applied system innovation

MDPI

Article

Big Data for Traffic Estimation and Prediction: A Survey of Data and Tools

Weiwei Jiang [1,*] and Jiayun Luo [2]

[1] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore; luoj0028@e.ntu.edu.sg
* Correspondence: jiangweiwei@mail.tsinghua.edu.cn

Abstract: Big data have been used widely in many areas, including the transportation industry. Using various data sources, traffic states can be well estimated and further predicted to improve the overall operation efficiency. Combined with this trend, this study presents an up-to-date survey of open data and big data tools used for traffic estimation and prediction. Different data types are categorized, and off-the-shelf tools are introduced. To further promote the use of big data for traffic estimation and prediction tasks, challenges and future directions are given for future studies.

Keywords: big data; call detail records; census data; GPS trajectory data; location-based service data; open data; public transport transaction data; road sensor data; survey data

1. Introduction

Big data can be traced back to almost eighty years ago, when people encountered the first attempts to quantify the growth rate in the volume of data or what has popularly been known as the "information explosion" (a term first used in 1941, according to the Oxford English Dictionary). The term "big data" first appeared in a publication named "Application-controlled demand paging for out-of-core visualization" written by Michael Cox and David Ellsworth in October 1997 [1]. The first definition of big data is "Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data".

With the development of computers, smartphones, the internet, and sensory equipment, data continue to increase at faster and faster speeds. Broadly speaking, big data can be defined as data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process with low latency. The characteristics of big data are summarized into five Vs, which represent volume, velocity, variety, veracity, and value [2]. The volume represents a large amount of data with unknown value coming from mobile devices, social media, the Internet of Things (IoT), and more.

Data can accumulate to tens of terabytes or perhaps hundreds of petabytes. The velocity means the fast rate at which data are received and cumulated and the need for acting on data at an increasing pace. The variety refers to different data types available, including unstructured or semi-structured data, such as text, audio, and video. These data types often require additional preprocessing. The veracity represents the quality and accuracy of the data. Data in the world can be messy, especially in the case of big data when data dimensions and data sources increase.

The last characteristic is value. Data have intrinsic value, but it would have no use if it is not transformed into a usable format or it is unable to extract information from it. With the development of big data, relevant tools are also constantly being developed and updated to fulfill growing needs. Major data processing and analytics tools include Hadoop, Spark, Flink, Cordova, Kafka, and Mahout. Hadoop HDFS, HBase, MongoDB,

Citation: Jiang, W.; Luo, J. Big Data for Traffic Estimation and Prediction: A Survey of Data and Tools. *Appl. Syst. Innov.* **2022**, *5*, 23. https://doi.org/10.3390/asi5010023

Academic Editor: Konstantin Nikolic

Received: 29 December 2021
Accepted: 26 January 2022
Published: 11 February 2022

Hive, and SQL are the in-memory databases used for storing big data. Spark, S4 distributed stream computing platform and Apache Streams are applications that support processing streaming data.

As one of the typical application scenarios, big data has been widely used in the transportation domain. Closely connected to intelligent transportation systems, the emerging infrastructures of the IoT, cyber physical systems (CPS) and smart cities have provided great opportunities for the collection of big data through static sensors, surveillance cameras, and mobile devices. The data size has grown from Trillionbyte to Petabyte levels. Big data have been used for various transportation modes with multiple purposes, e.g., strategic air traffic management [3], travel pattern modeling [4], road congestion pattern prediction [5], etc.

Big data are used for government policies, e.g., decision-making support tools for smart cities, transparent governance and critical operations. Cutting-edge technologies integrated with big data are also used for urban planning and smart cities, for example, big data, in-memory computing, deep learning, and GPUs (graphics processing units) are used in rapid transit systems [6]. For a broader discussion of the applications of big data in the transportation domain, interested readers may refer to the relevant surveys [7–10].

Among different applications, traffic estimation and prediction are the two most significant tasks, which are the focus of this study. While various data related to traffic situations can be obtained, traffic information, e.g., volume, speed, and travel time, is not available without some processing of the raw data in some cases. The process is referred to as the traffic estimation problem, whose target is to extract precise traffic information from any raw data. The aims of traffic estimation include the implementation of a method that can be used to extract precise traffic information and further calibrate a traffic model, such as the MMS-model [11] (MMS-model is a dedicated model name, instead of an abbreviation).

While estimating the traffic information can only give us the historical states, the aim of traffic prediction is to predict the future situation based on the historical input and adopt appropriate measures accordingly, e.g., traffic control. Various methods have been proposed for traffic estimation and prediction tasks, including statistical models, machine learning models, and deep-learning models, in which deep-learning models are becoming dominant because they show the best performance [12–15]. The success of deep-learning models is partially attributed to big data because these models rely on a large training dataset. Several open datasets also boost the development and fair comparison among new models.

To summarize, significant progress has been achieved in previous studies for traffic estimation and prediction with the appearance of big data. The additional benefits of using big traffic data are multi-fold. First, bigger data are the basis of mining longer and more complex patterns hidden in the transportation domain. Second, bigger data make the effective training of artificial intelligence models feasible, especially the deep neural networks, when the data are not due to selection problems and structural changes in data are not considered.

Last, bigger data from various sources are the basis of capturing the relationship among different transportation systems. However, there is a lack of an up-to-date summary and collection of open datasets and tools. Some of the relevant studies are based on private data, whose results are impossible to replicate. In this survey, we focus on open datasets, especially large-volume and multi-modal datasets. To further boost the relevant studies, we also release a processed GPS trajectory dataset that is collected from more than 20,000 taxi drivers in Beijing in three months, namely, November 2012, November 2014, and November 2015, which has been used in our previous studies [16,17]. The dataset is publicly available (The data would be available here: https://github.com/jwwthu/DL4Traffic).

Our contributions in this paper are summarized as follows:

- We summarize the different data types used for traffic estimation and prediction tasks.
- We summarize the latest collection of relevant open datasets.
- We contribute a new GPS trajectory dataset for the research community.

- We summarize the collection of relevant big data tools.
- We point out the challenges and future directions of utilizing big data for traffic estimation and prediction tasks.

The following parts of this paper are organized as follows. The data used for traffic estimation and prediction tasks are summarized in Section 2. The big data tools are collected and introduced in Section 3. The relevant challenges and future directions are pointed out in Section 4. The conclusion is drawn in Section 5.

2. Data-Driven Traffic Estimation and Prediction

In this section, we summarize the different types of data that can be used for traffic estimation and prediction. We also contribute an up-to-date collection of available open datasets for each data type as well as new GPS trajectory data for further studies. There are different ways of categorizing traffic big data. For example, traffic big data were previously divided into supervised and unsupervised types, in which supervised data are direct sources of traffic information, e.g., loop detectors and GPS traces, while unsupervised data are indirect sources that can be used to infer traffic information, e.g., call detail records and cell phone position data. In this study, we further divide traffic big data into more types based on data sources.

Open data policies in different countries vary greatly. Taking the U.S. government as an example, transportation related open data can be found from the National Transit Database (NTD) (https://www.transit.dot.gov/ntd/ntd-data), Federal Highway Administration (FHWA) (For example, Urban Congestion Reports https://ops.fhwa.dot.gov/perf_measurement/ucr/), Bureau of Economic Analysis (BEA) (https://www.bea.gov/data/), and American Community Survey (ACS) (https://www.census.gov/programs-surveys/acs/data.html).

The Next Generation Simulation (NGSIM) (https://ops.fhwa.dot.gov/trafficanalystools/ngsim.htm) is the most widely used open-source vehicle trajectory dataset for traffic flow studies [18]. In this study, we mainly focus on open datasets in academia. More data collections can be found online, which may be maintained by individuals and research institutes, e.g., mobility datasets (https://privamov.github.io/accio/docs/datasets.html), open traffic collection (https://github.com/graphhopper/open-traffic-collection/), and Beijing City Lab (https://www.beijingcitylab.com/data-released-1/).

2.1. Trip Surveys

Trip surveys are detailed questionnaires on mobility habits, which are usually collected by local authorities or researchers. Sometimes, this is only a way to accurately measure and understand people's changing daily travel patterns, when different travel modes are considered and location privacy is not violated. For example, in the "My Daily Travel Survey" (https://www.cmap.illinois.gov/data/transportation/travel-survey#My_Daily_Travel_Survey) by the Chicago Metropolitan Agency for Planning between August 2018 and April 2019, households in northeastern Illinois were asked to record the trips they made for work, school, shopping, errands, and socializing with family and friends.

Another example is the "California Household Travel Survey (CHTS)" (https://www.nrel.gov/transportation/secure-transportation-data/tsdc-california-travel-survey.html) by the National Renewable Energy Laboratory (NREL) between 2010 and 2012. As the largest such regional or statewise survey ever conducted in the United States, detailed travel behavior information was obtained from more than 42,500 households via multiple data-collection methods, including computer-assisted telephone interviewing, online and mail surveys, wearable (7574 participants) and in-vehicle (2910 vehicles) global positioning system (GPS) devices, and on-board diagnostic sensors that gathered data directly from a vehicle's engine.

Details of personal travel behavior were gathered within the region of residence, inter-regionally within the state, and in adjoining states and Mexico. The survey sampling plan was designed to ensure an accurate representation of the entire population of the state.

