

INVERSE DISCRETE CHOICE MODELLING: A FRAMEWORK FOR SOCIO-DEMOGRAPHIC ENRICHMENT OF BIG DATA

Yuanying Zhao

A thesis submitted as fulfilment of the requirements for the degree of
Doctor of Philosophy of Imperial College London

Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London

April 2021

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr Aruna Sivakumar, Dr Jacek Pawlak and Professor John W. Polak, who have provided me with constantly prompt feedbacks in accomplishing this PhD study as well as for their consistent encouragement, patience and support over this four-year period.

I owe many thanks to the researchers and fellow PhD students at Centre for Transport Studies and Urban Systems Labs for sharing their resources and ideas with me. Without their unreserved help, my PhD study would not proceed smoothly.

Furthermore, I am immensely indebted to my family and friends for being consistently and unconditionally supportive throughout my life. I treasure every little joyful moment when I feel immersed in their love and cares, making the endeavour not worth mentioning.

I am also especially grateful to my examiners, Prof Michel Bierlaire and Dr Christian Onof, for their comments and advice on my thesis.

My deepest gratefulness and respect go to Professor John W. Polak who have offered me such a great opportunity to undertake this research and to pursue the doctoral degree. Although no longer with us, he has provided invaluable insights and guidance to my work. His wisdom, enthusiasm and dedication to research will continue to inspire me in the rest of my life.

To my parents

Declaration

The research presented in this thesis is on my own, except where the work of others is referenced.

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are Licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

This thesis proposes an Inverse Discrete Choice Modelling (IDCM) framework for the enrichment of socio-demographic information for anonymous big datasets with predicable and interpretable enrichment performance. By addressing the research gaps in existing socio-demographic enrichment methods which fail to account for the underpinning microeconomic behaviour theory, the IDCM framework is applicable to and transferable between any enrichment contexts where behaviours of respondents can be obtained, and the socio-demographic information of the data is necessary yet unavailable.

Specifically, the IDCM approach postulates that a discrete choice model (DCM) which characterise the dependence between the socio-demographic attributes of people and their behaviour patterns can be inverted to estimate the explanatory socio-demographic information. Correspondingly, the IDCM performance theory expresses the enrichment performance of the IDCM approach as a function of the assumed recalibrated constant. Moreover, the IDCM performance theory establishes the link between the estimated enrichment performance and a developed metric of enrichment efficiency using the assumed recalibrated constant as a pivot, so as to characterise the variation in its enrichment performance due to changes in the data condition of the enriched sample.

Two empirical applications are conducted to validate the ability of the IDCM performance theory to forecast and interpret the IDCM enrichment performance in light of various data conditions. Whereas the IDCM approach performs comparably to logistic regression and support vector machines, the proven ability of the IDCM performance theory in the comparative analysis against the two supervised machine learning methods acknowledges the transferability of the IDCM framework.

Table of Contents

Acknowledgements.....	2
Declaration.....	4
Abstract.....	5
List of Figures	10
List of Tables	12
1 INTRODUCTION	15
1.1 Background	15
1.2 Motivation	17
1.3 Aims and objectives	19
1.4 Definition of common terms used.....	20
1.5 Thesis structure	21
2 LITERATURE REVIEW	24
2.1 Overview of existing semantic enrichment studies and research gaps	24
2.1.1 Techniques to processing raw transport big data into meaningful trips	25
2.1.2 Origin-destination (OD) matrices	27
2.1.3 Types of locations	29
2.1.4 Trip purposes and activities	31
2.1.5 Modes of transport	35
2.1.6 Socio-demographic attributes	37
2.1.7 Summary of previous semantic enrichment studies and research gap in socio-demographic enrichment.....	42
2.2 Transport big data sources for semantic enrichment.....	46
2.2.1 Global Navigation Satellite System (GNSS) data	47
2.2.2 Smartcard data	48
2.2.3 Mobile network data	49
2.2.4 Other sources for semantic enrichment.....	49
2.3 Machine learning.....	50
2.3.1 Typology of machine learning approaches	51

2.3.2	Representative supervised learning algorithms for enrichment of big data.....	55
2.3.2.1	Logistic regression.....	55
2.3.2.2	Support vector machines (SVMs).....	57
2.3.3	Machine learning versus statistical modelling.....	59
2.4	Summary	61
3	CONCEPTUAL FRAMEWORK OF IDCM METHODOLOGY	63
3.1	Discrete Choice Models (DCMs).....	63
3.1.1	Econometric formulation	64
3.1.2	Model components and structures	66
3.1.3	Parameter calibration	68
3.1.4	Goodness-of-fit (GOF) tests	71
3.2	Conceptual IDCM model based on microeconomic behavioural theories.....	73
3.3	Inverse problems	75
3.3.1	Definition and properties	75
3.3.2	Solution techniques for ill-posed inverse problems.....	78
3.4	IDCM analogy in information theory.....	82
3.5	Summary	87
4	INVERSE DISCRETE CHOICE MODELLING METHODOLOGY	89
4.1	IDCM approach.....	89
4.1.1	IDCM Econometric formulation.....	90
4.1.2	Requirements and procedures for IDCM implementation.....	95
4.2	IDCM Enrichment Performance	97
4.2.1	Enrichment performance assessment indicators	98
4.2.1.1	Aggregate (sample) level: sample enumeration and market share comparison.....	98
4.2.1.2	Disaggregate (individual) level: percentage correctly predicted.....	101
4.2.2	Explanatory power of variables	102
4.2.2.1	Adjusted McFadden's LRI	102
4.2.2.2	Mutual information, entropy, and channel efficiency	104
4.2.3	IDCM enrichment performance versus explanatory power of choice in enriched attribute	108

4.2.3.1	Enrichment context settings.....	109
4.2.3.2	Enrichment sample distribution versus assumed recalibrated constant	111
4.2.3.3	Enrichment efficiency versus assumed recalibrated constant	113
4.2.3.4	PCPs versus assumed recalibrated constant	122
4.2.3.5	PCPs versus enrichment efficiency	135
4.2.3.6	Comparison between PCPs using MLE and MAP estimators.....	139
4.3	Summary	140
5	EMPIRICAL APPLICATION: IDCM PERFORMANCE COMPARISON AND THEORETICAL VALIDATION	143
5.1	Data and variables	144
5.2	Experimental design.....	146
5.2.1	Modification of variables	146
5.2.2	Enrichment experiments	147
5.2.3	Setup of the approaches	149
5.3	Validation of IDCM performance theory.....	153
5.4	Comparative analysis of the IDCM approach against supervised machine learning methods	159
5.4.1	Enrichment performance.....	159
5.4.2	Transferability.....	172
5.5	Summary	176
6	EMPIRICAL APPLICATION: ENRICHMENT OF MOBILE NETWORK DATA FOR LONDON USING LONDON TRAVEL DEMAND SURVEY	179
6.1	Data	180
6.1.1	Geographies and spatial (geographical) aggregation.....	181
6.1.2	Data sources	183
6.1.3	Selection and modification of variables.....	185
6.2	Experimental design.....	188
6.3	Results and comparative analysis.....	197
6.3.1	IDCM enrichment performance versus data conditions	198
6.3.1.1	Quality of prior	198

6.3.1.2	Level of spatial detail	208
6.3.2	Capability of IDCM performance theory to estimate IDCM enrichment performance given data conditions.....	217
6.3.3	Comparison of IDCM framework against other socio-demographic enrichment methods	230
6.4	Summary	244
7	CHALLENGES TO IDCM APPLICATION IN TRANSPORT CONTEXTS.....	248
7.1	Issue description.....	249
7.2	Theoretical discussion.....	251
7.3	Monte Carlo (MC) enrichment experiments	254
7.3.1	Experimental design.....	255
7.3.2	Enrichment results and analysis.....	257
7.4	Summary	259
8	CONCLUSIONS	261
8.1	Summary	262
8.2	Key contributions	264
8.3	Implications.....	265
8.3.1	Privacy considerations	266
8.3.2	Transferability and generalisability	266
8.3.3	Operational efficiency.....	267
8.3.4	Customer experience.....	268
8.4	Limitations and ongoing works.....	269
	References.....	273
	Appendix 1: List of publications.....	291
	Appendix 2: Acronyms	293
	Appendix 3: Notations	295

List of Figures

Figure 1.1 Conceptual Comparison between Big and Small Datasets.....	16
Figure 1.2 Conceptual Comparison between Data Missingness and Semantic Enrichment ...	21
Figure 2.1 Procedure to Process Scattered GPS Data into Trip Data (Source: Wolf et al. (2014)).....	26
Figure 2.2 Classification of Methods for Trip Purpose Inference (Source: Gong et al. (2014))	32
Figure 2.3 Classification of Mode Detection Methods (Source: Gong et al. (2014)).....	35
Figure 2.4 Conceptual Classification of Machine Learning Algorithms	55
Figure 3.1 Conceptual Formulation of Forward and Inverse Problems.....	76
Figure 3.2 Conceptual Relationship between Linearity, Stochasticity and Ill-posedness	77
Figure 3.3 Components of Information Communication System	83
Figure 3.4 Example Choice-making Quasi-communication Channel	85
Figure 4.1 Selection of Algorithms to Solution of IDCM Approach	93
Figure 4.2 Selection of Required Variables for Socio-demographic Enrichment using IDCM Approach.....	95
Figure 4.3 Procedure of Implementing Socio-demographic Enrichment using IDCM Approach.....	96
Figure 4.4 Entropy of Binary Variable vs Probability of One Possible Outcome Value	106
Figure 4.5 Variation Pattern of η vs. β_0 when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$).....	115
Figure 4.6 Variation Pattern of $H(X)$ vs. β_0 when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$).....	118
Figure 4.7 Variation Pattern of $I(X, Y)$ vs. β_0 when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)	120
Figure 4.8 Comparison between Patterns of $H(X)$ and $I(X, Y)$ vs. β_0 when $p_y = 0.3$ and $\beta_1 = 1.792$	121
Figure 4.9 Variation Pattern of MLE-PCP vs. β_0 when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)	130
Figure 4.10 Variation Pattern of MAP-PCP vs. β_0 when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)	135
Figure 4.11 Variation Pattern of MLE-PCP vs. η when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)	136

Figure 4.12 Variation Pattern of MAP-PCP vs. η when $py = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)	138
Figure 4.13 Criteria for Selection between Estimation Methods for IDCM Enrichment.....	140
Figure 5.1 Market Shares of Gender and Inclusion in Specific Income Level Groups	145
Figure 5.2 Behaviour Features and Enriched Attributes for Enrichment using IDCM, LGR and SVM	148
Figure 5.3 Plots of Actual PCP vs. Expected PCP using MAP and MLE Estimators.....	158
Figure 5.4 Plots of MAP-PCP vs. LGR- and SVM-PCPs	164
Figure 5.5 Variation Pattern of MLE-PCP vs. η when $py = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)..	168
Figure 5.6 Plots of MLE-PCP vs. LGR- and SVM-PCPs	171
Figure 6.1 Inclusions between OAs, LSOAs and MSOAs (Source: OCSI (2019))	182
Figure 6.3 Market Shares of Gender and Age Groups for Early AM Observations.....	187
Figure 6.4 Example Comparison between $Pest(X)$ and $Perc(X)$ for Observations during Early AM	196
Figure 6.5 MAP-PCP vs MLE-PCP in ‘Similar Prior’ and ‘Dissimilar Prior’ Scenarios	207
Figure 6.6 Performance-driven Criteria for Selection between MLE and MAP Estimators of IDCM Approach given Prior Availability, Informativeness and Correctness	208
Figure 6.7 Actual MLE-PCP vs Expected MLE-PCP in ‘Similar’ & ‘Dissimilar Prior’ Scenarios	223
Figure 6.8 Actual MAP-PCP vs Expected MAP-PCPs for Gender Enrichment	228
Figure 6.9 Actual MAP-PCP vs Expected MAP-PCPs for Revised Gender Enrichment	229
Figure 6.10 MLE-PCP vs LGR- and SVM-PCPs in ‘Similar Prior’ Scenario	233
Figure 6.11 MLE-PCP vs LGR- and SVM-PCPs in ‘Dissimilar Prior’ Scenario	237
Figure 6.12 MAP-PCP vs LGR- and SVM-PCPs in ‘Similar Prior’ Scenario	240
Figure 6.13 MAP-PCP vs LGR- and SVM-PCPs in ‘Similar Prior’ Scenario	244

List of Tables

Table 2.1 Summary of Existing Studies for Enrichment of Behaviour Patterns	42
Table 2.2 Summary of Existing Attempts for Socio-demographic Enrichment.....	44
Table 2.3 Comparable Concepts in Machine Learning and Statistical Modelling	59
Table 4.1 Actual Joint Probability Distribution of Enrichment Sample <i>DErc</i>	124
Table 4.2 Joint Probability Distribution of <i>DErc</i> Enriched with <i>X</i> using the IDCM-MLE Approach.....	125
Table 5.1 Data Conditions and Results of Comparative Analysis for Actual and Expected MAP-PCPs.....	153
Table 5.2 Data Conditions and Results of Comparative Analysis for Actual and Expected MLE-PCPs	156
Table 5.3 Results of Comparative Analysis for MAP-PCP and LGR-PCP.....	159
Table 5.4 Results of Comparative Analysis for MAP-PCP and SVM-PCP	162
Table 5.5 Results of Comparative Analysis for MLE-PCP and LGR-PCP	165
Table 5.6 Results of Comparative Analysis for MLE-PCP and SVM-PCP	168
Table 6.1 Number Limits for People and Households in each OA, LSOA and MSOA.....	182
Table 6.2 Examples for Inclusion of MSOAs in OPCs given Spatial Aggregation	183
Table 6.3 Format of MSOA-Level MND on Trip Information	184
Table 6.4 Format of MSOA-Level MND for Dwelling Information.....	184
Table 6.5 Format of LTDS with Relevant Variables	185
Table 6.6 Labels and Names for Periods in 24 Hours of a Day.....	186
Table 6.7 One-proportion <i>Z</i> -test results for Comparison of $P(X)$ in Raw LTDS and MND	189
Table 6.8 Abbreviations for Factor Labelling of Experiment Design	197
Table 6.9 Number of Experiments Segmented by Attribute, Choice and Spatial Detail Level	197
Table 6.10 Test Results for Comparison of MLE-PCPs in ‘Similar’ & ‘Dissimilar Prior’ Scenarios	198
Table 6.11 Test Results for Comparison of MAP-PCPs in ‘Similar’ & ‘Dissimilar Prior’ Scenarios	202

Table 6.12 Test Results for Comparison of MLE- and MAP-PCPs in ‘Similar Prior’ Scenarios	203
Table 6.13 Test Results for Comparison of MLE- and MAP-PCPs in ‘Dissimilar Prior’ Scenarios	205
Table 6.14 Test Results for Comparison of MLE-PCPs in MSOA- & OPC-level Enrichment	208
Table 6.15 Test Results for Comparison of Enrichment Efficiency for Enrichment Samples with Same Spatial Detail Level between ‘Similar’ & ‘Dissimilar Prior’ Scenarios.....	211
Table 6.16 Test Results for Comparison of Enrichment Efficiency between Enrichment Samples Collected at MSOA & OPC Level	212
Table 6.17 Test Results for Comparison of MAP-PCPs in MSOA- & OPC-level Enrichment	214
Table 6.18 Test Results for Comparison of Actual and Expected MLE-PCPs in ‘Similar’ Scenarios	218
Table 6.19 Test Results for Comparison of Actual and Expected MLE-PCPs in ‘Dissimilar’ Scenarios	221
Table 6.20 Test Results for Comparison of Actual and Expected MAP-PCPs in ‘Similar’ Scenarios	223
Table 6.21 Test Results for Comparison of Actual and Expected MAP-PCPs in ‘Dissimilar’ Scenarios	224
Table 6.22 Test Results for Comparison of MLE- and LGR-PCPs in ‘Similar Prior’ Scenarios	230
Table 6.23 Test Results for Comparison of MLE- and SVM-PCPs in ‘Similar Prior’ Scenarios	231
Table 6.24 Test Results for Comparison of MLE- and LGR-PCPs in ‘Dissimilar Prior’ Scenarios	234
Table 6.25 Test Results for Comparison of MLE- and SVM-PCPs in ‘Dissimilar Prior’ Scenarios	235
Table 6.26 Test Results for Comparison of MAP- and LGR-PCPs in ‘Similar Prior’ Scenarios	238

Table 6.27 Test Results for Comparison of MAP- and SVM-PCPs in ‘Similar Prior’ Scenarios	239
Table 6.28 Test Results for Comparison of MAP- and LGR-PCPs in ‘Dissimilar Prior’ Scenarios	240
Table 6.29 Test Results for Comparison of MAP- and SVM-PCPs in ‘Dissimilar Prior’ Scenarios	242
Table 7.1 Statistics for Metrics of Information for LTDS (Chapters 5 and 6) and MND (Chapter 6)	249
Table 7.2 Probability Distribution and Corresponding Entropy of Enriched Socio- demographic Attribute in Each Series of Enrichment Experiments	255
Table 7.3 Number of Experiments in Each Series and Corresponding Maximum Mutual Information and Enrichment Efficiency Levels.....	256

1 INTRODUCTION

The PhD study documented in this thesis concerns the proposal and the development of an Inverse Discrete Choice Modelling (IDCM) socio-demographic enrichment framework. The background, motivation, aims and objectives of the current study are set out in this chapter, along with the definition of several common terms and the structure of this thesis.

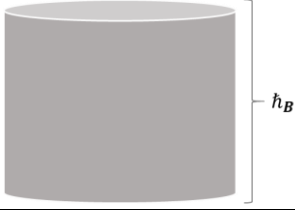
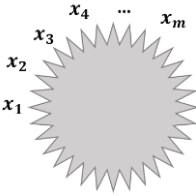
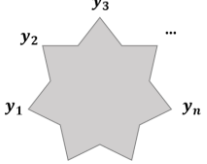
1.1 Background

Recent decades have witnessed a growth in the applications of so-called ‘big data’, viz. datasets containing more data than common software tools are able to capture, curate, manage or process within a reasonable elapsed time (Snijders, Matzat and Reips, 2013). Such data are often created or gathered by electronic equipment or devices frequently classified under the umbrella term of information and communications technologies. Examples of this include the movement data generated by mobile phones while the navigation of embedded positioning systems are used (Awange, 2012) and the vehicle plate number information collected by traffic monitoring cameras when checking the registration (licence) information of vehicles (Liu *et al.*, 2020). Whilst often collected for specific purposes, these big datasets are argued to be applicable in a wider variety of other contexts because they can provide insights into various aspects of people’s daily behaviour yet are collected at potentially much greater accuracy, across a larger sample and, most importantly, at a lower cost than traditional sources (Bohte and Maat, 2009).

For transport-related analysis and practices, one natural application of the subset of these types of data that contains spatial-temporal information is to understand people’s daily movement behaviour patterns. This subset of big data is hereafter referred to, from the transport modelling perspective, as *transport big data* in this thesis while they have also been named using other similar terms such as *trajectories* or *geotagged big data* (Giannotti and Nanni, 2007). Traditional methods of collecting such insights come with a high cost, not only in monetary or effort respects, but also in the burden on the respondents of manually recording details of corresponding trips, leading to possible recollection issues (Bricka and Bhat, 2006). The aforementioned benefits of big data sources have therefore led to them being used increasingly widely in many transport contexts (Chen *et al.*, 2016). This is significantly reflected in a trend towards developing models of travel behaviour and demand that rely on various big data

sources with spatial-temporal information of respondents. Attempts to infer travel-related behavioural variables, e.g. trip purposes, have sourced big data from positioning systems (e.g. GPS and GALILEO) (Chen *et al.*, 2010; Ectors *et al.*, 2017), payment transaction systems (e.g. London Oyster smartcard ticketing) (Kusakabe and Asakura, 2014; Han and Sohn, 2016) and mobile networks (Calabrese *et al.*, 2010).

In parallel to the enthusiasm created by the opportunities for new applications, researchers have also become increasingly aware of the limitations of using big data for transport analysis. Particularly, one direct shortcoming associated with big datasets is that they tend to be ‘thin’, i.e. containing only a handful of variables that can describe or interpret the data points themselves. For instance, a typical ‘big-but-thin’ dataset collected from a GPS logger often contains numerous data points of timestamped geographical coordinates yet lacks readily accessible and meaningful information about places, the respondents or their activities. This is as opposed to typical ‘small-but-thick’ datasets, e.g. travel surveys which contain limited data points while each data point includes a large number of variables characterising the respondents, their households and their travel behaviours. In the present thesis, the above descriptive information of a dataset is termed as its *semantic content* (Sheth, 1996). Accordingly, passive big datasets are thus recognised as possessing the property of ‘*low semantic richness*’.

Aspect	‘Small-but-thick’ Datasets (Rich Semantic Content)	‘Big-but-thin’ Datasets (Poor Semantic Content)
Overall Data Volume ¹		
Each Data Record ²		
Examples	Traditional travel surveys	Mobile network data

Note:

- 1) The height h of the cylinders represents the volume of the data, i.e. $h_S \ll h_B$;
- 2) m, n : number of variables contained in each data record ($m \gg n$);
 $\{x_1, x_2, \dots, x_m\}, \{y_1, y_2, \dots, y_n\}$: variables contained in each data record.

Figure 1.1 Conceptual Comparison between Big and Small Datasets

Figure 1.1 (above) conceptually compares ‘big’ and ‘small’ datasets (defined by the volume) which are respectively ‘poor’ and ‘rich’ in terms of semantic contents, as reflected by the number of variables that characterise each data point.

To date, the most prevailing practice to obtain absent semantic contents of datasets is via an accompanying survey (Schönfelder *et al.*, 2002). In most empirical applications of big data, however, follow-up surveys have become increasingly difficult or expensive to conduct, or even impossible due to the design of the data collection, or the privacy considerations set in legislation or expected by the public. For example, the General Data Protection Regulation (or the GDPR) represents the new legal framework for data protection responsibility in the European Union (Blackmer, 2016). Under the GDPR, data owners have significantly more legal responsibility in terms of data protection and usage for specific purposes, limiting the ways in which personally-identifiable information can be accessed or used. Such contextual information, e.g. trip purposes, the nature of the performed activities and the socio-demographic information, is nevertheless critical for travel demand modelling. The poor semantic richness of typical big data can hence hamper transport analysis in a number of ways. These range from the limited analytical depth of the insights, or problematic employment of data-hungry modelling, to failure to account for unobserved and confounding factors which possibly result in erroneous inferences or misguided policy implications. In light of the importance of contextual information and the difficulty associated with conventional methods to obtain the semantic contents, effective enrichment techniques for improving the semantic richness of anonymous, passive big datasets, without the need for a follow-up contact with the original respondents, are urgently needed to unlock the applicability of such data in not only travel behaviour modelling but also possibly in other domains, e.g. the identification of the vulnerable or voting pattern predictions (Gebru *et al.*, 2017; Steele *et al.*, 2017).

1.2 Motivation

In response to both the potential and limitations of emerging big data sources as outlined in Section 1.1, an extensive body of studies has recently emerged to develop or apply suitable methods to enrich the semantic contents of big datasets to thus enhance their applicability in various contexts. In transport modelling and travel behaviour analysis, for instance, substantial progress has been made in the enrichment of a number of travel-related patterns, e.g. modes of transport, trip destinations, the purposes and types of activities on-route or performed at the destination, from various big data sources where these variables are originally unavailable

(Brunauer *et al.*, 2013; Calastri *et al.*, 2019; Seo *et al.*, 2019). The enrichment of information about the socio-demographic characteristics of respondents and their households has lagged, however, and lacks a solid theoretical (behavioural as opposed to statistical) basis. A theoretical underpinning is important for the enrichment of socio-demographic information because microeconomic theory implies that people's behaviour is, or at least partly, driven by their socio-demographic attributes (Train and McFadden, 1978). This hence provides an additional principle to be followed by the development of any socio-demographic enrichment methods: to retain the dependence between the behaviour pattern(s) serving as the input feature of the enrichment (referred to as the input behaviour feature) and the socio-demographic attribute(s) to be enriched. One immediate consequence to neglect the underlying behavioural theories, as is reflected in existing socio-demographic enrichment methods, concerns their inability to predict or interpret their enrichment performance given the variation of data conditions. Such lack of a firm theoretical foundation would hamper the transferability of the approach under different data conditions or across application contexts. This motivates the current study to systematically establish a framework for predictable and interpretable enrichment of the socio-demographic attributes for anonymous big data sources based on the underlying data conditions.

Specifically, the present thesis draws upon the technique of discrete choice models (DCMs), a fundamental transport planning and policy-making tool for modelling discrete choices or behaviours in transport contexts, e.g. travel mode, route and destination. Relying on the known behavioural foundations, and assumptions firmly grounded in microeconomic consumer theory (Train and McFadden, 1978), DCMs establish the link between the probability of making decisions between a number of choice alternatives and the explanatory variables that can affect such choice behaviour (Ben-Akiva and Lerman, 1985). The latter explanatory variables, as implied by Lancaster (1966), consist of the properties of the choice alternatives and the characteristics of the decision-makers. DCMs have been developed and used over the past several decades, proving their versatility and robustness. Statistically capturing the theoretical correlation between the input features (behaviour patterns) and the output (socio-demographic attributes) of the intended enrichment, DCMs provide the building blocks to obtain the unobserved socio-demographic attribute from observations of the behaviour patterns while retaining the casual dependence between the two.

Motivated by the strong theoretical basis of DCMs as well as their transparent formulation and prevalent use in transport context, this PhD study, based on the idea proposed by Pawlak *et al.* (2015), explores the functionality of DCMs in the inverse: to enrich the semantic richness of

anonymous big data which explicitly or implicitly informs the behaviour patterns of respondents by supplementing additional information about the respondents and their households that are known to drive those behaviours.

1.3 Aims and objectives

Motivated by the need to improve the semantic richness of the increasingly available ‘big-but-thin’ data sources, the overall aim of the present thesis is to develop and validate an IDCM framework for enriching the socio-demographic information of passive and anonymous big data sources with predictable and interpretable enrichment performance.

The idea of inverting DCMs is originally conceptualised by Pawlak *et al.* (2015) and systematically formulated in the current study. Specifically, as is mentioned previously in Section 1.2, the IDCM framework developed in this study postulates that a DCM can serve as a mechanism to obtain the socio-demographic information absent from anonymous big datasets given knowledge of the behaviour patterns observed or extracted from the data. The quality, i.e. the performance, of such enrichment is driven by the theoretical dependence between the socio-demographic attributes and the behaviour features, and thus can be estimated and interpreted. In particular, the above aim includes the following research objectives (ROs):

- RO1:** to derive an econometric formulation of the IDCM approach which suggests the outcome (probability) value of the enriched socio-demographic attributes, given observations of the behaviour patterns for discrete choice alternatives.
- RO2:** to develop measures that quantify the enrichment performance of the IDCM approach at both disaggregate (individual) and aggregate (sample) levels; and to develop metrics for the explanatory power of the behaviour features in inferring the enriched socio-demographic attribute, which represents the data condition (structure and distribution) underlying the sample that requires enrichment.
- RO3:** to establish the IDCM performance theory that theoretically and mathematically links the disaggregate-level enrichment performance of the IDCM approach with respect to the metric of the explanatory power (as developed in RO2), so as to enable the estimation of the IDCM enrichment performance in relation to the data condition of the sample prior to the enrichment.

RO4: to validate the proposed IDCM performance theory in empirical application contexts so as to: (i) investigate the ability of the IDCM performance theory to estimate the IDCM enrichment performance; (ii) compare the IDCM framework against the methods extensively applied in previous attempts at socio-demographic enrichment in terms both of enrichment performance and transferability.

Several academic articles were also generated, as listed in Appendix 1 (Page 291), in addressing the above ROs.

1.4 Definition of common terms used

In order to facilitate the understanding of this PhD study, this section defines several terms commonly used in this thesis.

Transport big data: Transport data refers to the data, directly or indirectly obtained, to capture the spatio-temporal movements of the respondents. On the other hand, the term of ‘big data’, is used to describe datasets with high volumes and complex structures beyond the ability of commonly used software tools to process them in a reasonable elapsed time (Snijders, Matzat and Reips, 2013). Therefore, in this thesis, transport big data refers to

the novel sources of data which capture the spatial-temporal movements of respondents and are voluminous in terms of data points.

The proposed IDCM socio-demographic enrichment framework is argued to be applicable for any contexts where the observed behaviour patterns are driven by the enriched socio-demographic attributes. While this is not limited to the context of travel behaviour analysis, the behaviour patterns serving as the input features for the present analysis concern mainly travel-related choice behaviour. Transport big data thus acts as the major data source employed and discussed in the rest of this thesis.

Semantic richness: Semantic richness, or data semantics, concerns

the number of variables which characterise the data points of a dataset thus providing the meaning and describing the use of the data (Woods, 1975).

Hence, semantic richness can be viewed as a mapping between the information stored in the data and the real-world meaning they represent (Sheth, 1997), reflecting the extent to which the data have been interpreted (Smith, 1991). In the present study, the terms ‘low semantics’,

‘poor/low semantic richness’ or ‘lack of semantic contents’ refer to the data containing only a handful of meaningful variables describing the data points themselves.

Semantic enrichment: Semantic enrichment involves

the process of increasing or improving the semantic richness of a dataset by adding supplementary information completely unavailable from the original data using statistical matching approaches and/or other additional sources of information.

It should be noted that the role of the enrichment approaches mentioned or developed in the current study differs from the traditional meaning of data missingness, e.g. non-response in surveys (Andridge and Little, 2010).

Data Missingness							Semantic Enrichment						
Var. No.	A	B	C	...	Y	Z	Var. No.	A	B	C	...	Y	Z
1	3.39	4.00	9.24	1.16	3.03	7.14	1	3.39	4.00	9.24	1.16	3.03	7.14
2	7.24	12.00	6.08	2.14	3.30	7.04	2	7.24	12.00	6.08	2.14	3.30	7.04
3	2.22	8.53	3.14	8.17	4.18	8.20	3	2.22	8.53	3.14	8.17	4.18	8.20
...	6.00	3.77	12.23	4.44	9.35	10.25	...	6.00	3.77	12.23	4.44	9.35	10.25
7086	9.20	6.28	12.08	5.00	6.24	10.04	7086	9.20	6.28	12.08	5.00	6.24	10.04
7087	1.45	9.22	5.25	5.04	6.22	7.96	7087	1.45	9.22	5.25	5.04	6.22	7.96
...	13.30	6.18	4.03	4.99	6.99	3.81	...	13.30	6.18	4.03	4.99	6.99	3.81

Figure 1.2 Conceptual Comparison between Data Missingness and Semantic Enrichment

Figure 1.2 conceptualises the difference between traditional data missingness problems and the semantic enrichment addressed in the current study, with two example tables representing the corresponding datasets. Specifically, data missingness methods typically address the issue of missing values of an existing variable in a dataset so as to allow the analysis to be undertaken as if it were complete. In contrast, enrichment can be seen as an ‘extreme’ case of data missingness, i.e. where there are no observations available for the enriched variable(s), which is thus treated as absent variables.

1.5 Thesis structure

The rest of this thesis is structured as follows.

Chapter 2 documents an overview of existing literature related to the IDCM framework in terms of the enriched variable, source of big data and the method applied for the enrichment. Signifying the shortage in the enrichment of socio-demographic information for anonymous

big data sources, existing attempts at socio-demographic enrichment are specifically discussed, with a focus on the lack of theoretical underpinnings in the reviewed studies.

Chapter 3 improves and completes the conceptual framework of the IDCM approach initiated by Pawlak *et al.* (2015) by conceptualising the IDCM framework from various aspects, so as to enable an understanding of the proposed study. These aspects include:

- 1) a brief overview of discrete choice modelling theory which underlies the econometric formulation of the IDCM approach;
- 2) a conceptual, microeconomic behaviour model to explain the mechanism according to which the IDCM enrichment is formulated;
- 3) a brief introduction of inverse problems, to which the problem addressed by the IDCM approach corresponds, and thus reveals the challenge and value of the current study; and
- 4) an explained analogy between the choice-making process driven by socio-demographic attributes and the transmission of messages in an information communication system, which facilitates the feasibility to adapt metrics formally developed in information theory in establishing the IDCM performance theory.

Chapter 4 demonstrates the detailed establishment of the IDCM framework, consisting of the econometric formulation of the approach and the accompanying theory for estimating the enrichment performance of the approach. In particular, the econometric formulation expresses the value of the socio-demographic attribute(s) absent from a dataset given extracted observation(s) of behaviour pattern(s) of an individual or a group of individuals based on DCMs. This corresponds to the RO1 as explained in Section 1.3. On the other hand, this chapter develops metrics for the enrichment performance of the IDCM approach and the corresponding data conditions, which addresses the RO2. The link between the two metrics are then established as the IDCM performance theory, solving the RO3 as proposed in Section 1.3.

Chapter 5 outlines an empirical application of the IDCM framework using London Travel Demand Survey (LTDS) data, in which several socio-demographic attributes are enriched using the IDCM approach given knowledge of various travel behaviour patterns. The ability of the IDCM performance theory to estimate the IDCM enrichment performance is validated by assessing the consistency between the obtained, actual enrichment performance metric and the expected values calculated using the IDCM theory. In addition, the IDCM framework is compared against representative methods used in existing socio-demographic enrichment

attempts so as to demonstrate the transferability of the proposed IDCM framework. This chapter, together with Chapter 6, addresses the RO4 as described in Section 1.3.

Chapter 6 presents another empirical application of the IDCM framework by enriching an anonymised mobile network data from inverting a DCM estimated from the LTDS data. This extends the conclusions resulting from the application in Chapter 5, which solely exploits the LTDS data, by exploring the ability of the IDCM framework when applied to the socio-demographic enrichment of data collected from different sources. A comparative analysis against existing socio-demographic enrichment methods is then performed. This chapter also corresponds to the RO4 in Section 1.3 while providing a more comprehensive investigation in terms of data conditions as compared against Chapter 5.

Chapter 7 explores the enrichment effectiveness of the IDCM framework, which is raised as a substantive issue possibly presenting challenges to the applicability of the IDCM approach in transport contexts. The issue of the enrichment effectiveness of the IDCM framework is discussed theoretically based on the econometric formulation of the IDCM and analysed empirically using a set of exhaustive Monte Carlo enrichment experiments.

Chapter 8 provides the conclusions drawn from the development of the IDCM framework and the performed theoretical and empirical analysis, reflecting the rationale and the main implications of this PhD study. A brief summary of the study is presented, highlighting the limitations of the IDCM framework established so far, which gives rise to the avenues for further investigations and extensions of the proposed framework.

2 LITERATURE REVIEW

As explained in Chapter 1, the current study proposes and develops an Inverse Discrete Choice Modelling (IDCM) framework to improve the semantic richness of ‘big-but-thin’ data sources, thus enhancing the applicability of such datasets because of their increasing availability to researchers and practitioners. Specifically, the IDCM approach is performed to obtain the socio-demographic attributes of the respondents in anonymous big datasets based on the behaviour patterns observed or extracted from the data (especially those reflecting travel-related decision-making).

This chapter focuses on a review of existing literature related to the data requiring semantic enrichment and the methods to improving their poor semantic richness. This corresponds to the enrichment of, from the perspective of the proposed IDCM approach, the input behaviour features and the socio-demographic attribute given that both serve as significant elements of data semantics. In particular, Section 2.1 briefly introduces the techniques developed or applied to enable the use of transport big data for semantic enrichment, documents existing studies for enriching the aforementioned semantic contents, and highlights the research gap in these studies. Subsequently, Section 2.2 summarises the major transport big data sources enriched in previous studies. Section 2.3 outlines the research field of machine learning, given that machine learning algorithms are extensively applied in existing semantic enrichment attempts, while Section 2.4 provides a summary of the current chapter.

2.1 Overview of existing semantic enrichment studies and research gaps

To date, the most prevailing practice to obtain the absent semantic contents of ‘big-but-thin’ datasets is by an accompanying, or a follow-up, survey which is applicable only to cases where the respondents can be approached (Schönfelder *et al.*, 2002). In most instances, however, a follow-up survey is not feasible, e.g. on the grounds of the costs and burden of additional data collection, or simply because the original dataset was collected so long ago that the detailed recollection of particular behaviours would be dubious. Moreover, the growing demand for the preservation of privacy reflected in public sentiments and legislation, e.g. the General Data Protection Regulation (2016) in Europe and the Data Protection Act (2018) in the UK, have

made it increasingly difficult to approach respondents to obtain their socio-demographic information for these surveys. This has hence stimulated a growth of research seeking to develop or leverage suitable methods to supplement the otherwise semantically poor big data without the need for any follow-up contact with the original respondents.

In the context of transport big data, this trend in research has utilised the development of pre-processing techniques, e.g. map-matching and trajectory decomposition techniques, as explained in Section 2.1.1, to extract meaningful trips from raw transport big data. A plethora of studies have thus emerged to extract particular travel-related behaviour patterns from the obtained meaningful trips. Specifically, Sections 2.1.2 – 2.1.5 summarise the research seeking to infer, respectively, the locations, purposes and modes that reflect people’s travel behaviour patterns and the variety of methods developed or leveraged. In addition, the socio-demographic attributes as the essential exogenous determinant for the modelling of behaviour patterns also represent important elements of data semantics. Section 2.1.6 thus discusses the limited number of studies conducted to enrich the socio-demographic attributes of anonymous big data while the corresponding research gap in these studies is outlined in Section 2.1.7.

2.1.1 Techniques to processing raw transport big data into meaningful trips

Raw transport big data often consists of voluminous scattered geotags (timestamped geographical coordinates) rather than continuous tracks. A series of such geotagged data collected for a specific moving object (e.g. an individual) can be used to re-construct its movement patterns, which Giannotti *et al.* (2007) refers to as the ‘*trajectory*’. The re-constructed trajectories further requires additional processing techniques which transform them into meaningful trips. This corresponds eventually to the development of movement map-matching and trajectory decomposition techniques.

The flow chart presented in Figure 2.1 (below) illustrates the procedure to transform raw GPS points into meaningful trips. Although the procedure may vary slightly between different types of big data sources, to explain the techniques required to process raw geotags into meaningful trips for travel behaviour modelling purposes, the paragraphs below demonstrate the aforementioned two streams of pre-processing techniques developed for GPS data as an example.

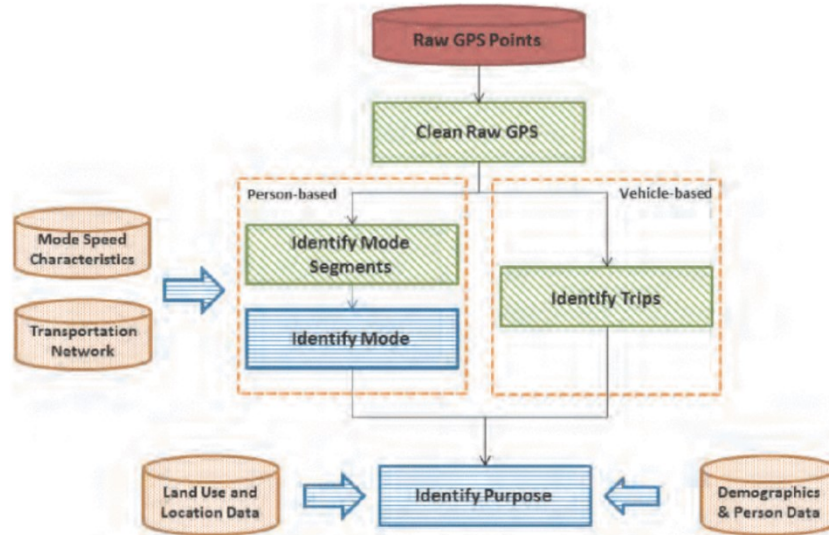


Figure 2.1 Procedure to Process Scattered GPS Data into Trip Data (Source: Wolf et al. (2014))

Map-matching refers to the process of fitting a sequence of observed coordinates onto a digital map. It is necessary for the semantic enrichment of transport big datasets because the raw data are sometimes imprecise, either due to measurement errors arising from the sensor, or sampling errors caused by the sampling rate (Pfoser and Jensen, 1999). An example of map-matching techniques is the *ST-Matching approach* proposed by Lou *et al.* (2009). The ST-Matching approach constructs a candidate graph to identify the best matching path sequence for low-sampling-rate GPS data, i.e. GPS data points collected at relatively long intervals, by combining the spatial geometric and topological structure of the road network with the temporal/speed constraints. The ST-Matching approach was found to outperform other map-matching techniques, e.g. the Average-Fréchet-Distance based map-matching algorithm (Mascret *et al.*, 2006) and the incremental algorithm (González *et al.*, 2007), in terms of accuracy and running time. Likewise, Bierlaire *et al.* (2013) proposed a *probabilistic map-matching approach* which selects from a set of potential paths the one with the maximum likelihood characterised by spatial-temporal information from GPS data. The robustness and effectiveness of their approach was validated in a real route choice modelling context.

The mapped trajectories require further decomposition techniques to segment them into meaningful trips before becoming applicable for the inference of travel-related behaviour patterns, which relies on the extraction of important movement features. For instance, Dodge *et al.* (2009) demonstrated a trajectory segmentation and feature extraction framework by developing an algorithm which decomposes the shapes generated for different movement parameters, e.g. velocity and acceleration, using variations in sinuosity and deviation from the median line. A case study presented that their framework is able to decompose the trajectories

into meaningful trips and to classify unknown moving objects into different types, e.g. humans, animals and vehicles. Alternatively, Hung *et al.* (2015) suggested a trajectory pattern mining framework named *Clustering and Aggregating Clues of Trajectories* to detect the frequent movement behaviour of a user. By identifying the so-called ‘clues’ revealed in the spatial-temporal features of mapped trajectories, they clustered the similar trajectories of a user into groups which represent the user’s movement patterns. Experimental validation using both synthetic and real-world datasets showed the effectiveness of their model in extracting meaningful trips from trajectories.

Enabling the identification of meaningful trips from raw transport big data, the above two strand of techniques have fostered a plethora of studies for improving the semantic richness of big datasets by inferring various travel-related behaviour patterns. These studies are explained in detail in Sections 2.1.2 – 2.1.5, classified according to the enriched variable, in terms of data source and developed/applied methodology.

2.1.2 Origin-destination (OD) matrices

Owing to the development of map-matching and trajectory decomposition techniques, which, as discussed in Section 2.1.1, can be used to derive meaningful trips, one of the most straightforward travel behaviour patterns that can be inferred from transport big data sources involves the locations associated with the derived trips. Specifically, the extracted location-related information can improve the semantic richness from two respects:

- 1) the enrichment of the locations of groups to derive origin-destination (OD) matrices; and
- 2) the enrichment of the types of visited locations (destination choice patterns).

The section below discusses the first stream of enrichment, concerning the derivation of OD matrices by aggregating individual-level locations enriched from different transport big data sources. OD matrices capture people’s movement flows in a certain area. Constituting the product of the second step of a typical four-step travel demand modelling process (McNally, 2007), OD matrices are necessary inputs for the planning, design and management of a transport system. The reliable estimation of OD matrices from traditional data sources can be challenging, however. This is because traditional approaches relying on either roadside and household surveys or traffic link counts involve restrictions such as limited sample sizes or low update frequencies, which can lead to sampling biases, reporting errors, or multiple non-unique

yet plausible OD matrices. The use of transport big data therefore represents an attractive opportunity to supplement these traditional data sources for constructing OD matrices.

For instance, Castillo *et al.* (2008) proposed a model for the estimation of paths and OD matrices using a combination of licence plate scanning data and traffic link counts. Subject to the conservation law stated for scanned links and to the observed traffic counts estimated according to the locations extracted from scanned plate data, their model involves a *quadratic function* of the weighted errors with respect to a given set of prior path flows. The importance of such data for the efficient estimation of OD matrices was demonstrated in an application of their model in the Nguyen–Dupuis network. Moreover, they developed a method for selecting the optimal set of links to be scanned for the prediction of traffic flows, which indicated an important practical implication for the effective use of such transport big data, given the high installation costs for licence plate readers. Other examples of using plate scanning data in the development of methods for OD matrix estimation can be found in the studies by Asakura *et al.* (2000), Rao *et al.* (2018) and Liu *et al.* (2020).

In more recent years, portable devices with embedded positioning systems, such as GPS-enabled smartphones, have been viewed as alternative sources for traffic data collection. GPS data is therefore also employed for the estimation of traffic flows and OD matrices in many applications. For example, Herrera *et al.* (2010) presented a field experiment evaluating the capability of GPS data to provide accurate measurements for traffic flows via a prototype traffic monitoring system. Such a system is established to exploit the locations derived from the extensive cellular network coverage and the high positioning accuracy provided by GPS devices. The results of their study indicated that 2% – 3% smartphone penetration can provide enough information for the accurate measurements of OD matrices.

High-quality OD matrices can also be estimated using GPS data combined with other transport big data sources. For instance, Munizaga *et al.* (2012) suggested a *rule-based model* to obtain the OD matrices for a multimodal public transport system based on alighting and activity stops extracted from a combination of GPS and smartcard data. When applying their method to two datasets collected at different times, they found that it could achieve a more than 80% accuracy for the enrichment of alighting positions, hence enabling the generation of OD matrices by aggregating the sample. In addition to the two examples discussed above, the work by Parry and Hazelton (2012), Morimura and Kato (2012), Sbaï *et al.* (2017), Ge and Fukuda (2016), and Moreira-Matias *et al.* (2016) also attempts to estimate OD matrices using data collected

from GPS-enabled devices. Similarly, smartcard data has also been exploited extensively to obtain OD matrix information. Examples of this include a bus-stop level OD matrix built by Zhang *et al.* (2007), rail passenger OD matrices constructed by Zhao *et al.* (2007), and a public transport OD matrix estimation method developed by Alsger *et al.* (2016).

While the applications of GPS data can be limited where the market share of GPS devices is low, e.g. in developing countries, mobile networks represent another reliable source of transport big data. This is owing to the facts that mobile phone users can provide approximate locations when making calls or sending text messages, and the global penetration of mobile phone use is high even in developing countries (Iqbal *et al.*, 2014). For instance, Iqbal *et al.* (2014) illustrated a methodology presenting the analysis of mobile phone call detail record (CDR) data to generate tower-to-tower transient OD matrices for a specific time period. They have further associated the generated tower-to-tower OD matrices with traffic network nodes to obtain node-to-node transient OD matrices. And the actual OD matrices were obtained by scaling up the node-to-node matrices according to the scaling factors determined from observed traffic counts. Validated using empirical CDR data collected from Dhaka, their method was observed to be effective, especially for generating complex OD matrices where land use patterns are heterogeneous and travel history data sources are limited. A number of subsequent investigations have also been performed to develop OD matrices using mobile network data, such as in the studies by Bonnel *et al.* (2015), Alexander *et al.* (2015) and Bachir *et al.* (2019).

Later in Section 2.1.7, Table 2.1 summaries the above studies as the enrichment of OD matrices for transport big data, together with the investigations attempting at enriching other travel-related behaviour variables, in terms of data source, methodology and accuracy (if available).

2.1.3 Types of locations

As described in Section 2.1.2, the second stream of analysis with respect to the locations obtained from transport big data sources relates to the type of locations visited by moving objects, or the choice of destinations. Works in this area have focused on the development of data mining algorithms to classify scattered locations observed in big databases.

Earlier investigations into this aspect of big data mining include, for example, the work by Ester *et al.* (1996), who presented a *Density-based Spatial Clustering of Applications with Noise (DBSCAN)* model that makes it possible to accommodate large spatial databases with minimal input parameters, clusters with arbitrary shapes, and good efficiency. As an extension

to classic clustering algorithms for spatial data classification (Beckmann *et al.*, 1990), the key to the DBSCAN involves recognising the existence of a typical density of points within a cluster which is usually considerably higher than outside the cluster. Handling up to 10% noise, the DBSCAN was found to outperform the well-known CLARANS algorithm in both efficiency and effectiveness at the time it was proposed.

Whilst the above DBSCAN model deals with the movement data purely in terms of their spatial characteristics, more recent research has shown an increasing awareness of the importance of analysing movement patterns in their spatio-temporal contexts, thus motivating approaches such as the *event-based conceptual model* developed by Andrienko *et al.* (2011). In particular, their model considers movements as a combination of spatial events and corresponding spatial-temporal contexts, and manages to extract event and context variables using *computations and interactive filters* as well as the spatial-temporal relations of the two. Validated by a prototype implementation in complex real-data context of wildlife movements, their model achieves approximately 90% accuracy in predicting the location type and was visualised using various basic visualisation tools, such as cartographic map display, space-time cube and scatterplots.

In addition to the methodologies outlined above, Andrienko *et al.* (2011) developed a *visual analytics procedure* by determining places-of-interest based on the time-varying characteristics of movements, i.e. certain repeated events in the movement data, of which the general procedure consists of: (i) event extraction and clustering; (ii) place extraction and spatial-temporal aggregation; and (iii) analysis of the aggregated data. Their procedure was validated in two empirical applications, respectively, which dealt with traffic congestion in Milan and flight dynamics in France, both showing plausible performance as assessed by domain experts.

One more instance of the studies attempting to extract the location types can be found in the study conducted by Gong *et al.* (2015) who improved the DBSCAN with 90% accuracy in identifying meaningful stop points by applying a time sequence constraint and a direction change constraint. Additionally, their approach contributes to combining a *support vector machine* (SVM) method to extract the duration of the stop, the mean distance to the centroid of a cluster at a stop location, and the shorter distance from the stop to home and to work, which distinguishes activity stops from non-activity stops with 96% confidence.

In terms of data source, various types of transport big data have been employed in the development or applications of approaches for the classification of visited locations. For instance, Lee *et al.* (2017) proposed a *rule-based method* for indoor-outdoor location

classification using GPS data. Applications of their method suggested 72% – 74% prediction rates for different seasons with the pure use of GPS data and an improvement of 12% – 14% with additional temperature information.

Mobile network data is another source of spatio-temporal big data that can be used to identify important locations. An example of this is the two-stage algorithm for detecting meaningful locations (e.g. home and work) suggested by Isaacman *et al.* (2011), which includes: (i) spatially *clustering* the cell towers contacted by a user; and (ii) the selection of important clusters by means of a *logistic regression model*. Applied to a context of commute distance estimation for carbon footprint analysis, their algorithm was found to perform equivalently to government census data and achieved 80% accuracy.

More recent studies have shown a trend to exploit multiple data sources for the inference of location types. For example, the practical use of the *time-and-grid-based clustering* technique proposed by Montoliu *et al.* (2013) in discovering places-of-interest is validated using the multimodal data collected through a client-server system installed in mobile phones. This particularly integrated the data collected via GPS, Wi-Fi, GSM and accelerometer sensors. Their model was found to be able to discover 63% of locations correctly, which substantially outperforms the use of GPS sensors alone.

2.1.4 Trip purposes and activities

The purpose or the type of activity performed in one or several trips can sometimes be inferred from one another with the type of visited locations (as explained in Section 2.1.2) (Ashbrook and Starner, 2003; Bennewitz *et al.*, 2005). For example, an activity understood as a commute (i.e. the purpose of a commute trip) can inform the category of the location to be visited as ‘work’ with high confidence. Likewise, regular visits to a primary school allows the conclusion that the traveller’s activity is education-related or involves escorting children. With additional information, such as the age, income level or marriage status of the traveller(s), the range of activities performed at a specific location can be narrowed and the inference confidence increases accordingly. Thus, the enrichment of activities and trip purposes according to the extraction of locations is explained below.

Relying on rather simple and limiting assumptions, early-stage activity recognition studies ignored either the type of places or the routes between places, which inevitably constrained the scope of these studies (see for examples the works by Ashbrook and Starner (2003) and

Bennewitz *et al.* (2005)). Successive researchers hence sought to relax these assumptions to achieve more flexible activity identification. For instance, Giannotti *et al.* (2007) suggested the concept of a *trajectory pattern* (T-pattern) to characterise movement behaviours by a bunch of independent routes sharing the same sequence of visited places and similar travel durations. Based on the in-route places and the travel durations between these places, they proposed *density-based algorithms* and *T-pattern mining algorithms* with either pre-defined regions or dynamically detected regions to automatically determine specific T-patterns. These algorithms were verified in an application of trip purpose enrichment for real-world GPS traces.

In addition to the type of visited locations, trip purposes and activities are also found to be significantly related to the origins of the attendees of corresponding events. Specifically, Calabrese *et al.* (2010) conducted an experimental study using geotagged traces collected from mobile networks to predict crowded events and their effects on urban systems. By analysing a combination of anonymised traces from a metropolitan area, and several selected events that attract considerable numbers of people, they found that events of the same type show similar spatial distribution patterns for the origins of the attendees. This is important from the point of view of appropriate management for large-scale events and activities.

In terms of the methodologies developed or applied for the enrichment of trip purposes and activities, Gong *et al.* (2014) provided a systematic categorisation, according to the type of algorithm and input data, as illustrated in Figure 2.2.

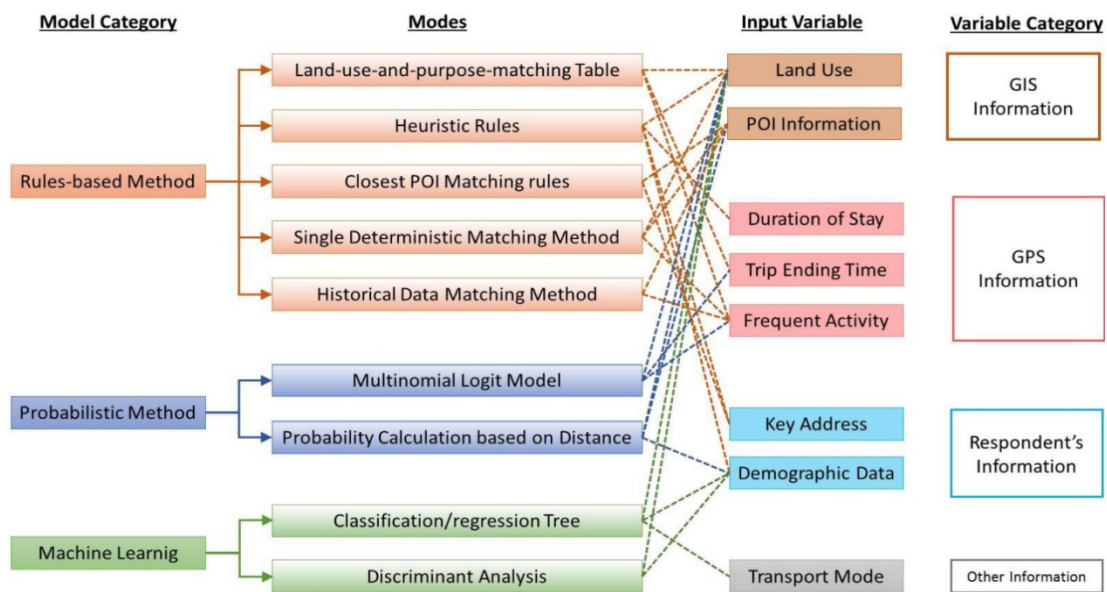


Figure 2.2 Classification of Methods for Trip Purpose Inference (Source: Gong *et al.* (2014))

Rule-based algorithms constitute the most popular means for trip purpose and activity identification. For instance, Wolf *et al.* (2001; 2004) developed a *rule-based algorithm* to enrich trip purposes and/or activities for GPS data by defining land use codes with pre-classified labels that characterise the time-of-day and duration-of-stay properties. Two car-based studies were conducted to evaluate their algorithm, which demonstrated that the enrichment performance of their model could match, or even exceed, the reporting quality from traditional participant-based survey data. This indicates the potential of using GPS logger data to replace traditional travel diaries completely for obtaining trip purpose and activity data.

To identify the type of activities with an optimal balance between improving model accuracy and retaining sufficient classification information, Ectors *et al.* (2017) developed an optimised classification methodology using a two-step data mining method. Specifically, their method involves: (i) searching for an optimal number and composition of the activity classes, and (ii) modelling the type of activity using probability and rule-based heuristic techniques. Their approach was verified using household travel-activity survey data and reached 79% accuracy. Importantly, their approach was found to outperform most previous studies by employing a suboptimal set of activity classes for higher classification accuracy yet preserving information from the original datasets.

Another important methodology for trip purpose and activity inference are probabilistic methods, which estimate the probability of each purpose using information derived from various sources such as GPS, Geographic Information System (GIS), and possibly the information about the respondents. An example can be found in the study conducted by Chen *et al.* (2010) who established a *probabilistic model* for trip end clustering and trip purpose detection, including time-of-day, history dependence and land use characteristics. Reporting success rates ranging between 60% and 95% in a case study based on GPS data, their investigation acknowledged the feasibility of using passive GPS travel survey collection methods to supplement or even replace conventional self-reported travel surveys in a complex urban environment with a multimodal transport network. This work is viewed as a benchmark in the approaches for trip purpose enrichment owing to its outperformance of most previous approaches, such as the study by Onay and Öztürk (2018), which often perform significantly better in mode detection than purpose recognition.

Assisted with GIS land use data and the attributes of the respondents, as well as features of travel modes, machine learning approaches can also be applied to classify different types of

trip purpose. For instance, Liao *et al.* (2007) demonstrated a hierarchically structured framework using a *conditional random field model* that is trained to identify and label significant activities (and corresponding locations) in which an individual is engaged over several weeks from GPS data. It is argued that their model can address some limitations of previous activity-and-place labelling approaches, e.g. interdependence between places and activities in terms of accuracy and flexibility, given that the higher levels of their framework are constructed based on the inference results from lower ones. The activity enrichment performance of their model was explored using GPS data, reaching 80% accuracy in the detection of activities and over 85% in place identification. Alternatively, Deng and Ji (2010) proposed a *C5.0 decision tree* method for trip purpose classification from GPS data. A proof-of-concept case study using GPS travel survey data was presented, showing that their approach achieved 88% accuracy in trip purpose enrichment.

In more recent investigations, Seo *et al.* (2019) proposed an interactive activity-travel survey method which employs an online estimation model for automatic estimation of the travel context, and an online machine learning method for dynamically updating the estimation model. Their method was found to be implementable on mobile devices to enhance the use of mobile phone trace data while reducing the burden on respondents. The interaction process of the approach allows the successful enrichment of the activity patterns in an empirical application. In addition, the effectiveness of their method was significantly demonstrated in long-term surveys, where high data quality is retained while traveller heterogeneity is also captured.

Whilst the content above mainly discusses trip purpose inference using GPS data, smartcard data have also been exploited to extract trip purpose information. An example of this is presented in the study by Kusakabe and Asakura (2014), who developed a *Bayes' probabilistic model* for data fusion of travel survey data for trip purpose detection. The empirical application of their model showed that 86% of the trip purposes were successfully enriched. An extension of their model, incorporated with an additional fare attribute, was conducted by Kusakabe *et al.* (2016), showing an average enrichment rate for major trip purpose of 88%. Another example of the use of smartcard data involves the methodology developed by Han and Sohn (2016), who enriched the sequence of activities for each trip chain using a continuous *hidden Markov unsupervised machine learning model*. Specifically, the algorithm is trained using the information about the trip chains revealed in smartcard data as well as the GIS land use characteristics. The importance of this model is recognised given the plausible estimation results for activity pattern inference and the advantage of not requiring a training dataset.

Beyond extracting only trip-related activities from transport data sources, another interesting study by Li *et al.* (2017) presented a *dictionary learning-based algorithm* to solve the statistical modelling framework for predicting digital activities from application-level traffic data collected from cellular network operators. Their work demonstrated the potential to predict activities conducted in the digital world from passive data, which can assist in the management of telecommunication networks. In a context of enrichment for real traffic records from China Mobile (2G/3G/4G) cellular networks, their method was compared to ARMA and Kalman filtering algorithms and was found to provide a unified solution for application-level traffic modelling and prediction with an appealing accuracy of approximately 95%.

2.1.5 Modes of transport

Understanding the mode choices of travellers is important due to the essential role of travel modes, especially public transport, in most transport-related policy-making (de Dios Ortuzar and Willumsen, 2011), e.g. the congestion charge applied in London (Leape, 2006). Therefore, the extraction of mode choice behaviour patterns from spatial-temporal data represents another crucial effort for improving the semantic richness of transport big data.

As is shown in Figure 2.3, Gong *et al.* (2014) categorised the methodologies for travel mode enrichment. To provide a clearer explanation for the progress made in travel mode detection, however, the contents below discuss the previous studies for travel mode enrichment according to data source and the input feature instead of the types of method employed.

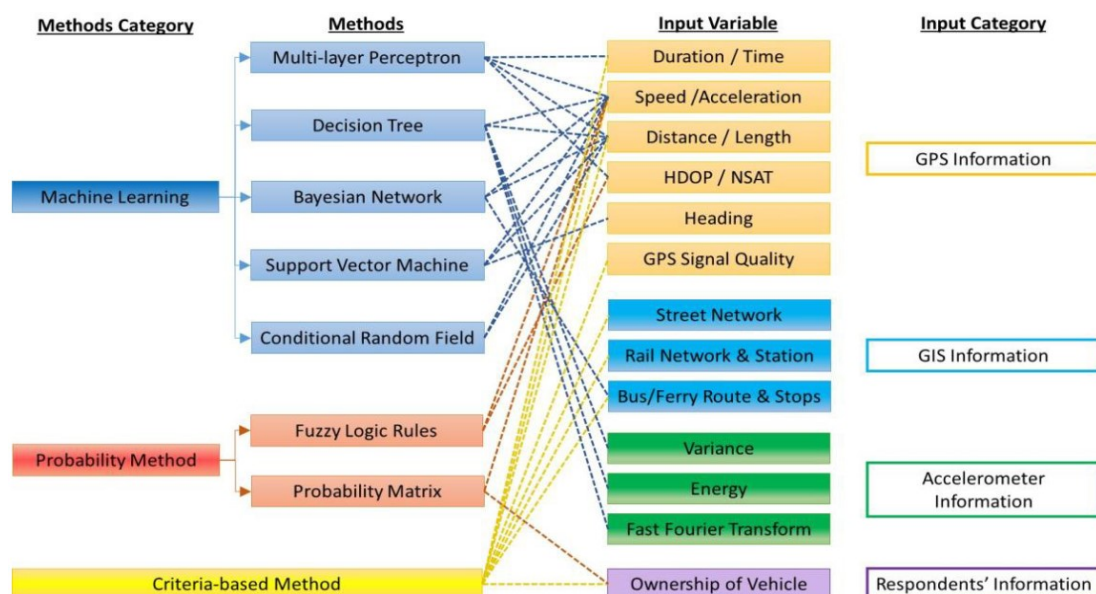


Figure 2.3 Classification of Mode Detection Methods (Source: Gong *et al.* (2014))

Early studies suggested that the kinematic properties of a moving object (e.g. average and maximum speeds/accelerations) that could be extracted from movement data facilitate discriminating between different modes of transport. For example, Byon *et al.* (2007) developed a *multi-layer perceptron neural network* for travel mode identification using the data collected from GPS-enabled mobile phones, with an accuracy rate as high as 98%. The impact of sampling rates on the detection accuracy was explored as the service was charged by queries, leading to the result that collecting data every 1.5 minutes can best accommodate the trade-off between the accuracy rate (80%) and the query cost.

It has been argued, however, that the kinematic characteristics on their own cannot offer ideally accurate mode identification results. Additionally, information from other data sources, such as GIS land use data, can be incorporated to improve the detection rate. Chung and Shalaby (2005), for instance, proposed a *rule-based elimination algorithm* to distinguish between four travel modes based on their attributes, with 92% accuracy. In their model, GPS travel data was supplemented with GIS land use data to establish the elimination rules for detecting and classifying road links. Examples of these elimination rules include whether the travelled links contain transit lines and whether the type of road infrastructure of the travelled links can accommodate corresponding mode.

Another example of using multiple data sources for travel mode enrichment can be seen in the study conducted by Stopher *et al.* (2008), who also applied a *rule-based elimination algorithm* to distinguish between walking, cycling and driving based on the movement properties. Additional information about the underlying street and public transport availability provided by the GIS layers was used to further detect the specific motorised vehicle, leading to a mode detection rate of 95%.

In addition to combining data from GPS logs and GIS technologies, Bohte and Maat (2009) developed an innovative approach to improve mode detection by providing the respondents with an *online interface* to indicate trip attributes as a validation stage. Applied to a large-scale empirical study, this approach was found to perform equivalently to travel survey results, showing that the enriched travel mode data gathered using their method is a reliable substitute for missing data from paper diaries.

Researchers also seek to enrich travel modes by incorporating GPS data with data types other than GIS data. For instance, Feng and Timmermans (2013) assessed the relative merits of combining accelerometer data with GPS data for travel mode detection, using a *Bayesian Belief*

Network. It was found that using accelerometer data itself outperforms GPS data alone, but that a combination of both can result in the best enrichment performance as compared with the sole use of either type of data. This confirms the importance of incorporating various data sources to making the best use of information for improving performance of mode choice enrichment.

Some recent studies, however, have moved on from combining data sources to highlight the potential of mining more features of moving patterns from a single data source. For example, Brunauer *et al.* (2013) suggested a *feed forward multilayer perceptron* which can extract and analyse distinct motion patterns characterised by the relative frequency distributions of movement features such as speed, acceleration and horizontal angular, deduced from GPS trajectories. Their method was able to identify modes of transport with 92% confidence.

Similarly, Dabiri and Heaslip (2018) implied that the shortcomings engendered by manual features can be overcome by utilising a *Convolutional Neural Network (CNN)* to derive high-level features automatically. Taking advantage of CNN architectures, they predicted travel modes purely from GPS trajectories with an accuracy rate of 85%, which they found to be superior to traditional machine learning methods.

In more recent literature, researchers have also managed to improve the enrichment rate by applying a combination of algorithms. An instance of this is the study conducted by Wang *et al.* (2018), in which a *random forest classifier* is presented and combined with a *rule-based approach* to distinguish between six different modes with a success rate of over 85% using GPS data only. This classifier was found to outperform SVMs and Artificial Neural Networks. It was also revealed that additional knowledge of the socio-demographic attributes of travellers can help improve the detection rate based on their approach.

2.1.6 Socio-demographic attributes

Compared to the success in extracting variables that reflect travel-related behaviour patterns from transport big data sources (Sections 2.1.2 – 2.1.5), the enrichment of information about the respondents and their households, such as their socio-demographic attributes, has remained relatively underexploited for anonymous big data. This can be attributed to the difficulties of conducting conventional data collection methods, e.g. follow-up travel surveys, and the growing awareness and demand in public sentiment and legislation for the protection of privacy, e.g. the GDPR (2016) in Europe and the Data Protection Act (2018) in the UK. Nonetheless, socio-demographic attributes are essential elements of the semantic contents of a dataset for

behaviour-based analysis given their role as the (typically) exogenous determinants of human behaviour patterns. This is illustrated, for example, in the studies by Wolf *et al.* (2001) and Axhausen *et al.* (2004), which show that socio-demographic information is required as an input for the corresponding analysis. Hence, the enrichment of big data sources—including but not limited to transport big datasets—with the socio-demographic attributes originally absent from those datasets remains a crucial area. Given the need for this data, but the challenges in accessing it, there has been an emerging field of research seeking to develop or leverage approaches to enrich the socio-demographic attributes for anonymous big datasets to improve their semantic richness and applicability in various contexts of behaviour-driven analysis.

The past few decades have seen dynamic developments in the understanding and modelling of people's spatial-temporal movement behaviour patterns, viz. travel behaviour analysis or human mobility analysis, in two broad yet distinct research fields (Chen *et al.*, 2016). Specifically, one field emerged from the needs of traditional travel behaviour analysis and modelling, while the other emerged from interests, often methodological, in data science, seeking to extract information from various big data sources. The extracted behaviour patterns then serve as the input features for understanding people's socio-demographic information. Thus, the following paragraphs discuss the so far limited number of socio-demographic enrichment attempts made in the two fields in terms of the data sources, employed methods and the type of socio-demographic attributes that are enriched, starting from the latter field of data science, where such enrichment is more extensively performed.

In data science and the relevant domains, there has been a trend to use mobile phone metadata, or CDRs, in the enrichment of socio-demographic attributes on account of the widespread coverage of mobile networks and the ability to extract both mobility and non-mobility (social) behaviour patterns of mobile network users. An example can be found in the investigation by Frias-Martinez *et al.* (2010), who proposed a *semi-supervised k-means clustering algorithm* for gender prediction from the behavioural, social and mobility patterns revealed in the CDR data collected from a developing economy. The results of their study suggested that the behavioural and social variables, including input/output calls and the degree of social network, significantly differ between males and females. Achieving a correct prediction rate of over 80%, their algorithm was found to outperform SVMs and Random Forests with boosting, which also confirms the applicability of CDRs for nonintrusive gender inference and extends the applications of mobile network data.

Herrera-Yagüe and Zufiria (2012) also used the communication patterns extracted from CDRs to predict the gender and age attributes of mobile phone users. A number of machine learning algorithms including Linear Discriminant Analysis, SVMs, Multilayer Perceptrons and decision trees were tested, based on both individual users and the entire networks. It was found that the enrichment performance once the communication patterns of isolated users were included (at a ca. 53% correction rate) could be improved by combining additional patterns at the network structure level for an accuracy of ca. 65%.

Other examples of using machine learning algorithms to extract socio-demographic attributes from mobile phone usage data include the works by Sarraute *et al.* (2014), Dong *et al.* (2014) and Jahani *et al.* (2017). Although all these studies involve the use of fundamentally similar machine learning algorithms for the prediction of age and gender based on communication patterns observed or extracted from mobile phone usage data, they use different communication patterns as the classification features for the corresponding enrichment.

Whereas the above studies rely extensively on the complexity of how the classification features are derived, Felbo *et al.* (2017) proposed that the enrichment approach can be developed to model temporal properties directly extracted using deep learning algorithms from raw mobile phone metadata. In particular, they proposed using a *convolutional network architecture* to capture the mobile phone usage pattern within a week, and three additional methods for aggregating patterns across weeks. Various machine learning algorithms are tested for the subsequent socio-demographic enrichment, with the SVM yielding the highest accuracy of 63% for age and 80% for gender. By not using any mobility features, the power of their algorithm by itself was validated as it presented equivalently accurate enrichment performance to the previous methods that typically include mobility patterns as the input feature. Their approach is proved effective for humanitarian efforts, e.g. in helping a non-profit organisation track displaced populations and deliver rapid aid to refugees in the 2015 Nepal earthquake.

In data science, attempts have also been made to enrich data other than mobile network data. For instance, Antenucci *et al.* (2014) managed to provide real-time predictions for the unemployment rate from job-related signals collected on Twitter, using a *recursive linear model* based on the social media indexes as the regressors. The effectiveness of their model was reflected in its ability to indicate the shifting Beveridge Curve¹ for the study area. Other

¹ A Beveridge curve is a graphical representation of the link between unemployment and job vacancy rate (Dow and Dicks-mireaux, 1958).

similar studies are presented, for instance, in the papers by Magno and Weber (2014), Llorente *et al.* (2015), Proserpio *et al.* (2016), and Tuomchomtam and Soonthornphisaj (2019), showing a plethora of research exploiting social media data from, e.g. Twitter, Facebook or web search, for mining the socio-demographic information of users.

Moreover, Gebru *et al.* (2017) proposed using Google Street View images to estimate the demographic makeup of the United States (U.S.) to a single-precinct resolution in terms of the income, race, education level and voting patterns. In particular, they demonstrated a deep-learning-based vision technique using convolutional neural networks to identify vehicles and recognise the features including the make, model and year. This information, coupled with additional aggregate-level attributes such as the number of cars per image and average car price, were deployed to estimate the demographic makeup of the country using a number of regression models. These include a *logistic regression model* for race and education, and a *ridge regression model* for income and voting preference, which are trained using U.S. Census and Presidential Election voting data. Although the high household car ownership rate in the U.S. assures the high accuracy (>80%) of their model, this in turn limited the applicability of their methods in areas where the penetration of auto vehicles is low.

Unlike the relatively large number of studies conducted for socio-demographic enrichment in data science, the investigations made in the context of transport modelling are rather limited. This can possibly be attributed to the fact that travel behaviour patterns (e.g. modes and trip purposes) are of primary interest of transport modellers and practitioners. However, as the socio-demographic attributes typically serve as the essential exogenous determinants of travel-related behaviours, the enrichment of such contextual information has also gained increasing attention in transport modelling and travel behaviour analysis.

As one of the earliest explorations in travel behaviour analysis and transport modelling, Auld *et al.* (2015) attempted to infer a number of individual- and household-level attributes from travel behaviour patterns using a variety of machine-learning-based and statistical-oriented choice modelling procedures. These methods consist of a *nested logit model* for work and education status, an ordinal logit model for age, a *binary logit model* for gender, a *partial decision tree* for licence possession, and a *C4.5 decision tree* for household-level information. Their methods yielded mixed results in respect to different types of socio-demographic data, however: specifically, work and education status, age, licence possession and presence of children were enriched with significant improvements (>60%), while the models for gender,

household size and number of vehicles resulted in only minor improvements (<15%). This indicates that these data-driven models cannot explain the reasons underlying such performance discrepancies, hence reducing the generalisability of these models. Auld *et al.*'s study, however, provides some key insights for the investigations with similar aims, which includes the current study. These insights encompass:

- 1) Multi-day data collection is necessary to average out the intra-person day-to-day variability that may confound the prediction for the socio-demographic attributes.
- 2) The enrichment performance can possibly be improved if individuals can be related to each other within their households.
- 3) Joint modelling of several attributes is challenging but helpful in improving model performance while reducing inconsistencies.

Then again, Blumenstock *et al.* (2010) proposed a *logistic regression model* to predict the demographic attributes based on CDR usage patterns, coupled with structured follow-up phone interviews. The best performing model in their work included five important CDR features and was able to enrich the gender attribute with an accuracy of 74%, which was found to be comparable to the results obtained from using Naive Bayes classifier and SVMs.

Similarly, Bwambale *et al.* (2019) proposed a hybrid trip generation model which estimated the demographic attributes in the first step as the latent variables for the use in subsequent steps. In particular, the demographic prediction model classifies the travellers into demographic groups based on phone usage variables such as the number of calls that are extracted from mobile network data as the regressors. A *multinomial logistic regression model* is applied to estimate the probability of a user falling into a specific demographic group defined by combinations of categories of gender, age group and work status. Their model also reached an accuracy of 74% in demographic classification.

Regardless of the relatively successful prediction results, the methods developed/used in the above studies typically model the absent socio-demographic attributes (response) as functions of the observed behaviour patterns as the predictor, which, as mentioned previously in Section 1.2, fails to account for the microeconomic behavioural principles that underlies the mutual dependence between the two. Serving as one of the questions addressed in this PhD study, the impact of this is discussed below in Section 2.1.7.

2.1.7 Summary of previous semantic enrichment studies and research gap in socio-demographic enrichment

This section summarises the previous attempts to improve the semantic richness of big data, which have been discussed extensively in Sections 2.1.2 – 2.1.6, for the enrichment of behaviour patterns and socio-demographic information, separately.

Specifically, Table 2.1 presents a summary of existing investigations into the development, or applications, of approaches to enrich behaviour pattern variables (mainly travel-related) in terms of the enriched variable, the type of transport big data source and/or the employed (or developed) methodology and the accuracy level (if available).

Table 2.1 Summary of Existing Studies for Enrichment of Behaviour Patterns

Enriched Variables	Data source and/or Methodology	Accuracy	Reference
OD matrices	Scanned plate data	—	Castillo <i>et al.</i> (2008), Asakura <i>et al.</i> (2000), Rao <i>et al.</i> (2018), and Liu <i>et al.</i> (2020)
	GPS data	—	Herrera <i>et al.</i> (2010), Parry & Hazelton (2012), Morimura & Kato (2012), Sbaï <i>et al.</i> (2017), Ge & Fukuda (2016), and Moreira-Matias <i>et al.</i> (2016)
	Smartcard data	—	Zhang <i>et al.</i> (2007), Zhao <i>et al.</i> (2007), and Alsgar <i>et al.</i> (2016)
	GPS data + Smartcard data	>80%	Munizaga <i>et al.</i> (2012)
	Mobile network CDR data	—	Iqbal <i>et al.</i> (2014), Bonnel <i>et al.</i> (2015), Alexander <i>et al.</i> (2015), and Bachir <i>et al.</i> (2019)
Types of location	Density-based Clustering	—	Ester <i>et al.</i> (1996)
	Event-based conceptual model	—	Andrienko <i>et al.</i> (2011; 2011)
	Density-based clustering + SVM	96%	Gong <i>et al.</i> (2015)
	GPS data + Temperature data Rule-based model	84% ~ 87%	Lee <i>et al.</i> (2017)
	Mobile network CDR data Clustering + logistic regression	—	Isaacman <i>et al.</i> (2011)

Trip purposes & types of activities	Time-and-grid-based clustering	63%	Montoliu <i>et al.</i> (2013)
	Density-based data mining algorithm	—	Giannotti <i>et al.</i> (2007)
	Mobile network CDR data Experimental study	—	Calabrese <i>et al.</i> (2010)
	GPS data Rule-based algorithm	Report quality	Wolf <i>et al.</i> (2001; 2004)
	Rule-based data mining heuristic	80%	Ectors <i>et al.</i> (2017)
	GPS data Probabilistic model	60% ~ 95%	Chen <i>et al.</i> (2010)
	GPS data + GIS land use data Conditional random fields	80%	Liao <i>et al.</i> (2007)
	Decision tree learning (C5.0)	88%	Deng & Ji (2010)
	Interactive passive online survey Machine learning methods	—	Seo <i>et al.</i> (2019)
	Smartcard data + travel survey Bayesian probabilistic model	88%	Kusakabe <i>et al.</i> (2014; 2016)
	Smartcard data + GIS land use data Continuous hidden Markov unsupervised machine learning model	—	Han & Sohn (2016)
	Dictionary learning-based algorithm	95%	Li <i>et al.</i> (2017)
Modes of transport	GPS data Multi-layer neural network perceptron	98%	Byon <i>et al.</i> (2007)
	GPS data + GIS land use data Rule-based elimination algorithm	92% ~ 95%	Chung & Shalaby (2005), Stopher <i>et al.</i> (2008), Bohte & Maat (2009)
	GPS data + accelerometer data Bayesian Belief Network	—	Feng & Timmermans (2013)
	GPS data Feed forward multilayer perceptron	92%	Brunauer <i>et al.</i> (2013)
	GPS data Convolutional Neural Network	85%	Dabiri & Heaslip (2018)
	GPS data Random forest + rule-based algorithm	>85%	Wang <i>et al.</i> (2018)

On the other hand, existing attempts for socio-demographic enrichment of various big data sources are summarised in Table 2.2 (below) with respect to the methodology, data source and the level of accuracy (if available).

Table 2.2 Summary of Existing Attempts for Socio-demographic Enrichment

Methodology	Data Source	Accuracy	Reference
Semi-supervised <i>k</i> -means clustering	CDRs	80%	Frias-Martinez <i>et al.</i> (2010)
Machine learning algorithms: Linear Discriminant Analysis, SVMs, Multilayer Perceptrons, and decision trees	CDRs	65%	Herrera-Yagüe & Zufiria (2012) Sarraute <i>et al.</i> (2014) Dong <i>et al.</i> (2014) Jahani <i>et al.</i> (2017)
Basic machine learning algorithm for predictions from features extracted using convolutional network deep learning	CDRs	60% ~ 80%	Felbo <i>et al.</i> (2017)
Machine learning algorithms	Social media	—	Antenucci <i>et al.</i> (2014) Magno and Weber (2014), Llorente <i>et al.</i> (2015), Proserpio <i>et al.</i> (2016), and Tuomchomtam & Soonthornphisaj (2019)
Deep-learning-based vision technique	Google street view	>80%	Gebbru <i>et al.</i> (2017)
Ad-hoc machine learning algorithms and statistical regression model	GPS	>60%	Auld <i>et al.</i> (2015)
Logistic regression model	CDRs	74%	Blumenstock <i>et al.</i> (2010)
Latent DCM	CDRs	74%	Bwambale <i>et al.</i> (2019)

Compared to the relative success in the extraction of various travel-related behaviour patterns, it is evident from the above tables that the enrichment of socio-demographic attributes for big data still remains underexploited. Additionally, from the up-to-date socio-demographic enrichment studies (Section 2.1.6 and Table 2.2), it is also noted that the approaches developed, or applied, in existing socio-demographic enrichment studies use either performance-based data mining and machine learning methods or statistical-fit-oriented models. Whereas these methods may be able to make reasonably successful predictions, they typically lack underpinning behavioural theories that can support or the postulated relationship (models) between the input behavioural patterns (features) and the output enriched attributes. Particularly, existing socio-demographic enrichment methods typically treat the socio-demographic attributes as the response variables that can vary with respect to the observed behavioural patterns as the predictors. Although this seemingly addresses the enrichment problem as formulated, they fail to account for the logic suggested by microeconomic theory,

as well as the extensive empirical research concerning people's behaviour, which states that people behave because of the type of person that they are but not *vice versa* (Mas-Colell, Whinston and Green, 1995; Lancaster, 2008). This means that, although the observed behaviour patterns of anonymous respondents may, to some extent, reflect their socio-demographic attributes due to the causality between the two, the change in their behaviours may not reflect the variation in the socio-demographic attributes because it can be caused by other factors, such as the change in the corresponding choice-making context. For example, the car ownership status of a choice-maker can influence their choice of travel modes yet the change in mode choice behaviour may not necessarily be caused by the change in the car ownership status. Instead, it may result from the implications of policy implementation, such as the congestion charge in London (Leape, 2006). Therefore, inferring car ownership status as a function of mode choice only is inaccurate. The reverse is also true, i.e. that the change in mode use preference may also be due to factors other than the status of car ownership. However, the relative positive or negative preference for mode choice due to car ownership is substantially less likely to vary due to either the change in the mode choice or the variation in the status of car ownership. This makes using an econometric model with a theoretical basis important in addressing the issue of predictable and interpretable socio-demographic enrichment and also in a more structured manner. A detailed explanation of the dependent relationship between people's behaviour pattern and the various explanatory variables (e.g. attributes of choice alternatives and the socio-demographic characteristics of respondents) is provided later in Section 3.1. Such a relationship provides an additional principle to be followed by any developed socio-demographic enrichment approaches, compared to the enrichment of other variables; namely that the dependence between the input behaviour patterns and the output enriched attributes should be retained when making predictions.

A most obvious consequence of not explicitly addressing the above dependent relationship, as reflected in existing socio-demographic enrichment methods, is their inability to understand the fundamental properties of the methods themselves, such as the enrichment performance under different data conditions (quality, variable types, sample sizes, correlation structure and contexts). However, the ability to predict and interpret the enrichment performance, in particular, is crucial to ascertain the generalisability and transferability of an enrichment approach under different data conditions or across application contexts. For instance, the ability to interpret the performance of an enrichment method is necessary in understanding the implications of policy interventions which are expected to induce changes in the population, in

terms of either composition or behavioural patterns. These changes in the underlying population can possibly render the enrichment performance of an existing, calibrated enrichment model questionable for the required post-intervention context unless theoretical interpretation can enable the assessment of the corresponding changes in enrichment performance (Pawlak, Zolfaghari and Polak, 2015). In other words, without known behavioural underpinnings to understand the performance of a socio-demographic enrichment approach, there is no obvious way to make assumptions or adjustments to the corresponding enrichment, hence restricting the transferability of the approach. This echoes a widely acknowledged remark from Moshe Ben-Akiva (1974) who maintained that a model that can duplicate the data perfectly might serve no useful purpose for prediction if it represents erroneous behavioural assumptions. Brathwaite *et al.* (2017) even argued that one reason why the prevailing machine learning methods have not made greater inroads amongst discrete choice modellers is that those models have not been linked to the economic theories underlying human decision-making. Yet to date, as is shown in the overview of the literature, the prevalent focus has been on maximising the predictive accuracy, without explicitly relating to the corresponding, underlying behavioural theory.

The above discussion provides the rationale for this study to address the aforementioned research gaps through the systematic development of an IDCM framework for the predictable and interpretable enrichment of socio-demographic attributes of anonymous big data.

2.2 Transport big data sources for semantic enrichment

In addition to classifying previous semantic enrichment attempts based on the enriched variable (as demonstrated in Sections 2.1.1 – 2.1.6), these existing investigations can also be categorised according either to the source of big data or to the method developed/applied for the enrichment. This section thus summarises the types of transport big data exploited in previous semantic enrichment studies while the following section discusses the methodologies.

Although lacking a unified definition, the term ‘big data’ usually refers to the datasets containing more data than common software tools are able to capture, curate, manage or process within a reasonable elapsed time (Snijders, Matzat and Reips, 2013). Big data is characterised and differentiated from traditional datasets, therefore, through three key concepts:

volume, variety and velocity. In particular, ‘volume’ implies that the size of big data is considerably large, i.e. usually larger than terabytes and petabytes as defined by (Sagiroglu and Sinanc, 2013). The variety and velocity of big data, meanwhile, suggest that the data can retain a variety of forms and can be generated at substantially fast speed (Kitchin and McArdle, 2016). Whereas other additional characteristics, e.g. veracity, have also been proposed to describe big data (Onay and Öztürk, 2018), exploring the common properties of big data is not the primary focus of the current study and is hence not explained any further in this thesis. It should also be noted that, while the term ‘big data’ can sometimes be used to define the research field in which advanced data analytics methods are employed to extract information from or otherwise deal with the big datasets that are too large or complex to be handled by traditional methods or software tools, in the present thesis, ‘*big data*’ refers solely to the datasets rather than the research field or the corresponding technologies.

As mentioned in Chapter 1, the essential shortcoming associated with big datasets is their semantic poverty, i.e. that they contain few variables that can characterise or interpret the voluminous data points. Unlike traditional data sources which are generated for specific research questions (e.g. travel surveys), most passive big datasets are collected as by-products of emerging digital technologies such as the information and communication technologies. This, along with the aforementioned three beneficial properties of big data, has spurred a plethora of investigations into extracting additional values from such data to improve its semantic richness, with one particularly prevalent strand of studies being committed to gain insights revealed in transport big data. This section thus discusses the sources of transport big data that are exploited to understand people’s travel-related behaviour patterns and/or information about the respondents as well as their households. Specifically, the following sections outline three example sources of transport big data that have been extensively and beneficially employed to gain insights into people’s travel behaviour patterns and their socio-demographic information, as well as a summary of the rest types of transport big data that are also used in existing enrichment attempts.

2.2.1 Global Navigation Satellite System (GNSS) data

GNSS is the standard generic term for satellite navigation systems that provide autonomous geo-spatial positioning with global coverage. Common GNSS Systems are GPS, GLONASS, Galileo, Beidou and other regional systems (‘GNSS-Global Navigation Satellite Systems: GPS,

GLONASS, Galileo, and more', 2008). In the contents below, GPS data is discussed as the representative of the big data collected from GNSS technologies.

As the world's most utilised navigation system, the Global Positioning System, or GPS, is a navigation system consisting of three fundamental components, called the segments, which work together to synchronise location and time information (Awange, 2012). In particular, the space segment involves the satellites circling the Earth to transmit signals of geographical positions and time of day to the users. Made up of control and monitoring stations, as well as ground antenna, the control segment performs control activities to ensure that the satellites operate properly. The user segment is composed of GPS receivers and transmitters with a variety of formats, such as those embedded in cars, phones and watches.

