Traffic congestion of large cities through a large-scale online platform

Vilhelm Verendel[1,*] **and Sonia Yeh**[2]

[1]Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg 41296, Sweden
[2]Department of Space, Earth and Environment, Division of Physical Resource Theory, Chalmers University of Technology, Gothenburg 41296, Sweden
[*]Correspondence and requests for materials should be addressed to V.V. (email: vive@chalmers.se)

ABSTRACT

Online real-time traffic data services have effectively delivered congestion information to people all over the world and provided great benefits to society. We study urban traffic data of 17 cities from a major online real-time traffic information provider. We sampled the online platform every 5 minutes over twelve months, in total more than 2 million samples covering over 170000 road segments. We use three variables to characterize **traffic information** for different cities: *data availability*, *duration*, and *reliability of travel delays*. Data availability measures the percentage of real-time information. Travel delays measure the average excess travel time for a given road segments (in travel time per kilometer). Reliability of travel delays is captured by the share of recurring congestion: the larger proportion of non-recurring congestion in total delays, the lower reliability of travel time estimates. We measure **traffic data quality** by developing methods of *filling in missing data* and compare that against data provider's "historical values;" and by *validating* against measurements of traffic sensors. Our goal is to make objective assessments of traffic data quality information more transparent and accessible. Future work is critically needed for collecting more case studies of ground truthing especially for developing countries.

Keywords: big data, urban traffic, non-recurrent congestion, travel delays, forecasting

1 Introduction

1.1 Background

While the rise of real-time and online traffic data has the potential to increase the availability of the data which form the basis of traffic planning and adaptive demand, individual travelers are known to react by selecting their transport routes and modes in response to traffic conditions with respect to certain specific characteristics: These include traffic route delays, the reliability of data, and ambiguity aversion[1,2], which warrants the study of the availability and recurring aspect of heterogeneous traffic conditions. The demand for this data may come also from others: Individuals or 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, conversely, city planners may be interested in getting an overall view of cities and the capability to quantify regular and unexpected traffic delays (recurring and non-recurring congestion, respectively).

HERE Traffic is one of the many data sources with the capability to collect and provide information about real-time traffic, incidents and accidents information globally in 83 countries to date, with "over 100% coverage for 57 countries", according to their website[1]. The data is available through an open 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. Taken together, this can allow travelers as well as city planners to get an overview of the traffic conditions in cities. Given the wide potential of using this data for

[1] http://www.here.com

both commercial and public use, there has been little research to date that provides independent evaluation of the data coverage and data availability. A scientific evaluation of this type of data that highlights the possibilities as well as the limitations is both timely and critical for researchers, practitioners, and private entities who can use the information to further models, tools, and make planning decisions for traffic in cities.

1.2 Performance of online real-time traffic data

Numerous studies including government reports and academic studies have examined the quality of the 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 traffic variables, typically speed and travel time,[3–9]. Some also look at other aspects such as the coverage of the road network[3,10] or timeliness to recognize jams[11–13]. It is suggested that theoretically mean point speed from sensors would always be greater than mean link speeds from FCD[3] and this has bored out in some empirical observations[3,6,7]. Jurewicz et al. (2018)[3] found FCD speeds are on average 23% lower than mean loop point-speeds. Others, however, found FCD speeds higher than fixed point measurements[8,9]. Others, however, found poor correlations between private data and ground truth and concluded that private sector data is not suitable for real-time measurements as they tended to show less variability though they could still be suitable for long-term trend analysis[11]. Most of these validation studies, however, are limited in scope (such as few routes in a particular city, or cover part of a city) for a number of days and many focus on developed countries. None have compared across countries for a long duration of time.

1.3 Measures to characterize traffic and congestion

We use three variables to characterize the **traffic information available** for the different cities: Data availability, travel delays, and reliability of travel delays. *Data availability* measures the quality of the traffic information provided to users, measured based on the percent of real-time information that each city has during the studied period. *Travel delays* for a road measures the average excess travel time for a given road segments given in overhead travel time per kilometer of road. *Reliability of travel delays* is captured by the share of recurring congestion. The larger proportion of non-recurring congestion in total delays, the lower reliability of travel time estimates.

Due to the lack of flow or volume data, we quantify congestion as the *travel delay* defined by the average travel delay per kilometer in a city as

$$\frac{60}{\sum_j l_j}\sum_i l_i\left(\frac{1}{min(s_i,f_i)}-\frac{1}{f_i}\right) \qquad \text{[min/km]} \tag{1}$$

for more details, see Supplementary Information.

