

Routledge
Taylor & Francis Group

Transportation Planning and Technology

ISSN: 0308-1060 (Print) 1029-0354 (Online) Journal homepage: http://www.tandfonline.com/loi/gtpt20

Inverse discrete choice modelling: theoretical and practical considerations for imputing respondent attributes from the patterns of observed choices

Yuanying Zhao, Jacek Pawlak & John W. Polak

To cite this article: Yuanying Zhao, Jacek Pawlak & John W. Polak (2018) Inverse discrete choice modelling: theoretical and practical considerations for imputing respondent attributes from the patterns of observed choices, Transportation Planning and Technology, 41:1, 58-79, DOI: 10.1080/03081060.2018.1402745

To link to this article: https://doi.org/10.1080/03081060.2018.1402745

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 14 Nov 2017.

Submit your article to this journal

Article views: 143

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=gtpt20

Download by: [Imperial College London Library] **Date:** 05 December 2017, At: 07:33

TRANSPORTATION PLANNING AND TECHNOLOGY, 2018
VOL. 41, NO. 1, 58–79
https://doi.org/10.1080/03081060.2018.1402745

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Inverse discrete choice modelling: theoretical and practical considerations for imputing respondent attributes from the patterns of observed choices

Yuanying Zhao, Jacek Pawlak and John W. Polak

Department of Civil and Environmental Engineering, Imperial College London, London, UK

ABSTRACT

The growing availability of geotagged big data has stimulated substantial discussion regarding their usability in detailed travel behaviour analysis. Whilst providing a large amount of spatio-temporal information about travel behaviour, these data typically lack semantic content characterising travellers and choice alternatives. The inverse discrete choice modelling (IDCM) approach presented in this paper proposes that discrete choice models (DCMs) can be statistically inverted and used to attach additional variables from observations of travel choices. Suitability of the approach for inferring socioeconomic attributes of travellers is explored using mode choice decisions observed in London Travel Demand Survey. Performance of the IDCM is investigated with respect to the type of variable, the explanatory power of the imputed variable, and the type of estimator used. This method is a significant contribution towards establishing the extent to which DCMs can be credibly applied for semantic enrichment of passively collected big data sets while preserving privacy.

ARTICLE HISTORY
Received 12 March 2017
Accepted 14 September 2017

KEYWORDS
Geotagged data; semantic enrichment; imputation; inverse problem; discrete choice; mutual information

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

1. Introduction

The growing availability of geotagged big data has stimulated substantial discussion regarding their usability for detailed travel behaviour analysis. Typically collected passively from information and communications technologies (ICTs) such as satellite positioning systems (e.g. GPS, GALILEO), payment transaction systems (e.g. London Oyster card) or mobile networks, geotagged data can provide information about spatio-temporal aspects of travel behaviour with greater accuracy and potentially at a lower cost than traditional travel surveys (Bohte and Maat 2009). Whilst extremely 'big' in terms of overall data volumes, geotagged data tend to be 'thin', that is, contain very few variables in each data point. The authors term this property as 'low semantic content', as opposed to 'thick' data sets with 'high semantic content' (e.g. travel surveys) in which the data contain a large number of variables describing travel behaviours as well as respondents and households.

In a typical 'thin' data set, such as that from a GPS logger, there are numerous records with accurate geographical coordinates and timestamps but no readily accessible semantic

CONTACT Yuanying Zhao yuanying.zhao14@imperial.ac.uk

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

information about the respondents who carry the GPS devices. The latter could only be obtained from an accompanying survey, which has been the most prevailing practice to date (Schönfelder and Antille 2002). In most contexts, however, follow-up surveys may be difficult or expensive to conduct, or even impossible due to privacy considerations. Since the contextual information, such as socioeconomic attributes of the traveller, purpose of travel and the nature of activities performed at the destination, is often critical for travel demand modelling, the low semantics of geotagged big data sources can hamper travel behaviour analysis in various ways. Such limitations include, for example, the hampering of exploration of heterogeneity in demand processes, barriers to applications of data-hungry modelling toolkits available to transport planners, or exacerbation of problems of unobserved or confounding factors which can lead to erroneous inferences and inefficient policy implications.

In response to both potential benefits and limitations of geotagged big data, recently there have been emergence of approaches seeking to enrich such semantically poor data sets, which have enabled effective imputation of additional variables unavailable in the original data set. It should be noted that the role of these enrichment approaches is different from imputation methods in the sense of traditional data missingness, for example, non-response in surveys (Andridge and Little 2010). While traditional methods are intended for dealing with missing values in a data set and to complete analysis as if it was complete, data missingness in the context of enrichment is 'extreme', that is, no observation contains the variable to be imputed. The purpose of enrichment approaches is therefore to attach additional variables to the original data set. The concept of statistical matching, of which enrichment is also a particular case, has been extensively discussed in D'Orazio, Di Zio, and Scanu (2006).

The inverse discrete choice modelling (IDCM) approach, proposed by Pawlak, Zolfaghari, and Polak (2015), and extended in this paper, aims at data enrichment by making use of the extensive body of empirical results developed in the field of discrete choice models (DCMs). DCMs are fundamental transport modelling and policy-making tools which take advantage of the natural prevalence of discrete choices (e.g. mode, route and destination) in transport contexts. They have been used and developed over the past decades, which proves their versatility and robustness. Relying on known behavioural foundations and assumptions firmly grounded in the microeconomic theory (Train and McFadden 1978), DCMs provide a way of linking attributes of individuals and discrete alternatives to specific decisions. We therefore seek to explore this functionality in an inverse way: to enrich data explicitly or implicitly capturing choice behaviours with additional variables characterising individuals.

Specifically, IDCM postulates that knowledge of choice sets, choices, and the preference structure captured in the form of a DCM provides a means of inferring attributes of the choice maker or choice alternatives. The probabilistic enrichment which IDCM leads to also has the side benefit of preserving privacy of a particular individual while obtaining the aggregate shares of attributes.

The application of IDCM in the context of transport modelling are various. For instance, enriching automatic number plate recognition camera data with travellers' socioeconomic attributes can enable the ability to advice on the best means of communicating a message to a particular user group, which can further increase the likelihood of more timely reaction and reduce exposure to the impacts of disruptions. As analysing

people's movement using OD matrices, where knowledge of people's demographics is essential, represents a vital tool for transport policy-making aimed at designing and operating a sustainable and equitable urban transport system, another application can be origin-destination (OD) matrix profiling. Such OD matrix profiling is desirable in cities in emerging countries where fixed monitoring infrastructures of regular data collection mechanisms are not always readily available. Moreover, enrichment of smart card data (e.g. London Oyster card data) can promote revenue generation from more detailed audience segmentation for on-board marketing, for example, advertisements on buses and undergrounds. And understanding of electric vehicle users will help develop more personalised charging services with different preferences revealed in their choice behaviours such as choices of charging durations and positions. The proposed methodology is also applicable in other domains such as humanitarian and disaster operations by enabling the identification of vulnerable individuals such as the elderly (de Montjoye et al. 2013).

In this paper, the theoretical foundations and consequent properties of the IDCM are systematically explored and formally codified. In addition, we present the first empirical application of the approach which was previously only used in a simulation study. Finally, the present contribution introduces the concept of mutual information (MI) and hence formalises the notion of explanatory power (EP) of a variable which lies at the heart of the IDCM performance.

This paper is structured as follows. Section 2 defines specific terms to facilitate understanding of the approach in the wider context of enrichment methods. Section 3 reviews the literature on existing data enrichment approaches and discusses solutions to inverse problems (IPs) in reference to IDCM. Section 4 formalises the IDCM using microeconomic and econometric foundations, develops research hypotheses based on previous studies, and introduces validation methods. Section 5 presents details about the enrichment exercise design using empirical data set and the IDCM, followed by discussing the findings in Section 6. Section 7 concludes the paper and provides suggestions for future research avenues.

2. Definition of terms

In order to facilitate understanding of this study, Section 2 provides the definition of several specific terms used in this paper.