The essential uses of GPS include determining positions (location), getting from one location to another (navigation), monitoring movements (tracking), creating maps (mapping), and taking precise time measurements (timing) (Awange, 2012). Given the widespread use of GPS for the above purposes, GPS devices today can generate an enormous amount of data related to the timestamped geolocations as by-products of these functions. This provides the opportunity to gain deeper insights into people's mobility patterns by developing and applying suitable data processing techniques (Giannotti and Nanni, 2007). Specifically, as documented in detail in Section 2.1, GPS data has been applied in an extensive number of studies seeking to extract information about movement patterns such as trip purposes and travel modes.

2.2.2 Smartcard data

Smartcards represent a type of physical electronic authorisation device, utilised to control the access to a resource by means either of a pattern of metal contacts to electrically connect to the internal chips or via radio-based (contactless) means (Hendry, 2007).

The applications of smartcards are various, but those associated with the extraction of travel behaviour patterns involve the automated fare collection systems that employ smartcards. In particular, smartcards are used by many public transit operators as transit passes or integrated ticketing. Examples include Singapore's CEPAS, Hong Kong's Octopus card and London's Oyster card. Since smartcards can provide personal identification information, their use for public transport or shared mobility can enable the passive collection of information concerning the ODs of trips. With additional GIS land use data and socio-demographic information, the travel patterns such as the purpose of trips can be obtained from smartcard data (Han and Sohn,

2016). A number of applications that increases the semantic richness of smartcard data have been discussed extensively in Section 2.1.

2.2.3 Mobile network data

Referring to the sets of data generated and collected by mobile network providers, mobile network data informs timestamped locations that can be generated by the process of a mobile phone communicating with the cellular network maintained by a mobile network provider. This is termed as cellular network-based positioning. Two types of mobile network data can be generated through the process of positioning, viz. CDR data and sightings data. CDR data represents the more preferred and widely employed type of data for semantic enrichment (Horak, 2006). Particularly, a CDR refers to a data record representing a phone call or other telecommunications transaction (e.g. text messages) with details about the caller, callee, the starting time and duration of the call, as well as the geographical coordinates of the tower that first detected the call (Horak, 2006).

In addition to revealing the spatial-temporal information, which can be processed to obtain the mobility behaviour patterns, the other reason for the wide application of CDR data in semantic enrichment is that it can possibly provide information regarding the non-mobility (social) behaviour patterns of mobile phone users, such as the number of input/output calls. Information about both the mobility and non-mobility behaviour patterns that can be observed or extracted from mobile network CDR data allows the application of such data as one of the major big data sources for semantic enrichment. Section 2.1 has thoroughly outlined example studies that seek to exploit the semantic richness of mobile network data. Moreover, the sufficient information regarding the behaviour patterns of mobile phone users also indicates that CDR data has the potential to act as the data source for the proposed IDCM approach, given that it reveals how the socio-demographic attributes of mobile phone users can affect their social and movement behaviour patterns (Frias-Martinez, Frias-Martinez and Oliver, 2010).

2.2.4 Other sources for semantic enrichment

Alongside the major sources for semantic enrichment of transport big data as discussed above, a number of other types of big datasets can be used to infer information about people as well as their mobility and social behaviour patterns.

For instance, traffic monitoring systems, such as automatic number plate recognition cameras and inductive loop detectors, can collect information about ODs and routes of vehicles (Liu *et al.*, 2020). Socio-demographic information and social behaviour patterns can also be extracted from social networks (e.g. Twitter data), Passenger Name Record data in the airline industry and Google Street View data (Antenucci *et al.*, 2014; Mottini and Acuna-Agost, 2016; Gebru *et al.*, 2017). Section 2.1 has also provided a detailed overview of these applications.

2.3 Machine learning

As mentioned at the beginning of Section 2.2, the method applied to improve the semantic richness of various big data sources represents another classification of existing semantic enrichment attempts. A detailed inspection of Table 2.1 and Table 2.2 (Section 2.1.7) suggests that a large majority of existing attempts for the semantic enrichment of big data sources employ, or at least involve the use of, machine learning methods. Moreover, machine learning algorithms play the most significant role as the applied methodology in the enrichment of socio-demographic information (Table 2.2). The research field of machine learning is thus explained in the current section. This is to provide insights into existing methods employed for socio-demographic enrichment, as well as to set the scene for the comparative analysis against the proposed IDCM framework (Chapters 5 and 6). Note that the current section presents only an overview of machine learning. More detailed review could be obtained by referring to the book by Mitchell (2006).

Seen as a subfield of artificial intelligence, machine learning can be most appropriately defined by the central problem it solves: the study of computer algorithms that automatically improve through experience and the fundamental laws that govern all learning processes (Mitchell, 2006). Typical machine learning algorithms establish mathematical models for making predictions or imputations without being explicitly programmed about how to do what they are supposed to do (Samuel, 2000).

Specifically, Section 2.3.1 demonstrates how machine learning algorithms are classified according to different principles. Then in Section 2.3.2, logistic regression and SVMs are explained as representatives of machine learning algorithms against which the IDCM framework is compared (given their use in previous socio-demographic enrichment attempts and given that their data requirements are similar to those of the IDCM framework). Section

2.3.3 intends to present a clearer view of the field of machine learning by conceptually comparing machine learning algorithms with traditional statistical modelling methods.

2.3.1 Typology of machine learning approaches

Machine learning algorithms can usually be classified according to either the characteristics of the data available to the learning system or the functional form of the learning algorithm (Mitchell, 2006; Alpaydin, 2014). Specifically, common classifications depending on the data availability categorise machine learning algorithms into three categories: supervised learning, unsupervised learning and reinforcement learning. On the other hand, different functional forms divide machine learning methods into parametric and nonparametric algorithms. Each of the above two classification systems is described below. In terms of the three categories based on the availability of data:

1) Supervised learning

Supervised learning refers to the branch of learning algorithms that produce an inferred function to map the input-and-output pairs by analysing labelled examples in a known dataset (training data). The produced function can hence be used to make predictions about unknown events (Russell and Norvig, 2010). An ideal supervised learning algorithm needs to generalise from the labelled training examples to unseen situations in a reasonable way (Hill, 1997). A typical supervised learning process consists of: (i) determining the type of training examples and corresponding data collection; (ii) the selection of the input features; and (iii) the structure of the learning algorithm; followed by (iv) the learning process over the collected training examples; and (v) hence evaluation of the learning performance.

According to the learning form and the type of output variables, types of supervised learning algorithms generally include active learning, regression and classification (Alpaydin, 2014). Specifically, active learning designates the learning algorithms that can communicate interactively with the user to label new data with more desired outputs, while regression and classification refer to algorithms used for predicting continuous and discrete output variables, respectively.

Examples of supervised learning algorithms include linear and logistic regression as well as support vector machines.

2) Unsupervised learning:

Unsupervised learning specifies those learning systems that infer functions to discover and describe hidden structures or patterns in unlabelled data, which itself can represent both the goal and the method. Instead of using human-labelled data, unsupervised learning approaches identify shared characteristics in the data and respond given the presence of the commonalities detected in new instances. This property makes it possible to model the probability densities over the input, underpinning a central application of unsupervised learning in the field of density estimation ('Unsupervised learning: Foundations of neural computation', 1999). Unlike supervised learning, there is no validation process in unsupervised learning as it is impossible to evaluate the correctness of outputs or labels.

Typical unsupervised learning approaches comprise clustering, neural networks and models including latent variables (e.g. expectation-maximisation algorithms). Principal component analysis (PCA) and cluster analysis are two major approaches used for unsupervised learning. In particular, PCA involves developing an orthogonal basis with a number of principal components, in which different dimensions of the data are correlated, and proceeding from finding the vector that minimise the average squared distance from a given set of data points so as iteratively to choose the next vector from the directions perpendicular to the previous one (Abdi and Williams, 2010). Cluster analysis, or clustering, meanwhile, performs segmentation tasks for the data according to their commonalities, such that the data in the same class are similar while data from different classes can be distinguished from one another based on the segmentation attributes (Everitt, 1980).

3) Reinforcement learning:

Reinforcement learning refers to those learning methods by which software agents maximise their performance in an environment where they are required to take actions. The performance is measured using the notion of cumulative reward, determined according to the scalar reinforcement signals (errors and rewards) obtained as feedback from the interactions with the environment (Bishop, 2006). The environment of reinforcement learning is usually stated as a Markov Decision Process (MDP) because most reinforcement learning algorithms use dynamic programming techniques (Van Otterlo and Wiering, 2012). It differs from conventional dynamic programming, however, in eliminating the requirement for an exact mathematical MDP model to accommodate large-scale situations where exact models are infeasible. Without a need for training data, the core goal of

reinforcement learning lies in finding a balance between exploration (of uncharted territory) and exploitation (of current knowledge) (Kaelbling, Littman and Moore, 1996).

MDP, Monte Carlo, and Q-learning are typical reinforcement learning models that are widely applied in applications such as robotics and business strategy planning (Van Otterlo and Wiering, 2012).

In more recent investigations, algorithms developed in the field of machine learning show a trend to fall simultaneously into several of the three categories above. Examples of this include (i) semi-supervised learning, for training examples with missing labels (Sinha, 2014); (ii) feature learning, which transforms inputs into useful features while retaining the level of information (Bengio, Courville and Vincent, 2013); and (iii) self-learning, where no external rewards or supervision is available (Bozinovski, 1982), etc. Moreover, it is argued that deep learning, as part of a broader family of machine learning methods based on artificial neural networks, has become the dominant approach for much ongoing work in the field of machine learning (Alpaydin, 2014). Deep learning, however, is beyond the scope of the current study, on account of the so-far limited inroads it has made into socio-demographic enrichment, as demonstrated in Section 2.1. Hence, it is not discussed any further in this thesis.

In the context of the different classes of machine learning algorithms discussed above, the IDCM framework developed in the present work is compared against supervised machine learning methods given the requirement of an estimation sample. This estimation sample is analogical to the training sample exploited in supervised learning. A detailed explanation of the estimation sample required for the IDCM framework, and its comparison to the training sample for supervised learning, is provided later in Section 4.1.2 (Chapter 4).

The other way of classifying machine learning algorithms is based on the functional form of the algorithm. Specifically, different machine learning algorithms can take different functional forms and thus employ different means of learning, which can significantly simplify the learning process at the cost of limiting what can be learned. This typology based on the functional form thus categorises machine learning methods into parametric and nonparametric algorithms. The difference between these two categories involves, as suggested by the names, whether the function takes a known form parameterised by some fixed, finite number of parameters (parametric) or not (nonparametric) (Wasserman, 2004).

1) Parametric learning algorithms:

Parametric machine learning typically comprise two stages: (i) selecting the functional form and (ii) estimating the coefficients (Russell and Norvig, 2010). Example parametric learning algorithms include linear regression, logistic regression and perceptron.

The primary benefit of parametric algorithms is their easier interpretability. Interpretability is defined, in mathematical logic, as a relation that conveys the possibility of explaining or translating one formal theory into another (Goodstein *et al.*, 1954). In machine learning, interpretability concerns the degree to which the model results can be consistently predicted (Kim, Khanna and Koyejo, 2016), or the degree to which the cause of the model results can be comprehended (Miller, 2019). Additionally, parametric learning is often speedy and requires less data for training.

Parametric learning is highly constrained to the employed functional form, however, and thus may be unable to handle complex problems (Russell and Norvig, 2010).

2) Nonparametric learning algorithms:

In contrast, nonparametric learning algorithms do not make explicit assumptions about the form of the mapping function (Russell and Norvig, 2010). By not making assumptions, they are free to learn any functional form from the training data. Nonparametric methods maintain the ability to generalise to unseen data while pursuing fitting performance. Decision trees and SVMs are two most popular nonparametric learning approaches.

Typical advantages of nonparametric learning algorithms include the flexibility to fit a large number of functional forms, and the relatively high prediction performance. Nonparametric learning algorithms usually require a large training sample, however, and more time to estimate the mapping function. Otherwise, they are more prone to overfitting (Russell and Norvig, 2010). Overfitting occurs when the algorithm models the training data too well, i.e. the details and noise in the training data are captured to a degree that negatively affects the performance when modelling new instances (Hawkins, 2004).

Figure 2.4 (below) conceptualises the relationship between the above two classification systems for machine learning algorithms. In particular, the different shades represent the three classes defined based on the availability of data while either kind of machine learning algorithms classified according to the functional form (viz. parametric and nonparametric algorithms) can be used for any of the three classes of learning.

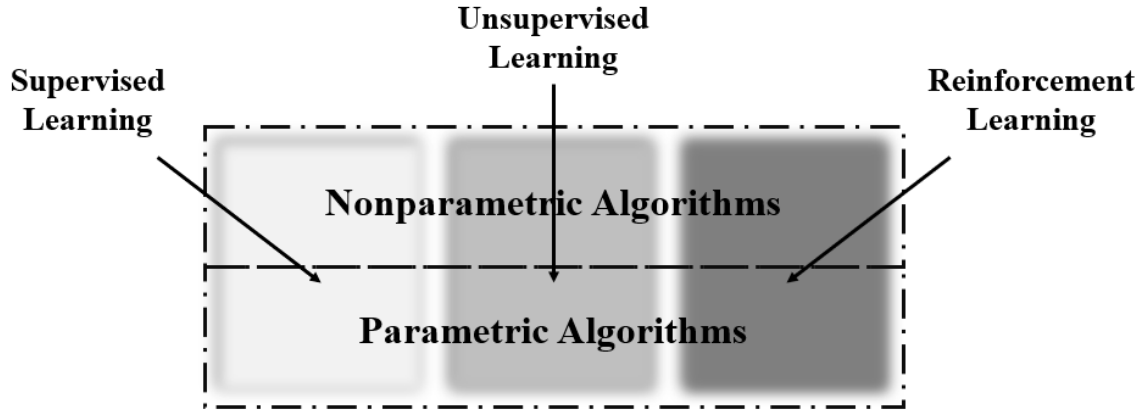


Figure 2.4 Conceptual Classification of Machine Learning Algorithms

2.3.2 Representative supervised learning algorithms for enrichment of big data

As mentioned previously in Section 2.3.1, the proposed IDCM socio-demographic enrichment framework is similar to supervised machine learning methods in terms of data requirements. Therefore, logistic regression and SVMs are outlined in Sections 2.3.2.1 and 2.3.2.2, respectively, as representatives of parametric and nonparametric supervised learning algorithms, in terms of their modelling structure and the learning process. This is particularly important because, later, in Chapters 4 – 6, the IDCM framework is compared theoretically and empirically against these representative supervised machine learning approaches in order to assess the merits of the developed IDCM framework.

2.3.2.1 Logistic regression

As an important model originating in statistical modelling, logistic regression refers to the statistical process of estimating the relationship between a binary dependent variable, which represents the occurrence of a certain event such as win/lose given a set of predictors. Specifically, this relationship is captured, in its basic form, by a logistic function which models the probability of the binary dependent variable. If the two possible outcome values (POVs) of the binary variable modelled are labelled as 0 and 1, and the probability of the random variable labelled as 1 is defined as p . Such a logistic function can be expressed as:

$$p = f(x) = \frac{1}{1 + b^{x\beta}}; \quad (2-1)$$

where b is the base of the logarithm;

X is the set of predictors affecting the probability;

β is the vector of parameters associating X with p which needs estimation.

Logistic regression therefore refers to the estimation of the parameters β of the logistic function (Cramer, 2005). In logistic regression, the log-odds (the logarithm of the odds) that the POV is labelled as 1 concerns a linear combination of one or more predictors which are either categorical or continuous:

$$\text{logit}(p) = \log_b \frac{p}{1-p} = X\beta. \quad (2-2)$$

The function that converts this log-odds to probability, i.e. Equation 2-1, is hence termed as the logistic function, which is what lends this method the name of logistic regression.

A logistic model itself is not a classifier, i.e. it does not perform statistical classification. The output probability produced by logistic regression can be used to make a classifier, however. The common practice for generating such a binary classifier involves selecting a threshold value which divides the probabilities greater and lower than it into two classes.

In light of the estimated probability of the output variable being labelled 0 or 1 given experimental data, the binary classification based on logistic regression represents an important supervised machine learning algorithm (Ng, 2000). Moreover, logistic regression can be generalised to model variables with more than two discrete outcome values. In statistics and machine learning, the method that generalises logistic regression to solve multiclass problems is referred to as multinomial logistic regression (Greene, 2012), although it is also known by some other names, such as SoftMax regression (Bridle, 1990) and maximum entropy classification (Malouf, 2002).

By definition, machine learning algorithms developed from the standard logistic regression correspond to parametric learning since such a logistic classifier has a fixed, finite set of parameters. It is also possible to perform nonparametric logistic regression, however. In particular, relaxation can be made in terms of the form of the log-odds, which is commonly linear, so as to model the nonparametric effect of the predictors (Hastie and Tibshirani, 1987). This might involve the use of smoothers such as smoothing splines. Conversely, the estimated nonparametric functions can suggest suitable parametric transformations for the covariates.

In this thesis, logistic regression is selected as the representative of supervised parametric learning methods in comparison against the IDCM framework since the logistic classifier has

been used to infer the socio-demographic attributes of anonymous big data sources previously in the context of travel behaviour modelling. A more thorough discussion of this has been provided in Section 2.1.6. Section 5.2.3 (Chapter 5), as well as Sections 5.4 and 6.3.3 (Chapters 5 – 6), will present the detailed theoretical and empirical comparisons of the IDCM framework against logistic regression, respectively.

2.3.2.2 Support vector machines (SVMs)

Support vector machines, also known as SVMs, are popular nonparametric supervised learning models with associated learning algorithms for classification and detection of outliers (Cortes and Vapnik, 1995). Specifically, an SVM algorithm is a trained non-probabilistic binary linear classifier which builds one, or a set of, hyperplanes in a high- or infinite-dimensional space to separate data points, each viewed as a n -dimensional vector with a $(n - 1)$ dimensional hyperplane. There can be more than one hyperplane that classifies the data, amongst which the optimal is achieved when the functional margin, viz. the distance from the hyperplane to the nearest data in any class, is maximised. If existing, such a hyperplane is known as the maximum-margin hyperplane. The hyperplanes defined as the linear combinations of the features and the parameters are termed as linear SVMs, which can be further classified into hard-margin and soft-margin classifiers. Given a set of n data points $(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)$ where y_i is either -1 or 1, each indicating a class for $\vec{x}_i$. Any hyperplane can be written as the set of $\vec{x}$ satisfying (Cortes and Vapnik, 1995):

$$\vec{w} \cdot \vec{x} - b = 0, \quad (2-3)$$

where $\vec{w}$ is the normal vector to the hyperplane.

b is the bias term.

Two parallel hyperplanes can be selected to separate the two classes of the data when the training data is linearly classifiable such that the distance reaches its maximum. For a normalised dataset, these selected hyperplanes can be defined as:

$$\begin{aligned} \vec{w} \cdot \vec{x} - b &= 1; \text{ and} \\ \vec{w} \cdot \vec{x} - b &= -1 \end{aligned} \quad (2-4)$$

The geometrical distance between the two classes is $\frac{2}{\|\vec{w}\|}$. Hence, the SVM seeking to minimise $\|\vec{w}\|$ (magnitude of the vector) for a maximum margin is defined as a hard-margin

SVM. Alternatively, the SVM soft-margin classifier is extended to accommodate problems in which the training data is not linearly separable, resulting in the hinge loss function:

$$\max(0, 1 - y_i(\vec{w} \cdot \vec{x}_i - b)), \quad (2-5)$$

where y_i is the i th target variable;
 $\vec{w} \cdot \vec{x}_i - b$ is the i th output.

The corresponding problem to be solved is thus effectively to minimise:

$$\frac{1}{2} \|\vec{w}\|^2 + c \left[\frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i(\vec{w} \cdot \vec{x}_i - b)) \right], \quad (2-6)$$

where c is the trade-off between increasing the margin size and ensuring that $\vec{x}_i$ lies on the correct side of the margin.

The hard-margin classifier can be viewed as a special case of the soft-margin when c is sufficiently small and the second term is thus neglectable.

In addition to linear classification, SVMs can efficiently perform nonlinear classification using a so-called ‘kernel trick’, which implicitly maps the inputs into high-dimensional feature spaces. In particular, a kernel trick refers to a set of instance-based learners which sums the weighted similarities of unlabelled inputs and training inputs quantified by a similarity function (Theodoridis and Koutroumbas, 2009). In the nonlinear classification of SVMs, the linear hyperplane discussed above is replaced by a similarity function (or a kernel) to form the nonlinear maximum-margin hyperplane in a transformed feature space. As the concept of an IDCM kernel is proposed in the current thesis (Chapter 4), the term ‘kernel’ is used to refer particularly to the IDCM kernel while the kernel of a SVM is referred as a ‘similarity function’ only, to distinguish between the two.

Classic SVMs can be either extended to solve multiclass classification or used for regression analysis and unsupervised learning, which respectively lead to multiclass SVMs, support vector regression and support vector clustering.

SVMs are widely employed in many real-world applications, such as classification of images, hand-written characters, and even of proteins in biological sciences (Cuingnet *et al.*, 2011). As documented in Section 2.1, SVMs are also extensively applied in the attempts to enrich various types of big data, including the socio-demographic attributes which serve as the enrichment

purpose of the proposed IDCM framework. For this reason, the SVM is chosen as the representative of nonparametric learning algorithms for the comparison against the IDCM framework in Section 5.2.3 (theoretical), and Sections 5.4 and 6.3.3 (empirical), respectively.

2.3.3 Machine learning versus statistical modelling

Given the shared objective to extract patterns by drawing knowledge or insights into the data, machine learning is often compared with the conventional research field of statistical modelling (Srivastava, 2015; Mendis, 2019; Stewart, 2019). The following discussion therefore explores the commonalities and differences between machine learning and statistical modelling. This is to facilitate the understanding with respect to the property of machine learning methods, which correspondingly results in the research gap as explained in Section 2.1.7 in existing socio-demographic enrichment studies applying machine learning algorithms.

Machine learning can be summarised as the process of developing a function that maps the input to the output of the data while neither the form of the function nor the input data are necessarily known (Samuel, 2000). Statistical modelling, meanwhile, refers to the mathematical representation of a theory concerning the relationship between a set of random, independent variables and other non-random, dependent variables (Vapnik, 1999). Whilst sometimes referred to by different names, the concepts in the above two research fields can be mapped, as demonstrated in the work by Bzdok, Altman and Krzywinski (2018). For instance, the input variables, which in machine learning are often called the inputs or the features, are referred to as explanatory variables, regressors or covariates in statistical modelling. Likewise, the dependent variables, or the responses, in statistical modelling are referred to as the outputs or the labels/classes in machine learning. Moreover, data points in statistical modelling are defined in machine learning as instances or examples, while statistical regression or classification is referred to as supervised learning in machine learning. Table 2.3 summarises example concepts and the different names used to term them respectively in the two disciplines.

Table 2.3 Comparable Concepts in Machine Learning and Statistical Modelling

Concepts	Terminologies	
	Machine Learning	Statistical Modelling
Input data	• Inputs; • Features.	• Explanatory variables; • Regressors; • Predictors; • Covariates.

Output data	• Labels; • Classes.	• Dependent variables; • Responses.
Data record	• Instances; • Examples.	• Data points.
Modelling	• Learning.	• Estimation.
Models (example)	• Supervised learning	• Statistical regression. • Statistical classification.

Despite the above commonalities, statistical modelling and machine learning can be distinguished primarily in terms of their key purposes (Bzdok, Altman and Krzywinski, 2018). Specifically, statistical models tend to focus on explaining the underlying relationships between the input and the output variables, viz. the interpretability of the model, while machine learning approaches are designed to make the most accurate predictions possible (Rajula *et al.*, 2020). This suggests that while statistical models can be used to make predictions, the predictive accuracy (measured by the rate of correctly predicted instance) may not be their main strength (*Machine Learning Versus Statistics: When to use each - healthcare.ai*, 2019). In contrast, machine learning methods require no assumptions concerning the underlying distribution of the data, i.e. they tend to treat their algorithms as a black box² as long as they work. As a consequence, machine learning methods can provide various degrees of interpretability ranging from the highly interpretable lasso regression (Santosa and Symes, 1986) to impenetrable neural networks (Hopfield, 1982), while they generally trade off interpretability for predictive power (Stewart, 2019).

The two areas are also differentiated with respect to their data requirements: machine learning generally requires more data than statistical modelling. This is particularly the case since the inference power of machine learning algorithms relies extensively on the presence of a tremendous number of observations while statistical models could make strong predictive models with a small number of observations; even though machine learning can possibly make desirable predictions on little data, and statistical models can be problematic on large samples (*Machine Learning Versus Statistics: When to use each - healthcare.ai*, 2019).

Recently, there has been a trend for the two fields to learn from each other and to converge theoretically and empirically (Mendis, 2019). This is reflected, for instance, in the use of many

² In science, computing and engineering, a black box refers to a device, system or object where the internal working mechanisms are unknown in terms of the inputs and outputs (Nizami, 2019).

statistical models, e.g. logistic regression, in machine learning applications, thus maximising the fit for making predictions while being able to set aside typical statistical modelling considerations (Rymarczyk *et al.*, 2019). However, the above discussion indicates that the flexibility and less interpretability of machine learning algorithms can lead to their inability to estimate and interpret their model performance. This not only applies when compared against statistical modelling methods but is also particularly the case in comparison to econometric models. This shortcoming of machine learning algorithms can limit their transferability in the context of socio-demographic enrichment, which eventually promotes the development of the IDCM framework (an econometric model).

2.4 Summary

This chapter has reviewed and discussed existing literature, from different perspectives, related to the development of the Inverse Discrete Choice Modelling (IDCM) framework with the shared purpose of improving the semantic richness of various ‘big-but-thin’ data sources.

In particular, Section 2.1 explained the techniques that transform raw transport big data points into meaningful trips. This was followed by an overview of existing studies attempting to enrich various behavioural and socio-demographic variables as the semantic contents originally absent from big datasets. The reviewed literature further led to the discussion with regard to the research gaps reflected in the shortage of socio-demographic enrichment attempts and the absent of theoretical underpinning in existing studies.

Based on the studies documented in Section 2.1, Section 2.2, on the other hand, described the sources of big data, focusing on transport big data which are extensively explored to gain insights into the travel-related behaviour patterns. Section 2.3, meanwhile, outlined the methodology developed or applied to obtain the socio-demographic information in previous investigations. Specifically, the research field of machine learning was briefly introduced given that machine learning algorithms serve as the mainstream methodology applied in existing socio-demographic enrichment attempts. According to the two classification systems set out for machine learning algorithms, two representative methods were selected and explained, for the comparative analysis against the proposed IDCM approach. In addition, the properties of machine learning algorithms were discussed with a conceptual comparison to statistical modelling approaches, hence facilitating the understanding of the inability of existing socio-demographic enrichment methods to account for their enrichment performance.

In summary, the present review made it possible to identify the research gaps in the previous attempts at the socio-demographic enrichment of anonymous big data, and the lack of theoretical justification for why existing socio-demographic enrichment methods perform better or worse for particular datasets or using specific parameters. Identification of these research gaps provides the rationale for the systematic development of the IDCM framework in this study, and the associated focus on the predictable enrichment performance according to a theoretical underpinning in microeconomic theory. This in turn underpins the establishment of the IDCM framework, ranging from the basic conceptual and econometric formulation developed in this thesis (as derived and explained in Chapters 3 – 4) to more complex extensions as proposed in Chapter 8. In addition, the discussion with respect to the source of big data justifies the use of mobile network data for the validation of the proposed IDCM framework in Chapter 6, while the description of the representative machine learning methods permits the comparative analysis in the two empirical enrichment applications as documented in Chapters 5 – 6.

3 CONCEPTUAL FRAMEWORK OF IDCM METHODOLOGY

This chapter conceptualises the Inverse Discrete Choice Modelling (IDCM) framework from a number of perspectives, so as to enhance understanding of the systematic establishment of the IDCM socio-demographic enrichment approach with predictable and interpretable enrichment performance in the following chapter.

The current chapter is particularly structured as follows. Section 3.1 briefly outlines discrete choice modelling theory, which serves as the foundation of the proposed IDCM framework in terms of both the econometric formulation and the accompanying performance theory. Section 3.2 then presents a conceptual, microeconomic model which incorporates microeconomic behavioural theories to explain the logic underlying the IDCM approach. The inversion mechanism captured in this conceptual model leads to the discussion around the research field of inverse problem in Section 3.3, which demonstrates the challenges of and the potential solution techniques for the IDCM approach. This is followed by Section 3.4 which conceptualises the analogy of choice-making behaviour to the transmission of messages in an information communication systems. Such an analogy motivates the subsequent use of several concepts developed in information theory to establish metrics for the IDCM performance theory. The chapter is summarised in Section 3.5 which also links up to the following chapter.

3.1 Discrete Choice Models (DCMs)

Discrete choice modelling represents the theoretical foundation for developing the econometric formulation of IDCM approach and the accompanying IDCM performance theory. This section hence briefly summarises the fundamentals of discrete choice models (DCMs) such that the establishment of the IDCM framework in Chapter 4 can be followed smoothly.

In particular, Section 3.1.1 presents the econometric framework underlying DCMs. Section 3.1.2 then outlines the components and structures of typical DCMs based on the underlying econometric framework. These jointly serve as the basic formulation upon which the current IDCM approach is developed. Subsequently, Section 3.1.3 explains the method for calibrating the parameters of DCMs and how the calibrated DCMs can be re-estimated to fit into different

modelling contexts. The explained procedures respectively represent a stage of the enrichment using the IDCM approach and the basic for the establishment of the IDCM performance theory. This is followed by Section 3.1.4, which demonstrates typical methods for testing the goodness-of-fit (GOF) of calibrated DCMs. These GOF measures provide ideas for developing the metrics for characterising the data distribution (condition) underlying the sample to be enriched by quantifying the mutual dependence between the input behaviour feature and the enriched attribute.

3.1.1 Econometric formulation

Whilst DCMs can take a wide variety of forms, they share the certain common principles³ that:

Individual decision-makers choose between *alternatives* which are characterised by a number of specific *attributes* according to some *decision rules* (Ben-Akiva and Lerman, 1985).

Microeconomic consumer theory provides the most common framework for establishing the compensatory decision rules for DCMs, by assuming that each decision-maker has a set of preferences over the alternatives in the choice set C such that the following properties hold (Mas-Colell, Whinston and Green, 1995):

- 1) Reflexivity: $j \succ j \quad \forall j \in C$;
- 2) Transitivity: $i \succ j$ and $j \succ k \Rightarrow i \succ k \quad \forall i, j, k \in C$;
- 3) Comparability: $i \succ j$ or $j \succ i \quad \forall i, j \in C$.

The operator $\succ$ is such that $i \succ j$ means the decision-maker prefers i to j or is indifferent between the two. Since C is finite, there always exists the most preferred alternative i^* :

$$\exists i^* \text{ s.t. } i^* \succ j \quad \forall j \in C. \quad (3-1)$$

According to the above three properties, there exists a function U :

$$U: C \rightarrow \mathbb{R} \text{ s.t. } i \succ j \Leftrightarrow U(i) \geq U(j) \quad \forall i, j \in C. \quad (3-2)$$

Therefore, the most preferred alternative i^* is the one that maximises U :

³ These principles also serve as the departures for studies of DCM extensions such as incorporation of latent variables (Morikawa, Ben-Akiva and McFadden, 2002) and group decision-making modelling (Zhang *et al.*, 2009).

$$i^* = \arg \max_{i \in C} U(i), \quad (3-3)$$

where U is a utility function which ranks the alternatives in preference order.

Lancaster put forward that the utility is related to the attributes of decision-makers and the properties of choice alternatives (Lancaster, 1966), i.e.:

$$U_a(i) = U(\mathbf{A}_{ai}, \mathbf{X}_a, \boldsymbol{\beta}_{ai}). \quad (3-4)$$

where $U_a(i)$ is the utility of choosing alternative i for decision-maker a

$\mathbf{A}_{ai}$ is a set of attributes characterising alternative i for decision-maker a ;

$\mathbf{X}_a$ is a set of attributes characterising decision-maker a ;

$\boldsymbol{\beta}_{ai}$ is a vector of parameters, often called tastes.

The above utility model is entirely deterministic in two senses:

- 1) Decision-makers are assumed to have no uncertainty about the attributes of the alternatives;
- 2) Analysts are assumed to have perfect knowledge about the utility of each decision-maker.

In real-world applications, however, the utility is stochastic due to a variety of reasons, such as the unobserved variations amongst the preference of decision-makers. Thus, probabilistic mechanisms are required to address such behaviour inconsistencies empirically (Edwards, 1962). An instance of such a probabilistic mechanism is random utility theory, which assumes that decision-makers have no uncertainty about the attributes of the choice alternatives while the analyst cannot completely observe the utility functions of decision-makers (Manski, 1977). The utility based on random utility theory can hence be decomposed into two parts: (i) an observed deterministic part (systematic utility) V dependent on the attributes and taste parameters, and (ii) an unobservable stochastic part (error term) ε :

$$U_a(i) = V_a(i) + \varepsilon_a(i) = V(\mathbf{A}_{ai}, \mathbf{X}_a, \boldsymbol{\beta}_{ai}) + \varepsilon_a(i). \quad (3-5)$$

The problem is then one of dealing with choice probabilities rather than deterministic choices. In particular, the probability that alternative i is chosen by individual a is effectively the probability that $U_a(i)$ is greater than or at least equal to the utilities of all the other alternatives in the choice set C_a of individual a . This can be expressed as:

$$P_a(i|C_a) = P[U_a(i) \geq U_a(j) \quad \forall i, j \in C_a]. \quad (3-6)$$

Equations 3-5 and 3-6 jointly yield the utility maximisation of random utility models (RUMs), which effectively involves that the decision-maker choose to maximise their utility in the presence of a random component. Note that not all DCMs are necessarily RUMs. There are also other theories addressing the stochasticity of decision-making behaviours. For example, the brand choice model proposed by Roberts and Urban (1988) is developed based on expected utility theory which assumes decision-makers' imperfect knowledge about the attributes of choice alternatives. These theories are, however, beyond the scope of the current study and are therefore not discussed in this thesis.

3.1.2 Model components and structures

DCMs can be categorised according to either the assumption made about the probability distribution of the error term ε or the number of alternatives modelled. Specifically, the most typical assumptions made with respect to the probability distribution of ε include Type 1 generalised extreme value (EV1) distribution and normal distribution, which leads to logit models and probit models, respectively. As for the number of alternatives modelled, binary and multinomial models are distinguished, respectively addressing 2 and 2+ alternatives. This section mainly discusses binary logit (BL) and multinomial logit (MNL) models, attributable to their tractability and subsequent use in the present analysis.

Following Equations 3-5 and 3-6, the probability that individual a choose alternative i from the choice set C_a can be manipulated as:

$$\begin{aligned} P_a(i|C_a) &= P[V_a(i) + \varepsilon_a(i) \geq V_a(j) + \varepsilon_a(j) \quad \forall i, j \in C_a] \\ &= P[\varepsilon_a(j) - \varepsilon_a(i) \leq V_a(i) - V_a(j) \quad \forall i, j \in C_a]. \end{aligned} \quad (3-7)$$

The above probability is a cumulative distribution, i.e. the probability that $\varepsilon_a(j) - \varepsilon_a(i)$ is below the observed quantity $V_a(i) - V_a(j)$. Denoting $f(\varepsilon_a)$ as the joint density function of the error term ε_a , the cumulative probability in Equation 3-7 can be rewritten as:

$$P_a(i|C_a) = \int \delta[\varepsilon_a(j) - \varepsilon_a(i) \leq V_a(i) - V_a(j) \quad \forall i, j \in C_a] f(\varepsilon_a) d\varepsilon_a, \quad (3-8)$$

where $\delta(\cdot)^4$ is an indicator function equalling 1 if $(\cdot)$ is true and 0 otherwise.

⁴ Conventional denotation I for indicator functions is used in the present study to denote the mutual information. Thus, δ is adopted in this thesis to represent indicator functions to distinguish the two.

The reformulated probability in Equation 3-8 is a multidimensional integral over the density function, which takes a closed form only for certain specifications of $f(\varepsilon_a)$. For example, a closed-form integral can be obtained for Equation 3-8 by assume that, for each decision-making unit a , the error components (ε_a) associated with all alternatives in the choice set (C_a) are independent (unnecessary for binary models) and identically distributed (i.i.d) according to an EV1 distribution (Ben-Akiva and Lerman, 1985). Such assumption about the i.i.d EV1 distribution of ε_a correspondingly lead to MNL models, enabling the tractability of logit models. Specifically, the density function of each error component assumed to follow the EV1 distribution takes the simplest form as:

$$f(\varepsilon_a) = e^{-\mu\varepsilon_a} e^{-e^{-\mu\varepsilon_a}}, \quad (3-9)$$

where μ is the scale parameter of common EV1 distribution such that $\text{Var}(\varepsilon) = \frac{\pi^2}{6\mu^2}$.

The corresponding cumulative distribution is:

$$F(\varepsilon_a) = e^{-e^{-\mu\varepsilon_a}}. \quad (3-10)$$

To derive the expression of the probability that individual a choose alternative i from the choice set C_a , Equation 3-7 is further reformulated as:

$$\begin{aligned} P_a(i|C_a) &= P[\varepsilon_a(j) - \varepsilon_a(i) \leq V_a(i) - V_a(j) \quad \forall i, j \in C_a] \\ &= P[\varepsilon_a(j) \leq V_a(i) - V_a(j) + \varepsilon_a(i) \quad \forall i, j \in C_a]. \end{aligned} \quad (3-11)$$

This hence indicates that $P_a(i|C_a)$ with a particular $\varepsilon_a(i)$ is the cumulative distribution over all $j \in C_a$ ($j \neq i$). As $\varepsilon_a(j)$ is i.i.d, such a cumulative distribution is effectively the product of the cumulative distributions of all $j \in C_a$ ($j \neq i$), i.e.:

$$P_a(i|C_a, \varepsilon_a(i)) = \prod_{\substack{j \in C_a \\ j \neq i}} e^{-e^{-\mu[V_a(i) - V_a(j) + \varepsilon_a(i)]}}. \quad (3-12)$$

As $\varepsilon_a(i)$ itself also follows the EV1 distribution, $P_a(i|C_a)$ is therefore the integral of the above function over all values of $\varepsilon_a(i)$ weighted by its density function:

$$P_a(i|C_a) = \int \left(\prod_{\substack{j \in C_a \\ j \neq i}} e^{-e^{-\mu[V_a(i) - V_a(j) + \varepsilon_a(i)]}} \right) e^{-\mu\varepsilon_a(i)} e^{-e^{-\mu\varepsilon_a(i)}} d\varepsilon_a(i). \quad (3-13)$$

By performing suitable algebraic manipulation, the above integral can produce a succinct, closed-form expression of $P_a(i|C_a)$ for BL and MNL models. This is expressed as:

$$P_a(i|C_a) = \frac{e^{\mu V_a(i)}}{\sum_{j \in C_a} e^{\mu V_a(j)}}. \quad (3-14)$$

Relaxation of the assumptions associated with BL and MNL models can further allow random taste variations and substitution patterns (Train, 2003). This has motivated the development of more complex DCM structures, e.g. the Generalised Extreme Value models (Train, 2003) and mixed logit models (Hensher and Greene, 2003). Despite the advances in those sophisticated structures, BL and MNL models have remained the workhorse of discrete choice modelling for decades, and still remain a benchmark against which more complex frameworks are evaluated. They are thus selected as the DCM structure explored in this study, serving as the departure point for the establishment of the IDCM socio-demographic enrichment framework.

3.1.3 Parameter calibration

Parameter calibration, or estimation, refers to the process of finding the parameters that can describe the distribution of the measured (observed) data (Pronzalo and Walter, 1997). Recall the components of the systematic utility V (Equation 3-5) of DCMs. It shows that the choice probability for a particular model structure can be calculated if the relevant attributes of choice alternatives and choice-makers are known together with the tastes. Therefore, parameter calibration (estimation) in discrete choice modelling refers effectively to determining the tastes.

The systematic utility function of DCMs can take various forms while most applications in choice modelling specify linear-in-parameters (LIP) utility functions (Bierlaire, 2003), i.e.:

$$V_a(i) = \beta_{i_0} + \beta_{ai_i} A_{ai} + \beta_{ai_a} X_a, \quad (3-15)$$

where β_{i_0} is a vector of alternative specific constants (ASCs⁵). For a model for n alternatives, only $(n - 1)$ ASCs can be estimated due to their inherently linear dependence.

⁵ ASCs are often included in DCMs to capture the average effect on the utility of all factors that are not included in the model (Train, 2003).

β_{ai_i} is a vector of parameters for alternative-specific attributes A_{ai} . β_{ai_i} are termed generic parameters if $\beta_{ai_i} = \beta_{aj_j} \forall j \in C_a$ holds and are termed as alternative-specific parameters otherwise.

β_{ai_a} is a vector of parameters for individual-specific attributes X_a . β_{ai_a} can be included in only $n - 1$ of the utility functions as only the difference between utilities matters.

The taste parameters of BL and MNL models in the LIP utility function presented in Equation 3-15 are calibrated using maximum likelihood methods (Ben-Akiva and Lerman, 1985).

In statistics, a likelihood function $\mathcal{L}(\theta)$ represents the joint probability distribution of the observed sample. For an i.i.d sample configuration x_i ($i = 1, \dots, n$), which is the case for logit models, the likelihood based on the probability density⁶ function $f(x; \theta)$ with parameters θ is defined as (Harris and Stocker, 1998):

$$\mathcal{L}(\theta) = \prod_i^n f(x_i; \theta). \quad (3-16)$$

The maximum likelihood estimation (MLE) method refers to the method of estimating the unknown parameters θ of a statistical model for given observed data by maximising the above likelihood (Fisher, 1912). For a set of observed data x_i ($i = 1, \dots, n$), the MLE estimator selects the set of model parameter values θ_{MLE}^* that maximises the likelihood function:

$$\theta_{MLE}^* = \arg \max_{\theta} \mathcal{L}(\theta; x_1, \dots, x_n) = \arg \max_{\theta} \prod_i^n f(x_i; \theta). \quad (3-17)$$

Bayesian inference suggests that MLE is a special case of maximum *a posteriori* estimation (MAP⁷). Specifically, the MAP estimator is the estimate of an unknown quantity in light of both the actual observations and any prior knowledge the analyst may have about the estimated quantities (Bassett and Deride, 2019). It assumes the knowledge of a prior distribution φ over parameters θ . Bayes' Theorem derives the MAP estimator for the model parameters as:

⁶ For the clarity of discussion, we follow the notation for continuous variables. A corresponding argument and derivation can be made, however, with respect to discrete variables, using the term 'probability mass function'.

⁷ MAP is used as the abbreviation for *maximum a posteriori* estimation rather than MAPE because MAPE has been conventionally used to represent the short for mean absolute percentage error.

$$\theta_{MAP}^* = \arg \max_{\theta} \prod_i^n \frac{f(x_i; \theta) \varphi(\theta)}{f(x_i)} = \arg \max_{\theta} \prod_i^n f(x_i; \theta) \varphi(\theta) \quad (3-18)$$

The term $f(x_i)$ in the denominator can be dropped from Equation 3-18 when it is independent of θ . Although this is assumed in this study, it needs to be noted that big datasets are typical exogenous datasets (Gandomi and Haider, 2015), in which the data is often correlated with θ , thus leading to $f(x_i)$ not being eliminated from the equation. The exogeneity of big data samples hence suggests a potential avenue for further investigations, i.e. to relax the current assumption regarding the independence of x_i with θ .

As a special case of the MAP, the MLE is the MAP where a uniform prior distribution is assumed for the model parameters to be estimated. It is employed when no prior information about the parameters is available, which is typical in parameter estimation.

DCMs are frameworks in which attributes of choice alternatives ($\mathbf{A}$), characteristics of choice-makers ($\mathbf{X}$) are mapped onto the probability space of the observed choice (Y) based on knowledge of the preferences captured in taste parameters ($\boldsymbol{\beta}$). Thus, Y is effectively a random discrete variable with the probability mass function denoted as $P(Y|\mathbf{X}, \mathbf{A}, \boldsymbol{\beta})$.

For a sample containing N choice-makers labelled $a = 1, 2, \dots, N$, each making L_a choices, the likelihood of observing this sample for BL and MNL models can be formulated, according to their structures as:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_a^N \prod_i^{C_a} \prod_l^{L_a} P_a(i|C_a) \delta_{ai}^l = \prod_a^N \prod_i^{C_a} \prod_l^{L_a} \left[\frac{e^{\mu V(A_{ai}, \mathbf{X}_a, \boldsymbol{\beta})}}{\sum_{j \in C_a} e^{\mu V(A_{aj}, \mathbf{X}_a, \boldsymbol{\beta})}} \right]^{\delta_{ai}^l}, \quad (3-19)$$

where δ_{ai}^l is an indicator equalling 1 if individual a choose alternative i in the l th choice and 0 otherwise.

C_a is the choice set of individual a consisting of all the possible values of Y .

The parameter calibration of a DCM thus involves finding $\boldsymbol{\beta}^*$ that maximises Equation 3-19. McFadden (1974) proved that the log-likelihood $\log \mathcal{L}(\boldsymbol{\beta})$ for BL and MNL models is globally concave in parameters $\boldsymbol{\beta}$ for LIP utilities. This provides the feasibility to compute $\boldsymbol{\beta}^*$ using simple numerical procedures such as Newton-Raphson search (Meeter, 1966).

The calibration of DCM parameters provides a mapping between the choice behaviour for discrete alternatives and the socio-demographic attributes of the choice-maker. This therefore

permits developing an approach to derive the socio-demographic attributes from observations of choice behaviours based on the DCM calibrated to map the correlation between the two. Moreover, parameter calibration also presents the idea for estimating the unknown component of statistical models by means of maximum likelihood methods, which serve as the solution methods to the developed IDCM econometric formulation, as explained later in Section 4.1.

In forecasting, it is often necessary to adjust the scale and ASCs, i.e. $\beta_{t,0}$ in Equation 3-15, of a DCM when the DCM is applied to a sample or population different in the area or year from the sample upon which this DCM is originally calibrated. This is to reflect that the unobserved factors can be different for different forecast areas/years. In other words, the constant recalibration refers to an appropriate adjustment to the DCM for the same variables with different probability distributions. The market share information about the explanatory attributes for the forecast area/year sample is used to recalibrate the ASCs (Train, 2003). Such a recalibrated model can then be used to predict the changed market share of the choices due to changes in explanatory factors while properly accounting for the true underlying probability distributions. The recalibration of constants also provides the idea to approximate the joint probability distribution underlying a sample when the marginal distribution of some of the variables are unknown. This allows to approximate the probability distribution of a data sample as a smooth function of the recalibrated constant, which is discussed later in Section 4.2.

3.1.4 Goodness-of-fit (GOF) tests

Assessing the GOF of models is necessary in choosing between model specifications with different structures or different components. The GOF of models is often measured in terms of both the overall model fit and the parameter estimation precision. In this section, the former overall model fit of models is discussed due to its relevance to the IDCM performance theory developed in Section 4.2. Specifically, there are two general approaches to develop the overall GOF metrics for DCM structures: (i) the log-likelihood and (ii) prediction tests (Ben-Akiva and Lerman, 1985). The following paragraphs discuss the former log-likelihood method since it is generally preferred and can be widely applied to any DCM.

The most common log-likelihood-based GOF metric is the McFadden's likelihood ratio index (*LRI*), which is developed as an analogue to *R*-squared, the coefficient of determination in conventional regression (McFadden, 1974), and is expressed as:

$$LRI = 1 - \frac{\ln \mathcal{L}(\boldsymbol{\beta})}{\ln \mathcal{L}(\mathbf{0})}, \quad (3-20)$$

where $\ln \mathcal{L}(\boldsymbol{\beta})$ is the value of the log-likelihood function at $\boldsymbol{\beta}^*$.

$\ln \mathcal{L}(\mathbf{0})$ is the value of the log-likelihood computed with ASCs only.

This metric is intuitively appealing in that it is bounded by 0 and 1. However, the LRI differs from conventional R -squared metric as it is unable to measure the proportion of the total variation in the predicted variable that is accounted for by the variation in the explanatory variables. There hence is no natural interpretation for the value of LRI between 0 and 1, making it a rather informal GOF metric (Greene, 2012). Another weakness of LRI is that it does not consider the complexity of the model specification. For instance, a model might fit the data well by capturing all the variation in the data with a large number of parameters whilst providing no useful generalisation. A modified LRI is hence proposed, taking into account the degrees of freedom (DOF) in the data used in the model, and is formulated as:

$$LRI_{Mod} = 1 - \frac{\ln \mathcal{L}(\boldsymbol{\beta}) - r}{\ln \mathcal{L}(\mathbf{0})}, \quad (3-21)$$

where r is the number of parameters in the model.

However, arguments have been made that the penalty to the DOF by the above measure is not sufficient (Amemiya, 1985). This thus promotes alternative approaches such as Akaike and Bayesian information criteria. Grounded in information theory, the Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby the relative quality of statistical models for a given set of data. By making a trade-off between the GOF and the simplicity of the model, the AIC estimates the relative amount of information lost by a given model and is defined as (Akaike, 1974):

$$AIC = 2r - 2 \ln \mathcal{L}(\boldsymbol{\beta}). \quad (3-22)$$

Similarly, as is defined in the equation below, the Bayesian information criterion (BIC) is also developed by introducing a penalty term for the number of parameters in the model dealing with the possible overfitting resulting from inclusion of additional parameters (Schwarz, 1978).

$$BIC = r \ln N - 2 \ln \mathcal{L}(\boldsymbol{\beta}). \quad (3-23)$$

where N is the number of observations in the sample.

Amongst a finite number of alternative models, the one with the lowest *AIC* or *BIC* is selected. LRI_{Mod} , *AIC* and *BIC* are indices incorporating different degrees-of-freedom corrections to the same measure of GOF. Studies have supported their use in measuring the GOF of discrete choice modelling. For instance, it is found that the index based on *AIC* favours more parsimonious models (Ben-Akiva and Swait, 1986).

In addition to evaluating the overall GOF of DCMs with different components, the complexity (structure) of the model specification is also crucial, which triggers the development of the notion of likelihood-ratio test. For instance, Equation 3-15 shows that the taste parameters for the attributes of choice alternatives in a LIP utility can be either generic or alternative-specific. With respect to the travel time attribute in the choice between different travel modes, for example, this means that the significance of a same amount of travel time can be either the same or different for different mode alternatives. The utility developed with the former generic taste parameters is termed as a restricted model while the latter with alternative-specific parameters is defined as a general model. The null hypothesis of this likelihood-ratio test assumes that the restricted model is more preferred as compared to the general model. The statistic for the null hypothesis, which take the form in Equation 3-24 below, is χ^2 distributed with r DOF where r indicates the number of alternative attributes in the model.

$$-2(\ln \mathcal{L}_R(\boldsymbol{\beta}) - \ln \mathcal{L}_G(\boldsymbol{\beta})), \quad (3-24)$$

where $\ln \mathcal{L}_R(\boldsymbol{\beta})$ is the log-likelihood of the restricted model;

$\ln \mathcal{L}_G(\boldsymbol{\beta})$ is the log-likelihood of the general model.

The GOF tests discussed above provide an idea to express the explanatory power of one variable in inferring the other using the log-likelihoods, which is treated as a way to quantify the quality or condition of the data, and is explained in detail later in Section 4.2.2.

3.2 Conceptual IDCM model based on microeconomic behavioural theories

As the first step to formulate the socio-demographic enrichment based on choice behaviour patterns, a conceptual, microeconomic model is presented in the following paragraphs to illustrate the logic underlying the IDCM approach. In particular, this model can be viewed as a counterpart to the microeconomic, behavioural foundations of discrete choice modelling,

drawing upon the time-allocation and goods-leisure paradigms (Train and McFadden, 1978; Jara-Díaz, 2007) as well as the place-visit model established by Pawlak *et al.* (2015). Whilst using choice behaviour between travel modes as an intuitive example, the logic of this model can be applied and extended to any type of, or a combination of, discrete choice behaviours.

Assume an individual traveller who wishes to select between different travel modes in a given set C of available mode alternatives by maximising his/her utility which characterises his/her indifference mapping between goods (affordable products/service) and leisure (available time). In the original Train and McFadden formulation (1978), the traveller is assumed to have a total amount of budget B and time T as the total good and leisure, respectively. The indirect utility is shaped by $\tilde{V}(B, T)$, representing the traveller's indifferent mapping between goods and leisure. Define the cost and time spent on mode $i \in C$ respectively as ς_i and t_i . The indirect utility function of choosing mode i can therefore be formulated as:

$$\tilde{V}_i(\varsigma_i, B - \varsigma_i, t_i, T - t_i). \quad (3-25)$$

The utility maximisation problem yielding from Equation 3-25 thus involves making a trade-off between the monetary cost and the travel time. In more general, however, as can be observed from Equation 3-15, a wider set of attributes, e.g. the properties of choice alternatives ($\mathbf{A}$) and the attributes of the decision-maker ($\mathbf{X}$) should be considered in forming such a utility function. In this sense, $T \in \mathbf{A}$ and $B \in \mathbf{X}$. And the aforementioned indirect utility $\tilde{V}_i$ can hence be extended accordingly as:

$$\tilde{V}_i(\varsigma_i, B - \varsigma_i, t_i, T - t_i | \mathbf{A}, \mathbf{X}, \boldsymbol{\beta}). \quad (3-26)$$

The choice of a most preferred mode i^* thus requires maximising this indirect utility as:

$$i^* = \arg \max_i \tilde{V}_i(\varsigma_i, B - \varsigma_i, t_i, T - t_i | \mathbf{A}, \mathbf{X}, \boldsymbol{\beta}). \quad (3-27)$$

In an idealised situation in which the choice i^* and the corresponding maximum systematic utility level ($\bar{V}$) are known, Equation 3-27 can be inverted directly to infer the attributes $\mathbf{X}$ of the traveller as:

$$\mathbf{X} = \tilde{V}^{-1}(\bar{V}, \varsigma_{i^*}, B - \varsigma_{i^*}, t_{i^*}, T - t_{i^*} | \mathbf{A}, \boldsymbol{\beta}). \quad (3-28)$$

It is noted from the above derivation that the IDCM socio-demographic enrichment approach, which is formulated based on the aforementioned conceptual model, involves the inversion of a DCM. This effectively corresponds to the question addressed in the umbrella area of inverse

problems. Moreover, as explained in Section 3.1.1 and Section 3.1.2, respectively, the choice-making behaviour in most real-world applications is stochastic and typical DCMs are non-linear. The stochastic and non-linear nature of DCMs yield correspondingly in the ill-posedness of this inverse problem, which, as will be demonstrated in Section 3.3.2, increases the difficulty in solving such inverse problems. Therefore, in the following section, the research field of inverse problems is discussed. This is to enable understanding of the mechanism underlying the operation of the socio-demographic enrichment using the IDCM approach and the associated challenges, highlighting the value of the current study.

3.3 Inverse problems

As discussed in Section 2.1.7, existing studies that seek to enrich the socio-demographic attributes of anonymous passive data utilise various approaches and algorithms that model the aimed socio-demographic attributes as the responses or classification labels as functions of the observed behaviour patterns as the predictors (features). Whereas this seemingly conforms with the aimed enrichment problem as stated and the data availability, these existing methods, as indicated in Section 1.2, fail to account for the behavioural principles suggested by microeconomic theory. In contrast, the IDCM socio-demographic enrichment approach proposed in the current study seeks to address this research gap by developing a methodology with behavioural (econometric) underpinnings grounded in DCMs. Maintaining the mutual dependence suggested by microeconomic behavioural theories acknowledges the need to perform suitable statistical inversion techniques on conventional DCMs. The solution to the IDCM approach corresponds to the umbrella term of inverse problems.

The research field of inverse problem is hence discussed in the present section to explain the mechanism underlying the IDCM approach and the associated challenges, as well as to describe possible solution methods. Specifically, Section 3.3.1 defines inverse problems and their characteristics, which reflect the associated challenges, while Section 3.3.2 overviews the solution techniques for inverse problems, thus offering potential solution method to the proposed IDCM socio-demographic enrichment framework.

3.3.1 Definition and properties

An inverse problem is defined as the process of inferring from a set of observations the causal factors that are believed to have produced them (Tarantola, 1987). This is relative to the concept

of forward problems which proceed from the causes to calculate the outcomes. If the mapping is from a functional space Q to another space J , the forward problem F can be written as:

$$j = F(q) \quad \text{for } q \in Q \text{ and } j \in J. \quad (3-29)$$

The corresponding inverse problem (F^{-1}) involves finding $q \in Q$ from knowledge of $j \in J$ such that Equation 3-29, or at least its approximation, holds (Bal, 2012). Such an inverse problem is therefore expressed as:

$$q \cong F^{-1}(j) \quad \text{for } q \in Q \text{ and } j \in J. \quad (3-30)$$

Inverse problems are crucial in science and engineering because they provide insights into parameters of the system that usually cannot be directly observed. Model estimation, as briefly discussed in Section 3.1.3, is a basic example of an inverse problem, in which the parameters of the model are found in order to maximise its fit to the observed data. Figure 3.1 provides a conceptual formulation of the inverse problem for model parameter estimation in comparison to the corresponding forward problem. Socio-demographic enrichment using the IDCM approach is also an inverse problem, in which socio-demographic attribute is estimated in order to reflect its effect in determining the observed choice behaviour pattern.

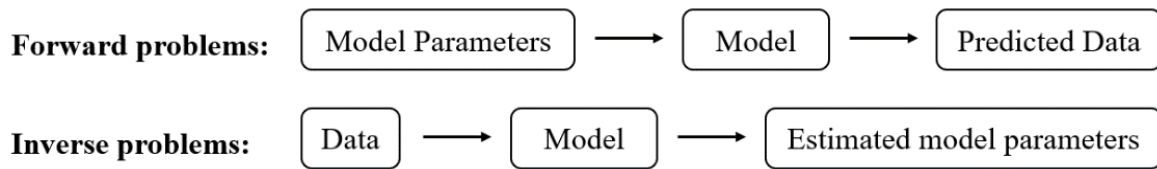


Figure 3.1 Conceptual Formulation of Forward and Inverse Problems

The key properties of inverse problems concern their linearity and stochasticity, which determine the complexity of the inversion process. In the sense of Hadamard (1902), a problem is considered ‘well-posed’ if it has a unique solution that continuously depends on the data. Correspondingly, problems resulting from either stochasticity or non-linearity are termed ‘ill-posed’ if either the uniqueness or the stability of well-posed problems is violated.

The relationship between linearity, stochasticity and ill-posedness presented in inverse problems is conceptualised in Figure 3.2 below (Page 77).

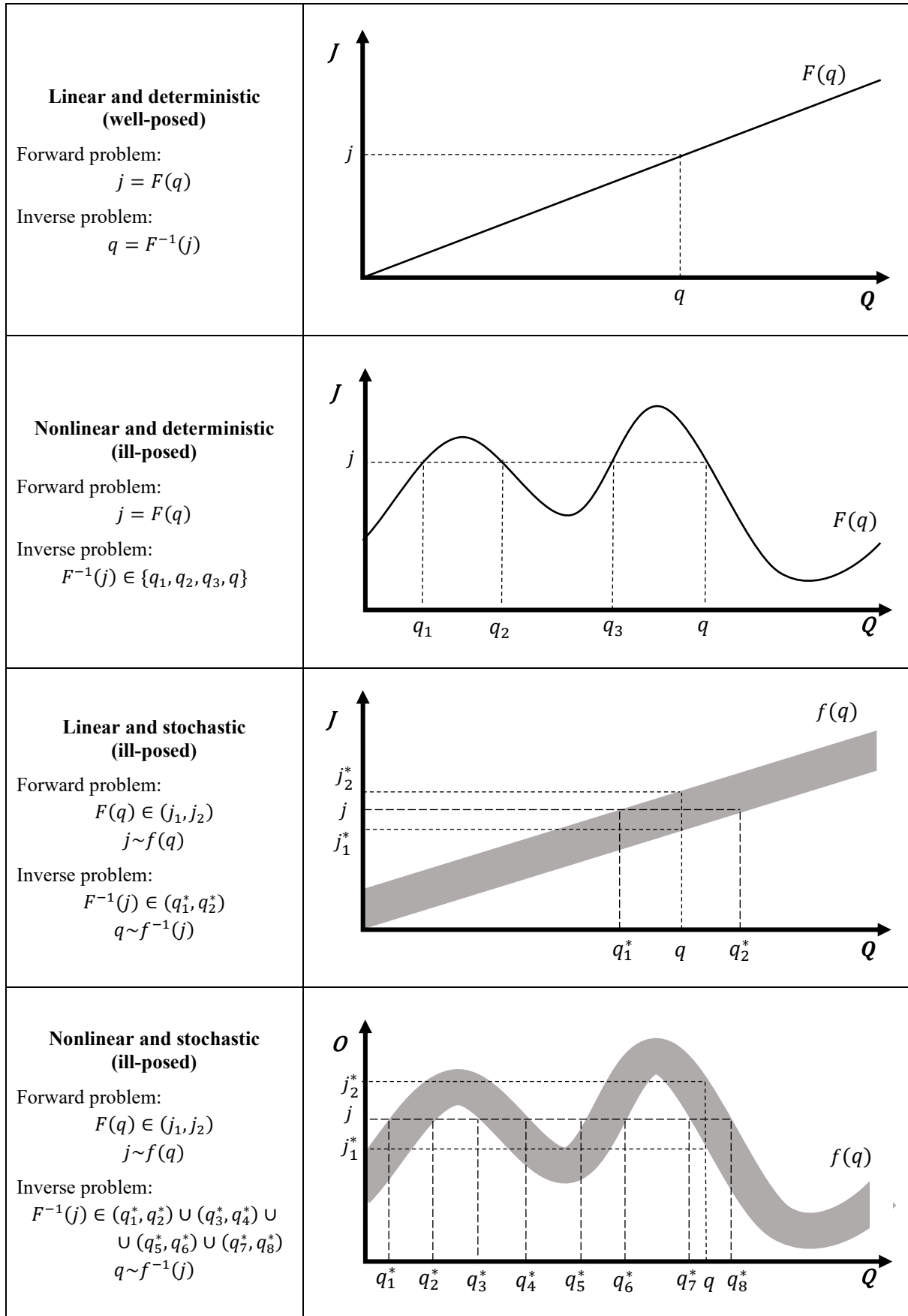


Figure 3.2 Conceptual Relationship between Linearity, Stochasticity and Ill-posedness

It is apparent that the properties of ill-posedness can reduce the relative ease of solving the corresponding inverse problems. This has provoked continuing efforts to develop robust estimation techniques for these increasingly complex modelling structures. It is also worth noting that, since inverse problems are often formulated in infinite dimensional spaces, the practical consideration of recovering only a finite number of unknown parameters when the number of measurements is limited can lead to the problem being recast in a discrete form, which corresponds to the case of a typical ill-conditioned inverse problem (O’Sullivan, 1986).

Choice behaviour models are stochastic due to either the imperfect knowledge of the researcher or unobserved inter- and intra-individual heterogeneity. A good example can be found in the RUMs that incorporate an error term to capture such stochasticity (Ben-Akiva and Lerman, 1985), as explained previously in Section 3.1. Additionally, choice models are typically non-linear (see for example the MNL models (Train, 2003)). Due to the stochastic and non-linear (i.e. ill-posed) nature of choice models, the inversion of a DCM, as in the current IDCM establishment, is often challenging. Therefore, the development of the solution techniques to various inverse problems (especially the ill-posed ones) are discussed in the following section.

3.3.2 Solution techniques for ill-posed inverse problems

A rich body of literature exists for addressing various inverse problems, ranging from simple deterministic linear ones to more complex stochastic or non-linear ones. Given the ill-conditioned nature of DCMs, suitable techniques are required to tackle the ill-posedness of corresponding inverse problem, i.e. the IDCM socio-demographic enrichment approach. The following contents hence review the methodologies targeting ill-posed or ill-conditioned issues, fitting the scope of the current IDCM establishment.

In the early development for the solutions to ill-posed inverse problems, Backus and Gilbert (1967) suggested an approach that takes linear combinations of the data to generate unique localised averages of the model. Such types of linear techniques relying on the calculation of the partial derivatives of parameters, however, have been found to be inapplicable to an increasing number of inherently more difficult, non-linear inverse problems (Oldenburg, 1984).

With enhanced computational power, novel methodologies that can accommodate more complex ill-conditioned inverse problems have emerged. For instance, Brute-force search (BFS), or exhaustive search, is a general problem-solving technique which systematically enumerates all possible candidate solutions and checks whether each candidate satisfies the

statement of the problem. Whilst being computationally expensive, BFS can assure the global optimality of the solution and is often applicable when the search space is finite and computationally manageable. It is hence preferred when the simplicity of implementation is more important than the solution speed. In applications, BFS is suitable for problems where search size is limited or where specific heuristics can be applied to reduce the set of candidate solutions to a manageable size.

Likewise, Bayesian methods are found to be significantly effective in processing incomplete and noisy data in the context of inverse problems (Tarantola, 1987). For example, regularisation can be applied to address the overfitting of ill-conditioned inverse problems in which a small change in the predictors can result in a large variability in the response variables. This is interpreted as a special case of Bayesian inference, which corresponds to a branch of Bayesian methods using Bayes' Theorem to obtain the probability of a hypothesis. Specifically, Bayesian inference derives the posterior probability as a result of a prior probability and a likelihood function based on a statistical model for observed data (Vapnik, 1999).

X is the sample with a set of observed data points $x_1, x_2, \dots, x_n$;

θ is the parameter of the distribution of data point X , i.e. $x \sim P(X|\theta)$;

ζ is the hyperparameter of the distribution of parameter θ , i.e. $\theta \sim P(\theta|\zeta)$;

As defined above, the prior is the distribution of the parameter before any data is observed, i.e. $P(\theta|\zeta)$. The sampling distribution, meanwhile, is the distribution of the observed data conditional on its parameters, i.e. $P(X|\theta)$. In the Bayesian framework, the complete solution to the inverse problem is the posterior distribution of the unknowns, which can be stated as:

$$P(\theta|X, \zeta) = \frac{P(X|\theta)P(\theta|\zeta)}{P(X|\zeta)} \propto P(X|\theta)P(\theta|\zeta). \quad (3-31)$$

Correspondingly, effective sampling strategies are needed to find the best-fitting posterior distributions (Calvetti, Kaipio and Somersalo, 2014). Recent decades have seen substantial developments in global optimisation technologies to solve the inverse problems that essentially require efficient probabilistic search procedures on large sample spaces. These new methods include, for instance, genetic algorithms (GAs) and genetics-based machine learning (Tominaga, Nobuto and Masahiro, 2000). In a GA, a population of candidate solutions to an optimisation problem are progressed towards better solutions according to an iterative process, starting from a population of randomly-generated individuals. The more fitted candidates in an

iteration are stochastically selected to perform the next iteration. The algorithm commonly terminates when either a maximum number of iterations has been completed, or a satisfactory fitness level has been reached (Whitley, 1994). Despite the widespread applications of GAs, several limitations are recognised. For example, difficulties have been found in using GAs to solve complex high-dimensional, multimodal problems. Moreover, GAs tend to converge towards local optima in many problems, because the key focus is on the functional fitness, and they cannot accommodate decision problems with single right/wrong measures effectively.

Some of the above shortcomings can be addressed by employing the Particle Swarm optimisation (PSO), which solves the optimisation of a problem by iteratively improving a candidate solution with regard to a given measure of quality (Eberhart and Kennedy, 1995). In particular, a PSO algorithm is a heuristic method that works based on a swarm (population) of candidate solutions by moving around the search space based on the best-known positions of the entire swarm. Such a swarm is termed as the particles, hence giving the name. Both GAs and the PSO are population-based search processes relying on the information shared by population members about a combination of deterministic and probabilistic rules. The PSO is more computationally efficient in terms of speed and memory, however (Hassan *et al.*, 2005).

Another dynamically developing set of sampling approaches involve random and pseudo-random exploration, i.e. Monte Carlo (MC) approaches (Hammersley and Handscomb, 1964). Representing a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results, MC methods generally use randomness to solve the problems that might be deterministic in principle. Although there is no standardised definition, MC methods almost always follow the common steps set out by Kalos and Whitlock (2008):

- 1) Define a domain of possible inputs;
- 2) Generate inputs randomly from a probability distribution over the domain;
- 3) Perform a deterministic computation on the inputs; and
- 4) Aggregate the results.

There are two important points in performing these steps. First, the randomly-generated input in Step 2) should be uniformly distributed. Secondly, the number of inputs should be large. The approximation would be poor if either point is violated. Suppose the integral of the function $f(x)$ is evaluated over the interval $[\lambda_1, \lambda_2]$. The idea underlying an MC simulation is:

- 1) to select a number of points randomly within the domain of the integration formed by $f(x)$;

- 2) to work out the fraction of the number of points that fall below the curve; and
- 3) to multiply this fraction by $f(x)$ to approximate the integral.

Assume that $\{x_1, x_2, \dots, x_n\}$ represents a sequence of N values drawn from the interval above. The integral can be approximated as:

$$\int_{\lambda_1}^{\lambda_2} f(x)dx \approx (\lambda_2 - \lambda_1) \frac{1}{N} \sum_{i=1}^N f(x_i). \quad (3-32)$$

The requirement of large amount of random numbers in the above MC methods have fostered the development of pseudo random number generators, which are substantially more efficient than the random numbers used for traditional statistical sampling (Barker, 2016). Gaps could still appear for random points generated by pseudo-Monte Carlo (PMC), however. Although these gaps can be filled in the long run, the efficiency of PMC is significantly reduced. This led to the development of quasi-Monte Carlo (QMC) methods, in which sequences are generated algebraically in order to retain some randomness while providing a more even coverage of the integration space than pure PMC sequences. For instance, Bhat used Halton sequences to achieve the same accuracy of approximation as the PMC yet with fewer draws and integrand evaluations, hence resulting in faster performance and greater accuracy across the same number of draws as compared to PMC (Bhat, 2003). MC methods are used in various disciplines, mainly for three types of problems: optimisation, numerical integration and generating draws from a probability distribution (Kroese *et al.*, 2014). For mixed logit models (a specific DCM structure), for example, in the context of travel behaviour analysis, PMC/QMC methods can be applied to obtain approximated estimation for choice probabilities.

Unsurprisingly, MC methods have been found to be widely applicable in solving inverse problems. As a powerful simulation technique for performing integration, Markov chain Monte Carlo (MCMC) has revolutionised the application of Bayesian methods in inverse problems (Kass *et al.*, 1997). In particular, MCMC methods perform statistical sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. They are mainly applied to calculate numerical approximations for multi-dimensional integrals. In Bayesian statistics, the recent development of MCMC methods has allowed crucial progress in computing large hierarchical models that require integrations over hundreds or even thousands of unknown parameters (Banerjee, Carlin and Gelfand, 2014). For instance, Bui-Thanh and Girolami (2014) implemented a Riemannian manifold Hamiltonian MC algorithm to solve inverse heat conduction problems. Their sampler

is able to address the weakness of previous MC algorithms by accommodating high dimensions and the exhibition of strong correlations. In particular, they exploited the Riemannian structure for the parameter space, which allows for exceedingly efficient sampling in high dimensions where different scaling is required for transient and stationary phases of the Markov chain. The performance of their method was assessed in an application exploring various complex models, ranging from logistic regression to stochastic volatility models. Significant improvement was observed in both time and sample size, as compared with other alternative sampling approaches.

The availability of a plethora of potential approaches for solving ill-posed or ill-conditioned inverse problems permits the solution of the proposed IDCM approach, which is established as an ill-conditioned inverse problem for DCM. In particular, the IDCM approach formalised so far, as will be demonstrated in Section 4.1, is developed in the enrichment context where individuals are assumed to make choices independently, and where the size of candidate solutions is manageable. Therefore, the BFS method discussed above can be applied to solve the corresponding inverse problem, as discussed in more detail later in Section 4.1 and implemented in the applications presented in Chapters 5 and 6.

For further investigations into the IDCM establishment, the inversion may be performed based on more complex DCM structures or take into account the correlation between the choices made by different individuals. In these extensions, the BFS method may fail to accommodate the corresponding inverse problems. However, the availability of the range of methodologies as outlined above can offer suitable solution techniques to these extensions by handling large sampling spaces and/or different types of variables. Figure 4.1 (Section 4.1) will also provide a guideline for selecting between different sampling techniques to solve the socio-demographic enrichment using the IDCM approach for different types of attribute variables.

3.4 IDCM analogy in information theory

The process of making choices (referred to as the choice-making process below by individual choice-makers characterised by a set of socio-demographic attributes can be compared to the transmission of messages via an information communication channel in an information communication system. Such an information communication system is the key objective studied in information theory. Therefore, before explaining the aforementioned analogy, information theory and the core problems it studies are outlined in the following paragraphs.

Information theory studies the quantification, storage, and communication of information. One of the fundamental issues addressed in information theory is to exactly or approximately reproduce a message transmitted over an information communication channel as the key object presented in an information communication system (Shannon, 1948). Figure 3.3 below (Page 83) illustrates the elements of such an information communication system. Specifically, the information source (1) (a person, an event or a phenomenon) generates the message s to be communicated. It is further encoded by the transmitter (2) to produce the suitable input for transmission through a communication channel (3). The receiver (4) obtains the transmitted channel output and reconstructs the message as s' and then sends it to the destination (5) for which the message is intended (Shannon, 1948; Cover and Thomas, 2005).

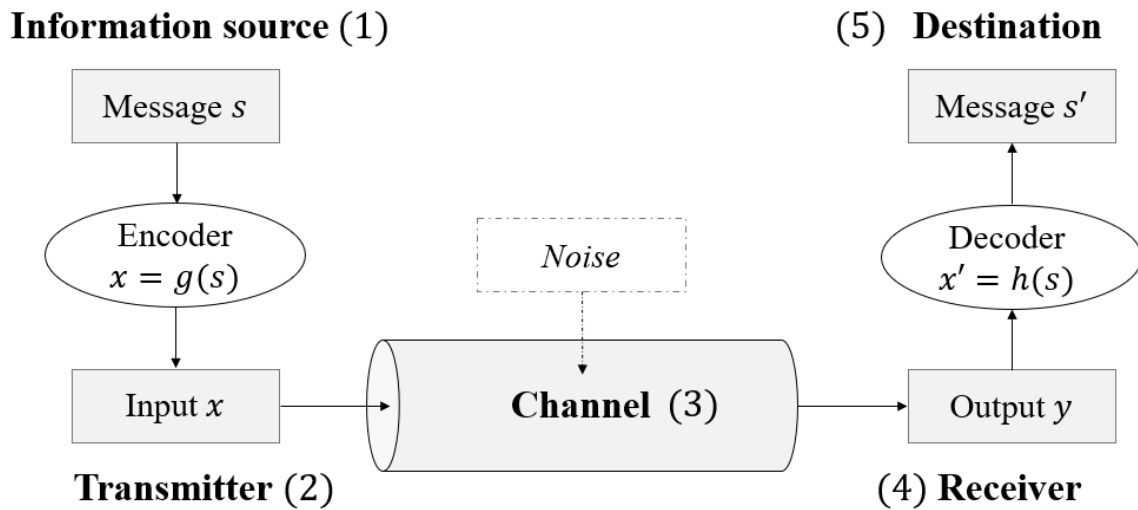


Figure 3.3 Components of Information Communication System
(Source: Shannon (1948); Stone (2015))

In both the input and the output of such a system, a message can be defined as an ordered sequence of outcome values of a specific random variable. For instance, a message s which informs the gender composition of the students of a class can be written as:

$$s = ('m', 'm', 'f', \dots, 'f', 'm'). \quad (3-33)$$

A message is encoded mainly for transmission efficiency purpose. In the above example, ' f ' and ' m ' are two symbols from a 26-long (English) alphabet, which can be replaced by 0 and 1 to remove the rest 24 redundant coding symbols. However, this specific consideration is not important for the present analysis and is therefore not discussed any further.

An information communication channel is defined as a noiseless channel if the output is always the same as the input. Otherwise, the channel is termed noisy, and the output may differ from

the input. For a typical noisy communication channel, the correctly conveyed messages constitute the information transmitted through the channel while the mistakenly communicated ones represent the noise. Adding noises during the transmission, which is the case for the information transmission over noisy channels, can induce errors in interpreting the output, and the decoder may consequently decode the output to a possibly incorrect message. The goal of information theory is then to reconstruct the original messages with low probability of error.

The framework demonstrated in Figure 3.3 is traditionally employed to characterise the communication process, in which a piece of information needs to be transmitted over a physical transmission medium, e.g. a wire for electric signals, a fluid such as air for sound waves or a logical connection such as telephone lines or routes through the Internet. Certain mapping rules are sometimes required when the input and the output represent different events or phenomena that are related by the information communication channel. A simple example of this involves the transmission of electric signal over a wire as the channel, in which the input is the operation of a switch while the output is the corresponding effect such as the light-up. Since the switch and the light are effectively two different objects and the events representing their operation patterns are different, an intuitive mapping rule can be defined as:

$$\begin{aligned} \text{Switch ON} &\rightarrow \text{Light UP} \\ \text{Switch OFF} &\rightarrow \text{Light OFF} \end{aligned} \tag{3-34}$$

It can be easily derived that the information is successfully transmitted when the electrical signal is consistent with the rule while the noise occurs if the rule is violated.

Recall Section 3.1.1 which describes microeconomic theory as indicating that people's choice behaviour is affected, or can even be entirely driven, by their socio-demographic attributes. This has also been proved by a plethora of empirical research regarding people's choice behaviour (Lancaster, 2008). The present study hence argues that the choice-making process of individual decision-makers can be conceptualised using the same framework of the transmission of messages over an information communication channel. Specifically, consider a situation where the original message contains the socio-demographic attributes of an individual choice-maker (information source) and the received message consists of the decisions (choices) made by this individual. The choice-making process in this interpretation can be viewed as a quasi-communication channel which transmits the socio-demographic attributes as the input to produce the choice behaviour as the output. All the other factors (if any) that can affect the choice (output) is treated as the noise given that they are unobserved or

not studied as the information source. In cases where multiple choices are made and observed, the input message would be a sequence of same values since the socio-demographic attribute of an individual, e.g. gender, usually remains the same while the output can be different, representing different choices for the same set of alternatives made by the individual.

To provide a direct, visual representation for the similarity between choice-making behaviour and the transmission of information over the information communication channel, Figure 3.4 below (Page 85) presents an example quasi choice-making communication channel, in which the input socio-demographic attribute takes whether the choice-maker is a student, and the output indicates the choice of whether attending a school. Note that, this quasi-communication channel is defined based on the intuition that students are generally more likely to attend school on a given day than those who are not students, which may be arguable but is presented for illustration purposes only.

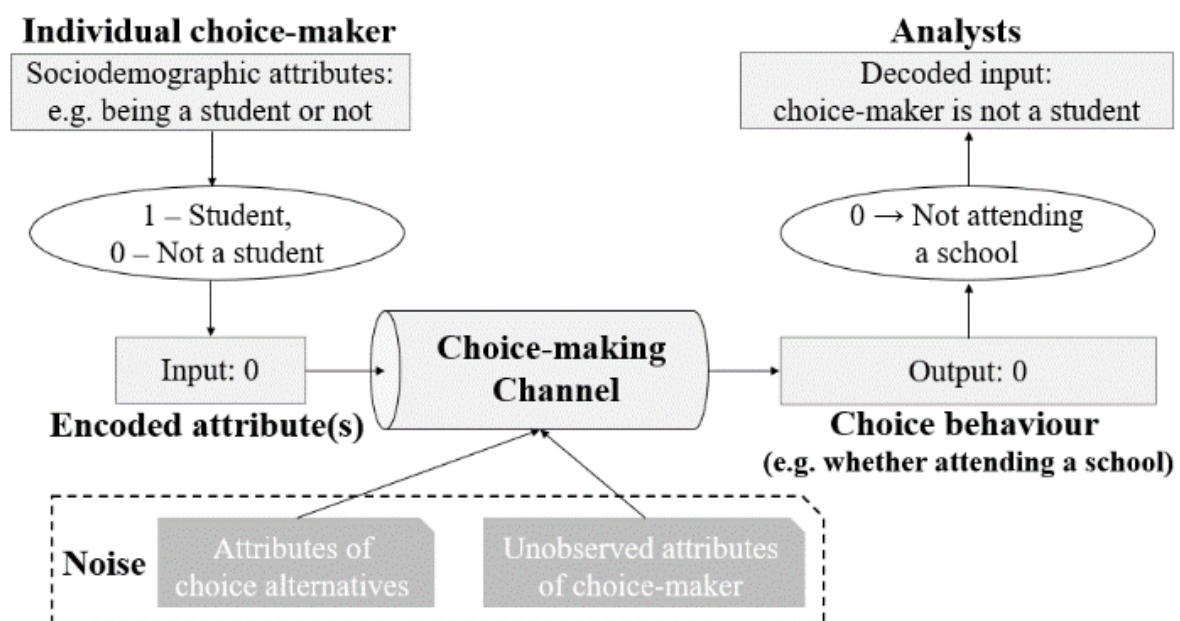


Figure 3.4 Example Choice-making Quasi-communication Channel

It is obvious that the quasi channel illustrated in Figure 3.4 is noisy because the dependence between the input and the output is not deterministic, i.e. not all the students may attend school on a given day (e.g. those are ill) and not all the individuals visiting schools are students (e.g. parents and teachers). More generally, quasi choice-making communication channels are typically noisy as the correspondence between a specific attribute-choice pair is rarely bijective, which can be attributed to a number of reasons as explained below.

- 1) The number of possible outcome values (POVs) of the input variable can differ from that of the output variable. For example, gender (input variable) may take only two reported POVs (male or female) while choices (output) can contain 2+ alternatives, e.g. travel modes which includes walking, cycling, public transport and driving, etc.
- 2) The attributes of choice alternatives can affect the decision-making. As the choice-maker plays the role of the information source, other factors that can affect the choice-making, e.g. the attributes of choice alternatives, would be treated as the noise and can possibly result in inconsistent choice behaviour patterns made by the same individual.
- 3) The choice-making process is stochastic, reflected by the presence of error terms, e.g. ε_a in RUMs as described in Section 3.1.1. This means that even for inputs and outputs having the same number of POVs without the influence of other external factors, the same input value can produce different output values and *vice versa*. Random utility theory explains this as either the choice-maker's imperfect knowledge of the choice alternatives, or the impact of other unobserved attributes of the choice-maker (Manski, 1977).

Mapping rules are also required for such quasi-communication channel as the socio-demographic attribute(s) (input) and the choice behaviour (output) are typically two variables. In the example presented in Figure 3.4, since the students are intuitively more likely to attend a school than other people, the corresponding mapping rule can be defined as:

$$\begin{aligned} \text{Student} &\rightarrow \text{Attend a school} \\ \text{Not a student} &\rightarrow \text{Not attend a school} \end{aligned} \tag{3-35}$$

Information and noise can also be defined correspondingly in a relative manner given whether the message is transmitted via the choice-making channel according to the mapping rule. Nevertheless, it would be tricky to distinguish between correct and incorrect messages for the quasi choice-making channel because, for instance, it is rather reckless to assert that a teacher attending a school has made the incorrect choice. This explains the reason why information and noise of such a choice-making channel are defined in a relative manner, i.e. information and noise for the quasi choice-making channel can be swapped by switching the definition of the mapping rule. However, the basic function of an information communication channel is to transmit information rather than to transmit the noise. Therefore, the convention to develop the mapping rules for such a quasi-communication channel is to define the POV of the output with a higher probability (e.g. attending a school) conditional on the specific POV of the input (e.g. being a student) as the information while the opposite as the noise.

Following the above analogy, the IDCM approach is equivalent to the decoding procedure in the transmission of information in the information communication system in presence of channel noise. Therefore, the challenges of reconstructing the original message can result from both the transmission capability of the quasi choice-making channel and the decoding capability of the decoder, i.e. the IDCM approach in the present context. This corresponds to the properties of ill-posed inverse problems, which has been discussed previously in Section 3.3. The enrichment performance of the IDCM approach is hence analogous to its decoding capability for a specific quasi choice-making channel.

Observing this analogy allows drawing upon the concepts and metrics developed in information theory, so as to quantify the information (noise) in the study choice-making system. These metrics can further enable the quantitative measure of the covariance structures between the input and the output for the intended enrichment in the big dataset to be enriched. This capability, on the other hand, also opens avenues to exploring how the socio-demographic enrichment performance of the IDCM approach varies depending on the properties (quality) of the enriched big dataset. The metrics employed and established for the IDCM framework are explained later in Section 4.2.

3.5 Summary

In this chapter, the conceptual and analytical framework underlying the Inverse Discrete Choice Modelling (IDCM) methodology (Chapter 4) has been discussed from four aspects.

The first perspective concerns the basics of discrete choice models (DCMs), which was briefly explained in Section 3.1 because of two reason. First, DCMs are the key and fundamental component in the econometric formulation of the IDCM approach. Meanwhile, the IDCM performance theory is established based on the concept of constant recalibration, which represents a common technique extensively performed in traditional discrete choice modelling. In light of the relevance to the established IDCM framework, DCMs were introduced in Section 3.1 in terms of the following four dimensions: (i) the econometric formulation (Section 3.1.1), (ii) model components and structures (Section 3.1.2), (iii) parameter calibration process (Section 3.1.3), as well as (iv) the tests for the goodness-of-fit of the model (Section 3.1.4).

This was followed by Section 3.2, which presented a microeconomic model, using an example in the context of travel behaviour analysis, to conceptually explain the logic underlying the econometric formulation of the IDCM approach. This conceptual model could be viewed as

the departure point of the further derivation and manipulation of the IDCM approach, as will be documented later in Section 4.1.

Moreover, the procedure presented by the microeconomic model in Section 3.2 also acknowledged the need for suitable inverse techniques in solving the socio-demographic enrichment using the IDCM approach. This therefore provoked an overview of the umbrella field of inverse problems, as surveyed in Section 3.3, with respect to the challenges and solution techniques. Specifically, the difficulties associated with the solution of ill-posed (or ill-conditioned) inverse problems (Section 3.3.1), corresponding to the scope of the IDCM approach, suggested the value of the current study. The discussion around the sufficient techniques available for solving various inverse problems (Section 3.3.2), on the other hand, assures that the various form of the IDCM approach, including the current one and further extensions, can be solved.

Finally, Section 3.4 conceptually explained the analogy of choice-making behaviour to the transmission of messages over an information communication channel. Since a number of metrics have been formally established in information theory to quantify the information with respect to the input and the output of the information communication channel, such an analogy enables the use of these metrics to measure the sample distributions of the enriched attribute and the choice patterns, as well as their mutual dependence patterns. This further permits developing the ability of the IDCM framework to estimate and explain its enrichment performance under different data conditions (distributions). This is termed as the IDCM enrichment performance theory, and is systematically proven in the binary enrichment context later in Section 4.2.

4 INVERSE DISCRETE CHOICE MODELLING METHODOLOGY

Representing the core theoretical contribution developed in the current PhD study, the methodology of the Inverse Discrete Choice Modelling (IDCM) socio-demographic enrichment framework consists of the following two components:

- 1) the econometric formulation, termed as the IDCM approach, which provides how the proposed socio-demographic enrichment could be expressed and solved mathematically.
- 2) the accompanying theory, termed as the IDCM performance theory, which mathematically estimates (forecasts) the enrichment performance using the IDCM approach based on a developed approximation of the data distribution underlying the enrichment sample.