This paper is organized as follows. In Section 2, we present descriptions of the data and key summaries regarding traffic information, including data description (Section 2.1), and data availability (Section 2.2). In Section 3, we describe the methods of filling in missing data (Section 3.1). We quantify the other traffic information in including travel delays (Section 3.2). The results are summarized in Section 4 and Section 5 offers discussions and suggestions for future work.

2 Data exploration

2.1 Data collection and volume

We collect traffic and accident data from 17 cities approximately every 5 minutes from Jan 1 2017 to August 31 2017, and because of slight changes in timing of the sampling and varying latency in network delays, we group these samples into 15 minute time windows (96 windows per day, in total 23136 samples per road segment in each city). After collecting all samples into the corresponding time window of the day (between 1 and 96), we average the measured traffic speeds in each bin. We thus use the same number of time windows per day throughout the sample period to simplify processing and the comparison of speeds between different times.

Roads are geographically represented by road segments as a sequence of edges with WGS 84/GPS coordinates. The cities were somewhat arbitrarily chosen to represent several different countries, different types of urban

environments, and areas believed to be both either highly congested or with relatively low congestion. City bounding box coordinates and other characteristics including the length of the total road network, number of roads with measurements, and number of reported accidents are provided in Table 1 and the geographies are illustrated in the Supplementary Information. We also included two cities with surrounding smaller cities: Amsterdam and Johannesburg/Pretoria (see the S.I.).

According to HERE, the traffic data comes from "billions of GPS data points every day and leverages over 100 different incident sources to provide a robust foundation for our traffic services."[2] 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[14]. Traffic data is asynchronously updated on HERE traffic network links in three minute intervals. The data has a typical delay between 1.5-3 minutes in relation to the real world state[14].

The real-time traffic data was obtained by using network access to the HERE Application Programming Interface (API) and computer programming to request and download the data every 5 minutes. In the *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 (real-time data availability), the direction of traffic, free flow traffic speed, traffic speed limit, and a geographical description of the road segments as a set of WGS84 polylines. Similarly, through a separate *incident* API, we requested a list of reported active incidents on particular road segments in the same areas. Traffic incidents are classified according to type (including accident, congestion, construction, planned event, weather), status, criticality, as well as start and end times (end times can be planned, in the case of road constructions) of incidents.

Cityname	Bounding box	Total road (km)	Real-time share	# roads	# roads w/ accidents	# accidents
Barcelona	41.1957,1.5202;41.6745,2.5983	6953	0.25	9759	527	1448
Gothenburg	57.50792,11.69806;57.85863,12.24774	1491	0.27	2105	347	1041
Stockholm	59.1502,17.5857;59.4761,18.6637	2513	0.30	3437	579	1653
Detroit	42.1166,-83.6382;42.5885,-82.5602	9686	0.33	7326	532	2096
Chicago	41.5954,-88.2710;42.0712,-87.1930	10181	0.37	10839	1947	5287
Florence	43.6654,10.9753;43.8959,11.5144	1856	0.50	1133	37	519
St Peterburg	59.5962,29.2278;60.2364,31.3838	7394	0.50	9471	0	0
New York	40.4940,-74.4653;40.9778,-73.3873	14738	0.52	18201	1757	5214
Berlin	52.3228,12.8499;52.7113,13.9279	4835	0.55	4071	814	1789
Sao Paulo	-24.00703,-46.82534;-23.358,-46.3652	6365.86	0.62	16677	592	787
Moscow	55.3854,36.5872;56.1042,38.7433	12121	0.65	23529	0	0
London	51.2907,-0.6269;51.6883,0.4511	8092	0.67	5114	336	2548
Istanbul	40.7868,28.4642;41.2685,29.5422	8019	0.69	27195	8	911
Rio	-23.08319,-43.81347;-22.77014,-43.10792	3234	0.70	6849	670	1750
Johannesburg Pretoria	-26.5836,27.2076;-25.4358,29.3637	16917	0.71	14051	921	3637
Amsterdam	51.9781,4.0265;52.7579,6.1826	10061	0.71	6712	1627	7095
Cape Town	-34.2084,18.1247;-33.6786,19.2027	4333	0.72	3889	469	1512

Table 1 Key summary of data for selected cities.

2.2 Data availability

There is variation both between cities and over time in what share of the road segments there is available real-time information. As outlined in in Section 1.3, the overall data availability at a given time is measured by the share of road segments where there is reported real-time information. Figure 1 shows box plots of the distribution of data availability in the different cities observed in the full sample period. This establishes both that there are large differences in the overall share and coverage of individual road segments with real time data. The two largest city areas (Greater Amsterdam and the region of Johannesburg Pretoria) are associated with more outliers; This is consistent with covering many roads outside the city with less possibility of traffic (see S.I.).