Geotagged data: Geotagging, or georeferencing, involves the process of adding geographical identification metadata to data otherwise containing no information about spatial meanings (Hill 2009), such as geographical coordinates, or other identifiers of a specific location, for example, name of the place, ordnance survey grid reference or postcode. Increasingly, geotagging is done passively through built-in location technologies including satellite navigation and mobile network triangulation. In the context of travel behavioural analysis, a series of geotagged data collected from a particular respondent can be used to re-construct individual movement pattern, also called a 'trajectory' (Giannotti et al. 2007), which can be further used to analyse corresponding activity patterns. The methods discussed in this paper are particularly aimed at geotagged data that arise as a by-product of the functioning of operational systems (e.g. public transport payment systems, navigation and fleet management systems) although geotagged data are also relevant to data collected as part of a deliberated research design.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

Data semantics: The semantic content of data refers to the variables which characterise these data points, thus providing the meaning and describing use of the data (Wood 1985). Data semantics can therefore be viewed as a mapping between the information stored in the data and the real-world objects they represent (Sheth 1997), reflecting the extent to which the data have been interpreted, that is, the meaning implicitly or explicitly represented by the data (Smith 1990). By 'low semantics' or 'of low semantic content', we refer to data that contain very few variables characterising the data points themselves.

Semantic enrichment: Semantic enrichment involves the process of increasing semantic content of particular data. 'Enrichment' means adding supplementary information to the original data set by using other sources of information, for example, other databases or pattern recognition rules.

Imputation: Imputation originates from statistics where it represents the process of replacing missing data with substituted values (Rubin 2004) derived from either external information sources or statistical modelling procedure. Broadly speaking, imputation serves as a mechanism for semantic enrichment. It is important because missing data can introduce a substantial amount of bias in data analysis. Imputation can occur at a level of the whole data point (unit imputation) or a particular component of it (item imputation). In the present analysis, we aim to enrich observed respondents' choice data with a set of socioeconomic which can be classified as a form of 'item imputation'.

3. Literature review

The literature review below consists of two parts: Section 3.1 discusses existing approaches to data semantics enrichment, while Section 3.2 scopes the wider domain of IPs which IDCM draws upon.

3.1. Enrichment of data semantics

Follow-up surveys, for example, questionnaires or diaries, are the most obvious ways of enriching geotagged transport big data sources. They are applicable only to cases where the respondents can be approached. In most instances, however, follow-up survey approaches are not feasible, for example, on the grounds of costs and burden of additional data collection, due to privacy regulations, or simply because the original data set was collected so long ago that detailed recollection of particular behaviour would be dubious. Therefore, a number of alternative approaches seeking to enrich the original data without the need for follow-up contact with original respondents have emerged.

These approaches were developed based on the availability of mature movement map-matching and trajectory decomposition techniques. For instance, Lou et al. (2009) proposed global map-matching algorithm *ST-Matching* for low-sampling-rate GPS trajectories, that is, GPS data points collected between relatively long intervals. Dodge, Weibel, and Forootan (2009) suggested a segmentation and feature extraction method that can classify trajectory data of unknown moving objects and assign to known moving objects with learned movement characteristics. With the availability of the aforementioned techniques, the semantics of travel behaviour data sets have been significantly enriched by imputing, for example, modes of transport, trip destinations and purposes, or activity types.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

Early activity recognition studies relied on rather limiting assumptions where either types of places or routes between places were eliminated from analyses (Bennewitz et al. 2005), which has inevitably constrained the scope of studies. Giannotti et al. (2007), therefore, documented 'trajectory pattern' that characterises a collection of independent routes sharing the same sequences of visited places with similar travel times. This has enabled consistent description of frequent travel behaviours.

Researchers are also interested in spatial occurrences of certain movement attributes such as tourism attraction visits. For example, Ester et al. (1996) presented *Density-based Spatial Clustering of Applications with Noise* to classify large spatial databases. Based on this study, Andrienko et al. (2011) suggested a visual analytics procedure that determined places of interest based on time-varying characteristics of movements. Montoliu, Blom, and Gatica-Perez (2013) proposed a framework that combined time-based and grid-based clustering techniques to discover places-of-interest from mobile phone data collected through multiple sources.

Regarding the detection of transport modes, the most common approach seeks to infer mode based on average and maximum speeds, derived from the underlying positioning data (Gong et al. 2014). Additional information, like GIS land-use data, has been incorporated to increase detection accuracy (Chung and Shalaby 2005; Stopher, FitzGerald, and Zhang 2008). Nevertheless, a recent study (Brunauer et al. 2013) suggested a GPS-only travel mode detection approach using *feed forward multilayer perceptrons* that extracted and analysed distinct motion patterns of different modes, for example, acceleration and horizontal angular speed.

In terms of trip purpose identification, detailed GIS land-use data is often used in the development of relevant enrichment procedures. Wolf, Guensler, and Bachman (2001, Wolf et al. 2004) conducted two car-based studies illustrating that trip purposes could be accurately extracted from GPS data used incorporated with GIS land-use databases. Chen et al. (2010) applied a probabilistic model where time of day, history dependence and land-use characteristics were considered in two models to predict home and non-home-based trips.

Whereas substantial progress has been made in geotagged trace data enrichment, enrichment with information on socioeconomic attributes of respondents has received relatively limited treatment. In one of the few studies to have addressed this issue, explored ad hoc approaches for the imputation of demographic characteristics from traditional travel diary surveys, with however, mixed results. Gebru et al. (2017) demonstrated, in a recent study, a machine vision framework based on convolutional neural networks, to determine the demographic makeup using Google Street View images where characteristics of motor vehicles encountered in particular neighbourhoods were recognised as explanatory variables in regression models. The shares of income, race, education level and vote patterns were found highly related to the makeups of vehicles in corresponding regions.

Signifying the gap and potentials in socioeconomic attribute identification, this study seeks to address through a systematic development of an enrichment procedure, firmly grounded in the existing microeconomic foundations and cutting-edge DCM capabilities.

3.2. IPs and solution techniques

Travel behaviour models typically serve to describe travel-related decisions and their implications under certain assumptions. This is captured in the form of a mathematical

function that relates attributes of the respondent and decision-making environment to variables describing corresponding behaviour. DCMs, for example, are functions where the aforementioned attributes are related to choice behaviour. The enrichment procedure is therefore an approach in which variables of actual behaviour are used to infer attributes of respondents or decision-making environment. This corresponds to the umbrella term of IPs, that is, the process of inferring from a set of observations the causal factors that are believed to have produced them (Tarantola 1987).

If the direct problem is denoted by M and the mapping is from a functional space Q to another space R, it can be written as

$$r = M(q) \quad \text{for } q \in Q \text{ and } r \in R. \tag{1}$$

The corresponding IP M^{-1} amounts to finding points $q \in Q$ from knowledge of $r \in R$ such that Equation (1) or at least its approximation holds (Bal 2012):

$$q \cong M^{-1}(r) \quad \text{for } q \in Q \text{ and } r \in R. \tag{2}$$

IPs are crucial as they enable insights into parameters of the system that usually cannot be observed directly. Model calibration (parameter estimation) is an example of where parameters of the model are found to maximise its fit to observed data, which is also a routine procedure carried out as part of the DCM toolkit. Semantic enrichment is also an IP in which respondent attributes are inferred from observed behaviour.

The two key attributes of IPs concern linearity and stochasticity, which determine the complexity of the inversion process. In a non-linear case, same system state can arise from different initial inputs while in a stochastic case, same initial input can lead to different solutions. The occurrence of either property increases the difficulty of solving corresponding IP, which is termed 'ill-posed' in the sense of Hadamard (1902, 28). He considered a problem 'well-posed' if it has a unique solution that continuously depends on the data. Figure 1 provides a conceptual representation of the relationship between linearity, stochasticity and well-posedness.

Travel behaviour models are, at least partly, assumed stochastic due to imperfect knowledge of the researcher or unobserved inter- and intra-individual heterogeneity. In addition, they are typically non-linear. A good example is the family of random utility-based DCMs which incorporate the error term to capture the aforementioned stochasticity (Ben-Akiva and Lerman 1985). Consequently, inversion of travel behaviour models is often challenging, resulting in continuing efforts to develop robust estimation techniques for increasingly complex modelling structures (Ben-Akiva and Lerman 1985).

