Summer 5-26-2017

Key Technology of Traffic Data Collection, Repair and Mining

Xiaoxia Wang
School of Traffic and Transportation, Beijing Jiaotong University, Beijing, 100044, P.R. China; MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing, 100044, P.R. China,
xxwang@bjtu.edu.cn

Zhanqiang Li
School of Traffic and Transportation, Beijing Jiaotong University, Beijing, 100044, P.R. China

Key Technology of Traffic Data Collection, Repair and Mining

Xiaoxia Wang [1,2*], *Zhanqiang Li*[1]

[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing, 100044, P.R. China

[2]MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing, 100044, P.R. China

Abstract: The big data revolution radically transforms how cities monitor, manage, and enhance the livability of our society. A successful smart city will be one in which the public agencies, businesses and people are able to make informed decisions and respond to dynamic conditions based on real-time sensing and data analytics. The paper listed some gathering methods for traffic data, and preprocessed the collected data according to their missing feathers; then conducted several repair methods and evaluated their repair effects; finally constructed an integrated platform for traffic data mining. The purpose of the paper is to present some technique for the exploration of traffic data in an intelligent transportation system.

Keywords: data collection, data repair, data mining, traffic data

1. INTRODUCTION

With the progress of urbanization in China, increasing people are pouring into metropolis, like Beijing, and the problem of traffic congestion becoming serious. Along with the advent of the era of big data and benefiting from their capabilities in allowing to process voluminous amounts of them created in real time, the application of data mining in the field of transportation is increasing and the technology is changing with each day.

This paper illustrated with the traffic data(the Traffic Performance Index and average speed for regions) crawled from Beijing Municipal Commission of Transport (BMCT, http://www.bjjtw.gov.cn) every 15 minutes with Visual Basic for Application (VBA), firstly completed data structure, then eliminated outliers and fix nulls with the candidate repair methods according to their features, and finally put forward a framework of steps with an integrated platform for traffic data mining.

2. PREPARE TRAFFIC DATA FOR PROCESSING

2.1 Data Acquisition

Collect urban data to reflect the real-time traffic information, for a long time from multiple running vehicles to a roadside base station via vehicle-to-vehicle and vehicle-to-infrastructure communications[1], with the objective to minimize the communication cost while satisfying the data collection time constraint; now from wireless sensor networks using smartphones as mobile base stations and leveraging human mobility[2]. Today with a wide variety of data on the websites, employ crawler machine, script language(e.g. Excel with VBA), and program with R or Python via a set of regular expressions and XPath for getting information in the XML document[3]. For example, apply the Beautiful Soup to parse the HTML and XML, find and modify the parse tree. For Beautiful Soup, urllib2 is superior to selenium and requests in stable and efficient to access the HTTPS website.

2.2 Division of Loss Type for the Raw Data

Since the original data occasionally lost, mostly due to network communication failure or the website maintenance, here extract a subset (some Sunday's data of the entire road network in 2015 from BMCT) as Figure 1 to demonstrate the following repair methods in view of the periodic characteristics of urban traffic. Crawler once per 15 minutes, so ideally 96 times for every day from 0:00 to 24:00. The intersection of each line

* Corresponding author. Email: xxwang@bjtu.edu.cn

at the X-axis implies data missing at that point. The X-axis and Y-axis are with the same meaning for the following graphs. Here the missing data fall into three categories. (1) little lost, like at 1:00 on date 9/6; (2) the absence of small amount of continuous data, such as 2:00-4:00 on date 8/30, 1:00-3:00 and 13:00-18:00 on date 9/13, or 0:00-1:30 on date 9/27; (3) the absence of a large amount of data, like date 9/20.