[2] https://www.here.com/en/products-services/here-traffic-suite/here-traffic-overview

We also see that data availability varies over time in the cities: In Figure 2 this can be seen for four of the cities (For rest, see S.I.). Scattering the data availability against time of day shows that day times and what are typical peak hours are associated with a higher level of measurements. This makes intuitive sense given that the traffic speeds are reported to be significantly based on traffic flows. This suggests that the data availability could be a smaller issue for measuring traffic delays as more real-time measurements are available when congestion can be expected to be high. The Figure also shows that there exists distinct states or levels at which degree a city is measured. That the measurements level cluster like this suggests that traffic in cities could be classified into a smaller number of distinct states.

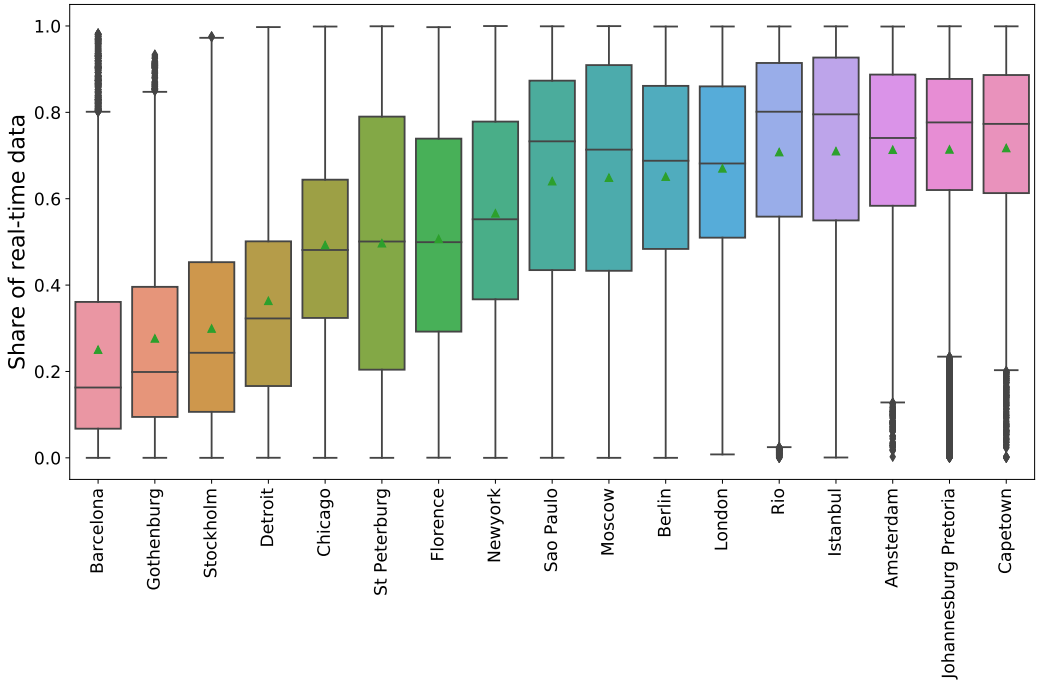

Figure 1. Data availability (% of real-time data) across the roads in different cities. Box plots show the 25th, 50th and 75th percentiles of the distribution, and triangles show mean values.

3 Data analysis

As all cities are characterized by some level of missing data, one could want to have a more complete view of traffic delays in cities. In the following, we proceed to evaluate imputation methods to fill in the missing data. Based on a more complete picture, we proceed to examine average traffic delays across cities, to which degree these delays are part of a recurring pattern, and whether the developments in these delays can be forecast on the short term.

3.1 Filling in missing data

For each 15 minute time window, if there was no real time information in any of the samples, we consider it to be a missing value (for technical details, see the S.I.). We are thus filling in a sparse matrix of type $(23136, n_i)$ where n_i is the number of road segments in city i. The cities vary not only in data availability but also several other characteristics, and it is not obvious whether some imputation method would work better relative to others across the cities. We choose to evaluate four different methods that represent several possible approaches. The methods that are evaluated are (i) A *mean*-based method that simply fills in missing values for a particular road segment with the observed mean speed value on the segment, (ii) A *correlations*-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, and (iii) A *k-nearest neighbors*-based method based on comparing the road segment to the k most similar