An extensive range of approaches have emerged to solve ill-posed IPs. For example, Backus and Gilbert (1967) suggested using linear combinations of data to generate unique localised averages of the model as possible solutions to ill-posed linear IPs. However, linear techniques rely on the calculation of partial derivatives of parameters, which was found inapplicable to an increasing number of freshly raised non-linear problems (Oldenburg 1984).

With increasing computational power, new methodologies that can accommodate more complex, ill-posed IPs have appeared. Particularly, Bayesian methods are effective in processing incomplete and noisy data occurring in such context (Tarantola 1987) where the complete solution of IPs is the posterior distribution of the unknown

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

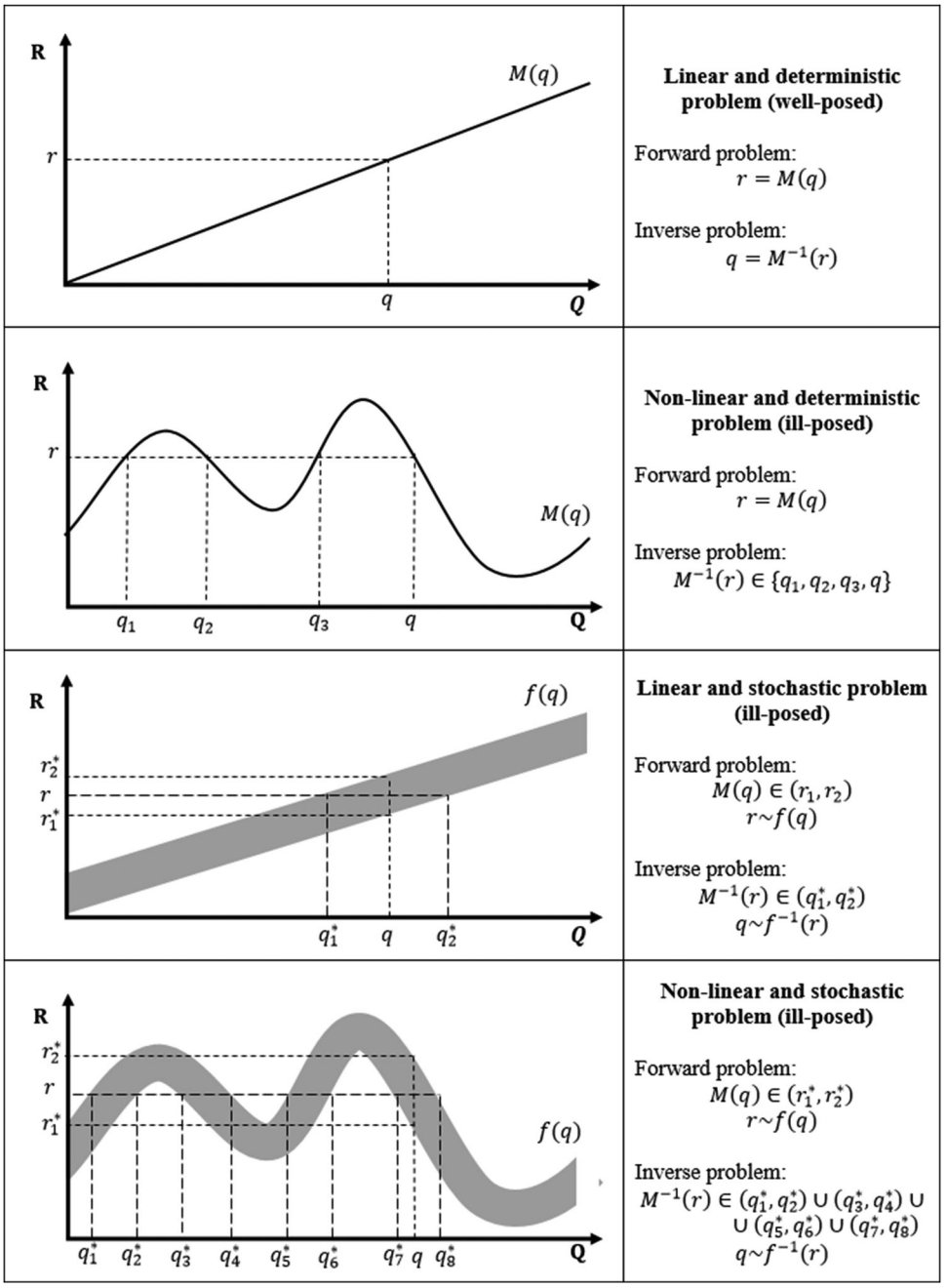


Figure 1. Conceptual representation of the relationship between linearity, stochasticity and well-posedness.

variables. Effective sampling strategies are therefore needed to find the best-fitting posterior distributions (Calvetti, Kaipio, and Somersalo 2014). The past decades have seen substantial developments in global optimisation technologies such as genetic algorithms (GAs) and genetics-based machine learning to address the problem which essentially

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

requires efficient probabilistic search procedures on large sample spaces (Goldberg and Holland 1988; Tominaga, Koga, and Okamoto 2000). Particle swarm optimisation (PSO) (Eberhart and Kennedy 1995) is a relatively recent heuristic search method based on collaborative behaviour of birds and fishes. Although both GA and PSO are population-based search processes relying on information sharing among population members with deterministic and probabilistic rules (Hassan et al. 2005), PSO is more computationally inexpensive.

Another rapidly developing set of approaches involve random and pseudo-random exploration, that is, so-called Monte Carlo (MC) approaches (Hammersley 2013). In the applied contexts, the MC process was found superior to gradient-descent and random search methods (Keilis-Borok and Yanovskaja 1967). Markov chain Monte Carlo is a powerful simulation technique for performing integration (Gilks 2005) that has revolutionised the application of Bayesian methods in IPs. For instance, Bui-Thanh and Girolami (2014) implemented it to solve heat conduction IPs.

Overall, there clearly exists a plethora of potential approaches to solving IPs, and exploration of their applicability in IDCM appears a warranted avenue of research which the current study is contributing to.

4. Methodology

Section 4 outlines the IDCM approach in detail. Specifically, Section 4.1 introduces two estimators and contexts that they fit in. Section 4.2 provides the theoretical foundations, derivation and mathematical expression of IDCM. Section 4.3 presents the hypothesis development for evaluating the proposed method while Section 4.4 discusses validation methods of the IDCM performance.

4.1. Estimation methods

In current contribution, two estimators corresponding to two statistical estimation approaches are used to solve the proposed IDCM. These two approaches are maximum likelihood estimation (MLE) and maximum a posteriori (MAP) estimation.

In statistics, a likelihood function $\mathcal{L}(\theta)$ represents the probability for the occurrence of an independent and identically distributed sample configuration $x_1, \ldots, x_n$ given the probability density $f(x; \theta)$ with known parameters θ (Harris and Stöcker 1998):

$$\mathcal{L}(\theta) = f(x_1; \theta) \ldots f(x_n; \theta). \tag{3}$$

MLE is an approach to estimating the parameters of a statistical model given observations. Specifically for a fixed set of observed data x_i ($i = 1, \ldots, n$), MLE selects the set of model parameter values θ^*_{MLE} that maximises the likelihood function (Fisher 1912):

$$\theta^*_{\mathrm{MLE}} = \arg\max_{\theta} \mathcal{L}(\theta; x_1, \ldots, x_n) = \arg\max_{\theta} \prod_{i}^{n} f(x_i; \theta). \tag{4}$$

On the other hand, the MAP estimator is an estimate of an unknown quantity given both the actual observations and any prior knowledge the researcher may have about the estimated quantities. MAP assumes that a prior distribution g over parameters θ is

known (Sorenson 1980). Hence Bayes' theorem gives the MAP estimate of model parameters:

$$\theta^*_{\text{MAP}} = \arg\max_{\theta} \prod_i^n \frac{f(x_i;\, \theta)g(\theta)}{f(x_i)} = \arg\max_{\theta} \prod_i^n f(x_i;\, \theta)g(\theta). \tag{5}$$

Note that $f(x_i)$ in the denominator is independent of θ and hence can be dropped. MLE can be seen as a special case of MAP where a uniform prior distribution is assumed for the model parameters to be estimated (Sorenson 1980). In general, the MAP estimator may be preferred to take advantage of the additional prior information.

