ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)
183, *Part* 1, *pp.* 251–280

'What drives commuter behaviour?': a Bayesian clustering approach for understanding opposing behaviours in social surveys

Laura C. Dawkins, Daniel B. Williamson, Stewart W. Barr and Sally R. Lampkin

University of Exeter, UK

[Received August 2018. Revised July 2019]

Summary. The city of Exeter, UK, is experiencing unprecedented growth, putting pressure on traffic infrastructure. As well as traffic network management, understanding and influencing commuter behaviour is important for reducing congestion. Information about current commuter behaviour has been gathered through a large on-line survey, and similar individuals have been grouped to explore distinct behaviour profiles to inform intervention design to reduce commuter congestion. Statistical analysis within societal applications benefit from incorporating available social scientist expert knowledge. Current clustering approaches for the analysis of social surveys assume that the number of groups and the within-group narratives are unknown *a priori*. Here, however, informed by valuable expert knowledge, we develop a novel Bayesian approach for creating a clear opposing transport mode group narrative within survey respondents, simplifying communication with project partners and the general public. Our methodology establishes groups characterizing opposing behaviours based on a key multinomial survey question by constraining parts of our prior judgement within a Bayesian finite mixture model. Drivers of group membership and within-group behavioural differences are modelled hierarchically by using further information from the survey. In applying the methodology we demonstrate how it can be used to understand the key drivers of opposing behaviours in any wider application.

Keywords: Bayesian modelling; Smart cities; Subjective priors; Survey analysis; Transport

1. Introduction

High levels of commuter congestion negatively impact both the quality of the environment and societal health and wellbeing. Reducing congestion is therefore a key challenge facing many UK cities in the 21st century. In particular, the city of Exeter, UK, is experiencing unprecedented economic and physical growth, with the population of greater Exeter set to increase by as much as 50% by 2026 (Exeter City Council, 2015). This growth will put further pressure on current infrastructure and presents a significant challenge in meeting and maintaining air quality standards (Exeter City Council, 2015). As a result, Exeter has become the test bed for a new 'smart cities' methodology which aims to reduce commuter congestion. This methodology is being developed within the Innovate UK 'Engaged smart transport' (EST) project, which is a collaboration between statisticians and social scientists at the University of Exeter, the City and County Councils, and a large consortium of industrial partners.

The concept of a smart city is to utilize smart technologies and city management systems to optimize sustainability, cost and service quality. In the context of reducing commuter congestion

Address for correspondence: Laura C. Dawkins, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4QF, UK.
E-mail: lauradawkins@hotmail.co.uk

this means optimizing the transport network over existing infrastructure; hence, as well as traffic network management in realtime, understanding and influencing commuter behaviour is an important part of any smart cities strategy. Information about current commuter behaviour in Exeter has been gathered through the large on-line 'Commute–Exeter' survey, containing over 40 questions, completed by over 3000 commuters. On the basis of these responses, similar surveyed individuals have been grouped to inform discussions about the design of group-specific interventions to reduce commuter congestion.

To achieve the aims of statistical analysis in societal applications most effectively it is important to incorporate social scientist expert knowledge where possible. In future stages of the EST project the survey analysis results will be used to inform workshop discussions with survey respondents from each group, to gain further insight into how transport behaviour may be influenced through group-specific behavioural interventions. In this application, therefore, the ability to communicate results with the general public clearly was extremely important. Through close collaboration with social scientists we identified that, to achieve this, creating clear and simply defined groups we could confidently address within these workshops was essential.

To meet this need, we present and apply a novel Bayesian methodology for analysing social surveys in which clear groups, characterizing a single opposing principal behaviour, are created based on a key multinomial survey question by constraining parts of our prior judgement within a Bayesian finite mixture model. Regression on additional survey questions is then used to understand the key drivers of these opposing behaviours. For the Commute–Exeter survey, this method groups survey respondents on the basis of an opposing transport mode narrative, e.g. 'cyclists', and explores which factors motivate mode choice within that group. These factors can then be the focus of further workshops, focus groups and potential behavioural interventions to reduce commuter car usage, and hence congestion.

Current statistical approaches for the analysis of social surveys assume that the number of groups and the within-group narratives are unknown *a priori*. These methodologies therefore focus on identifying these unknown groups. This can be a powerful approach in certain situations, but in some cases can create complicated and unclear group narratives *a posteriori*. When, as in this application, the survey analysis will be used to direct communication and to motivate future experiments, it can be important to provide a clear structural narrative to take forward to the next phase of the study.

In the social sciences, methods for forming groups, known as 'segmentation' (Barr and Prillwitz, 2011; Anable, 2005), often follow heuristic, non-model based approaches. One such approach, known as hierarchical agglomerative clustering, creates a hierarchy of clusters by initially treating each individual as a singleton cluster and subsequently successively merging (or agglomerating) the most similar clusters on the basis of a measure of distance (Wheeler *et al.*, 2004). The final groupings are created by identifying an appropriate level of the hierarchy to stop merging. A variety of other similar approaches exist, each often resulting in different groupings, with little systematic guidance available for determining the optimal number of groups or the most appropriate methods for the specific application (Fraley and Raftery, 2002). In addition, since these heuristic methods are not based on statistical models, formal inference is impossible and complex group narratives are created post segmentation. For example, Anable (2005) used principal component analysis and hierarchical clustering to group transport behaviour in survey respondents, identifying six groups differing subtly in their attitudes, requiring multiple descriptive paragraphs to explain each group narrative. To motivate the development, and to demonstrate the added value of our approach, in Dawkins *et al.* (2018), we presented the results of applying a classical heuristic, non-model-based segmentation procedure, similar to that of

Anable (2005), further demonstrating how such an approach leads to complex group narratives with no formal inference about group behaviours.

Alternatively, model-based clustering approaches for survey analysis allow for formal inference. The response variable on which the groups are based is viewed as coming from a mixture of underlying probability distributions, each representing a different group (Fraley and Raftery, 1998). In a Bayesian modelling framework this can be viewed as either a finite or infinite mixture of probability distributions. In most cases, Bayesian finite mixture models require the specification of the number of mixture components, or groups, *a priori*. Existing examples assume that the number of groups is unknown and therefore most commonly infer this prior specification based on marginal likelihoods or Bayes factors (Lau and Green, 2007). For example Fahey *et al.* (2007) used the Bayesian information criterion to determine the optimal number of groups when analysing dietary patterns based on the UK National Diet and Nutrition Survey. Again, this approach often leads to complex group narratives; for example Fahey *et al.* (2007) identified six groups, with narratives differing intricately in their consumption of multiple different food types. More recently, Malsiner-Walli *et al.* (2016) presented the 'sparse finite mixture model', in which the number of finite mixture components is determined during, rather than before, model fitting. This approach, however, requires a common group allocation probability and therefore does not facilitate regression on additional variables to explore which characteristics influence group membership.

Conversely, Bayesian infinite mixture models assume an infinite number of mixture components, with the total number of occupied groups inferred from the data during model fitting (Kim *et al.*, 2006). This approach often leads to a large number of groups containing a small percentage of individuals, complicating group narratives further. For example, using this approach, Crépet and Tressou (2011) identified 17 groups within the Individual and National Study on Food Consumption, differing in dietary consumption behaviour, with only three of the groups accounting for 98% of the individuals surveyed. Similarly, Muthukumarana and Swartz (2014) employed this approach to group students on the basis of surveyed course satisfaction, identifying 10 groups within just 75 surveyed students, with many groups differing only slightly in their satisfaction response profiles.

Within the Commute–Exeter survey, each individual was asked about their day-to-day commuting pattern; the number of days, of the 20 weekdays in a 4-week period, that they commute by five transport mode types; motor vehicle (MV), public transport (PT), bicycle, on foot or a combination of modes within one journey. We apply our novel prior constrained Bayesian finite mixture model methodology to create a simple five-group narrative in which each group contains survey respondents who principally commute by each of the five transport modes.

We recognize and model within-group heterogeneity by using the 'mixture-of-experts' extension of the finite mixture model, in which the model parameters are represented as functions of additional covariates, e.g. gender, income and personal attitudes. In doing so, we can identify the key drivers of membership within each of these clearly defined opposing behavioural groups and the key drivers of within-group behavioural differences. In a similar way, Gormley and Murphy (2008a, b) used a mixture-of-experts approach to explore the heterogeneity in voting behaviour within groups of the Irish electorate, Frühwirth-Schnatter (2011) to identify which factors had a significant effect on the risk of marijuana usage, Pamminger and Frühwirth-Schnatter (2010) and Frühwirth-Schnatter *et al.* (2016) to investigate how observable characteristics correlated with career mobility profile group membership, and Fahey *et al.* (2007) to understand the association between dietary pattern group membership and a selection of additional factors. In each case, the mixture-of-experts extension to the finite mixture model was used to provide greater insight into the relationships within the survey responses, as is particularly relevant in

this application for identifying possible intervention themes to influence commuter transport behaviour to reduce congestion.

This prior-constrained Bayesian methodology falls in between two opposing approaches: the commonly used unconstrained finite mixture model, and an *a priori* assignment approach, in which the respondents are grouped on the basis of their observed most frequent transportation mode and group behaviour is explored through multinomial logistic regression applied separately within each group. Both of these alternative approaches were implemented. When attempting to apply the unconstrained finite mixture model, Bayesian model fitting encountered the common issue of mixture component non-identifiability (as discussed by Jasra *et al.* (2005), Sperrin *et al.* (2010), Pamminger and Frühwirth-Schnatter (2010), Frühwirth-Schnatter (2011) and Frühwirth-Schnatter *et al.* (2012)), i.e., in the unconstrained model, equivalent prior distributions are placed on the parameters of each mixture component, resulting in identical marginal posterior distributions in each group. As a result, the statistical inference is invariant to permutations in the group labelling of the parameters and hence these labels can switch during Markov chain Monte Carlo (MCMC) sampling. For example, for part of the MCMC sampling chain, 'group 1' may characterize the 'predominantly commute by bicycle group', whereas another part of the chain characterizes the 'predominantly commute by PT group'. This makes it meaningless to draw inference directly from MCMC output (Jasra *et al.*, 2005) and hence requires an additional correction step (Frühwirth-Schnatter, 2011). In addition, this approach was found to encounter MCMC convergence issues making model fitting extremely challenging, and in all attempts the desired clear group narrative was unattainable. Conversely, the opposing *a priori* assignment approach is simple to fit but misses subtleties in group allocation achieved when using the full mixture model, i.e., in our approach, when grouping individuals on the basis of their principal behaviour, individuals at the intersection of two groups are not always allocated to the group characterizing the transport mode that they most frequently commute by, but the group that they are principally similar to in terms of all survey responses. Hence, these individuals will receive a more relevant group intervention.

Our methodology provides a general approach with which to understand the key drivers of opposing surveyed behaviours within any application, avoiding the complex group narratives that commonly arise when applying existing approaches. For example, it could be used to explore what influences individuals to shop primarily at competing supermarkets, to eat primarily healthy or unhealthy foods or to purchase primarily opposing brands of a product.

Alternative work in the field of Bayesian inference for survey data focuses on developing models that account for the disproportionality of the survey sample with respect to the target population, which is known as survey sample bias (e.g. Nandram *et al.* (2013), Si *et al.* (2015) and Wang *et al.* (2015)). As noted by Gelman (2007), Kunihama *et al.* (2016) and Kang and Bernstein (2016), there is an apparent disconnect in the literature between Bayesian modelling developments to account for sample bias, which are all based on simplistic single-variable survey response data, and the application of Bayesian models to analyse real world survey data. This is evident in the examples that were discussed throughout this introduction, in which the sampling mechanisms are either not mentioned, assumed to be random or previously corrected for by the survey administrator. This topic will be discussed further in the context of the Exeter commuter application in Section 6.

The remainder of the paper is organized as follows. The survey data are described in detail in Section 2. Section 3 presents our novel methodology including the social scientist expert informed constrained prior model, and model fitting approaches used. Section 4 presents the results and Section 5 the model validation. In Section 6 the issue of survey sample bias is discussed in detail in the context of the application. Finally, Section 7 concludes.

