Journal of Big Data

RESEARCH

Open Access

A fuel consumption-based method for developing local-specific CO_2 emission rate database using open-source big data

Linheng Li[1,2], Can Wang[1,2], Jing Gan[2,3*] and Dapeng Zhang[4]

*Correspondence:
jinggan1026@163.com

[1] School of Transportation, Southeast University, Nanjing 211189, China
[2] Institute On Internet of Mobility, Southeast University and University of Wisconsin-Madison, Southeast University, Nanjing 211189, Jiangsu, China
[3] School of Modern Posts, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
[4] School of Management Science and Engineering, Southwestern University of Finance and Economics, Chengdu 611130, China

Abstract

Emission data collection has always been a significant burden and challenge for Chinese counties to develop a CO_2 emission inventory. This paper proposed a fuel consumption-based method to develop a local-specific CO_2 emission rate database for Chinese counties using only open-source big data. Localized vehicle fuel consumption data is obtained through natural language processing (NLP) algorithm and large language model (LLM). The emission rates derived by our proposed method are consistent with field test results in literature. Besides, the CO_2 emission estimation results using local-specific traffic activity data indicate that our method could effectively improve the accuracy of vehicle emission assessment. Compared with conventional method, the novel approach proposed in this paper can provide a pathway for convenient, universal, and cost-saving assessment for local scale CO_2 emission rates. With this method, it is possible to formulate a local-specific CO_2 emission database in various Chinese counties using only open-access big data.

Keywords: Emission rates, Big data, Natural language processing, Localized fuel consumption

Introduction

Increasing demand for transportation activities has led to the combustion of more fossil fuels such as gasoline and diesel, which in turn has increased the release of CO_2. The issue of climate change caused by carbon emissions has become a hot spot of high concern in the international community. Road transport is a significant contributor to urban emissions, the carbon emission control of road transport has become an essential area for carbon reduction in every country around the world [17, 25, 42]. The Chinese government has also been committed to triggering the development of low-carbon road transport in recent years to reduce the CO_2 level in the atmosphere. These efforts not only contribute to environmental management but also promote effective progress in environmental management as a whole.

Accurate estimation of CO_2 emissions is a prerequisite for developing effective CO_2 emissions control strategies. To fight against climate change and take more effective climate-positive actions, some governments began to promote fine-grained low-carbon management, which needs higher resolution CO_2 emission rates database as data support. Many scholars show great passion for higher temporal resolution research on motor vehicle CO_2 emission [12, 20, 33]. However, emission data collection has always been a significant burden and challenge for Chinese counties to develop a comprehensive CO_2 emission inventory. Benefiting from the outstanding advantages of network big data and machine learning algorithms for data capture and analysis, the main goal of this paper is to explore a convenient, universal, and cost saving way for the establishment of a local scale CO_2 emission rate database for Chinese counties based on open-source network data.

Under the motto of "can we do more with less manpower and financial resources", this paper makes a contribution to the literature by developing a cost-saving method to construct a comprehensive CO_2 emission rate database using open-source big data and machine learning techniques. The use of machine learning techniques overcomes the limitations in big data acquisition, such as low efficiency, high manual cost, and small data volume. This method eliminates the most significant barrier for counties that are subjected to the lack of enough reliable laboratory or field test data, which could be further applied to compile the CO_2 emission inventories in any Chinese counties. Especially for counties suffering from data-sparse situations, our proposed methodology could be a good alternative method before a slew of reliable data is available from enormous field tests or laboratory tests.

Literature review

A comprehensive CO_2 emission rates hinges on the acquisition of vast amounts of emission data. The main approach adopted is to collect second-by-second vehicle dynamics data and tailpipe emissions data from various field-tests or laboratory-tests. Based on these big data, extensive and complicated statistical analysis is conducted to obtain the emission database. Undoubtedly, this is a massive undertaking that requires a significant investment of human and material resources [30, 36, 39, 50]. Zhang et al. [48] designed 16 dynamometer engine tests to collect exhaust emissions for one heavy-duty diesel vehicle under different test cycle and fuel property. A portable emissions measurement system (PEMS) was employed to collect second-by-second CO_2 emission data for heavy-duty transit buses in Beijing [47], light-duty passenger cars in Guangzhou, Beijing, and Macao [46], light-duty vehicles in Beijing [40, 45]. Gao et al. [13] selected 5 testing vehicles and established vehicle emission factors based on on-road measurement. Pechout et al. measured exhaust emission from four large motorcycles with portable FTIR [26]. Chandrashekar et al. [5] selected two diesel auto-rickshaws and collected their CO_2 emission factors in India using PEMS. However, these data are not publicly available for other researchers and are limited to the testing of certain specific vehicle categories and models. More importantly, these road tests can only provide a snapshot of driving conditions at one point in time. The data obtained by this method cannot cover a broad set of conditions drivers may experience throughout the year and represent the national average, real-world driving. Therefore, this method is only suitable for studying emissions in a specific city.

In European and American countries, a significant amount of work has been done by environmental research institutions and scholars to create a comprehensive CO_2

emission rate database. Based on a large amount of experimental data, they have developed some mature vehicle emission simulation models, such as Motor Vehicle Emission Simulator (MOVES) [8], International Vehicle Emission Model (IVE) [18], Greenhouse gases, Regulated Emissions and Energy use in Transportation(GREET) [14], EMission FACtor (EMFAC) [6] and COPERT [7].Among these models, MOVES developed by the U.S Environmental Protection Agency (EPA) has become one of the most widely used emission calculation models in the world, thanks to its open-source emission rate database, user-friendly interface and increased customization for international applications [1, 22, 24, 29, 38].

In China, strenuous and ceaseless efforts have been made to develop motor vehicle CO_2 emission inventory, such as Beijing, Shenzhen, Chengdu, Nanjing. But the limited on-road tests in a few megacities makes it challenging to develop a sophisticated model like MOVES. More importantly, most academic research on the modeling and analysis of road transport CO_2 emissions is also conducted at national or provincial levels [3, 41, 44]. Except for these economically developed cities, the establishment of CO_2 emission inventory for most small and medium-sized cities in China is still in its infancy. However, developing the CO_2 emission inventory for each county through the on-road tests method will be a massive project and almost unrealistic.

With a consideration of data collection burden for establishing CO_2 emission rate database in various Chinese counties, the MOVES database may be more suitable for our purposes than field data would be. However, MOVES model is explicitly developed based on local special traffic conditions, vehicle technical conditions, driving behavior, emission regulations, and climatic conditions. Hence it cannot be directly applied to Chinese cities, which may result in significant deviations in emission estimation if without calibration [28, 43].

Vehicle emission is determined by the emission rates and the vehicle's activity parameters (e.g., speed, acceleration, and vehicle travel kilometers). Speed is the most important factor affecting emission rates. By multiplying the vehicle's baseline emission rate (BER) with a speed-based adjustment value, the vehicle's emissions at different speeds can be estimated. This adjustment value is called the speed correction factor (SCF). BER and SCF modeling approach is adopted by many traffic emissions models (e.g., EMFAC model, and MOVES model) and scholars [27, 32] to describe the impact of vehicle speed on vehicle emissions. In order to obtain more accurate SCF curves, traditional methods involve extensive vehicle testing and complex statistical analysis calculations. The SCF function is usually obtained through extensive experimental data, complex classification processing, statistical analysis, and mathematical regression. The BER, likewise, is derived from an extensive collection of on-road vehicle testing data. Therefore, establishing a localized emission factor database for Chinese cities through traditional experimental methods is a time-consuming and laborious process, and from an economic perspective, it requires a significant amount of funding to support a sufficient number of real vehicle tests.

With the development of the Internet, network big data is experiencing explosive growth. In addition to some official websites of car manufacturers that publicly release car-related information data, more ordinary users are willing to share their real driving data online [10, 19, 34]. For example, the FUELLY website in the United States and the

BearOil app in China, where a large number of car owners will actively upload their real fuel consumption data. However, manually searching and analyzing this data requires a lot of manpower and resources. If it were possible to automatically obtain and utilize big data from the internet to construct localized CO_2 emission rate databases, it would eliminate the need for labor-intensive and complex data collection and processing, enabling convenient and cost-saving database construction.

In recent years, the explosive development of artificial intelligence algorithms has made it possible to automatically acquire massive amounts of data, thereby facilitating the resolution of this issue [4, 9, 16, 21, 31]. Especially with the rapid development of large language models such as the recent popular ChatGPT, more and more data can be utilized with less human intervention required [35, 37, 49].

Although the studies mentioned above offer valuable insights into the construction of a comprehensive CO2 emission rates database, our review has identified several critical gaps in the existing literature on emission rates. Firstly, much of the previous research has concentrated on collecting emission rate data from only a few vehicles within specific cities. This approach overlooks the need for a more efficient and universally applicable method that could facilitate carbon emission factor construction across various locales. Such reliance on limited data sources may not only be highly unreliable but also lead to diminished accuracy in emission assessments, undermining the potential for universal application across different geographic contexts. Secondly, there is a significant dependence on extensive on-road test data in previous studies. Conducting such expansive experiments in every city to establish a localized emission rate database represents a colossal challenge, both in terms of logistics and financial resources. This traditional method's practicality and feasibility are thus called into question, highlighting the necessity for innovative approaches that can circumvent these substantial barriers.

In response to these identified gaps, this paper introduces a novel methodology that employs natural language processing (NLP) algorithms and large language models (LLMs) to acquire localized vehicle fuel consumption data and derive the baseline emission rate (BER). Further, we utilize normalization processing of emission rates in MOVES, using baseline fuel consumption data to derive the speed correction factors (SCFs) of the emission rates under different speed bins. This innovative approach is a significant departure from traditional methods in the field of CO_2 emission rate database establishment. By leveraging advanced technologies, we aim to overcome the major hurdles faced by counties with insufficient field test data, thereby offering a pathway for a more convenient, universal, and cost-effective assessment of local-scale CO_2 emission rates. By addressing the critical research gaps identified through our literature review, our approach signifies a pivotal advancement towards establishing high-resolution, localized CO_2 emission databases. It underscores the potential of utilizing big data and artificial intelligence to enhance the accuracy and applicability of emission assessments, paving the way for more effective carbon management strategies on a local scale.