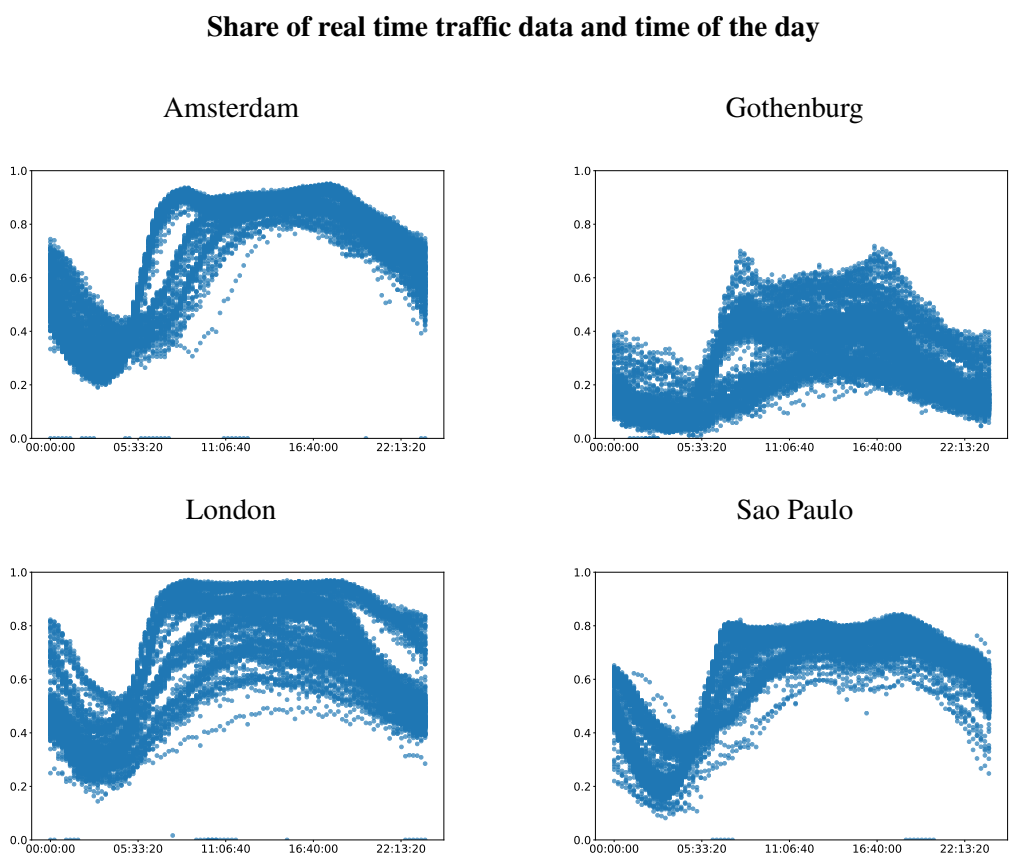

Figure 2. The share of road segments with real time traffic measurements varies with the time of the day: Examples from four cities over eight months. In general, day times are associated with a higher level of real time measurements, showing that larger traffic volume is significantly related to the number of real time measurements. Cities can be associated with several different levels of measurements, during different lengths of the day. The latter is possibly due to different peak periods in cities. Zeroes are outliers due to a small rate of measurement failures in this study.

City	mean	correlations	knn	knn window
Barcelona	7.63	7.46	4.79	4.95
Gothenburg	6.91	6.70	4.35	4.82
Stockholm	7.73	7.43	4.47	4.15
Detroit	5.26	5.11	3.20	3.26
Chicago	5.53	5.19	3.03	2.95
St Peterburg	7.84	7.33	4.74	4.52
Florence	7.08	6.53	4.23	4.12
New York	5.46	5.08	2.94	2.48
Sao Paulo	7.15	6.50	4.11	4.01
Moscow	6.59	6.23	4.21	4.06
Berlin	7.58	7.05	4.58	4.24
London	7.42	6.12	3.57	3.23
Rio	8.07	7.20	4.48	4.42
Istanbul	7.83	7.16	4.49	4.50
Amsterdam	8.22	7.72	4.73	4.54
Johannesburg Pretoria	8.16	7.59	4.91	4.88
Cape Town	8.86	7.84	4.89	4.61

Table 2 Evaluating different imputation methods on traffic speeds with the metric root mean square error (rmse) using repeated 10-fold cross-validation. Imputation strategy 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.

other road segments in the data, and (iv) A *sliding window k-nearest neighbors*-based method, using a time window of one month, which could possibly have the advantage to take difference between months of the year into account, while restricting the available data. The evaluation of each method was made with respect to the root mean squared error (rmse) and 10-fold cross validation. The results of these four methods are summarized in Table 2. The cities are ordered by an increasing mean availability of real time data and the numbers illustrate that there are several consistent results for these imputation methods.

First, consistently 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. Third, this improvement is consistently stronger than from historical correlations, which suggests that traffic speed relationships have a significant non-linear component. Fourth, the consistent (but relatively smaller) improvement from the time-dependent knn method could be consistent with the existence that the cities can be in different states during different times of the year: We see a consistent improvement when making predictions based on more recent observations. Fifth, and somewhat surprising, there is no obvious improvement by increasing data availability and the relative results of the imputation. It is currently unknown whether this depends on that the ground truth in the cities differ, or whether the road segment only partially cover the underlying road networks in different cities.

