Establishing Multisource Data-Integration Framework for Transportation Data Analytics

Zhiyong Cui, S.M.ASCE[1]; Kristian Henrickson, Ph.D.[2]; Salvatore Antonio Biancardo[3]; Ziyuan Pu, S.M.ASCE[4]; and Yinhai Wang, F.ASCE[5]

Abstract: In recent years, with the advancement in traffic sensing, data storage, and communication technologies, the availability and diversity of transportation data have increased substantially. When the volume and variety of traffic data increase dramatically, integrating multisource traffic data to conduct traffic analysis is becoming a challenging task. The heterogeneous spatiotemporal resolutions of traffic data and the lack of standard geospatial representations of multisource data are the main hurdles for solving the traffic data-integration problem. In this study, to overcome these challenges, a transportation data-integration framework based on a uniform geospatial roadway referencing layer is proposed. In the framework, on the basis of traffic sensors' locations and sensing areas, transportation-related data are classified into four categories, including on-road segment-based data, off-road segment-based data, on-road point-based data, and off-road point-based data. Four data-integration solutions are proposed accordingly. An iterative map conflation algorithm as a core component of the framework is proposed for integrating the on-road segment-based data. The overall integration performance of the four types of data and the efficiency of the iterative map conflation algorithm in terms of percentage of integrated roadway segments and computation time are analyzed. To produce efficient transportation analytics, the proposed framework is implemented on an interactive data-driven transportation analytics platform. Based on the implemented framework, several case studies of real-world transportation data analytics are presented and discussed.
DOI: 10.1061/JTEPBS.0000331.

Author keywords: Multisource transportation data; Data integration; Map conflation; Performance measurement; Travel time reliability; Transportation data analytics platform.

Introduction

In recent years, with advances in traffic sensing, data storage, and communication technologies, the availability and diversity of transportation data have increased substantially. Transportation data is normally generated by traffic sensors, such as inductive loop detectors, monitoring cameras, and Wi-Fi/Bluetooth sensors. With the help of the global position system (GPS), transportation data also can be collected by us through personal mobile computing devices and probe vehicles. Moreover, other transportation related data sources, such as weather data, traffic incident data, and vehicle emission data, also have great impact on the results of traffic analysis. Therefore, transportation data are very beneficial to public agencies and researchers for managing transportation systems and conducting traffic analysis.

In the context of modern traffic operations where highly accurate information is needed, single-source transportation data may not be sufficient. When the volume and variety of traffic data increase, using the existing multisource data to provide accurate traffic information is becoming a challenging task (El Faouzi et al. 2011). The hurdles of integrating and utilizing multisource data can be summarized as the following three aspects:

- The significant variability in the spatiotemporal resolution/granularity of multisource data;
- A lack of standard geospatial representations of traffic roadways/network for multisource data;
- A lack of a well-designed and widely accepted framework for traffic data integration.

Dealing with various spatiotemporal resolutions is one of the most challenging tasks for integrating multisource transportation data because data sets might be collected and formatted by various transportation-related practitioners for different purposes. The temporal resolution (the minimum time interval units) of traffic monitoring data may range from seconds to hours. The spatial granularity of traffic data also varies significantly. Some types of traffic data are collected to monitor arterials and urban corridors in downtown areas, and some others may be only applicable to freeways. Moreover, some data sets may overlap in terms of their spatial coverage, while other data sets may have complementary monitoring areas. These differences in the spatiotemporal resolutions normally lead to big hurdles if multiple data sets have to be incorporated into research studies and real applications. Hence, to solve this problem and facilitate further analysis and applications, a high-resolution geospatial referencing representation of a traffic network for connecting different data sources is utilized in this study.

[1]Ph.D. Candidate, Dept. of Civil and Environmental Engineering, Univ. of Washington, Seattle, WA 98195. ORCID: https://orcid.org/0000-0002-5780-4312. Email: zhiyongc@uw.edu

[2]Dept. of Civil and Environmental Engineering, Univ. of Washington, Seattle, WA 98195. Email: henr2237@uw.edu

[3]Assistant Professor, Dept. of Civil, Construction, and Environmental Engineering, Univ. of Naples Federico II, Naples 80125, Italy. ORCID: https://orcid.org/0000-0003-2567-7977. Email: salvatoreantonio.biancardo@unina.it

[4]Ph.D. Candidate, Dept. of Civil and Environmental Engineering, Univ. of Washington, Seattle, WA 98195. ORCID: https://orcid.org/0000-0002-9488-9175. Email: ziyuanpu@uw.edu

[5]Professor, Dept. of Civil and Environmental Engineering, Univ. of Washington, Seattle, WA 98195 (corresponding author). Email: yinhai@uw.edu

Note. This manuscript was submitted on February 25, 2019; approved on September 16, 2019; published online on February 19, 2020. Discussion period open until July 19, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Transportation Engineering, Part A: Systems***, © ASCE, ISSN 2473-2907.

The second challenge is how to utilize one standard data form or structure to represent multiple data with different geospatial representations. Using roadway geometric referencing layer is a normal option, but sometimes standard geometric referencing layers with correct and complete roadway geometric information are inaccessible or even nonexistent, especially for some urban areas. The OpenStreetMap (OSM) (Haklay and Weber 2008) provides a comprehensive worldwide map and detailed information on roadway networks. Although OSM roadway layers are editable with adequate roadway metadata, OSM roadway layers are still different from the roadway layers of the National Performance Management Research Data Set (NPMRDS) (Kaushik et al. 2015), which is used for supporting national highway system performance measurement and management activities. At the current stage, acquiring standard geometric representations or geographic information system (GIS) map layers for universal transportation analysis is difficult. If multiple data sets with different geospatial representations are used, a process of map conflation is normally needed to combine geographic information from overlapping sources so as to retain accurate data, minimize redundancy, and reconcile data conflicts (Longley et al. 2005). In a geospatial map layer, roadways are segmented into small pieces. Similarly, areas of traffic sensors can also be split into small areas according to the locations of those sensors. If the spatial resolution of traffic data is higher than that of the map layer, the original information in the traffic data will possibly be lost during the data-integration process. Therefore, a high-resolution geospatial map layer is critical for data integration. In this study, a map conflation algorithm based on a uniform high-resolution map layer is proposed to integrate transportation data.

Further, modern transportation analysis and management are in great need of extensible transportation data–integration algorithms and tools. Data integration refers to the combination of data residing at different sources and providing a unified view of these data to users (Lenzerini 2002). Designing data-integration systems is an important procedure in a variety of real-world applications, especially in intelligent transportation systems (ITSs). Other conventional problems in transportation modeling are also concerned with multisource processing, like planning problems, demand estimation, and traffic estimation (El Faouzi et al. 2011). Transportation data-integration frameworks and tools have been designed and implemented for multiple scenarios. Most of the existing studies integrate transportation data only when data integration is needed. However, data integration is an important topic for transportation-related research and practice. A review of the literature turned up few studies that proposed frameworks for multicategory transportation data integration. Thus, in this study, the main target is to design an extensible framework for integrating multisource transportation data to support further analysis and applications.

In this study, a transportation data-integration framework based on a uniform roadway referencing layer is proposed to support multisource data-based traffic analysis. According to data source characteristics, all transportation-related data sets are categorized into four types in terms of the sensor's location and sensing area. To make the transportation data-integration framework more efficient and extensible, an iterative map conflation algorithm as a component of the framework is proposed. In addition, the proposed traffic data-integration framework is implemented on an interactive online transportation analytics platform, the Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net) system (Cui et al. 2016; Ma et al. 2011). Several real-world transportation data analytics functions, including travel time analysis, freeway performance measurement, and traffic safety analysis, are implemented based on the proposed framework and discussed.

The originality and contribution of this work can be summarized as follows:

1. An extensible traffic data-integration framework is proposed based on a uniform roadway referencing layer. The proposed extensible framework support to integrate new potential data sources.
2. On the basis of traffic sensor locations and sensing areas, traffic data are classified into four categories: on-road segment-based data, off-road segment-based data, on-road point-based data, and off-road point-based data, and four data-integration solutions are proposed accordingly.
3. An iterative map conflation (IMC) algorithm for the on-road segment-based traffic data is proposed as the core component of the proposed data-integration framework.
4. The proposed framework is implemented on a data-driven transportation analytics platform. Several interactive analysis tools and case studies based on this framework are implemented and introduced.

The rest of the paper is organized as follows. The next section discusses relevant previous studies and applications. Then the data-integration framework is introduced in detail. Next, the experimental results are presented, followed by the introduction of several real applications and case studies based on the framework. The last section provides the discussion and conclusions of this study.

Literature Review

Technological advances in traffic sensing, telecommunications, data storage, and data processing have facilitated improvements in existing means of traffic data collection. To monitor traffic status and conduct traffic analysis, multiple types of traffic data have been collected by public agencies and private companies, including speeds, travel times, incident data, construction and work zone data, and event data (List et al. 2014). Speed or travel time data are normally generated by various traffic sensors, including global positioning satellite (GPS)-enabled fleets, mobile phones, Bluetooth/Wi-Fi devices, highway-embedded sensors, and video monitoring cameras (List et al. 2014). Other types of transporation-related data, like incident data and weather data, are normally collected by public agencies.

Since collected transportation data are normally in various formats, the raw data sets mostly need to be cleaned and processed before they can be used for further analysis. If there are missing values in these data sets, data imputation procedures, such as interpolation methods (Chow 2016), predictive matching methods (Henrickson et al. 2015), fuzzy theory–based methods (Tang et al. 2015), and deep learning–based methods (Duan et al. 2016), are also needed. Transportation data have been widely used in the fields of dynamic routing (Liu and Qu 2016), anormaly event detection (Zheng et al. 2018), traffic prediction (Cui et al. 2018), and transportation mode identification (Jahangiri and Rakha 2015), for example. Some existing studies were conducted using only one type of data. However, other types of traffic analysis studies, such as safety analysis (Thakali et al. 2016) and travel time reliability analysis (Zegeer et al. 2013), need multiple data sources.

To combine heterogeneous transportation data sources, different traffic data-integration technologies have been developed for specific applications. To visually manipulate and analyze urban traffic data, a framework for integrating different spatial and temporal levels of granularity was proposed (Claramunt et al. 2000). To estimate the traffic state on freeways and signalized arterial links, loop detector data and probe vehicle data were integrated (Valadkhani et al. 2017). According to a survey about data fusion

in ITS, data fusion approaches can be split into three groups: statistical, probabilistic, and artificial intelligence approaches (Fourati 2015). Algorithms and architectures designed for multisensory data fusion and integration are normally developed for some specific data sources. For example, the integration problem of in-vehicle information and loop detector data has been studied (Cremer and Schrieber 1996). To transfer data between two separate geographic information databases, Graettinger et al. (2009) proposed a linear referencing approach to combine information from the state route linear referencing system (LRS) with information from the local road LRS. The US DOT has also provided potential use cases for integrating emerging data sources into operational practice (Gettman et al. 2017). However, since most of the existing studies integrated a few types of data sources, it is valuable to design a comprehensive data-integration framework that supports integrating the majority of data sources.

