Bayesian Networks for travel demand generation: An application to Switzerland

Aurore Sallard*[1] and Dr. Miloš Balać[2]

[1]PhD student, Institute for Transport Planning and Systems, ETH Zürich, Switzerland
[2]Senior research assistant, Institute for Transport Planning and Systems, ETH Zürich, Switzerland

Short summary

Bayesian Networks (BNs) are probabilistic graphical models representing conditional dependencies existing between variables of interest. Recent studies have employed BNs for population synthesis and daily activity plan generation. Those studies highlight the ability of BNs to efficiently detect the causality links between variables in an easily interpretable way. This short paper aims to propose a further application of BNs for both population and daily activity plan synthesis in Switzerland. We show that understanding the dependency structure linking the population characteristics and its mobility behaviour is key to generating representative synthetic activity patterns. Furthermore, we lay the foundations for the development of temporally transferable travel demand models.

Keywords: Activity-based modeling; Bayesian networks; Synthetic Populations; Travel demand generation;

1 Introduction

According to Rasouli & Timmermans (2014), three main categories of activity-based models have been developed since their emergence in the late 60s (Chapin, 1968): constraint-based models (Jones et al., 1983), rule-based models (Arentze & Timmermans, 2000; Guan et al., 2003) and utility-maximization frameworks (Ben-Akiva & Bowman, 1998). Virtually all of them are dependent on a synthetic population: the quality of the outputs of the activity-based models is highly dependent on the quality of the input population. This is why Rasouli & Timmermans (2014) conclude their review paper by leading the research community towards the development of behaviourally rich models allowing the investigation of causality rules.

Approaches based on Markov processes, initially implemented for population synthesis (Farooq et al., 2013) are a step toward this direction. However, the approaches developed in this paper require the researchers to prepare manually the full set of conditional distributions. Sun & Erath (2015) address this issue and introduces Bayesian Networks (BNs) as an efficient tool for population synthesis. This first study was replicated and expanded by Joubert (2018). The first application of BNs for activity pattern generation was proposed by Joubert & De Waal (2020) and extended in de Waal & Joubert (2022). In these studies, the authors focus on the working population of Cape Town, South Africa, and show that their mobility behaviour is linked to the individuals' age, and employment status and to their owning a car and a driving license. All those studies show that BNs avoid over-fitting and scalability issues without being trapped by the curse of dimensionality. Moreover, they are easily interpretable and can detect complex dependency structures. They can create unobserved patterns, contrary to frequentist approaches, and allow the combination of multiple data sources into one single model. Thus, BNs appear as a promising approach in the domain of travel demand generation.

In this short paper, we propose to apply this methodology and open-source software to generate a synthetic population and its daily activity patterns. Our main contributions will be the following: first, develop a model linking population synthesis and activity chain generation using BNs; and, second, highlight the advantages of BNs compared to statistical matching in a "forecasting" experiment.

2 Methodology

Data

The Micro-census - Mobility and Transport (BFS, 2015) (referred to afterward as MZMV) is a travel survey conducted by the Swiss Federal Office of Statistics every five years. For this study, we focus on the 2010 and 2015 releases. For each edition, around 1% of the Switzerland resident population above the age of 6 is asked to report on their mobility on a certain day: for each of their trips, they have to provide information (among others) about the travel time and distance, the chosen transport mode and the trip purpose. Information about the respondents themselves and their households is collected too. Personal attributes include age, gender, driving license ownership, employment status and level of education, while the household description provide insights into the household size, structure and monthly income. Each observation (household, person, trip) is weighted. In the following, we consider the persons' weights as the focus lies on the individuals' complete activity chains. The trips data set allows us to reconstruct the individuals' activity chains.
After cleaning and removing incomplete trip data, the data set contains 50 576 personal records for 2015 and 57 087 for 2010. Those individuals reported 170 541 trips in 2015 and 190 308 in 2010 (average number of trips per respondent: 3.37 in 2015 and 3.33 in 2010). Both data sets contain around 3 000 distinct records of activity chains (2 880 in 2015 and 3 054 in 2010). The maximum activity chain length observed in the data set is 22, and 95.5% have a length lesser or equal to 7. Thus, to keep the BN structure concise and follow the approach of Joubert & De Waal (2020), we focus only on the observations where the activity chain length is not greater than 7.