This paper is organized as follows. "Preliminary analysis" analyze the possibility and conditions of using MOVES for constructing an emission rates database in China. "Methodology" presents the methodology. "Results and discussion" displays and discuss the model performance. "Applications and future scopes" presents the future scope of the method proposed in this paper and the conclusion is summarized in "Conclusions"

Preliminary analysis

The relationship between CO$_2$ emissions and fuel consumption

In order for an internal combustion engine to drive a vehicle along the road, it must convert the energy stored in the fuel into mechanical energy to propel the wheels. In the combustion reaction, the carbon from the fuel combines with oxygen from the air to produce CO$_2$. Hence, the amount of CO$_2$ a vehicle emits is directly related to the amount of fuel it consumes, showing a nearly linear relationship. This chemical principle is also the calculation principle of CO$_2$ emissions in MOVES. According to the existing literature [15, 23, 46, 47], the CO$_2$ emissions could be derived from the fuel consumption data through the carbon balance method, as shown in Eq. (1).

$$\overline{EF}_{i,g,y} = \frac{1}{100} \cdot \frac{44}{12} \cdot \overline{FC}_{i,g,y} \cdot \rho_g \cdot \gamma_c \tag{1}$$

where $\overline{EF}_{i,g,y}$ is the average CO$_2$ emission rates (*g/km*) for vehicle type i powered by gasoline and produced in year y; $\overline{FC}_{i,g,y}$ is the average fuel consumption (*L/100 km*) for vehicle type i powered by gasoline and produced in year y; ρ_g is the density of gasoline, 740 g/*L* is adopted in this study. γ_c is the carbon mass fraction of gasoline, 0.87 in this study.

To this end, this study adopted a fuel-consumption based method for developing a comprehensive local-specific CO$_2$ emission rates database in Chinese counties using the emission rates embedded in MOVES. The linear relationship between CO$_2$ emissions and fuel consumption is the basic principle of our fuel consumption-based approach.

The emission rates under different average speeds

The emission rates for every average speed are adjusted using the SCF that represents the relative changes in emission rates under different average speeds. Note that although the issue of the impact of acceleration and deceleration on fuel consumption arise throughout the modeling process, the average speed-based modeling method is generally used before a large amount of reliable data is available. Therefore, in this paper we will only address the SCF under different average speeds.

The SCF in MOVES can be obtained if we divide the emission rates for 16 average speed bins embedded in MOVES by the baseline emission rate that can be derived from the average (baseline) fuel consumption as shown in Eq. (1). This process is called the normalization of the MOVES emission rates. In this study, we assume that the average fuel consumption data corresponds to the average speed of the vehicle. In general, the SCF should be close to 1.0 when speed is equal to the average speed, and it is disaggregated by the vehicle type when modeling. Therefore, it is reasonable to assume that if the average speed of the vehicle driving on the same road type (e.g., urban roads, urban expressway, and highway) is similar between Chinese counties and the U.S., the CO$_2$ emission rate SCF of the same vehicle type used in MOVES could be transplanted to Chinese counties for the modeling of local-specific CO$_2$ emission rates. The requirement of a similar baseline speed is to ensure that the SCF should be close to 1.0 at the baseline speed in different cities. The requirement for the identical road type is also to dampen the influence of driving cycles with large differences. Note that although different driving cycles will produce different fuel consumption of different even though with the same average speed, the local average fuel consumption data can dampen this issue. Furthermore, compared with other pollutants (e.g., CO, HC), CO_2 is less sensitive to differences in the distribution of operating mode bins for cycles with similar average speed. The cycle correction factors (CCF) for CO_2 emissions for different cycles with similar average speed ranges from 0.2% to 6.9% in the findings of a project carried out by North Carolina State University and the University of Utah. The CO and HC CCFs for different cycles with similar average speed are 9% to 72% higher, and 3% to 63% higher, respectively [2, 11]. This is also the reason that the average-speed based method could be adopted for CO_2 emissions rates modeling. Further validation and discussion of this assumption are provided in "Applications and Future Scopes".

With the consideration of data availability, this research focuses on the emission rates of light duty passenger vehicles (LDPVs) under urban driving conditions (vehicle speed ranges from 0 to 60 km/h). We choose the urban fuel consumption data from laboratory operating condition test data (EPA Urban Dynamometer driving schedule FTP-75), which are sources of the urban emission rates data of MOVES. Its average speed is about 19.59mph (31.53 km/h), which is similar to the average speed on the urban roads of Chinese counties (approximately 30 km/h). Therefore, the urban fuel consumption data in the U.S (laboratory test data) and Chinese counties (real-world data) can be used to characterize differences in vehicle emission rates. It is worth noting that the urban dynamometer driving schedule of China Light-duty vehicle Test Cycle (CLTC) used in China is quite different from FTP-75. Besides, the test is at a constant speed for most of the time, which is quite different from the actual urban driving situation. This is the reason why we did not choose the fuel consumption data of the urban cycle from laboratory tests in China for local-specific emission rates modeling.

Methodology

Object of study

In MOVES, LDPVs are defined as cars with a gross vehicle weight rating of fewer than 4250 kg, including passenger cars and passenger trucks. The sedans, SUVs in China belong to the category of passenger cars and passenger trucks in MOVES, respectively. In all modes of road transport, the proportion of sedans and SUVs trips has always been at a high level in Chinese counties, due to the lagging development of public transportation and the lack of public transportation infrastructure. Therefore, it is particularly important to assess the CO_2 emissions estimation deviation of LDPVs applying MOVES in Chinese counties and find a simple method to establish emission rates model for LDPVs in Chinese counties using MOVES.

Automobile CO_2 emissions come from multiple processes during vehicle driving, such as the start process (i.e., emissions when starting a vehicle) and the running process (i.e., emissions after the vehicle is warm). Among all processes, CO_2 emissions during the running process from LDPVs contribute to most of the urban CO_2 emissions. Therefore, this study mainly focuses on CO_2 emissions during running operating modes emitted by LDPVs. In this study, MOVES3, the latest version is adopted to analyze the emission rates of LDPVs.

Research domain and framework

The county is the third administration level in China, where the administrative tiers are provincial, prefectural, county, and township level from high to low. In order to explore the characteristics of motor vehicle fleet composition, vehicle engine displacement distribution, and traffic activity in Chinese counties, four representative counties (Changxing county, Jintang county, Qingcheng county, and Wuan county) from the southeast and northwest of China are selected for the actual household survey and data collection. Changxing county, known for its balance between industrial activities and natural reserves, exemplifies regions undergoing rapid urbanization while maintaining significant green spaces. Jintang county, with its robust agricultural base, reflects the characteristics of rural economies transitioning towards more diversified economic structures. Qingcheng county, a tourist destination with a relatively high population density, represents urban areas with significant vehicular traffic due to tourism. Wuan county, with its extensive mining industry, illustrates areas where industrial activities predominantly influence traffic patterns and vehicle emissions. The topographic characteristics of these counties vary from mountainous regions in Qingcheng to the plains in Jintang, providing insights into how geography impacts vehicle usage and emissions. Economically, these counties span from high industrial activity in Changxing and Wuan to the more agriculture-focused economy in Jintang, offering a glimpse into the diverse economic factors influencing vehicle fleet composition and engine displacement distribution. This variety ensures that the study encompasses a broad spectrum of traffic activities, from urban congestion in Qingcheng to the more dispersed traffic in rural Jintang. The topographic characteristics and economic development levels of these four counties are different. Therefore, it is assumed that they can basically represent the general traffic situation of Chinese counties.

Therefore, it is assumed that these counties can fundamentally represent the general traffic situation of Chinese counties, encompassing the variability in motor vehicle fleet composition, vehicle engine displacement distribution, and traffic activity. This selection strategy allows for a nuanced understanding of CO2 emissions and the development of effective local environmental policies and carbon management strategies across different county contexts in China.

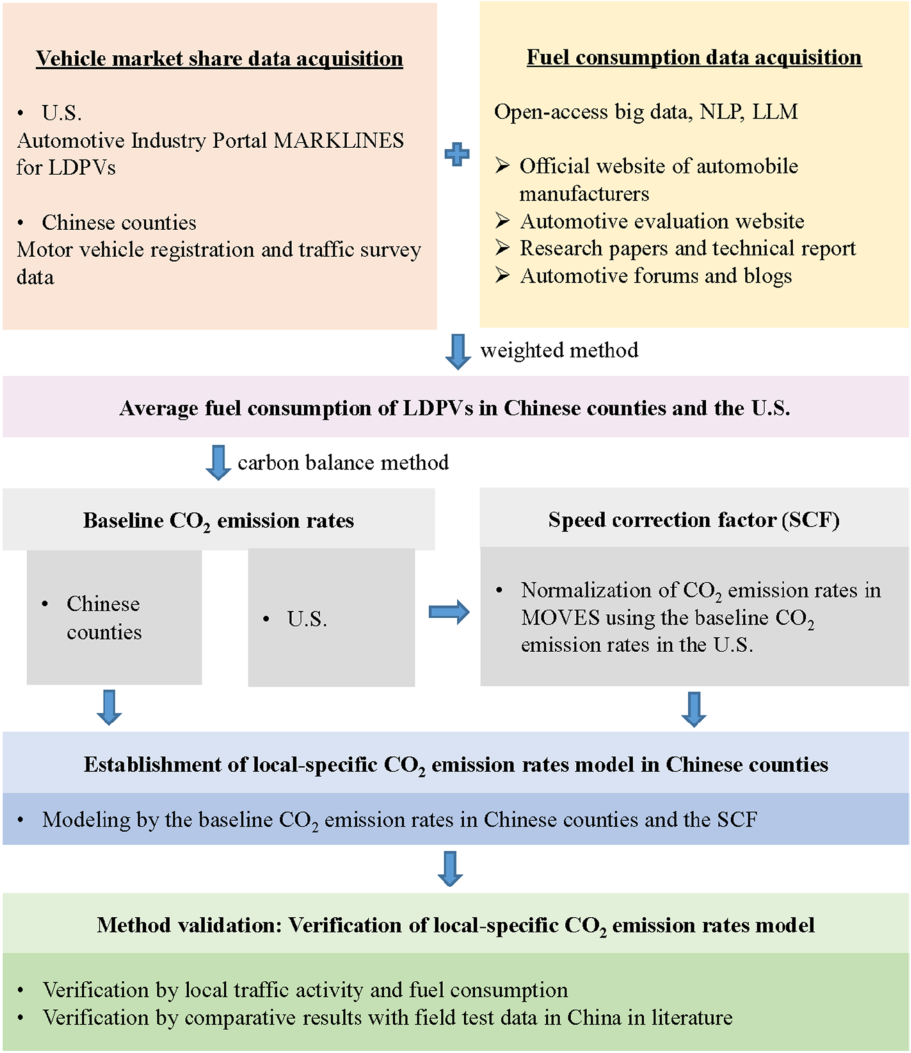
In this study, the local average CO_2 emission rates (baseline emission rate) and its corresponding speed correction factor (SCF) were applied to model local CO_2 emission rates in Chinese counties. First, leveraging the unique advantage of machine learning algorithms in acquiring data, this study utilized NLP algorithms and LLM to search and acquire fuel consumption data for various vehicle models from publicly available sources on the internet. Combining this data with the market share of each vehicle model, a weighted average method was employed to derive the localized average fuel consumption for each model. Then, the average (baseline) CO_2 emission rates were derived from the average fuel consumption through the carbon balance method. The SCF was derived by normalizing the emission rates in MOVES using the baseline CO_2 emission rates in the U.S., which will be transplanted to Chinese counties. And then, combining the baseline CO_2 emission rates in Chinese counties with the SCF, we developed a local-specific CO_2 emission rate model in Chinese counties. Finally, the proposed method was validated by local-specific traffic activity data and the emission rates findings from field tests in China by [46]. The flow diagram of the methodology described above is shown in Fig. 1.

