Identifying “slow” and “fast” movers in a travel preference space: Application of a synthetic pseudo-panel approach

Stanislav S. Borysov, Jeppe Rich

Department of Management, Technical University of Denmark, DTU, 2800 Kgs. Lyngby, Denmark

Abstract

We present a new approach to studying travel preference dynamics based on constructing a synthetic pseudo-panel (SPP) from repeated cross-sectional data. This is accomplished by creating a high-dimensional probabilistic model representation of the entire data set, which allows sampling from the probabilistic model in such a way that all of the intrinsic correlation properties of the original data are preserved. The key to this is the use of novel deep learning algorithms based on the Conditional Variational Autoencoder (CVAE) framework. We use the presented approach to reveal the dynamics of transport preferences for a fixed pseudo-panel of individuals based on a large Danish cross-sectional data set from 2006 to 2016. The model is utilized to classify individuals into 'slow' and 'fast' movers with respect to the speed of which their preferences change over time. It is found that the prototypical fast mover is a young woman who lives as single in a large city whereas the typical slow mover is a middle-aged man with high income from a nuclear family that lives in a detached house outside a city.

Introduction

Understanding preferences dynamics, whether these are related to transport or any other domain, is a fundamental research question which have impact not only for models and predictions but also for the way policies are designed and to whom they should be targeted. Examples include Vij et al. (2017) who consider modal preference shifts in the San Francisco area, the understanding of how value-of-time preferences change through a financial crisis as considered in Rich and Vandet (2019), understanding of technology uptake (Mau et al., 2008) and the dynamics of car ownership (Cirillo et al., 2016; Nolan, 2010) to mention a few.

Two model approaches remain popular for estimating dynamic behaviour: (i) panel methods (Kitamura, 1990) and (ii) pseudo-panel (Deaton, 1985) methods. Whereas the native panel approach has a number of theoretical advantages over pseudo-panel methods, it is often faced with severe practical challenges related to the collection of data (Kaplan and Atkins, 1987; Golob et al., 1997). In the transport community, they have been used mainly to explore car ownership dynamics (Dargay and Vythoulkas, 1999; Huang, 2007). Whereas pseudo-panel methods overcome many of the challenges related to the collecting of panel data, they imply other limitations for the model framework (Deaton, 1985; Gardes et al., 2005).

The aim of the paper is to facilitate the pseudo-panel analysis by constructing a synthetic pseudo-panel (SPP) from repeated cross-sectional data. This is achieved by utilizing newly developed machine learning algorithms which can mimic the properties of high-dimensional data. The models adopt a deep generative modelling approach from machine learning based on the Conditional Variational Autoencoder (CVAE) (Kingma et al., 2014; Sohn et al., 2015). The benefit of this approach is that the model can act as a “sampler” of individuals in such a manner that all of the

intrinsic correlation properties of the original high-dimensional data are preserved. This brings about a number of new possibilities in addition to the application to pseudo-panels and extents to a range of other application areas including, for instance, the generation of synthetic populations aimed at agent-based modelling and the tackling of data privacy issues. The approach can be applied to reveal transport preference dynamics over time from pure cross-sectional data on a very detailed level. It becomes possible to move a given pseudo-panel of individuals forward in time and investigate how their transport preferences evolve.

Methodology

A simple way of understanding the presented framework is alleviated by considering a definition of the data it applies to. Consider a repeated cross-sectional data with N_t individuals defined by their socio-economic profiles $s_{t,i}$ $(i = 1..N)$ and preferences $v_{t,i}$ for the time t $(t = 0..T)$, where $T + 1$ represents the number of time periods for which the survey has been collected. Each $s_{t,i}$ can be represented as a collection of M_s socio-economic attributes $s_{t,i,j}$ $(j = 1..M_s)$ and $v_{t,i}$ as a collection of M_v preference attributes $v_{t,i,j}$ $(j = 1..M_v)$.