Bayesian Networks

Bayesian Networks (Jensen et al., 1996) (BNs) are probabilistic graphical models consisting of two parts: a structure - one of a directed acyclic graph in which the vertices represent random variables, and the edges correspond to dependency links between the vertices - and parameters, which are joint probability tables encoding the probability distribution of the random variables. The structure and the parameters are estimated from the data or manually defined by an expert. The two approaches can complement each other. Learning the structure of a BN is an unsupervised learning problem. In this study, we use Python 3.9.7 and the library pgmpy (Ankan & Panda, 2015). An implementation of the Hill-Climb search algorithm (Selman & Gomes, 2006) based on the Bayesian Dirichlet equivalent uniform (BDEU) score (Heckerman et al., 1995) is used for the network structure estimation. The parameter learning is based on a maximum-likelihood estimator (White, 1982).

Learning the BN structure

Our goal is to compare the BN approach with the statistical matching algorithm (D'Orazio et al., 2006), which was implemented in the Switzerland eqasim scenario (Hörl & Balac, 2021a). Because of constraints inherent to this scenario, all activity chains not starting or not ending at home had to be removed from the data set. The base idea behind statistical matching is to link each agent with one activity chain record based on the weight and five socioeconomic attributes: age class, household size class, municipality type, sex, and marital status (Hörl & Balac, 2021b). Those attributes are usually obtained from national censuses, which only report a limited set of such socio-economic variables. For the BN estimation, three other attributes are included: the monthly household income, the respondent's employment status, and their ownership of a driver's license. Two main classes of variables are thus considered: the socioeconomic variables, among which are the five matching attributes, and the seven activities forming the activity chain.
Most of the BN's structure was estimated from the data, yet we imposed the following constraints. First, the five attributes mentioned above (age class, household size class, municipality type, sex, and marital status) are seen as "root" nodes in the network. This means that they cannot have parent variables. Second, we are interested in detecting how socioeconomic attributes influence activity chains. Consequently, we impose that an "activity" node can only influence the following activities. Finally, to sample a synthetic population from the BN, we generate a new activity chain for each observation – consisting of the set of socio-economic attributes – present in the training data set using the conditional probability tables estimated in the previous step.

3 Results and discussion

Replicating a given distribution

The first experiment aims to verify that the proposed travel demand generation approach is able to replicate a given activity chain distribution. Only the most recent release of the MZMV is used. The structure of the learned Bayesian Network is depicted in Figure 1. One can distinguish two "layers" in the network: on the top are socio-economic variables while the activities are on the bottom. The employment and the driving license ownership connect the two parts of the network, which is similar to the findings of Joubert & De Waal (2020). Figure 2a shows the comparison between the activity chain distributions. Dark blue bars represent the prevalence of activity chains sampled from the Bayesian Network while gray bars correspond to the distribution computed from the input data. Light blue bars represent the activity chain prevalence distribution obtained from statistical matching. The comparison shows that the statistical matching replicates almost perfectly the input distribution, which is confirmed by a Wasserstein distance (Vallender (1974))[1] between the two distributions of around 0.09. The distance between the distribution sampled from the BN and the input data is higher (around 0.12), which we can observe on Figure 2a by the larger gaps existing for some activity chains, such as the under-represented "h-w-s-h" or the over-represented "h-w-h-w-h", using the abbreviations of activity names introduced in Figure 2. Still, those differences disappear when we focus on the aggregated count of activities, as represented in Figure 2b.