2. The survey

The survey was developed on the web-based platform 'www.surveymonkey.com', containing 42 questions (which are available in supplementary material: `https://www.dropbox.com/s/95282pdtulf0ext/Dawkins_SupMat.zip?dl=0`). The aim of the survey was to understand the key drivers of commuting behaviour in Exeter. Individuals were therefore asked about their day-to-day commuting pattern, the number of days, of the 20 weekdays in a 4-week period, that they commute by each of the five transport mode types: MV, PT, bicycle, on foot or a combination of modes within one journey. Details about these choices were attained from further questions, e.g. the time that they make their transport mode choice, their attitudes towards weather and traffic congestion and values that they hold about cost, personal fitness and the environment. Additional questions were asked about simple demographics such as age, gender, employer and home and work postcodes.

Participation in the survey was voluntary, based on a widespread marketing campaign, incentivized by an iPad prize. The marketing campaign used two visual images and catchphrases: one targeted at road users, in particular MV users, shown in Fig. 1(a), and the other at non-road-users and the wider community, shown in Fig. 1(b). The non-road-user advertisement was concentrated in bus shelters, train stations and supermarkets, whereas the road user advertising was primarily placed on billboards on main commuting roads and central car parks. The survey was highly publicized on Twitter and by the local media via radio and newspaper.

The survey was open to the public for 7 weeks, from Monday, May 2nd, 2016, until Sunday, June 19th, 2016, receiving 3050 responses. These responses were explored for quality control to ensure sensible and consistent answers. Respondents were removed if they did not complete the key commuting behaviour question, responded as commuting fewer than 5 or more than 20 days in the 4-week period, or specified a work postcode outside the census local authority Exeter boundary. Where possible obvious mistakes were corrected for to minimize the loss of responses. After this quality control process 2648 responses remained in the analysis. Of these, 2500 were used to fit the model and 148 were withheld for model validation.

Survey skip logic, which takes the respondent to the next question on the basis of how they answered the current question, resulted in intentional missing data within the survey responses. For example those individuals who always commuted by a single transport mode missed questions related to day-to-day mode choice. These intentional missing data were therefore not imputed

(a) (b)

Fig. 1. Survey marketing material focused at (a) road users and (b) non-road-users

but treated as missing, meaning that individuals who skipped a given question were not included in the related parameter inference.

Responses to the survey questions were generally ordinal, i.e. selecting from five options on an increasing–decreasing scale, e.g. from 'strongly disagree' to 'strongly agree', or 'terrible'–'excellent'. Throughout, all responses, except for gender (which is treated as binary), were considered to be numeric and continuous to reduce the number of required parameters in this already complicated model framework. Treating this type of response scale, which is known as the Likert scale, in this way has become common practice in the social sciences literature (Jamieson, 2004). Although some argue that this simplification may produce errors in interpreting inference, many consider that using this approach is superior; see for example Kerlinger and Lee (2000) and Liang and Tsai (2008). As in these examples, we feel that we can make this simplification because, on the basis of the expertise of our collaborating social scientists, we believe that the survey respondents interpret these ordinal responses as being equally spaced in meaning, and hence that treating them as interval level measurements with linear effects is reasonable. In addition, any responses that did not obviously fit into the continuous scale were removed from the analysis, i.e. treated as missing. For example, when rating infrastructure around the city, 'not familiar' was removed from the continuous scale of responses 'terrible'–'excellent'.

The dimensionality and computational burden of model fitting were reduced by initially carrying out a 'variable classification' step. Social scientist expert judgement was used to choose which of the 42 survey questions are most relevant for inclusion within the mixture-of-experts model as either an explanatory variable for group membership or for the clustering variable.

Those classified by the social scientist as being relevant for explaining membership within the opposing behavioural groups (containing people who principally commute by each of the five transport modes) are termed 'group identifiers' (GIs), defined as those survey questions that best explain the differences between groups, and characterize factors that are non-influenceable in the intervention stage. The 12 survey questions classified as being GIs are presented in Table 1. Commute distance is an example of a selected GI, since individuals in the groups that are characterized by primarily walking or cycling will most probably have shorter commute distances, and home and work locations cannot be altered by interventions.

Those survey questions classified as being relevant for explaining the clustering variable, here the number of days that individuals commute by using each of the five transport modes, are termed 'behavioural influencers' (BIs) and are chosen as those that are thought to explain best the differences in behaviours between individuals within each group. Since our aim is to identify which of these factors have the most influence on transport mode usage and could therefore be used in the intervention stage of the project, BIs are chosen to characterize factors that could be influenceable in the intervention stage. The 16 survey questions that are classified as being BIs are presented in Table 2. Concern for the environment is an example of a selected BI, since individuals in a given group who are more concerned with the environment may use more sustainable transport modes more often, and environmental awareness could be encouraged through a social intervention.

We experienced memory and MCMC convergence issues when fitting the model using all classified GIs and BIs. We therefore used the automated Bayesian variable-selection approach of Kuo and Mallick (1998), applied to a subset of survey respondents, to reduce the number of covariates in the model. We allocated each respondent to a group by using a Bayesian finite mixture model as defined in Section 3 (excluding covariate information), and applied the Kuo and Mallick (1998) variable-selection approach to regression models for group allocation, and transport behaviour within each group separately. This approach embeds an indicator variable, associated with each covariate, within the regression equation. The indicator variables are treated

Table 1. Survey questions subjectively selected as GIs†

Number	*Survey question*	*Responses*	*z*
1	What is your gender?	Male; female	(0, 1)
2	What is your age?	Integers (17,. . .,70)	Integers (17,. . .,70)
3	When are you most likely to make the decision about how you commute to or from your place of work or study?	At time of leaving; in preceding hour before leaving; the night before; during the preceding weekend	(0,1,2,3)
4	How much flexibility do you have over the time you leave for your commute to and from your place of work or study?	None; a little; some; a lot; total	(0,1,2,3,4)
5	Do you attempt to avoid peak travel times?	Never; rarely; some of the time; most of the time; always	(0,1,2,3,4)
6	Which of the following best describes your place of work?	Large; medium; small; self-employed	(0,1,2,3)
7	Home and work postcodes are used to calculate commute distance in kilometres	Continuous (0,. . .,150)	Continuous (0,. . .,150)
8	Thinking about the parking facilities you have at or near your home, which *one* of the following best describes them?	Off- or on-road parking near your property; on-road parking immediately outside your property; off-road driveway or garage on your property	(0,1,2)
9	Thinking about the parking facilities that may be available at or near your place of work or study, please rate them below	No parking; not adequate; satisfactory; good; excellent	(0,1,2,3,4)
10	What is the highest formal qualifications level you have?	None; Certificate of Secondary Education, or O-level or General Certificate of Secondary Education; A-level or further education college; university or higher education undergraduate degree; postgraduate, Masters or doctoral qualification	(0,1,2,3,4)
11	Which one of the following income bands does your household fall into?	Under £15001; £15001–£30000; £30001–£45000; £45001–£60000; £60001–£75000; £75001–£90000; £90001–£105000; £105001–£120000; over £120000	(0,1,2,3,4,5,6,7,8)
12	Are you actively involved in any of the following?	Local place of worship; local community centre; community or volunteer group; special interest group or club; local gymnasium sports club or leisure centre; local fund raising events	Sum of all those ticked (0−6)

†The left-hand column states the survey question, numbered 1–12, the middle column shows the possible responses to the associated question, and the right-hand column, the GI covariate values given to each response.

as random within the model and the Bayesian posterior median of each of these binary inclusion parameters identifies which covariates are active within the model. The resulting selected GI and BI survey questions are shown in Table 3. The final model is therefore based on 14 parameters for group allocation (the second column of Table 3) and 17 parameters for transport behaviour (the fourth column of Table 3).

Table 2. Survey questions subjectively selected as BIs†

Number	*Survey question*	*Responses*	*x*
1	How much does receiving information about weather conditions influence your choice of travel mode to your place of work or study?	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
2	How much does receiving information about traffic congestion influence your choice of travel mode to your place of work or study? On a day when the following weather conditions occur, how much does each one influence your choice of travel mode to your place of work or study?	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
3	Ice	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
4	Rain	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
5	Wind	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
6	Storm	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
7	Snow	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
8	Cold	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
9	Warm	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
10	Dark mornings	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
11	Dark evenings	Never; occasionally; sometimes; often; always	(0,1,2,3,4)
12	Cost is a major consideration when you choose how to commute	Strongly disagree; disagree; neutral; agree; strongly agree	(−2,−1,0,1,2)
13	Being environmentally friendly in your choice of travel mode is important to you	Strongly disagree; disagree; neutral; agree; strongly agree	(−2,−1,0,1,2)
14	Having control over your commute is important to you	Strongly disagree; disagree; neutral; agree; strongly agree	(−2,−1,0,1,2)
15	Saving time on your commute is important to you	Strongly disagree; disagree; neutral; agree; strongly agree	(−2,−1,0,1,2)
16	Keeping fit and active is important to you	Strongly disagree; disagree; neutral; agree; strongly agree	(−2,−1,0,1,2)

†The structure of the table is the same as in Table 1.

There has been a recent increased interest in variable selection in model-based clustering due to the increasingly frequent use of high dimensional data sets in applied statistical research (Maugis *et al.*, 2009). Further, as discussed by Fop and Murphy (2018), it has been shown that including superfluous variables in model-based clustering can lead to identifiability problems and overparameterization. Indeed, even in data with moderate or low dimensionality, reducing the set of variables in the clustering process has been shown to be beneficial (Fowlkes *et al.*, 1988). In many cases, this variable selection is performed by using a so-called 'wrapper' method, in which variable selection is achieved simultaneously with model fitting (Fop and Murphy (2018)

Table 3. Survey questions selected for the model†

Group	*Selected GIs*	*Mode*	*Selected BIs*
1	(Used as baseline group)	2	2
		3	1
		4	1
		5	—
2	3; 9	2	1; 2; 13
		3	—
		4	—
		5	—
3	1; 3; 4; 5; 7	2	16
		3	1; 2; 9
		4	7; 13
		5	16
4	4; 5; 6; 7; 9	2	—
		3	5; 16
		4	1
		5	—
5	3; 9	2	12
		3	—
		4	—
		5	—

†For GIs, selection is based on the log-odds ratio of being in a given group rather than group 1 (predominantly MV users). For BIs, within each group, selection is based on the log-odds ratio of using a given transport mode rather than mode 1 (MV). The numbers of the selected GIs and BIs relate to the questions in Tables 1 and 2 respectively.

provide a comprehensive review of these approaches). Here, however, when a Bayesian shrinkage wrapper approach was attempted, the computational burden of including a large number of variables within the full model prohibited model fitting, necessitating that variable selection be carried out before model fitting. Since in this application we use the additional survey questions to understand behaviours within *a-priori*-defined groups, using a wrapper approach, in which variables are also selected on the basis of their importance in determining the group structure, is less important. A full understanding of these encountered model limitations will be an important area of future investigation.

3. The prior constrained Bayesian finite mixture model

Our methodology facilitates the understanding of the key drivers of r opposing principal behaviours within a large social survey, e.g. primarily shopping at $r = 4$ different supermarkets or, in this application, primarily commuting by $r = 5$ different transport modes.

Suppose that we have surveyed n individuals, each of which has responded to a key survey question, termed the structural response, e.g., in this application,

> 'How many days on average, in a typical four week period (Mondays to Fridays), do you commute to and from your place of work/study using some or all of the following travel mode options: Motor Vehicle; Public Transport; Bicycle; On Foot; A combination of modes?'.

Let the response to this question be denoted $y_i = (y_{i1}, \ldots, y_{ir})$, where y_{ij} is the frequency with which individual i behaves as in option j. For example, here y_{ij} is the number of days, out of a

possible 20, that individual i commutes by using transport mode j, taking $r = 5$ possible states (1, MV; 2, PT; 3, bicycle; 4, on foot; 5, combination).

Our methodology allocates the n surveyed individuals into one of $H = r$ groups, each characterizing principally behaving in one of the r opposing ways, by constraining key parts of the prior judgement about the structural response y. Hence, in this application, individuals are placed into one of $H = 5$ groups, each characterizing principally commuting by one of the five transport modes. Drivers of group membership and within-group behavioural differences are then inferred hierarchically on the basis of additional survey questions, selected by expert judgement and Bayesian variable selection, creating a clear narrative to aid communication and understanding, as described in Section 2.