4.2. The IDCM approach

DCMs comprise a large class of models, each making different assumptions about user behaviour, and have been applied to an enormously wide range of choice contexts, in transport and also in many other application domains (Small and Rosen 1981; Bhat 2003). Although DCMs take a wide variety of forms, they share the underlying concept: individual decision-makers choose between discrete alternatives according to some decision rules (Ben-Akiva and Lerman 1985). Microeconomic consumer theory (Mas-Colell, Whinston, and Green 1995) provides the most common framework for developing compensatory decision rules for DCMs, under which a utility function U can be derived to rank the choice alternatives in preference order (Train 2003).

Conventional DCMs are direct problems in which attributes of alternatives (A) and the respondent (X) are mapped onto the probability space of choices Y given the knowledge of preferences captured in taste parameters β:

$$P(Y|A, X, \beta). \tag{6}$$

Thus, the choice Y is effectively a random discrete variable with probability mass[1] function defined in Equation (6). A particular decision made by an individual, conditional on the factors captured by parameters A, X, β is simply a realisation of that random variable. In a typical transport modelling exercise, a researcher seeks to infer β based on a sample of observed choices. Bayes' theorem provides a means of establishing the probability of β attaining particular values given the choices and attributes for a single trip by an individual:

$$P(\beta|A, X, Y) = P(Y|A, X, \beta)\frac{P(\beta|A, X)}{P(Y|A, X)}. \tag{7}$$

Based on Equation (7), an MAP estimator is defined:

$$\beta^*_{\text{MAP}} = \arg\max_{\beta} P(Y|A, X, \beta)P(\beta|A, X). \tag{8}$$

In the absence of any knowledge about the prior distribution $P(\beta|A, X)$, which is typically the case, the MAP collapses to corresponding MLE estimator:

$$\beta^*_{\text{MLE}} = \arg\max_{\beta} P(Y|A, X, \beta). \tag{9}$$

Given that β are usually continuous, Newton's method-based gradient-descent algorithms are typically used to find β^*, thus providing the solution to this IP.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

The idea behind the IDCM enrichment follows a similar logic to that of the model calibration. Analogically to Equation (7), the likelihood that the decision-maker is characterised by a particular attribute given the observed choices, their attributes, and the preferences is defined:

$$L(X) = P(X|A,\, Y,\, \beta) = P(Y|A,\, X,\, \beta)\frac{P(X|A,\, \beta)}{P(Y|A,\, \beta)}. \tag{10}$$

Thus, the Bayesian MAP estimator of the attributes is

$$X^*_{\mathrm{MAP}} = \arg\max_X P(Y|A,\, X,\, \beta)P(X|A,\, \beta). \tag{11}$$

And the analogous MLE estimator:

$$X^*_{\mathrm{MLE}} = \arg\max_X P(Y|A,\, X,\, \beta). \tag{12}$$

In a sample of size N, each individual l can undertake M_l trips by one of the available modes of transport i. It should be noted that a joint log-likelihood should be applied across trips of each respondent to avoid different imputed values of a specific attribute for a same respondent based on different trips. Using Equation (10), the joint likelihood of observing the respondents being characterised by a specific set of attributes can be defined:

$$L(X) = \prod_{l=1}^{N}\prod_{m=1}^{M} P(X_l|A_{\mathrm{mi}},\, Y_{\mathrm{mi}},\ \beta)\,. \tag{13}$$

The conventional logarithmic transformation can be used to define the suitable log-likelihood which should be maximised to find the MAP or MLE estimates of X:

$$\max_X \sum_{l=1}^{N}\sum_{m=1}^{M}[\log P(Y_{\mathrm{mi}}|A_{\mathrm{mi}},\, X_l,\, \beta) + \log P(X_l|A_{\mathrm{mi}},\, \beta)]. \tag{14}$$

Assuming that individuals are independent of each other permits further effectively parallelisation of the maximisation procedure:

$$\sum_{l=1}^{N}\max_X \sum_{m=1}^{M}[\log P(Y_{\mathrm{mi}}|A_{\mathrm{mi}},\, X_l,\, \beta) + \log P(X_l|A_{\mathrm{mi}},\, \beta)]. \tag{15}$$

Equation (15) can thus be directly used in the enrichment procedure. The point to note, however, is that X is often of discrete nature: gender is discrete nominal, income level is often recorded on an ordinal scale. In such instances, optimisation algorithms relying on smoothness of the function and existence of derivatives may fail to converge. Therefore, the alternative approaches discussed in Section 3.2, or exhaustive search[2] ('brute force') need employing.

Although the MLE and MAP approaches give point-estimates for the imputed quantities, IDCM provides, almost as a by-product, probabilities associated with observing any value of the imputed attribute. These probabilities are essential in sample enumeration, where sample-level shares of particular attributes are obtained by summation of the respective probabilities across the sample. The benefit of obtaining such probability distribution of particular socioeconomic attributes is privacy preservation, that is, no

attribute is imputed with certainty, while retaining sample-level consistency with the observed values. This property is highly desirable in the age of growing concerns about personal data privacy implications resulting from passive data as well as the increasing ability to link and enrich datasets.

4.3. Hypothesis development

The study by Pawlak, Zolfaghari, and Polak (2015) presented MC experiments on simulated choice data. In their model, the quality of IDCM enrichment was assessed using the percentage correctly predicted (PCP). The sensitivity of the PCP was inspected in reference to the EP of a particular variable, as measured by the change in log-likelihood ($\bar{\rho}^2$[3]) due to inclusion of the imputed variable in corresponding DCM specification.

In their study, it is assumed that the choice of whether or not to visit a particular place (e.g. retail unit, car park or restaurant) depends only on the choice maker's attribute (X) with the associated coefficient β_X and a constant β_0. The authors conducted two series of experiments to respectively impute car ownership (binary variable) from choices of visiting a car park, and income level from restaurant choices. In each case, several scenarios were simulated where the EP of the imputed variable varies from almost unity to more real-life-like situations where choices become increasingly stochastic regarding the value of the attributes. The corresponding coefficients of each model can thus be estimated to further explore the influence of changing EP of the imputed attribute on performance of the IDCM approach.

The results of these simulation experiments showed that the higher PCP were produced as the EP of corresponding imputed variable increased, which provided a convenient way for the present analysis to derive the hypothesis to be tested using empirical data on choices, attributes of alternatives and decision-makers. In particular, we explore the imputation quality of the IDCM with respect to the change in the EP of the imputed variable.

4.4. Performance analysis

The proposed IDCM approach based on MLE and MAP estimates can be evaluated respectively in two ways. At the disaggregate (individual) level, the PCP is used to measure the imputation quality in terms of the proportion of individuals with correctly imputed attribute values. At an aggregate (sample) level, either the chi-square or Fisher's exact test can quantify the goodness of fit between the imputed and observed data by inspecting whether the shares of attributes in the imputed sample obtained through enumeration differ from those observed.

In order to better explore the link between the EP of imputed variables and performance of the IDCM approach, we propose to quantify the former using MI. Drawing on the information theory and the concept of entropy (Shannon 2001, 3–55), MI has stronger theoretical justification as a measure of EP than $\bar{\rho}^2$ which is a relatively informal goodness-of-fit metric. Particularly, MI quantifies the 'amount of information', typically in bits or nats, which can be inferred about one random variable through knowledge of the other random variable. The higher the MI, the more correlated are variables with respect to each other. Moreover, MI is more flexible with respect to treatment of different

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

variable types and handling non-linear relations than traditional correlation coefficients. The MI of two discrete random variables X and Y is formally defined as

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right), \tag{16}$$

where $p(x, y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are respectively the marginal probability distribution functions of X and Y.

5. Application of IDCM to imputation from the London travel demand survey

Section 5 demonstrates an application of the IDCM approach to imputing socioeconomic attributes of travellers from real-world observations of their travel mode choices. By removing respondent attribute data, the London Travel Demand Survey (LTDS) is used to mimic geotagged trace data in the imputation procedure. These attribute data, conversely, can act as a comparison to validate the imputation quality.