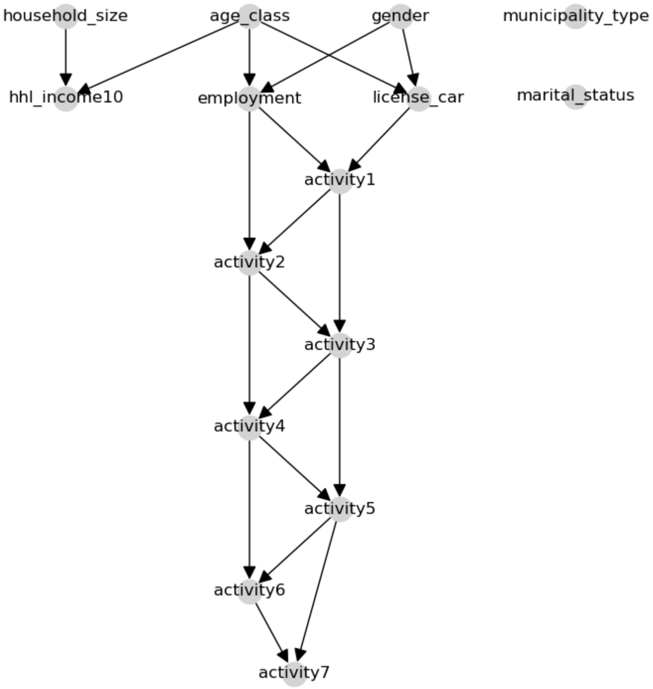
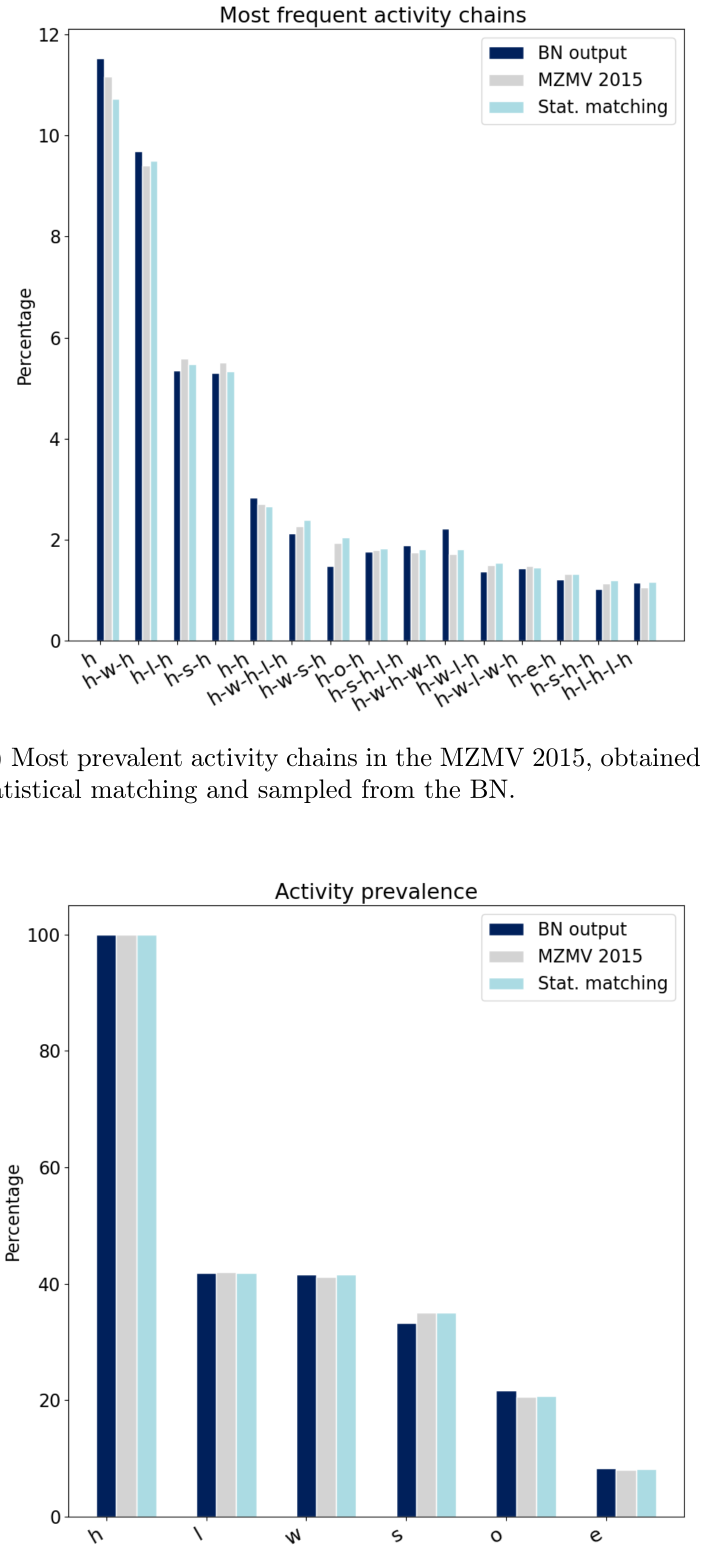


Figure 1: Bayesian Network structure.

Consequently, both Bayesian Networks and statistical matching are suitable methods when it comes to replicating a given distribution of activity chains. The relatively weaker performance of the BN, which we pointed out while computing the Wasserstein distances, can be explained by its ability to generate unobserved activity chains: looking at all the activity chains generated by the BN, regardless of their prevalence, 47.3% of them are absent from the training data set and were "created" by the BN. However, those activity chains are very rare and represent only 5.4% of the agents' daily plans. An advantage in favor of the BN approach yet seems to stand out when one combines data sources from different time contexts, as the next experiment shows.

