Influencing Transport Behaviour: A Bayesian Modelling Approach for Segmentation of Social Surveys

L. C. Dawkins,[1*] D. B. Williamson,[1] S. W. Barr[2] S. R. Lampkin[2]

[1] College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

[2] College of Life and Environmental Sciences, University of Exeter, Exeter, UK

[*] E-mail: L.C.Dawkins@exeter.ac.uk

Abstract

Current approaches for understanding and influencing transport behaviour often involve creating fixed, homogenous groups of similar surveyed individuals in order to explore specific behavioural profiles, an approach known as segmentation. Most commonly, segmentation is not based on a formal statistical model, but either a simple 'a priori' defined group classification narrative, failing to capture the complexity of varying group characteristics, or a 'post hoc' heuristic cluster analysis, applied to multi-dimensional behavioural variables, creating complex descriptive group narratives. Here we present an alternative, Bayesian finite mixture-modelling approach. A clear group narrative is created by constraining the Bayesian prior to group survey respondents based on the predominance of a single apposing transport behaviour, while a detailed insight into the behavioural complexity of each group is achieved using regression on multiple additional survey questions. Rather than assuming within group homogeneity, this creates a dynamic group structure, representing individual level probabilities of group membership and within group apposing travel behaviours. This approach also allows for numerical and graphical representation of the characteristics of these dynamic, clearly defined groups, providing detailed quantitative insight that would be unachievable using existing segmentation approaches. We present an application of this methodology to a large online commuting behaviour survey undertaken in the city of Exeter, UK. Survey respondents are grouped based on which transport mode type they predominantly commute by, and the key drivers of these predominant behaviours are modelled to inform the design of behavioural interventions to reduce commuter congestion in Exeter. Our approach allows us to prioritise the most effective intervention themes, and quantify their potential effect on motor vehicle usage. For example, we identify that individuals that predominantly commute by public transport, but also sometimes motor vehicle, do so on average up to one day per week less often, if they are strongly concerned about the environment, demonstrating how an intervention to promote environmental awareness could greatly reduce motor vehicle usage within this group.

1 Introduction

Influencing transport behaviour away from car and towards alternative, more sustainable modes of transport, has become very important for overcoming the growing environmental and societal issues related to car use, such as climate change and commuter congestion. This trend reflects a wider shift in environmental governance towards a focus on the role that behavioural change can play in delivering policy goals (Clarke et al. 2007; Jones and Whitehead 2013), as opposed to major investments in infrastructure or broader socio-economic adjustments to influence mobility patterns. Specifically, transport geographers have become progressively interested in understanding individual travel decision making (e.g. Anable 2005) and the implications this has for policy through developing segmentation models to characterise travellers and more effectively target policies for behavioural change (Department for Transport, 2011). This is normally undertaken by forming fixed groups within survey respondents, representing homogenous attitudes, behaviours and preferences, in order to design and promote targeted group specific interventions for behavioural change (Haustein and Hunecke, 2013).

These fixed groups are commonly based on one of four classes of variables: behavioural, e.g. frequency of travel or mode choice; geographical, e.g. location or settlement type; socio-demographic, e.g. age or income; or attitudinal, e.g. the desire to use an alternative mode. Most commonly, segmentation is not based on formal model inference, following one of two apposing approaches; 'a priori' segmentation, in which group profiles are well-defined from the outset, such that respondents can be clearly assigned to one of the postulated segments (Haustein and Hunecke, 2013), or 'post hoc' segmentation, in which individuals are grouped based on their similarity in a set of variables using heuristic cluster analysis (Anable, 2005; Cools et al., 2009), resulting in detailed, but complex, descriptive group narratives.

Here, we present an alternative, novel Bayesian finite mixture model approach for understanding and influencing transport behaviour through segmentation, which brings together the the clear group narrative advantage of the 'a priori' approach, aiding interpretation and communication to policy makers and the public, and the detailed group behavioural insight of the 'post hoc' approach, providing an in-depth analysis.

Firstly, a clear group narrative is created by constraining the Bayesian prior distributions to group survey respondents based on which of a set of apposing transport behaviours they predominantly follow. This is quantified through a key survey question, characterising the frequency with which each surveyed individual behaves in a set of apposing ways. For example, the number of times within a week individuals travel to a location using apposing routes, or the number of times within a month they commute to work using apposing transport modes, and survey respondents that predominantly behave in each of the apposing ways are grouped together. Secondly, the model-based nature of the approach allows for detailed group behavioural insight through regression on additional survey questions, formally quantifying how variables such as gender, income and personal attitudes, influence these predominant transport behaviours and membership within each predominant behavioural group. Rather than assuming within group homogeneity, this model-based approach therefore creates a dynamic group structure, representing individual level probabilities of predominant behavioural group membership and of within group apposing travel behaviours. In doing so, this allows us to identify the most important

factors for influencing the desired predominant transport behaviour (e.g. reducing car usage) and quantify how effective influencing these factors could be on changing this behaviour.

Throughout, we demonstrate how this approach can be used to understand and influence transport behaviour through an application to a large online commuting behaviour survey undertaken in the city of Exeter, UK. Exeter is experiencing unprecedented economic and physical growth, with the population 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). Initially, we motivate our novel segmentation approach by applying and highlighting the limitations of a commonly used 'post hoc' segmentation procedure. We then use our novel Bayesian approach to group survey respondents based on their predominant commuting transport mode type, and quantify the key drivers of group membership and apposing transport mode usage, based on additional survey questions, to inform discussions with Exeter commuters and policy makers about the design of behavioural interventions to reduce commuter motor vehicle usage in Exeter. In doing so we demonstrate the advantages of our Bayesian model-based segmentation approach in providing a clear group narrative, detailed behavioural insight and a formal model based representation of results, allowing for in-depth analysis that can be easily communicated to policy makers.

The remaining paper is organised as follows. Research context is discussed in Section 2. In Section 3, we present our novel methodology, firstly introducing the Exeter commuting behaviour survey in Section 3.1, a motivating application of classical segmentation analysis to the survey data in Section 3.2, and our novel constrained Bayesian statistical model in Section 3.3. In Section 4, we present the results of applying the methodology to the Exeter commuting behaviour survey, and finally, Section 5 provides a concluding discussion. Further statistical detail about this methodology is presented in the Technical Appendix at the end of this paper and in the related, more technical paper, Dawkins et al. (2017).

2 Research Context

Existing approaches for understanding and influencing transport behaviour most commonly employ either 'a priori' or 'post hoc' segmentation, as defined in Section 1.

Brög et al. (2009) used 'a priori' segmentation to group individuals based on their sustainable transport mode usage, as well as their interest in receiving information on alternative transport modes, in order to target specific intervention material towards different groups of people. Similarly, Scheiner (2006) grouped survey respondents based on 'a priori' defined settlement types to better understand mobility in the elderly. The 'a priori' segmentation approach is not based on formal model inference but provides a simple allocation rule and a clear group narrative. Anable (2005), however, argued that 'a priori' segmentation fails to adequately capture the complexity of varying group characteristics, and promoted the use of 'post hoc' segmentation for generating natural group associations within the survey respondents. Ryley (2006) used a 'post hoc' segmentation approach to identify life stage socio-demographic groups and develop group specific targeted policies to reduce car usage in Edinburgh.

In a similar way, Anable (2005) and Barr and Prillwitz. (2011) both used 'post hoc' segmen-

tation to group individuals based on multiple attitudinal variables, creating groups with differing psychographic profiles, to understand and influence sustainable transport behaviour in relation to leisure travel. Rather than basing the 'post hoc' segmentation directly on survey question responses, an alternative, more qualitatively orientated approach known as Q-methodology, groups individuals based on their rank-ordering of a set of behavioural and attitudinal statements, creating groups with similar viewpoints and preferences. For example, van Exel et al. (2011) used this approach to explore the commonalities and differences between travellers preferences for middle-distance travel by car and public transport, identifying four preference segments differing in their travel choices and motivations.

Most commonly, and as in Ryley (2006), Anable (2005), Barr and Prillwitz. (2011) and van Exel et al. (2011), heuristic cluster analysis methods are used to identify the fixed 'post hoc' groups within survey respondents. A variety of cluster analysis approaches exist (Wheeler et al., 2004), each often resulting in different groupings, and little systematic guidance is 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 also not based on statistical models, formal inference about group characteristics is impossible, resulting in complex, descriptive group narratives, created post segmentation. For example, Anable (2005) identified six groups differing in their behaviour and attitudes toward sustainable transport, requiring multiple descriptive paragraphs to explain each complex group narrative.