Quantifying fleet average fuel consumption

Acquiring the market share data

To explore the characteristics of LDPVs fleet composition and engine displacement distribution, we carried out an actual household survey and data collection in four representative counties (Changxing county, Jintang county, Qingcheng county, and Wuan county). And the vehicle registration data over the past decade (2009–2021) was obtained from the Vehicle Management Office in each county.

For the U.S., the LDPVs' registration data is not available for us, and the national sales data of sedans and SUVs were adopted, respectively. MarkLines provides membership information services for the automotive industry information around the world.



Vehicle market share data acquisition

- U.S.

Automotive Industry Portal MARKLINES for LDPVs

- Chinese counties

Motor vehicle registration and traffic survey data

Fuel consumption data acquisition

Open-access big data, NLP, LLM

- Official website of automobile manufacturers
- Automotive evaluation website
- Research papers and technical report
- Automotive forums and blogs

weighted method

Average fuel consumption of LDPVs in Chinese counties and the U.S.

carbon balance method

Baseline CO_2 emission rates

- Chinese counties
- U.S.

Speed correction factor (SCF)

- Normalization of CO_2 emission rates in MOVES using the baseline CO_2 emission rates in the U.S.

Establishment of local-specific CO_2 emission rates model in Chinese counties

- Modeling by the baseline CO_2 emission rates in Chinese counties and the SCF

Method validation: Verification of local-specific CO_2 emission rates model

- Verification by local traffic activity and fuel consumption
- Verification by comparative results with field test data in China in literature

Fig. 1 Flow diagram of the methodology

Information on the global automotive industry can be queried on its website, such as the production and sales data of each model of each manufacturer in more than 60 countries worldwide. Therefore, the market share data of each model of LDPVs in the U.S. can be collected.

Obtaining fuel consumption for each vehicle model

Localized fuel consumption data is the core basis for constructing a localized emission factor database. How to obtain localized fuel consumption data conveniently and accurately is crucial to ensure the effectiveness of the localized emission factor database and to improve the accuracy of road traffic carbon emission estimation. With the popularization of the Internet, automotive evaluation websites publicly disclose fuel consumption evaluation data, and there are more car owners willing to share their real driving fuel consumption data on some apps or websites.

The emergence of artificial intelligence technologies, such as open-source big data and natural language processing, presents opportunities for conveniently acquiring large-scale, localized fuel consumption data. This paper proposes a method for obtaining localized fuel consumption data based on NLP algorithms and LLM. The flowchart of the method is shown in Fig. 2. Compared with traditional manual search and statistical methods for obtaining fuel consumption data, the method has the following characteristics and advantages: firstly, it can collect a large amount of fuel consumption data from different sources of network reports, expand the data sources, ensure that the sample size of fuel consumption data is large enough, and improve the accuracy of fuel consumption level evaluation; secondly, it reduces the amount of manpower and material resources required for manual data collection; thirdly, it can easily achieve rolling updates of data. It should be noted that in the process of selecting open-access data sources for our analysis, we first undertook a manual evaluation process. This

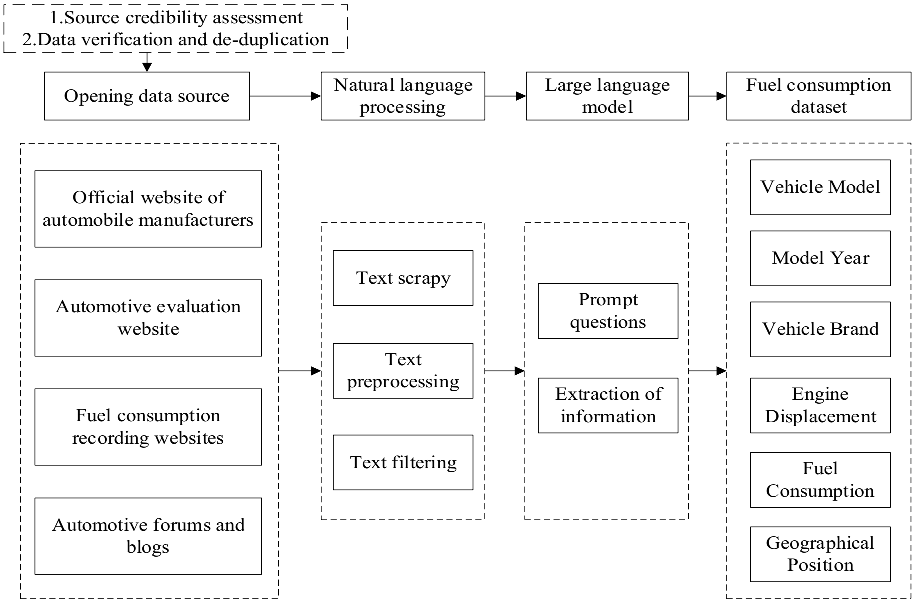

1.Source credibility assessment
2.Data verification and de-duplication
Opening data source
Natural language processing
Large language model
Fuel consumption dataset
Official website of automobile manufacturers
Automotive evaluation website
Fuel consumption recording websites
Automotive forums and blogs
Text scrapy
Text preprocessing
Text filtering
Prompt questions
Extraction of information
Vehicle Model
Model Year
Vehicle Brand
Engine Displacement
Fuel Consumption
Geographical Position

Fig. 2 Workflow of localized fuel consumption data acquisition method

process involves assessing the credibility, relevance, and comprehensiveness of the data provided by various platforms. We prioritize data sources based on their authority (e.g., government entities and reputable automotive research institutions), the frequency of updates, as well as the geographic and vehicle type coverage they offer. Additionally, the level of user engagement and the volume of available data are considered as indicators of the reliability and representativeness of the data sources. Moreover, to ensure that bias does not occur in the process of collecting open-source big data, we employed two methods for data preprocessing: source credibility assessment and data verification and de-duplication. In source credibility assessment, we conducted a comprehensive evaluation of the credibility of each data source, giving priority to data from official and established websites over secondary sources such as blogs. This approach helps to minimize the inclusion of duplicated or less reliable data. Subsequently, in data verification and de-duplication, we used advanced data processing techniques (including natural language processing algorithms) to identify and remove duplicate entries. This step ensures that identical data points, originating from different sources but referring to the same information, are counted only once.

Step 1: Use Scrapy to scrape text data from relevant websites. Fuel consumption data websites include car manufacturer official websites, car review websites (such as Car and Driver, MotorTrend, Autohome, etc.), fuel consumption recording websites (such as FUELLY, BearOil), car forums, and blogs (Use source credibility assessment and data verification and de-duplication simultaneously to ensure that bias does not occur in the process of collecting open-source big data).

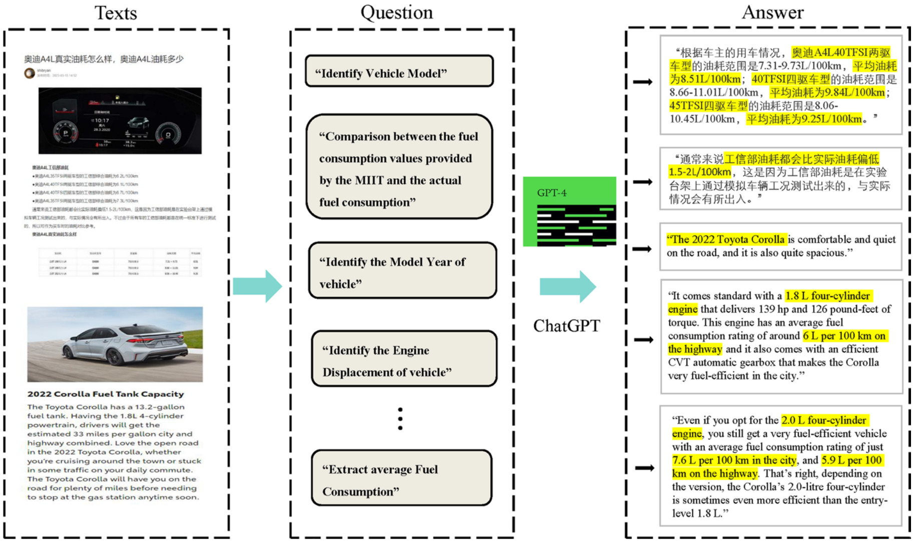
Step 2: Preprocess the text data, including removing HTML tags, JavaScript code, CSS styles, etc.

Step 3: Text filtering. Predefined vocabulary related to fuel consumption including "fuel consumption", "fuel efficiency", "*L/100 km*", and "fuel consumption rate". The TF-IDF algorithm is then applied to filter out reports or text snippets containing the predefined vocabulary.

Step 4: Information extraction: Designing effective prompt questions to extract car model, brand, and fuel consumption data from text using large language models such as ChatGPT. Figure 3 shows the flowchart of text information extraction using ChatGPT.

Step 5: Manual verification. Manually verify the extracted fuel consumption data, delete obviously incorrect data, and then calculate the fuel consumption of each car model.