Taken together, the knn methods work consistently best, despite the wide differences in the dimension and road network characteristics of the cities. Further work would be needed to expl

3.2 Travel delays

We now look at the overall picture of the imputed data in the different cities. Figures 3 and 4 illustrate, for two of the cities, the travel delay (minutes of delay per kilometer, on top of free flow speed, in the covered road network). This is shown for the first four weeks, as well as the full time period of eight months. We consider the first four regular

weeks from the first Monday of the year, and a total length of 28 days. We disregard the difference in magnitude of the total delay between cities, which can depend on what types of roads are covered in different cities: Our questions about whether the pattern is regular and can be forecast concern patterns of change over time. In the Figures, we can directly see at least four common properties of the time series (the last two properties are visible only in the full 8 months time series). First, there is a clear difference between weekdays versus weekends, and we typically see two peaks during day times. Second, the data shows complex seasonality: Recurring patterns are on several time scales (weekly, daily, and sub-daily). Third, the data has either a trend or different regimes: In both cities, dates starting around early July are related to common vacation periods in the cities, and a clear drop of mean delays is found in the series. During vacation periods, not only is the mean lower, but variance is lower as well. Fourth, outliers exist and some extremes are shown as clear drops in the data. In the case of sharp drops to zero, these correspond to sample errors (the availability of the HERE in the online traffic system appears to have been 100% during the period, but our network connectivity used for sampling was briefly down two times during the period). This illustrates common characteristics in the cities, but also the difficulty of frequently sampling a data source for a longer time and the need to track the source of outliers whenever possible (there are also outliers in terms or irregularly high delays, that need to be attributed to other sources).

These four findings lead to several questions about the different cities. First, to which degree are traffic delays regularly recurring versus irregular? Second, do these patterns of complex seasonality have the same underlying seasonality or not? Third, can they also be regularly ¡forecast on the short term?

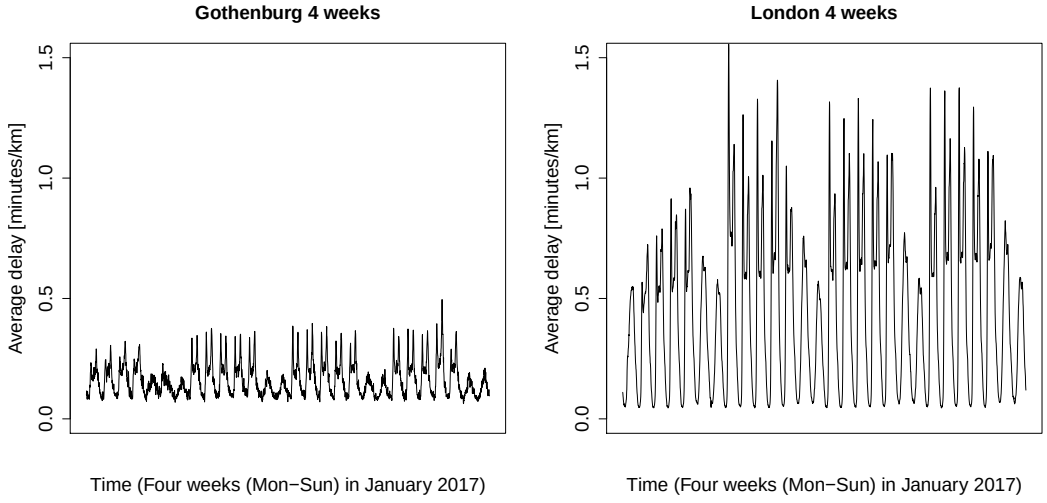

Figure 3. Time series for mean delays (minutes of extra travel time per km compared to free flow speed, across the city) in the first four weeks (starting the first Monday).

4 Results

We now summarize the key results from the methods and models that we presented above.

4.1 Average travel time delays

Figure 5 shows the average unit delay (minutes/km) by hour of the day, weekdays vs. weekend for four example cities, when delays have been quantified as in Eq. 1. Almost all cities have lower congestion on the weekends compared with weekdays. Most cities have both the morning peaks and afternoon peaks on the weekdays, and some only have the afternoon peaks on the weekends. In general, the city average afternoon peaks are higher than the morning peaks, except for in London. As a reference, a delay of 1 minutes per km is the same as traveling at 30 km/hour on a road that has a free flow speed of 60 km/hour, whereas traveling at 20 km/hr on a road with a free flow speed of 40 km/hr would have a delay of 1.5 minutes/km.

