Journal of Transport Geography 82 (2020) 102549

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Network–wide prediction of public transportation ridership using spatio–temporal link–level information

Stephan Karnberger[a], Constantinos Antoniou[b,*]

[a] Stadtwerke München, Emmy-Noether-Str. 2, 80992 Munich, Germany
[b] Technical University of Munich, Arcisstrasse 21, 80333 Munich, Germany

ARTICLE INFO

Keywords:
Public transportation
Ridership
Prediction
Inference
Machine learning

ABSTRACT

Public transportation is a key element to vivid city life. Understanding the dynamics and driving forces of public transportation ridership can be a very rewarding task. It is, however, a highly complex construct. In this research, we focus on a spatial viewpoint, which has seen little attention: the link level. It represents the trip of a vehicle between directly connected stations. Additionally, we put emphasis on the impact of exogenous events. In order to assess their spatio–temporal influences, a temporal resolution of 30 min complements the spatial link level. Ridership data for trams and buses is provided by Stadtwerke München (SWM), which is the operator of the public transportation network in Munich, Germany, including 82 bus and 17 tram lines. About 30% of trams and 50% of buses are equipped with automatic passenger counting sensors, which capture boarding and alighting at each individual station. The equipped vehicles are strategically placed by SWM to obtain a meaningful view on the whole system. The raw sensor data is cleaned and sanitized. The data we are using spans a 4–year period (2014–2017). Following a pre–processing step, ~59.79% of the data is considered, which equates to ~97 million observations. There are 693 tram links and 2944 bus links, which makes 3637 links in total. We distinguish the analysis in ridership prediction and inference. For prediction, we specify one model functional form and build this model for each link, using 5–fold cross–validation to avoid overfitting. We employ decision trees, combining them with bagging and boosting. We then perform inference, i.e. attempt to understand the relationship between the variables that emerged in the predictive models. Ridership is assessed for each link separately and visualized together in order to construct network views and maps. Conclusions are drawn, and recommendations for future research are formulated.

1. Introduction

Public transportation is a key element to vivid city life; for example, Cervero (2009) states that "replacement of elevated freeways with greenways, boulevards, and public transit can improve neighborhood quality and increase land values". Understanding the dynamics and driving forces of it always proves useful; for example, Jifeng et al. (2008) use system dynamics to model urban transportation systems and present a case study analyzing the public transportation network of the Dalian Central City in China. It is, however, a highly complex construct. A variety of factors come into play and the interactions between them are not simple (Tyrinopoulos and Antoniou, 2008). Furthermore, public transportation can be viewed at from different angles. From longterm developments for city planning purposes to short–term disturbances for real–time passenger information systems and everything in between; different spatial and temporal viewpoints exist and each of them demands a separate approach.

In this research, we focus on a spatial viewpoint, which has seen little attention: the link level. It represents the trip of a vehicle between directly connected stations. Additionally, we put emphasis on the impact of events. In order to assess their spatio–temporal influences, a temporal resolution of 30 min complements the spatial link level.

A large stream of literature discusses direct ridership models (DRM) (Gutiérrez et al., 2011; Jun et al., 2015; Chakour and Eluru, 2016; Peterson, 2011; Choi et al., 2012; Liu et al., 2014; Cervero et al., 2010). These models describe the long–term development of public transportation ridership using socio–economic, demographic and built–environment variables. Ridership is aggregated monthly or daily and analyzed at station level, i.e. the number of passengers boarding and/or alighting at a specific station is targeted. Built–environment variables refer to the properties of the station and its surroundings, while socio–economic and demographic variables may refer to the area around

* Corresponding author.
E-mail addresses: karnberger.stephan@swm.de (S. Karnberger), c.antoniou@tum.de (C. Antoniou).

https://doi.org/10.1016/j.jtrangeo.2019.102549
Received 1 May 2019; Received in revised form 13 August 2019; Accepted 24 September 2019

the station or to a more general context, e.g. growth domestic product. Another stream of literature investigates the effects of weather on ridership (Li et al., 2015, 2017; Stover and McCormack, 2012; Tao et al., 2016, 2018; Kashfi et al., 2015; Zhou et al., 2017; Trépanier et al., 2012). Other approaches are of course also considered, e.g. Chen et al. (2009) examine the relationship between the local environment of stations and their usage over day in New York City, using card data which records entrances at stations, Ceapa et al. (2012) address the regular weekday commute–based pattern of stations in London, Bhattacharya et al. (2013) propose a Gaussian process–based model to predict the number of people entering a bus at station level, Foell et al. (2015) investigate a two month period in Lisbon using individual histories about bus rides containing line, stop, and arrival time conducted by individual users. Despite their intuitive importance, only few papers address the effect events can have on public ridership. Rodrigues et al. (2013) developed a framework to use event data from the web for Singapore, and Rodrigues et al. (2017) propose a Bayesian additive model for the same problem statement.

