Journal of Big Data Analytics in Transportation
https://doi.org/10.1007/s42421-019-00007-7

ORIGINAL PAPER

Measuring Traffic in Cities Through a Large-Scale Online Platform

Vilhelm Verendel[1] · **Sonia Yeh**[2]

Received: 13 September 2018 / Revised: 16 July 2019 / Accepted: 4 September 2019

Abstract
Online real-time traffic data services could effectively deliver traffic information to people all over the world and provide large benefits to the society and research about cities. Yet, city-wide road network traffic data are often hard to come by on a large scale over a longer period of time. We collect, describe, and analyze traffic data for 45 cities from HERE, a major online real-time traffic information provider. We sampled the online platform for city traffic data every 5 min during 1 year, in total more than 5 million samples covering more than 300 thousand road segments. Our aim is to describe some of the practical issues surrounding the data that we experienced in working with this type of data source, as well as to explore the data patterns and see how this data source provides information to study traffic in cities. We focus on data availability to characterize how traffic information is available for different cities; it measures the share of road segments with real-time traffic information at a given time for a given city. We describe the patterns of real-time data availability, and evaluate methods to handle filling in missing speed data for road segments when real-time information was not available. We conduct a validation case study based on Swedish traffic sensor data and point out challenges for future validation. Our findings include (i) a case study of validating the HERE data against ground truth available for roads and lanes in a Swedish city, showing that real-time traffic data tends to follow dips in travel speed but miss instantaneous higher speed measured in some sensors, typically at times when there are fewer vehicles on the road; (ii) using time series clustering, we identify four clusters of cities with different types of measurement patterns; and (iii) a k-nearest neighbor-based method consistently outperforms other methods to fill in missing real-time traffic speeds. We illustrate how to work with this kind of traffic data source that is increasingly available to researchers, travellers, and city planners. Future work is needed to broaden the scope of validation, and to apply these methods to use online data for improving our knowledge of traffic in cities.

Keywords Big data · Urban traffic · Data availability · Travel delays · Time series clustering

Introduction

By mid-century, the global population could very well reach 10 billion, with three out of every four people likely to be living in highly urbanized places. Large efforts have already been made in intelligent traffic systems, real-time traffic information, traffic control management, accident alert systems, and it is widely believed that innovation in digital technologies and the availability of big traffic data in real-time can enable the construction of advanced traveler information systems for route choice, and improve the possibility for traffic planning to become even better for urban mobility (Lyons 2016) and provide significant value to individual travelers and urban planners in future cities (Levinson 2003; Xu and González 2017; Hensher 2018). Yet, congestion and traffic delays are still one of the largest challenges that big cities face today and city-level road network traffic data are often hard to come about (Barthelemy 2016). The societal costs of congestion are high; besides lost wages and inconvenience costs, there are also costs of extra fuel, accidents (including deaths and health costs), air pollution and many others (Arnott and Small 1994). A TRB report (Transportation Research Board 2009) concludes that in the US "In 2005, congestion cost travelers more than 4.2 billion hours and

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42421-019-00007-7) contains supplementary material, which is available to authorized users.

✉ Vilhelm Verendel
vive@chalmers.se

[1] Department of Computer Science and Engineering, Chalmers University of Technology, 41296 Gothenburg, Sweden

[2] Department of Space, Earth and Environment, Chalmers University of Technology, 41296 Gothenburg, Sweden

Published online: 25 October 2019

nearly $80 billion and resulted in the waste of approximately 3 billion gallons of fuel. One of the most significant impacts of congestion on the individual driver is the increasing difficulty of predicting how long a given trip will take. This lack of travel time reliability has both personal and economic costs". It is, thus, clear that traffic congestion disrupts economic activities on societal levels; and we need to know more about the modern sources of traffic data.

Background and Motivation

The rise of real-time and online traffic data has the potential to generally increase the availability of the data that form the basis of traffic planning and adaptive demand, but it is also known that individual travelers are known to react by selecting their transport routes and modes in response to particular conditions of traffic data; main factors include traffic route delays, the reliability of data, and ambiguity aversion (Ben-Elia and Avineri 2015; Chorus et al. 2006). This clearly motivates a study of the availability aspect of heterogeneous traffic conditions. Individuals and small groups of travelers may be mainly interested in single or some small set of particular routes and the reliability of traffic information for the route to their destination, but demand for these data may come also from others where, e.g., city planners may be interested in getting an overall view of the city or relevant urban area to quantify, track, and follow-up regular and unexpected traffic.

HERE Traffic (HERE) is one of the several large-scale data sources with the capability to collect and provide information about real-time traffic in at least 83 countries to date (for more about coverage, see http://www.here.com). The data are available through an open application programming interface (API) that is partially free, partially commercial, with access up to a certain data limit. In addition to traffic speed information, HERE also collects incident and accident information including location, duration, severity, as well as other data such as real-time weather information from multiple weather stations close to cities. Given the wide potential of using these data for both commercial and public use, there has been little research to date that provides independent evaluation of the data availability in one of these platforms that tracks traffic on the large scale. A scientific evaluation of this type of data that highlights the possibilities as well as the limitations is both timely and critical for travellers, researchers, practitioners, and private entities who can use the information to further models, tools, and make planning decisions for traffic in cities.

Traditional Sources and Online Real-Time Traffic Data

Reliable transportation information is arguably one of the most important services needed in an urban environment. Traditionally, speed data are collected by government agencies setting up fixed-point sensors in selected major arterial or freeways across cities or some rural roads. In recent years, traffic services such as Google Traffic, Tom Tom Traffic, Here Traffic, INTRIX Traffic, and Waze began offering traffic services including speed, travel time, congestion information, and accident reports that have now reached a much wider range of the public than never before. Most of these services rely on floating car data (FCD), or probe vehicle data where vehicle speed data are collected from a variety of connected vehicle sources such as in-built navigation services, commercial vehicle logistics and tracking devices, and from mobile phone applications (Jurewicz et al. 2018; Ambros and Jurewicz et al. 2017).

Numerous studies including government reports and academic studies have examined the quality of FCD. Most of these studies compare the level of similarity between FCD and a ground truth data source, typically stationary detector data, in terms of the relevant traffic variables, e.g., speed and travel time (Jurewicz et al. 2018; de Boer and Krootjes 2012; Clergue and Buttignol 2014; Clergue and Buttignol 2015; Hrubes and Blümelová 2015; Diependaele et al. 2015; Ambros et al. 2017). Some also look at other aspects such as the coverage of the road network (Jurewicz et al. 2018; Aarts et al. 2015) or timeliness to recognize jams (Hu et al. 2016; Kessler et al. 2018; Wang et al. 2014). It has been suggested that theoretically mean point speed from sensors would often be greater than mean link speeds from FCD (Jurewicz et al. 2018) and this has turned out to be the case in some empirical observations (Jurewicz et al. 2018; Clergue and Buttignol 2015; Hrubes and Blümelová 2015). Jurewicz et al. (2018) found FCD speeds are on average 23% lower than mean loop point-speeds. Others, however, found FCD speeds higher than fixed-point measurements (Diependaele et al. 2015; Ambros et al. 2017). Studies have also suggested poor agreements between private and ground truth data, and concluded that private sector data are not suitable for real-time measurements as they tended to show less variability though they could still be suitable for a longer-term trend analysis (Hu et al. 2016).