An alternative approach, used in this application, is to employ a model-based finite mixture model framework. This approach is often used in the context of clustering to overcome the limitations of these heuristic approaches (Fraley and Raftery, 2002), although very rarely used in the transport literature (Etienne and Latifa, 2013). In the model-based finite mixture model framework, responses to a key survey question, upon which the groups are based, are viewed as coming from a mixture of underlying probability distributions, each representing a different group, allowing for formal statistical inference. These probability distributions are specified based on the form of the observed data (e.g. Normally distributed) and contain a set of parameters that need to be estimated (e.g. the mean and variance). In addition, the model-based framework can be extended to include any available covariate information about each individual, either relating to group assignment or the cluster variable. Clustering is then based on this covariate information as well as the cluster variable, and the regression coefficients provide detail about the differences between groups and between people within each group.

Here, we extend this model-based approach to be implemented within a Bayesian framework. In the Bayesian framework the model parameters are treated as random variables, described by a posterior probability distribution, representing our uncertainty in the unknown parameters (Gelman et al., 2014), and allowing us to incorporate knowledge we have about the model prior to fitting it. The posterior distribution is obtained by multiplying the likelihood, or the probability of observing the data given specified values of the parameters, by the prior distributions, representing our knowledge and beliefs about the model parameters independent of the data. Prior distributions can be based on a range of sources, for example the judgement of experts in the field of application, previous related studies or historical data. Alternatively, if little is known about the parameter values before analysis, non-informative priors can be specified, allowing the inference to be primarily based on the data (Gelman et al., 2014).

Applications of Bayesian statistics in the social sciences have increased dramatically since the mid 1990's, due to the increased rate of complex societal data collection requiring multi-level statistical modelling, as well as recent breakthroughs in statistical computing, facilitating Bayesian model fitting (Gill, 2015; Lynch, 2007). Gelman et al. (2007) used a Bayesian model to explore the political voting attitudes of rich, middle-income, and poor voters in the USA, noting how Bayesian inference was crucial for allowing rich enough models for the exploration of state-to-state variation. Similarly, Kruschke (2010) identified several advantages of the Bayesian approach in the field of cognitive science, namely model flexibility, and the yield of rich information about parameters of interest.

The key advantage of using a Bayesian approach here, is our ability to structurally constrain the model through the Bayesian prior distributions. This allows us to create the clear group narrative in which survey respondents are grouped based on their predominant probability of following one of a set of a priori specified apposing transport behaviours, aiding model interpretation and communication to the public and policy makers, and the development of a targeted behavioural change intervention design. For example, in our commuting behaviour application, we use this approach to understand the most effective methods for reducing motor vehicle usage in Exeter commuters. We therefore constrain the Bayesian priors such that survey respondents are grouped based on the transport mode type they predominantly commute by (e.g. mainly commute by motor vehicle, mainly commute by public transport). This allows us to discuss each group in a clear way, avoiding the complex descriptive narratives of, for example, Anable (2005). In addition, the regression coefficients provide detail about what influences these clearly defined predominant behaviours, hence suggesting how we can encourage the predominant use of alternative modes of transport to motor vehicle. For example, concern for the environment is found to encourage the use of public transport in the clearly defined 'predominantly commute by public transport' group, suggesting that an intervention to promote environmental awareness will further encourage this predominant behaviour and therefore reduce motor vehicle usage.

Survey questions associated with group membership are termed "group identifiers" and characterise the differences between groups, e.g. age, while survey questions associated with within group behavioural differences are termed "behavioural influencers" and characterise the differences between individuals within each group, e.g. how much concern for the environment influences someone's transport behaviour.

This allows for clear graphical representation of the predominant behaviour group characteristics, as demonstrated in Figure 1, simplifying interpretation. The two levels of regression allow for graphical visualisation of both the relationship between each group identifier and the probability of being in each predominant behavioural group, as in Figure 1 (a), and the relationship between each behavioural influencer and the probability of each apposing predominant behaviour within each group, as in Figure 1 (b). This modelling framework therefore allows us to make statements such as "An individual aged 50 has a 0.5 probability of being in group h (e.g. predominantly commuting by public transport group)", and "An individual in group h, who is strongly concerned about the environment, has a 0.8 probability of commuting by public transport". Hence, each survey respondent has an individual probability of being a member of each group and of behaving in each way, related to their responses to the group identifier and behavioural influencer survey questions, providing a high level of quantitative detail. In addi-

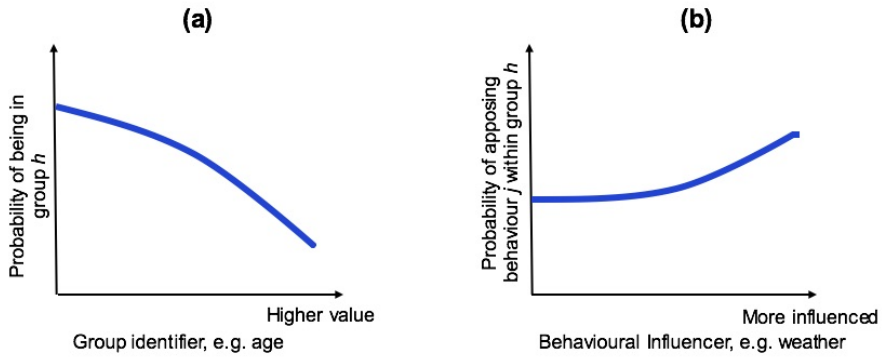

Figure 1: A schematic demonstrating how the characteristics and behavioural variation within each group can be represented graphically following this model-based Bayesian finite mixture model approach.

tion, the numerical quantification of these relationships allows for the identification of the most effective approaches for influencing each predominant travel behaviour and the quantification of their potential effectiveness on behaviour change, insight that cannot be made using existing segmentation approaches.

3 Methodology

3.1 The Commute-Exeter Survey

The Commute-Exeter survey was developed on the web-based platform "www.surveymonkey.com" and was open to the public for 7 weeks, from Monday 2nd May 2016 until Sunday 19th June 2016, receiving 3050 responses. The full survey contained 42 questions. However, survey skip logic, which takes the respondent to the next question based on how they answered the current question, meant each individual only answered questions relevant to their specified commuting behaviour, reducing the length of the survey.

The primary aims of the survey were to identify commute frequency and transport mode choice(s) and the influencers and decision-making processes behind these choices, to inform the design of interventions to reduce motor vehicle usage. The key survey question, used to group survey respondents, was therefore related to their day-to-day commuting pattern; the number of days, of the 20 weekdays in a 4 week period, that they travel by each of the five transport mode types; motor vehicle, public transport, bicycle, on foot or a combination of modes within one journey (e.g. Park&Ride). Details about these choices were attained from further questions. These included, for example, the time they make their transport mode choice, their attitudes towards weather and traffic congestion and values 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. Throughout, the question style was mainly tick box to facilitate flow. These were either multiple choice with one or multiple responses, or matrix/rating scales with one response per row or multiple responses for each case. The full survey is available in the supplementary material.

Participation in the survey was voluntary, based on a widespread marketing campaign to attract local people aged 17 and over who commute into the Exeter city centre, using a variety of transport modes and routes. As well as being a chance for commuters to "really do something" about congestion, participation was incentivised by an iPad prize. The campaign used two visual images and catch phrases, one targeted at road users, in particular motor vehicle users, shown in Figure 2 (a), and the other at non-road users and the wider community, shown in Figure 2 (b). The road user advertising was placed on billboard on main commuting roads and central car parks, while the non-road user advertisement was placed in bus shelters, train stations, large employers, supermarkets and on pub beer mats. In addition, the survey was highly publicised through twitter and the local media including radio and newspaper. Initially, marketing was focused within Exeter, including the distribution of leaflets at key locations including train stations and a promotional stall in the city centre, and was later spread to neighbouring settlements such as Exmouth, Taunton, Crediton, and Paignton.

Figure 2: Survey marketing material focused at (a) road users and (b) non-road users.

Responses to the survey were explored for quality control to ensure consistent and sensible answers. Of the 3050 responses, 2648 responses remained in the analysis after quality control. Of these, 2500 were used to fit the model and 148 were withheld for model validation (presented in Section 5 of Dawkins et al. (2017)).

When performing statistical 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 results (Pfeffermann, 1993; Gelman, 2007). This issue is discussed in great detail in the context of this application in Dawkins et al. (2017). The survey sample was shown to be representative of the 2011 census population in terms of the proportion of individuals commuting to Exeter for each census local authority districts. Hence, no sample bias is corrected for in this analysis, avoiding the complex application of sample weights within a hierarchical Bayesian framework.