We use the assumption that the available data is a realization from an underlying joint distribution, $s_{t,i}, v_{t,i} \sim P(S, V|x_{t,i}, t)$, where S and V are random variables of socio-economic profiles and transport-related preferences and $x_{t,i}$ is a measure of external information for each individual. Knowing this joint distribution gives rise to a lot of opportunities when analyse (and synthesise) populations and their preferences over time. In this paper, we focus on the problem of generating pseudo-panel data for a number of individuals with the socio-economic profiles $s_{0,i} \equiv s_{t=0,i}$ fixed at time $t = 0$. The probabilistic framework allows analysing the entire distribution of preferences for each individual, $P_{t,i}(V) \equiv P(V|S = s_{0,i}, x_{t,i}, t)$. Clearly, this assumes that the conditional distribution $P(V|S, x_t, t)$ can be estimated from the data in a sufficiently effective manner. When both M_s and M_v are small, several model approaches from generative modelling are available, for example, based on traditional probabilistic graphical models. However, when there are many dimensions, most approaches from machine learning cannot be applied due to scalability issues. To circumvent this problem, we propose the use of the Conditional Variational Autoencoder (CVAE) — a deep generative model which is briefly described below.

The framework for generating an SPP is depicted in Figure 1 and can be summarized as follows. A CVAE model is used to 'learn' preferences distributions from the travel diary data for all the years conditional on socio-economic and external attributes. Then, the pseudo-panel can be created by sampling a number of the preference realizations for every year for a fixed pool of individuals from the reference year $t = 0$, which in turn will form the corresponding preference distribution.

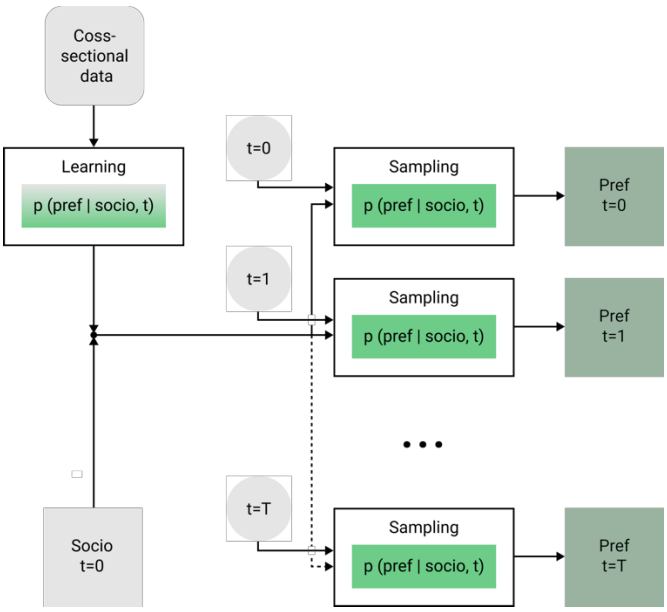

Figure 1. *Construction of a synthetic pseudo-panel.*

To model the conditional distribution $P(V|S, x_t, t)$, we use a Conditional Variational Autoencoder (Figure 2), which is a latent variable model. The main purpose of the CVAE is to estimate the probability distribution of the data through a sequence of nonlinear transformations, which are usually represented as a deep neural network, applied to a low-dimensional latent space Z with simple Gaussian properties, $Z = \mathcal{N}(0, \boldsymbol{I})$. During the training phase, the original data go through an encoding network, which maps data to the latent space. In the second stage, the data are reconstructed back to the original form using a decoding network. The objective is to find such parameters of the encoder and the decoder which jointly minimize (i) the reconstruction error between the encoder input and the decoder output and (ii) divergence between the data distribution projected to the latent space and the Gaussian prior. Once the CVAE has been trained, samples that mimic the distribution of the original data can be generated by doing Gaussian random sampling in the latent space and transforming these samples back to the original data space using the decoder network.

