CORRECTION OF MEASURED TAXICAB EXHAUST EMISSION DATA BASED ON CMEM MODLE

Qi Li [a], Tao Jia [a, *]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
2016202130018@whu.edu.cn, tao.jia@whu.edu.cn

Commission VI, WG VI/3

KEY WORDS: CMEM model, automobile exhaust emission, correlation coefficient, asynchronous correlation of time series, time delay error

ABSTRACT:

Carbon dioxide emissions from urban road traffic mainly come from automobile exhaust. However, the carbon dioxide emissions obtained by the instruments are unreliable due to time delay error. In order to improve the reliability of data, we propose a method to correct the measured vehicles' carbon dioxide emissions from instrument based on the CMEM model. Firstly, the synthetic time series of carbon dioxide emissions are simulated by CMEM model and GPS velocity data. Then, taking the simulation data as the control group, the time delay error of the measured carbon dioxide emissions can be estimated by the asynchronous correlation analysis, and the outliers can be automatically identified and corrected using the principle of DTW algorithm. Taking the taxi trajectory data of Wuhan as an example, the results show that (1) the correlation coefficient between the measured data and the control group data can be improved from 0.52 to 0.59 by mitigating the systematic time delay error. Furthermore, by adjusting the outliers which account for 4.73% of the total data, the correlation coefficient can raise to 0.63, which suggests strong correlation. The construction of low carbon traffic has become the focus of the local government. In order to respond to the slogan of energy saving and emission reduction, the distribution of carbon emissions from motor vehicle exhaust emission was studied. So our corrected data can be used to make further air quality analysis.

1. INTRODUCTION

The domestic and international research shows that the vehicle exhaust has become the main source of urban air pollution (Davics et al., 2006a). As one of the main ways of carbon emissions, urban traffic accounts for about 25% of the total carbon emissions of human activities (Cao et al., 2010). The construction of low carbon traffic has become the focus of the governments of the world (Castro et al., 2012). Carbon dioxide emissions from urban road traffic mainly come from automobile exhaust (Ge et al., 2011).

In order to respond to the slogan of energy saving and emission reduction, the distribution of carbon emissions from motor vehicle exhaust emission was studied. However, all these studies require accurate emission data.

He, et al. used CMEM to simulate the emission characteristics of light vehicles in Wuhan, compared with the measured data, it found that the overall trend is basically consistent (Chengwei et al., 2008a).

Lagged correlation analysis plays an important role in data mining based on time series (Zuojian et al., 2016a), which can be used extensively in real life such as weather forecast, stock market analysis, moving object tracking, network analysis, and so on. Hailin Li proposed a method of asynchronous correlation analysis based on dynamic time warping, the experimental results demonstrate that the proposed method expands the research of correlation analysis for time series and has a strong robustness (Hailin et al., 2014a).

Since most of the carbon dioxide emissions obtained by the instruments are unreliable due to time delay error (Brand et al., 2008a). To ensure the accuracy of the automobile exhaust emission analysis, we propose a method to correct the measured vehicles' carbon dioxide emissions from instrument based on the CMEM model (Vincenzo et al., 2016a). Firstly, the synthetic time series of carbon dioxide emissions are simulated by CMEM model and GPS velocity data (Zheng et al., 2017a). Then, we automatically identify missing data in the measured data, so the travel data is broken into a series of consecutive time intervals in seconds. Finally, taking the simulation data as the control group, the segment time delay errors of the measured carbon dioxide emissions can be estimated using the asynchronous correlation analysis (Duan et al., 2013a).

2. DATA

Taking the taxi trajectory data within the scope of Wuhan City tricyclic and the Tianhe airport high speed line. We measured for seven consecutive days from November 4th to 12th in 2016, and there are five to eight segments each day from 8:00 am to 9:00 pm, which can be seen in table 1. We used three instruments at the same time during the experiment, they are installed on one taxi, from which we can get three sets of measured data. The following elaborates

all these instruments and their measured data.

