清 华 大 学

综 合 论 文 训 练

**开 题 报 告**

题目：基于人工神经网络的潜变量

选择模型改善研究

系 别：工业工程

专 业：工业工程

姓 名：张普竣

指导教师：姜海 副教授

2018 年 12 月 21 日

# 引言

## 研究背景

离散选择模型最早起源于交通系统中的出行需求预测问题。出行需求预测是交通系统研究中的重要子问题，这一问题关注所有交通服务设施的使用者的出行需求，主要研究特定区域的住户对不同起点、不同出行方式、不同线路、不同目的地的出行需求，来为交通设施的规划与道路管理政策的制定提供理论指导与技术依据。

离散选择模型广泛引用于交通系统的规划问题中，并且这一方法较之前模型发挥出了更强大的作用。离散选择模型基于经济学与行为学的理论研究，采用随机效用最大化的方法从微观层面来对个体的选择行为进行研究建模。这种非集计的微观层面的方法，解决了旧模型的存在的不可移植的缺点，而且这一方法对于交通政策的敏感性更高，响应更明显，极大地提高了出行需求预测的准确性，也使其对交通政策制定的指导意义更加明显。也正是由于离散选择模型的这些优点，很多的交通规划机构与研究者都采用这一方法来进行出行需求的预测，在多式联运系统调度、定价考量、运营政策制定以及新建交通设施等问题中都发挥了重要作用。

随着社会的不断发展，起源于交通系统研究的离散选择模型在其他领域也有着非常广泛的应用。在市场研究中，离散选择模型被用于研究顾客需求，预测竞争业务的影响，帮助解决了很多诸如商品定价、产品开发以及需求估计等问题。在能源规划研究中，离散选择被用来预测住户或者公司对于不同供热系统、汽车燃耗等级的选择，以便制定更有效的能源政策来使更高效地利用能源。在社会学中，劳动经济学使用这一模型来研究不同家庭的就业情况、职业选择以及对各类职业训练项目的参与情况，以此作为依据制定出更高效的就业政策。可以说，现在社会方方面面都在涉及各种各样的选择，离散选择模型对于这类问题的解决提供了很完善的方法论，并且也有着充分的研究案例可供参考，在当今社会中发挥着很大作用。

离散选择的广泛应用也在不断促进着这一理论自身的发展，人们需要更好的离散选择模型来帮助他们解决各类问题，近年来比较有代表性的研究是在离散选择中引入潜变量。传统的离散选择只依据决策个体的特质与选择支的属性做出选择，而引入潜变量则是将人的知识、经历、认知等不能被直接观察到的抽象特质考虑到选择过程中。引入潜变量产生的新模型一方面代表了新的行为学假设，另一方面也改变了原来模型的结构，这使得新产生的潜变量选择模型不仅仅在理论上更让人信服，在实际应用中也表现出了更好的预测能力。

近年来一些其他领域的进步也对离散选择模型的发展产生着巨大的影响，尤其值得一提的是机器学习理论的新一轮热潮。随着近年来深度学习理论的不断发展与计算能力的不断提升，机器学习的方法被广泛应用在分类任务、机器视觉、自然语言处理等方面被广泛应用，产生了非常深远的影响。而机器学习的发展，也让很多学者们开始关心能否利用机器学习中涉及的一些方法来改善离散选择模型，从而进一步提高模型的预测能力。这一问题在近年来备受学者们关注。

## 研究内容与意义

本研究的核心点在于将机器学习中人工神经网络方法结合到离散选择模型中的潜变量选择模型中，来提高模型的预测能力，更好地解决出行需求预测以及其他选择的预测问题，从而为交通系统规划、产品定价等决策提供更有力、更高质量的理论指导与技术支持。

本研究的内容主要包括以下几个方面：

1. 对现有的离散选择模型与机器学习相结合的研究进行归纳，总结主要区别、适用方面和优缺点；
2. 对现有的住户调查的数据集进行分析，明确数据的特点，并以此为根据生成仿真数据；
3. 将人工神经网络与潜变量选择模型相结合，通过仿真数据进行实验，从而设计合适的模型结构来尽可能提高模型的预测能力；
4. 将算法应用到真实数据考察其预测能力，并利用交叉验证、时间价值等方法来评价模型的合理性。