3.2 The Classical Segmentation Approach

To motivate the development, and demonstrate the added value, of our novel Bayesian model-based approach, we first apply and present the results of a classical heuristic, non-model based 'post hoc' segmentation procedure, similar to that of Anable (2005). Initially, as in Anable (2005), we apply principle component analysis to the combined variable containing: the number of days in a month each individual commutes using each transport mode type (motor vehicle, public transport, bicycle, on foot, a combination of modes), and the 15 additional survey questions used in our Bayesian analysis (described in Table 1). By observing which of these survey questions have the largest principle component coefficients we identify that the first principle component (PC1), which accounts for the most variation in the data, has a positive relationship with the number of days commuting by motor vehicle and the influence of weather information and weather conditions, while the second principle component (PC2) has a positive relationship with the number of days commuting by bicycle and on foot, and a negative relationship with commute distance and the influence of traffic information. The interpretation of principle components is tricky and can be lengthy, especially if all (20 in this case) are analysed, as would be required to fully explore the data. This complexity is the first limitation of this commonly used approach.

Figure 3 (a) shows the cumulative proportion of variance accounted for by each principle component. This indicates that using the first 9 principle components to explain the data, accounts for 70% of the overall variance (a commonly used threshold for dimensionality reduction in practise). Following this, and as in Anable (2005), we apply heuristic hierarchical cluster analysis (Ward linkage method, see Wheeler et al. 2004) to the survey data, rotated to these 9 principle component axes, to find homogenous groups within the survey respondents. The resulting dendrogram, characterising how the hierarchical clustering procedure iteratively links survey respondents to create groups, is shown in Figure 3 (b). When the dendrogram is cut at a height of 26, 5 groups are created (shown in each rectangle). A scatter plot of the projection of each survey respondent onto the PC1 and PC2 axes, with group allocation indicated by colour, is then presented in Figure 3 (c), showing how these groups vary in terms of these two components which account for the most and second most variation in the data. Based on the principle component descriptions above, Figure 3 (c) identifies that Group 1 characterises people who travel by motor vehicle the most and are most influenced by weather, sometimes walk and cycle, have a range of commute distances, and are moderately influenced by traffic information.

Similar wordy and vague descriptions can be created for each of the five groups and ideally this exploration would be repeated for all 9 principle components used in the clustering algorithm to fully understand the characteristics of each of the groups. Furthermore, by describing the groups in this way we are treating them as homogenous and do not quantify the differences in individuals within groups. In particular, short of applying a logistic regression step after segmentation, using this non-model based approach means we have no numerical quantification of how responses to the survey questions vary with one another and therefore how we might most effectively influence commuter behaviour within each group.

In addition, and most crucially for communication with policy makers and the public, and

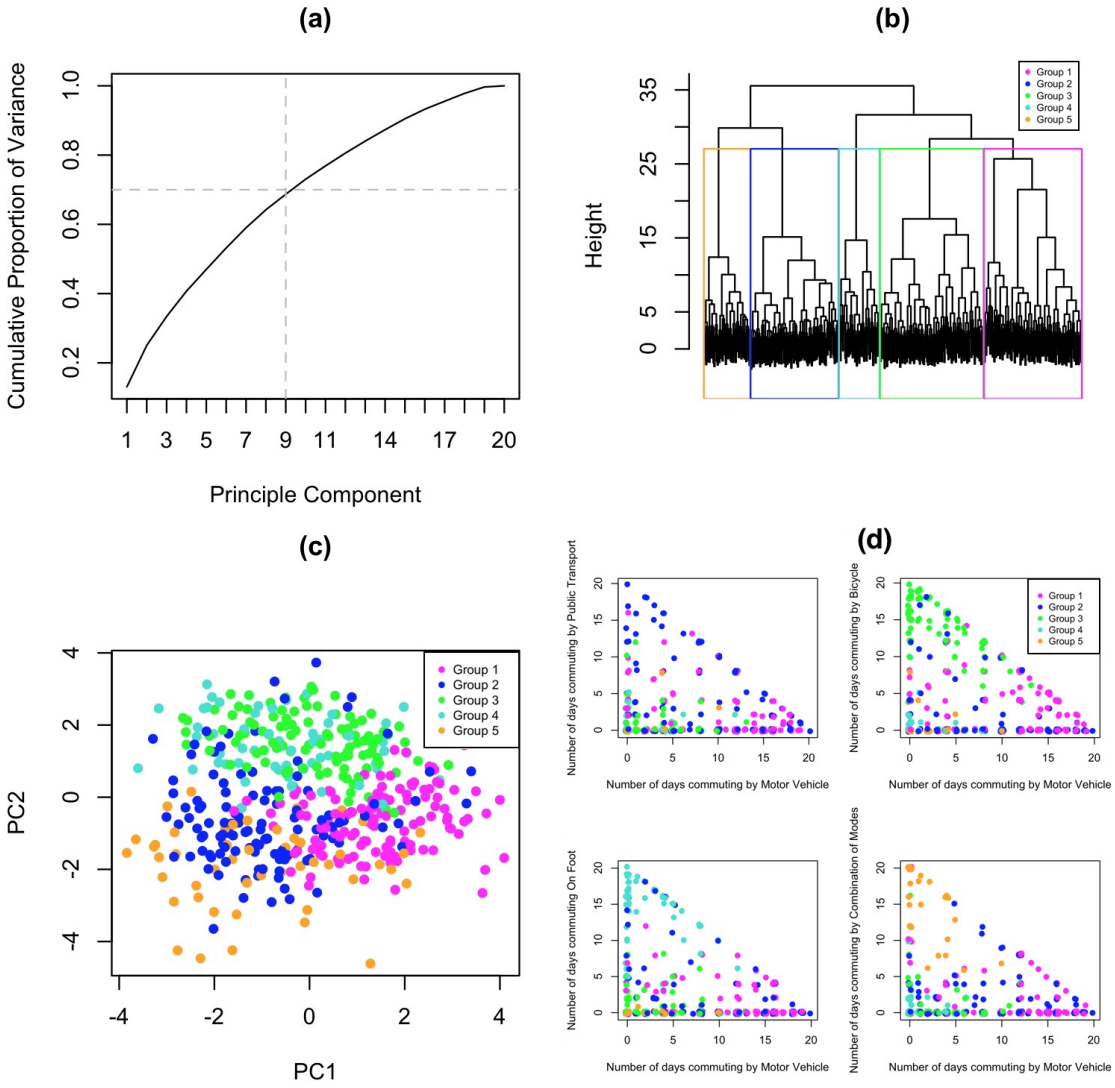

Figure 3: Following the described classical segmentation approach (a) The cumulative proportion of variance accounted for by each principle component, (b) The dendrogram created by applying Ward linkage hierarchical clustering, each of the five groups within the respectively coloured box, (c) A scatter plot of the projection of each survey respondent onto the PC1 and PC2 axes, with group allocation indicated by colour, and (d) Paired scatterplots showing the number of days each surveyed individuals commutes using each mode type, again with group allocation indicated by colour.

designing group specific interventions for behavioural change, creating 'post hoc' groups in this way leads to an unclear overall group narrative in terms of transport mode usage, as shown in Figure 3 (d). The number of days each surveyed individual commutes using each mode type is shown in 4 paired scatter plots. Groups 3, 4 and 5 generally characterise predominantly commuting by bicycle, on foot and by a combination of modes respectively, while groups 1 and 2 have highly varying transport mode usage. This unclear transport mode usage narrative, along with the vague and wordy group descriptions above, complicates intervention design and the dissemination of results to policy makers. Conversely, being able to address or discuss the group

of people who, for example, all "use public transport more than any other mode of transport" greatly eases this communication and collaboration.

As a final limitation, it should also be noted that this analysis would give different results if an alternative 'off the shelf' heuristic clustering approach were used. This creates the additional complication of having to arbitrarily choose the approach to use for a given analysis. Employing a careful, thought-through Bayesian analysis therefore presents a clear advantage.

3.3 A Constrained Bayesian Modelling Approach

Suppose we have n surveyed individuals, each of which has responded to the key apposing transport behavioural survey question upon which we want to base the groups. For individual i, denote their response to this key apposing transport behavioural survey question as $y_i = (y_{i1}, ..., y_{ir})$, where y_{ij} is the frequency with which individual i behaves as in option j. For example, in this commuting behaviour application y_{ij} is the number of days, out of a possible 20 weekdays in a month, individual i commutes to work using transport mode j, taking $r = 5$ possible states (1=motor vehicle, 2=public transport, 3=bicycle, 4=on foot, 5=combination). These n surveyed individuals are allocated to one of $H = 5$ groups, each characterising predominantly behaving in one of these $r = 5$ apposing ways.

We introduce a latent group indicator random variable, $S = (S_1, ..., S_n)$, representing the predominant behavioural group to which each individual $i = 1, ..., n$ is allocated. Since each S_i is discrete and categorical, each is modelled as a single draw from a multinomial distribution,

$$S_i \sim \text{Multinomial}(1; \eta_{i1}, ..., \eta_{iH}), \tag{1}$$

where η_{ih} is the probability of individual i being in group h, $h = 1, ..., H$.