Particularly, Section 5.1 introduces LTDS and how it is pre-processed to fit the scope of the analysis. Section 5.2 presents the procedure of calibrating the DCM of mode choice. Section 5.3 shows how the experiments of imputation are designed and conducted. The flow chart in Figure 2 illustrates the overall process of these experiments.

5.1. LTDS data and enrichment using Google distance API

Unlike medical research where protocols for conducting studies are well and strictly defined, for example, randomised controlled trials (Shepherd et al. 2002), there has been no guide on how to define variables or what process to follow in transport modelling practices, such as DCM estimation. Hence, the proposed IDCM initially requires a suitable DCM and then explore it in an inverse fashion, that is, to find attributes of travellers in the

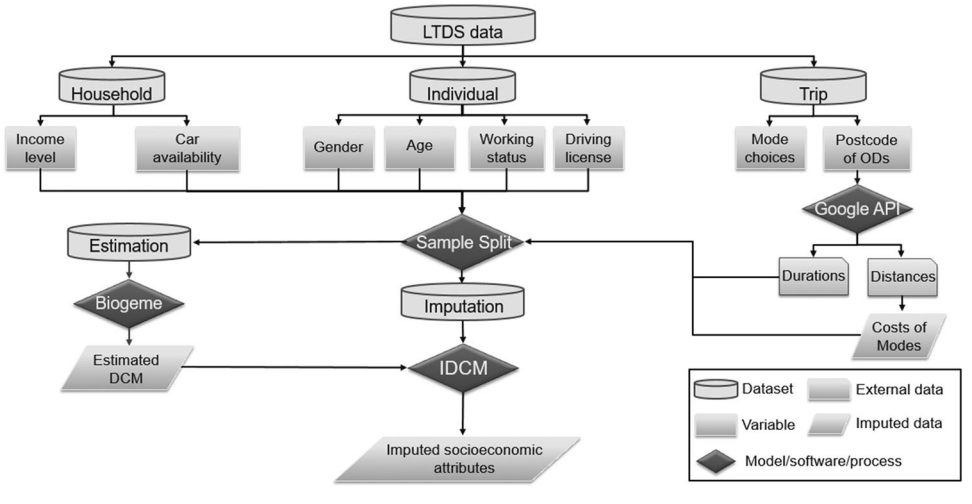


Figure 2. Procedure of the case study on LTDS data.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

sample that are most consistent with the observed choice patterns given known preference structure captured in DCM parameters. It should be noted that DCM estimation is not an essential procedure for all IDCM enrichment if a DCM is readily available from other situations or contexts. Neither a well-fitted DCM is required: for example, a model of overall low goodness of fit, but significant in terms of imputed variables is also desirable, which is related to the conclusion we draw later in Section 6.

To achieve this aim, we randomly split the selected sample into two subsamples: an estimation subsample containing 80% data records for pre-defining a suitable DCM and a 20% enrichment subsample for conducting and validating the IDCM approach. As a result, there is no problem of the endogeneity in the parameter estimation. To avoid coincidence caused by random sample split, a cross-validation in the form of a k-fold holdout method (Kohavi 1995) was conducted as a trade-off between robustness and computational demands, in which $k = 10$ as suggested by the results of Kohavi.

In this paper, LTDS data from the period 2011/2012 are employed. It is a continuous household survey which captures information on households, people, trips and vehicles covering all Greater London boroughs (Transport for London 2015).

Assuming the utility of the choice alternative i for individual a as $V(x_{ai}, s_a, \beta)$, the multinomial logit (MNL) model representing the probability for individual a to choose mode i, can be expressed as

$$P_a(i|C_a) = \frac{e^{V(x_{ai},s_a,\beta)}}{\sum_{h\in C_a} e^{V(x_{ah},s_a,\beta)}}, \tag{17}$$

where C_a is the choice set of individual a; x_{ai} is a vector of attributes that characterise mode i; s_a is a vector of attributes of individual a; β is a vector of taste parameters.

The specification will be discussed in detail in Section 5.2.

As the utility of each alternative in a DCM is a function of the attributes of the alternative and the decision-maker, only trips with complete information on ODs, mode choices, and taken by respondents with full information about their demographics (e.g. gender, age, working status and income level) are extracted from the raw data. Age and income are discretised and the average of each level is used.

Five modes of transport are considered in this study: walking, cycling, driving, bus and the underground. For each mode, travel duration and monetary cost are used as trip attributes. The LTDS is, however, a revealed preference data set and hence lacks necessary information, that is, travel durations and monetary costs, about the unchosen modes. To obtain such information, we employ Google Map Distance Matrix API (Google 2017), which can provide travel distance and time for a matrix of ODs based on the postcodes of OD and the travel mode.

With respect to the monetary cost of each mode: walking is free; cycling cost is inferred from Barclays cycle hiring price (Transport for London 2015); the prices of the underground and bus are determined by the fare guidance of Department of Transport(DfT) (2012) while the cost of driving according to the DfT rules for mileage expenses (Department for Transport 2016) is

$$C_{\text{Driving}} = 0.45 d_{\text{OD}}. \tag{18}$$

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

5.2. DCMs of mode choice

In each holdout repetition, the 80% estimation subsample is used to estimate the mode choice model. Linear-in-parameters utility V of each mode i is assumed for a respondent characterised by a set of socioeconomic attributes $X_\theta \in \Theta$:

$$V_i = \beta_{0i} + \beta_T T_i + \beta_C C_i + \sum_{\Theta} \beta_\theta X_\theta, \tag{19}$$

where T_i is the travel duration on mode i; C_i is the monetary cost of travel on mode i; X_θ is the value of the socioeconomic attribute of the respondent; β_{0i} is the alternative specific constant for mode i; β_T is the coefficient associated with travel duration; β_C is the coefficient associated with monetary cost of travel; β_θ is the coefficient associated with the socioeconomic attribute X_θ.

BIOGEME 1.8 (Bierlaire 2003) is used to estimate the MNL model and iteratively test various specifications. Table 1 provides the estimation results of all holdout samples, including the values of coefficients for attributes and corresponding significance levels.

These models are found to fit the data well, as indicated by the high adjusted rho-squared (0.491–0.505). All utility parameters are statistically significant at a 99% level, and intuitive in interpretation. In particular, all the parameters for travel duration and monetary cost are negative, which follows the fact that more costly modes, in terms of either money and or time, will be less preferred. The positive cycling-specific parameter for gender is consistent with the fact that males are more likely to ride a bicycle than females (Garrard, Rose, and Lo 2008). The working status 'employed' significantly affect both bus and the underground choices but in different ways: employed travellers prefer to take the underground while are less likely to choose buses, potentially because buses are likely to delay during peak hours. The parameter of drivers' license possession for car is positive, meaning that license holders tend to travel by car. And the fact that people earning more income tend to travel by the underground has been indicated by the positive underground-specific coefficient for income level.

Clearly these models contain a number of significant coefficients for socio-demographic variables which is crucial for the enrichment procedure. They are therefore accepted and chosen as the protocol for conducting and validating the IDCM imputation.

5.3. IDCM enrichment procedure

As is illustrated in Table 2, four series of enrichment experiment are conducted to respectively impute gender, working status 'employed', drivers' license possession, and income level, with both MLE and MAP estimators used. Specifically, gender is a nominal variable taking two values: 0 for 'female' and 1 for 'male'. Working status 'employed' is also a binary nominal variable consisting of 'employed' and 'not employed' (unemployed, retired and student). Drivers' license possession is another binary nominal variable while income is an ordinal variable, which is discretised into three levels, guaranteeing finite parameter space and thus globally optimum solution.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

Table 1. MNL model estimation results.

