Using posterior analysis to predict missing information in passively collected data

Azam Ali*[1], Thijs Dekker[2], Stephane Hess[3], and Charisma F. Choudhury[4]

[1]PhD Student, Choice Modelling Centre & Institute for Transport Studies, University of Leeds, UK
[2]Associate Professor, Choice Modelling Centre & Institute for Transport Studies, University of Leeds, UK
[3]Professor, Choice Modelling Centre & Institute for Transport Studies, University of Leeds, UK
[4]Professor, Choice Modelling Centre & Institute for Transport Studies, University of Leeds, UK

Short summary

Trip diaries are increasingly being collected using passive manner for instance via smartphone surveys, where travel information (e.g. travel modes, purposes) is *inferred* and the participants are asked to *validate* or *correct* the inferred information. However, many people neglect to validate their trip resulting in large shares of unvalidated or missing data. To better predict missing data for individuals who have validated some of their trips, we propose the use of posterior analysis. Posterior analysis makes use of the Bayes rule to find the likely location of an individual on a population distribution by conditioning on the individual's previously observed choices. Using a two-week long trip diary dataset collected in the UK, we find that by making use of posteriors, the average probability of inferring the chosen alternative in a trip purpose model on a testing dataset substantially increases from 0.47 to 0.57 compared to without using posteriors.

Keywords: Discrete choice modelling, Big data analytics, Missing Data, Posterior Analysis, Bayesian Estimation

1 Introduction

With the ubiquitous use of smartphones, many travel surveys are now carried out using smartphone-based travel survey applications. These survey applications mostly use algorithms to infer travel details such as modes and purposes from GPS tracks of individuals and the participants are asked to *validate* or *correct* the inferred information. While this makes it possible to collect high-resolution mobility data from larger samples with minimal user input, in many cases, the participants do not validate all the trips. This leads to a large proportion of unvalidated trips or missing data. For example, 70% of the users who downloaded the application for a time-use and travel survey in Switzerland did not validate their trips (Winkler et al., 2022). A large body of research has been carried out to model (and later predict) trip purposes using machine learning methods, often leveraging data on geo-spatial context (Servizi et al., 2021). However, one of the aspects which has been overlooked in previous studies is the fact that missing data is of two types, i.e. there are some users who validate some of their trips, while there are also some users who fail to validate any of their trips. The former type opens up the possibility of using previously observed choices of individuals to better predict future choices by making use of posterior analysis in mixed logit models.

Mixed multinomial logit models are considered as the workhorse of econometric modelling (Train, 2009). Mixed logit models are typically used to model random or unobserved heterogeneity across individuals by using a mixing distribution. In simpler words, instead of modelling a fixed taste parameter β for the complete population, it is considered that individuals have a taste distribution $g(\beta|\Omega)$ where Ω are the parameters of a distribution such as a normal distribution. Using Bayes Rules, the conditional distribution of an individual n based on the observed choices y which is denoted by $h(\beta_n|y_n, x_n, \Omega)$ can be estimated, which is called the posterior distribution, whereas the former distribution $g(\beta|\Omega)$ is called the unconditional distribution. As the posterior distribution makes use of more information including past observations, it is possible to better infer the likely

location of an individual on that distribution, i.e. with a tighter confidence interval (Train, 2009). This has also been referred to as individual level parameters by Train (2009). Posterior analysis has previously been used to better understand choice behaviour by carrying out cluster analysis on the posteriors (Train, 2009), estimating statistics such value of time and willingness to pay for sub-populations or individuals (Hess, 2010; Train, 2009; Sarrias, 2020), and better predicting new choices for individuals (Dumont et al., 2015; Train, 2009; Danaf et al., 2019; Song et al., 2018). Song et al. (2018) also make use of these posteriors to generate personalised recommendations for a travel survey application.

