated posterior matches the exact posterior (overlaid in red). The posterior distributions for other $\psi_i^{(I)}$ display similar patterns.

We test the sampler for $\psi_p^{(P)}$ similarly. The estimated posterior distributions for selected $\psi_p^{(P)}$ are shown in Figure A.6. Again, the estimated posterior matches to the exact posterior (overlaid in red). The posterior distributions for other $\psi_p^{(P)}$ display similar patterns.

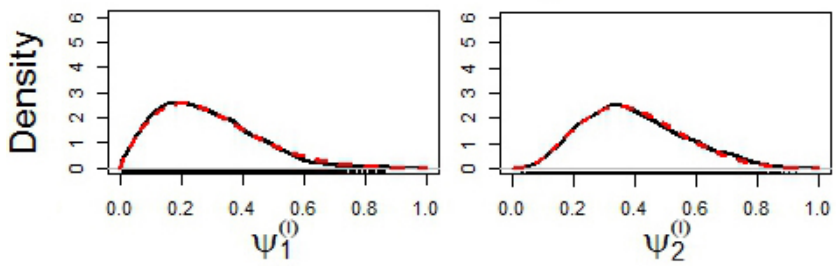


Figure A.5: Estimated posterior distributions for outlier-probability $\psi_i^{(I)}$ for items $i \in \{1, 2\}$, based on 10000 MCMC samples. The exact posterior is overlaid in red for each.

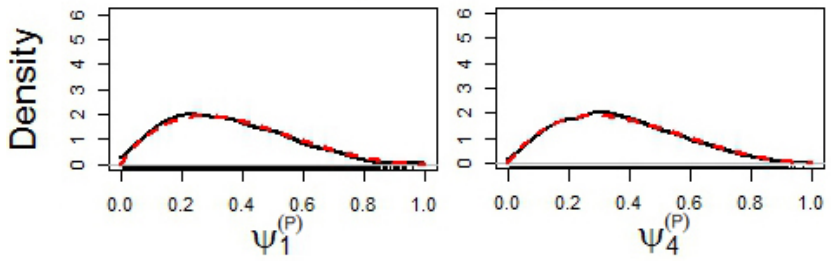


Figure A.6: Estimated posterior distributions for outlier-probability $\psi_p^{(P)}$ for profiles $p \in \{1, 4\}$, based on 10000 MCMC samples. The exact posterior is overlaid in red for each.

A.2 ℓ_2-Logistic Outlier Model

We now perform similar checks for each of the sampling methods from the algorithm described in Section 3.3 for the ℓ_2-logistic outlier model. This model, as specified in Section 2.3, has parameters $\Theta = (x, \mu, \gamma, \phi, \psi)$. We select the same values for the parameters $\{\Theta \backslash \gamma\}$ as we did in Section A.1, and set γ to the values leading to the chosen η fields. Using these parameter values, we simulate data y for profiles

$p \in \{1, \ldots, 50\}$ using the generative process for this model.

A.2.1 γ-Field Parameters

To test the γ-field sampler from the MCMC algorithm for the ℓ_2-logistic outlier model, we use the sampler to estimate the posterior distributions for the eight $\gamma_{x,i}$ vectors (one for each item) corresponding to the eight $\eta_{x,i}$ vectors randomly selected earlier. These posterior distributions are estimated with all other parameters (x, μ, ϕ, ψ and all other $\gamma_{x,i}$) fixed to their assigned values. We then compare the estimated posterior distributions to the true posterior distributions for the selected $\gamma_{x,i}$ vectors.

Figure A.7 shows that the estimated posterior for $\gamma_{11,5}$ converges (slowly) to the exact posterior distribution as the number of MCMC samples taken increases. Further evidence for convergence to the desired posterior distribution is provided in Figure A.8. We observe that the estimated marginal posterior distributions for all γ parameters match the exact distributions.

A.2.2 Form Usage Probabilities

When using the ℓ_2-logistic outlier model, we model γ-fields rather than η-fields. However, the η-fields can easily be derived from the γ-fields using Equation (2.26). Figure A.9 provides the equivalent plots to Figures A.7, following transformation into η-space. We can see that the estimated posterior matches the exact posterior.

A.2.3 Other Model Parameters

With the ℓ_2-logistic outlier model, the conditional posterior distributions for μ, ϕ, and ψ remain unchanged from those for the outlier model. The relevant samplers in the respective algorithms are also unchanged. Therefore, the results of sampler checks for those parameters will be the same as in Section A.1.

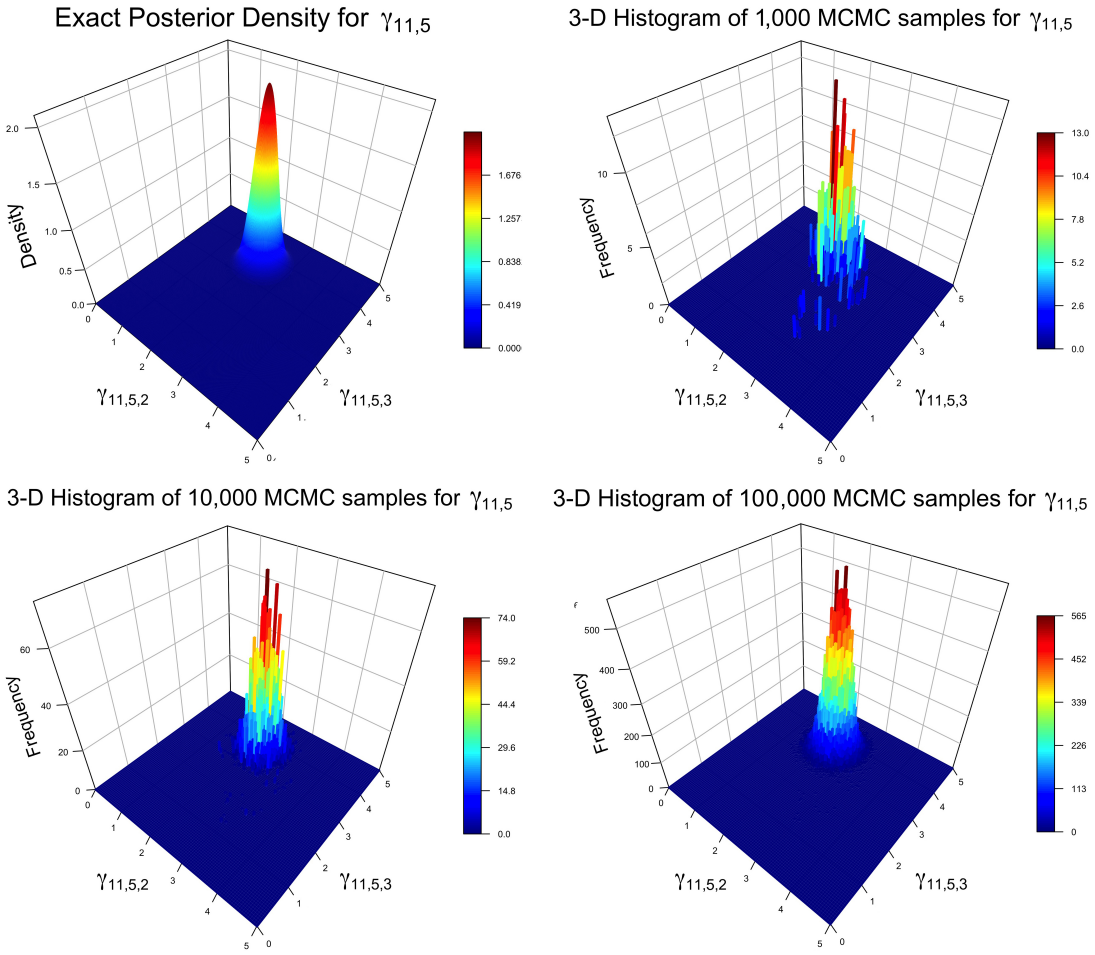


Figure A.7: The exact posterior distribution for $\gamma_{11,5}$ (i.e. the transformed form-usage probability vector for item $i = 5$ in cell $x = 11$) is displayed in the top left. 3-D histograms for $\gamma_{11,5}$ are displayed after 1000 (top right), 10000 (bottom left) and 100000 (bottom right) MCMC samples.

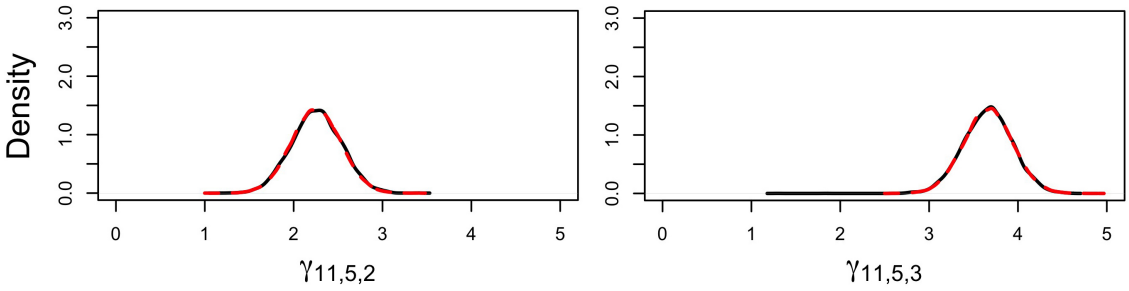


Figure A.8: Estimated marginal posterior distributions for transformed form-usage probabilities $\gamma_{11,5,2}$ (left) and $\gamma_{11,5,3}$ (right) based on 100,000 MCMC samples. The exact distributions are overlaid in red.