The GPS collects one data per second, mainly recording the vehicle's velocity and trajectory data measured every second during the measured time (Hengfeng et al., 2013a). There are about 30,600 locus points during 85 hours' collection. And the trajectories are about 1,200km in total. The driving details and trajectories are given below in Table 1 and Figure 1.

Date	Start time	End time	Distance
2016.11.04	8:01:35	20:30:03	183.7km
2016.11.05	8:03:30	20:32:43	168.4km
2016.11.06	7:52:03	20:03:03	203.9km
2016.11.09	8:00:34	19:04:16	143.8km
2016.11.10	7:53:05	20:54:23	201.9km
2016.11.11	8:20:32	20:23:53	160.6km
2016.11.12	8:28:23	20:12:23	156.5km

Table 1. Trajectory basic information

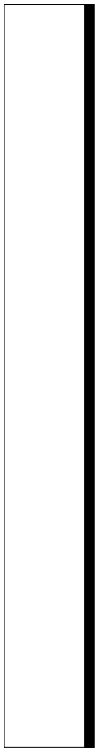

Figure 1. Trajectories

The Optima7 portable flue gas analyser is mainly used to monitor carbon dioxide in the automobile exhaust in this paper. The AFRISO flue gas analyser M60 is used to monitor exhaust gas flow rate. And use the formula below, we can get the data waiting to be corrected:

$$Vg_{co_2} = V_{gas} * \rho_{gas} * S_{pipe} \quad (1)$$

Where V_{gas} is the data obtained from the AFRISO flue gas analyser, ρ_{gas} is the exhaust gas density, and S_{pipe} is Cross sectional area of automobile exhaust duct.

3. METHODOLOGIES

This paper mainly used three methods to correct data error. We used CMEM model to simulate carbon dioxide emissions in real time, then the time delay error of the measured carbon dioxide emissions can be estimated by the asynchronous correlation analysis, and with the help of DTW algorithm, we can finally remove the delay error and extreme outliers of the CO2 emission data obtained by the Optima7 portable flue gas analyser and the AFRISO flue gas analyser M60. The data processing flow is shown in Figure 2.

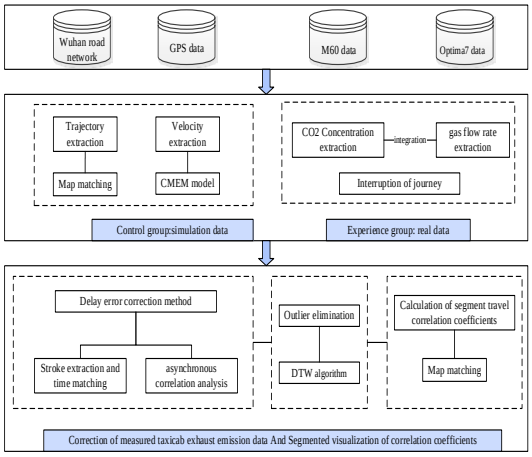

Figure 2. Data processing flow

3.1 CMEM Model

CMEM is a parametric analytical model based on exhaust emissions, it breaks down the emission process into many parts, and each part corresponds to the physical phenomena associated with emission during vehicle running. This process is decomposed and expressed by constructing characteristic parameters. CMEM the basic principle of calculating the content of each component in the exhaust gas is: Firstly, the engine power is calculated by vehicle dynamics principle, The fuel consumption rate is then obtained based on power, speed, and the current air-fuel ratio, The engine emission rate is calculated by the fuel consumption rate and the current burning ratio, Finally, the emission rate of the exhaust pipe is calculated by the engine emission rate and catalyst pass rate. The CMEM model is shown in figure 3.

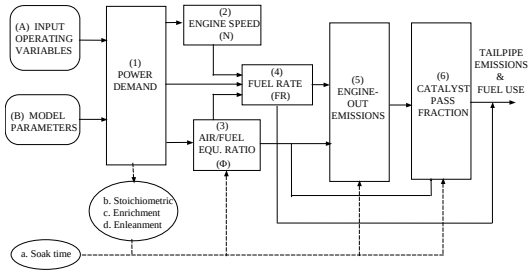

Figure 3. Modal Emissions Model Structure

As we can see in Figure 3, the CMEM model consists of six sub modules [6]. Respectively, the

engine power demand module, the engine speed module, the equivalent fuel to air ratio module, the fuel consumption module, the engine emission module, and the catalyst pass rate module.

