MLVis: Machine Learning Methods in Visualisation for Big Data (2019)
D. Archambault, I. Nabney and J. Peltonen (Editors)

Visual Analysis of Multivariate Urban Traffic Data Resorting to Local Principal Curves

Carla Silva[1,2], Pedro M. d'Orey[1,2] and Ana Aguiar[1,2]

[1]Instituto de Telecomunicações, Porto, Portugal
[2]Universidade do Porto, Portugal

Abstract

Traffic congestion causes major economic, environmental and social problems in modern cities. We present an interactive visualization tool to assist domain experts on the identification and analysis of traffic patterns at a city scale making use of multivariate empirical urban data and fundamental diagrams. The proposed method combines visualization techniques with an improved local principle curves method to model traffic dynamics and facilitate comparison of traffic patterns - resorting to the fitted curve with a confidence interval - between different road segments and for different external conditions. We demonstrate the proposed technique in an illustrative real-world case study in the city of Porto, Portugal.

CCS Concepts

• ***Human-centered computing*** → *Visual analytics; Empirical studies in visualization;*
• ***Computing methodologies*** → *Machine learning approaches; Modeling and simulation; Shape modeling;*

1. Introduction

Traffic congestion causes major economic, environmental and social problems in modern cities. Urban traffic is impaired by a number of spatio-temporal varying phenomena, namely travel demand (e.g. in peak hours) [OÇS*18], meteorological conditions, special events (e.g. soccer match) [WWL16], among other factors. Furthermore, several studies [LJZ17] [LLL*16] have demonstrated the direct and indirect spatial interactions between adjacent road segments leading, for instance, to the propagation of traffic jams.

Traffic can be monitored resorting to static (e.g. inductive loops) and/or mobile sensors (e.g. taxis [LLL*16], *crowdsensing* [GdA17]). Single inductive loops measure traffic volume, i.e. the number of vehicles transversing a given road segment r in time interval t. Mobile sensors record trajectory data to infer vehicular speed on different road segments. The current traffic state cannot be accurately described resorting to a single traffic variable. Fundamental diagrams (FDs) [AM16] are commonly used by domain experts to describe the traffic state of road segments. FDs describe pairwise relations between speed, volume or density variables. In this paper, we infer link-based fundamental diagrams by fusing data from multiple sources, namely inductive loop and taxi trajectory data. We can achieve a better estimation of the current traffic state on urban areas making the best use of both data sources.

Recently, much work has been devoted to understanding the main causes of traffic congestion [WWL16] and how congestion propagates in urban areas [LJZ17]. Visualization has also been used as a tool to better understand this phenomena [CGW15] [ZWC*16]. For instance, Wang et al. [WLY*13] [WYL*14] proposed interactive systems for visual analysis of traffic congestion (propagation) based on trajectory data or static *transportation cells*. Most works on visual analysis studied a single traffic variable (e.g. speed) and often use table-like pixel based visualization to reveal traffic congestion patterns, which techniques are not suitable for spatio-temporal sparse trajectory data.

We propose a novel visual analysis system to better understand traffic congestion in urban areas and the impact of externalities (e.g. weather). To tackle the shortcomings of the current state of the art, we resort to **multi-source fundamental diagrams** to model the relation between pairs of traffic variables. However, modeling the relations between traffic variables is specially challenging in **urban areas** due to several externalities (e.g. weather, parking, special events), road network design and operation (e.g. traffic lights), data sparsity, among others, that leads to noisy data clouds. To account for the aforementioned uncertainties and dynamics of urban scenarios, we extend the **Local Principles Curves (LPC)** method [OE11] to infer traffic patterns in urban areas, which has shown promising results in less dynamic scenarios (i.e. freeways [ED11]).

The proposed visual analysis system allows addressing challenging domain questions such as: 1) identification of evolving traffic patterns in urban areas, 2) detection of correlations between different road segments and 3) quantification of the impact of externalities on traffic patterns. To achieve these goals, we provide global and cell-based interactive views of the traffic state in urban areas with filtering mechanisms to assist on visual exploration.

DOI: 10.2312/mlvis.20191159

2. Related Work

Much work has been devoted to understanding the main causes of traffic congestion [WWL16] and how congestion propagates in urban areas [LJZ17] using machine learning methods. Silva et al. [SdA18] resorted to probabilistic graphical modeling to understand the associations between congestion and weather conditions.