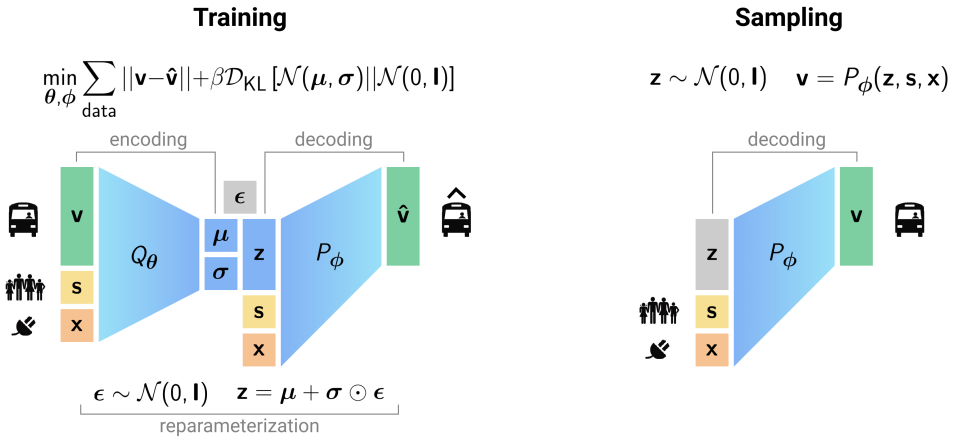

Figure 2. *Conditional Variational Autoencoder (CVAE).*

Case study

The case study for generating a super pseudo-panel of preferences is based on the Danish National Travel Survey (TU) [http://www.modelcenter.transport.dtu.dk/english/tvu], which represents a large continuous cross-sectional data set. It contains socio-economic characteristics of the participants, their geographic location as well as a detailed description of travel preferences throughout the day of the interview. The estimated preference changes are also filtered, to the extent possible, from changes to the infrastructure including extensions to the road and public transport network. However, since accounting for these changes to the infrastructure at the local level over a period of 11 years is complicated, a simplified approach based on transport zone accessibility scores estimated from the Danish national transport model is used. After removing the records with missing values, the data set contains 67419 records in total: 4345 (2006), 7010 (2007), 6606 (2008), 9885 (2009), 11966 (2010), 8354 (2011), 4783 (2012), 3600 (2013), 3541 (2014), 3172 (2015), and 4157 (2016).

Results and Discussion

For the case study, the CVAE model is used to estimate the preference distribution $P(V|S, x_t, g, t)$, where geographic location of the individuals g is included on a zone level. The model is trained using the all data available, while its hyper-parameters are tuned using a grid search on a separate validation set. To evaluate the quality of the modelled preference distribution, we generate 100,000 synthetic samples using the CVAE and compare their statistical properties to the properties of the observed data. The difference can be measured using standard metrics (e.g., a Standardized Mean Root Squared Error (SRMSE), Pearson correlation coefficient, Kullback-Leibler divergence, or Coefficient of determination) for the multidimensional histograms constructed for the synthetic and the observed distributions. The properties of the synthetic samples produced be the CVAE are in a very good agreement with the properties of the observed data (results are not shown).

Since the 4345 observations for 2006 are only around a half of the observations as in 2007 (7010), we use the 7010 individuals from 2007 as a base population s_0. This population, for which the socio-economic profiles and geographic locations g_i are known, is now moved forward to the years 2008 — 2016 and backward to 2006. For each individual i we sample travel preferences 1000 times for each year t to numerically estimate the joint distribution $P_{t,i}(V)$. The sampled distributions also take into account changes in the infrastructure through the conditioning on the respective accessibility scores $x_{t,i}$ for each year.

We use the constructed SPP to compare socio-economic profiles of people based on how much their travel preferences change during the observed period. The individuals are classified according to the speed by which their preferences change. In other words, a classification of 'slow' and 'fast' movers with respect to preference changes from 2006 to 2016. We do so by calculating the SRMSE distance between $P_{t=2006,i}(V)$ and $P_{t=2016,i}(V)$ for all individuals. Then, we range the individuals by this distance and define slow movers as those belonging to the first decile of the distance distribution whereas fast movers are those that belongs to the last decile. The marginal distributions of different socio-economics attributes for these two groups are shown in Figure 3 and described in Tables 1 and 2. According to the differences between these distributions, a few observations can be made. Firstly, the prototypical fast movers, defined by the distribution mode, are young single female adults living in cities, whereas slow movers are mainly represented by