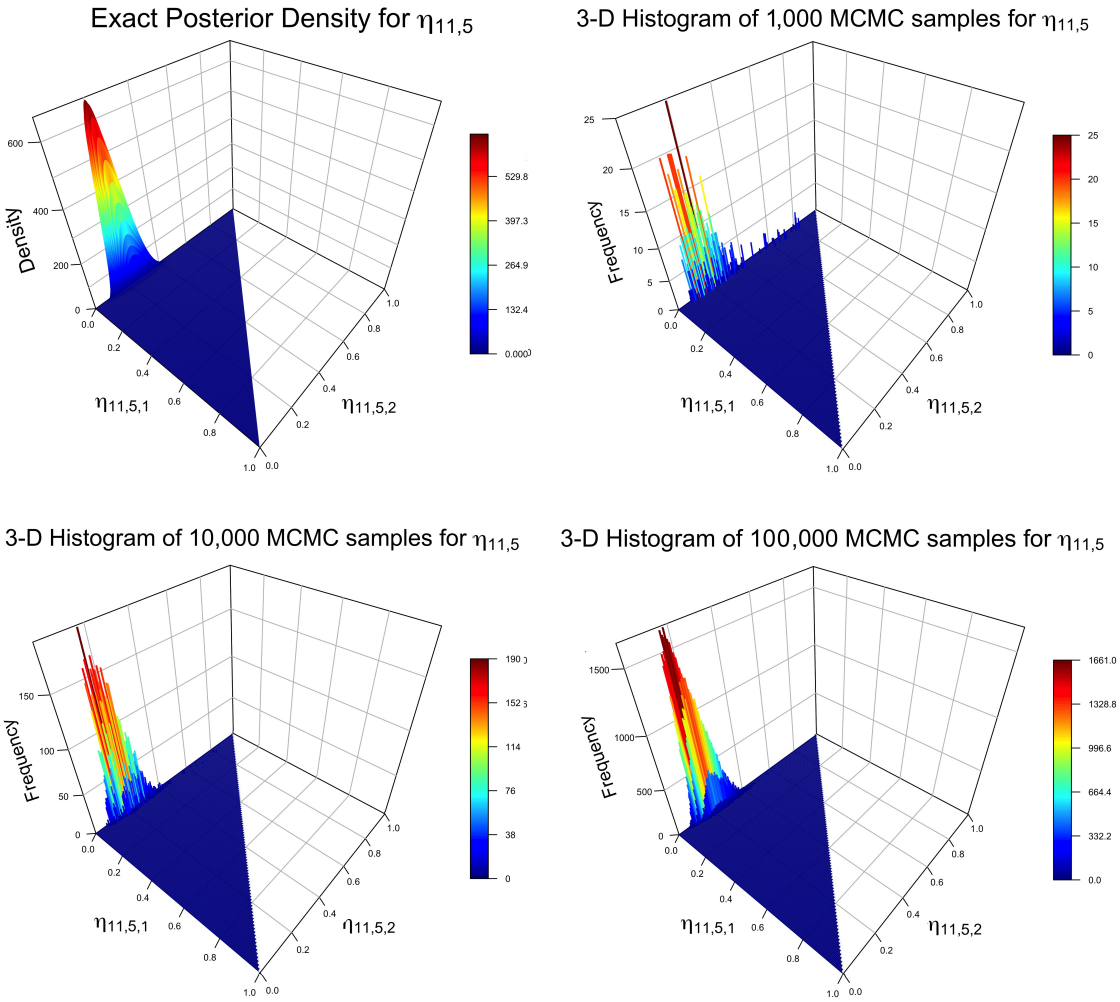


Figure A.9: The exact posterior distribution for $\eta_{11,5}$ (i.e. the form-usage probability vector for item $i = 5$ in cell $x = 11$) is displayed (top left). 3-D histograms for $\eta_{11,5}$ are shown after 1000 (top right), 10000 (bottom left) and 100000 (bottom right) MCMC samples.

A.3 MALA-Based γ-Sampler

In this section, we perform checks for the MALA-based γ-field sampler for the ℓ_2-logistic outlier model. We use the same parameter values and same simulated data as in the previous section, and compare the results obtained.

Figure A.10 shows that the estimated posterior for $\gamma_{11,5}$ converges (slowly) to the exact posterior distribution as the number of MCMC samples taken increases. Further evidence for convergence to the desired posterior distribution is provided in Figure A.11. We observe that the estimated marginal posterior distributions for all γ parameters match the exact posterior distributions.

Figure A.12 shows the equivalent plots to Figure A.10, following transformation into η-space. We can see that the estimated posteriors converge to the exact posterior distribution.

We are aiming to sample from the same (exact) posterior distributions for $\gamma_{11,5}$ (and thus $\eta_{11,5}$) as we were in Section A.2, and we indeed seem to be doing so. It can be inferred that our MALA-based sampling method for γ examined here and the original sampling method seem to lead to the same results.

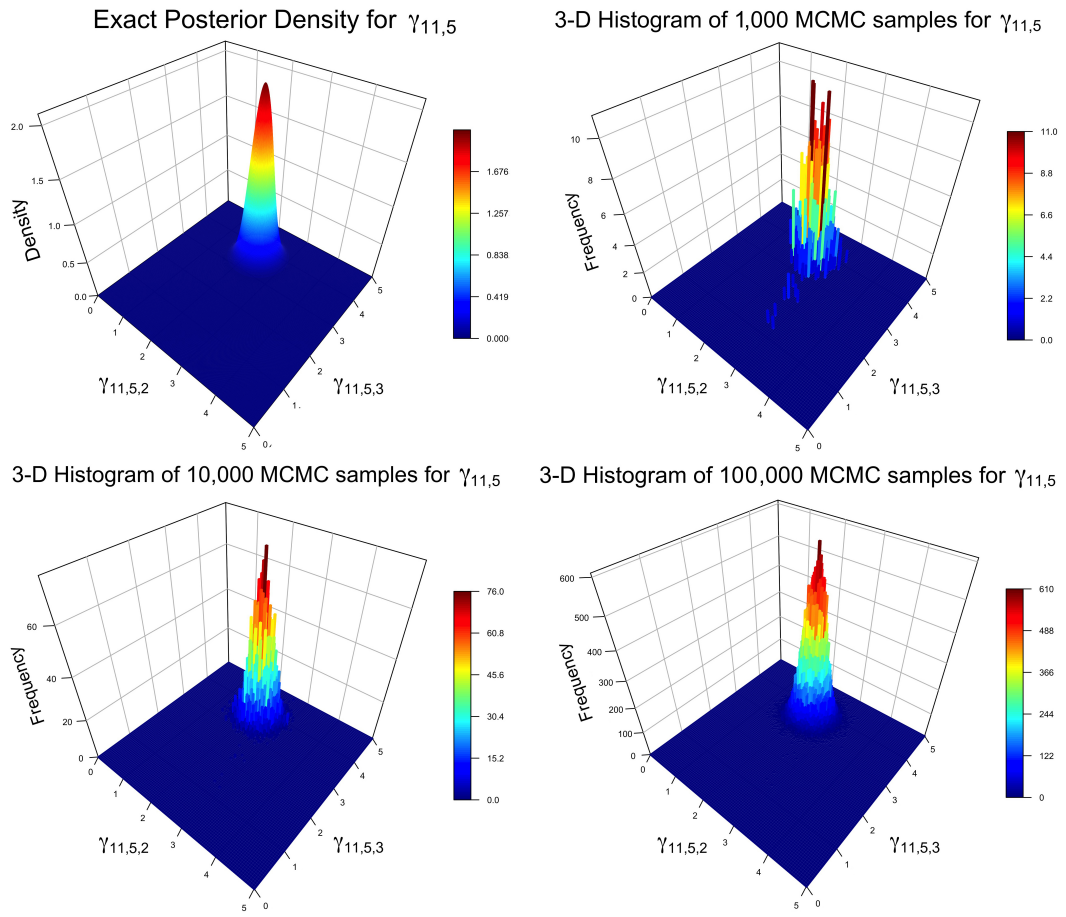


Figure A.10: The exact posterior distribution for $\gamma_{11,5}$ (i.e. the transformed form-usage probability vector for item $i = 5$ in cell $x = 11$) is displayed in the top left. 3-D histograms for $\gamma_{11,5}$ are displayed after 1000 (top right), 10000 (bottom left) and 100000 (bottom right) MCMC samples.

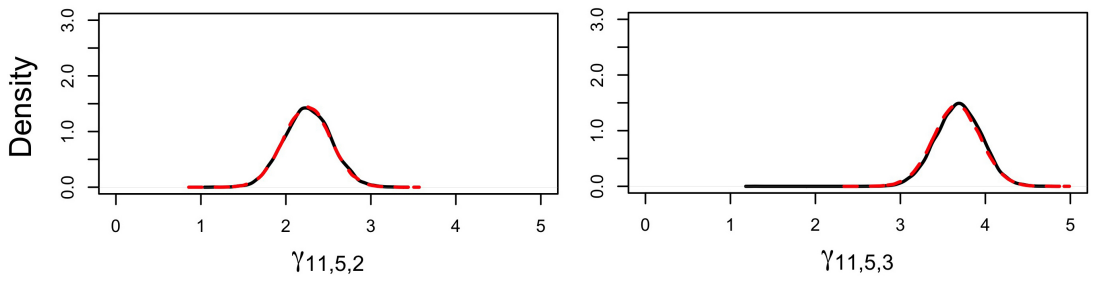


Figure A.11: Estimated marginal posterior distributions for transformed form-usage probabilities $\gamma_{11,5,2}$ (left) and $\gamma_{11,5,3}$ (right) based on 100,000 MCMC samples. The exact distributions are overlaid in red.