Through the process depicted in Fig. 3, it becomes evident that by employing advanced data processing techniques in our study, including the use of Natural Language Processing (NLP) algorithms and Large Language Models (LLMs), we can automatically extract and process vast amounts of data from open-access sources. This significantly reduces the manual labor traditionally associated with such tasks. Leveraging these cutting-edge tools enables us to efficiently handle large datasets, ensuring the data used for analysis is highly accurate and reliable. Furthermore, in terms of contributions to data processing and analysis, our approach not only simplifies the data collection and processing phases but also provides a robust framework for the analytical examination of CO_2 emissions, thereby enhancing the efficiency, accuracy, and scalability of emissions research. To visually demonstrate the advantages of our method, we conducted a



奥迪A4L真实油耗怎么样，奥迪A4L油耗多少

shibryan

奥迪A4L工信部油耗

- 奥迪A4L35TFSI两驱车型的工信部综合油耗为6.2L/100km
- 奥迪A4L40TFSI两驱车型的工信部综合油耗为6.1L/100km
- 奥迪A4L40TFSI四驱车型的工信部综合油耗为6.7L/100km
- 奥迪A4L35TFSI两驱车型的工信部综合油耗为7.3L/100km

通常来说工信部油耗都会比实际油耗偏低1.5-2L/100km，这是因为工信部油耗是在实验台架上通过模拟车辆工况测试出来的，与实际情况会有所出入。不过由于所有车的工信部油耗都是在统一标准下进行测试的，所以可作为买车时的油耗对比参考。

奥迪A4L真实油耗怎么样

2022 Corolla Fuel Tank Capacity

The Toyota Corolla has a 13.2-gallon fuel tank. Having the 1.8L 4-cylinder powertrain, drivers will get the estimated 33 miles per gallon city and highway combined. Love the open road in the 2022 Toyota Corolla, whether you're cruising around the town or stuck in some traffic on your daily commute. The Toyota Corolla will have you on the road for plenty of miles before needing to stop at the gas station anytime soon.

"Identify Vehicle Model"

"Comparison between the fuel consumption values provided by the MIIT and the actual fuel consumption"

"Identify the Model Year of vehicle"

"Identify the Engine Displacement of vehicle"

⋮

"Extract average Fuel Consumption"

"根据车主的用车情况，奥迪A4L40TFSI两驱车型的油耗范围是7.31-9.73L/100km，平均油耗为8.51L/100km；40TFSI四驱车型的油耗范围是8.66-11.01L/100km，平均油耗为9.84L/100km；45TFSI四驱车型的油耗范围是8.06-10.45L/100km，平均油耗为9.25L/100km。"

"通常来说工信部油耗都会比实际油耗偏低1.5-2L/100km，这是因为工信部油耗是在实验台架上通过模拟车辆工况测试出来的，与实际情况会有所出入。"

"The 2022 Toyota Corolla is comfortable and quiet on the road, and it is also quite spacious."

"It comes standard with a 1.8 L four-cylinder engine that delivers 139 hp and 126 pound-feet of torque. This engine has an average fuel consumption rating of around 6 L per 100 km on the highway and it also comes with an efficient CVT automatic gearbox that makes the Corolla very fuel-efficient in the city."

"Even if you opt for the 2.0 L four-cylinder engine, you still get a very fuel-efficient vehicle with an average fuel consumption rating of just 7.6 L per 100 km in the city, and 5.9 L per 100 km on the highway. That's right, depending on the version, the Corolla's 2.0-litre four-cylinder is sometimes even more efficient than the entry-level 1.8 L."

Fig. 3 Steps for processing text information using ChatGPT

comparison. We prepared 1000 entries of unprocessed open-source data using web scraping technology and compared this with manual selection; our method completed the processing of 1000 data entries in about 30 min, significantly outperforming the manual selection method (The manual selection method takes about half a minute on average to process each piece of data).

Estimating the fleet average fuel consumption

Due to the limitation of data acquisition, comparing fuel consumption data differences across all models is an almost impossible task that requires a lot of time and effort. Therefore, the fleet average fuel consumption is adopted in this study to assess the difference in the U.S. and Chinese counties. To evaluate the average fuel consumption of LDPVs (including Sedans and SUVs), the weighted average fuel consumption was adopted in this study according to fuel consumption data of each vehicle model and its market share. The fleet average fuel consumption for each vehicle category can be quantified through as shown in Eq. (2).

$$\overline{FC}_{i,y} = \sum_{n=1}^{N} \left(N_{i,n,y} \times \overline{FC}_{i,n,y}\right) / \sum_{n=1}^{N} N_{i,n,y} \tag{2}$$

where $\overline{FC}_{i,y}$ is the average fuel consumption for vehicle type i produced in year y; $N_{i,n,y}$ is the total annual sales of the nth-ranked model of vehicle type i in the U.S. in year y; $\overline{FC}_{i,n,y}$ is the average fuel consumption for the nth-ranked model of vehicle type i produced in year y.

For the four representative counties in China, the vehicle engine displacement distribution can be obtained through the vehicle registration data. The market share data is quantified by engine displacement distribution. Therefore, the weighted average fuel consumption can be calculated as Eq. (3)

$$\overline{FC}_{i,y} = \sum_{k=1}^{K} \left(N_{i,k,y} \times \overline{FC}_{i,k,y}\right) / \sum_{k=1}^{K} N_{i,k,y} \tag{3}$$

where $\overline{FC}_{i,y}$ is the average fuel consumption for vehicle type i produced in year y; $N_{i,k,y}$ is the total number of the kth displacement bin of vehicle type i in Chinese counties in year y; $\overline{FC}_{i,k,y}$ is the average fuel consumption for kth displacement bin of vehicle type i produced in year y.

Deriving the local-specific CO2 emission rates model from fuel consumption

Acquiring the CO2 emission rates in MOVES

In MOVES3, the output from an Emission Rate mode run is a set of emission rates, covering all the emissions processes, e.g., rate per distance, per vehicle, per hour, per start. The emissions rates during running processes are stored in the “rateperdistance” table. The rates generally vary by vehicle type, model year (MY), fuel type, temperature, road type, and speed bin. Among all the influencing factors, the average speed is the most important influencing factor of CO_2 emission rates. A total of 16 average speed bin are defined in MOVES, ranging from 2.5- to 72.5 + mph. The CO_2 emission rates are obtained through Emission rate mode runs for passenger cars and passenger trucks, respectively. Because the focus of this paper is on the emission of LDPVs powered by gasoline on urban roads, the urban unrestricted road is thus selected for road type, and gasoline is selected for fuel type.

Normalizing CO2 emission rates under different speed

As the analysis in “The emission rates under different average speeds”, in different cities, the absolute value of CO_2 emissions rates will be different due to varied vehicle attributes (e.g., displacement, air-compressor technology, gearbox type, gross weight). To get rid of the impact of these properties on CO_2 emission rates and to investigate the relative changing trend of CO_2 emission rates under different average speed, the method of normalizing emission rates under different speed is adopted, all the emission rates under different speed bin are divided by the baseline emission rate.

Combining existing literature [32, 46, 51], we choose the average fuel consumption of LDPVs in the U.S. described in “Estimating the fleet average fuel consumption” as the baseline fuel consumption and its corresponding speed 31.53 km/h (belongs to speed bin 5 in MOVES) as the baseline speed. The speed correction factor can be derived as Eq. (4) following.

$$\overline{EF}_{MOVES,i,g,y}(v) = \overline{EF}_{U.S.,i,g,y} \cdot SCF(v) \tag{4}$$

where $\overline{EF}_{MOVES,i,g,y}v$ is the MOVES average CO_2 emission rates (g/km) under speed v (km/h) for vehicle type i powered by gasoline and produced in year y; $\overline{EF}_{U.S.,i,g,y}$ is the average CO_2 emission rates (g/km) for vehicle type i powered by gasoline and produced in year y in the U.S. (the baseline CO_2 emission rate with an average speed of 31.53 km/h), which can be derived from the average fuel consumption data through the carbon balance method, as shown in Eq. (1). $SCF(v)$ denotes the speed correction factor for an average speed of v (km/h) at which the emissions are to be estimated.

Developing the local-specific CO2 emission rates model

As mentioned in "The emission rates under different average speeds", if the average speed of vehicles is basically consistent, the speed correction function $\alpha(v)$ can be migrated between different cities. Based on the SCF, the localspecific CO_2 emission rates model for LDPVs can be directly developed using the local-specific fuel consumption data, as Eq. (5) following.

$$\overline{EF}_{local,i,g,y}v = \overline{EF}_{local,i,g,y} \cdot SCF(v) \tag{5}$$

where $\overline{EF}_{local,i,g,y}v$ is the CO_2 emission rates (*g/km*) under speed v (*km/h*), for local vehicle type i powered by the fuel gasoline and produced in year y; $\overline{EF}_{local,i,g,y}$ is average CO_2 emission rates (*g/km*) for local vehicle type i powered by gasoline and produced in year y, which can be derived by Eq. (1) based on local real-world average fuel consumption data; $SCF(v)$ denotes the speed correction factor for the vehicle driving at a specific speed v (*km/h*).

Results and discussion

In this section, we address the research question of constructing localized carbon emission factors based on big data on fuel consumption, as posited in this study. We introduce and discuss the analysis of our experimental findings. Initially, an evaluation of MOVES assessment errors was conducted, demonstrating that a direct transplantation of this model would result in significant inaccuracies, thereby underscoring the necessity of our investigation. Building on this premise, we employed the normalization processing method introduced in "Normalizing CO2 emission rates under different speed" to derive speed correction factors. Utilizing actual average fuel consumption data from various counties in China, we established localized carbon dioxide emission factors specific to different speeds and conducted empirical validation, including the verification of our hypotheses and methods through comparative analysis with on-site testing data and local traffic activities. The outcomes indicate that the approach proposed in this manuscript can achieve the research objective of controlling carbon emission assessment errors within a 10% margin while minimizing human and material resources expenditure.

The comparative analysis of average fuel consumption in U.S. and Chinese counties

Passenger cars

Using the single-vehicle average fuel consumption collection method outlined in "Obtaining fuel consumption for each vehicle model", we collected average fuel consumption data for passenger cars and SUVs in China and the United States from MY2009 to MY2021. After manual verification and data cleaning, a total of 6121 fuel consumption data for the US and 5615 for China were obtained.

Passenger cars in MOVES are defined as any coupes, compacts, sedans or station wagons with the primary purpose of carrying passengers. As the weighted method of quantifying the average fuel consumption of one vehicle category introduced in "Estimating the fleet average fuel consumption", the gap of average fuel consumption of passenger cars between Chinese counties and the U.S. can be obtained, as shown in Fig. 4 below.