For each individual, $i = 1, ..., n$, the relationship between their probability of being in each predominant behavioural group, η_{ih}, and their response to a set of D additional 'group identifier' survey questions $z_i = (1, z_{i1}, ..., z_{iD})$, e.g. age and gender, is then modelled via multinomial logistic regression;

$$\log(\eta_{ih}/\eta_{i1}) = \alpha_{h0} + z_{i1}\alpha_{h1} + \cdots z_{iD}\alpha_{hD} \quad \text{for} \quad h = 2, ..., H, \tag{2}$$

For statistical identifiability, (i.e. to ensure that $\sum_{h=1}^{H} \eta_{ih} = 1$ for $i = 1, ..., n$), group 1 is used as the baseline group in the above regression model, requiring that $\alpha_1 = (0, .., 0)$, while $\alpha_2, ..., \alpha_H$ are group specific, unknown parameter vectors. Each of these vectors is made up of $D + 1$ regression coefficients, $\alpha_h = (\alpha_{h0}, \alpha_{h1}, ..., \alpha_{hD})$ for $h = 2, ..., H$, where α_{h0} represents the baseline value of the log-odds ratio of being in group h rather than group 1 (left-hand side of Equation 2) and the remaining D coefficients represent the effect of the D group identifier survey question responses on this log odds ratio.

These α regression coefficients are to be estimated from the data, hence the Bayesian framework requires us to specify a priori distributions for each of them, representing our prior beliefs about what values they may take. In this commuting behaviour application, since we do not know the baseline probability of group membership or what effect the D group identifier survey question will have on the probability of group membership prior to model fitting, we place a

relatively uninformative Normal prior (Gelman et al., 2014) on each α regression coefficient parameter (following the usual conjugate model discussed by Garthwaite et al. (2005)),

$$\alpha_{hd} \sim N(\gamma_{hd}, \xi_{hd}), \tag{3}$$

where $\gamma_{hd} = 0$ and $\xi_{hd} = 5$ for $h = 2, ..., H$ and $d = 0, 1, ..., D$.

The key apposing transport behavioural survey question response for individual i, y_i, represents the frequency with which individual i behaves in the r apposing ways over a number of occasions (m_i), and hence is modelled as m_i draws from a multinomial distribution. For example, in this commuting behaviour application m_i represents the overall number of days individual i commutes within the 20 weekday period, e.g. if $y_i = (10, 0, 8, 0, 0)$, individual i commutes for m_i=18 days, using motor vehicle for 10 of these days and bicycle for the remaining 8.

In the finite mixture model framework the response for each individual, y_i is modelled separately for membership within each group. Hence, when individual i is assigned to group h,

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

where θ_{ihj} is the probability of individual i, behaving as in option j, $j = 1, ..., r$, on any given occasion when assigned to group h, $h = 1, ..., H$. Hence, in this application θ_{ihj} represents the probability of individual i using transport mode j on any given day when assigned to group h.

For each individual, $i = 1, ..., n$, the relationship between their probability of behaving in the r apposing ways and C additional 'behavioural influencer' survey questions $x_i = (1, x_{i1}, ..., x_{iC})$, e.g. the influence of weather conditions, is also modelled via multinomial logistic regression;

$$\log(\theta_{ihj}/\theta_{ih1}) = \beta_{hj0} + x_{i1}\beta_{hj1} + \cdots + x_{iC}\beta_{hjC} \quad \text{for} \quad h = 1, ..., H, \quad j = 2, ..., r, \tag{5}$$

For statistical identifiability, here, behavioural option 1 is used as the baseline, requiring that $\beta_{h1} = (0, .., 0)$, while $\beta_{h2}, ..., \beta_{hr}$ are group and mode specific, unknown parameter vectors. Each of these vectors is made up of $C + 1$ regression coefficients, $\beta_{hj} = (\beta_{hj0}, \beta_{hj1}, ..., \beta_{hjC})$ for $h = 1, ..., H$ and $j = 2, ..., r$, where β_{hj0} represents the baseline value of the log-odds ratio of taking option j rather than option 1 in group h (left-hand side of Equation 5) and the remaining C coefficients quantify the effect of the C behavioural influencer survey questions on this log odds ratio. In this commuting behaviour application motor vehicle (transport mode option 1) is used as the baseline mode, hence β_{hj0}, for $j = 2, ..., r$, represents the baseline log-odds ratio of using transport mode j (e.g. bicycle) rather than motor vehicle in group h, and the remaining coefficients quantify the effect of the C behavioural influencer survey question responses on this log odds ratio of transport mode usage.

These β regression coefficients are to be estimated from the data, hence we place a prior on each β regression coefficient parameter, again using the usual conjugate Normal prior model of (Garthwaite et al., 2005),

$$\beta_{hjc} \sim N(\mu_{hjc}, \sigma_{hjc}) \tag{6}$$

Similar to the α regression coefficients, in the commuting behaviour application, we do not know the effect of the C behavioural influencer survey questions on the probability of mode usage prior to model fitting. We therefore place an uninformative Normal prior on each of

the β regression coefficient parameters that quantify the effect of the C behavioural influencer survey question responses on this log odds ratio of transport mode usage. That is $\mu_{hjc} = 0$ and $\sigma_{hjc} = 5$ for $h = 1, ..., H$, $j = 2, ..., r$ and $c = 1, ..., C$.

In our constrained Bayesian modelling framework we do, however, know something about the β regression coefficients that represent the baseline of this log-odds ratio, β_{hj0} in Eqn (5), prior to model fitting. Within each group we want the probability, and hence the log odds ratio, of following a single apposing transport behaviour to be greater than all other apposing behaviours. For example, in group 1, the probability of taking behavioural option 1 (i.e. motor vehicle) to be greater than the probability of taking any other option (i.e. public transport, bicycle, walking or a combination of modes), and equivalently for option 2 (i.e. public transport) in group 2, and so on. We therefore place a constraint on the prior distributions for the β regression coefficients to ensure this is achieved (described in more detail in the Technical Appendix A.1), and specify Normal priors for the β_{hj0} parameters to reflect these constrains as $\mu_{hj0} = -4$ for $h \neq j$ and $\mu_{hj0} = 0.6$ for $h = j$, and $\sigma_{hj0} = 0.1$, for $h = 1, ..., H$ and $j = 2, ..., r$. The approach taken for choosing these values is discussed in detail in Dawkins et al. (2017) (Section 3.3.1).

In the Bayesian modelling framework, fitting the statistical model in order to make statistical conclusions about the unknown parameters, here the regression coefficients α and β, involves quantifying their posterior distribution (left-hand side of Eqn 7), obtained by multiplying their prior distributions, $p(\alpha|\gamma, \xi)$ and $p(\beta|\mu, \sigma)$ by the likelihoods of the observed and latent variables y, $p(y|S, \beta, \mu, \sigma)$, and S, $p(S|\alpha, \gamma, \xi)$:

$$p(\alpha, \beta|y, \gamma, \xi, \mu, \sigma) \propto p(y|S, \beta, \mu, \sigma)p(S|\alpha, \gamma, \xi)p(\alpha|\gamma, \xi)p(\beta|\mu, \sigma). \tag{7}$$

Due to the complexity of this multilevel modelling framework this cannot be calculated analytically and must therefore be numerically approximated by sampling from the posterior distribution using Markov Chain Monte Carlo (MCMC) simulation methods (Brooks et al., 2011). MCMC methods are a general class of algorithm for sequentially drawing values of the target parameters from their approximate distributions, constructed based on a Markov chain that converges to the target posterior distribution (e.g. Eqn 7), resulting in m samples from this target posterior distribution (Gelman et al., 2014). In this application, MCMC is carried out using the R Bayesian modelling language Stan via the R package `rstan` (discussed in more detail in the Technical Appendix A.2).

The m MCMC posterior samples of α and β can then be used to calculate the probability of each individual being allocated to each group (see Appendix A.2). Individual i is then assigned to the group h that maximises the mean of this probability over all of the m posterior samples of α and β.

In addition, the m MCMC posterior samples of α then quantify the relationship between the group identifier survey questions and the probability of group membership (via Eqn. 2), while m posterior samples of β quantify the relationship between the behavioural influencer survey questions and the probability of each predominant transport behaviour within each group, (via Eqn. 5). These relationships can be used to graphically represent the group characteristics and identify the key drivers of transport behaviour, within these clear defined groups, informing behavioural intervention design.

3.3.1 Survey Question Variable Selection

The dimensionality and computational burden of model fitting is reduced by carrying out a variable selection modelling step. Firstly, each survey question is allocated as being either a group identifier (GI), describing differences between groups, or a behavioural influencer (BI), describing differences within groups. Next, the Bayesian finite mixture model is fit to a random subsample of 500 of the survey respondents (without including the GI and BI survey questions) to give a group allocation, S_i, for each of these $i = 1, ..., 500$ respondents. Finally, a regression model for S_i (Eqns. 1 and 2) is fit using all GI questions, and regression models for $y_i|S_i = h$ (Eqns. 4 and 5) are fit using all BI questions for each group ($h = 1, ..., H$) separately. Within each of these regression models a variable selection approach is used to identify which GIs and BIs are most statistically important for explaining relationships within the data. Just those survey questions selected as being important are then used within the full constrained Bayesian finite mixture model described in Section 3.3, reducing the computation burden of model fitting.