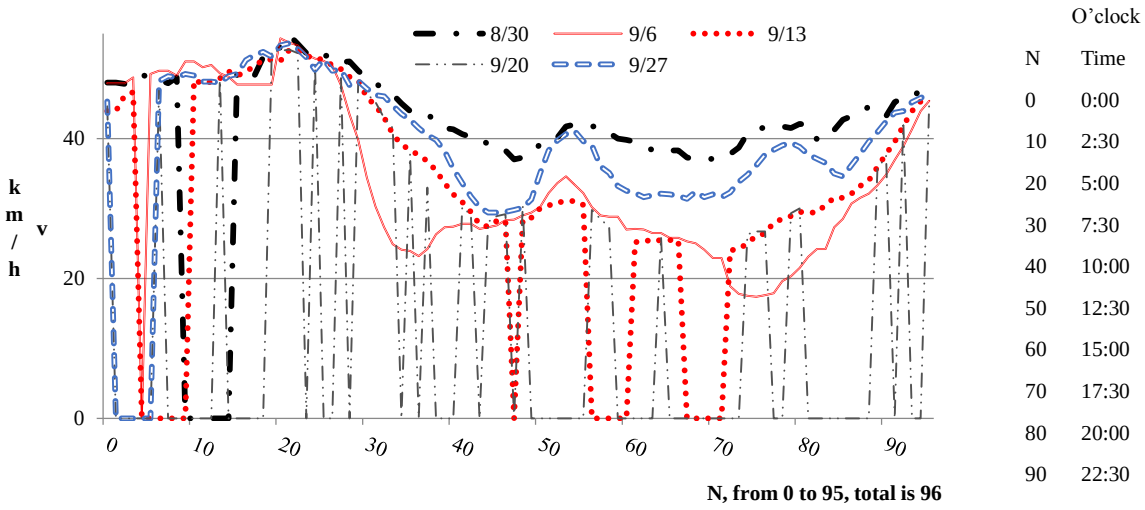

Figure 1. Raw data line charts on Sundays for the entire road network

2.3 Flow Chart for Data Repair

For analysis and mining, data must be correct, intact and consistent. Figure 2 outlines the repair steps.

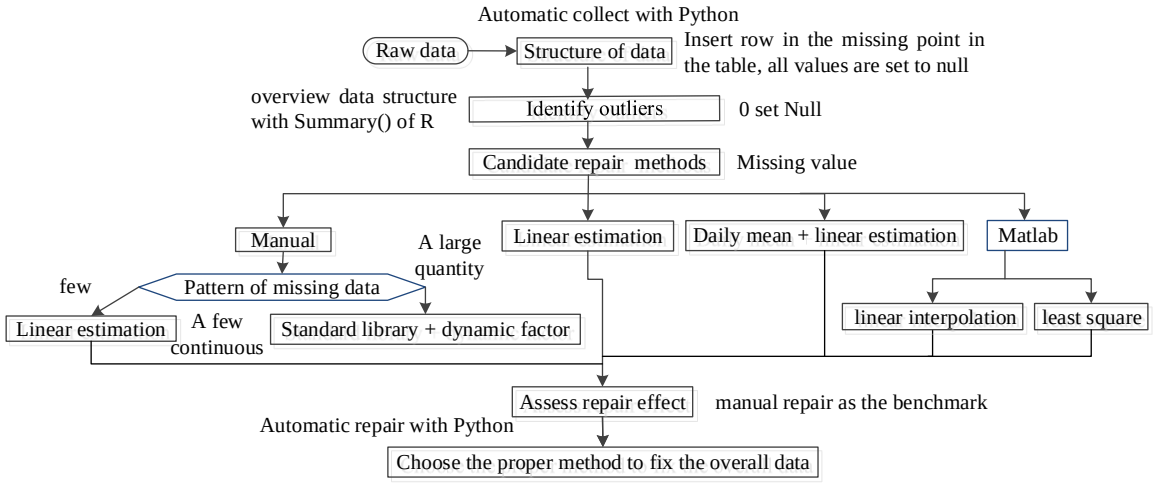

Figure 2. Flow chart for data repair

In the big data era, cloud computing transforms the traditional government services model, improves municipal services align to innovation with administration strategy, and creates intelligent executive network that encourages effective collaboration[4]. For instance, a generic and highly scalable cloud-based architecture SMASH[5] includes a distributed file system for capturing and storing data, a high performance computing engine to process large quantities of data, a reliable database system to optimize the indexing and querying of the data, and geospatial capabilities to visualize the resultant analyzed data. And cloud service providers deliver a satisfactory quality of experience for the user with net utility and service response time[6], with a scalable

visual tool for the analysis of high-throughput network traffic, power consumption and cluster-based system resource allocation[7].

3. IMPLEMENTATION OF DATA CLEAN AND REPAIR ACCORDING TO THEIR FEATURES

Repair missing traffic flow data usually built on correlation analysis of the traffic flow parameters series (between road cross-sections) with the co-integration theory of econometrics[8]. For distributed content-related data, detect inconsistencies and repair[9]. In view of algorithms, model repair diversification problem as a bi-criteria optimization problem [10].

3.1 Linear Estimation

For rare data loss, equation (1) considers the coherence.

$$t_j=[(j-i)\times t_k+(k-j)\times t_i]\times(k-i)^{-1} \quad (1)$$

t_j, the missing data; t_i and t_k, the accurate data close to the left and right of t_j. For 1:00 (date 9/6) as an example, equal to the average of 0:45 and 1:15. Figure 3 only address the accurate data before and after the missing.