Among the participating households, 42,436 completed the travel diary survey portion, 3871 completed the wearable GPS portion, and 1866 completed the vehicle GPS portion of the study. Trip details, including purpose, mode, travelers, tolls, departure time, arrival time, and distance were collected for both private vehicles and public transit trips. More trip surveys are available on the internet (https://www.nrel.gov/transportation/secure-transportation-data/tsdc-cleansed-data.html).

The pros of trip surveys include high-resolution and detailed information, which contain trip information directly and eliminates the need for traffic estimation. The cons of trip surveys include small sample size, limited spatial and temporal scale, self-reporting errors, and high cost to collect, which limits the availability of such data as well as real-time applications, e.g., they are less used for traffic prediction.

2.2. Census Data and Survey Data

Census data and survey data may also contain traffic information and other information useful for traffic problems, such as locations of residence and workplace. These data are usually collected by governments periodically. For example, in the Commuting Flows (https://www.census.gov/topics/employment/commuting/guidance/flows.html) by the US Census Bureau, the primary workplace location from the respondents are collected. In the Migration Data (https://www.irs.gov/statistics/soi-tax-stats-migration-data) by the Internal Revenue Service, both inflows and outflows of migration patterns are available for Filing Years 1991 through 2018 with the IRS. More survey data can be found in Table 1.

Table 1. The list of open survey data.

Name	Type	Spatial Range	Temporal Range	Download Link
GB Road Traffic Counts	Aggregated traffic counts	Great Britain	2000–2019	https://data.gov.uk/dataset/208c0e7b-353f-4e2d-8b7a-1a7118467acc/gb-road-traffic-counts
Highways England network journey time and traffic flow data	Traffic flow, travel time	Great Britain	2006–2020	https://data.gov.uk/dataset/9562c512-4a0b-45ee-b6ad-afc0f99b841f/highways-england-network-journey-time-and-traffic-flow-data
DataMall	Multiple types	Singapore	Frequently updated	https://datamall.lta.gov.sg/content/datamall/en/dynamic-data.html
Minnesota Department of Transportation Traffic Data	Traffic flow, traffic speed	Twin Cities, MN, USA	Since 2011	http://data.dot.state.mn.us/datatools/
Chicago Traffic Tracker	Traffic speed	Chicago, IL, USA	2011–2018	https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/77hq-huss/data
US-Accidents [19]	Traffic accident	USA	February 2016 to December 2020	https://smoosavi.org/datasets/us_accidents

The pros of census data and survey data are their large sample size and large coverage, which is usually the entire country. The cons are also obvious. The flows are aggregated at the municipality or county level, which is coarse. There is only limited traffic information. It is also expensive and time consuming to collect this type of data as trip surveys. Census data and survey data are not widely used for traffic estimation and prediction tasks with only a few exceptions [20].

2.3. Road Sensor Data

Static sensors can be deployed to collect traffic data, e.g., inductive loop detectors, magnetic and pneumatic tube sensors, radar sensors, infrared sensors, and acoustic sensors. Loop detectors are the most mature technology among road sensors and contribute many famous open datasets for traffic estimation and traffic prediction. A complete list of open

road sensor data used in previous studies is shown in Table 2. The pros of road sensor data are their abilities to collect large volume data continuously and automatically, usually in minutes or seconds.

The cons of these data are their fixed coverage in the spatial range as well as the missing data problem caused by sensor failures. Additionally, the traffic speed obtained from traffic sensors could be a point measurement and may not be suitable to calculate the average travel time across the road link.

Table 2. The list of open road sensor data.

Name	Type	Spatial Range	Temporal Range	Download Link
Performance Measurement System (PeMS) Data	Traffic speed, traffic flow	California, USA	2001–2019	http://pems.dot.ca.gov/
METR-LA [21]	Traffic speed, traffic flow	Los Angeles, USA	1 March to 30 June 2012	https://github.com/liyaguang/DCRNN
Seattle-Loop-Data [22]	Traffic Speed	Seattle, USA	1 – 31 January 2015	https://github.com/zhiyongc/Seattle-Loop-Data
Traffic Speed Guangzhou [23]	Traffic speed	Guangzhou, China	1 August 2016 to 30 September 2016	https://zenodo.org/record/1205229
Traffic Flow Madrid	Traffic flow	Madrid, Spain	Since 2013	http://datos.madrid.es
Traffic Flow Whitemud Drive	Traffic flow	Whitemud Drive, Canada	6 August 2015 to 28 August 2015	http://www.openits.cn/openData1/700.jhtml
UTD19 [24]	Traffic capacity	40 cities globally	Various time periods	https://utd19.ethz.ch/index.html
PORTAL	Traffic speed	Portland-Vancouver Metropolitan region, USA	Regularly updated	https://portal.its.pdx.edu/home

Traffic signals also belong to road sensor data, which are often seen in the literature. Different types of traffic signal data can be collected for traffic lights, traffic signal lanterns, traffic signal controller boxes, traffic signal poles, etc. For example, traffic light data are collected in those traffic lights with an internet connection, with a typical data type of numeric values for time to green. They also contain location and attributes of traffic controls located at each intersection. This type of data is more often used for traffic control relevant studies, instead of traffic estimation or prediction. For example, an open collection of the traffic light data is provided in [25] for the study of reinforcement learning based traffic control.

2.4. Call Detail Records

Call detail records (CDRs), cellular signaling data, and cell phone position data are all generated from the cellular network, which contains similar information on human mobility patterns. These data are produced by a phone carrier to store details of calls passing through a device and usually contain various attributes of the call, e.g., timestamp, source, destination, base transceiver station (BTS), duration. Then, the position and trajectory (sequence of locations) can be inferred from CDRs. Compared with GPS trajectory data with a higher spatial resolution, CDRs provide a coarse positioning approach. Data preprocessing is necessary not only for removing noise but also for identifying the key locations.

As one of the pioneering studies, cellular phone tracking data are used for traffic volume estimation and further evaluated with loop detector data in [26]. Afterwards, an increasing number of studies have used call detail records and cell phone position data for traffic status inference, even in recent years. One-month cellular signaling data (SD) are used to extract road-level human mobility with a multi-information fusion framework, which takes the SD uncertainty issue into consideration [27].

Based on a regression model, a cell probe (CP)-based method is proposed to estimate the vehicle speed with the normal location update (NLU) procedure and the consecutive handoff (HO) event as inputs, and the proposed method achieved a 97.63% accuracy [28]. Long-term evolution (LTE) access data are used as input for a deep-learning-based road traffic prediction system, to predict the real-time speed of traffic [29]. Characteristics of human mobility patterns were revealed by high-frequency cell phone position data in [30]. More usages can be found in recent surveys [31].

The publicly available CDR or similar data are shown in Table 3. The pros of using CDR or similar data are their ubiquitous availability (with mobile phones), large volume, and multiple dimensions (social, mobile, time, and demographics). However, the cons include poor positioning ability, noise, and preprocessing requirements. As the position can only be derived from the cellular BTSs, it is partially detected only when calls are made and known at the BTS level only.

The ping-pong effect between different BTSs can create noise in the collected data. Another concern when using CDR data is the potential privacy leakage issue for both the personal and location information. Several privacy protection solutions have been proposed. The personal information is anonymized and encrypted with a minimal possibility of tracing back to the individuals. The location information can be aggregated in a larger region without revealing the precise coordinates.

Table 3. The list of CDR data.

Name	Spatial Range	Temporal Range	Download Link
Telecommunications-SMS, call, internet-MI [32]	Milan, Italy	1 November 2013 to 1 Januray 2014	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EGZHFV
OpenCellID	Multiple cities globally	Regularly updated	https://www.opencellid.org

2.5. GPS Trajectory Data

GPS trajectory data are often collected from floating vehicles, which are equipped with on-board positioning systems and communication devices. Additionally, smartphones can also be used for GPS trajectory data collection, e.g., in a crowd-sourced approach. While GPS trajectory data are easy to collect, the preprocessing steps are necessary, e.g., location detection to group points into one meaningful location and trajectory segmentation to split a trajectory in sub-trajectories. Map matching is also used to map the GPS coordinates into the road network.

Due to advantages, such as fine granularity, GPS-based "Track and Trace" data were formally defined and highlighted for transportation modeling and policy-making in a recent survey [33]. Taxi trajectory data can also be used for trip purpose inference and travel pattern discovery [4]. More relevant studies are using GPS trajectory data for traffic estimation and prediction, e.g., travel time is estimated with the GPS trajectory data in [34] by combining the gradient boosting decision tree (GBDT) model and the deep neural network (DNN) model, and OD pairs are predicted with big GPS data in [35].

Due to privacy concerns, most of the open GPS trajectory data are collected from taxis, with only a few exceptions, as shown in Table 4. Additionally, in many cases, the raw GPS trajectory data are not shared, and only the estimated traffic states derived from the GPS trajectory are publicly available. The pros of using GPS trajectory data include the high spatial and temporal resolution as well as the tracking ability of the full traces. Taxis can also be seen as mobile probes, which reflect all road traffic situations. The cons are similar to CDRs, which include the preprocessing requirement and the noise in the data. Moreover, the GPS trajectory data are often collected at a high frequency, e.g., in seconds, so the huge volume of such data requires storage and computation abilities, which can only be fulfilled by big data tools.