In this commuting behaviour application, the D GIs are defined as those that are thought to best explain the differences between groups, and characterise factors that are non-influenceable in the intervention stage. The C BIs, associated with the key behavioural response, are chosen as those that are thought to best explain the differences between individuals within each group and characterise factors that could be influenceable in the intervention stage. There are a number of Bayesian variable selection methods for reducing the parameter space, see O'Hara and Sillanpää (2009) for a review. In this application we used the automated Bayesian variable selection approach of Kuo and Mallick (1998).

The multinomial logistic regression model structure means that resulting selected GIs are those that are important for explaining the log-odds ratio of being within each group; the predominantly public transport group (PT group), predominantly bicycle user group (Bicycle group), predominantly by foot group (Foot group), and the predominantly combination user group (Combo group); rather than the predominantly motor vehicle user group (MV group). Similarly, selected BIs are those that are important for explaining the log-odds ratio of taking each transport mode type; public transport (PT), bicycle (Bicycle), on foot (Foot) and a combination (Combo), rather than motor vehicle (MV), within each group. The resulting selected GI and BI survey questions are shown in Table 1 and represented graphically in Figures 5 and 6 in Section 4.

Table 1: Group Identifier (GI) and Behavioural Influencer (BI) survey questions selected as being important for explaining relationships within the survey data, and hence inclusion within the full model by the variable selection modelling step. The first column indicates whether the question is a GI or BI, the second column states the question as it appears within the survey, the third column presents the possible responses to the question, the fourth column specifies the values given to each response (i.e. z or x) in the multinomial logistic regression Eqns. (2) and (5), and the final column states which groups the GIs are selected for and which transport modes within groups the BIs are selected for.

Question Type	**Survey Question**	**Responses**	**z/x values**	**Selected for Group:Mode**
GI	What is your gender?	Male, Female	(0, 1)	Bicycle group
GI	When are you most likely to make the decision about how you commute to or from your place of work/study?	At time of leaving, In preceding hour before leaving, The night before, During the preceding weekend	(1,2,3,4)	PT, Bicycle and Combo groups
GI	How much flexibility do you have over the time you leave for your commute to and from your place of work/study?	None, A little, Some, A lot, Total	(1,2,3,4,5)	Bicycle and Foot groups
GI	Do you attempt to avoid peak travel times?	Never, Rarely, Some of the time, Most of the time, Always	(1,2,3,4,5)	Bicycle and Foot groups
GI	Which of the following best describes your place of work?	Large, Medium, Small, Self-employed	(1,2,3,4)	Foot group
GI	Home and work postcodes are used to calculate commute distance in kilometres	Continuous (0,...,150)	Continuous (0,...,150)	Bicycle and Foot groups
GI	Thinking about the parking facilities that may be available at or near your place of work/study, please rate them below.	No parking, Not adequate, Satisfactory, Good, Excellent	(1,2,3,4,5)	PT, Foot and Combo groups
BI	How much does receiving information about weather conditions influence your choice of travel mode to your place of work/study?	Never, Occasionally, Sometimes, Often, Always	(0,1,2,3,4)	MV group: Bicycle and Foot; PT group: PT; Bicycle group: Bicycle; Foot group: Foot
BI	How much does receiving information about traffic congestion influence your choice of travel mode to your place of work/study?	Never, Occasionally, Sometimes, Often, Always	(0,1,2,3,4)	MV group: PT; PT group: PT; Bicycle group: Bicycle
	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/study?	Never, Occasionally, Sometimes, Often, Always	(0,1,2,3,4)	
BI	Wind			Foot group: Bicycle
BI	Snow			Bicycle group: Foot
BI	Warm			Bicycle group: Bicycle
BI	Cost is a major consideration when you choose how to commute.	Strongly disagree, Disagree, Neutral, Agree, Strongly agree	(-2,-1,0,1,2)	Combo group: PT
BI	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)	PT group: PT; Bicycle group: Foot
BI	Keeping fit and active is important to you	Strongly disagree, Disagree, Neutral, Agree, Strongly agree	(-2,-1,0,1,2)	Bicycle group: PT and Combo; Foot group: Bicycle

4 Results

The $n = 2500$ modelled survey individuals were allocated into one of the $H = 5$ groups, as shown in Figure 4, based on the posterior distribution for S_i (Eqn.(13) in A.2). Of the

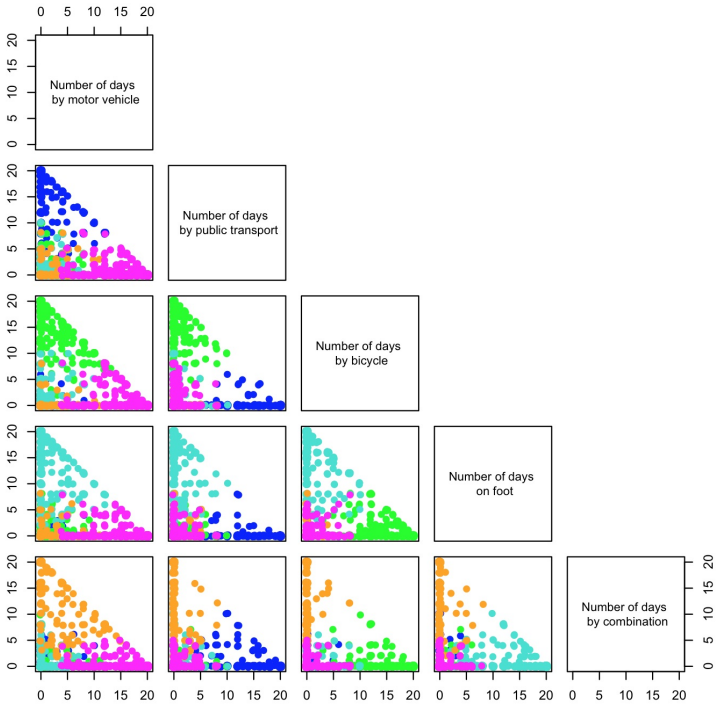

Figure 4: Pairwise scatter plots of the number of days, in a 20 weekday period, the $n = 2500$ modelled individuals commute using each transport mode type. The discrete points are jittered slightly to better represent the distribution of individuals in each plot. Group allocation is indicated by colour, described in the legend.

$n = 2500$ modelled individuals, 1099 are allocated to the predominantly motor vehicle user group (MV group), 269 to the predominantly public transport user group (PT group), 385 to the predominantly bicycle user group (Bicycle group), 475 to the predominantly on foot group (Foot group), and 272 to the predominantly combination user group (Combo group). Consistent with the prior constraints to target the five predominant behavioural groups, Figure 4 shows that, each of the groups represents individuals that always or predominantly use each of the five transport mode types, creating a clear group narrative, simplifying group interpretations. While each group represents a single apposing predominant transport mode, Figure 4 also shows that, within each group, individuals use a mixture of modes within the 20 weekday period. The higher density of points away from the axes in the first column of Figure 4 indicates that the secondary mode in all groups, other than the MV group, is motor vehicle. A proportion of individuals that predominantly use each of the alternative modes therefore also use motor vehicle for some of the days. It is thought that these individuals may be most influenceable within the intervention phase, since they have the means with which to commute by an alternative mode to motor vehicle, but do not yet do so all of the time. The interventions should therefore be focused on these individuals.

Rather than viewing these groups as homogenous, and qualitatively describing their characteristics, as in existing examples and in Section 3.2, our model-based Bayesian modelling approach allows us to numerically and graphically quantify the behavioural differences within and between these clearly defined groups, allowing us to prioritise the most effective behavioural intervention themes, and quantify their potential effect on predominant transport behaviour.