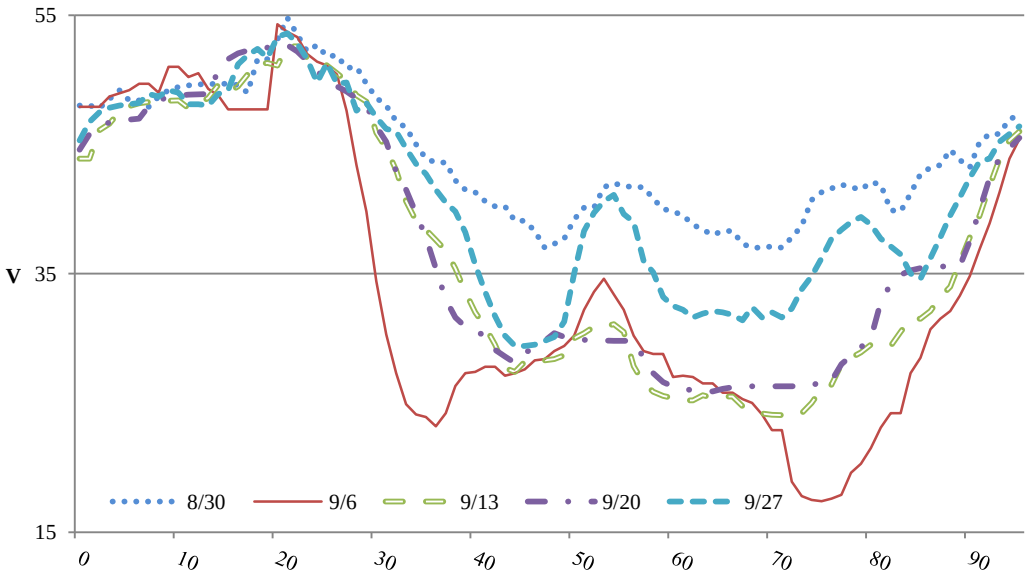

Figure 3. The results of linear estimation

3.2 Manual Repair Method: Standard Library + Dynamic Factor

Equation (2) and (3) estimate the missing data of the current period with historical data.

$$S_{D,t,L} = F_{D,L} \times S_{standard\ library,t,L} \quad t=1,\ 2,\ 3,\ \dots\ n \quad (2)$$

$$F_{D,L} = Average_{t=1}^{n} S_{D,t,L} / Average_{t=1}^{n} S_{standard\ library,t,L} \quad t=1,\ 2,\ 3,\ \dots\ n \quad (3)$$

$S_{D,t,L}$, the speed need to calculate at time t on day D in area L; $F_{D,L}$, the dynamic factor, equal to the ratio of the mean value during a certain time interval of the date that need to repair to the mean value of the same interval in the standard library; $S_{standard\ library,t,L}$, the speed in the standard library at time t in area L; n, time intervals exclusive nulls or zeroes.

3.2.1 The absence of a small amount of continuous data

In Figure 4(a), with 0:00-5:00(date 9/6) as a standard library, separately obtain the mean value for the corresponding period and divide as a dynamic factor. Then, multiplied by the value of the corresponding time of the standard library 9/6 to get the repaired for the other days. For the missing values on 11:00-18:00 (date 9/13),

with 11:00-18:00 (date 9/6) as a standard library, repeat the same steps and get the results as Figure 4(b).

