Cross-Domain Customer Profiling: Mining Passengers' Food Ordering Patterns From Transportation Habits

Xueou Wang
a0095911@u.nus.edu
National University of Singapore

Yifang Yin
idsyin@nus.edu.sg
National University of Singapore

Bryan Hooi
bhooi@comp.nus.edu.sg
National University of Singapore

See Kiong Ng
seekiong@nus.edu.sg
National University of Singapore

Wynne Hsu
whsu@comp.nus.edu.sg
National University of Singapore

Renrong Weng
renrong.weng@grabtaxi.com
Grab

Xiang Hui Nicholas Lim
nic.lim@grab.com
Grab

Rui Tan
rui.tan@grab.com
Grab

ABSTRACT

With the growing market for location-based services (LBS), research into spatiotemporal data is attracting increasing interest. Our real-life motivated problem of interest is: how can businesses make use of spatiotemporal data on hand when they enter a new business domain, as they face the dilemma of sparse data available for understanding their customers' behavior in the new domain? This requires cross-domain approaches which can understand customers' habits in a new target domain, with the help of data from a well-established source domain. Specifically, we study this problem in the context of exploring customers' temporal food ordering patterns via clustering analysis based on their daily transportation temporal behavior, using transportation data from a large technology company offering ride-hailing transportation, food delivery and payment solution services. Our work provides insights to business marketing research on cross-domain customer profiling.

CCS CONCEPTS

• **Computing methodologies** → *Cluster analysis*; *Topic modeling*; • **Applied computing** → **Transportation**; • **Information systems** → *Data mining*.

KEYWORDS

Data mining, Clustering, Temporal profiling, Cross-domain learning

ACM Reference Format:
Xueou Wang, Yifang Yin, Bryan Hooi, See Kiong Ng, Wynne Hsu, Renrong Weng, Xiang Hui Nicholas Lim, and Rui Tan. 2018. Cross-Domain Customer Profiling: Mining Passengers' Food Ordering Patterns From Transportation Habits. In *San Diego '20: The 9th SIGKDD International Workshop on Urban Computing August 24th, 2020, San Diego, California USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

UrbComp '20, August 24th, 2020

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Transportation behavior study has received increasing attention due to the thriving technology in location-based services (LBS) [1, 13, 14], and their potential for understanding the rich knowledge in passengers' behavior. Currently, most studies focus on public transportation via Automated Fare Collection (AFC) system, for example, Singapore's EZ-Link card system and London's Oyster smart-card system. Analyzing the spatiotemporal data collected through these systems not only reveals city residents' traveling lifestyles, but can also provide potential solutions in many areas such as city planning and economic behavioral study [6, 12]. Businesses, especially companies engaging in transportation services such as Uber, Didi and Grab, have plentiful transportation data, including trips' starting and ending timestamps, pick-up and drop-off locations, fares and many other variables. This transportation data can bring passengers' daily travelling behavior to light, and can further help in understanding passengers' behavioral patterns even in other business domains. This is of great importance to companies who want to expand into new market, since "understanding customers" is usually one of the keys to business success, while the lack of relevant data when expanding into new domains is a notorious obstacle faced by such companies.

In this work, we study a real-world application, where we perform a cross-domain clustering analysis by relating passengers' food ordering behavior to their transportation pattern, where the food ordering style may not be discoverable by looking into food ordering service alone. Our contributions in this work are as the following:

- We perform cross-domain customer profiling via clustering analysis, showing how underlying patterns in one domain can be related to other domains, while these patterns might be hidden without the assistance of another domain.
- We implement our method in a real business case, involving the understanding of passengers' food ordering habits using their transportation behavior.
- We analyse how the mined information can be utilised by business, and thus provide insights to other similar business research.

The rest of the work is organized as follows: Section 2 discusses related work; our detailed cross-domain customer profiling methodology with the case study is presented in Section 3; finally, Section 4 concludes and discusses further research direction.