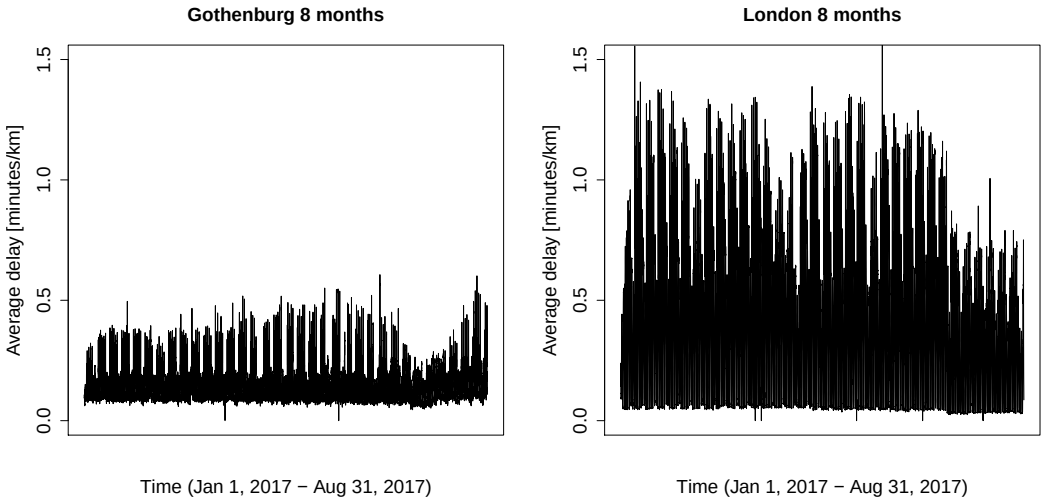

Figure 4. Time series for mean delays (minutes of extra travel time per km compared to free flow speed, across the city) in the full period.

4.2 Variability of travel delays

Travel delays not only are highest during daytime and peak hours, but the variability of delays also increases during these times (Figure 6). As shown earlier in Figure **??**, the number of accidents increases during day time and peak hours, i.e., with higher traffic volumes. On top of that, accidents increase both the level and the variability of travel delays, even more so during daytime and peak hours. As we discussed earlier in Section 2.3, one limitation is that the number of accidents are most likely under-reported in all cities as we compared the public records with our data. It is therefore possible that the variability of congestion associated with non-accidents is something that is smaller than shown above.