Because transportation data normally describe states of roadway segments or traffic networks, the characteristics of roadways, such as geospatial information, play an important role in the traffic data-integration process. Thus, multiple data-integration methods are designed based on roadway map conflation or GIS data combination. Following this strategy, a split-match-aggregate algorithm is proposed to integrate sidewalk data with transportation network data in GIS (Kang et al. 2015). Another study integrated a global position system and geographical information system for traffic congestion analysis (Taylor et al. 2000). Further, a conflation methodology for two roadway networks is proposed by taking advantage of GIS capabilities in the projection of transportation networks (Daneshgar et al. 2018). To create an integrated bike map, an optimized four-step geographic data conflation system is proposed to conflate data from an authoritative source (Los Angeles County Metropolitan Transportation Authority) and an open source (OpenStreetMap) (Li and Valdovinos 2017). However, most of these existing studies integrated a few types of transportation data but did not propose an extensible transportation data framework. The framework proposed in this study has high efficiency and adaptability so that it can accomplish data integration for various new data sources.

Methodology

Problem Statement

In this study, data integration refers to the integration of multiple spatiotemporal transportation data sets. The data-integration process contains two aspects: spatial integration and temporal integration. A set of n spatiotemporal transportation data sets that need to be integrated can be denoted by $\mathcal{X} = \{X_1, \ldots, X_n\}$. The spatiotemporal values in each data set is represented by $X_i \in \mathbb{R}^{T_i \times D_i}$, in which T_i and D_i are the temporal and spatial dimensions, respectively, of the ith data set. In addition, each data set X_i is associated with a geospatial/geometric representation of its corresponding roadway network, denoted by G_i. A uniform referencing layer covering the traffic networks of all the data sets is denoted by G_0. Then, given $\mathcal{X}$, $\mathcal{G} = \{G_1, \ldots, G_n\}$, and G_0, the geospatial integration process can be defined as a function F such that

$$F(\{(X_1, G_1), \ldots, (X_n, G_n)\}; G_0) = (\{(X_1, G_1, L_1), \ldots, (X_n, G_n, L_n)\}; G_0) \quad (1)$$

where L_i is generated linkage information that builds a link between G_i and G_0. The core of the spatial integration is to find the best linkage data set $\mathcal{L} = \{L_1, \ldots, L_n\}$. As for the temporal integration process, although transportation data sets may have different temporal resolutions, the data integration can be easily carried out using the minimum temporal resolutions among all data sets as the temporal resolution of the integrated data. Then the temporal resolution of the integrated data can be aggregated to a proper scale for further analysis, if needed. Thus, the key problem of the transportation data integration is the geospatial data-integration process.

Data Categorization

To build a general traffic data-integration framework, it is efficient to classify all the data sets and design the integration framework for each category, in which multiple data sets can be processed in a similar way. Transportation-related data are normally generated by location-fixed sensors, mobile devices, probe vehicles, or historical records reported by roadway management agencies. To describe the proposed data-integration framework, several representative types of data used in this study are introduced in detail in the experimental section, and their brief introductions are described as following:

- Location-fixed sensor data: loop detector–based traffic speed/volume data, Wi-Fi/Bluetooth-based travel time data, cellular mobile phone station–based traffic speed data, and others;
- Probe vehicle–based data: Google real-time traffic data, road segment–based travel time data, road segment–based traffic speed data, and others;
- Incident data: traffic accident data, construction or alert data, and others;
- Weather information data.

In this study, to describe the proposed data-integration framework in a concise and efficient way, several representative data sets are adopted. The main properties of these representative data sets, including covered area, sensing area, sensor location, estimated information, and time interval, are summarized and presented in Table 1. The sensing area describes whether a sensor monitors

Table 1. Data summarization and comparison

Data source	Collection method/tool	Covered area	Sensing area	Sensor location	Estimated information	Time interval
Loop data	Inductive loop detectors	Freeway main lanes and ramps	Point/milepost on roadway	On-road	• Speed • Volume • Occupancy	5 min
NPMRDS data	Probe vehicles	Freeway and urban corridors	Segment on roadway	On-road	• Speed (INRIX) • Travel time (HERE)	5 min
Verizon data	Mobile phone cellular towers	Freeway and Arterials	Segment	Off-road	• Speed	2.5 min
Incident data	Reported by police and agencies	Freeways	Point/milepost	On-road	• Incident info	flexible
Weather data	Weather stations	All roadways	Point/milepost	Off-road	• Weather info	5 min

J. Transp. Eng., Part A: Systems, 2020, 146(5): 04020024

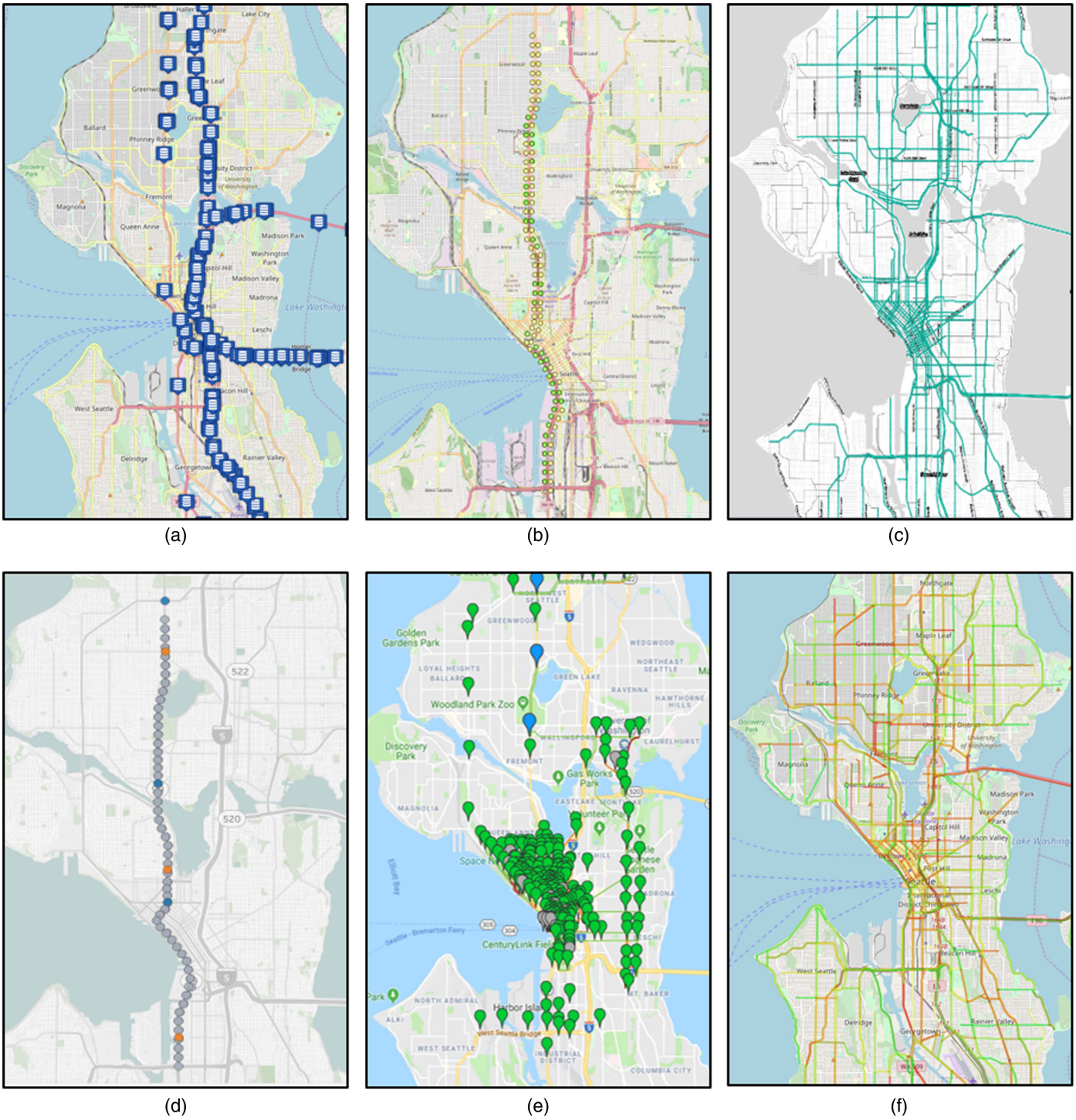

Fig. 1. Demonstration of several data sources on real map in Seattle area: (a) loop detector data; blue icons: sensor locations; (b) Verizon cellular data; colored circles: virtual sensor stations; (c) HERE road segment-based data; (d) camera-based traffic data; colored circles: camera locations; (e) bluetooth/Wi-Fi data; drops: sensor locations; and (f) INRIX road segment-based data. (Map data © OpenStreetMap contributors.)

traffic for a point or a segment of a roadway. The sensor location indicates whether a sensor is deployed on a roadway (on-road) or not (off-road). The time interval indicates the minimum or suitable time interval of a data source for analysis. Further, to demonstrate the various sensor coverage areas of these data sources, sample data sets in the Seattle area are illustrated on the real map as shown in Fig. 1. INRIX (Schrank et al. 2014) and HERE (Rafferty and Hankley 2014) are two leading companies in the field of managing traffic by analyzing data and providing traffic data and services. The INRIX data and HERE traffic data are road segment–based traffic data.

According to the Strategic Highway Research Program 2 (SHRP2) L02 product report (List et al. 2014), transportation data can be classified based on multiple principles, like data collection methods, types of information disseminated, and methods of information dissemination. However, based on a comparison, as shown in Table 1, geospatial attributes, i.e., sensor location and sensing area, are the most significant ones that distinguish the studied data

Table 2. Data categorization based on sensor location and sensing area

Sensing area and sensor location	Segment-based	Point-based
On-road	• NPMRDS data (HERE and INRIX) • Real-time traffic data, like Google Traffic	• Loop detector data • Incident data • Construction/event data • Surveillance camera data
Off-road	• Verizon speed data • Pedestrian/cycle path data	• Weather station data • Bluetooth/Wi-Fi data

sources from each other. Thus, the sensor location and the sensing area are chosen as the two main metrics to group data sources into four categories, as shown in Table 2. Given these two metrics, not only can the data sources used in this study be well classified, but other types of transportation data can also be categorized for data integration.