A latent indicator random variable $S = (S_1, \ldots, S_n)$ is introduced to represent the group allocation for each individual $i = 1, \ldots, n$. The probability density function for the structural response of individual i is modelled as a sum over these H groups as

$$p(y_i|x_i, z_i) = \sum_{h=1}^{H} \Pr(S_i = h|z_i, \alpha)\, p(y_i|x_i, \beta_h), \tag{1}$$

where z_i and x_i are responses to the additional survey questions that are used as covariates related to the group allocation and structural response model parameters respectively, and α and β are associated regression coefficients quantifying the characteristics of the groups.

Each discrete, categorical random variable S_i, $i = 1, \ldots, n$, is modelled as a single draw from a multinomial distribution:

$$S_i \sim \text{Multinom}(1; \eta_{i1}, \ldots, \eta_{iH}), \tag{2}$$

where η_{ih} is the probability that individual i is in group h, $h = 1, \ldots, H$. For each individual, the relationship between the additional selected survey questions and the probability of group membership is therefore modelled via multinomial logistic regression:

$$\log\left(\frac{\eta_{ih}}{\eta_{i1}}\right) = z_i^{\mathrm{T}} \alpha_h$$

$$\Rightarrow \eta_{ih} = \frac{\exp(z_i^{\mathrm{T}} \alpha_h)}{\sum_{l=1}^{H} \exp(z_i^{\mathrm{T}} \alpha_l)} \qquad \text{for } h = 1, \ldots, H, \tag{3}$$

where $z_i = (1, z_{i1}, \ldots, z_{iD})$ is the vector of responses to D survey questions, given by individual i, related to the group membership probabilities, previously introduced as GIs. For identifiability, group 1 is used as the baseline group, requiring that $\alpha_1 = (0, \ldots, 0)$, whereas $\alpha_2, \ldots, \alpha_H$ are group-specific, unknown parameter vectors. Each of these vectors is made up of $D + 1$ regression coefficients $\alpha_h = (\alpha_{h0}, \alpha_{h1}, \ldots, \alpha_{hD})$ for $h = 2, \ldots, H$, where α_{h0} represents the baseline log-odds ratio of being in group h rather than group 1 and the remaining coefficients quantify the effect of the D covariate responses on this log-odds ratio.

The structural response for individual i, y_i, is modelled as m_i draws from a multinomial distribution, where m_i is the number of times that individual i makes the choice between the r different behaviour options. In this application m_i represents the overall number of days that individual i commutes during a 20-weekday period. For example if $y_i = (12, 0, 4, 0, 0)$, individual i commutes for $m_i = 16$ days, using an MV for 12 of these days and a bicycle for the remaining 4 days.

For each individual, y_i is modelled separately for membership within each group. Hence, when individual i is assigned to group h,

$$y_i|S_i = h \sim \text{Multinom}(m_i; \theta_{ih1}, \ldots, \theta_{ihr}), \tag{4}$$

where θ_{ihj} is the probability that individual i behaves as in option j, $j = 1, \ldots, r$, on any given occasion when assigned to group h, $h = 1, \ldots, H$. Hence, in this application θ_{ihj} represents the probability that individual i uses transport mode j when assigned to group h. For each individual, the relationships between the additional selected survey questions and the r group parameters are also modelled via multinomial logistic regression:

$$\log\left(\frac{\theta_{ihj}}{\theta_{ih1}}\right) = x_i^{\mathrm{T}}\beta_{hj}$$

$$\Rightarrow \theta_{ihj} = \frac{\exp(x_i^{\mathrm{T}}\beta_{hj})}{\sum_{k=1}^{r}\exp(x_i^{\mathrm{T}}\beta_{hj})} \qquad \text{for } h = 1, \ldots, H, \quad j = 1, \ldots, r, \tag{5}$$

where $x_i = (1, x_{i1}, \ldots, x_{iC})$ is the vector of responses to C survey questions, given by individual i, related to the structural response probabilities, previously introduced as BIs. For identifiability, behavioural option 1 is used as the baseline, requiring that $\beta_{h1} = (0, \ldots, 0)$, whereas $\beta_{h2}, \ldots, \beta_{hr}$ are group- and mode-specific, unknown parameter vectors. Each of these vectors is made up of $C + 1$ regression coefficients $\beta_{hj} = (\beta_{hj0}, \beta_{hj1}, \ldots, \beta_{hjC})$ for $h = 1, \ldots, H$ and $j = 2, \ldots, r$, where β_{hj0} represents the baseline log-odds ratio of taking option j rather than option 1 in group h and the remaining coefficients quantify the effect of the C covariate survey questions on this log-odds ratio. In this application MV (transport mode option 1) is used as the baseline mode; hence β_{hj0}, for $j = 2, \ldots, r$, represents the baseline log-odds ratio of using transport mode j rather than MV in group h and the remaining coefficients quantify the effect of the C covariate survey questions on this log-odds ratio of transport mode usage.

3.1. Constrained prior modelling

To establish groups characterizing opposing behaviours, prior constraints are placed on the β regression coefficients. Let θ_{hj} represent the probability of any individual in group h taking behavioural option j. We specify these constraints such that $\theta_{hh} > \theta_{hj}$ for $h \neq j$, i.e., in group 1, the prior probability of taking behavioural option 1 is greater than the prior probability of taking any other option, equivalently for option 2 in group 2, and so on. For example, group 1 could characterize primarily shopping at supermarket chain 1, and group 2 primarily shopping at supermarket chain 2 etc. In our commuting application we create the opposing behavioural group narrative such that, in group 1, the prior probability of commuting by MV is greater than the prior probability of commuting by any other mode, equivalently for public transport in group 2, bicycle in group 3, on foot in group 4 and by a combination of modes in group 5.

Let $x = (1, x_1, \ldots, x_C)$ be any possible combination of responses to the C structural response covariate survey questions. By equation (5), the desired constraints that $\theta_{hh} > \theta_{hj}$ for $h \neq j$ can be represented in terms of β, for group 1 ($h = 1$), as

$$\begin{aligned} &\theta_{1j} < \theta_{11} \qquad \text{for } j = 2, \ldots, r, \\ &\Rightarrow \exp(x^{\mathrm{T}}\beta_{1j}) < \exp(x^{\mathrm{T}}\beta_{11}), \\ &\Rightarrow x^{\mathrm{T}}\beta_{1j} < 0, \end{aligned} \tag{6}$$

since $\beta_{h1} = (0, \ldots, 0)$ for $h = 1, \ldots, H$. Similarly for groups 2–5 ($h = 2, \ldots, 5$), using inequalities in terms of θ_{h1} to give simple positive or negative constraints,

$$\begin{aligned} &\theta_{hj} > \theta_{h1} \qquad \text{for } h = j, \\ &\Rightarrow \exp(x^{\mathrm{T}}\beta_{hj}) > \exp(x^{\mathrm{T}}\beta_{h1}), \\ &\Rightarrow x^{\mathrm{T}}\beta_{hj} > 0, \end{aligned} \tag{7}$$

and

$$\begin{aligned} &\theta_{hj} < \theta_{h1} \qquad \text{for } h \neq j, \\ &\Rightarrow \exp(x^{\mathrm{T}}\beta_{hj}) < \exp(x^{\mathrm{T}}\beta_{h1}) \\ &\Rightarrow x^{\mathrm{T}}\beta_{hj} < 0. \end{aligned} \tag{8}$$

These constraints create boundaries within the multi-dimensional β-space, beyond which samples from the priors of the β regression parameters are discarded within the Bayesian model fitting MCMC algorithm (Section 3.2.2).

As discussed in Section 1, the unconstrained finite mixture model can encounter non-identifiability issues from group label switching during MCMC sampling. Jasra *et al.* (2005) and Sperrin *et al.* (2010) reviewed various solutions to the problem falling into three main categories: identifiability constraints, deterministic relabelling algorithms and probabilistic relabelling algorithms. Identifiability constraints are inequalities placed on the parameters of the model during model fitting, breaking the symmetry in the likelihood, whereas deterministic and probabilistic relabelling algorithms are methods that are used to cluster the MCMC samples of each parameter, correcting the label switching, during or after model fitting.

The prior constraints equation (6)–(8) are a form of identifiability constraint, creating a non-symmetric prior so that the model is no longer invariant to a rearrangement of the group assignments, preventing label switching. Applying the alternative, relabelling algorithms, requires additional computation during or after MCMC sampling (Frühwirth-Schnatter, 2011); hence avoiding this is an additional, important computational advantage of our methodology when modelling large, complex survey data. These constraints relate directly to the desired predominant transport mode group structure, represented in terms of θ_{h1} (the probability of commuting by MV) to create simple inequalities above and below 0 (since $\beta_{h1} = (0, \ldots, 0)$ for $h = 1, \ldots, H$), which can be easily coded within the selected Bayesian modelling language, Stan. It may be, however, that only a subset of these constraints is required to create the desired predominant transport mode groups. For example, the constraints in equation (7) may be adequate for ensuring that each transport mode 2–5 is predominant in groups 2–5, or alternatively including constraints for four of the five groups may be adequate for constraining all five groups. In this analysis all constraints were included for completeness; however, valuable future work could involve exploring the effect of omitting selected constraints on the group structure, and the computational time that is saved in doing so.

3.2. Model fitting

3.2.1. Prior distributions

We place a normal prior on each α regression coefficient parameter, following the usual conjugate model that was discussed by Garthwaite *et al.* (2005):

$$\alpha_{hd} \sim N(\gamma_{hd}, \xi_{hd}), \qquad \text{for } h = 2, \ldots, H, \quad d = 0, 1, \ldots, D. \tag{9}$$

In this application, all α regression coefficients are assumed unknown and inferred from the data; hence we assign them relatively uninformative priors, such that $\gamma_{hd} = 0$ and $\xi_{hd} = 5$ for $h = 2, \ldots, H$, and $d = 1, \ldots, D$.

We also place a normal prior on each β regression coefficient parameter (Garthwaite *et al.*, 2005),

$$\beta_{hjc} \sim N(\mu_{hjc}, \sigma_{hjc}) \qquad \text{for } h = 1, \ldots, H, \quad j = 1, \ldots, r, \quad c = 0, 1, \ldots, C, \tag{10}$$

conditionally on prior constraints (6)–(8) (Section 3.1) being satisfied, with means and standard deviations elicited to reflect the constraints to reduce the number of discarded prior samples, as follows.

The coefficients $(\beta_{hj1}, \ldots, \beta_{hjC})$ represent the effect of the C mode influencer survey questions on the log-odds ratio of taking mode j rather than MV (mode 1). Since these effects are unknown *a priori*, to be inferred from the model, we specify relatively uninformative priors, such that, in equation (10), $\mu_{hjc} = 0$ and $\sigma_{hjc} = 5$ for $h = 1, \ldots, H$, $j = 2, \ldots, r$, $c = 1, \ldots, C$.

The remaining regression coefficients that are associated with the intercept terms, β_{hj0}, represent the baseline log-odds of using transport mode j rather than MV (mode 1) within group h. Hence these prior parameters μ_{hj0} and σ_{hj0} for $h = 1, \ldots, H$ and $j = 2, \ldots, r$ are elicited to reflect the prior constraints.

Social scientist expert judgement is used to define the overall group structure and to select application appropriate survey questions; however, as discussed by Garthwaite *et al.* (2005), experts often struggle to think of regression coefficients directly. Garthwaite *et al.* (2013) developed a method for eliciting priors from experts for coefficients in generalized linear models. For complex hierarchical Bayesian models such as ours, the role of expert and statistician is increasingly blurred. Many expert-subjective prior choices are made, not by those normally thought of as 'the experts' (in this case the social scientists), but by the statisticians who better understand the role and effect of certain parameters in the model. What are often described as 'modelling choices' are really the expert judgements of statisticians, leading to a prior uncertainty description that is owned by the scientist–statistician team, rather than either individual (see Williamson and Goldstein (2015) for a detailed discussion of these foundational concepts).

