Semantic Enrichment through Inverse Discrete Choice Modelling: A Mutual Information Approach

Yuanying Zhao[1], Jacek Pawlak, John W. Polak

Centre for Transport Studies, Imperial College London

Urban Systems Laboratory, Imperial College London

1. Background and Rationale

The growing availability of transport big data[2] from information and communications technologies has stimulated substantial discussion regarding their applicability for detailed travel behaviour analysis (Pawlak *et al.*, 2015). In parallel to the enthusiasm created by such opportunities, researchers have become increasingly aware of the major limitations of such data; particularly their lack of semantic content – typical big data sources containing little information on data point. For example, unless supplemented with an add-on survey, data from a GPS logger typically provide numerous data points with accurate geographical coordinates and timestamps but no readily accessible and meaningful information on the respondents or their activities. Since such contextual information is critical for travel behaviour analysis and policy-making practices, the weak semantics of transport big data can significantly limit analytical depth of the possible insights and can even lead to inferences erroneous and misguiding policy-making.

In response to the limitation outlined above, there has been a growth of research which aim to add information such as mode of transport (Brunauer *et al*., 2013), location visit pattern (Isaacman *et al*., 2011), trip purpose (Wolf *et al.*, 2001; 2004), and activity type (Calabrese *et al.*, 2010) to typical transport big data sources. A few attempts have also been made to attach socioeconomic attributes (Gebru *et al.*, 2017; Auld *et al.*, 2015; de Montjoye *et al.*, 2013). To our best knowledge, however, this area remains in its infancy with techniques that lack solid microeconomic behavioural theories, or theoretical understanding of their properties under different data conditions.

In addressing this gap, we proposed an inverse discrete choice modelling (IDCM) framework for data enrichment, drawing upon the extensive body of theoretical and empirical results developed in the field of discrete choice modelling (Zhao *et al.*, 2017a;

[1] Corresponding author: yuanying.zhao14@imperial.ac.uk

[2] Novel sources of data with voluminous data points, which are directly or indirectly obtained to have captured respondents' spatio-temporal movements.

Pawlak *et al.*, 2015). In particular, the IDCM approach proposes that demographics and preference structure captured in the form of a discrete choice model (DCM) can serve as a mechanism for inferring attributes of the decision maker from observed travel choices (e.g. time of day, route and, in certain circumstances, mode). Moreover, the probabilistic nature of the IDCM ensures individual-level privacy whilst retaining enough information for sample-level analysis, which can accommodate the growing need and policy trend for privacy preservation, e.g. the General Data Protection Regulation (Blackmer, 2016), yet capitalising on the richness of big data. This warrants further exploration into this approach.

2. Aims and Objectives

Findings from previous studies based on both simulated data (Pawlak *et al.*, 2015) and revealed preference data (Zhao *et al.*, 2017b) have shown that the performance of the IDCM approach is highly sensitive to the explanatory power (EP) of the imputed variable in the DCM specification. In particular, Pawlak *et al.* (2015) showed that the enrichment quality of the IDCM at individual level, measured by the 'percentage correctly predicted' (PCP) values of the imputed variables, improves as the EP increases. In addition, enrichment of discrete variables with multiple outcome categories was shown substantially more challenging than that for dichotomous ones. An extension of this study by Zhao *et al.* (2017b) used a more formal representation of the EP, i.e. the mutual information (MI) between the imputed variable and the choice (Cover & Thomas, 2012). The MI was employed because of its firmer theoretical grounding, in comparison to McFadden's Rho-squared, in information and probability theories as a more universal metric for measuring both linear and non-linear associations between variables. In addition to earlier findings, this empirical work suggested the existence of diminishing marginal improvement in PCP associated with growing MI for variables of the same type with the same number of outcome categories, which motivates the current study to develop a formal link between performance of the IDCM approach and the MI between observed choices and the imputed variable.

Specifically, we seek to achieve this aim through three objectives:

1) to explore variation in PCP and sample shares with respect to changes in MI values for specified discrete variable types with specified number of outcome categories by conducting detailed Monte Carlo (MC) experiments;
2) to generalise findings in 1) by relating them to the DCM and IDCM theories; and
3) to implement the findings in an empirical context of enriching a real-world transport big dataset.

3. Methodology

A general setting for the MC experiments for the aforementioned objectives is devised and presented in Table 1. In particular, three series of experiments will be conducted to impute nominal variables with respectively 2 and 3 outcome categories and an ordinal variable with 3 values. For controlling the number of independent variables, the number of choice alternatives is fixed. Variables with 3+ outcome categories are not considered as previous studies suggested a significant decrease from an additional category (Zhao *et al.*, 2017a; 2017b). Due to the nature of MC methods, the analysis could represent any case of imputation of corresponding variable types with the specified number of outcome categories.