It should be noted that GPS-based vehicle trajectory data can also be classified into on-road point-based data and integrated via similar methods. However, as a type of raw data for traffic performance measurement, trajectory data have different storage structures and analysis methodologies compared to other on-road point-based data. Thus, vehicle trajectory data are not included or discussed in this paper. In addition, some data sources may not be perfectly classified by these two metrics, considering that some sensors can be deployed on road or off road, like Bluetooth and Wi-Fi sensors.

Data-Integration Framework

A well-designed traffic data-integration framework should work effectively and efficiently for all types of data. To fulfill this requirement, a traffic data-integration framework based on a uniform roadway referencing layer is proposed. With this framework, only a roadway referencing layer and some linking layers/tables that link data sources to the referencing layer are needed for further analysis. In this way, all the data sources can retain their original data formats. Fig. 2 shows the traffic integration framework. Each data source contains its own roadway layer or geospatial information, and each data set connects to the roadway referencing layer via a linking table, acting as the linkage data, L_i, as shown in Eq. (1). To make terminologies clear in this study, the geometric layer of a data source that needs to be conflated is called the *conflated layer*, according to Chen et al. (2006). The *referencing layer* acts as an interface by which the analysis tool does not need to care about the geospatial information of each data set. When analyzing roadway performance, the analysis tool only needs to specify the analysis zone on the roadway referencing layer.

Because most state agencies focus on freeway performance measurement, and the data acquired for this study mostly measure freeway traffic conditions, the authors selected the freeway system in Washington state as the study area. In this study, the Washington State DOT (WSDOT) 24 k map layer (WSDOT 2018) was selected as the referencing layer. The 24 k map contains a routable freeway network, and each roadway segment contains multiple pieces of critical information for data integration, for example, segment ID, direction, route name, route ID, segment length, starting milepost, ending milepost, starting point ID, ending point ID, and segment geometric information.

The data-integration framework has the capability of integrating the four types of data, i.e., on-road point-based, on-road segment-based, off-road point-based, and off-road segment-based data. The detailed integration algorithms are introduced as follows.

On-Road Segment-Based Data

On-road segment-based data normally have their own roadway map layer, like INRIX and HERE data have their own traffic message channel (TMC)-based roadway layers. Because the definitions of segments in the referencing layer and the roadway layer of on-road segment-based data are totally different, these layers should be conflated for further data integration and analysis. To this end, an iterative map conflation algorithm for on-road segment-based roadway

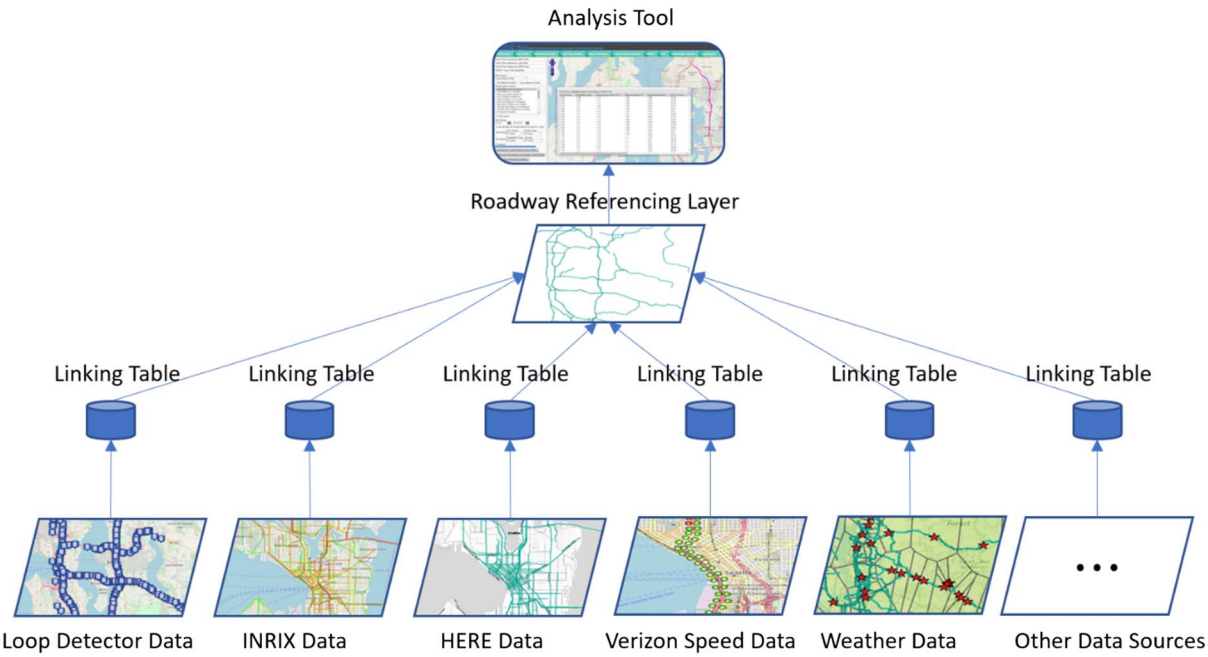

Fig. 2. Proposed traffic data-integration framework based on uniform roadway referencing layer. (Map data © OpenStreetMap contributors.)

J. Transp. Eng., Part A: Systems, 2020, 146(5): 04020024

layers is proposed. In this section, the HERE data are treated as representative on-road segment-based data, and the HERE map layer is treated as the conflated layer, for example, to demonstrate the details of the map conflation algorithm.

General Map Conflation Process. Fig. 3 shows the spatial structure of the referencing layer and the conflated layer on a real map. It is obvious that the referencing layer mainly contains freeways and the conflated layer covers more corridors, especially arterials and urban streets. The goal of the map conflation algorithm is to build a linking layer or generate a table of linkage data to connect the referencing and conflated layer.

The basic map conflation process is demonstrated in Fig. 4 and briefly summarized in the following steps:

1. Each roadway segment in the referencing layer with a referencing segment ID (RID) is broken into small segments (around 30 m). Each newly generated small segment shares the same RID. It should be noted that the length of those generated small segments can be set as the length of the shortest segment in both the referencing and conflated layers. However, if the length of the shortest segment is too small, the length of the small segments can be defined based on the properties of the specific integrated data sources, such as the traffic sensing mechanisms and the locations of the traffic sensing devices.
2. For each newly generated small segment, search and find the nearest segment with a conflation ID (CID) in the conflated layer to match based on several geometrical parameters. In this case, each newly generated small segment has the same CID.
3. Aggregate a group of newly generated small segments with the same CID and RID into a new link.
4. Create an entry for each new link in the linkage table, which contains the CID and RID information.

Iterative Map Conflation Algorithm. The search-and-match step in the map conflation process is one of the core steps. Several important thresholds are utilized to find the nearest roadway segment pairs between the referencing and conflated layers. These thresholds include the distance between the centroids of two segments in the referencing and conflated layers, D, and the angle between two segments in the referencing and conflated layers, A.

In addition, other factors also need to be considered, like lane type and the definitions of the roadway segment directions in different data sets. Fig. 5 demonstrates the thresholds and several search-and-match scenarios. During the matching process, a pair of nearest segments is selected from the referencing layer and the conflated layer, given the two thresholds D and A. According to the matching scenarios, as shown in Figs. 5(a–d), the D and A thresholds need to vary according to the scenario if an efficient search-and-match process is required.

Algorithm 1. Iterative Map Conflation Algorithm

1: Let $\mathcal{R} = \{R_0, \ldots, R_i, \ldots, R_M\}$ denote the set of all roadways segments in the referencing layer, where R_i is the RID of segment i.
2: Let $\mathcal{C} = \{C_0, \ldots, C_j, \ldots, C_N\}$ denote the set of all roadways segments in the conflated layer, where C_j is the CID of segment j.
3: for $i \leftarrow 0$ to M do
4: Split R_i into a set $R_i = \{R_i^0, \ldots, R_i^k, \ldots, R_i^K\}$ containing all newly generated segments
5: Initialize thresholds, distance D and angle A
6: while R_i is not null do
7: $K \leftarrow$ size of R_i
8: for $k \leftarrow 0$ to K do
9: Search and Match R_i^k from $\mathcal{C}$ based on thresholds (D, A)
10: Find the nearest pair (R_i^k, C_p) such that
11: $\text{Dist}(R_i^k, C_p) \le D$ and $\text{Angle}(R_i^k, C_p) \le A$
12: $\forall\, C_j \in \mathcal{C},\ \text{Dist}(R_i^k, C_p) \le \text{Dist}(R_i^k, C_j)$ and $\text{Angle}(R_i^k, C_p) \le \text{Angle}(R_i^k, C_j)$
13: $R_i.\text{pop}(R_i^k)$
14: end for
15: $(D, A) \leftarrow \text{Increase}(D, A)$
16: end while
17: end for

Thus, an iterative map conflation algorithm is proposed that can iteratively adjust the thresholds, D and A, to carry out the map conflation until all segments are conflated. The detailed algorithm is Algorithm 1. The thresholds, D and A, are initialized as the same for each segment in the referencing layer. If a newly generated segment cannot find the nearest segment in the conflated layer under this threshold pair, (D, A), the values of D and A will increase via the increase$(\cdot)$ function, shown in Line 15 in Algorithm 1, until all segments are paired. The increase$(\cdot)$ function can be a linearly increasing function, an exponentially increasing function, or other customized increasing function. In addition, the increase$(\cdot)$ function can let the values of D and A increase simultaneously or separately. Algorithm 1 is implemented via structured query language (SQL) scripts under the PostgreSQL environment. The angle, distance, and some other geospatial related parameters are calculated via the Postages tool. After running the map conflation algorithm, the linking layer, which connected the referencing layer and the conflated layer, will be automatically generated.

Off-Road Segment-Based Data

The off-road segment-based data are similar to the on-road segment-based data because they all measure traffic states of roadway segments. However, because the sensors are deployed off-road, integrating off-road segment-based data need to map sensors to roadways and utilize the proposed iterative map conflation algorithm. In this section, Verizon speed data are taken as an example to show the off-road sensor mapping process.