The current chapter therefore demonstrates how the IDCM framework is established in terms of the above two components. Specifically, Section 4.1 outlines the IDCM approach by presenting the detailed derivations for the econometric formulation and the guidelines (requirements and procedure) for implementing the IDCM enrichment in practice. On the other hand, Section 4.2 explains the IDCM performance theory which mathematically presents the link between the metrics measuring the enrichment performance of the IDCM approach and the data condition underlying the sample that requires socio-demographic enrichment, respectively. A summary of this chapter is provided in Section 4.3, which also relates the results developed in this chapter to the following chapters.

4.1 IDCM approach

This section presents the IDCM approach for enriching the socio-demographic attributes of anonymous big data in detail. In particular, Section 4.1.1 provides the detailed manipulation for deriving the econometric formulation of the IDCM approach based on the microeconomic, conceptual model in Section 3.2. In light of the derived econometric formulation, Section 4.1.2 then discusses the requirements and procedures for conducting the enrichment of socio-demographic attributes using the IDCM approach, which guides the empirical applications of the IDCM framework as documented later in Chapters 5 – 6.

4.1.1 IDCM Econometric formulation

As explained in Section 3.1.1 and restated below, random utility models (RUMs) assume that the full utility function U for choosing each choice alternative from a given choice set consist of a systematic utility V and an error component ε (Equation 3-5) such that the probabilistic nature of choice-making behaviour is permitted.

$$U_a(i) = V_a(i) + \varepsilon_a(i) = V(\mathbf{A}_{ai}, \mathbf{X}_a, \boldsymbol{\beta}_{ai}) + \varepsilon_a(i). \quad (4-1)$$

If the utility of each choice alternative in the choice set is not determined purely by the property of the alternative itself (or if the choice set contains only 2 choice alternatives), which is assumed in the following derivation, we can assume $V_i = 0$ for normalisation in that only the difference between utilities of alternatives matters. This leads to the probability of choosing an alternative $i \in C$ as expressed in Equation 3-6, and reformulated below as:

$$P_i = P(\varepsilon_j - \varepsilon_i \leq V_i - V_j \ \forall i, j \in C) = P(V_j \leq \varepsilon_i - \varepsilon_j \ \forall i, j \in C). \quad (4-2)$$

Section 3.1.2 states that different DCM structures exist due to the different assumptions made with regard to the distribution of the error term (ε). Generally, ε is defined to follow the probability density distribution $f(\cdot)$ which has a cumulative distribution function $F(\cdot)$. Consequently, Equation 4-2 can be further manipulated to obtain:

$$P_i = P(V_j \leq \varepsilon_i - \varepsilon_j \ \forall i, j \in C) = 1 - F(V_j) = S(V_j) \ \forall j \in C, i \neq j. \quad (4-3)$$

In light of the components of a typical linear-in-parameter (LIP) systematic utility V (Equation 3-15 in Section 3.1.3), P_i is effectively a function of the attributes of the choice alternatives ($\mathbf{A}$) and the attributes of the choice-maker ($\mathbf{X}$) mapped by the taste parameters ($\boldsymbol{\beta}$), i.e.:

$$P_i = S(\mathbf{A}, \mathbf{X}, \boldsymbol{\beta}). \quad (4-4)$$

The taste parameters $\boldsymbol{\beta}$ can be obtained by performing the parameter calibration upon a ‘thick’ dataset other than the sample that requires enrichment. Such a calibrated DCM is termed as an IDCM kernel. This indicates that, if P_i , $\mathbf{A}$ and $\boldsymbol{\beta}$ are known, $\mathbf{X}$ can be directly derived by inverting $S(\cdot)$, which is effectively computed as:

$$\mathbf{X} = S^{-1}(P_i | \mathbf{A}, \boldsymbol{\beta}). \quad (4-5)$$

In most empirical cases, however, the direct imputation of X using Equation 4-5 is cumbersome or even at all impossible because it is usually the choice Y instead of the probability $P_i = P(Y = i)$ that can be observed. This implies that Y is a random discrete variable distributed according to $S(\cdot)$ and parameterised by $\mathbf{A}, \mathbf{X}, \boldsymbol{\beta}$ as:

$$Y \sim S(\mathbf{A}, \mathbf{X}, \boldsymbol{\beta}). \quad (4-6)$$

In the current context, Equation 4-6 corresponds to the DCM for Y as a function of $\mathbf{A}$ and $\mathbf{X}$. Following the similar logic of the calibration of parameters as discussed previously in Section 3.1.3, the enrichment of $\mathbf{X}$ can thus be linked to the concept of likelihood maximisation. This effectively promotes establishing the econometric formulation of the IDCM approach as follows. Specifically, in light of a DCM with known $\boldsymbol{\beta}$ estimated upon a small ‘thick’ dataset, i.e. the IDCM kernel, Bayes’ Theorem provides a means to construct the likelihood of the choice-maker being characterised by a particular set of attributes $\mathbf{X}$ given a single observed Y and known $\mathbf{A}, \boldsymbol{\beta}$ as:

$$\mathcal{L}(\mathbf{X}) = P(\mathbf{X}|\mathbf{A}, Y, \boldsymbol{\beta}) = P(Y|\mathbf{A}, \mathbf{X}, \boldsymbol{\beta}) \frac{P(\mathbf{X}|\mathbf{A}, \boldsymbol{\beta})}{P(Y|\mathbf{A}, \boldsymbol{\beta})}. \quad (4-7)$$

As a result, the Bayesian maximum *a posteriori* (MAP) estimator of $\mathbf{X}$ can be derived as:

$$\mathbf{X}_{MAP}^* = \arg \max_{\mathbf{X}} P(Y|\mathbf{A}, \mathbf{X}, \boldsymbol{\beta}) P(\mathbf{X}|\mathbf{A}, \boldsymbol{\beta}). \quad (4-8)$$

Note that the term $P(Y|\mathbf{A}, \boldsymbol{\beta})$ is dropped from Equation 4-7 to obtain the MAP estimator of $\mathbf{X}$ in Equation 4-8 considering that this term is independent of $\mathbf{X}$. In cases where the prior with respect to $P(\mathbf{X})$ is unknown, the corresponding maximum likelihood (MLE) estimator can be derived to take the form:

$$\mathbf{X}_{MLE}^* = \arg \max_{\mathbf{X}} P(Y|\mathbf{A}, \mathbf{X}, \boldsymbol{\beta}). \quad (4-9)$$

Assume a sample containing a number of N choice-makers, in which each individual choice-maker a have made the choice between the available alternatives for a number of L_a times. As different choices made by a same individual a can possibly result in different enrichment outcome of $\mathbf{X}_a$ using the estimators established above, a joint log-likelihood is applied across all the L_a choice records of each choice-maker, to assure the uniqueness of the enriched attributes $\mathbf{X}_a$ for the same individual a . Based on Equation 4-7, the joint likelihood of a being characterised by $\mathbf{X}_a$ from the L_a choice observations can be derived as:

$$\mathcal{L}(\mathbf{X}_a) = \prod_{l=1}^{L_a} P(\mathbf{X}_a | \mathbf{A}_l, Y_l, \boldsymbol{\beta}). \quad (4-10)$$

Accordingly, the joint likelihood that the N individuals in the sample, each making L_a choices, are characterised by $\mathcal{X}$ (a vector of attributes $\mathbf{X}$ for the entire sample) can be derived as:

$$\mathcal{L}(\mathcal{X}) = \prod_{a=1}^N \prod_{l=1}^{L_a} P(\mathbf{X}_a | \mathbf{A}_l, Y_l, \boldsymbol{\beta}). \quad (4-11)$$

The conventional logarithmic transformation can be used to derive the suitable log-likelihood that should be maximised to find the solution to the MAP or MLE estimates of $\mathbf{X}$, i.e.:

$$\mathcal{X} = \max_{\mathbf{X}} \sum_{a=1}^N \sum_{l=1}^{L_a} [\log P(Y_l | \mathbf{A}_l, \mathbf{X}_a, \boldsymbol{\beta}) + \log P(\mathbf{X}_a | \mathbf{A}_l, \boldsymbol{\beta})] \quad (4-12)$$

It can be noted that the dimensionality of the optimisation problem presented in Equation 4-12 is $N \times |\mathcal{X}|$, where $|\mathcal{X}|$ denotes the cardinality of the vector $\mathcal{X}$. Therefore, the number of parameters to be estimated will grow with the sample, not adding degrees of freedom unless multiple choices made by same individuals are observed. This is different from the conventional sense in DCM parameter calibration where the optimisation dimensionality is fixed to $|\mathcal{X}|$. This hence provokes to assume that the individuals in the sample are independent of each other, i.e. independence of other individuals (IOI), which further enables the parallelisation of the maximisation procedure in Equation 4-12 as:

$$\begin{aligned} \mathcal{X} &= \max_{\mathbf{X}} \sum_{a=1}^N \sum_{l=1}^{L_a} [\log P(Y_l | \mathbf{A}_l, \mathbf{X}_a, \boldsymbol{\beta}) + \log P(\mathbf{X}_a | \mathbf{A}_l, \boldsymbol{\beta})] \\ &= \sum_a^N \max_{\mathbf{X}_a} \sum_{l=1}^{L_a} [\log P(Y_l | \mathbf{A}_l, \mathbf{X}_a, \boldsymbol{\beta}) + \log P(\mathbf{X}_a | \mathbf{A}_l, \boldsymbol{\beta})]. \end{aligned} \quad (4-13)$$

The right-hand side (RHS) of Equation 4-13 is thus directly adopted as the basic econometric formulation of the IDCM approach in the socio-demographic enrichment procedure.

It should be noted that the enriched socio-demographic attribute is sometimes of discrete nature. For example, gender is discrete nominal while income level is often recorded on an ordinal scale. This possibly makes the enrichment using the above IDCM approach either an

integer programming problem when all the socio-demographic attributes in $\mathbf{X}$ are of discrete nature or a mixed-integer programming problem when $\mathbf{X}$ consists of both discrete and continuous attributes. In such instances, the optimisation algorithms that rely on the smoothness of the function and existence of derivatives, e.g. the gradient-descent family (including Newton-Raphson and its derivatives), may fail to converge or even run when implementing Equation 4-13. This thus requires alternative methods to solve such problems.

Specifically, in circumstances where the attributes in $\mathbf{X}$ are purely discrete, the Brute-force search (BFS) can be applied if the number of candidate solutions that require evaluation is finite, which can also ensure the global optimality. Moreover, the IOI assumption and the consequent Equation 4-13 provide a way in which BFS can be implemented in high-performance, parallel computing environment, thus making it substantially manageable. However, the BFS is not feasible when $\mathbf{X}$ contain both continuous and discrete variables. In such cases, it is necessary to either discretise the attribute such that the BFS can be used or to employ approaches for solving mixed-integer programming problems, e.g. comparing conditional continuous optima for discrete candidate solutions or heuristics such as hill climbing or simulated annealing. Figure 4.1 summarises the criteria for selecting optimisation algorithms in performing the socio-demographic enrichment using Equation 4-13.

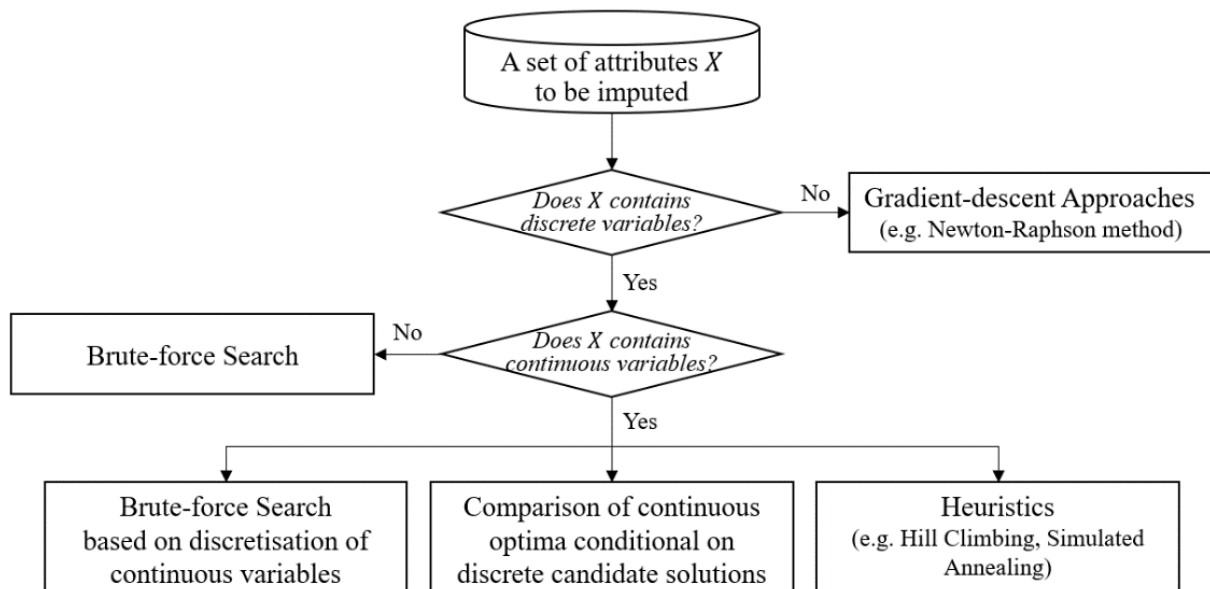


Figure 4.1 Selection of Algorithms to Solution of IDCM Approach

As is discussed above, the IOI assumption effectively enables the parallelisation of the maximisation procedure (Equation 4-13), making the enrichment using the IDCM approach substantially manageable. Nevertheless, it significantly simplifies people's behaviour patterns

by ignoring the dependence correlation of the choice-making behaviour between different individuals. For instance, the departure time-of-day choice made by the members of a household can depend on each other due to some correlated attributes, e.g. the number of vehicles owned by the household. Consequently, in the situations where the choices of a specific group of individuals can affect one another, the maximisation problem in Equation 4-12 can only be parallelised up to the level of groups, across which the choices are independent. Consider the case in which the above defined sample contains $\mathcal{H}$ households, each consisting of a number of N_h members who make choices that might be correlated. The corresponding expression for paralleled maximisation can therefore be derived from Equation 4-12 as:

$$\begin{aligned}\mathcal{X} &= \max_{\mathbf{X}} \sum_{a=1}^N \sum_{l=1}^{L_a} [\log P(Y_l | \mathbf{A}_l, \mathbf{X}_a, \boldsymbol{\beta}) + \log P(\mathbf{X}_a | \mathbf{A}_l, \boldsymbol{\beta})] \\ &= \sum_h^{\mathcal{H}} \max_{\mathbf{X}} \sum_a^{N_h} \sum_{l=1}^{L_a} [\log P(Y_l | \mathbf{A}_l, \mathbf{X}_a, \mathbf{X}_h, \boldsymbol{\beta}) + \log P(\mathbf{X}_a, \mathbf{X}_h | \mathbf{A}_l, \boldsymbol{\beta})].\end{aligned}\tag{4-14}$$

It is seen from the RHS of Equation 4-14 that a specific term $\mathbf{X}_h$ is presented to characterise the shared or correlated attributes between the members of the household to capture the effect of these attributes in the choice behaviour of individual decision-maker a . Whilst demonstrated using the household as an example for a group of individuals whose choices are not made independently from each other, it is argued that the formula in Equation 4-14 can address any other forms of group retaining dependence in choices, such as a class of students who visit a park to have a picnic or a person who chooses to go to a party as a companion of his/her friend.

In summary, Equation 4-12 represents the general econometric formulation of the IDCM socio-demographic enrichment approach while the most basic form of the econometric formulation can be found in Equation 4-13. The outcome for the enrichment performed based on Equation 4-13 (or its extensions in Equation 4-14) can be estimated using either the MAP (Equation 4-8) or the MLE (Equation 4-9) estimators, according to the availability of prior information about the enriched socio-demographic attribute. Based on the nature (i.e. type) of the enriched attribute, a guideline for selecting between different maximisation algorithm is provided (Figure 4.1) to solve the corresponding estimators for the IDCM enrichment.

4.1.2 Requirements and procedures for IDCM implementation

In light of the econometric formulation of the IDCM approach as well as the corresponding estimators as derived in Section 4.1.1, the current section summarises the socio-demographic enrichment for an anonymous ‘big-but-thin’ dataset using the IDCM approach as performing suitable inversion techniques to a DCM that is calibrated upon a ‘small-but-thick’ dataset from other sources. Such a DCM is termed as the IDCM kernel in the present study.

A typical socio-demographic enrichment process using the established IDCM approach consists of two steps: (1) the selection of variables (requirements), and (2) the implementation of the socio-demographic enrichment based on the derived econometric formulation (procedure). These two steps are explained respectively in the contents below.

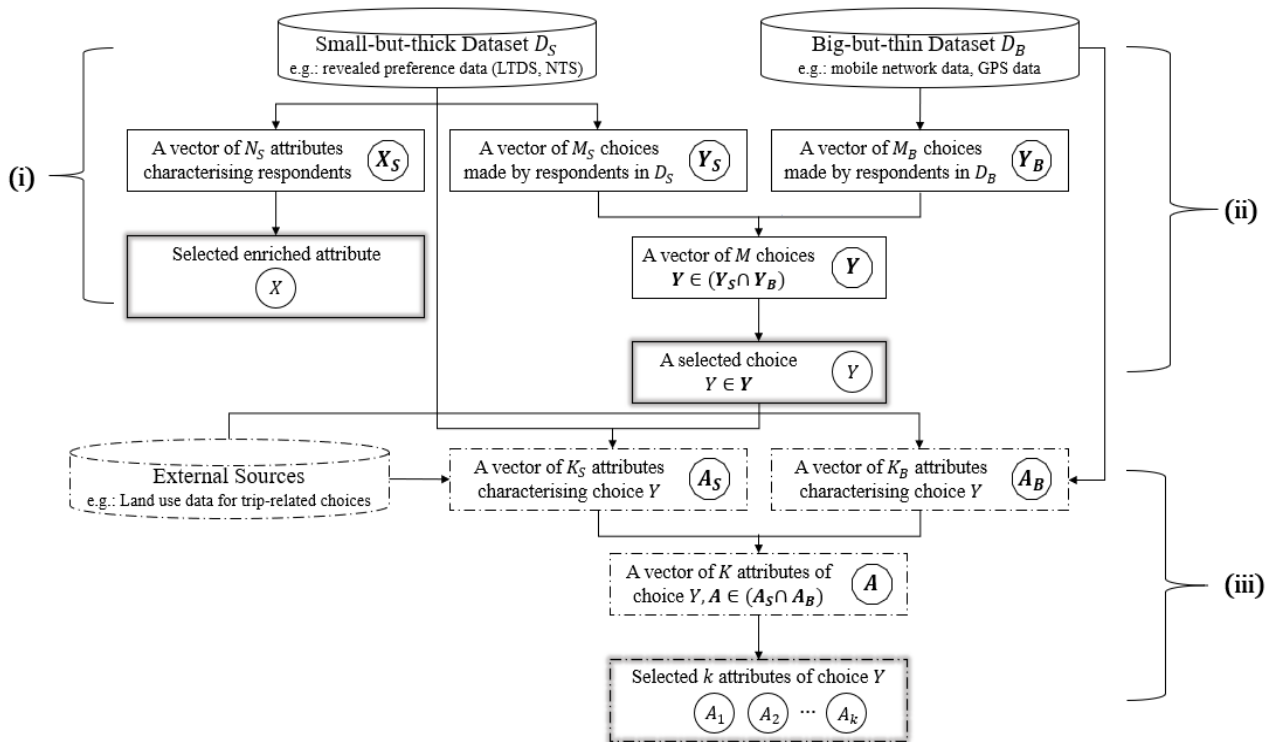


Figure 4.2 Selection of Required Variables for Socio-demographic Enrichment using IDCM Approach

(Key: boxes in dotted line refer to variables that are not compulsory for the enrichment)

First of all, the former variable selection step is necessary so as to determine the attribute(s) that could be enriched and the behaviour pattern(s) that could be extract as the input feature for the enrichment, based on the data availability. Figure 4.2 provides a flow chart for the first step, given the purpose of enriching the socio-demographic attributes of a ‘big-but-thin’ dataset (denoted as D_B) given an available ‘small-but-thick’ dataset (denoted as D_S). Particularly, the variables which need to be determined and selected during this step include:

- 1) The variable(s) $\mathbf{X}_S$ which could be enriched given the available data. $\mathbf{X}_S$ serves the set of attributes from which the actually enriched socio-demographic attribute X (if a single attribute is enriched at a time) are selected, i.e. $X \in \mathbf{X}_S$, as shown in **(i)** in Figure 4.2. The selection of $\mathbf{X}_S$ is determined purely given their availability in D_S .
- 2) The variable(s) $\mathbf{Y}$ that could be used as the input choice for the IDCM enrichment, which acts as the set of choices where the actual input choice variable Y (if a single choice variable is employed as the input) is selected, i.e. $Y \in \mathbf{Y}$. As can be seen in **(ii)** in Figure 4.2, $\mathbf{Y}$ should be extractable from both D_S and D_B .
- 3) The variable(s) that characterise the alternatives of the choice Y , i.e. the attributes (A) of choice alternatives, which can be extracted directly from D_S and D_B or could be obtained from external sources. They are optional for the IDCM enrichment and are thus boxed using the dotted line as shown in **(iii)** in Figure 4.2.

On the other hand, the second step for implementing the IDCM approach typically involves two stages: (i) the creation and estimation of an IDCM kernel, and (ii) the enrichment using the IDCM approach; which is as illustrated in Figure 4.3.

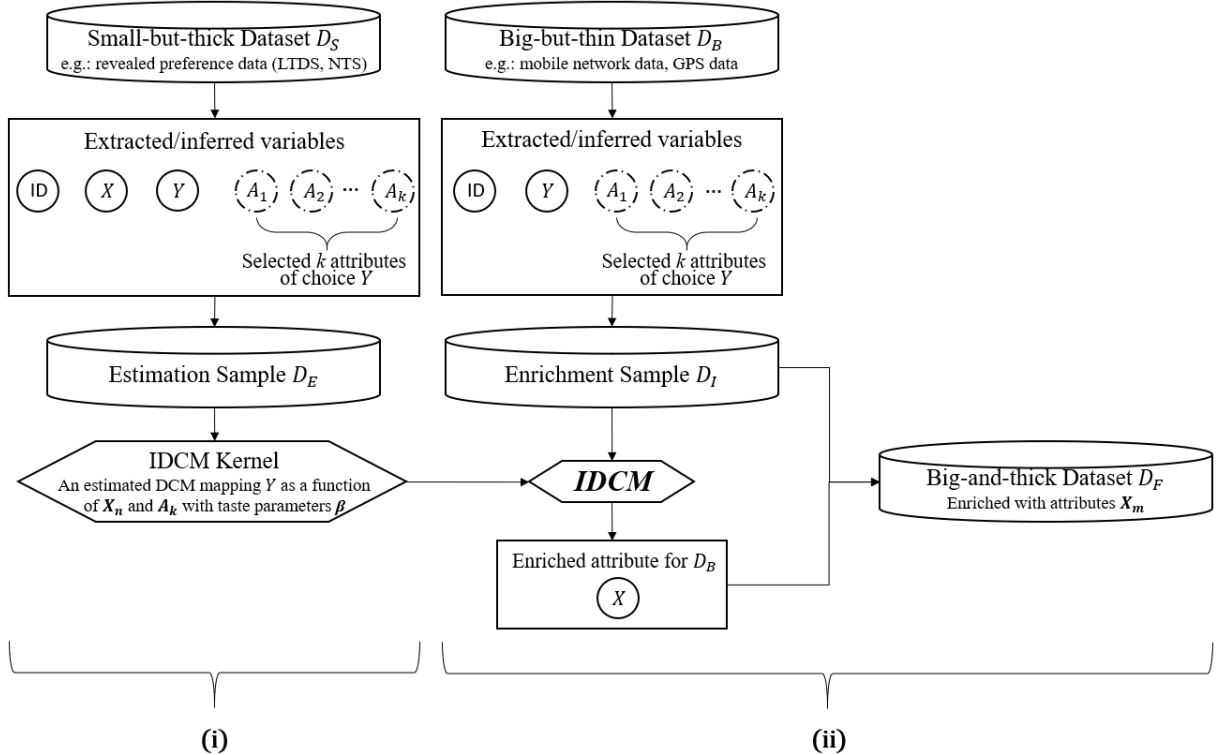


Figure 4.3 Procedure of Implementing Socio-demographic Enrichment using IDCM Approach

Specifically, the creation and estimation of an IDCM kernel, denoted by **(i)** in Figure 4.3, refers to the formulation and calibration of a DCM which captures the relationship between the

selected behaviour and enriched attribute. This can be achieved using an estimation sample (D_{Est}) which is extracted from D_S given the selected variables. In the second stage (ii), suitable estimation methods (as described previously in Section 4.1.1) are performed on the IDCM kernel to derive the socio-demographic attribute X for an enrichment sample (D_{Erc}) which contains observations of the choice behaviour Y extracted from D_B . Therefore, it can be concluded from Figure 4.2 and Figure 4.3 that there are three types of data involved in an IDCM enrichment process, including:

- 1) a ‘small-but-thick’ dataset (D_S), either revealed preference or stated preference, containing attributes of the respondents and the choices they have made, which allows the estimation of a DCM serving the IDCM kernel to be inverted in the subsequent enrichment process;
- 2) a ‘big-but-thin’ dataset (D_B) that must contain, or could be processed to infer, the same choice(s) as presented in the ‘small-but-thick’ D_S and requires enrichment to obtain one or more attributes captured in the estimated IDCM kernel; and
- 3) A ‘big-and-thick’, enriched dataset (D_F) which results from the ‘big-but-thin’ D_B with the values for the socio-demographic attributes enriched using the IDCM approach.

4.2 IDCM Enrichment Performance

As compared with other existing methods for socio-demographic enrichment, the advantage of the IDCM approach developed in Section 4.1 is assumed to rest on its solid grounding in microeconomic behavioural theories. This provides the potential to mathematically estimate its enrichment performance based on the specified data conditions or even across application contexts. The sections below hence outline how the enrichment performance of the IDCM approach is linked to the metrics representing the data condition, i.e. the covariance structure of the input behaviour feature and the enriched attribute underlying the enrichment sample.

Specifically, Section 4.2.1 defines the indicators for measuring the post-implementation enrichment performance of the IDCM approach at aggregate (sample) and disaggregate (individual) levels. On the other hand, Section 4.2.2 establishes a new set of metrics that can be used to quantify the relative importance of one variable in gaining insights into the other variable, which is termed as the explanatory power of the former in describing the latter. Section 4.2.3 then demonstrates how these metrics can be linked to each other through the systematic development of the IDCM enrichment performance theory.

4.2.1 Enrichment performance assessment indicators

The enrichment performance of the IDCM approach can be assessed at the aggregate (sample) level and the disaggregate (individual) level, respectively. In particular, as is documented in Section 4.2.1.1, the sample-level enrichment performance evaluates the level of similarity between the marginal probability distribution of the socio-demographic attribute enriched using the IDCM approach and the actual market share of the enriched attribute, if it were known, in the enrichment sample. In contrast, the disaggregate-level enrichment performance (Section 4.2.1.2) quantifies the proportion of the respondents in the enrichment sample whose socio-demographic attributes are correctly enriched using the IDCM approach.

4.2.1.1 Aggregate (sample) level: sample enumeration and market share comparison

At the aggregate sample level, the analysis concerning the IDCM enrichment performance can be conducted based on sample enumeration, which is the most straightforward aggregation technique commonly used in discrete choice modelling (Train, 2003).

Conventionally, sample enumeration is used to estimate the market share of the choice in the population based on an estimated DCM for this choice. In particular, for each of the choice alternatives, sample enumeration averages the sum of the probabilities of this alternative for all the weighted choice-makers across a sample drawn from the population, which represents the estimated market share for this choice alternative. Consider a sample of N_s choice-makers (labelled as $a = 1, 2, \dots, n_s$) drawn from the population with size N . Each choice-maker a is attached with a weight w_a to represent the number of similar choice-makers in the population. Suppose for each choice-maker a , the estimated DCM gives a probability $p_a(y_j)$ for choosing alternative y_j from the choice set Y . The corresponding sample-enumeration-based estimate for the market share of choice alternative y_i is defined as:

$$W(y_j) = \frac{1}{N} \sum_a^{N_s} w_a p_a(y_j), \quad (4-15)$$

where N is the size of the population;

$W(y_j)$ is the sample-enumeration-based market share estimate for alternative y_j .

If the above sample is drawn purely by simple random sampling, w_a is the same for all the choice-makers across the drawn sample, which equals N/N_s . The corresponding enumerated marked share is therefore be calculated according to:

$$W(y_j) = \frac{1}{N_s} \sum_a^{N_s} p_a(y_j). \quad (4-16)$$

A similar sample-enumeration-based estimate for the market share of the socio-demographic attribute to be enriched using the IDCM approach can be derived, by assuming that the representativeness, i.e. the weight, of all the anonymous respondents in the enrichment sample is the same. Particularly, as can be found from the expressions in Equations 4-8 and 4-9 (Section 4.1), the solutions to the MAP and the MLE estimators of the IDCM approach provide an exact outcome value of $\mathbf{X}$ instead of a probability distribution. This means that, for a specific discrete attribute X enriched using either of the estimators of the IDCM approach, the probability of a possible outcome values (POV) of X equals 1 if the approach derives the same outcome for this POV, and 0 otherwise. Therefore, for an enriched sample of N_s respondents, the enumerated marginal probability for each specific POV, i.e. $x_i \in X$, in the enriched sample using a particular estimation method can be expressed as:

$$W(x_i) = \frac{1}{N_s} \sum_a^{N_s} \delta_a(x_i), \quad (4-17)$$

where $\delta_a(x_i)$ is an indicator equalling 1 when respondent a is enriched with x_i .

The likelihood of X given by the MAP estimator (the RHS of Equation 4-8) also provides the opportunity for probabilistic analysis of the IDCM approach, i.e. to calculate the enumerated marginal probability distribution of X given the likelihood of each individual choice-maker being characterised by a specific POV of the attribute. For $x_i \in X$, such likelihood-based market share can be enumerated as:

$$W(x_i) = \frac{1}{N_s} \sum_a^{N_s} p_a(x_i|y_j) = \frac{1}{N_s} \sum_a^{N_s} p_a(y_j|x_i) \frac{p_a(x_i)}{p(y_j)}. \quad (4-18)$$

where $p_a(x_i|y_j)$ is the likelihood that respondent a is enriched with x_i for attribute X given observed y_j for the choice Y ;

- $p_a(y_j|x_i)$ is the probability that respondent a characterised by x_i chooses y_j ;
- $p_a(x_i)$ is the market share for the POV x_i of attribute X ;
- $p(y_j)$ is the market share for the alternative y_j of choice Y ;

It should be noticed, however, that the enumerated market share for the POV of the enriched socio-demographic attribute provided by Equation 4-18 is only applicable for the MAP estimator as deriving such an enumerated market share requires knowledge of information about the marginal probability distribution of the enriched attribute, viz. the prior.

The sample-enumeration-based estimate developed using a uniform weight is recommended if the enrichment sample is a good representative of the population (de Dios Ortuzar and Willumsen, 2011). This therefore suggests that the estimate for the market share of the socio-demographic attributed developed based on the above assumption could be biased if the enrichment sample is not randomly selected (collected) from the population. However, it is typical for big datasets to fail to account for the population as they are often collected for specific strata, e.g. mobile phone users. Acknowledging the possibility of biases, a same weight is assumed in the above development of the sample-enumeration-based estimate because the weight depends on the stratification sampling strategies adopted in collecting the data while the socio-demographic information which affects the stratification of the data is unavailable. This indicates that the aggregate-level enrichment performance of the IDCM framework evaluated based on the sample-enumerated market shares of the enriched attribute is not transferable between different datasets.

To evaluate the enrichment performance of the IDCM approach at the aggregate sample level, the enumerated market share of the attribute enriched using each estimator of the IDCM approach is compared to the actual marginal probability distribution of the attribute in the enrichment sample if it was observable. The level of consistency between the compared market shares can be formally measured using Pearson's Chi-squared test (Plackett, 1983) or Fisher's exact test (Fisher, 1922). The null hypothesis for these tests proposes that the market shares for the POVs of the attribute enriched using the IDCM approach are significantly similar to the actual ones revealed in the enrichment sample.

4.2.1.2 Disaggregate (individual) level: percentage correctly predicted

At the disaggregate individual level, the enrichment performance of the IDCM approach can be measured using the ‘percentage correctly predicted’ (PCP) metric, which effectively calculates the proportion of the enriched data points that are consistent with observed or simulated data in the sample (Scott, Hosmer and Lemeshow, 1991):

$$PCP = \frac{100}{N_s} \sum_a^{N_s} \delta_a, \quad (4-19)$$

where δ_a is an indicator equalling 1 if the enriched attributes for respondent a is consistent with that of the observed data and 0 otherwise.

In the current context where the performance of the IDCM socio-demographic enrichment approach is assessed, the PCP effectively refers to the respondents in the enrichment sample whose socio-demographic attribute are consistent with the corresponding value enriched using the IDCM approach. Note that if the considered number of the enriched attributes is more than 1, $\delta_a = 1$ when and only when all the study attributes are correctly enriched. Although the PCP represents an overall statistic for the entire sample, it essentially measures the enrichment performance of approaches at the individual level since it determines whether the enriched result for each respondent is correct while the composition of the sample is not considered.

The PCP can be viewed as a counterpart to the ‘percentage of correct classification’ or ‘percentage correctly classified’ (PCC) which are used in the assessment of accuracy estimation for classification of machine learning models (e.g. Winters-Hilt *et al.* (2006)). In particular, the PCC is calculated as the number of correct predictions divided by the total number of classifications performed. The PCP is hence used to measure the disaggregate-level enrichment performance of the other methods explored in the comparative analysis (Chapters 5 and 6). However, it should be noted that, for existing methods, PCPs can only be used to measure the post-implementation enrichment performance. In other words, such enrichment performance of existing methods cannot be predicted or estimated prior to the actual enrichment. For the proposed IDCM framework, on the other hand, the PCP is also used as the indicator for the enrichment performance in the IDCM performance theory, which is established later in Section 4.2.3 to enable the ability of the IDCM framework to estimate

(predict) and interpret (explain) its enrichment performance based on the data distribution underlying the enrichment sample.

4.2.2 Explanatory power of variables

The explanatory power of one variable in inferring the other measures the relative importance of knowing the value of the former variable in determining or understanding the value of the latter. In the context of the IDCM enrichment framework, the explanatory power of variables can be measured from two aspects:

- 1) The relative importance of the enriched attribute X in influencing the input behaviour feature Y . This is consistent with microeconomic theory which suggests that people's choice-making behaviour patterns is affected by their socio-demographic attributes; and
- 2) The relative importance of observing a choice behaviour Y in inferring the attribute X to be enriched. This is in keeping with the aim of this research as Y represents the input feature of the enrichment which can be obtained while X serves as the output that remains unknown.

Therefore, two metrics based on the above two aspects are created to measure the explanatory correlation between the input choice behaviour pattern and the enriched attribute using the IDCM enrichment. Specifically, Section 4.2.2.1 outlines the adjusted McFadden's *LRI* as an indicator for the explanatory power of the enriched attribute in affecting the behaviour pattern. On the other hand, the enrichment efficiency metric of the quasi choice-making communication channel is formulated in Section 4.2.2.2, based on several quantities that haven been formally established and employed in information theory, for measuring the explanatory power of the behaviour in inferring the socio-demographic attribute.

Although the developed IDCM approach (Equations 4-8 and 4-9) is applicable for the simultaneous enrichment of multiple attributes, the discussion below concerns the enrichment of a single attribute X from a single behaviour pattern in that the explanatory power defined above concerns the explanatory correlation only between a pair of variables.

4.2.2.1 Adjusted McFadden's *LRI*

For characterising the relative importance of the enriched attribute in affecting the behaviour pattern, the adjusted McFadden's *LRI* is established to describe the change in the log-likelihood due to the inclusion of the enriched socio-demographic attribute in the DCM

specification for the corresponding choice behaviour pattern. The compared DCMs are both estimated based upon the estimation sample. The adjusted LRI is termed as such because it originates from the McFadden's LRI (Equation 3-20), for measuring the overall GOF of DCMs. Representing the explanatory power of an enriched attribute X in influencing the choice behaviour Y , the metric of adjusted LRI can be derived, based on Equation 3-20, as:

$$LRI_{Adj} = 1 - \frac{\ln \mathcal{L}_{With_X}}{\ln \mathcal{L}_{Without_X}}, \quad (4-20)$$

where $\mathcal{L}_{With_X}$ is the maximum likelihood of the estimated DCM for Y with a specification including attribute X ;
 $\mathcal{L}_{Without_X}$ is the maximum likelihood of the estimated DCM for Y with a specification without attribute X .

The merit of this adjusted LRI metric is that it retains the dependence pattern between the enriched attribute and the choice behaviour. It has thus been used as the indicator for the explanatory power of the enriched socio-demographic attribute in the initial investigation into the enrichment performance of the IDCM approach by Zhao *et al.* (2018). Particularly, the results of the study conducted by Zhao *et al.* suggested that the enrichment performance of the IDCM approach can be improved as the adjusted McFadden's LRI grows. However, Section 3.1.4 has explained that the LRI is a relatively informal metric unable to quantify the proportion of the total variation that can be attributed to the inclusion of the enriched attribute in the DCM. Correspondingly, this also limits the use of adjusted LRI in quantitatively measuring the variation of the IDCM enrichment performance with respect to the represented explanatory power. Moreover, the adjusted LRI presents the explanatory correlation between the behaviour and the attribute in a way that differs from the purpose of the enrichment *per se*. Conversely, it would be ideal if the contribution of the observed values for the input behaviour feature in understanding the value of the enriched attribute can be identified, in accordance with the essential logic underlying the enrichment problem addressed by the IDCM approach. Therefore, a metric with firm theoretical underpinnings is needed to explore the postulated link between the two.

4.2.2.2 Mutual information, entropy, and channel efficiency

Following the analogy of choice-making behaviour to in an information communication system as outlined previously in Section 3.4, the metric representing the relative importance of the behaviour Y in understanding the enriched attribute X can be established. In particular, the analogous choice-making communication channel allows to borrow several concepts that have been formally and systematically developed in information theory based on statistical mechanics, to quantify the information transmitted over such a quasi-communication channel. These information-theoretic concepts as well as the established explanatory power metric of Y in inferring X are explained in the paragraphs below.

In information theory, mutual information describes the amount of information one can infer about one random variable based on knowledge of the other random variable (Cover and Thomas, 2005). The mutual information $I(X, Y)$ between X and Y quantifies the average amount of information (in bit) that can be gained about Y after observing a single value of X and *vice versa*. For discrete variables, $I(X, Y)$ is mathematically defined as:

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log_2 \left[\frac{p(x, y)}{p(x)p(y)} \right]. \quad (4-21)$$

where $p(x, y)$ are the joint probability density function for X and Y ;
 $p(x)$ and $p(y)$ are the marginal probability density functions for X and Y .

However, $I(X, Y)$ itself fails to describe the explanatory power of Y in understanding X as it measures the exact value of the amount of information that can be obtained about X from the observation of Y . This is because that $I(X, Y)$ measures the absolute amount of information that can be obtained about X given Y while the explanatory power depends also on the amount of information that remains needed, i.e. the amount of remaining uncertainty about X after observing Y . In other words, the amount of information required to infer different variables can be different in that the uncertainty (randomness) associated with the variables themselves can be distinguished from each other.

To understand this intuitively, consider two attributes X_1 and X_2 which represent gender and being a millionaire, respectively. Assume the mutual information between an observed behaviour Y and either X_1 or X_2 is the same, i.e.:

$$I(X_1, Y) = I(X_2, Y). \quad (4-22)$$

As gender is intuitively more evenly distributed across the population than being a millionaire, it can be derived that more uncertainty (randomness) is associated with gender as compared to being a millionaire. As a result, more additional information is required to obtain the gender of an individual than inferring whether he/she is a millionaire given the choice Y which however provides the same amount of information about X_1 and X_2 . It can be understood from the above example that, the relative importance (explanatory power) of knowing Y in understanding X_1 and X_2 , which is represented by the proportion of the total uncertainty shared by the mutual information, can vary between attributes due to the inherently different uncertainty of X_1 and X_2 . This motivates using the notions of entropy and conditional entropy in addition to mutual information to measure the explanatory power of the behaviour in inferring the attribute.

Shannon's entropy (typically referred to as just entropy) quantifies the average amount of uncertainty associated with a random variable according to the probability distribution of its POVs (Shannon, 1948). For a discrete random variable X having n categories distributed according to $P(X)$, the entropy (in bits) is defined as:

$$H(X) = - \sum_{i=1}^N p(x_i) \log_2 p(x_i). \quad (4-23)$$

The entropy of a variable also characterises its variability (randomness) over its POVs. In the limiting cases, the entropy reaches 0 (minimal) if one of its POVs is certain while the maximum is obtained when all POVs are identically distributed. For instance, the maximum entropy for binary variables is 1 bit, which can be calculated as:

$$H(X) = -2 \times 0.5 \log_2 0.5 = 1 \text{ bit}. \quad (4-24)$$

In other words, the maximum entropy is a situation in which the binary variable is evenly distributed, i.e. both POVs are equally likely. In that case, observing either of its POVs would produce 1 bit of information. Figure 4.4 (below) shows the variation of the entropy of a binary variable X versus the probability of one POV.

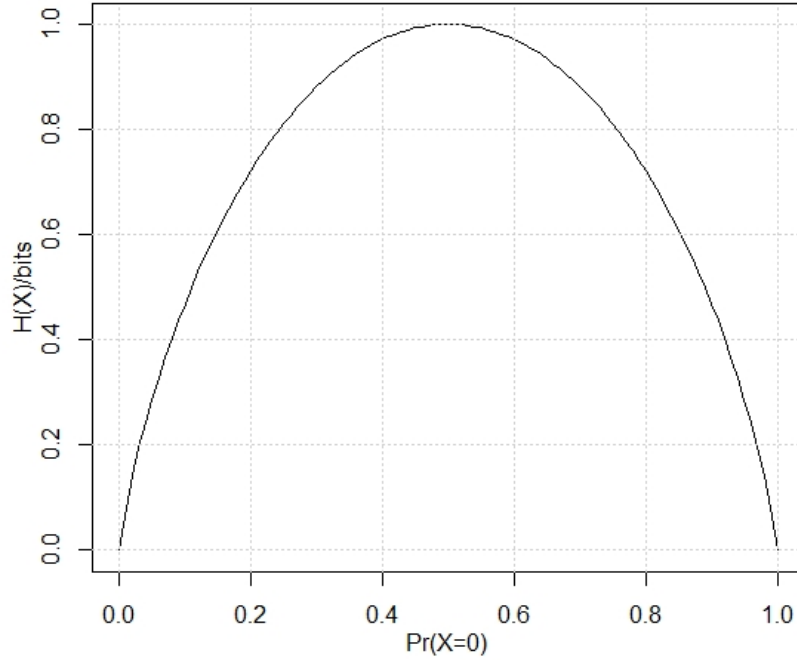


Figure 4.4 Entropy of Binary Variable vs Probability of One Possible Outcome Value

From the definition of $H(X)$, the relative certainty associated with a random variable X can be defined as the difference between the maximum uncertainty (entropy) that can be achieved based on the number of POVs and the actual uncertainty associated with the variable, i.e.:

$$CE(X) = H_{max} - H(X). \quad (4-25)$$

For binary variables, the maximum entropy is 1 bit as suggested by Equation 4-24. The corresponding certainty of a binary variable is hence equivalent to:

$$CE(X) = 1 - H(X). \quad (4-26)$$

On the other hand, conditional entropy quantifies the average amount of information that remains needed to describe the outcome of a random variable X as the value of another random variable Y is known (Stone, 2015). It is thus defined as the difference between the entropy of X and the mutual information between X and Y , i.e.:

$$H(X|Y) = H(X) - I(X, Y). \quad (4-27)$$

It is hence seen from Equations 4-21, 4-23, and 4-27 that, for the above intuitive example, even if the amount of information about X_1 obtained from Y equals that for X_2 , i.e. $I(X_1, Y) = I(X_2, Y)$, the average amount of information remaining needed to infer X_1 and X_2 are different:

$$H(X_1|Y) \neq H(X_2|Y). \quad (4-28)$$

This is interpreted by the difference in the entropies of X_1 and X_2 , i.e. $H(X_1) \neq H(X_2)$. It can thus be derived that the same amount of mutual information may provide different insights into different enriched attributes, depending on the inherent randomness of the enriched attributes themselves, i.e. whether the POVs of the enriched attribute are uniformly distributed, in which cases there will be a lot of insight, or whether they are already larger pre-determined (low entropy), in which case not much further information can be gained.

Based on the above information-theoretic notions, the proportion of the output entropy that is also shared by the input can be obtained, which measures how efficiently the information is communicated from the input to the output. This, in information theory, is defined as the transmission efficiency of the channel and can be calculated as (Stone, 2015):

$$\xi = \frac{I(X, Y)}{H(Y)} = \frac{\sum_{y \in Y} \sum_{x \in X} p(x, y) \log_2 \left[\frac{p(x, y)}{p(x)p(y)} \right]}{-\sum_{x \in X} p(y) \log_2 p(y)}, \quad (4-29)$$

where $H(Y)$ the entropy of the output variable Y of the information channel.

By quantifying the proportion of the output entropy which is composed respectively of information (input) and the noise, the metric of transmission efficiency ξ in Equation 4-29 captures the effectiveness of the corresponding communication channel in transmitting the intended information from the input to the output. However, in the current study where the IDCM approach serving as the decoder of the quasi choice-making communication channel, the primary interest is the decoding capability of the IDCM approach. In this regard, a similar quantity to ξ is thus analogously established, to measure the extent to which the output can be decoded into the input. Specifically, this metric is termed as the *enrichment efficiency*, denoted by η and expressed as:

$$\eta = 1 - \frac{H(X|Y)}{H(X)} = \frac{I(X, Y)}{H(X)} = \frac{\sum_{y \in Y} \sum_{x \in X} p(x, y) \log_2 \left[\frac{p(x, y)}{p(x)p(y)} \right]}{-\sum_{x \in X} p(x) \log_2 p(x)}. \quad (4-30)$$

where $H(X)$ is the entropy of the input variable of the information channel.

$H(X|Y)$ is the conditional entropy of the input variable X given the output Y .

$I(X, Y)$ is the mutual information between the input X and the output Y .

Taking into account both the variability of X and the mutual dependence between X and Y , the enrichment efficiency metric quantifies the explanatory power of Y in describing X in the context of the quasi choice-making communication channel. The enrichment performance of the IDCM approach, i.e. the capability of the IDCM decoder in decoding X from Y , is also explored with respect to this metric.

4.2.3 IDCM enrichment performance versus explanatory power of choice in enriched attribute

The ability of the IDCM framework to characterise (estimate and interpret) its enrichment performance can be enabled by its firm grounding in microeconomic theories and supported by information theory in light of the analogy of choice-making to information transmission. In the present thesis, this newly developed capability of the IDCM framework is named as the IDCM performance theory. The IDCM performance theory concerns the mathematical representation of the disaggregate-level enrichment performance (quantified by PCP) according to the underlying data condition, as measured by the explanatory power metric. The sections below explain in detail how the PCPs of the IDCM using both the MLE and the MAP estimators can be mathematically linked to the enrichment efficiency (η).

Specifically, Section 4.2.3.1 outlines the enrichment context for the current analysis, in terms of the type of the variables and format of the IDCM kernel, as well as how the estimators and the corresponding mapping rules are defined in the given enrichment context. Section 4.2.3.2 then suggests how the joint probability distribution $P(X, Y)$ underlying the enrichment sample can be approximated using reversed recalibration of constant. This is followed by Section 4.2.3.3 which derives the expression of η as a function of the assumed recalibrated constant ($\hat{\beta}_0$) and explores the corresponding variation pattern. Meanwhile, the representations of the MLE- and the MAP-PCPs using $\hat{\beta}_0$ are also derived in Sections 4.2.3.4, along with the discussion regarding their variation patterns. Section 4.2.3.5 subsequently demonstrates how the MLE-/MAP-PCP can be linked to the η metric using $\hat{\beta}_0$ as a pivot while Section 4.2.3.6 compares the patterns of the MLE- and the MAP-PCPs and hence proposes the strategy to select the best-performing yet information-effective estimation method between the two.

4.2.3.1 Enrichment context settings

For simplicity and tractability, the analysis performed in the current thesis is proven only in the binary enrichment context, i.e. the enrichment of a single binary socio-demographic attribute X given a single observation of a single binary choice behaviour Y (per individual).

Assume a ‘big-but-thin’ dataset D_B , which is processed to obtain the choice behaviour Y between two alternatives (y_0 and y_1) made by the anonymous respondents in D_B . This is treated as the enrichment sample D_{Erc} while the two choice alternatives are viewed as the two POVs of a random binary behaviour variable Y . Consider a binary socio-demographic attribute X to be enriched for D_B . The POVs of X are labelled 0 and 1 respectively to represent an indicator that captures the absence or presence of the categorical effect when these POV shifts:

$$X = \{x_0 = 0, x_1 = 1\} \text{ and } Y = \{y_0, y_1\}. \quad (4-31)$$

The IDCM kernel adopts the form of a binary logit (BL) model with LIP utilities, which characterises the probability of choosing each alternative (i.e. POV) of depending on the values of X . Based on the format of LIP utilities as presented in Equation 3-15, the utility functions for the current IDCM kernel including purely the binary attribute X can be expressed as:

$$\begin{aligned} V(y_0) &= 0 \\ V(y_1) &= \beta_0 + \beta_1 X. \end{aligned} \quad (4-32)$$

where $\beta = \{\beta_0, \beta_1\}$.

Accordingly, the form of logit models (Equation 3-14) leads to the current IDCM kernel as:

$$P(Y = y_j | X = x_i) = \frac{e^{V(y_j|x_i)}}{\sum_{y_j \in Y} e^{V(y_j|x_i)}}. \quad (4-33)$$

Suppose the availability of an estimation sample D_{Est} (a subset of D_S) which contains information about both X and Y . The taste parameters β of the above IDCM kernel are calibrated upon D_{Est} using standard DCM calibration procedures [45], which is subsequently inverted using the IDCM approach to obtain the absent X for D_{Erc} . Although the attributes of choice alternatives (A) can play a significant role in determining some choice behaviours, e.g. the travel time attribute in travel mode choices (Ben-Akiva, 1974), they are viewed as optional in the IDCM enrichment and not considered in the analysis below. The possible outcome of

this will be discussed later in Chapter 7. Hence, the MAP and the MLE estimators (Equations 4-8 and 4-9) are simplified for the enrichment of a particular socio-demographic X of an individual given a single choice observation Y as:

$$X_{MAP} = \arg \max_X P(Y|X)P(X). \quad (4-34)$$

$$X_{MLE} = \arg \max_X P(Y|X). \quad (4-35)$$

Recall the student and school attendance example (Section 3.4) which suggests the need for a mapping rule to define information and noise in a relative manner. As $\beta_1 > 0$ intuitively means that individuals being characterised by x_1 have higher propensity to choose y_1 , the information and noise when $\beta_1 > 0$ is signified as below.

$$\begin{aligned} \text{Information: } x_0 &\rightarrow y_0 \text{ and } x_1 \rightarrow y_1 \\ \text{Noise: } x_0 &\rightarrow y_1 \text{ and } x_1 \rightarrow y_0 \end{aligned} \quad (4-36)$$

The above mapping rule can be swapped when $\beta_1 < 0$ while the underlying logic is the same. Additionally, the IDCM kernel with a negative β_1 can always be expressed using an equivalent $\beta_1 > 0$ by swapping labels for the POVs of X or Y . Thus, $\beta_1 > 0$ is assumed in the following contents yet without loss of generality of the analysis. Similarly, the probability of Y when $Y = y_0$ is assumed to be no greater than 0.5, i.e. $p(y_0) \leq 0.5$. This is because the probability distribution of Y over $[0.5, 1.0]$ can be reflected equivalently by swapping the labels of the two POVs of Y .

Based on the above settings, the MAP estimator using the IDCM approach in Equation 4-34 for the current binary context can be transformed equivalently as:

$$\begin{aligned} X_{MAP} &= \arg \max_X P(Y|X)P(X) = \arg \max_{x_i \in X} p(y_j|x_i)p(x_i) \\ &= \begin{cases} \arg \max_{x_i \in X} \{p(x_0, y_0), p(x_1, y_0)\} & (Y = y_0) \\ \arg \max_{x_i \in X} \{p(x_0, y_1), p(x_1, y_1)\} & (Y = y_1) \end{cases} \end{aligned} \quad (4-37)$$

As $\beta_1 > 0$ is assumed, the corresponding MLE estimator defined by Equation 4-35 can be analogously derived for the current context as:

$$X_{MLE} = \arg \max_X P(Y|X) = \begin{cases} \arg \max_{x_i \in X} p(y_0|x_i) = x_0 & (Y = y_0) \\ \arg \max_{x_i \in X} p(y_0|x_i) = x_1 & (Y = y_1). \end{cases} \quad (4-38)$$

Finding the value of the MAP and MLE estimators in Equations 4-34 and 4-35 serve the complete solution to the IDCM approach for the binary enrichment context while the IDCM performance theory aims to express the enrichment performance obtained from using the above estimators. The settings above hence serve as the building blocks for the IDCM performance theory developed in the following sections.

4.2.3.2 Enrichment sample distribution versus assumed recalibrated constant

The expression of enrichment efficiency (defined in Equation 4-30 and restated below) reveals that the η metric is formulated as a function of the joint probability distribution of the input and the output of the enrichment as well as their marginal distributions.

$$\eta = \frac{\sum_{y \in Y} \sum_{x \in X} p(x, y) \log_2 \left[\frac{p(x, y)}{p(x)p(y)} \right]}{-\sum_{x \in X} p(x) \log_2 p(x)}. \quad (4-39)$$

In the context of a choice-making communication channel, η is effectively interpreted as a function of the joint probability distribution $P(X, Y)$ of the behaviour Y and the enriched attribute X , as well as their marginal distributions $P(X)$ and $P(Y)$, in the enrichment sample. Nevertheless, in reality, $P(X, Y)$ of a ‘big-but-thin’ dataset is hardly available because the respondents are anonymous and X cannot be observed. Furthermore, the representation of η as a function of the joint probabilities fails to demonstrate the smooth change in the η metric in respect to the variation of the data condition.

To address the above conundrum, the current study proposes that η can be linked to an approximation of $P(X, Y)$ that could be obtained based on the recalibration of constant. Recall Section 3.1.3 which describes the recalibration of constant for DCMs as a common technique used in traditional discrete choice modelling. In forecasting, a typical constant recalibration process involves re-calibrating the alternative specific constant (i.e. β_0 in Equation 4-32) in a DCM estimated from one sample for the choice prediction of the other sample. This can be performed by incorporating the knowledge of the market share information of the explanatory (socio-demographic) attribute $P(X)$ of the new sample. Such a recalibrated DCM can approximately reproduce the choice behaviour $P(Y|X)$ of the new sample, thus predicting an approximated $P(Y)$ of the new sample. The above constant recalibration also enables approximating $P(X, Y)$ of the new sample in that:

$$P(X, Y) = P(Y|X)P(X). \quad (4-40)$$

The recalibration of constant described above can be applied in a reverse manner when $P(Y)$ is obtainable while $P(X)$ remains unknown, which is the case of the IDCM enrichment. Specifically, information about the choice behaviour $P(Y|X)$ is treated as known in presence of the IDCM kernel and the market share of the choice $P(Y)$ can be obtained from the enrichment sample. An approximation of the recalibrated constant manipulated based on $P(Y|X)$ and $P(Y)$ can be further used to approximate the unknown $P(X)$. Since such a recalibrated constant is actually unknown, it is termed as the assumed recalibrated constant and denoted using $\hat{\beta}_0$. Treating $\hat{\beta}_0$ as a random continuous variable can enable the use of a single variable to approximate the smooth change in the data condition and thus the variation in $I(X, Y)$, $H(X)$ and η as derived in the following paragraphs.

The LIP utility for the current binary enrichment context (Equation 4-32) and the corresponding IDCM kernel with a BL structure (Equation 4-33) can jointly produce the equation set below.

$$\begin{cases} p(y_0|x_0) = \frac{p(x_0, y_0)}{p(x_0)} = \frac{e^{V(y_0|x_0)}}{e^{V(y_0|x_0)} + e^{V(y_1|x_0)}} = \frac{1}{1 + e^{\hat{\beta}_0}} \\ p(y_0|x_1) = \frac{p(x_1, y_0)}{p(x_1)} = \frac{e^{V(y_0|x_1)}}{e^{V(y_0|x_1)} + e^{V(y_1|x_1)}} = \frac{1}{1 + e^{\hat{\beta}_0 + \beta_1}}. \end{cases} \quad (4-41)$$

This equation set can further be reformulated to obtain:

$$\begin{cases} p(x_0, y_0) = \frac{p(x_0)}{1 + e^{\hat{\beta}_0}} \\ p(x_1, y_0) = \frac{p(x_1)}{1 + e^{\hat{\beta}_0 + \beta_1}}. \end{cases} \quad (4-42)$$

The above two joint probabilities sum to the marginal probability of Y when $Y = y_0$, i.e.:

$$p(y_0) = p(x_0, y_0) + p(x_1, y_0) = \frac{p(x_0)}{1 + e^{\hat{\beta}_0}} + \frac{p(x_1)}{1 + e^{\hat{\beta}_0 + \beta_1}}. \quad (4-43)$$

Since X is a binary variable, $p(x_1)$ in the above equation can be replaced by $1 - p(x_0)$, according to the nature of probabilities, to derive that:

$$p(y_0) = p(x_0, y_0) + p(x_1, y_0) = \frac{p(x_0)}{1 + e^{\hat{\beta}_0}} + \frac{1 - p(x_0)}{1 + e^{\hat{\beta}_0 + \beta_1}}. \quad (4-44)$$

As $P(Y)$ can be obtained by observing Y in the enrichment sample, the marginal probability distribution $P(X)$ can be approximated as a function of $\hat{\beta}_0$ given $P(Y)$ and β_1 as:

$$p(x_0) = \frac{(e^{\hat{\beta}_0} + 1)[(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1]}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)}, \quad (4-45)$$

where p_y is the marginal probability distribution of Y when $Y = y_0$, i.e. $p(y_0)$, in the enrichment sample;

β_1 is the taste parameter associated with attribute X in the utility function of the IDCM kernel (Equation 4-33).

Approximating $P(X)$ via the reversed recalibration of constant, Equation 4-45 can further be manipulated to obtain an approximation of the joint probability distribution of X and Y :

$$\begin{cases} p(x_0, y_0) = \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} \\ p(x_0, y_1) = \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\beta_1} - 1} \\ p(x_1, y_0) = \frac{1 - (e^{\hat{\beta}_0} + 1)p_y}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} \\ p(x_1, y_1) = \frac{e^{\beta_1}[1 - (e^{\hat{\beta}_0} + 1)p_y]}{e^{\beta_1} - 1}. \end{cases} \quad (4-46)$$

The above derivation thus provides a way to approximate the joint probabilities $P(X, Y)$ of an enrichment sample as functions of the assumed recalibrated constant $\hat{\beta}_0$, based on the knowledge of a DCM capturing the choice behaviour and the marginal distribution of the choice. In light of the approximated $P(X)$ and $P(X, Y)$ as well as the $P(Y)$ obtained from observing the enrichment sample, the enrichment efficiency metric can be expressed as a function of $\hat{\beta}_0$ for known p_y and β_1 . This also enables the representation of η with respect to the data distribution underlying the enrichment sample as a smooth function, eliminating the conundrum of less intuitive probabilities. The following section explores the variation pattern of enrichment efficiency with respect to $\hat{\beta}_0$.

4.2.3.3 Enrichment efficiency versus assumed recalibrated constant

As shown in Equations 4-21, 4-23, and 4-27, the metrics of mutual information, entropy as well as the enrichment efficiency of the quasi choice-making channel are defined according to

the joint probability distribution between X and Y . Based on the expression (Equation 4-46) of $P(X, Y)$ as derived in the previous section, these information-theoretic metrics are formulated below as functions of $\hat{\beta}_0$. Specifically, the expressions for mutual information, entropy and the enrichment efficiency are presented in Equations 4-47, 4-48 and 4-49, respectively, by replacing the joint probabilities with the corresponding expressions of $\hat{\beta}_0$.

$$I(X, Y) = -p_y \log_2 p_y - p_{y'} \log_2 p_{y'} - \frac{\lambda \omega}{\alpha(\tau - 1)} \log_2 \omega + \frac{\sigma v}{\alpha(\tau - 1)} \log_2 v \\ + p_{y'} \log_2 \alpha - \frac{\tau p_y \alpha}{(\tau - 1)} \log_2 \tau - \frac{\tau p_{y'}}{(\tau - 1)} \log_2 \tau, \quad (4-47)$$

where $p_{y'} = 1 - p_y$;

$$\tau = e^{\beta_1};$$

$$\alpha = e^{\hat{\beta}_0};$$

$$\omega = \alpha + 1 = e^{\hat{\beta}_0} + 1;$$

$$v = \tau \alpha + 1 = e^{\hat{\beta}_0 + \beta_1} + 1;$$

$$\lambda = \tau p_y \alpha - p_{y'} = p_y (e^{\hat{\beta}_0 + \beta_1} + 1) - 1; \text{ and}$$

$$\sigma = p_y \alpha - p_{y'} = p_y (e^{\hat{\beta}_0} + 1) - 1.$$

$$H(X) = \frac{\lambda \omega}{\alpha(\tau - 1)} \log \left[\frac{(\tau - 1)\alpha}{\lambda \omega} - 1 \right] - \log \left[1 - \frac{\lambda \omega}{\alpha(\tau - 1)} \right]. \quad (4-48)$$

$$\eta = \frac{-\alpha(\tau - 1)(p_y \ln p_y + p_{y'} \ln p_{y'}) - \lambda \omega \ln \omega + \sigma v \ln v + (\tau - 1)p_{y'} \alpha \ln \alpha - \tau \alpha (p_y \alpha + p_{y'}) \ln \tau}{\lambda \omega \log \left[\frac{(\tau - 1)\alpha}{\lambda \omega} - 1 \right] - \alpha(\tau - 1) \log \left[1 - \frac{\lambda \omega}{\alpha(\tau - 1)} \right]}. \quad (4-49)$$

The complexity of the enrichment efficiency (Equation 4-49) as a function of $\hat{\beta}_0$ can result in a cumbersome first derivative which does not yield in a closed-form expression for its root. However, solving the first order condition numerically allows determining existence of $\hat{\beta}_{0_\eta}^*$ at which η is maximised, i.e.:

$$\exists \hat{\beta}_{0_\eta}^* \text{ s.t. } \eta(\hat{\beta}_{0_\eta}^*) \geq \eta(\hat{\beta}_0) \text{ for } \hat{\beta}_0 \in \mathbb{R}. \quad (4-50)$$

To provide a direct demonstration of the continuous change of η as a function of $\hat{\beta}_0$, Figure 4.5 presents an example variation pattern of η with respect to $\hat{\beta}_0$, when the values of p_y and β_1 are specified as $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$). The value 0.3 is defined for p_y as it can reveal

a common circumstance where the observed choice is neither too deterministic nor too random. On the other hand, $\beta_1 = 1.792$ states the cases in which the choice depending on the socio-demographic attribute is observable yet not too deterministic, which, in the meanwhile gives an integer τ for reference convenience.

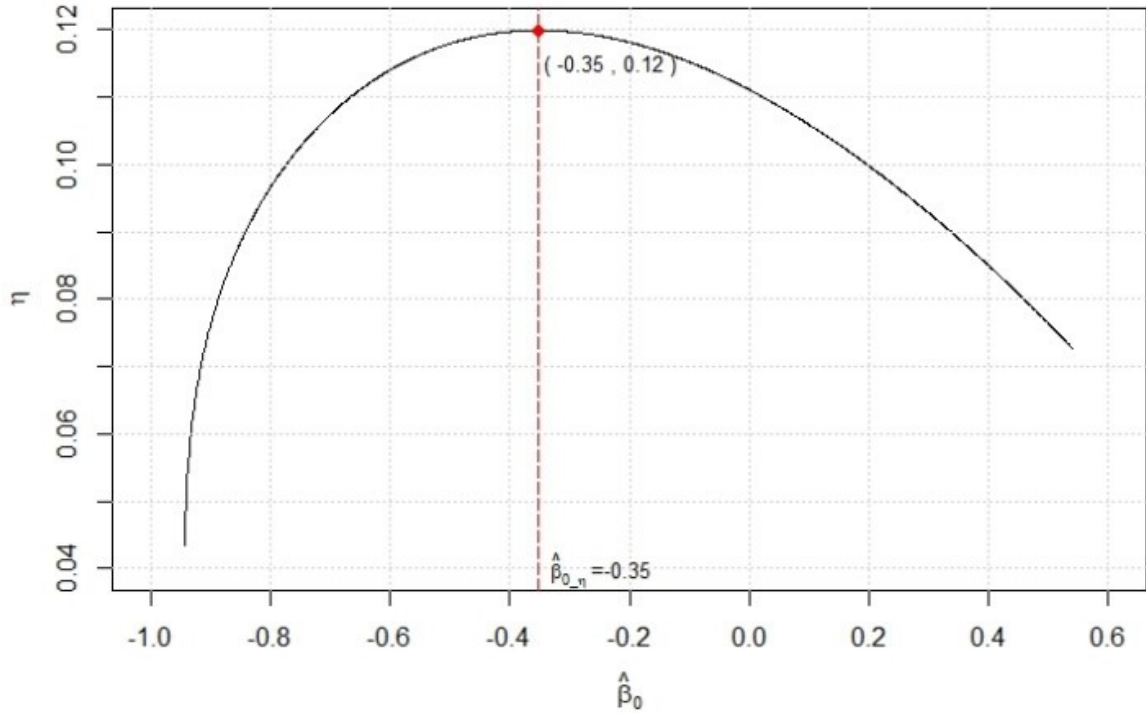


Figure 4.5 Variation Pattern of η vs. $\hat{\beta}_0$ when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)

It can be observed from Figure 4.5 that the variation of η with $\hat{\beta}_0$ follows the patterns below:

- 1) η increases as $\hat{\beta}_0$ increases when $\hat{\beta}_0 \leq \hat{\beta}_{0-\eta}^*$;
- 2) η decreases as $\hat{\beta}_0$ increases when $\hat{\beta}_0 \geq \hat{\beta}_{0-\eta}^*$; and
- 3) The pattern of η with respect to $\hat{\beta}_0$ is asymmetrical at the point where it is maximised.

To understand the above patterns, the respective variations of $H(X)$ and $I(X, Y)$ with respect to $\hat{\beta}_0$ are explored because η is, by definition, the ratio of the two. The following subsections provide the detailed derivation for this analysis.

Variation pattern of $H(X)$ with respect to $\hat{\beta}_0$

In particular, Equation 4-48 presents the expression of the entropy $H(X)$ of the enriched socio-demographic attribute X as a function of $\hat{\beta}_0$. Let:

$$\gamma = \frac{\lambda\omega}{\alpha(\tau - 1)} = \frac{(\tau p_y \alpha - p_{y'}) (\alpha + 1)}{\alpha(\tau - 1)} = \frac{(p_y e^{\hat{\beta}_0 + \beta_1} - 1 + p_y)(e^{\hat{\beta}_0} + 1)}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)}. \quad (4-51)$$

Comparing Equation 4-51 with Equation 4-45 suggests that γ is effectively the marginal probability distribution of X when labelled 0. Replacing the RHS of Equation 4-51 with γ , Equation 4-48 can then be reformulated as:

$$H(X) = \gamma \log_2 \left(\frac{1}{\gamma} - 1 \right) - \log_2(1 - \gamma). \quad (4-52)$$

The first derivative of $H(X)$ with respect to γ can be manipulated accordingly as:

$$\frac{\partial H(X)}{\partial \gamma} = \frac{1}{\ln 2} \ln \frac{1 - \gamma}{\gamma}. \quad (4-53)$$

By finding the root of the first derivative in Equation 4-53, the point at which the optimised $H(X)$ is achieved can be calculated as:

$$\frac{\partial H(X)}{\partial \gamma} = \ln \frac{1 - \gamma}{\gamma} = 0 \Rightarrow \gamma = 0.5. \quad (4-54)$$

Moreover, the second derivative of $H(X)$ with respect to γ is found to be permanently smaller than 0 (see the following equation), indicating that $H(X)$ achieves its minimum at $\gamma = 0.5$.

$$\frac{\partial^2 H(X)}{\partial \gamma^2} = -\frac{1}{\gamma(1 - \gamma)} < 0. \quad (4-55)$$

It can then be derived from the above analysis that:

- 1) $H(X)$ increases as γ grows when $\gamma < 0.5$;
- 2) $H(X)$ decreases as γ grows when $\gamma > 0.5$; and
- 3) $H(X)$ is maximised at the point $\gamma = 0.5$.

This is also in keeping with Figure 4.4, which suggests that the entropy of a binary variable is maximised when its POVs are equally distributed.

The above patterns offer the chance to explore $H(X)$ indirectly with respect to α by identifying how γ varies with α . Particularly, the first derivative of γ given α is formulated as:

$$\frac{\partial \gamma}{\partial \alpha} = \frac{\partial[(\tau p_y \alpha - p_{y'})(\alpha + 1)]}{\partial \alpha} = \frac{p_y \tau}{(\tau - 1)} + \frac{1 - p_y}{(\tau - 1)\alpha^2}. \quad (4-56)$$

The RHS of Equation 4-56 is found to be permanently greater than 0, suggesting the monotonical growth of γ as α increases. Consequently, the problem of finding α_{HX}^* for the maximum $H(X)$ is converted as to locate the value of α at which $\gamma = 0.5$, i.e.:

$$\gamma = \frac{(\tau p_y \alpha - p_{y'})(\alpha + 1)}{\alpha(\tau - 1)} = 0.5. \quad (4-57)$$

Solving the above equation leads to the value of α_{HX}^* at which the maximum $H(X)$ is reached:

$$\alpha_{HX}^* = \frac{(\tau + 1)(1 - 2p_y) + \sqrt{(1 - 2p_y)^2(\tau^2 + 1) - 2(4p_y^2 - 4p_y - 1)\tau}}{4p_y \tau}. \quad (4-58)$$

As $\alpha = e^{\hat{\beta}_0}$, it is also apparent that α increases monotonically with $\hat{\beta}_0$. Therefore, the value of $\hat{\beta}_0$ at which $H(X)$ is maximised can be calculated by finding the logarithm of α_{HX}^* as:

$$\hat{\beta}_{0_HX}^* = \ln \alpha_{HX}^* = \ln \left[\frac{(\tau + 1)(1 - 2p_y) + \sqrt{(1 - 2p_y)^2(\tau^2 + 1) - 2(4p_y^2 - 4p_y - 1)\tau}}{4p_y \tau} \right]. \quad (4-59)$$

The general pattern of $H(X)$ as a function of $\hat{\beta}_0$ for specific p_y and β_1 is thus summarised as:

- 1) $H(X)$ increases as $\hat{\beta}_0$ grows when $\hat{\beta}_0 < \hat{\beta}_{0_HX}^*$;
- 2) $H(X)$ decreases as $\hat{\beta}_0$ grows when $\hat{\beta}_0 > \hat{\beta}_{0_HX}^*$; and
- 3) $H(X)$ is maximised at the point $\hat{\beta}_0 = \hat{\beta}_{0_HX}^*$.

Figure 4.6 (below) visualises the above pattern for the example case when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$) are specified.

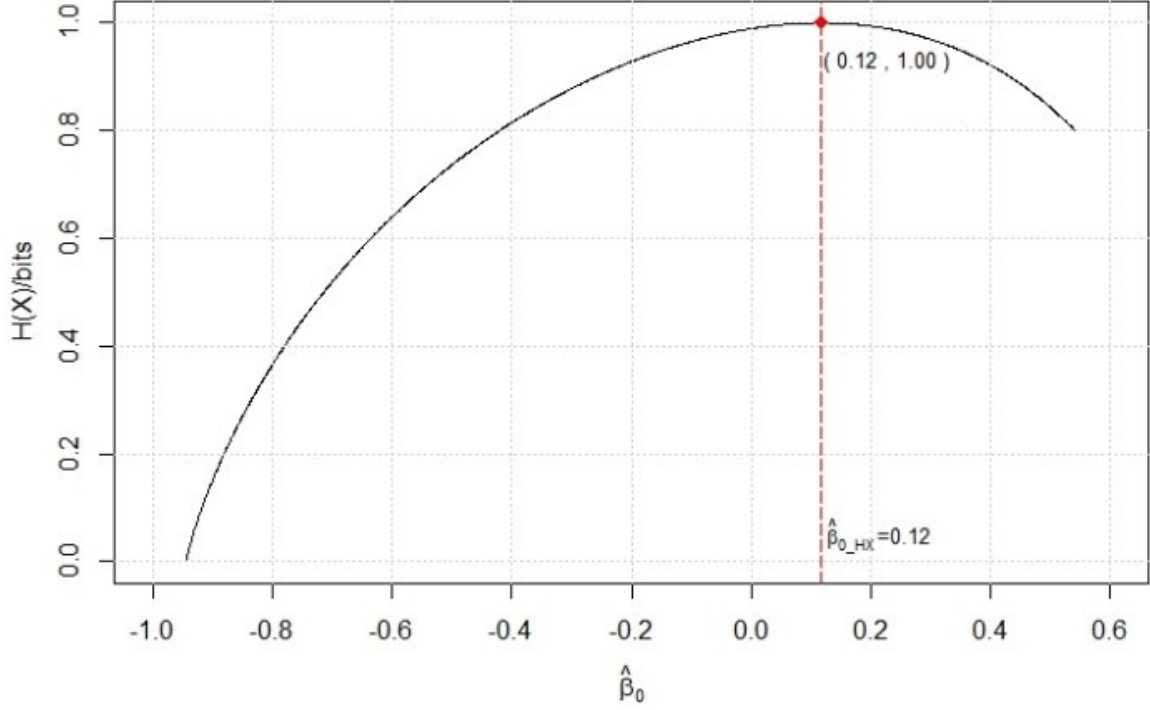


Figure 4.6 Variation Pattern of $H(X)$ vs. $\hat{\beta}_0$ when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)

Variation pattern of $I(X, Y)$ with respect to $\hat{\beta}_0$

Similarly, the variation pattern of $I(X, Y)$ with respect to α can be explored. In particular, the first derivative of Equation 4-47 is taken as:

$$\frac{\partial I(X, Y)}{\partial \alpha} = \frac{\tau p_y}{\tau - 1} \ln \frac{\tau \alpha + 1}{\alpha + 1} + \frac{p_{y'}}{(\tau - 1) \alpha^2} \ln \frac{\tau \alpha + 1}{\alpha + 1} - \frac{\tau p_y}{\tau - 1} \ln \tau. \quad (4-60)$$

Further manipulation of the above equation enables expressing it as:

$$\frac{\partial I(X, Y)}{\partial \alpha} = \ln \left[\frac{\tau \alpha + 1}{\tau(\alpha + 1)} \right] + \frac{\ln \tau}{\tau v \alpha^2 + 1}. \quad (4-61)$$

where $v = \frac{p_y}{p_{y'}}$.

Denote the first derivative of $I(X, Y)$ in Equation 4-61 as a function $\Phi(\cdot)$ of α , i.e.:

$$\Phi(\alpha) = \frac{\partial I(X, Y)}{\partial \alpha} = \ln \left[\frac{\tau \alpha + 1}{\tau(\alpha + 1)} \right] + \frac{\ln \tau}{\tau v \alpha^2 + 1}. \quad (4-62)$$

Equation 4-62 is found to have the properties below, given that $\alpha \in (0, +\infty)$.

$$\lim_{\alpha \rightarrow 0} \Phi(\alpha) = 0. \quad (4-63)$$

$$\lim_{\alpha \rightarrow +\infty} \Phi(\alpha) = 0. \quad (4-64)$$

To understand the variation of $\Phi(\alpha)$ over the domain of α , i.e. the values between the two limits, the second derivative of $I(X, Y)$ given α is derived:

$$\frac{\partial^2 I(X, Y)}{\partial \alpha^2} = \frac{\partial \Phi(\alpha)}{\partial \alpha} = \frac{\tau - 1}{(\tau \alpha + 1)(\alpha + 1)} - \frac{2\tau \nu \alpha \ln \tau}{(\tau \nu \alpha^2 + 1)^2}. \quad (4-65)$$

On the other hand, Equation 4-65 is found to retain the following patterns:

$$\lim_{\alpha \rightarrow 0} \frac{\partial \Phi(\alpha)}{\partial \alpha} = \tau - 1. \quad (4-66)$$

$$\lim_{\alpha \rightarrow +\infty} \frac{\partial \Phi(\alpha)}{\partial \alpha} = 0. \quad (4-67)$$

Based on the setting of the context that $\beta_1 > 0$ ($\tau > 1$) is assumed for the analysis, Equation 4-66 is greater than 0. This indicates that the slope of $\Phi(\alpha)$ is positive when α approaches 0. In contrast, observing the RHS of Equation 4-65 reveals that the first term, i.e. $\frac{\tau-1}{(\tau\alpha+1)(\alpha+1)}$, has order $O(\alpha^{-2})$ while the second term is about $O(\alpha^{-3})$. It is hence implied that Equation 4-65 is greater than 0 when α is large enough, meaning that Equation 4-67 is also positive.

Combining Equations 4-63 and 4-66 deduces that $\Phi(\alpha) > 0$ as α approaches 0 while Equations 4-64 and 4-67 jointly suggests that $\Phi(\alpha) < 0$ when α approaches positive infinity. The intermediate value theorem⁸ shows that there always exists α_{IXY}^* such that $\Phi(\alpha) = 0$ holds. This also indicates that:

$$\exists \alpha_{IXY}^* \text{ s.t. } I(\alpha_{IXY}^*) \geq I(\alpha) \text{ for } \alpha \in \mathbb{R}^+. \quad (4-68)$$

As α is a monotonically increasing function of $\hat{\beta}_0$, Equation 4-68 implies existence of $\hat{\beta}_{0_IXY}^*$ which yields the maximum $I(X, Y)$.

The variation pattern of $I(X, Y)$ given $\hat{\beta}_0$ is thus deduced as:

⁸ In mathematical analysis, the intermediate value theorem states that if f is a continuous function whose domain contains the interval $[a, b]$ then it takes on any given value between $f(a)$ and $f(b)$ at some point within the interval.

- 1) $I(X, Y)$ increases as $\hat{\beta}_0$ grows when $\hat{\beta}_0 < \hat{\beta}_{0_IXY}^*$;
- 2) $I(X, Y)$ decreases as $\hat{\beta}_0$ grows when $\hat{\beta}_0 > \hat{\beta}_{0_IXY}^*$; and
- 3) $I(X, Y)$ is maximised at the point $\hat{\beta}_0 = \hat{\beta}_{0_IXY}^*$.

Such patterns are illustrated in Figure 4.7 also using the case when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$) as an example.

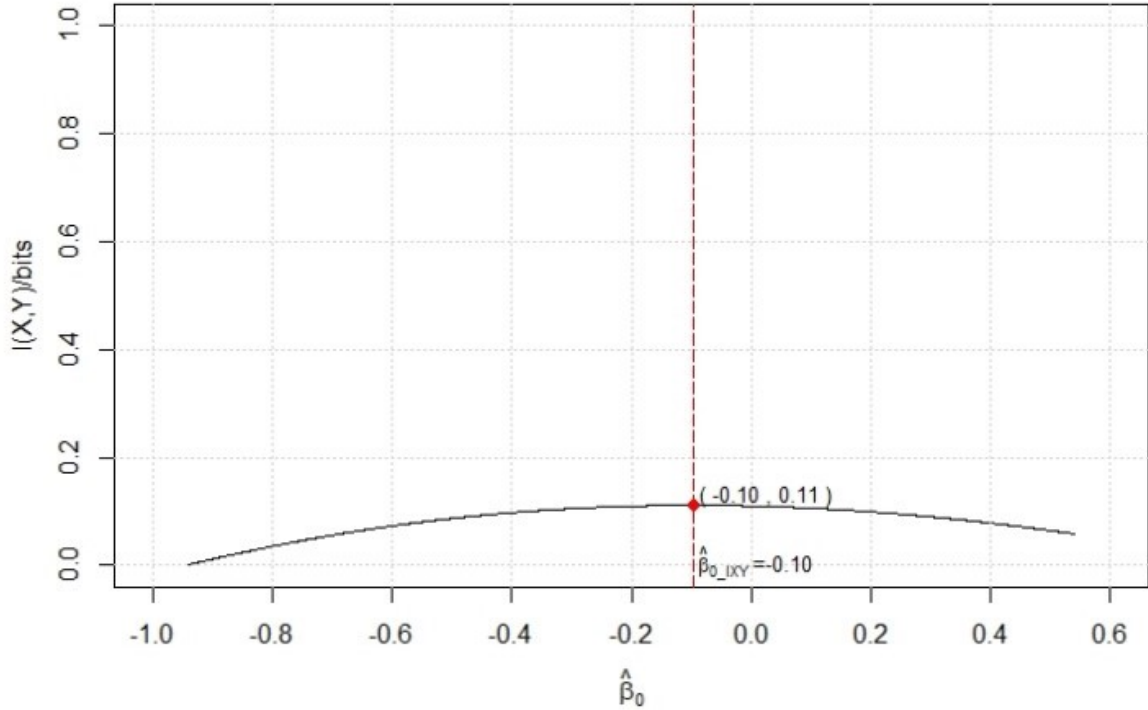


Figure 4.7 Variation Pattern of $I(X, Y)$ vs. $\hat{\beta}_0$ when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)

Comparison between the patterns of $H(X)$ and $I(X, Y)$

It can be concluded from the above derivations that $H(X)$ and $I(X, Y)$ achieve their maximum values respectively at $\hat{\beta}_0 = \hat{\beta}_{0_IXY}^*$ and $\hat{\beta}_0 = \hat{\beta}_{0_HX}^*$. Moreover, it is noticed that the two maxima are usually reached at different points, i.e. $\hat{\beta}_{0_IXY}^* \neq \hat{\beta}_{0_HX}^*$, unless they happen to numerically equal each other.

The different maximum points for $H(X)$ and $I(X, Y)$ can be interpreted by comparing their respective definitions (Equations 4-23 and 4-21). Specifically, the variation of $H(X)$ is purely driven by the change in $P(X)$ which can be any values between 0 and 1. For binary variables, $H(X)$ is maximised when the POVs of X are evenly distributed, i.e. $p(x_0) = p(x_1) = 0.5$. In contrast, $I(X, Y)$ is related, by definition, to the variability of both X and Y as well as the

correlation between the two. The different determinants for $H(X)$ and $I(X, Y)$ therefore lead to the discrepancy in terms of their variability and variation patterns.

Such discrepancy between the patterns of $H(X)$ and $I(X, Y)$ can also be explained from the perspective of the first derivative. In particular, Equation 4-59 shows that the value $\hat{\beta}_{0_HX}^*$ which maximises $H(X)$ can be derived by performing suitable manipulations to the first derivative of $H(X)$ with respect to $\hat{\beta}_0$. In contrast, the direct expression of corresponding $\hat{\beta}_{0_IXY}^*$ does not yield a close form. Whilst the above deductions prove existence of $\hat{\beta}_{0_IXY}^*$, the different forms of first derivatives for $H(X)$ and $I(X, Y)$ with respect to $\hat{\beta}_0$ also indicate the difference between the two in term of the points where they achieve their maxima.

To illustrate the above analysis directly and visually, Figure 4.8 compares the variations of $H(X)$ and $I(X, Y)$, as demonstrated respectively in Figure 4.6 and Figure 4.7, with respect to $\hat{\beta}_0$ when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$).

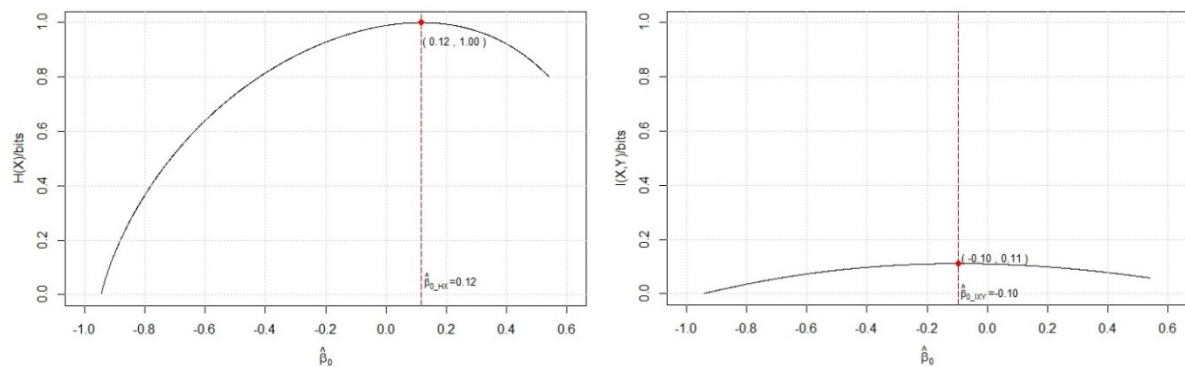


Figure 4.8 Comparison between Patterns of $H(X)$ and $I(X, Y)$ vs. $\hat{\beta}_0$ when $p_y = 0.3$ and $\beta_1 = 1.792$

It is observed from Figure 4.8 that:

- 1) When $\hat{\beta}_0 < \hat{\beta}_{0_IXY}^*$, both $I(X, Y)$ and $H(X)$ grows as $\hat{\beta}_0$ increases while the improvement rate of the two both decreases. Moreover, the general improvement rate of $H(X)$ is significantly higher than that of the $I(X, Y)$ because the variation of $I(X, Y)$ is subject to the relative value of β_1 to $\hat{\beta}_0$.
- 2) When $\hat{\beta}_{0_IXY}^* < \hat{\beta}_0 < \hat{\beta}_{0_HX}^*$, $I(X, Y)$ declines while $H(X)$ keeps growing. This corresponds to the interval over which η decreases with a higher rate than it is when $\hat{\beta}_{0_\eta}^* < \hat{\beta}_0 < \hat{\beta}_{0_IXY}^*$, shown by the steeper fall in the right wing in Figure 4.7 when $\hat{\beta}_0 > -0.10$.

- 3) Both $I(X, Y)$ and $H(X)$ decrease as $\hat{\beta}_0$ increases when $\hat{\beta}_0 > \hat{\beta}_{0_HX}^*$. Still, the fall in $H(X)$ is more significant than that in $I(X, Y)$ as the decreasing rate of both grows gradually after achieving their respective maximum.

The different declining rates in 1) and in 2), 3) of the above patterns is in that the two plots are maximised at different points such that $H(X)$ at one point is divided by a different $I(X, Y)$ if mapped symmetrically by $\hat{\beta}_0 = \hat{\beta}_{0_HX}^*$. Whilst illustrated with a particular instance of specified p_y and β_1 , the patterns above are generalisable, which is explained as follows.