3.2 Asynchronous correlation analysis of time series

For two time series Q and C, if the data at the time point t in time series Q has an influence on the data at the time point $t+\Delta t$ in time series C, then it is considered that the data point (Q,t) has a relationship with the data point $(C,t+\Delta t)$. For simplicity, let $(Q,t)=q_i,(C,t+\Delta t)=c_j$ and $r(i,j)=r(q_i,c_j)$ which represents the relationship between data point q_i and data point c_j.

According to the definition, there are three basic situations to reflect the relationship between different data points in time series: a) There is a correlation between synchronous time points, that is $q_i \Leftrightarrow c_j$;b) The data point q_i has an impact on the data point c_j ,that is $q_i \to c_j$;c) The data point c_j has an impact on the data point q_i, that is $c_j \to q_i$.

Therefore, the correlation coefficient of time series A and B can be defined as follows

$$ACA(Q,C)=\sum_{i=1}^{n}\sum_{k=1}^{K_i} r(q_i,c_{jk})=\sum_{j=1}^{m}\sum_{k=1}^{K_i} r(c_i,q_{jk}) \quad (2)$$

Where, K_i or K_j represents that the specific data point of a certain time series has an impact on K_i or K_j continuous data points of another time series. The working principle of ACA is to find two new sequences with equal length between the two time series Q' and C' to make the two time series accurately reflect the asynchronous correlation between the original time series Q and C.

3.3 DTW algorithm

Mining in time series data, dynamic time warping is a method of similarity measurement of strong robustness. Through the relationship between the points at different times, and adjust the sequence corresponding to the element, you can get a curved path. The optimal time series reflects the correlation between the minimum distance and the optimal path corresponding to the bending.

The definition assumes that there are two time series $Q=\{q_1,q_2,\cdots q_m\}$ and $C=\{c_1,c_2,\cdots c_m\}, m\neq n$,which means the relationship between the time series of DTW can deal with different length, and can find the optimal path in bending $P=\{p_1,p_2,\cdots p_k\}$ to obtain the minimum distance metric $DTW(Q,C)$, that is,

$$DTW(Q,C)=\min_P \sum_{l=1}^{k} d(p_l) \quad (3)$$

Where $d(p_l)=d(i,j)=d(q_i,c_j)=(q_i,c_j)^2$, in the distance values of the corresponding sequences, between the different elements of the curved path. At the same time, the curved path must satisfy the boundary and continuity and monotonicity, the optimal bending path starting and ending in sequence starting and end point position, and the cost matrix of p_{l+1} may only appear in the three cell in the upper left corner of the in p_l adjacent. In order to solve the equation (1), the dynamic programming method can be used to obtain the optimal bending path and minimum bending distance.

4. ANALYSIS AND RESULTS

4.1 Interruption of journey

In the process of recording data, data loss is bound to exist. Therefore, when there is a missing data, we interrupt the trip once. Firstly, we identify missing data in the GPS and interrupt the recorded interrupt location at the missing point. And then, on the basis of the interruption trip of the GPS record, the time discontinuity in the Vg_{co_2} is identified, the interruption position is recorded again, and the segmentation is counted. Take the first day's data as an example, section results are shown in Table 2.

Trip	Table Column Head		
	Segments	*Start Time*	*End Time*
1	47	8:17:39	9:40:04
2	20	11:06:18	12:22:32
3	1	13:20:26	14:22:26
4	31	14:44:25	15:26:57
5	100	15:47:20	17:19:20
6	24	18:50:53	20:47:49

Table 2. Segmentation results

As we can see in the Table 2, the trip is interrupted into many segments, and the amount of data in each section is greatly reduced, which provides great convenience for correlation analysis.

4.2 The elimination of segment delay error and the outliers

Take the first day's journey as an example, in Figure 4, its driving trajectory is showed by the red line, and the other six days' trajectory is showed by black lines. We can see that we have almost collected all the roads within the tricyclic Wuhan City.