Although the Verizon speed data are collected via cellular towers, the Verizon data provider defines multiple virtual traffic measuring segments on roadways between towers to acquire speed data with higher precision and spatial resolution. However, the starting and ending points of these virtual segments are not perfectly located

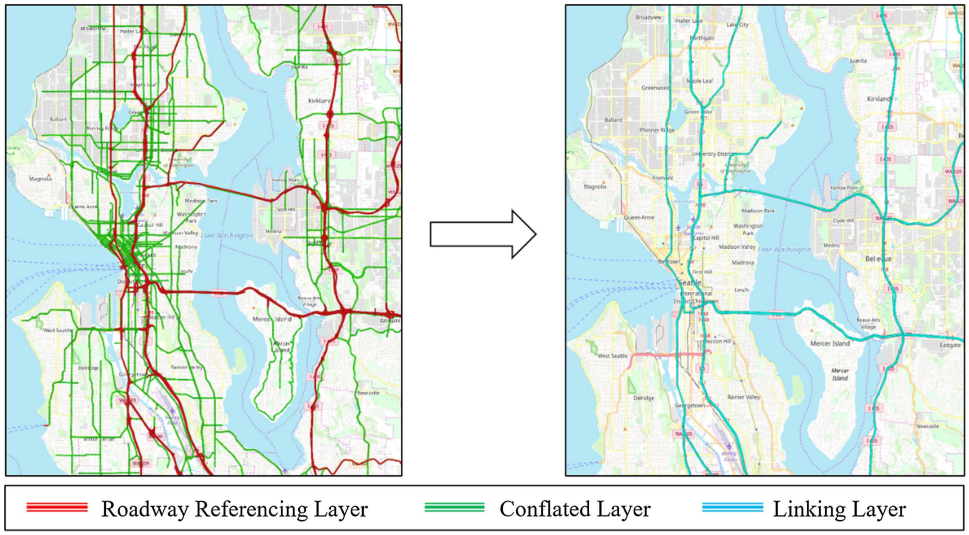

Fig. 3. Roadway referencing layer and conflated layer on real map. The goal of the map conflation algorithm is to build a linking layer. (Map data © OpenStreetMap contributors.)

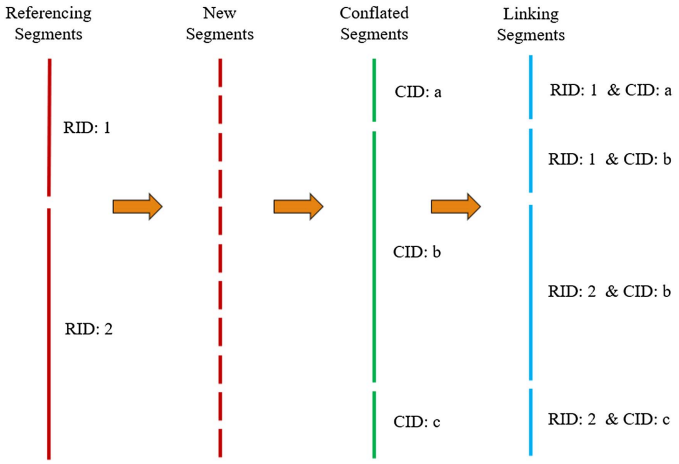

Fig. 4. Demo of map conflation process. RID = referencing segment ID; and CID = conflated segment ID.

on the roadway segments of the referencing layer. Thus, a mapping method is proposed and demonstrated in Fig. 6. The nearest point on the referencing layer is found for each of the virtual segment starting/ending points. These nearest points on the referencing layer are treated as new starting/ending points, and the new segments between these points are treated as new traffic monitoring segments. In the next step, these new segments can be integrated into the referencing layer using the proposed iterative map conflation algorithm. The process of finding the nearest point on a link/segment is carried out using the PostGIS versin 2.4 tool (Ramsey 2005).

On-Road Point-Based Data

The on-road point-based traffic data are the easiest type of data for data integration because they contain linear referencing information, i.e., the route number and mileposts of sensors. Such data, such as loop detector data and incident data, do not need their own geometric roadway layers to be conflated. Since the referencing layer contains milepost information of each roadway segment, by matching the milepost information, each on-road point-based sensor can be accurately located on the referencing layer. Hence, the integration of the on-road point-based traffic data is inherently completed using the route number and milepost information.

Off-Road Point-Based Data

As representative off-road point-based data, weather data are critical for roadway performance measurement, so they need to be integrated into the referencing layer. In this section, weather data are used an example to show the integration method for off-road point-based traffic data.

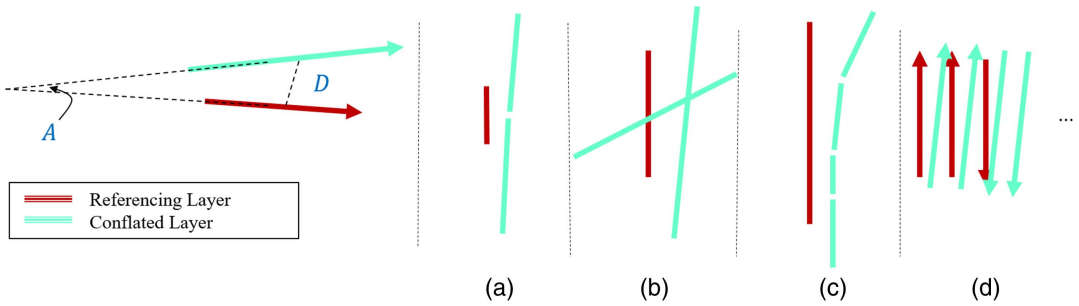

Fig. 5. Thresholds for searching-and-matching step in map conflation process. The subfigures show several matching scenarios that require different distance (d) and angle (a) thresholds.

J. Transp. Eng., Part A: Systems, 2020, 146(5): 04020024

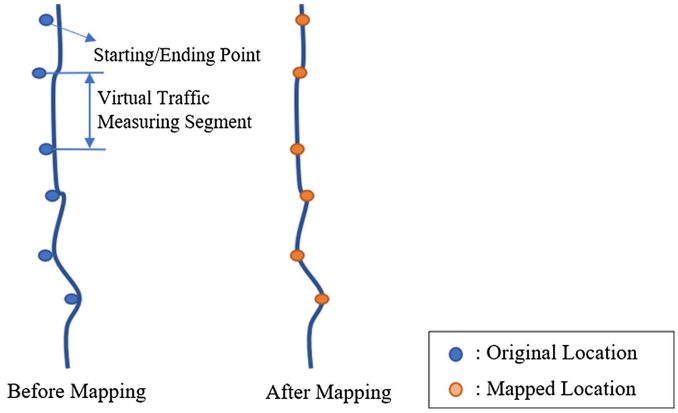

Fig. 6. Demo of off-road segment-based traffic data integration, taking Verizon speed data on SR-99 in Seattle as an example.

Weather data contain the coordinates of weather stations. However, weather stations are not perfectly located on or near freeways. In addition, the sizes of areas covered by weather stations vary. Because weather data cannot provide detailed coverage information, a mapping mechanism for data integration is proposed that maps each roadway segment to its nearest weather station.

To find the nearest weather station of a roadway segment, a spatial partitioning method based on the Voronoi diagram (Okabe et al. 2009) is adopted in this framework. All of Washington state is separated into Voronoi polygons, which are centered at the weather stations. Fig. 7 shows a map with all weather station–centered polygons and the WSDOT 24 k map layer. The edges between polygons are the intersected midlines between weather station pairs. In this case, the polygons perfectly partition the whole map, and each roadway segment in the referencing layer is covered by a polygon centered at the weather station. The red stars indicate the locations of the weather stations. In this way, each roadway segment in the referencing layer is covered by at least one polygon. If a roadway segment is covered by only one polygon, then the segment is mapped to the weather station at the center of this polygon. Otherwise, the segment will be mapped to a weather station whose corresponding polygon covers most of the roadway segment. The spatial partitioning and mapping process is carried out using the PostGIS tool.

Framework Summarization

Because the four types of data have different characteristics, the proposed traffic data-integration framework provides different solutions for them. Fig. 8 illustrates the proposed traffic data framework based on a uniform roadway referencing layer and the solutions for the four types of data. Compared to existing data-integration frameworks (Chen et al. 2006; Graettinger et al. 2009; Green et al. 2013; Kang et al. 2015; Luk and Yang 2003; Ma and Wang 2014; Memarian et al. 2018; Tarko and Rouphail 1997; Valadkhani et al. 2017), the proposed traffic data-integration framework has several advantages, as follows:

1. This framework is flexible in that it provides data-integration solutions to four types of traffic data and most of the transportation data can be categorized into those four types.
2. This framework is efficient because it only needs a uniform roadway referencing layer to manipulate all types of transportation data. The integration process is also very efficient in terms of the percentage of automatically integrated road segments and computation time.
3. This framework is extensible in that it can be applied to any other transportation data-integration tasks.

Experimental Results

Data Set Description

The data sets used in this study contain freeway loop data, NPMRDS data, Verizon speed data, incident data, and weather data. All the data were collected or provided by WSDOT and Seattle DOT. The selected study area mainly focuses on the freeway system in Washington state. The year 2015, which is covered by all data sets, is selected as the study period.

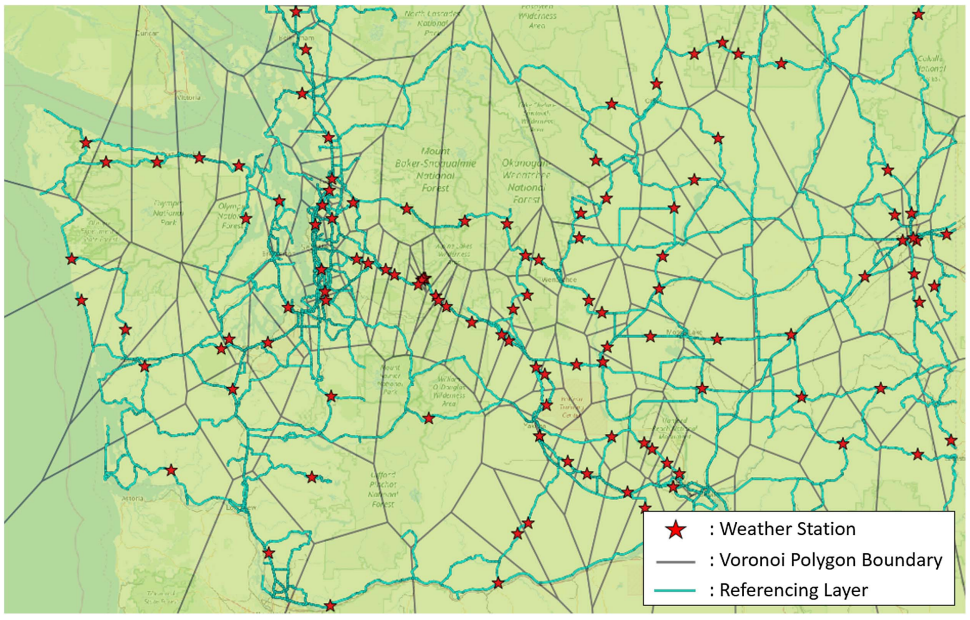

Fig. 7. Demo of weather station integration–based Voronoi map. (Map data © OpenStreetMap contributors.)