In terms of modelling trip purposes, many studies now use machine learning algorithms (Servizi et al., 2021) where it is not possible to carryout a similar posterior analysis. The use of mixed logit has been used in two studies (Hossain and Habib, 2021; Liu et al., 2023), but they do not utilise posterior analysis. Further, nearly all of the studies which previously have used posterior analysis for estimating discrete choice models, rely on stated preference data (Danaf et al., 2019; Dumont et al., 2015). However, no study to the best of our knowledge has made use of posterior analysis on revealed preference dataset with large number of repeated observations. This paper aims to fill in the research gap by utilising posterior analysis to predict missing or unvalidated data in a long duration panel dataset. The performance of the proposed method to predict missing trip purposes is tested using a holdout sample. We also propose to test the performance of the proposed framework by using the predicted trip purposes in a mode choice model and comparing the value-of-time and other outputs.

2 Data

The data used in this study is from a two-week long travel survey conducted via a smartphone based travel survey application in the UK (Calastri et al., 2020). The survey required validation from the users to ensure that the trip stops, modes, and purposes were accurately inferred by the survey application. A total of 41,210 trips were carried out by 580 individuals in the Yorkshire and Humber Region, UK. Of these, 11,130 trips had missing or unvalidated trip purposes. As indicated in figure 1, nearly 15% of the individuals validated all of their trips, whereas 10% of the individuals failed to validate any of their trips. The remaining individuals failed to validate around 10 to 30% of their trips. This reveals that there is a large proportion of trips in the dataset which can potentially be predicted using posteriors.

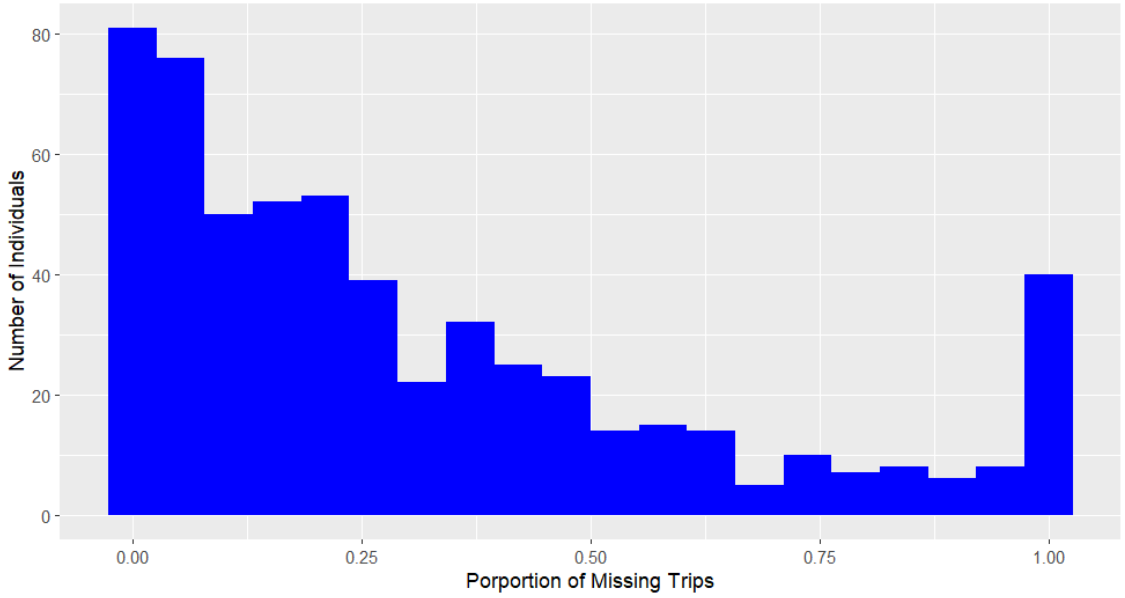

Figure 1: Histogram of Missing Trip Purpose in dataset

After removing missing data (including missing trip diaries and non-response to socio-demographic questions) there are a total of 24,134 trips carried out by 498 individuals. There were 31 trip

purposes in the original dataset; however, for analysis the trip purposes have been aggregated into 5 main purposes as indicated in table 1. The dataset was further augmented by adding the 20 nearest points of interest and distance to home.