Characterizing Traffic Information and Congestion, and Known Limitations

In this paper, we present a systematic overview of how a large-scale and high-availability online data source provides traffic information for 45 cities in different countries. Because of the nature of the commercial data source, algorithms and methods to compute some of the traffic information are proprietary, so the accuracy of this private sector data is not fully transparent. This creates challenges for broader use by the public, government agencies, private industry, researchers, and we can expect an increasing

demand for validation and performance reports to become public. In this paper, we, therefore, describe how cities have varying levels of available real-time data and examine how to fill in the gaps for roads where data are unavailable, but we also examine what challenges might arise when attempting to validate the information from the data source against ground truth road sensor data. This paper is organized as follows. In the next section, we present key summaries regarding traffic information, including data description, data availability, and data pre-processing. In the following section, we use time series clustering to discover and explore patterns of real-time traffic data availability. Moreover, we provide a case study of validating the traffic data against ground truth, and describe, evaluate, and compare methods to fill in missing data. The results are summarized in the last section where we also discuss and suggest future research.

Data Exploration

We collected traffic data from 50 cities during 12 months, and after filtering and pre-processing outliers and periods with measurement errors, ended up analyzing 45 cities during a 6-month period. This section outlines data characteristics and pre-processing together with a first look at different patterns in the data.

Data Collection, Data Volume, and Data Description

We collected traffic data from 50 cities approximately every 5 min from Jan 1 2018 to December 31 2018, and because of slight changes in timing of the sampling and varying network delays, we group these samples into 15-min time windows (96 windows per day, in total 35,040 samples per road segment in each city). After collecting all samples into the corresponding time window of the day (96 bins), we average the measured traffic speeds in each bin for each day. We, thus, use the same number of time windows per day throughout the sample period to simplify data processing and use this as a basis for comparison. We shifted time stamps to the local time zone for each city before further processing.

Roads are geographically represented by road segments as a sequence of edges with WGS 84/GPS coordinates. The cities were chosen to represent several different countries, different types of urban environments, and areas that should include both highly congested or with relatively low congestion. There were also countries and cities for which we were not able to obtain any data. City characteristics including number of roads with measurements, number of roads seen persistently throughout each month of the year, and other information are provided in Table 1, with all units being in km and traffic speeds in km/h. Our data collection covers mainly major cities, but we also did include two cities that can be characterized as greater urban areas: Amsterdam and Johannesburg/Pretoria. For details about data collection and filtering, see Supplementary Information A.1. For maps using osmnx (Boeing 2017), see Supplementary A.3.

According to HERE, the traffic data come from "billions of GPS data points every day and leverage over 100 different incident sources to provide a robust foundation for our traffic services" (see https://www.here.com/en/products-services/here-traffic-suite/here-traffic-overview). The information is collected from a variety of devices in the cities, including vehicle sensor data, smart phones, personal navigation devices, road sensors and connected cars, as well as public incident and accidents reports (HERE 2016). Traffic data are asynchronously updated in the HERE infrastructure in approximately 3-min intervals. The data have a typical delay between 1.5 and 3 min in relation to the real-world state (HERE 2016).

The real-time traffic data were obtained by purchasing and using network access to the HERE API and using computer programming to request and download the data every 5 min. More specifically, with the HERE flow API, each request gives an additional set of features besides traffic flow speed that includes the time when traffic information was last updated for the road segment, confidence score (whether the traffic speed data is a real-time measurement, or a historical estimate), the direction of traffic, free flow traffic speed, traffic speed limit, and a geographical description of the road segments as a sequence of WGS84 coordinates.

Data Availability, Persistent Roads, and Removal of Measurement Errors

There is variation over time both between cities and within a city in the number and share of the road segments with real-time information. For several cities, their road networks had small changes over the months: in some cases, some road segments have been added by HERE during the sample period, and in some cases road segments were removed during the year in a re-design of the network. We wanted to study the variation in traffic information availability under typical long-term conditions, so we filtered out the minority of road segments that were either added or removed during the year. For more details about the pre-processing, see Section A.1 (Supplementary) and the properties in Table 1 describes the 45 cities studied over the first 6 months of 2018.

We start by looking at how data availability varies over time in the cities: Fig. 1 shows variations across the day for four of the cities. The scatterplots of data availability against time of day shows that day times and what are typical peak hours are associated with a higher level of available traffic information. This makes intuitive sense given that the traffic speeds are significantly influenced by traffic flows. It suggests that, as more real-time measurements are available when traffic delays are high, the mean value of data

Table 1 Summary statistics for the 45 out of 50 cities that were analyzed for a 6-month period, after filtering out periods of measurement error in the second part of the year and non-persistent roads

City	Number of road segments	Percent persistent	Average length (km)	Min length (km)	Max length (km)
Amsterdam	7366	98	1.44	0.004	15.36
Auckland	4603	96	0.66	0.006	16.95
Bangalore	4283	94	0.53	0.005	10.40
Bangkok	12,222	98	0.50	0.002	10.18
Barcelona	9760	88	0.74	0.003	50.34
Berlin	4440	89	1.18	0.005	16.17
Buenos Aires	4716	96	0.36	0.004	6.29
Cape Town	3897	100	1.11	0.006	33.24
Chicago	14,809	98	0.85	0.002	9.57
Detroit	8117	98	1.25	0.005	12.95
Dublin	9087	99	0.28	0.003	10.41
Edinburgh	231	89	2.43	0.014	38.58
Florence	1141	87	1.75	0.007	20.22
Glasgow	562	78	1.97	0.012	21.84
Gothenburg	2118	98	0.73	0.005	11.83
Jakarta	12,880	100	0.46	0.002	9.71
Johannesburg Pretoria	14,533	96	1.20	0.007	98.51
Kuwait City	1796	99	0.55	0.008	6.56
London	5510	91	1.59	0.006	25.24
Los Angeles	12,833	96	0.78	0.005	7.94
Madrid	4532	92	0.56	0.003	14.80
Makkah	2041	97	0.79	0.006	16.99
Marseilles	754	47	0.99	0.007	11.47
Mexico City	3604	99	0.81	0.007	35.24
Moscow	23,648	99	0.52	0.004	34.35
Mumbai	4590	95	0.36	0.005	4.93
New York	20,855	96	0.77	0.005	21.88
Oslo	1138	84	0.83	0.007	8.71
Ottawa	1863	99	2.34	0.003	21.59
Oxford	98	82	2.45	0.039	7.70
Palermo	556	90	1.03	0.004	24.32
Paris	7661	26	1.33	0.008	14.22
Prague	5338	93	0.57	0.006	8.49
Rio	7025	96	0.47	0.004	11.48
Riyadh	7759	99	0.73	0.005	33.78
Rome	3427	94	0.69	0.003	9.44
San Francisco	3370	97	0.48	0.007	18.56
Sao Paulo	18,959	90	0.39	0.002	13.73
Sofia	11,656	94	0.15	0.003	5.34
St Petersburg	9466	99	0.78	0.006	17.62
Stockholm	3435	99	0.73	0.006	16.85
Sydney	9424	95	0.53	0.002	17.04
Tallinn	829	98	0.72	0.007	10.19
Vienna	4019	92	0.67	0.006	12.38
Warsaw	2980	97	0.57	0.006	7.33