While taxi GPS trajectory datasets are widely used in the literature, there is a concern that taxi drivers are not the best representatives since they are experienced drivers with a

higher possibility of finding optimal routes. In some studies, this wisdom from taxi drivers is used to build better navigation services [36]. On the contrary, the driving behavior gap between taxi drivers and private car drivers are becoming smaller as these navigation services are now widely used by both groups. In the studies for traffic estimation and prediction, taxi drivers are preferred as probes because they drive longer and provide more traffic situation measurements with a larger coverage.

Table 4. The list of open GPS trajectory and relevant data.

Name	Type	Spatial Range	Temporal Range	Download Link
SF Taxis or Cabspotting [37]	Taxi GPS trajectory	San Francisco, CA, USA	17 May 2008 to 10 June 2008	http://crawdad.org/epfl/mobility/20090224/index.html
Rome Taxis [38]	Taxi GPS trajectory	Rome, Italy	1 Feburary 2014 to 3 March 2014	http://crawdad.org/roma/taxi/20140717/index.html
Porto Taxis	Taxi GPS trajectory	Porto, Portugal	1 July 2013 to 30 June 2014	https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i/data
Geolife [39]	Taxi GPS trajectory	Beijing, China	April 2007 to August 2012	https://www.microsoft.com/en-us/download/details.aspx?id=52367
Mobile Data Challenge (MDC) [40]	Taxi GPS trajectory	Lake Geneva region	2009 to 2011	https://www.idiap.ch/dataset/mdc
TaxiCD	Taxi GPS trajectory	Chengdu, China	3 August to 30 2014	https://js.dclab.run/v2/cmptDetail.html?id=175
Grab-Posisi [41]	Grab Drive GPS trajectory	Southest Asia	8 April 2019 to 21 April 2019	Upon request
PrivateCarTrajectoryData	Private car GPS trajectory	Shenzhen, China	January 2016	https://github.com/HunanUniversityZhuXiao/PrivateCarTrajectoryData
TaxiBJ [42]	Taxi traffic flow	Beijing, China	Multiple time periods	https://github.com/Mouradost/ASTIR or https://www.jianguoyun.com/p/DesHv2UQs-HRBxi5gtYB
BikeNYC [43]	Bike traffic flow	NYC, USA	1 April 2014 to 30 Sepetmber 2014	https://github.com/Mouradost/ASTIR or https://www.jianguoyun.com/p/DesHv2UQs-HRBxi5gtYB
Traffic Speed Chengdu [44]	Taxi traffic speed	Chengdu, China	1 June 2015 to 15 July 2015	https://doi.org/10.6084/m9.figshare.7140209.v4
SHSpeed (Shanghai Traffic Speed) [45]	Taxi traffic speed	Shanghai, China	1 April to 30 2015	https://github.com/xxArbiter/grnn
TaxiSZ [46]	Taxi traffic speed	Shenzhen, China	1 January to 31 2015	https://github.com/lehaifeng/T-GCN

In addition to the existing datasets, we also contribute a new GPS trajectory dataset for the research community in this study. The GPS trajectory data were collected in Beijing during three time periods, namely, November 2012, November 2014, and November 2015. Each GPS data sample contains the following attributes: anonymous taxi identity, timestamp, latitude, longitude, azimuth, spot speed, and operation status (occupied, vacant, or stopped). The data were sampled with an interval of approximately one minute. The data summary is shown in Table 5. This dataset is publicly available (The data is available here: https://github.com/jwwthu/DL4Traffic).

Table 5. Data summary for the GPS trajectory dataset contributed in this study.

Time Periods	November 2012	November 2014	November 2015
Taxi Drivers	8879	17,749	20,067
Days	30	30	30

2.6. Location-Based Service Data

With the development of the mobile internet, location-based service (LBS) data have arisen with GPS functionality on smartphones. Various location-based big data are collected from location-based social networks, e.g, checkins, geotagged tweets, and micro-blogs, and Maps and Navigation Apps, e.g., Google Map and Baidu Map. These data are often collected in a crowd-sourced approach, and a traffic information extraction step is necessary. To effectively collect, integrate, and process crowd-sourced data, including mobile applications, webs, and external data sources, a prototype system was developed and validated in [47] to infer traffic conditions.

Traffic speed data of 29 cities across the world over a 40-day period were gathered from Google Map API and further used for the analysis of traffic congestion patterns [48]. Social media texts may also contain traffic information, e.g., geotagged tweets and micro-blogs. However, the challenge is to extract traffic-relevant information from natural languages. Deep-learning models have been proposed for conducting information extraction, e.g., an LSTM-CNN was proposed to extract traffic-relevant microblogs in [49], which outperformed the baselines, including the support vector machine model based on a bag of n-gram features. Social media data were further proven effective for traffic accident detection and reporting [50,51], traffic jam management [52].

Combining social media data from multiple social networks can further improve the traffic event detection accuracy, e.g., the case of combining Arabic and English data streams from Twitter and Instagram used in the SNSJam system [53]. However, location-based big data may be low-quality with noise. Taking Bluetooth speed data as the ground truth, the quality of Waze data in Sevierville, TN, USA was evaluated in [54], in which a kNN method managed to achieve a prediction accuracy of 84.5% and 82.9% for traffic speed estimation based on Waze data as the input.

The list of open location-based service data is shown in Table 6. The pros of using LBS data include their wide availability and semantic information, e.g., restaurants and malls. However, the cons are also clear. Human mobility can only be partially detected when check-ins or geotagged texts are made. The data are thus very sparse and can be sparser than CDRs. There is also self-selection bias, which would cause inaccurate traffic information.

Table 6. The list of open location-based service data.

Name	Type	Spatial Range	Temporal Range	Download Link
Q-Traffic or Baidu-Traffic [55]	Traffic speed from navigation apps	Beijing, China	1 April 2017 to 31 May 2017	https://github.com/JingqingZ/BaiduTraffic
Alibaba Cloud Tianchi Dataset	Travel time from navigation Apps	Guiyang, China	April 2017	https://tianchi.aliyun.com/competition/entrance/231598/information
Brightkite [56]	Check-ins	N/A	April 2008 to October 2010	https://snap.stanford.edu/data/loc-brightkite.html
Gowalla [56]	Check-ins	N/A	February 2009 to October 2010	https://snap.stanford.edu/data/loc-gowalla.html
Foursquare [57]	Check-ins	NYC, USA	24 October 2011 to 20 February 2012	https://sites.google.com/site/yangdingqi/home/foursquare-dataset

Table 6. *Cont.*

Name	Type	Spatial Range	Temporal Range	Download Link
Yelp	Check-ins	Multiple cities globally	Regularly updated	https://www.yelp.com/dataset
MapBJ [58]	Traffic congestion from navigation Apps	Beijing, China	March 2016 to June 2016	https://github.com/cxysteven/MapBJ
Tecent API	Traffic flow index	China	Since 2015	https://heat.qq.com/
Uber Movement	Travel time and speed	Multiple cities globally	Since 2017	https://movement.uber.com/

2.7. Public Transport Transaction Data

Transaction data can be collected in various public transport systems, especially those with automatic fare collection (AFC) systems, and further used for traffic estimation and prediction. For example, smart card data are often used for public transit origin-destination (OD) estimation [59]. Based on smart card and bus trajectory data, a two-stage transportation analysis approach is proposed to reconstruct the individual passenger trips and cluster these trips to identify the transit corridors in [60].

Based on 10 million taxi trip records in New York and Shenzhen and using spatio-temporal clustering, Bayesian probability, and Monte Carlo simulation, a two-layer framework, which consists of an activity inference model and a pairing journey model, was proposed to extract and predict travel patterns [61]. Using the gravity model and Bayesian rules, the purpose of dockless shared-bike users was inferred from a shared bike dataset in Shenzhen, China, and the introduction of activity type proportion and service capacity of point of interest (POIs) was proven effective for inference [62].

Based on the daily OD amount by transportation and by purpose and several surveys on time use and activities, an open people mass movement dataset named Open PFLOW was built in [63], which achieved a comparable accuracy with other approaches, e.g., commercial mobility data and traffic census.

Another approach of collecting public transport transaction data is based on internet of Things and counter devices installed in the vehicle doors, e.g., infrared or RGB image-based passenger counters. This passive approach of collecting data can be conducted silently without the burden of passenger involvement. However, more errors may exist in the collected due to the device fault or misjudgement caused by passenger's strange behaviors.

There are also many public transport transaction data that are available for the research community as shown in Table 7. The pros of using public transport transaction data are their wide availability and close connection with traffic states. The cons are the storage and computation requirements, which can be high if the data cover wide spatial and temporal ranges.

Another problem with public transport transaction data is that they can be incomplete. For example, in some countries and regions, tickets are only validated at the entrance to the vehicles, e.g., bus or subway. In those cases, only the inflow data can be collected, and the scope of follow-up studies is limited without being able to obtain the outflow or OD flow situations.

Table 7. The list of open public transport transaction data.