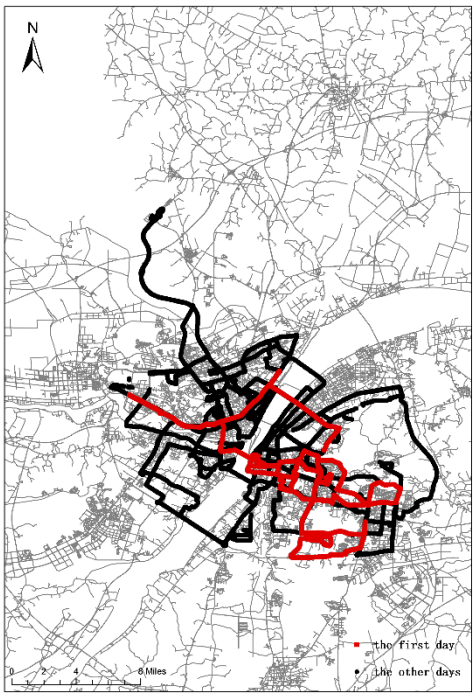

Figure 4. The first day's driving trajectory

For one day's data, after a number of interruption, we can refer to the simulated data obtained by the CMEM model, and doing asynchronous correlation analysis for each stroke. And then in each stroke, we figure out outliers and make modifications to recalculate the correlation coefficients between the two sets of data.

To make an analysis, the first day's first segment's asynchronous correlation analysis result is show in Figure 5. There are 47 asynchronous correlation coefficients between control group and the measured taxicab exhaust emission data. Each blue line represents a package trip. We found out that most of the segments have high correlation coefficient with the simulated data, which means no time delay error in these segments. There are only five segments need to eliminate time delay error. In fact, our instrument is sensitive, and when the correlation coefficient appears the first maximum, the data has reached a strong correlation, so we take the first peak as the delay error.

After the elimination of segment delay error, we found out the outlier and correct them. In Figure 6, before correcting outliers, the correlation coefficients is in the blue line. It is easy to see that there are five segments show low correlation with simulated data, which is consistent with the results shown in Figure 5. And after the outliers' eliminate, the correlation coefficients is in the red line. And at the same time, in the rejection of outliers, we control the corrected outliers, which account for less than 5% of the total data points. It shows that the correlation coefficients have been improved definitely and most segments reached strong correlation.

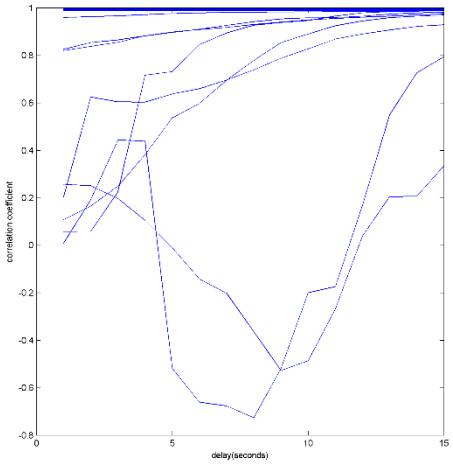

Figure 5. Asynchronous correlation coefficient

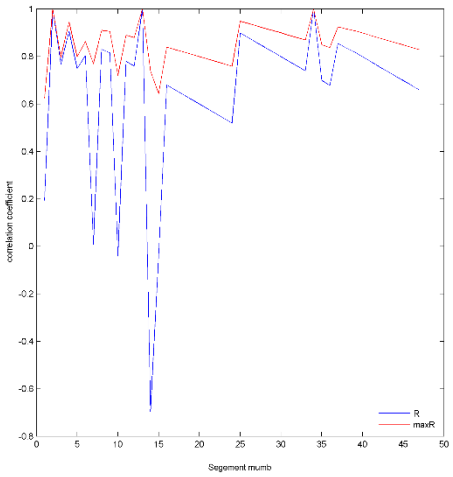

Figure 6. The outliers' elimination effect

4.3 Analysis of error correction results

In the process of data collection, there are a lot of reason to produce errors and it is impossible to avoid them. After the elimination of segment time delay error and the outliers, it is easy to see that we can get a better group of data.

