

Temporal Aggregated Analysis of GPS Trajectory Data Using Two-Fluid Model

Yunfei Zhang¹ , Allister Loder¹ , Felix Rempe² ,
and Klaus Bogenberger¹ 

Transportation Research Record
2023, Vol. 2677(5) 103–116
© National Academy of Sciences:
Transportation Research Board 2022



Article reuse guidelines:

sagepub.com/journals-permissions
DOI: 10.1177/03611981221128279
journals.sagepub.com/home/trr



Abstract

The market for on-demand mobility services is growing worldwide. These services include, for example, ride-hailing, ride-sharing, and car-sharing. Large-scale fleets of such services collect GPS trajectory (probe vehicle) data constantly everywhere in the network. At a certain penetration rate, this data becomes representative of the entire road network. It can give valuable insights into traffic dynamics and the evolution of congestion. In this paper, we use such GPS trajectory data from Chengdu, China, to investigate the stability and recurrence of macroscopic traffic patterns. Using the two-fluid theory, we find that the two-fluid coefficients are robust on between-day variation, not only supporting the theory itself but also emphasizing that the general evolution of traffic is a robust pattern. We investigate the deviations from the model using time series analysis of the residuals of the two-fluid model. Here, we find evidence for daily and weekly seasonality in the residuals, indicating that congestion patterns are convincingly recurring. These patterns can be used for network-wide traffic state prediction. We conclude that GPS trajectory data from large on-demand mobility fleets is a promising data source for observing traffic patterns in urban road networks once the data becomes representative.

Keywords

on-demand mobility, GPS trajectory, two-fluid model, macroscopic analysis

Measuring traffic in a metropolis can be costly for the traffic management center as it needs to install many stationary sensors and accompanying communication infrastructure. Since the advent of GPS probe vehicle data, traffic state information is reported by the moving vehicles, aggregated, and returned to all drivers (1). While stationary sensors commonly measure traffic flows well, the GPS probe vehicle data performs better on recording speeds (2). Thus, a fusion from both sources can improve the network-wide traffic state estimation (3). Currently, many probe vehicle data providers do not offer the original vehicle trajectories for privacy reasons, but rather aggregate the data to trip, origin–destination, traffic volumes, and speed data on road segments with a typical length of around 100 m. Consequently, no other trip-related information is available that could be informative for traffic state estimation. However, in recent years, on-demand mobility vehicles and taxis have turned into a large fleet of moving sensors that report at a large scale and in almost real time their trajectories, not only for traffic management but also for third-party applications and research. As this data is constantly collected, it

offers the opportunity for the first time to study the stability and recurrence of congestion patterns at a large scale both temporally and spatially. Predicting patterns in addition to speeds allows for the dimensionality to be reduced to its most essential dimensions, simplifying the prediction and allowing the explanation of them more comprehensibly.

The interest in network-level traffic dynamics models can be traced back to the late 1960s and can be categorized into three eras: (i) flow–speed relationships until 1979, (ii) two-fluid theory from 1979 to 2007, and (iii) the network macroscopic fundamental diagram (NMFD) from 2007 to the present (4). When discussing the suitability of a fleet of moving sensors for network traffic state estimation (5)—or arterial traffic state estimation

¹Chair of Traffic Engineering and Control, Technical University of Munich, Munich, Germany

²Foresight Services, BMW Group, Munich, Germany

Corresponding Author:

Yunfei Zhang, yunfei.zhang@tum.de

(6)—the use of the two-fluid theory seems intriguing as, compared with the other two approaches, it primarily relies on vehicle trajectories and speed measurements that are provided by such a fleet, and no flow measurements. Although its era has come to an end, it is still being used, sometimes together with NMF models, for example, based on taxi data (7) or drone data (8), or as a means for fusing data sources (9). As with the NMF (10, 11), the two-fluid parameters also depend on network topology and network features (12–14). The evidence further shows that the two-fluid parameters depend on driving behavior (aggressive/conservative) and crash rates, resulting from drivers' objective of maximizing the quality of their journeys by traveling fast and maintaining safety (15, 16). Once the vehicle fleet is large enough to be representative of the entire network, this fleet data can be used to detect and model the traffic patterns of the monitored road network. Working on patterns instead of using the full traffic data reduces the dimensionality of the problem (e.g., reducing thousands of streets to a few congestion patterns), which makes the complexity of dynamic urban traffic more comprehensible. Thus, these patterns act as a support for selecting adequate measures for traffic management, for example, if a specific pattern of traffic flows is linked to a bottleneck activation pattern. This network-level perspective has already been shown, for example, for loop detector data (17) and automated number-plate recognition system data (18), where the complexity of urban traffic dynamics has been reduced to a few clusters. It must not be limited to traffic state estimation, but can also be used to inform about other events such as weather (19), from which further measures for traffic management can be drawn. The advancement of deep learning techniques for congestion prediction in the big data age (20) may also support the development of high-resolution spatio-temporal pattern detection algorithms.

However, none of these analyses combine the questions of stability and recurrence of congestion patterns based on trajectory data over a long time period. Stability focuses on how congestion varies over time (range and severity) and how fast the network can recover from congestion, while recurrence studies the repeating patterns of congestion. In some cities, the transportation network company (TNC) already operates large fleets, but the intriguing question is whether such a data source can be used as a sensor for traffic management, in particular traffic state estimation and prediction. Here, we consequently investigate the fundamental suitability of the data source for such problems.

In this paper, we use an open-access trajectory data set from Chengdu, China that covers 30 days, which reports the waypoints of on average 1,250 vehicles circulating simultaneously in the city. From this data, we

estimate the two-fluid relationship (5) and show that the postulated relationship is indeed robust over several days. Also, we find a strong linear relationship as $R^2 \approx 0.98$. To recover as much variance as possible from the data, we then study the stability and recurrence of patterns not on the two-fluid relationship itself, but the residuals between the observed and predicted relationship. In the residual data, we find substantial daily and weekly seasonality. In other words, the deviation from the two-fluid relationship is recurrent and robust over the observation period. In future research, we will extend the time series analysis not only to a longer time period but also to more cities in our samples. This enriched data set will then be used to investigate and find explanatory variables for the distribution and seasonality of residuals. Nevertheless, it should be noted that we lack ground truth data and thus our analysis presents the first evidence that such large-scale fleet trajectory data is an appropriate basis for investigating the performance of urban road networks and the stability and recurrence of congestion patterns.

This paper is organized as follows. In the next section, we introduce the data used in this analysis. Thereafter, we present our methodology to investigate the robustness and recurrence of congestion patterns in Chengdu. We then proceed by presenting the results of our analysis, before closing the paper by discussing our findings.