2 RELATED WORK

Transportation data research mostly focuses on public transportation. These data are usually anonymous smart card data. Analyzing such data can provide insights in urban planning. With public AFC data, [4] clusters passengers based on continuous temporal patterns, and performs a longitudinal analysis through a five-year period on passengers' travel behavior evolution. [9] utilizes smart card data of Beijing commuters to analyze spatiotemporal pattern, and provide perspectives on city development. To better understand and assist in public transport planning, [15] studies commuting capacity by investigating Beijing smart-card data and household travel survey data. [8] integrates geo-demographical information with smart card data to mine passengers' travel behaviors. There are also works exploring different methodologies in pattern regularity discovery. In [3], the authors develop a Gaussian mixture model to cope with time continuity on temporal profiles of passengers. A simple DNN framework is implemented in [5] to segment Singapore commuters into different work scope groups. However, none of these works perform any cross-domain passenger profiling, neither do they provide any thorough analysis from a business point of view. Perhaps the most relevant work is [7]. They also detect temporal profiling of passengers, and they relate passenger profiling to socioeconomic study in the city of Paris. Their work is different from ours in that they make use of public transportation data and corresponding socioeconomic data without a cross-domain exploration, while we make use of ride-sharing transportation data and perform cross-domain learning which provides relevant insights from a business point of view.

3 CROSS-DOMAIN PASSENGER PROFILING

In this Section, we present our clustering methodology for our data, give details on experiments and provide thorough analysis on cross-domain passenger temporal profiling.

3.1 Dataset

Our dataset consists of ride-sharing and food ordering trips in the city of Jakarta during the period from July 2018 to December 2018, from Grab. Each passenger has a unique hashed string ID code. We first sanitize the dataset to remove outliers and keep passengers who have both transportation and food orders. We thus retain 12,471 passengers with 693,933 transportation rides and 418,521 food rides. Note the number of transportation rides is much more than the number of food rides. A snippet of an individual passenger's trips is shown in Table 1. For the `Hour` column, "10" in the first row means that the food order occurs during 10:00 to 10:59, and "21" in the fourth row means the transportation ride happens between 21:00 and 21:59. We are interested in how passenger profiling in the transportation domain can relate to the food ordering domain and what business insights it can indicate.

Table 1: An example of an individual passenger's rides

Month	Date	Day	Hour	Type
7	1	*Sunday*	10	*Food*
7	3	*Tuesday*	11	*Trans*[a]
7	17	*Tuesday*	8	*Trans*
7	18	*Wednesday*	21	*Food*
7	21	*Saturday*	7	*Trans*
...	...	...	...	...

[a] *Trans* is short for transportation.

3.2 Clustering Methodology

We adopt a similar clustering method as in [7]. We first construct a temporal profile for each passenger i, and then implement the mixture of unigrams model to perform clustering according to each passenger's temporal profile. We now articulate the details.

3.2.1 Passenger temporal profiling. We first construct a temporal representation for each passenger. Given a passenger i, we aggregate all his/her rides, extract attribute `Type`, `Day` and `Hour` of each trip as in Table 1, and combine them to build one "word". In this way, passenger i's transportation temporal profile, denoted as $\mathbf{u}_i^T$, can be represented by a set of `Type-Day-Hour` words. The example in Table 1 thus can be represented as:

$$\mathbf{u}_i^T = \{\texttt{FoodSunday10, TransTuesday11, TransTuesday8, FoodWednesday21, TransSaturday7, } \cdots\} \tag{1}$$

Formulating such profiles for all passengers, we obtain a corpus of each passenger's temporal profile. We build passengers' temporal profiles on a weekly basis since a week can be reasonably expected to be the shortest period cycle that can represent a passenger's lifestyle.