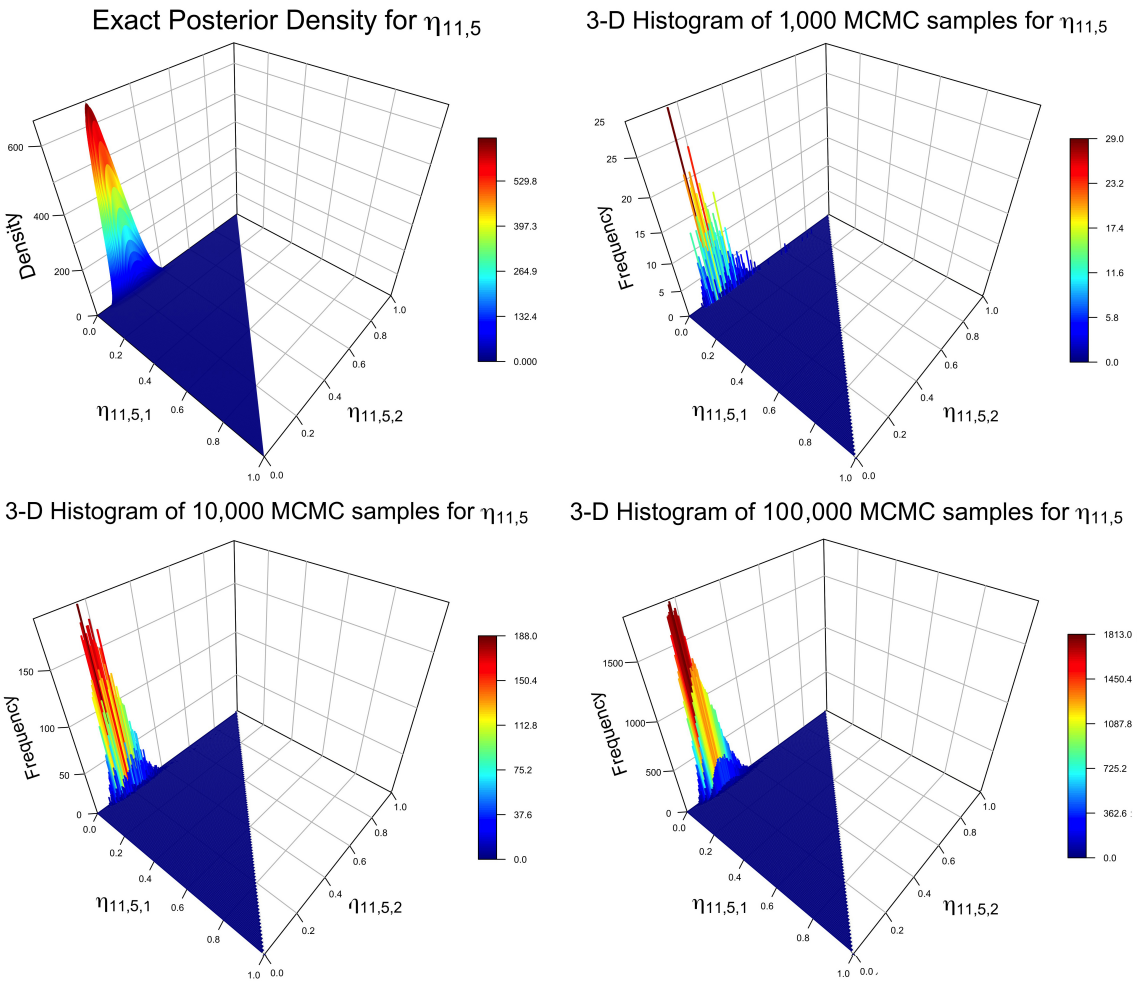


Figure A.12: The exact posterior distribution for $\eta_{11,5}$ (i.e. the form-usage probability vector for item $i = 5$ in cell $x = 11$), is displayed in the top left. 3-D histograms for $\eta_{11,5}$ are displayed after 1000 (top right), 10000 (bottom left) and 100000 (bottom right) MCMC samples.

APPENDIX B

Convergence Diagnostics Output

As mentioned in Chapter 5, three separate Markov chains were generated for each parameter in each analysis, with the algorithm started at different initial values for the parameters $\Theta = (x, \eta, \mu, \phi)$. We performed checks to make sure that these chains gave consistent results, as well as that there was no evidence for a lack of convergence.

In this appendix, we present an example of these convergence diagnostic checks. We consider the first analysis of Chapter 5, where in Section 5.2.2 we used the ℓ_1-Dirichlet zero-inflated model with the eight-item coarsened data subset $\mathbb{S}$ to locate 'new' profiles. We show here convergence diagnostic results for some of the parameters estimated for a randomly selected item, $i = 6$, which corresponds to the word 'much'. These results are indicative of those obtained when a wider view is taken beyond just these three parameters.

Robert & Casella (2010) describe how "a first empirical approach to convergence control is to draw pictures of the output of simulated chains". As an example, we provide these in Figure B.1 for the item usage rate μ_6, the zero-inflation probability

ϕ_6, and the form-usage probability $\eta_{190,6,1}$ (for the form 'MUCL' in cell 190). For each parameter, the three chains are exploring overlapping areas of the parameter space, and there is no evidence for a lack of convergence.

Figure B.2 shows the autocorrelations for these parameters to assess the mixing of the parallel chains. We can see that separate draws from each of these chains are essentially completely uncorrelated after around 100 iterations.

Next, we present the output from a range of standard convergence diagnostic methods. No evidence for a lack of stationarity or other problems with convergence were observed.

First, we consider the approach of Gelman & Rubin (1992), which uses the multiple chains to check for a lack of convergence. As Robert & Casella (2010) states, it is based on the idea that "convergence is diagnosed when the chains have 'forgotten' their initial values, and the output from all chains is indistinguishable". The Gelman-Rubin diagnostic is based on a criterion (the "shrink factor", or "potential scale reduction factor") which considers how the variance between the chains and the variance within the chains differ.

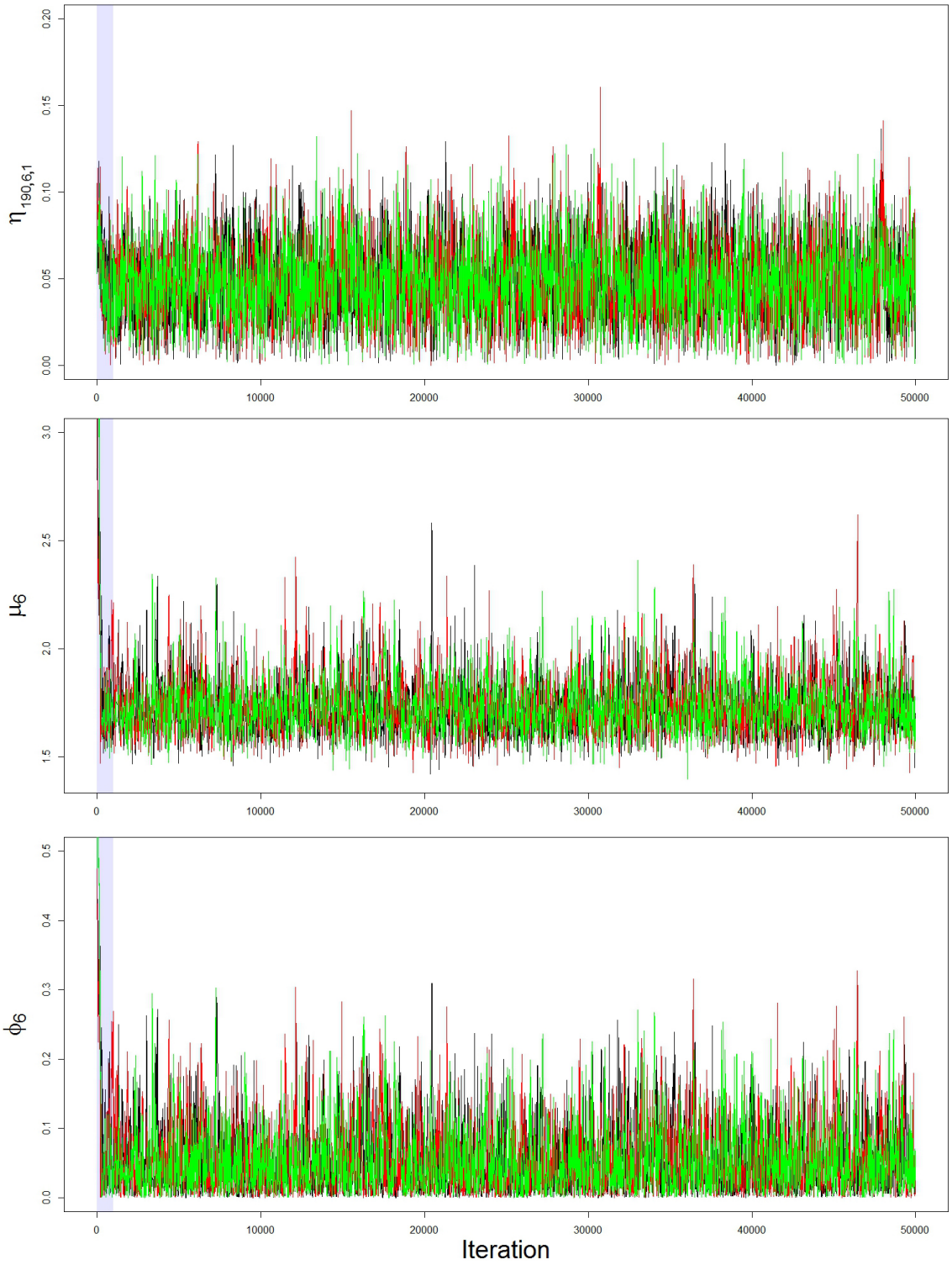


Figure B.1: Trace plots for the three parallel MCMC chains for $\eta_{190,6,1}$ (top), μ_6 (middle) and ϕ_6 (bottom. Results were obtained from modelling with the ℓ_1-Dirichlet zero-inflated model and the eight-item coarsened data subset $\mathbb{S}$. The burn-in region is shaded blue.

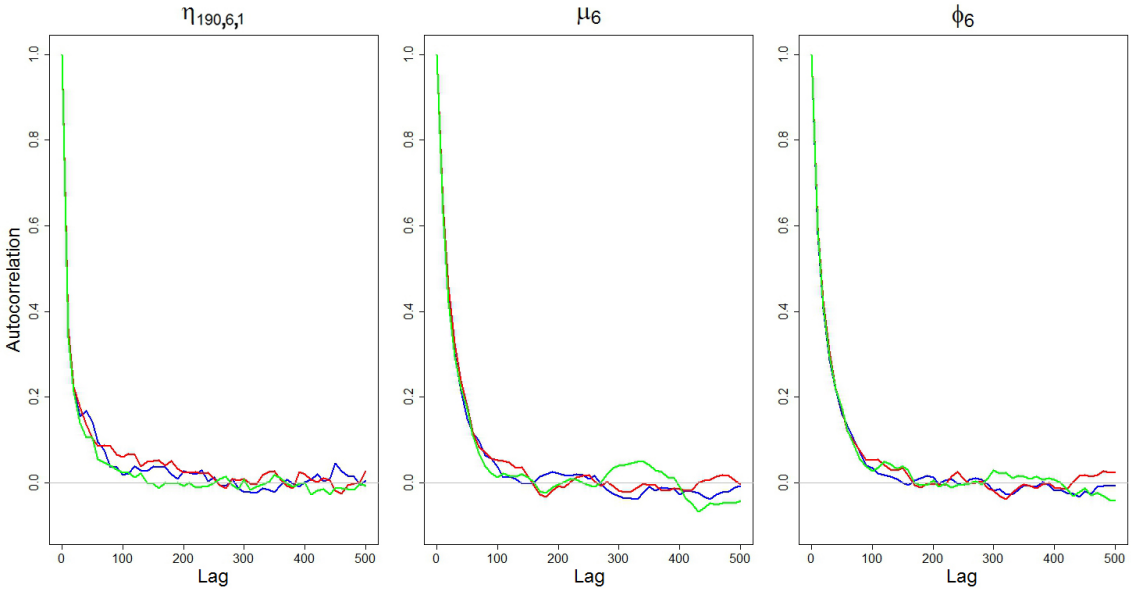


Figure B.2: Autocorrelation plots for three parallel MCMC chains for $\eta_{190,6,1}$ (left), μ_6 (middle) and ϕ_6 (right). Results were obtained from modelling with the ℓ_1-Dirichlet zero-inflated model and the eight-item coarsened data subset $\mathbb{S}$. The first 1000 iterations were discarded as a burn-in.