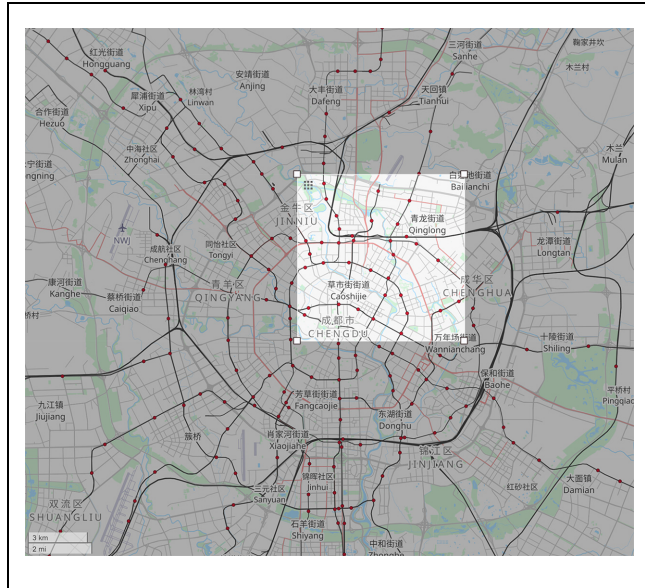
Data Set and Study Area

In this study, we utilize the GPS trajectory data set provided by the Didi Chuxing GAIA Open Dataset. Didi Chuxing is one of the biggest leading TNCs worldwide, providing transportation services such as ride-hailing and ride-sharing. Over 10 billion passenger trips are provided by the Didi platform per year (21). In this study, the data is only from ride-hailing services (22). Here, we use the GPS trajectory data set collected from Chengdu, China in 2016, which has been extensively used by other researchers in previous years. These studies involve different topics including data processing and outlier detection (23), demand prediction (24–29), order dispatching (30, 31), ride-splitting (32), traffic flow prediction (33, 34), and also travel time prediction (35, 36).

The GPS trajectory data was recorded in November 2016 with an average frequency of 3.11 s. The data include five variables: driver ID, order ID, timestamp, longitude, and latitude. In the analyzed data set, driver ID labels the identities of drivers, while order ID stands for individual orders. One driver can accept several orders in a single day, that is, the same driver ID is usually linked to several different order IDs within one day. Driver ID and order ID have already been anonymized for privacy. Examples of the GPS trajectory data are given in Table 1.

Table 1. Format of the Original Data Set

Variable	Example
Anonymous driver ID	389b1a63fca70651270be4d9e6446
Anonymous order ID	413994a1c492c8901d5db1baf1c7c
Timestamp	1477962003
Longitude	104.0579
Latitude	30.67172

**Figure 1.** Study area of the data set.

Note: Scale bar at the bottom left.

The data used in this study cover the area shown in Figure 1 with OpenStreetMap (OSM) as the background (37). It corresponds to the northeast corner of the area within the third ring road in Chengdu, which is covered with a high resolution. For example, for November 1, 2016, there are 32,155,517 GPS records in total, belonging to 181,172 orders and 35,449 drivers. Thus, each driver has roughly 5.11 orders per day, and each order contains 177.5 GPS records. Considering the average frequency as 3.11 s, the average trip duration is 9.2 min.

Figure 2 describes how GPS records are distributed per hour within one day (November 1, 2016). Ride-hailing services are concentrated mainly from 8:00 a.m. to 11:00 p.m. in Chengdu. The original data possess a GPS shift because of the unique Chinese geographical coordinate system (38), which we have fixed during pre-processing. To investigate the coverage of services, we visualize the GPS records as trajectories in Figure 3 with data aggregated from 9:00 a.m. to 9:05 a.m. and in Figure 4 from 9:00 a.m. to 9:01 a.m., on November 1, 2016. We find that a 1 min trajectory data aggregation might not be representative of the study area, while 5 min data

aggregation is able to cover most roads. Therefore, 5 min is selected as the aggregation level. In conclusion, the penetration rate of the ride-hailing fleet by Didi within 5 min is high enough for further aggregation and investigation of the traffic states in the city.

Methodology

We introduce this study's methodology step-wise. Following the introduction of the two-fluid theory, we perform a temporal aggregation of the data to extract macroscopic traffic indicators. Then, we use this aggregated data to estimate the two-fluid model parameters. After calculating the residuals between the real data and the estimation from the two-fluid model, we finally analyze the temporal correlations of the residuals by using time series analysis.

Two-Fluid Model

The two-fluid model is a concise model for urban road traffic developed by Herman and Prigogine (5). According to the model's assumptions, traffic consists of two fluids: moving and stopped vehicles. A speed threshold is selected to define whether a vehicle stops. This threshold might differ among different data sets: for high-frequency recordings of 1 s, a low threshold can capture the stopped state more accurately; for a lower frequency, a looser threshold can avoid misclassification. All involved variables of the two-fluid model are summarized in Table 2.

Based on the definition of the variables in Table 2, the following relationships can be set:

$$T_{min} = 60/v_{max} \quad T = 60/v \quad T = T_r + T_s \quad (1)$$

$$f_r + f_s = 1 \quad (2)$$

Here, trip time per unit distance equals the reciprocal of travel speed. Average trip time per unit distance T contains two parts: moving time T_r and stopped time T_s . The fractions of two fluids are denoted as f_r for the moving vehicles and f_s for the stopped vehicles. By definition, the sum of both fractions has to equal one.

The two-fluid model possesses three key assumptions. The first assumes a linear relationship between the fraction of moving vehicles f_r and the average speed v that exists as Equation 3.

$$v = v_r f_r + v_s f_s = v_r f_r + 0 f_s = v_r f_r \quad (3)$$

The average speed v is defined as the average speed from the moving and stopped fluids: $v = v_r f_r + v_s f_s$. Theoretically, $v_s = 0$.

The second assumption is based on the ergodic theory: the average speed of moving vehicles v_r is proportional to

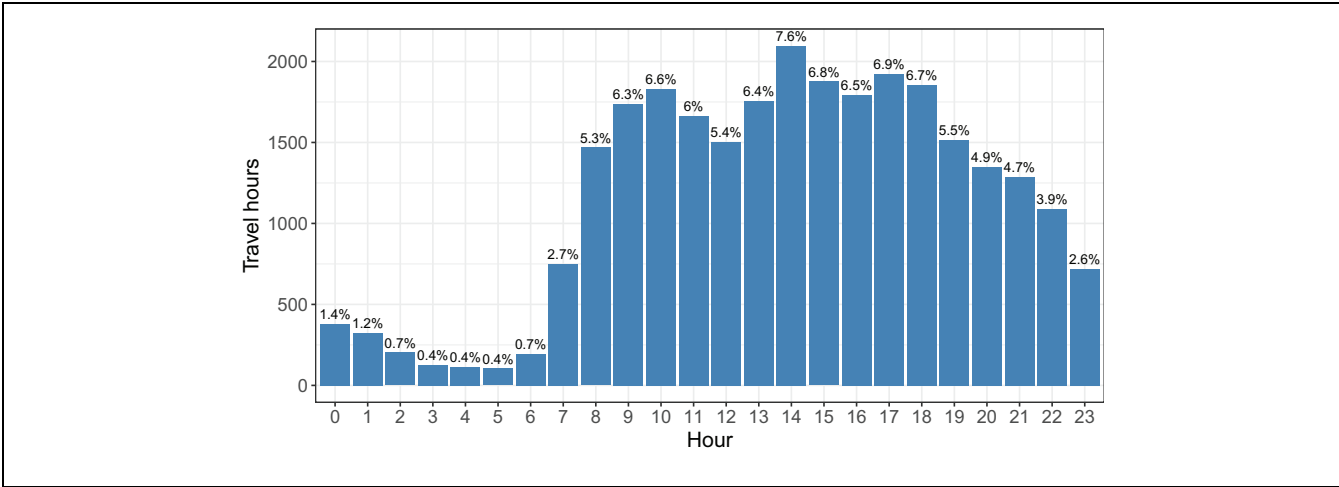


Figure 2. Distribution of global positioning system (GPS) records over hours of one day.