Visualization has been also used as a tool to better understand the complex traffic phenomena. We refer the reader to [CGW15] [ZWC*16] for a complete review of visualization techniques for urban and traffic data. Cruz et al. [CM16] use the figurative metaphor of pulsing blood vessels for visualizing traffic dynamics. Wang et al. [WLY*13] proposed an interactive system for visual analysis of traffic congestion based on trajectory data and the construction of traffic jam propagation graphs. Wang et al [WYL*14] presented a traffic visual analysis system based on static *transportation cells* that accurately record traffic volume and speed data, and study the correlations between cell patterns and route patterns. In this work, we combine machine learning methods with visualization tools to assist the domain expert (e.g. urban planner, traffic engineer) on detecting and comparing traffic patterns.

The estimation of fundamental diagrams has historically (i) focused mostly on highway or freeway scenarios (e.g. [QWZ15]), and (ii) made use of a single data source. Single-source FD estimation using trajectory data is challenging due to the dynamic human mobility patterns that traduces into variable probe penetration levels, spatio-temporal coverage, among others as show in [DRG16]. More recently, few works [AM16] [DRG16] estimated FDs in urban area. The number of studies estimating empirical fundamental diagrams in urban areas resorting to both data sources is very reduced. Geroliminis et al. [GD08] have demonstrated the existence of well-defined *macroscopic* fundamental diagrams in urban areas. [GS11b] has shown that the spatial variability of vehicle density can affect the shape, the scatter and the existence of a well-defined macroscopic fundamental diagram.

Our work distinguishes from the current state-of-the-art in 1) use of multi-source traffic and environmental data, 2) enhanced fundamental diagrams as a visualization tool to infer traffic patterns and 3) a machine learning method to model the relationship between traffic variables and scatter around defined local principle curves.

3. Visual Analysis of Multivariate Urban Traffic Data

3.1. Input Data

We consider that there exist static (e.g. inductive loop) and mobile (e.g. taxi) sensors measuring traffic variables, namely vehicular speed (v) and traffic volume (q). In addition there might exist additional sensors measuring urban data, such as meteorological conditions or pollutant emissions. Specifically, in this study, we use trajectory data collected by a fleet of taxis in the city of Porto, Portugal. Trajectory data refers to a sequence of ordered, timestamped geo-spatial position estimates obtained using GPS: $T = \{(t, \varphi, \lambda)\}$, where t is the timestamp, φ latitude and λ longitude. Road network data is used to match trajectory data to a sequence of road segments to estimate traffic variables. A road network is represented as a directed graph $G = (V, E)$, where V is the set of vertices (i.e. intersections) and E is the set of edges (i.e. roads). Traffic volume is acquired by inductive loops installed in key locations in the city.

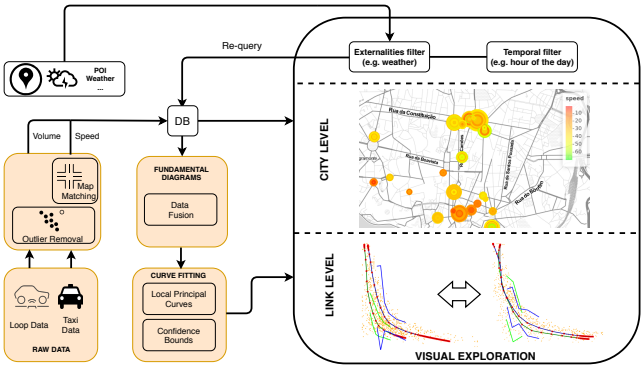


Figure 1: *Our system pipeline: raw data → preprocessing → modeling of fundamental diagrams using Local Principal Curves → visual exploration.*

3.2. Methodology

The main goal of this work is to provide an informative and intuitive tool for domain experts to better understand traffic congestion in urban areas through visualization techniques augmented by machine learning methods. We target the identification of traffic patterns in urban areas and improved understanding of the impact of externalities and spatial interrelations (i.e. adjacent roads) on these traffic patterns. Fig. 1. depicts the proposed methodology that consists of three main modules: (1) data pre-processing, (2) data fusion to infer and model fundamental diagrams through Local Principle Curves (LPC) and (3) multi-level and filterable visual exploration.