middle-aged men with high income that live in non-single households outside cities and in owned detached houses. It is also interesting to observe that elderly people over 70 years old change their preferences faster than middle-aged people. This can be related to the socio-economic developments (e.g., higher income) and accessibility improvements (e.g., better public transport). Secondly, personal income and household size are positively correlated with the probability of being a slow mover. To some extent, it is natural to expect people with high income to be less affected by societal and technological changes. Almost all students are fast movers, while employed persons are much more reluctant to change their preferences fast. Finally, almost all slow movers live in rural areas while fast movers are city dwellers. This is an expected observation given the highly dynamic changes in modern urban areas. Although no previous analysis has been carried out with the same degrees of details, it is the authors impression that the above results corresponds well with other findings from social science.

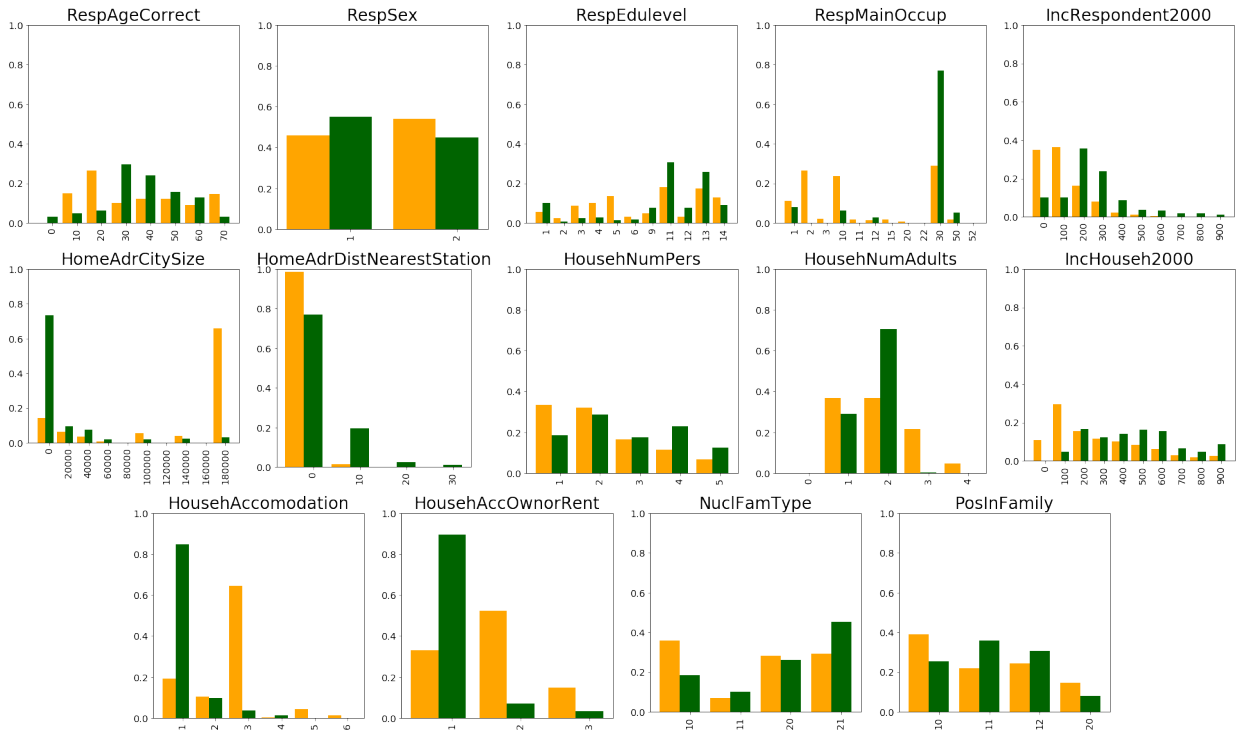

Figure 3. *Marginal distributions of socio-economic attributes of the fast (orange) and slow (green) movers in the travel preference space defined by the top / bottom deciles of the SRMSE distance between preference distributions of 2006 versus 2016. Young females in large cities changed their preferences most while old males in rural areas changed least. See more detailed description in Tables 1 and 2.*