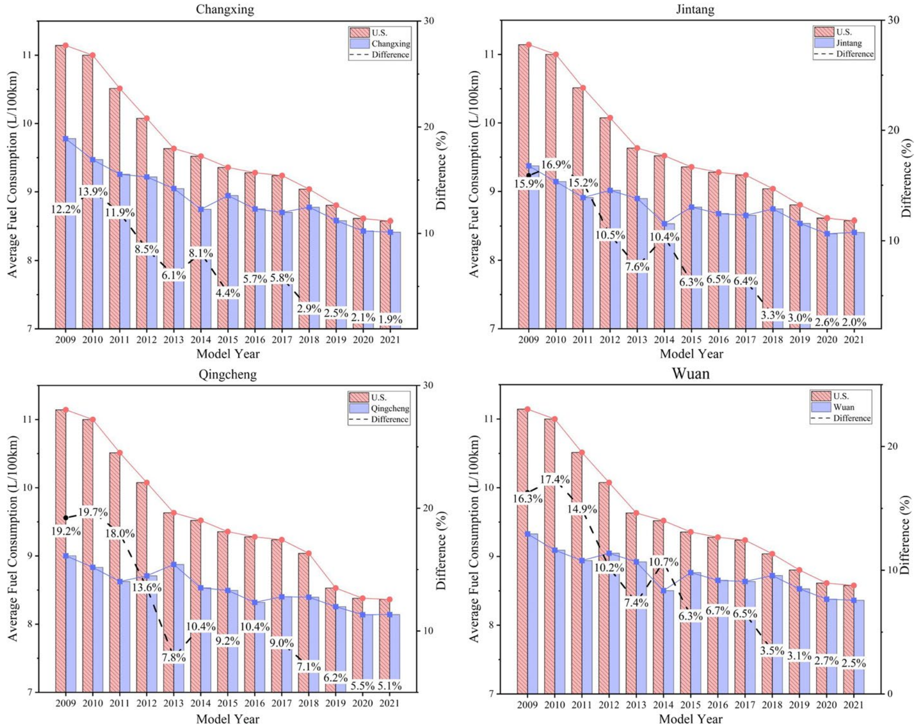


Fig. 4 The difference of passenger car fuel consumption between the U.S. and Chinese counties

It can be found that the average fuel consumption of sedans in the U.S. in the past decade has been declining year by year, mainly due to the improvement of single-vehicle fuel economy technology. However, from the point of view of actual fuel consumption levels of the four representative Chinese counties, the overall trend has been decreasing in the past decade, but there has been little fluctuation. Because fuel consumption indicators have always been a critical consideration for consumers when buying sedans, and small sedans with small-displacement remain the first choice for many families. Therefore, the fuel consumption gap between the U.S. and Chinese counties basically shows a downward trend year by year, ranging from 2 to 20%. This is why MOVES cannot be directly applied to estimate the CO_2 emissions of sedans in Chinese counties.

Passenger trucks

In MOVES, sport utility vehicles (SUVs) belongs to the classification of passenger trucks with the primary purpose of carrying passengers that includes pickups, SUVs, and minivans. Therefore, we choose the average fuel consumption of the passenger truck in the U.S. to compare with that of SUVs in Chinese counties. The weighted method in "Estimating the fleet average fuel consumption" is still adopted for passenger trucks. The gap of the average fuel consumption of passenger trucks between Chinese counties and the U.S. is shown in Fig. 5 below. It can be found that the changing trend of average fuel consumption level of SUVs in Chinese counties and passenger trucks in the U.S. in the past ten years is consistent with that of passenger cars, all benefit from government regulation and the improvement of vehicle fuel economy technology.

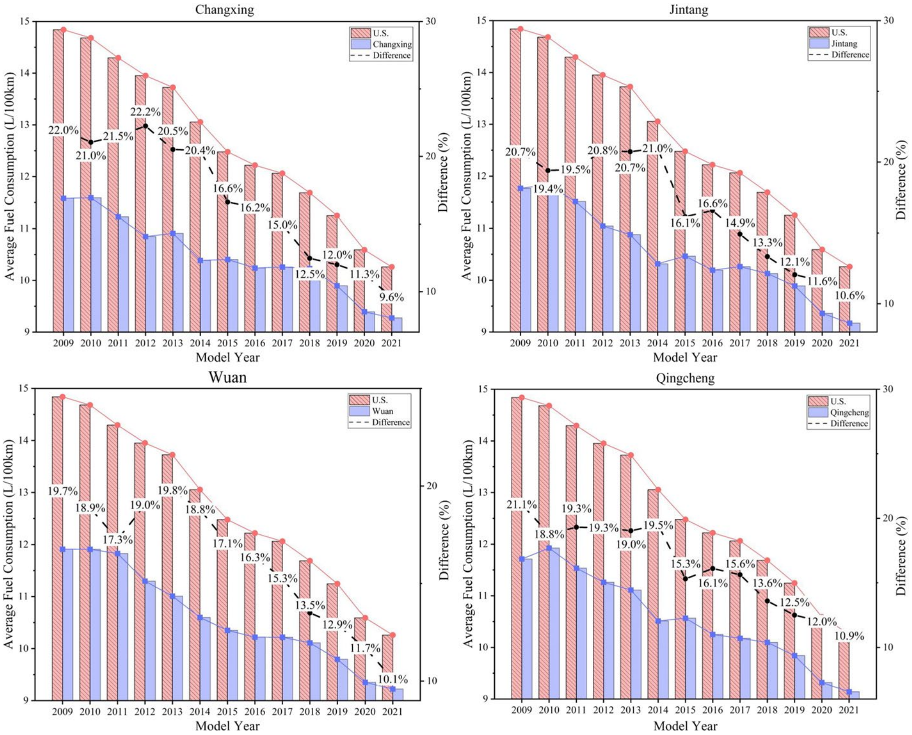


Fig. 5 The difference of passenger truck fuel consumption between the U.S. and Chinese counties

However, the fuel consumption gap is more significant than that of passenger cars between the U.S. and Chinese counties. This is because the passenger truck in MOVES contains pickups that are fuel hogs. However, in China, pickup trucks are classified as light trucks, which means that there will be all kinds of troubles in taxes, driver's licenses, and vehicle inspections. And pickup trucks are not allowed to be driven into urban areas in many cities. Secondly, many families will not consider buying SUVs with high displacement due to their high fuel consumption level, so the SUVs with a displacement of over 3.0L have a deficient proportion in the passenger truck market in China. The big difference between passenger trucks attributes in the U.S. and Chinese counties are the reason why the emission rates in MOVES must be localized for international applications.

The relative changing trend of CO2 emission rates in MOVES

As the method acquiring CO_2 emission rates of MOVES mentioned in "Acquiring the CO2 emission rates in MOVES", the CO_2 emission rates for each vehicle type are obtained through several emission rate mode runs for passenger car and passenger truck, respectively. Figure 6 depicts the changing trend of CO_2 emission rates concerning average speed for MY2009-MY2021.

As for the average speed, for different vehicle types and models, it can be easily found that the changing trend of CO_2 emission rates with respect to average speed is basically the same, even if their absolute values are different. CO_2 emission rates decrease with the increase of average speed before speed bin 13. After that, rates will instead increase as the speed increases, which is called "sweet spot". Also, before the speed bin 3, the

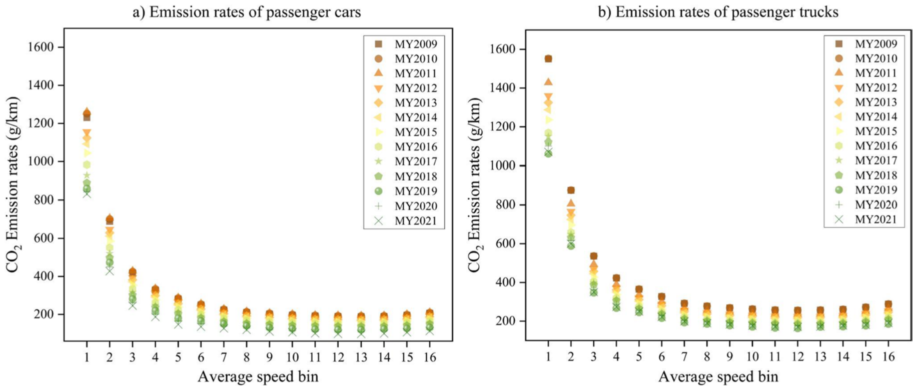


Fig. 6 Running CO2 emission rates by average speed bin in MOVES for passenger cars and passenger trucks for the model year 2009–2021

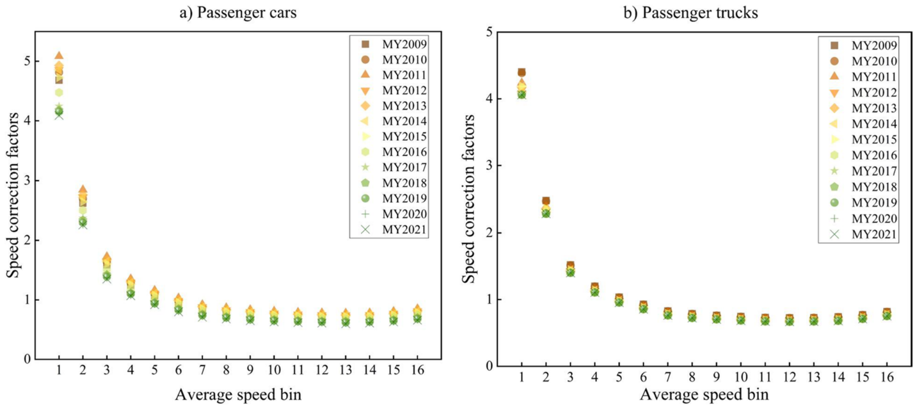


Fig. 7 Average SCFs by average speed bin in MOVES for passenger cars and passenger trucks for MY2009-2021

emission factors show a sharp decline trend as the speed increases, which indicate that the idling state should be avoided as much as possible in the driving process.

The local-specific CO_2 emission rates model for Chinese counties

From the analysis of average fuel consumption in "The comparative analysis of average fuel consumption in U.S. and Chinese counties", the absolute values of average fuel consumption in the U.S. and Chinese counties are different because of the difference in vehicle attributes. This is the most crucial reason why MOVES cannot be directly applied to Chinese counties to estimate the CO_2 emissions, even if we input localized data for activity parameters. This is also the reason why a local-specific emission rates model should be established.

As the normalization method mentioned in "Normalizing CO2 emission rates under different speed", the average fuel consumption results of LDPVs in the U.S. in Sect. 5.1 was applied as the baseline fuel consumption, the baseline emission rate can be derived by Eq. (4), and then the relative emission rate (SCF) can thus be derived by Eq. (5).