3.3. M1: Preprocessing

In this stage, we clean and calculate traffic metrics from urban data. Speed estimation is performed resorting to taxi trajectory data conducting the following steps: (1) sensor-related outlier removal (e.g. arising from GPS multipath errors), (2) map matching of trajectory data to a sequence of road segments of the network graph, (3) smooth speed time series by applying an Hampel filter to remove additional data outliers and (4) speed estimation in different road segments and time intervals (v_r^t) by aggregating the corresponding sub-trajectories.

Volume estimation is performed making use of data collected by inductive loops. Data collected by these static sensors is often corrupted and noisy (e.g. due to sensor malfunction, parked vehicle). To improve data quality, we apply two outlier detection and filtering mechanisms in sequence, namely *Hampel filter* for removing local outliers and *Tukey's filter* to filter extreme values.

3.4. M2: Modeling of the Fundamental Diagrams (FD)

FD estimation: Road traffic is characterized by a state defined by the flow rate (q), mean vehicle speed (v) and density (k). The traffic state can be described graphically by three fundamental diagrams of traffic flow (i.e. $q-v$, $q-k$ and $v-k$ diagrams) inferred by fusing speed and traffic volume data from multiple sources. Trajectory data collected from mobile probes allows accurately determining the mean vehicle speed in road segment r given sufficient sampling rate. On the other hand, single inductive loops provide accurate traffic volume data but these are sparsely deployed in the

city. Merging pre-processed data collected by different sensors allows improving the accuracy of the fundamental diagrams estimation by making the best use of both datasets. Since no occupancy data is available, we infer vehicle density through the following fundamental traffic theory (approximate) relation $k = \frac{q}{v}$. This step generates a 2D point cloud for each fundamental diagram type.

FD modeling using LPC: we model the fundamental diagrams describing the traffic state using the LPC method. Principal Curves are smooth curves passing through the middle of the distribution of a data cloud [ED11]. The LPC method is described in detail in Algorithm 1. After variable normalization, this algorithm *iteratively* calculates local centers of mass and a first local principal component updating x until the convergence criteria is met (i.e., μ^x remains approximately constant). The calculation of local center of mass and the principal component is weighted by w_i^x, where H is a bandwidth matrix and K_H is a d-dimensional kernel function,

$$w_i^x = K_H \frac{(x_i - x)}{\sum_{i=1}^{n}(x_i - x)} \quad (1)$$

The resulting principal curve is composed by the series of local centers of mass μ^x. The LPC input parameters with critical importance on the system performance are 1) starting point x_0, 2) step length t_0 and 3) bandwidth h. Given the dynamicity of urban traffic flow the input parameters must be tuned for the different road segments, contrary to what is mentioned in [ED11] for freeways. The parameter x_0 is selected automatically based on a local density estimate. The parameters t_0 and h are selected according to an automatic method proposed in [Ein11]. Angle penalization α is not considered because the data clouds do not form crossings locally.

To understand varying traffic phenomena (e.g. traffic hysteresis [GS11a]), we extend the LPC algorithm to also model data dispersion around the defined Principal Curve. The proposed method described in Algorithm 2 is composed of two main parts: (1) determination of the closest center of mass for each data cloud point using an euclidean distance metric (steps 2-12) and (2) computation of confidence bounds of the LPC curve (steps 13-16) resorting to a variability measure (e.g. n^{th} quantile) based on the set of euclidean distances between a center of mass μ_x and all its associated data points. Fig. 2 shows an illustrative example of the calculated LPC and the association of data points to the closest center of mass.

Algorithm 1 Modified Local principal Curves (LPC)

Input: x_0, t_0, h, scaled = True
Output: fitted curve within a confidence interval

1: **procedure** LPC(x_n) ▷ data cloud
2: $x \leftarrow x_0, x_0 \in \mathbf{R}^2$ and $t_0 > 0$ ▷ Initialization
3: **repeat**
4: $\mu^x \leftarrow \sum_{i=1}^{n} w_i^x x_i$ ▷ Calculate local centre of mass
5: $\sum^x \leftarrow (\sigma_{jk}^x) \in \mathbf{R}^{2\times 2}$ ▷ Calculate covariance matrix
6: via $\sigma_{jk}^x \leftarrow \sum_{i=1}^{n} w_i^x (x_{ij} - \mu_j^x)(x_{ik} - \mu_k^x)$
7: $\gamma^x \leftarrow ev(\sum^x)$ ▷ Calculate 1^{st} eigenvector of $\sum^x$
8: $x \leftarrow \mu^x + t_0\gamma^x$ ▷ New center of mass
9: **until** μ^x remains constant ▷ End of data cloud
10: BOUNDS(μ^x, x_n) ▷ Confidence bounds calculation
11: **end procedure**