3.2.2 Clustering with mixture of unigrams model. We now introduce the mixture of unigrams model. Mixture of unigrams is a generative probabilistic model for discrete data with latent structure, which is widely applied in topic modeling [2]. It assigns a single topic to each document. Assume there are N text documents, $\mathbf{d}_i, i = 1, 2, \ldots, N$. Each document $\mathbf{d}_i$ comprises a bag of N_i words, $\mathbf{w}_{ij}, j = 1, 2, \ldots, N_i$, where the vocabulary size is W. The sequence of words is ignored. Every document belongs to a topic, $\mathbf{z}_i$, where $\mathbf{z}_i$ is its topic index among $1, 2, \ldots, K$, where K is the total number of topics. A hierarchical generative model for a document $\mathbf{d}_i$ is then constructed as the following:

$$\begin{aligned} \mathbf{z}_i &\sim \mathcal{M}(1, \boldsymbol{\pi}), \\ \mathbf{d}_i|\mathbf{z}_i = k &\sim \mathcal{M}(N_i, \boldsymbol{\phi}_k), \end{aligned} \tag{2}$$

where $\mathcal{M}(N, \boldsymbol{p})$ denotes a multinomial distribution over N trials and event probabilities $\boldsymbol{p}$, and $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ gives the probabilities for sampling each topic, and $\boldsymbol{\phi}_k = (\phi_{1k}, \phi_{2k}, \ldots, \phi_{Wk})$ gives the probabilities for sampling each word given topic k. Then we have

$$p(\mathbf{d}_i) = \sum_{\mathbf{z}_i=1}^{K} p(\mathbf{z}_i) \prod_{j=1}^{N_i} p(\mathbf{w}_{ij}|\mathbf{z}_i). \tag{3}$$

Table 2: Food ordering demand decomposition based on weekday/weekend v.s. hour. Each number gives the percentage of each cluster food ordering demand in weekdays or weekends for a given hour.

	Cluster	Hour																							Overall
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	
Weekday	1	9.6	5.26	4.72	3.7	5.56	6.2	28.25	30.72	25.51	26.26	14.27	16.85	22.52	25.82	27.14	27.27	28.36	28.78	27.12	25.57	23.27	20.68	16.16	12.43	29.95
	2	8.62	6.76	6.11	5.76	7.14	6.98	20.45	28.51	31.99	30.7	47.48	42.73	37.74	34.11	32.19	33.99	33.23	31.74	26.91	22.29	18.44	16.27	10.97	10.47	20.02
	3	1.24	0.75	0.28	0.41	0	1.55	3.25	5.37	6.59	8.12	20.77	22.1	15.58	11.82	11.79	11.21	7.77	6.04	6.57	6.43	5.33	4.29	2.62	1.79	11.3
	4	76.89	86.04	87.22	89.3	86.51	65.89	42.53	27.41	27.78	28.62	12.29	13.03	18.28	21.99	22.46	21.14	22.71	24.56	30.44	37.0	44.8	51.73	64.46	70.12	29.82
	5	3.64	1.2	1.67	0.82	0.79	19.38	5.52	7.99	8.14	6.3	5.18	5.29	5.88	6.27	6.42	6.38	7.92	8.88	8.97	8.71	8.16	7.02	5.79	5.19	8.91
Weekend	1	13.23	12.2	5.62	0	3.85	12.5	28.07	25.98	26.72	23.14	17.98	20.1	22.69	23.6	25.21	24.71	25.78	25.83	25.16	24.52	21.9	17.82	14.78	9.5	29.95
	2	8.95	6.5	10.11	11.36	15.38	4.17	19.3	27.94	26.96	25.57	45.13	39.02	33.09	30.07	28.34	28.95	29.22	29.96	24.53	19.81	18.82	15.57	11.18	11.09	20.02
	3	0.78	4.07	0	0	0	0	0	2.94	4.41	5.29	8.71	9.08	8.17	8.25	7.95	7.99	6.97	7.49	6.75	5.91	4.77	5.32	1.67	0.9	11.3
	4	70.82	73.98	79.78	86.36	80.77	79.17	43.86	33.33	31.37	33.43	20.94	23.57	27.94	29.53	30.47	29.06	29.45	27.47	33.76	40.86	47.17	54.39	68.77	75.34	29.82
	5	6.23	3.25	4.49	2.27	0	4.17	8.77	9.8	10.54	12.57	7.25	8.24	8.11	8.55	8.03	9.29	8.59	9.26	9.8	8.91	7.35	6.91	3.6	3.17	8.91

The parameters $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ can be estimated with Expectation-Maximization (EM) algorithm. In our case, each passenger's temporal profile in (1) is modeled as a document, the `Type-Day-Hour` combination are words in the document, and the cluster of the passenger corresponds to the topic.