Given the project time constraints and the relative experiences of our team with Bayesian hierarchical models of this type *versus* complex transport surveys of this size, it was decided that prior choices for the covariate coefficient parameters would be made by the statisticians in the team, Dawkins and Williamson. To assist in making these specifications we developed an interactive web application via the RStudio Shiny package (RStudio, 2013). This Shiny package enables the user to explore how the probability of each category within a multinomial logistic regression model changes depending on the number and range of covariates and the values of the normal prior parameters for the regression coefficients, using sliders and drop-down menus. Once these values have been specified, the Shiny application simulates a large number (default 1000) of regression coefficients from the specified normal priors and uses these along with every possible integer combination of the specified covariates to produce a distribution for these multinomial probabilities. A screen shot of this R Shiny web application is presented in Fig. 2, with $r = 5$ multinomial categories and $C = 3$ covariates: two in the range $(0, \ldots, 4)$ and one in the range $(-2, \ldots, 2)$. The R code that is required to run this Shiny application is available from

`http://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`.

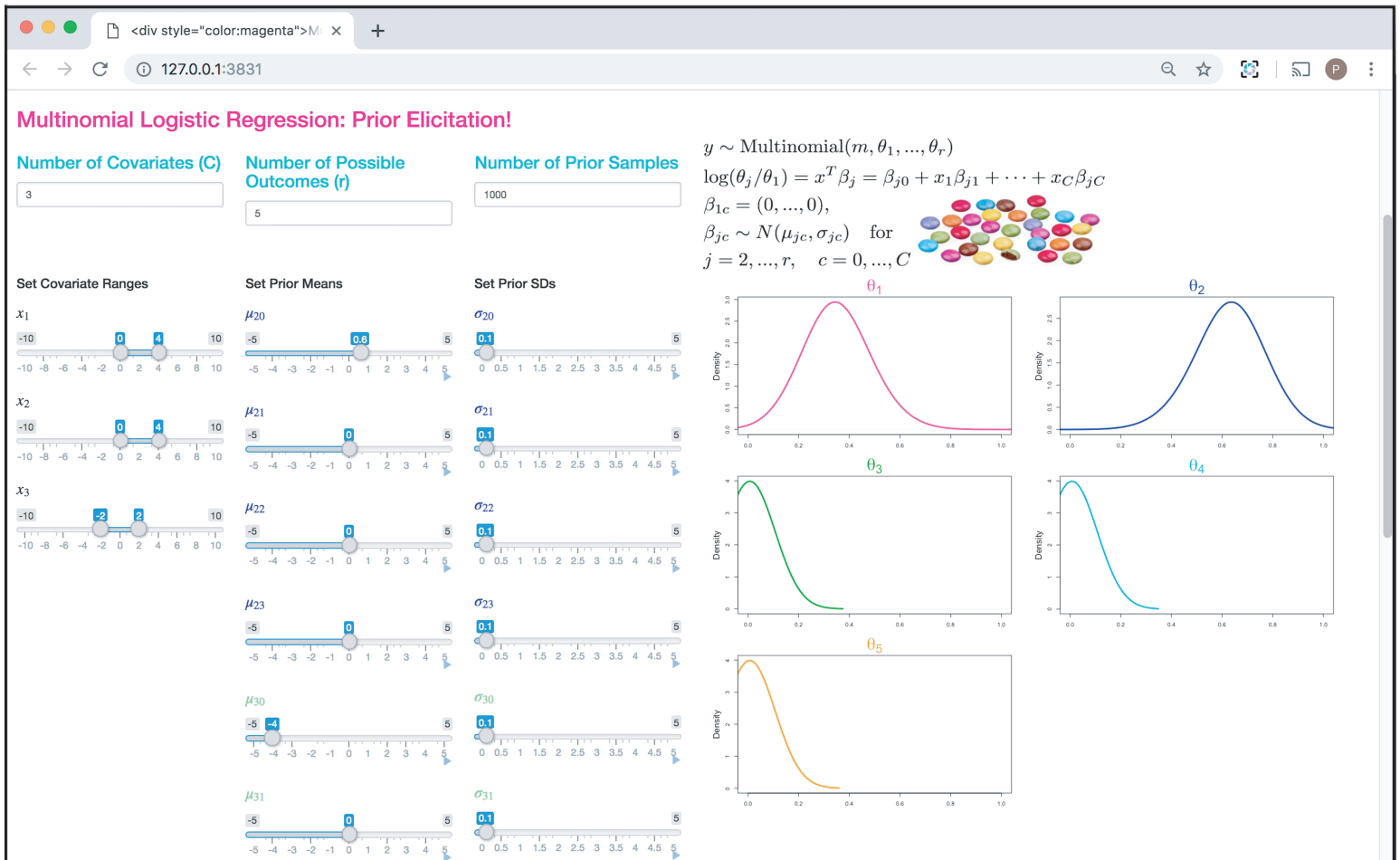

Fig. 2. Screen shot of the Shiny interactive web application used for prior elicitation: the sliders for μ_4, μ_5, σ_4 and σ_5 are viewed by scrolling down the web page; the R code required to run this Shiny application is available from https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets

For each of the five groups individually ($h = 1, \ldots, 5$), this Shiny application was used to select values for the prior parameters μ_{hj0} and σ_{hj0} for $j = 2, \ldots, r$. We based our choices on the principle that the resulting model would represent the predominant transport mode of that group, i.e. $\theta_{hh} > \theta_{hj}$ for $h \neq j$, when all other regression coefficients were approximately 0 (i.e. $\mu_{hj0} = 0$ for $j \neq h$). The values for the parameters in question were varied via the sliders in Fig. 2 and the resulting distributions for $\theta_1, \ldots, \theta_5$, representing the probability of using each of the five transport modes within the group of interest, were observed. For example, Fig. 2 shows a prior specification for the 'predominantly commute by PT' group, since a multinomial random variable simulated by using values of $\theta_1, \ldots, \theta_5$ from the resulting distributions (shown on the right-hand side) would most probably characterize predominantly using transport mode 2 (PT), as required. Using this criterion, our intercept and regression coefficient prior parameters are specified as $\mu_{hj0} = -4$ for $h \neq j$ and $\mu_{hj0} = 0.6$ for $h = j$, and $\sigma_{hj0} = 0.1$, for $h = 1, \ldots, H$ and $j = 2, \ldots, r$.

To explore the effect of the prior choices on our posterior inferences, these specifications were varied within the choice principles described above. The computational cost of model fitting restricted this sensitivity study to a small number of parameter settings, all of which gave consistent posterior inference. Specifically, each combination of $\mu_{hj0} = (-5, -3)$ for $h \neq j$, $\mu_{hj0} = (0.4, 0.8)$ for $h = j$, and $\sigma_{hj0} = (0.1, 0.2)$ was tested, resulting in model parameters that are no more than 10% greater or less than the original posterior spread, giving overall equivalent inference when plotted as in Figs 3, 4 and 5.

3.2.2. Posterior inference

Inference is carried out by using the Bayesian modelling language Stan via rstan (Stan Development Team, 2016). Stan samples from the posterior by using Hamiltonian Monte Carlo sampling. Since the Hamiltonian Monte Carlo algorithm evolves by using Hamilton's differential equations, it does not provide sampling for discrete parameters (StanDevelopment Team, 2016). Therefore, the posterior of the discrete group allocation indices $S = (S_1, \ldots, S_n)$ cannot be sampled directly and must be integrated out of the model calculations. Group allocation is then carried out *a posteriori*. The regression coefficient parameters α and β are therefore sampled from their joint posterior, integrating over S:

$$p(\alpha, \beta | y, \gamma, \xi, \mu, \sigma) \propto \int p(y|S, \beta, \mu, \sigma) p(S|\alpha, \gamma, \xi) p(\alpha|\gamma, \xi) p(\beta|\mu, \sigma) \mathrm{d}S.$$

Since S is discrete, this is equivalent to

$$p(\alpha, \beta | y, \gamma, \xi, \mu, \sigma) \propto \prod_{i=1}^{n} \left\{ \sum_{h=1}^{H} \Pr(S_i = h|\alpha, \gamma, \xi) p(y_i|\beta_h, \mu_h, \sigma_h) \right\} p(\alpha|\gamma, \xi) p(\beta|\mu, \sigma).$$

For group assignment, the probability that individual i is assigned to each group, $h = 1, \ldots, H$, is calculated for all posterior samples of α and β by using the posterior distribution of S_i:

$$\Pr(S_i = h | y_i, \alpha, \gamma, \xi, \beta_h, \mu_h, \sigma_h) = \frac{\Pr(S_i = h|\alpha, \gamma, \xi) p(y_i|\beta_h, \mu_h, \sigma_h)}{\sum_{h=1}^{H} \Pr(S_i = h|\alpha, \gamma, \xi) p(y_i|\beta_h, \mu_h, \sigma_h)}. \tag{11}$$

Individual i is assigned to the group h that maximizes the mean of this probability over all posterior samples, and posterior samples of the regression coefficient parameters α and β are used to explore the key characteristics of each group.

Model fitting took approximately 1 week when allocated 25 Gbytes of memory on one node of a cluster computer. The MCMC chain was run for 160000 iterations to reach convergence and a further $M = 20000$ iterations were retained as the posterior samples.

4. Results

The M posterior samples of α and β are used to allocate the n modelled individual into one of the $H = 5$ groups, based on the posterior distribution for S_i (equation (11)). Fig. 3 presents pairwise marginal plots of the structural response (the number of days commuting by each transport mode type), together with the corresponding group allocations.

Of the $n = 2500$ modelled individuals, 1099 are allocated to the primarily MV user group (MV group), 269 to the primarily PT user group (PT group), 385 to the primarily bicycle user group (bicycle group), 475 to the primarily on foot group (foot group) and 272 to the primarily

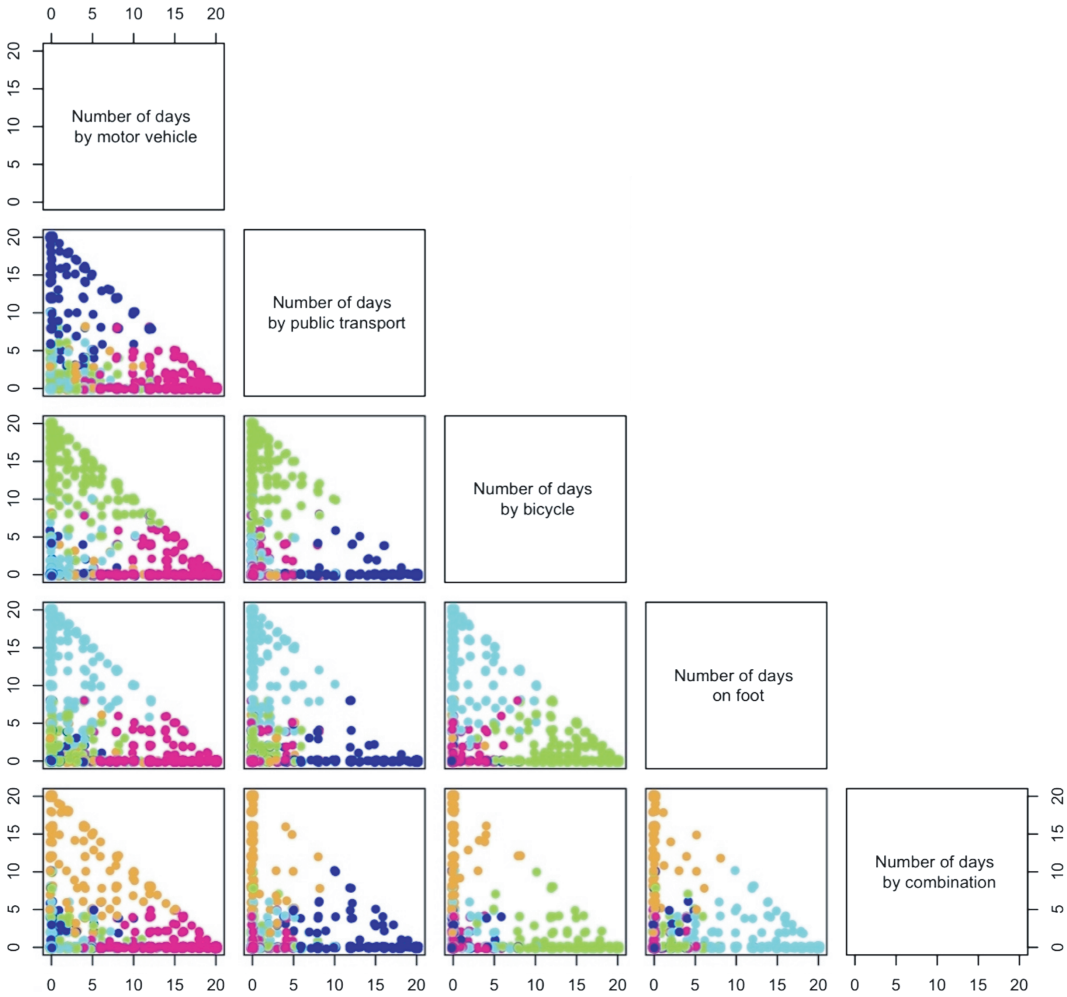

Fig. 3. Pairwise marginal plots of the structural response (number of days commuting by each transport mode type) (the discrete points are jittered slightly to represent the distribution of individuals in each plot better): ●, MV group; ●, PT group; ●, bicycle group; ●, foot group; ●, combo group