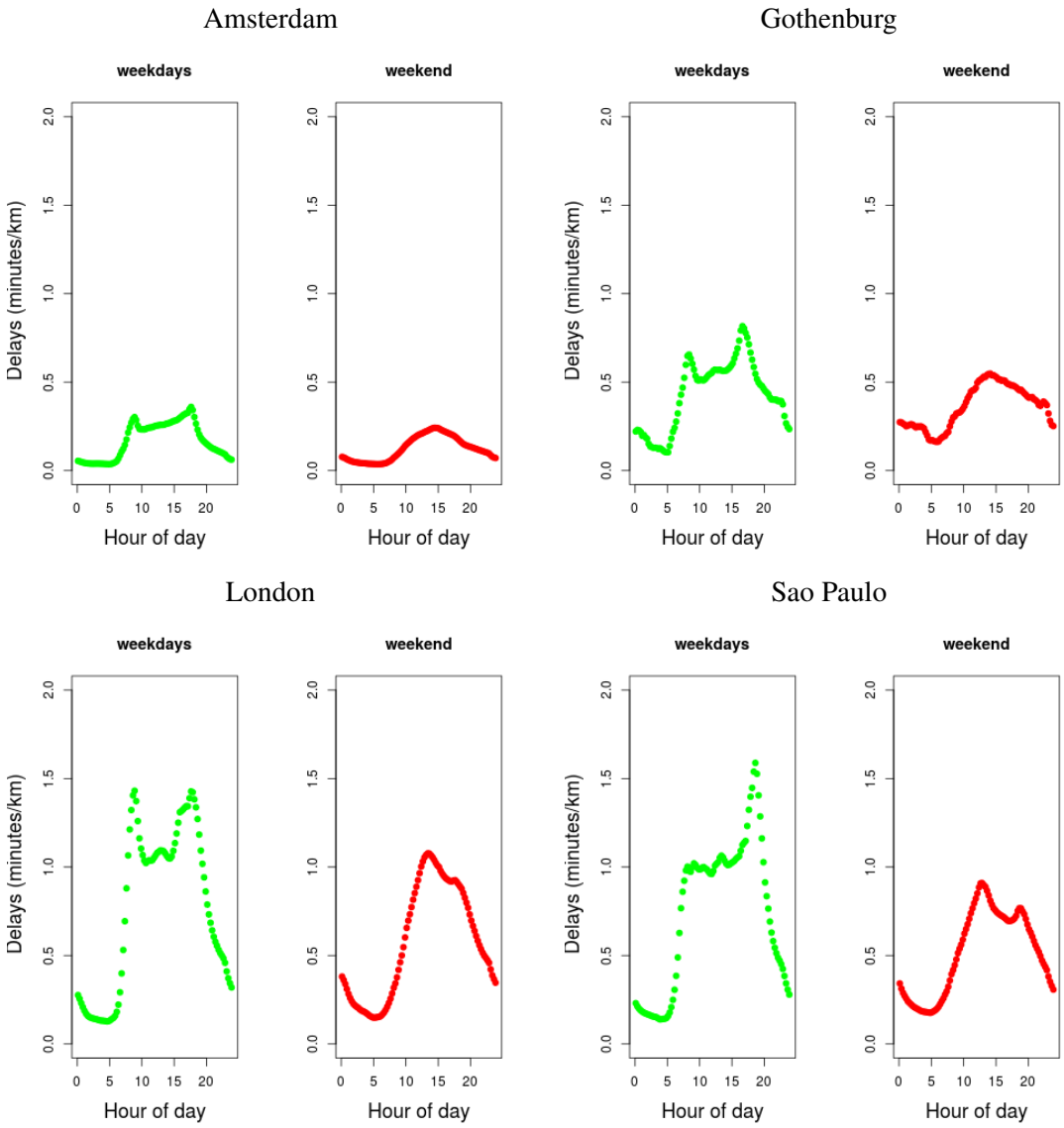

Figure 5. City average delay (minutes/km) by hour of the day, weekdays vs. weekend.