Figure 7 depicts the average speed correction factors of passenger cars and passenger trucks after the normalization process. Comparing with Fig. 6, it can be found that although their absolute value of CO_2 emission rates varies greatly, the SCFs are basically the same for each vehicle type and model after normalization, except for a few model years of passenger cars. Therefore, an exciting conclusion could be draw that the speed correction factors may be transplanted between different cities, even with different vehicle model distributions. This is the principle of using the CO_2 emissions rates embedded in MOVES to establish CO_2 emission rates inventory in Chinese counties.

It is worth noting that the emission factor in MOVES3 will decrease monotonically with the model year, and the emission rates after MY2017 are predicted according to EPA's Light Duty GHG MY 2017+rules (U.S. Environmental Protection Agency, 2012). However, the actual vehicle emissions may not meet the regulations, and fuel consumption will not be significantly reduced sharply every year due to technological advancement as expected. Therefore, each model year's SCF obtained according to the above method mentioned in "Normalizing CO2 emission rates under different speed" will be different if there are outliers in MOVES.

In Fig. 7, we can observe that the variance of the SCFs for different MY after normalization is significantly higher for passenger cars than for passenger trucks. Particularly, there are some outliers in the model year before 2012. This can be explained by the data source of the pre 2012 vehicle emission factors in MOVES. The CO_2 emission rates of passenger cars increased from MY2009 to MY2011 and reached its peak in MY2011. CO_2 emission rates for MY2009 were based on EPA's "Light-Duty Automotive Technology, Carbon Dioxide Emissions, and Fuel Economy Trends: 1975 through 2009", while the rates from MY 2011 to MY 2016 were derived from the Light-Duty Greenhouse Gas (LD GHG) rulemaking analysis. Since they are from different data sources, it's not surprising that they are slightly inconsistent. And the MY2010 rate was interpolated as a midpoint between the model years 2009 and 2011. That's why there is a peak at MY2011 for passenger cars when looking at CO_2 rate (g/km) from MOVES output. Therefore, to avoid the influence of outliers, we choose the average of SCFs of passenger trucks at each average speed bin as the SCF of LDPVs in Chinese counties. The specific values are shown in Table 1.

Based on the average speed correction factor in Table 1, the local-specific running CO_2 emission rates of LDPVs can thus be derived by Eq. (5) in "Developing the local-specific CO2 emission rates model", using the estimated real-world fuel consumption data in Chinese counties obtained in "Object of study"1.

Verification of local-specific emission rates model

CO_2 emission rates are the core basis for accurately estimating the CO_2 emission level of a region, but the straightforward verification of CO_2 emission rates is usually challenging. The assumption mentioned in "The emission rates under different average

Table 1 Average speed correction factors of CO_2 emission rates under different speed bin

Average speed bin	1	2	3	4	5	6	7	8
Speed correction factor	4.1129	2.3184	1.4212	1.1221	0.9685	0.8682	0.7746	0.7358
Average speed bin	9	10	11	12	13	14	15	16
Speed correction factor	0.7134	0.6964	0.6833	0.67487	0.6818	0.6923	0.7212	0.7649

speeds" that the SCF could be transplanted to different cities if the average speed of the urban road network is consistent will inevitably introduce uncertainties in the emission rates model establishment. To verify the rationality of our assumption and validate the local-specific CO_2 emission rates model, we compare our results with the field test data findings from the literature and apply the local traffic activity data to the assessment of CO_2 emissions.

Verification by comparative results with local field test data

In China, the sedans and SUVs belong to the category of light-duty passenger vehicles (LDPVs). Zhang et al. [46] carried out on-road emissions tests for 60 gasoline LDPVs in three Chinese cities. They evaluated the impacts of changes in average speed on the relative fuel consumption of gasoline LDPVs in Beijing and obtained ideal results, which have been applied in many other Chinese cities (e.g., Nanjing, Shenzhen). In this paper, the comparison of our results with their results was carried out for method validation.

In MOVES, the speed was divided by bins (i.e., interval), and each speed bin corresponds to a specific emission rate. To ensure the validity of the comparison, we adopted the same binning method and averaged the SCF obtained by Zhang et al. [46] according to the speed range defined by the speed bin in MOVES. And because the field test by Zhang et al. is conducted in the urban area, the comparative work is thus carried out only in 1–8 speed bins (< 60 km/h).

If the SCF we derived from MOVES and that derived from field tests are in good agreement, the method proposed in this study can be regarded as an alternative and effective way to establish the local-specific CO_2 emission rate database.

Figure 8a) illustrates the SCF under different speed bins of MOVES and the local field test findings by Zhang et al. The results show that they have a good consistency except for the speed bin 1 and speed bin 2. That might be caused by the few measurement data obtained by Zhang et al. in low-speed zones (< 10 km/h).

The Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) are the two most commonly used indicators to measure the accuracy of variables, and they are also two crucial scales for evaluating model performance. In addition to these two popular indicators, Normalized Mean Square Error (NMSE) is also adopted to assess the

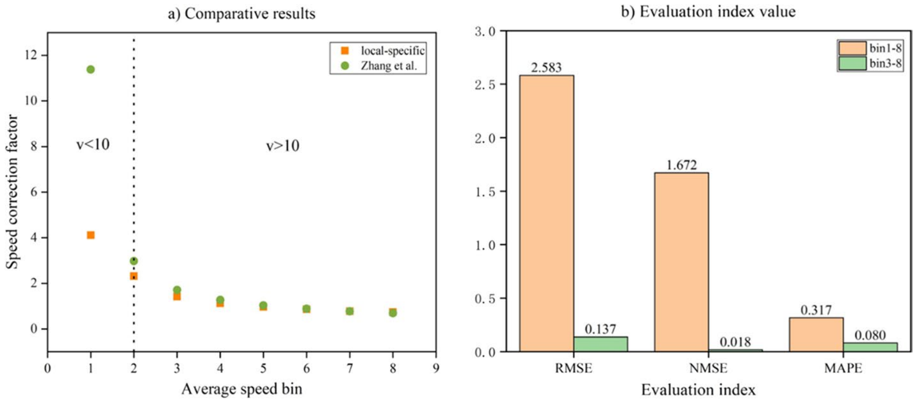


Fig. 8 The comparative results of the SCF of local-specific and Zhang et al.

difference between two emission rates database for a more comprehensive comparison. The evaluation index values are plotted in Fig. 8b).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(FC_i - \widehat{FC}_i\right)^2}{n}} \tag{6}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{FC_i - \widehat{FC}_i}{n}\right| \tag{7}$$

$$NMSE = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(FC_i - \widehat{FC}_i\right)^2}{FC_i \cdot \overline{\widehat{FC}_i}} \tag{8}$$

where $\widehat{FC}_i$ and FC_i are the normalized fuel consumption (relative to the benchmark fuel consumption) in MOVES and the findings by Zhang et al., respectively. i denotes the speed bin.

The RMSE, NMSE, and MAPE for average speed bin 1–8 are 2.583, 1.672, and 0.317, respectively. In general, it should be expected that the MAPE value is much smaller than the RMSE value. Because for the root mean square error RMSE, each error is square. This means that the single error increases quadratically and has different effects on the final RMSE value. Hence the RMSE is more affected by outliers, and it can be seen that larger outliers have a more significant impact on RMSE. The large errors mainly come from average speed bin 1 and 2, where few data were collected by Zhang.

If the average speed bin 1 and 2 where few data was collected by Zhang were excluded, the RMSE, NMSE, MAPE would decrease to 0.137, 0.018, and 0.080, respectively. For NMSE, the value lower than 0.5 is generally regarded as an allowable upper limit [40]. Therefore, such small errors indicate that it is effective to apply the speed correction factors we derived in Table 1 to establish a local CO_2 emission rate database in Chinese counties.

Verification by local traffic activity and fuel consumption

As mentioned in 5.2, traffic activity surveys were conducted in four representative counties, and link-specific traffic data in different time periods were obtained by the video data and bayonet system data from the Traffic Police Department.

Average speed distribution is the main activity parameter in MOVES, and the average fuel consumption can be derived by the product of the vehicle average speed distribution and the emission rates at that that speed. Since the traffic data we obtained from the Traffic Police Department does not include speed information, the method of simulation by VISSIM was adopted to get the average speed distribution on urban roads. Since the fuel consumption data we obtained is the real-world average fuel consumption under various driving scenarios, the corresponding activity parameter should also be an average one. To get the average speed distribution of urban roads (including truck roads, secondary truck roads, and branch roads), we simulated the average speed under different road type with various traffic conditions (different traffic volume at different

Table 2 The estimation performance for sedans MY2021 in four counties

	Method	MOVES	Zhang et al.	Local-specific	Real-world
Changxing	CO_2 emissions (g/km)	195.52	239.09	215.92	203.28
	Deviation (%)	− 3.97%	+ 14.98%	+ 5.86%	–
Jintang	CO_2 emissions (g/km)	198.99	244.20	219.29	203.07
	Deviation (%)	− 2.05%	+ 16.84	+ 7.39%	–
Qingcheng	CO_2 emissions (g/km)	203.24	244.77	216.60	196.65
	Deviation (%)	+ 3.25%	+ 19.66%	+ 9.21%	–
Wuan	CO_2 emissions (g/km)	195.49	237.32	214.58	202.03
	Deviation (%)	− 3.34%	+ 14.86%	+ 5.84	–

Table 3 The estimation performance for SUVs MY2021 in four counties

	Method	MOVES	Zhang et al.	Local-specific	Real-world
Changxing	CO_2 emissions (g/km)	259.17	257.53	232.56	218.95
	Deviation (%)	+ 15.52%	+ 14.98%	+ 5.85%	–
Jintang	CO_2 emissions (g/km)	263.38	260.38	233.82	216.52
	Deviation (%)	+ 17.79%	+ 16.84%	+ 7.39%	–
Qingcheng	CO_2 emissions (g/km)	268.65	268.61	237.70	215.80
	Deviation (%)	+ 19.67%	+ 19.66%	+ 9.21%	–
Wuan	CO_2 emissions (g/km)	259.16	255.76	231.25	217.74
	Deviation (%)	+ 15.98%	+ 14.87%	+ 5.85%	–

times) and finally took the average as the county's activity parameter input. The CO_2 emissions (g/km) thus can be calculated, as shown in Eq. (9).

$$CO_2 emissions\ g/km = \sum_{j=1}^{16} EF\left(v_j\right) \cdot \overline{ASD}\left(j\right) \tag{9}$$

where $EF(v_j)$ is the emission rates at speed j, $\overline{ASD}(j)$ is the average speed distribution at speed j.

