Machine learning-based framework for data reviewing of a national household travel survey

Lisa Ecke*[1], Miriam Magdolen[1], Sina Jaquart[1], Peter Vortisch[1]

[1] Institute for Transport Studies, Karlsruhe Institute of Technology, Germany

SHORT SUMMARY

Machine learning techniques have mainly been applied to physical measurement data in the past. In this paper, machine learning is applied to survey data of everyday travel behavior provided by the German Mobility Panel (MOP). The presented model framework supports trained staff in checking trips collected in a trip diary. To this aim, four algorithms are applied and tested further. The neural network (NN) shows the most appropriate results. By using the NN for the individual trip checks, the time effort for the trained staff can be reduced by 20.4 %. In addition, it decreases the number of data samples where all reported trips must be checked. Our study shows that machine learning can support the process of data checking in the MOP leading to significant time reduction.

Keywords: Big Data, Data checking, Data quality, German Mobility Panel, Machine learning

1. INTRODUCTION

High data quality is a key resource for better understanding and forecasting travel behavior. To ensure data quality, data must be checked and implausible data must be processed so that people can trust them for their analyses. The German Mobility Panel (MOP) is a longitudinal national household travel survey of the German population. Since 1994, the data has been checked each year individually by trained staff based on defined rules. This procedure ensures a consistently high data quality for further use. However, the data checking is time-consuming and thus costly, as around 70,000 trips per year have to be checked individually. Over the years, a large data pool of raw and verified data has been built up, allowing the individual checking techniques of trained staff to be analyzed.

This paper investigates machine learning techniques to support the review process of the MOP data. Using the data sets that have been checked in the past, an appropriate model can learn the patterns and human procedures based on historical data. Our approach presents a model that can reproduce human checking routines. Therefore, we demonstrate how feature engineering was implemented for the dataset, how appropriate models were selected, and which input variables contribute to model improvements.

2. LITERATURE

Wherever data is processed, ensuring data quality is essential for performing reliable analyses. This applies to measured data as well as survey data. However, data reviewing methods involving machine learning can be found in the literature, especially for measured data. To the authors' knowledge, studies in travel behavior research use machine learning techniques predominantly for recognizing transportation modes in GPS data (Bolbol et al. 2012; Feng und Timmermans 2013; Zhou et al. 2016; Wang et al. 2018; Yazdizadeh et al. 2019). Further studies that use machine learning techniques focus on identifying trip purposes in GPS data (McGowen und McNally 2007; Deng und Ji 2010; Montini et al. 2014; Yazdizadeh et al. 2019). Summing up, the literature focuses mainly on supporting modern survey methods during and after the fieldwork and less on data checking processes of traditional household travel surveys.

3. DATA

The data basis of this paper originates from the MOP. Since 1994, the MOP has been collecting data on the travel behavior of the German population. Households are asked to provide information on their everyday travel for three consecutive years. The participants report their travel behavior in a trip diary for one week. Thus, the trip diary contains information about all their trips taken during seven consecutive days, including distances, means of transport, trip purposes and start and arrival times. Participants also provide information on sociodemographic characteristics of the household and characteristics relevant to travel behavior, such as driver's license or car ownership (Ecke et al. 2020). For our analysis, we rely on five years of data (2015-2019). In total, information is available on 7,389 households, 15,217 individuals, and 284,299 trips.

4. METHODOLOGY

This study applies machine learning methods to support and automate the individual data checking and editing of trips after the data collection process. Examples for implausibilities in the data are e.g. high speeds for walking trips or a missing return trip back home. During the first steps of the work, it became clear that it is impossible to automate the whole checking and editing process due to the data structure. Thus, the focus was shifted to automate the process of data checking as much as possible to minimize the workload for the human reviews by trained staff. For this purpose, the implemented model approaches are applied to examine the individual trips of the trip diaries for input errors. Trips that may need to be further adjusted should be labelled accordingly. Our methodology aims to classify the trips into two classes: The class of plausible reported trips (OK) and the class of trips that may need to be edited (P). Subsequently, the trips of class P are further checked. Data preparation for the machine learning models to check the trips includes adding labels and constructing features at different levels of detail to optimize the performance of the models. To optimize the classification of trips, we apply four machine learning methods and compare their performance: A binary Neural Network (NN), a decision tree, a random forest and a Support Vector Machine (SVM).

The models are implemented in Python 3.8.0 release. Due to the complexity of machine learning and the optimization of the training algorithms, Pandas and NumPy are used. Furthermore, we use Matplotlib and Seaborn for data visualization and TensorFlow, Keras, Scikit-Learn and TensorBoard for machine learning.