	Holdout sample number									
Attribute coefficients	1	2	3	4	5	6	7	8	9	10
Travel duration	−0.172**	−0.175**	−0.178**	−0.174**	−0.178**	−0.181**	−0.179**	−0.175**	−0.178**	−0.177**
Travel cost	−0.372**	−0.380**	−0.386**	−0.379**	−0.247**	−0.259**	−0.230**	−0.378**	−0.243**	−0.262**
Gender (male): cycling	1.370**	1.130**	1.220**	1.190**	1.090**	1.260**	1.160**	1.210**	1.240**	1.150**
Employed: bus	−0.661**	−0.662**	−0.717**	−0.736**	−0.706**	−0.677**	−0.825**	−0.739**	−0.692**	−0.501**
Employed: underground	0.294**	0.364**	0.447**	0.268**	0.252**	0.322**	0.296**	0.349**	0.404**	0.476**
License possession: driving	0.766**	0.800**	0.788**	0.823**	0.807**	0.853**	0.713**	0.811**	0.864**	0.908**
Income level: underground	0.156**	0.155**	0.150**	0.139**	0.154**	0.149**	0.149**	0.148**	0.132**	0.141**
ASC: cycling	−5.700**	−5.580**	−5.670**	−5.590**	−5.520**	−5.680**	−5.580**	−5.710**	−5.690**	−5.620**
ASC: driving	−3.000**	−3.070**	−3.080**	−3.050**	−3.230**	−3.310**	−3.220**	−3.100**	−3.190**	−3.380**
ASC: bus	−1.130**	−1.160**	−1.110**	−1.100**	−1.290**	−1.290**	−1.280**	−1.110**	−1.310**	−0.914**
ASC: underground	−2.820**	−2.870**	−2.940**	−2.690**	−2.950**	−2.990**	−3.010**	−2.850**	−2.980**	−2.510**

Significance: **.01 *.05.

Table 2. Experiment design.

No.	Attribute to be imputed	Type	Estimator	Repetition
1	Gender	Nominal	MLE and MAP	10
2	Working status 'employed'	Nominal	MLE and MAP	10
3	Drivers' license possession	Nominal	MLE and MAP	10
4	Income level	Ordinal	MLE and MAP	10

In the MLE approach as defined in Equation (12), it is assumed that no prior knowledge on the parameters is available, that is, equal probabilities occur in all categories of either socioeconomic attributes or mode choices. Hence, the joint log-likelihood that should be maximised to find the MLE estimates is defined as the summation of the logarithms of the probabilities of all trips undertaken by the individual. For a sample consisting of N respondents independent of each other, it is equivalent to Equation (20):

$$X^*_{\mathrm{MLE}} = \sum_{n}^{N} \max_{X_n} \sum_{m=1}^{M} \log P(Y_{m_n i} | A_{m_n i}, X_n, \beta). \tag{20}$$

In terms of the MAP approach, the MAP estimator of the attribute to be imputed has been defined on the basis of Equations (8) and (12), leading to the expression in Equation (21):

$$X^*_{\mathrm{MAP}} = \sum_{n}^{N} \max_{X_n} \sum_{m=1}^{M} [\log P(Y_{m_n i} | A_{m_n i}, X_n, \beta) + logP(X_n | A_{m_n i}, \beta)]. \tag{21}$$

The prior $P(X_n|A_{m_n i}, \beta)$ can be derived from corresponding estimation subsample. The process is achieved using exhaustive search over the given parameter space which is computationally manageable due to the independence between respondents, and finite thanks to the discrete nature of imputed variables. This furthermore guarantees the solution to be the global optimum.

6. Results

Section 6 discusses findings from the four series of imputation experiments over 10 repetitions performed on the 20% imputation subsamples as outlined in Section 5.3. PCPs of each imputed variable of all holdout samples based on MLE and MAP estimates are presented in Figure 3.

We are also interested in the shares of individuals characterised by specific attributes in imputed and observed samples by conducting sample enumeration, which is a standard

Table 3. Chi-squared test results (*P*-values) based on MLE and MAP estimates.

No.	Est.	1	2	3	4	5	6	7	8	9	10
1	MLE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	MAP	0.608	0.693	0.665	0.098	0.875	0.937	0.478	0.098	0.753	0.906
2	MLE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.009	0.000
	MAP	0.238	0.537	0.507	0.307	0.168	0.077	0.615	0.854	0.614	0.178
3	MLE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	MAP	0.959	0.254	0.946	0.885	0.665	0.615	0.129	0.324	0.780	0.680
4	MLE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	MAP	0.479	0.834	0.617	0.867	0.088	0.259	0.560	0.706	0.586	0.801

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

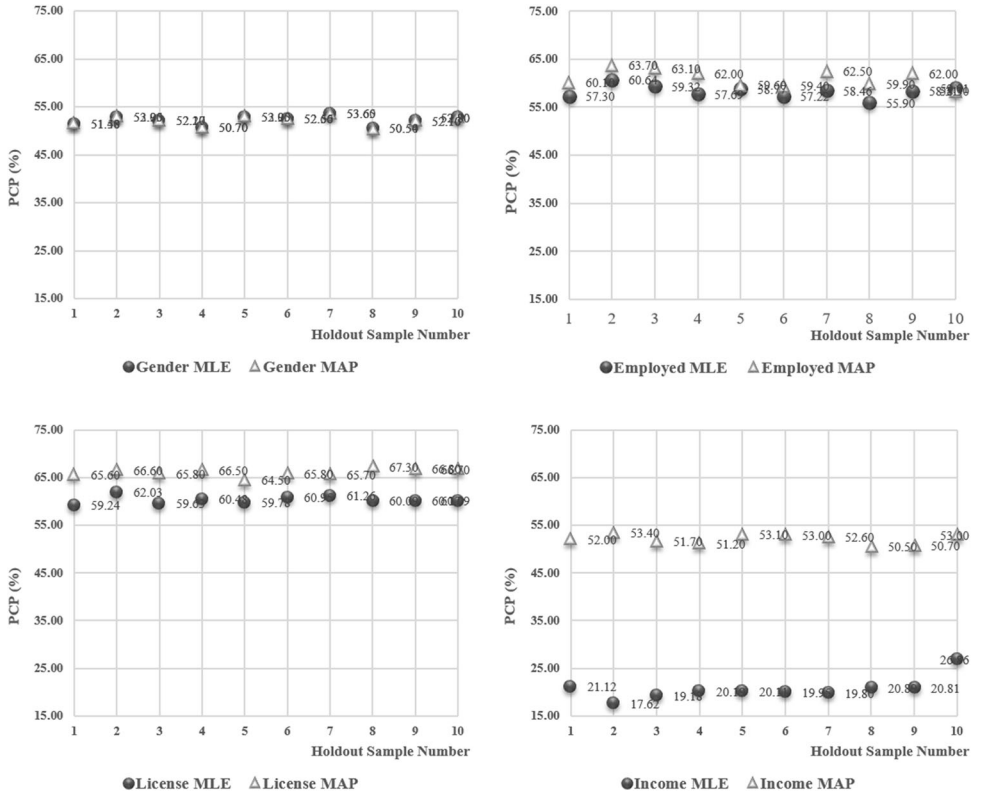


Figure 3. Comparison of PCPs using MLE and MAP estimators.

technique used frequently in investigating performance of DCMs (de Dios Ortúzar and Willumsen 1994). Due to large sample sizes, chi-squared test is used to validate the approach on aggregate level. And Table 3 shows the *p*-values of chi-squared test based on the two estimators.

It can be seen from Figure 3 and Table 3 that the PCP and the chi-squared test statistics of each imputed variable using same estimators are stable across randomly selected subsamples, which provides a reasonable premise for the following analyses.

Figure 3 shows that the MAP estimator generally performs no worse than its MLE counterpart on the individual level as it improves PCPs of the MLE-based imputation. Particularly in the first graph, performance of the MAP estimator almost equals to that of the MLE estimator. This is probably due to the equal distribution across genders, which have led to similar information of the uniform prior distribution of the MLE estimator. It is also noticed that the IDCM model is generally better at predicting nominal attributes than ordinal attributes. This is reasonably understandable as ordinal attributes (e.g. income) consist of more categories and therefore more input information is required for more accurate prediction. Moreover, ordinal attributes are imputed using the average value of each category for simplification, which introduce extra noise that reduces the prediction accuracy.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

The results of chi-squared test in Table 3 demonstrate that the MAP also significantly improves the imputation quality on the overall sample level. This is indicated by all p-values over .05, showing that imputed samples do not differ from corresponding observed sample at a 95% confidence level. This is intuitive in the sense that MAP estimators usually involve more relevant information than MLE estimators.

The findings above are also related to the hypothesis developed in Section 4.3. Rather than $\overline{\rho}^2$, we calculate MIs (in bits) between travel modes and each attribute of all folds so as to explore the relationship between the EP and imputation quality of IDCM. Results are presented in Figure 4.