Attribute	Values	Description	Fast	Slow
RespSex	1	Male	**0.459**	**0.550**
	2	Female	**0.540**	**0.449**
RespEdulevel	1	1st-7th form	0.057	*0.101*
	2	8th form	0.022	0.007
	3	9th form	0.087	0.025
	4	10th form	0.099	0.027
	5	Upper secondary certificate, higher preparatory certificate	*0.135*	0.014
	6	Higher commercial certificate, higher technical certificate, business college	0.032	0.018
	9	Other schooling	0.049	0.075
	11	Vocational (certificate of apprenticeship, etc.)	**0.179**	**0.305**
	12	Short-term further education (1.5 - 2 years)	0.031	0.078
	13	Medium-term further education (2 - 5 years)	**0.174**	**0.256**
	14	Long-term further education (minimum 5 years)	0.129	0.089
RespMainOccup	1	Pupil	0.111	**0.081**
	2	Student	**0.265**	0.001
	3	Apprentice, trainee	0.021	0.001
	10	Retired person, state pension, early retirement pension	*0.235*	*0.062*
	11	Unemployed	0.018	0.001
	12	Receiver of pre-retirement pay	0.014	0.028
	15	Social assistance, rehabilitation, long-term ill	0.018	0
	20	Full-time housewife/-husband, otherwise out of work	0.007	0.001
	22	National serviceman	0	0
	30	Employee	**0.289**	**0.770**
	50	Self-employed	0.018	0.051
	52	Assisting spouse (of self-employed person)	0	0
HousehAccomodation	1	Detached single-family house	**0.191**	**0.848**
	2	Terraced house, linked house	*0.104*	**0.098**
	3	Block of flats	**0.643**	*0.038*
	4	Farm	0.002	0.012
	5	Student residence	0.044	0
	6	Other	0.014	0.001
HousehAccOwnorRent	1	Owner-occupied dwelling	**0.329**	**0.895**
	2	Rent	**0.522**	**0.071**
	3	Cooperative	*0.148*	*0.032*
NuclFamType	10	Single	**0.358**	*0.185*
	11	Single with child/children	0.068	0.099
	20	Couple	*0.281*	**0.262**
	21	Couple with child/children	**0.292**	**0.452**
PosInFamily	10	Single	**0.390**	*0.253*
	11	Older in couple	*0.221*	**0.358**
	12	Younger in couple	**0.242**	**0.306**
	20	Child in nuclear family (under 25 years of age)	0.145	0.081

Table 1. *Distributions of the categorical attributes of the fast and slow movers in the travel preference space. Modes are highlighted with bold font and underlined. Second and third most frequent values are highlighted with bold and italic fonts, respectively.*