J. Transp. Eng., Part A: Systems, 2020, 146(5): 04020024

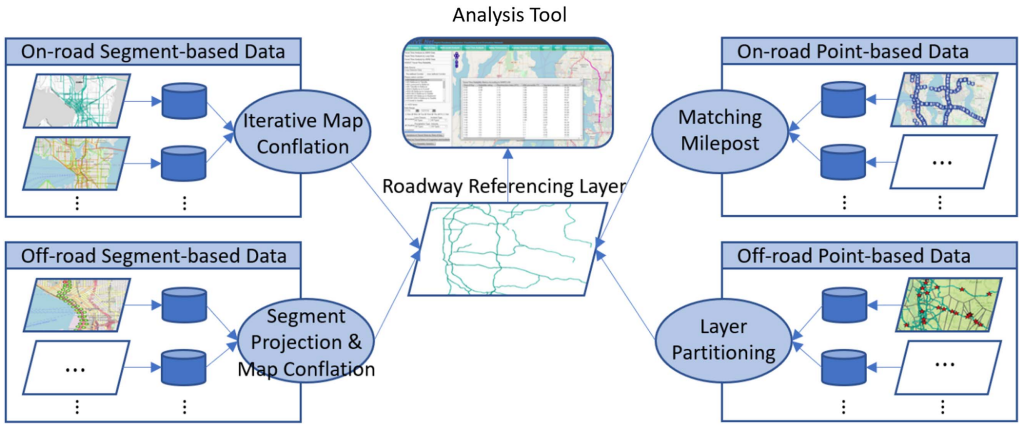

Fig. 8. Proposed traffic data-integration framework based on uniform roadway referencing layer for four types of traffic data. (Map data © OpenStreetMap contributors.)

The proposed framework's ability to integrate data sources is not restricted to the aforementioned data sets. However, the sensing areas of some other data sources, like Bluetooth/Wi-Fi data and surveillance camera data, are quite limited in that they do not cover most of the freeway system. Thus, they are not used in this study. Detailed information on each data source is described in the following subsection.

Freeway Loop Data

Inductive loop detectors are widely used to monitor freeway performance in the United States because of their reliability and durability (Klein et al. 2006). An inductive loop detector is a conductive coil embedded in pavement that detects moving vehicles passing over it with electromagnetics. Speed, volume, and occupancy are three key indicators that traffic detectors can collect during a fixed time interval.

NPMRDS Data

NPMRDS data are used by states to monitor system performance. INRIX was selected as the current NPMRDS data provider since 2017; HERE was the previous provider. INRIX combines multiple data sources, including GPS-equipped devices and cell phones and aggregate heterogeneous speed data based on a series of statistical models. INRIX data cover almost the entire roadway network in Washington, including freeways, highways, and most arterials and side streets. INRIX has adopted the TMC code, which is used to identify a specific road segment. The INRIX speed data were aggregated into 5-min intervals.

Like INRIX data, HERE data combine data sources from multiple categories, including phone and auto GPS navigation devices. HERE data are collected separately from trucks and other vehicles, thereby making it possible to obtain data for both trucks and passenger vehicles. HERE also adopted the TMC as its base network, but the TMC network used by HERE is slightly different from that of INRIX. For each TMC, instead of providing speed data, HERE provides travel time data. The time interval of HERE data is also 5 min.

Verizon Speed Data

Verizon developed an innovative technology to estimate traffic speed using cellular communication signals between cellular towers and mobile phones. The mechanism of acquiring street-level average speed values adopted by Verizon can be briefly described in the following three steps:

- First, communication messages between mobile devices carried by a probe vehicle and surrounding cellular towers are recorded. At the same time, the speed values of the probe vehicle are also archived.
- Then the attributes of the recorded communication messages and the corresponding archived vehicle speed values on specific roadways are matched. The matched communication messages are defined as "signatures" of moving cellular devices. After enough signatures are collected and validated, a model for estimating vehicle speed can be built based on historical signatures.
- Finally, when vehicles carrying Verizon mobile devices traverse a specific roadway surrounded by cellular towers, the live signature sequences received by the cellular towers can be matched with the historical signatures. The vehicle speed information can be estimated using the developed model.

Verizon also defines multiple virtual sensor stations on each road and provides average speed information between consecutive stations, as shown by the green and yellow dots in Fig. 1(b). In this study, the Verizon data provide speed information between virtual stations during a fixed time interval (2.5 min).

Incident Data

Washington State's Incident Response (IR) team collects and maintains traffic incident data in the Washington Incident Tracking System. Incident data include most incidents that happen on freeways and Washington state highways. For each incident, the Washington State IR team logs details such as incident location (route and milepost), notification time, clear time, and closure lanes.

Weather Data

Weather data are retrieved from WSDOT. The retrieved raw data include information about precipitation type, precipitation intensity, and precipitation start and end times. Each weather station is associated with a latitude and longitude pair. In this case, weather data can be visualized in a mapping system.

Experimental Settings

In this study, all the spatiotemporal data sets characterizing roadway traffic properties are stored in a Microsoft SQL server database on a Windows Server 2012 after preprocessing. Their associated roadway geometric data sets are stored in a PostgreSQL database on a Linux server. The spatial database extender, PostGIS, is used to support geospatial queries during the data-integration process.

Table 3. Data-integration performance of the four categories of data

Data type	Data set	Data scale	Percentage of integrated data	Running time
On-road segment-based	HERE data	36,522 segments	95.4	18.2 min
Off-road segment-based	Verizon data	93 segments	100	2.3 s
On-road point-based	Incident data	51,117 incidents	100	Automatically
Off-road point-based	Weather data	124 stations	100	0.2 s

All the experiments are tested on a Windows 10 computer with 32 GB memory.

In this study, the WSDOT 24 k map is selected as the referencing layer for the whole data-integration framework. The experimental settings of the framework's four components are listed as follows:

On-Road Segment-Based Data Integration

The HERE data are treated as representative on-road segment-based data, and the HERE map layer is treated as the conflated layer. The proposed IMC algorithm for on-road segment-based data integration is implemented mainly using Python, SQL, and PostGIS. In the map conflation process, 10,482 referencing map segments and 28,007 HERE segments are utilized for the segment matching process. The initial distance thresholds D of the IMC algorithm is set at $D = 0.0005$, which is around 40 m. The unit of D is a spatial reference identifier (SRID) equaling 4,326 in the PostGIS environment, and A is initialized as 0.25 rad. A linearly increasing function, Increase(D, A), which contains two steps, is defined as

$$\begin{aligned} &(1)\ A = A + \Delta a, \quad \textbf{if } A \le 0.65 \\ &(2)\ D = D + \Delta d,\ A = 0.25, \quad \textbf{if } A > 0.65 \end{aligned} \tag{2}$$

where $\Delta a = 0.2$ and $\Delta d = 0.0001$.

Off-Road Segment-Based Data Integration

The experimental settings are identical to those of on-road segment-based data integration.

On-Road Point-Based Data Integration

Since this type of data containing linear referencing information is inherently integrated into the referencing layer, no setting is needed.

Off-Road Point-Based Data Integration

The proposed Voronoi polygon method is implemented mainly based on SQL and PostGIS.

Data-Integration Performance

The integration performance of the four types of data are shown in Table 3. Except for the on-road segment-based data, all other types of data are 100% integrated. The on-road segment-based data-integration method also performs very well in that more than 95% of the segments overlapping with the referencing layer were automatically integrated. The running time of the on-road segment-based data integration was around 18 min. Considering that processing and integrating all these data sets may take days or weeks in practice, the integration of all four types of data is very fast. Although the incident data contain more than 50,000 records, the on-road point-based data require no integration method because they contain the referencing information, including route number and milepost. Because the integrations of both on-road and off-road segment-based data are designed based on the proposed IMC algorithm, that algorithm is analyzed in the next section taking the on-road segment-based data as an example. In summary, the map conflation algorithm works well in terms of time consumption and conflation accuracy for the roadway segments in the whole state.

Analysis of Map Conflation Algorithm

In this section, the proposed map conflation algorithm is analyzed taking the off-road segment-based data (HERE data) as an example. Fig. 9 shows the distributions of the length of roadway segments of the conflated and referencing layers. The trends of the two subfigures are both similar to a Poisson distribution. However, the number of segments in the conflated layer far exceeds that in the referencing layer, since the HERE data cover urban corridors and the referencing layer only covers freeways.

In this study, the efficiency of the conflation algorithm is measured in terms of percentage of integrated segments and computation time. There are two influencing factors as described in the methodology section: angle and distance thresholds between conflated segments and referencing segments. Fig. 10(a) shows the integration performance when the distance threshold is fixed at 40 m and the angle threshold increases from 5° to 75°. The integrated percentage increases slightly and the computation time curves almost remain unchanged. Thus, although increasing the angle threshold to large values will not result in more computation time, it will not improve map conflation performance. Fig. 10(b) illustrates the integration performance when the angle threshold is fixed at 0.25 rad $\approx 28.6°$ and the distance threshold increases from 5 m to around 50 m. Unlike the flat curves in Fig. 10(a), the running time curve keeps increasing as the distance distance threshold increases. However, the integration percentage curve shows that when the distance threshold exceeds 10, increasing the distance threshold cannot dramatically improve the integration performance. In the map conflation process, the two key parameters, angle and distance, significantly affect the final results.

To measure how these two parameters interact with each other during the conflation process, Fig. 11 shows a heatmap of the amount of successfully conflated HERE segments with respect to different angle and distance thresholds after the conflation process. When the angle threshold between referencing segments and conflated segments is small, say, 5° or 10°, increasing the angle threshold can greatly enlarge the amount of conflated segments. When the angle threshold increases to 20° or higher, the conflation results cannot achieve significant improvement. When the distance threshold is enlarged, the amount of successfully conflated segments gradually increases. In summary, the heatmap shows that increasing the angle and distance thresholds can improve conflation performance. It is better to set the angle threshold to larger than 15°.