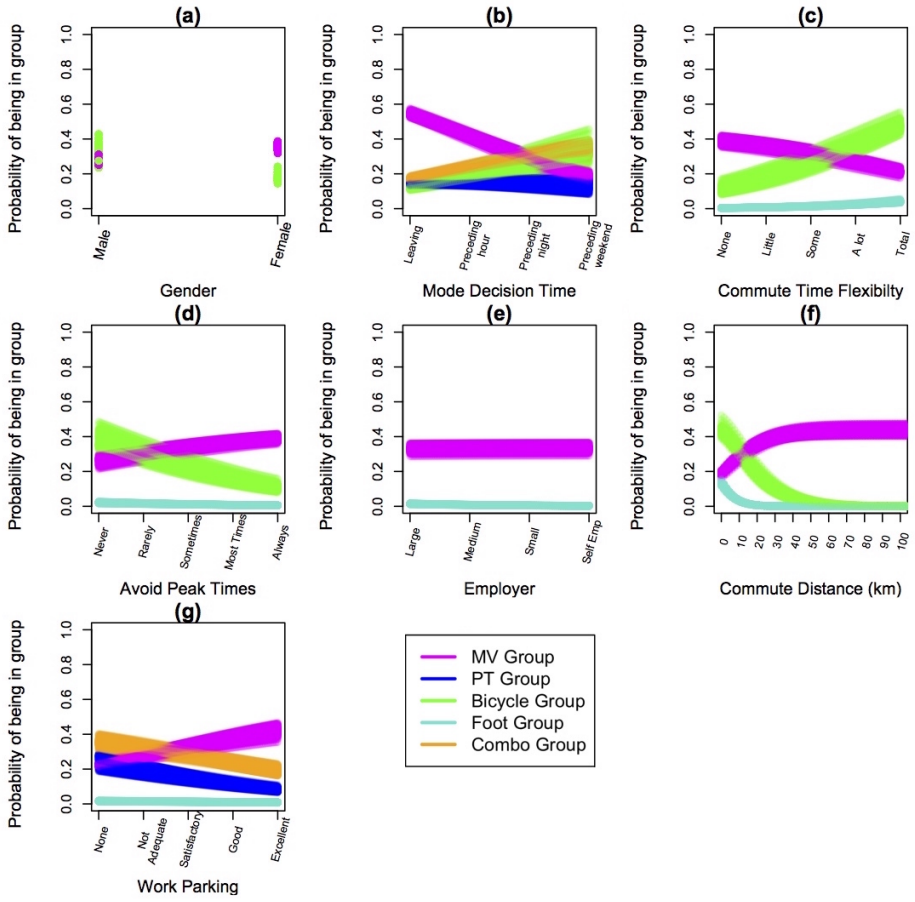

Figure 5: The relationship between the probability of being in each group and group identifiers (a) gender, (b) transport mode decision time, (c) commute time flexibility, (d) avoidance of peak congested times, (e) employer, (f) commute distance and (g) work parking facilities. These relationships are shown in a different colours for each group.

The posterior samples of α quantify the relationship between the group identifier survey questions (Table 1) and the probability of group membership (Eqn. 2), as shown in Figure 5. The groups present in each plot of Figure 5 reflect the group identifiers selected in the variable selection process in Table 1. The absence of a group in a given figure is informative in of itself, identifying that the log-odds ratio of being in that group rather than the MV group is not strongly related to the associated group identifier. The spread in each of the relationships in Figure 4 represents the full posterior distribution of α, quantifying our uncertainty in the results.

The relationships in Figure 5 provide a clear graphical and numerical representation of how the probability of group membership varies with group identifier response, providing a detailed

insight into the differences between the five groups. This insight can be used to inform and guide discussions with the public and policy makers about the design of group specific interventions to influence transport behaviour to reduce commuter congestion in Exeter. For example, Figure 5 (b) shows that the probability of being in the MV group is much higher for individuals who decide which mode to commute by just before leaving. This suggests that interventions designed to reduce motor vehicle usage within the MV group, should be targeted at the hour preceding their commute. In addition, Figure 5 (c) shows that individuals have the greatest probability of being in the Bicycle group if they have a lot or total commute time flexibility, suggesting that interventions designed for this group could promote commuting at alternative times to avoid peak commuter traffic, reducing commuter congestion as a result. Figure 5 (g) shows the individuals with inadequate work parking facilities have the greatest probability of being in the Combo group, while those with excellent work parking have greatest probability of being in the MV group. This suggests that putting greater restrictions on work place parking facilities could be a very effective way of reducing the number of people predominantly driving into the city, and increase the number predominantly using a combination of modes (i.e. using Park&Ride) to reach work.

The posterior samples of β are used to quantify the relationship between the behavioural influencer survey questions (Table 1) and the probability of transport mode usage within each group (Eqn. 5), as shown in Figure 6. As in Figure 5, the absence of a group and mode in a given plot in Figure 6 is informative in identifying that the log-odds ratio of using that mode rather than motor vehicle within that group is not strongly related to the associated mode influencer. For example, the only selected mode influencer associated with the Combo group is 'importance of cost' in explaining the probability of using public transport rather than motor vehicle, suggesting that commuting behaviour may be less influenceable in the Combo group. The spread in each of the relationships in Figure 6 represents the full posterior distribution of β, again quantifying our uncertainty in the results.

The relationships in Figure 6 are not necessarily causal, however they can be used to suggest the most important drivers of predominant commuting behaviour, again to inform intervention design to reduce predominant motor vehicle usage. For example, the first and second rows of Figure 6 (g) suggest that individuals in the PT group who also commute using motor vehicle some of the time, i.e. within the group of commuters that may be most influenceable away from commuting by motor vehicle (identified in Figure 4), use motor vehicle less often and public transport more often (up to 18%, equivalent to 1 day per week of weekdays) if they are more concerned with the environment. Therefore, an intervention reinforcing the positivity of environmentally sustainable commuting could increase public transport usage and reducing motor vehicle usage in the PT group. In addition, the first and third rows of Figure 6 (a) show that individuals in the MV group, who also sometimes commute using bicycle, use bicycle more often and motor vehicle less often (up to 30%, equivalent to 1.5 days per week of weekdays) if they are more influenced by receiving weather information. This suggests that an intervention which provide reliable weather information related to cycling could reduce motor vehicle usage and increase bicycle usage within the MV group.

In addition, this graphical and numerical representation of the effect of different information, attitudes and personal values on transport mode usage allows for direct comparison, and hence

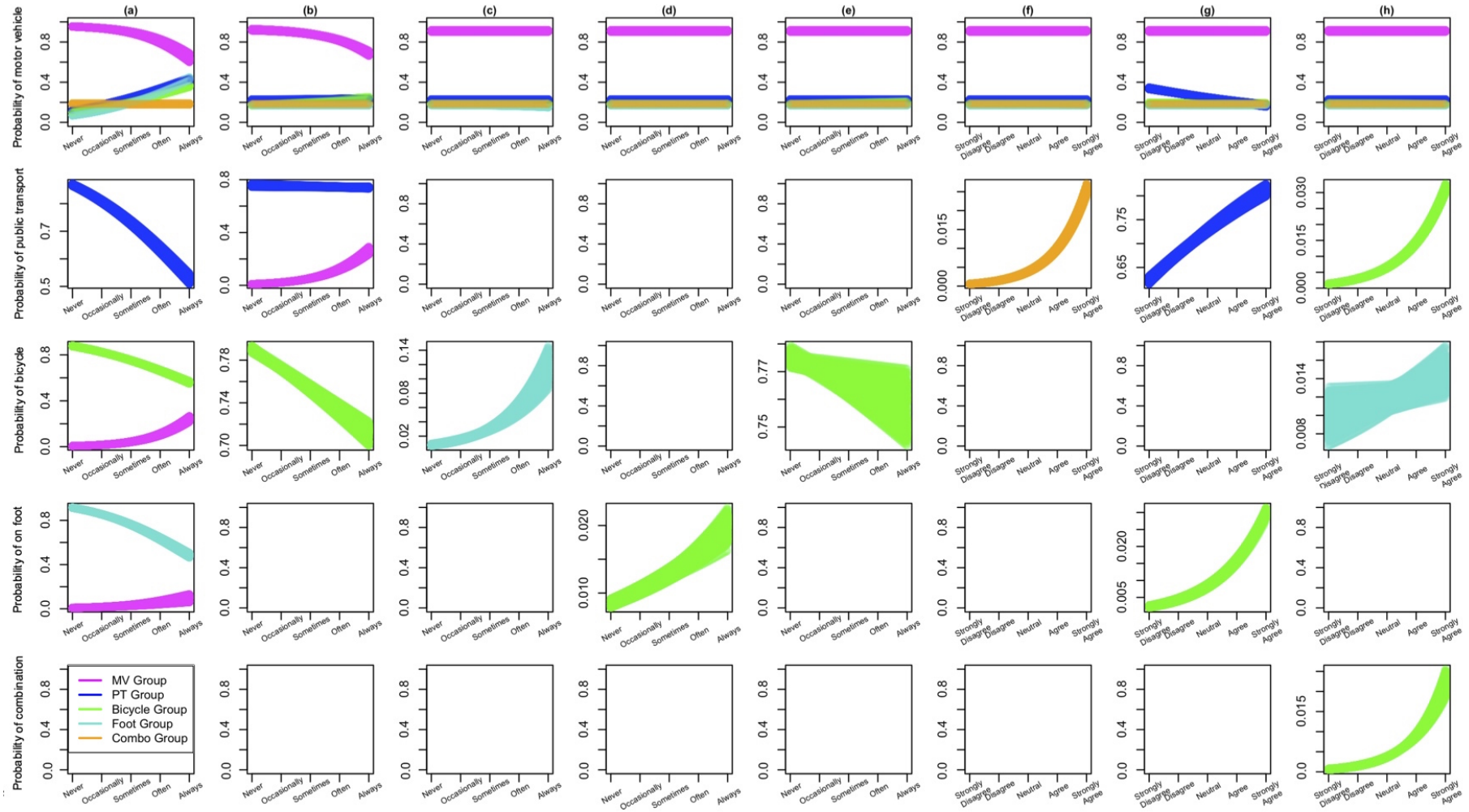

Figure 6: The relationship between mode influencers (a) influence of receiving weather information, (b) influence of receiving traffic congestion information, (c) wind, (d) snow, (e) warm day, (f) importance of cost, (g) importance of the environment, and (h) importance of fitness, and the probability of using (row 1) motor vehicle, (row 2) public transport, (row 3) bicycle, (row 4) on foot, (row 5) a combination of modes in one journey. These relationships are shown in a different colours for each group. The y axis label is given on the far left for each row.