Attribute	Values	Description	Fast	Slow
RespAgeCorrect	[0, 10)	Age; years	0.001	0.031
	[10, 20		**0.151**	0.049
	[20, 30)		**0.263**	0.064
	[30, 40)		0.101	**0.295**
	[40, 50)		0.121	**0.239**
	[50, 60)		0.122	*0.158*
	[60, 70)		0.091	0.129
	≥ 70		*0.146*	0.031
IncRespondent2000	[0, 100)	The respondent's gross income, price index 2000; 1000 DKK	**0.350**	0.099
	[100, 200)		**0.363**	*0.101*
	[200, 300)		*0.164*	**0.356**
	[300, 400)		0.078	**0.239**
	[400, 500)		0.022	0.085
	[500, 600)		0.011	0.035
	[600, 700)		0.004	0.032
	[700, 800)		0.001	0.018
	[800, 900)		0.001	0.018
	≥ 900		0.001	0.011
HomeAdrCitySize	[0, 20)	Home, town size; 1000 people	**0.141**	**0.733**
	[20, 40)		*0.062*	**0.095**
	[40, 60)		0.035	*0.075*
	[60, 80)		0.007	0.019
	[80, 100)		0	0
	[100, 120)		0.054	0.019
	[120, 140)		0	0
	[140, 160)		0.041	0.024
	[160, 180)		0	0
	≥ 180		**0.657**	0.031
HomeAdrDistNearestStation	[0, 10)	Home, distance to nearest station; km	**0.984**	**0.770**
	[10, 20)		**0.014**	**0.194**
	[20, 30)		0	*0.025*
	≥ 30		*0.001*	0.009
HousehNumPers	1	Number of persons in the household; persons	**0.333**	*0.185*
	2		**0.319**	**0.285**
	3		*0.164*	0.174
	4		0.115	**0.229**
	≥ 5		0.067	0.125
HousehNumAdults	0	Number of adults (age ≥ 18) in the household; persons	0	0
	1		**0.368**	**0.289**
	2		**0.369**	**0.706**
	3		*0.216*	*0.004*
	≥ 4		0.045	0
IncHouseh2000	[0, 100)	The household's gross income, price index 2000; 1000 DKK	0.109	0.001
	[100, 200)		**0.295**	0.047
	[200, 300)		**0.155**	**0.165**
	[300, 400)		*0.116*	0.122
	[400, 500)		0.101	0.139
	[500, 600)		0.082	**0.164**
	[600, 700)		0.061	*0.156*
	[700, 800)		0.031	0.065
	[800, 900)		0.018	0.048
	≥ 900		0.027	0.088

Table 2. *Distributions of the discretized numerical attributes of the fast and slow movers in the travel preference space. Modes are highlighted with bold font and underlined. Second and third most frequent values are highlighted with bold and italic fonts, respectively.*

References

Cirillo, C., Xu, R., Bastin, F., 2016. A dynamic formulation for car ownership modeling. Transportation Science 50 (1), 322–335.

Dargay, J. M., Vythoulkas, P. C., 1999. Estimation of a dynamic car ownership model: A pseudo-panel approach. Journal of Transport Economics and Policy 33 (3), 287–301.

Deaton, A., 1985. Panel data from time series of cross-sections. Journal of Econometrics 30 (1), 109 – 126.

Golob, T. F., Kitamura, R., Long, L., 1997. Panels for Transportation Planning, 1st Edition. Transportation Research, Economics and Policy. Springer US.

Huang, B., Jun. 2007. The Use of Pseudo Panel Data for Forecasting Car Ownership. MPRA Paper 7086, University Library of Munich, Germany.

Kaplan, R. M., Atkins, C. J., 1987. Selective attrition causes overestimates of treatment effects in studies of weight loss. Addictive Behaviors 12 (3),
297 – 302.

Kitamura, R., 1990. Panel analysis in transportation planning: An overview. Transportation Research Part A: General 24 (6), 401 – 415.

Kingma, D. P., Rezende, D. J., Mohamed, S., Welling, M., 2014. Semi-supervised learning with deep generative models. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS'14. MIT Press, Cambridge, MA, USA, pp. 3581–3589.

Mau, P., Eyzaguirre, J., Jaccard, M., Collins-Dodd, C., Tiedemann, K., 2008. The neighbor effect: Simulating dynamics in consumer preferences for new vehicle technologies. Ecological Economics 68 (1), 504 – 516.

Nolan, A., 2010. A dynamic analysis of household car ownership. Transportation Research Part A: Policy and Practice 44 (6), 446 – 455.

Rich, J., Vandet, C. A., 2019. Is the value of travel time savings increasing? Analysis throughout a financial crisis. Transport Research Part A: Policy and Practice (submitted).

Sohn, K., Yan, X., Lee, H., 2015. Learning structured output representation using deep conditional generative models. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2. NIPS'15. MIT Press, Cambridge, MA, USA, pp. 3483–3491.

Vij, A., Gorripaty, S., Walker, J. L., 2017. From trend spotting to trend splaining: Understanding modal preference shifts in the san francisco bay area. Transportation Research Part A: Policy and Practice 95, 238 – 258.