Further details are found in Sections A.1 and A.3 (Supplementary Information)

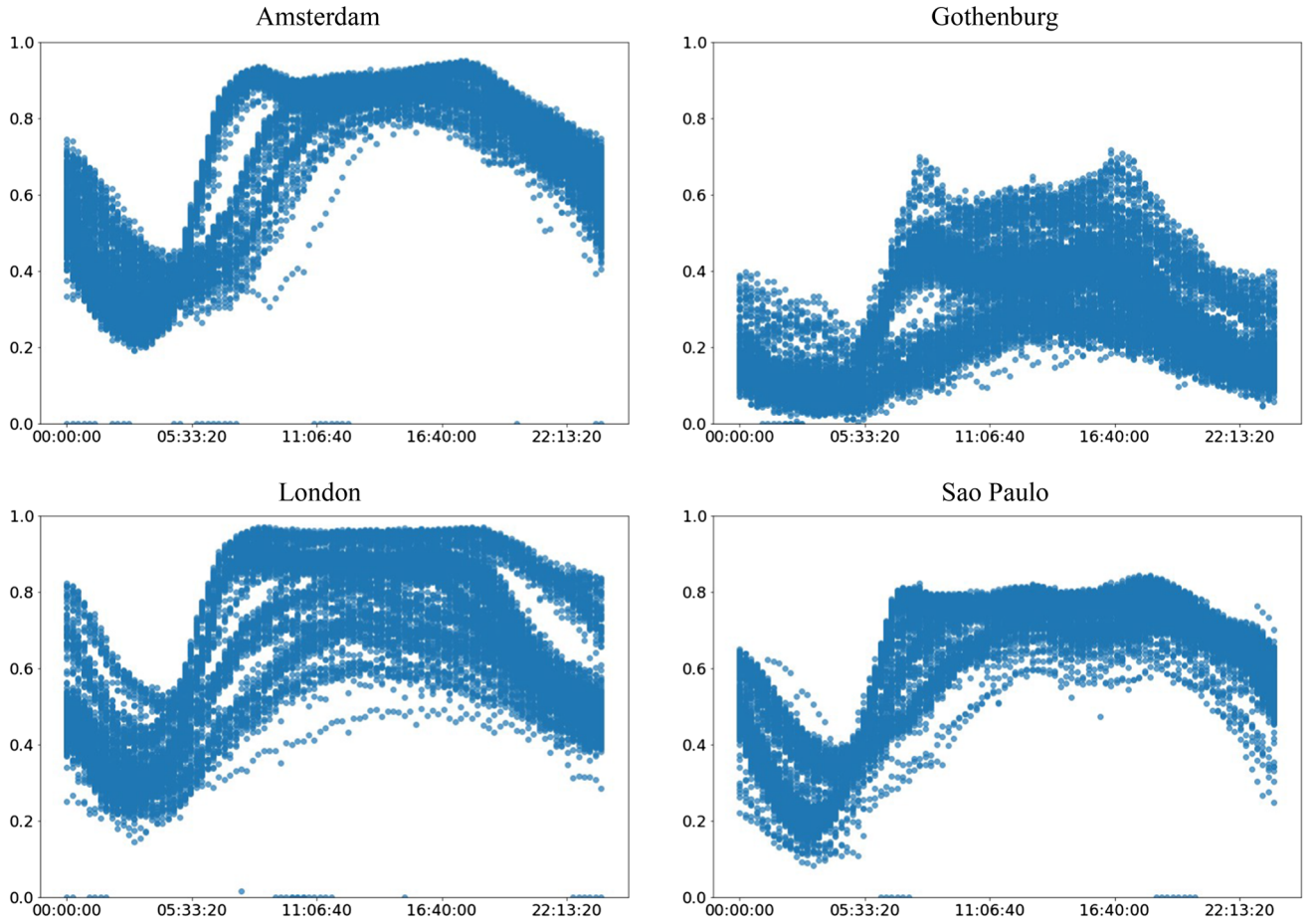


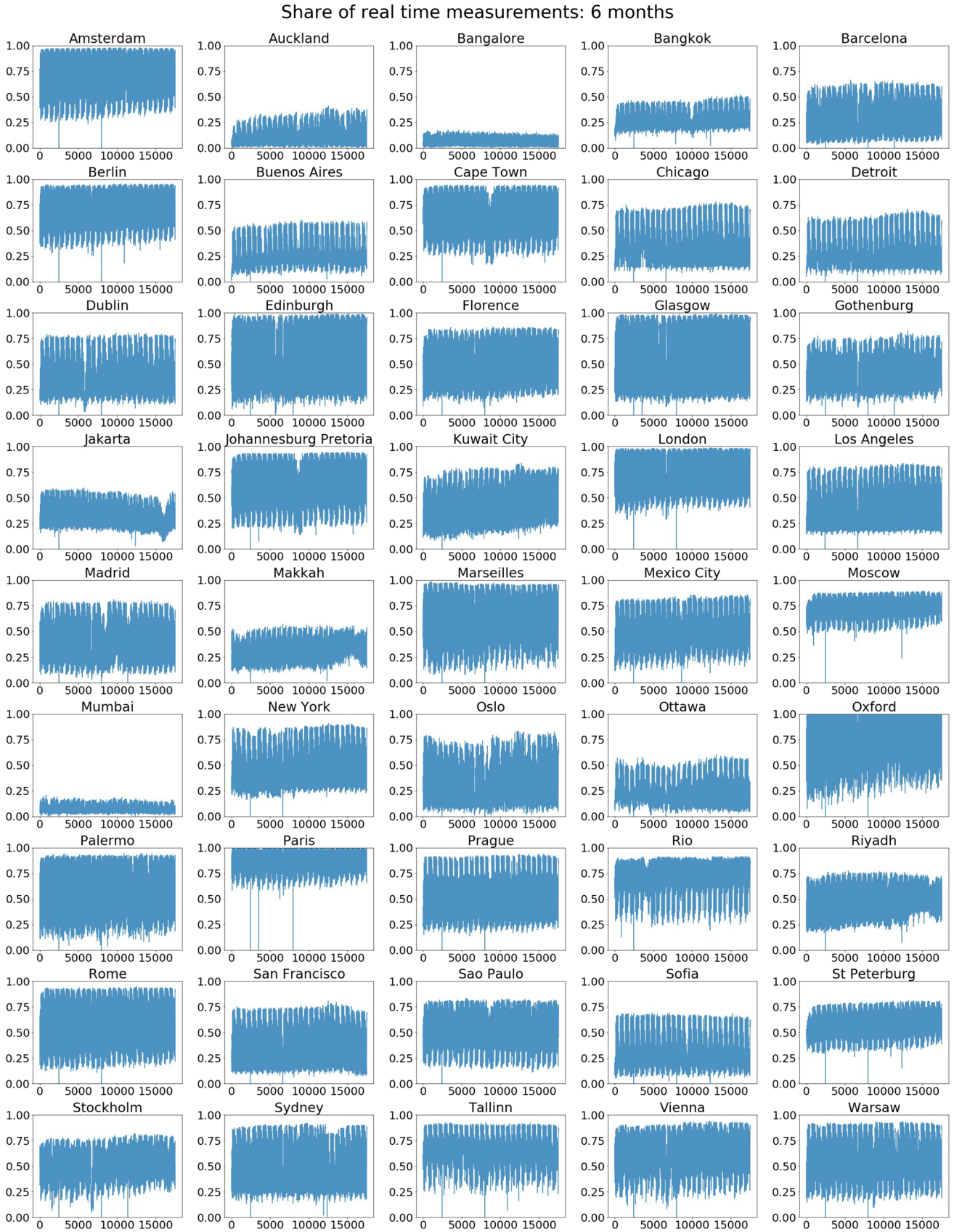
Fig. 1 Share of real-time traffic data, distributed across time of day. Traffic data with real-time information (vertical axis) as share of roads, vs. time of day (horizontal axis). Examples from four cities during 6 months, where time has been adjusted to the local time zone. In general, day times are associated with a higher share of real-time measurements, suggesting that larger traffic volume is significantly related to the number of real-time measurements. For each city, there can be several distinct levels of measurement at the same time of day (partially explained by the differences between weekdays and weekends). Zeroes are outliers due to a small rate of measurement failures in this study