As $I(X, Y) \leq H(X)$ always holds by definition, the growth rate of $H(X)$ is always no less than that of $I(X, Y)$ over the interval described in 1). In addition, the relative improvement rate of $I(X, Y)$ as compared to $H(X)$ also declines because the variability of $I(X, Y)$ cannot exceed that of $H(X)$ for the choice Y that is not deterministic given X . As a consequence, there exists a point at which such improvement rate reduces to 0. And this is the point, i.e. $\hat{\beta}_0 = \hat{\beta}_{0_n}^*$, where η as the ratio of the two is maximised and beyond which the ratio η decreases. Given that $H(X)$ is purely driven by the variability of X while $I(X, Y)$ is subject to the dependence between X and Y , unless the dependence between X and Y is deterministic, $I(X, Y)$ would always start to fall not later than the point at which $H(X)$ turns to decrease. And this results in the general variation asymmetrical pattern (Figure 4.7) of η at $\hat{\beta}_{0_n}^*$ for the maximum η .

4.2.3.4 PCPs versus assumed recalibrated constant

The disaggregate enrichment performance (i.e. PCPs) of the IDCM approach using the MAP and the MLE estimators (referred to as the MAP-PCP and the MLE-PCP) can be expressed using $\hat{\beta}_0$ for known p_y and β_1 given that the PCPs is also linked to $P(X, Y)$. The detailed derivation is explained as below.

Recall Equations 4-34 and 4-35 which respectively define the MAP and the MLE estimators for the binary enrichment context. It is noticed that both the MAP and the MLE estimators seek to find the POV of the socio-demographic attribute X that can maximise the corresponding likelihood function while their expressions are different due to presence or absence of the prior of X . Specifically, as is indicated by Equation 4-34, the MAP estimator includes the term $P(X)$ to account for the prior information yet this is not considered in the expression of the MLE estimator (Equation 4-35), yielding the MLE estimator as a special case of the MAP estimator where X is assumed to be evenly distributed. For the binary enrichment context, $P(X)$ for both

POVs of X in the enrichment using the MLE estimator is assumed as 0.5, thus can be eliminated from the expression for the MLE estimator.

Representation of MLE-PCP using $P(X, Y)$

To start from the special case where the distribution of $P(X)$ is assumed to be uniform, Equation 4-35 indicates that the MLE estimator of X selects the POV of X given observed Y such that $P(Y|X)$ is maximised. Such enrichment using the MLE estimator is thus found to be independent of β_0 , which is formally defined in Lemma 1.

Lemma 1. The IDCM enrichment of a binary attribute X from single observation of the choice Y using the maximum likelihood estimator is independent of the constant β_0 in the binary logit kernel with linear-in-parameter utilities.

The proof of Lemma 1 is provided as follows.

Consider y_0 is observed for the choice Y . The prediction between x_0 and x_1 to maximise $P(Y|X)$ can then be performed by comparing $p(y_0|x_0)$ and $p(y_0|x_1)$ and finding their difference. This can be manipulated based on the IDCM kernel (Equations 4-32 and 4-33) taking the form of a BL model as:

$$\begin{aligned} \Delta p(y|x) &= p(y_0|x_0) - p(y_0|x_1) = \frac{e^{V(y_0|x_0)}}{\sum_{i=0}^{i=1} e^{V(y_i|x_0)}} - \frac{e^{V(y_0|x_1)}}{\sum_{i=0}^{i=1} e^{V(y_i|x_1)}} \\ &= \frac{1}{1 + e^{\beta_0 + \beta_1 x_0}} - \frac{1}{1 + e^{\beta_0 + \beta_1 x_1}}. \end{aligned} \quad (4-69)$$

As the POVs of X are labelled to make an indicator (Equation 4-31), the above equation can further be simplified as:

$$\begin{aligned} \Delta p(y|x) &= \frac{1}{1 + e^{\beta_0 + \beta_1 x_0}} - \frac{1}{1 + e^{\beta_0 + \beta_1 x_1}} \\ &= \frac{1}{1 + e^{\beta_0}} - \frac{1}{1 + e^{\beta_0 + \beta_1}} = \frac{e^{\beta_0}(e^{\beta_1} + 1)}{(e^{\beta_0} + 1)(e^{\beta_0 + \beta_1} + 1)}. \end{aligned} \quad (4-70)$$

The sign of the RHS of Equation 4-70 is found to be purely determined by the numerator. Specifically, the denominator is positive for any real values of β_0 and β_1 . In the case of the numerator, the sign is determined by value of the expression in the parentheses, itself

depending on β_1 . This proves the assertion that the MLE enrichment using an IDCM kernel with a BL structure is purely determined by β_1 and is independent of β_0 .

As $\beta_1 > 0$ is defined in Section 4.2.3.1 for the current context, the numerator in the RHS of Equation 4-70 is also positive, leading to the enrichment output below:

$$X_{MLE}^* = \arg \max_{x_i \in X} p(y_0|x_i) = x_0. \quad (4-71)$$

The enrichment output of the MLE estimation for observations of y_1 can be derived accordingly, which yield the complete MLE enrichment for both POVs of Y as:

$$\begin{cases} X = \arg \max_{x_i \in X} p(y_0|x_i) = x_0, & Y = y_0 \\ X = \arg \max_{x_i \in X} p(y_1|x_i) = x_1, & Y = y_1. \end{cases} \quad (4-72)$$

The above deductions reveal, from the perspective of the quasi choice-making channel, that the output message of each POV of Y , which can be transmitted from either x_0 or x_1 , is however decoded into a certain POV of X . This can be expressed as the decoding scheme below.

$$\begin{aligned} Y = y_0 &\rightarrow X_{MLE}^* = x_0 \\ Y = y_1 &\rightarrow X_{MLE}^* = x_1 \end{aligned} \quad (4-73)$$

Based on the above decoding scheme of the MLE estimator, the MLE-PCP can also be related to $P(X, Y)$. To explain the representation of the MLE-PCP using a combination of $P(X, Y)$ for specific POVs of X and Y with a more straightforward illustration, assume that the true joint probability distribution in the enrichment sample D_{Erc} is as demonstrated below in Table 4.1.

Table 4.1 Actual Joint Probability Distribution of Enrichment Sample D_{Erc}

$P(X, Y)$	$X = x_0$	$X = x_1$	$P(Y)$
$Y = y_0$	$p(x_0, y_0)$	$p(x_1, y_0)$	$p(x_0, y_0) + p(x_1, y_0)$
$Y = y_1$	$p(x_0, y_1)$	$p(x_1, y_1)$	$p(x_0, y_1) + p(x_1, y_1)$
$P(X)$	$p(x_0, y_0) + p(x_0, y_1)$	$p(x_1, y_0) + p(x_1, y_1)$	

Whereas the marginal probability distribution $P(Y)$ listed in the rightmost column of Table 4.1 can be obtained from D_{Erc} , the values of $P(X, Y)$ and $P(X)$ are greyed out to indicate that they remain unknown as X is not presented in D_{Erc} . They are, however, presented only as a reference against which the mathematical expressions of the PCPs are derived.

According to the decoding scheme of the MLE estimator (Equation 4-73), it is derived that the enrichment sample D_{Erc} enriched with the attribute X using the MLE estimator of the IDCM

approach (referred to as the IDCM-MLE) holds the joint probability distribution $P'(X, Y)$ as illustrated in Table 4.2.

Table 4.2 Joint Probability Distribution of D_{Erc} Enriched with X using the IDCM-MLE Approach

$P(X, Y)$	$X = x_0$	$X = x_1$	$P(Y)$
$Y = y_0$	$p(x_0, y_0) + p(x_1, y_0)$	0	$p(x_0, y_0) + p(x_1, y_0)$
$Y = y_1$	0	$p(x_0, y_1) + p(x_1, y_1)$	$p(x_0, y_1) + p(x_1, y_1)$
$P(X)$	$p(x_0, y_0) + p(x_1, y_0)$	$p(x_0, y_1) + p(x_1, y_1)$	

By comparing Table 4.1 and Table 4.2, the overlapped value of the two matrices in each box, i.e. the value of the joint probability in $P'(X, Y)$ which is also presented in the same category of $P(X, Y)$, can jointly represent the proportion of the sample that is enriched with the correct value of X . For example, the joint probability for individuals who are characterised by x_0 choosing y_0 in the true distribution of the enrichment sample (Table 4.1) is $p(x_0, y_0)$, i.e.:

$$P(X = x_0, Y = y_0) = p(x_0, y_0). \quad (4-74)$$

As the enriched sample (Table 4.2) for the same category is $p(x_0, y_0) + p(x_1, y_0)$, the correctly enriched proportion for this category is their overlapped value $p(x_0, y_0)$. Summing up such correctly enriched proportions in all categories of the sample produce the overall MLE-PCP:

$$PCP_{MLE} = p(x_0, y_0) + p(x_1, y_1). \quad (4-75)$$

Expression of MLE-PCP as a function of $\hat{\beta}_0$

Replacing the RHS of Equation 4-75 with the corresponding expressions of $P(X, Y)$ using $\hat{\beta}_0$, the representation of the MLE-PCP as a function of $\hat{\beta}_0$ is enabled. Specifically, the expression of the MLE-PCP using $\hat{\beta}_0$ for known p_y and β_1 can be manipulated and derived as:

$$\begin{aligned}
 PCP_{MLE} &= p(x_0, y_0) + p(x_1, y_1) \\
 &= \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} + \frac{e^{\beta_1}[1 - (e^{\hat{\beta}_0} + 1)p_y]}{e^{\beta_1} - 1} \\
 &= \frac{p_y(1 - e^{2\hat{\beta}_0 + \beta_1}) + e^{\hat{\beta}_0 + \beta_1} - 1}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} \\
 &= \frac{p_y(1 - \tau\alpha^2) + \tau\alpha - 1}{\alpha(\tau - 1)}.
 \end{aligned} \quad (4-76)$$

As is outlined below, two constraints are required for the expression of the MLE-PCP in Equation 4-75 above.

- 1) $PCP_{MLE} \geq 50\%$: This is defined such that the information transmitted through this quasi choice-making channel is more than the noise. In particular, If $PCP_{MLE} < 50\%$, an equivalent enrichment with $PCP_{MLE} > 50\%$ can be derived by swapping information and noise defined by the mapping rule in Equation 4-36. This is also in line with the intuition that the enrichment should perform no worse than a random guess (50% correct).
- 2) $0 < p(x, y) < 1 \quad \forall x \in X, y \in Y$: This reflects the stochasticity or the probabilistic nature of the analysis. If the reverse were true, i.e. any joint probability or marginal probability equals 0 or 1, the channel would be, at least partly, noiseless and the decoding would be bijective, hence losing the meaning for developing the enrichment approach and for exploring the enrichment (decoding) performance.

The above two constraints jointly define the domain of $\hat{\beta}_0$, of which the derivation is provided as follows. Specifically, Constraint 1 can be mathematically expressed as:

$$PCP_{MLE} = \frac{p_y(1 - \tau\alpha^2) + \tau\alpha - 1}{\alpha(\tau - 1)} \geq 50\%. \quad (4-77)$$

Suitable manipulation provides the equivalent of the above equation as:

$$-2p_y\tau\alpha^2 + (\tau + 1)\alpha + 2(p_y - 1) \geq 0. \quad (4-78)$$

The solution to the above inequality involves finding the root of the quadratic equation of α , which provides the domain of α satisfying Constraint 1 as:

$$\alpha \in \left[\frac{\tau+1 - \sqrt{\tau^2 + 2[1+8p_y(p_y-1)]\tau+1}}{4p_y\tau}, \frac{\tau+1 + \sqrt{\tau^2 + 2[1+8p_y(p_y-1)]\tau+1}}{4p_y\tau} \right]. \quad (4-79)$$

As α is a monotonic function of $\hat{\beta}_0$, the corresponding domain of $\hat{\beta}_0$ given Constraint 1 can be found by simply performing the logarithm transformation to Equation 4-79:

$$\hat{\beta}_0 \in \left[\ln \frac{\tau+1 - \sqrt{\tau^2 + 2[1+8p_y(p_y-1)]\tau+1}}{4p_y\tau}, \ln \frac{\tau+1 + \sqrt{\tau^2 + 2[1+8p_y(p_y-1)]\tau+1}}{4p_y\tau} \right]. \quad (4-80)$$

As for Constraint 2, given that the joint probabilities in Equation 4-46 are calculated based on $P(X)$, the second constraint holds equivalently when $0 < p(x_0) < 1$, i.e.:

$$\begin{cases} p(x_0) = \frac{(e^{\hat{\beta}_0} + 1)[(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1]}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} > 0 \\ p(x_0) = \frac{(e^{\hat{\beta}_0} + 1)[(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1]}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} < 1. \end{cases} \quad (4-81)$$

The simultaneous equations above are solved to obtain the domain required Constraint 2 as:

$$\hat{\beta}_0 \in \left(\ln \frac{1 - p_y}{p_y} - \beta_1, \ln \frac{1 - p_y}{p_y} \right). \quad (4-82)$$

As a result, the domain of $\hat{\beta}_0$ can be derived jointly from the intersection of the two separate domains (Equations 4-79 and 4-82) defined by the two constraints. Specifically, the lower bound of the domain of $\hat{\beta}_0$ can be determined by finding the relatively lower value between the two lower bounds in Equations 4-79 and 4-82. The corresponding difference between the two is expressed as:

$$\begin{aligned} & \ln \frac{1 - p_y}{\tau p_y} - \ln \frac{\tau + 1 - \sqrt{\tau^2 + 2[1 + 8p_y(p_y - 1)]\tau + 1}}{4p_y\tau} \\ &= \ln \frac{4(1 - p_y)}{\tau + 1 - \sqrt{\tau^2 + 2[1 + 8p_y(p_y - 1)]\tau + 1}}. \end{aligned} \quad (4-83)$$

Let $\epsilon = 1 + 8p_y(p_y - 1)$. Equation 4-83 can be reformulated as:

$$\begin{aligned} & \ln \frac{1 - p_y}{\tau p_y} - \ln \frac{\tau + 1 - \sqrt{\tau^2 + 2[1 + 8p_y(p_y - 1)]\tau + 1}}{4p_y\tau} \\ &= \ln \frac{4(1 - p_y)}{\tau + 1 - \sqrt{\tau^2 + 2[1 + 8p_y(p_y - 1)]\tau + 1}} \\ &= \ln \frac{4(1 - p_y)}{\tau + 1 - \sqrt{\tau^2 + 2\epsilon\tau + 1}} \end{aligned} \quad (4-84)$$

$$= \ln \frac{4(1-p_y)}{\frac{2(1-\epsilon)\tau}{\tau+1+\sqrt{(\tau+\epsilon)^2+1-\epsilon^2}}}.$$

As $p_y \in (0, 0.5]$, the domain of ϵ can be derived as:

$$\epsilon = 1 + 8p_y(p_y - 1) \in [-1, 1). \quad (4-85)$$

It is thus obtained from the RHS of Equation 4-83 that:

$$\ln \frac{4(1-p_y)}{\frac{2(1-\epsilon)\tau}{\tau+1+\sqrt{(\tau+\epsilon)^2+1-\epsilon^2}}} > \ln \frac{4(1-p_y)}{\frac{2(1-\epsilon)\tau}{2\tau}} = \ln \frac{4(1-p_y)}{1-\epsilon}. \quad (4-86)$$

The RHS of Equation 4-86 is minimised when $p_y = 0.5$ and $\gamma = -1$, i.e. $\ln \frac{4(1-p_y)}{1-\epsilon} = 0$. This shows that the left-hand side (LHS) of Equation 4-86, i.e. the difference of the lower bounds of Equations 4-79 and 4-82, is greater than 0, i.e.:

$$\begin{aligned} & \ln \frac{1-p_y}{\tau p_y} - \ln \frac{\tau+1-\sqrt{\tau^2+2[1+8p_y(p_y-1)]\tau+1}}{4p_y\tau} \\ &= \ln \frac{4(1-p_y)}{\tau+1-\sqrt{\tau^2+2[1+8p_y(p_y-1)]\tau+1}} > 0. \end{aligned} \quad (4-87)$$

Therefore, the lower bound of the domain for $\hat{\beta}_0$ jointly subject to Constraints 1 and 2 is determined as the larger value, i.e. $\ln \frac{1-p_y}{\tau p_y}$.

On the other hand, the upper bound of the domain can be found by comparing the two upper bounds in Equations 4-79 and 4-82. Similarly, the difference of the two is taken as:

$$\begin{aligned} & \ln \frac{1-p_y}{p_y} - \ln \frac{\tau+1+\sqrt{\tau^2+2[1+8p_y(p_y-1)]\tau+1}}{4p_y\tau} \\ &= \ln \frac{4\tau(1-p_y)}{\tau+1-\sqrt{\tau^2+2[1+8p_y(p_y-1)]\tau+1}}. \end{aligned} \quad (4-88)$$

The equation below derives that the second term in the LHS of Equation 4-88 is smaller, i.e.:

$$\begin{aligned}
& \ln \frac{4\tau(1-p_y)}{\tau+1-\sqrt{\tau^2+2[1+8p_y(p_y-1)]\tau+1}} \\
& > \ln \frac{4\tau(1-p_y)}{2\tau} = \ln[2(1-p_y)] \geq 0.
\end{aligned} \tag{4-89}$$

The domain of $\hat{\beta}_0$ which satisfies both constraints can hence be defined as:

$$\hat{\beta}_0 \in \left(\ln \frac{1-p_y}{\tau p_y}, \ln \frac{\tau+1+\sqrt{\tau^2+2[1+8p_y(p_y-1)]\tau+1}}{4\tau p_y} \right]. \tag{4-90}$$

Note that, although the domain of $\hat{\beta}_0$ in Equation 4-90 is derived based on the enrichment using the MLE estimator, it also applies to the MAP-PCP discussed in the following paragraphs, and the expressions of $H(X)$, $I(X, Y)$ and η which are developed previously in Section 4.2.3.2.

As per Constraint 1 (Equation 4-77), the minimum MLE-PCP for the given domain is 50%:

$$PCP_{MLE_{min}}^* = 50\% \quad \text{when} \quad \hat{\beta}_{0_{MLE_{min}}}^* = \ln \frac{\tau+1+\sqrt{\tau^2+2[1+8p_y(p_y-1)]\tau+1}}{4p_y\tau}. \tag{4-91}$$

On the other hand, the maximum MLE-PCP can be derived by finding $\hat{\beta}_{0_{MLE_{max}}}^*$ at which the first derivative of Equation 4-76 equals 0. As α is a monotonically increasing function of $\hat{\beta}_0$, this is equivalent to find $\alpha_{0_{MLE_{max}}}^*$ where the first derivative given α equals 0, which simplifies the calculation as:

$$\frac{\partial PCP_{MLE}}{\partial \alpha} = \frac{(\tau-1)(1-p_y-p_y\tau\alpha^2)}{(\tau-1)^2\alpha^2} = 0. \tag{4-92}$$

Equation 4-92 is solved to find the point at which the maximum MLE-PCP can be reached:

$$PCP_{MLE_{max}}^* = \frac{[\tau-2\sqrt{\tau p_y(1-p_y)}]}{(\tau-1)} \quad \text{when} \quad \alpha_{0_{MLE_{max}}}^* = \sqrt{\frac{1-p_y}{p_y\tau}} \tag{4-93}$$

The value of $\hat{\beta}_0$ that produces the maximum MLE-PCP is hence derived correspondingly as:

$$\hat{\beta}_{0_{MLE_{max}}}^* = -\frac{1}{2}\beta_1 + \frac{1}{2}\ln \frac{1-p_y}{p_y}. \tag{4-94}$$

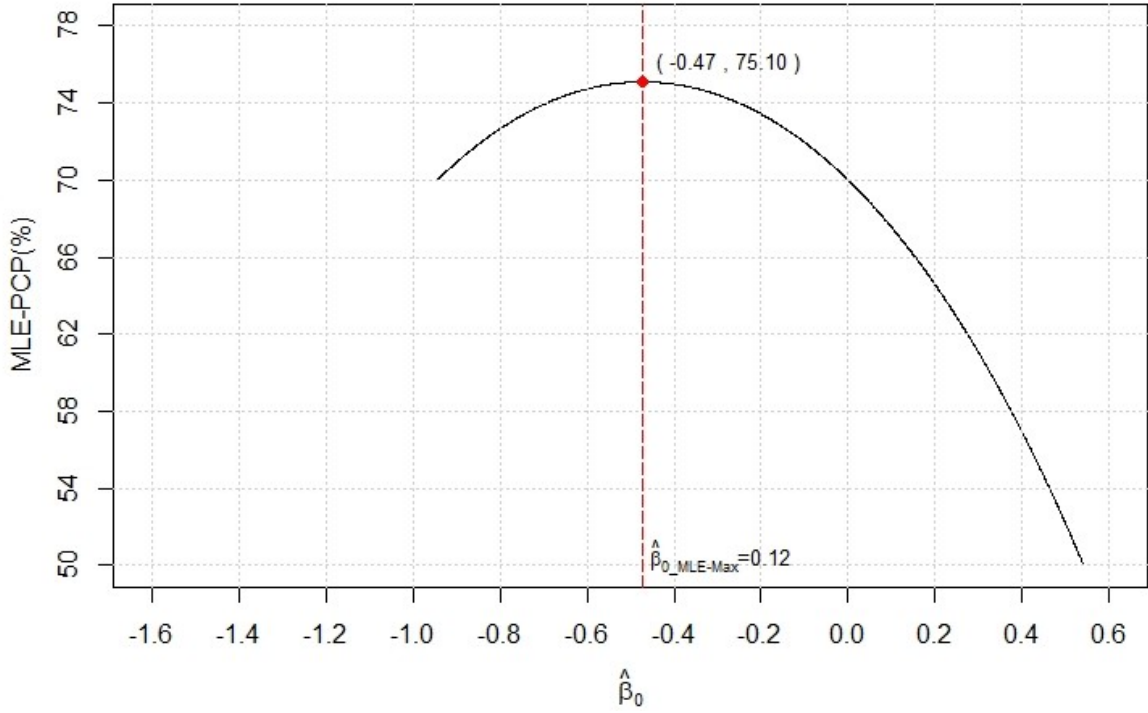


Figure 4.9 Variation Pattern of MLE-PCP vs. $\hat{\beta}_0$ when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)

Therefore, the variation of the MLE-PCP with respect to the assumed recalibrated constant $\hat{\beta}_0$ can be concluded to follow the patterns below:

- 1) The MLE-PCP is improved as $\hat{\beta}_0$ increases when $\hat{\beta}_0 \leq \hat{\beta}_{0_MLEmax}^*$; and
- 2) The MLE-PCP reduces as $\hat{\beta}_0$ increases when $\hat{\beta}_0 \geq \hat{\beta}_{0_MLEmax}^*$.

Example variation of the MLE-PCP with respect to $\hat{\beta}_0$ when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$) is shown in Figure 4.9 to provide a clearer visualisation of the above patterns.

Representation of MAP-PCP using $P(X, Y)$

Similarly, the MAP-PCP can be formulated as a function of $\hat{\beta}_0$ if it can be expressed as specific combinations of $P(X, Y)$. In particular, Equation 4-34 suggests that the prediction of a specific POV of X is effectively to find the value of X which can maximise $\mathcal{L}(X|Y)$ as the product of $P(Y|X)$ and $P(X)$. This is equivalent to find the maximum joint probability $p(x_i, y_i)$ for the observed choice condition, i.e.:

$$X_{MAP}^* = \arg \max_X P(Y|X)P(X) = \arg \max_X P(X, Y). \quad (4-95)$$

In the binary enrichment context, the prediction of a more likely POV of a given choice $y_j \in Y$ concerns selecting $x_i \in X$ which produces the larger value between $p(x_i, y_j)$. This can be obtained by finding the difference between $p(x_0, y_j)$ and $p(x_1, y_j)$, i.e.:

$$\Delta p(x, y) = p(x_0, y_j) - p(x_1, y_j). \quad (4-96)$$

For any choice $y_j \in Y$, the output prediction of X is x_0 if $\Delta p(x, y) > 0$ and x_1 otherwise. Hence, the selection (enrichment) of X for each possibly observed choice can be expressed by:

$$X_{MAP}^* = \begin{cases} X = \arg \max_{x_i \in X} p(x_i, y_0), & Y = y_0 \\ X = \arg \max_{x_i \in X} p(x_i, y_1), & Y = y_1 \end{cases} \quad (4-97)$$

The decoding scheme using the MAP estimator can be expressed accordingly as:

$$\begin{aligned} Y = y_0: & \begin{cases} p(y_0|x_0)p(x_0) > p(y_0|x_1)p(x_1) \rightarrow X_{MAP}^* = x_0 \\ p(y_0|x_0)p(x_0) < p(y_0|x_1)p(x_1) \rightarrow X_{MAP}^* = x_1 \end{cases} \\ Y = y_1: & \begin{cases} p(y_1|x_0)p(x_0) > p(y_1|x_1)p(x_1) \rightarrow X_{MAP}^* = x_0 \\ p(y_1|x_0)p(x_0) < p(y_1|x_1)p(x_1) \rightarrow X_{MAP}^* = x_1 \end{cases} \end{aligned} \quad (4-98)$$

It can be noted that the decoding scheme of the MLE estimator is also a special case of the decoding scheme for the MAP estimator (Equation 4-98) when:

$$\begin{cases} p(x_0) = p(x_1) = 0.5 \\ p(y_0|x_0) > p(y_0|x_1) \\ p(y_1|x_0) < p(y_1|x_1) \end{cases} \quad (4-99)$$

Note that the latter two conditions are subject to the setting that $\beta_1 > 0$.

Equation 4-98 indicates that, for a given observed value of Y , say y_j , the MAP selects the POV of X for the largest $p(x_i, y_j)$, i.e. x_i is predicted for all observations of y_j . This demonstrates the proportion of correct imputation for this choice condition ($Y = y_j$) over the whole sample is $p(x_i, y_j)$. Hence, the PCP using the MAP estimator can be derived by summing up such proportions over all possibly observed POVs of the choice as:

$$PCP_{MAP} = \max_{x_i \in X} p(x_i, y_0) + \max_{x_i \in X} p(x_i, y_1). \quad (4-100)$$

For different conditions presented in Equation 4-98, the MAP-PCP involves the summation of different combinations of joint probabilities, which effectively consists of three segments:

$$\begin{aligned}
& PCP_{MAP} \\
& = \begin{cases} p(x_1, y_0) + p(x_1, y_1), p(x_0, y_0) < p(x_1, y_0) \text{ and } p(x_0, y_1) < p(x_1, y_1) \\ p(x_0, y_0) + p(x_1, y_1), p(x_0, y_0) > p(x_1, y_0) \text{ and } p(x_0, y_1) < p(x_1, y_1) \\ p(x_0, y_0) + p(x_0, y_1), p(x_0, y_0) > p(x_1, y_0) \text{ and } p(x_0, y_1) > p(x_1, y_1) \end{cases} \quad (4-101)
\end{aligned}$$

Note that the case where $p(x_0, y_0) < p(x_1, y_0)$ and $p(x_0, y_1) > p(x_1, y_1)$ is dropped by setting $\beta_1 > 0$.

Expression of MAP-PCP as a function of $\hat{\beta}_0$

Similarly to the analysis of the MLE estimator, the variation of the MAP-PCP can be explored with respect to the recalibrated constant $\hat{\beta}_0$. Recall 4-46 in which the joint probabilities are expressed as functions of $\hat{\beta}_0$ given the knowledge of the MAP-PCP and the corresponding conditions can also be expressed using $\hat{\beta}_0$. The calculation is performed as below (Equations 4-102 – 4-110).

When $p(x_0, y_0) < p(x_1, y_0)$ and $p(x_0, y_1) < p(x_1, y_1)$, the MAP-PCP can be expressed as:

$$\begin{aligned}
PCP_{MAP} &= p(x_1, y_0) + p(x_1, y_1) \\
&= \frac{1 - (e^{\hat{\beta}_0} + 1)p_y}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} + \frac{e^{\beta_1}[1 - (e^{\hat{\beta}_0} + 1)p_y]}{e^{\beta_1} - 1} \\
&= \frac{(\tau\alpha + 1)[1 - (\alpha + 1)p_y]}{\alpha(\tau - 1)}. \quad (4-102)
\end{aligned}$$

The corresponding condition above can be manipulated as:

$$\begin{cases} p(x_0, y_0) > p(x_1, y_0) \\ p(x_0, y_1) < p(x_1, y_1) \end{cases} \Leftrightarrow \begin{cases} \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} < \frac{1 - (e^{\hat{\beta}_0} + 1)p_y}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} \\ \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\beta_1} - 1} < \frac{e^{\beta_1}[1 - (e^{\hat{\beta}_0} + 1)p_y]}{e^{\beta_1} - 1} \end{cases} \quad (4-103)$$

This gives the corresponding domain of $\hat{\beta}_0$ as:

$$\hat{\beta}_0 < \ln \frac{2p_y}{p_y(1 + \tau)}. \quad (4-104)$$

In terms of the second condition when $p(x_0, y_0) > p(x_1, y_0)$ and $p(x_0, y_1) < p(x_1, y_1)$, the MAP-PCP in Equation 4-101 can be formulated as:

$$\begin{aligned}
PCP_{MAP} &= p(x_0, y_0) + p(x_1, y_1) \\
&= \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} + \frac{e^{\beta_1}[1 - (e^{\hat{\beta}_0} + 1)p_y]}{e^{\beta_1} - 1} \\
&= \frac{p_y(1 - \tau\alpha^2) + \tau\alpha - 1}{\alpha(\tau - 1)}.
\end{aligned} \tag{4-105}$$

Manipulating the corresponding condition gives that:

$$\begin{cases} p(x_0, y_0) > p(x_1, y_0) \\ p(x_0, y_1) < p(x_1, y_1) \end{cases} \Leftrightarrow \begin{cases} \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} > \frac{1 - (e^{\hat{\beta}_0} + 1)p_y}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} \\ \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\beta_1} - 1} < \frac{e^{\beta_1}[1 - (e^{\hat{\beta}_0} + 1)p_y]}{e^{\beta_1} - 1} \end{cases}. \tag{4-106}$$

This thus results in the corresponding domain of $\hat{\beta}_0$ as:

$$\ln \frac{2p_y}{p_y(1 + \tau)} < \hat{\beta}_0 < \ln \frac{p_y(1 + \tau)}{2p_y\tau}. \tag{4-107}$$

The third condition in Equation 4-101 gives that the MAP-PCP can be represented as:

$$\begin{aligned}
PCP_{MAP} &= p(x_0, y_0) + p(x_0, y_1) \\
&= \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} + \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\beta_1} - 1} \\
&= \frac{(\alpha + 1)[(\tau\alpha + 1)p_y - 1]}{\alpha(\tau - 1)}.
\end{aligned} \tag{4-108}$$

Accordingly, the condition is equivalent to:

$$\begin{cases} p(x_0, y_0) > p(x_1, y_0) \\ p(x_0, y_1) > p(x_1, y_1) \end{cases} \Leftrightarrow \begin{cases} \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} > \frac{1 - (e^{\hat{\beta}_0} + 1)p_y}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)} \\ \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1}{e^{\beta_1} - 1} > \frac{e^{\beta_1}[1 - (e^{\hat{\beta}_0} + 1)p_y]}{e^{\beta_1} - 1} \end{cases}. \tag{4-109}$$

The consequent domain of $\hat{\beta}_0$ is thus derived as:

$$\hat{\beta}_0 > \ln \frac{(1 - p_y)(1 + \tau)}{2p_y\tau}. \tag{4-110}$$

As the domain of $\hat{\beta}_0$ (Equation 4-90) also applies to the MAP estimator, the complete solution to the MAP-PCP for the binary context can be expressed based on the calculations above as:

$$PCP_{MAP} = \begin{cases} \frac{(e^{\hat{\beta}_0 + \beta_1} + 1)[1 - (e^{\hat{\beta}_0} + 1)p_y]}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)}, & \hat{\beta}_0 \in (\hat{\beta}_{0_Lower}, \hat{\beta}_{0_1}) \\ \frac{p_y(1 - e^{2\hat{\beta}_0 + \beta_1}) + e^{\hat{\beta}_0 + \beta_1} - 1}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)}, & \hat{\beta}_0 \in (\hat{\beta}_{0_1}, \hat{\beta}_{0_2}) \\ \frac{(e^{\hat{\beta}_0} + 1)[(e^{\hat{\beta}_0 + \beta_1} + 1)p_y - 1]}{e^{\hat{\beta}_0}(e^{\beta_1} - 1)}, & \hat{\beta}_0 \in (\hat{\beta}_{0_2}, \hat{\beta}_{0_Upper}) \end{cases} \quad (4-111)$$

where $\hat{\beta}_{0_Lower} = \ln \frac{1-p_y}{\tau p_y}$;

$$\hat{\beta}_{0_1} = \ln \frac{2(1-p_y)}{p_y(1+\tau)};$$

$$\hat{\beta}_{0_2} = \ln \frac{(1-p_y)(1+\tau)}{2p_y\tau}; \text{ and}$$

$$\hat{\beta}_{0_Upper} = \ln \frac{\tau+1+\sqrt{\tau^2+2\tau+16p_y(1-p_y)+1}}{4\tau p_y}.$$

Solving the first derivative of Equation 4-111 with respect to $\hat{\beta}_0$ by segments reveals that:

- 1) the MAP-PCP decreases with respect to $\hat{\beta}_0$ on $(\hat{\beta}_{0_Lower}, \hat{\beta}_{0_2})$; and
- 2) the MAP-PCP is an increasing function of $\hat{\beta}_0$ on $(\hat{\beta}_{0_2}, \hat{\beta}_{0_Upper})$.

Hence, the MAP-PCP for the given enrichment context is minimised when $\hat{\beta}_{0_MAP_min}^* = \hat{\beta}_{0_2}$.

Substituting the value into Equation 4-111, the minimum MAP-PCP is thus derived as:

$$PCP_{MAP_min}^* = 50\% + \frac{p_y(\tau - 1)}{2(\tau + 1)}. \quad (4-112)$$

Alternatively, the best MAP-PCP can be derived by comparing the value of the MAP-PCP respectively at $\hat{\beta}_0 = \hat{\beta}_{0_Lower}$ and $\hat{\beta}_0 = \hat{\beta}_{0_Upper}$. It is found that the higher PCP_{MAP} is achieved when $\hat{\beta}_0 = \hat{\beta}_{0_Lower}$, at which $PCP_{MAP} = 100\%$. This also keeps with the intuition that the marginal distribution of X is deterministic when $\hat{\beta}_0 = \hat{\beta}_{0_Lower}$, as represented by $p(x_0) = 0$. Considering that the domain of $\hat{\beta}_0$ (Equation 4-90) does not include this point, the MAP-PCP is maximised to 100% when $\hat{\beta}_0$ infinitely approaches $\hat{\beta}_{0_Lower}$, i.e.:

$$PCP_{MAP_max}^* \rightarrow 100\% \text{ as } \hat{\beta}_0 \rightarrow \ln \frac{1-p_y}{\tau p_y} \quad (4-113)$$

An instance pattern of the MAP-PCP varying with respect to $\hat{\beta}_0$ when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$) is presented in Figure 4.10.

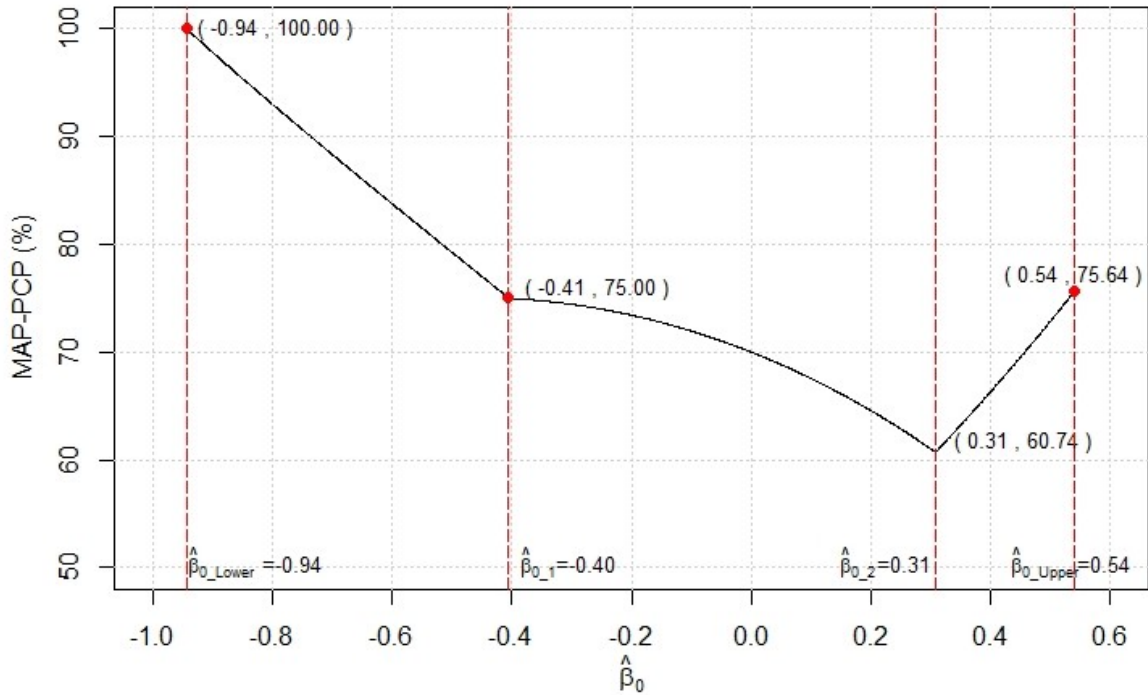


Figure 4.10 Variation Pattern of MAP-PCP vs. $\hat{\beta}_0$ when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)

4.2.3.5 PCPs versus enrichment efficiency

As is implied from Sections 4.2.3.3 and 4.2.3.4, the MLE- and the MAP-PCPs as well as the enrichment efficiency η can be approximated as a continuous function of the assumed recalibrated constant $\hat{\beta}_0$ over a specified domain for known, fixed p_y and β_1 . This offers the chance to mathematically link the disaggregate-level IDCM enrichment performance with the data condition of the enrichment sample by mathematically relating the PCPs to η .

Due to the complexity of the representation of enrichment efficiency with $\hat{\beta}_0$, the direct expression of the PCPs as a function of the η metric can be algebraically intractable. However, the deduction in the above sections still presents a way to link the two by means of the assumed recalibrated constant $\hat{\beta}_0$ as the pivot. Specifically, this can be achieved by numerically mapping each pair of MLE-/MAP-PCP and η expressed using the same pivot $\hat{\beta}_0$ for fixed and specified values of p_y and β_1 .

For demonstration of the following analysis, examples of the commensurate variations of the MLE- and the MAP-PCPs with respect to η when $p_y = 0.3$ and $\beta_1 = 1.792$ are illustrated in Figure 4.11 and Figure 4.12 (below) respectively.

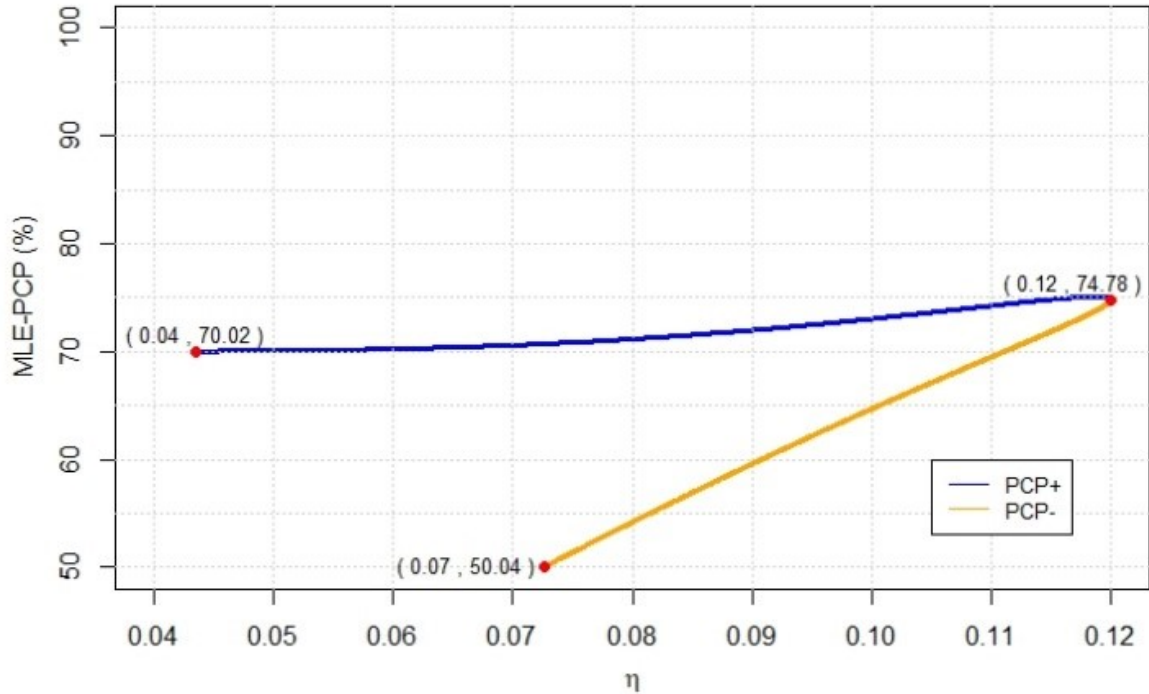


Figure 4.11 Variation Pattern of MLE-PCP vs. η when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)

The variation pattern of the MLE-PCP with respect to η can be divided by the point where $\hat{\beta}_0 = \hat{\beta}_{0-\eta}^*$ into two parts: the upper division (PCP_{MLE}^+) and the lower division (PCP_{MLE}^-), which are represented by the blue and orange curves respectively. For a particular η level, the MLE-PCP can take up to two possible values, each located in one of the two divisions. The defined domain of $\hat{\beta}_0$ also allows existence of a single MLE-PCP value for a given η level.

Inspecting Figure 4.11 reveals that both PCP_{MLE}^+ and PCP_{MLE}^- generally improve as η increases and converge at the maximum η level. This acknowledges the positive role of the explanatory power in improving the socio-demographic enrichment performance of the IDCM approach using the MLE estimator. On the other hand, the variability and variation patterns in the two divisions are found to be different: PCP_{MLE}^+ (blue) is relatively higher in terms of the absolute value while the improvement rate is slower. An opposite pattern is indicated in PCP_{MLE}^- (orange) where the MLE-PCP increases fast, yet the overall absolute values are low.

Mathematically, the discrepancy in the improvement rate of the two divisions can be explained by their respective variation patterns with respect to $\hat{\beta}_0$. Particularly, as can be observed in Figure 4.6, η is found to be asymmetrical at $\hat{\beta}_0 = \hat{\beta}_{0_ \eta}^*$: the change in the left wing is significantly steeper than that in the right wing. The result of this is that the variation of the MLE-PCP over the interval of $\hat{\beta}_0$ for the left wing of η is relatively slower than that over the interval for the right wing, revealed by the flatter PCP_{MLE}^+ in Figure 4.11. This also comply with the logic that the variability of $I(X, Y)$ becomes wider when $H(X)$ is larger and *vice versa*.

In addition, the difference in the absolute values of PCP_{MLE}^+ and PCP_{MLE}^- for the same η level in turn has its source in the transmission efficiency (Equation 4-29) of the quasi choice-making channel. Specifically, the decoding capability of the IDCM approach using the MLE estimator (measured by the MLE-PCP) depends also on the inherent transmission performance of the quasi channel, which is measured by the transmission efficiency. In other words, using the decoding scheme provided by the MLE estimator, the message transmitted over a channel with a higher transmission efficiency is more likely to be reproduced when the explanatory power of the output variable is known. It is therefore drawn that the enrichment performance using the MLE estimator can generally be improved with the growth in the explanatory power of the choice behaviour in inferring the enriched socio-demographic attribute. However, the variability and variation rate of the MLE-PCP are determined jointly by the explanatory power and the transmission efficiency of the choice-making communication channel.

The variation of the MAP-PCP can also be explored with respect to η using $\hat{\beta}_0$ as the pivot. Analogously divided into the upper and lower divisions at the maximum η , the pattern of the MAP-PCP is found to vary in a different manner in each division as compared to the MLE-PCP. Particularly, it is found from Figure 4.12 (below) that:

- 1) In the upper division, PCP_{MAP}^+ decreases as η increases and the maximum η produces the smallest PCP_{MAP}^+ ; and
- 2) In the lower division, PCP_{MAP}^- shows a V-shape by decreasing over $(\eta(\hat{\beta}_{0_upper}^*), \eta(\hat{\beta}_{0_2}^*))$ while increasing over $(\eta(\hat{\beta}_{0_2}^*), \eta(\hat{\beta}_{0_ \eta}^*))$.

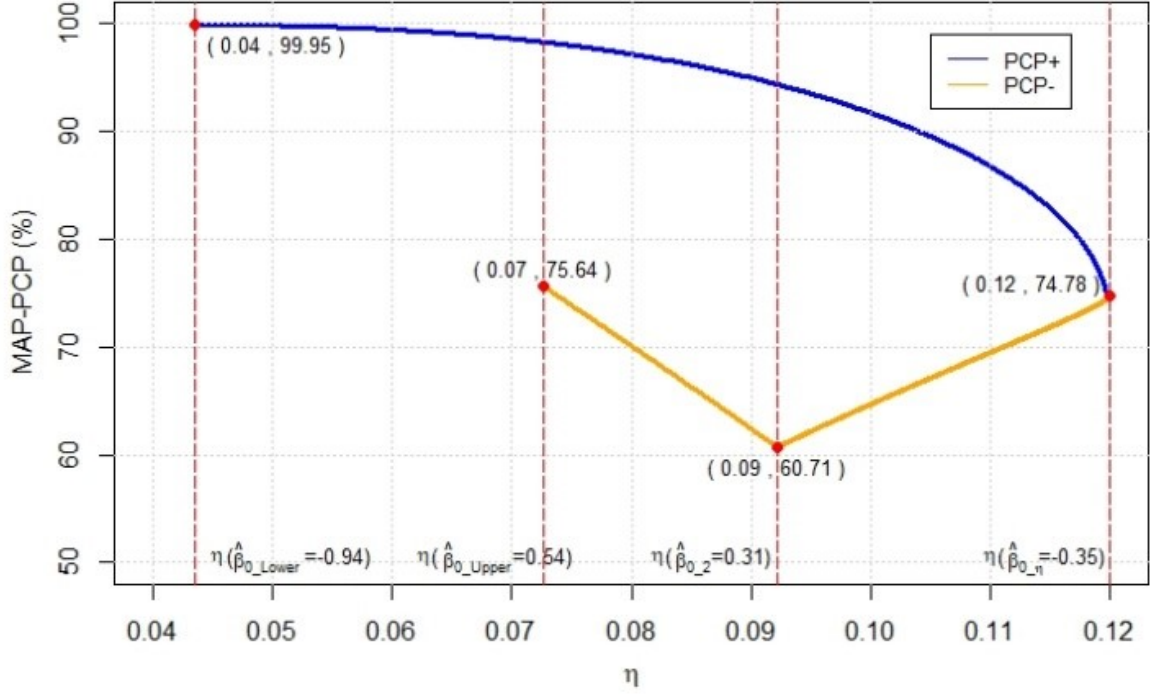


Figure 4.12 Variation Pattern of MAP-PCP vs. η when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)

The aforementioned pattern can be jointly attributed to η and $CE(X)$, which respectively reflect the explanatory power of the choice and the informativeness of the prior. Specifically, the MAP-PCP is found to equal the larger marginal probability $P(X)$ for X when $\hat{\beta}_0 \in (\hat{\beta}_{0_Lower}, \hat{\beta}_{0_1}) \cup (\hat{\beta}_{0_2}, \hat{\beta}_{0_Upper})$. This corresponds to the intervals over which the prior is informative enough such that the prior itself can virtually exclusively drive the MAP-PCP. These intervals also keep with the range over which $H(X)$ is low (Figure 4.6), i.e. high $CE(X)$. Interestingly, it is noted that the indifferent MLE- and MAP-PCP not only happen in the limit where $CE(X)$ is minimised and the MAP estimator collapses to its MLE counterpart. In contrast, the MAP-PCP is found to equal the MLE-PCP when $\hat{\beta}_0 \in (\hat{\beta}_{0_1}, \hat{\beta}_{0_2})$. This represents the domain where the prior is not informative enough, reflected by the high $H(X)$ in Figure 4.6. The significance of the explanatory power is hence revealed over this interval, as reflected by the growth in the MAP-PCP with respect to η .

Note that, whilst the examples are presented with specific instances for specific p_y and β_1 , the analysis in this and the following section is argued to be generalisable to any p_y and β_1 as the respective generalisability of the PCPs and the η with respect to the pivot ($\hat{\beta}_0$) have been proven in Sections 4.2.3.3 and 4.2.3.4.

4.2.3.6 Comparison between PCPs using MLE and MAP estimators

By comparing the expressions of the MLE-PCP and the MAP-PCP with respect to $\hat{\beta}_0$ (Equations 4-76 and 4-111), it is easily noticed that:

$$PCP_{MLE} = PCP_{MAP} \text{ when } \hat{\beta}_0 \in (\hat{\beta}_{0-1}, \hat{\beta}_{0-2}). \quad (4-114)$$

In contrast, the MAP-PCP is found to outperform the MLE-PCP over the rest of the domain. This confirms the intuition and prior simulation and empirical work (Zhao, Pawlak and Polak, 2018, 2019) that the MAP estimator performs no worse than the MLE as it incorporates additional information about the prior distribution of the enriched variable. Moreover, this also indicates that, for the proposed IDCM enrichment, the impact of the prior information is effective only when the amount of informativeness, which can be measured by $CE(X)$, is large enough. In other words, the effect of the prior included in the MAP estimator in improving the performance of the IDCM can be ignorable when $\hat{\beta}_0 \in (\hat{\beta}_{0-1}, \hat{\beta}_{0-2})$.

The findings above provide a guideline for identifying the enrichments that can be improved by using the MAP estimator instead of the MLE, defined as functions of $\hat{\beta}_0$, i.e.:

$$\hat{\beta}_0 \in \left(\ln \frac{1-p_y}{\tau p_y}, \ln \frac{2(1-p_y)}{p_y(1+\tau)} \right) \cup \left(\ln \frac{(1-p_y)(1+\tau)}{2p_y\tau}, \ln \frac{\tau+1+\sqrt{\tau^2+2\tau+16p_y(1-p_y)+1}}{4\tau p_y} \right). \quad (4-115)$$

The corresponding values for $P(X)$ can be determined based on the above equation given the expression of $p(x_0)$ with $\hat{\beta}_0$ (Equation 4-45). Replacing $\hat{\beta}_0$ with the bounds of the interval above, it can be found that the market share of X that can play an effective role in providing the prior information can be calculated as:

$$p(x_0) < \frac{(\tau-1)p_y+2}{2(\tau+1)} \text{ or } p(x_0) > \frac{1}{2} + \frac{p_y(\tau-1)}{2(\tau+1)}. \quad (4-116)$$

Over the interval of $p(x_0)$ other than the equation above, the individual-level enrichment performance is indifferent between using the MLE and the MAP estimator. In these cases, the MLE estimator can be used to perform the IDCM approach regardless of the availability of the prior information to make an effective yet information-efficient estimation method.

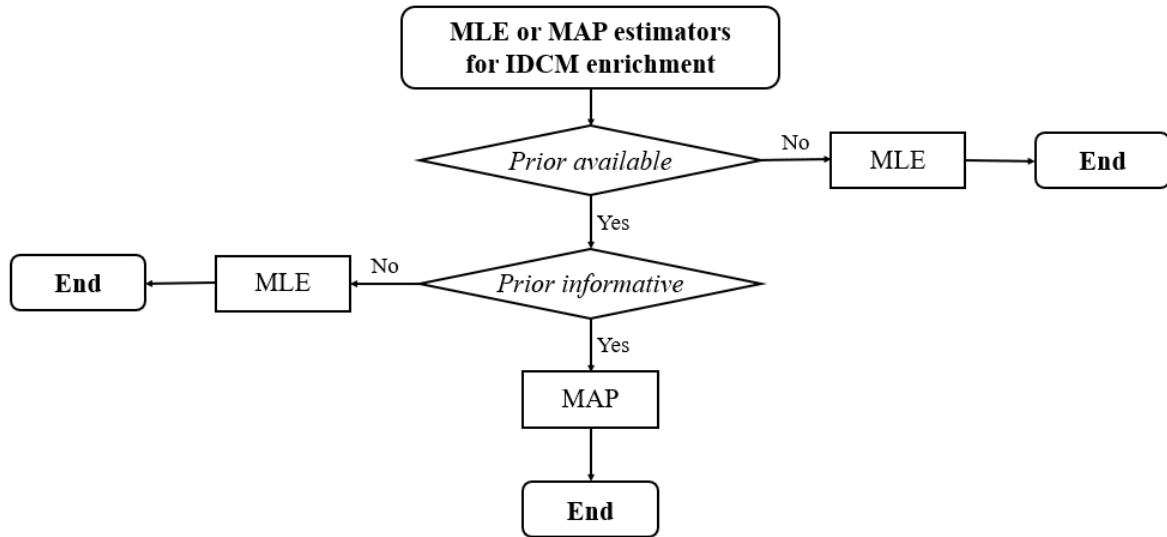


Figure 4.13 Criteria for Selection between Estimation Methods for IDCM Enrichment

The flow chart in Figure 4.13 demonstrates the criteria for selecting the better estimation methods between the MLE and the MAP estimators to solve the IDCM approach so as to, based on the theoretical analysis in Section 4.2.3, make the effective yet efficient use of the available information about X . When the IDCM approach is applied to empirical contexts, however, the socio-demographic enrichment performance using the two estimators may change due to the variation in data conditions, e.g. quality of the prior. This will be discussed later in Chapters 5 – 6, in which the above theoretical guideline is supplemented based on two empirical studies.

4.3 Summary

In this chapter, the methodology of the Inverse Discrete Choice Modelling (IDCM) socio-demographic enrichment framework was systematically established in terms of the enrichment approach and the accompanying performance theory.

The IDCM socio-demographic enrichment approach was developed in Section 4.1 respectively from the theoretical and the empirical aspects. Specifically, Section 4.1.1 presented the detailed derivations for the econometric formulation of the IDCM approach. This includes the general econometric formulation which implies the solution to the optimisation problem that the likelihood of all the individuals in the sample being characterised by specific attributes in presence of the observed behaviour patterns is maximised. Based on this general formulation, the basic econometric formulation of the IDCM approach was developed, by assuming that the choice behaviour of all the individuals in the sample is independent of each other, which

enables the parallelisation of the maximisation procedure. Moreover, an extended formulation was also established to capture the correlation between a group of individuals in the sample by including an attribute that characterise the commonality of each group and by performing the paralleled computation up to the group-level only. Two estimators corresponding to the maximum *a posteriori* (MAP) and the maximum likelihood (MLE) estimation methods were proposed to solve the developed econometric formulation respectively given presence and absence of prior information. Different types of socio-demographic attribute can be enriched using the developed econometric formulation. A guideline was thus provided, for selecting the suitable estimation algorithm for enriching different, or combinations of, socio-demographic attribute. Section 4.1.2 then demonstrated the requirements and procedure for implementing the econometric formulation in practice. This, along with the econometric formulation, constitute the theoretical and empirical aspects of the developed IDCM approach.

On the other hand, Section 4.2 documented how the IDCM enrichment performance theory is established to characterise the link between the metrics respectively measuring the enrichment performance of the IDCM approach and the data condition of the enrichment sample. In particular, Section 4.2.1 described the indicators that can be used to measure the enrichment performance of the IDCM approach, as well as other enrichment methods, respectively at the aggregate sample level and the disaggregate individual level. The ‘percentage correctly predicted’ (PCP) for evaluating the disaggregate-level enrichment performance was also employed as the metric for enrichment performance in the IDCM performance theory. On the other hand, Section 4.2.2 discussed the metrics that can quantify the covariance structure underlying the enrichment sample, as well as explained the reason to develop and utilise the enrichment efficiency η metric in the IDCM performance theory. The mathematical link between the PCP and η was then established in Section 4.2.3. Particularly, the IDCM performance theory was developed for characterising the aforementioned link in the enrichment of a single binary socio-demographic attribute X from single observations of single binary choice behaviour pattern Y for each individual, which was termed as the binary enrichment context. A concept of assumed recalibrated constant ($\hat{\beta}_0$) was proposed, based on the recalibration of constant commonly applied in traditional discrete choice modelling, to mathematically approximate the joint probability distribution $P(X, Y)$ underlying the enrichment sample. By showing that both the PCP and the η are expressed as functions of $P(X, Y)$, the link between the two metrics was then effectively established as the IDCM

performance theory using $\hat{\beta}_0$ as the pivot. By comparing the respective variation patterns of the MLE- and the MAP-PCPs with respect to η , Section 4.2.3 also presented a theoretical guideline for selecting the best-performing yet information-effective estimation method between the MLE and the MAP estimators.

In the following chapters, the enrichment performance of the IDCM approach as well as the capability of the IDCM performance theory are explored in two empirical enrichment contexts and compared with representatives of other existing enrichment methods.

5 EMPIRICAL APPLICATION: IDCM PERFORMANCE COMPARISON AND THEORETICAL VALIDATION

The Inverse Discrete Choice Modelling (IDCM) socio-demographic enrichment framework which consists of the enrichment approach and the accompanying performance theory has been systematically established in Chapter 4. This chapter assesses the capability of that framework in an empirical application by enriching a number of socio-demographic attributes based on observations of the travel-related behaviour patterns revealed in London Travel Demand Survey (LTDS) data. Specifically, the proposed IDCM enrichment framework is validated in the following ways, according to the two components outline above:

- 1) Whether the IDCM enrichment performance theory, as proposed and formulated in Section 4.2.3, is able to estimate and interpret the socio-demographic enrichment performance of the IDCM approach (Section 4.1) given different data conditions. This is examined by assessing the consistency of the actual enrichment performance level achieved in the empirical application with the expected values derived using the IDCM performance theory.
- 2) The relative capability of the IDCM framework compared to other existing socio-demographic enrichment methods in terms of the enrichment performance and the transferability, i.e. the applicability of the approach under different data conditions or across application contexts. This is achieved via a comparative analysis between the IDCM framework and logistic regression (LGR) and support vector machines (SVMs).

Based on the above two dimensions, this chapter is structured as follows. Section 5.1 describes the data exploited in this application, and how the variables explored in the empirical enrichment are selected and modified to fit the scope of the present analysis. Section 5.2 explains how the empirical experiments are designed, in light of the aforementioned aims to validate the IDCM framework. This is followed by the results and corresponding analysis for the two streams of assessments separately: Section 5.3 demonstrates the evaluation of the ability of the IDCM performance theory to estimate the IDCM enrichment performance, while

Section 5.4 provides the comparative analysis between the IDCM framework and LGR and SVM. A brief summary of the current chapter is presented in Section 5.5.

5.1 Data and variables

The 2015/16 LTDS is used to perform this empirical enrichment application. LTDS is a continuous annual household survey capturing data on the travel habit information of approximately 8,000 randomly-selected households covering all Greater London boroughs and the area outside Greater London but within the city's orbital motorway, the M25 (Transport for London, 2011). It also captures the demographic information of the interviewed respondents and their households, and information about the trips they have made on specified dates, including both weekdays and weekends. The fact that the LTDS data contains sufficient socio-demographic information as well as stated trip-related behaviour patterns makes it suitable for the present analysis. In particular, two household-level socio-demographic attributes (income level and car ownership) and four individual-level attributes (gender, work status, driving licence possession and entitlement to a pass for reduced fares on public transport (PT)) are selected as the socio-demographic attributes to be enriched in the current application, since these are available in the LTDS dataset. In terms of the input behaviour features for the enrichment, the choice of departure time-of-day, travel mode choice and the frequency of travel are three trip-related behavioural variables that can be straightforwardly obtained from the LTDS data. Since these behaviour patterns can also be easily extracted from transport big data sources using existing techniques, as introduced in Sections 2.1.1 – 2.1.5, the feasibility of using these patterns to mimic those could be obtained from typical transport big datasets is verified. It is hence feasible to employ the LTDS data as an analogue to the datasets that can be processed from raw transport big data sources by stripping out the corresponding socio-demographic information.

Existing studies have shown that the departure time-of-day and travel mode are more significantly driven by the attributes of the choice alternatives themselves, e.g. travel time and monetary cost, than the socio-demographic attributes of the choice-makers (Train and McFadden, 1978). Despite the relatively insignificant influence that socio-demographic characteristics have on these choices, they are still adopted in the present analysis because the aim of the current application is to assess whether the socio-demographic enrichment performance using the IDCM approach can be predicted using the IDCM performance theory.

It is argued that, once this ability of the IDCM framework is verified, it can be applicable to the behaviour features to which the explanatory power of the enriched socio-demographic attribute is more vital. The possible effect of using a behaviour pattern that is relatively insignificantly related to the enriched attribute as the input feature on the enrichment performance using the IDCM approach is discussed later in Chapter 7.

The above variable selections are also supported by the IDCM enrichment performance theory, as explained previously in Section 4.2.3, which indicates the link between the metrics measuring the IDCM enrichment performance and the data condition of the enrichment sample, respectively. In particular, the IDCM performance theory suggest that the ‘percentage correctly predicted’ (PCP) using the IDCM approach can be related to the entropy $H(X)$ of the enriched attribute (referred to as the entropy), and the enrichment efficiency (η) as the ratio of the mutual information $I(X, Y)$ between the socio-demographic attribute and the behaviour feature (referred to as the mutual information) and the entropy. This implies that the validation of the IDCM performance theory requires sufficient variation in the entropy of the enriched socio-demographic attribute, as well as in the mutual information. The postulated impact of the entropy is thus explored by enriching the socio-demographic attributes having different probability distribution patterns, ranging from the nearly evenly distributed ones (e.g. gender) to the less evenly distributed ones (e.g. income levels). This can be more directly observed from Figure 5.1, which illustrates the respective market shares of gender and of falling into one of the ten income levels defined in the 2015/16 LTDS.

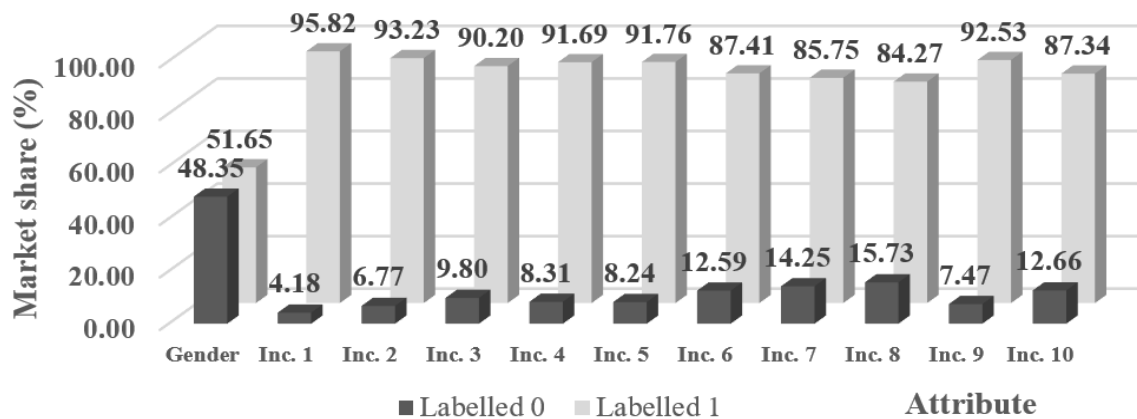


Figure 5.1 Market Shares of Gender and Inclusion in Specific Income Level Groups

Supported by existing studies, the mutual information between the socio-demographic attribute and the input behaviour feature also varies in light of the above selections. Specifically, for instance, the consensus in literature is that the possession of a driving licence and car ownership

can be positively related to the choice of car travel over other modes of transport (Ben-Akiva and Lerman, 1974; Ding *et al.*, 2017). Similarly, the elasticity to peak-hour travel decisions of commuters and students is different from that for people not employed, *ceteris paribus* (Abkowitz, 1981), while, by economic rationality, travellers holding a reduced-fare PT pass are more prone to travel by PT than those without such passes. Moreover, it is found that the number of travel activities (frequencies) can be promoted by either the possession of auto vehicles or a higher income level (Chang *et al.*, 2014). It is therefore established that the use of the three travel-related behaviour patterns stated above as the input features for the enrichment of the six selected socio-demographic attributes encompasses not only a sufficient number of examples for the comparative analysis, but also sufficient variation in entropy and mutual information, permitting a comprehensive validation of the IDCM performance theory.

Whilst only trip-related behaviour patterns are investigated, the results of this analysis are applicable in any behaviour contexts including, but not limited to, travel behaviour analysis, since microeconomic consumer theory suggests that people's behaviours are driven, or can be affected, by their socio-demographic attributes.

5.2 Experimental design

The experimental design for the current empirical application is described in this section from the following three perspectives:

- 1) To modify the selected variables to fit the scope of the present analysis;
- 2) To determine the task of each experiment, i.e. the input feature (behaviour pattern) and the output (enriched attribute), and how to generate the samples required to perform each enrichment experiment; and
- 3) To define the setup of the IDCM econometric formulation for the current application, as well as of the LGR and the SVM, explaining their selection as the representatives of existing socio-demographic enrichment methods in the comparative analysis with the IDCM framework.

5.2.1 Modification of variables

As is explained in Section 5.1, the variables selected for the empirical validation of the IDCM framework need to be binary given that the IDCM performance theory (Section 4.2.3) is proven

for the binary enrichment context only. Consequently, work status is adjusted to represent whether the traveller is unemployed or not (i.e. all the other status except ‘unemployment’). In contrast, the ten income levels are revised into ten binary attributes. Note that the ten income levels are correlated, making it questionable to directly consider the revised binary variable as whether the household’s income belongs to a specific income level. However, as the IDCM performance theory requiring validation in this empirical context is only applicable to binary enrichment, these ten binary income-level attributes are enriched separately in the subsequent enrichment experiments, each reflecting whether the income of the individual’s household lies out of a specific income category or not. The remaining four socio-demographic attributes, viz. gender, car ownership, driving licence possession and entitlement to a PT pass, are inherently binary and thus require no additional modification.

Likewise, the three behavioural variables reflecting the travel-related behaviour features are also adjusted into binary variables. Specifically, the choice behaviours regarding the departure time-of-day and the travel mode are separated to represent three binary choice variables:

- 1) whether the trip is made during peak hours;
- 2) whether travelling by PT; and
- 3) whether travelling by car.

Meanwhile, travel frequencies are adapted to represent three binary travel frequency patterns:

- 1) whether any trips have been made or not;
- 2) whether more than a single trip (excluding one trip only) has been made or not; and
- 3) whether two or more trips (excluding two trips only) have been made or not.

The behaviour patterns and socio-demographic attributes as modified above respectively represent the input features and the output enriched attributes for this analysis.

5.2.2 Enrichment experiments

In light of the above discussion, the variables selected for this empirical application include three binary choice variables and three binary ordinal travel frequency variables as the input features as well as 15 binary socio-demographic attributes as the enriched socio-demographic attributes (output). This results in 90 combinations $((3 + 3) \times 15)$ of input-and-output pairs, which constitute the set of 90 enrichment experiments, each uniquely involving the enrichment

of one of the 15 attributes from one of the six behaviour features using the IDCM approach and the two supervised machine learning (ML) methods.

Figure 5.2 conceptualises the aforementioned experimental design of this empirical application, which also includes the modification of selected variables.

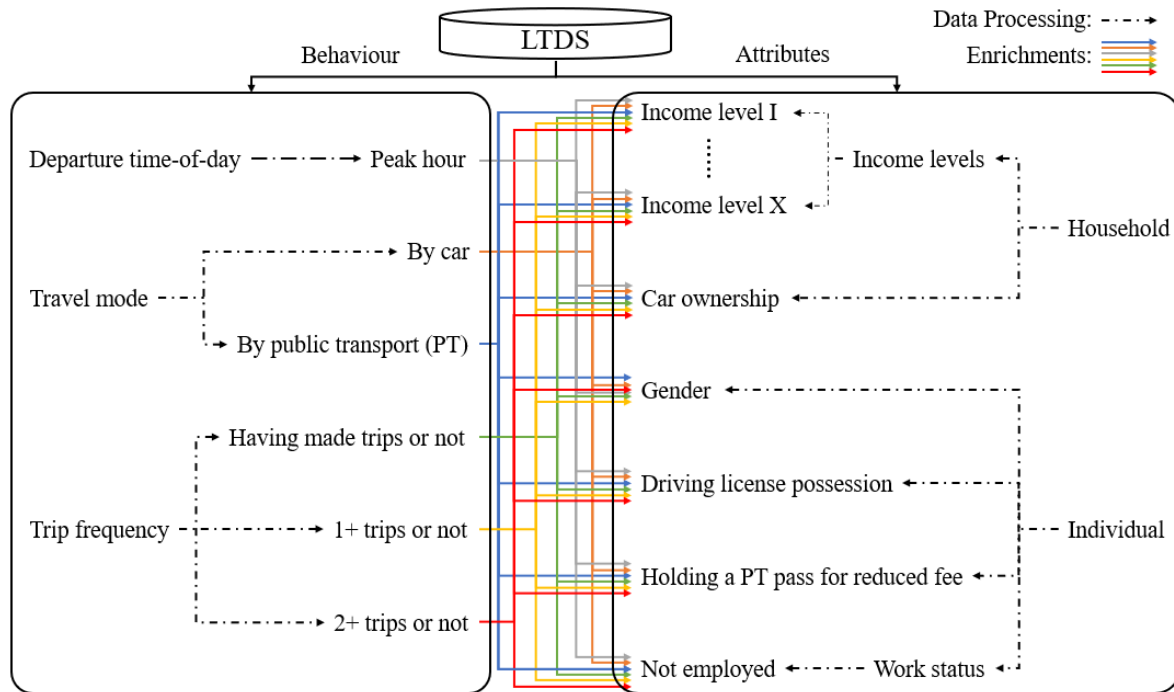


Figure 5.2 Behaviour Features and Enriched Attributes for Enrichment using IDCM, LGR and SVM

Recall that Section 3.2.1 suggests the need for two types of samples for an IDCM enrichment:

- 1) an *enrichment sample* where the socio-demographic information is absent and thus requires enrichment; and
- 2) an *estimation sample* upon which the IDCM kernel(s) can be estimated for its subsequent enrichment using the IDCM approach.

Similarly to the estimation sample required by the IDCM approach, a *training sample* is also needed for the two supervised ML methods in order to train the respective algorithms for the prediction of new instances. Hence, the employed LTDS dataset is divided randomly into two subsamples: (i) 20% as the ‘small-but-thick’ estimation (training) subsample; and (ii) 80% as the ‘big-but-thin’ enrichment subsample. The proportion is defined to mimic the relative sample sizes of the two type of data sources. To avoid sample specificity due to such a random split, a cross-validation in the form of a k -fold holdout method is employed where $k = 20$ (Kohavi, 1995). This means that the random split is repeated 20 times in each set of enrichment

experiments such that 20 estimation subsamples and enrichment subsamples are generated accordingly. To imitate anonymous ‘big-but-thin’ data, the socio-demographic information in each enrichment subsample is removed while the stated travel behaviour patterns are preserved, serving as the input feature for the enrichment. These enrichment subsamples maintain the structure and content of the data that can be extracted and processed from transport big datasets. Conversely, the removed socio-demographic attributes are used as the reference against which the accuracy of the enriched socio-demographic attributes is assessed to derive the PCPs of the methods performed in these enrichment experiments.

For each of the 90 set of designed enrichment experiments, a specified socio-demographic attribute is enriched for all the individual respondents in the 20 enrichment subsamples in respect to each particular stated behaviour pattern, using the IDCM approach (both the MLE and MAP estimators) and the LGR and SVM. The PCPs using the three approaches across the 20 repetitions in each set of experiments, in conjunction with the average PCP and the standard deviation, are collected for the comparative analysis of the IDCM approach and the two supervised ML methods. To validate the developed IDCM enrichment performance theory, the expected PCPs are calculated based on the IDCM performance theory, and their consistency to the corresponding actual MLE- and MAP-PCPs collected from each sets of enrichment experiments using the IDCM approach is compared.

5.2.3 Setup of the approaches

This section specifies the setup for the econometric formulation of the IDCM approach implemented according to the above experimental design, thus supplementing the methodology explained in Section 4.1. Also, this section explains the selection and setup of the LGR and SVM as the representative methods for the comparative analysis with the IDCM approach, building on the methodologies set out in Section 2.3.2.1 and Section 2.3.2.2, respectively.

In terms of the IDCM econometric formulation, the components for the linear-in-parameter utility function of the binary logit IDCM kernel need to be determined such that the IDCM kernel can be estimated upon the 20% estimation subsample. In traditional travel behaviour modelling using discrete choice models (DCMs), as is mentioned in Section 3.1.4, the selection of a model specification depends partially on the overall goodness-of-fit (GOF). As the pilot investigation into the IDCM approach has shown, however, the performance when enriching a single socio-demographic attribute using the IDCM performed on a specific behaviour pattern

cannot be improved by using an IDCM kernel with a higher GOF due to the inclusion of any other variables (Zhao, Pawlak and Polak, 2018). For instance, when using a DCM for travel mode choice, increasing the overall GOF by adding car ownership as an explanatory variable would not improve the enrichment performance of the corresponding IDCM enrichment of the gender attribute as compared to using an IDCM kernel including only gender as the predictor. In the current analysis, therefore, the explanatory variables other than the enriched socio-demographic attribute are excluded from the IDCM kernel. The utility function of such an IDCM kernel, which takes the simplest form to reduce additional computational efforts for DCM estimation, is defined as:

$$\begin{aligned} V(y_0) &= 0 \\ V(y_1) &= \beta_0 + \beta_1 X \end{aligned} \tag{5-1}$$

where y_0, y_1 possible outcome values (POVs) of the choice Y used for the enrichment;
 x_0, x_1 POVs of the enriched attribute X which are labelled 0 and 1 respectively;
 β_0, β_1 taste parameters of the DCM kernel to be estimated.

The corresponding MLE and MAP estimators for the present analysis, given the present binary enrichment context, can be correspondingly expressed as:

$$X_{MLE} = \arg \max_X P(Y|X). \tag{5-2}$$

$$X_{MAP} = \arg \max_X P(Y|X)P(X). \tag{5-3}$$

With respect to the methods selected for the comparative analysis, the literature reviewed in Section 2.1.6 suggests using ML methods as the representative of existing methods because they constitute the mainstream methods used in existing attempts for enriching socio-demographic attributes. According to the classification of ML methods as discussed in Section 2.3.1, the data requirement and enrichment purpose of the IDCM framework jointly imply that the comparison should be undertaken against supervised ML methods for binary classification to fit the current research scope. In order to conduct a comprehensive comparative analysis, therefore, the LGR and the SVM are selected to represent, respectively, parametric and nonparametric learning algorithms as two types of supervised learning approaches classified based on their functional form.

Specifically, LGR is chosen because it preserves the same functional form as the IDCM kernel in the current binary enrichment context. More importantly, it is employed to predict the socio-demographic information in one of the few studies, to the author's best knowledge, in the context of travel behaviour modelling (Bwambale, Choudhury and Hess, 2019), the details of which are documented previously in Section 2.1.6. On the other hand, the SVM is employed because of its widespread use in classification tasks for socio-demographic enrichment in many contexts (e.g. Gebru *et al.* (2017) and Auld *et al.* (2015), which are also discussed previously in Section 2.1.6). The setups for these two supervised ML methods are provided below.

In terms of the setups, the logistic function in Equation 2-1 (Section 2.3.2.1) expresses the probability p that the predicted variable is labelled 1, as reproduced below.

$$p = f(x) = \frac{1}{1 + b^{x\beta}}; \quad (5-4)$$

where b the base of the logarithm;

X the set of predictors affecting the probability;

β The vector of parameters associating X with p which needs estimation.

In the current enrichment context, the proposed LGR classifier involves inferring the socio-demographic attribute X (response) from the observed choice Y (predictor). Hence, p represents the probability of $X = x_1$, i.e. $p(x_1)$. Using the natural base of logarithm, $p(x_1)$ can be expressed as:

$$p(x_1 = 1) = \frac{1}{1 + e^{\theta_0 + \theta_1 Y}}. \quad (5-5)$$

where θ_0, θ_1 model parameters (distinguished from the tastes of the IDCM kernel).

Label the POVs of Y , i.e. y_0 and y_1 , as 0 and 1, respectively. The probability of X retaining each of the POVs conditional on the value of Y can be derived as:

$$P(X|Y) = \begin{cases} p(x_0|y_0) = \frac{e^{\theta_0}}{1 + e^{\theta_0}} \\ p(x_1|y_0) = \frac{1}{1 + e^{\theta_0}} \\ p(x_0|y_1) = \frac{e^{\theta_0 + \theta_1}}{1 + e^{\theta_0 + \theta_1}} \\ p(x_1|y_1) = \frac{1}{1 + e^{\theta_0 + \theta_1}} \end{cases} \quad (5-6)$$

The threshold that turns the LGR function above into a binary classifier is set out as 0.5, i.e. x_i is predicted such that the probability of X conditional on the observed Y , $p(x_i|Y)$, is greater than 0.5. In the binary enrichment context, this is equivalent to selecting the maximum $P(X|Y)$ for a given Y . Thus, the binary classifier based on LGR above can be expressed as:

$$X_{LGR} = \arg \max_X P(X|Y) = \arg \max_X \frac{P(Y|X)P(X)}{P(Y)} = \arg \max_X P_T(Y|X)P_T(X). \quad (5-7)$$

where $P_T(Y|X)$ probability of observing Y given X in the estimation (training) sample.

$P_T(X)$ probability of X in the estimation (training) sample.

Although the LGR classifier in Equation 5-7 retains a similar form to the MAP estimator of the IDCM approach (Equation 5-3), the probability distribution of X presented in the two equations have different meanings. Specifically, $P_T(X)$ in the LGR classifier characterises the probability distribution of X in the estimation (training) sample, while $P(X)$ in the MAP estimator can refer to any prior information regarding the distribution of X in the enrichment sample, which can be, but is not necessarily, the same as $P_T(X)$.

On the other hand, the general introduction to SVMs in Section 2.3.2.2 indicates that an SVM is a trained non-probabilistic binary linear classifier which builds one, or a set of, hyperplanes in a high- or infinite-dimensional space to separate data points, each viewed as a n -dimensional vector with a $(n - 1)$ dimensional hyperplane (Cortes and Vapnik, 1995). It can perform linear or nonlinear classification in light of different sample sizes and feature spaces. For data that are not linearly classifiable, a similarity function can be used to implicitly map their inputs into high-dimensional feature spaces (Theodoridis and Koutroumbas, 2009). Particularly, a non-linear classifier with a Gaussian similarity function is recommended by Ng for samples with intermediate to large size while the number of features is small (Ng, 2000), which fits the current implementation where only one feature (behaviour pattern) is employed as the predictor. The Gaussian similarity function takes the form:

$$k(\vec{Y}_i, \vec{Y}_j) = e^{(-\gamma \|\vec{Y}_i - \vec{Y}_j\|^2)}, \quad (5-8)$$

where $\vec{Y}_i$ the vector of features for the i th data point in a sample;

γ kernel-specific parameter.

When training an SVM with the similarity function in Equation 5-8, two parameters c and γ must be considered. Common to all the SVM similarity functions, c trades off between

increasing the margin size and misclassification, while γ defines the importance of each single training example as a specific parameter of the function. Either a small c or a small γ can result in high bias yet low variance, and *vice versa*. Thus, various c and γ are tested in this study to find the relatively best-performing values given the training sample. The sets of possible values for both parameters are defined as $\{0.01, 0.1, 1, 10\}$ (Pedregosa *et al.*, 2011), which hence leads to 16 combinations of c and γ to be tested for the performance of the SVM in this application.

5.3 Validation of IDCM performance theory

This section assesses the ability of the IDCM performance theory to estimate (predict) the enrichment performance of the IDCM approach using both the MLE and the MAP estimators given specified data conditions. This is achieved by comparing the actual PCPs obtained from conducting the designed experiments and the corresponding expected values calculated based on the entropy $H(X)$ and the enrichment efficiency (η) using the IDCM performance theory. This assessment is performed across the 20 repetitions in each set of enrichment experiments, using either the paired t -test or the Wilcoxon signed-rank test, depending on whether the PCPs are normally distributed. The null hypothesis made for these tests proposes that there is no difference between the mean of the actual and expected PCPs across the 20 repetitions in each set of enrichment experiments.

Table 5.1 and Table 5.2 (below) present the $H(X)$ and η for each of the 90 sets of enrichment experiments, along with the results for the comparison between the actual and the expected values for the MAP- and the MLE-PCPs, respectively.

Table 5.1 Data Conditions and Results of Comparative Analysis for Actual and Expected MAP-PCPs

Exp NO.	Behaviour Pattern	Attribute	$H(X)$ (%)	η (%)	MAP-PCP (%)		Sd. ^a (%)	Test ^b	p -value
					Actual	Expected			
1	Peak	Gender	99.90	0.00	51.85	52.07	0.53	W	0.109
2	Peak	Not employed	83.79	1.33	73.17	73.17	0.00	T	1.000
3	Peak	PT Pass	88.07	0.55	69.93	69.93	0.00	T	1.000
4	Peak	Ownership	88.34	0.03	69.79	69.79	0.00	T	1.000
5	Peak	Licence	91.11	0.00	67.40	67.40	0.00	T	1.000
6	Peak	Income (1)	25.42	0.28	95.79	95.79	0.00	W	1.000
7	Peak	Income (2)	36.02	0.20	93.15	93.15	0.00	T	1.000
8	Peak	Income (3)	45.02	0.02	90.48	90.48	0.00	T	1.000
9	Peak	Income (4)	40.51	0.03	91.88	91.88	0.00	T	1.000
10	Peak	Income (5)	41.14	0.01	91.75	91.75	0.00	T	1.000

11	Peak	Income (6)	54.28	0.00	87.50	87.50	0.00	T	1.000
12	Peak	Income (7)	60.77	0.04	85.16	85.16	0.00	T	1.000
13	Peak	Income (8)	62.52	0.04	84.49	84.49	0.00	T	1.000
14	Peak	Income (9)	37.55	0.05	92.54	92.54	0.00	T	1.000
15	Peak	Income (10)	54.94	0.01	87.19	87.19	0.00	T	1.000
16	PT	Gender	99.92	0.01	51.64	51.72	0.18	W	0.109
17	PT	Not employed	78.49	0.38	76.57	76.57	0.00	T	1.000
18	PT	PT Pass	88.37	0.02	69.85	69.85	0.00	T	1.000
19	PT	Ownership	92.42	1.91	65.58	65.98	0.12	T	0.100
20	PT	Licence	92.08	0.29	66.36	66.36	0.00	T	1.000
21	PT	Income (1)	24.43	0.02	95.94	95.94	0.00	T	1.000
22	PT	Income (2)	35.73	0.04	93.26	93.26	0.00	T	1.000
23	PT	Income (3)	45.79	0.00	90.39	90.39	0.00	W	1.000
24	PT	Income (4)	41.25	0.03	91.75	91.75	0.00	W	1.000
25	PT	Income (5)	41.08	0.00	91.72	91.72	0.00	T	1.000
26	PT	Income (6)	54.71	0.02	87.39	87.39	0.00	T	1.000
27	PT	Income (7)	59.52	0.03	85.61	85.61	0.00	T	1.000
28	PT	Income (8)	62.80	0.01	84.21	84.21	0.00	T	1.000
29	PT	Income (9)	38.48	0.01	92.52	92.52	0.00	T	1.000
30	PT	Income (10)	55.07	0.01	87.23	87.23	0.00	T	1.000
31	Car	Gender	99.93	0.00	51.55	51.75	0.20	W	0.100
32	Car	Not employed	78.66	0.13	76.54	76.54	0.00	T	1.000
33	Car	PT Pass	88.39	0.06	69.89	69.89	0.00	T	1.000
34	Car	Ownership	92.46	7.64	66.03	66.03	0.00	T	1.000
35	Car	Licence	92.13	1.54	66.41	66.41	0.00	T	1.000
36	Car	Income (1)	24.10	0.53	95.92	95.92	0.00	T	1.000
37	Car	Income (2)	35.68	0.45	93.26	93.26	0.00	T	1.000
38	Car	Income (3)	45.60	0.02	90.38	90.38	0.00	T	1.000
39	Car	Income (4)	41.51	0.09	91.73	91.73	0.00	T	1.000
40	Car	Income (5)	41.11	0.00	91.74	91.74	0.00	T	1.000
41	Car	Income (6)	54.92	0.00	87.36	87.36	0.00	T	1.000
42	Car	Income (7)	59.23	0.08	85.61	85.61	0.00	T	1.000
43	Car	Income (8)	62.80	0.12	84.22	84.22	0.00	T	1.000
44	Car	Income (9)	38.45	0.07	92.51	92.51	0.00	T	1.000
45	Car	Income (10)	55.19	0.05	87.21	87.21	0.00	T	1.000
46	0/1	Gender	99.99	0.02	50.36	50.36	0.00	W	1.000
47	0/1	Not employed	88.68	2.67	69.40	69.40	0.00	T	1.000
48	0/1	PT Pass	90.97	0.20	67.95	67.95	0.00	T	1.000
49	0/1	Ownership	92.07	0.18	66.45	66.45	0.00	T	1.000
50	0/1	Licence	96.93	1.01	60.10	60.10	0.00	T	1.000
51	0/1	Income (1)	28.54	0.20	94.87	94.87	0.00	T	1.000

52	0/1	Income (2)	40.26	0.34	92.19	92.19	0.00	T	1.000
53	0/1	Income (3)	47.33	0.03	89.96	89.96	0.00	T	1.000
54	0/1	Income (4)	42.42	0.07	91.29	91.29	0.00	T	1.000
55	0/1	Income (5)	41.77	0.01	91.55	91.55	0.00	T	1.000
56	0/1	Income (6)	54.92	0.00	87.39	87.39	0.00	T	1.000
57	0/1	Income (7)	60.05	0.01	85.21	85.21	0.00	T	1.000
58	0/1	Income (8)	58.55	0.12	85.69	85.69	0.00	T	1.000
59	0/1	Income (9)	36.91	0.05	93.03	93.03	0.00	T	1.000
60	0/1	Income (10)	50.92	0.21	88.66	88.66	0.00	T	1.000
61	1+	Gender	99.99	0.01	50.44	50.44	0.00	T	1.000
62	1+	Not employed	88.77	2.02	69.43	69.43	0.00	T	1.000
63	1+	PT Pass	90.79	0.20	67.87	67.87	0.00	T	1.000
64	1+	Ownership	92.18	0.13	66.49	66.49	0.00	T	1.000
65	1+	Licence	96.72	0.83	59.87	59.87	0.00	W	1.000
66	1+	Income (1)	29.58	0.29	94.85	94.85	0.00	T	1.000
67	1+	Income (2)	39.66	0.14	92.28	92.28	0.00	T	1.000
68	1+	Income (3)	46.54	0.04	90.00	90.00	0.00	T	1.000
69	1+	Income (4)	42.84	0.01	91.33	91.33	0.00	T	1.000
70	1+	Income (5)	41.89	0.01	91.54	91.54	0.00	T	1.000
71	1+	Income (6)	55.17	0.01	87.35	87.35	0.00	T	1.000
72	1+	Income (7)	60.57	0.00	85.31	85.31	0.00	T	1.000
73	1+	Income (8)	59.46	0.08	85.64	85.64	0.00	T	1.000
74	1+	Income (9)	36.86	0.01	93.05	93.05	0.00	T	1.000
75	1+	Income (10)	51.46	0.15	88.65	88.65	0.00	T	1.000
76	2+	Gender	99.99	0.08	51.11	51.11	0.00	W	1.000
77	2+	Not employed	89.01	0.08	69.40	69.40	0.00	T	1.000
78	2+	PT Pass	90.27	0.12	68.00	68.00	0.00	T	1.000
79	2+	Ownership	92.17	0.26	66.49	66.49	0.00	T	1.000
80	2+	Licence	97.03	1.59	60.20	60.20	0.00	T	1.000
81	2+	Income (1)	29.62	0.17	94.86	94.86	0.00	T	1.000
82	2+	Income (2)	40.03	0.15	92.24	92.24	0.00	T	1.000
83	2+	Income (3)	46.59	0.03	89.89	89.89	0.00	T	1.000
84	2+	Income (4)	41.96	0.03	91.35	91.35	0.00	T	1.000
85	2+	Income (5)	42.20	0.07	91.55	91.55	0.00	T	1.000
86	2+	Income (6)	54.37	0.02	87.40	87.40	0.00	T	1.000
87	2+	Income (7)	61.04	0.00	85.30	85.30	0.00	T	1.000
88	2+	Income (8)	59.43	0.04	85.73	85.73	0.00	T	1.000
89	2+	Income (9)	36.69	0.07	93.04	93.04	0.00	T	1.000
90	2+	Income (10)	50.96	0.30	88.62	88.62	0.00	T	1.000

^aThe standard deviation for the difference between the actual MAP-PCP and the expected MAP-PCP obtained across the 20 repetitions in each set of enrichment experiments.