Other topic models such as Latent Dirichlet Allocation (LDA) [2] are also available. We tried the LDA model, but mixture of unigrams was more effective for our data. The number of clusters K needs to be decided by the analysts, and it is a nontrivial question. There are some metrics to evaluate K such as coherence score [11] and PMI-score [10], but human judgement is often still necessary. Depending on the specific case and data, we suggest a cross-validation approach for selecting K. We vary K from 3 to 20 and find that $K = 5$ performs well in our case.

3.2.3 Cross-domain passenger profiling. After grouping the passengers into 5 clusters, we look into each cluster to analyze passengers' cross-domain behavior, by visualizing their traveling and food ordering patterns. We conduct detailed analysis on cluster profiling together with its business implication in Section 3.2.4.

3.2.4 Result analysis. We first aggregate all passengers' food ordering patterns, as shown in Fig. 1. Lunch ordering tends to be more popular than dinner and supper, as people tend to cook in the evening. The most intensive demand starts from about 11am, especially on weekdays, with Wednesday noon the highest, and Thursday noon the lowest.

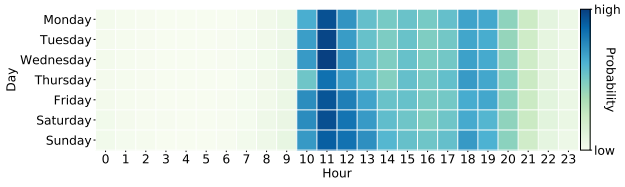

Figure 1: Aggregated passengers' food ordering pattern

Fig. 2 displays the clustered passengers' temporal profile for transportation and food order services. The five heatmaps in the left column (red) show transportation temporal profile, and the right column (green) shows the food ordering temporal pattern of each cluster. The five clusters each have their own characteristics. Cluster 1 exhibits a "flat" commuting style, spreading over a week, with more inclination towards weekday daytime. Looking into their food ordering pattern, it also spreads over any day and time of a week. Cluster 2 displays a weekend traveling pattern. Part-time working people such as part-time teachers or coaches are likely in this cluster. Their food ordering occurs more during lunch time for both weekdays and weekends. Cluster 3 consists of regular daytime commuters, going to work in the morning and leaving off in the late afternoon during weekdays. Interestingly, this cluster manifests a strong weekday lunch delivery demand, while the demand at other times is much weaker than the other clusters. The traveling pattern of Cluster 4 is somewhat similar to Cluster 2, with leaning more towards late times. This cluster may consist more "night travelers", such as security and pub staff. Inspecting their food ordering profile, we observe that they have a higher demand for dinner and supper compared to the other clusters, while their lunch demand is slightly lower (more on this point later when we analyze Table. 2). Cluster 5 passengers are "early morning travelers". These may include people like students, and parents who need send their kids to schools. Their food ordering style is diffuse but lunch and dinner at normal times show more weight.

Table 2 further decomposes food ordering demand (in %) on a weekday/weekend v.s. hour basis among the five clusters. The last "Overall" column is the cluster proportion. Interesting observations can be discovered in this table. Firstly, Cluster 2 is slightly smaller than Cluster 1, but displays a stronger demand for food delivery services than Cluster 1. These two combined constitute around half of the demand, which almost agrees with their overall percentage sum. Businesses may take this into consideration when designing marketing strategies. Next, by inspecting demand from 22:00 to 5:59 on the next day, we find that Cluster 4 takes up only 30% of all passengers; however, it drives even more than half of the late night demand. Hence, for new "night traveling" passengers who have yet to use food delivery services, businesses can consider promoting late night and early morning food delivery services to them. Fig. 3 shows the density of weekday food delivery demand over weekend demand, on a daily basis, for all five clusters. All of them are right-skewed and have peaks greater than 1, indicating that people use more food delivery services on weekdays. Note that Cluster 3 has a slightly fatter tail than the other clusters, suggesting that ordinary working people constitute the heaviest of weekday food orders.