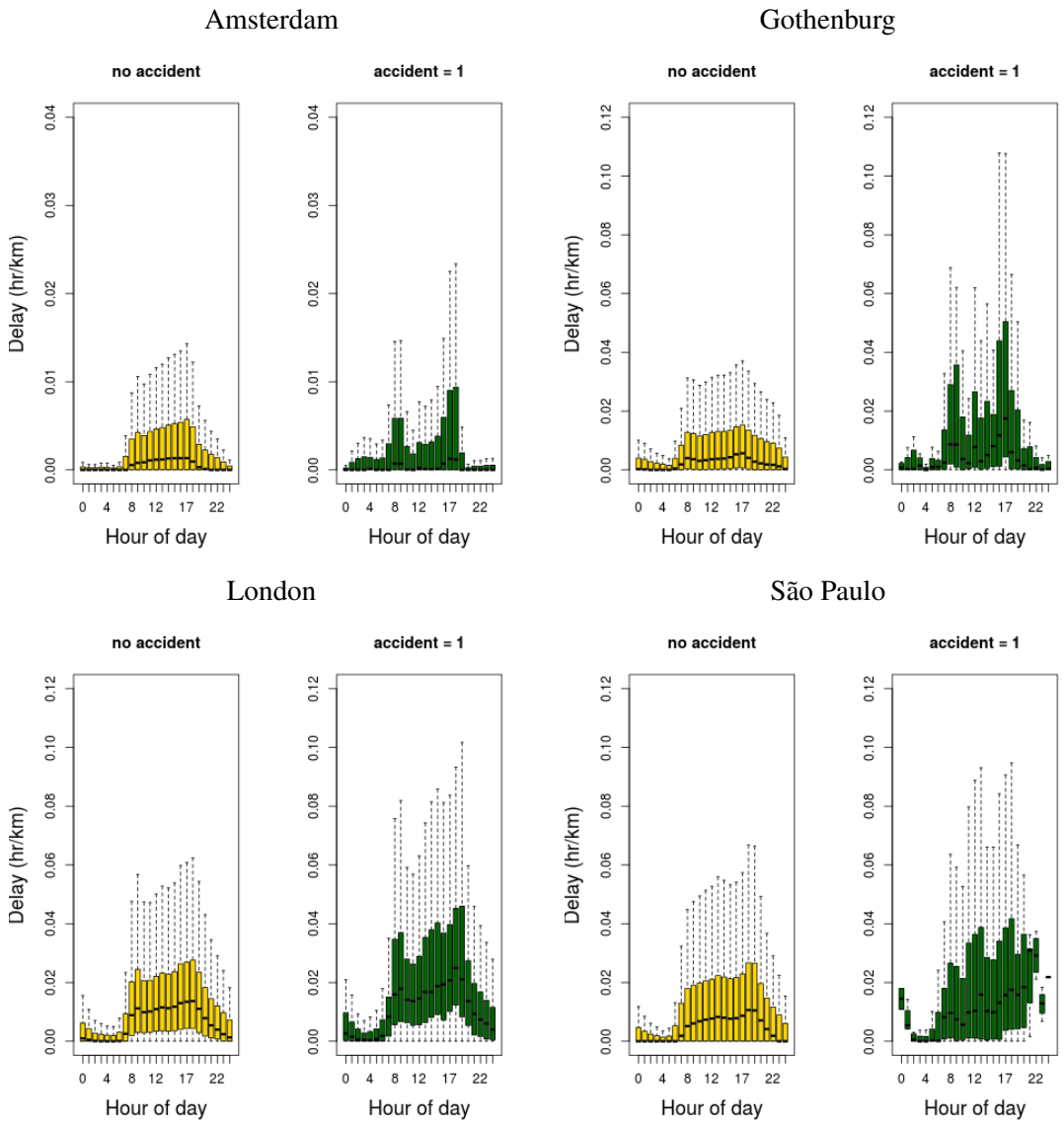

Figure 6. Boxplots of delays (hr/km) by hour of day, with and without recorded accidents for the studied cities. Note that the number of accidents are known to be under-reported.

5 Conclusion and Discussion

We have examined eight months of traffic data from 17 large city regions as covered in one of the new large-scale online platforms available to travelers, consumers, and policy-makers interested in city traffic around the world. We describe our findings in several areas where cities may vary and examine the data coverage. Despite varying characteristics of the cities in the data such as different characteristic road segment length, shares of covered accidents and of real time measurements, and varying levels of recurring traffic delays, other common characteristics emerge out of our observations and results. These are with respect to expected traffic patterns with traffic peaks during different times of the day, what methods that are best in filling in missing data, and the possibility to forecast travel delays on the short term. That these show consistent levels of improvement reveals some common patterns that generalize, and show a consistent approach to get a better and more complete view of the city data. This could not have been obviously expected, as the sparsity of the data and the differences such as the different dimension of the data arising from the varying number of roads makes the studied city regions rather different.

These findings raise a number of research questions. Future work can quite likely improve on some of these results by better methods and adding more data such as as weather and information about incidents to the forecasting. One possibility is to further zoom in on particular properties and regions 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 the ground truth: Using publicly available data sources for validation is also an important and needed direction of future research.

References

1. Ben-Elia, E. & Avineri, E. Response to travel information: A behavioural review. *Transport Reviews* **35**, 352–377 (2015).

2. Chorus, C., Molin, E. & Van Wee, B. Use and effects of advanced traveller information services (ATIS): A review of the literature. *Transport Reviews* **26**, 127–149 (2006).

3. Jurewicz, C. *et al.* Use of connected vehicle data for speed management in road safety. In *28th ARRB International Conference – Next Generation Connectivity*.

4. de Boer, G. & Krootjes, P. The quality of floating car data benchmarked: An alternative to roadside equipment? In *19th ITS World Congress*.

5. Clergue, L. & Buttignol, V. Using gps data in favour of traffic knowledge. In *Transport Research Arena*.

6. Clergue, L. & Buttignol, V. Probe data and its application in traffic studies. In *2015 IPWEA/IFME Conference*.

7. Hrubeš, P. & Blümelová, J. Comparative analysis for floating car and loop detectors data. In *22nd ITS World Congress*.

8. Diependaele, K., Riguelle, F. & Temmerman, P. Speed behavior indicators based on floating car data: Results of a pilot study in belgium. *Transportation Research Procedia* **14**, 2074–2082 (2015).

9. Ambros, J. *et al.* Improving the self-explaining performance of czech national roads. *Transportation Research Record* **2635**, 62–70 (2017).

10. Aarts, L. T., Bijleveld, F. D. & Stipdonk, H. L. 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 (2015). URL https://www.swov.nl/sites/default/files/publicaties/rapport/r-2015-03.pdf.

11. Hu, J., Fontaine, M. D. & Ma, J. Quality of private sector travel-time data on arterials. **142**, 04016010 (2016). URL https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29TE.1943-5436.0000815. DOI doi:10.1061/(ASCE)TE.1943-5436.0000815.

12. Kessler, L., Huber, G., Kesting, A. & Bogenberger, K. Comparing speed data from stationary detectors against floating-car data. *IFAC-PapersOnLine* **51**, 299–304 (2018). URL http://www.sciencedirect.com/science/article/pii/S2405896318307729. DOI https://doi.org/10.1016/j.ifacol.2018.07.049.

13. Wang, Y., Araghi, B. N., Malinovskiy, Y., Corey, J. & Cheng, T. Error assessment for emerging traffic data collection devices. Tech. Rep., Washington State Department of Transportation.

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

Acknowledgement

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.