Algorithm 2 Confidence Bounds Calculation

1: **procedure** BOUNDS(μ^x, x_n) ▷ Centers of mass and points
2: $v \leftarrow \{\}$
3: **for** `i = 1 to` $|x_n|$ **do** ▷ Assign each x to closest μ^x
4: $distance \leftarrow \infty$
5: **for** `j = 1 to` $|\mu^x|$ **do**
6: $tmp \leftarrow dist(\mu^x, x_n, \text{"euclidean"})$
7: **if** tmp < distance **then**
8: $distance \leftarrow tmp$
9: **end if**
10: **end for**
11: v.append($distance$, x_n, μ^x)
12: **end for**
13: **for** z = 1 to $|\mu^x|$ **do**
14: $d \leftarrow quantile(distance, n^{th})$
15: add d to y coordinate of μ^x
16: **end for**
17: **return** x_n, μ^x ▷ Coordinates within a confidence interval
18: **end procedure**

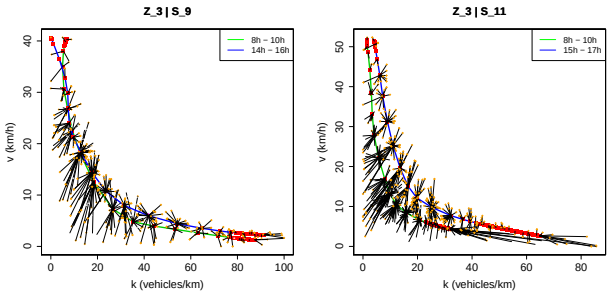


Figure 2: *Confidence Bounds Calculation (e.g. $v - q$ FD).*

3.5. M3: Visual Exploration

We consider a visual exploration stage with three main steps:

- **city-level exploration (Fig. 3)**: presents a high level view of the current traffic situation in the city in terms of vehicular speed and flow in a given time period. This stage allows the user to identify road segments or city zones for further exploration.
- **road-level exploration (e.g. Fig 5a)**: presents the traffic state at each individual road resorting to three fundamental diagrams ($q - v$, $q - k$ and $v - k$). This view allows the user to detect abnormal traffic patterns and to assess - through filtering - how externalities (e.g. weather) and temporal aspects (e.g. time of the day) impact the traffic patterns.
- **zone-level exploration (e.g. Fig 5)**: focuses on the comparison of the traffic patterns in adjacent or close by road segments. The main aim is to understand if these traffic patterns co-evolve or not under certain conditions through the application of the temporal and externalities filters.

We resort to the following visualization techniques:

- **city-level exploration**: provides a city-level **map view** of the traffic state. For each road, we represent the traffic state using a colored and variable radius circle in which the traffic speed and volume is encoded by color (red and green color represents low

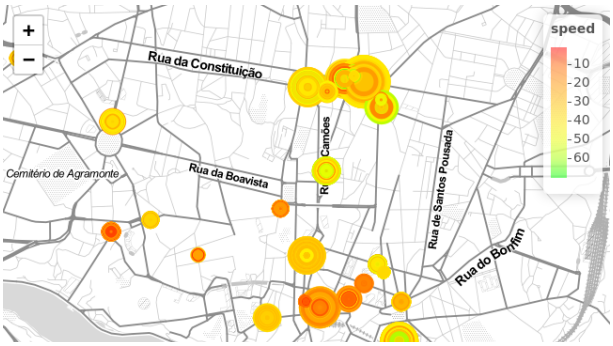


Figure 3: *Global Visual Exploration View: traffic speed and volume represented by circle color and radius, respectively.*

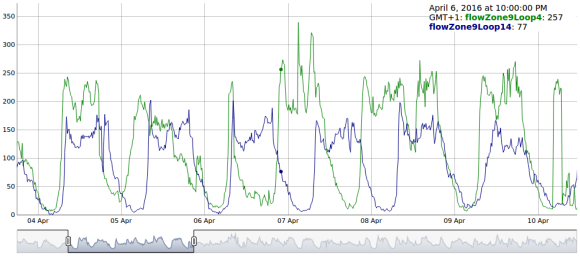


Figure 4: *Temporal Filtering (e.g. Traffic Volume Data).*

and high speed, respectively) and circle radius (larger radius for higher traffic volume), respectively. The map view is updated by modifying the temporal and externalities (e.g. weather) filter.