As is seen in Figure 4, the nominal variables show a pattern that higher PCP is produced by higher EP. This is in line with the results of the MC experiments by Pawlak, Zolfaghari, and Polak (2015). In particular, the relationship shows diminishing

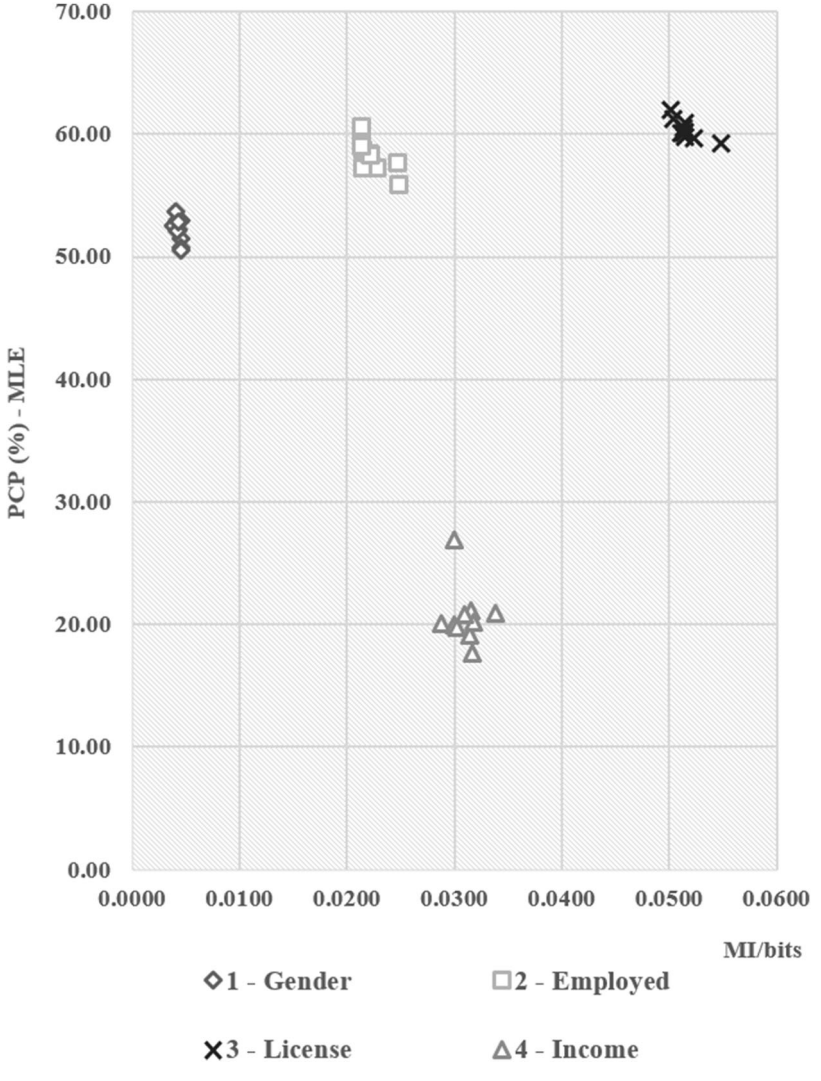


Figure 4. Correlation pattern between MI and PCPs.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

marginal improvement in PCP as the MI grows, which seems to be logarithmic or square root.

The ordinal variable, however, does not follow the trend of nominal variables. There are two potential explanations for this phenomenon. First, the income attribute can have three or more different values, as opposed to two in the nominal variables above, which increase the complexity of imputing the exact value of the variable from same amount of input information, that is, single discrete choice. Moreover, income is discretised and grouped into three levels with the average value of each level used in the IDCM imputation, hence reducing the MI with the choice. The exact reason will be explored in the future. It should be noted that we do not discuss the correlation between MI and PCPs using the MAP. This is because the amount of information contained in MAP-based PCPs has to some extent been 'polluted' by the prior information.

7. Conclusions

This paper formalises and extends a data enrichment approach which uses IDCM to infer socioeconomic characteristics of travel decision-makers from observations of discrete travel choices. Performance of the IDCM applied to a mode choice model based on real-world data set is explored. The empirical results are compared to that from the earlier MC experiment which employs the same inversion mechanisms.

It is observed that performance of the IDCM is highly sensitive to the EP of the imputed variable, measured by MI between the variable and discrete choices, in corresponding DCM. Specifically, performance of imputing the same type of variables using the proposed method is improved as the EP increases. Moreover, the nature of the imputed variable also plays a significant role. Particularly, attributes with numerical meanings or having more than two potential values, such as ordinal variables, can be more difficult to impute than nominal variables and two-category categorical variables. The exact reason for this will be investigated in the future.

This study can be viewed as an important step towards bridging the gap between travel behaviour analysis and data collected by ICT devices. The substantial benefit of using the increasingly available geotagged data to substitute traditional surveys while preserving individual privacy provide sufficient rationale to continue developing such enrichment approaches. For further investigation, MC experiment will be expanded to explore the role of DCM structures and EP of variables in determining imputation quality. It is hoped that this avenue will lead to new theoretical and empirical insights enabling more effective and robust enrichment procedures.

Notes

1. For the clarity of discussion, we follow the notation for discrete variables. A corresponding argument and derivation can be made, however, with respect to continuous variables, and using probability density functions.
2. In computer science, brute-force search or exhaustive search is a general problem-solving technique that systematically enumerates all possible candidates for the solution and checking whether each candidate satisfies the statement of the problem.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

3. $\overline{\rho}^2$ can be calculated according to $\Delta\overline{\rho}^2 = 1 - {LL_Full}/{LL_without_X}$, where *LL_Full* is the log-likelihood of the full model while $LL_without_X$ is the log-likelihood of the model excluding attribute *X*.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Andridge, Rebecca R., and Roderick J. A. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review* 78 (1): 40–64.

Andrienko, Gennady, Natalia Andrienko, Christophe Hurter, Salvatore Rinzivillo, and Stefan Wrobel. 2011. *From Movement Tracks Through Events to Places: Extracting and Characterizing Significant Places from Mobility Data*. Providence, RI: IEEE.

Backus, G. E., and J. F. Gilbert. 1967. "Numerical Applications of a Formalism for Geophysical Inverse Problems." *Geophysical Journal International* 13 (1–3): 247–276.

Bal, Guillaume. 2012. "Introduction to Inverse Problems." Lecture Notes-Department of Applied Physics and Applied Mathematics, Columbia University, New York.

Ben-Akiva, Moshe E. and Steven R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. Vol. 9. Cambridge, MA: MIT Press.

Bennewitz, Maren, Wolfram Burgard, Grzegorz Cielniak, and Sebastian Thrun. 2005. "Learning Motion Patterns of People for Compliant Robot Motion." *The International Journal of Robotics Research* 24 (1): 31–48.

Bhat, Chandra R. 2003. "Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences." *Transportation Research Part B: Methodological* 37 (9): 837–855.

Bierlaire, Michel. 2003. "*BIOGEME: A Free Package for the Estimation of Discrete Choice Models*." Paper presented at the Swiss Transport Research Conference, Ascona, Switzerland, March 19–21.

Bohte, Wendy, and Kees Maat. 2009. "Deriving and Validating Trip Purposes and Travel Modes for Multi-Day GPS-Based Travel Surveys: A Large-Scale Application in the Netherlands." *Transportation Research Part C: Emerging Technologies* 17 (3): 285–297.

Brunauer, Richard, Michael Hufnagl, Karl Rehrl, and Andreas Wagner. 2013. *Motion Pattern Analysis Enabling Accurate Travel Mode Detection from Gps Data Only*. The Hague: IEEE.

Bui-Thanh, Tan, and Mark Girolami. 2014. "Solving Large-Scale PDE-Constrained Bayesian Inverse Problems with Riemann Manifold Hamiltonian Monte Carlo." *Inverse Problems* 30 (11): 114014.

Calvetti, Daniela, Jari P. Kaipio, and Erkki Somersalo. 2014. "Inverse Problems in the Bayesian Framework." *Inverse Problems* 30 (11): 110301.