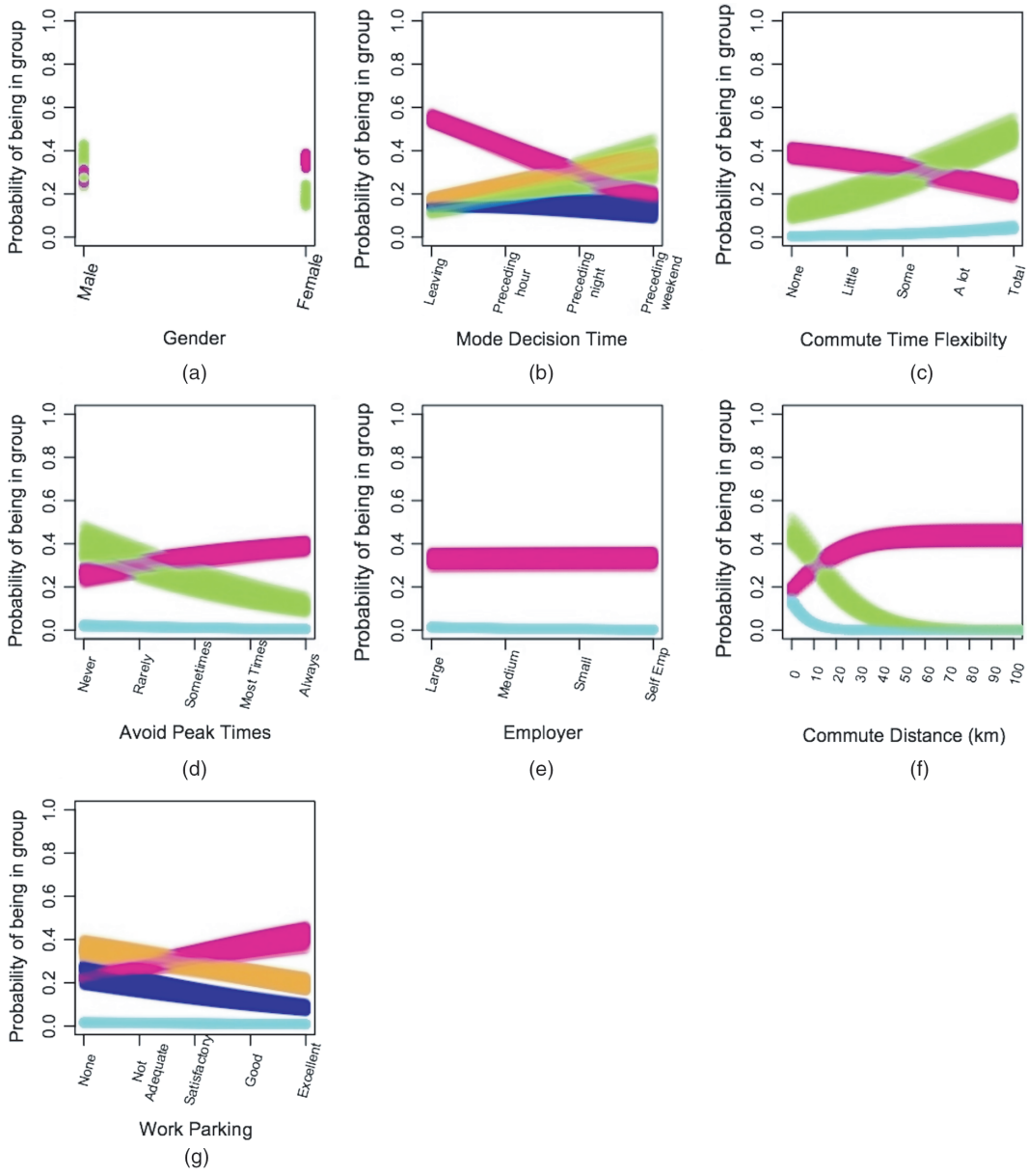

Fig. 4. Relationship between the probability of being in each group and GIs (a) gender, (b) transport mode decision time, (c) commute time flexibility, (d) avoidance of peak congested times, (e) employer, (f) commute distance and (g) work parking facilities: ——, MV group; ——, PT group; ——, bicycle group; ——, foot group; ——, combo group

combination user group (combo group). Consistent with the prior constraints to target the five opposing behavioural groups, Fig. 3 shows that each of the groups represents individuals who, in general, principally use each of the five transport mode types.

Posterior samples of α quantify the relationship between GIs and the probability of group allocation, identifying the key characteristics that differ between the five groups. These relation-

ships are presented in Fig. 4 as continuous, and isolated for each group and GI, by calculating the probability of group membership over the full numerical range of the specific GI while holding all other GIs constant at their mean value.

In Fig. 4, the groups in each plot reflect the GIs that were selected in the variable-selection process (Table 3 in supplementary material: `https://www.dropbox.com/s/95282pdtulf0ext/Dawkins_SupMat.zip?dl=0`). The absence of a group in a given figure is informative in itself, identifying that the log-odds ratio of being in that group rather than the MV group is not strongly related to the associated GI. Since the MV group is the baseline in the multinomial logistic regression model, the probability of being in the MV group, η_{i1}, can be calculated for all GIs $d = 1, \ldots, D$, based on the α_{hd} regression coefficients associated with all groups for which GI d is selected (equation (3)). The spread in each of these relationships represents the full posterior distribution of α.

The relationships in Fig. 4 can be used to inform and guide discussions about, and the design of, group-specific interventions to reduce commuter congestion. For example, Fig. 4(b) shows that individuals who decide which mode to commute by just before leaving have a much greater probability of being in the MV group than in any other group. This suggests that interventions that are designed to reduce MV usage within the primarily commute by MV group should be targeted at the hour preceding their commute, whereas for other groups the intervention should occur earlier (e.g. on the previous day). Similar intervention design guidance can be taken from the other GIs; for example individuals with greater commute time flexibility are most likely to be in the bicycle group, suggesting that interventions that are designed for this group should encourage utilizing this flexibility to avoid congestion; and those with poor work parking are most likely to be in the combo group, suggesting that interventions to restrict work parking could encourage an increase in combination commutes (e.g. 'park and ride').

Posterior samples of β quantify the relationship between BIs and the probability of transport mode usage within each group, identifying the key drivers of within-group behavioural differences. For each of the groups we select one BI which is found to be the most influential and present these in Fig. 5. As with the GIs, these relationships are isolated for each mode influencer individually and calculated as a continuous function. A plot showing all the BI relationships is included in the supplementary material: `https://www.dropbox.com/s/95282pdtulf0ext/Dawkins_SupMat.zip?dl=0`.

The relationships in Fig. 5 are not necessarily causal; however, as in Fig. 4, they can be used to inform discussions about, and the design of, group-specific interventions. Fig. 5(a) suggests that individuals within the MV group who are more influenced by receiving traffic information have a lower probability of commuting by MV, with this decrease reflected in an increase in the probability of commuting by public transport. This suggests that a group intervention to encourage a greater influence from traffic information could result in fewer commutes by MV and more by PT. Within the survey respondent workshop it will therefore be important to understand what type of traffic information is influencing individuals and how this can be communicated most effectively. In a similar way, Fig. 5(b) indicates that individuals in the PT group who are more concerned with the environment have a lower probability of commuting by MV and a higher probability of commuting by PT, suggesting that an intervention to encourage environmental concern within this group could result in fewer commutes by car. Within the workshop with this group it will be important to understand what kind of environmental concern is contributing to this change in behaviour. For both the bicycle and the foot groups (Figs 5(c) and 5(d)) those individuals who are more influenced by weather information have a lower probability of commuting by bicycle and on foot respectively, which is reflected in a greater probability of commuting by MV. These relationships indicate that information concerning bad weather (e.g. rain) influence

Fig. 5. For each of the five groups, the relationship between one key mode influencer and the probability of commuting by the given principal transport mode and an alternative, for columns (a) the MV group (——), (b) the PT group (——), (c) the bicycle group (——), (d) the foot group (——) and (e) the combo group (——)

people within these groups to drive rather than to take their alternative mode; hence an intervention to encourage resilience to bad weather may reduce MV commutes within these groups. It will be important to understand how this could be achieved within the group workshops, e.g. through promotion of appropriate wet weather clothing. Finally, Fig. 5(e) suggests that individuals within the combo group who are more concerned with cost have a slightly lower probability of commuting by a combination of modes and slightly higher probability of commuting by PT, suggesting that information about the cost of different transport options could influence transport behaviour within this group. Again, understanding exactly what information would be of interest to these commuters will be important insight from the group workshops.

The possible effect of a given intervention, in terms of reducing MV usage, can be very approximately quantified, giving a rough indication of how successful a given intervention could be. For example, suppose that we design an intervention to provide useful traffic information to the 214 individuals within the MV group who also sometimes commute by an alternative mode of transport (e.g. on foot). Suppose that this intervention increases the influence of traffic information in these individuals by two increments (i.e. occasionally becomes often etc.), decreasing their probability of commuting by MV as in Fig. 5(a). Then, within a 20-weekday period, their 2954 combined days commuting by MV reduces to approximately 2600 days, meaning approximately 15 fewer of these 214 individuals commuting by MV per day. Using the UK 2011 census figures for the number of people commuting to Exeter from each local authority district, we can weight these individuals to be representative of the whole population, indicating that this intervention could result in approximately 430 fewer people commuting to Exeter by MV per day. Similar calculations can be made for the other groups. This demonstrates how even small behavioural nudges to encourage individuals to reduce their commute by motor vehicle by just 1 or 2 days fewer per month could have a large combined effect on the number of people using the highly overprescribed road networks.

As can be seen in Fig. 3, in some cases individuals on the boundary between two groups (e.g. commute 11 days by one mode of transport and 9 days by another) are allocated to their non-predominant mode group. For example, 24 individuals who commute by MV slightly more often than bicycle (in most cases 12:8 days) are allocated to the bicycle group rather than to the MV group. Exploration of the characteristics of these individuals (as presented in the supplementary material) identifies that these 24 individuals behave and have values that are more characteristic of the bicycle group. Specifically, most have high commuting flexibility and a short commute distance, and those who commute by MV more often are more sensitive to weather information. As a result, interventions that are designed for the bicycle group are likely to be more effective at influencing commuter behaviour in these 24 individuals, demonstrating the advantage of using this constrained mixture model approach, as opposed to the *a priori* allocation to the predominant mode group approach that was described in Section 1.

5. Model validation

Of the 2648 quality-controlled survey respondents, $n_{\text{val}} = 148$ were selected as a stratified sample over the five groups for model validation. This validation quantifies how well the model represents these n_{val} individuals in terms of their predicted group allocation variable and structural response; the two components of the hierarchical finite mixture model (equation (1)).