First, the asynchronous correlation coefficients before error correction are added to the attribute fields of the trajectory data. Then, the asynchronous correlation coefficients after error correction are added to the attribute fields of the trajectory data. Use different colours to indicate the different degree of correlation, so we can see the correct effect more intuitive.

In Figure 8, there are 14 pictures. The picture in the same row represents the same day's data. The first to seventh row representing the first day to the seventh day respectively. The first column represents the correlation coefficient distribution before data correcting, and the second column represents the correlation coefficient distribution before data

correcting. In each picture, there are three kinds of lines in different colour. The red one represents low correlation coefficients, which means the correlation coefficients is lower than 0.4. The green one represents moderate correlation coefficients, which means the correlation coefficients is 0.4-0.6. The red one represents high correlation coefficients, which means the correlation coefficients is 0.6-1.

We have statistics on the two sets of data. One set is before the error correction, the proportion of routes with less asynchronous correlation. The other group is after error correction, the percentage of routes with less asynchronous correlation. The result shows in figure7.

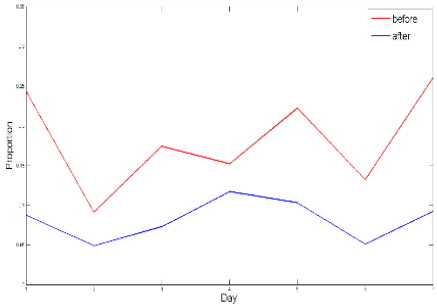

Figure 7. The data correction result

As we can see in Figure 7, more than half of the low correlation coefficient segments have been well corrected. Then we visualize the result in Figure 8. After correcting the data, the red trajectories are obviously reduced, and the green and blue trajectories are increased equally.

It is obvious that we can get higher correlation coefficients after data correction. The trend of the measured data is closer to the theoretical simulation data, which indicates that the availability of actual measurement data is greatly improved.

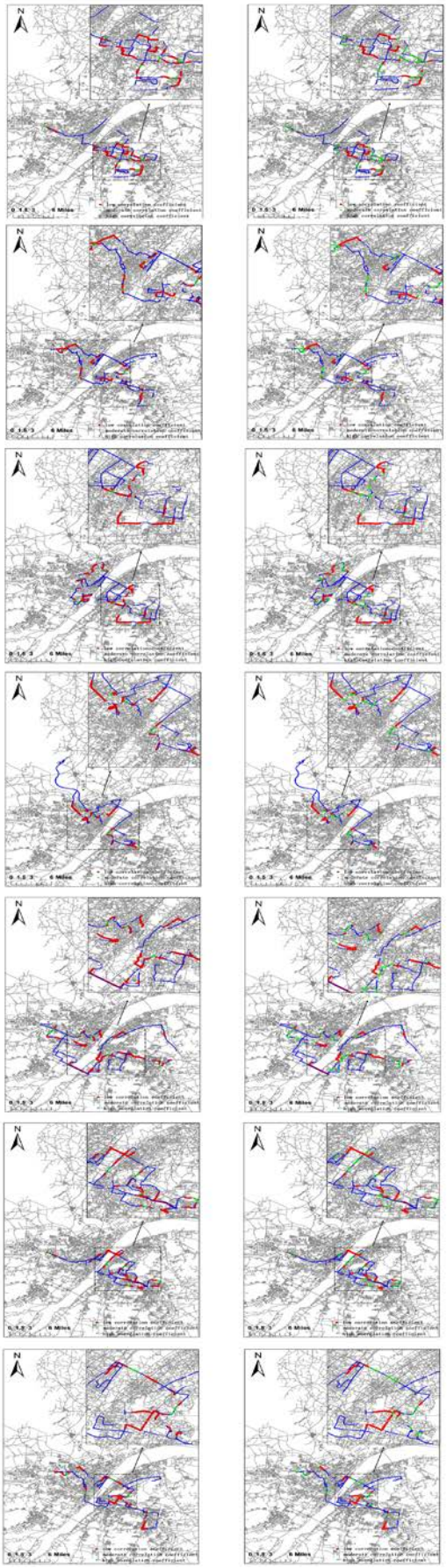

Figure 8. Trajectory data comparison

5. CONCLUSION

In this study, we propose a method to correct the system delay error based on asynchronous correlation analysis. By applying this method to our measured data, we found that the correlation coefficient between the time series of the corrected data and the control group were significantly improved. The availability of actual measurement data is greatly improved, which suggested the feasibility of our method. The corrected emission data can be used to study the spatiotemporal visualization analysis, which will benefit the sustainable urban planning, such as urban traffic planning and environmental protection.