Chen, Cynthia, Hongmian Gong, Catherine Lawson, and Evan Bialostozky. 2010. "Evaluating the Feasibility of a Passive Travel Survey Collection in a Complex Urban Environment: Lessons Learned from the New York City Case Study." *Transportation Research Part A: Policy and Practice* 44 (10): 830–840.

Chung, Eui-Hwan, and Amer Shalaby. 2005. "A Trip Reconstruction Tool for GPS-Based Personal Travel Surveys." *Transportation Planning and Technology* 28 (5): 381–401.

de Dios Ortúzar, Juan, and Luis G. Willumsen. 1994. *Modelling Transport*. Hoboken, NJ: Wiley.

de Montjoye, Yves-Alexandre, Jordi Quoidbach, Florent Robic, and Alex Pentland. 2013. *Predicting Personality Using Novel Mobile Phone-Based Metrics*. Heidelberg: Springer.

Department for Transport. "Staff Guide to Fares and Tickets from 2 January 2012." https://www.whatdotheyknow.com/request/98541/response/246731/attach/4/Staff%20guide%20to%20fares%20and%20ticketing%20January%202012.pdf.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

Department for Transport. "WebTAG: Transport Analysis Guidance Data Book." https://www.gov.uk/government/publications/webtag-tag-data-book-july-2016.
Dodge, Somayeh, Robert Weibel, and Ehsan Forootan. 2009. "Revealing the Physics of Movement: Comparing the Similarity of Movement Characteristics of Different Types of Moving Objects." *Computers, Environment and Urban Systems* 33 (6): 419–434.
D'Orazio, Marcello, Marco Di Zio, and Mauro Scanu. 2006. *Statistical Matching: Theory and Practice*. New York: Wiley.
Eberhart, Russ C., and James Kennedy. 1995. *A New Optimizer using Particle Swarm Theory*. Nagoya: IEEE.
Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. *A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*.
Fisher, R. A. 1912. "The Maximum–Likelihood–Method." *Messenger in Mathematics* 41: 155–160.
Garrard, Jan, Geoffrey Rose, and Sing Kai Lo. 2008. "Promoting Transportation Cycling for Women: The Role of Bicycle Infrastructure." *Preventive Medicine* 46 (1): 55–59.
Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. "Using Deep Learning and Google Street View to Estimate the Demographic Makeup of the US." arXiv Preprint arXiv: 1702. 06683.
Giannotti, Fosca, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. "Trajectory Pattern Mining." ACM.
Gilks, Walter R. 2005. *Markov Chain Monte Carlo*. Wiley Online Library.
Goldberg, David E., and John H. Holland. 1988. "Genetic Algorithms and Machine Learning." *Machine Learning* 3 (2): 95–99.
Gong, Lei, Takayuki Morikawa, Toshiyuki Yamamoto, and Hitomi Sato. 2014. "Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies." *Procedia-Social and Behavioral Sciences* 138: 557–565.
Google. 2017. "GoogleMaps Distance Matrix API." https://developers.google.com/maps/documentation/distance-matrix/.
Hadamard, Jacques. 1902. "Sur Les Problèmes Aux Dérivées Partielles Et Leur Signification Physique." *Princeton University Bulletin* 13 (49–52): 28.
Hammersley, John. 2013. *Monte Carlo Methods*. New York: Springer Science & Business Media.
Harris, John W., and Horst Stöcker. 1998. *Handbook of Mathematics and Computational Science*. New York: Springer Science & Business Media.
Hassan, Rania, Babak Cohanim, Olivier De Weck, and Gerhard Venter. 2005. "A Comparison of Particle Swarm Optimization and the Genetic Algorithm".
Hill, Linda L. 2009. *Georeferencing: The Geographic Associations of Information*. Cambridge, MA: MIT Press.
Keilis-Borok, V. I., and T. B. Yanovskaja. 1967. "Inverse Problems of Seismology (Structural Review)." *Geophysical Journal International* 13 (1–3): 223–234.
Kohavi, Ron. 1995. "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection." Stanford, CA.
Lou, Yin, Chengyang Zhang, Yu Zheng, Xing Xie, Wei Wang, and Yan Huang. 2009. "Map-Matching for Low-Sampling-Rate GPS Trajectories". ACM.
Mas-Colell, Andreu, Michael Dennis Whinston, and Jerry R. Green. 1995. *Microeconomic Theory*. Vol. 1. New York: Oxford University Press.
Montoliu, Raul, Jan Blom, and Daniel Gatica-Perez. 2013. "Discovering Places of Interest in Everyday Life from Smartphone Data." *Multimedia Tools and Applications* 62 (1): 179–207.
Oldenburg, Doug W. 1984. "An Introduction to Linear Inverse Theory." *IEEE Transactions on Geoscience and Remote Sensing* GE-22 (6): 665–674.
Pawlak, J., A. Zolfaghari, and J. W. Polak. 2015. "Imputing Socioeconomic Attributes for Movement Data by Analysing Patterns of Visited Places and Google Places Database: Bridging between Big Data and Behavioural Analysis." Austin, TX, 11–13 July.
Rubin, Donald B. 2004. *Multiple Imputation for Nonresponse in Surveys*. Vol. 81. New York: Wiley.
Schönfelder, Stefan, and Nicolas Antille. 2002. *Exploring the Potentials of Automatically Collected GPS Data for Travel Behaviour Analysis: A Swedish Data Source*. Zürich: ETH, Eidgenössische

Downloaded by [Imperial College London Library] at 07:33 05 December 2017

Technische Hochschule Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau IVT.
Shannon, Claude E. 2001. "A Mathematical Theory of Communication." *ACM SIGMOBILE Mobile Computing and Communications Review* 5 (1): 3–55.
Shepherd, James, Gerard J. Blauw, Michael B. Murphy, Edward LEM Bollen, Brendan M. Buckley, Stuart M. Cobbe, Ian Ford, et al. 2002. "Pravastatin in Elderly Individuals at Risk of Vascular Disease (PROSPER): A Randomised Controlled Trial." *The Lancet* 360 (9346): 1623–1630.
Sheth, Amit. 1997. "Panel: Data Semantics: What, Where and How?" In *Database Applications Semantics*, 601–610. Copenhagen: Springer.
Small, Kenneth A., and Harvey S. Rosen. 1981. "Applied Welfare Economics with Discrete Choice Models." *Econometrica: Journal of the Econometric Society* 49 (1): 105–130.
Smith, Gary W. 1990. *Modeling Security-Relevant Data Semantics*. Oakland, CA: IEEE.
Sorenson, Harold Wayne. 1980. *Parameter Estimation: Principles and Problems*. Vol. 9. M. Dekker.
Stopher, Peter, Camden FitzGerald, and Jun Zhang. 2008. "Search for a Global Positioning System Device to Measure Person Travel." *Transportation Research Part C: Emerging Technologies* 16 (3): 350–369.
Tarantola, Albert. 1987. "Inverse Problem Theory: Methods for Data Fitting and Parameter Estimation".
Tominaga, Daisuke, Nobuto Koga, and Masahiro Okamoto. 2000. *Efficient Numerical Optimization Algorithm Based on Genetic Algorithm for Inverse Problem*. Las Vegas, ND: Morgan Kaufmann.
Train, Kenneth. 2003. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.
Train, Kenneth, and Daniel McFadden. 1978. "The Goods/Leisure Tradeoff and Disaggregate Work Trip Mode Choice Models." *Transportation Research* 12 (5): 349–353.
Transport for London. 2015. "Cycle Hire Charges to be Simplified in the New Year." https://tfl.gov.uk/info-for/media/press-releases/2014/december/cycle-hire-charges-to-be-simplified-in-the-new-ye
Wolf, Jean, Stacey Bricka, T. Ashby, and C. Gorugantua. 2004. "Advances in the Application of GPS to Household Travel Surveys".
Wolf, J., R. Guensler, and W. Bachman. 2001. *Elimination of the Travel Diary: An Experiment to Derive Trip Purpose from GPS Travel Data*. Washington, DC: Transportation Research Board.
Wood, J. 1985. "What's in a Link?" Readings in Knowledge Representation. Morgan Kaufmann.

Downloaded by [Imperial College London Library] at 07:33 05 December 2017