availability could be misleading. Moreover, interestingly, Fig. 1 also shows that there exist distinct states or levels of how well a city is measured. The cities can be in a few different distinct states at the same time of the day; this partially depends on the differences between weekdays and weekends.

Figure 2 shows longer-term time series of data availability, and that there are large differences in the levels of the share of road segments with real-time measurements. Some even have a slowly increasing trend, which may be connected to slow changes in technology. The short-term dips are consistent with different types of public holidays and shorter vacation periods. The few spikes downwards to a level of zero illustrate network outages on our side of the data collection process. Supplementary Figure A.8 shows a weekly time series for the first week in 2018, and that besides the level we have a complex seasonality with repeated variation across the day, between weekdays/weekends, and between weeks. This suggests that part of the explanation for the distinct levels in Fig. 1 is the variations of days and weeks. As for whether there is a grouping of cities beyond the obvious mean levels of data availability will be addressed below with time series clustering.

Traffic Data Analysis

As we have seen above, most roads in all cities have some degrees of missing data. We address the following research questions: (i) Are there regular recurring patterns of data availability that differ between different cities? (ii) Are the real-time data measurements valid, and what challenges can arise with validation? (iii) What methods are efficient in filling in the missing data when real-time measurements are not available for some road segments? This section addresses these three questions to assist with using the data source for further analysis.


Share of real time measurements: 6 months
Amsterdam
Auckland
Bangalore
Bangkok
Barcelona
Berlin
Buenos Aires
Cape Town
Chicago
Detroit
Dublin
Edinburgh
Florence
Glasgow
Gothenburg
Jakarta
Johannesburg Pretoria
Kuwait City
London
Los Angeles
Madrid
Makkah
Marseilles
Mexico City
Moscow
Mumbai
New York
Oslo
Ottawa
Oxford
Palermo
Paris
Prague
Rio
Riyadh
Rome
San Francisco
Sao Paulo
Sofia
St Peterburg
Stockholm
Sydney
Tallinn
Vienna
Warsaw
0.00
0.25
0.50
0.75
1.00
0
5000
10000
15000

◂**Fig. 2** Data availability for 45 of the cities in 15-min time windows during 6 months: The share of road segments in each city with real-time information. Spikes (to zero) represent network problems as part of our data collection process (HERE data was always available as far as we could observe; so measurement errors were on our side.)

Patterns of Data Availability

To find patterns of data availability beyond those seen visually in Fig. 2 and on aggregate forms such as in Table 1, we consider the shape of the time series. To focus on variation of the shape, we standardize the time series (shift with mean and divide by standard deviation) to capture similar patterns over time. We thus chose to disregard the absolute level to find whether there are similarities in how availability varies over time.

After standardization, we still have significant variation in the time series; for an illustration of this, see Fig. 3 that shows 4 weeks of data (May–June 2018). Some of the variations among standardized time series are shifts in time (despite that the series have been adjusted to local time zone); different cultural habits might mean that peaks occur at different times of the day, i.e., that they can be the similar patterns but shifted in time. Moreover, daylight conditions and using the same time zone in a large country might mean that there are differences in cities that are far apart, affecting their peak hours. We would also like to take into account these shifts to group cities that have similar patterns throughout the day, even if they are shifted by a lag (we expect this to be on the scale of minutes to hours).

To address this property of the series, we compute a similarity metric between every pair of time series using dynamic time warping (DTW) (Aghabozorgi et al. 2015; Sardá-Espinosa 2019). This will find the best possible match between a pair of time series by considering the different possible shifts to minimize the distance (how well the sequences can be aligned optimally with each other to minimize the sum of absolute distances between pairs of aligned indices). We then use this distance metric to cluster time series into groups of similar series. One instance of hierarchical clustering (using Ward's method) is shown in Fig. 4.

With a hierarchical clustering, the number of clusters is not obvious; it just gives us many different possible ways to partition cities into different numbers of clusters. To address this arbitrariness of choice, we evaluate the quality of different clusters ranging from 2 to 8 clusters. Using six quantitative metrics that have been proposed to measure quality for time series clustering algorithms including Silhouette index (Rousseeuw 1987), Calinski–Harabasz index (Johnson and Wichern 1988), DB index (Davies and Bouldin 1979), Modified DB index (Kim and Ramakrishna 2005), Dunn index (Dunn 1974), and the COP index (Arbelaitz et al. 2013), the results are shown in Table 2. From these results, we judge that the cases with either four or two clusters are most interesting as these cases have a majority of the two best scores for the different clustering metrics.

The case with four clusters is shown in Fig. 5 (the same data as in Fig. 4, but the series are now grouped in their clusters).

- Cluster 1: Amsterdam, Berlin, Cape Town, Edinburgh, Florence, Glasgow, Johannesburg/Pretoria, London, Marseilles, Mexico City, Moscow, Oxford, Palermo, Paris, Rio, Rome, Sao Paulo, St Petersburg, Vienna, Warszawa.
- Cluster 2: Auckland, Buenos Aires, Chicago, Detroit, Dublin, Ottawa, Sofia.
- Cluster 3: Bangalore, Jakarta, Kuwait City, Makkah, Mumbai, Riyadh.
- Cluster 4: Bangkok, Barcelona, Gothenburg, Los Angeles, Madrid, New York, Oslo, Prague, San Francisco, Stockholm, Sydney, Tallinn.

We also note that clustering often keeps cities that are culturally related together, e.g., as being in the same or neighboring countries. This suggests that the clustering picks up groups of cities that are similar in shape partially because of spatial dependence. Future work would be needed to form an explanation for these differences and to explore what these factors precisely are.

Validation: A Case Study Based on Swedish Traffic Sensor Data

The wide geographical and temporal span of the data means that full validation would require a coordination of large research efforts. A first step towards more validation can be a case study of validation with sensor data that we can access. The primary aim with the following analysis is to provide a comparison between the HERE data and the ground truth, and to provide a qualitative understanding of questions that might emerge when comparing the HERE data to the ground truth. Our attempt is to conduct analyses that are simple and clear to explore both possibilities and challenges for future validation efforts.