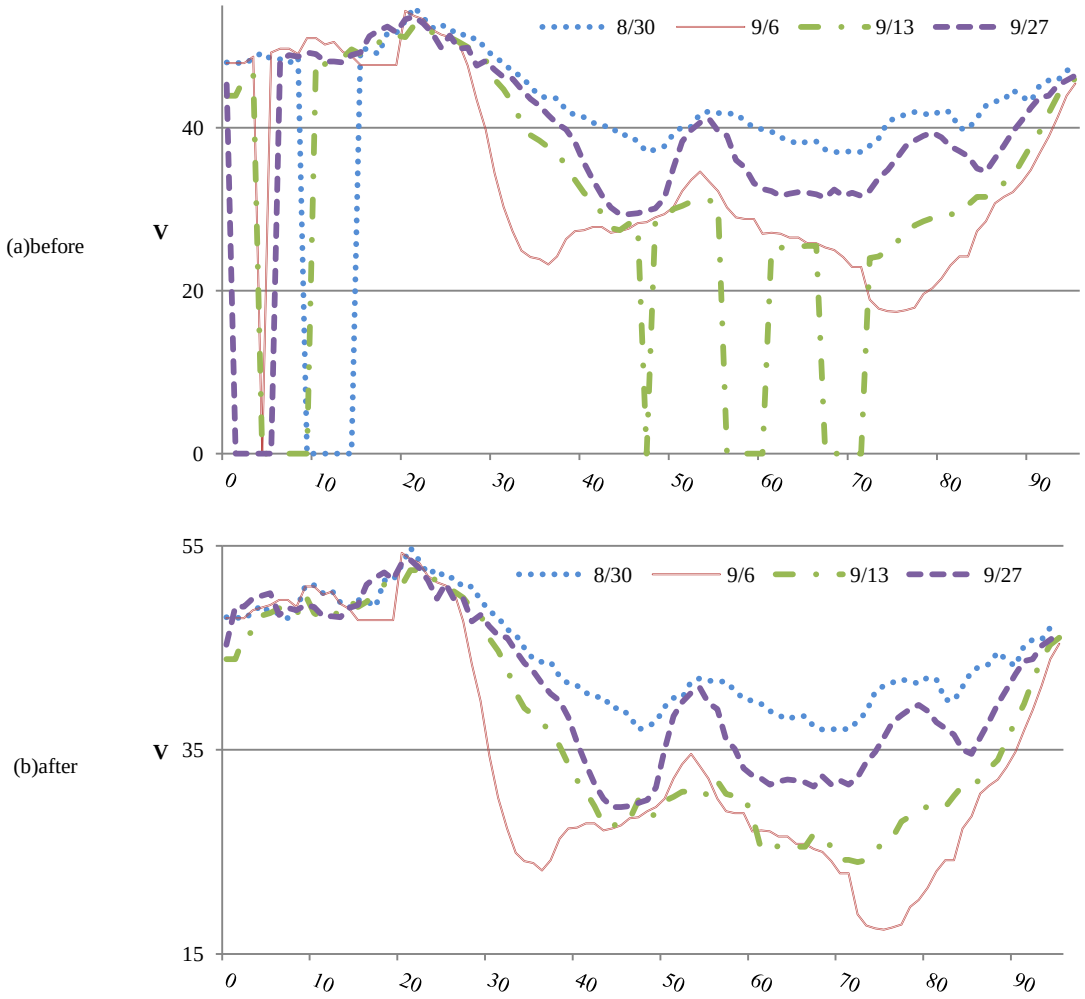

Figure 4. Repair the absence of a small amount of continuous data

3.2.2 The absence of a large amount of data

Pick the repaired 9/13 as a standard library for the similarity of the existing trend of date 9/20 to them. Following the above procedure, repair date 9/20 as Figure 5.

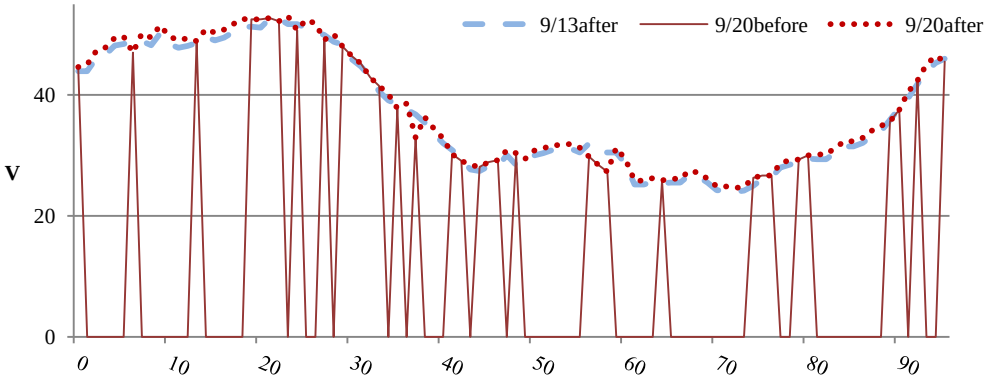

Figure 5. Repair the absence of a large amount of continuous data

3.3 Daily Mean + Linear Estimation Method

As equation (4) reflects the mean value of the same day the data need to be repaired. The functional condition is the random distribution of the missing data. Namely, the missing data don't affect the distribution and the average value of the speed, and the average speed of the day mirror the mediocre level of the day.

$$y_t = (y_{t+n} + y_{t-1} + y_{mean})/3 \quad (4)$$

y_t, the speed of time t to repair; y_{t-1}, y_{t+1}, the speed before and after the time t; y_{mean}, the mean value of the original data on that day to fix. For the continuous missing data, with the time sequence, y_{t-1} is certainly not null; If y_{t+1} is null, then continue to look down until find a non-null of y_{t+n}. The results as Figure 6.