# 文献综述

本章我们将对过去的与我们的问题相关的一些研究进行介绍，主要包括潜变量选择模型以及一些机器学习方法在离散选择模型中的应用。

## 潜变量选择模型

近几年来对于潜变量选择模型的研究有很多，在模型结构、参数估计、模型评价，适用场景以及实际应用案例等方面都有着非常多的进步。这一节我们会对潜变量选择模型各个方面的研究进行综述。

### 潜变量选择模型的提出与推广

上世纪80年代，[McFadden (1986)](#_ENREF_13" \o "McFadden, 1986 #44)首次提出将潜变量引入到离散选择模型中。他在他的研究中提到，传统的离散选择模型很少使用反映人们偏好和认知的心理测量学数据，但是这些数据实际可以对效用函数产生非常大的影响，因为在市场研究中，这些因素更能够直接影响人们的选择。他的研究构建出了一个潜变量选择模型的二元原型，对应的变量表示见表格 1，第一列为涉及的变量，第二列为对应变量的说明。

表格 1 潜变量选择模型二元原型变量表示

|  |  |
| --- | --- |
| 变量 | 变量含义 |
|  | 测量到的成本 |
|  | 产品观测到的属性向量（包括成本） |
|  | 潜在的产品质量 |
|  | 被感知到的质量指标向量 |
|  | 被观测到的外部因素（比如教育、收入） |
|  | 潜在的成本意识 |
|  | 态度指标向量 |
|  | 潜在的购买效用（不购买的效用定为0） |
|  | 被观测到的需求（购买取1，不购买取0） |

在这一模型中，各个变量之间的关系可被建模为以下等式：

|  |  |  |
| --- | --- | --- |
|  |  | ( 2‑1 ) |
|  |  | ( 2‑2 ) |
|  |  | ( 2‑3 ) |
|  |  | ( 2‑4 ) |
|  |  | ( 2‑5 ) |
|  |  | ( 2‑6 ) |
|  |  | ( 2‑7 ) |

在这个模型中，等式( 2‑1 )和( 2‑2 )将被观测到的变量通过线性函数与反映态度与认知的潜变量关联起来；等式( 2‑4 )和( 2‑5 )将潜变量与观测到的指标关联起来；等式( 2‑3 )反映了一个线性补偿的效用形式，购买产品的效用由产品的客观属性与人的主观偏好共同决定；等式( 2‑6 )和( 2‑7 )则反映了顾客追求效用最大化这一假设。之后的潜变量选择模型的定义方式有很多，但基本都是在这一定义方式上进行微调。

[Ben-Akiva et al. (2002)](#_ENREF_2)的研究对潜变量选择模型的推广有着非常大的影响，这一研究系统地介绍了以下问题：

* 如何在选择模型中引入心理测量学数据进行建模；
* 潜变量选择模型的估计方法；
* 潜变量选择模型与传统离散选择模型相比的优势所在。

这一研究系统地提出了潜变量选择模型的框架，如图 1 所示，模型主要分为两部分：潜变量子模型和离散选择子模型。每个子模型都由一个或多个结构方程和测量方程构成。除此之外，这一研究给出了这一模型的积分似然函数形式。