We have access to traffic sensor data collected by the Swedish Transport Administration. The sensors are located at major roads and highways to measure traffic speeds and vehicle volumes in several large Swedish cities. The data includes vehicle counts as well as mean vehicle speeds every 1 min throughout 2018. We compare HERE traffic speeds from road segments with data from these traffic sensors located on the same roads; as a first step, we pair HERE road segments with high availability to roads where there are also traffic sensors with high availability (some missing data was also the case for some of the road sensors). More details about location can be found in the Supplementary

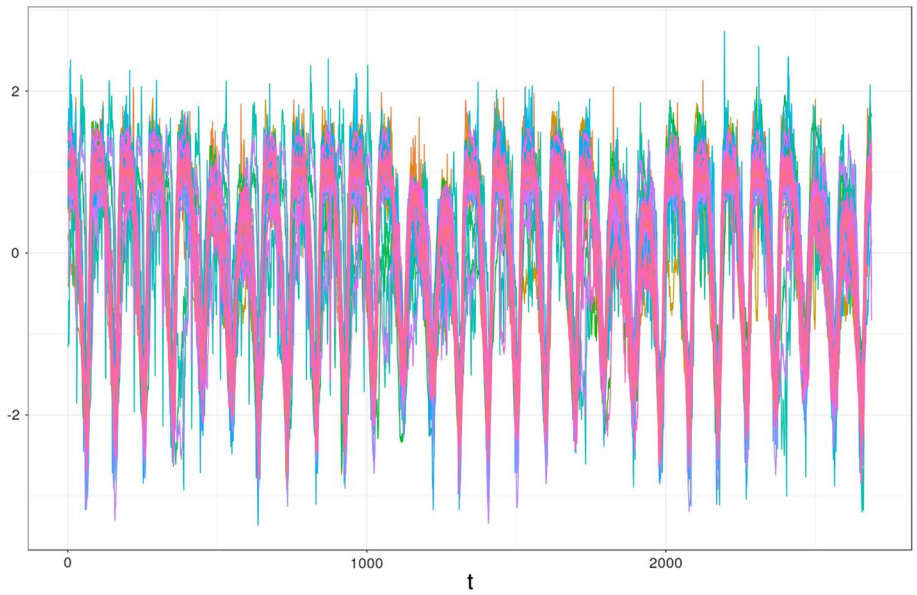


Fig. 3 The standardized time series for the 45 cities used for shape-based time series clustering. 4 weeks from May to June 2018

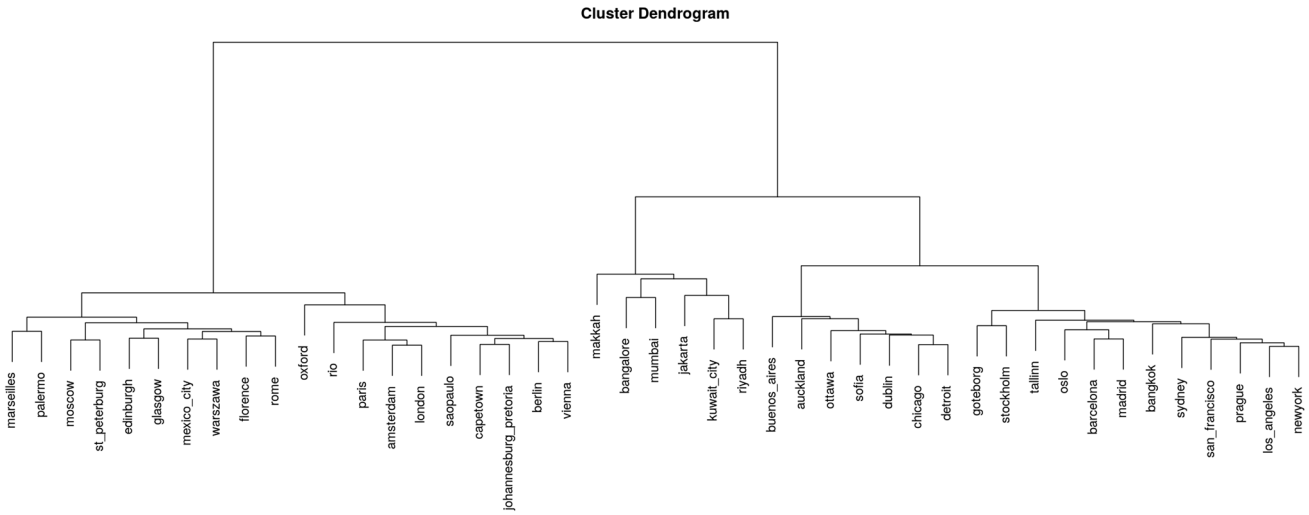


Fig. 4 Hierarchical clustering of the 45 time series, using Dynamic Time Warping to compute the distance metric between each pair of series, and using Ward's method for hierarchical clustering. The tree shows that the candidates for the number of clusters for grouping cities tends to group cities with geographical proximity

Table 2 Clustering the standardized time series for 45 cities: comparing different metrics to help discover an appropriate number of clusters

Number of clusters	Silhouette index	Calinski–Harabasz index	DB index	DB* index	Dunn index	COP index
2	**0.226**	34.963	**1.349**	**1.349**	0.221	0.481
3	0.171	20.388	1.873	1.890	0.230	0.402
4	**0.177**	15.129	**1.415**	**1.428**	**0.268**	0.326
5	0.107	10.597	1.850	1.992	0.182	0.323
6	0.047	8.819	1.765	1.981	**0.235**	0.306
7	0.059	**7.085**	1.682	1.996	0.223	**0.295**
8	−0.013	**5.796**	2.001	2.674	0.197	**0.292**

DB* index stands for modified DB index

For each index proposed to measure quality of a clustering, the two best (either smallest or largest) values are printed in bold

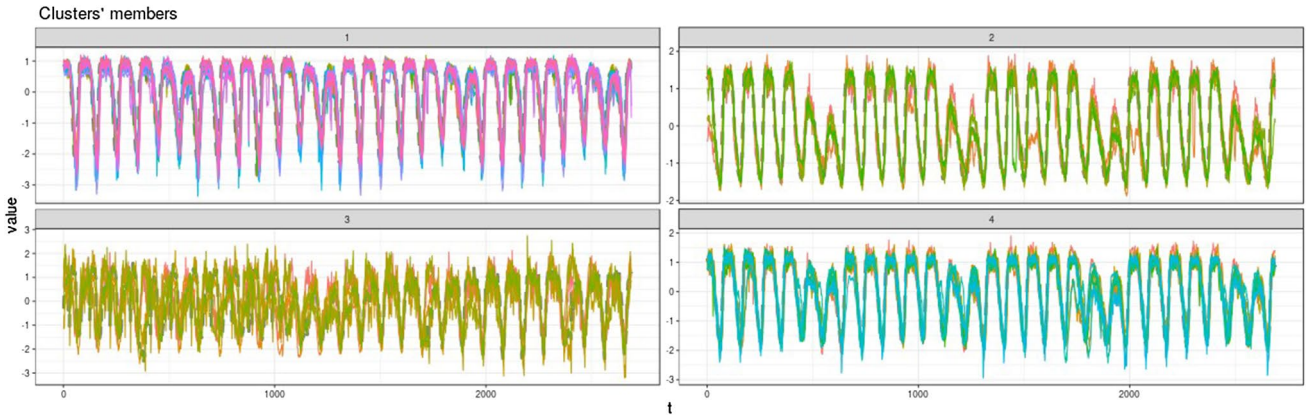


Fig. 5 Four clusters of time series, hierarchically clustered with dynamic time warping distance and Ward's method. Showing four weeks out of the 6-month time series. Cluster 1 (top left): weekdays and weekends are similar. Cluster 2 (top right): weekdays and weekends differ notably, with an additional dip on Sundays. Cluster 3 (lower left): noisy measurements with few weekday/weekend patterns. Cluster 4 (lower right): a weaker dip on Sundays