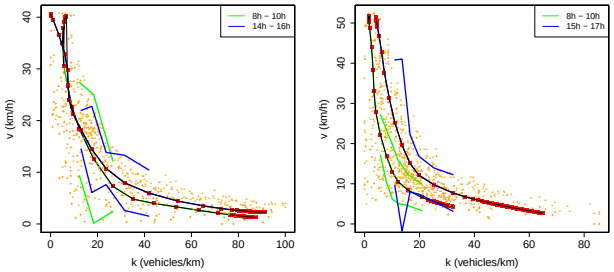
- **road and zone-level**: **fundamental diagrams** represent the traffic behaviour in individual road segments. Each point in the data cloud represents the observed traffic state in terms of v, q and k pairs for a given time interval (in this paper we consider a 15-min time interval). The LPC approximates the data cloud distribution and dispersion (through the confidence bound) to facilitate comparison of: 1) for a given road segment (e.g. to compare peak and non-peak hours - see 2 curves in Fig. 5b) or 2) between different road links (e.g. compare curves in Fig. 5a and Fig. 5b for the same time interval) using the aforementioned filters.

When the user applies filtering, the processing pipeline is recomputed, and the map views and the fundamental diagrams are updated. We consider the following filters for visual exploration:

- **temporal:** to assist the user in defining critical temporal periods, namely 1) filtering by the *hour of the day* (e.g. 8-10 h to represent peak traffic), 2) *day of the week* (e.g. weekdays vs weekend) to isolate traffic patterns, 3) *temporal window* to study or detect, for instance, special events, among many other possibilities.
- **externalities:** to assess the impact of externalities (e.g. weather) on the traffic patterns and existence of local phenomena. For instance, the user could compare conventional traffic patterns with the ones from extreme weather events (e.g. heavy snow).

4. Case Study

In this section, we present a simple illustrative use case demonstrating the proposed visual analysis system. Assume that the user is interested comparing the traffic patterns between peak and non-peak periods. First, the user could make use of the map view to



(a) **(b)**

Figure 5: *Road and Zone-level Visual Exploration (e.g. $v-k$ diagrams inferred in the city of Porto, Portugal)*

select two road segments of interest according to an expert-defined criteria. Afterwards, the expert would resort to time series data (e.g. similar to Fig 4) to define the peak (e.g. 8-10 h) and non-peak hours (e.g. 14-16h) for applying the temporal filter (type: *hour of the day*). Following, the visual exploration system would be triggered to update the fundamental diagram views for the selected road segments.

Fig. 5 depicts the the $k-v$ fundamental diagrams for two road segments in the city of Porto, Portugal, that were approximated by LPC and the corresponding confidence bounds. The $k-v$ diagram shows how sharply the vehicular speed (v) decreases for increasing vehicle density (k). Typically, the speed reaches the lowest values when the density equals the jam density (i.e.when a large number of vehicles are very close and unable to move or moving very slowly). This diagram is particular useful to translate the traffic condition of a segment. The shape of the FDs depends on network topology and control parameters (e.g. traffic light settings).

Analyzing a given road segment (e.g. Fig 5a) we observe that there exist similarities between the LPC curves but the decay rate of LPC and the data dispersion is considerably higher for the morning peak period. This results is expected given the more complex traffic dynamics during peak hours. Comparing both road segments (Fig 5a vs Fig 5b) we clearly see that the traffic patterns of one road segment is clearly more impacted during peak hours. A domain expert or further visual exploration (e.g. applying different data filtering) could provide insights for this discrepancy.

5. Conclusions

We have presented an interactive visualization tool to analyze traffic patterns at a city scale resorting to multivariate urban (traffic) data. The proposed method combined visualization techniques with the local principle curves method to facilitate visual exploration and comparison of traffic data patterns. Filtering mechanism support the discovery of relations between traffic variables and external factors (e.g., weather).

Acknowledgements

This work is a result of the projects *MobiWise* (POCI-01-0145-FEDER-016426), funded by the European Regional Development Fund (FEDER), through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) and by National Funds (OE), through Fundação para a Ciência e Tecnologia, I.P., *S2MovingCity* (CMUP-ERI/TIC/0010/2014) and UID/EEA/50008/2019 funded by the applicable financial framework (FCT/MCTES) (PIDDAC).