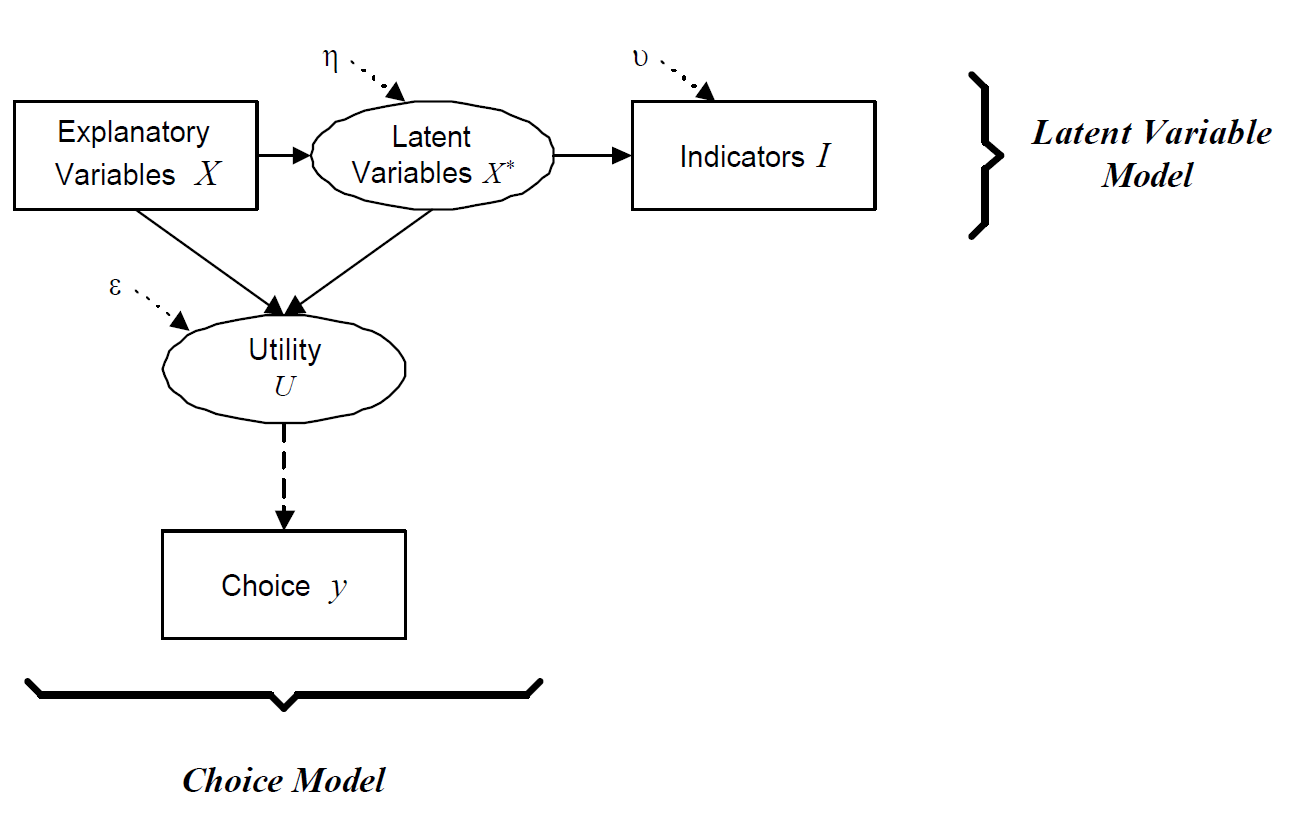


图 1 潜变量选择模型框架

对于模型的参数估计，这一研究采用极大似然法。但是因为模型的似然函数包括复杂的多重积分，所以必须采用数值积分或者仿真方法([Kenneth E Train, 2009](#_ENREF_23))来进行估计，这需要一些特殊的软件和工具的帮助。近年来，优化算法与计算能力的提升也使得这一问题得以解决，PythonBiogeme([Bierlaire, 2016](#_ENREF_3))和Mplus([Muthén & Muthén, 2004](#_ENREF_14)) 这些估计工具为潜变量选择模型的估计做出了非常重要的贡献。

这一研究中还展示了潜变量模型在三个现实案例中的应用，这些案例研究也表现出潜变量选择模型相比于传统的离散选择模型的一些优点。首先，引入潜变量提升了模型整体的拟合优度；其次，潜变量的系数在统计上都具有较高的显著性；最后，从行为学角度来看，潜变量选择模型能够更好地解释人的选择过程。

### 潜变量选择模型的估计方法

对于潜变量选择模型的估计，一般有两种方法：顺序估计与同步估计。顺序估计指的是值得是先对潜变量子模型进行极大似然估计，然后再将潜变量带入到离散选择子模型中进行估计。同步估计则是直接对整个模型进行极大似然估计。

两种估计方法都被广泛采用。[Ben-Akiva et al. (2002)](#_ENREF_2)提出顺序估计会产生不一致的估计结果，所以一些学者更倾向于使用同步估计([Raveau et al., 2011](#_ENREF_19); [Raveau et al., 2012](#_ENREF_20))，但是顺序估计较为简单，反而成了学界最被广泛使用的方法([Domarchi et al., 2008](#_ENREF_5); [Galdames et al., 2011](#_ENREF_6); [Maldonado-Hinarejos et al., 2014](#_ENREF_12))。