Name	Type	Spatial Range	Temporal Range	Download Link
SHMetro [64]	Metro	Shanghai, China	1 July 2016 to 30 September 2016	https://github.com/ivechan/PVCGN
HZMetro [64]	Metro	Hangzhou, China	January 2019	https://github.com/ivechan/PVCGN
Hangzhou Metro	Metro	Hangzhou, China	1 January 2019 to 25 January 2019	https://tianchi.aliyun.com/competition/entrance/231708/information
Bike Bay Area [65]	Shared bike	Bay Area, USA	1 September 2015 to 31 August 2016	https://github.com/TwinkleBill/babs_open_data_year_3
BikeNYC	Shared bike	NYC, USA	1 2013 to 12 December 2016	https://www.citibikenyc.com/system-data
BikeDC	Shared bike	Washington, DC, USA	2011–2016	https://www.capitalbikeshare.com/system-data
BikeChicago	Shared bike	Chicago, IL, USA	2015–2020	https://www.divvybikes.com/system-data
Bike Chattanooga Trip Data	Shared bike	Chattanooga, Tennessee, USA	23 July 2012 to 9 April 2020	https://data.chattlibrary.org/
Mobike Beijing [66]	Shared bike	Beijing, China	10 May 2017 to 24 May 2017	https://github.com/SharingBikeNNU/Riding-Modes_Tucker
Ride Austin	Ride sharing	Austin, USA	2 June 2016 to 13 April 2017	https://data.world/ride-austin
UberNYC	Ride sharing	NYC, USA	from April to September 2014	https://github.com/fivethirtyeight/uber-tlc-foil-response
Didi GAIA Open Data	Ride sharing	Chengdu, Xi'an, and Haikou, China	Various time periods	https://outreach.didichuxing.com/research/opendata/
TaxiNYC	Taxi	NYC, USA	Since 2009	http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

2.8. Surveillance and Airborne Digital Cameras

Closed circuit television (CCTV) cameras are widely used for monitoring traffic patterns and helping police forces in large cities. By collecting these surveillance videos, traffic information can be estimated and further used for prediction. For example, traffic states with CCTV systems were analyzed in [67]. However, the challenge is the heavy workload of processing digital multimedia data. Transportation video data management was considered in [68], and a high-performance computing architecture was developed with distributed files and distributed computing systems.

Edge computing is also applied for traffic flow detection in real-time videos. A specific edge device Jetson TX2 platform was used in [69], with the YOLOv3 (You Only Look Once) model for vehicle detection and the optimized DeepSORT (Deep Simple Online and Realtime Tracking) algorithm for vehicle tracking. The edge device managed to achieve an average processing speed of 37.9 frames per second, with an average accuracy of 92.0%.

In addition to static traffic surveillance cameras, remote sensing tools, e.g., airborne digital cameras, and unmanned aerial vehicle (UAV) cameras can also be used for traffic monitoring. For example, the average traffic speed and density were estimated with an airborne optical 3K camera system in [70], and the traffic flow parameters were calculated from UAV aerial videos in [71].

Some open video data are summarized in Table 8. The pros of video data are their continuous observation, wide coverage for road traffic, and multiple usages for traffic estimation, prediction, accident analysis and even vehicle tracking. The cons of video data are their high storage and computation requirements. Video streams also have many engineering challenges, e.g., compression artifacts, blurring, and hardware faults.

Table 8. The list of open video data.

Name	Spatial Range	Temporal Range	Download Link
MIT Traffic Dataset [72]	One camera in MIT	90 min long	http://mmlab.ie.cuhk.edu.hk/datasets/mit_traffic/index.html
Video Surveillance Data [73]	One camera in Chelyabinsk, Russia	982 video frames	https://github.com/alnfedorov/traffic-analysis

2.9. License-Plate Recognition Data

License plate recognition (LPR) data are emerging data sources for vehicle re-identification as well as traffic estimation and prediction. Large-scale LPR data collection systems are being developed all over the world, e.g., the automatic number-plate recognition (ANPR) systems. However, due to data privacy concerns, there are few open datasets and few relevant studies.

In [74], a six-day LPR dataset from a small road network in Langfang, China, was used to estimate and predict link-based traffic states, and the feasibility was validated by using a more comprehensive link-level field experiment dataset. A similar type of data is electronic registration identification (ERI), which was also used in traffic flow prediction [75]. The pros of LPR or ERI data are their precise tracking abilities of individual vehicles. However, this feature also causes data privacy concerns. Another disadvantage is that there are no publicly available LPR or ERI datasets.

2.10. Toll Ticket Data

Electronic toll collection (ETC) systems are widely applied in toll roads, HOV lanes, toll bridges, etc. For example, more than 90 percent of vehicles on expressways in China were equipped with the ETC system by the end of 2019. Toll ticket data (TTD) collected from ETC systems provide a high-coverage and low-cost approach for expressway traffic estimation.

Based on the TTD from the Shandong Expressway Toll System in China, a simulation-based dynamic traffic assignment algorithm was proposed to obtain traffic flow in [76] and achieved a high examination accuracy (approximately 5% MAPE). The only public source of toll ticket data used for traffic flow prediction that the authors know is the Amap data for KDD CUP 2017 (https://tianchi.aliyun.com/dataset/dataDetail?dataId=60). The pros of tool ticket data include their high coverage and low cost. However, the cons are clear. Toll ticket data can only be collected in places where ETC systems are applicable.

2.11. External Data

In addition to the above data that are used to extract and predict traffic states, some external data are also used for these problems as supplementary data, e.g., weather data, calendar information, emissions of air pollutants, and noise. These data do not have a large volume and are easily available. There are many public sources for weather data, e.g., Weather Underground API (https://www.wunderground.com/weather/api/), MesoWest (http://mesowest.utah.edu/). Calendar information is also widely accessible from the internet, e.g., OfficeHolidays (https://www.officeholidays.com/). To name some relevant studies that use these external data, 48 weather forecasting factors were analyzed and used for traffic flow prediction based on a regression model in [77], and rainfall data have also been used for similar purposes [78,79].

This section is concluded with Table 9, which shows the relationship between the potential data sources with possible data attributes. This table is far from being perfect but is a preliminary reference.

Table 9. A summary for the linking between data sources and possible results.

	Road Traffic Volume	**Road Traffic Speed**	**Cyclist Speed**	**Bicycle Volume**	**Passenger Flow in Public Transport Systems**	**Pedestrian Speed**	**Pedestrian Volume**	**Pedestrian Route**
Traffic signal controllers & Road sensors	+	+						
GPS sensors	+	+						
Bicycle counters				+				
Passenger counters & AFC systems					+			
On-board computers					+			
Cellular phones & Smartphones						+	+	+
CCTV cameras	+	+	+	+	+	+	+	+
ANPR systems	+	+(1)						
ETC systems	+(2)							

(1) only segmental speed measurement; (2) only volume measurement in the entrance.

3. Big Data Tools

There are already some preliminary trials of applying the latest big data tools for traffic big data, as traditional big data tools are not optimized for big data in the transportation domain. Some preliminary efforts have been proposed for expanding the off-the-shelf tools for a better support of spatiotemporal big data, e.g., Hadoop has been expanded with the capacities of spatiotemporal indexing [80] and trajectory analytics [81]. Based on PostgreSQL and PostGIS, an open-source mobility database named MobilityDB [82] was proposed for moving object geospatial trajectories, e.g., GPS traces [83].

For analyzing and visualizing spatiotemporal big data, new data processing algorithms and methods should be implemented as new or extensions of existing commercial or open-source platforms, for ease of use and a better integration, e.g., the latest SuperMap GIS platform provides full support to the Spark computing framework [84]. Offline and online trajectory analyses are often separated in existing systems.

To overcome this shortcoming, a Spark-based hybrid and efficient framework named Dragoon [85] was proposed for both offline and online trajectory data analytics. Dragoon manages to decrease storage overhead up to doubled compared with the state-the-art big trajectory data management system Ultraman [86], while maintaining a similar offline trajectory query performance. Dragoon also manages to achieve at least 40% improvement of scalability over Flink and Spark Streaming and achieves an average doubled performance improvement for online trajectory data analytics.

Based on cellular data (2G/3G/4G), base station data, user information, and road network data, a real-time urban mobility monitoring and traffic management system was proposed in [87], leveraging big data tools, including Hive, Spark, and Hbase. The system proposed a total of 600 TB cellphone data collected over 3 million people daily for 3 years in a field case study in Guiyang, China.

In this section, we collect and summarize the off-the-shelf big data storage and computing tools that are already used or can be further leveraged in traffic estimation and prediction tasks. While these tools have not been widely used yet, there is already the trend of more and more data being collected and used in traffic problems, e.g., the large volume of GPS data may exceed the level of TB for a metropolitan area. The storage and computation resources of traditional tools may not be sufficient in the near future.

3.1. Existing Tools

3.1.1. Apache Hadoop

Apcahe Hadoop [88], first released in April 2006, is a sub-project of Lucene (a collection of industrial-strength search tools), under the umbrella of the Apache Software Foundation. Hadoop is written in Java and is a great big data tool because it can process structured and unstructured data from different sources. It parallelizes data processing utilizing computer clusters, accelerating large computations and hiding I/O latency through increased concurrency. It can also leverage its distributed file system to cheaply and reliably replicate chunks of data to nodes (computers) in the cluster, making data locally available on the machine that is processing it. Additionally, Hadoop provides, to the application programmer, the abstraction of map and reduce operations.

3.1.2. Apache Pig

Apache Pig [89], first released in September 2008, was initially developed by Yahoo's researchers for executing MapReduce jobs on large datasets on a high-level of abstraction. It provides Pig Latin, a high-level scripting language, for writing data analysis code. Users write a script using the Pig Latin Language to process data in HDFS. The Pig Engine, a component of Apache Pig, converts all the scripts into a map and reduces the task for processing. However, this is a totally internal process, and thus the programmers would not see the procedures. The result of Pig would be stored in HDFS after finishing. Compared to MapReduce, the Apache pig reduces the time of development using the multi-query

approach. The same job could be completed in much less code using Pig Latin compared to using Java.