Adding labels

First, the data is labelled by comparing the raw data with the checked data. This is done according to how the trip was edited during the checking processes in the past. Binary labels are used to determine whether the trip's purpose, distance, transport mode, or start or end time was edited. With that and for each trip feature, a binary label is added whether this feature was edited. It is also labelled whether a trip is deleted or added before or after the trip.

Adjustment of data format and categories

Format adjustments (e.g. date) must be made to keep the feature construction process traceable. In particular, to apply the model to all years, the categories of the trip characteristics purpose and transport mode must be consistent for all years. Each variable is passed to the model as a separate binary feature.

Construction and selection of required features

Features are selected based on their importance to the performance of the model. The number of features and the complexity can be reduced for the algorithm. Most approaches for evaluating the feature importance, especially for an NN, are based on the final weights. According to Kralj Novak et al. (2019), the VIANN method is used to measure the variance of these weights and obtain the relative variable importance for NNs. In Kralj Novak et al. (2019), the weights are changed during the training phase. This updating of weights is repeated until the model reaches its final state.
Due to the randomization of the weights in an NN, the sum of the weight changes may differ depending on the training cycle; thus, the VIANN score may vary. However, in our data, after several trainings, it becomes clear that the features start time, end time, duration, and velocity have vanishingly minor importance for the NN despite their values.

Implementation of the models

To optimize the classification of trips, we implement four models (binary NN, decision tree, random forest and SVM) and compare their performance. All models support binary classification. The NN (Recall= 0.82, Precision= 0.33) is the most effective model in our study.

5. RESULTS AND DISCUSSION

The impact of using machine learning to support the checking of the trip diaries is presented in the following. The results of the NN show, that 49,253 out of 274,273 trips are assigned to class P and must be checked manually. The exact number of trips may differ slightly depending on the training run of the model and the random division into training and test data. Thus, only 18% of the original total number of trips have to be checked by the trained staff.

In addition, the trips can be distinguished concerning the degree of difficulty of the check. If the purpose of a trip, the means of transport or the distance of a trip has to be checked, it may be necessary to also look at the trip history of a person over seven days. This makes the manual review more laborious. This is the case for 8,753 trips. If the trip is incorrectly assigned to class P, this step is also needed. For example, if for a trip assigned to class P ,it is not obvious that the start time overlaps with the end time of the previous trip, then the other possible reasons must be excluded.

Accordingly, it is essential to reduce the number of trips and the number of individuals to be checked. The 49,253 trips to be checked were reported from 9,462 different participants. This results in an effort reduction of 20.4% at the level of participants (Figure 1).

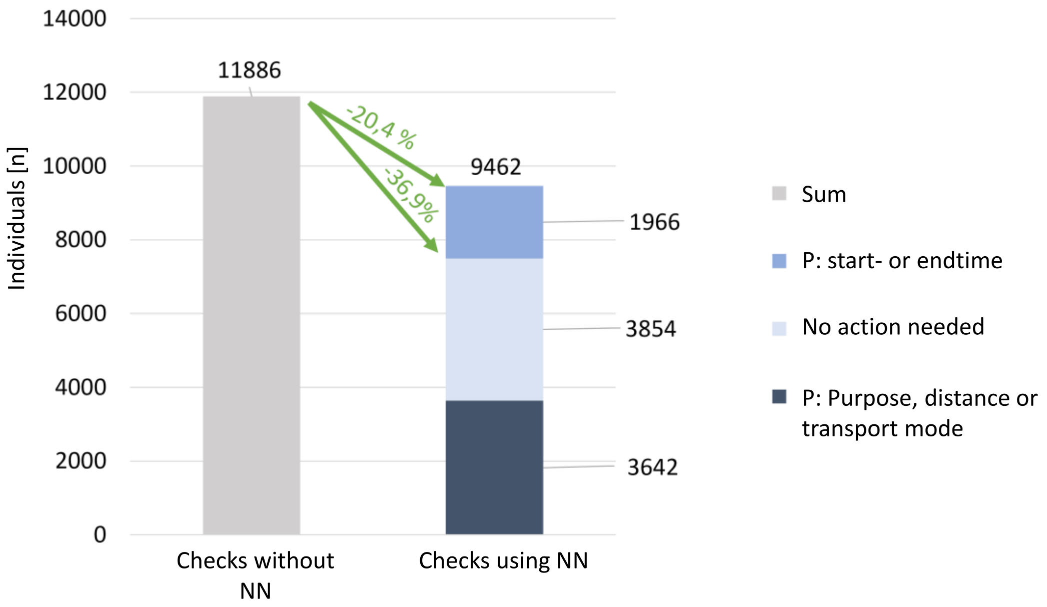

Figure 1 Reduction in effort in terms of checked participants in 2015, 2016, 2018, and 2019