Table 3 Characteristics of the road segments, lanes and sensors in the validation case study

Road segment	Direction	Lane type	Sensors	Description
E_10353−	Northbound	Slow	5476, 5479, 5482, 5485, 5487	Highway, 2.92 km
E_10353−	Northbound	Fast	5475, 5478, 5481, 5484, 5486	Highway, 2.92 km
E_10352+	Southbound	Slow	5520, 5517, 5514, 5511, 5508	Highway, 2.03 km
E_10352+	Southbound	Fast	5519, 5516, 5513, 5510, 5507	Highway, 2.03 km
E_10381−	Eastbound	Slow	35,027	Highway, 0.54 km
E_10381−	Eastbound	Fast	35,028	Highway, 0.54 km
E_10380+	Westbound	Slow	35,025	Highway, 0.83 km
E_10380+	Westbound	Fast	35,026	Highway, 0.83 km

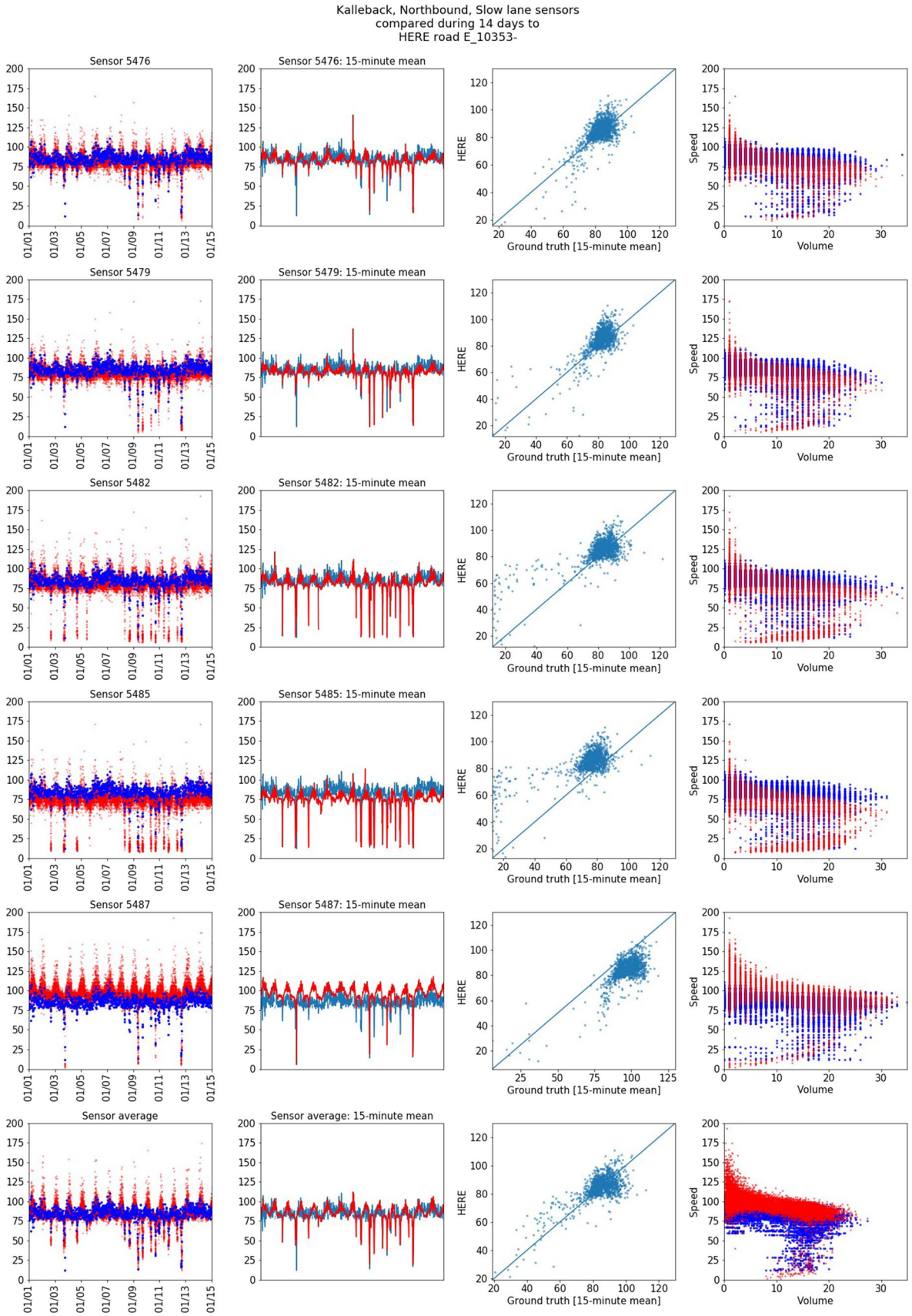
Validation includes sensors from in total 8 lanes intersecting with 4 HERE road segments, and comparisons with 24 traffic road sensors both on the lane and road segment

Information. Supplementary Figures A.9 and A.10 show a road network map of Gothenburg, Sweden, including where HERE data had availability share greater than 0.9, and the sensors had availability share greater than 0.99. The HERE road segments and number of sensors are summarized in Table 3.

We illustrate the HERE data (grouped into 15-min intervals) directly to the sensor data (collected and averaged in 1 min intervals), but besides this, we also make the comparison where the readings for each traffic sensor was aggregated into 15-min time intervals by taking the mean speeds in each interval. Measuring a road segment at only one point (one sensor) has obvious limitations especially as the HERE road segments have lengths on the scale of a kilometer; another issue is that the sensors measure traffic in particular lanes, but the HERE data treat a two-lane as one road segment (in our case with these particular roads). We, therefore, study several sensors along the same road segment and also consider taking the speed average along consecutive sensors in the same lane, and consider whether there are differences between sensors that are placed in parallel in two lanes. For the case study we had available (i) two longer road segments with two lanes where there are high-availability sensor measurements at more than one point along the road segment, and (ii) two road segments where there is a set of high-availability sensor measurements for two lanes sitting in "parallel" in practically the same location (metres of each other).

Figure 6 illustrates the results for traffic measurements from a slow lane and a corresponding HERE road segment with a length of 2.9 km. This road has two lanes in each direction, and five sensors along each lane. So, we consider in total five sensors; these are from the bottom five locations from the road network illustrated in Supplementary Figure A.10 and the speed readings are shown on the different rows in Fig. 6. The figure shows a comparison between sensor data and the corresponding HERE speeds for a period of the first 14 days in 2018.