References

[AM16] AMBÜHL L., MENENDEZ M.: Data fusion algorithm for macroscopic fundamental diagram estimation. *Transportation Research Part C: Emerging Technologies 71* (2016), 184 – 197. 1, 2

[CGW15] CHEN W., GUO F., WANG F.: A survey of traffic data visualization. *IEEE Transactions on Intelligent Transportation Systems 16*, 6 (Dec 2015), 2970–2984. 1, 2

[CM16] CRUZ P., MACHADO P.: Pulsing blood vessels: A figurative approach to traffic visualization. *IEEE Computer Graphics and Applications 36*, 2 (Mar 2016), 16–21. 2

[DRG16] DU J., RAKHA H., GAYAH V.: Deriving macroscopic fundamental diagrams from probe data: Issues and proposed solutions. *Transportation Research Part C: Emerging Technologies 66* (2016), 136 – 149. 2

[ED11] EINBECK J., DWYER J.: Using principal curves to analyse traffic patterns on freeways. *Transportmetrica 7*, 3 (2011), 229–246. 1, 3

[Ein11] EINBECK J.: Bandwidth selection for mean-shift based unsupervised learning techniques: a unified approach via self-coverage. *Journal of pattern recognition research. 6*, 2 (2011), 175–192. 3

[GD08] GEROLIMINIS N., DAGANZO C. F.: Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B 42*, 9 (2008), 759 – 770. 2

[GdA17] GIL D. S., D'OREY P. M., AGUIAR A.: On the challenges of mobile crowdsensing for traffic estimation. In *ACM Conference on Embedded Network Sensor Systems (SenSys)* (New York, NY, USA, 2017), ACM, pp. 52:1–52:2. 1

[GS11a] GEROLIMINIS N., SUN J.: Hysteresis phenomena of a macroscopic fundamental diagram in freeway networks. *Transportation Research Part A: Policy and Practice 45*, 9 (2011), 966 – 979. 3

[GS11b] GEROLIMINIS N., SUN J.: Properties of a well-defined macroscopic fundamental diagram for urban traffic. *Transportation Research Part B 45*, 3 (2011), 605 – 617. 2

[LJZ17] LIANG Y., JIANG Z., ZHENG Y.: Inferring traffic cascading patterns. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2017), ACM, pp. 2:1–2:10. 1, 2

[LLL*16] LIU Z., LI Z., LI M., XING W., LU D.: Mining road network correlation for traffic estimation via compressive sensing. *IEEE Transactions on Intelligent Transportation Systems 17*, 7 (July 2016), 1880–1893. 1

[OÇS*18] OLMOS L. E., ÇOLAK S., SHAFIEI S., SABERI M., GONZÁLEZ M. C.: Macroscopic dynamics and the collapse of urban traffic. *Proceedings of the National Academy of Sciences 115*, 50 (2018), 12654–12661. 1

[OE11] OZERTEM U., ERDOGMUS D.: Locally defined principal curves and surfaces. *J. Mach. Learn. Res. 12* (July 2011), 1249–1286. 1

[QWZ15] QU X., WANG S., ZHANG J.: On the fundamental diagram for freeway traffic: A novel calibration approach for single-regime models. *Transportation Research Part B: Methodological 73* (2015), 91 – 102. 2

[SdA18] SILVA C., D'OREY P. M., AGUIAR A.: Interpreting traffic congestion using fundamental diagrams and probabilistic graphical modeling. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (Nov 2018), pp. 580–587. 2

[WLY*13] WANG Z., LU M., YUAN X., ZHANG J., WETERING H. V. D.: Visual traffic jam analysis based on trajectory data. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec. 2013), 2159–2168. 1, 2

[WWL16] WU F., WANG H., LI Z.: Interpreting traffic dynamics using ubiquitous urban data. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (New York, NY, USA, 2016), ACM, pp. 69:1–69:4. 1, 2

[WYL*14] WANG Z., YE T., LU M., YUAN X., QU H., YUAN J., WU Q.: Visual exploration of sparse traffic trajectory data. *Visualization and Computer Graphics, IEEE Transactions on 20* (12 2014), 1813–1822. 1, 2

[ZWC*16] ZHENG Y., WU W., CHEN Y., QU H., NI L. M.: Visual analytics in urban computing: An overview. *IEEE Transactions on Big Data 2*, 3 (Sep. 2016), 276–296. 1, 2