Table 1. Experiment Settings for Monte Carlo Simulation

Exp. No.	Type of Var.	Number of Cat.	Number of Exp.	Repe-titions	Influential factor to IDCM performance	
1	Nominal	2	4045	20	Number of outcome categories	—
2	Nominal	3	115225	20		Type of discrete variables
3	Ordinal	3	115225	20	—	

In each experiment, a set of probabilities $P(Y)$ and joint probabilities $P(X,Y)$ are defined and thus fix the MI level. The corresponding conditional probabilities $P(Y|X)$ which characterise the correlation between the imputed variable X and choices Y in the form of a DCM can be calculated. In information and probability theories, the MI between two discrete variables X and Y is formalised as (Cover & Thomas, 2012):

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \qquad \text{(Eq. 1)}$$

where

$p(x,y)$ joint probability distribution function of X and Y;

$p(x)$ and $p(y)$ marginal probability distribution functions of X and Y.

By invoking Bayes' theorem, Eq. 1 can be transformed as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x|y)p(y) \log \frac{p(y|x)}{p(y)} \qquad \text{(Eq. 2)}$$

Eqs. 1-2 show that the MI does not rely on specific distributional assumptions and it is non-linear. A same MI can be produced by different conditional probabilities $P(X|Y)$ that infer different levels of IDCM enrichment quality. Moreover, deriving $P(Y|X)$

which is required for DCM parameter estimation and data simulation for a particular value of MI would be problematic. Nevertheless, it is simpler in an implicit way, i.e. computing MI levels given various specified $P(Y|X)$, which can be calculated given $P(Y)$ and $P(X,Y)$. By changing $P(Y)$ and $P(X,Y)$ by 0.1 and 0.01 respectively in each experiment, number of experiments can be designed respectively for variables with 2 and 3 outcome categories (Table 2).

The sample simulated for each experiment is randomly split into 2 subsamples respectively used for computing the MI and the IDCM enrichment. A cross-validation in the form of a k-fold holdout method (Kohavi, 1995) with $\mathrm{k} = 20$ as suggested by Kohavi is employed account for possible sample specificity.

4. Expected Findings and Applications

Given Eq.1, the upper and lower bounds of MI can be determined by the limiting correlation between X and Y. The MI reaches its minimum when X and Y are independent of each other, i.e. $p(x,y) = p(x)p(y)$, as:

$$I_{\min}(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x)p(y)\log 1 = 0 \qquad \text{(Eq. 3)}$$

While the dependence between X and Y is deterministic, i.e. $p(x|y) = p(y|x)$, the maximum MI is achieved as Equation 1 reduces to:

$$I_{\max}(X;Y) = -\sum_{x\in X} p(x)\log p(x) = -\sum_{y\in Y} p(y)\log p(y) \qquad \text{(Eq. 4)}$$

A formal mathematical relationship between the EP and performance of the IDCM, represented by $P(X|Y)$, in between the limiting cases will be explored starting from Equation 2. For the imputation of a dichotomous variable $X = \{x, \neg x\}$ from a binary choice $Y = \{y, \neg y\}$ through the IDCM approach, Equation 2 is equivalent to:

$$\begin{aligned} I(X;Y) = p(x|y)\,p(y)[\log\frac{p(y|x)}{p(y|\neg x)} + \frac{p(\neg y|x)}{p(y|x)}\log\frac{p(\neg y|x)}{p(\neg y|\neg x)}] \\ + p(y)\log\frac{p(y|\neg x)}{p(y)} + p(\neg y)\log\frac{p(\neg y|\neg x)}{p(\neg y)} \end{aligned} \qquad \text{(Eq. 5)}$$

$P(Y)$ and $P(Y|X)$ are knowable in empirical contexts. The $p(x|y)$ in Equation 5 is hence a linear function of the MI:

$$p(x|y) = kI(X;Y) + b \qquad p(x|y) \in [0.5,1] \qquad \text{(Eq. 6)}$$

where

$$k = \frac{1}{p(y)}[\log\frac{p(y|x)}{p(y|\neg x)} + \frac{p(\neg y|x)}{p(y|x)}\log\frac{p(\neg y|x)}{p(\neg y|\neg x)}]^{-1}$$

$$b = -\frac{p(y)\log\frac{p(y|\neg x)}{p(y)} + p(\neg y)\log\frac{p(\neg y|\neg x)}{p(\neg y)}}{p(y)[\log\frac{p(y|x)}{p(y|\neg x)} + \frac{p(\neg y|x)}{p(y|x)}\log\frac{p(\neg y|x)}{p(\neg y|\neg x)}]}$$

As the coefficient k can be proved non-negative, the likelihood $p(x|y)$ of identifying a choice maker is characterised by attribute x from his/her choice y through IDCM enrichment increases as the MI grows. However, this representation fails for variables with 2+ categories to capture the link as $P(X|Y)$ is asymmetric over its domain. We hereby introduce a new quantity for measuring the imputation quality (IQ) for variables with 2+ categories using the IDCM.