Since the HERE geometric roadway layer is complex in terms of the number and topology of the segments, some segments inevitably need to be postprocessed manually. After postprocessing, the on-road segment-based data can also be fully integrated. Fig. 12 shows an example of the conflated HERE layer. The HERE layer covers most freeways and some arterials and urban streets, while the referencing layer only contains Washington freeways. Thus, the

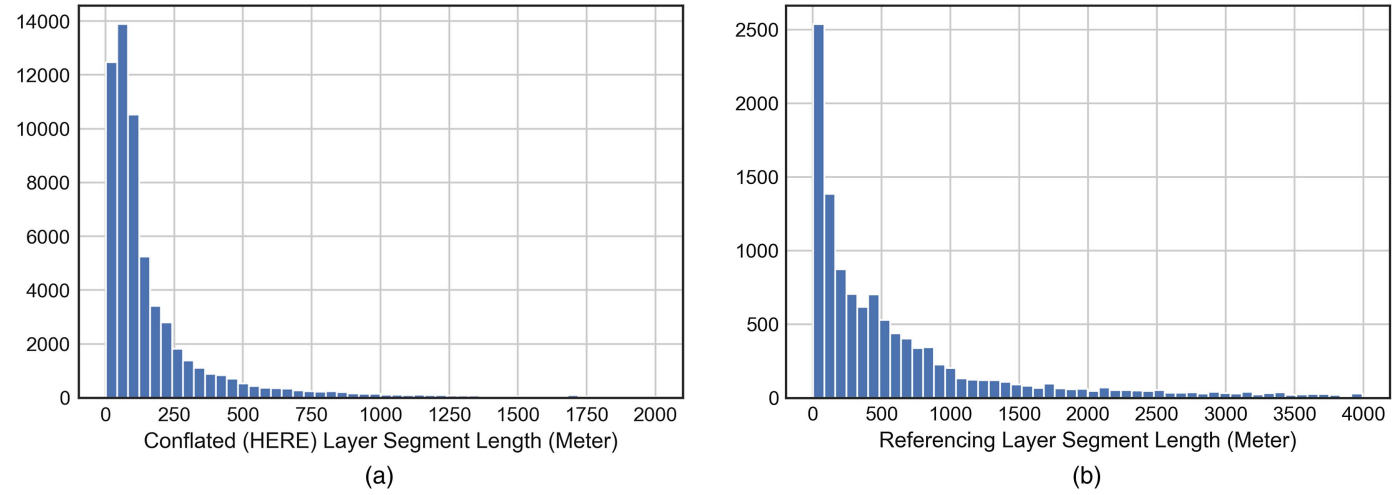

Fig. 9. Distributions of road segment length: (a) distribution of road segment length in conflated layer; and (b) distribution of road segment length in referencing layer.

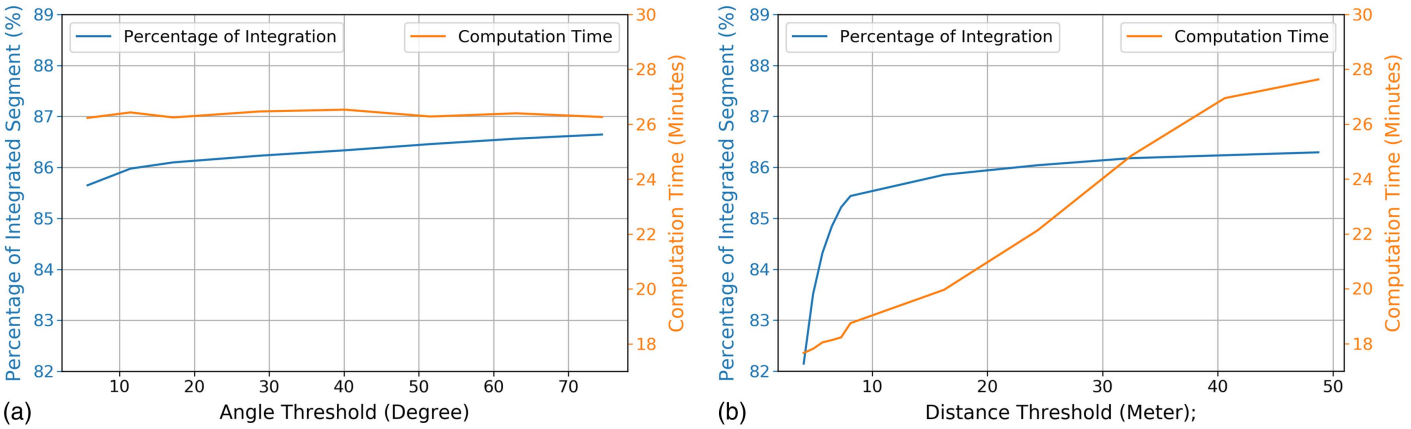

Fig. 10. Map conflation algorithm efficiency analysis on angle threshold and distance threshold in terms of percentage of integrated segments and computation time: (a) efficiency analysis on angle threshold; and (b) efficiency analysis on distance threshold.

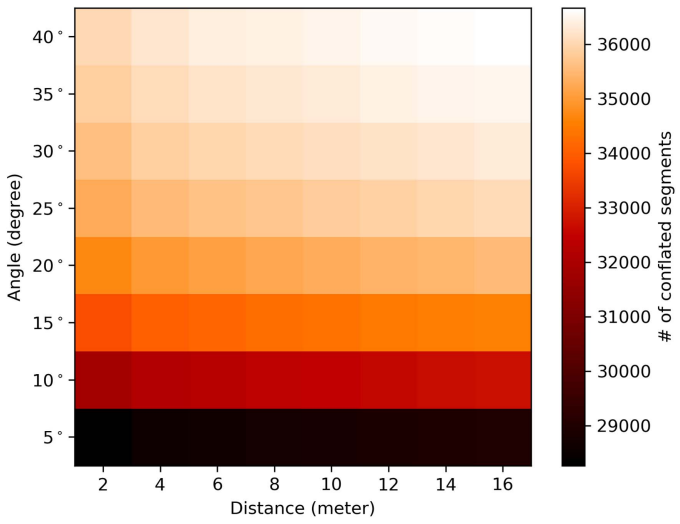

Fig. 11. Heatmap of total amount of successfully conflated segments with respect to different angle and distance thresholds after conflation process.

generated linking roadway layers are mostly located on freeways, shown by green links in Fig. 12. In the rural areas, some corridors are not covered by both referencing and conflated layers, and therefore, they are not included in the linking layer.

Applications and Case Studies

Based on the data-integration framework, multiple data sources can be combined together by means of a referencing roadway geometric layer. Fig. 13 demonstrates the architecture of the potential traffic analysis applications based on the data-integration framework. Each type of transportation data is connected to the geospatial referencing database via a linkage table or database. The set of the linkage tables/databases, which are the desired set $\mathcal{L} = \{L_1, \ldots, L_n\}$ described in the problem statement section, are the key acquirements of this study. In the data extraction modules, multiple analysis parameters can be selected for different specific analysis. The roadway selection can be fulfilled by using multiple advanced spatial query/manipulation functions. The date and time information or other types of parameters can be directly specified from the original transportation databases. Given all the

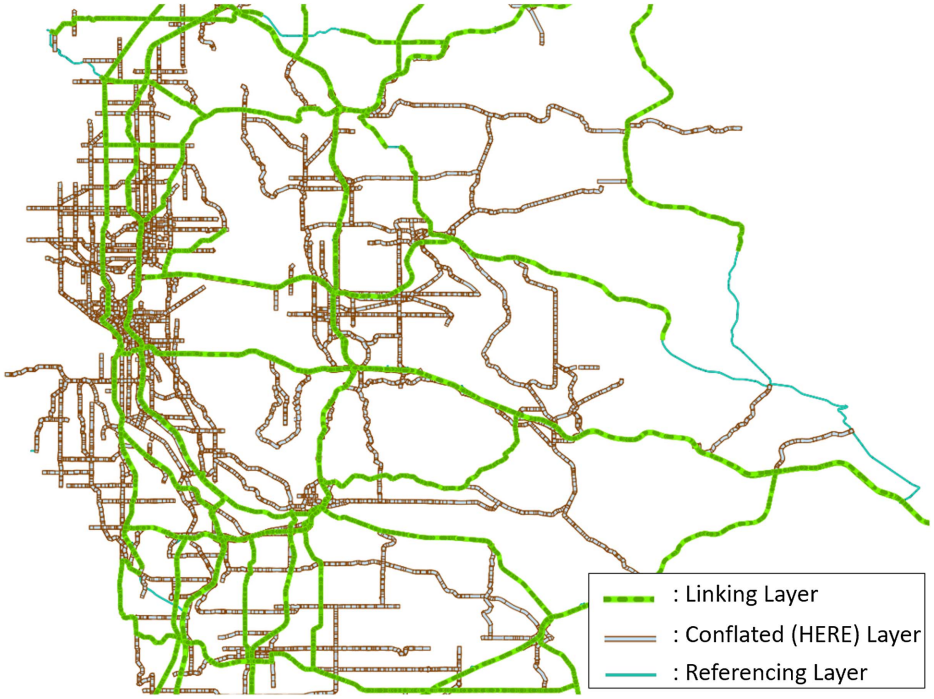

Fig. 12. Visualized map conflation results for on-road segment-based data, taking the HERE data as an example.

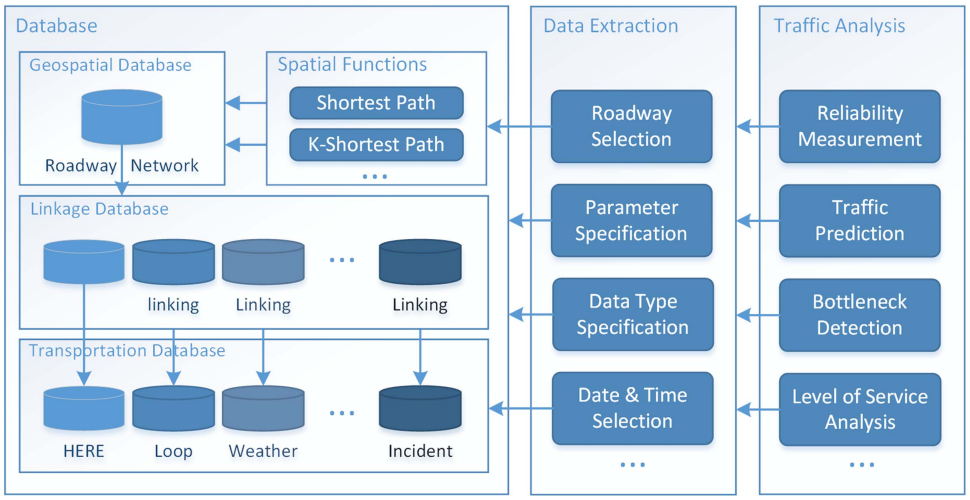

Fig. 13. Architecture of applications based on multisource data-integration framework.

required data extraction parameters are specified, multiple traffic analysis modules can be carried out efficiently using a single system, such as travel time reliability measurement, traffic prediction, traffic network bottleneck detection, level of service analysis, etc. Guided by the SHRP2 L02 product report (List et al. 2014) and L08 product report (Zegeer et al. 2013), several traffic analysis functions are implemented based on this architecture and introduced in the following case studies.