Table 2 and 3 presents the comparison results between different emission rates database for sedans and SUVs of MY2021 in four representative counties (i.e., Changxing, Jintang, Qingcheng, and Wuan), respectively. For passenger cars, MOVES performs well for model year 2021 but may overestimate or underestimate emission levels. However, for pickup trucks, MOVES still significantly overestimates actual emission levels. The deviations for SUVs are more significant than that for sedans because of the larger difference in fuel consumption between Chinese counties and the U.S., as discussed in "Passenger trucks". This is why we should establish a localized emission rates database. If using our proposed local-specific emission rates model, there is a significant improvement by approximately 10% of the estimation performance compared to MOVES for SUVs. The reason why there is no substantial improvement for sedans lies in the fact that the sedans difference between Chinese counties and the U.S. has been shrinking in the past decade, and the MY2021 emission rates in MOVES are close to the situation in Chinese counties, as discussed in "Passenger cars". Overall, the deviation with the real-world does not exceed 10% both for sedans and SUVs in four representative counties, using our proposed method. The results prove the rationality of our assumption in "The emission rates under different average speeds" that speed correction factor could be transplanted between different counties if their average speed is almost consistent and demonstrate the effectiveness of our established CO_2 emission rate database.

To better show the estimation performance comparison between MOVES and our proposed method, the estimation deviation using different emission rate database for MY2009-2021 was plotted in Fig. 9. It can be found that MOVES has a more significant estimation deviation for the older vehicles, even more than 30% for SUVs of MY 2009 and sedans of MY2011. The estimation deviation for the latest MY is smallest, but it still remains at + 20% for SUVs in Jintang and Qingcheng. However, the local-specific CO_2 emission rates established in this paper can well estimate the CO_2 emissions for LDPVs of the different MY, with a deviation within 10% for all representative counties. The maximum deviation (9.2%) for Qingcheng is associated with the highest fraction of time spent at lower speed in average speed bin 4 (speed ranges from 20 km/h to 28 km/h). Furthermore, compared with MOVES, the estimation improvement ranges from 2 to 29%, with an average of 17.9% for all MY. And the estimation performance is better than the emission rates model proposed by Zhang et al. Furthermore, in terms of the time cost of the method, the MOVES-based approach requires approximately 1–2 h to process data for a single county, owing to the necessity for the input of

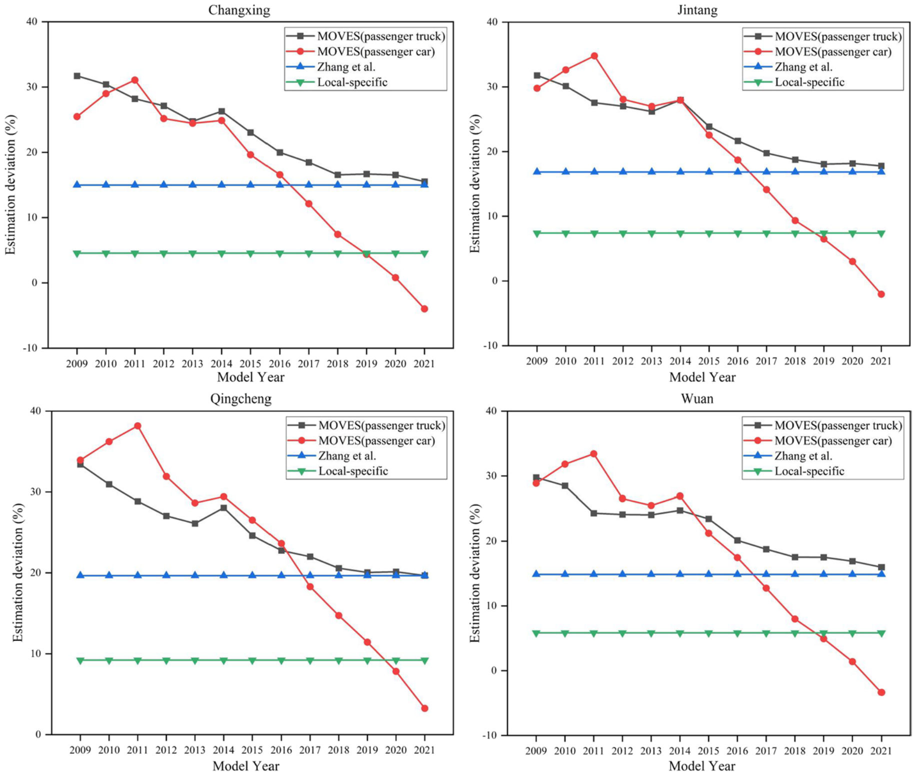


Fig. 9 The estimation deviation with real-world data of the different database

extensive ancillary environmental configuration information such as weather, humidity, distribution of speed, proportion of vehicle age, etc. In contrast, the method proposed in this paper, along with the approach by Zhang et al., can be completed in a matter of seconds. Therefore, it can be concluded that the method introduced herein holds certain advantages in efficiency, accuracy, and cost.

Applications and future scopes

This research illustrates how to develop a local-specific CO_2 emission rate database in Chinese counties using open-access data sources, especially for those counties with scarce field test data. Actually, the method proposed in this study is suitable for various Chinese cities (e.g., metropolis, medium, and small cities), not only limited to the counties. This is because the market share data and fuel consumption data of each vehicle category in the U.S. could be accessed publicly through multiple data statistic services. The usage of open-access market share data and fuel consumption data can provide fine-scale average fuel consumption data of each vehicle category without great efforts in data collection or strict requirements for the portable emissions measurement system. This fuel-consumption based method eliminates the most significant barrier for counties that are subjected to the lack of enough field test data for establishing a high-resolution CO_2 emission rate database.

Moreover, this method can provide dynamically updated fuel consumption data with the change of fuel-efficiency technology every year instead of using the outdated test data a few years ago. If we want to update the emission rate database through the field test method, it is necessary to choose a representative car model with the latest fuel-efficiency technology for testing every year, which will be a considerable work in no doubt. Therefore, the method proposed in this paper can fill the defect that field test data cannot be easily updated in real-time.

Most importantly, for the field test data method, it was necessary to impute rates for cells for which no data was available, i.e., "holes". Regarding the vehicle speed bins, empty cells usually occur for low-speed modes or high-speed modes not covered by available data, such as meaning operating modes with speed lower than about 10 km/h or higher than 60 km/h. Therefore, the speed correction factors can be applied to fix the "holes" of the CO_2 emission rates data obtained by field tests. For instance, if the field-test data ranging from speed bin 3–5 can only be easily obtained, we can use the speed correction factors in Table 1 as the "hole-filling" coefficient to estimate the CO_2 emission rates in other speed bin. It should be noted that if a user wants to estimate emissions at a specific intermediate speed, linear interpolation can provide an approximation. For example, an interpolation between the rates for the 25mph and 30mph average speed bins can be used to represent emissions at a 28.3mph average speed.

However, how to control the deviation of the average fuel consumption data estimation for each vehicle category is a significant challenge for our method, as there are too many factors affecting the fuel consumption level of a vehicle. Therefore, the future study may focus on the more accurate estimation method of the average fuel consumption of each vehicle category. Nevertheless, the general applicability of our approach in various cities is a significant advantage for high-resolution CO_2 emission rate database establishment without great efforts of data acquisition through field tests in each city.

To further underscore the significance of our research, we have expanded our discussion to highlight the potential impact on local environmental policies and carbon management strategies. The development of a localized CO_2 emission rate database offers critical insights for policymakers and urban planners to design targeted interventions for reducing emissions. By providing detailed, local emission data, our approach facilitates the formulation of precise carbon reduction targets, enhances the efficiency of resource allocation in carbon management efforts, and supports sustainable urban development initiatives. This enhanced capacity for detailed environmental analysis and policy formulation represents a substantial contribution to efforts aimed at mitigating the environmental impacts of transportation emissions at a local level.

Conclusions

Thanks to the unique advantages of big data and artificial intelligence algorithms in web data extraction, this paper employs natural language processing and large language model algorithms to automatically acquire large-scale vehicle fuel consumption data. Building on this, we propose a fuel consumption-based methodology for the customization analysis of emission rates in MOVES for developing a county-level CO_2 emission database in Chinese counties. The speed correction factors derived from normalization processing of CO_2 emission rates under different average speeds show a good consistency with field test data from the literature. Besides, the CO_2 emission estimation results using different emission rates database with the local-specific traffic activity data indicate that our proposed method can well estimate the CO_2 emissions.

There are other possibilities to construct a high-resolution CO_2 emission rate database. Comparing with the conventional field test method for emission rate database establishment, the method proposed in this paper can derive a more comprehensive database and is much more convenient and cost-saving. It is expected that the findings from this study would be helpful for better application of MOVES in Chinese counties.

This paper illustrates how to develop a local-specific CO_2 emission rate database, and we believe it may open a door toward an alternative to the field test method for many tasks. It is possible to formulate a local-specific CO_2 emission database in various Chinese counties using only open-access big data. With the refined emission rate database, refined management of low-carbon road transportation and the traffic environment can be achieved in the future. Furthermore, the potential applications of our methodology extend far beyond the scope of this study. The methodology developed herein holds substantial promise for broader application across a variety of urban settings, not limited to the Chinese context. The adaptability of our approach to diverse data landscapes and regulatory environments suggests its utility in both developed and other developing countries. By tailoring the methodology to account for local vehicular patterns, emission regulations, and available datasets, it could significantly contribute to the creation of customized CO_2 emission databases worldwide. This would not only facilitate more accurate and localized environmental policy making but also enhance the global effort towards sustainable urban development and carbon management. Future research could explore these applications in detail, assessing the method's effectiveness in different geographical and socio-economic contexts to provide a comprehensive perspective on its global utility.

Acknowledgements

This study was carried out with the support of the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY222030), the National Natural Science Foundation of China (Grant No. 52202408) and the China Postdoctoral Science Foundation (No. 2022M720719). In addition, the authors would like to express their appreciation to the MOVES Team for providing the MOVES3 documentation and for their technical assistance regarding emission regulations and measurements.

Author contributions

Linheng Li: Investigation, Writing—review& editing. Can Wang: Data curation, Writing—original draft, Jing Gan: Data curation, Writing—original draft. Dapeng Zhang: Data curation, Writing—original draft, Visualization.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The data analysis code related to this article can be accessed through the following link: https://github.com/linhenglee/ChatGPT-based-method.

Declarations

Competing interests

This paper has not been submitted elsewhere in identical or similar form, nor will it be during the first three months after its submission to the Publisher.