prioritisation, of the most important effects, indicating which themes should be focused on in the interventions. The possible effect of a given intervention, in terms of reducing motor vehicle usage, can also be approximately quantified, giving an indication of how successful a given intervention could be. For example, suppose we design an intervention to promote environmental concern within the PT group, which increases concern for the environment within the group by two increments (i.e. strongly disagree becomes neutral, and disagree becomes agree etc.). Assuming this decreases the probability of using motor vehicle for each of the 269 individuals within this group as in the first row of Figure 6 (g), in 20 weekdays, 410 days of commuting by motor vehicle becomes approximately 300 days, meaning approximately 6 less of these 269 individuals commuting by motor vehicle 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 269 individuals to be representative of the whole population, indicating that this intervention could result in approximately 150 less people commuting to Exeter by motor vehicle per day. This level of insight into the effect of potential interventions on the frequency of each predominant transport mode usage could not be achieved without applying our novel Bayesian constrained finite mixture-model.

5 Discussion and Conclusion

We have presented a Bayesian finite mixture model approach for understanding and influencing transport behaviour through segmentation. This approach brings together the advantages of the existing 'a priori' and 'post hoc' approaches, in providing a clear group narrative and detailed insight into group characteristics, as well as the additional advantages of a model-based formal statistical inference in providing a dynamic group structure and numerical and graphical representation of results. This development has major implications for the ways in which transport researchers and policy makers can understand and work with travel behaviour data.

First, the grouping of surveyed individuals is based on a single key behavioural question, quantifying the frequency with which each individual behaves in a set of apposing ways. The Bayesian prior is used to structurally constrain the mixture model such that survey respondents that predominantly behave in each of the apposing ways are grouped together. The key drivers of group membership and behavioural differences are modelled via Bayesian regression on additional "group identifier" and "behavioural influencer" survey questions. Rather than creating homogenous, fixed groups, as in existing segmentation approaches, this approach therefore recognises that, while individuals may be allocated to a specific group, individuals within each group behave differently. The model-based approach provides numerical and graphical representation of relationships within the data, simplifying group interpretation and allowing for the identification of the key drivers of transport behaviour, indicating which intervention themes should be focused on. This also allows for numerical quantification of how potentially effective a given intervention theme could be on changing apposing transport behaviours, providing a strong evidence base for pursuing a given strategy for behavioural change.

Second, throughout we have demonstrated how this Bayesian finite mixture model approach can be applied to a transport behaviour survey to aid in the design of group specific interven-

tions to influence travel behaviour away from using motor vehicle. In this application we apply the modelling approach to a large, online commuting behaviour survey undertaken in the city of Exeter, UK, to address the growing issue of commuter congestion and the associated environmental impacts. Groups are based on the predominant use of five apposing transport mode types and the relationship between additional survey questions, characterising demographics, attitudes and personal values, and the probability of group membership and mode usage are modelled to identify the most effective intervention themes within each group. In applying our modelling approach we are able to gain far greater insight into the relationships within the survey data than would be achievable using existing segmentation approaches, allowing us to pursue intervention themes for reducing commuter congestion with a detailed quantitative understanding of each group and how effective interventions could be.

In this project this detailed analysis was used to inform discussions with survey respondents within each group to gain further insight into the survey results and their implications for policy and behavioural intervention design. The clearly defined predominant behavioural groups meant we could avoid lengthly group narrative explanations within these discussions, which could instead be conducted with confidence, addressing individuals with a known, simple similarity. These discussions were directed and focused based on the identification of the key drivers of transport mode usage and group membership, meaning we were able to delve deeper into the detailed motivations behind each transport predominant behaviour. More detail about the specific implications of this approach within this application, in terms of conducting these discussions, designing the interventions and addressing policy issues, will be discussed within the forthcoming publication, Lampkin et al. (2017).

The methodology described in this paper has a number of significant implications for travel behaviour researchers and policy makers. First, the model outlined in this paper provides a much greater level of analytical sophistication and insight into the complexity of travel behaviour segments. Researchers who have advocated segmentation (e.g. Anable 2005; Barr and Prillwitz. 2011) have demonstrated the conceptual value of clustering participants to derive segments for understanding travel behaviour. However, the innovation in the model presented here is an implicit recognition that individuals in a given segment hold a plurality of practices and that these can be quantified. Accordingly, it is possible to categorise on the basis of predominance of reported behaviours, accepting that participants will use other modes.

This intellectual innovation, enabled through the Bayesian modelling techniques outlined in this paper, also has major implications for policy, and specifically the kinds of segmentation frameworks developed by the Department for Transport (2011). Through the quantification of mode choice and related mode influencers, it becomes possible to explore the kinds of interventions that might be plausible for different segments. For example, understanding that environmental concern is a key influencer for those who are more likely to avoid car and use public transport for commuting can provide direct links to possible interventions, the success of which can be assessed on a day-to-day basis, enabling a quantification of the potential for measures to reduce the number of car journeys. As such, the methodology described here provides an innovative and dynamic means of exploring predominant travel behaviour and the impacts of potential interventions that move beyond static forms of segmentation.

Acknowledgements

This work was funded by Innovate-UK project 'Engaged Smart Travel' NE/N007328/1. We would like to thank Nicolas G. Walding, University of Exeter, and Simon Notley, Dynniq, for their Geographic Information System (GIS) work which contributed to preparing the survey data for analysis, as well as Lisa Cole, University of Exeter, for designing the graphics used to promote the commute-exeter survey.

References

Anable, J. (2005). "Complacent car addicts" or "aspiring environmentalists"? identifying travel behaviour segments using attitude theory. *Transport Policy*, 12(1):65–78.

Barr, S. and Prillwitz., J. (2011). Green travellers? exploring the spatial context of sustainable mobility styles. *Applied Geography*, 32:798–809.

Betancourt, M. and Girolami, M. (2013). Hamiltonian monte carlo for hierarchical models. *arXiv, 1312.0906*, 217(481).

Brög, W., Erl, E., Ker, I., Ryle, J., and Wall, R. (2009). Evaluation of voluntary travel behaviour change: experiences from three continents. *Transport Policy*, 16:281–292.

Brooks, S., Gelman, A., Jones, G., and Meng, X., editors (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC.

Clarke, J., Newman, J., Smith, N., Vidler, E., and Westmarland, L. (2007). *Creating citizen-consumers: Changing publics and changing public services*. Pine Forge Press.

Cools, M., Moons, E., Janssens, B., and Wets, G. (2009). Shifting towards environment-friendly modes: profiling travelers using q-methodology. *Transportation*, 36(4):437–453.

Dawkins, L. C., Williamson, D. B., Barr, S. W., and Lampkin, S. R. (2017). "What drives commuter behaviour?": Bayesian analysis for apposing predominant behaviours in social surveys. *Annals of Applied Statistics, In review*.

Department for Transport (2011). Climate change and transport choices. Segmentation model: a framework for reducing CO2 emissions from personal travel. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/49971/climate-change-transport-choices-full.pdf.

Etienne, C. and Latifa, O. (2013). Model-based count series clustering for bike sharing system usage mining: A case study with the Vélib' system of Paris. *ACM Transactions on Intelligent Systems and Technology*, 5(3):39–59.

Exeter City Council (2015). Exeter City Council, Air Quality Action Plan, 2011-2016. https://exeter.gov.uk/media/1221/air-quality-action-plan-2011-2016.pdf.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of American Statistical Association*, 100(470):680–701.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian Data Analysis, 3rd Ed.* Chapman & Hall Ltd (London; New York).