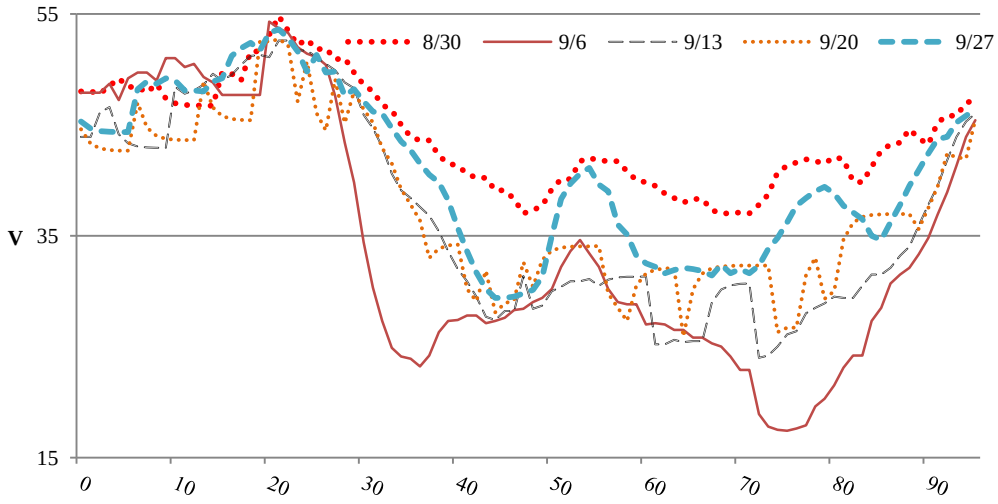

Figure 6. The results of daily mean + linear estimation method

3.4 Matlab with Spline and Least Square Method

Here automatically fix with MATLAB spline (Figure 7) and least square function (Figure 8). Obviously, it's troublesome for some negative values and sharp fluctuations. With 'interp' and 'ploytit' to fill the blank, replace nulls before carry out time series and regional analyses and cluster the time window[11, 12].