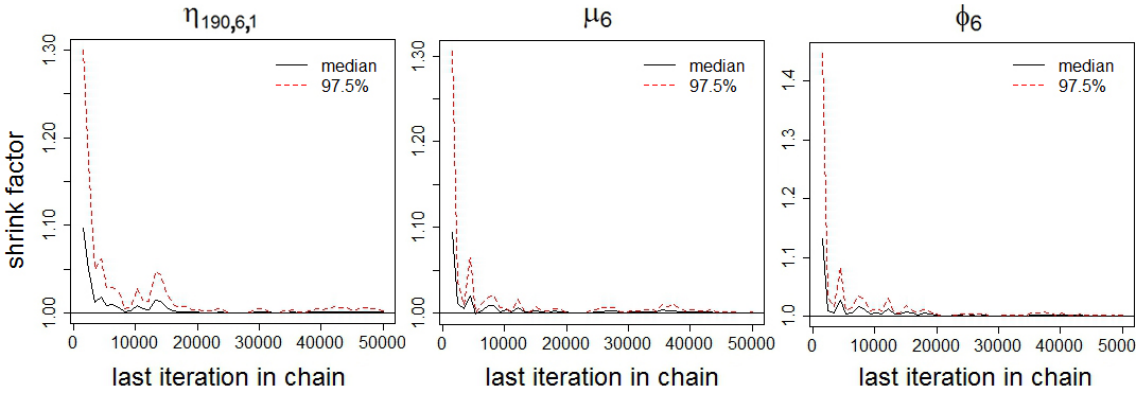


Figure B.3: Gelman-Rubin plots for $\eta_{190,6,1}$ (left), μ_6 (middle) and ϕ_6 (right). Results were obtained from modelling with the ℓ_1-Dirichlet zero-inflated model and the eight-item coarsened data subset $\mathbb{S}$. The first 1000 iterations were discarded as a burn-in.

Figure B.3 displays this shrink factor for the three parameters considered in Figures B.1 and B.2, calculated after varying number of iterations of the chains. The shrink factor (in black) is plotted along with upper confidence limits (in red), which rely on the target distribution being close to normal. In practice, we look for the upper limit to take values close to 1, as they do in these plots (values greater than 1 indicate a lack of convergence).

For each parameter, we next consider the Geweke (1992) test for stationarity of the MCMC chains. This test is based on the idea that the means of the first and last segment of each chain will be equal if the sampled values come from a stationary distribution. A test for this equality is based on the Geweke Z-score (the difference between the means of the two segments, divided by the estimated standard error).

Figure B.4 shows what happens to this Z-score when successively more iterates are discarded from the first segment of the chain. Nearly all of the plotted Z-scores correspond to failures to reject the test hypothesis of equal means, so there is little evidence that the chains for these parameters are not stationary.

Lastly, we present convergence diagnostic results based on the work of Heidelberger & Welch (1983). For each chain of each parameter, a test statistic is calculated to accept or reject the null hypothesis that the sampled values are from a stationary distribution. The test is initially applied to the entire chain, and then with 10% of the chain discarded successively. The test concludes when the null hypothesis has

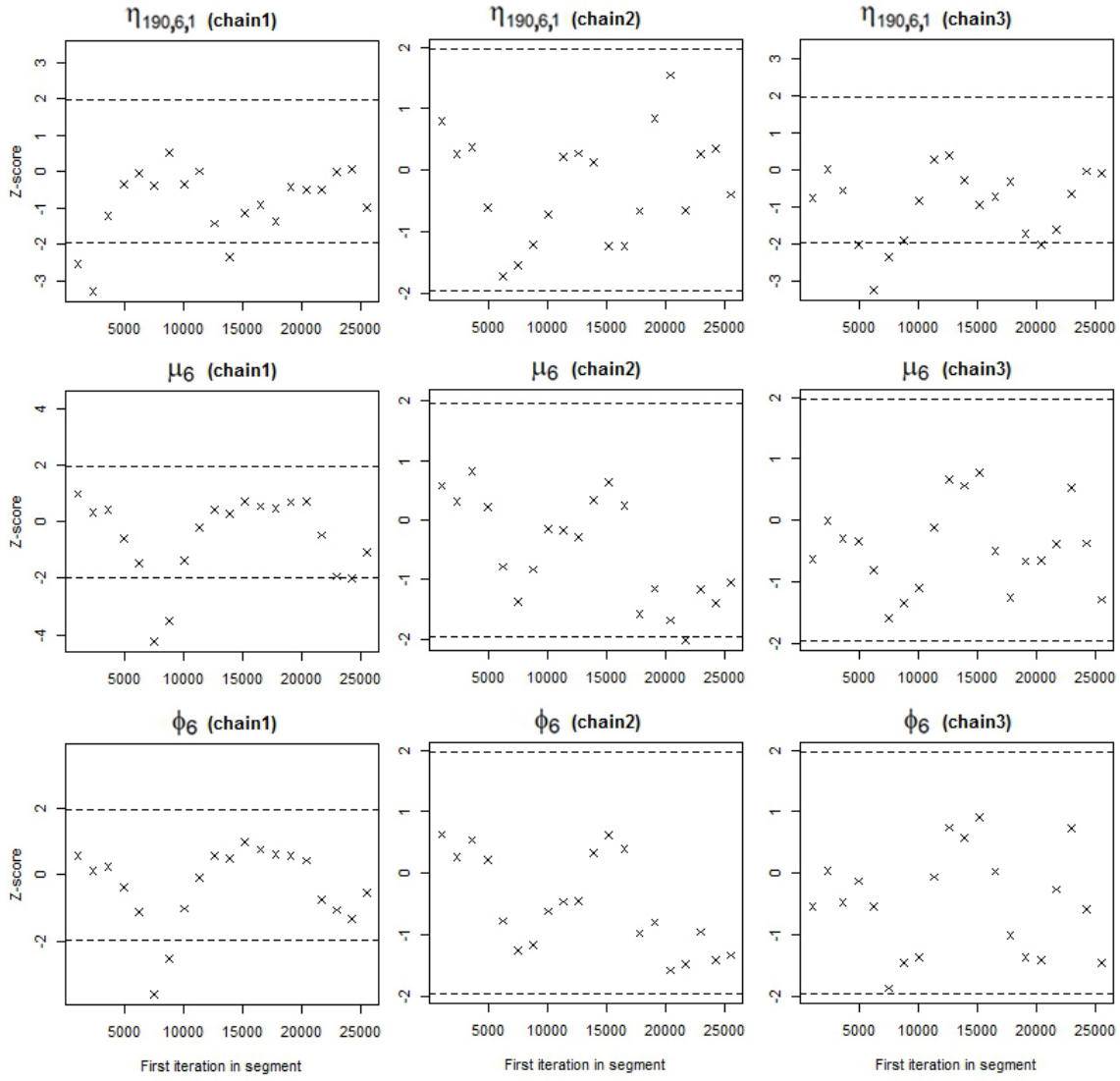


Figure B.4: Geweke-Brooks plots for the three parallel MCMC chains for $\eta_{190,6,1}$ (top), μ_6 (middle) and ϕ_6 (bottom). Results were obtained from modelling with the ℓ_1-Dirichlet zero-inflated model and the eight-item coarsened data subset $\mathbb{S}$. The first 1000 iterations were discarded as a burn-in.

been accepted, or when half of the chain has been discarded (which corresponds to rejecting the null hypothesis).

In the event of the chain passing the stationarity test, the section of the chain not discarded is used for the 'halfwidth test'. A 95% credible interval for the mean of this chain is calculated, and this mean is compared to half the width of the interval. The halfwidth test is passed if the ratio between the half-width and the mean is lower than 0.1. If this is not the case, then the chain is deemed insufficiently long to estimate the mean with sufficient accuracy, and the test is failed.

The output in Table B.1 shows that all chains of all three parameters pass both tests.

Parameter	**Chain**	**Stationarity Test**		**Halfwidth Test**		
		Test Result	p-value	Test Result	Mean	Halfwidth
$\eta_{190,6,1}$	1	Passed	0.099	Passed	0.049	0.001
$\eta_{190,6,1}$	2	Passed	0.939	Passed	0.048	0.001
$\eta_{190,6,1}$	3	Passed	0.190	Passed	0.049	0.001
μ_6	1	Passed	0.667	Passed	1.73	0.008
μ_6	2	Passed	0.489	Passed	1.74	0.008
μ_6	3	Passed	0.609	Passed	1.74	0.008
ϕ_6	1	Passed	0.633	Passed	0.054	0.003
ϕ_6	2	Passed	0.464	Passed	0.055	0.003
ϕ_6	2	Passed	0.615	Passed	0.054	0.003

Table B.1: Heidelberg stationarity and halfwidth test results for the three parallel MCMC chains for $\eta_{190,6,1}$, μ_6 and ϕ_6. Results were obtained from modelling with the ℓ_1-Dirichlet zero-inflated model and the eight-item coarsened data subset $\mathbb{S}$.