^b‘T’ refers to the paired t-test; ‘W’ represents the Wilcoxon signed-rank test.

Table 5.2 Data Conditions and Results of Comparative Analysis for Actual and Expected MLE-PCPs

Exp NO.	Behaviour Pattern	Attribute	$H(X)$ (%)	η (%)	MLE-PCP (%)		Sd. ^a (%)	Test ^b	p-value
					Actual	Expected			
1	Peak	Gender	99.90	0.00	50.89	50.89	0.00	T	1.000
2	Peak	Not employed	83.79	1.33	52.91	52.91	0.00	T	1.000
3	Peak	PT Pass	88.07	0.55	53.05	53.05	0.00	T	1.000
4	Peak	Ownership	88.34	0.03	55.59	55.59	0.00	T	1.000
5	Peak	Licence	91.11	0.00	55.91	55.91	0.00	T	1.000
6	Peak	Income (1)	25.42	0.28	64.63	64.63	0.00	T	1.000
7	Peak	Income (2)	36.02	0.20	63.44	63.44	0.00	T	1.000
8	Peak	Income (3)	45.02	0.02	62.92	62.92	0.00	T	1.000
9	Peak	Income (4)	40.51	0.03	63.57	63.57	0.00	T	1.000
10	Peak	Income (5)	41.14	0.01	63.64	63.64	0.00	T	0.330
11	Peak	Income (6)	54.28	0.00	62.33	62.33	0.00	T	0.330
12	Peak	Income (7)	60.77	0.04	62.15	62.15	0.00	T	0.163
13	Peak	Income (8)	62.52	0.04	61.92	61.92	0.00	W	1.000
14	Peak	Income (9)	37.55	0.05	64.59	64.59	0.00	T	1.000
15	Peak	Income (10)	54.94	0.01	62.63	62.63	0.00	T	1.000
16	PT	Gender	99.92	0.01	51.34	51.34	0.00	T	1.000
17	PT	Not employed	78.49	0.38	63.20	63.20	0.00	T	0.330
18	PT	PT Pass	88.37	0.02	62.44	62.44	0.00	T	1.000
19	PT	Ownership	92.42	1.91	65.57	65.57	0.00	T	1.000
20	PT	Licence	92.08	0.29	62.08	62.08	0.00	T	1.000
21	PT	Income (1)	24.43	0.02	77.42	77.42	0.00	T	1.000
22	PT	Income (2)	35.73	0.04	75.99	75.99	0.00	T	1.000
23	PT	Income (3)	45.79	0.00	74.02	74.02	0.00	T	1.000
24	PT	Income (4)	41.25	0.03	74.99	74.99	0.00	W	0.317
25	PT	Income (5)	41.08	0.00	74.84	74.84	0.00	T	0.999
26	PT	Income (6)	54.71	0.02	72.44	72.44	0.00	T	1.000
27	PT	Income (7)	59.52	0.03	70.75	70.75	0.00	T	0.577
28	PT	Income (8)	62.80	0.01	70.09	70.09	0.00	T	1.000
29	PT	Income (9)	38.48	0.01	75.07	75.07	0.00	T	1.000
30	PT	Income (10)	55.07	0.01	71.76	71.76	0.00	T	1.000
31	Car	Gender	99.93	0.00	51.40	51.40	0.00	T	0.330
32	Car	Not employed	78.66	0.13	68.41	68.41	0.00	T	1.000
33	Car	PT Pass	88.39	0.06	61.80	61.80	0.00	T	0.999
34	Car	Ownership	92.46	7.64	50.19	50.19	0.00	T	0.330
35	Car	Licence	92.13	1.54	55.63	55.63	0.00	T	0.330
36	Car	Income (1)	24.10	0.53	78.83	78.83	0.00	T	1.000

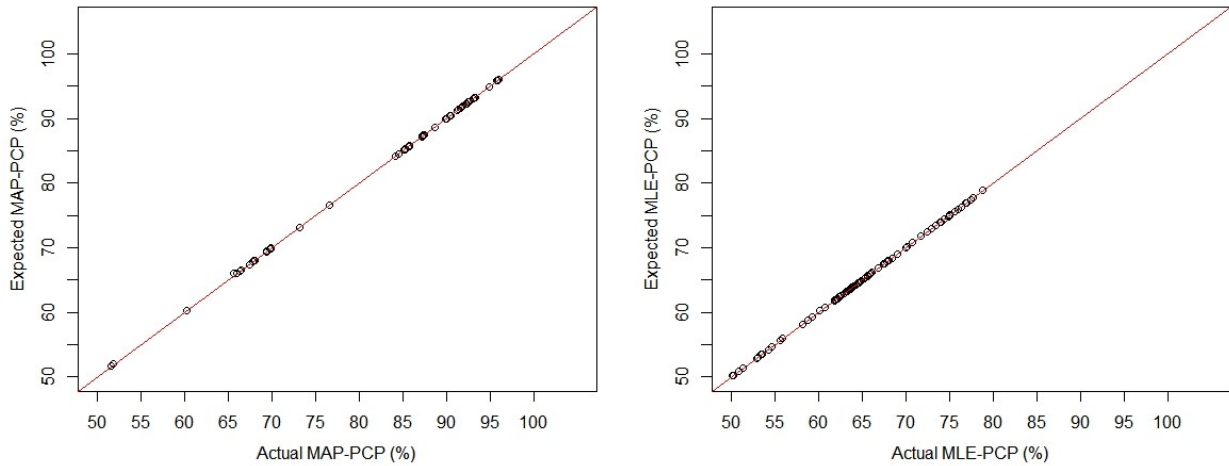
37	Car	Income (2)	35.68	0.45	76.89	76.89	0.00	T	0.163
38	Car	Income (3)	45.60	0.02	75.63	75.63	0.00	T	1.000
39	Car	Income (4)	41.51	0.09	76.31	76.31	0.00	T	0.330
40	Car	Income (5)	41.11	0.00	76.93	76.93	0.00	T	1.000
41	Car	Income (6)	54.92	0.00	73.92	73.92	0.00	T	1.000
42	Car	Income (7)	59.23	0.08	73.46	73.46	0.00	T	0.163
43	Car	Income (8)	62.80	0.12	72.88	72.88	0.00	T	0.330
44	Car	Income (9)	38.45	0.07	77.70	77.70	0.00	T	0.330
45	Car	Income (10)	55.19	0.05	74.40	74.40	0.00	T	1.000
46	0/1	Gender	99.99	0.02	50.28	50.28	0.00	T	0.100
47	0/1	Not employed	88.68	2.67	67.57	67.57	0.00	T	1.000
48	0/1	PT Pass	90.97	0.20	61.88	61.88	0.00	T	0.330
49	0/1	Ownership	92.07	0.18	58.82	58.82	0.00	T	0.163
50	0/1	Licence	96.93	1.01	60.19	60.19	0.00	T	1.000
51	0/1	Income (1)	28.54	0.20	69.99	69.99	0.00	T	1.000
52	0/1	Income (2)	40.26	0.34	69.08	69.08	0.00	T	0.330
53	0/1	Income (3)	47.33	0.03	67.48	67.48	0.00	T	0.100
54	0/1	Income (4)	42.42	0.07	67.98	67.98	0.00	T	0.186
55	0/1	Income (5)	41.77	0.01	67.86	67.86	0.00	T	0.171
56	0/1	Income (6)	54.92	0.00	65.81	65.81	0.00	T	0.793
57	0/1	Income (7)	60.05	0.01	64.87	64.87	0.00	T	0.697
58	0/1	Income (8)	58.55	0.12	64.23	64.23	0.00	T	0.100
59	0/1	Income (9)	36.91	0.05	68.06	68.06	0.00	T	0.330
60	0/1	Income (10)	50.92	0.21	65.43	65.43	0.00	T	0.330
61	1+	Gender	99.99	0.01	50.15	50.15	0.00	W	0.161
62	1+	Not employed	88.77	2.02	65.84	65.84	0.00	T	0.330
63	1+	PT Pass	90.79	0.20	60.73	60.73	0.00	T	1.000
64	1+	Ownership	92.18	0.13	58.12	58.12	0.00	T	0.330
65	1+	Licence	96.72	0.83	59.26	59.26	0.00	T	0.330
66	1+	Income (1)	29.58	0.29	67.80	67.80	0.00	T	1.000
67	1+	Income (2)	39.66	0.14	66.87	66.87	0.00	T	1.000
68	1+	Income (3)	46.54	0.04	65.33	65.33	0.00	T	0.100
69	1+	Income (4)	42.84	0.01	65.88	65.88	0.00	T	0.330
70	1+	Income (5)	41.89	0.01	65.92	65.92	0.00	T	0.100
71	1+	Income (6)	55.17	0.01	63.92	63.92	0.00	T	0.275
72	1+	Income (7)	60.57	0.00	63.23	63.23	0.00	T	0.369
73	1+	Income (8)	59.46	0.08	62.62	62.62	0.00	T	0.100
74	1+	Income (9)	36.86	0.01	66.04	66.04	0.00	T	0.867
75	1+	Income (10)	51.46	0.15	63.65	63.65	0.00	T	0.330
76	2+	Gender	99.99	0.08	51.40	51.40	0.00	T	0.100
77	2+	Not employed	89.01	0.08	54.24	54.24	0.00	T	0.100

78	2+	PT Pass	90.27	0.12	54.64	54.64	0.00	T	0.100
79	2+	Ownership	92.17	0.26	53.56	53.56	0.00	T	0.163
80	2+	Licence	97.03	1.59	53.46	53.46	0.00	T	0.330
81	2+	Income (1)	29.62	0.17	65.66	65.66	0.00	T	0.104
82	2+	Income (2)	40.03	0.15	64.72	64.72	0.00	T	0.100
83	2+	Income (3)	46.59	0.03	64.08	64.08	0.00	T	0.100
84	2+	Income (4)	41.96	0.03	64.56	64.56	0.00	T	0.100
85	2+	Income (5)	42.20	0.07	64.56	64.56	0.00	T	0.100
86	2+	Income (6)	54.37	0.02	63.25	63.25	0.00	T	0.100
87	2+	Income (7)	61.04	0.00	62.95	62.95	0.00	T	0.270
88	2+	Income (8)	59.43	0.04	63.83	63.83	0.00	T	0.100
89	2+	Income (9)	36.69	0.07	66.12	66.12	0.00	T	0.187
90	2+	Income (10)	50.96	0.30	65.28	65.28	0.00	T	0.330

^aThe standard deviation for the difference between the actual MLE-PCP and the expected MLE-PCP obtained across the 20 repetitions in each set of enrichment experiments.

^b‘T’ refers to the paired t-test; ‘W’ represents the Wilcoxon signed-rank test.

It can be found from Table 5.1 and Table 5.2 (above) that the comparison between the actual and expected PCPs using either the MAP or the MLE estimators lie out of the reject region for the null hypothesis, as indicated by the p -values that are greater than 0.1, at the given degrees of freedom ($df = 19$). This reveals that the difference between the actual and expected PCPs using both estimators is negligible for all the 90 sets of enrichment experiments.



(a) Actual MAP-PCP vs. Expected MAP-PCP

(b) Actual MLE-PCP vs. Expected MLE-PCP

Figure 5.3 Plots of Actual PCP vs. Expected PCP using MAP and MLE Estimators

So as to provide a direct visual illustration of the significant similarity, Figure 5.3(a) and Figure 5.3(b) show the plots for the patterns of actual PCPs versus the expected values using, respectively, the MAP and the MLE estimators. As can be seen, both exhibit virtually linear patterns with a 45° slope. The above analysis proves that the proposed IDCM performance

theory is able to forecast the enrichment performance of the IDCM approach and to interpret the variation in the enrichment performance under different data conditions in the binary enrichment context.

5.4 Comparative analysis of the IDCM approach against supervised machine learning methods

In this section, the IDCM framework is compared against the LGR and the SVM in terms of the enrichment performance (Section 5.4.1) and the transferability (Section 5.4.2).

5.4.1 Enrichment performance

The current section presents and discusses the results of the comparative analysis for the enrichment performance of the IDCM approach, respectively using the MAP and the MLE estimators, with the enrichment performance of the LGR and the SVM.

The enrichment performance of the IDCM approach using the MAP estimator is compared against that of the LGR and the SVM based on Table 5.3 and Table 5.4 (below), which respectively show the results of the comparative analysis in all the 90 sets of enrichment experiments, using either the paired *t*-test or the Wilcoxon signed-rank test depending on whether the PCPs are normally distributed. Specifically, the paired *t*-test is used when both compared PCPs are normally distributed across the 20 repetitions in the compared set, otherwise the Wilcoxon signed rank test is adopted.

Table 5.3 Results of Comparative Analysis for MAP-PCP and LGR-PCP

Exp NO.	Behaviour Pattern	Attribute	$H(X)$ (%)	η (%)	PCP (%)		Sd. ^a (%)	Test ^b	<i>p</i> -value
					MAP	LGR			
1	Peak	Gender	99.90	0.00	51.85	51.85	0.00	W	1.000
2	Peak	Not employed	83.79	1.33	73.17	73.17	0.00	T	1.000
3	Peak	PT Pass	88.07	0.55	69.93	69.93	0.00	T	1.000
4	Peak	Ownership	88.34	0.03	69.79	69.79	0.00	T	1.000
5	Peak	Licence	91.11	0.00	67.40	67.40	0.00	T	1.000
6	Peak	Income (1)	25.42	0.28	95.79	95.79	0.00	W	1.000
7	Peak	Income (2)	36.02	0.20	93.15	93.15	0.00	T	1.000
8	Peak	Income (3)	45.02	0.02	90.48	90.48	0.00	T	1.000
9	Peak	Income (4)	40.51	0.03	91.88	91.88	0.00	T	1.000
10	Peak	Income (5)	41.14	0.01	91.75	91.75	0.00	T	1.000

11	Peak	Income (6)	54.28	0.00	87.50	87.50	0.00	T	1.000
12	Peak	Income (7)	60.77	0.04	85.16	85.16	0.00	T	1.000
13	Peak	Income (8)	62.52	0.04	84.49	84.49	0.00	T	1.000
14	Peak	Income (9)	37.55	0.05	92.54	92.54	0.00	T	1.000
15	Peak	Income (10)	54.94	0.01	87.19	87.19	0.00	T	1.000
16	PT	Gender	99.92	0.01	51.64	51.64	0.00	W	1.000
17	PT	Not employed	78.49	0.38	76.57	76.57	0.00	T	1.000
18	PT	PT Pass	88.37	0.02	69.85	69.85	0.00	T	1.000
19	PT	Ownership	92.42	1.91	65.58	65.58	0.00	T	1.000
20	PT	Licence	92.08	0.29	66.36	66.36	0.00	T	1.000
21	PT	Income (1)	24.43	0.02	95.94	95.94	0.00	T	1.000
22	PT	Income (2)	35.73	0.04	93.26	93.26	0.00	T	1.000
23	PT	Income (3)	45.79	0.00	90.39	90.39	0.00	W	1.000
24	PT	Income (4)	41.25	0.03	91.75	91.75	0.00	W	1.000
25	PT	Income (5)	41.08	0.00	91.72	91.72	0.00	T	1.000
26	PT	Income (6)	54.71	0.02	87.39	87.39	0.00	T	1.000
27	PT	Income (7)	59.52	0.03	85.61	85.61	0.00	T	1.000
28	PT	Income (8)	62.80	0.01	84.21	84.21	0.00	T	1.000
29	PT	Income (9)	38.48	0.01	92.52	92.52	0.00	T	1.000
30	PT	Income (10)	55.07	0.01	87.23	87.23	0.00	T	1.000
31	Car	Gender	99.93	0.00	51.55	51.55	0.00	W	1.000
32	Car	Not employed	78.66	0.13	76.54	76.54	0.00	T	1.000
33	Car	PT Pass	88.39	0.06	69.89	69.89	0.00	T	1.000
34	Car	Ownership	92.46	7.64	66.03	66.03	0.00	T	1.000
35	Car	Licence	92.13	1.54	66.41	66.41	0.00	T	1.000
36	Car	Income (1)	24.10	0.53	95.92	95.92	0.00	T	1.000
37	Car	Income (2)	35.68	0.45	93.26	93.26	0.00	T	1.000
38	Car	Income (3)	45.60	0.02	90.38	90.38	0.00	T	1.000
39	Car	Income (4)	41.51	0.09	91.73	91.73	0.00	T	1.000
40	Car	Income (5)	41.11	0.00	91.74	91.74	0.00	T	1.000
41	Car	Income (6)	54.92	0.00	87.36	87.36	0.00	T	1.000
42	Car	Income (7)	59.23	0.08	85.61	85.61	0.00	T	1.000
43	Car	Income (8)	62.80	0.12	84.22	84.22	0.00	T	1.000
44	Car	Income (9)	38.45	0.07	92.51	92.51	0.00	T	1.000
45	Car	Income (10)	55.19	0.05	87.21	87.21	0.00	T	1.000
46	0/1	Gender	99.99	0.02	50.36	50.36	0.65	W	1.000
47	0/1	Not employed	88.68	2.67	69.40	69.40	0.00	T	1.000
48	0/1	PT Pass	90.97	0.20	67.95	67.95	0.00	T	1.000
49	0/1	Ownership	92.07	0.18	66.45	66.45	0.00	T	1.000
50	0/1	Licence	96.93	1.01	60.20	60.20	0.00	T	1.000
51	0/1	Income (1)	28.54	0.20	94.87	94.87	0.00	T	1.000

52	0/1	Income (2)	40.26	0.34	92.19	92.19	0.00	T	1.000
53	0/1	Income (3)	47.33	0.03	89.96	89.96	0.00	T	1.000
54	0/1	Income (4)	42.42	0.07	91.29	91.29	0.00	T	1.000
55	0/1	Income (5)	41.77	0.01	91.55	91.55	0.00	T	1.000
56	0/1	Income (6)	54.92	0.00	87.39	87.39	0.00	T	1.000
57	0/1	Income (7)	60.05	0.01	85.21	85.21	0.00	T	1.000
58	0/1	Income (8)	58.55	0.12	85.69	85.69	0.00	T	1.000
59	0/1	Income (9)	36.91	0.05	93.03	93.03	0.00	T	1.000
60	0/1	Income (10)	50.92	0.21	88.66	88.66	0.00	T	1.000
61	1+	Gender	99.99	0.01	50.44	50.44	0.72	T	1.000
62	1+	Not employed	88.77	2.02	69.43	69.43	0.00	T	1.000
63	1+	PT Pass	90.79	0.20	67.87	67.87	0.00	T	1.000
64	1+	Ownership	92.18	0.13	66.49	66.49	0.00	T	1.000
65	1+	Licence	96.72	0.83	59.87	59.87	0.00	W	1.000
66	1+	Income (1)	29.58	0.29	94.85	94.85	0.00	T	1.000
67	1+	Income (2)	39.66	0.14	92.28	92.28	0.00	T	1.000
68	1+	Income (3)	46.54	0.04	90.00	90.00	0.00	T	1.000
69	1+	Income (4)	42.84	0.01	91.33	91.33	0.00	T	1.000
70	1+	Income (5)	41.89	0.01	91.54	91.54	0.00	T	1.000
71	1+	Income (6)	55.17	0.01	87.35	87.35	0.00	T	1.000
72	1+	Income (7)	60.57	0.00	85.31	85.31	0.00	T	1.000
73	1+	Income (8)	59.46	0.08	85.64	85.64	0.00	T	1.000
74	1+	Income (9)	36.86	0.01	93.05	93.05	0.00	T	1.000
75	1+	Income (10)	51.46	0.15	88.65	88.65	0.00	T	1.000
76	2+	Gender	99.99	0.08	51.11	51.11	0.00	W	1.000
77	2+	Not employed	89.01	0.08	69.40	69.40	0.00	T	1.000
78	2+	PT Pass	90.27	0.12	68.00	68.00	0.00	T	1.000
79	2+	Ownership	92.17	0.26	66.49	66.49	0.00	T	1.000
80	2+	Licence	97.03	1.59	60.20	60.20	0.00	T	1.000
81	2+	Income (1)	29.62	0.17	94.86	94.86	0.00	T	1.000
82	2+	Income (2)	40.03	0.15	92.24	92.24	0.00	T	1.000
83	2+	Income (3)	46.59	0.03	89.89	89.89	0.00	T	1.000
84	2+	Income (4)	41.96	0.03	91.35	91.35	0.00	T	1.000
85	2+	Income (5)	42.20	0.07	91.55	91.55	0.00	T	1.000
86	2+	Income (6)	54.37	0.02	87.40	87.40	0.00	T	1.000
87	2+	Income (7)	61.04	0.00	85.30	85.30	0.00	T	1.000
88	2+	Income (8)	59.43	0.04	85.73	85.73	0.00	T	1.000
89	2+	Income (9)	36.69	0.07	93.04	93.04	0.00	T	1.000
90	2+	Income (10)	50.96	0.30	88.62	88.62	0.00	T	1.000

^aThe standard deviation for the difference between the MAP-PCP and the LGR-PCP obtained across the 20 repetitions in each set of enrichment experiments.

^b‘T’ refers to the paired t-test; ‘W’ represents the Wilcoxon signed-rank test.

Table 5.4 Results of Comparative Analysis for MAP-PCP and SVM-PCP

Exp NO.	Behaviour Pattern	Attribute	$H(X)$ (%)	η (%)	PCP (%)		Sd. ^a (%)	Test ^b	p-value
					MAP	SVM			
1	Peak	Gender	99.90	0.00	51.85	52.07	0.53	W	0.109
2	Peak	Not employed	83.79	1.33	73.17	73.17	0.00	T	1.000
3	Peak	PT Pass	88.07	0.55	69.93	69.93	0.00	T	1.000
4	Peak	Ownership	88.34	0.03	69.79	69.79	0.00	T	1.000
5	Peak	Licence	91.11	0.00	67.40	67.40	0.00	T	1.000
6	Peak	Income (1)	25.42	0.28	95.79	95.79	0.00	W	1.000
7	Peak	Income (2)	36.02	0.20	93.15	93.15	0.00	T	1.000
8	Peak	Income (3)	45.02	0.02	90.48	90.48	0.00	T	1.000
9	Peak	Income (4)	40.51	0.03	91.88	91.88	0.00	T	1.000
10	Peak	Income (5)	41.14	0.01	91.75	91.75	0.00	T	1.000
11	Peak	Income (6)	54.28	0.00	87.50	87.50	0.00	T	1.000
12	Peak	Income (7)	60.77	0.04	85.16	85.16	0.00	T	1.000
13	Peak	Income (8)	62.52	0.04	84.49	84.49	0.00	T	1.000
14	Peak	Income (9)	37.55	0.05	92.54	92.54	0.00	T	1.000
15	Peak	Income (10)	54.94	0.01	87.19	87.19	0.00	T	1.000
16	PT	Gender	99.92	0.01	51.64	51.72	0.18	W	0.109
17	PT	Not employed	78.49	0.38	76.57	76.57	0.00	T	1.000
18	PT	PT Pass	88.37	0.02	69.85	69.85	0.00	T	1.000
19	PT	Ownership	92.42	1.91	65.58	65.98	0.12	T	0.100
20	PT	Licence	92.08	0.29	66.36	66.36	0.00	T	1.000
21	PT	Income (1)	24.43	0.02	95.94	95.94	0.00	T	1.000
22	PT	Income (2)	35.73	0.04	93.26	93.26	0.00	T	1.000
23	PT	Income (3)	45.79	0.00	90.39	90.39	0.00	W	1.000
24	PT	Income (4)	41.25	0.03	91.75	91.75	0.00	W	1.000
25	PT	Income (5)	41.08	0.00	91.72	91.72	0.00	T	1.000
26	PT	Income (6)	54.71	0.02	87.39	87.39	0.00	T	1.000
27	PT	Income (7)	59.52	0.03	85.61	85.61	0.00	T	1.000
28	PT	Income (8)	62.80	0.01	84.21	84.21	0.00	T	1.000
29	PT	Income (9)	38.48	0.01	92.52	92.52	0.00	T	1.000
30	PT	Income (10)	55.07	0.01	87.23	87.23	0.00	T	1.000
31	Car	Gender	99.93	0.00	51.55	51.75	0.20	W	0.100
32	Car	Not employed	78.66	0.13	76.54	76.54	0.00	T	1.000
33	Car	PT Pass	88.39	0.06	69.89	69.89	0.00	T	1.000
34	Car	Ownership	92.46	7.64	66.03	66.03	0.00	T	1.000
35	Car	Licence	92.13	1.54	66.41	66.41	0.00	T	1.000
36	Car	Income (1)	24.10	0.53	95.92	95.92	0.00	T	1.000

37	Car	Income (2)	35.68	0.45	93.26	93.26	0.00	T	1.000
38	Car	Income (3)	45.60	0.02	90.38	90.38	0.00	T	1.000
39	Car	Income (4)	41.51	0.09	91.73	91.73	0.00	T	1.000
40	Car	Income (5)	41.11	0.00	91.74	91.74	0.00	T	1.000
41	Car	Income (6)	54.92	0.00	87.36	87.36	0.00	T	1.000
42	Car	Income (7)	59.23	0.08	85.61	85.61	0.00	T	1.000
43	Car	Income (8)	62.80	0.12	84.22	84.22	0.00	T	1.000
44	Car	Income (9)	38.45	0.07	92.51	92.51	0.00	T	1.000
45	Car	Income (10)	55.19	0.05	87.21	87.21	0.00	T	1.000
46	0/1	Gender	99.99	0.02	50.36	50.36	0.58	W	1.000
47	0/1	Not employed	88.68	2.67	69.40	69.40	0.00	T	1.000
48	0/1	PT Pass	90.97	0.20	67.95	67.95	0.00	T	1.000
49	0/1	Ownership	92.07	0.18	66.45	66.45	0.00	T	1.000
50	0/1	Licence	96.93	1.01	60.20	60.18	0.14	T	0.100
51	0/1	Income (1)	28.54	0.20	94.87	94.87	0.00	T	1.000
52	0/1	Income (2)	40.26	0.34	92.19	92.19	0.00	T	1.000
53	0/1	Income (3)	47.33	0.03	89.96	89.96	0.00	T	1.000
54	0/1	Income (4)	42.42	0.07	91.29	91.29	0.00	T	1.000
55	0/1	Income (5)	41.77	0.01	91.55	91.55	0.00	T	1.000
56	0/1	Income (6)	54.92	0.00	87.39	87.39	0.00	T	1.000
57	0/1	Income (7)	60.05	0.01	85.21	85.21	0.00	T	1.000
58	0/1	Income (8)	58.55	0.12	85.69	85.69	0.00	T	1.000
59	0/1	Income (9)	36.91	0.05	93.03	93.03	0.00	T	1.000
60	0/1	Income (10)	50.92	0.21	88.66	88.66	0.00	T	1.000
61	1+	Gender	99.99	0.01	50.44	50.44	0.85	W	0.100
62	1+	Not employed	88.77	2.02	69.43	69.43	0.00	T	1.000
63	1+	PT Pass	90.79	0.20	67.87	67.87	0.00	T	1.000
64	1+	Ownership	92.18	0.13	66.49	66.49	0.00	T	1.000
65	1+	Licence	96.72	0.83	59.87	60.18	0.55	W	0.100
66	1+	Income (1)	29.58	0.29	94.85	94.85	0.00	T	1.000
67	1+	Income (2)	39.66	0.14	92.28	92.28	0.00	T	1.000
68	1+	Income (3)	46.54	0.04	90.00	90.00	0.00	T	1.000
69	1+	Income (4)	42.84	0.01	91.33	91.33	0.00	T	1.000
70	1+	Income (5)	41.89	0.01	91.54	91.54	0.00	T	1.000
71	1+	Income (6)	55.17	0.01	87.35	87.35	0.00	T	1.000
72	1+	Income (7)	60.57	0.00	85.31	85.31	0.00	T	1.000
73	1+	Income (8)	59.46	0.08	85.64	85.64	0.00	T	1.000
74	1+	Income (9)	36.86	0.01	93.05	93.05	0.00	T	1.000
75	1+	Income (10)	51.46	0.15	88.65	88.65	0.00	T	1.000
76	2+	Gender	99.99	0.08	51.11	51.11	0.83	W	1.000
77	2+	Not employed	89.01	0.08	69.40	69.40	0.00	T	1.000

78	2+	PT Pass	90.27	0.12	68.00	68.00	0.00	T	1.000
79	2+	Ownership	92.17	0.26	66.49	66.49	0.00	T	1.000
80	2+	Licence	97.03	1.59	60.20	60.20	0.00	T	1.000
81	2+	Income (1)	29.62	0.17	94.86	94.86	0.00	T	1.000
82	2+	Income (2)	40.03	0.15	92.24	92.24	0.00	T	1.000
83	2+	Income (3)	46.59	0.03	89.89	89.89	0.00	T	1.000
84	2+	Income (4)	41.96	0.03	91.35	91.35	0.00	T	1.000
85	2+	Income (5)	42.20	0.07	91.55	91.55	0.00	T	1.000
86	2+	Income (6)	54.37	0.02	87.40	87.40	0.00	T	1.000
87	2+	Income (7)	61.04	0.00	85.30	85.30	0.00	T	1.000
88	2+	Income (8)	59.43	0.04	85.73	85.73	0.00	T	1.000
89	2+	Income (9)	36.69	0.07	93.04	93.04	0.00	T	1.000
90	2+	Income (10)	50.96	0.30	88.62	88.62	0.00	T	1.000

^aThe standard deviation for the difference between the MAP-PCP and the SVM-PCP obtained across the 20 repetitions in each set of enrichment experiments.

^b‘T’ refers to the paired t-test; ‘W’ represents the Wilcoxon signed-rank test.

Table 5.3 and Table 5.4 (above) show that the IDCM approach using the MAP estimator performs equivalently to either the LGR or the SVM in all the 90 sets of enrichment experiments (p -values > 0.1). Additionally, it can be seen that the LGR-PCP are identical to the MAP-PCP in all the 90 sets of enrichment experiments, reflected by p -values = 1.000. This is attributed to the use of the prior obtained from the estimation (training) subsample in the MAP estimator. Particularly, the mathematical expressions for, respectively, the MAP estimator (Equation 5-3) and the LGR classifier (Equation 5-7) converge to the same formulation when $P_T(X)$ is identical to the prior, which is the case in the present analysis.

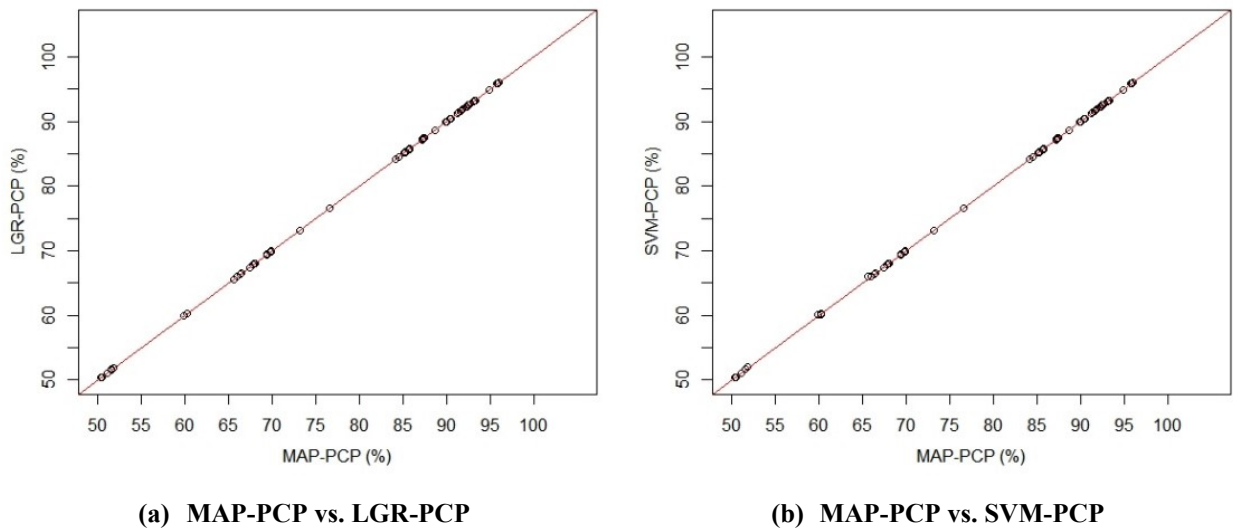


Figure 5.4 Plots of MAP-PCP vs. LGR- and SVM-PCPs

The comparison of the MAP-PCP with the LGR- and the SVM-PCP in the above analysis is visualised in the plots provided in Figure 5.4. The fact that the dots lie substantially around the 45° line is a direct illustration of the comparable enrichment performance of the IDCM MAP estimator and the LGR and SVM in the given enrichment context.

Similarly, the enrichment performance of the IDCM approach using the MLE estimator is also compared against the two supervised ML methods, with the results set out in Table 5.5 and As for the comparison against the SVM, **Error! Not a valid bookmark self-reference.** indicates that the MLE-PCP and the SVM-PCP are significantly different (p -values < 0.1) in 89 sets of enrichment experiments, with the only exception being set No. 50 (in red) in which the difference between the two is negligible (p -values > 0.1).

Table 5.6 (below), respectively.

Table 5.5 Results of Comparative Analysis for MLE-PCP and LGR-PCP

Exp NO.	Behaviour Pattern	Attribute	$H(X)$ (%)	η (%)	PCP (%)		Sd. ^a (%)	Test ^b	p -value
					MLE	LGR			
1	Peak	Gender	99.90	0.00	50.89	51.85	0.42	W	0.000
2	Peak	Not employed	83.79	1.33	52.91	73.17	0.14	T	0.000
3	Peak	PT Pass	88.07	0.55	53.05	69.93	0.14	T	0.000
4	Peak	Ownership	88.34	0.03	55.59	69.79	0.14	T	0.000
5	Peak	Licence	91.11	0.00	55.91	67.40	0.17	T	0.000
6	Peak	Income (1)	25.42	0.28	64.63	95.79	0.17	W	0.000
7	Peak	Income (2)	36.02	0.20	63.44	93.15	0.15	T	0.000
8	Peak	Income (3)	45.02	0.02	62.92	90.48	0.14	T	0.000
9	Peak	Income (4)	40.51	0.03	63.57	91.88	0.16	T	0.000
10	Peak	Income (5)	41.14	0.01	63.64	91.75	0.16	T	0.000
11	Peak	Income (6)	54.28	0.00	62.33	87.50	0.15	T	0.000
12	Peak	Income (7)	60.77	0.04	62.15	85.16	0.18	T	0.000
13	Peak	Income (8)	62.52	0.04	61.92	84.49	0.19	W	0.000
14	Peak	Income (9)	37.55	0.05	64.59	92.54	0.17	T	0.000
15	Peak	Income (10)	54.94	0.01	62.63	87.19	0.12	T	0.000
16	PT	Gender	99.92	0.01	51.34	51.64	0.14	W	0.000
17	PT	Not employed	78.49	0.38	63.20	76.57	0.08	T	0.000
18	PT	PT Pass	88.37	0.02	62.44	69.85	0.10	T	0.000
19	PT	Ownership	92.42	1.91	65.57	65.58	0.04	T	0.330
20	PT	Licence	92.08	0.29	62.08	66.36	0.06	T	0.000
21	PT	Income (1)	24.43	0.02	77.42	95.94	0.07	T	0.000
22	PT	Income (2)	35.73	0.04	75.99	93.26	0.07	T	0.000
23	PT	Income (3)	45.79	0.00	74.02	90.39	0.09	W	0.000

24	PT	Income (4)	41.25	0.03	74.99	91.75	0.08	W	0.000
25	PT	Income (5)	41.08	0.00	74.84	91.72	0.07	T	0.000
26	PT	Income (6)	54.71	0.02	72.44	87.39	0.10	T	0.000
27	PT	Income (7)	59.52	0.03	70.75	85.61	0.09	T	0.000
28	PT	Income (8)	62.80	0.01	70.09	84.21	0.09	T	0.000
29	PT	Income (9)	38.48	0.01	75.07	92.52	0.11	T	0.000
30	PT	Income (10)	55.07	0.01	71.76	87.23	0.08	T	0.000
31	Car	Gender	99.93	0.00	51.40	51.55	0.15	W	0.002
32	Car	Not employed	78.66	0.13	68.41	76.54	0.09	T	0.000
33	Car	PT Pass	88.39	0.06	61.80	69.89	0.07	T	0.000
34	Car	Ownership	92.46	7.64	50.19	66.03	0.08	T	0.000
35	Car	Licence	92.13	1.54	55.63	66.41	0.09	T	0.000
36	Car	Income (1)	24.10	0.53	78.83	95.92	0.06	T	0.000
37	Car	Income (2)	35.68	0.45	76.89	93.26	0.10	T	0.000
38	Car	Income (3)	45.60	0.02	75.63	90.38	0.08	T	0.000
39	Car	Income (4)	41.51	0.09	76.31	91.73	0.06	T	0.000
40	Car	Income (5)	41.11	0.00	76.93	91.74	0.06	T	0.000
41	Car	Income (6)	54.92	0.00	73.92	87.36	0.09	T	0.000
42	Car	Income (7)	59.23	0.08	73.46	85.61	0.07	T	0.000
43	Car	Income (8)	62.80	0.12	72.88	84.22	0.07	T	0.000
44	Car	Income (9)	38.45	0.07	77.70	92.51	0.07	T	0.000
45	Car	Income (10)	55.19	0.05	74.40	87.21	0.07	T	0.000
46	0/1	Gender	99.99	0.02	50.28	50.36	0.51	W	0.126
47	0/1	Not employed	88.68	2.67	67.57	69.40	0.24	T	0.000
48	0/1	PT Pass	90.97	0.20	61.88	67.95	0.22	T	0.000
49	0/1	Ownership	92.07	0.18	58.82	66.45	0.19	T	0.000
50	0/1	Licence	96.93	1.01	60.19	60.20	0.10	T	0.000
51	0/1	Income (1)	28.54	0.20	69.99	94.87	0.18	T	0.000
52	0/1	Income (2)	40.26	0.34	69.08	92.19	0.22	T	0.000
53	0/1	Income (3)	47.33	0.03	67.48	89.96	0.19	T	0.000
54	0/1	Income (4)	42.42	0.07	67.98	91.29	0.23	T	0.000
55	0/1	Income (5)	41.77	0.01	67.86	91.55	0.23	T	0.000
56	0/1	Income (6)	54.92	0.00	65.81	87.39	0.19	T	0.000
57	0/1	Income (7)	60.05	0.01	64.87	85.21	0.20	T	0.000
58	0/1	Income (8)	58.55	0.12	64.23	85.69	0.16	T	0.000
59	0/1	Income (9)	36.91	0.05	68.06	93.03	0.25	T	0.000
60	0/1	Income (10)	50.92	0.21	65.43	88.66	0.20	T	0.000
61	1+	Gender	99.99	0.01	50.15	50.44	0.50	W	0.005
62	1+	Not employed	88.77	2.02	65.84	69.43	0.27	T	0.000
63	1+	PT Pass	90.79	0.20	60.73	67.87	0.25	T	0.000
64	1+	Ownership	92.18	0.13	58.12	66.49	0.29	T	0.000

65	1+	Licence	96.72	0.83	59.26	59.87	0.37	W	0.000
66	1+	Income (1)	29.58	0.29	67.80	94.85	0.19	T	0.000
67	1+	Income (2)	39.66	0.14	66.87	92.28	0.19	T	0.000
68	1+	Income (3)	46.54	0.04	65.33	90.00	0.19	T	0.000
69	1+	Income (4)	42.84	0.01	65.88	91.33	0.21	T	0.000
70	1+	Income (5)	41.89	0.01	65.92	91.54	0.15	T	0.000
71	1+	Income (6)	55.17	0.01	63.92	87.35	0.17	T	0.000
72	1+	Income (7)	60.57	0.00	63.23	85.31	0.20	T	0.000
73	1+	Income (8)	59.46	0.08	62.62	85.64	0.16	T	0.000
74	1+	Income (9)	36.86	0.01	66.04	93.05	0.18	T	0.000
75	1+	Income (10)	51.46	0.15	63.65	88.65	0.19	T	0.000
76	2+	Gender	99.99	0.08	51.40	51.11	0.59	W	0.156
77	2+	Not employed	89.01	0.08	54.24	69.40	0.25	T	0.000
78	2+	PT Pass	90.27	0.12	54.64	68.00	0.31	T	0.000
79	2+	Ownership	92.17	0.26	53.56	66.49	0.16	T	0.000
80	2+	Licence	97.03	1.59	53.46	60.20	0.47	T	0.000
81	2+	Income (1)	29.62	0.17	65.66	94.86	0.26	T	0.000
82	2+	Income (2)	40.03	0.15	64.72	92.24	0.18	T	0.000
83	2+	Income (3)	46.59	0.03	64.08	89.89	0.19	T	0.000
84	2+	Income (4)	41.96	0.03	64.56	91.35	0.22	T	0.000
85	2+	Income (5)	42.20	0.07	64.56	91.55	0.17	T	0.000
86	2+	Income (6)	54.37	0.02	63.25	87.40	0.25	T	0.000
87	2+	Income (7)	61.04	0.00	62.95	85.30	0.18	T	0.000
88	2+	Income (8)	59.43	0.04	63.83	85.73	0.20	T	0.000
89	2+	Income (9)	36.69	0.07	66.12	93.04	0.25	T	0.000
90	2+	Income (10)	50.96	0.30	65.28	88.62	0.14	T	0.000

^aThe standard deviation for the difference between the MLE-PCP and the LGR-PCP obtained across the 20 repetitions in each set of enrichment experiments.

^b‘T’ refers to the paired t-test; ‘W’ represents the Wilcoxon signed-rank test.

It can be observed from Table 5.5 that the MLE-PCP significantly differs from the LGR-PCP in 87 out of the 90 sets of enrichment experiments, suggested by the p -values that are substantially smaller than 0.1. A further investigation into the PCP values in these 87 sets of experiments suggests that the LGR-PCP surpasses the MLE-PCP in all 87 sets of experiments. Apart from these 87 sets of experiments, the difference between the MLE-PCP and the LGR-PCP in the remaining three sets of enrichment experiments (No. 19, 46 and 76) is found to be insignificant (p -values > 0.1 in red).

As LGR-PCPs are proved above to be identical to corresponding MAP-PCPs in the current context, the comparison between the MLE- and the LGR-PCPs are equivalent the comparison

between the MLE- and the MAP-PCP. In particular, the trivial difference between the MLE- and the LGR-PCPs in sets No. 19, 46, and 76 can be attributed to either the relatively high enrichment efficiency η (No. 19) or the significantly high attribute entropy $H(X)$ (No. 46 and No. 76) as marked in red. Recall Section 4.2.3.5, which indicates that the MLE-PCP varies as a result of the η in the two divisions representing different attribute entropy levels. This is specifically demonstrated by the orange and blue curves in Figure 4.8 (i.e. as reproduced in Figure 5.5), suggesting that the MLE-PCP can perform better with an increasing η , or for the same η level, the relatively lower the level of $H(X)$, which is revealed by the blue curve compared against the orange curve.

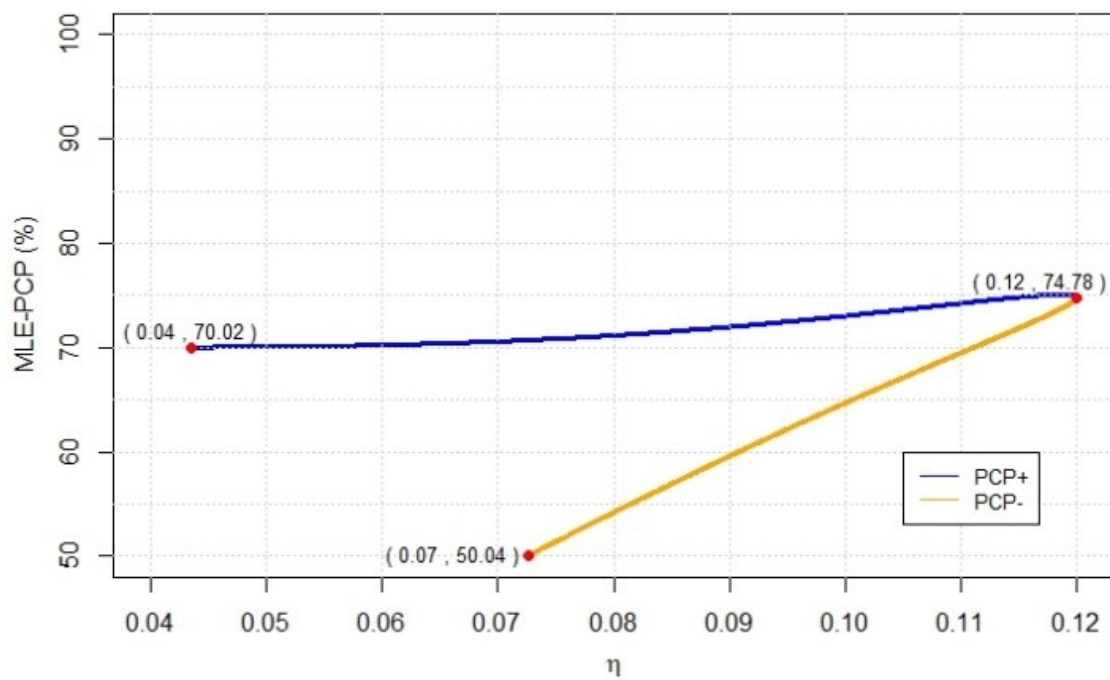


Figure 5.5 Variation Pattern of MLE-PCP vs. η when $p_y = 0.3$ and $\beta_1 = 1.792$ ($\tau = 6$)

As for the comparison against the SVM, **Error! Not a valid bookmark self-reference.** indicates that the MLE-PCP and the SVM-PCP are significantly different (p -values < 0.1) in 89 sets of enrichment experiments, with the only exception being set No. 50 (in red) in which the difference between the two is negligible (p -values > 0.1).

Table 5.6 Results of Comparative Analysis for MLE-PCP and SVM-PCP

Exp NO.	Behaviour Pattern	Attribute	$H(X)$ (%)	η (%)	PCP (%)		Sd. ^a (%)	Test ^b	p -value
					MLE	SVM			
1	Peak	Gender	99.90	0.00	50.89	52.07	0.18	W	0.000
2	Peak	Not employed	83.79	1.33	52.91	73.17	0.14	T	0.000
3	Peak	PT Pass	88.07	0.55	53.05	69.93	0.14	T	0.000

4	Peak	Ownership	88.34	0.03	55.59	69.79	0.14	T	0.000
5	Peak	Licence	91.11	0.00	55.91	67.40	0.17	T	0.000
6	Peak	Income (1)	25.42	0.28	64.63	95.79	0.17	W	0.000
7	Peak	Income (2)	36.02	0.20	63.44	93.15	0.15	T	0.000
8	Peak	Income (3)	45.02	0.02	62.92	90.48	0.14	T	0.000
9	Peak	Income (4)	40.51	0.03	63.57	91.88	0.16	T	0.000
10	Peak	Income (5)	41.14	0.01	63.64	91.75	0.16	T	0.000
11	Peak	Income (6)	54.28	0.00	62.33	87.50	0.15	T	0.000
12	Peak	Income (7)	60.77	0.04	62.15	85.16	0.18	T	0.000
13	Peak	Income (8)	62.52	0.04	61.92	84.49	0.19	W	0.000
14	Peak	Income (9)	37.55	0.05	64.59	92.54	0.17	T	0.000
15	Peak	Income (10)	54.94	0.01	62.63	87.19	0.12	T	0.000
16	PT	Gender	99.92	0.01	51.34	51.72	0.08	W	0.000
17	PT	Not employed	78.49	0.38	63.20	76.57	0.08	T	0.000
18	PT	PT Pass	88.37	0.02	62.44	69.85	0.10	T	0.000
19	PT	Ownership	92.42	1.91	65.57	65.98	0.09	T	0.000
20	PT	Licence	92.08	0.29	62.08	66.36	0.06	T	0.000
21	PT	Income (1)	24.43	0.02	77.42	95.94	0.07	T	0.000
22	PT	Income (2)	35.73	0.04	75.99	93.26	0.07	T	0.000
23	PT	Income (3)	45.79	0.00	74.02	90.39	0.09	W	0.000
24	PT	Income (4)	41.25	0.03	74.99	91.75	0.08	W	0.000
25	PT	Income (5)	41.08	0.00	74.84	91.72	0.07	T	0.000
26	PT	Income (6)	54.71	0.02	72.44	87.39	0.10	T	0.000
27	PT	Income (7)	59.52	0.03	70.75	85.61	0.09	T	0.000
28	PT	Income (8)	62.80	0.01	70.09	84.21	0.09	T	0.000
29	PT	Income (9)	38.48	0.01	75.07	92.52	0.11	T	0.000
30	PT	Income (10)	55.07	0.01	71.76	87.23	0.08	T	0.000
31	Car	Gender	99.93	0.00	51.40	51.75	0.06	W	0.000
32	Car	Not employed	78.66	0.13	68.41	76.54	0.09	T	0.000
33	Car	PT Pass	88.39	0.06	61.80	69.89	0.07	T	0.000
34	Car	Ownership	92.46	7.64	50.19	66.03	0.08	T	0.000
35	Car	Licence	92.13	1.54	55.63	66.41	0.09	T	0.000
36	Car	Income (1)	24.10	0.53	78.83	95.92	0.06	T	0.000
37	Car	Income (2)	35.68	0.45	76.89	93.26	0.10	T	0.000
38	Car	Income (3)	45.60	0.02	75.63	90.38	0.08	T	0.000
39	Car	Income (4)	41.51	0.09	76.31	91.73	0.06	T	0.000
40	Car	Income (5)	41.11	0.00	76.93	91.74	0.06	T	0.000
41	Car	Income (6)	54.92	0.00	73.92	87.36	0.09	T	0.000
42	Car	Income (7)	59.23	0.08	73.46	85.61	0.07	T	0.000
43	Car	Income (8)	62.80	0.12	72.88	84.22	0.07	T	0.000
44	Car	Income (9)	38.45	0.07	77.70	92.51	0.07	T	0.000

45	Car	Income (10)	55.19	0.05	74.40	87.21	0.07	T	0.000
46	0/1	Gender	99.99	0.02	50.28	49.43	0.51	W	0.000
47	0/1	Not employed	88.68	2.67	67.57	69.40	0.24	T	0.000
48	0/1	PT Pass	90.97	0.20	61.88	67.95	0.22	T	0.000
49	0/1	Ownership	92.07	0.18	58.82	66.45	0.19	T	0.000
50	0/1	Licence	96.93	1.01	60.19	60.18	0.21	T	0.741
51	0/1	Income (1)	28.54	0.20	69.99	94.87	0.18	T	0.000
52	0/1	Income (2)	40.26	0.34	69.08	92.19	0.22	T	0.000
53	0/1	Income (3)	47.33	0.03	67.48	89.96	0.19	T	0.000
54	0/1	Income (4)	42.42	0.07	67.98	91.29	0.23	T	0.000
55	0/1	Income (5)	41.77	0.01	67.86	91.55	0.23	T	0.000
56	0/1	Income (6)	54.92	0.00	65.81	87.39	0.19	T	0.000
57	0/1	Income (7)	60.05	0.01	64.87	85.21	0.20	T	0.000
58	0/1	Income (8)	58.55	0.12	64.23	85.69	0.16	T	0.000
59	0/1	Income (9)	36.91	0.05	68.06	93.03	0.25	T	0.000
60	0/1	Income (10)	50.92	0.21	65.43	88.66	0.20	T	0.000
61	1+	Gender	99.99	0.01	50.15	49.80	0.67	W	0.025
62	1+	Not employed	88.77	2.02	65.84	69.43	0.27	T	0.000
63	1+	PT Pass	90.79	0.20	60.73	67.87	0.25	T	0.000
64	1+	Ownership	92.18	0.13	58.12	66.49	0.29	T	0.000
65	1+	Licence	96.72	0.83	59.26	60.18	0.24	W	0.000
66	1+	Income (1)	29.58	0.29	67.80	94.85	0.19	T	0.000
67	1+	Income (2)	39.66	0.14	66.87	92.28	0.19	T	0.000
68	1+	Income (3)	46.54	0.04	65.33	90.00	0.19	T	0.000
69	1+	Income (4)	42.84	0.01	65.88	91.33	0.21	T	0.000
70	1+	Income (5)	41.89	0.01	65.92	91.54	0.15	T	0.000
71	1+	Income (6)	55.17	0.01	63.92	87.35	0.17	T	0.000
72	1+	Income (7)	60.57	0.00	63.23	85.31	0.20	T	0.000
73	1+	Income (8)	59.46	0.08	62.62	85.64	0.16	T	0.000
74	1+	Income (9)	36.86	0.01	66.04	93.05	0.18	T	0.000
75	1+	Income (10)	51.46	0.15	63.65	88.65	0.19	T	0.000
76	2+	Gender	99.99	0.08	51.40	49.93	0.60	W	0.000
77	2+	Not employed	89.01	0.08	54.24	69.40	0.25	T	0.000
78	2+	PT Pass	90.27	0.12	54.64	68.00	0.31	T	0.000
79	2+	Ownership	92.17	0.26	53.56	66.49	0.16	T	0.000
80	2+	Licence	97.03	1.59	53.46	60.20	0.47	T	0.000
81	2+	Income (1)	29.62	0.17	65.66	94.86	0.26	T	0.000
82	2+	Income (2)	40.03	0.15	64.72	92.24	0.18	T	0.000
83	2+	Income (3)	46.59	0.03	64.08	89.89	0.19	T	0.000
84	2+	Income (4)	41.96	0.03	64.56	91.35	0.22	T	0.000
85	2+	Income (5)	42.20	0.07	64.56	91.55	0.17	T	0.000

86	2+	Income (6)	54.37	0.02	63.25	87.40	0.25	T	0.000
87	2+	Income (7)	61.04	0.00	62.95	85.30	0.18	T	0.000
88	2+	Income (8)	59.43	0.04	63.83	85.73	0.20	T	0.000
89	2+	Income (9)	36.69	0.07	66.12	93.04	0.25	T	0.000
90	2+	Income (10)	50.96	0.30	65.28	88.62	0.14	T	0.000

^aThe standard deviation for the difference between the MLE-PCP and the SVM-PCP obtained across the 20 repetitions in each sets of enrichment experiments.

^b‘T’ refers to the paired t-test; ‘W’ represents the Wilcoxon signed-rank test.

The results in the 89 sets of enrichment experiments where a significant difference is revealed show that the SVM outperforms the MLE estimator in most experiments except sets No. 46, 61 and 76, where the MLE estimator performs slightly better. The relatively worse performance of the SVM in sets No. 46, 61 and 76, as well as the indifferent SVM-PCP in set No. 50 can probably be attributed to the fact that the specified SVM parameters (viz. c and γ) trained on the estimation sample cannot maximise the SVM enrichment performance in the corresponding enrichment samples, indicating that the enrichment performance of the SVM substantially depends on the choice and training of the two parameters.

The patterns described above are plotted in Figure 5.6 to provide a straightforward visualisation for the relatively better or worse enrichment performance using the MLE estimator as compared against the LGR and SVM, respectively. Specifically, the general better enrichment performance of both the LGR and SVM is reflected by that most dots in Figure 5.6 lie above the 45° line. However, some of the dots in the bottom left part of Figure 5.6(b) are found to lie below the 45° line, complying with the above findings that the SVM performs slightly worse than the MLE estimator in the corresponding enrichment experiments (i.e. No. 46, 61 and 76).

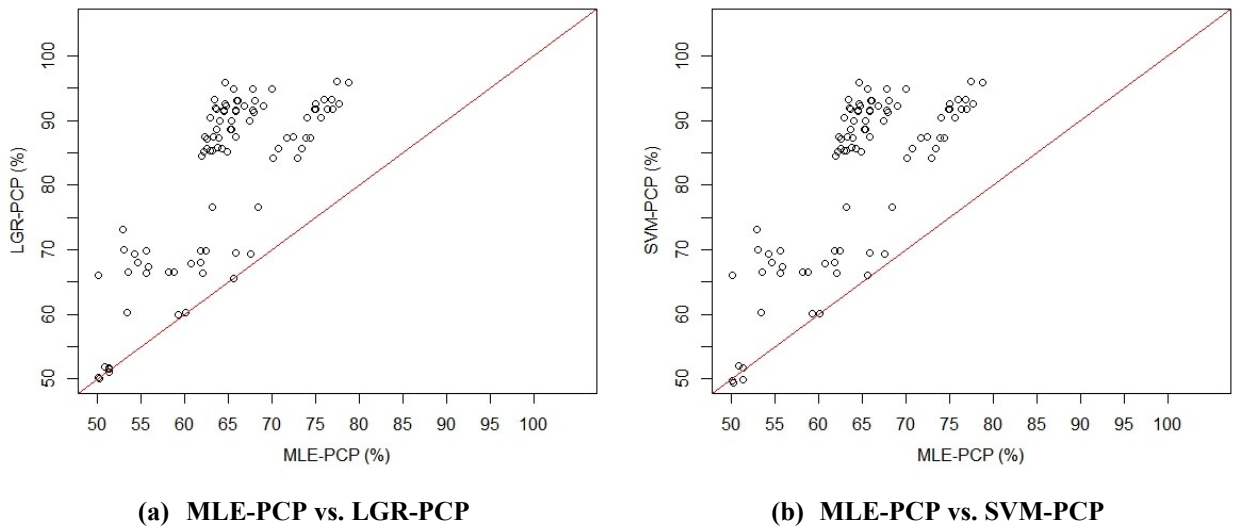


Figure 5.6 Plots of MLE-PCP vs. LGR- and SVM-PCPs

It can be derived from the above analysis and Figure 5.6 that LGR generally performs no worse than the IDCM approach using the MLE estimator, but that the SVM is highly likely to perform better than the IDCM with the MLE estimator. The relatively worse performance of the MLE estimator as compared to the two supervised ML methods can be linked to the way in which the estimation (training) subsamples and the enrichment subsamples are generated. To be specific, as is introduced in Section 5.2.2, the estimation and enrichment subsamples are randomly split in a 20/80 proportion from a single LTDS dataset. The respective variables (choices and enriched attributes) in the two subsamples are hence analogously distributed, by the property of the construction of random subsamples given their same sources in the LTDS data. Although the training process is conducted solely upon the training subsample, they inevitably leverage the information concerning the probability distribution of the enriched attribute in the enrichment subsample as exposed in the distribution of the training subsample. Such information, in the sense of the enrichment using the IDCM approach, is the prior knowledge to the probability distribution of the enriched attribute in the enrichment sample. Conversely, the IDCM approach using the MLE estimator intentionally excludes the prior from the estimation subsample by assuming a uniform distribution for the enriched attribute, and thus possibly performs less effectively as compared with the other approaches that make ‘better’ use of the available data by incorporating the informative prior. This interpretation is supported by observing that the relatively small discrepancy between the PCPs using the MLE estimator and LGR (or the SVM) occurs in experiments where the distribution of the attributes is close to uniform, e.g. gender enrichment in experiments (No. 1, 16, 31, 46, 61 and 76).

5.4.2 Transferability

The aforementioned issue regarding the similar covariance structures of the estimation and the enrichment subsamples also gives rise to the question concerning the transferability of the discussed approaches, i.e. their applicability for the enrichment of samples that retain different distributions as compared to the distribution in the estimation/training sample.

Transferability is a crucial measure when evaluating an enrichment approach because it determines the extent to which the methodology can be applied across various application contexts. Unlike the currently employed estimation and enrichment subsamples that have similar distributions, in many real-world applications, the marginal distribution of either the input behaviour pattern Y , or the output enriched socio-demographic attribute X , can vary

between the ‘thick’ estimation (training) sample and the ‘thin’ dataset which requires enrichment. This can be due to various factors. The contents below explain this issue using two examples corresponding, respectively, to the difference in $P(Y)$ and $P(X)$, and also discuss their impacts on the transferability of the LGR classifier as compared to the IDCM approach.

Example 1: Transferability of LGR and IDCM for the changes in behaviour

Consider a task to enrich the gender attribute (X) given mode choice observations (Y) in an anonymous dataset collected after implementing an increased bus charge (denoted as Z) yet where the only ‘thick’ data of the study area available for estimating the IDCM kernel (or training the LGR classifier) were collected before implementing the policy. The assumption made here is that the increased bus charge would reduce the market share of bus choice while the gender distribution remains stable before and after the policy implementation (which is clearly questionable (Hersch, 1996) yet is for demonstration purpose only). X is assumed to be independent of Z , i.e. $P(X, Z) = P(X)P(Z)$.

In light of the thick ‘before’ sample, the trained LGR classifier can be expressed according to its definition (Equation 5-7) as:

$$X_{LGR} = \arg \max_X P_{Bef}(Y|X)P_{Bef}(X). \quad (5-9)$$

Based on the independence assumed between X and Z , the actual condition probability that describes the thin ‘after’ sample, i.e. $P_{Aft}(X|Y, Z)$, can be manipulated as:

$$\begin{aligned} P_{Aft}(X|Y, Z) &= \frac{P_{Aft}(X, Y, Z)}{P_{Aft}(Y, Z)} = \frac{P_{Aft}(X, Y|Z)}{P_{Aft}(Y|Z)} = \frac{P_{Aft}(Y|X, Z)P_{Aft}(X|Z)}{P_{Aft}(Y|Z)} \\ &= \frac{P_{Aft}(Y|X, Z)P_{Aft}(X)}{P_{Aft}(Y|Z)}. \end{aligned} \quad (5-10)$$

This gives the expected (true) LGR classifier for the ‘after’ sample as:

$$X_{LGR-E} = \arg \max_X P_{Aft}(X|Y, Z) = \arg \max_X P_{Aft}(Y|X, Z) P_{Aft}(X). \quad (5-11)$$

The LGR classifier in Equation 5-11 is termed as an ‘expected LGR classifier’ which cannot be actually trained because the true distribution of the ‘after’ sample is not available. The transferability of LGR in this case hence effectively involves the applicability of the LGR classifier trained using the ‘before’ sample in performing the defined enrichment task for the

‘after’ sample, which can be explored by comparing the difference between Equations 5-9 and 5-11. Specifically, the gender distributions in both samples are assumed to be approximately equal, i.e. $P_{Bef}(X) \approx P_{Aft}(X)$. However, $P_{Bef}(Y|X)$ and $P_{Aft}(Y|X, Z)$ which represent people’s mode choice preferences, respectively before and after implementing the policy, are different according to the assumption that the additional charge would make bus travel less attractive. This can possibly result in the worse performance of Equation 5-9 in the enrichment of gender for the ‘after’ sample, which is neither recognised nor quantifiable by the LGR classifier itself. Additionally, LGR cannot explain such a reduction in performance, making its enrichment performance uninterpretable and restricting its value to the applications that require transferability across various data conditions.

Analogously, the IDCM approach using the MLE and the MAP estimators in this example can be expressed, respectively, as:

$$X_{MLE} = \arg \max_X P_{Bef}(Y|X). \quad (5-12)$$

$$X_{MAP} = \arg \max_X P_{Bef}(Y|X)P_I(X). \quad (5-13)$$

where $P_I(X)$ refers to the prior information of the distribution of X in the enrichment sample that can be available to the analysts;

In this example, the prior can be obtained from the estimation sample given $P_{Bef}(X) \approx P_{Aft}(X)$. $P_I(X)$ can thus be approximated using $P_{Bef}(X)$ since no other informative prior is available or needed. The properties below are found from the equations regarding the two enrichment methods above (LGR and the IDCM):

- 1) The MAP estimator (Equation 5-13) can perform comparably to the LGR (Equation 5-9) as they take the same form when $P_{Bef}(X) \approx P_{Aft}(X)$.
- 2) The MLE estimator (Equation 5-12) performs no better than the LGR and the MAP estimator when the prior information is excluded purposely by the MLE itself.
- 3) The MLE and the MAP estimators, as well as the LGR classifier, can perform equivalently when the prior is not informative enough, e.g. when $P_{Bef}(X)$ is uniformly distributed across the sample.
- 4) The market share information of the mode choices, i.e. $P(Y|Z)$, in the ‘after’ sample can be employed to perform constant recalibration, as explained in Section 4.2.3. The

recalibrated constant $\hat{\beta}_0$, which effectively captures the impact of unobserved factors including Z , can then serve as the pivot to link the PCPs with η using the IDCM performance theory. This enables the prediction for the PCPs using the IDCM approach before performing the enrichment of the socio-demographic attribute for the ‘after’ sample. Moreover, it also permits the interpretation of any variation in the enrichment performance due to the changes in the joint probability distribution underlying the enrichment sample.

Example 2: Transferability of LGR and IDCM for the changes in attribute distribution

Consider another case in which the employment rate of a given area increases substantially due to some economy-stimulating policy (Z) while the market share of mode choices remains similar (assumed for explanation only). The task involves enriching the work status (whether employed) (X) from mode choices (Y) of an anonymous dataset. Again the only ‘thick’ data available in the studied area were collected before implementing the policy.

The LGR classifiers trained upon the ‘thick’ data (‘before’ sample) and expected to capture the new distribution features for the thin data (‘after’ sample) maintain the same expressions as demonstrated in Equations 5-9 and 5-11. The enrichment performance of the LGR classifier in Equation 5-9 in this enrichment task is also found to decline when comparing the two equations. Unlike in Example 1, however, the present decline is explained by the different distributions of X , i.e. $P_B(X) \neq P_A(X)$. Yet, the LGR classifier *per se* provides no prediction for or explanation of such decline. In contrast, the IDCM approach using the MLE and the MAP estimators for this task also take the same form presented in Equations 5-12 and 5-13. In terms of their usability, the MLE estimator is applicable when no informative prior of X is available, and the corresponding PCP level can be estimated using the IDCM performance theory. In cases where any informative prior $P_I(X)$ is known concerning the distribution of X in the ‘after’ sample, e.g. government statistics, it can be utilised to improve the enrichment, which is also tractable given the IDCM performance theory.

The above two examples explain how the transferability of LGR can be restricted due to the changes in $P(Y)$ and $P(X)$, respectively. In some empirical applications, the two probability distributions can vary simultaneously, thus limiting the transferability of LGR under different data conditions or across application contexts.

As for the representative of nonparametric supervised ML methods, the SVM retains the flexibility that is the general advantage of nonparametric algorithms, yet the analysis with respect to its transferability cannot be performed based on the functional form of the algorithm, thus requiring a different perspective. As is explained in Section 5.2.3, the implementation of a nonlinear SVM requires determining the parameters c and γ to trade-off between bias and variance. The enrichment performance of the SVM can vary depending on different parameters, even if the results in the current application are not affected by their values, reflected by the same enrichment performance of the SVM using different combinations of c and γ . Determining fitted values of these parameters relies substantially on the training and testing upon ‘thick’ samples. Extensive applications haven shown that SVM algorithms trained on different samples can result in different optimised parameters (Pedregosa *et al.*, 2011). Hence, parameters of an SVM trained effectively from one sample can possibly act less productively in the enrichment for another sample.

Similarly to the LGR classifier discussed above, the SVM itself is unable to interpret the deterioration in its enrichment performance given its nonparametric nature. This inability, as well as the need to find the fittest SVM parameters, inevitably hampers the transferability of the SVM in contexts where it might be necessary to evaluate scenarios that may involve changes in the underlying covariance structures between the variables, e.g. policy contexts affecting variables of interest.

In contrast, the IDCM framework developed in Chapter 4 of this thesis presents a way to enrich the socio-demographic attributes while retaining the way in which the mutual dependence between the enriched attributes and the corresponding behaviour patterns are held. As a result of this, the IDCM framework provides a means to estimate and interpret the IDCM enrichment performance using the established performance theory. This IDCM performance theory makes the possible deterioration in the enrichment performance due to the variation in the covariance structure underlying the enrichment sample predictable, in turn enabling an understanding of whether the variation of the enrichment performance is acceptable, and supporting the transferability of the IDCM approach between data samples or application contexts.

5.5 Summary

In summary, this chapter presented an empirical application using the London Travel Demand Survey (LTDS) data to demonstrate the socio-demographic enrichment using the developed

Inverse Discrete Choice Modelling (IDCM) framework. In this application various binary socio-demographic attributes modified based on the available socio-demographic information in the LTDS were enriched from several adapted behaviour patterns in the data. The IDCM framework was validated from two perspectives: (i) the capability of the IDCM performance theory to estimate (predict) the IDCM enrichment performance; and (ii) a comparative analysis against the methods used in previous socio-demographic enrichment attempts.

Specifically, the IDCM performance theory is found to be able to predict and interpret the enrichment performance of the IDCM approach using both the MLE and the MAP estimators under various data conditions, reflected by the enrichment of various socio-demographic attributes given different behaviour patterns in the specified binary enrichment context.

In terms of the comparative analysis, the IDCM framework is compared against logistic regression (LGR) and support vector machines (SVMs), as representatives of existing methods employed for previous socio-demographic enrichment attempts from two senses, respectively regarding the enrichment performance and the transferability. The results show that LGR and the SVM perform generally no worse than the IDCM approach using the MLE estimator, while the IDCM approach using the MAP estimator perform comparably to the two supervised machine learning (ML) methods in the given binary enrichment context.

LGR and the SVM, however, are found to be limited in terms of their transferability, which is revealed by the fact that neither method is able to forecast or interpret the variation in their enrichment performance as the data conditions vary. For LGR, this can be attributed to the difference in the probability distribution of variables between estimation and imputation samples. As for the SVM, the main factor restricting its transferability involves the variation in the model parameters, the setting of which relies extensively upon the ability to assess the model goodness-of-fit post-training. The above analysis underlines the necessity of having theoretical underpinnings for enrichment approaches so as to be able to interpret their enrichment performance under different data conditions and thus sustain their transferability across contexts. Unlike the two supervised ML methods, the verified capability of the IDCM socio-demographic enrichment framework to estimate the enrichment performance of the IDCM approach permits its applicability under various data conditions or across application contexts. Moreover, the availability of the two estimation methods using the MLE and the MAP estimators allows the IDCM approach to be used for socio-demographic enrichment given different data availabilities. These two estimators, in conjunction with the IDCM performance

theory, ensure the widespread applicability and the transferability of the IDCM framework under various data conditions.

Owing to the firm grounding in microeconomic theories, the transferability of the IDCM socio-demographic enrichment framework warrants further exploration and development of the framework for various extensions such as the enrichment of a single socio-demographic attribute from multiple choices. A limitation of the empirical validation presented in this chapter is the similar probability distributions in the estimation and the enrichment samples resulting from the sole use of the LTDS data. In the following chapter, a second empirical application is provided to explore: (i) the variation in the enrichment performance of the IDCM approach in the presence of data collected from different sources (thus possibly having different probability distributions); and (ii) the capability of the IDCM performance theory to estimate the corresponding IDCM enrichment performance.

6 EMPIRICAL APPLICATION: ENRICHMENT OF MOBILE NETWORK DATA FOR LONDON USING LONDON TRAVEL DEMAND SURVEY

Chapter 5 has validated the capability of the Inverse Discrete Choice Modelling (IDCM) socio-demographic enrichment framework in an empirical context using the London Travel Demand Survey (LTDS) data. The results in Section 5.4, however, indicate that socio-demographic enrichment based on estimation and enrichment samples derived from the same data source may limit the insight into transferability of the approach due to the similar data distributions in the two subsamples. Moving beyond the results from Chapter 5, the present chapter aims to extend the investigation with respect to the transferability of the proposed framework by implementing the IDCM approach in the enrichment of the socio-demographic attributes of anonymised mobile network data (MND) collected in London, this time using the IDCM kernels estimated from data collected from a different source, viz. the LTDS. Specifically, the aim of enriching the MND using the LTDS is achieved with the following objectives:

- 1) To understand the pattern of variation in the enrichment performance of the IDCM approach using the maximum likelihood (MLE) and the maximum *a posteriori* (MAP) estimators, as well as to identify the more preferred estimation method, under different data conditions by exploring how the MLE- and the MAP-PCPs vary given different qualities of priors or for enriching samples collected at different levels of geographical detail.
- 2) To evaluate the capability of the IDCM performance theory to estimate the enrichment performance of the IDCM approach using the MLE and the MAP estimators given the various specified data conditions as described in 1).
- 3) To compare the IDCM framework against existing socio-demographic enrichment methods in terms of enrichment performance and the ability to forecast the enrichment performance,

as well as to interpret possible variations in the respective enrichment performance under different data conditions or across application contexts.

This empirical application is demonstrated in the following way. Section 6.1 introduces the data used in the present analysis in terms of their sources, the variables selected or extracted for the enrichment, as well as how they are modified to capture the change in the data conditions described above in objective 1). Based on Section 6.1, Section 6.2 demonstrates the design for the experiments conducted in the current application. Section 6.3 then presents the results as well as the corresponding analysis, which shows how the above research objectives are achieved. Section 6.4 concludes this chapter.

6.1 Data

As is mentioned in Section 4.1.2, a typical IDCM socio-demographic enrichment process requires two types of data samples:

- 1) an *enrichment sample* where the socio-demographic information is absent and thus requires enrichment; and
- 2) an *estimation sample* upon which the IDCM kernel(s) can be estimated for its subsequent use in the enrichment using the IDCM approach.

In most real-world applications, the probability distributions (i.e. covariance structures) of the estimation and the enrichment samples are different. This may be attributed to various external factors that can change the composition of the socio-demographic attributes of the population or their behaviour patterns, such as the ones that have been discussed in Section 5.4.2. Another factor that may cause a difference in the distribution between the estimation and the enrichment samples is the different sources from which the two samples are collected. Such differences may lead to the variation in the enrichment performance. The transferability of an enrichment approach can be limited without an awareness or interpretability to such variations.

Measuring the extent to which an enrichment approach can be applied under different data conditions (distributions) or across different application contexts, transferability is a crucial metric for evaluating the applicability of an enrichment approach. In the present application, therefore, the enrichment performance and the performance estimation capability of the developed IDCM framework are explored under various data conditions using two data samples collected from different data sources, i.e. a mobile network provider and the LTDS.

Although different, these two sources still cover the same geographical area during a relatively close time period, thus avoiding, at least for the current study, issues associated with either spatial or temporal transferability.

This section describes the data employed in this empirical application. Specifically, Section 6.1.1 explains the geographies, defining the levels of geographical (spatial) detail at which the data used in this study are collected and how they can be linked to each other. Section 6.1.2 outlines the two data sources from which the estimation and the enrichment samples are derived. This is followed by Section 6.1.3 which illustrates the variables that are extracted to serve as either the input or the output in the current enrichment application, as well as the reasons for selecting these variables and how they are modified to fit the current scope.

6.1.1 Geographies and spatial (geographical) aggregation

This section explores the spatial aggregation of the enrichment samples used in the present analysis. This is important because, in reality, the samples that require socio-demographic enrichment, i.e. the enrichment samples, can be collected at different levels of spatial detail due to the difference in the capabilities of data collection devices, or different privacy-preservation requirements, while the choice of corresponding estimation sample(s) can be limited. The paragraphs below hence explain two levels of spatial detail defined by two coding systems used in the UK and how they are linked (aggregated or disaggregated) to each other.

A number of coding systems are adopted by different authorities and organisations in the UK to spatially divide and define the geographical areas of the country. According to the levels of spatial detail at which the datasets exploited in this application are collected, two specific systems are explained in this section. Postal codes, also known as postcodes, are alphanumeric codes devised and adopted by the General Post Office to define the geographies in UK (Royal mail, 2004). Designating an area with several addresses or a single major delivery point, a full postcode consists of two parts separated by a single space. In particular, the former part of a postcode is termed as the outward postcode (OPC), which includes one or two letters defining the postcode area, followed by one or two digits indicating the postcode district within the area. The latter part, called the inward postcode, contains three characters for assisting in the delivery of posts within the postal district defined by the OPC (*A short history of the postcode*, 2005; Office for National Statistics, 2016b). For instance, ‘SW7 2AZ’ is the postcode uniquely designating the area of Imperial College London South Kensington Campus, of which ‘SW’

stands for the south-west area of London and ‘SW7’ defines the South Kensington District (Royal mail, 2004).

Alternatively, a variety of geographical units, such as counties, districts and Census Output Areas (OAs) and Super Output Areas (SOAs), constitute the main geographies associated with the vast majority of place-based open data, e.g. the UK Census (Office for National Statistics, 2011, 2016a). Among these geographies, OAs are the smallest geographic units for Census data releases, i.e. the lowest geographical level. OAs can be aggregated to produce most geographies for the Census data by best-fitting to the required output geographical detail level. In particular, OAs typically consist of adjacent unit postcodes with approximately 310 residents and tend to be constrained by obvious geographical boundaries such as major roads. SOAs are a set of geographical areas built up from aggregations of adjacent OAs to improve the reporting of small area statistics (Office for National Statistics, 2011). According to the population size, SOAs are hierarchically aggregated as two tiers named as Lower Layer Super Output Areas (LSOAs) and Middle Layer Super Output Areas (MSOAs) which contain around 1,500 and 7,200 people (650 and 4,000 households), respectively. Table 6.1 presents the range for the number of people and households for the above OA-based geographies.

Table 6.1 Number Limits for People and Households in each OA, LSOA and MSOA
(Source: Office for National Statistics (2019))

Geographical Area	Lower threshold		Upper threshold	
	People	Households	People	Households
OA	100	40	625	250
LSOA	1,000	400	3,000	1,200
MSOA	5,000	2,000	1,5000	6,000

In addition, Figure 6.1 presents a direct illustration of how OAs nestle within the LSOA and how LSOAs nestle within the MSOA.

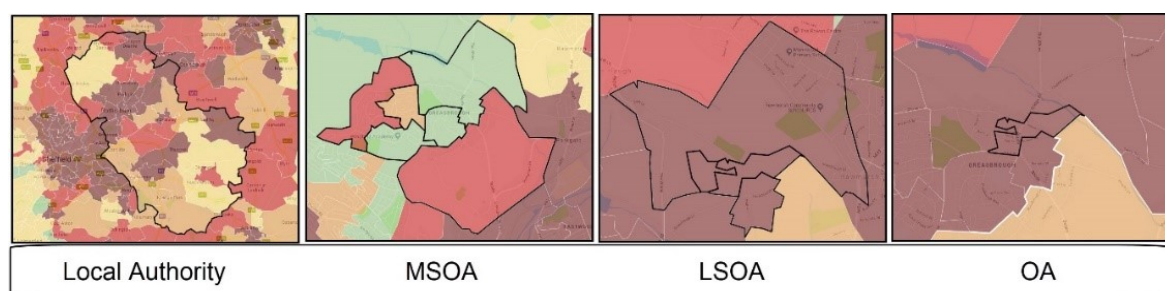


Figure 6.1 Inclusions between OAs, LSOAs and MSOAs (Source: OCSI (2019))

The Office for National Statistics of UK defines the above geographies using a code with a fixed length of nine characters. Particularly, the first three characters indicate the geographical detail level (e.g. LSOA or MSOA) and the following six digits define the individual unit. For example, the MSOA zone labelled ‘Kensington and Chelsea 001’ is coded as ‘E02000577’ where the prefix ‘E02’ refers to the MSOA-level geography in England.

The MND and LTDS data exploited in the current application are both collected from the Greater London area. According to the two coding systems above, the Greater London area can be geographically divided either into 228 OPC zones or 1,092 MSOAs. These OPC and MSOA zones can be linked using spatial aggregation or disaggregation, which is performed in this study to accommodate the different levels of spatial detail between the estimation and the enrichment samples. In particular, the 1,092 MSOAs can be aggregated to OPC-level zones, each containing 1 – 18 adjacent MSOAs. One challenge in such spatial aggregation involves the mismatch between the boundaries of MSOA and OPC zones due to the fact that they are defined by non-nested coding systems. It could be addressed by using the geographical centroid of each MSOA to determine its inclusion in an OPC zone, which can be questionable yet is convenient for the present analysis. Examples of this spatial aggregation linking a number of MSOA zones to a specific OPC zone are illustrated in Table 6.2.

Table 6.2 Examples for Inclusion of MSOAs in OPCs given Spatial Aggregation

OPC	List of Included MSOA Zones
WC1X	E02000189
WC1N	E02000190, E02000192
RM6	E02000002, E02000003, E02000769
...	...
CR0	E02000206, E02000209, E02000210, E02000211, E02000212, E02000213, E02000214, E02000215, E02000216, E02000217, E02000218, E02000219, E02000220, E02000221, E02000223, E02000225, E02000227, E02000229

6.1.2 Data sources

The enrichment sample for this empirical application is extracted from the datasets capturing the movements processed from mobile network call detail records collected by a mobile network provider. These MND datasets contain information about the MSOA-level movement patterns of mobile network users in London for 28 continuous days including both day-of-

weeks (DOWs), i.e. 20 weekdays and four weekends, in October and November 2016. In particular, the MND consists of two datasets:

- 1) A ‘flow’ dataset, which contains estimates of average hourly flows between MSOA zones of specific origin-destination (OD) pairs within the studied Greater London area; and
- 2) A ‘location’ dataset, which reports the average number of hourly dwellings (i.e. the number of people occur during a specific hour) in each MSOA zone.

The number of movements and dwellings in the two datasets are segmented into cohorts of at least ten individuals defined by a combination of gender and age group (20-24, 25-59, and 60+). According to such segmentation, the socio-demographic information can be preserved without identifying individual respondents, presenting a typical and realistic scenario that researchers and practitioners are likely to face in light of the prevailing privacy preservation legislation and industry best-practice. Examples for the formats of these two datasets are provided in Table 6.3 and Table 6.4, respectively.

Table 6.3 Format of MSOA-Level MND on Trip Information

DOW	Hour	Origin	Destination	Gender	Age Group	Rounded trips¹
Weekday	10	E02000001	E02000003	Male	Elderly	15
Weekday	8	E02000002	E02000002	Male	Young	20
Weekday	15	E02000002	E02000769	Female	Elderly	10
...	...	...	...	...	...	...
Weekend	23	E02000983	E02000960	Female	Young	5
Weekend	23	E02000983	E02000983	Male	Elderly	2

¹ The average number of trips during the 1-hour period between specified OD by respondents characterised by particular socio-demographic attributes on each DOW over the study period.

Table 6.4 Format of MSOA-Level MND for Dwelling Information

DOW	Hour	MSOA	Gender	Age Group	Rounded dwellings¹
Weekday	0	E02000001	Male	Young	73
Weekday	0	E02000001	Female	Young	62
Weekday	0	E02000001	Male	Middle-aged	153
...	...	...	...	...	...
Weekend	23	E02000983	Male	Elderly	55
Weekend	23	E02000983	Female	Elderly	46

¹ The average number of people (footfalls) who are characterised by particular socio-demographic attributes observed during the 1-hour period in the specified OD on a specific DOW.

On the other hand, the estimation sample for the current application is extracted from the 2015/16 LTDS, as introduced previously in Section 5.1. The LTDS is selected as the estimation sample because the data were collected from a different source than the MND, which serves as the enrichment sample. This is desirable for the present analysis so as to be able to explore the enrichment performance of the IDCM approach using samples with possibly different covariance structures, in contrast to the analysis in Chapter 5, where the data were derived from the same source. Furthermore, the LTDS data contain rich information about the socio-demographic attributes of the respondents in London, as well as their revealed travel-related choice behaviour patterns at the OPC level. This permits the estimation of various discrete choice models (DCMs) that can serve as the IDCM kernel(s) for the subsequent socio-demographic enrichment using the MND. An example of the format of the LTDS data with relevant information for the present application is provided in Table 6.5.

Table 6.5 Format of LTDS with Relevant Variables

Person ID	Trip ID	DOW	Time	Origin	Destination	Gender	Age
1500102101	150010210101	Weekday	07:59	W2	W2	Male	Middle-aged
1500102101	150010210102	Weekday	11:23	WC1X	NW4	Male	Middle-aged
1500102102	150010210201	Weekday	09:07	SW11	SW7	Female	Young
...	...	...	...			...	...
1500102103	150010210301	Weekend	16:20	RM6	RM6	Male	Elderly
1500102105	150010210301	Weekday	23:20	W6	SE17	Female	Middle-aged

6.1.3 Selection and modification of variables

A typical socio-demographic enrichment using the IDCM approach requires, as shown by the flow chart in Figure 4.2 (Section 4.1.2), the selection of the socio-demographic attributes that can be enriched using the IDCM approach and the input behaviour feature based on the available datasets.

Observing Table 6.3 and Table 6.4 suggests that the MND, from which the enrichment sample is extracted, contains only a limited number of variables. Hence, the input behaviour feature and the enriched attribute for the current application are selected on the basis of their availability in the employed MND. This is to ensure that the explored methods can be validated, i.e. that the enriched attribute can be compared with the actual observation of that attribute in the MND. The most promising discrete choice behaviour variable that can be directly and unambiguously obtained from the MND is the hourly departure time-of-the-day (TOD), i.e. the

hourly time period during which the group of respondents chooses to start their trips. This is therefore selected as the input behaviour feature for the enrichment using both the IDCM approach and the other methods applied in the comparative analysis. Since the IDCM performance theory developed so far has been proven only for the enrichment based on a single binary choice (per respondent), the TOD behaviour pattern is hereby re-coded to represent the binary behaviour variables describing whether the group of respondents chooses to depart during a specific hourly period or not. In this manner, 24 binary choice variables can be derived. Such binary TOD choice can also be distinguished by weekdays and weekends, according to the expected differences in activity-travel patterns between DOWs. This thus results in a total number of 48 hourly TOD choice variables, each exploited independently as an input behaviour feature for an IDCM enrichment process. During each hourly period which corresponds to a specific choice variable, the respondents observed to make footfalls (dwellings) in an MSOA zone constitute the enrichment sample for the hourly behaviour. In such an enrichment sample, the choice to depart is made by the number of respondents observed to travel during this period while the rest constitutes those who choose not to travel.