Case Study: Travel Time Analysis

Given multiple transportation data integrated, variations in travel time in different data sets can be compared. Fig. 14 demonstrates the variations in travel time by time of day measured by loop detector data and HERE data, respectively, in 2015. The study area is the high-occupancy vehicle (HOV) lane on I-405 from Bellevue to Lynnwood in Washington state. It can be found that the variations in travel time in the two data sets have similar patterns. The travel times influenced by incident and weather are also compared, respectively. Since the geospatial representations of HERE data and loop detectors are identical after the map conflation process, obvious differences in the two travel time distributions can still be observed that the average and variance of the travel time measured by HERE data are both higher than that measured by loop detector data. Hence, data quality can efficiently be compared for different data sources based on the proposed data-integration framework.

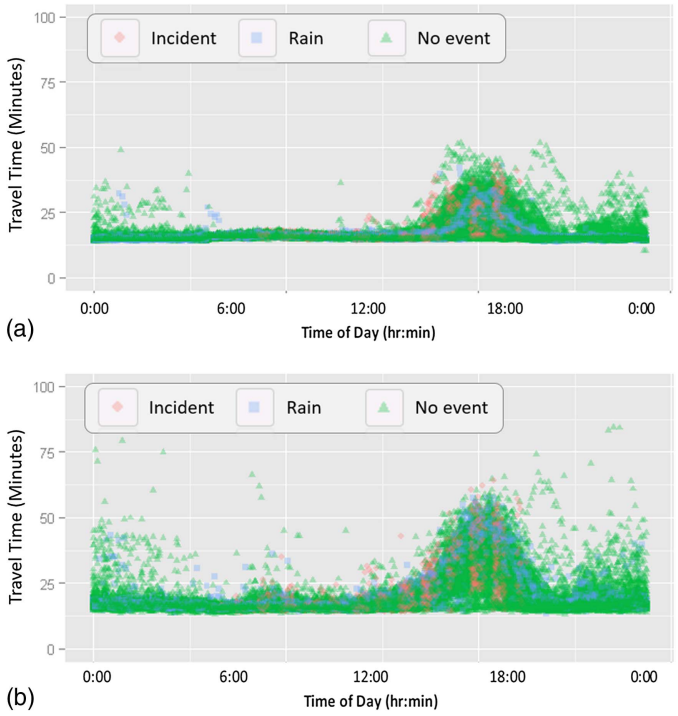

Fig. 14. Variations in travel time by time of day measured by (a) loop detector data; and (b) HERE data. The study area is the high-occupancy vehicle lane on I-405 from Bellevue to Lynnwood in Washington state.

Case Study: Performance Measurement

Using the proposed traffic data-integration framework, a variety of reliability performance measurement metrics, such as reliability rating, planning time index (PTI), and 80th percentile travel time index (TTI), can be analyzed based on different data sources. PTI is defined as the 95th percentile travel time divided by the free-flow travel time. Fig. 15 illustrates the PTI distribution by time of day on the I-405 freeway from the city of Bellevue to the city of Lynwood in 2015, in which three data sets, the general-purpose (GP) lane loop detector data, the HOV lane loop detector data, and the HERE data are compared. The PTI distributions of HERE data and GP lane loop detector data are similar, and the PTI distribution of HOV lane loop detector data has a more obvious evening peak. Based on the data-integration framework, multisource-based data analysis can be easily and efficiently fulfilled using a transportation data analysis tool. The travel time reliability analysis module is implemented on the DRIVE Net platform.

Case Study: Web-Based Analysis Tool

The proposed traffic data-integration framework is fully developed and implemented on an online interactive publicly accessible transportation data platform, DRIVE Net (WSDOT 2018). On the DRIVE Net platform, multiple traffic analysis modules, such as travel time analysis, travel time reliability analysis, transportation emission analysis, and traffic safety analysis, are implemented (Fig. 16). Because the framework integrated multiple data sources together via a uniform referencing layer, the referencing layer is used by the tool to provide convenient route selection modules, including predefined route selection and user-defined route selection. In addition, the tool can provide customized analysis periods by letting users select the date and day of the week. Further, the integrated data sets, which have significant effects on travel time reliability, such as weather and incident data, can be easily combined in the analysis module. By implementing the multisource transportation data-integration framework, the use of such an efficient transportation data analysis tool can dramatically reduce workforce and computation investment and cost.

Discussions and Conclusions

In this study, to overcome the main hurdles of analyzing multisource transportation data, a transportation data-integration framework based on a uniform roadway referencing layer is proposed. Four types of traffic analysis–related data, including on-road segment-based data, on-road point-based data, off-road segment-based data, and off-road point-based data, are categorized from the perspective of sensor locations and sensing areas to deal more efficiently with multisource traffic data. Meanwhile, an iterative map conflation algorithm is proposed mainly for integrating on-road segment-based data. By implementing the data-integration framework on a transportation big data platform, several real-world case studies and applications for traffic analysis are realized and presented. The experimental results show that the proposed framework performs well in terms of accuracy and efficiency.

The main advantages of the proposed framework are flexibility and adaptability. With the data-integration framework, when a new data source needs to be integrated into the data analytical system, the same referencing layer can be used to integrate and manipulate the new data source. In the experiments, nearly all types of data can be 100% integrated. However, because data sources have different data formats, some special cases inevitably exist during the integration process. The framework sometimes cannot automatically integrate 100% of all data sources, taking the HERE data-integration in this study as an example. Despite the existence of extreme cases,

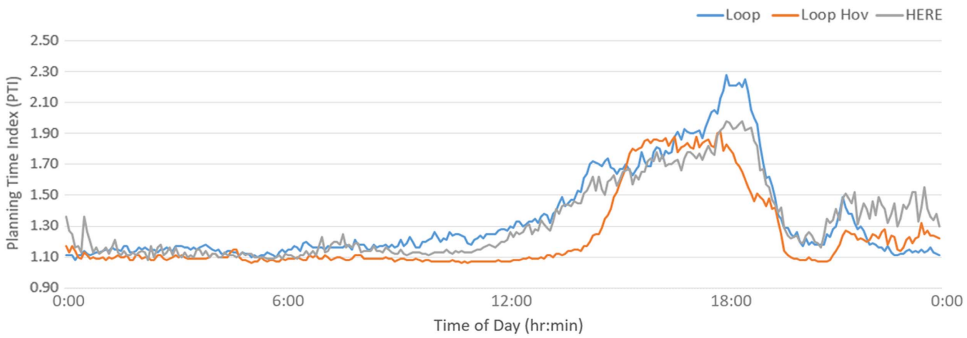

Fig. 15. PTI distributions by time of day on I-405 freeway corridor starting in Bellevue to Lynwood. Loop detector data on GP lanes and HOV lane; HERE data are compared.

J. Transp. Eng., Part A: Systems, 2020, 146(5): 04020024

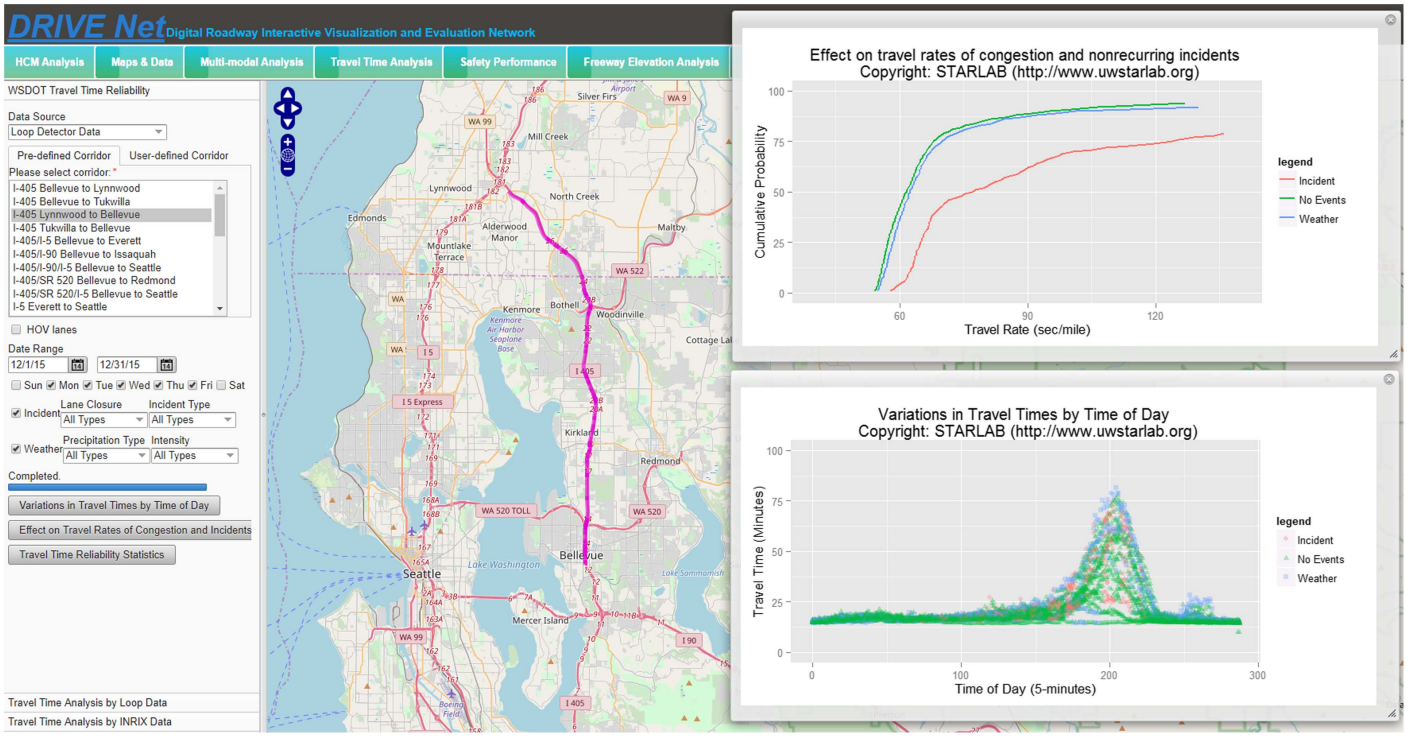

Fig. 16. Interface and functionality modules of web-based transportation data analytics platform. Two figures were generated showing travel rates, i.e., travel time reliability, and travel time variations are demonstrated on platform interface. (Base image courtesy of DRIVE Net, STAR Lab, used with permission.)

the framework can still work perfectly after manually checking and correcting the unintegrated data to facilitate further analysis in research work and real applications.