APPENDIX C

Derivation of $\nabla \ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\}))$

To perform the Metropolis Adjusted Langevin Algorithm (MALA) with either of the ℓ_2-logistic models defined in Section 2.3, we needed the gradient $\nabla \ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\}))$, where

$$\nabla \ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})) = \begin{pmatrix} \frac{\partial}{\partial \gamma_{x,i,1}} \ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})) \\ \frac{\partial}{\partial \gamma_{x,i,2}} \ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})) \\ \vdots \\ \frac{\partial}{\partial \gamma_{x,i,F_i}} \ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})) \end{pmatrix}. \tag{C.1}$$

In this appendix, we provide a derivation of this gradient for the ℓ_2-logistic outlier model. The gradient for the ℓ_2-logistic zero-inflated model then follows, since this model is nested within the ℓ_2-logistic outlier model (and is obtained by setting $\psi^{(P)} = 0$ and $\psi^{(I)} = 0$ for all $p \in \{1, \ldots, P\}$ and $i \in \{1, \ldots, I\}$).

C.1 Conditional Log-Posterior Distribution for $\gamma_{x,i}$

First, we define the conditional log-posterior distribution for $\gamma_{x,i}$, given the data y and the values of all other model parameters $\Theta \backslash \{\gamma_{x,i}\}$, where $\Theta = (\gamma, \mu, x, \phi, \psi)$.

From Equation (2.33), for some proportionality constant Z_γ, we have that:

$$\ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})) = Z_\gamma + \left\{ \sum_{p \in P_{x,i}} \ln \left((1 - \psi_p^{(P)} \psi_i^{(I)}) \left[\prod_{f=1}^{F_i} \phi_i (1 - y_{p,i,f}) + (1 - \phi_i) \beta_{p,i,f} \right] \right. \right.$$
$$\left. \left. + \psi_p^{(P)} \psi_i^{(I)} \left[\prod_{f=1}^{F_i} \phi_i (1 - y_{p,i,f}) + (1 - \phi_i) \bar{\beta}_{p,i,f} \right] \right) \right\}$$
$$- \frac{1}{2} (\tilde{\gamma}_{x,i} - \lambda_i)^T \Sigma_i^{-1} (\tilde{\gamma}_{x,i} - \lambda_i) - \frac{1}{\sigma_i^2} \left(\sum_{f=1}^{F_i} \sum_{x' \in \mathcal{N}(x)} (\gamma_{x,i,f} - \gamma_{x',i,f})^2 \right). \tag{C.2}$$

Recall, $P_{x,i} = \{p : \sum_f y_{p,i,f} > 0, x_p = x\}$ is the set of profiles p located in cell x in which item i was used at least once. Each pair of neighbours is only counted once here, compared to twice in Equation (2.33), thus the factor of $\frac{1}{2}$ from the ℓ_2-term is removed.

C.2 Problem Formulation

For a given form $f^* \in \{2, \ldots, F_i\}$ of item i, we have that

$$\frac{\partial}{\partial \gamma_{x,i,f^*}} \ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\})) = \left\{ \sum_{p \in P_{x,i}} \frac{\partial}{\partial \gamma_{x,i,f^*}} \ln\left(F_1(i,p)\right) \right\}$$
$$- \frac{1}{2} \frac{\partial}{\partial \gamma_{x,i,f^*}} (\tilde{\gamma}_{x,i} - \lambda_i)^T \Sigma_i^{-1} (\tilde{\gamma}_{x,i} - \lambda_i)$$
$$- \frac{2}{\sigma_i^2} \left(|N(x)| \gamma_{x,i,f} - \sum_{x' \in \mathcal{N}(x)} \gamma_{x',i,f} \right), \tag{C.3}$$

where $|N(x)|$ is the number of neighbouring cells to x, and

$$F_1(i,p) = \left\{ (1 - \psi_p^{(P)} \psi_i^{(I)}) \left[\prod_{f=1}^{F_i} \phi_i (1 - y_{p,i,f}) + (1 - \phi_i) \beta_{p,i,f} \right] \right.$$
$$\left. + \psi_p^{(P)} \psi_i^{(I)} \left[\prod_{f=1}^{F_i} \phi_i (1 - y_{p,i,f}) + (1 - \phi_i) \bar{\beta}_{p,i,f} \right] \right\}. \tag{C.4}$$

C.3 Deriving $\frac{\partial}{\partial \gamma_{x,i,f^*}}\{(\tilde{\gamma}_{x,i}-\lambda_i)^T\Sigma_i^{-1}(\tilde{\gamma}_{x,i}-\lambda_i)\}$

Using that λ_i is an vector of $F-1$ zeroes, and that Σ_i^{-1} is symmetric,

$$\begin{aligned}
\frac{\partial}{\partial \gamma_{x,i,f^*}}(\tilde{\gamma}_{x,i})^T\Sigma_i^{-1}(\tilde{\gamma}_{x,i}) &= \frac{\partial}{\partial \gamma_{x,i,f^*}}\left(\sum_{a=1}^{F_i-1}\sum_{b=1}^{F_i-1}(\gamma_{x,i,a+1})(\gamma_{x,i,b+1})[\Sigma_i^{-1}]_{a,b}\right) \\
&= \frac{\partial}{\partial \gamma_{x,i,f^*}}\left\{\left(\sum_{a\in\{2,\ldots,F_i\}\setminus f^*}(\gamma_{x,i,a})(\gamma_{x,i,f^*})[\Sigma_i^{-1}]_{a-1,f^*-1}\right)\right. \\
&\quad + \left(\sum_{b\in\{2,\ldots,F_i\}\setminus f^*}(\gamma_{x,i,b})(\gamma_{x,i,f^*})[\Sigma_i^{-1}]_{f^*-1,b-1}\right) \\
&\quad \left.+\gamma_{x,i,f}^2[\Sigma_i^{-1}]_{f^*-1,f^*-1}\right\} \\
&= \frac{\partial}{\partial \gamma_{x,i,f^*}}\left\{2\left(\sum_{a\in\{2,\ldots,F_i\}\setminus f^*}(\gamma_{x,i,a})(\gamma_{x,i,f^*})[\Sigma_i^{-1}]_{a-1,f^*-1}\right)\right. \\
&\quad \left.+\ \gamma_{x,i,f}^2[\Sigma_i^{-1}]_{f^*-1,f^*-1}\right\} \\
&= 2\left\{\left(\sum_{a\in\{2,\ldots,F_i\}\setminus f^*}(\gamma_{x,i,a})[\Sigma_i^{-1}]_{a-1,f^*-1}\right)+\gamma_{x,i,f}[\Sigma_i^{-1}]_{f^*-1,f^*-1}\right\}.
\end{aligned} \tag{C.5}$$

C.4 Deriving $\frac{\partial}{\partial \gamma_{x,i,f^*}}\{\ln(F_1(i,p))\}$

We have that

$$\frac{\partial}{\partial \gamma_{x,i,f^*}}\{\ln(F_1(i,p))\} = \left\{\frac{1}{F_1(i,p)}\right\}\frac{\partial}{\partial \gamma_{x,i,f^*}}F_1(i,p), \tag{C.6}$$

Further,

$$\begin{aligned}
\frac{\partial}{\partial \gamma_{x,i,f^*}}F_1(i,p) = \Big\{&(1-\psi_p^{(P)}\psi_i^{(I)})\frac{\partial}{\partial \gamma_{x,i,f^*}}F_2(i,p) \\
&+\psi_p^{(P)}\psi_i^{(I)}\frac{\partial}{\partial \gamma_{x,i,f^*}}F_3(i,p)\Big\},
\end{aligned} \tag{C.7}$$

where

$$F_2(i,p) = \prod_{f=1}^{F_i} \phi_i(1-y_{p,i,f}) + (1-\phi_i)\beta_{p,i,f}$$
$$F_3(i,p) = \prod_{f=1}^{F_i} \phi_i(1-y_{p,i,f}) + (1-\phi_i)\bar{\beta}_{p,i,f}. \tag{C.8}$$

C.4.1 Deriving $\frac{\partial}{\partial\gamma_{x,i,f^*}}\{F_2(i,p)\}$

$$\frac{\partial}{\partial\gamma_{x,i,f^*}}F_2(i,p) = \sum_{f=1}^{F_i}\left[\left(\frac{\partial}{\partial\gamma_{x,i,f^*}}F_4(i,p,f)\right)\prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f}}F_4(i,p,g)\right], \tag{C.9}$$

where

$$F_4(i,p,f) = \phi_i(1-y_{p,i,f}) + (1-\phi_i)\beta_{p,i,f}. \tag{C.10}$$

First, consider the case for a form $f \neq f^*$. We have that

$$\begin{aligned}
\frac{\partial}{\partial\gamma_{x,i,f^*}}F_4(i,p,f) &= (1-\phi_i)\left\{\frac{\partial}{\partial\gamma_{x,i,f^*}}\beta_{p,i,f}\right\} \\
&= (1-\phi_i)\frac{\partial}{\partial\gamma_{x,i,f^*}}\left\{y_{p,i,f}(1-e^{-\mu_i\eta_{x_p,i,f}}) + (1-y_{p,i,f})e^{-\mu_i\eta_{x_p,i,f}}\right\} \\
&= (1-\phi_i)(1-2y_{p,i,f})\left\{\frac{\partial}{\partial\gamma_{x,i,f^*}}e^{-\mu_i\eta_{x_p,i,f}}\right\} \\
&= -(1-\phi_i)(1-2y_{p,i,f})\mu_i e^{-\mu_i\eta_{x_p,i,f}}\left\{\frac{\partial}{\partial\gamma_{x,i,f^*}}\eta_{x_p,i,f}\right\}
\end{aligned} \tag{C.11}$$

Using that $x_p = x$,

$$\begin{aligned}
\frac{\partial}{\partial\gamma_{x,i,f^*}}\eta_{x_p,i,f} &= \frac{\partial}{\partial\gamma_{x,i,f^*}}\left\{e^{\gamma_{x,i,f}}\Big/\left(e^{\gamma_{x,i,f^*}} + \sum_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f^*}} e^{\gamma_{x,i,g}}\right)\right\} \\
&= \left\{-e^{\gamma_{x,i,f}}e^{\gamma_{x,i,f^*}}\Big/\left(\sum_{g\in\{1,\ldots,F_i\}} e^{\gamma_{x,i,g}}\right)^2\right\} \\
&= -\eta_{x,i,f}\eta_{x,i,f^*}
\end{aligned} \tag{C.12}$$