However, there are still some deficiencies in the current research. In the actual measurement, time delay of the measured time series is dynamic, but we assumed that they are not changing in one segment. To be exact, the time series correction method based on dynamic time warping needs further investigation.

ACKNOWLEDGMENTS

This research was funded by the National Natural Science Foundation of China (Grant NO. 41401453).

I express my gratitude to Chu Ma, Yuqian Li, Kai Chen and Xuesong Yu, who have given me constant help and informative suggestions. My sincere thanks also go to the anonymous referees for their careful work and thoughtful suggestions which will help me improve my study substantially.

REFERENCES

Brand C, Boardman B, J., 2008a. Taming of the few—The unequal distribution of greenhouse gas emissions from personal travel in the UK. *Energy Policy*, 36(1), pp. 224-238.

Cao L, Krumm, 2010. From GPS traces to a routable road map. *The Workshop on Advances in Geographic Information Systems*, pp 3-12.

Castro P S, Zhang D, Li S, 2012. Urban Traffic Modelling and Prediction Using Large Scale Taxi GPS Traces. *Pervasive Computing*, pp. 57-72.

Chengwei Xu, Chaozhong Wu, Xiumin Chu, Jing Gong, Zhang k, J., 2008a. Light vehicle in Wuhan city based on CMEM model of average emission factors of. *Traffic and computer*, 04, pp. 185-188.

Davics J J, Beresford A R, Hopper A, J., 2006a. Scalable, Distributed, Real-Time Map Generation. *Pervasive Computing IEEE*, 5(4), pp. 47-54.

DuanHong, YangLuo, J., 2013a. A Gesture Trace Detection Method Using DTW. *Applied Mechanics and Materials*, 2617(380), pp 72-73.

Ge Y, Liu C,Xiong H, et al, 2011. A taxi business intelligence system. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, Ca, Usa, August, pp 735-738.

Hailin Li J., 2014a. Asynchronous correlation analysis of time series based on dynamic bending. *Computer application research*, 07, pp. 1976-1979.

Hengfeng Li, Lars Kulik, Kotagiri Ramamohanarao, J., 2015. Robust inferences of travel paths from GPS trajectories. *International Journal of Geographical Information Science*, 29(12), pp. 233-250.

Jing Wang, Chaoliang Wang, Xianfeng Song, Venkatesh Raghavan, J., 2016. Automatic intersection and traffic rule detection by mining motor vehicle GPS trajectories. *Environment and Urban Systems*, pp. 31-32.

Vincenzo Barrile, Maria Nadia Postorino, J., 2016a. GPS and GIS Methods to Reproduce Vehicle Trajectories in Urban Areas. *Procedia Social and Behavioral Sciences*, 223, pp. 123-124.

Zhe Jiang, Michael Evans, Dev Oliver, Shashi Shekhar, J., 2016a. Identifying K Primary Corridors from urban bicycle GPS trajectories on a road network. *Information Systems*,57, pp. 34-35.

Zheng Zhang, Romain Tavenard, Adeline Bailly, Xiaotong Tang, Ping Tang, Thomas Corpetti, J., 2017a. Dynamic Time Warping under limited warping path length. *Information Sciences*, 393, pp. 111-132.

Zuojian Zhou, Wanchun Dou,Guochao Jia,Chunhua Hu, Xiaolong Xu, Xiaotong Wu, Jingui Pan, J., 2016a. A Method for Real time Trajectory Monitoring to Improve Taxi Service Using GPS Big Data. *Information and Management*, pp. 21-24.