Table 1: Aggregated Trip Purposes in cleaned dataset

Trip Purposes	**Count**	**Percentage %**
Commute (Work/Education)	3711	15.38
Business	1685	6.98
Return to Home	7579	31.4
Shopping/Services/Caring/other	6863	26.37
Leisure (Entertainment, Visit Friends, Hobbies)	4796	19.87

To find the impact of using posteriors, the dataset was split into a training and testing dataset. Two different types of missing data were considered in the testing dataset. In order to have some observations for individuals who have all of their trip purposes unvalidated, 10% of the users were set aside in the testing dataset which accounts for 2,414 observations from 49 individuals. For the remaining individuals, 20% of their total observed trips were randomly set aside which leads to 4,342 observations. The training dataset consists of 17,378 trips from 449 individuals.

3 Methodology

The basis of multinomial logit models is utility maximisation theory which states that each decision maker assigns a utility to each alternative in the choice set and chooses the alternative with the highest utility. However, as utility is not fully observable to the modeller, a probabilistic approach is considered which leads to the formation of random utility theory (Train, 2009). A decision maker n associates a utility $U_{n,i}$ for the alternate i which is the sum of a deterministic or observable utility $V_{n,i}$ and a random or error component $\varepsilon_{n,i}$ which is type 1 extreme value distribution in the case of multinomial logit model (MNL). The observable utility is a function of taste parameter β and explanatory variables x, i.e. $f(\beta, x)$. The probability of n choosing alternative i is given by eq 1 and the model parameters i.e. β are estimated by maximising the log-likelihood (LL).

$$U_{n,i} = V_{n,i} + \varepsilon_{n,i} \tag{1}$$

$$P_{n,i} = \frac{e^{\beta\ x_{n,i}}}{\sum_{j=1}^{J} e^{\beta\ x_{n,j}}} \tag{2}$$

In mixed multinomial logit models (MMNL), β is considered to be a distribution which leads to the probability equation as 3. Since the probability for MMNL does not have a closed form solution and requires an integration over all β, simulation approach is required. There are further two methods to estimate MMNL, i.e. the classical approach using draws or Bayesian estimation. In this study, a Bayesian approach is considered where the posteriors β_n are estimated by default and is often faster compared to the classical approach when there are a high number of random parameters. For more details readers are referred to Train (2009).

$$P_{n,i} = \int \frac{e^{\beta\ x_{n,i}}}{\sum_{j=1}^{J} e^{\beta\ x_{n,j}}} f(\beta) d(\beta) \tag{3}$$

In order to calculate the probability for new choices, i.e. x_{nT+1} in a mixed logit model, there are two methods, namely either using posteriors or relying on the sample level distribution only, as mentioned in eq 4 and 5, respectively. Posteriors can only be used in the case where there are some previously observed choices for the individual, whereas the unconditional probability can be used for all cases. Moreover, there are two ways to predict probabilities for mixed logit model, i.e. by using point averages of β or β_n or by using the complete distribution, i.e. $g(\beta|\Omega)$ or $h(\beta_n|y_n, x_n, \Omega)$ and then averaging the result. The former is problematic as this would indicate that there are point estimates for individual whereas we can only estimate the likely distribution of an individual given their previously observed choices. Even though the first approach is wrong, we report both method's results.

$$P(i|x_{n_{T+1}}, y_n, x_n, \theta) = \int L_{n_{T+1}}(i|\beta)h(\beta_n|y_n, x_n, \theta)d\beta \tag{4}$$

$$P(i|x_{n_{T+1}}, \theta) = \int L_{n_{T+1}}(i|\beta)g(\beta|\theta)d\beta \tag{5}$$

4 Results and discussion

In order to model trip purpose, a mixed multinomial model was estimated using the Hierarchical Bayesian approach. All of the parameters were considered as random (normally distributed) with a full covariance matrix. 60,000 burn-in and 40,000 post burn-in draws were selected after inspecting the Markov chains.