[Raveau et al. (2010)](#_ENREF_18)对两种方法的估计结果进行了更细致的研究。研究结果表明，两种方法都能产生参数的无偏估计，但是顺序估计在涉及到时间价值等评价指标时会有较为不好的表现，这会影响其对交通政策的评估。

### 潜变量选择模型的可确定性

由于潜变量选择模型结构的特殊性，模型的可确定性也是学者们关注的焦点。潜变量选择模型必须在估计时必须引入一定的限制，才能保证估计结果的唯一性。

[Bollen (1989)](#_ENREF_4)首次系统地提出了潜变量模型的可确定性准则，[Walker (2001)](#_ENREF_26)在她的研究中提出了离散选择模型的可确定性准则，[Ben-Akiva et al. (2002)](#_ENREF_2)结合两者提出的准则，归纳出一套充分不必要条件，来分三步判定潜变量选择模型的可确定性。除此之外，他还提出了一套可在实际中应用的检验模型可确定性的方法，这一方法使用蒙特卡洛实验来生成特定结构模型的合成数据，然后比较极大似然估计的结果与真实参数之间的差异，如果差异不显著，则可确定这一模型结构是可确定的。

对于在估计时如何引入约束，目前主要有两种方法，一种是在结构方程处引入，一种是在测量方程中引入。[Raveau et al. (2012)](#_ENREF_20)通过一些实验说明了在实际情况中，在结构方程中引入约束更能够避免严重的偏差。

### 蒙特卡洛模拟产生合成数据

在研究潜变量选择模型时，使用真实的数据只能得到模型参数的估计值，我们并不知道真正的参数是多少，这给相关研究带来了极大的不便。[Williams and Ortúzar (1982)](#_ENREF_27)提出了一套使用仿真方法来测试行为学模型的方法论，之后的学者们利用这一套方法论开展了很多关于模型估计与模型可确定性的研究([Raveau et al., 2012](#_ENREF_20); [Vij & Walker, 2016](#_ENREF_25))。

为了生成对研究更有意义的数据，我们需要总结归纳真实世界的数据特征，很多学者也在这一方面做出了贡献([Raveau et al., 2011](#_ENREF_19); [Yáñez et al., 2010](#_ENREF_28))。对于不同出行方式选择支的出行时间与出行成本有以下相关关系：

* 对于私人出行方式（汽车），出行时间与出行成本具有较高的相关性；
* 对于公交出行方式，出行时间与出行成本的相关性适中；
* 对于非机动出行方式（自行车，步行），出行时间与出行成本基本独立。

对于社会经济学属性（年龄、收入、受教育程度等），也都存在一定的相关关系。

对于潜变量子模型中的测量方程需要使用的态度指标，也存在一定的相关性，并且这一相关性可以通过探索性因子分析得到([Atasoy et al., 2013](#_ENREF_1); [Hurtubia et al., 2010](#_ENREF_9); [Popuri et al., 2011](#_ENREF_16))。

另外，[Vij and Walker (2016)](#_ENREF_25)还根据[Williams and Ortúzar (1982)](#_ENREF_27)提出的方法论与[Raveau et al. (2010)](#_ENREF_18)列出的方法归纳出生成数据时需要注意的条件。

* 模型必须是理论上可确定的；
* 各个效用项在量级上应该是相当的；
* 选择的错误率应该在25%左右。

了解了数据的特征后，我们就可以按照这些特征来生成我们需要的数据了。截断正态分布与伯努利分布为我们经常使用的分布，[Robert (1995)](#_ENREF_21)研究了如何生成截断正态分布的随机数，[Madsen and Birkes (2013)](#_ENREF_11)提出了生成具有不同分布的相关随机数的方法。

### 时间价值与其他衡量指标

对于离散选择模型，时间价值与频率价值使我们经常使用的指标([Gallen et al., 2001](#_ENREF_7))。对于潜变量选择模型的时间价值与频率价值，也有非常多的研究案例可以参考([Atasoy et al., 2013](#_ENREF_1); [Raveau et al., 2018](#_ENREF_17))。