5.1. Group allocation

Denote the mth draw from the M posterior samples of α and β, available after model fitting, as

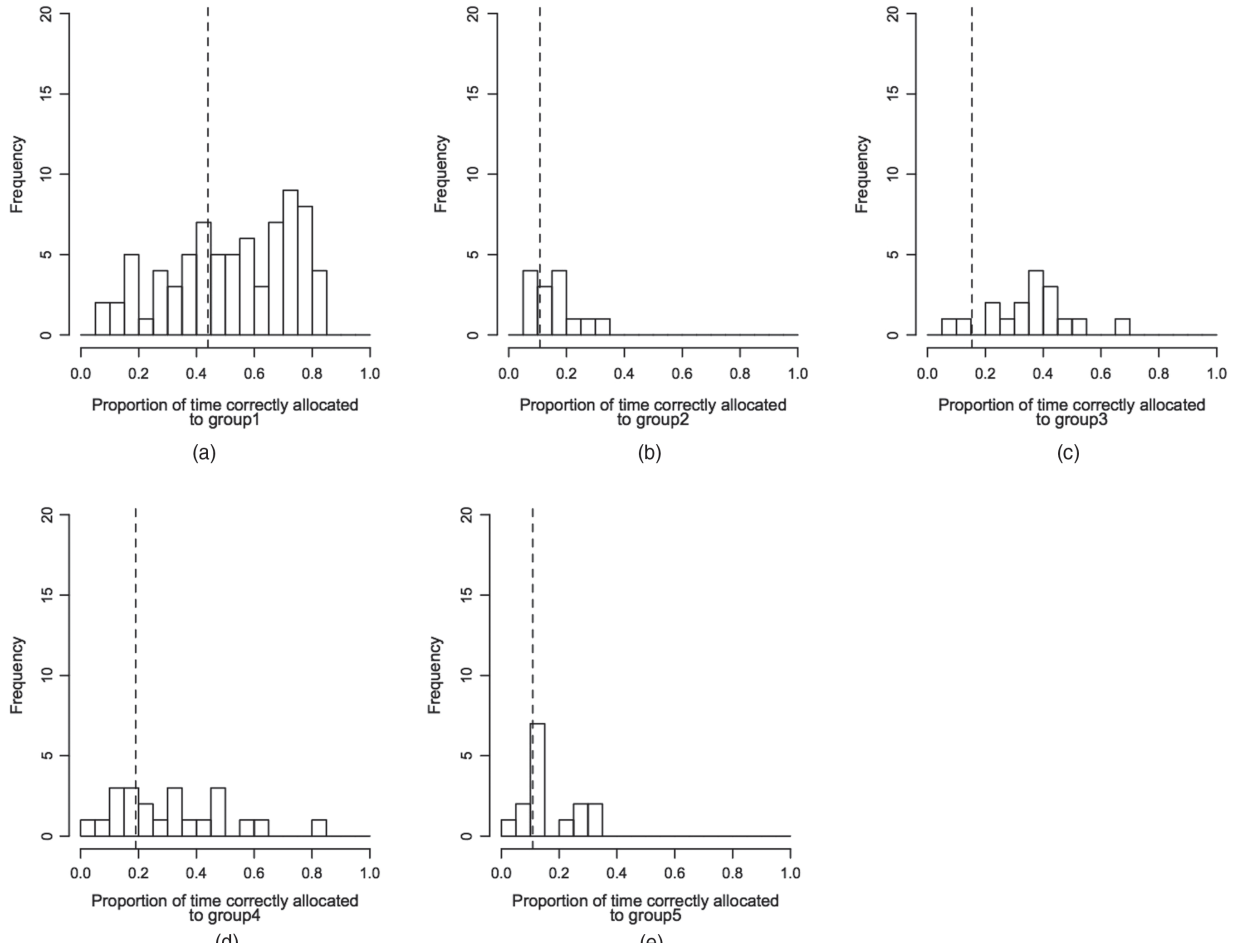

Fig. 6. Histograms of the proportion of posterior samples for which each of the 148 validation individuals is allocated to their 'true' group 1–5 in plots (a)–(e) respectively: ¦, proportion of the 2500 modelled individuals assigned to each group, approximately representing the random chance of being allocated to each group (i.e. non-skilled allocation)

$(\alpha^{(m)}, \beta^{(m)})$. The group allocation component of the model is validated in terms of the proportion of time that the model can allocate individuals i to their 'true' group. Since group allocation is not an observed variable the true group is represented by the predominant transport mode that is used by each individual. For example, if an individual uses an MV more than any other mode their true group is specified as group 1. This definition of the true group, however, does not capture the subtlety in allocating individuals on the boundary of the groups (as discussed in Sections 1 and 4), which may affect the validation results. This subtlety is impossible to identify empirically and hence we consider this definition of the true group to be optimal.

For validation individual $i = 1, \ldots, n_{\text{val}}$, the proportion of correct allocations is quantified by sampling a group allocation variable $\tilde{S}_i^{(m)}$ for each posterior draw $m = 1, \ldots, M$:

$$\tilde{S}_i^{(m)} | \alpha^{(m)} \sim \text{Multinom}(1; \eta_{i1}^{(m)}, \ldots, \eta_{iH}^{(m)}), \tag{12}$$

where

$$\eta_{ih}^{(m)} = \frac{\exp(z_i^{\mathrm{T}} \alpha_h^{(m)})}{\sum_{l=1}^{H} \exp(z_i^{\mathrm{T}} \alpha_l^{(m)})} \qquad \text{for } h = 1, \ldots, H,$$

and observing the proportion of times that $\tilde{S}_i^{(m)}$ is equal to the estimated true group S_i, i.e. $\Sigma_{m=1}^{M} \mathbf{1}\{\tilde{S}_i^{(m)} = S_i\}/M$, where $\mathbf{1}\{x\}$ is an indicator function which takes the value 1 when x is true and 0 otherwise, shown in Fig. 6.

Allocation skill, which is presented in Fig. 6, is quantified in relation to the probability of being assigned to that group by random chance (i.e. non-skilled allocation), based on the proportion of modelled individuals who are assigned to each group (the broken curve). High group allocation skill would be characterized by a histogram that is skewed towards 1. Because of the reduced dimensionality of the model and the relationships that were presented in Fig. 4, group allocation skill is relatively low. This is because an individual with intermediate responses to GI survey questions will have roughly equal multinomial probabilities of being in each group in equation (12) and will therefore be allocated to many different groups over the M posterior samples. Allocation skill is poorest for true groups 2 (predominantly commute by PT) and 5 (predominantly commute by a combination of modes) since the probability of being in each of these groups is calculated on the basis of two selected GI survey questions, 'mode decision time' and 'work parking' (as shown as Figs 4(b) and 4(g) and Table 3). Conversely, since group 1 (predominantly commute by MV) is used as the baseline group in the multinomial logistic regression, allocation to that group is based on all of the selected GI survey questions (as shown as Fig. 4). As a result, allocation skill is highest for this group, i.e. the histogram is skewed towards 1. For all groups, allocation skill is greater than random chance, indicating that the model has some skill in predicating group membership.

5.2. Structural response

The structural response component of the model is validated in terms of the how successful the model is at representing the known structural response (transport mode usage) for each validation individual, given that they are allocated to their true group. For validation individual $i = 1, \ldots, n_{\text{val}}$, this is quantified by sampling a structural response $\tilde{y}_i^{(m)}$, conditional on S_i, for each posterior draw $m = 1, \ldots, M$:

$$\tilde{y}_i^{(m)} | S_i = h, \beta_h^{(m)} \sim \text{Multinom}(m_i; \theta_{ih1}^{(m)}, \ldots, \theta_{ihr}^{(m)}),$$

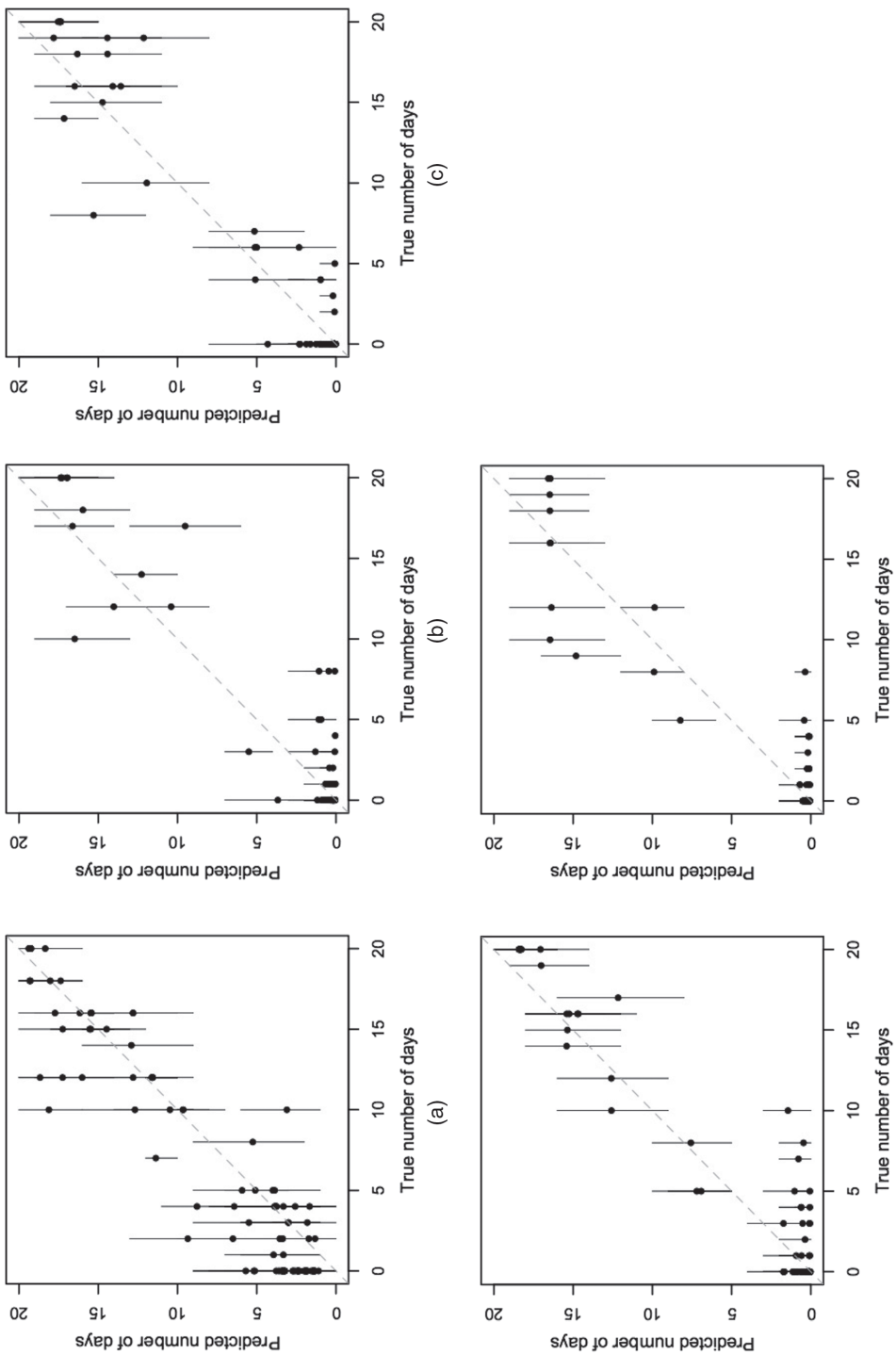

Fig. 7. Comparison of the mean (•) and 90% prediction interval (|) of $\tilde{y}_j$, and the known surveyed structural response value y_j for the 148 validation individuals, characterized by the number of days by using (a) MV, (b) PT, (c) bicycle, (d) on foot and (e) a combination of transport modes on one day: ⋰, $y = x$ line

where m_i is the number of days validation individual i commutes and

$$\theta_{ihj}^{(m)} = \frac{\exp(x_i^{\mathrm{T}} \beta_{hj}^{(m)})}{\sum_{k=1}^{r} \exp(x_i^{\mathrm{T}} \beta_{hj}^{(m)})} \qquad \text{for } h = 1, \ldots, H, \quad j = 1, \ldots, r,$$

and comparing the 90% prediction interval of $(\tilde{y}_i^{(1)}, \ldots, \tilde{y}_i^{(M)})$ with the known surveyed structural variable y_i, shown in Fig. 7.