Thus,

$$\frac{\partial}{\partial \gamma_{x,i,f^*}} F_4(i,p,f) = (1-\phi_i)(1-2y_{p,i,f})\mu_i \eta_{x,i,f}\eta_{x,i,f^*} e^{-\mu_i \eta_{x,i,f}} \tag{C.13}$$

Next, consider when $f = f^*$. Then, we have

$$\frac{\partial}{\partial \gamma_{x,i,f^*}} F_4(i,p,f^*) = -(1-\phi_i)(1-2y_{p,i,f^*})\mu_i e^{-\mu_i \eta_{x,i,f^*}} \left\{ \frac{\partial}{\partial \gamma_{x,i,f^*}} \eta_{x,i,f^*} \right\}. \tag{C.14}$$

Given that

$$\begin{aligned}
\frac{\partial}{\partial \gamma_{x,i,f^*}} \eta_{x_p,i,f^*} &= \frac{\partial}{\partial \gamma_{x,i,f^*}} \left\{ e^{\gamma_{x,i,f^*}} \Big/ \left(e^{\gamma_{x,i,f^*}} + \sum_{\substack{g \in \{1,\ldots,F_i\} \\ g \neq f^*}} e^{\gamma_{x,i,g}} \right) \right\} \\
&= \left\{ \left[\left(\sum_{g \in \{1,\ldots,F_i\}} e^{\gamma_{x,i,g}} \right) e^{\gamma_{x,i,f^*}} - (e^{\gamma_{x,i,f^*}})^2 \right] \Big/ \left(\sum_{g \in \{1,\ldots,F_i\}} e^{\gamma_{x,i,g}} \right)^2 \right\} \\
&= \eta_{x,i,f^*}(1-\eta_{x,i,f^*}),
\end{aligned} \tag{C.15}$$

we therefore have

$$\frac{\partial}{\partial \gamma_{x,i,f^*}} F_4(i,p,f^*) = (1-\phi_i)(2y_{p,i,f^*}-1)\mu_i \eta_{x,i,f^*}(1-\eta_{x,i,f^*}) e^{-\mu_i \eta_{x,i,f^*}}. \tag{C.16}$$

Using Equations (C.13) and (C.16), we therefore have that

$$\begin{aligned}
\frac{\partial}{\partial \gamma_{x,i,f^*}} F_2(i,p) = &\left\{ \left(\frac{\partial}{\partial \gamma_{x,i,f^*}} F_4(i,p,f^*) \right) \prod_{\substack{g \in \{1,\ldots,F_i\} \\ g \neq f^*}} F_4(i,p,g) \right\} \\
&+ \sum_{\substack{f \in \{1,\ldots,F_i\} \\ f \neq f^*}} \left[\left(\frac{\partial}{\partial \gamma_{x,i,f^*}} F_4(i,p,f) \right) \prod_{\substack{g \in \{1,\ldots,F_i\} \\ g \neq f}} F_4(i,p,g) \right],
\end{aligned}$$

and thus,

$$
\begin{aligned}
&= (1-\phi_i)(2y_{p,i,f^*}-1)\mu_i\eta_{x,i,f^*}(1-\eta_{x,i,f^*})e^{-\mu_i\eta_{x,i,f^*}} \\
&\quad \times \prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f^*}} \Big\{\phi_i(1-y_{p,i,g}) + (1-\phi_i)\beta_{p,i,g}\Big\} \\
&\quad + \sum_{\substack{f\in\{1,\ldots,F_i\}\\ f\neq f^*}} \Bigg[(1-\phi_i)(1-2y_{p,i,f})\mu_i\eta_{x,i,f}\eta_{x,i,f^*}e^{-\mu_i\eta_{x,i,f}} \\
&\qquad \times \prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f}} \{\phi_i(1-y_{p,i,g}) + (1-\phi_i)\beta_{p,i,g}\}\Bigg].
\end{aligned} \tag{C.17}
$$

C.4.2 Deriving $\frac{\partial}{\partial\gamma_{x,i,f^*}}\{F_3(i,p)\}$

We have that

$$
\begin{aligned}
\frac{\partial}{\partial\gamma_{x,i,f^*}}F_3(i,p) = &\left\{\left(\frac{\partial}{\partial\gamma_{x,i,f^*}}F_5(i,p,f^*)\right)\prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f^*}}F_5(i,p,g)\right\} \\
&+ \sum_{\substack{f\in\{1,\ldots,F_i\}\\ f\neq f^*}}\left[\left(\frac{\partial}{\partial\gamma_{x,i,f^*}}F_5(i,p,f)\right)\prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f}}F_5(i,p,g)\right],
\end{aligned} \tag{C.18}
$$

where

$$
F_5(i,p,f) = \phi_i(1-y_{p,i,f}) + (1-\phi_i)\bar{\beta}_{p,i,f}. \tag{C.19}
$$

When $f = f^*$, we have

$$
\begin{aligned}
\frac{\partial}{\partial\gamma_{x,i,f^*}}F_5(i,p,f^*) &= (1-\phi_i)\frac{\partial}{\partial\gamma_{x,i,f^*}}\left\{y_{p,i,f^*}(1-e^{-\mu_i\bar{\eta}_{i,f^*}}) + (1-y_{p,i,f^*})e^{-\mu_i\bar{\eta}_{i,f^*}}\right\} \\
&= -(1-\phi_i)(1-2y_{p,i,f^*})\mu_i e^{-\mu_i\bar{\eta}_{i,f^*}}\left\{\frac{\partial}{\partial\gamma_{x,i,f^*}}\bar{\eta}_{i,f^*}\right\}.
\end{aligned} \tag{C.20}
$$

Using Equation (C.15), we have that

$$\begin{aligned}\frac{\partial}{\partial\gamma_{x,i,f^*}}\bar{\eta}_{i,f^*} &= \frac{\partial}{\partial\gamma_{x,i,f^*}}\left\{\frac{1}{C}\sum_{x=1}^{C}\left[e^{\gamma_{x,i,f^*}}/\left(e^{\gamma_{x,i,f^*}}+\sum_{g\neq f^*}e^{\gamma_{x,i,g}}\right)\right]\right\}\\ &= \frac{1}{C}\sum_{x=1}^{C}\eta_{x,i,f^*}(1-\eta_{x,i,f^*})\\ &= \bar{\eta}_{i,f^*}-\frac{1}{C}\sum_{x=1}^{C}(\eta_{x,i,f^*})^2, \end{aligned} \tag{C.21}$$

and thus

$$\frac{\partial}{\partial\gamma_{x,i,f^*}}F_5(i,p,f^*) = (1-\phi_i)(2y_{p,i,f^*}-1)\mu_i\left\{\bar{\eta}_{i,f^*}-\frac{1}{C}\sum_{x=1}^{C}(\eta_{x,i,f^*})^2\right\}e^{-\mu_i\bar{\eta}_{i,f^*}}. \tag{C.22}$$

If instead $f \neq f^*$, then

$$\frac{\partial}{\partial\gamma_{x,i,f^*}}F_5(i,p,f) = (1-\phi_i)(2y_{p,i,f}-1)\mu_i e^{-\mu_i\bar{\eta}_{i,f}}\left\{\frac{\partial}{\partial\gamma_{x,i,f^*}}\bar{\eta}_{i,f}\right\}. \tag{C.23}$$

Using Equation (C.12), we have that

$$\begin{aligned}\frac{\partial}{\partial\gamma_{x,i,f}}\bar{\eta}_{i,f} &= \frac{\partial}{\partial\gamma_{x,i,f}}\left\{\frac{1}{C}\sum_{x=1}^{C}\left[e^{\gamma_{x,i,f}}/\left(e^{\gamma_{x,i,f^*}}+\sum_{g\neq f^*}e^{\gamma_{x,i,g}}\right)\right]\right\}\\ &= \frac{-1}{C}\sum_{x=1}^{C}\eta_{x,i,f}\eta_{x,i,f^*} \end{aligned} \tag{C.24}$$

and so

$$\frac{\partial}{\partial\gamma_{x,i,f^*}}F_{10}(i,p,f) = \frac{(1-\phi_i)}{C}(1-2y_{p,i,f})\mu_i e^{-\mu_i\bar{\eta}_{i,f}}\left\{\sum_{x=1}^{C}\eta_{x,i,f}\eta_{x,i,f^*}\right\}. \tag{C.25}$$

Combining Equations (C.22) and (C.25) with Equation (C.18) gives us that

$$\begin{aligned}\frac{\partial}{\partial\gamma_{x,i,f^*}}F_3(i,p) = (1-\phi_i)(2y_{p,i,f^*}-1)\mu_i\left\{\bar{\eta}_{i,f^*}-\frac{1}{C}\sum_{x=1}^{C}(\eta_{x,i,f^*})^2\right\}e^{-\mu_i\bar{\eta}_{i,f^*}}\\ \times\prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f^*}}\phi_i(1-y_{p,i,g})+(1-\phi_i)\bar{\beta}_{p,i,g}\\ +\sum_{\substack{f\in\{1,\ldots,F_i\}\\ f\neq f^*}}\left[\frac{(1-\phi_i)}{C}(1-2y_{p,i,f})\mu_i e^{-\mu_i\bar{\eta}_{i,f}}\left\{\sum_{x=1}^{C}\eta_{x,i,f}\eta_{x,i,f^*}\right\}\right.\\ \left.\times\prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f}}\phi_i(1-y_{p,i,g})+(1-\phi_i)\bar{\beta}_{p,i,g}\right]. \qquad \text{(C.26)}\end{aligned}$$

C.4.3 Combining Results to give $\frac{\partial}{\partial\gamma_{x,i,f^*}}\{F_1(i,x)\}$