Pig Latin can be extended using user-defined functions (UDFs) that the user can write in Java, Python, JavaScript, Ruby, or Groovy and then call directly from the language.

3.1.3. Apache Mahout

Apache Mahout [90], first released in Apr 2009, is an open-source project to provide service for the application of distributed and scalable machine learning algorithms focused on linear algebra and statistics. It is written in Java and Scala. Mahout operates in addition to Hadoop, which allows users to apply machine learning algorithms through distributed computing on Hadoop. Mahout's core algorithms include recommendation mining, clustering, classification, and frequent item-set mining. Developers are still working on Mahout but a number of algorithms have been implemented.

3.1.4. Apache Spark

Apache Spark [91] started as a research project at the UC Berkeley AMPLab in 2009 and was open-sourced in early 2010. It is a cluster-computing technology designed for fast computation with its own cluster management. As an extension of the Hadoop MapReduce model, it allows more types of computation, including batch applications, iterative algorithms, interactive queries, and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application by reducing the number of read/write operations to disk. Spark provides built-in APIs in Java, Scala, or Python that enable writing applications in different programming languages.

3.1.5. Apache Kafka

Apache Kafka [92], initially released in January 2011, is an open-source distributed publish-subscribe messaging system created by LinkedIn and written in Scala and Java. It aims to provide a unified, high-throughput, and low-latency platform for handling and messaging high-volume real-time data. Kafka utilizes a binary TCP-based protocol that is optimized for efficiency and group messages to reduce the overhead of the network roundtrip. Kafka is also fault-tolerant as the messages persist on the disk and replicate within the cluster and are built on top of the ZooKeeper synchronization service.

3.1.6. Apache Flink

Apache Flink [93], initially released in May 2011, is an open-source, unified stream-processing and batch-processing framework developed by the Apache Software Foundation. The main part of Apache Flink is a distributed streaming dataflow engine that is written in Java and Scala. It executes tasks in a parallel and pipelined way and thus can perform bulk/batch processing programs, stream processing programs, and iterative algorithms. In addition to providing a high-throughput, low-latency streaming engine, Flink also supports event-time processing and state management. Applications developed using Flink are fault tolerant when machines fail, as Flink relies on external fault tolerance services for maintaining configuration information and distributed synchronization.

3.1.7. Apache Storm

Apache Storm [94], initially released in September 2011, is a distributed stream processing computation framework originally created by Nathan Marz and the team at BackType. Storm was later acquired and open-sourced by Twitter and became a standard for distributed real-time processing systems in a short amount of time. Storm is written predominantly in the Clojure programming language. It uses custom created "spouts" and "bolts" to define the data manipulation process to allow batch, distributed processing of streaming data. "Spouts" would take in data and distribute to "bolts", and the functions and codes users write would be processed in bolts. The whole storm application composed

of "spouts" and "bolts" could be represented by a directed acyclic graph (DAG) and form a "topology".

Data flow through edges on the graph are named streams. Together, the topology acts as a data transformation pipeline. The general topology structure is similar to a MapReduce job but the data are processed in real time instead of in individual batches. Additionally, Storm topologies run indefinitely until being shut down or encounter failure, while a MapReduce job DAG must eventually end.

3.1.8. Apache HDFS

The Hadoop Distributed File System (HDFS) [95], first released on 4 September 2007, is a distributed file system designed to run on low-cost commodity hardware. HDFS can store and process large application datasets by storing data files across multiple machines and through parallel processing. It is highly fault tolerant as it stores data redundantly across data nodes. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. Originally built as infrastructure for the Apache Nutch web search engine project, HDFS is now an Apache Hadoop subproject.

3.1.9. Apache HBase

Apache HBase [96], initially released in March 2008, is an open-source, column-oriented, non-relational database that evolved from Google's Bigtable. It is written in Java and sits on top of the HDFS or Alluxio, providing Hadoop with Bigtable functionalities. It is capable of storing a large amount of sparse data in a fault-tolerant way. In addition, HBase provides compression, in-memory operations, and Bloom filter features on a per-column basis.

Through the Java API and many other APIs, users can input tables from HBase into MapReduce jobs run in Hadoop, and MapReduce jobs can also output in HBase tables format. Due to the lineage with Hadoop and HDFS, HBase has been adopted worldwide. HBase cannot replace a classic SQL database; however, the SQL layer and JDBC driver are provided by the Apache Phoenix project to enable the use of HBase with analytics and business intelligence applications.

3.1.10. MongoDB

MongoDB [97], first released in February 2009, is an open-source document-oriented database. Classified as a NoSQL database program, MongoDB does not use regular tables and rows to store data but instead uses JSON-like documents. These documents support embedded fields and related data can be stored within them. MongoDB has optional schemas; thus, users are not required to specify the number or type of columns before inserting data. MongoDB was developed by MongoDB Inc. and licensed under the Server Side Public License (SSPL).

3.1.11. Apache Hive

Apache Hive [98], first released in October 2010, is a data warehouse infrastructure initially developed by Facebook for querying and analyzing data. It is built on top of Apache Hadoop and provides users with an SQL-like interface for manipulating data stored in databases and file systems that integrate with Hadoop. Hive solves the problem of having to use Java API for executing SQL applications and queries over distributed data by providing its own SQL type language for querying called HiveQL or HQL. Since most data warehousing applications work with SQL-based querying languages, Hive facilitates the use of SQL-based applications to Hadoop.

3.2. Comparison and Recommendation

Since these tools are designed for general purposes and not dedicated for traffic data and problems, it is necessary to compare the tools with similar functionalities and give a

recommendation for choosing suitable big data tools. The advantages and disadvantages of big data databases and tools are summarized in Table 10.

While more recent tools have new features and functionalities supported, they are also prone to potential errors. In this study, we recommend more mature tools that are already widely used in other fields. Specifically, we recommend the combination of Apache HDFS and Apache Hadoop for those with little or no previous experience of using big data tools and Apache Spark for those with some experience.

Table 10. A comparison of big data tools.

Tool	Main Purpose	Advantage	Disadvantage
Apache Hadoop	Distributed computing	Mature and reliable.	Difficult to use.
Apache Pig	Distributed computing	High-level interface.	The need to learn Pig Latin language.
Apache Mahout	Distributed machine learning	Support for machine learning algorithms.	Performance bottleneck of default models.
Apache Spark	Distributed computing	In-memory cluster computing. Easy to use.	No automatic optimization process.
Apache Kafka	Distributed messaging	Low latency. High throughput.	Reduced performance.
Apache Flink	Distributed computing	High efficiency. Easy to use.	Immature and lack of API support.
Apache Storm	Distributed computing	Fast and fault tolerant.	Difficult to learn and use.
Apache HDFS	Database	Fault tolerant. Integrated with Apache Hadoop.	Difficult to use.
Apache HBase	Database	Real-time querying. Suitable for sparse data. Low-latency operation.	No SQL-like interface.
MongoDB	Database	Document-oriented database.	Transactions are not supported. Limited data size. High memory usage.
Apache Hive	Database	SQL-like interface.	Not a full database. No real-time querying.

4. Challenges and Future Directions

While there are many successful application cases of big data for traffic estimation and prediction tasks, challenges still exist. For now, the available data sources presented in this study are still limited and can be used in statistical production only in a very limited way. There are still severe difficulties in the data generating process, which requires joint cooperation among academia, industry and government. Legislation is another matter to consider, since many companies do not want to share and would rather sell the data or use their monopoly, e.g., with mobile data in their countries.

Data richness varies greatly for different transportation modes and the problems of data sparsity, high missing data ratios, data noise, and lack of data still exist. Data quality, privacy, and policies have not been fully considered in previous studies. For example, crowd-sourced data have the problems of low data quality, noise removal difficulty, and privacy concerns. Some efforts have been made to address these challenges, e.g., sparse Bayesian learning was used for traffic state estimation with under-sampled data [99]. Existing data for traffic estimation and prediction tasks are heterogeneous in the spatial and temporal ranges. Cross-scale data fusion by integrating various sources is still challenging in the transportation domain.

Another challenge is the lack of “real” big data in the transportation domain, especially open data. While we reviewed many open data sources in this study, some of them have a data volume that is difficult to define as “big". The time range of some available datasets is not long enough for training effective deep-learning models. This challenge is partially caused by the high-cost and time-consuming process of collecting some types of data.

The other reason is the concern of location privacy leakage, which prohibits the collection of fine-grained data. For existing big datasets, e.g., GPS trajectories, the existing big data tools discussed in Section 3 are not yet fully exploited. With the popularity of using graph data for traffic prediction, e.g., transportation networks in graph neural networks, the existing graph processing tools cannot fully meet the computing requirements; thus, new tools are needed.

To address these challenges, some future directions are pointed out in this section.

4.1. Heterogeneous Data Fusion

A single data source may not be enough for traffic tasks, when different transportation modes are entangled together. Data from multiple sources can be combined to make a better estimation and prediction for traffic situations. Heterogeneous data fusion is driven by this observation, which often uses deep learning for urban big data fusion [100]. There are already some relevant studies. For example, cellular data and loop detectors were integrated for freeway traffic speed estimation in [101]. Sparse stationary traffic sensor data were combined with GPS trace data to estimate the traffic flow in the entire road network in [102].