As analysed above, different clusters show different temporal preferences for food ordering. Based on different needs and profiles, businesses can allocate resources more wisely and efficiently, and provide customized services for target customers.

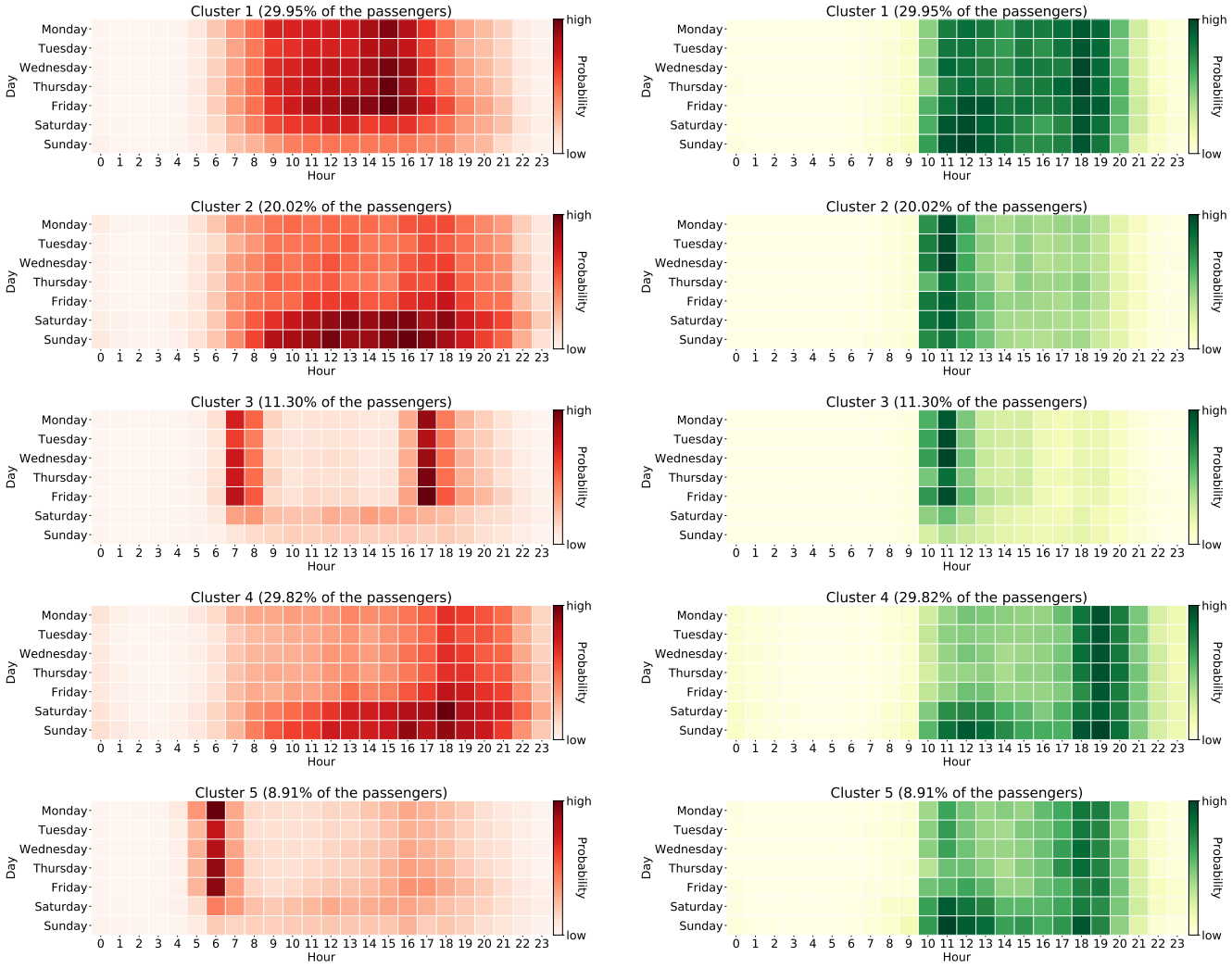

Figure 2: Cross-domain passenger temporal profiling. The left column (red) show 5 transportation temporal clusters. The right column (in green) shows the food ordering temporal profiling for the corresponding transportation cluster on the left.