Table 2 shows the predictive performance of the mixed multinomial logit model on the testing dataset. It can be observed that in the testing dataset 1, which consists of individuals who do not have any previously observed choices, the performance of the MMNL model is better compared to the MNL model. It was considered necessary to have this testing dataset to benchmark the predictive performance of the model on individuals who are not present in the training dataset. Table 2 further shows the impact of using posteriors in testing dataset 2. It can be observed that by using posteriors, the average probability of the chosen alternative (APCA) increases from 0.466 to 0.569 and accuracy increases from 54.31 to 66.51%, in the case of using point estimates, compared to without using posteriors. There is a loss in the predictive performance when the complete posterior distribution is used, with a drop from 0.569 to 0.532, and from 66.51 to 62.47% in APCA and accuracy respectively. This is expected as by making use of the complete distribution, there is more uncertainty across a person taste coefficients. It was further observed that by using posterior analysis, 63.2% of the observations have an increase in average probability of the chosen alternative by an average of 0.25 compared to without using posteriors. For the remaining observations, there is a decrease in the APCA by 0.15 compared to without using posteriors. Both of these results are similar to Train (2009) where nearly a quarter of choices had a worse off probability compared to without using posteriors but the gain in probability is twice as large as the decrease in probability. At an individual level, 80% (351 out of 437) of individuals had an increase in the total probability of chosen alternatives when posteriors are used. This indicates that for most individuals, there is a gain in the predictive performance by making use of their previously observed choices. It was observed that individuals who benefited from posteriors had an average of 52.2 trips or observations in the training dataset whereas for individuals whose predictive performance decreased due to posteriors had an average of 39.4 trips. Therefore, posterior analysis should be used when the total number of observations per individual are high.

In terms of the estimates of the trip purpose model, most of the coefficients as mentioned in table 3 are intuitive. The predictive accuracy of the model is satisfactory compared to other studies, as the model with the highest accuracy is 66.17% which is similar to Ermagun et al. (2017) trip purpose model that has an accuracy of 64.17%. Nevertheless, the main finding of this study is the application of posterior analysis to predict missing data.

Table 2: Predictive Performance of estimated model on Testing Datasets

Metric	MNL	MMNL - Using point estimates		MMNL - Using distribution	
		Without Posteriors	With Posteriors	Without Posteriors	With Posteriors
Testing Dataset 1 Consists of individuals with previous choices unobserved (n=49, Observations 2,414)					
APCA	0.424	0.478	-	0.478	-
Accuracy(%)	56.88	54.68	-	54.45	-
Testing Dataset 2 Consists of individuals with previous choices observed (n=437, Observations 4,342)					
APCA	0.412	0.466	0.569	0.466	0.532
Accuracy(%)	54.88	54.31	66.51	54.92	62.47

Table 3: Estimates for Mixed Multinomial Logit Model

Parameter	Commute		Business		Leisure		Shopping	
	Post means	Post std	Post means	Post std	Post means	Post std	Post means	Post std
asc	-13.2	0.6	-9.95	0.55	0.82	0.13	-0.56	0.14
β *weekday*	4.22	0.22	3.78	0.24	-0.43	0.04	0.002	0.025
β *occupation*	0.6	0.09	1.73	0.15	-	-	-	-
β *distance*	1.65	0.07	1.47	0.07	0.95	0.04	0.73	0.03
β *time 0 to 9 hrs*	6.58	0.38	3.98	0.26	0.21	0.11	1.48	0.1
β *time 9 to 16 hrs*	3.78	0.26	3.32	0.24	-0.03	0.09	0.91	0.05
β *time 16 to 19 hrs*	0.44	0.2	0.6	0.26	-0.34	0.08	0.19	0.08
β *restaurants*	0.09	0.04	0.102	0.013	0.14	0.02	0.099	0.009
β *services*	0.22	0.03	0.0142	0.001	-0.043	0.01	0.028	0.01
β *transportation*	0.138	0.01	0.012	0.015	-0.022	0.013	0.09	0.01
β *shopping*	0.05	0.06	0.043	0.01	0.3	0.05	0.43	0.05
β *entertainment*	0.42	0.05	0.17	0.06	0.24	0.03	0.027	0.013
β *office*	-0.32	0.09	-0.172	0.012	-0.013	0.012	0.32	0.03
β *industrial*	0.07	0.08	0.013	0.01	-	-	-	-
β *green*	-	-	-	-	0.37	0.12	0.037	0.016
β *education*	0.088	0.009	-	-	-	-	-	-
Initial Log-likelihood	-27,968.81							
Final Log-likelihood	-17,020.86							