License plate recognition data and cellular signaling data were combined for traffic pattern and population distribution correlation analysis in [103]. License plate recognition data and GPS trajectory data are combined for traffic flow estimation in [104]. Geomagnetic detector data, floating car data, and license plate recognition data were combined for average link travel time extraction [105]. Bus transit schedule data, real-time bus location data, and cell phone data from geographical mapping software were combined to predict bus delays, and a mean absolute percentage error (MAPE) of approximately 6% was achieved [106]. The success of these attempts demonstrates both the necessity and the prospect of heterogeneous data fusion techniques.

4.2. Hybrid Computing and Learning Modes

With various data collection sources, different computing modes have been used to collect and process different types of data, e.g., cloud computing, mobile computing, edge computing, fog computing, etc. Different computing modes have different computation and communication capacities and how to integrate and utilize these computing modes with existing big data infrastructures remains challenging. Some studies are exploring in this direction. For example, based on IoT devices and fog computing, a low-cost vehicular traffic monitoring system was developed in [107], which collects vehicle GPS traces and uses fog devices to process the collected data and extract traffic information.

Traffic estimation and prediction problems are usually formulated as supervised problems from the perspective of machine learning. However, other learning modes can be leveraged for solving the challenges in these problems, e.g., transfer learning [108] and generative adversarial learning [109]. Transfer learning is a potential solution for the data sparsity problem in different locations. A POI-embedding mechanism was proposed in [110] to fuse human mobility data and city POI data. Furthermore, CNN and LSTM were combined to capture both spatiotemporal and geographical information, and mobility knowledge was transferred from one city to another, which was proven effective for improving the prediction performance for the target city with only limited data available.

Existing traffic estimation and prediction methods are developed with the assumption that the traffic infrastructures remain the same. This may not hold in some cases with new urban planning. A novel off-deployment traffic estimation problem was proposed and defined in [111], and a traffic generative adversarial network approach named TrafficGAN was further proposed to solve this problem, which was able to describe how the traffic patterns change with the travel demand and underlying road network structures.

4.3. Distributed Solutions and Platforms

As seen in Section 3, big data tools are often operated in a distributed manner, or designed with the real-time parallel processing capability for streaming data, which is missing in the current store and then processes paradigms for traffic estimation and prediction studies.

Some efforts for building and applying distributed solutions and platforms have been made in recent years, however, mainly in the industry, e.g., Didi Brain (https://www.didiglobal.com/science/brain) for travel-demand prediction and JD Urban Spatio-Temporal Data Engine (JUST) (http://just.urban-computing.com/) for traffic trajectory and order management. In addition to further adopting the existing tools, other new technologies may also be applied for traffic estimation, traffic prediction, or other relevant problems. For example, blockchain is used as a promising solution for traffic big data exchange trust and privacy protection [112] and federated learning is used for privacy-preserving traffic flow prediction [113].

5. Conclusions

With the development of big data, an increasing number of tools for collecting, processing, storing, and utilizing data have become available. This study focused on the tasks of traffic estimation and prediction and presents an up-to-date collection of the available datasets and tools as a reference for those who seek public resources. The collection and usage of external data are also encouraged, e.g., weather data, calendar information, emissions of air pollutants, and noise, as these provide valuable information for better estimating and predicting traffic states.

While there are multiple data sources available, the data richness varies greatly, and the data volumes are not sufficiently large with limited spatial and temporal ranges. Off-the-shelf big data tools have not been widely used in previous studies; however, there is a trend that more relevant tools would be needed with the accumulation of heterogeneous data that are beyond the abilities of traditional tools. To change this situation, challenges and future directions were indicated with the aim of promoting the application of big data in the transportation domain.

Author Contributions: Conceptualization, W.J. and J.L.; methodology, W.J. and J.L.; software, W.J. and J.L.; validation, W.J. and J.L.; formal analysis, W.J. and J.L.; investigation, W.J. and J.L.; resources, W.J. and J.L.; data curation, W.J. and J.L.; writing—original draft preparation, W.J. and J.L.; writing—review and editing, W.J. and J.L.; visualization, W.J. and J.L.; supervision, W.J. and J.L.; project administration, W.J. and J.L.; funding acquisition, W.J. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data would be available here: https://github.com/jwwthu/DL4Traffic.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cox, M.; Ellsworth, D. Application-controlled demand paging for out-of-core visualization. In Proceedings of the Visualization'97 (Cat. No. 97CB36155), Phoenix, AZ, USA, 24 October 1997; pp. 235–244.
2. The 5 V's of Big Data [Online]. 17 September 2016. Available online: https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/ (accessed on 20 December 2021).
3. Xie, J.; Reddy Kothapally, A.; Wan, Y.; He, C.; Taylor, C.; Wanke, C.; Steiner, M. Similarity search of spatiotemporal scenario data for strategic air traffic management. *J. Aerosp. Inf. Syst.* **2019**, *16*, 187–202. [CrossRef]
4. Gong, L.; Liu, X.; Wu, L.; Liu, Y. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartogr. Geogr. Inf. Sci.* **2016**, *43*, 103–114. [CrossRef]

5. He, Z.; Zheng, L.; Chen, P.; Guan, W. Mapping to cells: A simple method to extract traffic dynamics from probe vehicle data. *Comput.-Aided Civ. Infrastruct. Eng.* **2017**, *32*, 252–267. [CrossRef]
6. Aqib, M.; Mehmood, R.; Alzahrani, A.; Katib, I.; Albeshri, A.; Altowaijri, S.M. Rapid transit systems: smarter urban planning using big data, in-memory computing, deep learning, and GPUs. *Sustainability* **2019**, *11*, 2736. [CrossRef]
7. Torre-Bastida, A.I.; Del Ser, J.; Laña, I.; Ilardia, M.; Bilbao, M.N.; Campos-Cordobés, S. Big Data for transportation and mobility: recent advances, trends and challenges. *IET Intell. Transp. Syst.* **2018**, *12*, 742–755. [CrossRef]
8. Moharm, K.I.; Zidane, E.F.; El-Mahdy, M.M.; El-Tantawy, S. Big data in ITS: Concept, case studies, opportunities, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 3189–3194. [CrossRef]
9. Zhu, L.; Yu, F.R.; Wang, Y.; Ning, B.; Tang, T. Big data analytics in intelligent transportation systems: A survey. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 383–398. [CrossRef]
10. Neilson, A.; Daniel, B.; Tjandra, S. Systematic review of the literature on big data in the transportation domain: Concepts and applications. *Big Data Res.* **2019**, *17*, 35–44. [CrossRef]
11. Allström, A. Highway Traffic State Estimation and Short-Term Prediction. Ph.D. Thesis, Linköping University, Linköping, Sweden, 2016.
12. Jiang, W.; Zhang, L. Geospatial data to images: A deep-learning framework for traffic forecasting. *Tsinghua Sci. Technol.* **2018**, *24*, 52–64. [CrossRef]
13. Jiang, W.; Luo, J. Graph Neural Network for Traffic Forecasting: A Survey. *arXiv* **2021**, arXiv:2101.11174.
14. Jiang, W. Applications of deep learning in stock market prediction: Recent progress. *Expert Syst. Appl.* **2021**, *184*, 115537. [CrossRef]
15. Jiang, W. Graph-based Deep Learning for Communication Networks: A Survey. *Comput. Commun.* **2022**, *185*, 40–54. [CrossRef]
16. Jiang, W.; Lian, J.; Shen, M.; Zhang, L. A multi-period analysis of taxi drivers' behaviors based on GPS trajectories. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 1–6.
17. Jiang, W.; Zhang, L. The impact of the transportation network companies on the taxi industry: Evidence from Beijing's GPS taxi trajectory data. *IEEE Access* **2018**, *6*, 12438–12450. [CrossRef]
18. Li, L.; Jiang, R.; He, Z.; Chen, X.M.; Zhou, X. Trajectory data-based traffic flow studies: A revisit. *Transp. Res. Part C Emerg. Technol.* **2020**, *114*, 225–240. [CrossRef]
19. Moosavi, S.; Samavatian, M.H.; Parthasarathy, S.; Ramnath, R. A countrywide traffic accident dataset. *arXiv* **2019**, arXiv:1906.05409.
20. Katranji, M.; Kraiem, S.; Moalic, L.; Sanmarty, G.; Khodabandelou, G.; Caminada, A.; Selem, F.H. Deep multi-task learning for individuals origin–destination matrices estimation from census data. *Data Min. Knowl. Discov.* **2020**, *34*, 201–230. [CrossRef]
21. Li, Y.; Yu, R.; Shahabi, C.; Liu, Y. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
22. Cui, Z.; Ke, R.; Wang, Y. Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction. *arXiv* **2018**, arXiv:1801.02143.
23. Chen, X.; Chen, Y.; He, Z. Urban Traffic Speed Dataset of Guangzhou, China. 2018. Available online: https://crawdad.org/epfl/mobility/20090224/ (accessed on 25 Janurary 2022).
24. Loder, A.; Ambühl, L.; Menendez, M.; Axhausen, K.W. Understanding traffic capacity of urban networks. *Sci. Rep.* **2019**, *9*, 16283. [CrossRef]
25. Reinforcement Learning for Traffic Signal Control: Benchmark Dataset [Online]. Available online: https://traffic-signal-control.github.io/dataset.html (accessed on 20 December 2021).
26. Caceres, N.; Romero, L.M.; Benitez, F.G.; del Castillo, J.M. Traffic flow estimation models using cellular phone data. *IEEE Trans. Intell. Transp. Syst.* **2012**, *13*, 1430–1441. [CrossRef]
27. Song, Y.; Liu, Y.; Qiu, W.; Qin, Z.; Tan, C.; Yang, C.; Zhang, D. MIFF: Human Mobility Extractions with Cellular Signaling Data under Spatio-temporal Uncertainty. *ACM Interact. Mob. Wearable Ubiquitous Technol.* **2020**, *4*, 1–19. [CrossRef]
28. Chen, C.H. A cell probe-based method for vehicle speed estimation. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2020**, *103*, 265–267. [CrossRef]
29. Ji, B.; Hong, E.J. Deep-learning-based real-time road traffic prediction using long-term evolution access data. *Sensors* **2019**, *19*, 5327. [CrossRef] [PubMed]
30. Zhao, C.; Zeng, A.; Yeung, C.H. Characteristics of human mobility patterns revealed by high-frequency cell-phone position data. *EPJ Data Sci.* **2021**, *10*, 5. [CrossRef]
31. Ghahramani, M.; Zhou, M.; Wang, G. Urban sensing based on mobile phone data: Approaches, applications, and challenges. *IEEE/CAA J. Autom. Sin.* **2020**, *7*, 627–637. [CrossRef]
32. Italia, T. Telecommunications-SMS, Call, Internet-MI. 2015. Available online: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EGZHFV (accessed on 20 December 2021).
33. Harrison, G.; Grant-Muller, S.M.; Hodgson, F.C. New and emerging data forms in transportation planning and policy: Opportunities and challenges for "Track and Trace" data. *Transp. Res. Part C Emerg. Technol.* **2020**, *117*, 102672. [CrossRef]
34. Zou, Z.; Yang, H.; Zhu, A.X. Estimation of travel time based on ensemble method with multi-modality perspective urban big data. *IEEE Access* **2020**, *8*, 24819–24828. [CrossRef]