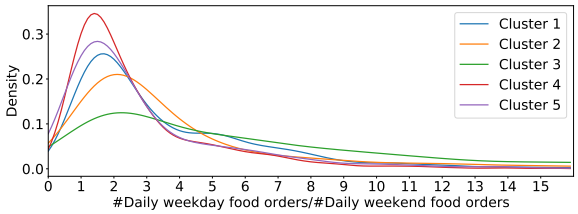

Figure 3: Density plot of daily weekday food ordering demand over daily weekend food ordering demand

4 CONCLUSION

In this work, we showed how to apply clustering analysis to perform cross-domain customer profiling within a real case study, where transportation temporal profiling can help in understanding customers' food ordering behavior, so that businesses can provide better services and resource planning. There are other interesting research questions we can investigate. One direction is that if we take temporal profile words' sequence into consideration, we may apply an LSTM model to model passengers' temporal profile evolution. Another intriguing aspect is in to apply spatial clustering to capture spatial patterns. With location labels, such as residential and shopping malls attached to pick-up and drop-off locations, more complete and informative customer profiling can be unveiled to help businesses understand customer needs.

ACKNOWLEDGMENTS

This work was funded by the Grab-NUS AI Lab, a joint collaboration between GrabTaxi Holdings Pte. Ltd. and National University of Singapore.

REFERENCES

[1] Muhammad Alam, Joaquim Ferreira, and José Fonseca. 2016. Introduction to intelligent transportation systems. In *Intelligent Transportation Systems*. Springer, 1–17.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] Anne-Sarah Briand, Etienne Côme, K Mohamed, and Latifa Oukhellou. 2016. A mixture model clustering approach for temporal passenger pattern characterization in public transport. *International Journal of Data Science and Analytics* 1, 1 (2016), 37–50.

[4] Anne-Sarah Briand, Etienne Côme, Martin Trépanier, and Latifa Oukhellou. 2017. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies* 79 (2017), 274–289.

[5] Dacheng CHEN, Ruizhi YANG, Lei SHI, Ying Kiat THAM, David LIM Tian Hui, Jasper KUAN Hon Whye, and See Kiong NG. 2018. Traveler Segmentation using Smart Card Data with Deep Learning on Noisy Labels. (2018).

[6] John W Dickey. 2018. *Metropolitan transportation planning*. Routledge.

[7] Mohamed Khalil EL MAHRSI, Etienne COME, Johanna BARO, and Latifa OUKHELLOU. 2014. Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A case study in Rennes, France. In *ACM SIGKDD Workshop on Urban Computing*.

[8] Yunzhe Liu and Tao Cheng. 2020. Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science* 16, 1 (2020), 76–103.

[9] Xiaolei Ma, Congcong Liu, Huimin Wen, Yunpeng Wang, and Yao-Jan Wu. 2017. Understanding commuting patterns using transit smart card data. *Journal of Transport Geography* 58 (2017), 135–145.

[10] David Newman, Edwin V Bonilla, and Wray Buntine. 2011. Improving topic coherence with regularized topic models. In *Advances in neural information processing systems*. 496–504.

[11] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 952–961.

[12] Vukan Vuchic. 2017. *Transportation for livable cities*. Routledge.

[13] Simon Washington, Matthew G Karlaftis, Fred Mannering, and Panagiotis Anastasopoulos. 2020. *Statistical and econometric methods for transportation data analysis*. CRC press.

[14] Xinhu Zheng, Wei Chen, Pu Wang, Dayong Shen, Songhang Chen, Xiao Wang, Qingpeng Zhang, and Liuqing Yang. 2015. Big data for social transportation. *IEEE Transactions on Intelligent Transportation Systems* 17, 3 (2015), 620–630.

[15] Jiangping Zhou, Enda Murphy, and Ying Long. 2014. Commuting efficiency in the Beijing metropolitan area: An exploration combining smartcard and travel survey data. *Journal of Transport Geography* 41 (2014), 175–183.