Table 2. Summary of Variables

Category	Variable	Description	Unit
Two-fluid	n	Indicator of network performance	na
	T_{min}	Minimum trip time per unit distance	min/km
Fraction	v_{max}	Maximum average vehicle speed	km/h
	f_r	Fraction of moving vehicles	na
	f_s	Fraction of stopped vehicles	na
Speed	v	Average speed	km/h
	v_r	Average speed of moving vehicles	km/h
	v_s	Average speed of stopped vehicles	km/h
Trip time	T	Average trip time per unit distance	min/km
	T_r	Average running time per unit distance	min/km
	T_s	Average stopped time per unit distance	min/km

Note: na = not applicable.

the fraction of moving vehicles f_r in the entire network as formalized in Equation 4 (5). Here, v_{max} is the maximum average vehicle speed for the whole fleet and n is a model parameter that characterizes the network performance.

$$v_r = v_{max}(1 - f_s)^n = v_{max}(f_r)^n \quad (4)$$

The third assumption states that the ratio of stopped time T_s to average trip time T is approximately equal to the fraction of stopped vehicles in the whole network.

$$\frac{T_s}{T} = f_s \quad (5)$$

Based on these three assumptions, the two-fluid theory's main relationship is finalized in Equation 6.

$$T_s = T - T_{min}^{\frac{1}{n+1}} \cdot T^{\frac{n}{n+1}} \quad (6)$$

where T_{min} and n are network-specific model parameters. As first published by Ardekani and Herman (12), this

can be transformed into a linear relationship between the logarithms of the average trip time T and running time T_r in Equation 7.

$$\log T_r = \frac{1}{n + 1} \log T_{min} + \frac{n}{n + 1} \log T \quad (7)$$

Data Aggregation

To estimate two-fluid parameters, we need first to aggregate raw individual GPS records. Following the findings from Figure 3, we set the aggregation interval to 5 min. Then, we use function *distm* from R package *geosphere* to calculate the distance between two records, which is then divided by the time interval to calculate the vehicle's instant speed.

To calculate the fraction of stopped vehicles f_s , a definition of a speed threshold is required. A vehicle is labeled as "stopped" when its speed is below this threshold. By plotting the density curve of speeds lower than

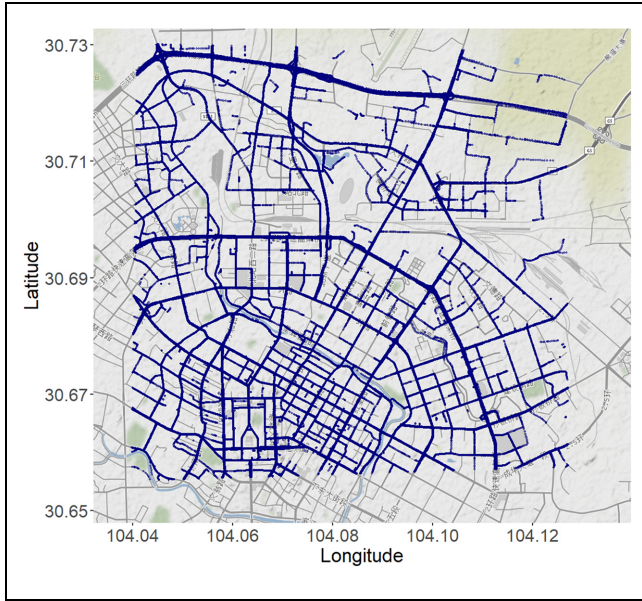


Figure 3. Global positioning system (GPS) trajectories in a 5 min period (9:00–9:05 a.m., November 1, 2016).

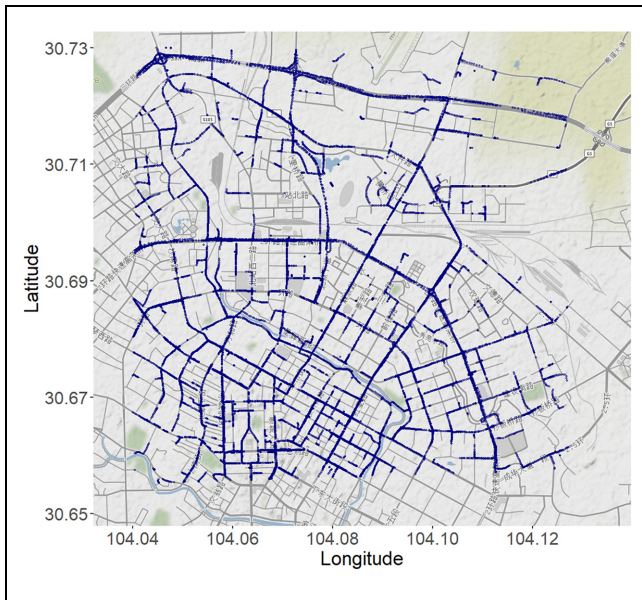


Figure 4. Global positioning system (GPS) trajectories in a 1 min period (9:00–9:01 a.m., November 1, 2016).

20 km/h in Figure 5, we conclude that 5 km/h serves as a good threshold for stopped vehicles. When the speed is equal to zero, the density is the highest. It keeps decreasing between 0 and 5 km/h and becomes stable afterward. Considering the average interval as 3.11 s, the vehicle with an average speed of 5 km/h has traveled roughly 4.5 m between two consecutive observations, which can be assumed to be stopped given the current GPS accuracy.

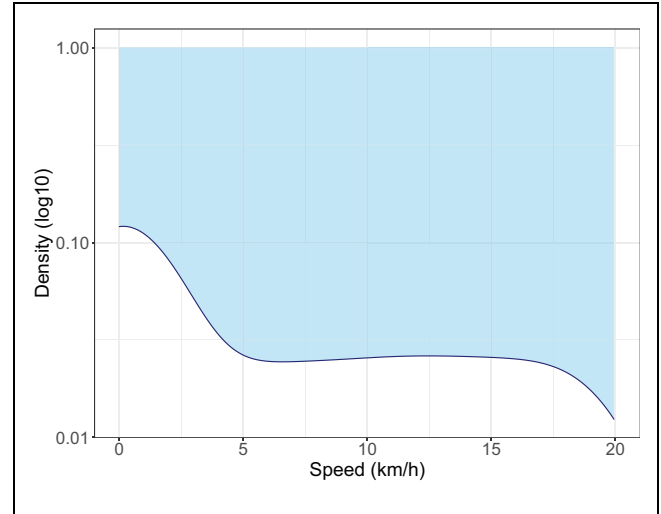


Figure 5. Density plot (logarithm) of speeds lower than 20 km/h in one day.

For each aggregation period, the variables listed in Table 3 are then generated. The speed v is calculated as the average speed of all vehicles in each interval, while the fraction of stopped vehicles f_s is determined as the number of observation points of stopped vehicles' overall trajectory measurements. All observations receive the same weight. Vehicle number n_{veh} is the number of unique vehicles running inside the study area during the aggregation interval.