Putting Equations (C.7), (C.17) and (C.26) together gives us that

$$\begin{aligned}\frac{\partial}{\partial\gamma_{x,i,f^*}}F_1(i,p) = (1-\psi_p^{(P)}\psi_i^{(I)})\Bigg((1-\phi_i)(2y_{p,i,f^*}-1)\mu_i\eta_{x,i,f^*}(1-\eta_{x,i,f^*})e^{-\mu_i\eta_{x,i,f^*}}\\ \times\prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f^*}}\Big\{\phi_i(1-y_{p,i,g})+(1-\phi_i)\beta_{p,i,g}\Big\}\\ +\sum_{\substack{f\in\{1,\ldots,F_i\}\\ f\neq f^*}}\Big[(1-\phi_i)(1-2y_{p,i,f})\mu_i\eta_{x,i,f}\eta_{x,i,f^*}e^{-\mu_i\eta_{x,i,f}}\\ \times\prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f}}\{\phi_i(1-y_{p,i,g})+(1-\phi_i)\beta_{p,i,g}\}\Big]\Bigg)\\ +\psi_p^{(P)}\psi_i^{(I)}\left(\left((1-\phi_i)(2y_{p,i,f^*}-1)\mu_i\left\{\bar{\eta}_{i,f^*}-\frac{1}{C}\sum_{x=1}^{C}(\eta_{x,i,f^*})^2\right\}e^{-\mu_i\bar{\eta}_{i,f^*}}\right.\right.\\ \times\prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f^*}}\{\phi_i(1-y_{p,i,g})+(1-\phi_i)\bar{\beta}_{p,i,g}\}\\ +\sum_{\substack{f\in\{1,\ldots,F_i\}\\ f\neq f^*}}\left[\frac{(1-\phi_i)}{C}(1-2y_{p,i,f})\mu_i e^{-\mu_i\bar{\eta}_{i,f}}\left\{\sum_{x=1}^{C}\eta_{x,i,f}\eta_{x,i,f^*}\right\}\right.\\ \left.\left.\left.\times\prod_{\substack{g\in\{1,\ldots,F_i\}\\ g\neq f}}\phi_i(1-y_{p,i,g})+(1-\phi_i)\bar{\beta}_{p,i,g}\right]\right). \qquad \text{(C.27)}\end{aligned}$$

C.5 Combining Results to give $\nabla \ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\}))$

Define $\nabla_f = \frac{\partial}{\partial \gamma_{x,i,f}} \ln(\pi_\gamma(\gamma_{x,i} \mid y, \Theta \backslash \{\gamma_{x,i}\}))$ for $f \in \{2, \ldots, F_i\}$. We have that $\nabla_1 = 0$, and putting Equations (C.3), (C.5) and (C.27) together gives us that

$$
\begin{aligned}
\nabla_{f^*} = \Bigg\{ \sum_{p \in P_{x,i}} \Bigg(& \frac{1}{(1-\psi_p^{(P)}\psi_i^{(I)}) \left[\prod_{f=1}^{F_i} \phi_i(1-y_{p,i,f}) + (1-\phi_i)\beta_{p,i,f}\right] + \psi_p^{(P)}\psi_i^{(I)} \left[\prod_{f=1}^{F_i} \phi_i(1-y_{p,i,f}) + (1-\phi_i)\bar{\beta}_{p,i,f}\right]} \\
& \times \Bigg\{ (1-\psi_p^{(P)}\psi_i^{(I)}) \Bigg((1-\phi_i)(2y_{p,i,f^*}-1)\mu_i \eta_{x,i,f^*}(1-\eta_{x,i,f^*}) e^{-\mu_i \eta_{x,i,f^*}} \\
& \times \prod_{\substack{g \in \{1,\ldots,F_i\} \\ g \neq f^*}} \left\{ \phi_i(1-y_{p,i,g}) + (1-\phi_i)\beta_{p,i,g} \right\} + \sum_{\substack{f \in \{1,\ldots,F_i\} \\ f \neq f^*}} \Big[(1-\phi_i)(1-2y_{p,i,f})\mu_i \eta_{x,i,f} \eta_{x,i,f^*} \\
& \times e^{-\mu_i \eta_{x,i,f}} \prod_{\substack{g \in \{1,\ldots,F_i\} \\ g \neq f}} \{\phi_i(1-y_{p,i,g}) + (1-\phi_i)\beta_{p,i,g}\} \Big] \Bigg) \\
& + \psi_p^{(P)}\psi_i^{(I)} \Bigg((1-\phi_i)(2y_{p,i,f^*}-1)\mu_i \left\{ \bar{\eta}_{i,f^*} - \frac{1}{C}\sum_{x=1}^{C} (\eta_{x,i,f^*})^2 \right\} e^{-\mu_i \bar{\eta}_{i,f^*}} \\
& \times \prod_{\substack{g \in \{1,\ldots,F_i\} \\ g \neq f^*}} \{\phi_i(1-y_{p,i,g}) + (1-\phi_i)\bar{\beta}_{p,i,g}\} + \sum_{\substack{f \in \{1,\ldots,F_i\} \\ f \neq f^*}} \Bigg[\frac{(1-\phi_i)}{C}(1-2y_{p,i,f})\mu_i e^{-\mu_i \bar{\eta}_{i,f}} \\
& \times \left\{ \sum_{x=1}^{C} \eta_{x,i,f}\eta_{x,i,f^*} \right\} \prod_{\substack{g \in \{1,\ldots,F_i\} \\ g \neq f}} \phi_i(1-y_{p,i,g}) + (1-\phi_i)\bar{\beta}_{p,i,g} \Bigg] \Bigg) \Bigg\} \Bigg) \\
& - \left\{ \left(\sum_{a \in \{2,\ldots,F_i\} \backslash f^*} (\gamma_{x,i,a}) [\Sigma_i^{-1}]_{a-1,f^*-1} \right) + \gamma_{x,i,f} [\Sigma_i^{-1}]_{f^*-1,f^*-1} \right\} \\
& - \frac{2}{\sigma_i^2} \left(|N(x)| \gamma_{x,i,f} - \sum_{x' \in \mathcal{N}(x)} \gamma_{x',i,f} \right). \qquad \text{(C.28)}
\end{aligned}
$$

References

Abrol, S. & Khan, L. (2010). Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining. In *SOCIALCOM '10: Proceedings of the 2010 IEEE Second International Conference on Social Computing*. Washington, DC, USA: IEEE Computer Society, pp. 153–160.

Ahmed, A., Hong, L. & Smola, A. J. (2013). Hierarchical geographical modeling of user locations from social media posts. In *WWW '13: Proceedings of the 22nd International Conference on World Wide Web*. New York, NY, USA: ACM, pp. 25–36.

Aitchison, J. & Shen, S. M. (1980). Logistic-normal distributions: some properties and uses. *Biometrika* **67**, 261–272.

Amos, W. & Manica, A. (2006). Global genetic positioning: Evidence for early human population centers in coastal habitats. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 820–824.

Backstrom, L., Sun, E. & Marlow, C. (2010). Find me if you can: Improving geographical prediction with social and spatial proximity. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*. New York, NY, USA: ACM, pp. 61–70.

Benskin, M. (1988). The numerical classification of languages, and dialect maps for the past. In *Spatial and Temporal Distributions, Manuscript Constellations*. Amsterdam, Netherlands: Benjamins, pp. 13–38.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research* **3**, 993–1022.

Bo, H., Baldwin, P. & Cook, T. (2012). Geolocation prediction in social media data by finding location indicative words. In *COLING 2012: Proceedings of the*

24th International Conference on Computational Linguistics. Mumbai, India: The COLING 2012 Organizing Committee, pp. 1045–1062.

Bradburd, G. S., Ralph, P. L. & Coop, G. M. (2016). A spatial framework for understanding population structure and admixture. *Public Library of Science Genetics* **12**, e1005703.

Bryc, K., Auton, A., Nelson, M., Oksenberg, J., Hauser, S., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S. & Bustamante, C. (2010). Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences* **107**, 786–791.

Chandra, S., Khan, L. & Muhaya, F. (2011). Estimating Twitter user location using social interactions - a content based approach. In *Privacy, Security, Risk and Trust: 2011 IEEE Third International Conference on Social Computing*. Washington, DC, USA: IEEE Computer Society, pp. 838–843.

Chang, H., Lee, D., Eltaher, M. & Lee, J. (2012). @Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *ASONAM 2012: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining*. Washington, DC, USA: IEEE Computer Society, pp. 111–118.

Chen, C.-Y. & Grauman, L. (2011). Clues from the beaten path: Location estimation with bursty sequences of tourist photos. In *2011 IEEE Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, pp. 1569–1576.

Chen, J., Zheng, H., Bei, J.-X., Sun, L., .Jia, W.-H., Li, T., Zhang, F., Seielstad, M., Zeng, Y.-X., Zhang, X. & Liu, J. (2009). Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *The American Journal of Human Genetics* **85**, 775–785.

Cheng, Z., Caverlee, J. & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, pp. 759–768.

Crandall, D., Backstrom, L., Huttenlocher, D. & Kleinberg, J. (2009). Mapping the world's photos. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web*. New York, NY, USA: ACM, pp. 761–770.

Cressie, N. & Kornack, J. (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science* **18**, 436–456.

Dalvi, N., Kumar, R. & Pang, B. (2012). Object matching in tweets with spatial models. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, pp. 43–52.

Davis, C., Pappa, G., de Oliveira, D. & de L Arcanjo, F. (2011). Inferring the location of Twitter messages based on user relationships. *Transactions in GIS* **15**, 735–751.

Eisenstein, J., Ahmed, A. & Xing, E. (2011). Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning*. St Louis, MO, USA: The International Machine Learning Society, pp. 1041–1048.

Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. P. (2010). A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA, USA: Association for Computational Linguistics, pp. 1277–1287.

Elhaik, E., Tatarinova, T., Chebotarev, D., Piras, I., Calò, C., Montis, A. D., Atzori, M., Marini, M., Tofanelli, S., Francalacci, P., Pagani, L., Tyler-Smith, C., Xue, Y., Cucca, F., Schurr, T., Gaieski, J., Melendez, C., Vilar, M., Owings, A., Gomez, R., Fujita, R., Santos, F., Comas, D., Balanovsky, O., Balanovska, E., Zalloua, P., Soodyall, H., Pitchappan, R., Prasad, A., Hammer, M., Matisoo-Smith, L., Wells, R. & Consortium, T. G. (2014). Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nature Communications* **5**, 3513.

Engelhardt, B. & Stephens, M. (2010). Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *Public Library of Science Genetics* **6**, e1001117.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2014). *Bayesian Data Analysis*, vol. 2. Boca Raton, FL, USA: Chapman & Hall/CRC.

Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.

Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geweke, J. (1992). *Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments*. Oxford, UK: Oxford University Press.