[1] Here, instead of considering the entire range of activity chains, we take into account only the 50 most prevalent activity chains, so as to ensure that all activity chains are observed a minimal number of times.



(a) Most prevalent activity chains in the MZMV 2015, obtained from statistical matching and sampled from the BN.

(b) Percentage of activity chains containing at least one activity of each purpose.

Figure 2: Activity chain distribution and prevalence of each activity type. The six activities considered in this study are the following: home (h), work (w), education (e), leisure (l), shopping (s) and other (o).

Table 1: Accuracy, precision and F-score of the Bayesian Networks and of the Statistical matching approach

Chain	MZMV 2015 prevalence	BN prevalence	Statistical matching prevalence	BN			Statistical matching		
				Accuracy	Precision	F-score	Accuracy	Precision	F-score
h	10.7%	10.3%	10.6%	82.7%	18.0%	17.6%	82.2%	16.4%	16.3%
h-w-h	9.5%	8.5%	8.1%	85.0%	17.6%	16.6%	84.7%	14.5%	13.3%
h-l-h	5.5%	5.3%	5.5%	90.0%	6.9%	6.7%	89.8%	6.8%	6.8%
h-s-h	5.3%	5.7%	5.9%	90.1%	9.5%	9.9%	89.8%	8.1%	8.5%
h-h	2.6%	2.5%	2.7%	95.0%	2.8%	2.8%	94.9%	2.9%	2.9%
h-w-h-l-h	2.3%	2.0%	2.2%	95.9%	3.9%	3.6%	95.6%	3.4%	3.3%
h-w-h-w-h	1.8%	2.6%	2.4%	95.8%	2.7%	3.2%	95.9%	2.7%	3.1%
h-w-s-h	1.8%	1.3%	1.9%	97.0%	2.9%	2.3%	96.5%	4.7%	4.7%
h-s-h-l-h	1.7%	1.7%	1.7%	96.7%	3.5%	3.5%	96.7%	1.5%	1.5%
h-e-h	1.6%	1.4%	1.3%	97.3%	10.6%	9.9%	97.3%	7.3%	6.7%
Average				90.3%	9.8%	9.6%	90.1%	8.6%	8.4%

Towards temporally transferable travel demand generation models

In this second experiment, the BN is estimated using the older release of MZMV, dating back to 2010, while, as in the previous experiment, the population for which we generate the activity chains is the set of respondents of MZMV 2015. The obtained network has the same structure as the one depicted in Figure 1; however, the conditional probability tables changed as the training data set is different. To generate the corresponding data with the statistical matching algorithm, activity chains extracted from MZMV 2010 were matched to MZMV 2015 respondents.

Aggregated performance indicators: Similarly as before, the aggregated performance of both approaches can be measured with the Wasserstein distance. The distance between the activity chain distribution estimated from the BN and the one from the MZMV 2015 is 0.172, almost exactly the same as the distance from the statistical matching distribution and the reference data, which is 0.171. This shows that, although the BN method has a disadvantage because it is able to generate unseen activity patterns, it can compensate it by better reacting to changes in the socio-economic structure of the population, such as those one can observe between the two releases of MZMV.

Disaggregated performance indicators: Beyond the Wasserstein distance, disaggregated indicators such as accuracy, precision and F-score can be used to compare the performance of the two approaches. Those indicators are presented in Table 1 for the 10 most prevalent activity chains; the averages are related to the 15 most prevalent activity chains. The three indicators show similar values to the ones presented in Joubert & De Waal (2020) and de Waal & Joubert (2022). Moreover, they show that the BN approach outperforms in almost all cases the statistical matching algorithm.

Discussion

A detailed analysis shows that those better results are mostly linked to the fact that the BN captures with a higher accuracy the links between the population characteristics and its activity chains. More precisely, the statistical matching algorithm cannot be implemented using more than a few matching attributes, as explained in Hörl & Balac (2021b), and those attributes must be chosen by the researchers. Figure 1 highlighted that the employment status and the ownership of a driving license, which are not used for matching in Hörl & Balac (2021b), are directly influencing the mobility behavior. In the previous experiment, the fact that those variables were extracted from the most recent release of the MZMV thus led to improved results. To confirm this, a similar experiment was realized without sampling those two attributes from MZMV 2015: the distributions were kept unchanged compared to MZMV 2010. The average accuracy, precision and F-score are presented in Table 2. The table shows that, in this case, the BN approach is outperformed by statistical matching. Consequently, the BN identified which attributes are linking the population characteristics with its observed mobility behavior, which results in a more representative synthetic travel demand.