From the figure, we can see that for this particular road segment (i) sensors can be biased toward both higher (sensor 5487) or lower speeds (sensor 5485) compared to the HERE data, (ii) the average of the different sensors tend to follow


Kalleback, Northbound, Slow lane sensors
compared during 14 days to
HERE road E_10353-
Sensor 5476
Sensor 5476: 15-minute mean
Sensor 5479
Sensor 5479: 15-minute mean
Sensor 5482
Sensor 5482: 15-minute mean
Sensor 5485
Sensor 5485: 15-minute mean
Sensor 5487
Sensor 5487: 15-minute mean
Sensor average
Sensor average: 15-minute mean
HERE
Ground truth [15-minute mean]
Speed
Volume
01/01
01/03
01/05
01/07
01/09
01/11
01/13
01/15

◂**Fig. 6** Five sensors cover a northbound highway to Gothenburg, Sweden. Sensor data (red) vs HERE data (blue) from the first 14 days of 2018. Speeds (km/h) vs time on the two left-most plots. Ground truth speed vs HERE speeds in the third column. Volume given in vehicle counts (vehicles/minute). Each of the top five rows corresponds to one sensor along the road segment. Bottom row: data based on averaging the five sensors. The right-most column has speed–volume plots that show traffic volumes vs. speeds using the sensor speeds (red) and the HERE speeds (blue) (the volume data from the sensors)

the profile of the HERE speeds except for the high speeds, (iii) HERE speeds tend to follow dips (down-wards) in real time to a larger degree than spikes upwards (iv) traffic in individual locations can be much slower than the HERE road segments (third column, rows 3 and 4). A likely explanation for (iii) and (iv) can be seen in the speed–volume plots on the right in general there are fewer vehicles that can be measured when the speed is high. Taken together, this seems consistent with the idea that HERE real-time traffic information tends to describe an average traffic speed in several locations along a road for a number of vehicles.

Similar findings showing more traffic sensors matched with the HERE data for both fast and slow lanes along this parallel stretch of highway, in both directions, are found in Supplementary Figs. A.11, A.12, and A.13. These different cases illustrate similar patterns for a majority of the sensors, generalizing to longer time periods, suggesting the hypothesis that the HERE system captures a majority of the speed dips on major roads (averages in bottom row of each plot).

One issue can also be demonstrated when we turn to another road segment with two lanes and one sensor in each lane (at the rightmos location in Supplementary Fig. A.9). Supplementary Figures A.14 and A.15 show pairs of sensors are located nearly in parallel on the same stretch of road for 7 days; we observe that the fast lane at the same location has on average higher speeds as expected, and the fast lane seems mainly used during day time, but besides that it looks similar to other sensor readings. However, consider the time series over the longer time span of 180 days: Supplementary Figs. A.16 and A.17 reveal an artifact. The dips to the far right of the series show a period of low traffic speed where there are no correspondent values in the HERE traffic data. The explanation for the low sensor readings was planned road works during the time with restricted access to one of the lanes; this shows that there are certain kinds of speed dips that can go undetected also on shorter road segments. To learn more about the scale at which measurements from moving vehicles become representative of one point on the road is a question for further validation efforts.

Taken together, our findings in this case study is that HERE traffic data tends to agree with the ground truth when averaged across several sensors along a HERE road segment, and it seems that HERE can detect and report on many dips in traffic speed in real time. It is important to note that the limitations of our conclusions are: (i) based only on a few roads, (ii) only highways with large volumes of traffic, (iii) longer road segments, and (iv) a case study in one city. We expect that any further efforts to validate the information could add more nuance to these findings, e.g., for different types of roads in the road network of a city. Future work is needed along these different aspects of validation, and we have outlined several questions that can be investigated in future studies.

Filling in Missing Data: Evaluating and Comparing Methods

In this section, we address that the road segments do not have real-time traffic information at some of the time windows. We compare different statistical methods to fill in the gaps with missing data, based on relationships in the available data, and filling in the data in retrospect and not in real time.

For each 15-min time window in the HERE data, if there was no real-time measurement available we consider it a missing value. We, thus, fill in a matrix of type (17,472, n_i), reflecting all 15-min time windows in the first 6 months of 2018 with n_i as the number of persistent road segments in city i (Table 1). As we have seen above, the cities vary not only with respect to the level and profile of data availability but also with several other characteristics, and it is not obvious whether some method to fill in missing values would work better in one city than others in another. We evaluate four different methods to fill in missing values: (i) A mean-based method that simply fills in all missing values for a particular road segment with the observed mean speed value of the segment (we can see this as a simple baseline), (ii) A correlation-based method depending on the previously observed correlations between pair-wise real-time measurements for each pair of road segments (resulting in a linear regression, with the predictors at a given time chosen from the observed roads at that time), (iii) A k-nearest neighbors-based method (knn) that fills in a value for road i in a given row based on an average of k most similar other rows in the data where there are real-time measurements for i, and (iv) A sliding window k-nearest neighbors-based method, using a time window of 1 month as basis for filling in the missing value (knn window). In the last case, the method is the same as for the k-nearest neighbors, but with a temporal restriction of the data. The latter could possibly have the advantage to take into account factors such as the differences between months of the year, while restricting the available data to a more relevant time period. This, however, carries with it the tradeoff between the size of available data versus choosing more recent data. The evaluation of each method was made with respect to the root mean squared error (rmse) and tenfold cross-validation: For each city and each method, a random 10% of the known real-time measurements are held

Table 4 Evaluating different methods on filling in missing traffic speeds value with the metric average root mean square error using tenfold cross-validation

City	mean	correlations	knn	knn window
Barcelona	7.63	7.46	4.79	4.95
Berlin	7.58	7.05	4.58	4.24
Cape Town	8.86	7.84	4.89	4.61
Chicago	5.53	5.19	3.03	2.95
Detroit	5.26	5.11	3.20	3.26
Florence	7.08	6.53	4.23	4.12
Gothenburg	6.91	6.70	4.35	4.82
London	7.42	6.12	3.57	3.23
Moscow	6.59	6.23	4.21	4.06
New York	5.46	5.08	2.94	2.48
Sao Paulo	7.15	6.50	4.11	4.01
Stockholm	7.73	7.43	4.47	4.15
St Petersburg	7.84	7.33	4.74	4.52
Rio	8.07	7.20	4.48	4.42

Methods from left to right: (i) mean, (ii) historical correlations, (iii) k-nearest neighbors (full data), (iv) k-nearest neighbor (restrict to same month). The k-nearest neighbors were run with $k=10$ and consistently out-perform the naive and correlations-based methods

out from fitting the method and the method is evaluated by predicting these. The results of evaluating the four methods are summarized in Table 4 for data from 14 cities during 6 months, and the scores show that there are several consistent results across the methods.

First, across all cities, using historical correlations improves on the naive mean-based method. Second, the knn methods consistently give better results than the other methods. This improvement is also larger in size than what was gained from historical correlations, which suggests that traffic speed has a important non-linear dependence. Third, the typical (but relatively smaller) improvement with the time-dependent knn method indicates that the cities can be in different states during different times of the year and that estimation can be improved by taking this into account; we do not always see an improvement when estimating the missing value based on more recent observations, but it does not tend to make the result much worse.