## 机器学习方法在离散选择模型中的应用

离散选择模型从广义角度来看其实属于分类问题，而分类问题在机器学习领域也属于一个较大的分支。正是因为存在这种相似性，研究者开始探讨两者之间的关系，并尝试将两者集成起来以发挥更强的作用。本节我们将对这些研究进行综述。

### 离散选择模型与机器学习见关系的研究

[Lhéritier et al. (2018)](#_ENREF_10)在研究航班选择问题中，分析了离散选择模型与机器学习方法各自的优劣，并通过一些实验来比较他们在具有不同特性的数据上的表现。这一研究指出，以MNL为代表的离散选择模型结构简单，在行为分析上的可解释性强，但是在处理共线性的属性变量时则略显吃力，需要专业的知识才能刻画不同属性的非线性影响。而机器学习的方法擅长的是高准确率的分类，它能够自动将出行者归类并且将不同属性对最终选择的非线性影响考虑进来。

[Hagenauer and Helbich (2017)](#_ENREF_8)利用7种分类器来对出行方式选择进行分类，包括MNL模型和一些机器学习中的分类模型。这一研究给出了不同分类器的预测准确性与不同变量对选择的影响大小，其中随机森林的表现最好，出行距离与拥有的汽车数量对选择的影响最大。

### 将离散选择模型与机器学习相结合的研究

[Otsuka and Osogami (2016)](#_ENREF_15)在离散选择模型中引入深度学习，通过这种方式来创建不同选择的平均特征向量，再将这些向量作为输入引入选择模型。

[Yang et al. (2017)](#_ENREF_29)通过两种半监督的机器学习方法来研究选择问题。这一研究将期望最大化算法和聚类标签算法，根据选择建模问题的条件进行了调整。除此之外，这一研究还基于聚类标记算法开发了两个新算法。新算法使用贝叶斯信息准则来评估聚类的效果，并以此自动调整聚类数目。

[Sifringer et al. (2018)](#_ENREF_22)使用全连接神经网络来改进MNL模型。首先，他在研究中提出了MNL模型在神经网络中的等效形式，以便在现代机器学习的软件包中实现MNL模型。他通过全连接神经网络将数据中所有未被使用的变量转化为一个新的变量添加到MNL模型中，通过这种方式显著提升了MNL模型的预测能力。

# 工作计划

根据之前提到的研究内容与对问题进行的文献调研，我们制定了综合论文训练的工作计划，见表格 2。

表格 2 工作计划

|  |  |
| --- | --- |
| 时间段 | 工作内容 |
| 2018秋12-14周 | 文献阅读与开题报告的撰写 |
| 2018秋15-16周 | 收集问题需要的数据，根据真实数据特征生成合成数据 |
| 2019春1-3周 | 利用深度学习框架实现模型，并在合成数据上进行实验 |
| 2019春4-6周 | 调整模型参数，挑选合适模型结构；撰写中期报告，完成中期答辩 |
| 2019春7-9周 | 对模型进行时间价值，交叉验证等评价，验证其合理性 |
| 2019春10-12周 | 在真实数据上进行实验，并对结果进行分析评价 |
| 2019春13-15周 | 整理所有结果并进行分析，撰写论文 |
| 2019春16周 | 准备论文答辩 |

Atasoy, B., Glerum, A., & Bierlaire, M. (2013). Attitudes towards mode choice in Switzerland. *disP - The Planning Review, 49*(2), 101-117. doi:10.1080/02513625.2013.827518

Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T., & Polydoropoulou, A. (2002). Integration of choice and latent variable models. 431-470.

Bierlaire, M. (2016). PythonBiogeme: a short introduction.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley Series in Probability and Mathematical Statistics: John Wiley & Sons.

Domarchi, C., Tudela, A., & González, A. (2008). Effect of attitudes, habit and affective appraisal on mode choice: an application to university workers. *Transportation, 35*(5), 585-599. doi:10.1007/s11116-008-9168-6

Galdames, C., Tudela, A., & Carrasco, J.-A. (2011). Exploring the Role of Psychological Factors in Mode Choice Models by a Latent Variables Approach. *Transportation Research Record, 2230*(1), 68-74. doi:10.3141/2230-08

Gallen, S., Swissmetro, O., Lausanne, E., Bierlaire, M., Abay, G., & Meier, A. (2001). *The acceptance of modal innovation: The case of Swissmetro*.

Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications, 78*, 273-282. doi:10.1016/j.eswa.2017.01.057

Hurtubia, R., Atasoy, B., Glerum, A., Curchod, A., & Bierlaire, M. (2010). *Considering latent attitudes in mode choice: The case of Switzerland.* Paper presented at the World Conference on Transport Research.

Lhéritier, A., Bocamazo, M., Delahaye, T., & Acuna-Agost, R. (2018). Airline itinerary choice modeling using machine learning. *Journal of Choice Modelling*. doi:<https://doi.org/10.1016/j.jocm.2018.02.002>

Madsen, L., & Birkes, D. (2013). Simulating dependent discrete data. *83*(4), 677-691.

Maldonado-Hinarejos, R., Sivakumar, A., & Polak, J. W. (2014). Exploring the role of individual attitudes and perceptions in predicting the demand for cycling: a hybrid choice modelling approach. *Transportation, 41*(6), 1287-1304. doi:10.1007/s11116-014-9551-4

McFadden, D. J. M. s. (1986). The choice theory approach to market research. *5*(4), 275-297.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide: Statistical analysis with latent variables: User'ss guide*: Muthén & Muthén.

Otsuka, M., & Osogami, T. (2016, 2016). *A Deep Choice Model.* Paper presented at the Thirtieth AAAI Conference on Artificial Intelligence.

Popuri, Y., Proussaloglou, K., Ayvalik, C., Koppelman, F., & Lee, A. (2011). Importance of traveler attitudes in the choice of public transportation to work: findings from the Regional Transportation Authority Attitudinal Survey. *Transportation, 38*(4), 643-661. doi:10.1007/s11116-011-9336-y

Raveau, S., Álvarez-Daziano, R., Yáñez, M. F., Bolduc, D., & de Dios Ortúzar, J. (2018). Sequential and Simultaneous Estimation of Hybrid Discrete Choice Models. *Transportation Research Record: Journal of the Transportation Research Board, 2156*(1), 131-139. doi:10.3141/2156-15

Raveau, S., Daziano, R., Yanez, M., Bolduc, D., & Ortúzar, J. d. D. (2010). *Sequential and Simultaneous Estimation of Hybrid Discrete Choice Models: Some New Findings* (Vol. 2156).

Raveau, S., Muñoz, J. C., & de Grange, L. (2011). A topological route choice model for metro. *Transportation Research Part A: Policy and Practice, 45*(2), 138-147. doi:10.1016/j.tra.2010.12.004

Raveau, S., Yáñez, M. F., & Ortúzar, J. d. D. (2012). Practical and empirical identifiability of hybrid discrete choice models. *Transportation Research Part B: Methodological, 46*(10), 1374-1383. doi:10.1016/j.trb.2012.06.006

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing, 5*(2), 121-125. doi:10.1007/BF00143942

Sifringer, B., Lurkin, V., & Alahi, A. (2018). *Enhancing Discrete Choice Models with Neural Networks.* Paper presented at the hEART 2018–7th Symposium of the European Association for Research in Transportation conference.

Train, K. E. (2009). *Discrete choice methods with simulation*: Cambridge university press.

Train, K. E., McFadden, D. L., & Goett, A. A. (1987). Consumer Attitudes and Voluntary Rate Schedules for Public Utilities. *The Review of Economics and Statistics, 69*(3), 383-391. doi:10.2307/1925525

Vij, A., & Walker, J. L. (2016). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological, 90*, 192-217. doi:10.1016/j.trb.2016.04.021

Walker, J. L. (2001). *Extended discrete choice models: integrated framework, flexible error structures, and latent variables.* Massachusetts Institute of Technology,

Williams, H., & Ortúzar, J. d. D. (1982). Behavioural theories of dispersion and the mis-specification of travel demand models. *Transportation Research Part B: Methodological, 16*(3), 167-219.

Yáñez, M. F., Mansilla, P., & Ortúzar, J. d. D. (2010). The Santiago Panel: measuring the effects of implementing Transantiago. *Transportation, 37*(1), 125-149. doi:10.1007/s11116-009-9223-y

Yang, J., Shebalov, S., & Klabjan, D. (2017). Semi-supervised Learning for Discrete Choice Models. *arXiv preprint arXiv:1702.05137*.