35. Wang, Y.; Xu, D.; Peng, P.; Xuan, Q.; Zhang, G. An urban commuters' OD hybrid prediction method based on big GPS data. *Chaos Interdiscip. J. Nonlinear Sci.* **2020**, *30*, 093128. [CrossRef]
36. Zheng, Y.; Yuan, J.; Xie, W.; Xie, X.; Sun, G. Drive smartly as a taxi driver. In Proceedings of the 2010 7th International Conference on Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing, Xi'an, China, 26–29 October 2010; pp. 484–486.
37. Piorkowski, M.; Sarafijanovic-Djukic, N.; Grossglauser, M. CRAWDAD Data Set Epfl/Mobility (v. 24 February 2009). 2009. Available online: https://www.semanticscholar.org/paper/CRAWDAD-dataset-epfl%2Fmobility-(v.2009-02-24)-Pi%C3%B3rkowski-Sarafijanovic-Djukic/84b87fff23a4a60586f3382279b81ff54d5eb003 (accessed on 20 December 2021)
38. Bracciale, L.; Bonola, M.; Loreti, P.; Bianchi, G.; Amici, R.; Rabuffi, A. CRAWDAD Dataset Roma/Taxi. 2014. Available online: http://crawdad.org/roma/taxi/20140717/ (accessed on 20 December 2021).
39. Zheng, Y.; Xie, X.; Ma, W.Y. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.* **2010**, *33*, 32–39.
40. Kiukkonen, N.; Blom, J.; Dousse, O.; Gatica-Perez, D.; Laurila, J. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS Berlin* **2010**, *68*, 7.
41. Huang, X.; Yin, Y.; Lim, S.; Wang, G.; Hu, B.; Varadarajan, J.; Zheng, S.; Bulusu, A.; Zimmermann, R. Grab-posisi: An extensive real-life gps trajectory dataset in southeast Asia. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility, Chicago, IL, USA, 5 November 2019; pp. 1–10. Available online: https://dl.acm.org/doi/10.1145/3356995.3364536 (accessed on 20 December 2021).
42. Zhang, J.; Zheng, Y.; Qi, D. Deep spatio-temporal residual networks for citywide crowd flows prediction. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 1655–1661.
43. Mourad, L.; Qi, H.; Shen, Y.; Yin, B. ASTIR: Spatio-Temporal Data Mining for Crowd Flow Prediction. *IEEE Access* **2019**, *7*, 175159–175165. [CrossRef]
44. Guo, F.; Zhang, D.; Dong, Y.; Guo, Z. Urban link travel speed dataset from a megacity road network. *Sci. Data* **2019**, *6*, 1–8. [CrossRef] [PubMed]
45. Wang, X.; Chen, C.; Min, Y.; He, J.; Yang, B.; Zhang, Y. Efficient metropolitan traffic prediction based on graph recurrent neural network. *arXiv* **2018**, arXiv:1811.00740.
46. Zhao, L.; Song, Y.; Zhang, C.; Liu, Y.; Wang, P.; Lin, T.; Deng, M.; Li, H. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 3848–3858. [CrossRef]
47. Mai-Tan, H.; Pham-Nguyen, H.N.; Long, N.X.; Minh, Q.T. Mining Urban Traffic Condition from Crowd-Sourced Data. *SN Comput. Sci.* **2020**, *1*, 1–16. [CrossRef]
48. Nair, D.J.; Gilles, F.; Chand, S.; Saxena, N.; Dixit, V. Characterizing multicity urban traffic conditions using crowdsourced data. *PLoS ONE* **2019**, *14*, e0212845.
49. Chen, Y.; Lv, Y.; Wang, X.; Li, L.; Wang, F.Y. Detecting traffic information from social media texts with deep learning approaches. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 3049–3058. [CrossRef]
50. Wan, X.; Lucic, M.C.; Ghazzai, H.; Massoud, Y. Empowering Real-Time Traffic Reporting Systems With NLP-Processed Social Media Data. *IEEE Open J. Intell. Transp. Syst.* **2020**, *1*, 159–175. [CrossRef]
51. Ali, F.; Ali, A.; Imran, M.; Naqvi, R.A.; Siddiqi, M.H.; Kwak, K.S. Traffic accident detection and condition analysis based on social networking data. *Accid. Anal. Prev.* **2021**, *151*, 105973. [CrossRef]
52. Wang, Y.; He, Z.; Hu, J. Traffic Information Mining From Social Media Based on the MC-LSTM-Conv Model. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 1132–1144. [CrossRef]
53. Alkouz, B.; Al Aghbari, Z. SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks. *Inf. Process. Manag.* **2020**, *57*, 102139. [CrossRef]
54. Hoseinzadeh, N.; Liu, Y.; Han, L.D.; Brakewood, C.; Mohammadnazar, A. Quality of location-based crowdsourced speed data on surface streets: A case study of Waze and Bluetooth speed data in Sevierville, TN. *Comput. Environ. Urban Syst.* **2020**, *83*, 101518. [CrossRef]
55. Liao, B.; Zhang, J.; Wu, C.; McIlwraith, D.; Chen, T.; Yang, S.; Guo, Y.; Wu, F. Deep sequence learning with auxiliary information for traffic prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 537–546.
56. Cho, E.; Myers, S.A.; Leskovec, J. Friendship and mobility: User movement in location-based social networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1082–1090.
57. Yang, D.; Zhang, D.; Yu, Z.; Wang, Z. A sentiment-enhanced personalized location recommendation system. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, Paris, France, 1–3 May 2013; pp. 119–128.
58. Cheng, X.; Zhang, R.; Zhou, J.; Xu, W. Deeptransport: Learning spatial-temporal dependency for traffic condition forecasting. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.
59. Hussain, E.; Bhaskar, A.; Chung, E. Transit OD matrix estimation using smartcard data: Recent developments and future research challenges. *Transp. Res. Part C Emerg. Technol.* **2021**, *125*, 103044. [CrossRef]