In addition to the hourly split, the LTDS also suggests that the 24 hours in a day can be divided into six periods given the distinguishable travel pattern revealed in each period. This, therefore, represents another type of input behaviour pattern for this empirical application, which are again revised into six binary period choices, each indicating whether the group of respondents chooses to depart during a specific period or not. This leads to another twelve choice variables once weekdays and weekends are considered separately. Table 6.6 illustrates the labels for the above six period choices corresponding to their characteristics.

Table 6.6 Labels and Names for Periods in 24 Hours of a Day

Period No.	Period Label	Time Period
1	Early AM	04:00 a.m. ~ 06:59 a.m.
2	AM Peak	07:00 a.m. ~ 09:59 a.m.
3	Inter Peak	10:00 a.m. ~ 03:59 p.m.
4	PM Peak	04:00 p.m. ~ 06:59 p.m.
5	Evening	07:00 p.m. ~ 23:59 p.m.
6	Night	12:00 a.m. ~ 03:59 a.m.

In terms of the enriched attributes, gender and age group are two socio-demographic attributes directly reported in the MND. To fit the scope of the current binary enrichment context, the three groups (20 – 24, 25 – 59, 60+) dividing age are modified as three separate attributes, each

implying whether the respondents fall into a specific age group or not. Conversely, gender is reported in the MND as a binary variable and requires no additional recoding. In total, therefore, four socio-demographic attributes are extracted as the attributes enriched in the present application, which are respectively named as ‘Gender’, ‘Young group’ (20 – 24), ‘Middle-aged group’ (25 – 59), and ‘Elderly group’ (60+).

The IDCM performance theory (Section 4.2.3) implies that the enrichment performance of the IDCM approach for socio-demographic attributes with different inherent randomness, i.e. $H(X)$ can vary with respect to the enrichment efficiency (η) in a different manner. Specifically, as is revealed by the blue curves in Figure 4.11 and Figure 4.12, the enrichment of an attribute with low certainty retains relatively low ‘percentage correctly predicted’ (PCP) levels, although these can be improved rapidly as η increases. In contrast, the enrichment of an attribute with low uncertainty, i.e. in the high-certainty group (orange curves in Figure 4.11 and Figure 4.12) usually maintain high PCP levels while the improvement rate and the variability due to η are both relatively minor. Gender and age-based attributes can represent the high- and low-certainty groups because gender is almost evenly distributed in the MND while the probability of falling into an age-group is significantly more deterministic. An example of this is illustrated in Figure 6.2, which presents the market shares of these four attributes using the observations from early AM (Period 1) in the enrichment sample. This explains the reason why both gender and the three age-based attributes are included as the enriched attributes for the present analysis.

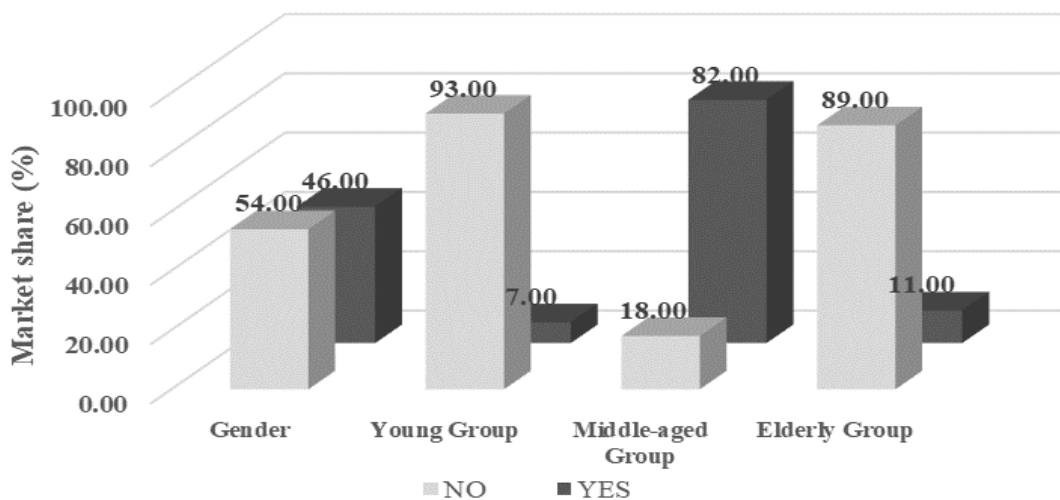


Figure 6.2 Market Shares of Gender and Age Groups for Early AM Observations

Note that, for each choice variable, i.e. each hourly/period slot, during which the choice of departure or not is made, the study sample of respondents can be different from the samples for the other slots because the group of respondents observed during one slot cannot be re-

identified during another slot. Hence, the market share of the enriched attribute $P(X)$ is different for different choices (slots). The input behaviour features, and enriched attributes are also extracted from the LTDS data to constitute the estimation sample upon which the IDCM kernels are calibrated. The variables extracted from the LTDS are also modified accordingly to fit the binary scope.

6.2 Experimental design

Based on the research objectives outlined in the introduction of this chapter, as well as the variables determined in Section 6.1.3, the enrichment experiments performed in the current MND enrichment application are designed as follows.

As per the first objective of this empirical application, the present analysis aims to explore the socio-demographic enrichment performance of the IDCM approach under different data conditions. The variation in the data condition involves two aspects: (i) the difference in the quality of the prior information with respect to the marginal probability distribution of the enriched attribute; and (ii) the change in the spatial detail level at which the enrichment sample is collected. The paragraphs below hence define the former prior quality by introducing the concepts of *similar* and *dissimilar priors*.

Recall Section 5.2.3 which formulates two estimation methods, viz. the MAP and the MLE estimators, to obtain the value of a binary socio-demographic attribute X using the IDCM approach for the binary enrichment context when neglecting the impact of the attributes of choice alternatives. Equations 6-1 and 6-2 below reproduce the corresponding formulations for the MAP and the MLE estimators as:

$$X_{MAP} = \arg \max_X P(Y|X)P(X); \quad (6-1)$$

$$X_{MLE} = \arg \max_X P(Y|X). \quad (6-2)$$

Specifically, the term $P(X)$ in the right-hand side (RHS) of Equation 6-1 represents any prior information informing the marginal probability distribution of X . Given that, in the current application, the only available prior involves the marginal probabilities that can be obtained from the estimation sample, such prior $P_{est}(X)$ is referred to as the *estimation prior*. Note, however, that the prior included in the MAP estimator need not necessarily be the estimation prior. Instead, it can come from any information available to the analysts with regard to $P(X)$.

As is argued in Section 5.4, the substantial outperformance of the IDCM approach using the MAP estimator, as well as the non-IDCM logistic regression (LGR) and support vector machines (SVMs) approaches, compared to the IDCM approach using the MLE estimator, can be explained by the significant similarity in the covariance structures of the estimation and the enrichment samples. It is thus assumed in the present analysis that the quality of the prior can be defined by its similarity to the true probability distribution of the enriched attribute in the enrichment sample, i.e. $P_{erc}(X)$. Accordingly, the impact of prior quality on the IDCM enrichment performance can be studied by defining two scenarios of different prior conditions depending on the level of similarity of the estimation prior to $P_{erc}(X)$, which leads to the notion of similar and dissimilar priors. In particular, one sample can be defined as similarly distributed to another if the difference in the marginal distributions of X in the two samples is not statistically significant given one-proportion Z-test. For the current analysis, an estimation prior that is similarly distributed to the enrichment sample is referred to as a *similar prior*. The opposite instance thus defines a *dissimilar prior*, which is not similarly distributed as compared to the enrichment sample. The two scenarios corresponding to the conditions of whether to use a similar or a dissimilar prior are therefore referred to as the ‘Similar Prior’ Scenario and the ‘Dissimilar Prior’ Scenario. The ‘Similar Prior’ Scenario is analogous to the empirical validation in Chapter 5, where both the estimation and the enrichment samples are generated from the same 2015/16 LTDS dataset. It is therefore treated as the benchmark against which the ‘Dissimilar Prior’ Scenario (if it exists) is compared in order to understand the impact that the quality of the prior has on the IDCM enrichment performance.

Table 6.7 presents the results of the one-proportion Z-test when comparing $P(X)$ in the raw LTDS and MND data.

Table 6.7 One-proportion Z-test results for Comparison of $P(X)$ in Raw LTDS and MND

DOW	Socio-demographic Attribute	Choice Type	Choice NO.	$P(X = 0)$ (%)		<i>p</i> -value
				LTDS	MND	
Weekdays	Gender	Hour	0	56.36	48.16	0.000
Weekdays	Gender	Hour	1	56.36	48.34	0.000
Weekdays	Gender	Hour	2	56.36	48.53	0.000
Weekdays	Gender	Hour	3	56.36	48.56	0.000
Weekdays	Gender	Hour	4	56.36	48.58	0.000
Weekdays	Gender	Hour	5	56.36	48.62	0.000
Weekdays	Gender	Hour	6	56.36	48.88	0.000
Weekdays	Gender	Hour	7	56.36	49.16	0.000

Weekdays	Gender	Hour	8	56.36	48.97	0.000
Weekdays	Gender	Hour	9	56.36	49.22	0.000
Weekdays	Gender	Hour	10	56.36	49.31	0.000
Weekdays	Gender	Hour	11	56.36	49.32	0.000
Weekdays	Gender	Hour	12	56.36	49.35	0.000
Weekdays	Gender	Hour	13	56.36	49.35	0.000
Weekdays	Gender	Hour	14	56.36	49.33	0.000
Weekdays	Gender	Hour	15	56.36	49.32	0.000
Weekdays	Gender	Hour	16	56.36	49.31	0.000
Weekdays	Gender	Hour	17	56.36	49.20	0.000
Weekdays	Gender	Hour	18	56.36	49.06	0.000
Weekdays	Gender	Hour	19	56.36	48.88	0.000
Weekdays	Gender	Hour	20	56.36	48.73	0.000
Weekdays	Gender	Hour	21	56.36	48.65	0.000
Weekdays	Gender	Hour	22	56.36	48.50	0.000
Weekdays	Gender	Hour	23	56.36	48.24	0.000
Weekdays	Young	Hour	0	93.28	92.52	0.103
Weekdays	Young	Hour	1	93.28	92.79	0.103
Weekdays	Young	Hour	2	93.28	92.98	0.103
Weekdays	Young	Hour	3	93.28	93.05	0.103
Weekdays	Young	Hour	4	93.28	93.05	0.103
Weekdays	Young	Hour	5	93.28	93.05	0.103
Weekdays	Young	Hour	6	93.28	93.08	0.103
Weekdays	Young	Hour	7	93.28	93.17	0.103
Weekdays	Young	Hour	8	93.28	93.30	0.103
Weekdays	Young	Hour	9	93.28	93.27	0.103
Weekdays	Young	Hour	10	93.28	93.30	0.103
Weekdays	Young	Hour	11	93.28	93.30	0.103
Weekdays	Young	Hour	12	93.28	93.30	0.103
Weekdays	Young	Hour	13	93.28	93.31	0.103
Weekdays	Young	Hour	14	93.28	93.30	0.103
Weekdays	Young	Hour	15	93.28	93.29	0.103
Weekdays	Young	Hour	16	93.28	93.29	0.103
Weekdays	Young	Hour	17	93.28	93.28	0.103
Weekdays	Young	Hour	18	93.28	93.28	0.103
Weekdays	Young	Hour	19	93.28	93.21	0.103
Weekdays	Young	Hour	20	93.28	93.19	0.103
Weekdays	Young	Hour	21	93.28	93.09	0.103
Weekdays	Young	Hour	22	93.28	92.83	0.103
Weekdays	Young	Hour	23	93.28	92.49	0.000
Weekdays	Elderly	Hour	0	30.97	17.20	0.000

Weekdays	Middle-aged	Hour	1	30.97	17.59	0.000
Weekdays	Middle-aged	Hour	2	30.97	17.84	0.000
Weekdays	Middle-aged	Hour	3	30.97	17.94	0.000
Weekdays	Middle-aged	Hour	4	30.97	17.95	0.000
Weekdays	Middle-aged	Hour	5	30.97	17.95	0.000
Weekdays	Middle-aged	Hour	6	30.97	17.99	0.000
Weekdays	Middle-aged	Hour	7	30.97	18.06	0.000
Weekdays	Middle-aged	Hour	8	30.97	18.16	0.000
Weekdays	Middle-aged	Hour	9	30.97	18.08	0.000
Weekdays	Middle-aged	Hour	10	30.97	17.99	0.000
Weekdays	Middle-aged	Hour	11	30.97	18.00	0.000
Weekdays	Middle-aged	Hour	12	30.97	18.01	0.000
Weekdays	Middle-aged	Hour	13	30.97	18.00	0.000
Weekdays	Middle-aged	Hour	14	30.97	18.00	0.000
Weekdays	Middle-aged	Hour	15	30.97	18.01	0.000
Weekdays	Middle-aged	Hour	16	30.97	18.05	0.000
Weekdays	Middle-aged	Hour	17	30.97	18.14	0.000
Weekdays	Middle-aged	Hour	18	30.97	18.17	0.000
Weekdays	Middle-aged	Hour	19	30.97	18.08	0.000
Weekdays	Middle-aged	Hour	20	30.97	18.01	0.000
Weekdays	Middle-aged	Hour	21	30.97	17.81	0.000
Weekdays	Middle-aged	Hour	22	30.97	17.48	0.000
Weekdays	Middle-aged	Hour	23	30.97	17.09	0.000
Weekdays	Elderly	Hour	0	75.74	90.28	0.000
Weekdays	Elderly	Hour	1	75.74	89.62	0.000
Weekdays	Elderly	Hour	2	75.74	89.18	0.000
Weekdays	Elderly	Hour	3	75.74	89.01	0.000
Weekdays	Elderly	Hour	4	75.74	89.01	0.000
Weekdays	Elderly	Hour	5	75.74	88.99	0.000
Weekdays	Elderly	Hour	6	75.74	88.93	0.000
Weekdays	Elderly	Hour	7	75.74	88.77	0.000
Weekdays	Elderly	Hour	8	75.74	88.54	0.000
Weekdays	Elderly	Hour	9	75.74	88.65	0.000
Weekdays	Elderly	Hour	10	75.74	88.71	0.000
Weekdays	Elderly	Hour	11	75.74	88.70	0.000
Weekdays	Elderly	Hour	12	75.74	88.69	0.000
Weekdays	Elderly	Hour	13	75.74	88.69	0.000
Weekdays	Elderly	Hour	14	75.74	88.70	0.000
Weekdays	Elderly	Hour	15	75.74	88.71	0.000
Weekdays	Elderly	Hour	16	75.74	88.66	0.000
Weekdays	Elderly	Hour	17	75.74	88.58	0.000

Weekdays	Elderly	Hour	18	75.74	88.55	0.000
Weekdays	Elderly	Hour	19	75.74	88.70	0.000
Weekdays	Elderly	Hour	20	75.74	88.80	0.000
Weekdays	Elderly	Hour	21	75.74	89.10	0.000
Weekdays	Elderly	Hour	22	75.74	89.70	0.000
Weekdays	Elderly	Hour	23	75.74	90.41	0.000
Weekdays	Gender	Period	1	56.36	48.70	0.000
Weekdays	Gender	Period	2	56.36	49.12	0.000
Weekdays	Gender	Period	3	56.36	49.33	0.000
Weekdays	Gender	Period	4	56.36	49.19	0.000
Weekdays	Gender	Period	5	56.36	48.63	0.000
Weekdays	Gender	Period	6	56.36	48.43	0.000
Weekdays	Young	Period	1	93.28	93.06	0.103
Weekdays	Young	Period	2	93.28	93.25	0.103
Weekdays	Young	Period	3	93.28	93.30	0.103
Weekdays	Young	Period	4	93.28	93.29	0.103
Weekdays	Young	Period	5	93.28	93.01	0.103
Weekdays	Young	Period	6	93.28	92.87	0.103
Weekdays	Middle-aged	Period	1	30.97	17.96	0.000
Weekdays	Middle-aged	Period	2	30.97	18.10	0.000
Weekdays	Middle-aged	Period	3	30.97	18.00	0.000
Weekdays	Middle-aged	Period	4	30.97	18.12	0.000
Weekdays	Middle-aged	Period	5	30.97	17.76	0.000
Weekdays	Middle-aged	Period	6	30.97	17.69	0.000
Weekdays	Elderly	Period	1	75.74	88.98	0.000
Weekdays	Elderly	Period	2	75.74	88.65	0.000
Weekdays	Elderly	Period	3	75.74	88.70	0.000
Weekdays	Elderly	Period	4	75.74	88.60	0.000
Weekdays	Elderly	Period	5	75.74	89.24	0.000
Weekdays	Elderly	Period	6	75.74	89.43	0.000
Weekend	Gender	Hour	0	50.52	48.21	0.000
Weekend	Gender	Hour	1	50.52	48.28	0.000
Weekend	Gender	Hour	2	50.52	48.39	0.000
Weekend	Gender	Hour	3	50.52	48.40	0.000
Weekend	Gender	Hour	4	50.52	48.41	0.000
Weekend	Gender	Hour	5	50.52	48.41	0.000
Weekend	Gender	Hour	6	50.52	48.44	0.000
Weekend	Gender	Hour	7	50.52	48.52	0.007
Weekend	Gender	Hour	8	50.52	48.55	0.007
Weekend	Gender	Hour	9	50.52	48.61	0.007
Weekend	Gender	Hour	10	50.52	48.67	0.007

Weekend	Gender	Hour	11	50.52	48.71	0.007
Weekend	Gender	Hour	12	50.52	48.72	0.007
Weekend	Gender	Hour	13	50.52	48.74	0.007
Weekend	Gender	Hour	14	50.52	48.74	0.007
Weekend	Gender	Hour	15	50.52	48.71	0.007
Weekend	Gender	Hour	16	50.52	48.71	0.007
Weekend	Gender	Hour	17	50.52	48.69	0.007
Weekend	Gender	Hour	18	50.52	48.68	0.007
Weekend	Gender	Hour	19	50.52	48.67	0.007
Weekend	Gender	Hour	20	50.52	48.66	0.007
Weekend	Gender	Hour	21	50.52	48.63	0.007
Weekend	Gender	Hour	22	50.52	48.55	0.007
Weekend	Gender	Hour	23	50.52	48.36	0.000
Weekend	Young	Hour	0	92.34	92.30	0.253
Weekend	Young	Hour	1	92.34	92.68	0.026
Weekend	Young	Hour	2	92.34	92.94	0.026
Weekend	Young	Hour	3	92.34	93.04	0.026
Weekend	Young	Hour	4	92.34	93.03	0.026
Weekend	Young	Hour	5	92.34	93.04	0.026
Weekend	Young	Hour	6	92.34	93.05	0.026
Weekend	Young	Hour	7	92.34	93.08	0.026
Weekend	Young	Hour	8	92.34	93.16	0.026
Weekend	Young	Hour	9	92.34	93.19	0.026
Weekend	Young	Hour	10	92.34	93.22	0.026
Weekend	Young	Hour	11	92.34	93.23	0.026
Weekend	Young	Hour	12	92.34	93.24	0.026
Weekend	Young	Hour	13	92.34	93.25	0.026
Weekend	Young	Hour	14	92.34	93.26	0.026
Weekend	Young	Hour	15	92.34	93.26	0.026
Weekend	Young	Hour	16	92.34	93.26	0.026
Weekend	Young	Hour	17	92.34	93.27	0.026
Weekend	Young	Hour	18	92.34	93.27	0.026
Weekend	Young	Hour	19	92.34	93.26	0.026
Weekend	Young	Hour	20	92.34	93.23	0.026
Weekend	Young	Hour	21	92.34	93.11	0.026
Weekend	Young	Hour	22	92.34	92.85	0.026
Weekend	Young	Hour	23	92.34	92.42	0.253
Weekend	Middle-aged	Hour	0	29.15	17.01	0.000
Weekend	Middle-aged	Hour	1	29.15	17.46	0.000
Weekend	Middle-aged	Hour	2	29.15	17.77	0.000
Weekend	Middle-aged	Hour	3	29.15	17.87	0.000

Weekend	Middle-aged	Hour	4	29.15	17.88	0.000
Weekend	Middle-aged	Hour	5	29.15	17.89	0.000
Weekend	Middle-aged	Hour	6	29.15	17.92	0.000
Weekend	Middle-aged	Hour	7	29.15	17.91	0.000
Weekend	Middle-aged	Hour	8	29.15	17.95	0.000
Weekend	Middle-aged	Hour	9	29.15	18.03	0.000
Weekend	Middle-aged	Hour	10	29.15	18.08	0.000
Weekend	Middle-aged	Hour	11	29.15	18.09	0.000
Weekend	Middle-aged	Hour	12	29.15	18.10	0.000
Weekend	Middle-aged	Hour	13	29.15	18.11	0.000
Weekend	Middle-aged	Hour	14	29.15	18.10	0.000
Weekend	Middle-aged	Hour	15	29.15	18.08	0.000
Weekend	Middle-aged	Hour	16	29.15	18.08	0.000
Weekend	Middle-aged	Hour	17	29.15	18.08	0.000
Weekend	Middle-aged	Hour	18	29.15	18.06	0.000
Weekend	Middle-aged	Hour	19	29.15	18.03	0.000
Weekend	Middle-aged	Hour	20	29.15	17.99	0.000
Weekend	Middle-aged	Hour	21	29.15	17.80	0.000
Weekend	Middle-aged	Hour	22	29.15	17.46	0.000
Weekend	Middle-aged	Hour	23	29.15	17.08	0.000
Weekend	Elderly	Hour	0	78.52	90.69	0.000
Weekend	Elderly	Hour	1	78.52	89.86	0.000
Weekend	Elderly	Hour	2	78.52	89.29	0.000
Weekend	Elderly	Hour	3	78.52	89.09	0.000
Weekend	Elderly	Hour	4	78.52	89.09	0.000
Weekend	Elderly	Hour	5	78.52	89.07	0.000
Weekend	Elderly	Hour	6	78.52	89.04	0.000
Weekend	Elderly	Hour	7	78.52	89.01	0.000
Weekend	Elderly	Hour	8	78.52	88.89	0.000
Weekend	Elderly	Hour	9	78.52	88.78	0.000
Weekend	Elderly	Hour	10	78.52	88.71	0.000
Weekend	Elderly	Hour	11	78.52	88.68	0.000
Weekend	Elderly	Hour	12	78.52	88.66	0.000
Weekend	Elderly	Hour	13	78.52	88.64	0.000
Weekend	Elderly	Hour	14	78.52	88.64	0.000
Weekend	Elderly	Hour	15	78.52	88.66	0.000
Weekend	Elderly	Hour	16	78.52	88.66	0.000
Weekend	Elderly	Hour	17	78.52	88.65	0.000
Weekend	Elderly	Hour	18	78.52	88.68	0.000
Weekend	Elderly	Hour	19	78.52	88.71	0.000
Weekend	Elderly	Hour	20	78.52	88.79	0.000

Weekend	Elderly	Hour	21	78.52	89.09	0.000
Weekend	Elderly	Hour	22	78.52	89.69	0.000
Weekend	Elderly	Hour	23	78.52	90.50	0.000
Weekend	Gender	Period	1	50.52	48.42	0.000
Weekend	Gender	Period	2	50.52	48.56	0.007
Weekend	Gender	Period	3	50.52	48.71	0.007
Weekend	Gender	Period	4	50.52	48.69	0.007
Weekend	Gender	Period	5	50.52	48.59	0.007
Weekend	Gender	Period	6	50.52	48.33	0.000
Weekend	Young	Period	1	92.34	93.04	0.026
Weekend	Young	Period	2	92.34	93.14	0.026
Weekend	Young	Period	3	92.34	93.24	0.026
Weekend	Young	Period	4	92.34	93.26	0.026
Weekend	Young	Period	5	92.34	93.03	0.026
Weekend	Young	Period	6	92.34	92.79	0.026
Weekend	Middle-aged	Period	1	29.15	17.90	0.000
Weekend	Middle-aged	Period	2	29.15	17.97	0.000
Weekend	Middle-aged	Period	3	29.15	18.09	0.000
Weekend	Middle-aged	Period	4	29.15	18.07	0.000
Weekend	Middle-aged	Period	5	29.15	17.73	0.000
Weekend	Middle-aged	Period	6	29.15	17.58	0.000
Weekend	Elderly	Period	1	78.52	89.06	0.000
Weekend	Elderly	Period	2	78.52	88.89	0.000
Weekend	Elderly	Period	3	78.52	88.67	0.000
Weekend	Elderly	Period	4	78.52	88.66	0.000
Weekend	Elderly	Period	5	78.52	89.25	0.000
Weekend	Elderly	Period	6	78.52	89.63	0.000

Table 6.7 (above) shows that the probability distribution of the four socio-demographic attributes follows a significantly dissimilar distribution between the LTDS and the MND data for most choice variables (p -values < 0.1). Exceptions, highlighted in red, are observed in a few cases where p -values are greater than 0.1, revealing that the difference between the market shares of the ‘Young group’ in the two data sources is not significant at the 90% level for specific choice behaviour patterns. Nonetheless, the relatively low p -values (0.103 and 0.253) still suggest that the $P(X)$ in the LTDS and the MND are not similar.

The illustration in Figure 6.3 (below) directly presents the difference in the probability distribution of $P(X)$ in the LTDS and the MND data, taking the observations during the early AM (Period 1) as an example.

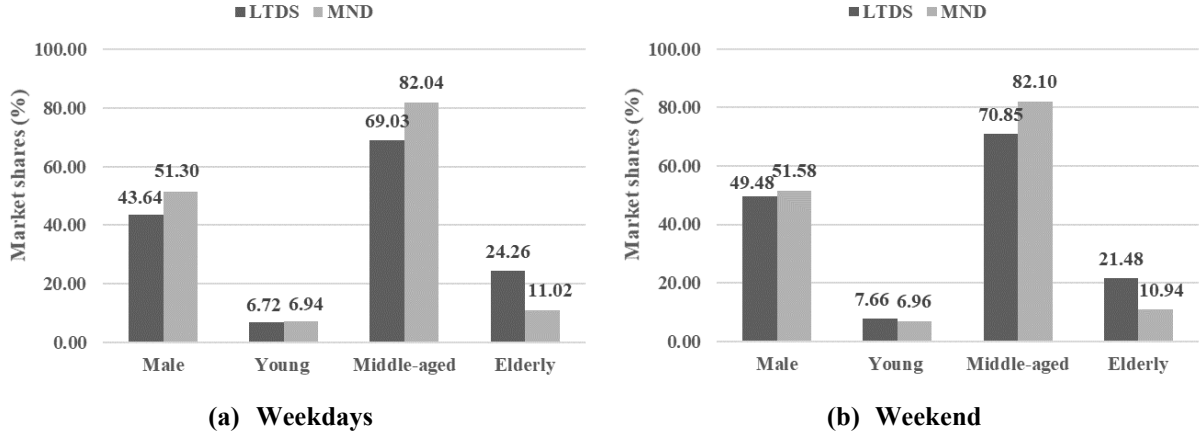


Figure 6.3 Example Comparison between $P_{est}(X)$ and $P_{erc}(X)$ for Observations during Early AM

It can thus be derived from the above analysis that the data extracted from the raw LTDS and MND data can naturally play the role of the estimation and the enrichment samples for the ‘Dissimilar Prior’ Scenario. In contrast, the corresponding estimation and enrichment samples for the ‘Similar Prior’ Scenario require additionally selecting a random subsample from the LTDS such that the obtained $P_{est}(X)$ is significantly similar to the $P_{erc}(X)$ in the MND.

In each of the two scenarios described above, 32 ($2 \times 4 \times 4$) sets of enrichment experiments are conducted, each corresponding to the enrichment of one of the four binary socio-demographic attributes (‘Gender’, ‘Young group’, ‘Middle-aged group’ and ‘Elderly group’), given observations of one of the four choice patterns (weekday-hourly, weekend-hourly, weekday-period and weekend-period choices) for the anonymised enrichment MND sample collected at one of the two spatial detail levels (the MSOA and OPC levels). In each of the two sets of experiments involving hourly TOD choice patterns (weekday-hourly and weekend hourly choices), 24 enrichment experiments are performed, each relating to one of the 24 hours during which the departure TOD choices are observed. On the other hand, for each of the two sets of experiments given period TOD choice patterns (weekday-period and weekend-period choices) (Table 6.6), six experiments are conducted to enrich, respectively, each socio-demographic attribute from a binary choice indicating whether departure is made during a period. For reference convenience, the experiment sets are labelled in the manner below:

$$(\text{Spatial-level})_(\text{DOW})_(\text{Type-of-Choice})_(\text{Attribute}) \quad (6-3)$$

The variables defined in the brackets (Equation 6-3) are replaced with the specific abbreviations according to Table 6.8 (below) to complete the code for each enrichment experiment set.

Table 6.8 Abbreviations for Factor Labelling of Experiment Design

Variables	Values	Abbreviations
DOW	Weekday	WKDY
	Weekend	WKND
Type of Choice	Hourly choice	HR
	Period choice	PR
Attribute	Gender	GEN
	Young	YNG
	Middle-aged	MDA
	Elderly	EDL

For instance, the set of experiments for the enrichment of gender, given hourly TOD choice patterns made on weekdays at the MSOA level can be coded as MSOA_WKDY_HR_GEN.

Table 6.9 summarises the number of enrichment enrichments conducted in each set of experiments segmented by the attribute, the choice and the spatial detail level.

Table 6.9 Number of Experiments Segmented by Attribute, Choice and Spatial Detail Level

Enriched Attribute	WKDY_HR	WKND_HR	WKDY_PR	WKND_PR
Gender	MSOA: 24	MSOA: 24	MSOA: 6	MSOA: 6
	OPC: 24	OPC: 24	OPC: 6	OPC: 6
Young group	MSOA: 24	MSOA: 24	MSOA: 6	MSOA: 6
	OPC: 24	OPC: 24	OPC: 6	OPC: 6
Middle-aged group	MSOA: 24	MSOA: 24	MSOA: 6	MSOA: 6
	OPC: 24	OPC: 24	OPC: 6	OPC: 6
Elderly group	MSOA: 24	MSOA: 24	MSOA: 6	MSOA: 6
	OPC: 24	OPC: 24	OPC: 6	OPC: 6

6.3 Results and comparative analysis

The sections below present the results and corresponding comparative analysis for the current empirical application according to the outlined objectives. Particularly, Section 6.3.1 presents the analysis concerning the variation of the IDCM enrichment performance under different data conditions caused by the different probability distribution between the estimation and the enrichment samples caused by different factors. Correspondingly, the capability of the IDCM performance theory to estimate the enrichment performance of the IDCM approach based on specified data conditions is investigated and analysed in Section 6.3.2, while Section 6.3.3

compares the IDCM framework against logistic regression (LGR) and support vector machines (SVMs) as representatives of the methods that have been extensively applied in previous socio-demographic enrichment attempts.

6.3.1 IDCM enrichment performance versus data conditions

The current section presents the enrichment performance (i.e. PCP) when conducting the enrichment experiments that are designed in Section 6.2 using the IDCM approach. In particular, Section 6.3.1.1 explores the variation of the IDCM enrichment performance with respect to the different qualities of the prior, reflecting the variation in the estimation sample. On the other hand, Section 6.3.1.2 discusses how the IDCM enrichment performance varies with regard to the different levels of spatial detail at which the enrichment sample is collected.

6.3.1.1 Quality of prior

To investigate the impact of the quality of the estimation prior ('Similar' or 'Dissimilar') on the enrichment performance of the IDCM approach, the PCPs obtained using the IDCM approach for each paired set of experiments in the 'Similar Prior' Scenario and the 'Dissimilar Prior' Scenario (i.e. the enrichment of the same socio-demographic attribute given the same type of choice for the enrichment sample collected at the same spatial detail level) are compared using either paired *t*-test or Wilcoxon signed-rank test, depending on whether the samples of PCPs are normally distributed. In particular, a paired *t*-test is performed when the PCPs obtained from both of the compared sets of enrichment experiments are normally distributed, and otherwise the Wilcoxon signed-rank test.

The enrichment performance of the IDCM approach using the MLE and the MAP estimators are analysed respectively in the following paragraphs. In particular, Table 6.10 presents the test results of comparing the MLE-PCPs between the 'Similar Prior' and the 'Dissimilar Prior' Scenarios while the comparison for MAP-PCPs is demonstrated later in Table 6.11.

Table 6.10 Test Results for Comparison of MLE-PCPs in 'Similar' & 'Dissimilar Prior' Scenarios

Exp. NO.	Series Label	MLE-PCP (%)		Sd. ^a (%)	Test ^b	p-value
		Similar	Dissimilar			
1	MSOA_WKDY_HR_GEN	50.55	49.45	1.12	T	0.000
2	MSOA_WKDY_HR_YNG	86.00	86.00	0.00	T	1.000
3	MSOA_WKDY_HR_MDA	76.47	76.47	0.00	T	1.000

4	MSOA_WKDY_HR_EDL	82.18	82.18	0.00	T	1.000
5	MSOA_WKDY_PR_GEN	50.08	49.49	1.02	T	0.217
6	MSOA_WKDY_PR_YNG	85.73	85.73	0.00	T	1.000
7	MSOA_WKDY_PR_MDA	76.13	76.13	0.00	T	1.000
8	MSOA_WKDY_PR_EDL	81.80	81.80	0.00	T	1.000
9	MSOA_WKND_HR_GEN	51.01	49.15	0.67	W	0.000
10	MSOA_WKND_HR_YNG	87.31	87.31	0.00	W	1.000
11	MSOA_WKND_HR_MDA	77.56	77.56	0.00	W	1.000
12	MSOA_WKND_HR_EDL	83.39	83.39	0.00	W	1.000
13	MSOA_WKND_PR_GEN	50.71	49.30	1.12	W	0.050
14	MSOA_WKND_PR_YNG	76.40	76.40	0.00	W	1.000
15	MSOA_WKND_PR_MDA	69.59	69.59	0.00	W	1.000
16	MSOA_WKND_PR_EDL	73.70	73.70	0.00	W	1.000
17	OPC_WKDY_HR_GEN	50.55	49.45	1.12	T	0.000
18	OPC_WKDY_HR_YNG	86.18	86.18	0.00	T	1.000
19	OPC_WKDY_HR_MDA	76.60	76.60	0.00	T	1.000
20	OPC_WKDY_HR_EDL	82.35	82.35	0.00	T	1.000
21	OPC_WKDY_PR_GEN	50.07	49.49	1.03	T	0.224
22	OPC_WKDY_PR_YNG	85.91	85.91	0.00	T	1.000
23	OPC_WKDY_PR_MDA	76.27	76.27	0.00	T	1.000
24	OPC_WKDY_PR_EDL	81.97	81.97	0.00	T	1.000
25	OPC_WKND_HR_GEN	51.02	49.15	0.67	W	0.000
26	OPC_WKND_HR_YNG	87.47	87.47	0.00	W	1.000
27	OPC_WKND_HR_MDA	77.69	77.69	0.00	W	1.000
28	OPC_WKND_HR_EDL	83.55	83.55	0.00	W	1.000
29	OPC_WKND_PR_GEN	50.71	49.30	1.13	W	0.050
30	OPC_WKND_PR_YNG	76.45	76.45	0.00	W	1.000
31	OPC_WKND_PR_MDA	69.64	69.64	0.00	W	1.000
32	OPC_WKND_PR_EDL	73.75	73.75	0.00	W	1.000

^a Standard deviation of the difference between the MLE-PCPs obtained from each set of enrichment experiments in the ‘Dissimilar Prior’ and the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

Specifically, it can be observed from Table 6.10 that the MLE-PCPs obtained, respectively, from the ‘Similar prior’ and ‘Dissimilar prior’ Scenarios are not significantly different in all 24 sets of experiments for enriching the three age-based attributes, shown by *p*-values that are all greater than 0.1. This keeps consistency with the logic that the enrichment performance using the MLE estimator is not affected by the quality of the prior since the MLE estimator itself purposely excludes any prior information regarding the enriched attribute. Nevertheless, it is

also noted, from the results marked in red, that in six out of the eight sets of enrichment experiments for gender, the MLE-PCPs are slightly different between scenarios. These include:

- 1) MSOA_WKDY_HR_GEN (No. 1);
- 2) MSOA_WKND_HR_GEN (No. 9);
- 3) MSOA_WKND_PR_GEN (No. 13);
- 4) OPC_WKDY_HR_GEN (No. 17);
- 5) OPC_WKND_HR_GEN (No. 25); and
- 6) OPC_WKND_PR_GEN (No. 29).

This contradicts the aforementioned intuition that the enrichment performance of the MLE estimator is not affected by the quality of the prior. To explain this discrepancy, the property associated with the correctness of the prior needs to be defined, in addition to the informativeness of the prior as the other property described and discussed in Section 4.2.3. Specifically, the dissimilarity of the estimation prior compared to the true attribute distribution in the enrichment sample, as reflected in the ‘Dissimilar Prior’ Scenario, can result in two types of prior effect which are referred, according to their relatively positive or negative impact in informing the actual $P_{erc}(X)$, to here as ‘correct’ and ‘incorrect’, respectively. These are further defined as follows.

For binary attributes, a ‘correct’ prior means that the difference between the market shares of the two possible outcome values (POVs) of the enriched attribute has the same sign in the estimation sample and the enrichment sample, i.e.:

$$[p_{est}(X = 1) - p_{est}(X = 0)] \cdot [p_{erc}(X = 1) - p_{erc}(X = 0)] > 0 \quad (6-4)$$

where $p_{est}(X = i)$ is the marginal probability for $X = i$ in the estimation sample.

$p_{erc}(X = i)$ is the marginal probability for $X = i$ in the enrichment sample.

For example, if the proportion of males in the enrichment sample is greater than that of females, a correct estimation prior means that the proportion of males is also greater than that of female in the estimation sample. Otherwise, the estimation prior is referred to as an incorrect prior. The prior correctness for binary attributes defined in Equation 6-4 can also be extended to accommodate categorical or ordinal attributes: the estimation prior is referred to as correct if the relatively large or small amount of the probability distribution for the POVs of the attribute can be ranked in the same order in both the estimation and the enrichment samples. Otherwise, the estimation prior is denoted as an incorrect prior.

In light of the above definition, Table 6.7 (Page 189) indicates that the prior information provided by the LTDS data is incorrect for gender as compared to the MND data. This implies that the estimation samples in the ‘Dissimilar Prior’ Scenario typically provide the enrichment of gender with incorrect prior information as both samples are extracted directly from the two data sources. In particular, the market share of female respondents is found to be greater than that of the male respondents in the estimation sample (LTDS), while the opposite is observed for the enrichment sample (MND). In contrast, the estimation sample is found to provide correct priors for the enrichment of age-based attributes. This is reflected by the same sign of the difference between the market shares of the two POVs for each age-based attribute. The incorrect priors for gender in the ‘Dissimilar Prior’ Scenario could result in the mapping rules being swapped between these two samples in the corresponding sets of enrichment experiments. As is suggested by Lemma 1 (Section 4.2.3), the enrichment using the MLE estimator is purely determined by the β_1 of the IDCM kernel, which also defines the mapping rule and decoding scheme. The incorrectness of the prior thus provides an incorrect mapping rule for the decoding scheme of the MLE estimator, which consequently deteriorates the MLE-PCP obtained in the ‘Dissimilar Prior’ Scenario as compared to that in the ‘Similar Prior’ Scenario. The effect of this prior incorrectness is, however, found to be insignificant (p -values > 0.1) for the enrichment of gender given weekday-period choices, i.e.:

- 1) MSOA_WKDY_PR_GEN (No. 5); and
- 2) OPC_WKDY_PR_GEN (No. 21).

Detailed investigations suggest that the indifferent MLE-PCP between the ‘Similar Prior’ and the ‘Dissimilar Prior’ Scenarios in these two sets of enrichment experiments (greyed out in Table 6.7) can be interpreted by the relatively non-significant informativeness of the employed estimation prior. In particular, recall Equation 3-114 (Section 4.2.3.6), which defines the domain for $P(X)$ that can provide an informative prior given the probability distribution of the choice $P(Y)$ and the β_1 in the IDCM kernel, as restated below:

$$p(x_0) < \frac{(\tau - 1) p_y + 2}{2(\tau + 1)} \quad \text{or} \quad p(x_0) > \frac{1}{2} + \frac{p_y(\tau - 1)}{2(\tau + 1)}. \quad (6-5)$$

According to this definition, over two-thirds of the estimation sample in the above two exceptional sets of enrichment experiments provide an uninformative prior, whereas those in the remaining 28 sets of experiments are found to be informative. This indicates that the impact of the prior correctness on the MLE-PCP is negligible for uninformative estimation priors.

Table 6.11 Test Results for Comparison of MAP-PCPs in ‘Similar’ & ‘Dissimilar Prior’ Scenarios

Exp. NO.	Series Label	MAP-PCP (%)		Sd. ^a (%)	Test ^b	p-value
		Similar	Dissimilar			
1	MSOA_WKDY_HR_GEN	50.98	49.21	0.80	W	0.000
2	MSOA_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
3	MSOA_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000
4	MSOA_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000
5	MSOA_WKDY_PR_GEN	50.76	49.21	0.85	T	0.006
6	MSOA_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
7	MSOA_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
8	MSOA_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
9	MSOA_WKND_HR_GEN	51.35	48.87	0.29	T	0.000
10	MSOA_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
11	MSOA_WKND_HR_MDA	82.13	82.10	0.18	W	0.317
12	MSOA_WKND_HR_EDL	89.07	89.07	0.00	W	1.000
13	MSOA_WKND_PR_GEN	51.40	49.30	1.06	W	0.050
14	MSOA_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
15	MSOA_WKND_PR_MDA	82.11	82.11	0.00	T	1.000
16	MSOA_WKND_PR_EDL	89.03	89.03	0.00	T	1.000
17	OPC_WKDY_HR_GEN	50.98	49.21	0.80	W	0.000
18	OPC_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
19	OPC_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000
20	OPC_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000
21	OPC_WKDY_PR_GEN	50.76	49.21	0.85	T	0.006
22	OPC_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
23	OPC_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
24	OPC_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
25	OPC_WKND_HR_GEN	51.35	48.86	0.29	T	0.000
26	OPC_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
27	OPC_WKND_HR_MDA	82.13	82.10	0.17	W	0.317
28	OPC_WKND_HR_EDL	89.07	89.07	0.00	W	1.000
29	OPC_WKND_PR_GEN	51.40	49.30	1.06	W	0.050
30	OPC_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
31	OPC_WKND_PR_MDA	82.11	82.11	0.00	T	1.000
32	OPC_WKND_PR_EDL	89.03	89.03	0.00	T	1.000

^a Standard deviation of the difference between the MAP-PCPs obtained from each set of experiments in the ‘Dissimilar Prior’ and the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

On the other hand, the results of the comparative analysis for the MAP-PCPs between ‘Similar Prior’ and ‘Dissimilar Prior’ Scenarios, as provided in Table 6.11, indicate that the IDCM enrichment performance using the MAP estimator (MAP-PCPs) in the ‘Similar prior’ Scenario surpasses that in the ‘Dissimilar prior’ Scenario for all the eight sets of gender enrichment experiments (p -values < 0.1). This reveals that an incorrect prior can deteriorate the MAP-PCP compared with using a correct prior. In the remaining 24 sets of enrichment experiments for age-based attributes, however, the MAP estimator is found to perform comparably in ‘Similar prior’ and ‘Dissimilar prior’ Scenarios, indicating that the correct priors can provide an equivalent level of enrichment performance, regardless of their different magnitude of similarity to the true distribution of the enriched attribute in the enrichment sample.

The impact of the qualities of prior on the enrichment performance of the IDCM approach can be investigated from another dimension, to identify the estimation method that performs best (i.e. the MLE or the MAP estimator) given different prior conditions. Table 6.12 presents the tests result for the comparison of the MLE- and the MAP-PCPs in the ‘Similar Prior’ Scenarios.

Table 6.12 Test Results for Comparison of MLE- and MAP-PCPs in ‘Similar Prior’ Scenarios

Exp. NO.	Series Label	MLE-PCP (%)	MAP-PCP (%)	Sd. ^a (%)	Test ^b	p -value
1	MSOA_WKDY_HR_GEN	50.55	50.98	0.47	W	0.000
2	MSOA_WKDY_HR_YNG	86.00	93.11	4.66	W	0.000
3	MSOA_WKDY_HR_MDA	76.47	82.10	3.83	W	0.000
4	MSOA_WKDY_HR_EDL	82.18	88.99	4.55	W	0.000
5	MSOA_WKDY_PR_GEN	50.08	50.76	0.95	T	0.131
6	MSOA_WKDY_PR_YNG	85.73	93.13	5.19	T	0.017
7	MSOA_WKDY_PR_MDA	76.13	82.06	4.23	T	0.019
8	MSOA_WKDY_PR_EDL	81.80	88.93	5.04	T	0.018
9	MSOA_WKND_HR_GEN	51.01	51.35	0.28	W	0.000
10	MSOA_WKND_HR_YNG	87.31	93.07	3.24	W	0.000
11	MSOA_WKND_HR_MDA	77.56	82.13	2.49	W	0.000
12	MSOA_WKND_HR_EDL	83.39	89.07	2.98	W	0.000
13	MSOA_WKND_PR_GEN	50.71	51.40	0.84	W	0.050
14	MSOA_WKND_PR_YNG	76.40	93.08	29.84	W	0.050
15	MSOA_WKND_PR_MDA	69.59	82.11	21.78	W	0.050
16	MSOA_WKND_PR_EDL	73.70	89.03	26.38	W	0.050
17	OPC_WKDY_HR_GEN	50.55	50.98	0.47	W	0.000
18	OPC_WKDY_HR_YNG	86.18	93.11	4.55	W	0.000
19	OPC_WKDY_HR_MDA	76.60	82.10	3.74	W	0.000

20	OPC_WKDY_HR_EDL	82.35	88.99	4.46	W	0.000
21	OPC_WKDY_PR_GEN	50.07	50.76	0.96	T	0.141
22	OPC_WKDY_PR_YNG	85.91	93.13	5.07	T	0.017
23	OPC_WKDY_PR_MDA	76.27	82.06	4.14	T	0.019
24	OPC_WKDY_PR_EDL	81.97	88.93	4.93	T	0.018
25	OPC_WKND_HR_GEN	51.02	51.35	0.27	W	0.000
26	OPC_WKND_HR_YNG	87.47	93.07	3.13	W	0.000
27	OPC_WKND_HR_MDA	77.69	82.13	2.41	W	0.000
28	OPC_WKND_HR_EDL	83.55	89.07	2.88	W	0.000
29	OPC_WKND_PR_GEN	50.71	51.40	0.84	W	0.050
30	OPC_WKND_PR_YNG	76.45	93.08	30.00	W	0.050
31	OPC_WKND_PR_MDA	69.64	82.11	21.91	W	0.050
32	OPC_WKND_PR_EDL	73.75	89.03	26.53	W	0.050

^a Standard deviation of the difference between the MAP-PCPs and the MLE-PCPs obtained from each set of experiments in the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

It can be seen from Table 6.12 that the MLE estimator generally performs no better than its MAP counterpart, revealed by the greater (or at least equivalent) values of the average MAP-PCP in each set of enrichment experiments. Specifically, as is indicated by the *p*-values being smaller than 0.1, the MAP-PCP surpasses the MLE-PCP in 30 sets of experiments with 90% confidence, except for the enrichment of gender from period choices made on weekdays, i.e.:

- 1) MSOA_WKDY_PR_GEN (No. 5); and
- 2) OPC_WKDY_PR_GEN (No. 21).

The MLE- and MAP-PCPs in the above two sets of enrichment experiments are found to be indifferent, as revealed by the corresponding *p*-values (in red) that are greater than 0.1. This can be interpreted by the fact that over two-thirds of the priors included in the MAP estimator for the two exceptional sets are not sufficiently informative to make a difference in the PCP, thus leading to the comparable enrichment performance when using the MLE and the MAP estimators. This also aligns with the findings, from the previous theoretical (Section 4.2.3) and empirical investigations (Section 5.4) into the IDCM performance theory, that using an MAP estimator can improve the MLE-PCP only when the prior is informative enough according to the domain defined in Equation 6-5.

Meanwhile, the test results for comparing the MLE- and the MAP-PCPs in the ‘Dissimilar Prior’ Scenario are demonstrated in Table 6.13 (below).

Table 6.13 Test Results for Comparison of MLE- and MAP-PCPs in ‘Dissimilar Prior’ Scenarios

Exp. NO.	Series Label	MLE-PCP (%)	MAP-PCP (%)	Sd. ^a (%)	Test ^b	p-value
1	MSOA_WKDY_HR_GEN	49.45	49.21	0.36	W	0.004
2	MSOA_WKDY_HR_YNG	86.00	93.11	4.66	W	0.000
3	MSOA_WKDY_HR_MDA	76.47	82.10	3.83	W	0.000
4	MSOA_WKDY_HR_EDL	82.18	88.99	4.55	W	0.000
5	MSOA_WKDY_PR_GEN	49.49	49.21	0.34	T	0.102
6	MSOA_WKDY_PR_YNG	85.73	93.13	5.19	T	0.017
7	MSOA_WKDY_PR_MDA	76.13	82.06	4.23	T	0.019
8	MSOA_WKDY_PR_EDL	81.80	88.93	5.04	T	0.018
9	MSOA_WKND_HR_GEN	49.15	48.87	0.63	W	0.007
10	MSOA_WKND_HR_YNG	87.31	93.07	3.24	W	0.000
11	MSOA_WKND_HR_MDA	77.56	82.10	2.55	W	0.000
12	MSOA_WKND_HR_EDL	83.39	89.07	2.98	W	0.000
13	MSOA_WKND_PR_GEN	49.30	49.30	0.00	W	1.000
14	MSOA_WKND_PR_YNG	76.40	93.08	29.84	W	0.050
15	MSOA_WKND_PR_MDA	69.59	82.11	21.78	W	0.050
16	MSOA_WKND_PR_EDL	73.70	89.03	26.38	W	0.050
17	OPC_WKDY_HR_GEN	49.45	49.21	0.36	W	0.004
18	OPC_WKDY_HR_YNG	86.18	93.11	4.55	W	0.000
19	OPC_WKDY_HR_MDA	76.60	82.10	3.74	W	0.000
20	OPC_WKDY_HR_EDL	82.35	88.99	4.46	W	0.000
21	OPC_WKDY_PR_GEN	49.49	49.21	0.34	T	0.103
22	OPC_WKDY_PR_YNG	85.91	93.13	5.07	T	0.017
23	OPC_WKDY_PR_MDA	76.27	82.06	4.14	T	0.019
24	OPC_WKDY_PR_EDL	81.97	88.93	4.93	T	0.018
25	OPC_WKND_HR_GEN	49.15	48.86	0.63	W	0.007
26	OPC_WKND_HR_YNG	87.47	93.07	3.13	W	0.000
27	OPC_WKND_HR_MDA	77.69	82.10	2.47	W	0.000
28	OPC_WKND_HR_EDL	83.55	89.07	2.88	W	0.000
29	OPC_WKND_PR_GEN	49.30	49.30	0.00	W	1.000
30	OPC_WKND_PR_YNG	76.45	93.08	30.00	W	0.050
31	OPC_WKND_PR_MDA	69.64	82.11	21.91	W	0.050
32	OPC_WKND_PR_EDL	73.75	89.03	26.53	W	0.050

^a Standard deviation of the difference between the MAP-PCPs and the MLE-PCPs obtained from each set of experiments in the ‘Dissimilar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

It can be implied from Table 6.13 that the MLE- and MAP-PCPs are significantly different (p -values < 0.1) in the majority of the enrichment experiments (28 sets) except for the four sets below (marked in red in the table).

- 1) MSOA_WKDY_PR_GEN (No. 5);
- 2) MSOA_WKND_PR_GEN (No. 13);
- 3) OPC_WKDY_PR_GEN (No. 21); and
- 4) OPC_WKND_PR_GEN (No. 29).

This is attributed to the fact that a large number of the priors included in the MAP estimator for the above four sets of experiments are not informative enough given Equation 6-5. The remaining 28 sets of enrichment experiments, however, confirm that an informative prior makes a difference to the PCPs resulting from using the MAP estimator compared with those from using the MLE estimator. Interestingly, however, it can be noted from the average PCP values that, whilst the MAP-PCP still outperforms the MLE-PCP in all the experiments for enriching age-based attributes in this ‘Dissimilar Prior’ Scenario, the MLE-PCP exceeds the MAP-PCP in the four sets of gender enrichment experiments given hourly choice patterns, i.e. where the estimation prior is informative yet incorrect:

- 1) MSOA_WKDY_HR_GEN (No. 1);
- 2) MSOA_WKND_HR_GEN (No. 9);
- 3) OPC_WKDY_HR_GEN (No. 17); and
- 4) OPC_WKND_HR_GEN (No. 25).

As compared to the ‘Similar Prior’ Scenario, where the MLE estimator performs no better than the MAP estimator, the outperformance of the MLE estimator in the above four sets of experiments can be explained by the incorrect estimation prior for gender enrichment. This highlights the role that can be played by the correctness of the prior in determining the enrichment performance using the MAP estimator. The indifferent performance of the two estimators in the remaining four sets of gender enrichment experiments (No. 5, 13, 21, and 29) shows that the effect of an incorrect yet uninformative prior can be neglected. In contrast, the correct estimation priors for all the enrichment experiments for age-based attributes maintain their better performance when using the MAP estimator, despite the dissimilarity of the estimation priors to the true $P(X)$ in the enrichment samples.

Straightforward visualisations for comparing the relative better or worse performance of the two estimators in the ‘Similar Prior’ and the ‘Dissimilar Prior’ Scenarios are illustrated in Figure 6.4(a) and Figure 6.4(b), respectively.

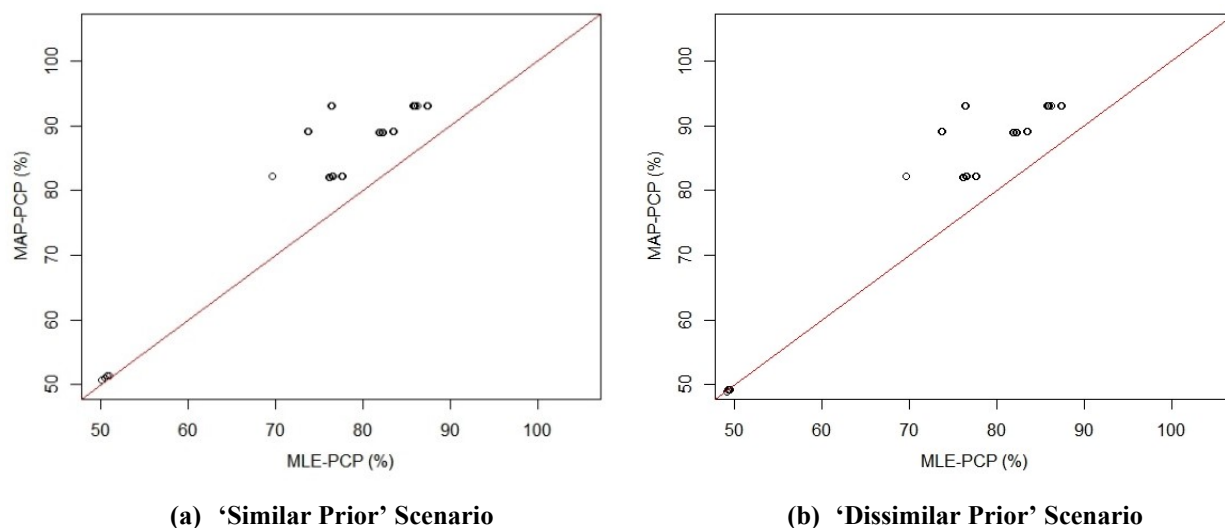


Figure 6.4 MAP-PCP vs MLE-PCP in ‘Similar Prior’ and ‘Dissimilar Prior’ Scenarios

The above analysis reveals that the relative enrichment performance using the MAP estimator of the IDCM approach, as compared with its MLE counterpart, is affected by both the informativeness and the correctness of the prior included in the MAP estimator. In particular, the MLE- and the MAP-PCP are indifferent when the prior included in the MAP estimator is uninformative as defined by the domain of an informative prior in Equation 6-5, regardless of its correctness. In contrast, the MAP estimator can improve the enrichment performance compared to the MLE estimator when, and only when, the included prior is both informative and correct. In cases where an informative but incorrect prior is employed, the enrichment performance of using the MLE estimator can be deteriorated by using the MAP estimator. Thus, the MAP estimator represents the more preferred estimation method for the IDCM approach when an informative and correct prior is available, while the MLE estimator is recommended when no prior, or only an incorrect prior, is available. The choice between the MLE and the MAP estimators is indifferent in circumstances where the available prior is uninformative for the enrichment sample, however, since the two estimators perform effectively the same.

As is summarised by the flow chart in Figure 6.5 (below), these results supplement the criteria for selecting between the two estimation methods of the IDCM approach to make better yet information-effective enrichment, based on the theoretical results derived in Section 4.2.3.6 (Figure 4.13) and the findings in the current section.

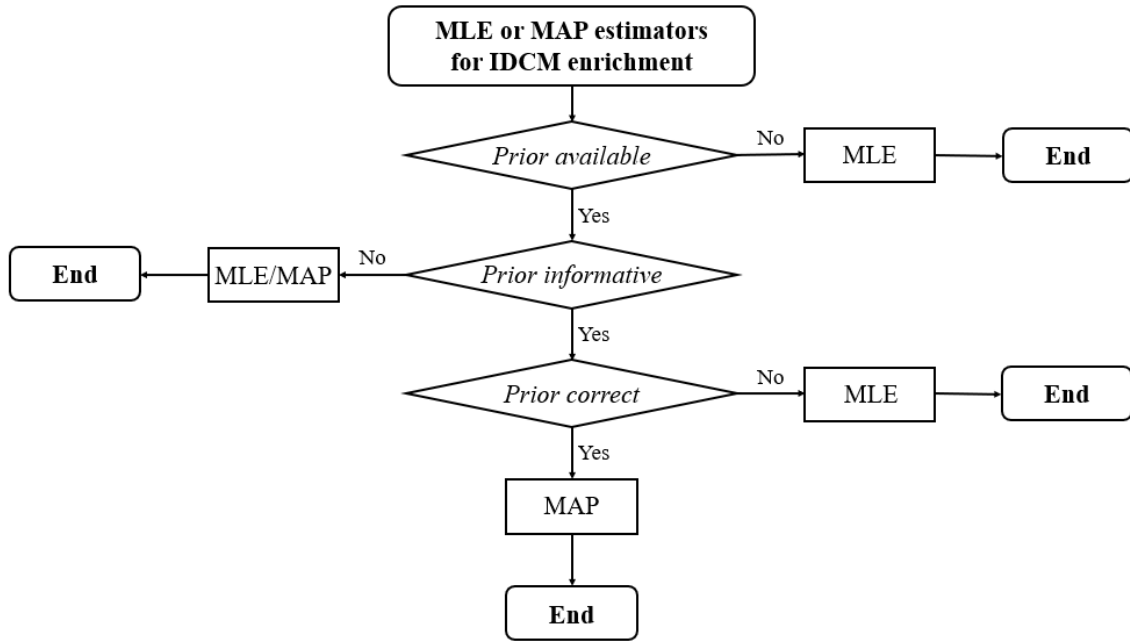


Figure 6.5 Performance-driven Criteria for Selection between MLE and MAP Estimators of IDCM
Approach given Prior Availability, Informativeness and Correctness

6.3.1.2 Level of spatial detail

In this section, the pattern of variation in the enrichment performance of the IDCM approach, based on a single estimation sample (thus the same IDCM kernel), is studied with respect to the different levels of spatial detail at which the enrichment samples are collected (or aggregated). The following paragraphs present the comparative analysis performed on the enrichment results at two levels of spatial detail using the MLE and the MAP estimators. Specifically, Table 6.14 demonstrates the test results for comparing the MLE-PCPs obtained at the MSOA level and at the aggregated OPC levels while the corresponding comparison between the MAP-PCPs is analysed later in Table 6.17.

Table 6.14 Test Results for Comparison of MLE-PCPs in MSOA- & OPC-level Enrichment

Scenario	Exp NO. (Paired)	Series Label	MLE-PCP (%)		Sd. ^a (%)	Test ^b	p-value
			MSOA	OPC			
Similar	1, 17	WKDY_HR_GEN	50.55	50.55	0.01	T	0.162
Similar	2, 18	WKDY_HR_YNG	86.00	86.18	0.11	T	0.000
Similar	3, 19	WKDY_HR_MDA	76.47	76.60	0.09	T	0.000
Similar	4, 20	WKDY_HR_EDL	82.18	82.35	0.10	T	0.000
Similar	5, 21	WKDY_PR_GEN	50.08	50.07	0.01	T	0.175
Similar	6, 22	WKDY_PR_YNG	85.73	85.91	0.13	T	0.018

Similar	7, 23	WKDY_PR_MDA	76.13	76.27	0.10	T	0.034
Similar	8, 24	WKDY_PR_EDL	81.80	81.97	0.11	T	0.043
Similar	9, 25	WKND_HR_GEN	51.01	51.02	0.00	W	1.000
Similar	10, 26	WKND_HR_YNG	87.31	87.47	0.11	W	0.000
Similar	11, 27	WKND_HR_MDA	77.56	77.69	0.08	W	0.000
Similar	12, 28	WKND_HR_EDL	83.39	83.55	0.09	W	0.000
Similar	13, 29	WKND_PR_GEN	51.02	51.03	0.00	T	0.363
Similar	14, 30	WKND_PR_YNG	76.40	76.45	0.19	W	<u>0.414</u>
Similar	15, 31	WKND_PR_MDA	69.59	69.64	0.14	W	<u>0.577</u>
Similar	16, 32	WKND_PR_EDL	73.70	73.75	0.17	W	<u>0.496</u>
Dissimilar	1, 17	WKDY_HR_GEN	49.45	49.45	0.01	T	0.162
Dissimilar	2, 18	WKDY_HR_YNG	86.00	86.18	0.11	T	0.000
Dissimilar	3, 19	WKDY_HR_MDA	76.47	76.60	0.09	T	0.000
Dissimilar	4, 20	WKDY_HR_EDL	82.18	82.35	0.10	T	0.000
Dissimilar	5, 21	WKDY_PR_GEN	49.49	49.49	0.01	T	1.000
Dissimilar	6, 22	WKDY_PR_YNG	85.73	85.91	0.13	T	0.018
Dissimilar	7, 23	WKDY_PR_MDA	76.13	76.27	0.10	T	0.034
Dissimilar	8, 24	WKDY_PR_EDL	81.80	81.97	0.11	T	0.043
Dissimilar	9, 25	WKND_HR_GEN	49.15	49.15	0.01	W	1.000
Dissimilar	10, 26	WKND_HR_YNG	87.31	87.47	0.11	W	0.000
Dissimilar	11, 27	WKND_HR_MDA	77.56	77.69	0.08	W	0.000
Dissimilar	12, 28	WKND_HR_EDL	83.39	83.55	0.09	W	0.000
Dissimilar	13, 29	WKND_PR_GEN	49.30	49.30	0.01	W	1.000
Dissimilar	14, 30	WKND_PR_YNG	76.40	76.45	0.19	W	<u>0.414</u>
Dissimilar	15, 31	WKND_PR_MDA	69.59	69.64	0.14	W	<u>0.577</u>
Dissimilar	16, 32	WKND_PR_EDL	73.70	73.75	0.17	W	<u>0.496</u>

^a Standard deviation of the difference between the MLE-PCPs obtained from each set of experiments at the MSOA and the OPC level.

^b 'T' denotes the paired *t*-test; 'W' refers to the Wilcoxon signed-rank test.

It is shown in Table 6.14 that the MLE-PCP varies with respect to the spatial detail level in different ways for the enrichment of different socio-demographic attributes. In particular, the MLE-PCP is found to be comparable in the enrichment performed at the MSOA and the OPC levels in all sets of experiments for gender, as indicated by the *p*-values that are greater than 0.1 (highlighted in red in the table). For the enrichment of age-based attributes, meanwhile, the MLE estimator is found to perform comparably at the two levels of spatial detail for only the six compared sets of enrichment experiments (*p*-values underlined in red) given weekend-period choice patterns, i.e.:

- 1) WKND_PR_YNG (No. 14 and 30);

- 2) WKND_PR_MDA (No. 15 and 31); and
- 3) WKND_PR_EDL (No. 16 and 32).

In the remaining 18 compared sets of enrichment experiments, meanwhile, the MLE estimator performed substantially better at the OPC level (p -values < 0.1) than at the MSOA level.

To interpret the above patterns, the variation in the underlying probability distribution of the enrichment sample is investigated with respect to the spatial aggregation performed so as to obtain the OPC-level enrichment sample which is originally collected at the MSOA level. This is supported by the IDCM performance theory developed in Section 4.2.3, which suggests that the enrichment performance of the IDCM approach can be mathematically linked to the properties of the corresponding quasi choice-making communication channel, itself determined by the underlying covariance structure of the enrichment sample. Specifically, these channel properties involve the inherent randomness (uncertainty) associated with the enriched attribute *per se*, measured by its entropy $H(X)$, and the explanatory power of the choice pattern in inferring the enriched attribute, which is quantified by the developed metric of enrichment efficiency (η). It is therefore essential to understand the variation pattern of the IDCM enrichment performance by recognising the effect of the performed spatial aggregation in altering the corresponding $H(X)$ and η .

In particular, the one-proportion Z-test results (p -values = 1.000) for comparing the marginal probability distribution between the MSOA- and the OPC-level enrichment samples show that $P(X)$ remains unchanged in light of the performed spatial aggregation. This can be explained by the fact that the performed spatial aggregation merely modifies the variables informing geographical details, such as the ODs and dwelling locations, by combining them to a more aggregated spatial detail level, while the socio-demographic attribute cohort which segments the respondents remains the same, given that different attribute cohorts cannot merge with each other. As a result, the entropy $H(X)$ of the enriched attribute also remains the same irrespective of the spatial aggregation. On the other hand, the market shares of choice pattern variables in the enrichment sample, i.e. $P(Y)$, are altered by the spatial aggregation because the trips (i.e. the choices of departures) made between a pair of different MSOAs aggregated within the same OPC zone would be classified as ‘staying’ (the choice to not travel) in the OPC-level enrichment sample. This further modifies the underlying covariance structure $P(X, Y)$ and the choice behaviour pattern $P(Y|X)$, resulting in the variation in the corresponding η .

Based on the above finding that $H(X)$ remains unchanged with respect to the performed spatial aggregation, the difference in the MLE-PCPs for the enrichment samples collected at different spatial levels is explored solely with regard to the variation in η . Table 6.15 presents the test results of comparing the enrichment efficiency between the ‘Similar Prior’ and the ‘Dissimilar Prior’ Scenarios for the 16 paired sets of enrichment experiments at the same spatial level.

Table 6.15 Test Results for Comparison of Enrichment Efficiency for Enrichment Samples with Same Spatial Detail Level between ‘Similar’ & ‘Dissimilar Prior’ Scenarios

Spatial Level	Exp NO. (Paired)	Series Label	η (%)		Sd. ^a (%)	Test ^b	p-value
			Similar	Dissimilar			
MSOA	1, 17	WKDY_HR_GEN	0.0499	0.0499	0.00	W	1.000
MSOA	2, 18	WKDY_HR_YNG	0.0319	0.0319	0.00	W	1.000
MSOA	3, 19	WKDY_HR_MDA	0.0549	0.0549	0.00	W	1.000
MSOA	4, 20	WKDY_HR_EDL	0.1373	0.1373	0.00	W	1.000
MSOA	5, 21	WKDY_PR_GEN	0.0476	0.0476	0.00	W	1.000
MSOA	6, 22	WKDY_PR_YNG	0.0230	0.0230	0.00	W	1.000
MSOA	7, 23	WKDY_PR_MDA	0.0649	0.0649	0.00	W	1.000
MSOA	8, 24	WKDY_PR_EDL	0.1492	0.1492	0.00	T	1.000
MSOA	9, 25	WKND_HR_GEN	0.0238	0.0238	0.00	W	1.000
MSOA	10, 26	WKND_HR_YNG	0.0792	0.0792	0.00	W	1.000
MSOA	11, 27	WKND_HR_MDA	0.0136	0.0136	0.00	W	1.000
MSOA	12, 28	WKND_HR_EDL	0.0828	0.0828	0.00	T	1.000
MSOA	13, 29	WKND_PR_GEN	0.0239	0.0239	0.00	T	1.000
MSOA	14, 30	WKND_PR_YNG	0.0695	0.0695	0.00	T	1.000
MSOA	15, 31	WKND_PR_MDA	0.0115	0.0115	0.00	W	1.000
MSOA	16, 32	WKND_PR_EDL	0.0817	0.0817	0.00	W	1.000
OPC	1, 17	WKDY_HR_GEN	0.0496	0.0496	0.00	W	1.000
OPC	2, 18	WKDY_HR_YNG	0.0308	0.0308	0.00	W	1.000
OPC	3, 19	WKDY_HR_MDA	0.0544	0.0544	0.00	W	1.000
OPC	4, 20	WKDY_HR_EDL	0.1363	0.1363	0.00	W	1.000
OPC	5, 21	WKDY_PR_GEN	0.0474	0.0474	0.00	W	1.000
OPC	6, 22	WKDY_PR_YNG	0.0222	0.0222	0.00	W	1.000
OPC	7, 23	WKDY_PR_MDA	0.0644	0.0644	0.00	W	1.000
OPC	8, 24	WKDY_PR_EDL	0.1488	0.1488	0.00	T	1.000
OPC	9, 25	WKND_HR_GEN	0.0235	0.0235	0.00	W	1.000
OPC	10, 26	WKND_HR_YNG	0.0773	0.0773	0.00	W	1.000
OPC	11, 27	WKND_HR_MDA	0.0133	0.0133	0.00	W	1.000
OPC	12, 28	WKND_HR_EDL	0.0815	0.0815	0.00	T	1.000
OPC	13, 29	WKND_PR_GEN	0.0236	0.0236	0.00	T	1.000

OPC	14, 30	WKND_PR_YNG	0.0679	0.0679	0.00	T	1.000
OPC	15, 31	WKND_PR_MDA	0.0114	0.0114	0.00	T	1.000
OPC	16, 32	WKND_PR_EDL	0.0805	0.0805	0.00	W	1.000

^a Standard deviation of the difference between the enrichment efficiency obtained from each set of experiments in the ‘Dissimilar Prior’ and the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

It can be seen from Table 6.15 that the enrichment samples collected at the same spatial detail level (MSOA- or OPC-level) in the ‘Similar Prior’ Scenario retain the same level of enrichment efficiency as the corresponding enrichment samples in the ‘Dissimilar Prior’ Scenario. This is revealed by the *p*-values that are all equal to 1.000. This finding is reasonable since the variation of the probability distribution underlying the enrichment samples is irrelevant to the quality of the prior, which is effectively determined by the probability distribution in the estimation sample. The following discussion with respect to the variation in η can therefore be conducted without mentioning the scenarios concerning the quality of the prior.

Inspecting Table 6.16, which provides the comparison of the enrichment efficiency between the MSOA- and the OPC-level enrichment samples, indicates that the η varies with respect to the spatial detail level, following a similar pattern to the variation of the MLE-PCP as shown previously in Table 6.14 (Page 208).

Table 6.16 Test Results for Comparison of Enrichment Efficiency between Enrichment Samples Collected at MSOA & OPC Level

Exp NO. (Paired)	Series Label	η (%)		Sd. ^a (%)	Test ^b	<i>p</i> -value
		MSOA	OPC			
1, 17	WKDY_HR_GEN	0.0499	0.0496	0.0011	W	0.440
2, 18	WKDY_HR_YNG	0.0319	0.0308	0.0011	W	0.000
3, 19	WKDY_HR_MDA	0.0549	0.0544	0.0014	W	0.002
4, 20	WKDY_HR_EDL	0.1373	0.1363	0.0017	W	0.002
5, 21	WKDY_PR_GEN	0.0476	0.0474	0.0008	W	0.600
6, 22	WKDY_PR_YNG	0.0230	0.0222	0.0008	W	0.075
7, 23	WKDY_PR_MDA	0.0649	0.0644	0.0009	W	0.173
8, 24	WKDY_PR_EDL	0.1492	0.1488	0.0007	T	0.232
9, 25	WKND_HR_GEN	0.0238	0.0235	0.0005	W	1.000
10, 26	WKND_HR_YNG	0.0792	0.0773	0.0023	W	0.000
11, 27	WKND_HR_MDA	0.0136	0.0133	0.0003	W	0.004
12, 28	WKND_HR_EDL	0.0828	0.0815	0.0016	T	0.001
13, 29	WKND_PR_GEN	0.0239	0.0236	0.0004	T	0.484
14, 30	WKND_PR_YNG	0.0695	0.0679	0.0020	T	0.100

15, 31	WKND_PR_MDA	0.0115	0.0114	0.0003	W	<u>0.345</u>
16, 32	WKND_PR_EDL	0.0817	0.0805	0.0016	W	<u>0.116</u>

^a Standard deviation of the difference between the enrichment efficiency obtained from each set of experiments at the MSOA and the OPC spatial detail levels.

^b 'T' denotes the paired *t*-test; 'W' refers to the Wilcoxon signed-rank test.

Specifically, the variation in η due to the spatial aggregation is negligible in all sets of enrichment experiments for gender (shown by the *p*-values that are greater than 0.1, and marked in red in the table). In terms of the age-based attributes, the change in η only remains vital in the enrichment of the 'Middle-aged' and the 'Elderly' groups given period choices (*p*-values underlined in red) yet are found to vary significantly for all the remaining sets of enrichment experiments (*p*-values < 0.1).

Since the variation in the probability distribution of the enrichment samples between different spatial detail levels is fundamentally driven by the change in $P(Y)$ which, in the sense of spatial aggregation, is determined by lengths of trips, the aforementioned variation patterns of η can be explained by the variation in the trip length given different socio-demographic attributes. Specifically, for the two POVs of gender (viz. male and female), the length of trips remains consistently long (or short) on both weekdays and weekends, thus leading to the indifferent η at the two spatial levels for the enrichment of gender. In contrast, the trip length categorised by age groups is only comparable for period choice observations, while those observed by hours can be distinguished significantly for different age groups.

Comparing the respective variation patterns of the MLE-PCP and η between the two spatial detail levels reveals that the MLE-PCPs differ significantly at the two spatial detail levels in all the compared sets of experiments where a substantial difference in η is also observed between the MSOA- and the OPC-level enrichment samples. This indicates that the MLE-PCP can vary between enrichment samples that are aggregated at different spatial levels when the enrichment efficiency of the sample is changed due to the spatial aggregation. Conversely, however, it is found that the significant difference in the MLE-PCP obtained at the two spatial detail levels may also result from the enrichment samples with similar η , specifically in the enrichment of the 'Middle-aged' and the 'Elderly' groups given weekday-period choices in both scenarios (greyed out in Table 6.14 and Table 6.15), i.e.:

- 1) WKDY_PR_MDA (No. 7 and No. 23); and
- 2) WKDY_PR_EDL (No. 8 and No. 24).

In the two pairs of compared sets of experiments, the variation in η is trivial at the MSOA and the OPC levels (p -values > 0.1) while that in the MLE-PCP is significant (p -values < 0.1). This further suggests that the MLE-PCP is also likely to vary for the enrichment of samples where the similarity in η is insignificant with relatively small p -values. Together, this demonstrates that the changes in the η of the enrichment sample due to the spatial aggregation are the reason contributing to the variation in the MLE-PCP conducted at different spatial levels.

Another finding from the comparative analysis for the MLE-PCPs at different spatial detail levels is that the MLE-PCPs obtained when enriching the OPC-level sample are no worse than those obtained at the MSOA level. Given that the estimation samples are extracted from the LTDS, which was originally collected at the OPC level, the outperformance of the MLE-PCP in the OPC-level enrichment suggests that, given the same IDCM kernel, the MLE estimator performs better for the enrichment sample that is collected at a closer spatial detail level to the estimation sample as this estimator typically captures more similar behaviour patterns given the enriched attribute, i.e. $P(Y|X)$.

On the other hand, the IDCM enrichment performance using the MAP estimator is found to be insensitive to the performed spatial aggregation in the current context, indicated by insignificant statistics (p -values > 0.1) in the MAP-PCP comparison tests conducted at the MSOA and the OPC levels. The full results of this are as reported in Table 6.17.

Table 6.17 Test Results for Comparison of MAP-PCPs in MSOA- & OPC-level Enrichment

Scenario	Exp NO. (Paired)	Series Label	MAP-PCP (%)		Sd. ^a (%)	Test ^b	p -value
			MSOA	OPC			
Similar	1, 17	WKDY_HR_GEN	50.98	50.98	0.01	W	1.000
Similar	2, 18	WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
Similar	3, 19	WKDY_HR_MDA	82.10	82.10	0.00	W	1.000
Similar	4, 20	WKDY_HR_EDL	88.99	88.99	0.00	W	1.000
Similar	5, 21	WKDY_PR_GEN	50.60	50.60	0.01	T	0.363
Similar	6, 22	WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
Similar	7, 23	WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
Similar	8, 24	WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
Similar	9, 25	WKND_HR_GEN	51.36	51.37	0.00	T	0.328
Similar	10, 26	WKND_HR_YNG	93.07	93.07	0.00	W	1.000
Similar	11, 27	WKND_HR_MDA	82.13	82.13	0.00	W	1.000
Similar	12, 28	WKND_HR_EDL	89.07	89.07	0.00	W	1.000
Similar	13, 29	WKND_PR_GEN	51.40	51.40	0.00	T	1.000

Similar	14, 30	WKND_PR_YNG	93.08	93.08	0.00	T	1.000
Similar	15, 31	WKND_PR_MDA	82.11	82.11	0.00	T	1.000
Similar	16, 32	WKND_PR_EDL	89.03	89.03	0.00	T	1.000
Dissimilar	1, 17	WKDY_HR_GEN	49.21	49.21	0.01	W	1.000
Dissimilar	2, 18	WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
Dissimilar	3, 19	WKDY_HR_MDA	82.10	82.10	0.00	W	1.000
Dissimilar	4, 20	WKDY_HR_EDL	88.99	88.99	0.00	W	1.000
Dissimilar	5, 21	WKDY_PR_GEN	49.21	49.21	0.01	T	0.363
Dissimilar	6, 22	WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
Dissimilar	7, 23	WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
Dissimilar	8, 24	WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
Dissimilar	9, 25	WKND_HR_GEN	48.87	48.86	0.01	T	1.000
Dissimilar	10, 26	WKND_HR_YNG	93.07	93.07	0.00	W	1.000
Dissimilar	11, 27	WKND_HR_MDA	82.10	82.10	0.00	W	1.000
Dissimilar	12, 28	WKND_HR_EDL	89.07	89.07	0.00	W	1.000
Dissimilar	13, 29	WKND_PR_GEN	49.30	49.30	0.01	W	1.000
Dissimilar	14, 30	WKND_PR_YNG	93.08	93.08	0.00	T	1.000
Dissimilar	15, 31	WKND_PR_MDA	82.11	82.11	0.00	T	1.000
Dissimilar	16, 32	WKND_PR_EDL	89.03	89.03	0.00	T	1.000

^a Standard deviation of the difference between the MAP-PCPs obtained from each set of experiments at the MSOA and the OPC level.

^b 'T' denotes the paired *t*-test; 'W' refers to the Wilcoxon signed-rank test.

As can be observed from Table 6.17, the comparable MAP-PCPs in the MSOA- and the OPC-level enrichment in all compared sets of experiments imply that the IDCM enrichment performance using the MAP estimator is not affected by the spatial aggregation. This can be explained by the use of the same prior employed in the MAP estimators for different spatial levels. Specifically, regardless of the correctness of the prior, the MAP estimator using a same informative prior can perform equivalently in enriching the socio-demographic attributes for the enrichment samples that are collected or aggregated at different spatial detail levels. As for the uninformative priors, it is implied from Table 6.16 that the η remains stable for the two sets of enrichment experiments where uninformative priors are involved, i.e. WKDY_PR_GEN (No. 5 and No. 21). This indicates that the variation in the MAP-PCP with respect to the change in η in the presence of an uninformative prior is not captured in the current analysis. It can be derived accordingly, however, that the variation pattern of this MAP-PCP is analogous to that of the MLE-PCP, because the IDCM enrichment performance theory (Section 4.2.3) suggests that the enrichment performance using the MAP estimator collapses to the MLE-PCP when the

prior is uninformative, given Equation 6-5. And the MAP-PCP obtained from using an uninformative prior should therefore retain the same property as the corresponding MLE-PCPs.

In summary, the above analysis concerning the enrichment of the socio-demographic attribute for MSOA- and OPC-level enrichment samples indicates that the enrichment performance of the IDCM approach using the MLE and the MAP estimators responds in different ways to the difference in the spatial detail levels at which the enrichment samples are collected or aggregated. Specifically, the enrichment performance using the MLE estimator of the IDCM approach varies with respect to the change or the difference in the enrichment efficiency of the enrichment samples caused by the spatial aggregation, or simply because the enrichment samples are collected at different spatial detail levels. Moreover, the performance of the MLE estimator can be improved as the spatial detail level of the estimation sample gets closer to that of the enrichment sample, and *vice versa*. Conversely, the impact of spatial aggregation on the enrichment performance using the MAP estimator is insignificant when the included prior can be defined as informative according to Equation 6-5. Since the $P(X)$ of the enrichment samples remains unchanged in the presence of the spatial aggregation, the use of a similarly informative prior for the MAP estimator can result in the same level of performance in enriching samples aggregated at different spatial levels. The MAP-PCP is argued to vary in a similar manner to the MLE-PCP when an uninformative prior is employed because the IDCM performance theory suggests the indifferent performance of the MLE estimator and the MAP estimator with an uninformative prior.

The above conclusions can be extended to any form of aggregation that can modify the probability distribution of the observed choice patterns $P(Y)$ given which the socio-demographic attributes are enriched using the IDCM approach. Consider the aggregation of the behaviour pattern Y of travel mode choice, from the more detailed level which involves the choice of whether to travel by bus or not (Y_1), to the relatively more aggregated level that captures the choice of whether to travel by public transport (Y_2), thus incorporating modes such as rail, subway and light rail. The respective covariance structures of Y_1 and Y_2 with a socio-demographic attribute X (e.g. gender), can be different, thus possibly resulting in the difference in the corresponding dependence between the choice pattern and the gender, i.e. $P(Y|X)$, as well as in the enrichment efficiency (η). In this case, the MLE-PCP would likely vary if the η of the enrichment sample capturing the information about the joint probability distribution of X and Y , viz. $P(X,Y)$, is significantly changed (or not substantially similar) due to the

aforementioned aggregation. Whether the variation is to an improved or a deteriorated level depends on whether the enrichment efficiency is changed to be closer to or further from that of the estimation sample upon which the employed IDCM kernel is estimated. The MAP-PCP would also vary in the same manner as the MLE-PCP if the prior included in the MAP estimator is uninformative. Otherwise, the MAP-PCP can remain stable irrespective of the performed aggregation and the correctness of the prior. It is also argued that, in addition to the above example, the conclusions drawn in this section can be generalised to any context of behaviour patterns, other than travel-related choice behaviours where the behaviour is driven, or can be affected, by the socio-demographic attributes that require enrichment.

6.3.2 Capability of IDCM performance theory to estimate IDCM enrichment performance given data conditions

The enrichment performance of the IDCM approach can vary in the currently explored MND enrichment application due to the variation in the conditions of the data. As is discussed in Section 6.3.1 above, such variation in data conditions can arise in either the estimation sample used for generating the IDCM kernel, or the enrichment sample that requires the enrichment of socio-demographic attributes, reflected respectively by the quality of the estimation prior or the difference in the spatial detail level at which the enrichment sample is collected. Essentially, the former quality of the estimation prior represents the informativeness and the correctness of the estimation prior, i.e. $P_{est}(X)$ in the estimation sample, as compared to the true distribution $P_{erc}(X)$ in the enrichment sample, while the latter spatial-level variation effectively involves the difference in $P(Y)$ between the enrichment samples collected at different spatial detail levels. Identifying the possible effect of the variation in the data conditions can have on the enrichment performance of the IDCM approach, the current section evaluates the applicability of the IDCM socio-demographic enrichment framework by investigating the capability of the IDCM performance theory to estimate (forecast) the enrichment performance of the IDCM approach using the MLE and the MAP estimators in light of the variation in data conditions in the current MND enrichment context.

Specifically, Table 6.18 and Table 6.19 (below) present the test results for the comparison between the MLE-PCP obtained from the actual enrichment (Actual MLE-PCP) and that calculated using the IDCM performance theory (Expected MLE-PCP) in the ‘Similar Prior’ and the ‘Dissimilar Prior’ Scenarios, respectively.

Table 6.18 Test Results for Comparison of Actual and Expected MLE-PCPs in ‘Similar’ Scenarios

Exp. NO.	Series Label	MLE-PCP (%)		Sd. ^a (%)	Test ^b	p-value	Max(Diff) ^c (%)
		Actual	Expected				
1	MSOA_WKDY_HR_GEN	50.55	50.51	1.06	T	0.863	2.31
2	MSOA_WKDY_HR_YNG	86.00	86.10	0.28	T	0.062	0.80
3	MSOA_WKDY_HR_MDA	76.47	76.61	0.85	T	0.408	2.14
4	MSOA_WKDY_HR_EDL	82.18	82.35	0.62	T	0.190	1.68
5	MSOA_WKDY_PR_GEN	50.08	50.90	0.70	T	0.034	1.68
6	MSOA_WKDY_PR_YNG	85.73	85.81	0.21	T	0.361	0.29
7	MSOA_WKDY_PR_MDA	76.13	76.25	0.91	T	0.778	1.89
8	MSOA_WKDY_PR_EDL	81.80	82.01	0.77	T	0.538	1.75
9	MSOA_WKND_HR_GEN	51.01	51.00	0.45	W	0.951	0.81
10	MSOA_WKND_HR_YNG	87.31	87.38	0.35	W	0.520	0.84
11	MSOA_WKND_HR_MDA	77.56	77.80	0.30	W	0.001	1.06
12	MSOA_WKND_HR_EDL	83.39	83.62	0.34	W	0.004	1.19
13	MSOA_WKND_PR_GEN	51.02	51.08	0.50	T	0.782	0.95
14	MSOA_WKND_PR_YNG	76.40	76.57	0.24	W	0.173	0.45
15	MSOA_WKND_PR_MDA	69.59	69.61	0.28	W	0.345	0.53
16	MSOA_WKND_PR_EDL	73.70	73.66	0.42	W	0.345	0.89
17	OPC_WKDY_HR_GEN	50.55	50.52	1.03	T	0.885	2.26
18	OPC_WKDY_HR_YNG	86.18	86.29	0.27	T	0.070	0.78
19	OPC_WKDY_HR_MDA	76.60	76.75	0.83	T	0.403	2.08
20	OPC_WKDY_HR_EDL	82.35	82.52	0.61	T	0.182	1.64
21	OPC_WKDY_PR_GEN	50.07	50.88	0.69	T	0.035	1.64
22	OPC_WKDY_PR_YNG	85.91	85.99	0.21	T	0.388	0.29
23	OPC_WKDY_PR_MDA	76.27	76.38	0.88	T	0.766	1.84
24	OPC_WKDY_PR_EDL	81.97	82.18	0.75	T	0.533	1.71
25	OPC_WKND_HR_GEN	51.02	51.01	0.44	W	0.951	0.79
26	OPC_WKND_HR_YNG	87.47	87.54	0.34	W	0.529	0.82
27	OPC_WKND_HR_MDA	77.69	77.92	0.29	W	0.001	1.04
28	OPC_WKND_HR_EDL	83.55	83.77	0.33	W	0.003	1.17
29	OPC_WKND_PR_GEN	51.03	51.09	0.49	T	0.783	0.93
30	OPC_WKND_PR_YNG	76.45	76.62	0.23	W	0.173	0.44
31	OPC_WKND_PR_MDA	69.64	69.65	0.28	W	0.345	0.52
32	OPC_WKND_PR_EDL	73.75	73.72	0.41	W	0.345	0.87

^a Standard deviation of the difference between the expected and actual MLE-PCPs obtained from each set of experiments in the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

^c Max(Diff) represents the maximum absolute value of the difference between the actual and expected PCPs in each set of enrichment experiments.