Estimating the Two-Fluid Parameters and Residuals

Using the linear model from Equation 7, two-fluid parameters n and T_{min} can be estimated using ordinary least squares. Defining that $a = n/(n + 1)$ and $b = 1/(n + 1) \log T_{min}$, the two parameters of interest can be derived as shown in Equations 8 and 9.

$$n = \frac{a}{1 - a} \quad (8)$$

$$T_{min} = 10^{\frac{b}{1-a}} \quad (9)$$

The residuals e are calculated in Equation 10 where \hat{T} is the predicted value from the two-fluid model.

$$e = T - \hat{T} \quad (10)$$

The residuals capture all the variations and trends in the data that are not described by the two-fluid model. The advantage of using the residuals instead of the observed values is that the expectable part is removed from the data and only the information of the deviation part is used for the time series. In describing time series patterns, there is usually a distinction of three

Table 3. Variables of One Aggregated Period

Variable	Symbol	Example	Meaning
Start time	t_s	1477929722	Timestamp of the earliest record
End time	t_e	1477930022	Timestamp of the latest record
Observations	N	2289	Number of available GPS records
Average speed	v	23.92	Average speed (km/h)
Stop fraction	f_s	0.29	Percentage of stopped vehicles
Vehicle number	n_{veh}	13	Number of vehicles in the interval

Note: GPS = global positioning system.

components (39). *Trend* refers to a long-term change in data. *Cycle* occurs when these are repeated patterns such as rises and falls of a non-fixed frequency, while *seasonality*, corresponding to its name, is always of a fixed frequency. In time series analysis, the term trend is also used to combine both trend and cycle as just defined.

Here, we focus on the seasonality in the residuals of the two-fluid model. We assume that seasonality exists in the residuals because of day-of-week (DoW) and hour-of-day factors. Therefore, the residuals are not stationary. To make the time series stationary, we keep differentiating residuals until certain conditions have been met. The number of differentiation steps is called degree d . Here, we use an autoregressive integrated moving average (ARIMA) model, where *integrated* refers to this differentiation process. In addition, ARIMA includes an autoregressive component and a moving average component. The former forecasts the variable using a linear combination of past p values of the variable and the latter using past q forecast errors. An ARIMA model has the following form:

$$y'_t = c + \phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{11}$$

where y'_t is the differentiated time series with degree d , ϕ_i and θ_i are the model coefficients. y'_{t-i} stands for the past prediction and ε_{t-i} is for the past forecast error. Here, the error term ε_{t-i} is assumed to be independent and identically distributed, which follows a normal distribution with mean of zero and variance σ^2 .

Results and Analysis

In this section, we present the results following the same sequence as in the Methodology section.

Data Filtering

Before estimating the two-fluid model, the empirical data has to be checked for outliers as current GPS measurements come with errors. For example, the maximum

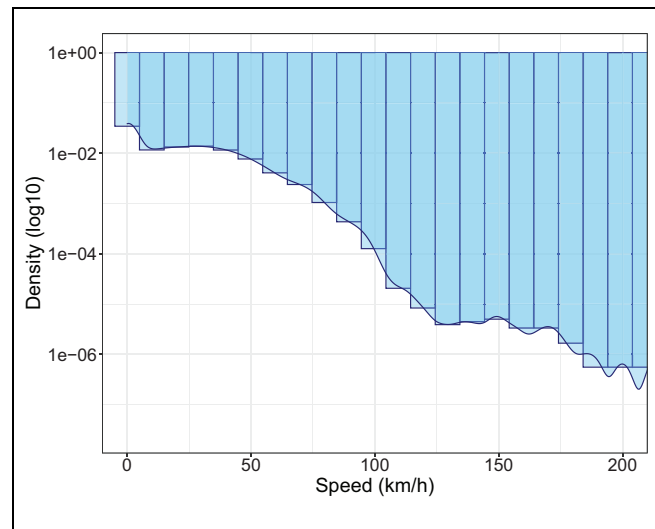


Figure 6. Unfiltered speed distribution of the entire data set.

speed is 288 km/h, which is unreasonable and unfeasible. Therefore, we investigated the speed distribution of all vehicles in the downtown area. Figure 6 shows the distribution of speed values. In total, 0.03% of observations are greater than 80 km/h. Speeds that exceed this limit are consequently rare. Considering the trajectories with speeds over 120 km/h we find that most of them result from sudden GPS drifts. Contrarily, speeds in the range from 80 to 120 km/h still look reasonable but may indicate speeding during non-peak hours as supported by the findings from a single day shown in Figure 7. Thus, we decided to remove all trajectory parts from the data that exceed 120 km/h to avoid an impact of clear erroneous measurements on our results. For the remaining observations, the aggregation of the data to a macroscopic network state may equal the small errors or GPS drift, assuming that the error process itself is unbiased (Gaussian process, etc.).

The next step is to remove all observations with a very small fleet size and most likely an unrepresentative fleet. We remove all observations from 11:59 p.m. to 6:00 a.m. for all days as these time periods are either without

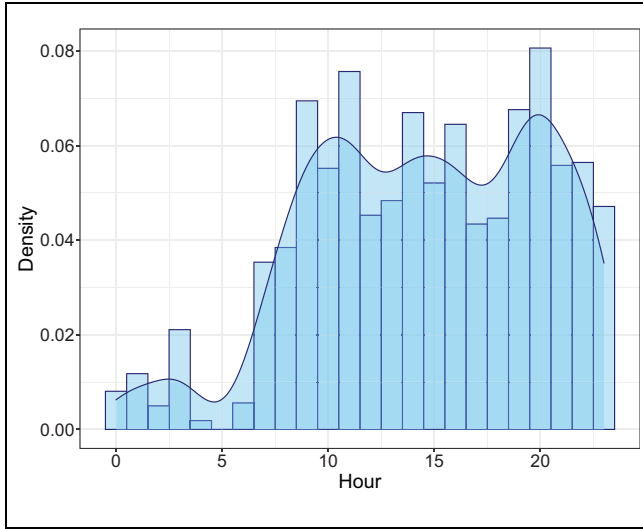


Figure 7. Frequency distribution of 80 to 120 km/h observations on November 1, 2016.

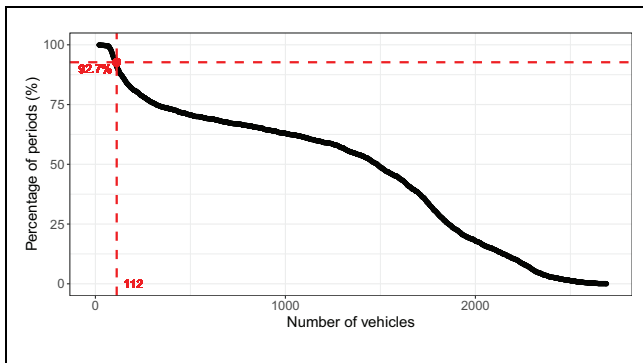


Figure 8. Percentage of observations versus available vehicles.

congestion or with free flow on every road but also with smaller fleet size compared with daytime hours. Further, we define a threshold for the number of vehicles in the fleet. We set this threshold based on the relationship shown in Figure 8 between the number of vehicles in the aggregation period and its (inverted) cumulative share at the location of the steepest slope. This location is identified using differentiation. The threshold is selected as 112 vehicles and we filter out the observations that have fewer vehicles in one aggregation period. An aggregation interval of 5 min results in 288 available intervals per day. Therefore, the 30-day data set contains 8,640 intervals. From these 8,640 observations, 93.0% (i.e., 8,031 observations) are kept for further analysis.