In Figs 7(a)–7(e) respectively, true mode usage lies within the 90% prediction interval for 90%, 92%, 93%, 92% and 85% of validation individuals. This indicates a good level of skill in predicting the use of all transport modes other than combination, when individuals are assigned to their true group. Similarly to the group allocation validation, underperformance of the model is due to dimensionality reduction through variable selection. Only one selected BI is used to calculate the probability of using a combination (see the plot of all BIs in the on-line supplementary material), the importance of fitness within the cycling group, and none quantify the use of a combination within the combo group. Therefore, for the combo group, this prediction is based solely on the baseline regression coefficient, explaining why the prediction intervals are very similar for all individuals with a high use of a combination of modes.

6. Discussion

When performing inference based on a sample of survey respondents, known or expected disproportionality of the sample with respect to the target population should be accounted for to avoid biased inference (Pfeffermann, 1993; Gelman, 2007). When conducting a survey it is common to use sampling designs based on population strata, to achieve similar proportions of characteristics within the sample to those in the population (Gelman and Carlin, 2002). Survey weights are then constructed to adjust for the representativeness of each stratum within the survey analysis. There is a rich literature on including survey weight adjustments in estimation (see Rao (2011) for a detailed review). However, as discussed by Kunihama *et al.* (2016), the vast majority of such methods are not appropriate in model-based inferences, particularly under Bayesian frameworks, and those methods that have been proposed for Bayesian analysis require highly complex models. As identified by Gelman (2007), Kunihama *et al.* (2016) and Kang and Bernstein (2016), this has led to a disconnect between the analysis of survey data in practice and methods that have been developed to account for survey sample bias in Bayesian inference.

In this application, participation in the survey was voluntary, incentivized by an iPad prize, resulting in what is known as a non-probability-based convenience sample. Voluntary Internet-based surveys, like this, have seen a rapid recent increase in their implementation due to their reduced cost and increased speed of data collection, compared with more conventional stratified sampling methods (Morrissey *et al.*, 2016). Correcting for sample bias is complicated by the lack of information about the sampling mechanism and how it should be represented in model inference. For example, it may be expected that the iPad prize, which is used to encourage survey participation, may cause an underrepresentation or overrepresentation of particular categories of the population. Specifically, Statista, which is an on-line portal for statistics, suggests that in 2015 approximately 20% of people in the UK aged 25–34 years owned a tablet, compared with only 10% of people aged 55–64 years (Statista, 2019). The preferred approach for adjusting non-probability samples in practice is to weight each surveyed individual by using post-stratification weights (Dever *et al.*, 2008; Loosveldt and Sonck, 2008; Baker *et al.*, 2013). This approach adjusts the sample to fit known aspects of the population. For example, suppose that the number of individuals in the population in stratum m (e.g. males) is equal to N_m and

the number of individuals in the sample in stratum m is equal to n_m; then all males in the sample would be weighted by the ratio of the proportion of males in the population and in the sample: $(N_m/N)/(n_m/n)$.

Applying this approach to correct for survey sample bias within this application is complicated by the unknown target population. A census of the UK is taken every 10 years, with the most recent being held on March 27th, 2011. A number of variables are known for the population of Exeter; however, the only available census information related to the commuting population is the location of usual residence (by local authority district). Since the age distribution of the Exeter commuting population is unknown, the potential sample bias that is associated with the iPad prize cannot be explored.

Therefore, to gain some insight into how representative of the target population the sample may be, we can compare the proportion of the sample who live in each local authority district (calculated on the basis of the location of the surveyed home postcode, converted to longitude–latitude, shown in Fig. 8(a)) with the proportion of the Exeter commuter population who live in each local authority district. This comparison is shown in Fig. 8(b).

Fig. 8(b) indicates that the proportion of individuals commuting from each region into Exeter is consistent between the sample and 2011 population. There is a slight undersampling of East Devon and Torbay and a slight oversampling of Mid Devon and Teignbridge; however, in general there is close agreement. This census information was, however, 5 years old and therefore may not be representative of the population at the time of the survey (April–June 2016). The city of Exeter has changed greatly, with the growth of businesses and organizations, such as the Met Office and the University of Exeter, bringing more professionals and academics to the area, and the redevelopment of the Princesshay shopping precinct increasing retail employment.

If this potentially outdated census information were considered to be informative about the current Exeter commuter population, it could be used to calculate post-stratification weights to correct for the small sampling biases that are found in Fig. 8. The weight for individual i is calculated as $w_i = (N_m/N)/(n_m/n)$ for $i \in D_m$, where D_m is the local authority district from which individual i commutes and N_m and n_m are the number of individuals in the population and sample respectively who commute from D_m.

As a simplistic sensitivity study, an approach similar to the frequentist pseudo-maximum-likelihood estimation for population parameter inference that was used by Pfeffermann (1993) and Rao *et al.* (2010) is used to weight the likelihood for each individual, such that

$$f^{\mathrm{w}}(y_i|x_i, z_i) = w_i \left\{ \sum_{h=1}^{H} \Pr(S_i = h|z_i, \alpha)\, p(y_i|x_i, \beta_h) \right\}, \tag{13}$$

used in the model inference as presented in Section 3.2.2. The results of this sensitivity study, equivalent to Figs 3 and 4 and the plot of all BIs in the supplementary material (`https://www.dropbox.com/s/95282pdtulf0ext/Dawkins_SupMat.zip?dl=0`), are also presented in the supplementary material and show how applying post-stratification weights has no effect on model inference. Therefore the conclusions of the analysis, which relate to guiding discussions for intervention design, remain unchanged.

Alternatively, if additional census information about the target population were available in this application, the multilevel regression and post-stratification approach for adjusting non-probability samples could have been employed here. This approach is an extension of classical post-stratification which overcomes the assumption of post-stratification cell simple random sampling by firstly fitting a multilevel regression model to individual responses, using a large number of demographic variables, to estimate the response in each cell, and secondly averaging

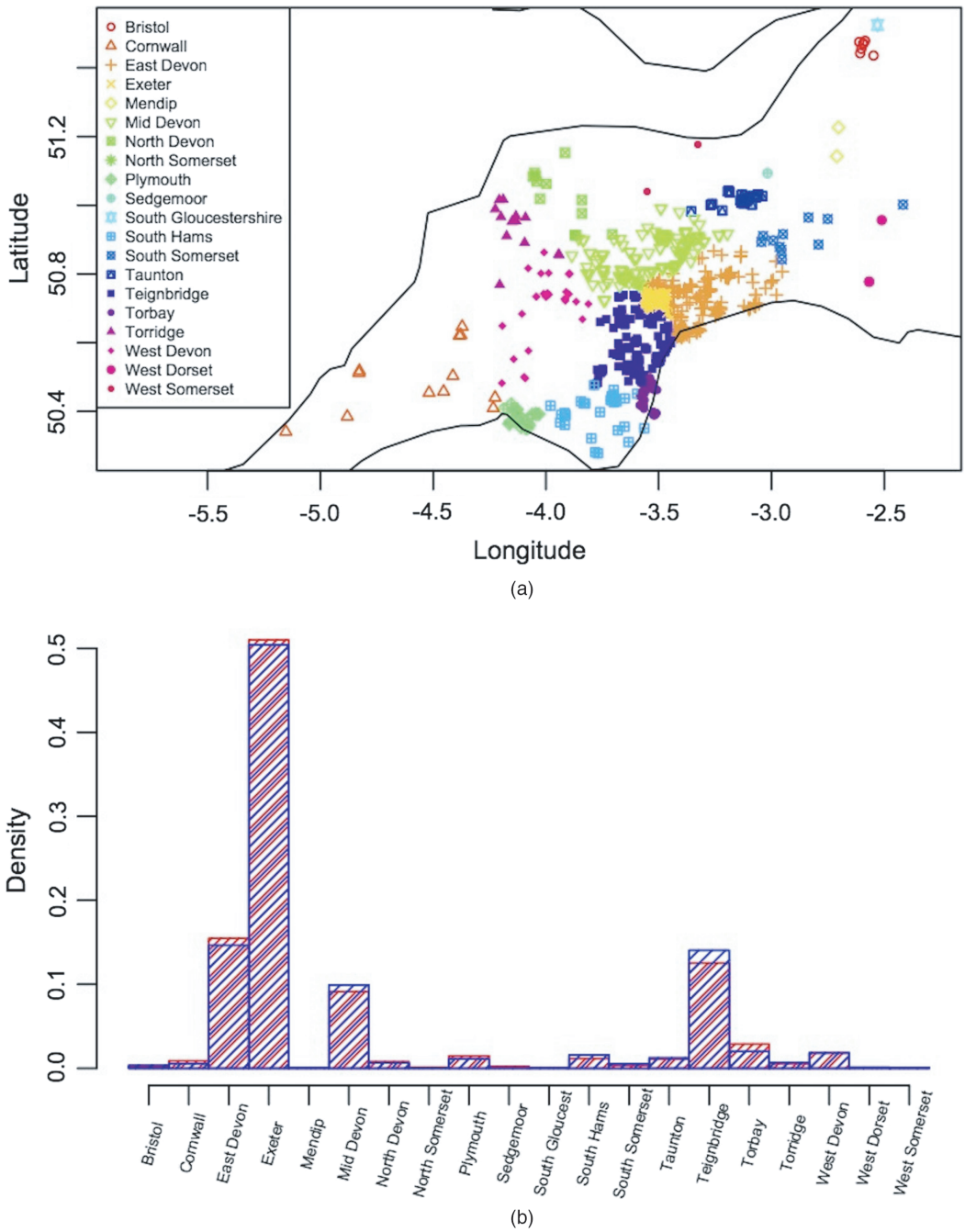

Fig. 8. (a) Home location of modelled individuals by local authority district and (b) proportion of individuals commuting to Exeter from these local authorities in the sample (▨) and the population (▨)

over post-stratification cells to obtain an overall estimate. This approach was first introduced by Park *et al.* (2004), who used the approach to produce state level estimates for USA election results based on non-probability pre-election polls and found that it outperformed standard survey-weighted estimates when estimating state level outcomes. More recently, this approach has been employed with the same success for equivalent USA election applications by Lax and Phillips (2009), Ghitza and Gelman (2013) and Wang *et al.* (2015), based on various non-probability pre-election polls, including, for example, a highly biased sample collected by using an opt-in poll on the Xbox gaming platform (Wang *et al.*, 2015). This class of approach shows great promise for application when adjusting for non-probability samples where adequate census information is available; however, since only one census variable is available here, such approaches could not be applied.

In addition, the fully Bayesian approach for accounting for survey sample bias is currently infeasible in this application because of the complexity of the hierarchical finite mixture model and the unknown sampling mechanism and target population. The Bayesian approach treats the non-surveyed population as missing data and infers the population parameters by sampling the missing data via the posterior predictive distribution of the missing population, given the observed sample, as discussed in chapter 8 of Gelman *et al.* (2014).

Here, these already complex missing data models would be made more complex by the multivariate structural response variable and additional incomplete covariate information. Suppose, within the likelihood of these missing data models, that we wished to include a vector of k covariates for each individual $i = 1, \ldots, N$, $x_i = (x_{i1}, \ldots, x_{ik})$; then this likelihood would include a conditional model for each incomplete variable:

$$p(y, x_1, \ldots, x_k) = p(y|x_1, \ldots, x_k)\, p(x_1|x_2, \ldots, x_k)\, p(x_2|x_3, \ldots, x_k) \ldots p(x_k),$$

requiring various modelling assumptions to be made which could result in unreliable model inference. In addition, the added computational burden of the increasingly complex model could become too challenging for the already computationally demanding model fitting.

To address the disconnect in accounting for sample bias in Bayesian analysis, Kunihama *et al.* (2016) proposed a more simple mixture model approach, in which the standard mixture weights are adjusted to represent the full population, based on the survey weights. Parameter inference is then carried out with a simple modification to the MCMC algorithm that is used for standard mixture models. Directly applying this approach to correct for sample bias in this application would, however, result in additional mixture-specific parameters that do not relate to the allocated groups, complicating the interpretation of the modelling results.

This gap in methodology therefore still remains open and further simplistic methods, that apply to a variety of modelling frameworks and sampling mechanisms, are required.