When looking at feature importance, it is noticeable that the model learns a lot from features unrelated to history or at most related to the trip before. These include, for example, checking whether the start and end times of the trips are correct, whether it is the first or last trip of the day, whether multiple trips can be combined into a round trip or the purpose and mode of transport of the previous trip. This means that if the check of these correlations does not give any reason for a plausibility check of the trip, human staff has to look at the participant's history.

6. CONCLUSIONS

This paper investigates the applicability of machine learning for checking the quality of reported trips in the MOP. Due to the long history of the MOP and many years of manual data checking by human staff, data is provided to support the training of the model.

For the checking, trips are classified first binary. One class is for trips that do not need to be edited because they were correctly reported. The second class is for trips that may need further editing and should be checked accordingly by trained staff. The classification is constructed in such a way as to reduce the workload for the trained staff. Independent editing by the model is not possible because the single edits occur only very rarely so that the model cannot learn it.

We applied four different machine learning models to label the reported trips. A comparison of the models shows that the NN is the most effective model compared to a decision tree, random forest, and SVM. The application of the NN leads to a considerable reduction in the number of trips to be checked. The future steps for this project is to create a more consistent training data set to improve the consistency of the data. Overall, it can be concluded that machine learning is good support for the checking processes of travel behavior data.

REFERENCES

Bolbol, Adel; Cheng, Tao; Tsapakis, Ioannis; Haworth, James (2012): Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. In: *Computers,* *Environment and Urban Systems* 36 (6), S. 526–537. DOI: 10.1016/j.compenvurbsys.2012.06.001.

Deng, Zhongwei; Ji, Minhe (Hg.) (2010): Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach. International Conference on Traffic and Transportation Studies (ICTTS). Kunming, China, 3.-5. Aug 2010.

Ecke, Lisa; Chlond, Bastian; Magdolen, Miriam; Vortisch, Peter (2020): Deutsches Mobilitätspanel (MOP) - Wissenschaftliche Begleitung und Auswertungen. Bericht 2019/2020: Alltagsmobilität und Fahrleistung. Hg. v. Institut für Verkehrswesen (IFV). Karlsruher Institut für Technologie (KIT). Karlsruhe.

Feng, Tao; Timmermans, Harry J.P. (2013): Transportation mode recognition using GPS and accelerometer data. In: *Transportation Research Part C: Emerging Technologies* 37, S. 118–130. DOI: 10.1016/j.trc.2013.09.014.

Kralj Novak, Petra; Šmuc, Tomislav; Džeroski, Sašo (Hg.) (2019): Variance-Based Feature Importance in Neural Networks. 22nd International Conference, DS 2019. Split, Croatia, 28.10.2019-30.10.2019: Springer International Publishing; Imprint Springer.

McGowen, P.; McNally, M. (2007): Evaluating the potential to predict activity types from GPS and GIS data. Submitted for Consideration to the Western Regional Science Association. Newport Beach, CA.

Montini, Lara; Rieser-Schüssler, Nadine; Horni, Andreas; Axhausen, Kay W. (2014): Trip Purpose Identification from GPS Tracks. In: *Transportation Research Record* 2405 (1), S. 16–23. DOI: 10.3141/2405-03.

Wang, Bao; Gao, Linjie; Juan, Zhicai (2018): Travel Mode Detection Using GPS Data and Socioeconomic Attributes Based on a Random Forest Classifier. In: *IEEE Trans. Intell. Transport. Syst.* 19 (5), S. 1547–1558. DOI: 10.1109/TITS.2017.2723523.

Yazdizadeh, Ali; Patterson, Zachary; Farooq, Bilal (2019): An automated approach from GPS traces to complete trip information. In: *International Journal of Transportation Science and Technology* 8 (1), S. 82–100. DOI: 10.1016/j.ijtst.2018.08.003.

Zhou, Xiaolu; Yu, Wei; Sullivan, William C. (2016): Making pervasive sensing possible: Effective travel mode sensing based on smartphones. In: *Computers, Environment and Urban Systems* 58, S. 52–59. DOI: 10.1016/j.compenvurbsys.2016.03.001.