The Two-Fluid Relationship

Each aggregated period contains trip time T and the fraction of stopped vehicles f_s to compute the logarithm

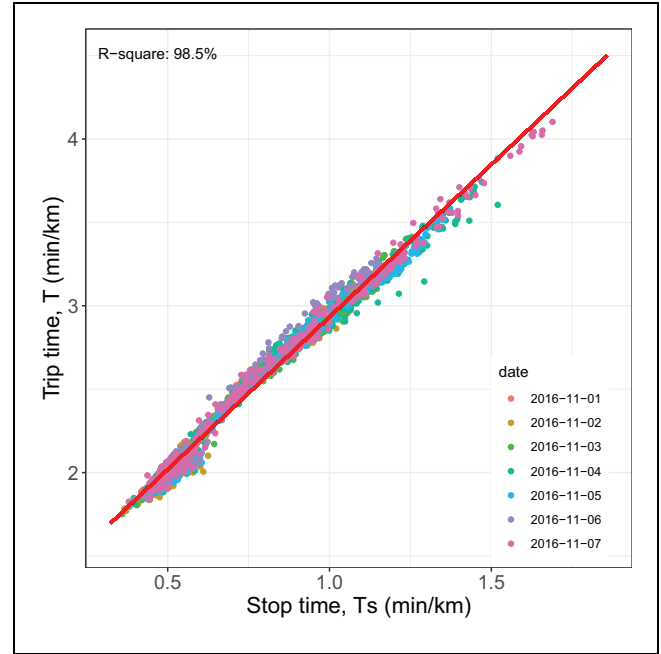


Figure 9. Trip time T versus stop time T_s .

of travel time $\log T$, stop time T_s (see Equation 5), running time T_r (see Equation 1) and its logarithm $\log T_r$. Figure 9 shows the relationship between trip time T and stop time T_s . With the increase of stop time (units: min/km), the average speed is reduced, and thus, trip time T increases. Similarly, a clear linear relation is apparent in Figure 10 as expected from Equation 6, although there are still some outliers for certain days, probably because of non-recurrent events. Consequently, the available data generally supports the application of the two-fluid theory.

Model Estimates Analysis

For linear regression, we do not create one single model for all days but 30 single-day models to find day-specific slopes and intercepts, resulting in series of two-fluid model coefficients n and T_{min} (Equations 8 and 9). Examples from the first seven days are given in Table 4. We conclude with the following findings:

- All models have $R^2 \approx 0.98$, indicating strong linear relationships between dependent variable $\log T_r$ and independent variable $\log T$.
- The two coefficients for the two-fluid model vary slightly, which arguably depends on the interactions of local road network structure and traffic demand.
- Compared with weekdays, weekends, especially Sundays (bold in Table 4), have a substantial change in both coefficients. The minimal trip time

Table 4. Example of Output From Linear Regression Estimation

Date	Day-of-week	Observations	Intercept	Slope	R ²	n	T _{min}
2016-11-01	Tue	265	-0.0284	0.680	0.986	2.13	0.814
2016-11-02	Wed	248	-0.0227	0.667	0.987	2.00	0.855
2016-11-03	Thu	256	-0.0213	0.661	0.980	1.95	0.865
2016-11-04	Fri	257	-0.0152	0.642	0.979	1.80	0.907
2016-11-05	Sat	277	-0.0244	0.664	0.987	1.98	0.846
2016-11-06	Sun	274	-0.0349	0.700	0.986	2.34	0.765
2016-11-07	Mon	256	-0.0167	0.653	0.989	1.89	0.895

Note: Bold font highlights substantial change in both coefficients on Sunday.

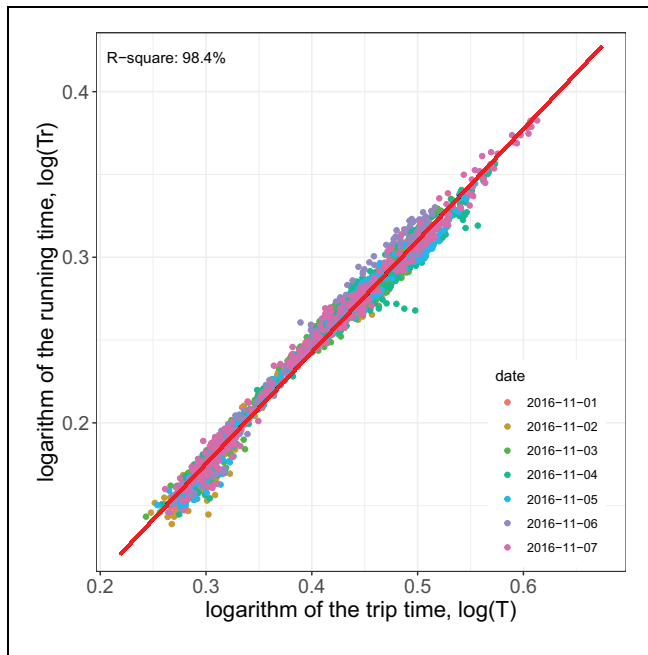


Figure 10. Logarithm of running time $\log T_r$, versus logarithm of trip time $\log T$.

T_{min} is the lowest on Sunday, indicating the highest maximum speed.

Figure 11 compares how the model coefficients fluctuate over time: the red line shows n and the blue line represents t_{min} . Table 5 presents the sample summary statistics of these two coefficients. Though there are minor differences, both variables show a correlation coefficient of -0.991 . This value indicates that these two coefficients are strongly correlated. In the following, we analyze the data by exploiting the time series seasonality patterns that cannot be represented by a linear regression model. Future research can then investigate how temporal demand patterns influence this distribution.

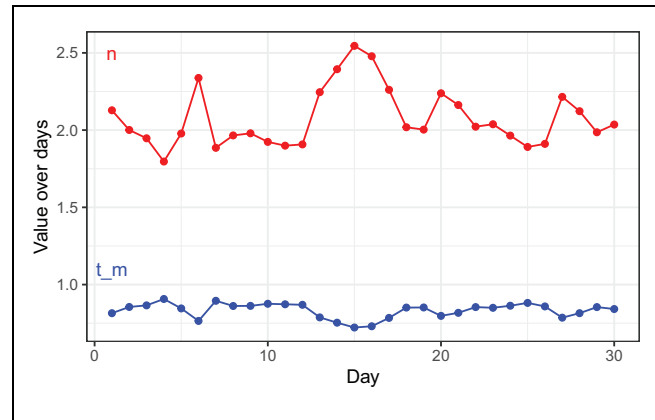


Figure 11. Fluctuation of model coefficients over 30 days.

Table 5. Two-Fluid Theory Parameters

Variable	Sample size	Mean	Variance	Standard deviation
n	30	2.706	0.035	0.186
T_{min}	30	0.833	0.002	0.048

Residual Analysis

Residuals are calculated from the model estimations and visualized in Figure 12 (absolute values) and Figure 13 (scaled residuals). The red horizontal dashed line indicates zero and “DoW” represents day-of-week. In both figures, significant recurring daily patterns can be observed that we can extrapolate to a similar weekly pattern.