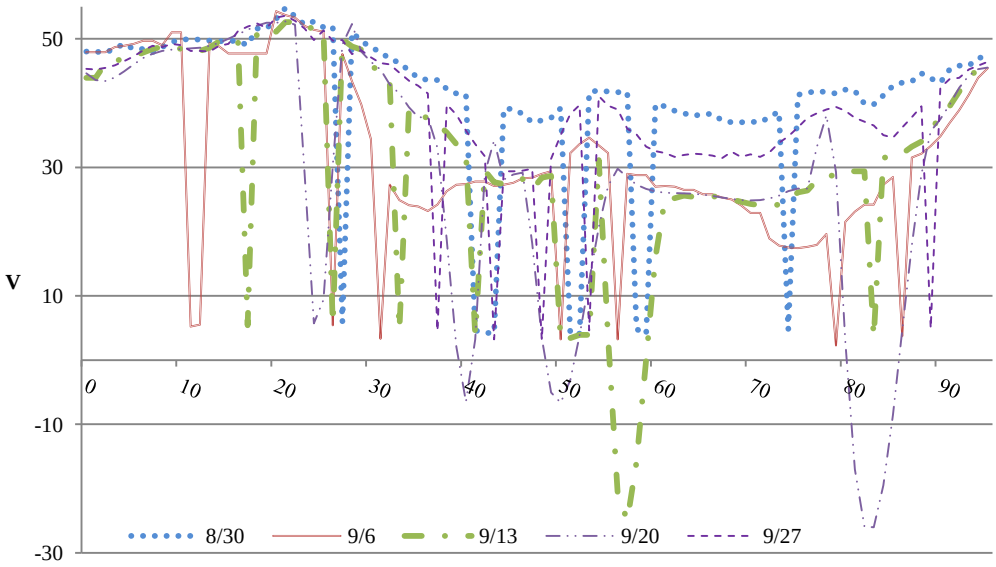

Figure 7. The results of MATLAB Spline

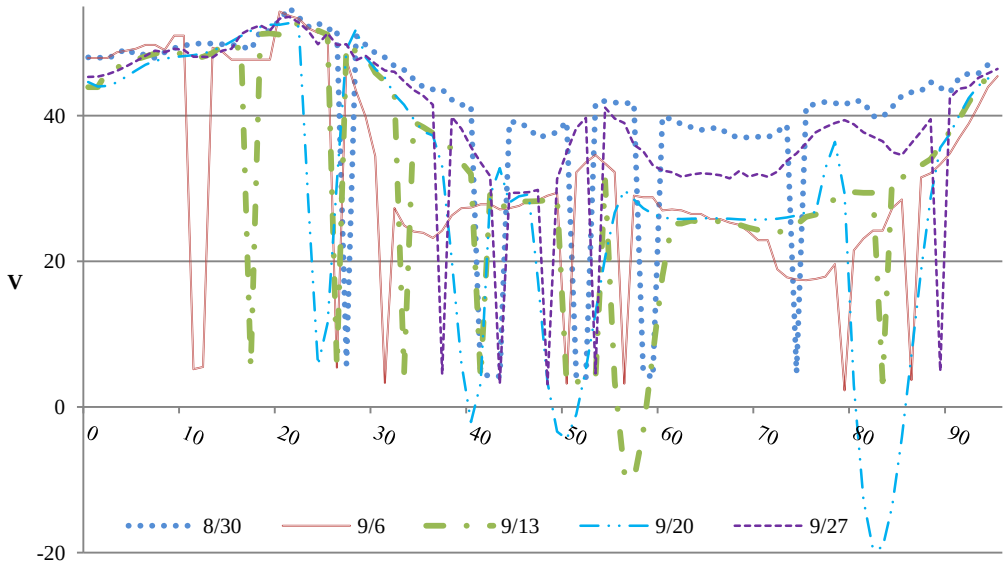

Figure 8. The results of MATLAB least square

3.5 Test Repair Effect

In summary Figure 9 intuitively figures out the general effect of each method.

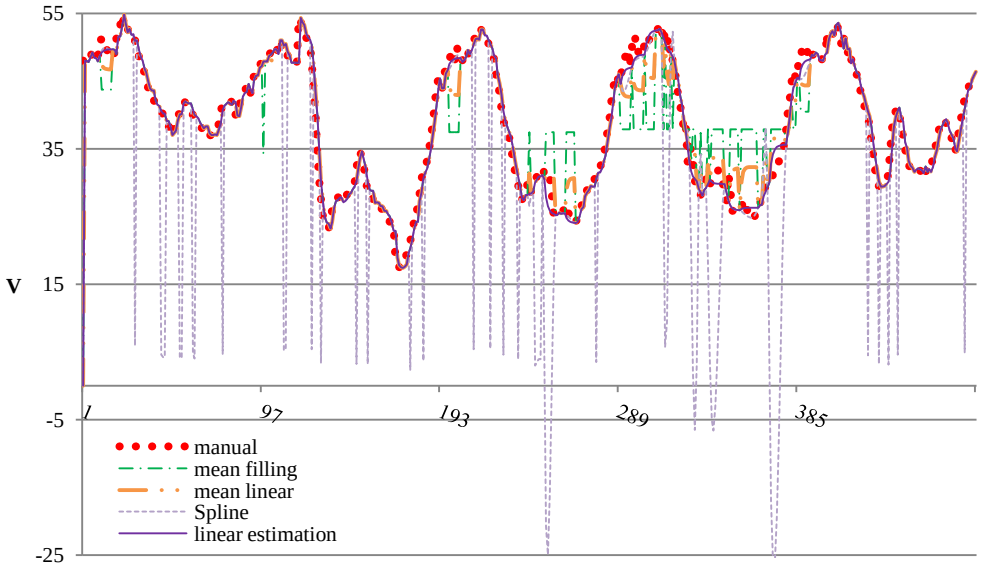

Figure 9. Comparison between repair methods (manual repair as the benchmark)

And Table 1 assesses the accuracy and reliability of these methods.

Table 1. Error evaluation between above repair methods

Error evaluation method	Mean filling	Mean linear	Spline	Linear estimation
Mean Square Error	4.2678	2.0275	12.6090	0.8779
Mean Absolute Error	1.7590	0.7959	4.5122	0.3014
Mean Square Percentage Error	0.0061	0.0014	0.0529	0.00026
Mean Absolute Percentage Error	0.0321	0.0145	0.0823	0.0055

Linear estimation is the best. Nevertheless, only considered the mean is worse than the combination with linear estimation, because of the coherence of traffic speed. Namely the value at one moment is close to before and after that moment, which is relatively stable and will not sharply rise or decrease.