Received: 16 November 2023 Accepted: 7 May 2024

Published online: 17 May 2024

References

1. Boddi Reddy SA, Ahmed S, Arocho I. Simulation of real-time operational level emissions from nonroad equipment: case study of a construction site. Pract Period Struct Des Constr. 2023;28:05022008.
2. Boroujeni BY, Frey HC. Road grade quantification based on global positioning system data obtained from real-world vehicle fuel use and emissions measurements. Atmos Environ. 2014;85:179–86.
3. Cao X, Li X, Zhou Y. Greenhouse gas inventory evaluation of low-carbon cities in the context of spatial planning: a case study of Shenzhen, China. Environ Eng Sci. 2022. https://doi.org/10.1089/ees.2022.0268.
4. Carniel T, Cazenille L, Dalle JM, Halloy J. Using natural language processing to find research topics in living machines conferences and their intersections with bioinspiration & biomimetics publications. Bioinspir Biomim. 2022;17:065008.
5. Chandrashekar C, Chatterjee P, Pawar DS. Estimation of CO_2 and co emissions from auto-rickshaws in indian heterogeneous traffic. Transp Res Part D Transp Environ. 2022;104:103202.
6. EMFAC. 2023. https://arb.ca.gov/emfac/.
7. Emisia. 2023. https://www.emisia.com/utilities/copert/.
8. EPA. 2023. https://www.epa.gov/moves.
9. Erdengasileng A, Han Q, Zhao T, Tian S, Sui X, Li K, Wang W, Wang J, Hu T, Pan F, et al. Pre-trained models, data augmentation, and ensemble learning for biomedical information extraction and document classification. Database. 2022. https://doi.org/10.1093/database/baac066.
10. Fedorov A, Nikolskaia K, Ivanov S, Shepelev V, Minbaleev A. Traffic flow estimation with data from a video surveillance camera. J Big Data. 2019;6:1–15.
11. Frey HC. Development and evaluation of a simplified version of moves for coupling with a traffic simulation model. Annu Meet Transp Res Board. 2013;31:32.
12. Gao C, Gao C, Song K, Xing Y, Chen W. Vehicle emissions inventory in high spatial– temporal resolution and emission reduction strategy in harbin-changchun megalopolis. Process Saf Environ Prot. 2020;138:236–45.
13. Gao C, You H, Gao C, Na HM, Xu QJ, Li XJ, Liu HT. Analysis of passenger vehicle pollutant emission factor based on on-board measurement. Atmos Pollut Res. 2022;13:101421.
14. GREET. 2023. https://greet.es.anl.gov/.
15. Herndon SC, Nelson DD Jr, Wood EC, Knighton WB, Kolb CE, Kodesh Z, Allen DT. Application of the carbon balance method to flare emissions characteristics. Ind Eng Chem Res. 2012;51(39):12577–85.
16. Hussein H, Abbas E, Keshavarzi S, Fazelzad R, Bukhanov K, Kulkarni S, Au F, Ghai S, Alabousi A, Freitas V. Supplemental breast cancer screening in women with dense breasts and negative mammography: a systematic review and meta-analysis. Radiology. 2023;306:e221785.
17. IEA. 2022. Global CO_2 emissions from transport by subsector. 2000–2030. https://www.iea.org/data-and-statistics/charts/global-CO2-emissions-from-transport-by-subsector-2000-2030.
18. IVE. 2023. http://www.issrc.org/ive/.
19. Kumar S, Tiwari P, Zymbler M. Internet of Things is a revolutionary approach for future technology enhancement: a review. J Big Data. 2019;6(1):1–21.
20. Khazini L, Kalajahi MJ, Blond N. An analysis of emission reduction strategy for light and heavy-duty vehicles pollutions in high spatial–temporal resolution and emission. Environ Sci Pollut Res. 2022;29(16):23419–35.
21. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl. 2023;82:3713–44.

22. Lei J, Yang C, Fu Q, Chao Y, Dai J, Yuan Q. An approach of localizing moves to estimate emission factors of trucks. Int J Transp Sci Technol. 2023. https://doi.org/10.1016/j.ijtst.2023.02.002.
23. Li X, Xie Y, Li C, Wang Z, Hopke PK, Xue C. Using the carbon balance method based on fuel-weighted average concentrations to estimate emissions from household coal-fired heating stoves. Chemosphere. 2022;307:135639.
24. Li Y, Wang G, Murphy C, Kleeman MJ. Modeling expected air quality impacts of oregon's proposed expanded clean fuels program. Atmos Environ. 2023. https://doi.org/10.1016/j.atmosenv.2023.119582.
25. Lin S, Liu Y, Chen H, Wu S, Michalaki V, Proctor P, Rowley G. Impact of change in traffic flow on vehicle non-exhaust pm2. 5 and pm10 emissions: a case study of the m25 motorway, UK. Chemosphere. 2022;303:135069.
26. Pechout M, Jindra P, Hart J, Vojtisek-Lom M. Regulated and unregulated emissions and exhaust flow measurement of four in-use high performance motorcycles. Atmos Environ X. 2022;14:100170.
27. Perugu H. Emission modelling of light-duty vehicles in India using the revamped VSP-based moves model: the case study of hyderabad. Transp Res Part D Transp Environ. 2019;68:150–63.
28. Ramezani H, Lu XY, Shladover SE. 2019. Calibration of motor vehicle emission simulator (MOVES) using real heavy-duty truck data. Technical Report.
29. Romero Y, Chicchon N, Duarte F, Noel J, Ratti C, Nyhan M. Quantifying and spatial disaggregation of air pollution emissions from ground transportation in a developing country context: case study for the lima metropolitan area in peru. Sci Total Environ. 2020;698:134313.
30. Rosero F, Fonseca N, Mera Z, L'opez JM. Assessing on-road emissions from urban buses in different traffic congestion scenarios by integrating real-world driving, traffic, and emissions data. Sci Total Environ. 2023;863:161002.
31. Samaras A, Bekiaridou A, Papazoglou AS, Moysidis DV, Tsoumakas G, Bamidis P, Tsigkas G, Lazaros G, Kassimis G, Fragakis N, et al. Artificial intelligence-based mining of electronic health record data to accelerate the digital transformation of the national cardiovascular ecosystem: design protocol of the cardiomining study. BMJ Open. 2023;13:e068698.
32. Song G, Yu L, Wu Y. Development of speed correction factors based on speed-specific distributions of vehicle specific power for urban restricted-access roadways. J Transp Eng. 2016;142:04016001.
33. Sun S, Sun L, Liu G, Zou C, Wang Y, Wu L, Mao H. Developing a vehicle emission inventory with high temporal-spatial resolution in Tianjin, China. Sci Total Environ. 2021;776:145873.
34. Shepelev V, Aliukov S, Glushkov A, Shabiev S. Identification of distinguishing characteristics of intersections based on statistical analysis and data from video cameras. J Big Data. 2020;7:1–23.
35. Wang Q, Fang Y, Ravula A, et al. Webformer: the web-page transformer for structure information extraction. 2022. https://doi.org/10.48550/arXiv.2202.00217.
36. Wang X, Song G, Wu Y, Zhai Z. Road grade estimation based on power demand difference of heavy-duty diesel trucks for emission estimation. Transp Res Rec. 2023;2677:129–50.
37. Wang Z, Guo Y, Xu Y, Xue Y, Liu Y, Shen H, Cheng X. Scient: a semanticfeature- based framework for core information extraction from web pages. In: Tanveer M, Agarwal S, Ozawa S, Ekbal A, Jatowt A, editors. Neural information processing: 29th international conference, ICONIP 2022, virtual event, November 22–26, 2022, proceedings, Part III. New York: Springer; 2023. p. 311–23.
38. Wang Z, Wu G, Scora G. MOVESTAR: an open-source vehicle fuel and emission model based on USEPA moves. arXiv. 2020. https://doi.org/10.48550/arXiv.2008.04986.
39. Wei T, Frey HC. Intermodal comparison of tailpipe emission rates between transit buses and private vehicles for on-road passenger transport. Atmos Environ. 2022;281:119141.
40. Wu Y, Song G, Yu L. Sensitive analysis of emission rates in moves for developing sitespecific emission database. Transp Res Part D Transp Environ. 2014;32:193–206.
41. Xu Y, Liu Z, Xue W, Yan G, Shi X, Zhao D, Zhang Y, Lei Y, Wang J. Identification of on-road vehicle CO_2 emission pattern in China: a study based on a high-resolution emission inventory. Resour Conserv Recycl. 2021;175:105891.
42. Xu Z, Jiang T, Zheng N. Developing and analyzing eco-driving strategies for on-road emission reduction in urban transport systems-a vr-enabled digital-twin approach. Chemosphere. 2022;305:135372.
43. Yao R, Wang X, Xu H, Lian L. Emission factor calibration and signal timing optimisation for isolated intersections. IET Intel Transport Syst. 2018;12:158–67.
44. Zhang B, Yin S, Lu X, Wang S, Xu Y. Development of city-scale air pollutants and greenhouse gases emission inventory and mitigation strategies assessment: a case in Zhengzhou, central china. Urban Climate. 2023;48:101419.
45. Zhang L, Song G, Zhang Z. Heterogeneity analysis of operating mode distribution for modeling energy consumption of light-duty vehicles. Transp Res Rec. 2023;2677:93–109.
46. Zhang S, Wu Y, Liu H, Huang R, Un P, Zhou Y, Fu L, Hao J. Real-world fuel consumption and CO_2 (carbon dioxide) emissions by driving conditions for light-duty passenger vehicles in China. Energy. 2014;69:247–57.
47. Zhang S, Wu Y, Liu H, Huang R, Yang L, Li Z, Fu L, Hao J. Real-world fuel consumption and CO_2 emissions of urban public buses in Beijing. Appl Energy. 2014;113:1645–55.
48. Zhang T, Jin T, Qi J, Liu S, Hu J, Wang Z, Li Z, Mao H, Xu X. Influence of test cycle and fuel property on fuel consumption and exhaust emissions of a heavy-duty diesel engine. Energy. 2022;244:122705.
49. Zheng O, Abdel-Aty M, Wang Z, Ding S, Wang D, Huang Y. AVOID: autonomous vehicle operation incident dataset across the globe. 2023. arXiv preprint arXiv:2303.12889.
50. Zheng X, He L, He X, Zhang S, Cao Y, Hao J, Wu Y. Real-time black carbon emissions from light-duty passenger vehicles using a portable emissions measurement system. Engineering. 2022;16:73–81.
51. Zhang Z, Song G, Zhang L, Zhai Z, He W, Yu L. How do errors occur when developing speed correction factors for emission modeling? Transp Res Part D Transp Environ. 2021;101:103094.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.