7. Conclusions

We have presented novel Bayesian methodology for analysing social surveys, developed to meet the needs of the EST project in effectively communicating survey results with project partners and the general public to reduce commuter congestion in Exeter. Directed by social scientist expert judgement, our methodology established a simple five-group narrative within survey respondents, in which each group contained individuals who primarily commuted by one of five transport types, by structurally constraining key parts of our prior within a Bayesian finite mixture model. Drivers of group membership and within-group behavioural differences were modelled hierarchically by using further survey questions, to inform discussions about

interventions to reduce MV usage and hence commuter congestion. The resulting model inference provided a simple narrative of the key characteristics of each group and the key factors that influence primarily commuting by each of the five transport types, clearly suggesting possible approaches for influencing commuting behaviour away from using MVs within each group.

Our methodology provides a general approach with which to understand clearly the key drivers of opposing surveyed behaviours within any application, e.g. which factors influence individuals to shop primarily at competing supermarkets, avoiding the complex group narratives that commonly arise when applying existing approaches for grouping survey respondents.

Inference was carried out by using the Bayesian modelling language Stan. This modelling choice was motivated by Stan's robustness and largely unsupervised learning of the optimal mixing proposal through Hamiltonian Monte Carlo sampling, meaning that the software that was created for this survey could be used by our social scientist colleagues on future survey projects. However, in conducting this inference we found that Stan is inefficient for large data sets and large numbers of parameters. Hence, if time had allowed, we feel that it would have been beneficial to explore alternative modelling languages. Although the focus of this study is not the implementation of the modelling approach, we hope that readers will learn from our experience and, with access to our data and code (available from `https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`), will be able to make a judgement about whether a Stan implementation is appropriate for their application, or use our experience as a reason to explore alternative languages from the outset.

Extensions to this methodology could incorporate interactions between covariates within each regression model to identify whether the effect of a given covariate on the response variable is different at different values of another covariate. Since, in this application, group-specific interventions will be designed for each group as a whole, not separately for those with a different response to specific questions, this level of detail is not needeeds; however, it would be an interesting insight into the groups. In addition, since individuals commuting via routes that are best served by PT or cycle paths, for example, may be more likely to choose these alternative modes, it would be interesting to extend this framework to explore how the 'route to work or study' influences transport mode choice, and hence commuter group allocation. Modelling all possible routes was, however, beyond the scope of this project as it would require a so far not developed integrated smart city data set on employers, road, bus and rail networks, and park-and-ride routes throughout the county.

The issue of survey sample bias was discussed in detail. The disconnect between the analysis of survey data in practice and methods that have been developed to account for survey sample bias in Bayesian inference, identified in previous literature, was demonstrated through the exploration of incorporating post-stratification sample weights within the model. This exploration demonstrated the infeasibility of applying the currently available fully Bayesian approaches, because of the complexity of the hierarchical finite mixture model and the unknown sampling mechanism and target population. Post-stratification weights for each individual were created based on the 2011 census of the UK and, when used to weight each individual within the model likelihood, the results, and therefore conclusion of the analysis, were unchanged.

Further phases of the EST project will involve discussing the survey analysis results with group-specific panels of survey respondents, and lastly testing the success of the actioned interventions in terms of reducing MV usage, and therefore commuter congestion.

Acknowledgements

This work was funded by Innovate-UK EST project NE/N007328/1. We thank Nicolas Walding,

University of Exeter, and Simon Notley, Dynniq, for their geographic information system work which contributed to preparing the survey data for analysis. The data and R code that were used in this paper are available from `https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets`.

References

Anable, J. (2005) "Complacent car addicts" or "aspiring environmentalists"?: Identifying travel behaviour segments using attitude theory. *Transprt Poly*, **12**, 65–78.
Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P. and Dever, J. A. (2013) Summary report of the AAPOR task force on non-probability sampling. *J. Surv. Statist. Methodol.*, **1**, 90–143.
Barr, S. and Prillwitz, J. (2011) Green travellers?: Exploring the spatial context of sustainable mobility styles. *Appl. Geog.*, **32**, 798–809.
Crépet, A. and Tressou, J. (2011) Bayesian nonparametric model for clustering individual co-exposure to pesticides found in the French diet. *Baysn Anal.*, **6**, 127–144.
Dawkins, L. C., Williamson, D. B., Barr, S. W. and Lampkin, S. R. (2018) Influencing transport behaviour: a Bayesian modelling approach for segmentation of social surveys. *J. Transprt Geog.*, **70**, 91–103.
Dever, J., Rafferty, A. and Valliant, R. (2008) Internet surveys: can statistical adjustments eliminate coverage bias? *Surv Res. Meth.*, **2**, 47–62.
Exeter City Council (2015) Exeter City Council, air quality action plan, 2011-2016. Exeter City Council, Exeter. (Available from `https://exeter.gov.uk/media/1221/air-quality-action-plan-2011-2016.pdf`.)
Fahey, M. T., Thane, C. W., Bramwell, G. D. and Coward, W. A. (2007) Conditional Gaussian mixture modelling for dietary pattern analysis. *J. R. Statist. Soc.* A, **170**, 149–166.
Fop, M. and Murphy, T. B. (2018) Variable selection methods for model-based clustering. *Statist. Surv.*, **12**, 18–65.
Fowlkes, E. B., Gnanadesikan, R. and Kettenring, J. R. (1988) Variable selection in clustering. *J. Classificn*, **5**, 205–228.
Fraley, C. and Raftery, A. E. (1998) How many clusters?: Which clustering method?; Answers via model-based cluster analysis. *Comput. J.*, **41**, 578–588.
Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Ass.*, **97**, 611–631.
Frühwirth-Schnatter, S. (2011) Panel data analysis: a survey on model-based clustering of time series.. *Adv. Data Anal Classifcn*, **5**, 251–280.
Frühwirth-Schnatter, S., Pamminger, C., Weber, A. and Winter-Ebmer, R. (2012) Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *J. Appl. Econmetr.*, **27**, 1116–1137.
Frühwirth-Schnatter, S., Pamminger, C., Weber, A. and Winter-Ebmer, R. (2016) Mothers' long-run career patterns after first birth. *J. R. Statist. Soc.* A, **179**, 707–725.
Garthwaite, P. H., Al-Awadhi, S. A., Elfadaly, F. G. and Jenkinson, D. J. (2013) Prior distribution elicitation for generalised linear and piecewise-linear models. *J. Appl. Statist.*, **40**, 59–75.
Garthwaite, P. H., Kadane, J. B. and O'Hagan, A. (2005) Statistical methods for eliciting probability distributions. *J. Am. Statist. Ass.*, **100**, 680–701.
Gelman, A. (2007) Struggles with survey weighting and regression modeling. *Statist. Sci.*, **22**, 153–164.
Gelman, A. and Carlin, J. B. (2002) Poststratification and weighting adjustments. In *Survey Nonresponse* (eds R. M. Groves, D. A. Dillman, J. A. Eltinge and R. J. A. Little), pp. 289–302. New York: Wiley.
Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2014) *Bayesian Data Analysis*, 3rd edn. New York: Chapman and Hall.
Ghitza, Y. and Gelman, A. (2013) Deep interactions with MRP: election turnout and voting patterns among small electoral subgroups. *Am. J. Polit. Sci.*, **57**, 762–776.
Gormley, I. C. and Murphy, T. B. (2008a) Exploring voting blocs within the Irish electorate: a mixture modeling approach. *J. Am. Statist. Ass.*, **103**, 1014–1027.
Gormley, I. C. and Murphy, T. B. (2008b) A mixture of experts model for rank data with applications in election studies. *Ann. Appl. Statist.*, **2**, 1452–1477.
Jamieson, S. (2004) Likert scales: how to (ab)use them. *Med. Educ.*, **38**, 1212–1218.
Jasra, A., Holmes, C. C. and Stephens, D. A. (2005) Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.*, **20**, 50–67.
Kang, J. and Bernstein, K. (2016) On bayesian inference with complex survey data. *Biometr Biostatist Int. J.*, **3.**
Kerlinger, F. N. and Lee, H. B. (2000) *Foundations of Behavioral Research*, 4th edn. Fort Worth: Harcourt.
Kim, S., Mahlet, T. G. and Vannucci, M. (2006) Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **93**, 877–893.
Kunihama, T., Herring, A. H., Halpern, C. T. and Dunson, D. B. (2016) Nonparametric Bayes modeling with sample survey weight. *Statist. Probab. Lett.*, **113**, 41–48.

Kuo, L. and Mallick, B. (1998) Variable selection for regression models. *Sankhya* B, **60**, 65–81.
Lau, J. W. and Green, P. J. (2007) Bayesian model-based clustering procedures. *J. Computnl Graph. Statist.*, **16**, 526–558.
Lax, J. R. and Phillips, J. H. (2009) How should we estimate public opinion in the states? *Am. J. Polit. Sci.*, **53**, 107–121.
Liang, J.-C. and Tsai, C.-C. (2008) Internet self-efficacy and preferences toward constructivist internet-based learning environments: a study of pre-school teachers in Taiwan. *J. Educ. Technol. Soc.*, **11**, 226–237.
Loosveldt, G. and Sonck, N. (2008) An evaluation of the weighting procedures for an online access panel survey. *Surv. Res. Meth.*, **2**, 93–105.
Malsiner-Walli, G., Frühwirth-Schnatter, S. and Grün, B. (2016) Model-based clustering based on sparse finite Gaussian mixtures. *Statist. Comput.*, **26**, 303–324.
Maugis, C., Celeux, G. and Martin-Magniette, M. L. (2009) Variable selection for clustering with Gaussian mixture models. *Biometrics*, **65**, 701–709.
Morrissey, K., Kinderman, P., Pontin, E., Tai, S. and Schwannauer, M. (2016) Web based health surveys: using a two step Heckman model to examine their potential for population health analysis. *Socl Sci. Med.*, **163**, 45–53.
Muthukumarana, S. and Swartz, T. B. (2014) Bayesian analysis of ordinal survey data using the Dirichlet process to account for respondent personality traits. *Communs Statist. Simuln Computn*, **43**, 82–98.
Nandram, B., Bhatta, D., Bhadra, D. and Shen, G. (2013) Bayesian predictive inference of a finite population proportion under selection bias. *Statist. Methodol.*, **11**, 1–21.
Pamminger, C. and Frühwirth-Schnatter, S. (2010) Model-based clustering of categorical time series. *Baysn Anal.*, **5**, 345–368.
Park, D. K., Gelman, A. and Bafumi, J. (2004) Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Polit. Anal.*, **12**, 375–385.
Pfeffermann, D. (1993) The role of sampling weights when modeling survey data. *Int. Statist. Rev.*, **61**, 317–337.
Rao, J. N. K. (2011) Impact of frequentist and Bayesian methods on survey sampling practice: a selective appraisal. *Statist. Sci.*, **26**, 240–256.
Rao, J. N. K., Verret, F. and Hidiroglou, M. A. (2010) A weighted estimating equations approach to inference for two-level models for survey data. In *Proc. Survey Methods Section Statistical Society of Canada A. Meet.*, *Québec*.
RStudio (2013) Easy web applications in R. RStudio, Boston. (Available from `http://www.rstudio.com/shiny`.)
Si, Y., Pillai, N. S. and Gelman, A. (2015) Bayesian nonparametric weighted sampling inference. *Baysn Anal.*, **10**, 605–625.
Sperrin, M., Jaki, T. and Wit, E. (2010) Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statist. Comput.*, **20**, 357–366.
Stan Development Team (2016) RStan: the R interface to Stan. Stan Development Team. (Available from `http://mc-stan.org`.)
Statista (2019) Distribution of tablet computer users in the United Kingdom (UK) from 2011 to 2017, by age. Statista, Hamburg. (Available from `https://www.statista.com/statistics/272201/distribution-of-tablet-computer-user-in-the-united-kingdom-uk-by-age/`.)
Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015) Forecasting elections with non-representative polls. *Int. J. Forecast.*, **31**, 980–991.
Wheeler, D., Shaw, G. and Barr, S. (2004) *Statistical Techniques in Geographical Analysis*. Chichester: Wiley.
Williamson, D. B. and Goldstein, M. (2015) Posterior belief assessment: extracting meaningful subjective judgements from Bayesian analyses with complex statistical models. *Baysn Anal.*, **10**, 877–908.