Figures 14 and 15 present the daily and weekly seasonality extracted from a time series decomposition, where the frequency is set to one day and one week. Daily seasonality shows a “W”-shaped pattern with two drop-downs that might be caused by two traffic peaks and corresponding demand increases. Similarly, weekly

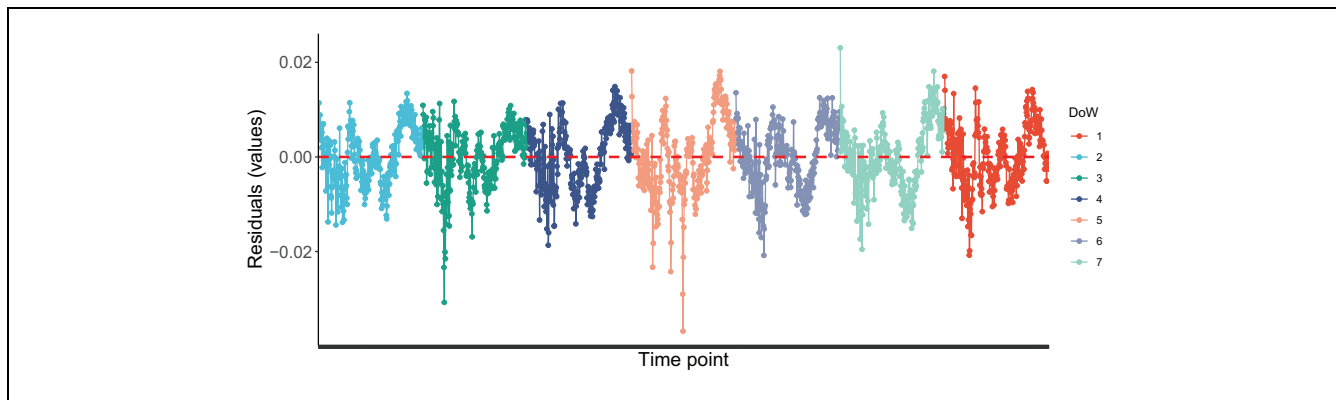


Figure 12. Residual values over one week.

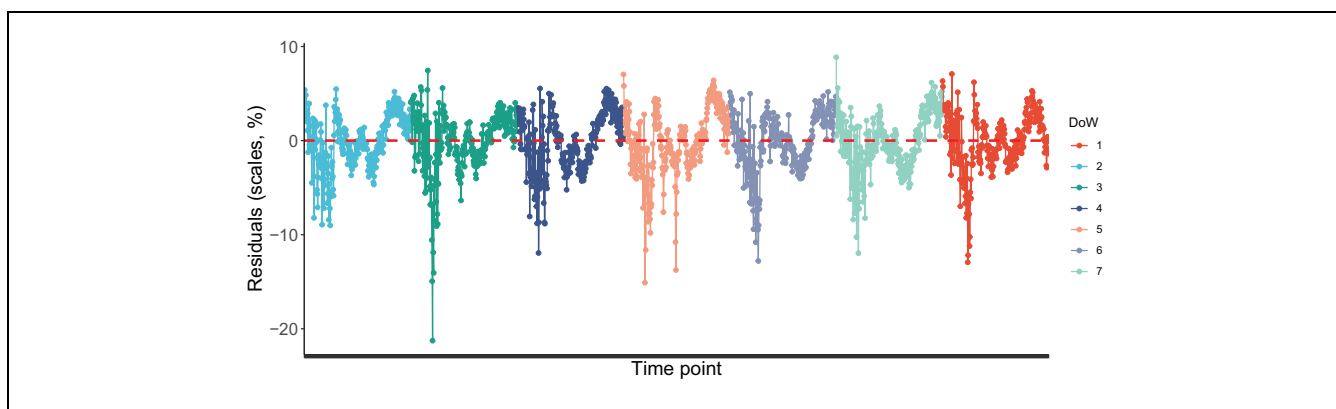


Figure 13. Residual scales (percentage) over one week.

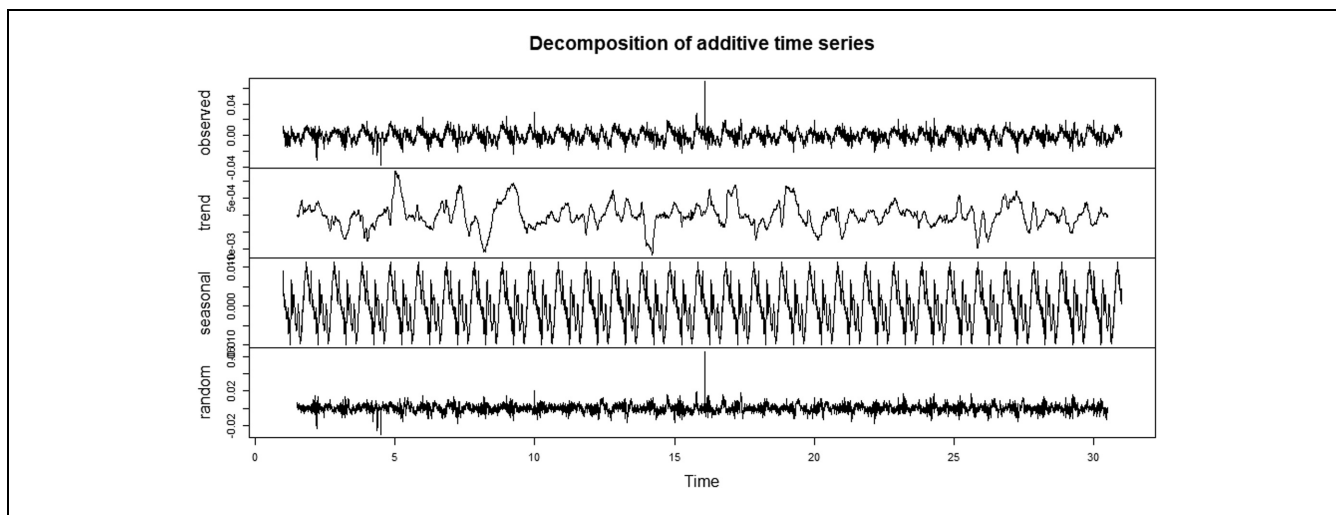


Figure 14. Seasonality from time series decomposition: daily pattern.

seasonality can also be viewed as a combination of seven consecutive daily patterns.

For modeling time series data, an ARIMA model is used. Since seasonality exists in the data, the observed

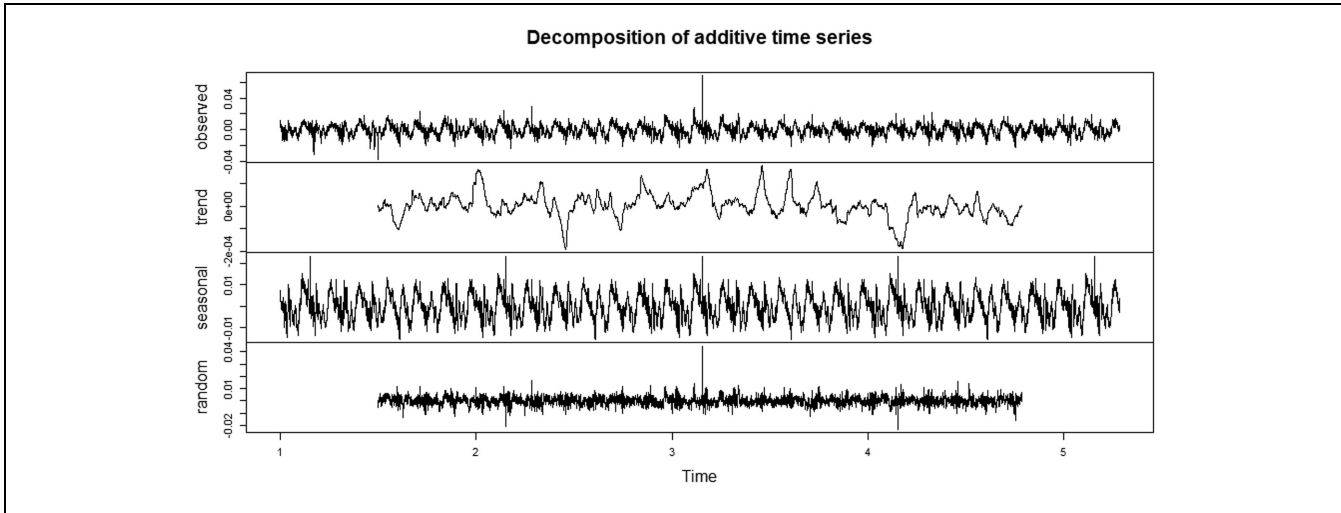


Figure 15. Seasonality from time series decomposition: weekly pattern.