5 Conclusions and Next Steps

In this research, we demonstrate the benefits of using posterior analysis to predict missing data in a 2-week long panel dataset collected in the UK. Posterior analysis makes use of previously observed choices of an individual to better find the likely location of an individual over a random distribution of the coefficients. A trip purpose model is developed by using a mixed multinomial logit modelling framework that is estimated by the hierarchical Bayesian approach. It was observed that there is a substantial improvement in predictive performance of mixed logit when the posterior distributions are used as the average probability of the selected alternative increases from 0.47 to 0.53, and the accuracy increases from 54.31% to 62.5% in a testing dataset compared to the case when posteriors are not used. The results advocate the benefits of using posterior methods for prediction.

For the full paper, we will improve the specification of the preliminary model and compare it with a mixed multinomial probit model estimated using the Hierarchical Bayesian approach. We will also do more rigorous validation by using the probabilities obtained from the trip purpose model in a mode choice model. This would demonstrate the benefits of improved prediction on valuation, i.e. the value of time for different purposes.

References

Calastri, C., Crastes dit Sourd, R., and Hess, S. 2020. We want it all: experiences from a survey seeking to capture social network structures, lifetime events and short-term travel and activity planning. *Transportation*, 47(1):175–201.

Danaf, M., Becker, F., Song, X., Atasoy, B., and Ben-Akiva, M. 2019. Online discrete choice models: Applications in personalized recommendations. *Decision Support Systems*, 119:35–45.

Dumont, J., Giergiczny, M., and Hess, S. 2015. Individual level models vs. sample level models: contrasts and mutual benefits. *Transportmetrica A: Transport Science*, 11(6):465–483.

Ermagun, A., Fan, Y., Wolfson, J., Adomavicius, G., and Das, K. 2017. Real-time trip purpose prediction using online location-based search and discovery services. *Transportation Research Part C: Emerging Technologies*, 77:96–112.

Hess, S. 2010. Conditional parameter estimates from Mixed Logit models: distributional assumptions and a free software tool. *Journal of Choice Modelling*, 3(2):134–152.

Hossain, S. and Habib, K. N. 2021. Inferring the Purposes of using Ride-Hailing Services through Data Fusion of Trip Trajectories, Secondary Travel Surveys, and Land Use Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2675(9):558–573.

Liu, Y., Miller, E. J., and Habib, K. N. 2023. Inferring Trip Destination Purposes for Trip Records Collected through Smartphone Apps. *Journal of Transportation Engineering, Part A: Systems*, 149(2):04022145.

Sarrias, M. 2020. Individual-specific posterior distributions from Mixed Logit models: Properties, limitations and diagnostic checks. *Journal of Choice Modelling*, 36:100224.

Servizi, V., Pereira, F. C., Anderson, M. K., and Nielsen, O. A. 2021. Mining User Behaviour from Smartphone data: a literature review. *European Transport Research Review*, 13(1):57. arXiv:1912.11259 [cs, stat].

Song, X., Danaf, M., Atasoy, B., and Ben-Akiva, M. 2018. Personalized Menu Optimization with Preference Updater: A Boston Case Study. *Transportation Research Record: Journal of the Transportation Research Board*, 2672(8):599–607.

Train, K. E. 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2nd edition.

Winkler, C., Meister, A., Schmid, B., and Axhausen, K. W. 2022. TimeUse+: Testing a novel survey for understanding travel, time use, and expenditure behavior. page 17 p.