Table 2: Average accuracy, precision and F-score, over the 15 most prevalent activity chains, obtained with the statistical matching algorithm, the BN method using non-matching attributes and without using the non-matching attributes.

	Accuracy	Precision	F-score
BN with employment status and driving license	90.3%	9.8%	9.6%
BN without employment status and driving license	90.0%	8.4%	8.2%
Statistical matching	90.1%	8.6%	8.4%

4 Conclusions

This short paper presents an application of Bayesian Networks for synthetic population and travel demand generation. Contrary to Joubert & De Waal (2020), who focus on the employed population living in Cape Town, we can synthesize agents representative to the Swiss people, except for children below six years of age. This study is based on open-source software. We used both aggregated and disaggregated indicators to evaluate our approach. We showed that BNs could accurately replicate a given activity chain distribution and outperform the statistical matching algorithm when combining two data sources in a "forecasting" experiment. We highlighted that the ability of BN to identify the critical socio-economic attributes influencing the activity chain is of great help to generating representative synthetic population and travel demand. Several points indicate a potential for future research: this study will first be extended to include experiment results obtained from the 2005 release of MZMV. Moreover, the networks were estimated using a combination of data-driven, machine-learned methods and expert knowledge, as some constraints about the links' directions were imposed. It would thus be relevant to conduct sensitivity analyses to evaluate the impact of imposing these constraints. A third possible future research direction is to estimate networks specific to given socio-economic categories. In this way, one could capture and represent the differences in dependency structures and ultimately contribute to the understanding of social mechanisms leading to heterogeneity in mobility behaviour.

Acknowledgements

This study is part of a National Research Program (NFP78) funded by the Swiss National Science Foundation.

References

Ankan, A., & Panda, A. (2015). pgmpy: Probabilistic graphical models using python. In *Proceedings of the 14th python in science conference (scipy 2015).*

Arentze, T., & Timmermans, H. (2000). *Albatross: a learning based transportation oriented simulation system.* Citeseer.

Ben-Akiva, M. E., & Bowman, J. L. (1998). Activity based travel demand model systems. In *Equilibrium and advanced transportation modelling* (pp. 27–46). Springer.

BFS. (2015). *Mikrozensus Mobilität und Verkehr.* `https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/erhebungen/mzmv.html`. (Accessed: 2022-05-03)

Chapin, F. S. (1968). Activity systems and urban structure: A working schema. *Journal of the American Institute of Planners*, *34*(1), 11–18.

de Waal, A., & Joubert, J. W. (2022). Explainable Bayesian networks applied to transport vulnerability. *Expert Systems with Applications*, *209*, 118348.

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice.* John Wiley & Sons.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263.

Guan, J., Roorda, M. J., & Miller, E. J. (2003, January). Approximation of 24 Hour Travel Times in the Greater Toronto Area. *Presented at the 82nd Annual Meeting of the Transportation Research Board, Washington DC*.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, *20*(3), 197–243.

Hörl, S., & Balac, M. (2021a). Introducing the eqasim pipeline: From raw data to agent-based transport simulation. *Procedia Computer Science*, *184*, 712–719.

Hörl, S., & Balac, M. (2021b). Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C: Emerging Technologies*, *130*, 103291.

Jensen, F. V., et al. (1996). *An introduction to Bayesian networks* (Vol. 210). UCL press London.

Jones, P. M., Dix, M. C., Clarke, M. I., & Heggie, I. G. (1983). *Understanding travel behaviour* (No. Monograph).

Joubert, J. W. (2018). Synthetic populations of South African urban areas. *Data in brief*, *19*, 1012–1020.

Joubert, J. W., & De Waal, A. (2020). Activity-based travel demand generation using Bayesian networks. *Transportation Research Part C: Emerging Technologies*, *120*, 102804.

Rasouli, S., & Timmermans, H. (2014). Activity-based models of travel demand: promises, progress and prospects. *International Journal of Urban Sciences*, *18*(1), 31–60.

Selman, B., & Gomes, C. P. (2006). Hill-climbing search. *Encyclopedia of Cognitive Science*, *81*, 82.

Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, *61*, 49–62.

Vallender, S. (1974). Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, *18*(4), 784–786.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 1–25.