Ghobrial, L., Lankester, F., Kiyang, J., Akih, A., Vries, S. D., Fotso, R., Gadsby, E., Jenkins, P. & Gonder, M. (2010). Tracing the origins of rescued chimpanzees reveals widespread chimpanzee hunting in Cameroon. *BMC Ecology* **10**, 1.

Giorgi, E. & Diggle, P. (2014). On the inverse geostatistical problem of inference on missing locations. *Spatial Statistics* **11**, 35–44.

Gopalan, P., Hao, W., Blei, D. M. & Storey, J. D. (2015). Scaling probabilistic models of genetic variation to millions of humans. *bioRxiv preprint*, 013227.

Guillot, G., Leblois, R., Coulon, A. & Frantz, A. (2009). Statistical methods in spatial genetics. *Molecular Ecology* **18**, 4734–4756.

Han, B., Cook, P. & Baldwin, T. (2014). Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research* **49**, 451–500.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hauff, C. & Houben, G.-J. (2012). Placing images on the world map: A microblog-based enrichment approach. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, pp. 691–700.

Hays, J. & Efros, A. (2008). IM2GPS: Estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008*. Washington, DC, USA: IEEE Computer Society, pp. 1–8.

Hecht, B., Hong, L., Suh, B. & Chi, E. (2011). Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, pp. 237–246.

Heidelberger, P. & Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research* **31**, 1109–1144.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. & Tsioutsiouliklis, K. (2012). Discovering geographical topics in the Twitter stream. In *WWW '12: Proceedings of the 21nd International Conference on World Wide Web*. New York, NY, USA: ACM, pp. 769–778.

Ikawa, Y., Enoki, M. & Tatsubori, M. (2012). Location inference using microblog messages. In *WWW '12: Proceedings of the 21st International Conference on World Wide Web*. New York, NY, USA: ACM, pp. 687–690.

Intagorn, S. & Lerman, K. (2014). Placing user-generated content on the map with confidence. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Dallas, TX, USA: ACM, pp. 413–416.

Jaccard, P. (1901). Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles* **37**, 241–272.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA, USA: The AAAI Press, pp. 273–282.

Kass, R. & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kinsella, S., Murdock, V. & O'Hare, N. (2011). I'm eating a sandwich in Glasgow: Modeling locations with tweets. In *Proceedings of the Third International Workshop on Search and Mining User-Generated Contents*. New York, NY, USA: ACM, pp. 61–68.

Laere, O. V., Quinn, J., Schockaert, S. & Dhoedt, B. (2014). Spatially aware term selection for geotagging. *IEEE Transactions on Knowledge and Data Engineering* **26**, 221–234.

Laere, O. V., Schockaert, S. & Dhoedt, B. (2012). Georeferencing Flickr photos using language models at different levels of granularity: an evidence based approach. *Web Semantics: Science, Services and Agents on the World Wide Web* **16**, 17–31.

Laere, O. V., Schockaert, S. & Dhoedt, B. (2013). Georeferencing Flickr resources based on textual meta-data. *Information Sciences* **238**, 52–74.

Laing, M. (1991). Anchor texts and literary manuscripts in early Middle English. In *Essays Celebrating the Publication of the Linguistic Atlas of Late Mediaeval English*. Suffolk, UK: D.S. Brewer, pp. 27–52.

Lao, O., Lu, T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A., Gieger, C., Wichmann, H., Ruther, A., Schreiber, S., Becker, C., Nurnberg, P., Nelson, M., Krawczak, M. & Kayser, M. (2008). Correlation between genetic and geographic structure in Europe. *Current Biology* **18**, 1241–1248.

Li, R., Wang, S., Deng, H., Wang, R. & Chang, K. (2012). Towards social user profiling: Unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, pp. 1023–1031.

Liu, L., Huang, Y., Wang, J., Tang, Z., Lu, L., Wu, R. & Lei, Q. (2013). Study on discriminating flue-cured tobacco by volatile compounds related to geographical origin and cultivar. *Asian Journal of Chemistry* **25**, 7587–7592.

Lovins, J. B. (1968). Development of a stemming algorithm. *Translation and Computational Linguistics* **11**, 22–31.

McGee, J., Caverlee, J. & Cheng, Z. (2013). Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge management*. New York, NY, USA: ACM, pp. 459–468.

McIntosh, A. (1987). *Guide to A Linguistic Atlas of Late Mediaeval English*. Aberdeen, UK: Aberdeen University Press.

McIntosh, A., Samuels, M., Benskin, M., Laing, M. & Williamson, K. (1986). *A Linguistic Atlas of Late Mediaeval English*. Aberdeen, UK: Aberdeen University Press.

McIntosh, A., Samuels, M., Benskin, M., Laing, M. & Williamson, K. (2013). *An Electronic Version of A Linguistic Atlas of Late Mediaeval English*. http://www.lel.ed.ac.uk/ihd/elalme/elalme.html. Accessed 2016-12-20.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Mondol, S., Sridhar, V., Yadav, P., Gubbi, S. & Ramakrishnan, U. (2014). Tracing the geographic origin of traded leopard body parts in the Indian subcontinent with DNA-based assignment tests. *Conservation Biology* **29**, 556–564.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A., Auton, A., Indap, A., King, K., Bergmann, S., Nelson, M., Stephens, M. & Bustamante, C. (2008). Genes mirror geography within Europe. *Nature* **456**, 98–101.

OHare, N. & Murdock, V. (2013). Modeling locations with social media. *Information Retrieval* **16**, 30–62.

Petkova, D., Novembre, J. & Stephens, M. (2015). Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics* **48**, 94–100.

Pope, L., Butlin, R., Wilson, G., Woodroffe, R., Erven, K., Conyers, C., Franklin, T., Delahay, R., Cheeseman, C. & Burke, T. (2007). Genetic evidence that culling increases badger movement: Implications for the spread of bovine tuberculosis. *Molecular Ecology* **16**, 4919–4929.

Price, A., Helgason, A., Palsson, S., Stefansson, H., Clair, D., Andreassen, O., Reich, D., Kong, A. & Stefansson, K. (2009). The impact of divergence time on the nature of population structure: An example from Iceland. *Public Library of Science Genetics* **5**, e1000505.

Priedhorsky, R., Culotta, A. & Valle, S. D. (2014). Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. New York, NY, USA: ACM, pp. 1523–1536.

Pritchard, J., Stephens, M. & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Puckett, E. E. & Eggert, L. S. (2016). Comparison of SNP and microsatellite genotyping panels for spatial assignment of individuals to natal range: A case study using the American black bear (Ursus americanus). *Biological Conservation* **193**, 86–93.

Ranola, J., Novembre, J. & Lange, K. (2014). Fast spatial ancestry via flexible allele frequency surfaces. *Bioinformatics* **30**, 2915–2922.

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York, NY, USA: Springer Science & Business Media, 2nd ed.

Robert, C. & Casella, G. (2010). *Introducing Monte Carlo Methods with R*. New York, NY, USA: Springer, 1st ed.

Roberts, G. & Rosenthal, J. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 255–268.

Roberts, G. & Tweedie, R. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–363.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B. & Baldridge, J. (2012). Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1500–1510.

Rout, D., Bontcheva, K., Preoţiuc-Pietro, D. & Cohn, T. (2013). Where's @wally?: A classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. New York, NY, USA: ACM, pp. 11–20.

Sadilek, A., Kautz, H. & Bigham, J. P. (2012). Finding your friends and following them to where you are. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, pp. 723–732.

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers* **18**, 401–409.

Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J. & Mühlhäuser, M. (2013). A multi-indicator approach for geolocalization of tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA, USA: The AAAI Press, pp. 573–582.

Serdyukov, P., Murdock, V. & Zwol, R. V. (2009). Placing Flickr photos on a map. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, pp. 484–491.

Sokal, R. & Michener, C. (1958). A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin* **38**, 1409–1438.

Stadje, W. (1990). The collector's problem with group drawings. *Advances in Applied Probability* **22**, 866–882.

Tian, C., Kosoy, R., Lee, A., Ransom, M., Belmont, J., Gregersen, P. & Seldin, M. (2008). Analysis of East Asia genetic substructure using genome-wide SNP arrays. *Public Library of Science One* **3**, e3862.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika* **17**, 401–419.

Verdinelli, I. & Wasserman, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Annals of Statistics* **26**, 1215–1241.

Wasser, S., Brown, L., Mailand, C., Mondol, S., Clark, W., Laurie, C. & Weir, B. (2015). Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots. *Science* **349**, 84–87.

Wasser, S., Clark, W., Drori, O., Kisamo, E., Mailand, C., Mutayoba, B. & Stephens, M. (2008). Combating the illegal trade in African elephant ivory with DNA forensics. *Conservation Biology* **22**, 1065–1071.

Wasser, S., Clark, W. & Laurie, C. (2009). The ivory trail. *Scientific American* **301**, 68–76.

Wasser, S., Mailand, C., Booth, R., Mutayoba, B., Kisamo, E., Clark, B. & Stephens, M. (2007). Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 4228–4233.

Wasser, S., Shedlock, A., K.Comstock, Orstrander, E., Mutayoba, B. & Stephens, M. (2004). Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 14847–14852.

Wing, B. & Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 955–964.

Xing, J., Watkins, W., Shlien, A., Walker, E., Huff, C., Witherspoon, D., Zhang, Y., Simonson, T., Weiss, R., Schiffman, J., Malkin, D., Woodward, S. & Jorde, L. (2010). Toward a more uniform sampling of human genetic diversity: A survey of worldwide populations by high-density genotyping. *Genomics* **96**, 199–210.

Xing, J., Watkins, W., Witherspoon, D., Zhang, Y., Guthery, S., Thara, R., Mowry, B., Bulayeva, K., Weiss, R. & Jorde, L. (2009). Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Research* **19**, 815–825.

Yang, W.-Y., Novembre, J., Eskin, E. & Halperin, E. (2012). A model-based approach for analysis of spatial structure in genetic data. *Nature Genetics* **44**, 725–731.

Yang, W.-Y., Platt, A., Chiang, C. W.-K., Eskin, E., Novembre, J. & Pasaniuc, B. (2014). Spatial localization of recent ancestors for admixed individuals. *G3: Genes, Genomes, Genetics* **4**, 2505–2518.

Yang, Y. & Pedersen, J. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 412–420.