Inspecting Table 6.18 suggests that, in the ‘Similar Prior’ Scenario, the difference between the actual and the expected MLE-PCPs is statistically insignificant (p -values > 0.1) in 24 sets of enrichment experiments. In the remaining eight sets of enrichment experiments (listed below), however, the prediction using the expected MLE-PCPs is found to be less statistically robust, indicated by the relatively small p -values, which are lower than 0.1 (marked in red in the table).

- 1) MSOA_WKDY_HR_YNG (No. 2);
- 2) MSOA_WKDY_PR_GEN (No. 5);
- 3) MSOA_WKND_HR_MDA (No. 11);
- 4) MSOA_WKND_HR_EDL (No. 12);
- 5) OPC_WKDY_HR_YNG (No. 18);
- 6) OPC_WKDY_PR_GEN (No. 21);
- 7) OPC_WKND_HR_MDA (No. 27); and
- 8) OPC_WKND_HR_EDL (No. 28).

The discrepancy between the actual MLE-PCP and their predicted values in the above experiment sets can be explained by the assumption upon which the IDCM performance theory is developed: the dependence of the observed choice behaviour pattern on the enriched socio-demographic attribute, i.e. $P(Y|X)$, is similar in samples extracted from different data sources yet with the same spatial and temporal coverage (i.e. the estimation and the enrichment samples in the current empirical application). Building on this assumption, Section 4.2.3 explains in detail how the constant recalibration can be applied in a reversed manner to reproduce the $P(Y|X)$ for the enrichment sample by modifying the constant term in the IDCM kernel estimated upon the estimation sample. The IDCM performance theory then employs this recalibrated DCM with the recalibrated constant $\hat{\beta}_0$ to approximate the $P(X)$ and hence $P(X, Y)$ of the enrichment sample, given the observed $P(Y)$ in the enrichment sample. Recall Equation 4-75 which suggests that the MLE-PCP can be expressed as the summation of a combination of two joint probabilities:

$$PCP_{MLE} = p(x_0, y_0) + p(x_1, y_1). \quad (6-6)$$

The capability of the IDCM performance theory to estimate the MLE-PCP hence involves the capability of $\hat{\beta}_0$ to express the $P(X, Y)$ in the enrichment sample, which is effectively determined by the respective similarity of $P(Y|X)$ and $P(X)$ in the estimation and the enrichment samples, given that:

$$P(X, Y) = P(Y|X)P(X). \quad (6-7)$$

In the ‘Similar Prior’ Scenario (Table 6.18), $P(X)$ is similar in the estimation and the enrichment samples given the experimental setting. The difference between the actual and expected MLE-PCPs is therefore mainly caused by the disparity in $P(Y|X)$ between the estimation and the enrichment samples. This is support by the one-proportion Z -test results which lead to p -values that are significantly smaller than 0.1 for comparing the $P(Y|X)$ in the estimation sample and the corresponding enrichment sample for the above-listed eight sets of experiments where the actual and the expected MLE-PCPs are statistically different.

This significant difference in the $P(Y|X)$ between the estimation and the enrichment samples, and thus the reduced prediction accuracy for MLE-PCPs, contradicts the underlying assumption made in the IDCM performance theory. There may be two reasons for this.

- 1) The random extraction of data from the LTDS to deliver an estimation sample with a similar prior to that of the enrichment MND sample unintentionally alters the probability distribution of choice patterns given the enriched socio-demographic attributes, which interferes with the original behaviour patterns underlying the data.
- 2) The difference in mobile phone usage behaviour patterns when travelling. This is related to the respective collection process and sample size of the LTDS and the MND data. Specifically, the LTDS is a travel dairy survey conducted purposely to obtain information about the travel-related behaviour patterns of respondents with a small sample size. This may return a relatively accurate representation of behaviour but may possibly lack the ability to represent the entire population of the study area. In contrast, the MND data is collected passively from mobile phone users when making/receiving calls or sending/receiving text messages. Although a vast majority of records can be collected, and the mobile phone ownership rate is considerably high (Poushter, 2016), the travel behaviour pattern revealed in MND is highly related to the way mobile phones are used when travelling. For instance, the elderly may choose to travel without a mobile phone, thus indicating a higher dwelling rate. Another example can be reflected in the less mobile phone usage when driving as compared with that on bus or train.

The corresponding test results for the comparison between the actual and expected MLE-PCPs in the ‘Dissimilar Prior’ Scenario are as demonstrated in Table 6.19 (below).

Table 6.19 Test Results for Comparison of Actual and Expected MLE-PCPs in ‘Dissimilar’ Scenarios

Exp. NO.	Series Label	MLE-PCP (%)		Sd. ^a (%)	Test ^b	p-value	Max(Diff) ^c (%)
		Actual	Expected				
1	MSOA_WKDY_HR_GEN	49.45	49.49	1.05	T	0.869	2.31
2	MSOA_WKDY_HR_YNG	86.00	86.10	0.28	T	0.066	0.80
3	MSOA_WKDY_HR_MDA	76.47	76.62	0.85	T	0.387	2.14
4	MSOA_WKDY_HR_EDL	82.18	82.36	0.62	T	0.183	1.68
5	MSOA_WKDY_PR_GEN	49.49	50.04	0.97	T	0.222	1.68
6	MSOA_WKDY_PR_YNG	85.73	85.81	0.21	T	0.361	0.29
7	MSOA_WKDY_PR_MDA	76.13	76.25	0.91	T	0.778	1.89
8	MSOA_WKDY_PR_EDL	81.80	82.01	0.77	T	0.534	1.75
9	MSOA_WKND_HR_GEN	49.15	49.29	0.42	W	0.140	0.81
10	MSOA_WKND_HR_YNG	87.31	87.37	0.34	W	0.539	0.78
11	MSOA_WKND_HR_MDA	77.56	77.78	0.30	W	0.002	1.06
12	MSOA_WKND_HR_EDL	83.39	83.63	0.32	W	0.002	1.20
13	MSOA_WKND_PR_GEN	49.30	49.57	0.42	W	0.249	0.95
14	MSOA_WKND_PR_YNG	76.40	76.57	0.23	W	0.172	0.43
15	MSOA_WKND_PR_MDA	69.59	69.61	0.29	W	0.345	0.53
16	MSOA_WKND_PR_EDL	73.70	73.66	0.42	W	0.345	0.89
17	OPC_WKDY_HR_GEN	49.45	49.47	1.03	T	0.886	2.26
18	OPC_WKDY_HR_YNG	86.18	86.29	0.27	T	0.073	0.78
19	OPC_WKDY_HR_MDA	76.60	76.75	0.83	T	0.388	2.08
20	OPC_WKDY_HR_EDL	82.35	82.52	0.61	T	0.175	1.65
21	OPC_WKDY_PR_GEN	49.49	50.02	0.96	T	0.230	1.64
22	OPC_WKDY_PR_YNG	85.91	85.99	0.21	T	0.388	0.29
23	OPC_WKDY_PR_MDA	76.27	76.38	0.88	T	0.767	1.84
24	OPC_WKDY_PR_EDL	81.97	82.18	0.75	T	0.529	1.71
25	OPC_WKND_HR_GEN	49.15	49.27	0.41	W	0.136	0.79
26	OPC_WKND_HR_YNG	87.47	87.54	0.33	W	0.558	0.77
27	OPC_WKND_HR_MDA	77.69	77.90	0.29	W	0.002	1.03
28	OPC_WKND_HR_EDL	83.55	83.78	0.32	W	0.002	1.17
29	OPC_WKND_PR_GEN	49.30	49.55	0.41	W	0.249	0.93
30	OPC_WKND_PR_YNG	76.45	76.62	0.23	W	0.172	0.42
31	OPC_WKND_PR_MDA	69.64	69.65	0.28	W	0.345	0.52
32	OPC_WKND_PR_EDL	73.75	73.72	0.41	W	0.344	0.87

^a Standard deviation of the difference between the expected and actual MLE-PCPs obtained from each set of experiments in the ‘Dissimilar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

^c Max(Diff) represents the maximum absolute value of the difference between the actual and expected PCPs in each set of enrichment experiments.

For the ‘Dissimilar Prior’ Scenario, Table 6.19 reveals that there are also six sets of enrichment experiments in which the difference between the actual and the expected MLE-PCP is statistically significant with 90% confidence. These (highlighted in red in the table) include:

- 1) MSOA_WKDY_HR_YNG (No. 2);
- 2) MSOA_WKND_HR_MDA (No. 11);
- 3) MSOA_WKND_HR_EDL (No. 12);
- 4) OPC_WKDY_HR_YNG (No. 18);
- 5) OPC_WKND_HR_MDA (No. 27); and
- 6) OPC_WKND_HR_EDL (No. 28).

In terms of the ‘Dissimilar Prior’ Scenario (Table 6.19), all the six listed sets of enrichment experiments demonstrating statistically different actual and expected MLE-PCPs are also observed to account for the eight exceptions in the ‘Similar Prior’ Scenarios. This suggests that the difference in $P(Y|X)$ between the estimation and the enrichment samples can hamper the ability of the IDCM performance theory to predict the MLE-PCP, regardless of the similarity of the prior. Interestingly, there are two sets of enrichment experiments for which the difference between the actual and the expected MLE-PCPs is found to be statistically significant in the ‘Similar Prior’ Scenario but insignificant in the ‘Dissimilar Prior’ Scenario. These are the two sets of experiments for the enrichment of gender given weekday-period choices, i.e.:

- 1) MSOA_WKDY_PR_GEN (No.5); and
- 2) OPC_WKDY_PR_GEN (No. 21).

Examining all the experiments in these two sets reveals that the differences in $P(Y|X)$ and $P(X)$ coincidentally compensate the errors caused by each other so as jointly to produce a similar approximation of $P(X, Y)$ to the enrichment sample, hence resulting in a comparable prediction for the MLE-PCP.

As is discussed above, a statistically significant difference is noted between the actual and the expected MLE-PCP in a considerable number of experiment sets (eight sets for the ‘Similar Prior’ Scenario and six sets for the ‘Dissimilar Prior’ Scenario) based on the results of either the paired t -test or the Wilcoxon signed-rank test. Observing the rightmost column in Table 6.18 and Table 6.19, however, demonstrates that the maximum absolute value of the difference between the actual MLE-PCPs and their expected values in these experiments is smaller than 2%, which can be considered as manageably small errors in application contexts. Although the

results of the statistical test appear discouraging at first sight, closer inspection supports the overall capability of the IDCM performance theory to estimate the enrichment performance of the IDCM approach using the MLE estimator in light of different data conditions (either the quality of the estimation prior or the probability distribution underlying the choice patterns in the enrichment sample).

Figure 6.6 provides a direct visualisation of the comparison between the actual and the expected MLE-PCPs for all the enrichment experiments in ‘Similar Prior’ and ‘Dissimilar Prior’ Scenarios respectively, showing the successful prediction of the IDCM performance theory, with the data lying substantially close to or on the 45° line.

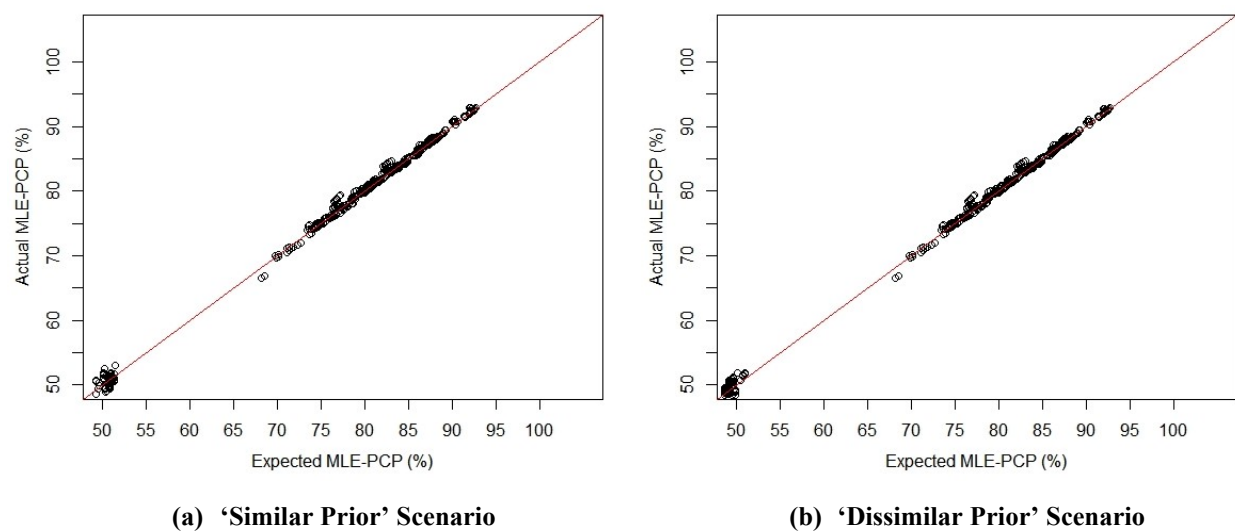


Figure 6.6 Actual MLE-PCP vs Expected MLE-PCP in ‘Similar’ & ‘Dissimilar Prior’ Scenarios

Similarly, the test results for the comparative analysis conducted between the actual and the expected MAP-PCPs in the ‘Similar Prior’ and ‘Dissimilar Prior’ Scenarios are provided in Table 6.20 and Table 6.21 (below), respectively.

Table 6.20 Test Results for Comparison of Actual and Expected MAP-PCPs in ‘Similar’ Scenarios

Exp. NO.	Series Label	MAP-PCP (%)		Sd. ^a (%)	Test ^b	p-value	Max(Diff) ^c (%)
		Actual	Expected				
1	MSOA_WKDY_HR_GEN	50.98	51.37	0.65	W	0.012	2.31
2	MSOA_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000	0.00
3	MSOA_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000	0.00
4	MSOA_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000	0.00
5	MSOA_WKDY_PR_GEN	50.60	51.41	0.94	T	0.091	2.10
6	MSOA_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000	0.00
7	MSOA_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000	0.00
8	MSOA_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000	0.00

9	MSOA_WKND_HR_GEN	51.36	51.50	0.25	T	0.015	0.76
10	MSOA_WKND_HR_YNG	93.07	93.07	0.00	W	1.000	0.00
11	MSOA_WKND_HR_MDA	82.13	82.13	0.00	W	1.000	0.00
12	MSOA_WKND_HR_EDL	89.07	89.07	0.00	W	1.000	0.00
13	MSOA_WKND_PR_GEN	51.40	51.45	0.13	T	0.363	0.31
14	MSOA_WKND_PR_YNG	93.08	93.08	0.00	T	1.000	0.00
15	MSOA_WKND_PR_MDA	82.11	82.11	0.00	T	1.000	0.00
16	MSOA_WKND_PR_EDL	89.03	89.03	0.00	T	1.000	0.00
17	OPC_WKDY_HR_GEN	50.98	51.36	0.64	W	0.012	2.26
18	OPC_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000	0.00
19	OPC_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000	0.00
20	OPC_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000	0.00
21	OPC_WKDY_PR_GEN	50.60	51.40	0.94	T	0.092	2.11
22	OPC_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000	0.00
23	OPC_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000	0.00
24	OPC_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000	0.00
25	OPC_WKND_HR_GEN	51.37	51.49	0.24	T	0.016	0.74
26	OPC_WKND_HR_YNG	93.07	93.07	0.00	W	1.000	0.00
27	OPC_WKND_HR_MDA	82.13	82.13	0.00	W	1.000	0.00
28	OPC_WKND_HR_EDL	89.07	89.07	0.00	W	1.000	0.00
29	OPC_WKND_PR_GEN	51.40	51.45	0.12	T	0.363	0.30
30	OPC_WKND_PR_YNG	93.08	93.08	0.00	T	1.000	0.00
31	OPC_WKND_PR_MDA	82.11	82.11	0.00	T	1.000	0.00
32	OPC_WKND_PR_EDL	89.03	89.03	0.00	T	1.000	0.00

^a Standard deviation of the difference between the expected and actual MAP-PCPs obtained from each set of experiments in the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

^c Max(Diff) represents the maximum absolute value of the difference between the actual and expected PCPs in each set of enrichment experiments.

Table 6.21 Test Results for Comparison of Actual and Expected MAP-PCPs in ‘Dissimilar’ Scenarios

Exp. NO.	Series Label	MAP-PCP (%)		Sd. ^a (%)	Test ^b	<i>p</i> -value	Max(Diff) ^c (%)
		Actual	Expected				
1	MSOA_WKDY_HR_GEN	49.21	50.31	0.61	W	0.000	2.06
2	MSOA_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000	0.00
3	MSOA_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000	0.00
4	MSOA_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000	0.00
5	MSOA_WKDY_PR_GEN	49.21	50.32	0.67	W	0.050	1.76
6	MSOA_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000	0.00
7	MSOA_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000	0.00

8	MSOA_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000	0.00
9	MSOA_WKND_HR_GEN	48.87	50.03	1.23	W	0.000	3.19
10	MSOA_WKND_HR_YNG	93.07	93.07	0.00	W	1.000	0.00
11	MSOA_WKND_HR_MDA	82.10	82.12	0.11	W	0.317	0.53
12	MSOA_WKND_HR_EDL	89.07	89.07	0.00	W	1.000	0.00
13	MSOA_WKND_PR_GEN	49.30	49.46	0.26	W	0.344	0.35
14	MSOA_WKND_PR_YNG	93.08	93.08	0.00	T	1.000	0.00
15	MSOA_WKND_PR_MDA	82.11	82.11	0.00	T	1.000	0.00
16	MSOA_WKND_PR_EDL	89.03	89.03	0.00	T	1.000	0.00
17	OPC_WKDY_HR_GEN	49.21	50.29	0.61	W	0.000	2.06
18	OPC_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000	0.00
19	OPC_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000	0.00
20	OPC_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000	0.00
21	OPC_WKDY_PR_GEN	49.21	50.31	0.67	T	0.010	1.76
22	OPC_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000	0.00
23	OPC_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000	0.00
24	OPC_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000	0.00
25	OPC_WKND_HR_GEN	48.86	50.02	1.23	W	0.000	3.19
26	OPC_WKND_HR_YNG	93.07	93.07	0.00	W	1.000	0.00
27	OPC_WKND_HR_MDA	82.10	82.12	0.11	W	0.317	0.52
28	OPC_WKND_HR_EDL	89.07	89.07	0.00	W	1.000	0.00
29	OPC_WKND_PR_GEN	49.30	49.45	0.25	W	0.344	0.34
30	OPC_WKND_PR_YNG	93.08	93.08	0.00	T	1.000	0.00
31	OPC_WKND_PR_MDA	82.11	82.11	0.00	T	1.000	0.00
32	OPC_WKND_PR_EDL	89.03	89.03	0.00	T	1.000	0.00

^a Standard deviation of the difference between the expected and actual MAP-PCPs obtained from each set of experiments in the ‘Dissimilar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

^c Max(Diff) represents the maximum absolute value of the difference between the actual and expected PCPs in each set of enrichment experiments.

It can be found from Table 6.20 and Table 6.21 that, in both the ‘Similar Prior’ and ‘Dissimilar Prior’ Scenarios, the prediction of the actual MAP-PCP with the expected values derived from the IDCM performance theory is statistically successful (p -values > 0.1) in 26 sets of enrichment experiments, including all 24 sets of experiments for the enrichment of the three age-based attributes. Strikingly, all of the remaining six sets of experiments where the expected MAP-PCPs differ significantly from the corresponding actual MAP-PCPs (p -values < 0.1 and marked in red in the table) relate the experiments for the enrichment of gender, i.e.:

- 1) MSOA_WKDY_HR_GEN ((No. 1);

- 2) MSOA_WKDY_PR_GEN (No. 5);
- 3) MSOA_WKND_PR_GEN (No. 9);
- 4) OPC_WKDY_HR_GEN (No. 17);
- 5) OPC_WKDY_PR_GEN (No. 21); and
- 6) OPC_WKND_PR_GEN (No. 25);

Additionally, the rightmost columns in Table 6.20 and Table 6.21 also suggest the relatively obvious errors in the prediction using the expected MAP-PCP for the above six sets of experiments, shown by the maximum differences of 2.31% and 3.19% respectively. The interpretation of the above unsuccessful prediction cases requires the mathematical expression of the MAP-PCP. Similarly to the MLE-PCP, the MAP-PCP can also be represented as the summation of a combination of joint probabilities as derived in Equation 4-101 (Section 4.2.3):

$$\begin{aligned}
 & PCP_{MAP} \\
 &= \begin{cases} p(x_1, y_0) + p(x_1, y_1), p(x_0, y_0) < p(x_1, y_0) \text{ and } p(x_0, y_1) < p(x_1, y_1) \\ p(x_0, y_0) + p(x_1, y_1), p(x_0, y_0) > p(x_1, y_0) \text{ and } p(x_0, y_1) < p(x_1, y_1). \\ p(x_0, y_0) + p(x_0, y_1), p(x_0, y_0) > p(x_1, y_0) \text{ and } p(x_0, y_1) > p(x_1, y_1) \end{cases} \quad (6-8)
 \end{aligned}$$

Moreover, the availability of the three segments in Equation 6-8 also indicates that the expression of the MAP-PCP as a specific combination of the joint probabilities depends on the informativeness of the prior, as defined by Equation 6-5. Particularly, the MAP-PCP is equivalent to the larger marginal probability between the two POVs of the enriched socio-demographic attributes with an informative prior, while collapsing to the same expression with the MLE-PCP otherwise. This therefore implies that the capability of the IDCM performance theory to predict the MAP-PCP depends on the respective similarity of $P(Y|X)$ and $P(X)$ in the estimation and the enrichment samples, as well as on the informativeness of the prior.

For the ‘Similar Prior’ Scenario, the difference in the $P(Y|X)$ between the two samples acts as the main contributor to the prediction error in the expected MAP-PCP, which can be attributed to either the random selection of the LTDS data or the en-route mobile phone usage pattern, as already explained previously in the analysis of the MLE-PCP. The reason for the relatively larger error values compared to the prediction of the MLE-PCP relates to the fact that the estimation priors employed in all the enrichment experiments for gender consist of a mixture of informative and uninformative priors, meaning that, in each set of gender enrichment experiments, the predicted MAP-PCP falls into different segments in Equation 6-8 according

to whether the estimation prior is informative or not. This would result in statistical significance for the mean of the difference between the actual MAP-PCPs and their expected values. This can be attributed to that they are effectively drawn from different sample distributions. In particular, the expected MAP-PCP for the enrichment using an uninformative prior is equivalent to the expected MLE-PCP, thus sharing the same prediction error due to the difference in $P(Y|X)$, as discussed in the analysis for the MLE-PCP. In contrast, the MAP-PCP predicted using an informative (and correct) prior would be significantly similar to the actual MAP-PCP as they essentially equal to the larger value of $P_{erc}(X)$, which can be obtained from the estimation prior based on the experimental setting of the ‘Similar Prior’ Scenario. This also explains the trivial difference between the actual and the expected MAP-PCPs for the enrichment of the three age-based attributes. Specifically, it can be derived that, since the included estimation prior is informative and correct in all the enrichment experiments for age-based attributes, the difference in the $P(Y|X)$ between the estimation and the enrichment samples that contributes to the discrepancy between the actual and the expected MLE-PCP in the six sets of experiments (i.e. No. 2, 11, 12, 18, 27 and 28) is compensated, leading to the similarity of the expected MAP-PCP values to their actual values. Given that, in the analysis of the prediction of the MLE-PCP for gender, the difference between the actual and expected values is either statistically insignificant (i.e. No. 1, 9, 17 and 25) or acceptable in the application context (No. 5 and 21), it is reasonable to argue that the prediction for the MAP-PCP with the expected values derived using the IDCM performance theory is successful when the employed estimation prior for all the compared experiments is either informative or uninformative, but not so when it is a mixture of the two.

As compared to the ‘Similar Prior’ Scenario, the larger difference between the actual and the expected MAP-PCPs for the ‘Dissimilar Prior’ Scenario in the six sets of gender enrichment experiments (marked in red in Table 6.21) can be explained by the incorrectness of the included estimation prior, in addition to the statistically significant difference between the actual and expected MAP-PCPs in the presence of a mixture of informative and uninformative priors. Since the enrichment using the MAP estimator with an informative and incorrect prior would enrich the attribute with the POV associated with the larger marginal probability in the estimation sample, i.e. the POV actually related to the smaller marginal probability in the enrichment sample, Equation 6-8 indicates that the MAP-PCP from using an informative but incorrect prior is equivalent to the smaller marginal probability between the POVs of the

enriched attribute. This hence exacerbates the prediction error for the six sets of gender enrichment experiments using the IDCM performance theory in the ‘Dissimilar Prior’ Scenario.

Figure 6.7 illustrates visually the relative deviation of the expected MAP-PCPs from their actual values obtained in all the enrichment experiments for gender in both the ‘Similar Prior’ and the ‘Dissimilar Prior’ scenarios, to exhibit the relatively more obvious deviation in the latter than in the former.

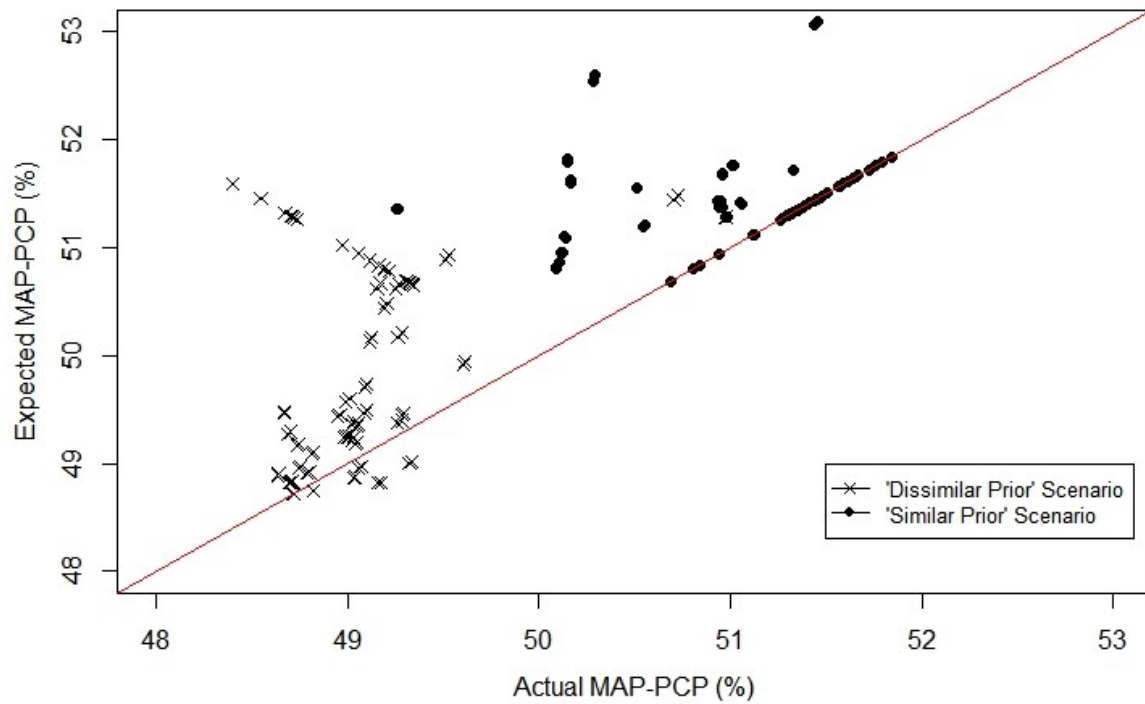


Figure 6.7 Actual MAP-PCP vs Expected MAP-PCPs for Gender Enrichment

Whereas the capability of the IDCM performance theory is significantly reduced in the original forecasting of the MAP-PCP using an informative yet incorrect prior, the estimation can be improved by swapping the mapping rule for the original enrichment using the MAP estimator, thus swapping the labels of the POVs of the attribute in the binary enrichment context. For instance, in the current context, where the prior with respect to gender is incorrect, the MAP-PCP can be improved by enriching the gender as ‘female’ if it is predicted as ‘male’ using the MAP estimator, and *vice versa*. This is shown in Figure 6.8 (below) which presents the comparison between the actual and expected MAP-PCPs for gender after revising the mapping rules for informative yet incorrect priors.

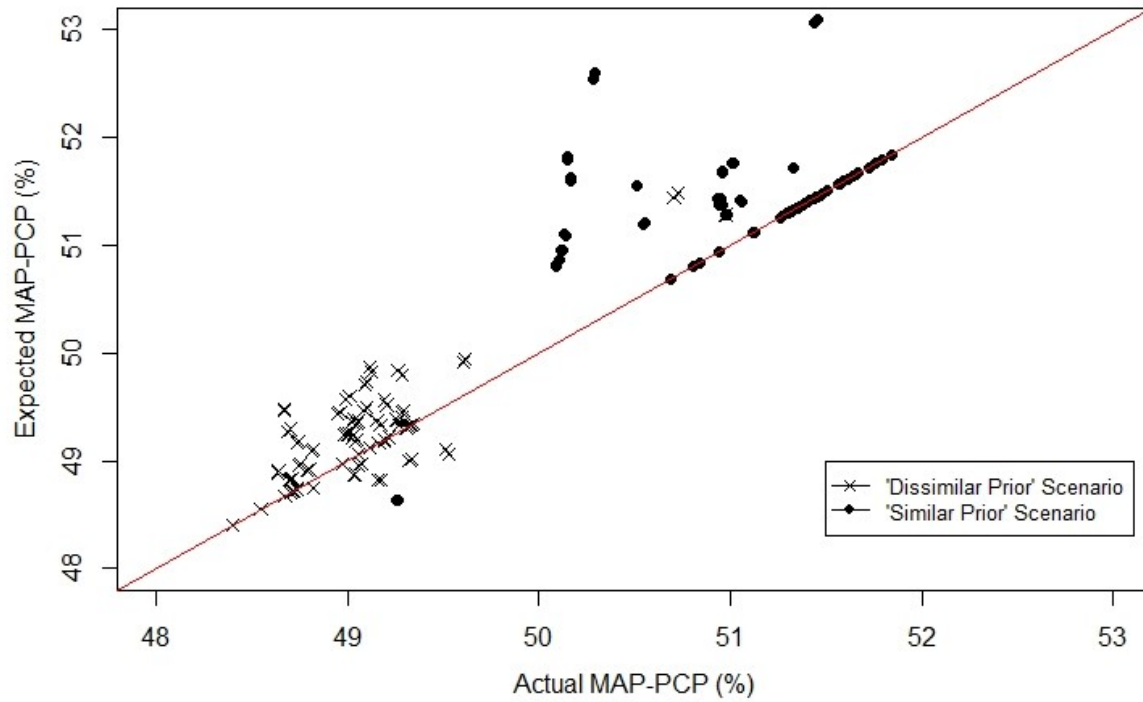


Figure 6.8 Actual MAP-PCP vs Expected MAP-PCPs for Revised Gender Enrichment

To summarise, the above analysis validates the capability of the IDCM enrichment performance theory to estimate (forecast) the enrichment performance of the IDCM approach using both the MLE and the MAP estimators in the presence of different data conditions. Specifically, the IDCM performance theory can successfully predict the enrichment performance using the MLE estimator with a prediction error that is either statistically insignificant or of which the absolute value is smaller than 2%. Regardless of the difference in the behaviour patterns between the estimation and the enrichment samples which cause the statistical significance between the predicted MLE-PCPs and their actual values, the error of $\pm 2\%$ is considerably acceptable in practical application and therefore acknowledges the capability of the IDCM performance theory to forecast the IDCM enrichment performance using the MLE estimator. On the other hand, the capability of the IDCM performance theory is found to vary when predicting the MAP-PCP given priors with different level of informativeness and correctness. Specifically, the IDCM performance theory is as capable as it is for predicting the MLE-PCP when making predictions with an uninformative estimation prior while the prediction when using an informative and correct prior is even stronger. In addition, whereas the capability of the IDCM performance theory is significantly reduced in the original forecasting of the MAP-PCP using an informative yet incorrect prior, the prediction

can be improved by swapping the mapping rule of the MAP estimator in the current binary enrichment context to make an acceptable prediction.

The extension of the above analysis for enriching the socio-demographic attributes with more than two POVs constitutes the ongoing investigation of the IDCM framework.

6.3.3 Comparison of IDCM framework against other socio-demographic enrichment methods

In Chapter 5, the IDCM socio-demographic enrichment framework is assessed in an empirical enrichment context using the LTDS data against two representative methods that have been extensively utilised in previous socio-demographic enrichment attempts, viz. the LGR and the SVM. As is discussed in Section 5.2.3, the IDCM approach can be compared with supervised machine learning (ML) methods given their extensive use in existing socio-demographic enrichment attempts and the similar data requirements with the current IDCM framework .

In this section, the IDCM approach is also compared against these two supervised ML methods in the current MND enrichment context to evaluate their respective enrichment performance across varied data conditions. The following contents present the comparative results and analysis for the MLE and the MAP estimators separately.

Specifically, Table 6.22 and Table 6.23 (below) demonstrate the comparative analysis for the PCP using the MLE estimator with LGR and the SVM, respectively (i.e. MLE-PCP vs LGR-PCP and MLE-PCP vs SVM-PCP) in the ‘Similar Prior’ Scenario.

Table 6.22 Test Results for Comparison of MLE- and LGR-PCPs in ‘Similar Prior’ Scenarios

Exp. NO.	Series Label	MLE-PCP (%)	LGR-PCP (%)	Sd. ^a (%)	Test ^b	p-value
1	MSOA_WKDY_HR_GEN	50.55	50.98	0.48	W	0.000
2	MSOA_WKDY_HR_YNG	86.00	93.11	4.66	W	0.000
3	MSOA_WKDY_HR_MDA	76.46	82.10	3.83	W	0.000
4	MSOA_WKDY_HR_EDL	82.18	82.47	24.38	W	0.003
5	MSOA_WKDY_PR_GEN	50.08	50.60	0.72	T	0.135
6	MSOA_WKDY_PR_YNG	85.73	93.13	5.19	T	0.017
7	MSOA_WKDY_PR_MDA	76.13	82.06	4.23	T	0.019
8	MSOA_WKDY_PR_EDL	81.80	88.93	5.04	T	0.018
9	MSOA_WKND_HR_GEN	51.01	51.23	0.73	W	0.003
10	MSOA_WKND_HR_YNG	87.31	93.07	3.24	W	0.000

11	MSOA_WKND_HR_MDA	77.56	82.13	2.49	W	0.000
12	MSOA_WKND_HR_EDL	83.39	82.55	23.62	W	0.003
13	MSOA_WKND_PR_GEN	51.02	51.40	0.21	T	0.007
14	MSOA_WKND_PR_YNG	76.40	93.08	29.84	W	0.050
15	MSOA_WKND_PR_MDA	69.60	82.11	21.79	W	0.050
16	MSOA_WKND_PR_EDL	73.70	89.03	26.38	W	0.050
17	OPC_WKDY_HR_GEN	50.55	50.98	0.47	W	0.000
18	OPC_WKDY_HR_YNG	86.18	93.11	4.55	W	0.000
19	OPC_WKDY_HR_MDA	76.60	82.10	3.75	W	0.000
20	OPC_WKDY_HR_EDL	82.35	82.47	24.32	W	0.003
21	OPC_WKDY_PR_GEN	50.08	50.60	0.72	T	0.137
22	OPC_WKDY_PR_YNG	85.91	93.13	5.07	T	0.018
23	OPC_WKDY_PR_MDA	76.27	82.06	4.14	T	0.019
24	OPC_WKDY_PR_EDL	81.97	88.93	4.94	T	0.018
25	OPC_WKND_HR_GEN	51.02	51.23	0.72	W	0.003
26	OPC_WKND_HR_YNG	87.47	93.07	3.14	W	0.000
27	OPC_WKND_HR_MDA	77.69	82.13	2.41	W	0.000
28	OPC_WKND_HR_EDL	83.55	82.55	23.57	W	0.003
29	OPC_WKND_PR_GEN	51.03	51.40	0.21	T	0.007
30	OPC_WKND_PR_YNG	76.45	93.08	30.00	W	0.050
31	OPC_WKND_PR_MDA	69.64	82.11	21.91	W	0.050
32	OPC_WKND_PR_EDL	73.75	89.03	26.53	W	0.050

^a Standard deviation of the difference between the LGR-PCPs and the MLE-PCPs obtained from each set of experiments in the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

Table 6.23 Test Results for Comparison of MLE- and SVM-PCPs in ‘Similar Prior’ Scenarios

Exp. NO.	Series Label	MLE-PCP (%)	SVM-PCP (%)	Sd. ^a (%)	Test ^b	<i>p</i> -value
1	MSOA_WKDY_HR_GEN	50.55	50.94	0.44	W	0.001
2	MSOA_WKDY_HR_YNG	86.00	93.11	4.66	W	0.000
3	MSOA_WKDY_HR_MDA	76.46	81.67	4.14	W	0.000
4	MSOA_WKDY_HR_EDL	82.18	88.99	4.55	W	0.000
5	MSOA_WKDY_PR_GEN	50.08	50.13	0.13	T	0.363
6	MSOA_WKDY_PR_YNG	85.73	93.13	5.19	T	0.017
7	MSOA_WKDY_PR_MDA	76.13	82.06	4.23	T	0.019
8	MSOA_WKDY_PR_EDL	81.80	88.93	5.04	T	0.018
9	MSOA_WKND_HR_GEN	51.01	51.39	0.26	W	0.000
10	MSOA_WKND_HR_YNG	87.31	93.07	3.24	W	0.000

11	MSOA_WKND_HR_MDA	77.56	82.10	2.55	W	0.000
12	MSOA_WKND_HR_EDL	83.39	89.07	2.98	W	0.000
13	MSOA_WKND_PR_GEN	51.02	51.45	0.11	T	0.000
14	MSOA_WKND_PR_YNG	76.40	93.08	29.84	W	0.050
15	MSOA_WKND_PR_MDA	69.60	82.11	21.79	W	0.050
16	MSOA_WKND_PR_EDL	73.70	89.03	26.38	W	0.050
17	OPC_WKDY_HR_GEN	50.55	50.94	0.43	W	0.001
18	OPC_WKDY_HR_YNG	86.18	93.11	4.55	W	0.000
19	OPC_WKDY_HR_MDA	76.60	81.69	4.06	W	0.000
20	OPC_WKDY_HR_EDL	82.35	88.99	4.46	W	0.000
21	OPC_WKDY_PR_GEN	50.08	50.13	0.13	T	0.363
22	OPC_WKDY_PR_YNG	85.91	93.13	5.07	T	0.018
23	OPC_WKDY_PR_MDA	76.27	82.06	4.14	T	0.019
24	OPC_WKDY_PR_EDL	81.97	88.93	4.94	T	0.018
25	OPC_WKND_HR_GEN	51.02	51.39	0.25	W	0.000
26	OPC_WKND_HR_YNG	87.47	93.07	3.14	W	0.000
27	OPC_WKND_HR_MDA	77.69	82.10	2.47	W	0.000
28	OPC_WKND_HR_EDL	83.55	89.07	2.88	W	0.000
29	OPC_WKND_PR_GEN	51.03	51.45	0.11	T	0.000
30	OPC_WKND_PR_YNG	76.45	93.08	30.00	W	0.050
31	OPC_WKND_PR_MDA	69.64	82.11	21.91	W	0.050
32	OPC_WKND_PR_EDL	73.75	89.03	26.53	W	0.050

^a Standard deviation of the difference between the SVM-PCPs and the MLE-PCPs obtained from each set of experiments in the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

It is revealed from Table 6.22 and Table 6.23 that, in the ‘Similar Prior’ Scenario, both the LGR-PCP and the SVM-PCP are significantly different (p -values < 0.1) from the MLE-PCP in 30 sets of enrichment experiments. A further investigation into the average PCP values suggests that the two supervised ML methods significantly outperform the IDCM approach using the MLE estimator in these 30 sets of enrichment experiments. Since $P_{est}(X)$ and $P_{erc}(X)$ are similarly distributed in the ‘Similar Prior’ Scenario, according to the experimental setting, the trained LGR and SVM using the estimation sample inevitably learn information about the attribute distribution of the enrichment sample, while the MLE estimator purposely excludes such information. This hence results in the better performance of the two supervised ML methods as compared with the MLE-PCP. In contrast, as is shown by the p -values that are greater than 0.1, the MLE-PCP is found to be indifferent from either the LGR-PCP or the SVM-

PCP in the remaining two sets of enrichment experiments, which are highlighted in red in Table 6.22 and Table 6.23, i.e.:

- 1) MSOA_WKDY_PR_GEN (No. 5); and
- 2) OPC_WKDY_PR_GEN (No. 21).

As is discussed previously in Section 6.3.1.1, the large majority of the estimation priors included in the MAP estimator for the above two sets of experiments are not informative enough to make a change between the MLE-PCP and the MAP-PCP. The indifferent MLE- and LGR-/SVM-PCPs observed in these two sets of experiments thus suggest that, in the ‘Similar Prior’ Scenario, LGR and the SVM cannot perform better than the IDCM approach using the MLE estimator by including additional information about $P_{erc}(X)$ through training if the training sample provides uninformative priors according to the definition in Equation 6-5.

The above results align with the findings obtained from the previous empirical application in Chapter 5, indicating that the LGR and the SVM generally perform no worse than the MLE estimator of the IDCM approach when the distribution of the enriched attribute in the estimation is similar to the true attribute distribution in the enrichment sample, while whether they can outperform the MLE estimator depends on the informativeness of the estimation prior.

Figure 6.9 visualises the relative enrichment performance of LGR and the SVM against the MLE estimator for the 32 sets of enrichment experiments in the ‘Similar Prior’ Scenario.

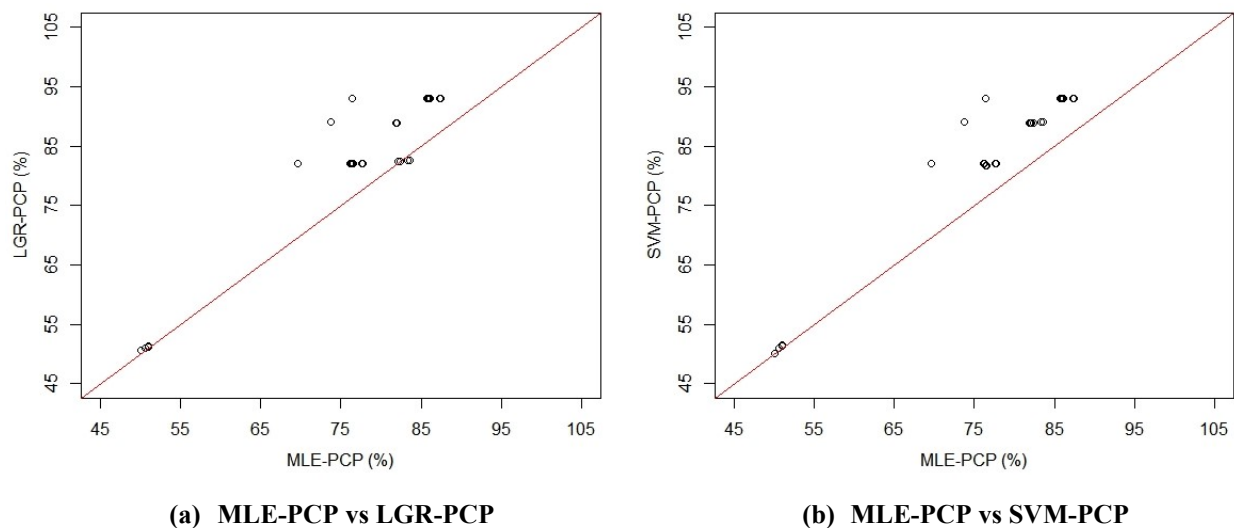


Figure 6.9 MLE-PCP vs LGR- and SVM-PCPs in ‘Similar Prior’ Scenario

Similarly, the results for the comparative analysis of the MLE-PCP with the LGR- and the SVM-PCPs for the 32 sets of enrichment experiments in the ‘Dissimilar Prior’ Scenario are shown in Table 6.24 and Table 6.25 (below), respectively.

Table 6.24 Test Results for Comparison of MLE- and LGR-PCPs in ‘Dissimilar Prior’ Scenarios

Exp. NO.	Series Label	MLE-PCP (%)	LGR-PCP (%)	Sd.^a (%)	Test^b	p-value
1	MSOA_WKDY_HR_GEN	49.45	49.21	0.36	W	0.006
2	MSOA_WKDY_HR_YNG	86.00	93.11	4.66	W	0.000
3	MSOA_WKDY_HR_MDA	76.46	82.10	3.83	W	0.000
4	MSOA_WKDY_HR_EDL	82.18	82.47	24.38	W	0.003
5	MSOA_WKDY_PR_GEN	49.49	49.21	0.34	T	0.102
6	MSOA_WKDY_PR_YNG	85.73	93.13	5.19	T	0.017
7	MSOA_WKDY_PR_MDA	76.13	82.06	4.23	T	0.019
8	MSOA_WKDY_PR_EDL	81.80	88.93	5.04	T	0.018
9	MSOA_WKND_HR_GEN	49.16	48.99	0.91	W	0.074
10	MSOA_WKND_HR_YNG	87.31	93.07	3.24	W	0.000
11	MSOA_WKND_HR_MDA	77.56	82.10	2.55	W	0.000
12	MSOA_WKND_HR_EDL	83.39	85.81	17.19	W	0.000
13	MSOA_WKND_PR_GEN	49.30	49.30	0.00	W	1.000
14	MSOA_WKND_PR_YNG	76.40	93.08	29.84	W	0.050
15	MSOA_WKND_PR_MDA	69.60	82.11	21.79	W	0.050
16	MSOA_WKND_PR_EDL	73.70	89.03	26.38	W	0.050
17	OPC_WKDY_HR_GEN	49.45	49.21	0.36	W	0.006
18	OPC_WKDY_HR_YNG	86.18	93.11	4.55	W	0.000
19	OPC_WKDY_HR_MDA	76.60	82.10	3.75	W	0.000
20	OPC_WKDY_HR_EDL	82.35	82.47	24.32	W	0.003
21	OPC_WKDY_PR_GEN	49.48	49.21	0.34	T	0.100
22	OPC_WKDY_PR_YNG	85.91	93.13	5.07	T	0.018
23	OPC_WKDY_PR_MDA	76.27	82.06	4.14	T	0.019
24	OPC_WKDY_PR_EDL	81.97	88.93	4.94	T	0.018
25	OPC_WKND_HR_GEN	49.15	48.98	0.91	W	0.074
26	OPC_WKND_HR_YNG	87.47	93.07	3.14	W	0.000
27	OPC_WKND_HR_MDA	77.69	82.10	2.47	W	0.000
28	OPC_WKND_HR_EDL	83.55	85.81	17.15	W	0.000
29	OPC_WKND_PR_GEN	49.30	49.30	0.00	W	1.000
30	OPC_WKND_PR_YNG	76.45	93.08	30.00	W	0.050
31	OPC_WKND_PR_MDA	69.64	82.11	21.91	W	0.050

32	OPC_WKND_PR_EDL	73.75	89.03	26.53	W	0.050
----	-----------------	-------	-------	-------	---	-------

^a Standard deviation of the difference between the LGR-PCPs and the MLE-PCPs obtained from each set of experiments in the ‘Dissimilar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

Table 6.25 Test Results for Comparison of MLE- and SVM-PCPs in ‘Dissimilar Prior’ Scenarios

Exp. NO.	Series Label	MLE-PCP (%)	SVM-PCP (%)	Sd. ^a (%)	Test ^b	p-value
1	MSOA_WKDY_HR_GEN	49.45	49.21	0.36	W	0.006
2	MSOA_WKDY_HR_YNG	86.00	93.11	4.66	W	0.000
3	MSOA_WKDY_HR_MDA	76.46	82.10	3.83	W	0.000
4	MSOA_WKDY_HR_EDL	82.18	88.99	4.55	W	0.000
5	MSOA_WKDY_PR_GEN	49.49	49.35	0.22	T	0.180
6	MSOA_WKDY_PR_YNG	85.73	93.13	5.19	T	0.017
7	MSOA_WKDY_PR_MDA	76.13	82.06	4.23	T	0.019
8	MSOA_WKDY_PR_EDL	81.80	88.93	5.04	T	0.018
9	MSOA_WKND_HR_GEN	49.16	49.13	0.44	W	0.161
10	MSOA_WKND_HR_YNG	87.31	93.07	3.24	W	0.000
11	MSOA_WKND_HR_MDA	77.56	82.10	2.55	W	0.000
12	MSOA_WKND_HR_EDL	83.39	89.07	2.98	W	0.000
13	MSOA_WKND_PR_GEN	49.30	49.20	0.25	W	0.050
14	MSOA_WKND_PR_YNG	76.40	93.08	29.84	W	0.050
15	MSOA_WKND_PR_MDA	69.60	82.11	21.79	W	0.050
16	MSOA_WKND_PR_EDL	73.70	89.03	26.38	W	0.050
17	OPC_WKDY_HR_GEN	49.45	49.21	0.36	W	0.006
18	OPC_WKDY_HR_YNG	86.18	93.11	4.55	W	0.000
19	OPC_WKDY_HR_MDA	76.60	82.10	3.75	W	0.000
20	OPC_WKDY_HR_EDL	82.35	88.99	4.46	W	0.000
21	OPC_WKDY_PR_GEN	49.48	49.34	0.23	T	0.180
22	OPC_WKDY_PR_YNG	85.91	93.13	5.07	T	0.018
23	OPC_WKDY_PR_MDA	76.27	82.06	4.14	T	0.019
24	OPC_WKDY_PR_EDL	81.97	88.93	4.94	T	0.018
25	OPC_WKND_HR_GEN	49.15	49.12	0.44	W	0.161
26	OPC_WKND_HR_YNG	87.47	93.07	3.14	W	0.000
27	OPC_WKND_HR_MDA	77.69	82.10	2.47	W	0.000
28	OPC_WKND_HR_EDL	83.55	89.07	2.88	W	0.000
29	OPC_WKND_PR_GEN	49.30	49.20	0.24	W	0.050
30	OPC_WKND_PR_YNG	76.45	93.08	30.00	W	0.050

31	OPC_WKND_PR_MDA	69.64	82.11	21.91	W	0.050
32	OPC_WKND_PR_EDL	73.75	89.03	26.53	W	0.050

^a Standard deviation of the difference between the SVM-PCPs and the MLE-PCPs obtained from each set of experiments in the ‘Dissimilar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

It is noted from Table 6.24 and Table 6.25 that, in the ‘Dissimilar Prior’ Scenario, the LGR- and the SVM-PCPs are also significantly different from the MLE-PCP in 28 sets of enrichment experiments, as indicated by the *p*-values that are smaller than 0.1. In the remaining four sets of enrichment experiments, meanwhile, LGR and the SVM are found to perform comparably to the MLE estimator. Specifically, for the comparison between LGR and the MLE estimator, the four sets of enrichment experiments (marked in red in Table 6.24) in which the difference between the respective PCPs is negligible encompass:

- 1) MSOA_WKDY_PR_GEN (No. 5);
- 2) MSOA_WKND_PR_GEN (No. 13);
- 3) OPC_WKDY_PR_GEN (No. 21); and
- 4) OPC_WKND_PR_GEN (No. 29).

In contrast, the difference between the SVM-PCP and the MLE-PCP is found to be negligible in the following four sets of enrichment experiments, as highlighted in red in Table 6.25.

- 1) MSOA_WKDY_PR_GEN (No. 5);
- 2) MSOA_WKND_PR_GEN (No. 9);
- 3) OPC_WKDY_PR_GEN (No. 21); and
- 4) OPC_WKND_PR_GEN (No. 25).

The above results imply that LGR and the SVM could respond differently to the informativeness of the estimation prior. In particular, the significance of the difference between the LGR-PCP and the MLE-PCP depends on whether the estimation prior can be defined as informative, given Equation 6-5. This is reflected by the observation that the four sets of experiments where the indifferent difference is detected for the LGR- and MLE-PCPs are exactly the four sets of experiments for equivalent MLE- and MAP-PCPs as demonstrated in Table 6.13 (Page 205). This is also intuitively logical since the expressions for the MAP estimator of the IDCM approach and LGR, as shown in Equations 6-9 and 6-10 below, would retain the same form when the estimation prior is employed in the MAP estimator.

$$X_{MAP} = \arg \max_X P(Y|X)P(X); \quad (6-9)$$

$$X_{LGR} = \arg \max_X P_T(Y|X)P_T(X). \quad (6-10)$$

The different sets of experiments for the comparable SVM- and MLE-PCPs suggest that the extent (informativeness) to which the incorrect information about the attribute distribution in the estimation (training) sample can affect the SVM-PCP is different from that for LGR, as well as for the IDCM MAP estimator. This can be attributed to the selection of the specific parameters c and γ for the similarity function of the SVM, which is interpreted in more detail in the following comparative discussion with respect to the MAP estimator.

An additional finding in terms of the comparative analysis of the IDCM approach using the MLE estimator with the two supervised ML methods for the ‘Dissimilar Prior’ Scenario is that, in the 26 sets of enrichment experiments where a significant difference is found in the enrichment performance of the compared methods, the MLE-PCP outperforms LGR and the SVM in the two sets of experiments for gender enrichment yet is outperformed by the two ML methods in the remaining 24 sets of enrichment experiments for age-based attributes. This signifies that the enrichment performance using the MLE estimator of the IDCM approach exceeds that of the LGR and the SVM when the estimation prior is informative yet incorrect. While the estimation (training) sample can provide informative and correct priors, however, the two supervised ML methods are more capable than the IDCM approach using the MLE estimator. Figure 6.10 visualises the comparison between the MLE-PCP and respectively the LGR-PCP and the SVM-PCP for the 32 sets of experiments in the ‘Dissimilar Prior’ Scenario.

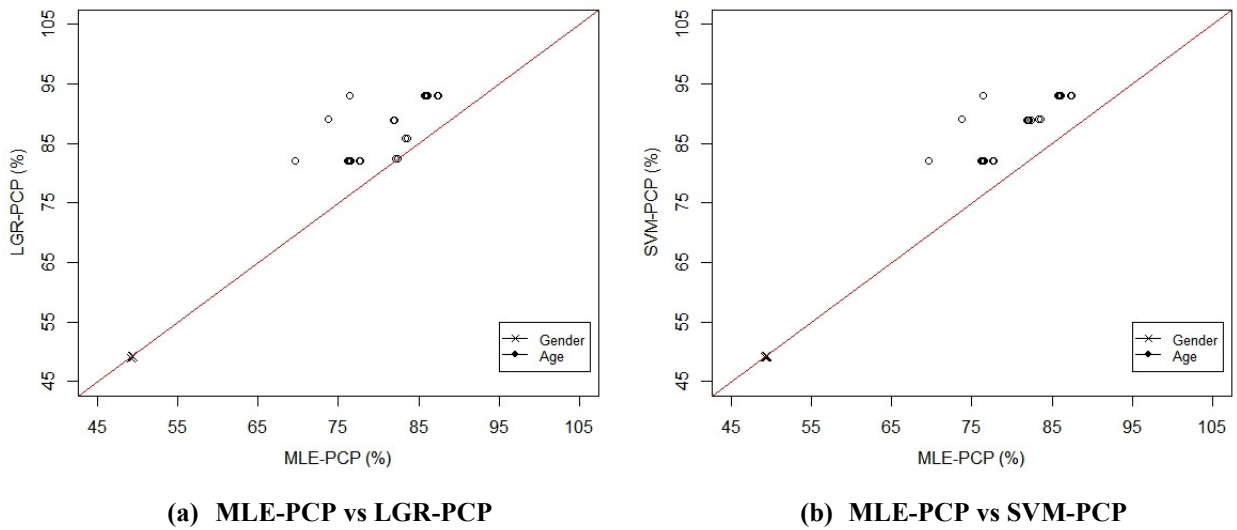


Figure 6.10 MLE-PCP vs LGR- and SVM-PCPs in ‘Dissimilar Prior’ Scenario

Similarly, the enrichment performance of the IDCM approach using the MAP estimator is also studied by comparing with the two supervised ML methods. In particular, Table 6.26 and Table 6.27 (below) demonstrate the test results for the comparative analysis of the MAP-PCP with the LGR-PCP and the SVM-PCP, respectively, for the 32 sets of enrichment experiments in the ‘Similar Prior’ Scenario.

Table 6.26 Test Results for Comparison of MAP- and LGR-PCPs in ‘Similar Prior’ Scenarios

Exp. NO.	Series Label	MAP-PCP (%)	LGR-PCP (%)	Sd.^a (%)	Test^b	p-value
1	MSOA_WKDY_HR_GEN	50.98	50.98	0.00	W	1.000
2	MSOA_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
3	MSOA_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000
4	MSOA_WKDY_HR_EDL	88.99	82.47	22.08	W	0.180
5	MSOA_WKDY_PR_GEN	50.60	50.60	0.00	T	1.000
6	MSOA_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
7	MSOA_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
8	MSOA_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
9	MSOA_WKND_HR_GEN	51.36	51.23	0.65	W	0.317
10	MSOA_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
11	MSOA_WKND_HR_MDA	82.13	82.13	0.00	W	1.000
12	MSOA_WKND_HR_EDL	89.07	82.55	22.07	W	0.157
13	MSOA_WKND_PR_GEN	51.40	51.40	0.00	T	1.000
14	MSOA_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
15	MSOA_WKND_PR_MDA	82.11	82.11	0.00	T	1.000
16	MSOA_WKND_PR_EDL	89.03	89.03	0.00	T	1.000
17	OPC_WKDY_HR_GEN	50.98	50.98	0.00	W	1.000
18	OPC_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
19	OPC_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000
20	OPC_WKDY_HR_EDL	88.99	82.47	22.08	W	0.180
21	OPC_WKDY_PR_GEN	50.60	50.60	0.00	T	1.000
22	OPC_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
23	OPC_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
24	OPC_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
25	OPC_WKND_HR_GEN	51.37	51.23	0.65	W	0.317
26	OPC_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
27	OPC_WKND_HR_MDA	82.13	82.13	0.00	W	1.000
28	OPC_WKND_HR_EDL	89.07	82.55	22.07	W	0.157

29	OPC_WKND_PR_GEN	51.40	51.40	0.00	T	1.000
30	OPC_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
31	OPC_WKND_PR_MDA	82.11	82.11	0.00	T	1.000
32	OPC_WKND_PR_EDL	89.03	89.03	0.00	T	1.000

^a Standard deviation of the difference between the LGR-PCPs and the MAP-PCPs obtained from each set of experiments in the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

Table 6.27 Test Results for Comparison of MAP- and SVM-PCPs in ‘Similar Prior’ Scenarios

Exp. NO.	Series Label	MAP-PCP (%)	SVM-PCP (%)	Sd. ^a (%)	Test ^b	<i>p</i> -value
1	MSOA_WKDY_HR_GEN	50.98	50.94	0.12	W	0.180
2	MSOA_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
3	MSOA_WKDY_HR_MDA	82.10	81.67	1.45	W	0.180
4	MSOA_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000
5	MSOA_WKDY_PR_GEN	50.60	50.13	0.75	T	0.185
6	MSOA_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
7	MSOA_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
8	MSOA_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
9	MSOA_WKND_HR_GEN	51.36	51.39	0.09	T	0.169
10	MSOA_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
11	MSOA_WKND_HR_MDA	82.13	82.10	0.18	W	0.317
12	MSOA_WKND_HR_EDL	89.07	89.07	0.00	W	1.000
13	MSOA_WKND_PR_GEN	51.40	51.45	0.13	T	0.363
14	MSOA_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
15	MSOA_WKND_PR_MDA	82.11	82.11	0.00	T	1.000
16	MSOA_WKND_PR_EDL	89.03	89.03	0.00	T	1.000
17	OPC_WKDY_HR_GEN	50.98	50.94	0.13	W	0.180
18	OPC_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
19	OPC_WKDY_HR_MDA	82.10	81.69	1.40	W	0.180
20	OPC_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000
21	OPC_WKDY_PR_GEN	50.60	50.13	0.75	T	0.185
22	OPC_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
23	OPC_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
24	OPC_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
25	OPC_WKND_HR_GEN	51.37	51.39	0.09	T	0.168
26	OPC_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
27	OPC_WKND_HR_MDA	82.13	82.10	0.17	W	0.317

28	OPC_WKND_HR_EDL	89.07	89.07	0.00	W	1.000
29	OPC_WKND_PR_GEN	51.40	51.45	0.13	T	0.363
30	OPC_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
31	OPC_WKND_PR_MDA	82.11	82.11	0.00	T	1.000
32	OPC_WKND_PR_EDL	89.03	89.03	0.00	T	1.000

^a Standard deviation of the difference between the SVM-PCPs and the MAP-PCPs obtained from each set of experiments in the ‘Similar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

Table 6.26 and Table 6.27 reveal that the MAP-PCP is comparable with the LGR- and the SVM-PCPs in the ‘Similar Prior’ Scenario for all 32 sets of enrichment experiments (p -values < 0.1). This aligns with the findings in Section 5.4.1 that the MAP estimator can perform equivalently to LGR and the SVM in the binary enrichment contexts where the estimation sample can provide the same, or similar, prior information with the true distribution of the enriched attribute in the enrichment sample. Figure 6.11 illustrates the comparison between the MAP-PCP and the LGR- and SVM-PCPs in the ‘Similar Prior’ Scenario.

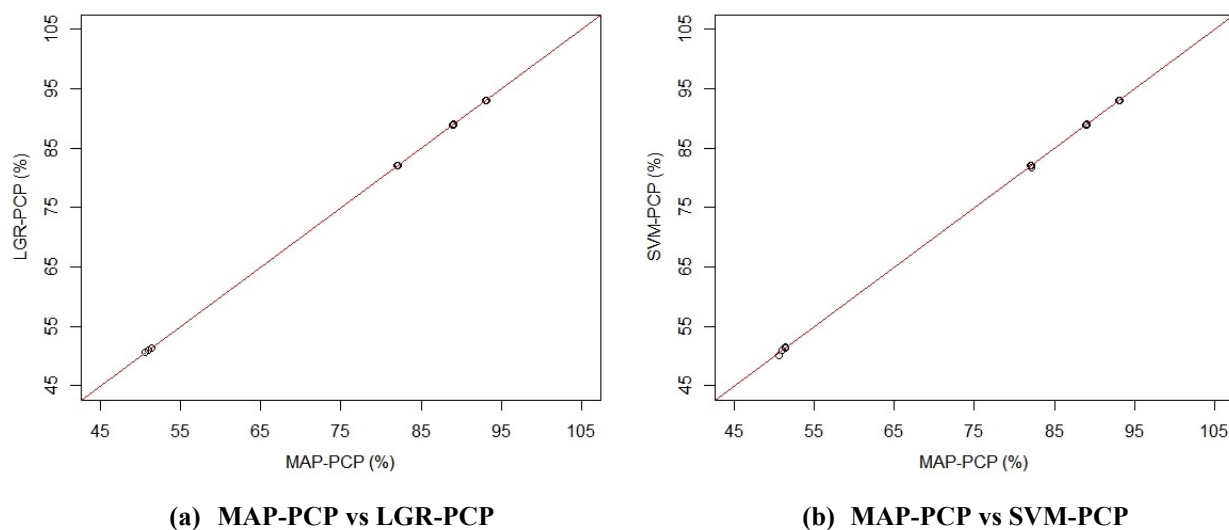


Figure 6.11 MAP-PCP vs LGR- and SVM-PCPs in ‘Similar Prior’ Scenario

As for the ‘Dissimilar Prior’ Scenario, Table 6.28 and Table 6.29 (below) report the test results for comparing the LGR-PCP with the LGR- and the SVM-PCPs, respectively.

Table 6.28 Test Results for Comparison of MAP- and LGR-PCPs in ‘Dissimilar Prior’ Scenarios

Exp. NO.	Series Label	MAP-PCP (%)	LGR-PCP (%)	Sd. ^a (%)	Test ^b	<i>p</i> -value
1	MSOA_WKDY_HR_GEN	49.21	49.21	0.00	W	1.000
2	MSOA_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
3	MSOA_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000

4	MSOA_WKDY_HR_EDL	88.99	82.47	22.08	W	0.180
5	MSOA_WKDY_PR_GEN	49.21	49.21	0.00	T	1.000
6	MSOA_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
7	MSOA_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
8	MSOA_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
9	MSOA_WKND_HR_GEN	48.87	48.99	0.60	W	0.317
10	MSOA_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
11	MSOA_WKND_HR_MDA	82.10	82.10	0.00	W	1.000
12	MSOA_WKND_HR_EDL	89.07	85.81	15.96	W	0.317
13	MSOA_WKND_PR_GEN	49.30	49.30	0.00	W	1.000
14	MSOA_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
15	MSOA_WKND_PR_MDA	82.11	82.11	0.00	T	1.000
16	MSOA_WKND_PR_EDL	89.03	89.03	0.00	T	1.000
17	OPC_WKDY_HR_GEN	49.21	49.21	0.00	W	1.000
18	OPC_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
19	OPC_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000
20	OPC_WKDY_HR_EDL	88.99	82.47	22.08	W	0.180
21	OPC_WKDY_PR_GEN	49.21	49.21	0.00	T	1.000
22	OPC_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
23	OPC_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
24	OPC_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
25	OPC_WKND_HR_GEN	48.86	48.98	0.60	W	0.317
26	OPC_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
27	OPC_WKND_HR_MDA	82.10	82.10	0.00	W	1.000
28	OPC_WKND_HR_EDL	89.07	85.81	15.96	W	0.317
29	OPC_WKND_PR_GEN	49.30	49.30	0.00	W	1.000
30	OPC_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
31	OPC_WKND_PR_MDA	82.11	82.11	0.00	T	1.000
32	OPC_WKND_PR_EDL	89.03	89.03	0.00	T	1.000

^a Standard deviation of the difference between the LGR-PCPs and the MAP-PCPs obtained from each set of experiments in the ‘Dissimilar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

In particular, the MAP-PCP is found to be comparable to the LGR-PCP in all 32 sets of enrichment experiments, as suggested by the fact that *p*-values are greater than 0.1. This indicates that the enrichment performance of the IDCM approach using the MAP estimator using the estimation prior is equivalent to that of the LGR regardless of the correctness or the informativeness of the prior in the binary enrichment context, in line with the same expression held by the two. Since the ability of the IDCM performance theory to predict and interpret the

MAP-PCP is verified in Section 6.3.2, the comparable enrichment performance using the MAP estimator and LGR in the current application signifies the ability of the IDCM performance theory in recognising, forecasting and explaining the variation in the enrichment performance using LGR in the binary context.

Table 6.29 Test Results for Comparison of MAP- and SVM-PCPs in ‘Dissimilar Prior’ Scenarios

Exp. NO.	Series Label	MAP-PCP (%)	SVM-PCP (%)	Sd.^a (%)	Test^b	p-value
1	MSOA_WKDY_HR_GEN	49.21	49.21	0.00	W	1.000
2	MSOA_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
3	MSOA_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000
4	MSOA_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000
5	MSOA_WKDY_PR_GEN	49.21	49.35	0.33	T	0.363
6	MSOA_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
7	MSOA_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
8	MSOA_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
9	MSOA_WKND_HR_GEN	48.87	49.13	0.78	W	0.141
10	MSOA_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
11	MSOA_WKND_HR_MDA	82.10	82.10	0.00	W	1.000
12	MSOA_WKND_HR_EDL	89.07	89.07	0.00	W	1.000
13	MSOA_WKND_PR_GEN	49.30	49.20	0.25	W	0.050
14	MSOA_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
15	MSOA_WKND_PR_MDA	82.11	82.11	0.00	T	1.000
16	MSOA_WKND_PR_EDL	89.03	89.03	0.00	T	1.000
17	OPC_WKDY_HR_GEN	49.21	49.21	0.00	W	1.000
18	OPC_WKDY_HR_YNG	93.11	93.11	0.00	W	1.000
19	OPC_WKDY_HR_MDA	82.10	82.10	0.00	W	1.000
20	OPC_WKDY_HR_EDL	88.99	88.99	0.00	W	1.000
21	OPC_WKDY_PR_GEN	49.21	49.34	0.33	T	0.363
22	OPC_WKDY_PR_YNG	93.13	93.13	0.00	T	1.000
23	OPC_WKDY_PR_MDA	82.06	82.06	0.00	T	1.000
24	OPC_WKDY_PR_EDL	88.93	88.93	0.00	T	1.000
25	OPC_WKND_HR_GEN	48.86	49.12	0.78	W	0.141
26	OPC_WKND_HR_YNG	93.07	93.07	0.00	W	1.000
27	OPC_WKND_HR_MDA	82.10	82.10	0.00	W	1.000
28	OPC_WKND_HR_EDL	89.07	89.07	0.00	W	1.000
29	OPC_WKND_PR_GEN	49.30	49.20	0.24	W	0.050
30	OPC_WKND_PR_YNG	93.08	93.08	0.00	T	1.000
31	OPC_WKND_PR_MDA	82.11	82.11	0.00	T	1.000

32	OPC_WKND_PR_EDL	89.03	89.03	0.00	T	1.000
----	-----------------	-------	-------	------	---	-------

^a Standard deviation of the difference between the SVM-PCPs and the MAP-PCPs obtained from each set of experiments in the ‘Dissimilar Prior’ Scenarios.

^b ‘T’ denotes the paired *t*-test; ‘W’ refers to the Wilcoxon signed-rank test.

In contrast, the results of the comparative analysis between the MAP- and the SVM-PCP for the ‘Dissimilar Prior’ Scenario (Table 6.29) imply that the enrichment performance of the two is comparable in 30 sets of experiment enrichments, with the only exceptions being the following two sets (p -values < 0.1) which are marked in red in the table.

- 1) MSOA_WKND_PR_GEN (No. 13); and
- 2) OPC_WKND_PR_GEN (No. 29).

Inspecting the PCP values in these two sets of enrichment experiments shows that the MAP estimator slightly outperforms the SVM. This is consistent with the previous analysis for the MLE-PCP, which suggests that the SVM performs equivalently to the MLE estimator, and with the findings in Section 6.3.1.1 that reveal that the MAP estimator outperforms its MLE counterpart in the above two sets of experiments. This can be explained by the values specified for the SVM parameters, i.e. c and γ , indicating that different parameters would show different sensitivity to the informativeness of the estimation prior and hence lead to different enrichment performance. This also implies that the combination of c and γ that may improve the SVM-PCP can be trained. When various combinations of c and γ are tested, however, it is apparent that the best-performing parameters trained using the estimation sample cannot always maximise the SVM-PCP for the enrichment sample. As is discussed previously in Section 5.4.2, this is typical for various applications using the SVM, therefore limiting the transferability of this method for socio-demographic enrichment purposes across different data conditions or application contexts.

The enrichment performance of the IDCM approach using the MAP estimator as compared against LGR and the SVM is visualised in Figure 6.12(a) and Figure 6.12(b) (below), respectively.

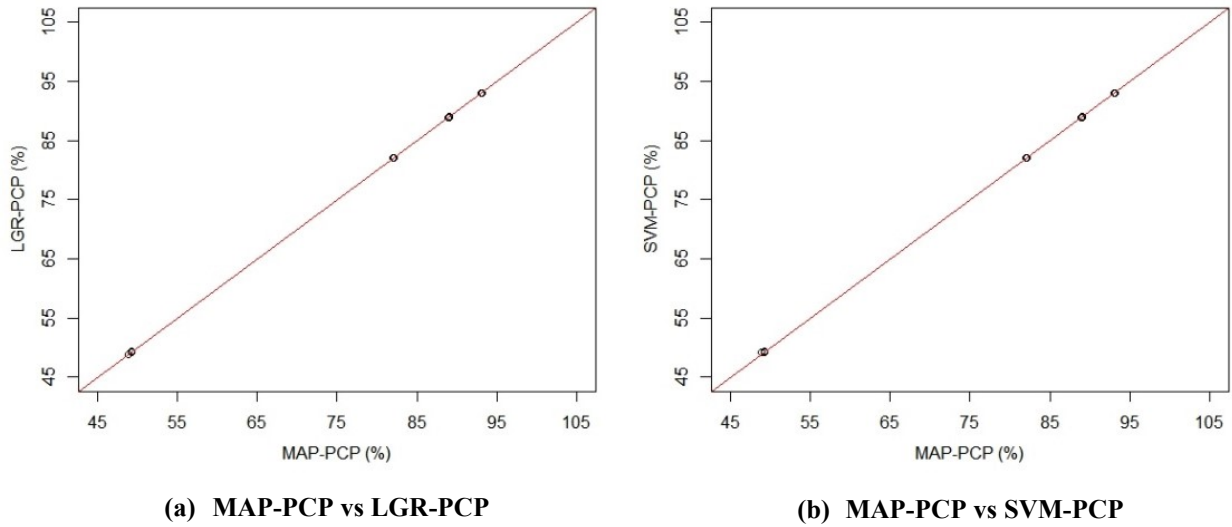


Figure 6.12 MAP-PCP vs LGR- and SVM-PCPs in ‘Similar Prior’ Scenario

6.4 Summary

The Inverse Discrete Choice Modelling (IDCM) framework was explored in this chapter via an empirical application of socio-demographic enrichment for anonymised mobile network data (MND) from stated travel behaviour patterns in London Travel Demand Survey (LTDS). The MND and the LTDS data were collected respectively by a mobile network provider and Transport for London, covering the Greater London area and collected during a close period. By assuming that the behaviour pattern given specific socio-demographic attributes remains similar in data collected from different sources for the same (spatial and temporal) population, the objectives of this empirical application included:

- 1) To understand the applicability of the IDCM framework for the enrichment sample with different data conditions with the estimation sample;
- 2) To validate the capability of the IDCM performance theory to predict and interpret the IDCM enrichment performance in light of the presented difference in data conditions; and
- 3) To compare the IDCM framework against representatives of existing socio-demographic enrichment methods to understand the relative transferability of the proposed framework under different data conditions.

To achieve the above objectives, Section 6.1 introduced the LTDS and the MND respectively serving as the estimation and the enrichment samples for the present study, as well as how the datasets were modified to fit the current scope. Section 6.2, meanwhile, demonstrated the experimental design based on the above objectives. The results and analysis for the performed

enrichment experiments were provided in Section 6.3. Specifically, as per the first objective, the difference in the data distribution of the estimation (LTDS) and the enrichment (MND) samples were explored and discussed in Section 6.3.1 from the following two dimensions:

- 1) Quality of the estimation prior, which indicates whether the prior provided by the estimation sample can inform the correct information with respect to the marginal distribution of the enriched attribute in the enrichment sample.
- 2) Level of spatial detail in the enrichment sample, which implies the level of minimum geography at which the enrichment sample was collected and its similarity to the spatial detail level of the estimation sample.

In terms of the first dimension, the enrichment performance of the IDCM approach was found to be determined by the informativeness and the correctness of the estimation prior. In particular, the IDCM approach using the maximum likelihood (MLE) and the maximum *a posteriori* (MAP) estimators can perform equivalently when including an uninformative prior. In the presence of an informative prior, the MAP estimator performs better than the MLE when the prior is correct while the MLE estimator surpasses the MAP when the prior is incorrect. This finding supplemented the criteria theoretically developed in Section 4.2.3.6 for selecting the preferred estimation methods for the IDCM approach. Generally, an informative while incorrect prior would cause the deterioration in the enrichment performance of both estimators while the negative effect to the MAP estimator is more significant. In contrast, the informative and correct priors for either estimator would lead to similar enrichment performance, suggesting that the enrichment performance of either estimator is irrelevant the similarity between the attribute distributions in the estimation and the enrichment samples.

On the other hand, the comparative analysis with respect to the level of spatial detail implied that the variation in the enrichment performance of the IDCM approach due to the spatial aggregation or disaggregation performed to the enrichment sample is effectively driven by the altered enrichment efficiency. In particular, the IDCM enrichment performance using the MLE estimator would vary with a significant change in the enrichment efficiency underlying the enrichment sample. In addition, the MLE estimator would perform better at the spatial detail level closer to that of the estimation sample. In contrast, the enrichment performance of the MAP estimator would vary in a similar way as that of the MLE estimator when the prior is uninformative. However, the MAP would perform stably in light of informative priors. This finding not only applies to the context of spatial-level aggregation/disaggregation, but is also

feasible for any aggregation that change the market share of the behaviour pattern to alter the enrichment efficiency underlying the enrichment sample.

Based on the observed variation in the IDCM enrichment performance due to the change in data condition, Section 6.3.2 then presented the results and analysis regarding the assessment of the IDCM performance theory to estimate and interpret such performance variation. This corresponds to the second objective of this application. Specifically, the IDCM performance theory was found to be able to estimate the enrichment performance using the MLE estimator with statistically insignificant or empirically acceptable errors ($\pm 2\%$). As for the MAP estimator, the following patterns were found:

- 1) The IDCM performance theory is able to predict the MAP-PCP when an informative and correct prior is employed.
- 2) The prediction for the MAP-PCP can be unsuccessful with an informative yet incorrect prior, but can be improved by modifying the mapping rule of the MAP estimator.
- 3) The IDCM performance theory is as capable as it is in the prediction of the MLE-PCP while using an uninformative prior, i.e. statistically insignificant or empirically acceptable errors.

The last stream of analysis concerned the comparison between the IDCM framework with the methods applied in existing socio-demographic enrichment attempts, as required by the third objective of the current application. Particularly, the IDCM framework was compared against logistic regression (LGR) and support vector machines (SVMs) as representatives of parametric and nonparametric supervised machine learning algorithms. Section 6.3.3 documented the test results and the corresponding analysis, which can be summarised as:

- 1) The IDCM approach using the MAP estimator were found to perform equivalently to the LGR under various data conditions in the present enrichment context. Similar enrichment performance were also revealed by the SVM. However, the SVM was found to perform slightly worse than the MAP estimator in some cases due to the specified SVM parameters.
- 2) Both the LGR and the SVM were found to outperform the IDCM approach using the MLE estimator in light of an informative and correct estimation prior, while the MLE estimator performs better than the two machine learning algorithms in cases of an informative yet incorrect prior. The IDCM approach using the MLE estimator was found to perform comparably to the two machine learning algorithms only when the estimation prior is

uninformative while the sensitivity to prior informativeness is different between the LGR and the SVM, which could be attributed to the specified SVM parameters.

In addition to confirming the findings in Section 5.4, the results analysed in this chapter extended the empirical findings about the enrichment performance of the IDCM approach under the variation in the data conditions in either the estimation or the enrichment sample, as well as given the difference between the data conditions of the estimation and the enrichment samples. Based on the analysis presented in this chapter, it can be concluded that the predictable enrichment of the socio-demographic attributes in the anonymous MND data can be enabled using the IDCM enrichment framework, which can possibly improve the usability of the extensive passive MND from two senses:

- 1) The socio-demographic attributes enriched using the IDCM approach can assist mobile network providers to gain a better understanding about their users, and thus provide more personalised services to their clients without disturbing them or infringing their privacy.
- 2) The MND with enriched socio-demographic attributes using the IDCM approach can enable the applications of the MND in various contexts, such as travel behaviour analysis and transport demand modelling, as well as other behaviour analysis contexts where the socio-demographic attributes are required.

The widespread applications that can be enabled by the enrichment of the socio-demographic attributes for various big data sources including but not limited to MND guarantee further investigations into and improvements for the IDCM framework.

7 CHALLENGES TO IDCM

APPLICATION IN TRANSPORT CONTEXTS

It is found from the two empirical applications in Chapters 5 and 6 that the relatively low enrichment performance of the Inverse Discrete Choice Modelling (IDCM) approach are characterised by either the high entropy of the enriched socio-demographic attribute (e.g. gender) or the low mutual information with the input behaviour feature. From the point of view of transport modellers, the low mutual information can be interpreted by the relatively weak explanatory power of the enriched socio-demographic attribute in affecting the behaviour as compared against other factors. This is reflected in the relative insignificant absolute value of the coefficient associated with the enriched attribute in comparison to the parameters of other explanatory variables in the corresponding discrete choice model (DCM), i.e. the IDCM kernel.

The main focus of this thesis has remained on establishing the theory for the predictability of the IDCM enrichment performance rather than on the level of the enrichment performance *per se*. However, the relatively low enrichment efficiency in the empirical applications may lead to a question of whether the IDCM approach can perform better than as shown in the presented empirical applications when the mutual correlation between the input behaviour feature and the enriched socio-demographic attribute is stronger.

Therefore, the aim of the present chapter is to explore this potential applicability challenge, using the proposed IDCM performance theory, in order to establish the extent to which the IDCM approach can prevail in real-world contexts given different levels of correlation between the input and the output of the enrichment. In particular, Section 7.1 provides a detailed outline of the challenge including the influential factors. Section 7.2 draws upon the proposed IDCM performance theory to theoretically discuss the link to and the impact of this substantive issue on the enrichment using the IDCM approach. The applicability of the theoretical discussion is assessed and validated in Section 7.3 by presenting the design, results and analysis of to a set of Monte Carlo (MC) enrichment experiments performed to validate the while Section 7.4 summarises the chapter.

7.1 Issue description

The IDCM framework established in Chapter 4 and the two real-world applications in Chapters 5 and 6 explain the operation of the IDCM framework from the theoretical and empirical standpoints. The consistent finding from both perspectives is that the enrichment performance of the IDCM approach varies correspondingly with respect to two factors:

- 1) The inherent entropy of the attribute which reflect the uncertainty (randomness) associated with the socio-demographic attribute to be enriched. Specifically, a higher level of uncertainty of an enriched attribute, reflected in its larger entropy, would, *ceteris paribus*, lead to a poorer absolute level of performance level, i.e. ‘percentage correctly predicted’ (PCP), of the IDCM approach.
- 2) The mutual information between the behaviour pattern (input feature of the enrichment) and the enriched attribute. Particularly, such mutual information partially determines the proposed metric of enrichment efficiency which is formulated as the ratio of the mutual information and the entropy of the enriched attribute. Representing the explanatory power of the behaviour feature in informing the variation in the enriched attribute, the enrichment performance of the IDCM approach is positively correlated with the mutual information, all else equal. This is especially visible when the entropy of the enriched attribute is high, i.e. the attribute is inherently uncertain, and the behaviour feature can provide additional information on the likely value of the enriched attribute.

Table 7.1 summarises the statistics for the entropy $H(X)$, mutual information $I(X, Y)$ and the enrichment efficiency η captured by the data in the London Travel Demand Survey (LTDS) and the mobile network data (MND), which serve as the datasets employed for the two empirical applications in Chapters 5 and 6.

Table 7.1 Statistics for Metrics of Information for LTDS (Chapters 5 and 6) and MND (Chapter 6)

Source	Metrics	Minimum	Maximum	Mean	Median
LTDS	$H(X)$	0.2410	0.9999	0.6138	0.5501
	$I(X, Y)$	0.0000	0.0826	0.0043	0.0009
	η (%)	0.00	7.64	0.31	5.00
MND	$H(X)$	0.0000	0.9999	0.6347	0.5858
	$I(X, Y)$	0.0000	0.0042	0.0003	0.0001
	η (%)	0.00	0.82	0.06	0.02

According to the two factors discussed above to affect the enrichment performance of the IDCM approach, ideally, the empirical data exploited to show the generality of the conclusions drawn for the enrichment performance of the IDCM framework need to cover the variation of the entropy and the mutual information. Observing Table 7.1 indicates that the socio-demographic attributes contained in both the LTDS and the MND range from nearly deterministic ones ($H(X)$ approaches 0.00) to almost completely random ones ($H(X)$ approaches 1.00). This hence provides sufficient data for investigating the variation of the entropy of the enriched socio-demographic attribute. At the same time, however, the mutual information between the input behaviour feature and the enriched attribute, i.e. $I(X, Y)$, captured in both datasets is significantly low (less than 0.10 in the LTDS and below 0.05 in the MND). This correspondingly leads to the low enrichment efficiency in both datasets (below 10.00% and 1.00% respectively). This further gives rise to the doubt that whether the empirical results from these two applications are reliable enough to validate the proposed IDCM framework given that the data cannot comprehensively cover one of the contributing factors to the IDCM enrichment performance. In other words, the low range of mutual information and enrichment efficiency captured by the LTDS and the MND results in the question of whether the enrichment effectiveness of the IDCM approach has been exerted extensively under these defective data situations. In the following contents, these data conditions are termed as the low-mutual-information context while the rest of the domain is referred to as the high-mutual-information context in a relative manner.

In addition, the benefit of the IDCM framework stems from the ability to relate not only its theory but also observed empirical findings to the mature field of discrete choice modelling in transport. This provides the opportunity to explain the above concern from the perspective of DCMs. In particular, the relatively low level of mutual information observed in the presented application contexts can be interpreted in relation to the empirical results drawn from an extensive number of traditional discrete choice modelling efforts (e.g. Ben-Akiva and Lerman (1974) and Goetzke (2008)). Specifically, these studies imply that most travel-related behaviour patterns are generally more significantly driven by the properties of choice alternatives rather than the socio-demographic attributes of travellers unless the attribute is closely related to travel or the attribute of trips. For example, the behaviour for the choice of travel mode depends more significantly on the travel time and monetary cost of the trip rather than most socio-demographic attribute (e.g. gender and age), with the exception of certain attributes such as car availability, which can be closely related to the trip (Lucarotti, 1977;

Supernak and Schoendorfer, 1985), or the income level given its association with cost-sensitivity. Consider a DCM solely adopting the socio-demographic attribute to be enriched as the explanatory variable in the utility function, which is stated as the IDCM kernel in Equations 4-32 and 5-1, and re-expressed below as:

$$V(y_0) = 0 \quad (7-1)$$

$$V(y_1) = \beta_0 + \beta_1 X$$

where y_0, y_1 possible outcome values (POVs) of the choice Y used for the enrichment;
 x_i POVs of the enriched attribute X where $i = 0, 1$;
 β_0, β_1 taste parameters of the IDCM kernel to be estimated.