The main contribution of this study is the proposed data-integration framework, which is of practical importance, although there is still much room for improvement in the methodology sections. Future studies will explore more intelligent methods to build connections between the geospatial metadata of different data sources and design more efficient and accurate map conflation methods. When further complex analysis is required in the future, more data sources, such as roadway elevation data, vehicle trajectory data, and connected vehicle related data, will be tested for integration based on the proposed framework.

Data Availability Statement

Some or all data, models, or code used during the study were provided by a third party, such as the following:

1. Data sources used for data integration in the experiments of this study.
2. The shape file of the geometric referencing layer.

Direct requests for these materials may be submitted to the provider as indicated in the acknowledgments. Some or all data, models, or code generated or used during the study are available from the corresponding author by request, such as the source code of the proposed data-integration framework/algorithm.

Acknowledgments

This study was supported by the SHRP2 Reliability Data and Analysis Tools (L38) project from WSDOT. Thanks to WSDOT for providing the raw data and the referencing layer for this study.

References

Chen, C.-C., C. A. Knoblock, and C. Shahabi. 2006. "Automatically conflating road vector data with orthoimagery." *GeoInformatica* 10 (4): 495–530. https://doi.org/10.1007/s10707-006-0344-6.

Chow, A. 2016. "Heterogeneous urban traffic data and their integration through kernel-based interpolation." *J. Facil. Manage.* 14 (2): 165–178. https://doi.org/10.1108/JFM-08-2015-0025.

Claramunt, C., B. Jiang, and A. Bargiela. 2000. "A new framework for the integration, analysis and visualisation of urban traffic data within geographic information systems." *Transp. Res. C: Emerging Technol.* 8 (1): 167–184. https://doi.org/10.1016/S0968-090X(00)00009-7.

Cremer, M., and S. Schrieber. 1996. "Monitoring traffic load profiles with heterogeneous data source configurations." In *Proc., 13th Int. Symp. on Transportation and Traffic Theory*, 21–40. Lyon, France: International Symposium on Transportation and Traffic Theory.

Cui, Z., R. Ke, and Y. Wang. 2018. "Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction." Preprint, submitted January 7, 2018. https://arxiv.org/abs/1801.02143.

Cui, Z., S. Zhang, K. C. Henrickson, and Y. Wang. 2016. "New progress of DRIVE net: An E-science transportation platform for data sharing, visualization, modeling, and analysis." In *Proc., IEEE Int. Smart Cities Conf. (ISC2)*, 2. Washington, DC: IEEE.

Daneshgar, F., K. Farokhi Sadabadi, and A. Haghani. 2018. "A conflation methodology for two GIS roadway networks and its application in performance measurements." *Transp. Res. Rec.* 2672 (45): 284–293. https://doi.org/10.1177/0361198118793000.

Duan, Y., Y. Lv, Y. L. Liu, and F. Y. Wang. 2016. "An efficient realization of deep learning for traffic data imputation." *Transp. Res. C: Emerging Technol.* 72 (Nov): 168–181. https://doi.org/10.1016/j.trc.2016.09.015.

El Faouzi, N.-E., H. Leung, and A. Kurian. 2011. "Data fusion in intelligent transportation systems: Progress and challenges: A survey." *Inf. Fusion* 12 (1): 4–10. https://doi.org/10.1016/j.inffus.2010.06.001.

Fourati, H. 2015. *Multisensor data fusion: From algorithms and architectural design to applications*, 639. Boca Raton, FL: CRC Press.

Gettman, D., A. Toppen, K. Hales, A. Voss, S. Engel, and D. El Azhari. 2017. *Integrating emerging data sources into operational practice: Opportunities for integration of emerging data for traffic management and TMCs*. Final Rep. No. FHWA-JPO-18-625. Washington, DC: DOT.

Graettinger, A., X. Qin, G. Spear, S. Parker, and S. Forde. 2009. "Combining state route and local road linear referencing system information." *Transp. Res. Rec.* 2121 (1): 152–159. https://doi.org/10.3141/2121-17.

Green, E., J. Ripy, M. Chen, and X. Zhang. 2013. "Conflation methodologies to incorporate consumer travel data into state HPMS datasets." In *Proc., 92nd Transportation Research Board Annual Meeting*, 15. Washington, DC: Transportation Research Board.

Haklay, M., and P. Weber. 2008. "Openstreetmap: User-generated street maps." *IEEE Pervasive Comput.* 7 (4): 12–18. https://doi.org/10.1109/MPRV.2008.80.

Henrickson, K., Y. Zou, and Y. Wang. 2015. "Flexible and robust method for missing loop detector data imputation." *Transp. Res. Rec.* 2527 (1): 29–36. https://doi.org/10.3141/2527-04.

Jahangiri, A., and H. A. Rakha. 2015. "Applying machine learning techniques to transportation mode recognition using mobile phone sensor data." *IEEE Trans. Intell. Transp. Syst.* 16 (5): 2406–2417. https://doi.org/10.1109/TITS.2015.2405759.

Kang, B., J. Y. Scully, O. Stewart, P. M. Hurvitz, and A. V. Moudon. 2015. "Split-match-aggregate (SMA) algorithm: Integrating sidewalk data with transportation network data in GIS." *Int. J. Geogr. Inf. Sci.* 29 (3): 440–453. https://doi.org/10.1080/13658816.2014.981191.

Kaushik, K., E. Sharifi, and S. E. Young. 2015. "Computing performance measures with national performance management research data set." *Transp. Res. Rec.* 2529 (1): 10–26. https://doi.org/10.3141/2529-02.

Klein, L. A., M. K. Mills, and D. R. P. Gibson. 2006. *Traffic detector handbook: Volume II*. Rep. No. FHWA-HRT-06-139. Washington, DC: DOT, Federal Highway Administration.

Lenzerini, M. 2002. "Data integration: A theoretical perspective." In *Proc., 21st ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems*, 233–246. New York: Association for Computing Machinery.

Li, L., and J. Valdovinos. 2017. "Optimized conflation of authoritative and crowd-sourced geographic data: Creating an integrated bike map." In *Information Fusion and Intelligent Geographic Information Systems (IF&IGIS'17)*, 227–241. New York: Association for Computing Machinery.

List, G. F., B. Williams, and N. Rouphail. 2014. *Establishing monitoring programs for travel time reliability*. Strategic Highway Research Program (SHRP 2) Rep. No. S2-L02-RR-2. Washington, DC: Transportation Research Board.

Liu, S., and Q. Qu. 2016. "Dynamic collective routing using crowdsourcing data." *Transp. Res. B: Methodol.* 93 (Nov): 450–469. https://doi.org/10.1016/j.trb.2016.08.005.

Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2005. *Geographic information systems and science*. Hoboken, NJ: Wiley.

Luk, J. Y. K., and C. Yang. 2003. "Comparing driver information systems in a dynamic modeling framework." *J. Transp. Eng.* 129 (1): 42–50. https://doi.org/10.1061/(ASCE)0733-947X(2003)129:1(42).

Ma, X., and Y. Wang. 2014. "Development of a data-driven platform for transit performance measures using smart card and GPS data." *J. Transp. Eng.* 140 (12): 04014063. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000714.

Ma, X., Y.-J. Wu, and Y. Wang. 2011. "DRIVE Net: E-science transportation platform for data sharing, visualization, modeling, and analysis." *Transp. Res. Rec.* 2215 (1): 37–49. https://doi.org/10.3141/2215-04.

Memarian, A., S. P. Mattingly, J. M. Rosenberger, J. C. Williams, S. A. Ardekani, and H. Hashemi. 2018. "Modeling framework to identify an affected area for developing traffic management strategies." *J. Transp. Eng.* 144 (10): 04018059. https://doi.org/10.1061/JTEPBS.0000182.

Okabe, A., B. Boots, K. Sugihara, and S. N. Chiu. 2009. *Spatial tessellations: Concepts and applications of voronoi diagrams*, 690. Chichester, UK: Wiley.

Rafferty, P., and C. Hankley. 2014. *National performance management research data set (NPMRDS). Wisconsin traffic operations and safety laboratory*, 12. Sunnyvale, CA: HERE Technologies.

Ramsey, P. 2005. *Postgis manual*. Victoria, BC, Canada: Refractions Research Inc.

Schrank, D., E. Bill, and L. Tim. 2014. *Urban mobility report: Powered by Inrix traffic data*. Rep. No. SWUTC/15/161302-1. Kirkland, Washington: INRIX Analytics Company.

Tang, J., G. Zhang, Y. Wang, H. Wang, and F. Liu. 2015. "A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation." *Transp. Res. C: Emerging Technol.* 51 (Feb): 29–40. https://doi.org/10.1016/j.trc.2014.11.003.

Tarko, A. P., and N. M. Rouphail. 1997. "Intelligent traffic data processing for ITS applications." *J. Transp. Eng.* 123 (4): 298–307. https://doi.org/10.1061/(ASCE)0733-947X(1997)123:4(298).

Taylor, M. A. P., J. E. Woolley, and R. Zito. 2000. "Integration of the global positioning system and geographical information systems for traffic congestion studies." *Transp. Res C: Emerging Technol.* 8 (1-6): 257–285. https://doi.org/10.1016/S0968-090X(00)00015-2.

Thakali, L., L. Fu, and T. Chen. 2016. "Model-based versus data-driven approach for road safety analysis: Do more data help?" *Transp. Res. Rec.* 2601 (1): 33–41. https://doi.org/10.3141/2601-05.

Valadkhani, A. H., Y. Hong, and M. Ramezani. 2017. "Integration of loop and probe data for traffic state estimation on freeway and signalized arterial links." In *Proc., IEEE 20th Int. Conf. on Intelligent Transportation Systems (ITSC)*, 1–6. Washington, DC: IEEE.

WSDOT (Washington State Department of Transportation). 2018. "Washington state highways LRS (linear reference systems)." Accessed December 31, 2008. http://www.wsdot.wa.gov/mapsdata/geodatacatalog/maps/NOSCALE/DOT_TDO/LRS/WSDOT_LRS.htm.

Zegeer, J., J. A. Bonneson, R. G. Dowling, P. Ryus, M. Vandehey, and W. Kittelson. 2013. *Incorporating travel time reliability into the highway capacity manual*. Strategic Highway Research Program (SHRP 2) Rep. No. S2-L08-RW-1. Washington, DC: Transportation Research Board.

Zheng, Z., C. Wang, P. Wang, Y. Xiong, F. Zhang, and Y. Lv. 2018. "Framework for fusing traffic information from social and physical transportation data." *PLoS One* 13 (8): e0201531. https://doi.org/10.1371/journal.pone.0201531.