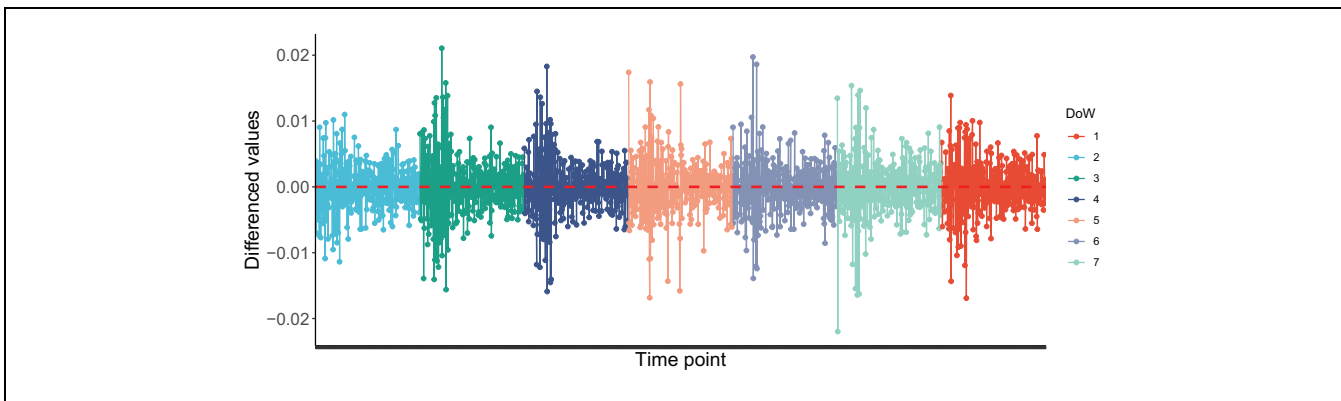


Figure 16. Differenced residual time series over one week.

Note: DoW = day-of-week.

residual time series data is not yet stationary. To test statistically for this, we utilize the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (40), a *unit root test*, with the null hypothesis being that the data is stationary. After one degree differencing, the p-value of the test is as low as 0.0022, proving that the null hypothesis must be rejected. Thus, ARIMA can be applied on the differenced residual time series as shown in Figure 16.

The *auto.arima* function from R package *forecast* has been used to search iteratively for the best ARIMA model. The best ARIMA has been proposed with a dimension of $(0,0,4)$. Considering the one degree of differencing already made, this leads to the final ARIMA model with $p = 0$, $d = 1$, and $q = 4$. This model is a special case of the moving average model. We then checked

the residuals from the ARIMA model (i.e., not the input residuals) shown in Figure 17. The results suggest that the residuals follow a normal distribution with zero mean. There is no further significant correlation in the residuals' time series as seen in the auto-correlation function. However, the time plot of the residuals exhibits a sharp peak in the middle of the time series, which might be from a special non-recurrent event in the road network. Except for this sharp peak, the variation of the residuals stays within the same limits over time, that is, the variance is constant.

From the stable two-fluid model coefficients and the revealed seasonality, we can conclude that our approach to using TNC vehicles for macroscopic traffic state estimation with the two-fluid theory seems robust.

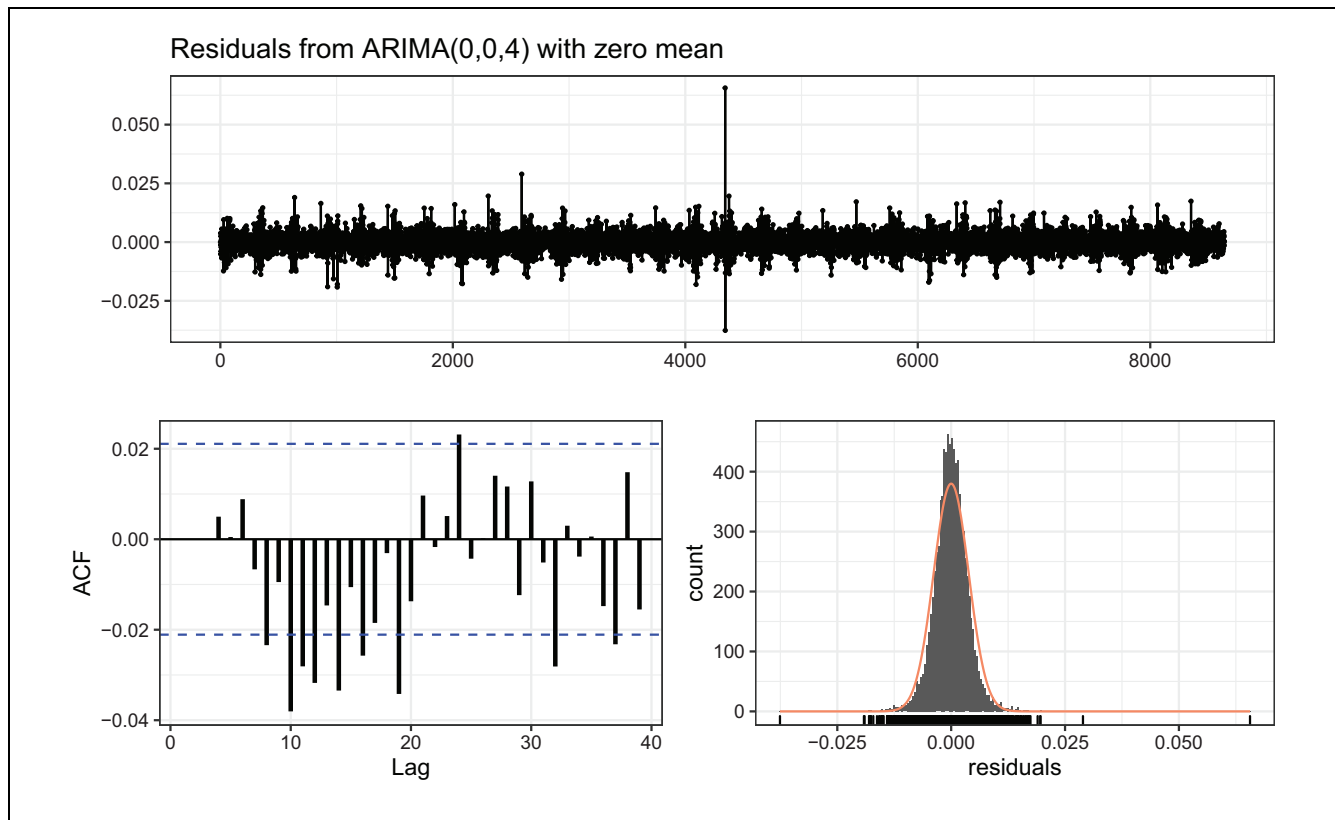


Figure 17. Checks of the autoregressive integrated moving average (ARIMA) model.

Note: ACF = auto-correlation function.

Consequently, we can utilize such data and models to forecast traffic states with recurrent and non-recurrent components.