4. TRAFFIC DATA MINING SUPPORT TRAFFIC DECISIONS

Traditionally, explore traffic data for roadway safety issues, such as accident analysis and prevention with studies of driver behavior and safety[13], work zone safety[14], road safety simulation models; knowing how traffic behaves in various zones(like cities, regions, etc.) with passive and active sensors[15] or microscopic lane-changing model[16], the connection between real-time traffic characteristics and freeway crashes occurrence for real-time crash prediction[17]. Now the ongoing application like mitigation adverse weather impacts on road mobility with Intelligent Transportation System(ITS) innovations to incorporate diverse data sources and perform proactive and reactive maintenance activities[18]; or dynamic scene understanding(automatically interpret activities, as well as detect unusual events) through the video via temporal association rules with mining algorithm[19].

Along with the growth of data volume, enterprises should choose the appropriate platform for data integration and maximize the value of data mining.[3] When utilize data mining techniques, to discover maritime traffic patterns with the potential use of open source data mining tools R [20]; or find trajectory patterns of frequent behaviors with GSM data[21]. As for mining algorithms, such as a map-matching method based on multi-criteria genetic algorithm with dynamic time wrapping[22]; for data stream, clustering and classification for ubiquitous usage[23]; a Bayesian network, decision trees and artificial neural networks to detect the influence of factors on car accidents and injury risk[24]. Figure 10 shows the overall framework and an integrate platform for traffic data mining.

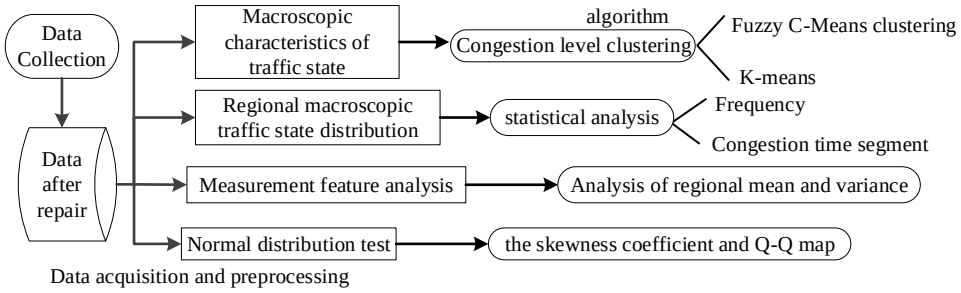

Figure 10. Framework of steps of data mining

After repair, the area data can be utilized for analysis and mining state and characteristics of regional macro traffic, such as the analysis of regional congestion level and congestion time by clustering, traffic speed distribution and features of different regions. Further, thanks to quickly data visualized tools, it is becoming easier to identify correlations and conceive of innovative and unanticipated usage for existing information. Urban planners, administration, travelers, and drivers can conduct their diverse knowledge discovery tasks with direct semantic and visual assists. For example, a method of visualizing spatiotemporal events by multi-layered geo-locational word clouds representation from an automat geo-located microblog stream.[25] Another example, a visual analytics system with interactive visualization tools to reflect urban mobility patterns and trends by discover and analyze the hidden knowledge of massive taxi trajectory data within a city, which include taxi topic maps, topic routes, street clouds and parallel coordinates, to visualize the probability-based topical information. [26]

5. CONCLUSIONS

The acquisition and processing of big traffic data lay the foundation for the subsequent analysis and mining which promote the development of the ITS. With the fusion and development of "internet + traffic", big traffic data and cloud computing are improving the service of public transportation system via ITS and transform urban residents travel experience.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant 71303018 and by the Center of Cooperative Innovation for Beijing Metropolitan Transportation. The authors would thank Lijuan Zhuge for her graduation design in my research group under my guidance.

REFERENCES

[1] He, Z., Zhang, D. (2017). Cost-efficient traffic-aware data collection protocol in VANET. Ad Hoc Networks, 55, 28–39.

[2] Can, Z., Demirbas, M. (2015). Smartphone-based data collection from wireless sensor networks in an urban environment. Journal of Network and Computer Applications, 58, 208–216.

[3] Wang, X., Li, Z. (2016). Integrated Platform for Smart Traffic Big Data. Proceedings of the 6th IEEE International Conference on Logistics, Informatics and Services Sciences. Beijing, 345–350.

[4] Wang, X., Li, Z. (2016). Traffic and transportation smart with cloud computing on big data. International Journal of Computer Science and Applications, 13(1), 1–16.

[5] Sinnott, R. O., Morandini, L., Wu, S. (2016). SMASH: A Cloud-Based Architecture for Big Data Processing and Visualization of Traffic Data. 2015 IEEE International Conference on Data Science and Data Intensive Systems. Sydney: IEEE Computer Society, 53–60.