Taken together, the knn methods consistently perform better than the simpler methods, despite the clear differences in the dimension and road network characteristics of the cities. Different directions would be possible to pursue to improve on the results, and it would be possible to including knowledge about the geography of roads and the structure of the road network. This demonstrates that it is possible to fill in the missing values to have an average root mean square error less than 10 km/h. Further work would be needed to explore the tradeoffs between sample size and other performance indicators that are important for applications.

Discussion and Conclusion

We analyze 6 months of traffic data from 45 large cities/urban regions available in one of the large-scale online platforms available to travelers, policy-makers, and researchers interested in city traffic around the world. We examine several areas where cities may vary and examine the data availability. Despite varying characteristics of the cities such as different road segment length, shares of real-time measurements, and difference in the patterns of data availability, a few common characteristics emerged from of our observations and results.

The findings include (i) using time series clustering, we identify four clusters of cities with distinct groups of data availability patterns, with the main difference being how availability changes in weekends and (ii) k-nearest neighbor based methods consistently improve on other methods to fill in missing values for traffic speeds. Moreover, (iii) the validation case study with ground truth for one city shows that the HERE data can follow dips in traffic speed quite well in real time, more so than sudden increases in speed that happen with fewer vehicles on the road. Taken together, this data source can be a basis for further research leading to a more complete view of city traffic compared with sensor data at fewer locations. We also found some challenges further research using this scale of data over time, including changes in the road networks by the addition or removal of road segments between months. It can also be important to take into account granularity of ground truth data with respect to different lanes on a road, and there tends to be better agreement between online data and ground truth when averaging several sensors.

Future work could be done in the following directions. Adding more data such as weather, socio-economic variables, and information about traffic incidents could be used to understand the variability in traffic data. Another possibility is to zoom in on particular parts and properties of the road networks to find similarities and differences in the cities. An important issue will be to improve our understanding of the coverage and validation against more types of ground truth data; to study the possibility of using different publicly available data sources for large-scale validation could be important, as validation with traffic sensor data on a large scale can be both expensive and difficult. Understanding the reasons behind the identified clusters of cities having similar data availability patterns could add to our understanding of why traffic information is available on different levels and at different times in the different cities, and the relationship to traffic congestion/delays.

In this study, we find patterns in how and when traffic data is available, and show that there is a sound basis for further studies that are directly related to applications of traffic

data such as studies of traffic congestion and traffic delays. Congestion and traffic delays continue to affect many cities around the world, and, by understanding the availability of traffic data from new data sources like large-scale online platforms, and developing methods that fill in the gaps of missing data and address the new challenges with this data source, we see a promising basis to improve our knowledge about traffic in current and future cities.

Acknowledgements Open access funding provided by Chalmers University of Technology. This research is funded by the Areas of Advance in Transport, Energy, and Information and Communication Technology at Chalmers University of Technology. We appreciate data support and assistance provided by David Donk, Miho Ishii, and Petter Djerf from HERE Technologies.

References

Aarts LT, Bijleveld FD, Stipdonk HL (2015) Usefulness of floating car speed data for proactive road safety analyses: analysis of tomtom speed data and comparison with loop detector speed data of the provincial road network in the Netherlands. Report r-2015-3. Report, SWOV. https://www.swov.nl/sites/default/files/publicaties/rapport/r-2015-03.pdf

Aghabozorgi S, Shirkhorshidi AS, Wah TY (2015) Time-series clustering—a decade review. Inf Syst 53:16–38. https://doi.org/10.1016/j.is.2015.04.007

Ambros J, Jurewicz C (2017) From big data to speed and safety: a review of surrogate safety measures based on speeds from floating car data. In: 2017 Australasian road safety conference, Perth

Ambros J et al (2017) Improving the self-explaining performance of czech national roads. Transp Res Rec 2635:62–70

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recognit 46:243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Arnott R, Small K (1994) The economics of traffic congestion. Am Sci 82:446–455

Barthelemy M (2016) The structure and dynamics of cities. Cambridge University Press, Cambridge

Ben-Elia E, Avineri E (2015) Response to travel information: a behavioural review. Transp Rev 35:352–377

Boeing G (2017) OSMnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput Environ Urban Syst 65:126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004

Chorus C, Molin E, Van Wee B (2006) Use and effects of advanced traveller information services (ATIS): a review of the literature. Transp Rev 26:127–149

Clergue L, Buttignol V (2014) Using GPS data in favour of traffic knowledge. In: Transport research arena

Clergue L, Buttignol V (2015) Probe data and its application in traffic studies. In: 2015 IPWEA/IFME conference

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1:224–227. https://doi.org/10.1109/TPAMI.1979.4766909

de Boer G, Krootjes P (2012) The quality of floating car data benchmarked: an alternative to roadside equipment? In: 19th ITS World Congress

Diependaele K, Riguelle F, Temmerman P (2015) Speed behavior indicators based on floating car data: results of a pilot study in belgium. Transp Res Proc 14:2074–2082

Dunn JC (1974) Well-separated clusters and optimal fuzzy partitions. J Cybern 4:95–104. https://doi.org/10.1080/01969727408546059

Hensher DA (2018) Tackling road congestion—what might it look like in the future under a collaborative and connected mobility model? Transp Policy 66:A1–A8

HERE (2016) Speed data v1.3 specification version 1.0. Report, HERE Global B.V.

Hrubes P, Blümelová J (2015) Comparative analysis for floating car and loop detectors data. In: 22nd ITS world congress

Hu J, Fontaine MD, Ma J (2016) Quality of private sector travel-time data on arterials. J Transp Eng 142:04016010. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000815

Johnson RA, Wichern DW (eds) (1988) Applied multivariate statistical analysis. Prentice-Hall Inc, Upper Saddle River

Jurewicz C et al (2018) Use of connected vehicle data for speed management in road safety. In: 28th ARRB international conference—next generation connectivity

Kessler L, Huber G, Kesting A, Bogenberger K (2018) Comparing speed data from stationary detectors against floating-car data. IFAC-PapersOnLine 51:299–304. https://doi.org/10.1016/j.ifacol.2018.07.049

Kim M, Ramakrishna R (2005) New indices for cluster validity assessment. Pattern Recognit Lett 26:2353–2363. https://doi.org/10.1016/j.patrec.2005.04.007

Levinson D (2003) The value of advanced traveler information systems for route choice. Transp Res Part C Emerg Technol 11:75–87

Lyons G (2016) Getting smart about urban mobility—aligning the paradigms of smart and sustainable. Transp Res Part A Policy Pract 115:4–14

Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sardá-Espinosa A (2019) Time-series clustering in R using the dtwclust package. R J. https://doi.org/10.32614/RJ-2019-023

Transportation Research Board (2009) Implementing the results of the second strategic highway research program: saving lives, reducing congestion, improving quality of life. Report

Wang Y, Araghi BN, Malinovskiy Y, Corey J, Cheng T (2014) Error assessment for emerging traffic data collection devices. Tech. Rep., Washington State Department of Transportation

Xu Y, González MC (2017) Collective benefits in traffic during mega events via the use of information technologies. J R Soc Interface 14:20161041