60. Zhang, T.; Li, Y.; Yang, H.; Cui, C.; Li, J.; Qiao, Q. Identifying primary public transit corridors using multi-source big transit data. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 1137–1161. [CrossRef]
61. Gong, S.; Cartlidge, J.; Bai, R.; Yue, Y.; Li, Q.; Qiu, G. Extracting activity patterns from taxi trajectory data: A two-layer framework using spatio-temporal clustering, Bayesian probability and Monte Carlo simulation. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 1210–1234. [CrossRef]
62. Li, S.; Zhuang, C.; Tan, Z.; Gao, F.; Lai, Z.; Wu, Z. Inferring the trip purposes and uncovering spatio-temporal activity patterns from dockless shared bike dataset in Shenzhen, China. *J. Transp. Geogr.* **2021**, *91*, 102974. [CrossRef]
63. Kashiyama, T.; Pang, Y.; Sekimoto, Y. Open PFLOW: Creation and evaluation of an open dataset for typical people mass movement in urban areas. *Transp. Res. Part C Emerg. Technol.* **2017**, *85*, 249–267. [CrossRef]
64. Liu, L.; Chen, J.; Wu, H.; Zhen, J.; Li, G.; Lin, L. Physical-Virtual Collaboration Modeling for Intra-and Inter-Station Metro Ridership Prediction. *IEEE Trans. Intell. Transp. Syst.* **2020**, 1–15. [CrossRef]
65. Yi, P.; Huang, F.; Peng, J. A Rebalancing Strategy for the Imbalance problem in Bike-sharing systems. *Energies* **2019**, *12*, 2578. [CrossRef]
66. Cao, M.; Huang, M.; Ma, S.; Lü, G.; Chen, M. Analysis of the spatiotemporal riding modes of dockless shared bicycles based on tensor decomposition. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 2225–2242. [CrossRef]
67. Bui, K.H.N.; Yi, H.; Cho, J. A multi-class multi-movement vehicle counting framework for traffic analysis in complex areas using cctv systems. *Energies* **2020**, *13*, 2036. [CrossRef]
68. Hao, Q.; Qin, L. The design of intelligent transportation video processing system in big data environment. *IEEE Access* **2020**, *8*, 13769–13780. [CrossRef]
69. Chen, C.; Liu, B.; Wan, S.; Qiao, P.; Pei, Q. An Edge Traffic Flow Detection Scheme Based on Deep Learning in an Intelligent Transportation System. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1840–1852. [CrossRef]
70. Leitloff, J.; Rosenbaum, D.; Kurz, F.; Meynberg, O.; Reinartz, P. An operational system for estimating road traffic information from aerial images. *Remote Sens.* **2014**, *6*, 11315–11341. [CrossRef]
71. Brkić, I.; Miler, M.; Ševrović, M.; Medak, D. An Analytical Framework for Accurate Traffic Flow Parameter Calculation from UAV Aerial Videos. *Remote Sens.* **2020**, *12*, 3844. [CrossRef]
72. Wang, X.; Ma, X.; Grimson, W.E.L. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 539–555. [CrossRef] [PubMed]
73. Fedorov, A.; Nikolskaia, K.; Ivanov, S.; Shepelev, V.; Minbaleev, A. Traffic flow estimation with data from a video surveillance camera. *J. Big Data* **2019**, *6*, 1–15. [CrossRef]
74. Zhan, X.; Li, R.; Ukkusuri, S.V. Link-based traffic state estimation and prediction for arterial networks using license-plate recognition data. *Transp. Res. Part C Emerg. Technol.* **2020**, *117*, 102660. [CrossRef]
75. Zheng, L.; Yang, J.; Chen, L.; Sun, D.; Liu, W. Dynamic spatial-temporal feature optimization with ERI big data for Short-term traffic flow prediction. *Neurocomputing* **2020**, *412*, 339–350. [CrossRef]
76. Yao, E.; Wang, X.; Yang, Y.; Pan, L.; Song, Y. Traffic Flow Estimation Based on Toll Ticket Data Considering Multitype Vehicle Impact. *J. Transp. Eng. Part A Syst.* **2021**, *147*, 04020158. [CrossRef]
77. Lee, J.; Hong, B.; Lee, K.; Jang, Y.J. A prediction model of traffic congestion using weather data. In Proceedings of the 2015 IEEE International Conference on Data Science and Data Intensive Systems, Sydney, NSW, Australia, 11–13 December 2015; pp. 81–88.
78. Dunne, S.; Ghosh, B. Weather adaptive traffic prediction using neurowavelet models. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 370–379. [CrossRef]
79. Jia, Y.; Wu, J.; Xu, M. Traffic flow prediction with rainfall impact using a deep learning method. *J. Adv. Transp.* **2017**, *2017*, 6575947. [CrossRef]
80. Li, Z.; Huang, Q.; Carbone, G.J.; Hu, F. A high performance query analytical framework for supporting data-intensive climate studies. *Comput. Environ. Urban Syst.* **2017**, *62*, 210–221. [CrossRef]
81. Bakli, M.; Sakr, M.; Soliman, T.H.A. HadoopTrajectory: A Hadoop spatiotemporal data processing extension. *J. Geogr. Syst.* **2019**, *21*, 211–235. [CrossRef]
82. MobilityDB [Online]. Available online: https://github.com/MobilityDB/MobilityDB. (accessed on 20 December 2021).
83. Zimányi, E.; Sakr, M.; Lesuisse, A. MobilityDB: A Mobility Database Based on PostgreSQL and PostGIS. *ACM Trans. Database Syst.* **2020**, *45*, 1–42. [CrossRef]
84. Wang, S.; Zhong, Y.; Wang, E. An integrated GIS platform architecture for spatiotemporal big data. *Future Gener. Comput. Syst.* **2019**, *94*, 160–172. [CrossRef]
85. Fang, Z.; Chen, L.; Gao, Y.; Pan, L.; Jensen, C.S. Dragoon: A hybrid and efficient big trajectory management system for offline and online analytics. *VLDB J.* **2021**, *30*, 287–310. [CrossRef]
86. Ding, X.; Chen, L.; Gao, Y.; Jensen, C.S.; Bao, H. Ultraman: A unified platform for big trajectory data management and analytics. *Proc. VLDB Endow.* **2018**, *11*, 787–799. [CrossRef]
87. Yao, Z.; Zhong, Y.; Liao, Q.; Wu, J.; Liu, H.; Yang, F. Understanding Human Activity and Urban Mobility Patterns From Massive Cellphone Data: Platform Design and Applications. *IEEE Intell. Transp. Syst. Mag.* **2020**, 206–219. [CrossRef]
88. Apcahe Hadoop [Online]. Available online: https://hadoop.apache.org/ (accessed on 20 December 2021).
89. Apache Pig [Online]. Available online: https://pig.apache.org/ (accessed on 20 December 2021).

90. Apache Mahout [Online]. Available online: https://mahout.apache.org (accessed on 20 December 2021).
91. Apache Spark [Online]. Available online: https://spark.apache.org/ (accessed on 20 December 2021).
92. Apache Kafka [Online]. Available online: https://kafka.apache.org/ (accessed on 20 December 2021).
93. Apache Flink [Online]. Available online: https://flink.apache.org/ (accessed on 20 December 2021).
94. Apache Storm [Online]. Available online: http://storm.apache.org/ (accessed on 20 December 2021).
95. Hadoop Distributed File System (HDFS) [Online]. Available online: https://hadoop.apache.org/hdfs/ (accessed on 20 December 2021).
96. Apache HBase [Online]. Available online: https://hbase.apache.org/ (accessed on 20 December 2021).
97. MongoDB [Online]. Available online: https://docs.mongodb.com/manual/introduction/ (accessed on 20 December 2021).
98. Apache Hive [Online]. Available online: https://hive.apache.org/ (accessed on 20 December 2021).
99. Babu, C.N.; Sure, P.; Bhuma, C.M. Sparse Bayesian Learning Assisted Approaches for Road Network Traffic State Estimation. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1733–1741. [CrossRef]
100. Liu, J.; Li, T.; Xie, P.; Du, S.; Teng, F.; Yang, X. Urban big data fusion based on deep learning: An overview. *Inf. Fusion* **2020**, *53*, 123–133. [CrossRef]
101. Zhang, J.; He, S.; Wang, W.; Zhan, F. Accuracy analysis of freeway traffic speed estimation based on the integration of cellular probe system and loop detectors. *J. Intell. Transp. Syst.* **2015**, *19*, 411–426. [CrossRef]
102. Gkountouna, O.; Pfoser, D.; Züfle, A. Traffic Flow Estimation using Probe Vehicle Data. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, Sydney, NSW, Australia, 6–9 October 2020; pp. 579–588.
103. Chen, H.; Cai, M.; Xiong, C. Research on human travel correlation for urban transport planning based on multisource data. *Sensors* **2021**, *21*, 195. [CrossRef] [PubMed]
104. Wang, P.; Lai, J.; Huang, Z.; Tan, Q.; Lin, T. Estimating traffic flow in large road networks based on multi-source traffic data. *IEEE Trans. Intell. Transp. Syst.* **2020**, 5672–5683. [CrossRef]
105. Guo, Y.; Yang, L. Reliable Estimation of Urban Link Travel Time Using Multi-Sensor Data Fusion. *Information* **2020**, *11*, 267. [CrossRef]
106. Shoman, M.; Aboah, A.; Adu-Gyamfi, Y. Deep Learning Framework for Predicting Bus Delays on Multiple Routes Using Heterogenous Datasets. *J. Big Data Anal. Transp.* **2021**, *2*, 1–16. [CrossRef]
107. Vergis, S.; Komianos, V.; Tsoumanis, G.; Tsipis, A.; Oikonomou, K. A low-cost vehicular traffic monitoring system using fog computing. *Smart Cities* **2020**, *3*, 138–156. [CrossRef]
108. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
109. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *arXiv* **2014**, arXiv:1406.2661.
110. Jiang, R.; Song, X.; Fan, Z.; Xia, T.; Wang, Z.; Chen, Q.; Cai, Z.; Shibasaki, R. Transfer Urban Human Mobility via POI Embedding over Multiple Cities. *ACM Trans. Data Sci.* **2021**, *2*, 1–26. [CrossRef]
111. Zhang, Y.; Li, Y.; Zhou, X.; Kong, X.; Luo, J. Off-Deployment Traffic Estimation—A Traffic Generative Adversarial Networks Approach. *IEEE Trans. Big Data* **2020**. [CrossRef]
112. Hassija, V.; Gupta, V.; Garg, S.; Chamola, V. Traffic jam probability estimation based on blockchain and deep neural networks. *IEEE Trans. Intell. Transp. Syst.* **2020**, 3919–3928. [CrossRef]
113. Liu, Y.; James, J.; Kang, J.; Niyato, D.; Zhang, S. Privacy-preserving traffic flow prediction: A federated learning approach. *IEEE Internet Things J.* **2020**, *7*, 7751–7763. [CrossRef]