[6] Calyam, P., Rajagopalan, S., Seetharam, S., Selvadhurai, A., Salah, K., Ramnath, R. (2014). VDC-Analyst: Design and verification of virtual desktop cloud resource allocations. Computer Networks, 68, 110–122.

[7] Bharadwaj, K., Flores, S., Rodriguez, J., Long, L., & Marai, G. E. (2016). Developing a scalable SNMP monitor. Proceedings - IEEE 28th International Parallel and Distributed Processing Symposium Workshops. Chicago: IEEE Computer Society, 1043–1047.

[8] Heng, L., Zhengyu, D., Xiaofa, S. (2013). Correlation Analysis and Data Repair of Loop Data in Urban Expressway Based on Co-integration Theory. Procedia - Social and Behavioral Sciences, 96(CICTP), 798–806.

[9] Du, Y., Member, S. (2016). Content-Related Repairing of Inconsistencies in Distributed Data. Journal of Computer Science and Technology, 31(4), 741–758.

[10] He, C., Tan, Z., Chen, Q., Sha, C. (2016). Repair diversification : A new approach for data repairing. Information Sciences, 346–347, 90–105.

[11] Cui, Y., Wang, X. (2016). Analysis of Congestion on Space and Time of Beijing 's Traffic Data. Proceedings of The 16th Cota International Conference of Transportation Professionals. Shanghai: ASCE, 1342–1353.

[12] Cui, Y., Wang, X. (2016). Research on the Distribution of Freight with Time Windows in Consideration of Traffic Congestion. The 15th Wuhan International Conference on E-Business. Wuhan: ALFRED UNIV., 303–310.

[13] Carsten, O., Kircher, K., Jamson, S. (2013). Vehicle-based studies of driving in the real world: The hard truth? Accident Analysis and Prevention, 58, 162–174.

[14] Yang, H., Ozbay, K., Ozturk, O., Xie, K. (2015). Work Zone Safety Analysis and Modeling: A State-of-the-Art Review. Traffic Injury Prevention, 16(4), 387–396.

[15] Castillo, E., Grande, Z., Calvino, A., Szeto, W. Y., Lo, H. K. (2015). A State-of-the-Art Review of the Sensor Location, Flow Observability, Estimation, and Prediction Problems in Traffic Networks. Journal of Sensors.

[16] Rahman, M., Chowdhury, M., Xie, Y., He, Y. (2013). Review of Microscopic Lane-Changing Models and Future Research Opportunities. IEEE Transactions on Intelligent Transportation Systems, 14(4), 1942–1956.

[17] Roshandel, S., Zheng, Z., Washington, S. (2015). Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. Accident Analysis and Prevention, 79, 198–211.

[18] Dey, K. C., Mishra, A., Chowdhury, M. (2015). Potential of Intelligent Transportation Systems in Mitigating Adverse Weather Impacts on Road Mobility: A Review. IEEE Transactions on Intelligent Transportation Systems, 16(3), 1107–1119.

[19] Talha, A. M., Junejo, I. N. (2014). Dynamic scene understanding using temporal association rules. Image and Vision Computing, 32(12), 1102–1116.

[20] Hadzagic, M., Webb, S., Shahbazian, E. (2013). Maritime Traffic Data Mining Using R. 16th International Conference on Information Fusion. Istanbul: 2041–2048.

[21] Elragal, A., Raslan, H. (2014). Analysis of Trajectory Data in Support of Traffic Management : A Data Mining Approach. 14th Industrial Conference on Data Mining. St Petersburg: 174–188.

[22] Jovi, J. (2017). Implementation of generic algorithm in map-matching model. Expert Systems With Applications, 72, 283–292.

[23] Nguyen, H. L., Woon, Y. K., Ng, W. K. (2015). A survey on data stream clustering and classification. Knowledge and Information Systems, 45(3), 535–569.

[24] Castro, Y., Kim, Y. J. (2016). Data mining on road safety : factor assessment on vehicle accidents using classification models using classification models. International Journal of Crashworthiness, 21(2), 104–111.

[25] Itoh, M., Yoshinaga, N., Toyoda, M. (2016). Word-clouds in the sky: Multi-layer spatio-temporal event visualization from a geo-parsed microblog stream. Proceedings of the International Conference on Information Visualisation. Lisbon: 282–289.

[26] Chu, D., Sheets, D. A., Zhao, Y., Wu, Y., Yang, J., Zheng, M., & Chen, G. (2014). Visualizing hidden themes of taxi movement with semantic transformation. IEEE Pacific Visualization Symposium. Yokohama: 137–144.