Gelman, A., Shor, B., Bafumi, J., and Park, D. (2007). Rich state, poor state, red state, blue state: What's the matter with connecticut? *Quarterly Journal of Political Science*, 2:345–367.

Gill, J. (2015). *Bayesian Methods: A Social and Behavioral Sciences Approach, Third Edition*. CRC Press.

Haustein, S. and Hunecke, M. (2013). Identifying target groups for environmentally sustainable transport: assessment of different segmentation approaches. *Current Opinion in Environmental Sustainability*, 5(2):197–204.

Jones, R. Pykett, J. and Whitehead, M. (2013). *Changing Behaviours: On the Rise of Psychological State*. Edward Elgar Publishing.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, 14(7):293–300.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya B*, 60(1):65–81.

Lampkin, S. R., Barr, S., Dawkins, L. C., and Williamson, D. B. (2017). An innovative approach to segmentation in travel bahaviour research. *In preparation for Journal of Transport Geography*.

Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4(1):85–118.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Int Stat Rev*, 61(2):317–337.

Ryley, T. (2006). Use of non-motorised modes and life stage in edinburgh. *Journal of Transport Geography*, 14:367–375.

Scheiner, J. (2006). Does the car make elderly people happy and mobile? settlement structures, car availability and leisure mobility of the elderly. *European Journal of Transport and Infrastructure Research*, 2:151–172.

Stan Development Team (2016). RStan: the R interface to Stan.

van Exel, N. J. A., de Graaf, G., and Rietveld, P. (2011). "i can do perfectly well without a car!" an exploration of stated preferences for middle-distance travel. *Transportation*, 38:383–407.

Wheeler, D., Shaw, G., and Barr, S. (2004). *Statistical Techniques in Geographical Analysis*. John Wiley & Sons, Inc.

A Technical Appendix

A.1 The constrained Bayesian finite mixture model

Let $S = (S_1, ..., S_n)$, represent the predominant behavioural group to which each individual $i = 1, ..., n$ is allocated, η_{ih} be the probability of individual i being in group h, $h = 1, ..., H$, and $z_i = (1, z_{i1}, ..., z_{iD})$ be individual i's response to a set of D 'group identifier' survey questions, e.g. age and gender, then:

$$\begin{aligned} S_i &\sim \text{Multinomial}(1; \eta_{i1}, ..., \eta_{iH}), \\ \log(\eta_{ih}/\eta_{i1}) &= z_i^T \alpha_h, \\ \Rightarrow \eta_{ih} &= \frac{\exp(z_i^T \alpha_h)}{\sum_{l=1}^{H} \exp(z_i^T \alpha_l)}, \quad \text{for} \quad i = 1, ..., n, \quad h = 1, ..., H, \\ \alpha_{hd} &\sim N(0, 5), \quad \text{for} \quad h = 2, ..., H, \quad d = 0, 1, ..., D. \end{aligned}$$

Let $y_i = (y_{i1}, ..., y_{ir})$, represents the frequency with which individual i behaves in r apposing ways over m_i occasions, θ_{ihj} be the probability of individual i, behaving as in option j, $j = 1, ..., r$, on any given occasion when assigned to group h, $h = 1, ..., H$, and $x_i = (1, x_{i1}, ..., x_{iC})$ be individual i's response to a set of C 'behavioural influencer' survey questions, e.g. the influence of weather conditions, then:

$$\begin{aligned} y_i | S_i = h &\sim \text{Multinomial}(m_i; \theta_{ih1}, ..., \theta_{ihr}), \\ \log(\theta_{ihj}/\theta_{ih1}) &= x_i^T \beta_{hj}, \\ \Rightarrow \theta_{ihj} &= \frac{\exp(x_i^T \beta_{hj})}{\sum_{k=1}^{r} \exp(x_i^T \beta_{hj})}, \quad \text{for} \quad i = 1, ..., n, \quad h = 1, ..., H, \quad j = 1, ..., r, \qquad (8) \\ \beta_{hjc} &\sim N(0, 5), \quad \text{for } h = 1, ..., H, \quad j = 1, ..., r, \quad c = 1, ..., C, \\ \beta_{hj0} &\sim N(\mu_{hj0}, 0.1), \quad \text{for } h = 1, ..., H, \quad j = 1, ..., r, \\ \mu_{hj0} &= \begin{cases} 0.6, & \text{if } h = j, \\ -4, & \text{otherwise.} \end{cases} \end{aligned}$$

A.1.1 Constraining the Bayesian Prior

The Bayesian prior distribution for the β regression coefficients are used to structurally constrain the model to group individuals based on apposing predominant behaviours, creating the clear group narrative. Let θ_{hj} represent the probability of any individual in group h taking behavioural option j. We specify these parameters constraints such that $\theta_{hh} > \theta_{hj}$ for $h \neq j$. That is, 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, in this commuting behaviour application we create the predominant behavioural group narrative such that, in group 1, the prior probability of commuting by motor vehicle (transport mode 1) 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 $x = (1, x_1, ..., x_C)$ be any possible combination of responses to the C behavioural influencers survey questions. By Eqn (8), the constraints that $\theta_{hh} > \theta_{hj}$ for $h \neq j$ are represented in terms of the regression parameter to be inferred, β, as, for group 1 ($h = 1$),

$$\begin{aligned} & \theta_{1j} < \theta_{11} \quad \text{for } j = 2, ..., r, \\ \Rightarrow \quad & \exp(x^T \beta_{1j}) < \exp(x^T \beta_{11}), \\ \Rightarrow \quad & x^T \beta_{1j} < 0, \end{aligned} \tag{9}$$

since $\beta_{h1} = (0, ..., 0)$ for $h = 1, ..., H$. Similarly, for groups 2-5 ($h = 2, ..., 5$),

$$\begin{aligned} & \theta_{hj} > \theta_{h1} \quad \text{for } h = j, \\ \Rightarrow \quad & \exp(x^T \beta_{hj}) > \exp(x^T \beta_{h1}), \\ \Rightarrow \quad & x^T \beta_{hj} > 0, \end{aligned} \tag{10}$$

and,

$$\begin{aligned} & \theta_{hj} < \theta_{h1} \quad \text{for } h \neq j, \\ \Rightarrow \quad & \exp(x^T \beta_{hj}) < \exp(x^T \beta_{h1}) \\ \Rightarrow \quad & x^T \beta_{hj} < 0. \end{aligned} \tag{11}$$

These constraints create boundaries within the multidimensional β parameters space, beyond which MCMC samples of the β regression parameters are discarded within the Bayesian model fitting MCMC algorithm (see A.2).

A.2 Model Fitting in `rstan`

Model fitting is carried out using the Bayesian modelling language Stan via `rstan`. Stan samples from the posterior using Hamiltonian Monte Carlo (HMC), a Markov chain Monte Carlo (MCMC) method that uses techniques from differential geometry to generate efficient sampling transitions, spanning the full marginal variance of the target distribution (Betancourt and Girolami, 2013). Since the HMC algorithm evolves using Hamilton's differential equations, it does not provide sampling for discrete parameters (Stan Development Team, 2016). Therefore, the posterior of the discrete group allocation indices, $S = (S_1, ..., S_n)$, cannot be sampled directly and must be integrated out of the model calculations. Group allocation is then carried out a posteriori. The unknown regression coefficient parameters α and β are therefore sampled from their joint posterior, integrating over S,

$$p(\alpha, \beta | y, \gamma, \xi, \mu, \sigma) \propto \int p(y|S, \beta, \mu, \sigma) p(S|\alpha, \gamma, \xi) p(\alpha|\gamma, \xi) p(\beta|\mu, \sigma) dS$$

Since S is discrete, this is equivalent to

$$p(\alpha, \beta|y, \gamma, \xi, \mu, \sigma) \propto \prod_{i=1}^{n} \left(\sum_{h=1}^{H} Pr(S_i = h|\alpha, \gamma, \xi) p(y_i|\beta_h, \mu_h, \sigma_h) \right) p(\alpha|\gamma, \xi) p(\beta|\mu, \sigma) \tag{12}$$

For group assignment, the probability of individual i being assigned to each group, $h = 1, ..., H$, is calculated for all posterior samples of α and β using the posterior predictive distribution of S_i,

$$\Pr(S_i = h|y_i, \alpha, \gamma, \xi, \beta_h, \mu_h, \sigma_h) = \frac{Pr(S_i = h|\alpha, \gamma, \xi) p(y_i|\beta_h, \mu_h, \sigma_h)}{\sum_{h=1}^{H} (Pr(S_i = h|\alpha, \gamma, \xi) p(y_i|\beta_h, \mu_h, \sigma_h))}. \tag{13}$$

Individual i is assigned to the group h that maximises 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.

In this application, the MCMC chain was run for 160,000 iterations to reach convergence and a further $M = 20,000$ iterations were retained as the posterior samples of α and β, used in the presentation of results in Section 4.