For a number of variables with a fixed summation, e.g. $\sum_{x\in X} p(x|y) = 1$, their product reaches its maximum when the values of these variables are equal. In such case, however, the IQ is minimised as the likelihood of inferring any outcome category of X from observed y is purely random. As one $p(x|y)$ $(x \in X)$ increases infinitely close to 1 while the others decreases almost to 0, their product, on the other hand, reaches infinitely to its minimum and the IQ is almost maximised. Hence the negative of the product is used as a measurement of the IQ:

$$IQ(X|Y) = -\prod_{x\in X} p(x|y) \quad y \in Y \qquad \text{(Eq. 7)}$$

As part of the contribution, we will explore extension of the link between IQ and the MI.

In terms of the MC experiments, we will report results showing the sensitivity of PCP and sample shares with respect to the MI to explore the reason for diminishing marginal PCP improvement with MI growth revealed in the previous work (Zhao *et al.*, 2017b). Additionally, the enrichment quality with respect to the type and the number of outcome categories of the imputed variable will also be explored according to Table 1. Eqs. 4-7 jointly with the MC experiment outcome link the forward DCM specification with expected performance of the IDCM enrichment. This closes the chain between a priori information possessed by the researcher and the expected quality of enrichment and hence enables assessment of the expected outcome before conducting the enrichment, which is novel as compared to existing approaches where performance is established by post-enrichment cross-validation.

As a final step, we will seek to demonstrate applicability of the IDCM approach in the context of enriching anonymous mobile network data with the socioeconomic information of the anonymised mobile users, e.g. their age, gender, and income level. Because of the stochastic nature of the IDCM, it reveals only the probability distribution of the random variable describing the imputed variable which reduces intrusion into individual privacy while still permitting aggregate, sample-level analysis.

It is expected that this study will further enable understanding robustness and properties of inferring an attribute through observed choice behaviours using the IDCM approach. The contribution is hence an important step towards establishing statistical properties and thus credibility of the IDCM approach under a wide set of data conditions. This is crucial given absence of such rigour in enrichment studies to date. In addition, we note this as an essential pre-requisite for further extensions of the IDCM approach, such as using multiple choices or multiple individual-level models.

References

Auld, J., Mohammadian, A., Simas Oliveira, M., Wolf, J. & Bachman, W. (2015) Demographic Characterization of Anonymous Trace Travel Data. Transportation Research Record: Journal of the Transportation Research Board. (2526), 19-28.

Blackmer, W. (2016) GDPR: Getting Ready for the New EU General Data Protection Regulation. Information Law Group, InfoLawGroup LLP, Retrieved. 22 (08), 2016.

Brunauer, R., Hufnagl, M., Rehrl, K. & Wagner, A. (2013) Motion pattern analysis enabling accurate travel mode detection from gps data only. 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). , IEEE. pp.404-411.

Calabrese, F., Pereira, F. C., Di Lorenzo, G., Liu, L. & Ratti, C. (2010) The geography of taste: analyzing cell-phone mobility and social events. International Conference on Pervasive Computing. , Springer. pp.22-37.

Cover, T. M. & Thomas, J. A. (2012) Elements of information theory. , John Wiley & Sons.

de Montjoye, Y., Quoidbach, J., Robic, F. & Pentland, A. (2013) Predicting Personality Using Novel Mobile Phone-Based Metrics. SBP. , Springer. pp.48-55.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L. & Fei-Fei, L. (2017) Using Deep Learning and Google Street View to Estimate the Demographic Makeup of the US. arXiv Preprint arXiv:1702.06683.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J. & Varshavsky, A. (2011) Identifying important places in people’s lives from cellular network data. International Conference on Pervasive Computing. , Springer. pp.133-151.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection.   International Joint Conference on Artificial Intelligence. , Stanford, CA. pp.1137-1145.

Pawlak, J., Zolfaghari, A. & Polak, J. W. (2015) Imputing Socioeconomic Attributes for Movement Data by analysing Patterns of Visited Places and Google Places Database: Bridging between Big Data and Behavioural Analysis. 4th International Choice Modelling Conference. 11-13 July, Austin, TX, USA.

Wolf, J., Bricka, S., Ashby, T. & Gorugantua, C. (2004) Advances in the application of GPS to household travel surveys. National Household Travel Survey Conference, Washington DC.

Wolf, J., Guensler, R. & Bachman, W. (2001) Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from GPS Travel Data. Proceedings of the 80th Annual Meeting of the Transport Research Board. Washington, D.C.

Zhao, Y., Pawlak, J. & Polak, J. W. (2017a) Privacy-preserving socioeconomic attribute enrichment for mapping of passively-derived OD matrices. The 2017 RGS-IBG Annual International Conference. 27 August - 1 September, London.

Zhao, Y., Pawlak, J. & Polak, J. W. (2017b) Inverse Discrete Choice Models (IDCM): Theoretical and practical considerations for imputing respondent attributes from the patterns of observed choices. 49th Universities' Transport Studies Group Conference. 4-6 January, Dublin, Ireland.