Two methods to collect ridership data are individual smart cards and automatic passenger counting (APC) systems. A smart card is a personalized card, which has to be used at the origin and destination stations. Thus, the smart card records the origin and destination for each trip made by a certain individual, along with further information, e.g. time of trip and travel mode used. Automatic passenger counting systems capture the number of people entering and exiting vehicles at stations. These may be implemented e.g. as infrared sensors or treadle mat sensors (Pinna et al., 2010).

Table 1 provides a taxonomy of a number of relevant papers according to their spatiotemporal viewpoint and the data being used. It is apparent that station level ridership has been studied extensively, but no study could be identified which considers the link level. The difference between station and link level, however, is substantial. On a link level, the ridership present in a certain vehicle is assessed, whereas on a station level the boardings/alightings at a certain station are considered. When only the boardings and alightings are considered, the aspect of direction is ignored. A link inherently includes direction, as it is defined by a start station and an end station. The level of crowdedness can vary significantly for a vehicle depending on the direction it is heading towards. For example, a station may have a lot of boardings in the morning, but it could be that the crowdedness of vehicles heading into different directions from this station differ. However, inspecting the link level makes it a lot harder to relate environmental factors with ridership. This is because, when considering the station level, each station is fixed and can be described by a fixed set of environmental variables (e.g. built environment, land–mix). Links, on the other hand, cannot be abstracted by these variables. One could imagine to describe a link by its start and end stations' characteristics, but, as a link inherently is part of a bigger route, this description does not work.

Few papers address actual prediction of ridership (Bhattacharya et al., 2013; Ceapa et al., 2012; Rodrigues et al., 2013, 2017). As mentioned, much more emphasis is placed on the aspect of inference. One possible reason for this may be that inference is possibly seen as more valuable than prediction in the context of public ridership. If the driving forces behind public transportation usage are known, these can be consciously influenced. Another possible reason may be that the importance of influence factors change when shifting from a more coarse grained temporal view, like weekly or daily, to a more fine grained temporal resolution like hourly, as a diurnal usage pattern has to be addressed. Classical variables, like socio–economic features, lose importance in this setting.

Finally, the impact of events on ridership has been largely neglected, which can be attributed to poor information availability and the difficulty of measuring their impact (Rodrigues et al., 2013). The few available references (Rodrigues et al., 2013, 2017), which address the effect of events, focus on certain venues and the impact on arriving passengers for these locations. To the best of the authors' knowledge, impact of events on the whole network has not been studied.

Consolidating the above mentioned points, in this research we concentrate on a fine grained temporal (30 min) and spatial resolution (the link level). Furthermore, we quantify and assess the impact of localised events on network–wide ridership. Specifically, we want to address prediction of and inference about the actual number of passengers in a vehicle belonging to a line and heading in a given direction, for a certain point in time and location, and how it may (or may not) be affected by various events in the network.

Table 1
Categorization of relevant research.

	Paper	Temporal[b]					Spatial[c]				Data[d]		
		M	W	D	TD	m	N	R	S	OD	APC	SC	Other/NA
DRM[a]	Gutiérrez et al. (2011)	X							X				X
	Jun et al. (2015)			X					X				X
	Chakour and Eluru (2016)				X				X		X		
	Peterson (2011)		X						X				X
	Choi et al. (2012)				X					X		X	
	Liu et al. (2014)			X					X				X
	Cervero et al. (2010)			X					X				X
Weather	Li et al. (2017)			X				X					X
	Li et al. (2015)			X				X				X	
	Stover and McCormack (2012)			X			X						X
	Tao et al. (2016)			X						X		X	
	Tao et al. (2018)					60	X		X			X	
	Zhou et al. (2017)	X					X		X			X	
Other	Chen et al. (2009)				X				X				X
	Ceapa et al. (2012)					10			X			X	
	Bhattacharya et al. (2013)					60			X				X
	Kashfi et al. (2015)			X			X					X	
	Foell et al. (2015)			X								X	
	Rodrigues et al. (2013)					30			X			X	
	Rodrigues et al. (2017)					30			X			X	

[a] Direct ridership models.
[b] Y = Yearly, M = Monthly, W = Weekly, D = Daily, TD = Time of Day, m = minute resolution.
[c] N = Network, R = Route, S = Station, OD = Origin–Destination.
[d] APC = Automatic Passenger Counting, SC = Smart Card.