When the dependence of the choice behaviour on the socio-demographic attribute is weak, one immediate consequence of using such a DCM to estimate this behaviour pattern is the potentially erroneous estimation of β_1 , i.e. the biased estimation of the taste parameter associated with the socio-demographic attribute. A number of existing studies, e.g. Haaf *et al.* (2014) and Lancsar *et al.* (2017) have shown that a poorly specified model can lead to bias estimates, in terms of either the sign or the magnitude of the estimated parameters. In the current content, this is because the relative absolute value of β_1 is trivial when compared with the average effect of the other unobserved factors as reflected by β_0 . This is referred to in the following contents using the ratio between the two, i.e. $|\frac{\beta_1}{\beta_0}|$. This thus effectively describe the aforementioned information-theoretic, substantive concern from the perspective of discrete choice modelling, resulting in the same research question of whether the proposed IDCM framework has operated to its best effectiveness when based on a behaviour pattern that is not strongly driven by the socio-demographic attribute to be enriched as the input feature. This substantive issue is hence termed as the enrichment effectiveness of the IDCM framework.

7.2 Theoretical discussion

This section theoretically discusses the substantive issue about the enrichment effectiveness of the IDCM framework, as described in Section 7.1, based on the econometric formulation of the IDCM approach. Specifically, the enrichment effectiveness of the IDCM framework refers to the extent to which the enrichment performance using the IDCM approach can be exerted to its maximum. As the empirical applications presented in Chapters 5 and 6 have validated the

ability of the IDCM performance theory to estimate the IDCM enrichment performance for the low-mutual-information context, in the current chapter, the enrichment effectiveness of the IDCM framework is explored by assessing whether the IDCM approach can outperform the expected value derived from the theory in the high-mutual-information context. In the paragraphs below, whether the effectiveness of the IDCM framework would be affected by the relatively weak dependence between the input behaviour feature and the enriched attribute is theoretically analysed based on the econometric formulation of the IDCM approach using the maximum *a posteriori* (MAP) and the maximum likelihood (MLE) estimators as well as their respective decoding schemes.

In particular, the MAP and the MLE estimators as developed in Section 4.1 and simplified in Section 4.2 are as restated below.

$$X_{MAP} = \arg \max_X P(Y|X)P(X). \quad (7-2)$$

$$X_{MLE} = \arg \max_X P(Y|X). \quad (7-3)$$

The corresponding decoding schemes of the MAP and the MLE estimators for the binary enrichment context are also reformulated below as:

$$Y = y_0: \begin{cases} p(y_0|x_0)p(x_0) > p(y_0|x_1)p(x_1) & \rightarrow X_{MAP}^* = x_0 \\ p(y_0|x_0)p(x_0) < p(y_0|x_1)p(x_1) & \rightarrow X_{MAP}^* = x_1 \end{cases} \quad (7-4)$$

$$Y = y_1: \begin{cases} p(y_1|x_0)p(x_0) > p(y_1|x_1)p(x_1) & \rightarrow X_{MAP}^* = x_0 \\ p(y_1|x_0)p(x_0) < p(y_1|x_1)p(x_1) & \rightarrow X_{MAP}^* = x_1 \end{cases}$$

$$\begin{aligned} Y = y_0 & \rightarrow X_{MLE}^* = x_0 \\ Y = y_1 & \rightarrow X_{MLE}^* = x_1 \end{aligned} \quad (7-5)$$

Consider first the special case when the enriched socio-demographic attribute is assumed to be evenly distributed, prompting the use of the MLE estimator. Inspecting Equation 7-3 and the corresponding decoding scheme (Equation 7-5) indicate that the enrichment result using the MLE estimator is purely associated with the relatively positive or negative effect of the enriched socio-demographic attribute on the behaviour, i.e. the sign of β_1 . implies that the magnitude of the taste parameter β_1 does not influence the relative better or worse performance of the IDCM approach in enriching this attribute from observations of the behaviour pattern, using the IDCM approach with the MLE estimator. Nevertheless, if the association between the behaviour feature and the enriched attribute is weak, and the number and effect of

unobserved factors the IDCM kernel are considerably large, the sign of β_1 can be incorrectly estimated. This would hamper the enrichment effectiveness of the IDCM approach using the MLE estimator. In addition, the absolute enrichment performance is still low, in light of the low enrichment efficiency ('weak explanatory power' in the sense of DCMs). Therefore, it follows that, in order to obtain an effective enrichment (utmost enrichment performance versus the expected value) of the IDCM approach using the MLE estimator, the behaviour pattern selected for enriching a specific socio-demographic needs to be strongly associated with that attribute, which is intuitive and expected. In the information-theoretic point of view, this is represented by the relatively high level of mutual information between the behaviour pattern and the enriched socio-demographic attribute. Alternatively, in the DCM perspective, this would coincide with the relatively large $|\frac{\beta_1}{\beta_0}|$, indicating that the relative importance of the unobserved factors is less significant as compared to that of the enriched socio-demographic attribute. This prompts the need of a model that is correctly specified, i.e. by including more explanatory variables, e.g. the attributes of the choice alternatives, as the components of the IDCM kernel. The inclusion of additional explanatory variables are expected to increase the goodness-of-fit of the model while also ensuring the accurate estimation for sign of the taste parameter associated with the enriched attribute.

Relaxing the assumption of evenly distributed attribute motivates the use of the MAP estimator. In such instances, the econometric formulation (Equation 7-2) and the decoding scheme (Equation 7-4) also confirm the importance of a strong correlation between the behaviour pattern and the enriched socio-demographic attribute. The difference when compared to the MLE case concerns the role and the impact of the informativeness and correctness of the prior included in the MAP estimator, which can either improve or deteriorate the enrichment effectiveness. Specifically, previous theoretical (Chapter 4) and empirical results (Chapters 5 and 6) suggest that an informative and correct prior can improve the enrichment results and possibly compensate for the errors induced from using an ineffective IDCM kernel (i.e. when the sign of β_1 is incorrectly estimated). However, the opposite can be true for an informative but incorrect prior, which could possibly deteriorate the enrichment results by introducing errors to the otherwise correctly enriched attribute using the MLE estimator (i.e. purely the IDCM kernel itself).

In summary, the above theoretical discussion suggests that, in order to be effective (resulting in as higher enrichment performance as can be expected by the IDCM performance theory),

the enrichment using the IDCM approach should be undertaken in the contexts meeting the following conditions:

Condition (CON) 1: The behaviour pattern(s) used as the input feature are relatively strongly correlated with the socio-demographic attribute to be enriched, as reflected by either a high mutual information between the two or a high ratio between the absolute values of β_1 and β_0 ;

Condition (CON) 2: Additional information with respect to other explanatory variables in the DCM for the behaviour pattern(s), beyond the enriched socio-demographic attribute, is available to ensure a correctly specified IDCM kernel; and

Condition (CON) 3: An estimation sample (or other data sources) is available that can provide informative and correct prior of the enriched socio-demographic attribute.

In circumstances where not all the above criteria can be satisfied, Conditions 1 and 2 can compensate each other for the enrichment using both the MLE and the MAP estimator while CON 3 can improve or adjust the negative effects of Conditions 1 and 2 for the enrichment using the MAP estimator. Another possibility involves using the IDCM kernels derived from joint models of behaviour or from repeated choices, which while beyond the scope of the present thesis, constituting one of the most important recommended avenues for further research associated with the IDCM as mentioned later in Chapter 8.

7.3 Monte Carlo (MC) enrichment experiments

In this section, the above theoretical discussion is validated in a set of detailed MC enrichment experiments that can reflect the data variation as described in the three conditions as proposed in Section 7.2. In particular, the MC experiments are performed to exhaustively explore the variation of the enrichment performance, i.e. the ‘percentage correctly predicted’ (PCP), of the IDCM approach with respect to the covariance structures of the input and output underlying the data sample. The comprehensive insights of these performed enrichment experiments are assured by simulating the variables (viz. the input behaviour pattern and the enriched attribute) that can represent the different levels of entropy, mutual information as well as the enrichment efficiency over their respective domains. In the contents below, Section 7.3.1 describes the experimental design of the MC simulation while Section 7.3.2 discusses the results.

7.3.1 Experimental design

The current MC simulation is conducted in the same binary enrichment context as previous chapters, i.e. the binary enrichment context in which the enrichment of a single binary socio-demographic attribute is undertaken based on a single observation of a binary variable reflecting a specific behaviour pattern of each individual. Denote the enriched socio-demographic attribute as X and the input behaviour pattern as Y .

To explore the variation of the entropy of the enriched socio-demographic attribute, i.e. $H(X)$, ten series of enrichment experiments are conducted, each involving the enrichment of a binary attribute with a specific level of entropy ranging from the almost deterministic level ($H(X) \rightarrow 0.00$) to the evenly distributed level ($H(X) \rightarrow 1.00$). The variation of $H(X)$ is permitted by simulating different market shares for the POVs of X incrementally by 0.05 for each series of enrichment experiments. Table 7.2 presents the probability distribution for the two POVs of the enriched socio-demographic attribute in each of the ten series of enrichment experiments as well as the corresponding entropy level.

Table 7.2 Probability Distribution and Corresponding Entropy of Enriched Socio-demographic Attribute in Each Series of Enrichment Experiments

Series No.	No. 1		No. 2		No. 3		No. 4		No. 5	
Probs.	0.05	0.95	0.10	0.90	0.15	0.85	0.20	0.80	0.25	0.75
$H(X)$ (bits)	0.2864		0.4690		0.6098		0.7219		0.8113	
Series No.	No. 6		No. 7		No. 8		No. 9		No. 10	
Probs.	0.30	0.70	0.35	0.65	0.40	0.60	0.45	0.55	0.50	0.50
$H(X)$ (bits)	0.8813		0.9341		0.9710		0.9928		1.0000	

On the other hand, the variation of the mutual information, i.e. $I(X, Y)$, between the behaviour pattern and the enriched socio-demographic attribute in each series of enrichment experiments is allowed by changing the joint probability distribution $P(X, Y)$ of the two incrementally by 0.01 in each experiment. As a result, there are in total 7869 experiments designed for all the ten series of enrichment experiments. Given that the level of entropy in each series is different, the enrichment efficiency (η) for X given Y are also varied for different experiment across series. Particularly, the lowest $I(X, Y)$ and η levels in all the ten series of enrichment experiments are 0.00 bits and 0% respectively while their respective maxima vary across series. Therefore, Table 7.3 (below) demonstrates the number of experiments designed for each of the

ten series of enrichment experiments given the variation of $P(X, Y)$ as described above, along with the corresponding maximum mutual information and enrichment efficiency.

Table 7.3 Number of Experiments in Each Series and Corresponding Maximum Mutual Information and Enrichment Efficiency Levels

Series No.	No. 1	No. 2	No. 3	No. 4	No. 5
Number of Exp.	188	401	588	751	888
$I_{\max}(X, Y)$ (bits)	0.1703	0.3408	0.4784	0.5871	0.6741
$\eta_{\max}$ (%)	59.45	73.10	78.45	81.32	83.09
Series No.	No. 6	No. 7	No. 8	No. 9	No. 10
Number of Exp.	1001	1088	1151	1188	625
$I_{\max}(X, Y)$ (bits)	0.7424	0.7940	0.8301	0.8515	0.8586
$\eta_{\max}$ (%)	84.24	85.01	85.49	85.77	85.86

The lowest mutual information in each series of enrichment experiments is 0.00 bits. It can therefore be noted from Table 7.3 that the simulated data captures the mutual information over $[0.0000, 0.8586]$, which is a broad coverage as compared with its domain by definition, i.e. $[0.0000, 1.0000]$. Analogically, the MC simulation also provides a comprehensive coverage for the enrichment efficiency between 0% and 85.86%. There are 4588 and 1157 out of the 7869 enrichment experiments that respectively reflect the variation of the mutual information in the LTDS and MND. These represent the realistic scenarios (low-mutual-information context) faced by the present study. In contrast, the remaining enrichment experiments make up for the high-mutual-information context (i.e. Condition 1 towards the end of Section 7.2). The thorough coverage for the variation of the three metrics, viz. entropy, mutual information and enrichment efficiency, assures the detailed and exhaustive investigation into the enrichment effectiveness of the IDCM framework from the view of information theorists.

For each experiment, a sample of 500 respondents are generated, each having a particular POV for the binary socio-demographic attribute and a specific POV for the binary behaviour variable according to the designed probability distribution. The sample is randomly split into two subsamples: (i) a 20% estimation subsample for creating the IDCM kernel; and (ii) an 80% enrichment subsample for implementing the IDCM approach. The random split is repeated for 20 times to avoid sample specificity caused by such random division. As the estimation and the enrichment subsamples are derived from the same samples, their covariance structures are similar, resulting in informative and correct priors, similar to the empirical application presented in Chapter 5. This corresponds to Condition 3 as documented in Section 7.2. Note

that the second condition discussed in Section 7.2 is not investigated in this MC simulation as the other explanatory variables are not related to the mutual dependence between the input behaviour feature and the enriched attribute.

Given the above design, the IDCM kernel retaining the form of a binary logit model with the utility function expressed as Equation 7- for each repetition of each enrichment experiment can be estimated. This also permits understanding the enrichment effectiveness of the IDCM approach given the relative importance of X in affecting Y , i.e. from the perspective of discrete choice modelling. As the empirical applications in Chapters 5 and 6 have validated the capability of the IDCM performance theory in estimating the IDCM enrichment performance, such an understanding can be achieved by assessing the consistency between the actual enrichment performance obtained from conducting the MC enrichment experiments and the corresponding expected values with respect to $|\frac{\beta_1}{\beta_0}|$.

7.3.2 Enrichment results and analysis

Given the experimental design in Section 7.3.1, the absolute value of the ratio between β_1 and β_0 is found to vary over the following domain:

$$|\frac{\beta_1}{\beta_0}| \in [0.000, +\infty). \quad (7-6)$$

The minimum value of $|\frac{\beta_1}{\beta_0}|$ exists when $\beta_1 = 0$ while $\beta_0 \neq 0$. Intuitively, in these cases, the inclusion of the enriched socio-demographic attribute in the IDCM kernel is found to have completely no effect to the occurrence of the behaviour feature. This is also confirmed from the information-theoretic perspective as the mutual information and the enrichment efficiency reach 0.0000 and 0% respectively.

On the other hand, $|\frac{\beta_1}{\beta_0}|$ approaches $+\infty$ when $\beta_0 = 0$. In such situations, the behaviour is entirely determined by the enriched attribute. It is noted, however, that:

$$I(X, Y) \neq 1.00 \text{ bit}. \quad (7-7)$$

This can be explained by the assumption that the behaviour pattern is only observed to be related to the enriched socio-demographic attribute when establishing the IDCM kernel. As a

result of this, the probability of the POVs of the behaviour feature is equal when the enriched attribute is labelled 0 as the systematic utilities of both alternatives are 0, i.e.:

$$V(Y = 0|X = 0) = V(Y = 1|X = 0) = 0; \quad (7-8)$$

$$P(Y = 0|X = 0) = P(Y = 1|X = 0) = 0.5. \quad (7-9)$$

Equation 7-9 hence implies that the behaviour given X is random when labelled 0. As the mutual information metric characterises the average amount of information about one variable given the knowledge of the other, the deterministic correlation between X and Y when $X = 1$ and the randomness of Y when $X = 0$ jointly suggest that Equation 7-7 holds.

The enrichment effectiveness of the IDCM approach is explored with respect to both the mutual information (information-theoretic perspective) and $|\frac{\beta_1}{\beta_0}|$ (discrete choice modelling perspective), with the former expressing the quantitative upper limit for the latter in each series. The contents below discuss the results of such analysis for the IDCM approach respectively using the MLE and the MAP estimators.

By comparing the actual MLE-PCP obtained from the 7869 enrichment experiments with their respective expected values, it is found that, in 7861 out of the 7869 enrichment experiments, the difference between the actual and expected MLE-PCPs is negligible, i.e. lower than 2%. In the rest 8 exceptional cases, the difference varies between 3% ~ 7%. A more detailed inspection reveals that the mutual information in these 8 enrichment experiments equals 0 while $|\frac{\beta_1}{\beta_0}|$ remains at a substantially low level by ranging between 0.01 ~ 0.04. On the other hand, it is interesting to note that the difference between the actual and the expected MLE-PCP is trivial when $|\frac{\beta_1}{\beta_0}| = 0$, which can be explained as follows. As is mentioned in the previous paragraphs, $|\frac{\beta_1}{\beta_0}| = 0$ occurs when $\beta_1 = 0$. In such cases, the inclusion of the enriched attribute in the IDCM kernel plays an ignorable role in affecting the behaviour pattern, which can be mathematically expressed as:

$$\begin{aligned} P(Y = 0|X = 0) &= P(Y = 0|X = 1); \\ P(Y = 1|X = 0) &= P(Y = 1|X = 1). \end{aligned} \quad (7-10)$$

A comparison between Equations 7-10 and 7-3 indicates that the enrichment using the MLE estimator is purely random when $\beta_1 = 0$. This implies that the actual MLE-PCP would

approach closely to 0.5, which is its expected value, when the sample size is large enough. The above analysis confirms the theoretical analysis in Section 7.2, which suggests that the enrichment effectiveness of the IDCM approach using the MLE estimator can possibly be deteriorated when the mutual dependence between the behaviour pattern and the enriched attribute is trivial. Moreover, it also explains the existence of the errors when estimating the MLE-PCP using the IDCM performance theory in the empirical applications in Chapters 5 and 6. In particular, as the mutual information between the input feature and the enriched socio-demographic attribute remains at a relatively low level in both applications, it is possible that the sign of β_1 in the IDCM kernel is not correctly estimated and subsequently hampers the enrichment effectiveness using the MLE estimator. On the other hand, the impact of such low correlation between the input and the output of the enrichment can also be viewed as insignificant, reflected by the acceptable errors in both empirical applications.

In terms of the MAP estimator, the difference between the actual MAP-PCP obtained from the MC enrichment experiment and their expected values derived using the IDCM performance theory is found to be insignificant ($< 2\%$) for all the 7869 enrichment experiments. This also keeps with the theoretical conclusions drawn in Section 7.2 that the informative and correct prior can compensate the relative weakness of the MLE estimator when the mutual relationship between the behaviour and the attribute is low. In addition, such insignificant difference between the actual and the expected MAP-PCP also aligns with the results in Chapter 5 as well as the ‘Similar Prior’ Scenario and the enrichment of age-based attribute in ‘Dissimilar Prior’ Scenario in Chapter 6 where the priors are informative and correct.

The above analysis with respect to the impact of the weak mutual correlation between the input behaviour feature and the enriched socio-demographic attribute on the IDCM enrichment respectively using the MLE and the MAP estimators show the stability of the enrichment effectiveness of the IDCM framework to the quality of the input available for the enrichment.

7.4 Summary

This chapter explored the enrichment effectiveness of the Inverse Discrete Choice Modelling (IDCM) framework with respect to the level of correlation between the input behaviour feature and the enriched socio-demographic attribute (as explained in Section 7.1). In particular, the enrichment effectiveness is defined as the extent to which the IDCM approach can perform as compared to its maximum performance level. This issue arouses from the observation that the

correlation between the behaviour feature and the enriched attribute is weak in the two empirical applications conducted in this study to validate the IDCM framework. From the information-theoretic perspective, such low weak correlation is reflected by the low mutual information while represented by the relatively small value of $|\frac{\beta_1}{\beta_0}|$ from the view of discrete choice modellers.

The above substantive issue was investigated theoretically (Section 7.2) and empirically (Section 7.3) based respectively on the econometric formulation of the IDCM approach and on the results of a set of exhaustive Monte Carlo (MC) enrichment experiments simulating different levels of mutual dependence levels. The results from both aspects revealed the stability of the enrichment effectiveness of the IDCM framework at different levels of correlation between the input behaviour feature and the enriched attribute. Small fluctuations were observed in the enrichment performance when such correlation is significantly weak, showing a slight deterioration in the enrichment effectiveness of the proposed framework due to the weak correlation. This therefore recommended the use of the behaviour patterns that are strongly correlated with the enriched attribute as the input of the enrichment.

When presented with a behaviour pattern that is not significantly related to the enriched attribute, the theoretical discussion in Section 7.2 suggested that the inclusion of other significant explanatory variables (e.g. the attributes of the choice alternatives) into the IDCM kernel can improve the enrichment effectiveness of the IDCM approach using the maximum likelihood (MLE) estimator. This is because the MLE-PCP is purely driven by the sign of β_1 in the IDCM kernel while the inclusion of other explanatory variables can improve the goodness-of-fit of the IDCM kernel and hence adapt the accuracy of β_1 . In addition to the inclusion of extra explanatory variables, the use of an informative and correct prior correspond to the other way to retain the enrichment effectiveness of the IDCM approach using the maximum *a posteriori* (MAP) estimator. In summary, the overall high and stable level of enrichment effectiveness of the IDCM framework as shown by the theoretical and empirical analysis in this chapter confirms the applicability of the proposed framework in light of different levels of correlation presented between the input behaviour feature and the enriched socio-demographic attribute.

8 CONCLUSIONS

Enabled by the latest advances in information and communication technologies, a growing amount of passive data with high volume, i.e. the ‘so-called’ big data, is being generated. The opportunities to extract additional values from this big data, and the limitation of its poor semantic richness, have jointly led to a massive growth of studies seeking to capitalise on advanced data analytics methods to discover information from, or otherwise deal with, these increasingly available big datasets.

Transport big data represents one of the mainstream big data sources, offering potential access to a rich number of travel-related behavioural variables, e.g. trip purposes and travel modes, that allow insights into people’s daily behaviour patterns. In contrast to the relative success in extracting travel-related behaviour patterns, however, the enrichment of the socio-demographic attributes of respondents and their households from/for the passive big datasets has remained underexploited. Additionally, existing approaches/algorithms applied to enrich the socio-demographic information of big data sources typically lack solid behavioural underpinnings in microeconomics and/or sociology to support and justify the relationships (models) between the input behaviour features and the output socio-demographic attributes. In particular, these existing methods model the behaviour patterns of respondents as the explanatory variable in order to predict their socio-demographic attributes as the dependent variable. While this conforms with the established enrichment problem, it fails to account for the microeconomic behavioural theory suggesting that people behave because of the type of person (socio-demographic attributes) that they are, and not *vice versa*. This lack of theoretical foundation to existing socio-demographic enrichment methods renders them incapable to either estimate (forecast) their enrichment performance prior to the enrichment *per se*, or to interpret the variability of, or any variation in, their enrichment performance, given different data conditions.

The above research gaps motivate the Inverse Discrete Choice Modelling (IDCM) framework proposed and developed in the present PhD study for the enrichment of the socio-demographic attributes for anonymous big data with predictable and interpretable enrichment performance. The current chapter concludes this research from four perspectives: the findings and key contributions of this thesis are summarised in Sections 8.1 and 8.2, respectively. Section 8.3 then discusses the potential implications of the developed IDCM framework, while Section 8.4

outlines the limitations of the results and analysis presented in this thesis, leading to the possible avenues for further investigations to extend the current framework, as well as to address the limitations.

8.1 Summary

The IDCM framework is established to enrich the socio-demographic attributes for anonymous big data given behaviour features extracted from big data, with predictable and interpretable enrichment performance. This thus corresponds to the two components of the framework, i.e. the econometric formulation of the approach and the accompanying theory for predicting and interpreting the enrichment performance.

Specifically, the econometric formulation is developed by postulating that the causal relationship in discrete choice models (DCMs) between the socio-demographic attributes X and the behaviour Y can be inverted. In other words, the proposed IDCM approach obtains the explanatory socio-demographic attributes in the model from observations of the behaviour features originally modelled as the response, hence the name IDCM framework. This inverted DCM is named the IDCM kernel. The econometric formulation then consists of the two estimators, corresponding respectively to the maximum likelihood (MLE) and the maximum a posterior (MAP) estimation methods, to solve the above inversion.

On the other hand, the accompanying theory, termed as the IDCM performance theory, suggests that the enrichment performance, i.e. the percentage correctly predicted (PCP) of the IDCM approach can be estimated (and therefore predicted) as a function of an assumed recalibrated constant given known market share of the input behaviour pattern and an IDCM kernel. In addition, the theory also proposes to mathematically link the PCP to the concept of enrichment efficiency, which is established in light of an analogy of the IDCM framework in information theory. In particular, the enrichment efficiency metric quantifies the explanatory power of the input feature in inferring the enriched attribute to therefore represent the covariance structure underlying the data. The IDCM performance theory thus mathematically links it to the PCP using the assumed recalibrated constant as the pivot, presenting a way to interpret the variability of, or variation in, the enrichment performance of the IDCM approach.

Two empirical applications of the IDCM framework in real-world socio-demographic enrichment contexts are performed, indicating that the IDCM performance theory is able to

forecast the enrichment performance of the IDCM approach (using both estimation methods) prior to the enrichment *per se* with either statistically insignificant or empirically acceptable errors ($\pm 2\%$). In addition, it is found to be able to detect and to interpret the variability and variation in its enrichment performance in light of different data conditions. These applications also supplement the theoretical establishment of the IDCM approach with respect to the selection criteria between the MLE and the MAP estimators given the availability, informativeness and correctness of the priors.

In the two empirical applications, comparative analyses are performed between the IDCM framework and two supervised machine learning methods, i.e. logistic regression (LGR) and support vector machines (SVMs), as representatives of the methods that have been extensively applied in existing socio-demographic enrichment attempts. These analyses implies that:

- 1) The IDCM approach performs comparably to the LGR and SVM when using the MAP estimator, or using the MLE estimator when the prior about the enriched socio-demographic attribute in the estimation sample is defined as uninformative.
- 2) When the estimation prior is informative, the relative enrichment performance of these methods is determined by the correctness of the estimated prior: the IDCM approach using the MLE estimator outperforms the IDCM-MAP estimator as well as the LGR and SVM when the estimation is incorrect, while the LGR and SVM outperforms the MLE estimator when a correctly estimated prior is employed.
- 3) Neither the LGR nor the SVM is able to predict their respective enrichment performance or to detect, estimate or interpret the variability and variation in their enrichment performance due to the variation in the data conditions of either the training or the enrichment samples. This significantly limits the transferability of those methods under various data conditions or across application contexts.
- 4) The proved capability of the IDCM framework to identify, estimate and explain its enrichment performance using both the MLE and the MAP estimators given various data distributions assures its applicability for enriching the socio-demographic attributes of anonymous big data sources with predictable enrichment performance. Furthermore, this supports the transferability of the IDCM framework under different data conditions or across application contexts.

8.2 Key contributions

Given the results and findings of this thesis as summarised in Section 8.1, the primary contributions of the present PhD study can be concluded as outlined below:

- 1) Developed the econometric formulation of the IDCM approach for enriching socio-demographic attributes in anonymous big data given the extracted discrete variables reflecting the behaviour patterns of the anonymous respondents in the data, while retaining the causal dependence between the attribute and the behaviour patterns suggested by microeconomic consumer theories.
- 2) Expressed the joint probability distribution of the enriched socio-demographic attribute and the behaviour pattern underlying the enrichment sample as a function of the assumed recalibrated constant $\hat{\beta}_0$ for a binary DCM using a reversed constant recalibration technique given the information with regard to the distribution of the behaviour pattern in the enrichment sample.
- 3) Represented the disaggregate, individual-level performance of the IDCM approach, measured by the PCP, for the enrichment of a single binary socio-demographic attribute given a single pattern exhibited once (per individual) as a specific combination of the joint probabilities expressed in 2), thus as a function of $\hat{\beta}_0$.
- 4) Analogised the behaviour-making process to the transmission of messages in an information communication system, enabling the use of the metrics developed in information theory to quantitatively interpret the dependence pattern between the behaviour and the enriched attribute and hence to measure the properties of the IDCM approach.
- 5) Formulated the metric of enrichment efficiency (η) to characterise the explanatory power of the observed behaviour pattern in inferring the enriched attribute as the quotient of the mutual information between the two and the entropy of the enriched attribute, given the analogy of behaviour-making to information transmission as described in 4).
- 6) Established the IDCM performance theory which links the MAP- and the MLE-PCP of the IDCM approach to the η that represents the data distribution in the enrichment sample, using $\hat{\beta}_0$ as the pivot, for the binary enrichment context as explained in 3).

- 7) Validated the ability of the IDCM performance theory to estimate the enrichment performance of the IDCM approach according to specified data conditions, and to detect and interpret any changes in the performance arising from variations in the data conditions.
- 8) Suggested the criteria for selecting between estimation methods (the MAP or the MLE) to solve the econometric formulation of the IDCM approach, in light of the availability and the quality (informativeness and the correctness) of the prior information.

8.3 Implications

Broadly, using big data analytics technologies to exploit the value and information that can be provided by big data is increasingly become a new strategic objective in many contexts. This can be evidenced by the survey analysis conducted by Gartner (2015), who suggested that three quarters of companies have invested, or planned to invest, in the utilisation of big data technologies since 2015. According to Borgi *et al.* (2017), the trending applications of big data analytics range from prospecting fresh profit streams, enhancing customer experience and creating active marketing and campaigns, to reducing fraud and improving security intelligence, and research, planning and development.

Transport modelling is one of the most ideal research fields to benefit from the big data that are generated to capture the spatial-temporal movement patterns of respondents, which have been seen as cost-effective alternatives or supplements to conventional transport modelling datasets. Borgi *et al.* (2017) argued that the untapped potential of such big data for industrial or service sectors can usually be classified into three aspects: (i) operational efficiency; (ii) customer experience; and (iii) new business models. As a newly-developed statistical model, the IDCM approach itself can be treated as a novel business model when applied for business purposes. Furthermore, in the following sections, the implications of the IDCM framework proposed and established in this PhD study are discussed, according to both the definition by Borgi *et al.* (2017) and the key contributions of this study, from the following four dimensions: privacy considerations (Section 8.3.1), transferability and generalisability across application contexts (Section 8.3.2), operational efficiency (Section 8.3.3) and customer experience (Section 8.3.4).

8.3.1 Privacy considerations

As is extensively discussed in Chapters 1 and 2, the shortage in socio-demographic enrichment, compared to the enrichment of other behaviour patterns from big data sources, is fundamentally attributable to the difficulties associated with traditional data collection methods, and the growing desire for preservation of privacy by the public and reflected in increasingly forbidding legislative regulations for the protection of privacy. The lack of socio-demographic information within anonymous big data has, however, significantly limited the applicability of such data in behaviour-driven analysis contexts, due to the essential role of socio-demographic attributes as exogenous determinants of human behaviour.

The outcome of the socio-demographic enrichment using the proposed IDCM approach is a probability distribution over the possible outcome values (POVs) of the enriched socio-demographic attributes. This probabilistic nature of the IDCM approach provides a means to preserve individual-level privacy whilst retaining enough information for sample-level analysis. This offers a valuable tool to address the need to accommodate the increasingly stringent policies surrounding privacy in big data use while still capitalising on the opportunities for novel applications of such data. It also allows a convenient way to enrich data that have been aggregated to arbitrary large groups, e.g. to maintain k -anonymity.

The above discussion suggests that the proposed IDCM approach can be employed effectively to infer the socio-demographic attributes of respondents in a large number of contexts where their behaviour patterns are observed, without infringing the privacy protection regulations.

8.3.2 Transferability and generalisability

As is argued and extensively discussed across this thesis, the proposed IDCM framework differs from the other existing socio-demographic enrichment approaches in terms of its sound basis in microeconomic behaviour theory, by which it retains the causal dependence between the input behaviour features and the enriched socio-demographic attributes. This contributes to two properties of the IDCM socio-demographic enrichment framework, as explained below.

- 1) The enrichment results using the IDCM approach are able to reflect the actual variation pattern between either the input or the output, given the change in the other, or according to other external factors. This allows more accurate enrichment results than can be delivered by other methods which ignore the causal dependence between the two.

- 2) The IDCM framework is capable of detecting, estimating and interpreting its socio-demographic enrichment performance given the data condition underlying the enrichment sample based on the established the IDCM performance theory.

The above two properties of the developed IDCM socio-demographic enrichment framework assures the effectiveness of the socio-demographic enrichment using the IDCM approach, as well as its transferability and generalisability under various data conditions and across application contexts. The transferability and generalisability of an enrichment approach is important for the following two reasons.

- 1) The increasing difficulties associated with the collection of ‘thick’ data limits their availability for the estimation of the IDCM kernel, as well as for training other socio-demographic enrichment methods such as machine learning algorithms. This therefore requires an understanding of the performance of the enrichment methods given different data conditions and therefore to comprehend the feasibility of using corresponding methods for specific data conditions.
- 2) The transferability and generalisability determines the extent to which a socio-demographic enrichment approach can be employed across different application contexts, which is an important measurement for assessing the value of an enrichment method.

In light of the discussion above, the proved transferability and generalisability of the proposed IDCM socio-demographic enrichment framework acknowledges its value in different application contexts and warrants further investigations seeking to extend and improve it.

8.3.3 Operational efficiency

In the past, transport planning was performed based on inefficient resources such as historical averages or even personal experience (Borgi, Zoghلامي and Abed, 2017). The advances in novel big data analytics, however, such as automatic data processing, facilitate the process of real-time information from sources such as sensors, hence optimising resource consumption and permitting better decision-making capabilities and improved operational performance (Lin and Zheng, 2013). This provides the opportunity to enhance the operational efficiency of transport and logistic modelling by taking advantage of the high volume and velocity of transport big data as well as the advanced modelling analytics. This in turn can significantly

reduce the risks involved in long-term investments and increase the predicative values (Sivan, Johns and Venugopal, 2014).

Since microeconomic theory suggests that people's behaviour patterns are at least partly driven by their socio-demographic attributes, an understanding of the personal information about an individual, or a group of respondents, through the enrichment of socio-demographic attributes using the IDCM approach can effectively facilitate the prediction of their current and future behaviour. This better prediction feeds through to the increased reliability of the corresponding transport models and better transport planning decision-making, enabling the effectiveness of novel big data analytical tools adopted in transport planning and management and enhancing the capability of the modern, self-organising transport system (Borgi, Zoghلامي and Abed, 2017).

8.3.4 Customer experience

Another important implication of extracting additional values from big data sources concerns the acquisition of customer insights. Many meaningful variables, such as sales activities, complaint management details or customer service inquiries, can be inferred from the data sourced from the distribution and transport network. This information can support the management of customer relations and the understanding of the demands (Borgi, Zoghلامي and Abed, 2017). Moreover, this is not only important in transport and logistics but is also crucial for any other service providers, e.g. mobile network providers and shopping malls, in which customer loyalty management and customer feedback analysis can be achieved in presence of the various big data sources and corresponding analytics tools and capabilities.

In addition, the socio-demographic enrichment using the IDCM approach furnishes service providers with a more comprehensive understanding of the personalities and preferences of their users, hence enabling them to deliver more personalised services to improve customer satisfaction and minimise customer attrition or churn (Coussement, Lessmann and Verstraeten, 2017). This suggests that the use of the IDCM approach in the inference of the socio-demographic attributes of customers or clients has potential for any service sectors.

8.4 Limitations and ongoing works

Despite the substantial contributions and implications that may result from the use of the IDCM framework for the socio-demographic enrichment of big data sources, the framework developed so far contains a number of limitations. These motivate the corresponding avenues for further development and are discussed in the following paragraphs.

1) Number of choice observations per individual addressed by IDCM theory

For trackability purposes, the IDCM performance theory established so far is proven only in the enrichment of a single binary socio-demographic attribute, given a single observation of a specific binary behaviour pattern for each individual. This is unrealistic for most applications, however, since individuals' choice behaviours can be observed more than once, even if only a single type of behaviour pattern is observed. This limitation then loses some of the meaning to extract additional values from big data, and does not make effective use of the big datasets that typically contain a massive number of observations. The first step to improve the IDCM performance theory therefore involves extending the enrichment so as to accommodate multiple observations of a single type of behaviour pattern per individual.

The challenge involved with the above extension concerns the trade-off between the tractability of the IDCM performance theory and the realistic behaviour patterns. For a single type of choice behaviour made multiple times by each individual, as is proposed in Section 4.1, a joint likelihood is considered to avoid different enrichment results for the same socio-demographic attribute of the same respondent. In terms of the econometric formulation, the outcome value of the enriched socio-demographic attribute that is obtained by maximising such a joint likelihood can be derived equivalently by maximising the corresponding log-likelihood function, which is addable. In contrast, the IDCM performance theory involves estimating the probability of each POV of the enriched socio-demographic attribute given the POVs of the choice behaviour made for various times by different respondents. This may increase the difficulty in expressing the enrichment performance of the IDCM approach and the enrichment efficiency of the corresponding quasi choice-making channel as functions of the assumed recalibrated constant. This in turn presents a forbidding challenge in terms of generalising the link between the enrichment performance and the underlying data distribution. Nonetheless, this extension is of great importance to increase the practical value of the IDCM framework.

2) Number of possible outcome values of behaviour pattern addressed by IDCM theory

The IDCM performance theory is currently proven in the binary enrichment context only. In reality, however, either the input behaviour feature or the enriched socio-demographic attribute can have more than just two possible outcome values (POVs). Although any variables can be aggregated to obtain a binary variable, such aggregation of POVs of discrete variables, or the discretisation of continuous variables, would typically lead to a loss of information in the corresponding variables, in the sense of Shannon's entropy. Hence, the other stream of ongoing work to extend the IDCM performance theory entails addressing the involved variables, i.e. the input behaviour features and the enriched attributes, of different nature (e.g. continuous or discrete and with more than two POVs). One of the most direct ways to deal with behaviour patterns with more than two POVs could be to model them as multiple binary choice variables.

3) Generalisation of the IDCM kernel structure to the GEV family

For trackability purposes, multinomial logit (MNL) and binary logit (BL) models are adopted as the structure of the IDCM kernel for the econometric formulation of the IDCM approach developed so far. Although MNL and BL have remained the workhorse structure of discrete choice modelling for several decades and still serves as the benchmark against which more complex frameworks are evaluated, a number of more complicated DCM structures have been developed to address some of the restrictions of MNL and BL. These give greater flexibility for modelling the dependence pattern of the input behaviour features on the properties of choice alternatives, as well as the attributes of the choice-maker.

Specifically, one of the major limitations of MNL and BL models is their inability to accommodate modelled choice alternatives that have common unobserved attributes, resulting from the assumption that the unobservable components of utility are independent. This is termed as the '*independence from irrelevant alternatives*' (IIA) property of MNL and BL models, which can lead to their failure to capture possible correlations between choice alternatives. The structures developed to address the IIA of MNL and BL models include:

- Nested logit model;
- Generalised Extreme Value (GEV) models; and
- Mixture (mixed) logit model.

Note that both the MNL and the BL models are special cases of the nested logit model, while all these models, viz. MNL, BL and nested logit models, are particular forms of the GEV model.

Still, the availability of sufficient extensions (generalisation) of the MNL and the BL models, as listed above, allows the currently established IDCM formulation to be extended by employing a more flexible DCM structure as the IDCM kernel. This would accommodate the possible correlations between choice alternatives (if any) and therefore capture the dependence pattern between the choice behaviour features and the enriched attribute in a way that captures the unobserved correlation between choice alternatives, enabling the use of a wider range of choice behaviour features as the inputs for the enrichment. Moreover, using an IDCM kernel with a more accurate structure would increase the overall goodness-of-fit of the model, as well as possibly increase the accuracy at which the sign of the taste parameters in this DCM is estimated. This would further enhance the effectiveness of the IDCM approach, which is driven by the accuracy of the sign of the taste parameter for the enriched attribute in the IDCM kernel, as explained in Chapter 7 of this thesis.

Correspondingly, the accompanying IDCM performance theory needs also to be extended to accommodate the above generalisation for estimating the enrichment performance of the IDCM approach using an IDCM kernel taking the form of the GEV model.

4) Number of input choice behaviour features

Although the econometric formulation of the IDCM approach developed so far can accommodate multiple observations of a specific choice behaviour pattern, the sole inclusion of a single behaviour pattern as the input feature for the enrichment can possibly limit the enrichment performance using the IDCM approach. This is because the pattern for a single behaviour feature may vary slightly as the number of observations increases, as reflected by the mutual information underlying its correlation with the enriched attribute, and can be restricted *ceteris paribus*. This implies that additional information about the socio-demographic attributes can only be obtained from other behaviour patterns of an individual.

The above limitation of the currently established IDCM formulation leads to another possible extension of the proposed enrichment approach, which involves the simultaneous adoption of multiple IDCM kernels for more than only a single behaviour pattern, each capturing the joint likelihood function for an individual decision maker making the same type of choice behaviour any number of times. Unlike existing methods which model multiple behaviour features as predictors for enriching the absent attributes, the simultaneous consideration of a number of IDCM kernels using the IDCM socio-demographic enrichment approach allows several input

behaviour features to be modelled simultaneously, while retaining the microeconomic theories underlying the proposed enrichment. This is supported by the comparative analysis presented in Chapters 5 and 6 of this thesis, which indicates that the enrichment performance of the IDCM approach using the MAP estimator is similar to that of LGR and the SVM in the binary enrichment context, using a single behaviour feature as the input of the enrichment. This demonstrates the trivial difference between these approaches in terms of the enrichment performance when only a single input feature is modelled. When multiple input behaviour features are involved, however, the enrichment performance using the IDCM approach is expected to differ from that of the other existing methods, such as LGR and the SVM, owing to the different manner in which the input features enter the enrichment formulation. Given that multiple behaviour patterns are assumed to provide possibly more information about the enriched socio-demographic attribute than only a single feature, the proposed extension to the econometric formulation of the IDCM approach is expected to improve its enrichment performance, while still maintaining the capability to estimate the performance prior to the enrichment *per se*, since the theoretical underpinnings are preserved.

Similarly to the enrichment given a single behaviour pattern with more than two POVs, or made more than once, the challenge involved with the simultaneous inversion of multiple IDCM kernels also concerns the trade-off between the tractability of IDCM performance theory and the actual probability distributions. Although the outcome value of the enriched socio-demographic attribute can be obtained easily by maximising the corresponding joint log-likelihood function, developing the corresponding IDCM performance theory is challenging. Particularly, the corresponding theory involves the estimation of the probability distribution of the enriched socio-demographic attribute given different combinations of the POVs of a number of behaviour patterns possibly observed for various times. This could increase the complexity of the expressions for the enrichment performance in the IDCM approach, and in the corresponding data conditions, as well as the link between the two. Despite the challenge involved with this extension, a solution is required in order to improve the applicability and generalisability of the IDCM socio-demographic enrichment framework.

References

- A short history of the postcode* (2005) *The Independent*.
- Abdi, H. and Williams, L. J. (2010) 'Principal component analysis', *Wiley Interdisciplinary Reviews: Computational Statistics*. doi: 10.1002/wics.101.
- Abkowitz, M. D. (1981) 'An analysis of the commuter departure time decision', *Transportation*. doi: 10.1007/BF00148464.
- Akaike, H. (1974) 'A New Look at the Statistical Model Identification', *IEEE Transactions on Automatic Control*. doi: 10.1109/TAC.1974.1100705.
- Alexander, L. *et al.* (2015) 'Origin-destination trips by purpose and time of day inferred from mobile phone data', *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2015.02.018.
- Alpaydin, E. (2014) 'Introduction to Machine Learning Ethem Alpaydin.', *Introduction to Machine Learning, Third Edition*.
- Alsger, A. *et al.* (2016) 'Validating and improving public transport origin-destination estimation algorithm using smart card fare data', *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2016.05.004.
- Amemiya, T. (1985) *Advanced econometrics*. Harvard University Press.
- Andridge, R. R. and Little, R. J. A. (2010) 'A review of hot deck imputation for survey non-response', *International Statistical Review*. doi: 10.1111/j.1751-5823.2010.00103.x.
- Andrienko, G. *et al.* (2011) 'From movement tracks through events to places: Extracting and characterizing significant places from mobility data', *VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings*, pp. 161–170. doi: 10.1109/VAST.2011.6102454.
- Andrienko, G., Andrienko, N. and Heurich, M. (2011) 'An event-based conceptual model for context-aware movement analysis', *International Journal of Geographical Information Science*. doi: 10.1080/13658816.2011.556120.
- Antenucci, D. *et al.* (2014) 'Using Social Media to Measure Labor Market Flows', *National Bureau of Economic Research*. doi: 10.3386/w20010.
- Asakura, Y., Hato, E. and Kashiwadani, M. (2000) 'Origin-destination matrices estimation model using automatic vehicle identification data and its application to the Han-Shin

- expressway network', *Transportation*. doi: 10.1023/A:1005239823771.
- Ashbrook, D. and Starner, T. (2003) 'Using GPS to learn significant locations and predict movement across multiple users', *Personal and Ubiquitous Computing*. doi: 10.1007/s00779-003-0240-0.
- Auld, J. *et al.* (2015) 'Demographic Characterization of Anonymous Trace Travel Data', *Transportation Research Record: Journal of the Transportation Research Board*, 2526, pp. 19–28. doi: 10.3141/2526-03.
- Awange, J. L. (2012) 'The global positioning system', in *Environmental Science and Engineering (Subseries: Environmental Science)*. doi: 10.1007/978-3-540-88256-5_3.
- Axhausen, K. *et al.* (2004) 'Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information', *Transportation Research Record*, 1870(August), pp. 46–54. doi: 10.3141/1870-06.
- Bachir, D. *et al.* (2019) 'Inferring dynamic origin-destination flows by transport mode using mobile phone data', *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2019.02.013.
- Backus, G. E. and Gilbert, J. F. (1967) 'Numerical application of a formalism for geophysical inversion', *Geophysical Journal of the Royal Astronomical Society*, 13, pp. 247–276.
- Bal, G. (2012) 'Introduction to inverse problems.', in *Lecture Notes-Department of Applied Physics and Applied Mathematics, Columbia University*. New York.
- Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014) *Hierarchical Modeling and Analysis for Spatial Data, Hierarchical Modeling and Analysis for Spatial Data*. doi: 10.1201/b17115.
- Barker, E. (2016) 'Recommendation for Key Management – Part 1: General', *NIST Special Publication 800-57*. doi: 10.6028/NIST.SP.800-57pt3r1.
- Bassett, R. and Deride, J. (2019) 'Maximum a posteriori estimators as a limit of Bayes estimators', *Mathematical Programming*. doi: 10.1007/s10107-018-1241-0.
- Beckmann, N. *et al.* (1990) 'The R*-tree: an efficient and robust access method for points and rectangles', in *Proceedings of the 1990 ACM SIGMOD international conference on Management of data*.
- Ben-Akiva, M. (1974) 'Structure of passenger travel demand models', *Transportation Research Record*.
- Ben-Akiva, M. and Lerman, S. R. (1974) 'Some estimation results of a simultaneous model of auto ownership and mode choice to work', *Transportation*. doi: 10.1007/BF00167966.

- Ben-Akiva, M. and Lerman, S. R. (1985) *Discrete Choice Analysis. Theory and Application to Travel Demand*, MIT Press series in transportation studies. 9. doi: 10.1016/0191-2607(88)90064-7.
- Ben-Akiva, M. and Swait, J. (1986) 'AKAIKE LIKELIHOOD RATIO INDEX.', *Transportation Science*. doi: 10.1287/trsc.20.2.133.
- Bengio, Y., Courville, A. and Vincent, P. (2013) 'Representation learning: A review and new perspectives', *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/TPAMI.2013.50.
- Bennewitz, M. *et al.* (2005) 'Learning motion patterns of people for compliant robot motion', *International Journal of Robotics Research*. doi: 10.1177/0278364904048962.
- Bhat, C. R. (2003) 'Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences', *Transportation Research Part B: Methodological*. doi: 10.1016/S0191-2615(02)00090-5.
- Bierlaire, M. (2003) 'BIOGEME: A free package for the estimation of discrete choice models', in *Swiss transport research conference*.
- Bierlaire, M., Chen, J. and Newman, J. (2013) 'A probabilistic map matching method for smartphone GPS data', *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2012.08.001.
- Bishop, C. M. (2006) *Pattern Recognition and Machine Learning, Information Science and Statistics*.
- Blackmer, W. S. (2016) *GDPR: getting ready for the new EU general data protection regulation*.
- Blumenstock, J. E., Gillick, D. and Eagle, N. (2010) 'Who's calling? Demographics of mobile phone use in Rwanda', in *AAAI Spring Symposium - Technical Report*.
- Bohte, W. and Maat, K. (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*. Elsevier Ltd, 17(3), pp. 285–297. doi: 10.1016/j.trc.2008.11.004.
- Bonnel, P. *et al.* (2015) 'Passive mobile phone dataset to construct origin-destination matrix: Potentials and limitations', in *Transportation Research Procedia*. doi: 10.1016/j.trpro.2015.12.032.
- Borgi, T., Zoghlami, N. and Abed, M. (2017) 'Big data for transport and logistics: A review',

in *Proceedings of International Conference on Advanced Systems and Electric Technologies, IC_ASET 2017*. doi: 10.1109/ASET.2017.7983742.

Bozinovski, S. (1982) 'A self-learning system using secondary reinforcement', *Cybernetics and Systems Research*, pp. 397–402.

Brathwaite, T., Vij, A. and Walker, J. L. (2017) 'Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice'.

Bricka, S. and Bhat, C. R. (2006) 'Comparative analysis of global positioning system-based and travel survey-based data', *Transportation Research Record*. doi: 10.3141/1972-04.

Bridle, J. S. (1990) 'Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters', *Advances in Neural Information Processing Systems*.

Brunauer, R. *et al.* (2013) 'Motion pattern analysis enabling accurate travel mode detection from GPS data only', in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*. doi: 10.1109/ITSC.2013.6728265.

Bui-Thanh, T. and Girolami, M. (2014) 'Solving large-scale PDE-constrained Bayesian inverse problems with Riemann manifold Hamiltonian Monte Carlo', *Inverse Problems*. doi: 10.1088/0266-5611/30/11/114014.

Bwambale, A., Choudhury, C. F. and Hess, S. (2019) 'Modelling trip generation using mobile phone data: A latent demographics approach', *Journal of Transport Geography*. doi: 10.1016/j.jtrangeo.2017.08.020.

Byon, Y.-J., Abdulhai, B. and Shalaby, A. S. (2007) 'Impact of Sampling Rate of GPS-Enabled Cell Phones on Mode Detection and GIS Map Matching Performance', in *Transportation Research Board 86th Annual Meeting*.

Bzdok, D., Altman, N. and Krzywinski, M. (2018) 'Statistics versus machine learning', *Nature Methods*. doi: 10.1038/nmeth.4642.

Calabrese, F. *et al.* (2010) 'The geography of taste: Analyzing cell-phone mobility and social events', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-642-12654-3_2.

Calastri, C. *et al.* (2019) 'Mode choice with latent availability and consideration: Theory and a case study', *Transportation Research Part B: Methodological*. doi: 10.1016/j.trb.2017.06.016.

- Calvetti, D., Kaipio, J. P. and Somersalo, E. (2014) ‘Inverse problems in the Bayesian framework’, *Inverse Problems*. doi: 10.1088/0266-5611/30/11/110301.
- Castillo, E., Menéndez, J. M. and Jiménez, P. (2008) ‘Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations’, *Transportation Research Part B: Methodological*. doi: 10.1016/j.trb.2007.09.004.
- Chang, J. S. *et al.* (2014) ‘Comparative analysis of trip generation models: Results using home-based work trips in the Seoul metropolitan area’, *Transportation Letters*. doi: 10.1179/1942787514Y.0000000011.
- Chen, C. *et al.* (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*. doi: 10.1016/j.tra.2010.08.004.
- Chen, C. *et al.* (2016) ‘The promises of big data and small data for travel behavior (aka human mobility) analysis’, *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2016.04.005.
- Chung, E.-H. and Shalaby, A. (2005) ‘A Trip Reconstruction Tool for GPS-based Personal Travel Surveys’, *Transportation Planning and Technology*, 28(5), pp. 381–401. doi: 10.1080/03081060500322599.
- Cortes, C. and Vapnik, V. (1995) ‘Support-Vector Networks’, *Machine Learning*. doi: 10.1023/A:1022627411411.
- Coussement, K., Lessmann, S. and Verstraeten, G. (2017) ‘A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry’, *Decision Support Systems*. doi: 10.1016/j.dss.2016.11.007.
- Cover, T. M. and Thomas, J. A. (2005) *Elements of Information Theory, Elements of Information Theory*. doi: 10.1002/047174882X.
- Cramer, J. S. (2005) ‘The Origins of Logistic Regression’, *SSRN Electronic Journal*. doi: 10.2139/ssrn.360300.
- Cuingnet, R. *et al.* (2011) ‘Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome’, *Medical Image Analysis*. doi: 10.1016/j.media.2011.05.007.
- Dabiri, S. and Heaslip, K. (2018) ‘Inferring transportation modes from GPS trajectories using a convolutional neural network’, *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2017.11.021.

- Deng, Z. and Ji, M. (2010) 'Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach.', *Traffic and Transportation Studies*, pp. 768–777.
- Ding, C. *et al.* (2017) 'Exploring the influence of built environment on travel mode choice considering the mediating effects of car ownership and travel distance', *Transportation Research Part A: Policy and Practice*. doi: 10.1016/j.tra.2017.04.008.
- de Dios Ortuzar, J. and Willumsen, L. G. (2011) *Modelling Transport*. John Wiley & Sons.
- Dodge, S., Weibel, R. and Forootan, E. (2009) 'Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects', *Computers, Environment and Urban Systems*. doi: 10.1016/j.compenvurbsys.2009.07.008.
- Dong, Y. *et al.* (2014) 'Inferring user demographics and social strategies in mobile social networks', in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi: 10.1145/2623330.2623703.
- Dow, J. C. R. and Dicks-mireaux, L. A. (1958) 'The excess demand for labour a study of conditions in Great Britain, 1946-56', *Oxford Economic Papers*. doi: 10.1093/oxfordjournals.oep.a040791.
- Eberhart, R. and Kennedy, J. (1995) 'New optimizer using particle swarm theory', in *Proceedings of the International Symposium on Micro Machine and Human Science*. doi: 10.1109/mhs.1995.494215.
- Ectors, W. *et al.* (2017) 'Developing an optimised activity type annotation method based on classification accuracy and entropy indices', *Transportmetrica A: Transport Science*. doi: 10.1080/23249935.2017.1331275.
- Edwards, W. (1962) 'Utility, subjective probability, their interaction, and variance preferences', *Journal of Conflict Resolution*. doi: 10.1177/002200276200600106.
- Ester, M. *et al.* (1996) 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise', *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. doi: 10.1.1.71.1980.
- Everitt, B. (1980) 'Cluster analysis', *Quality and Quantity*. doi: 10.1007/BF00154794.
- Felbo, B. *et al.* (2017) 'Modeling the temporal nature of human behavior for demographics prediction', in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, pp. 140–152.
- Feng, T. and Timmermans, H. J. P. (2013) 'Transportation mode recognition using GPS and

- accelerometer data', *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2013.09.014.
- Fisher, R. A. (1912) 'The Maximum-Likelihood-Method', *Messenger in Mathematics*, 41, pp. 155–160.
- Fisher, R. A. (1922) 'On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P', *Journal of the Royal Statistical Society*. doi: 10.2307/2340521.
- Frias-Martinez, V., Frias-Martinez, E. and Oliver, N. (2010) 'A gender-centric analysis of calling behavior in a developing economy using Call Detail Records', in *AAAI Spring Symposium - Technical Report*.
- Gandomi, A. and Haider, M. (2015) 'Beyond the hype: Big data concepts, methods, and analytics', *International journal of information management*. Elsevier, 35(2), pp. 137–144.
- Gartner (2015) 'Survey Analysis: Practical Challenges Mount as Big Data Moves to Mainstream.', *Gartner Symposium/ITxpo*.
- Ge, Q. and Fukuda, D. (2016) 'Updating origin-destination matrices with aggregated data of GPS traces', *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2016.06.002.
- Gebru, T. *et al.* (2017) 'Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1700035114.
- Giannotti, F. *et al.* (2007) 'Trajectory pattern mining', in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*. doi: 10.1145/1281192.1281230.
- Giannotti, F. and Nanni, M. (2007) 'Trajectory pattern mining', ... *Discovery and Data Mining*, p. 339. doi: 10.1145/1281192.1281230.
- 'GNSS-Global Navigation Satellite Systems: GPS, GLONASS, Galileo, and more' (2008) *Choice Reviews Online*. doi: 10.5860/choice.45-6185.
- Goetzke, F. (2008) 'Network effects in public transit use: Evidence from a spatially autoregressive mode choice model for New York', *Urban Studies*. doi: 10.1177/0042098007085970.
- Gong, L. *et al.* (2014) 'Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies', *Procedia - Social and Behavioral Sciences*. Elsevier B.V., 138(0), pp. 557–565. doi: 10.1016/j.sbspro.2014.07.239.

- Gong, L. *et al.* (2015) 'Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines', *Journal of Modern Transportation*. doi: 10.1007/s40534-015-0079-x.
- González, G. *et al.* (2007) 'Embedding emotional context in recommender systems', in *Proceedings - International Conference on Data Engineering*. doi: 10.1109/ICDEW.2007.4401075.
- Goodstein, R. L. *et al.* (1954) 'Undecidable Theories', *The Mathematical Gazette*. doi: 10.2307/3611179.
- Greene, W. W. H. . (2012) *Econometric analysis 7th Ed*, Prentice Hall.
- Haaf, C. G. *et al.* (2014) 'Sensitivity of Vehicle Market Share Predictions to Discrete Choice Model Specification', *Journal of Mechanical Design, Transactions of the ASME*. doi: 10.1115/1.4028282.
- Hadamard, J. (1902) 'Sur les problèmes aux dérivées partielles et leur signification physique', *Princeton Uni. Bull.* (1902).
- Hammersley, J. M. and Handscomb, D. C. (1964) *Monte Carlo Methods*, *Monte Carlo Methods*. doi: 10.1007/978-94-009-5819-7.
- Han, G. and Sohn, K. (2016) 'Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model', *Transportation Research Part B: Methodological*. doi: 10.1016/j.trb.2015.11.015.
- Harris, J. W. and Stocker, H. (1998) *Handbook of Mathematics and Computational Science*, *Handbook of Mathematics and Computational Science*. doi: 10.1007/978-1-4612-5317-4.
- Hassan, R. *et al.* (2005) 'A comparison of particle swarm optimization and the genetic algorithm', in *Collection of Technical Papers - AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*. doi: 10.2514/6.2005-1897.
- Hastie, T. and Tibshirani, R. (1987) 'Non-Parametric Logistic and Proportional Odds Regression', *Applied Statistics*. doi: 10.2307/2347785.
- Hawkins, D. M. (2004) 'The Problem of Overfitting', *Journal of Chemical Information and Computer Sciences*. doi: 10.1021/ci0342472.
- Hendry, M. (2007) *Multi-application smart cards: technology and applications*. Cambridge university press.
- Hensher, D. A. and Greene, W. H. (2003) 'The mixed logit model: The state of practice', *Transportation*. doi: 10.1023/A:1022558715350.

- Herrera-Yagüe, C. and Zufiria, P. J. (2012) ‘Prediction of telephone user attributes based on network neighborhood information’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-642-31537-4_50.
- Herrera, J. C. *et al.* (2010) ‘Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment’, *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2009.10.006.
- Hersch, J. (1996) ‘Smoking, seat belts, and other risky consumer decisions: Differences by gender and race’, *Managerial and Decision Economics*. doi: 10.1002/(SICI)1099-1468(199609)17:5<471::AID-MDE789>3.0.CO;2-W.
- Hill, M. (1997) ‘Learning Inductive learning Analytical learning’, *Machine Learning*.
- Hopfield, J. J. (1982) ‘Neural networks and physical systems with emergent collective computational abilities.’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.79.8.2554.
- Horak, R. (2006) *Telecommunications and Data Communications Handbook*, *Telecommunications and Data Communications Handbook*. doi: 10.1002/9780470127247.
- Hung, C. C., Peng, W. C. and Lee, W. C. (2015) ‘Clustering and aggregating clues of trajectories for mining trajectory patterns and routes’, *VLDB Journal*. doi: 10.1007/s00778-011-0262-6.
- Iqbal, M. S. *et al.* (2014) ‘Development of origin-destination matrices using mobile phone call data’, *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2014.01.002.
- Isaacman, S. *et al.* (2011) ‘Identifying important places in people’s lives from cellular network data’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-642-21726-5_9.
- Jahani, E. *et al.* (2017) ‘Improving official statistics in emerging markets using machine learning and mobile phone data’, *EPJ Data Sci.* The Author(s). doi: 10.1140/epjds/s13688-017-0099-3.
- Jara-Díaz, S. (2007) *Transport economic theory*.
- Kaelbling, L. P., Littman, M. L. and Moore, A. W. (1996) ‘Reinforcement learning: A survey’, *Journal of Artificial Intelligence Research*. doi: 10.1613/jair.301.

- Kalos, M. H. and Whitlock, P. A. (2008) ‘What is Monte Carlo?’, in *Monte Carlo Methods*. doi: 10.1002/9783527626212.ch1.
- Kass, R. E. *et al.* (1997) ‘Markov Chain Monte Carlo in Practice.’, *Journal of the American Statistical Association*. doi: 10.2307/2965438.
- Kim, B., Khanna, R. and Koyejo, O. (2016) ‘Examples are not enough, learn to criticize! Criticism for interpretability’, in *Advances in Neural Information Processing Systems*.
- Kitchin, R. and McArdle, G. (2016) ‘What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets’, *Big Data and Society*. doi: 10.1177/2053951716631130.
- Kohavi, R. (1995) ‘A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection’, in *Appears in the International Joint Conference on Artificial Intelligence (IJCAI)*. doi: 10.1067/mod.2000.109031.
- Kroese, D. P. *et al.* (2014) ‘Why the Monte Carlo method is so important today’, *Wiley Interdisciplinary Reviews: Computational Statistics*. doi: 10.1002/wics.1314.
- Kusakabe, T. and Asakura, Y. (2014) ‘Behavioural data mining of transit smart card data: A data fusion approach’, *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2014.05.012.
- Kusakabe, T., Tsubota, T. and Bhaskar, A. (2016) ‘Validation study of naïve Bayes probabilistic model for transit passengers’ trip purpose estimation : A case study exploiting detailed Brisbane household travel survey data’, *Transport Research Board*.
- Lancaster (2008) ‘A New Approach to Consumer Theory Author (s): Kelvin J . Lancaster Source : The Journal of Political Economy , Vol . 74 , No . 2 (Apr ., 1966), pp . 132-157 Published by : The University of Chicago Press Stable URL : <http://www.jstor.org/stable/182883>’, 74(2), pp. 132–157.
- Lancaster, K. J. (1966) ‘A New Approach to Consumer Theory’, *Journal of Political Economy*. doi: 10.1086/259131.
- Lancsar, E., Fiebig, D. G. and Hole, A. R. (2017) ‘Discrete Choice Experiments: A Guide to Model Specification, Estimation and Software’, *PharmacoEconomics*. doi: 10.1007/s40273-017-0506-4.
- Leape, J. (2006) ‘The London congestion charge’, *Journal of Economic Perspectives*. doi: 10.1257/jep.20.4.157.
- Lee, B., Lim, C. and Lee, K. (2017) ‘Classification of indoor-outdoor location using

combined global positioning system (GPS) and temperature data for personal exposure assessment', *Environmental Health and Preventive Medicine*. doi: 10.1186/s12199-017-0637-4.

Li, R. *et al.* (2017) 'The Learning and Prediction of Application-Level Traffic Data in Cellular Networks', *IEEE Transactions on Wireless Communications*. doi: 10.1109/TWC.2017.2689772.

Liao, L., Fox, D. and Kautz, H. (2007) 'Extracting places and activities from GPS traces using hierarchical conditional random fields', *International Journal of Robotics Research*. doi: 10.1177/0278364907073775.

Lin, X. and Zheng, X. (2013) 'A cloud-enhanced system architecture for logistics tracking services', in *International Conference on Computer, Networks and Communication Engineering (ICCNCE 2013)*. Atlantis Press.

Liu, J. *et al.* (2020) 'A dynamic OD prediction approach for urban networks based on automatic number plate recognition data', in *Transportation Research Procedia*. doi: 10.1016/j.trpro.2020.03.137.

Llorente, A. *et al.* (2015) 'Social Media Fingerprints of Unemployment', *PLoS ONE*. doi: 10.1371/journal.pone.0128692.

Lou, Y. *et al.* (2009) 'Map-matching for low-sampling-rate GPS trajectories', in *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. doi: 10.1145/1653771.1653820.

Lucarotti, P. S. K. (1977) 'Car availability-the fundamental modal split', *Transportation Planning and Technology*. doi: 10.1080/03081067708717107.

Machine Learning Versus Statistics: When to use each - healthcare.ai (2019) *Data Science Blog*. Available at: <https://healthcare.ai/machine-learning-versus-statistics-use/> (Accessed: 1 November 2021).

Magno, G. and Weber, I. (2014) 'International gender differences and gaps in online social networks', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-319-13734-6_9.

Malouf, R. (2002) 'A comparison of algorithms for maximum entropy parameter estimation', in. doi: 10.3115/1118853.1118871.

Manski, C. F. (1977) 'The structure of random utility models', *Theory and Decision*. doi:

10.1007/BF00133443.

Mas-Colell, A., Whinston, M. D. and Green, J. R. (1995) *Microeconomic theory*. New York: Oxford University Press.

Mascret, A. *et al.* (2006) ‘Coastline matching process based on the Discrete Fréchet Distance’, in *Progress in Spatial Data Handling - 12th International Symposium on Spatial Data Handling, SDH 2006*. doi: 10.1007/3-540-35589-8_25.

McFadden, D. L. (1974) ‘Conditional Logit Analysis of Qualitative Choice Behavior’, in *Frontiers in Econometrics*.

McNally, M. G. (2007) ‘The Four-Step Model’, in. doi: 10.1108/9780857245670-003.

Meeter, D. (1966) ‘Introduction to Numerical Analysis’, *Technometrics*. doi: 10.1080/00401706.1966.10490427.

Mendis, A. (2019) *Statistical Modelling vs Machine Learning*, *KDnuggets*. Available at: <https://www.kdnuggets.com/2019/08/statistical-modelling-vs-machine-learning.html> (Accessed: 1 November 2021).

Miller, T. (2019) ‘Explanation in artificial intelligence: Insights from the social sciences’, *Artificial Intelligence*. doi: 10.1016/j.artint.2018.07.007.

Mitchell, T. M. (2006) ‘The Discipline of Machine Learning’, *Machine Learning*. doi: 10.1080/026404199365326.

Montoliu, R., Blom, J. and Gatica-Perez, D. (2013) ‘Discovering places of interest in everyday life from smartphone data’, *Multimedia Tools and Applications*, 62(1), pp. 179–207. doi: 10.1007/s11042-011-0982-z.

Moreira-Matias, L. *et al.* (2016) ‘Time-evolving O-D matrix estimation using high-speed GPS data streams’, *Expert Systems with Applications*. doi: 10.1016/j.eswa.2015.08.048.

Morikawa, T., Ben-Akiva, M. and McFadden, D. (2002) ‘Discrete choice models incorporating revealed preferences and psychometric data’, *Advances in Econometrics*. doi: 10.1016/S0731-9053(02)16003-8.

Morimura, T. and Kato, S. (2012) ‘Statistical Origin-destination generation with multiple sources’, in *Proceedings - International Conference on Pattern Recognition*.

Mottini, A. and Acuna-Agost, R. (2016) ‘Relative Label Encoding for the Prediction of Airline Passenger Nationality’, in *IEEE International Conference on Data Mining Workshops, ICDMW*. doi: 10.1109/ICDMW.2016.0100.

Munizaga, M. A. and Palma, C. (2012) ‘Estimation of a disaggregate multimodal public

- transport Origin-Destination matrix from passive smartcard data from Santiago, Chile’, *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2012.01.007.
- Ng, A. (2000) ‘CS229 Lecture notes’, *CS229 Lecture notes*. doi: 10.1111/j.1466-8238.2009.00506.x.
- Nizami, L. (2019) ‘Glanville’s “Black box”: What can an Observer know?’, in *2019 AISB Convention*.
- O’Sullivan, F. (1986) ‘A statistical perspective on ill-posed inverse problems’, *Statistical Science*, 1(4), pp. 502–518. doi: 10.1214/ss/1177013530.
- OCSI (2019) *LSOAs, LEPs and lookups : A beginner’s guide to statistical geographies*. Available at: <https://ocsi.uk/2019/03/18/lsoas-leps-and-lookups-a-beginners-guide-to-statistical-geographies/> (Accessed: 8 December 2020).
- Office for National Statistics (2011) *2011 Census: Population Estimates for the United Kingdom, UK Census*.
- Office for National Statistics (2016a) *Census Geography*. Available at: <https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography>.
- Office for National Statistics (2016b) *UK population by postcode*. Available at: <https://cy.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/ukpopulationbypostcode>.
- Office for National Statistics (2019) *Census geography - Office for National Statistics, Office for National Statistics*.
- Oldenburg, D. W. (1984) ‘An introduction to linear inverse theory’, *IEEE Transactions on Geoscience and Remote Sensing*, GE-22(6), pp. 665–674.
- Onay, C. and Öztürk, E. (2018) ‘A review of credit scoring research in the age of Big Data’, *Journal of Financial Regulation and Compliance*. doi: 10.1108/JFRC-06-2017-0054.
- Van Otterlo, M. and Wiering, M. (2012) ‘Reinforcement learning and markov decision processes’, in *Adaptation, Learning, and Optimization*. doi: 10.1007/978-3-642-27645-3_1.
- Parry, K. and Hazelton, M. L. (2012) ‘Estimation of origin-destination matrices from link counts and sporadic routing data’, *Transportation Research Part B: Methodological*. doi: 10.1016/j.trb.2011.09.009.
- Pawlak, J., Zolfaghari, A. and 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’, in *International Choice Modelling Conference*

2015.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine learning in Python', *the Journal of machine Learning research*, 12, pp. 2825–2830.

Pfoser, D. and Jensen, C. S. (1999) 'Capturing the uncertainty of moving-object representations', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/3-540-48482-5_9.

Plackett, R. L. (1983) 'Karl Pearson and the Chi-Squared Test', *International Statistical Review / Revue Internationale de Statistique*. doi: 10.2307/1402731.

Poushter, J. (2016) *Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies*, Pew Research Center.

Pronzalo, L. and Walter, E. (1997) *Identification of parametric models from experimental data*, Springer, Berlin Wentzell PD, Andrews DT, Hamilton

Proserpio, D., Counts, S. and Jain, A. (2016) 'The psychology of job loss: Using social media data to characterize and predict unemployment', in *WebSci 2016 - Proceedings of the 2016 ACM Web Science Conference*. doi: 10.1145/2908131.2913008.

Rajula, H. S. R. *et al.* (2020) 'Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment', *Medicina (Lithuania)*. doi: 10.3390/medicina56090455.

Rao, W. *et al.* (2018) 'Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data', *Transportation Research Part C: Emerging Technologies*. doi: 10.1016/j.trc.2018.07.002.

Roberts, J. H. and Urban, G. L. (1988) 'Modeling Multiattribute Utility, Risk, and Belief Dynamics for New Consumer Durable Brand Choice', *Management Science*. doi: 10.1287/mnsc.34.2.167.

Royal mail (2004) *Address Management Guide (4th Edition)*.

Russell, S. and Norvig, P. (2010) *Artificial Intelligence A Modern Approach Third Edition*, Pearson. doi: 10.1017/S0269888900007724.

Rymarczyk, T. *et al.* (2019) 'Logistic regression for machine learning in process tomography', *Sensors (Switzerland)*. doi: 10.3390/s19153400.

Sagiroglu, S. and Sinanc, D. (2013) 'Big data: A review', in *Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013*. doi: 10.1109/CTS.2013.6567202.

- Samuel, A. L. (2000) ‘Some studies in machine learning using the game of checkers’, *IBM Journal of Research and Development*. doi: 10.1147/rd.441.0206.
- Santosa, F. and Symes, W. W. (1986) ‘Linear Inversion of Band-Limited Reflection Seismograms’, *SIAM Journal on Scientific and Statistical Computing*. doi: 10.1137/0907087.
- Sarraute, C., Blanc, P. and Burrioni, J. (2014) ‘A study of age and gender seen through mobile phone usage patterns in Mexico’, in *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. doi: 10.1109/ASONAM.2014.6921683.
- Sbaï, A. *et al.* (2017) ‘Estimation of an Urban OD Matrix Using Different Information Sources’, in *International Conference on Computational Science and Its Applications*.
- Schönfelder, S. *et al.* (2002) ‘Exploring the Potentials of Automatically Collected GPS Data for Travel Behaviour Analysis - A Swedish Data Source’, *GI-Technologien für Verkehr und Logistik, Institut für Geoinformatik, Universität Münster*. doi: 10.3929/ethz-a-004403386.
- Schwarz, G. (1978) ‘Estimating the Dimension of a Model’, *The Annals of Statistics*. doi: 10.1214/aos/1176344136.
- Scott, A. J., Hosmer, D. W. and Lemeshow, S. (1991) ‘Applied Logistic Regression.’, *Biometrics*. doi: 10.2307/2532419.
- Seo, T. *et al.* (2019) ‘Interactive online machine learning approach for activity-travel survey’, *Transportation Research Part B: Methodological*. doi: 10.1016/j.trb.2017.11.009.
- Shannon, C. E. (1948) ‘A mathematical theory of communication’, *The Bell System Technical Journal*. doi: 10.1145/584091.584093.
- Sheth, A. (1996) ‘Data Semantics : what , where and how ?’, in *Data Semantics (IFIP Transactions)*. doi: 10.1007/978-0-387-34913-8_26.
- Sheth, A. (1997) ‘Panel: Data Semantics: what, where and how?’, in *Data Semantics (IFIP Transactions)*, pp. 601–610. doi: 10.1007/978-0-387-34913-8_26.
- Sinha, K. (2014) ‘Semi-supervised learning’, in *Data Classification: Algorithms and Applications*. doi: 10.1201/b17320.
- Sivan, A. P., Johns, J. and Venugopal, J. (2014) ‘Big Data Intelligence in Logistics Based On Hadoop And Map Reduce’, *IJIRSET*, 3, pp. 2634–2640.
- Smith, G. W. (1991) ‘Modeling Security-Relevant Data Semantics’, *IEEE Transactions on Software Engineering*. doi: 10.1109/32.106974.
- Snijders, C., Matzat, U. and Reips, U. (2013) ““Big Data”: Big Gaps of Knowledge in the

Field of Internet Science', *International Journal of Internet Science*.

Srivastava, T. (2015) *Difference between Machine Learning & Statistical Modeling*.

Available at: <https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/> (Accessed: 1 November 2021).

Steele, J. E. *et al.* (2017) 'Mapping poverty using mobile phone and satellite data', *Journal of The Royal Society Interface*, 14(127), p. 20160690. doi: 10.1098/rsif.2016.0690.

Stewart, M. (2019) *The Actual Difference Between Statistics and Machine Learning*.

Available at: [https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3#:~:text=The major difference between machine,about the relationships between variables.&text=Statistics is the mathematical study of data.](https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3#:~:text=The%20major%20difference%20between%20machine,about%20the%20relationships%20between%20variables.&text=Statistics%20is%20the%20mathematical%20study%20of%20data.) (Accessed: 1 November 2021).

Stone, J. V. (2015) *Information theory: a tutorial introduction*. Sebtel Press.

Stopher, P., FitzGerald, C. and Zhang, J. (2008) 'Search for a global positioning system device to measure person travel', *Transportation Research Part C: Emerging Technologies*, 16(3), pp. 350–369. doi: 10.1016/j.trc.2007.10.002.

Supernak, J. and Schoendorfer, D. (1985) 'Automobile Availability and Its Application In Transportation Studies', *Transportation Research Record*, 1037, p. 73.

Tarantola, A. (1987) 'Inverse problem theory: methods for data fitting and model parameter estimation.', *Inverse problem theory: methods for data fitting and model parameter estimation*. doi: 10.1016/0031-9201(89)90124-6.

Theodoridis, S. and Koutroumbas, K. (2009) *Pattern Recognition (Fourth Edition)*, *Pattern Recognition*.

Tominaga, D. ., Nobuto, K. and Masahiro, O. (2000) 'Efficient numerical optimization algorithm based on genetic algorithm for inverse problem', in *Proceedings of the 2nd Annual Conference on Genetic and Evolutionary Computation*. Morgan Kaufmann Publishers Inc., pp. 251–258.

Train, K. E. (2003) *Discrete choice methods with simulation*, *Discrete Choice Methods with Simulation*. doi: 10.1017/CBO9780511753930.

Train, K. E. and McFadden, D. (1978) 'The goods/leisure tradeoff and disaggregate work trip mode choice models', *Transportation Research*, 12(5), pp. 349–353. doi: [http://dx.doi.org/10.1016/0041-1647\(78\)90011-4](http://dx.doi.org/10.1016/0041-1647(78)90011-4).

Transport for London (2011) *Travel in London, Supplementary Report: London Travel*

Demand Survey (LTDS).

Tuomchomtam, S. and Soonthornphisaj, N. (2019) ‘Automatic demographic and personality extraction from anonymous social media’, in *Frontiers in Artificial Intelligence and Applications*. doi: 10.3233/FAIA190181.

UK Government (2018) ‘Data Protection Act 2018’, *www.Gov.uk*.

‘Unsupervised learning: Foundations of neural computation’ (1999) *Computers & Mathematics with Applications*, 38(5–6), p. 256. doi: 10.1016/s0898-1221(99)90165-7.

Vapnik, V. N. (1999) ‘An overview of statistical learning theory’, *IEEE Transactions on Neural Networks*. doi: 10.1109/72.788640.

Wang, B., Gao, L. and Juan, Z. (2018) ‘Travel Mode Detection Using GPS Data and Socioeconomic Attributes Based on a Random Forest Classifier’, *IEEE Transactions on Intelligent Transportation Systems*. doi: 10.1109/TITS.2017.2723523.

Wasserman (2004) ‘All of Statistics : A Concise Course in Statistical Inference Brief Contents’, *Simulation*. doi: 10.1007/978-0-387-21736-9.

Whitley, D. (1994) ‘A genetic algorithm tutorial’, *Statistics and Computing*. doi: 10.1007/BF00175354.

Winters-Hilt, S. *et al.* (2006) ‘Support Vector Machine implementations for classification & clustering’, *BMC Bioinformatics*. doi: 10.1186/1471-2105-7-S2-S4.

Wolf, J. (2004) ‘Applications of new technologies in travel surveys’, *Travel Survey Methods - Standards and Future Directions*. doi: 10.1108/9780080464015-029.

Wolf, J. *et al.* (2014) ‘Applying GPS Data to Understand Travel Behavior, Volume II: Guidelines’, in *Applying GPS Data to Understand Travel Behavior, Volume II: Guidelines*. doi: 10.17226/23436.

Wolf, J., Guensler, R. and Bachman, W. (2001) ‘Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data’, *Transportation Research Record*, 1768(1), pp. 125–134. doi: 10.3141/1768-15.

Woods, W. A. (1975) ‘WHAT’S IN A LINK: Foundations for Semantic Networks A2 - BOBROW, DANIEL G.’, in *Representation and Understanding: Studies in Cognitive Science*.

Zhang, J. *et al.* (2009) ‘Modeling household discrete choice behavior incorporating heterogeneous group decision-making mechanisms’, *Transportation Research Part B: Methodological*. doi: 10.1016/j.trb.2008.05.002.

Zhang, L. *et al.* (2007) 'Study on the method of constructing bus stops OD matrix based on IC card data', in *2007 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2007*. doi: 10.1109/WICOM.2007.780.

Zhao, J., Rahbee, A. and Wilson, N. H. M. (2007) 'Estimating a rail passenger trip origin-destination matrix using automatic data collection systems', *Computer-Aided Civil and Infrastructure Engineering*. doi: 10.1111/j.1467-8667.2007.00494.x.

Zhao, Y., Pawlak, J. and Polak, J. W. (2018) 'Inverse discrete choice modelling: theoretical and practical considerations for imputing respondent attributes from the patterns of observed choices', *Transportation Planning and Technology*. Taylor & Francis, 41(1), pp. 58–79. doi: 10.1080/03081060.2018.1402745.

Zhao, Y., Pawlak, J. and Polak, J. W. (2019) 'Enrichment of transport big data: exploring performance of the inverse discrete choice modelling approach using Monte Carlo simulation', in *Transportation Research Board 98th Annual Meeting*.

Appendix 1: List of publications

As part of this study, the following academic papers were generated:

- 1) Zhao, Y., Pawlak, J. & Polak, J.W. (2017). Inverse Discrete Choice Models (IDCM): Theoretical and Practical Considerations for Imputing Respondent Attributes from the Patterns of Observed Choices. Podium presentation at 49th Universities' Transport Study Group Conference, Trinity College Dublin, Dublin, Ireland.
- 2) Zhao, Y., Pawlak, J. & Polak, J.W. (2017). Privacy-preserving socioeconomic attribute enrichment for mapping of passively derived OD matrices. Podium presentation at 2017 RGS-IBG Annual International Conference, Royal Geographical Society, London, UK.
- 3) Zhao, Y., Pawlak, J. & Polak, J.W. (2018). Inverse Discrete Choice Models (IDCM): Theoretical and Practical Considerations for Imputing Respondent Attributes from the Patterns of Observed Choices. Peer-reviewed paper published by the Journal of Transportation Planning and Technology.
- 4) Zhao, Y., Pawlak, J. & Polak, J.W. (2018). Semantic Enrichment through Inverse Discrete Choice Modelling: A Mutual Information Approach. Extended abstract accepted by 15th International Conference on Travel Behaviour Research, Santa Barbara, California, US.
- 5) Zhao, Y., Pawlak, J. & Polak, J.W. (2018). Semantic enrichment through inverse discrete choice modelling: a mutual information approach. Podium presentation at 7th Symposium of the European Association for Research in Transportation Conference, Athens, Greece.
- 6) Zhao, Y., Pawlak, J. & Polak, J.W. (2019). Enrichment of transport big data: exploring performance of the Inverse Discrete Choice Modelling approach using Monte Carlo simulation. Poster presentation at 98th Annual meeting of Transportation Research Board, Washington, D.C., US.
- 7) Zhao, Y., Pawlak, J. & Polak, J.W. (2019). Sociodemographic Enrichment of Anonymised Mobile Network Data: Application of the Inverse Discrete Choice Modelling Theory. Podium presentation at 2019 International Choice Modelling Conference, Kobe, Japan.
- 8) Zhao, Y., Pawlak, J., Sivakumar, A. & Polak, J.W. (2020) Exploring the link between spatial aggregation of Mobile Network Data and the quality of sociodemographic

enrichment using the Inverse Discrete Choice Modelling approach. Paper accepted by 12th International Conference on Transport Survey Methods.

- 9) Zhao, Y., Pawlak, J. & Sivakumar, A (2021) Theoretical Underpinnings for Performance of Sociodemographic Enrichment using Inverse Discrete Choice Modelling Approach. Paper submitted to Transportation Research Part B.

Appendix 2: Acronyms

Acronym	Meaning
IDCM	Inverse discrete choice modelling
ICT	Information and communication technologies
GDPR	General Data Protection Regulation
DCM	discrete choice model
RO	research objective
LTDS	London Travel Demand Survey
MDP	Markov Decision Process
POV	possible outcome value
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GIS	Geographic Information System
CDR	call detail record
OD	origin-and-destination
DBSCAN	Density-based Spatial Clustering of Applications with Noise
CNN	Convolutional Neural Network
BFS	Brute-force search
GA	Genetic algorithm
PSO	Particle Swarm optimisation
MC	Monte Carlo
PMC	pseudo-Monte Carlo
QMC	quasi-Monte Carlo
MCMC	Markov chain Monte Carlo
GOF	goodness-of-fit
RUM	random utility model
EV1	Type 1 generalised extreme value
BL	binary logit
MNL	multinomial logit
i.i.d	independent and identically distributed
LIP	linear-in-parameters
ASC	alternative specific constants

MLE	maximum likelihood estimation
MAP	maximum <i>a posteriori</i> estimation
LRI	likelihood ratio index
AIC	Akaike information criterion
BIC	Bayesian information criterion
IOI	independence of other individuals
RHS	right-hand side
PCP	percentage correctly predicted
LTDS	London Travel Demand Survey
LGR	logistic regression
SVM	support vector machine
PT	public transport
ML	machine learning
MND	mobile network data
OPC	outward postcode
OA	Output Area
SOA	Super Output Area
MSOA	Middle Layer Super Output Area
LSOA	Lower Layer Super Output Area
DOW	day-of-week
TOD	time-of-the-day
IIA	independence from irrelevant alternatives
GEV	Generalised Extreme Value

Appendix 3: Notations

Notation	Meaning
p	probability.
b	the base of the logarithm.
X	a vector of predictors of a function; a vector of socio-demographic attributes in the context of enrichment.
β	a vector of parameters associating the response with the predictors; a vector of taste parameters for discrete choice models.
$\vec{w}$	the normal vector to the hyperplane of an SVM
c	the parameter of SVMs which represent the trade-off between increasing the margin size and ensuring that the predicted outcome using the SVM lie on the correct side of the margin.
Q, J	functional spaces.
θ	the parameters of the distribution of a set of data points.
ζ	the hyperparameter of the distribution of parameter θ , i.e. $\theta \sim P(\theta \zeta)$;
N	the number of observations in a sample.
i, j, k	choice alternatives.
$\mathcal{C}$	choice set.
U	a utility function which ranks the alternatives in preference order.
A	a set of attributes characterising the choice alternatives.
a	a decision-maker.
V	systematic utility, i.e. the deterministic part of a utility function.
ε	error term, i.e. the stochastic part of a utility function.
$\delta(\cdot)$	indicator function equalling 1 if $(\cdot)$ is true and 0 otherwise.
μ	scale parameter of the common EV1 distribution such that $\text{Var}(\varepsilon) = \frac{\pi^2}{6\mu^2}$.
$\mathcal{L}$	likelihood function.
$\varphi(\cdot)$	prior distribution of the parameter(s) of a model.
Y	a behavioural variable for discrete choice alternatives.
L_a	the number of observations for a specific behaviour made by individual a .
P	probability distribution.
r	the number of parameters in the model.

D_{Est}	estimation sample, i.e. the sample used to estimate the IDCM kernel.
D_{Erc}	enrichment sample, i.e. the sample for which the socio-demographic requires enrichment.
B	the total amount of budget available to an individual.
T	the total amount of time available to an individual
ς	the cost spent on a specific item.
t	the time spent on a specific item.
$\tilde{V}$	indirect utility function.
s	Message.
s'	message obtained by receiver of an information communication channel transmitted from original message s' .
S	manipulated cumulative function.
$\mathcal{H}$	the number of households in a sample.
X	a specific socio-demographic attribute, $X \in \mathbf{X}$.
W	market share.
$I(X, Y)$	the mutual information between two variables X and Y .
$H(X)$	the (Shannon's) entropy of a variable X .
ξ	transmission efficiency of an information communication channel.
η	enrichment efficiency of an information communication channel.
y_0, y_1	the two POVs of the choice variable Y .
x_0, x_1	the two POVs of the socio-demographic attribute X .
$\hat{\beta}_0$	the assumed recalibrated constant, i.e. the pivot.
β_1	the taste parameter associated with the enriched attribute in the IDCM kernel.
p_y	the marginal probability of Y when $Y = y_0$ in the enrichment sample.
θ_0, θ_1	model parameters of a logistic regression function.
γ	kernel-specific parameter of the SVM similarity function.