Conclusion

This research retrieved one-month GPS trajectory data from Didi Chuxing, a TNC providing large amounts of on-demand mobility services in China. The original data set comprised more than 30 million GPS records per day in Chengdu, China. The original data offer a good coverage of the road network and recurrent patterns, and the data set thus is capable of representing the general traffic state. We aggregated the data to 5 min intervals and estimated traffic states that are required for the two-fluid model of urban traffic (5). Our results showed, first, that traffic in Chengdu indeed exhibits convincingly the relationship postulated by the two-fluid theory with an average $R^2 \approx 0.98$. Second, estimating daily two-fluid theory parameters (Table 5) revealed that they are robust concerning between-day variation. Third, we calculated the residuals of the two-fluid model to understand the fluctuations around the general model trend and to further investigate its temporal patterns. We found that the

residuals exhibit a strong daily and weekly pattern, which can be modeled by an ARIMA model, supporting the idea that the congestion patterns are robust and recurrent, and thus they can be used for predictions. Based on this, in future research, we will then establish the relationship between spatio-temporal traffic patterns and the two-fluid model parameters as well as the model's residuals.

The presented analysis will be developed further in several directions to address its limitations. First, the data is still biased, likely caused by the single data source from ride-hailing vehicles. It is reasonable that ride-hailing vehicles tend to travel more within the entertainment and restaurants area, causing a biased estimation of the total average speed for the whole network. To solve this, we plan in the next steps to introduce more data sources, like loop detector data, and use data fusion techniques to understand and then minimize the potential biases. In future research we will also extend the sample not only to a longer time period but also to include more cities. This will enable further study on how representative TNC vehicles are as a sensor for traffic management and how stable and recurrent congestion patterns are across cities.

In closing, our analysis has three important implications. First, the results presented in this paper contribute to recent work on estimating the parameters for the two-fluid theory model. On the one hand, these results support earlier findings that taxi GPS data can indeed be used for two-fluid theory model parameter estimation and subsequent network monitoring (7), but at a much larger scale. Building on the multi-modal extension of the two-fluid model presented by Paipuri et al. (8), our findings (robust parameters and predictable seasonality) underline our motivation that such taxi vehicles can be used as moving sensors to inform about the multi-modal traffic state once a multi-modal speed model like the two-fluid model is calibrated, for example, based on drone data. This link will be investigated in future research. Second, having a large fleet of moving sensors is a promising tool for monitoring and predicting the performance of urban road networks. Our results have shown that the data from on-demand mobility services can be used to inform about the network-wide traffic state in its dynamics and stability of patterns over time. Although we lack a ground truth reference to assess the representativeness of the data, we can infer that revealed patterns can be used for predicting comparative traffic patterns. Consequently, traffic management centers should have an interest in obtaining such trajectory data for improving their traffic state estimation, in particular when complemented with a pattern prediction. As many cities already have a fleet of vehicles for many services, they could in principle rely on them as moving sensors if they do not wish to rely on (commercial) on-demand mobility trajectory data. Third, as trajectory data is now available at a large scale to estimate the two-fluid models in almost every city, this model should be further exploited to understand which factors drive the network performance, similar to an analysis based on stationary detector data (11).

Acknowledgments

The authors would like to thank Didi Chuxing for providing the GPS trajectory data (Data source: Didi Chuxing GAIA Initiative, <https://gaia.didichuxing.com>).

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Y. Zhang, A. Loder, F. Rempe, K. Bogenberger; data collection: Y. Zhang; analysis and interpretation of results: Y. Zhang, A. Loder; draft manuscript preparation: Y. Zhang, A. Loder. All authors reviewed the results and approved the final version of the manuscript.


Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding


The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Yunfei Zhang acknowledges the support from the German Federal Ministry for Digital and Transport (BMVI) for the funding of the project TEMPUS (Testbed Munich - Pilot test of urban automated road traffic), grant no. 01MM20008K. Allister Loder acknowledges the support from the German Federal Ministry for Digital and Transport (BMVI) for the funding of the project KIVI (Artificial Intelligence in Ingolstadt's Transportation System), grant no. 45KI05A011.

ORCID iDs

Yunfei Zhang  <https://orcid.org/0000-0003-1902-1816>

Allister Loder  <https://orcid.org/0000-0003-3102-6564>

Felix Rempe  <https://orcid.org/0000-0002-8007-8152>

Klaus Bogenberger  <https://orcid.org/0000-0003-3868-9571>

References

1. Bickel, P., C. Chen, J. Kwon, J. Rice, E. van Zwet, and P. Varaiya. Measuring Traffic. *Statistical Science*, Vol. 22, No. 4, 2007, pp. 581–597.
2. Ambühl, L., A. Loder, M. Menendez, and K. W. Axhausen. Empirical Macroscopic Fundamental Diagrams: New Insights From Loop Detector and Floating Car Data. In *TRB 96th Annual Meeting Compendium of Papers*. Transportation Research Board, 2017, p. 17–03331.
3. Ambühl, L., and M. Menendez. Data Fusion Algorithm for Macroscopic Fundamental Diagram Estimation. *Transportation Research Part C: Emerging Technologies*, Vol. 71, 2016, pp. 184–197. <https://doi.org/10.1016/j.trc.2016.07.013>.
4. Johari, M., M. Keyvan-Ekbatani, L. Leclercq, D. Ngoduy, and H. S. Mahmassani. Macroscopic Network-Level Traffic Models: Bridging Fifty Years of Development Toward the Next Era. *Transportation Research Part C: Emerging Technologies*, Vol. 131, 2021, p. 103334. <https://doi.org/10.1016/J.TRC.2021.103334>.
5. Herman, R., and I. Prigogine. A Two-Fluid Approach to Town Traffic. *Science*, Vol. 204, No. 4389, 1979, pp. 148–151.
6. Jones, E. G., and W. Farhat. Validation of Two-Fluid Model of Urban Traffic for Arterial Streets. *Transportation Research Record: Journal of the Transportation Research Board*, 2004. 1876: 132–141.
7. Lu, S., V. L. Knoop, and M. Keyvan-Ekbatani. Using Taxi GPS Data for Macroscopic Traffic Monitoring in Large Scale Urban Networks: Calibration and MFD Derivation. *Transportation Research Procedia*, Vol. 34, 2018, pp. 243–250.
8. Paipuri, M., E. Barmounakis, N. Geroliminis, and L. Leclercq. Empirical Observations of Multi-Modal Network-Level Models: Insights From the pNEUMA Experiment. *Transportation Research Part C: Emerging Technologies*, Vol. 131, 2021, p. 103300. <https://doi.org/10.1016/J.TRC.2021.103300>.

9. Loder, A., L. Bressan, M. J. Wierbos, H. Becker, A. Emmonds, M. Obee, V. L. Knoop, M. Menendez, and K. W. Axhausen. How Many Cars in the City Are too Many? Towards Finding the Optimal Modal Split for a Multi-Modal Urban Road Network. *Frontiers in Future Transportation*, Vol. 2, 2021, p. 665006. <https://doi.org/10.3389/ffutr.2021.665006>.
10. Daganzo, C. F., and N. Geroliminis. An Analytical Approximation for the Macroscopic Fundamental Diagram of Urban Traffic. *Transportation Research Part B: Methodological*, Vol. 42, 2008, pp. 771–781.
11. Loder, A., L. Ambühl, M. Menendez, and K. W. Axhausen. Understanding Traffic Capacity of Urban Networks. *Scientific Reports*, Vol. 9, No. 1, 2019, pp. 1–10.
12. Ardekani, S. A., and R. Herman. A Comparison of the Quality of Traffic Service in Downtown Networks of Various Cities Around the World. *Traffic Engineering & Control*, Vol. 26, No. 12, 1985, pp. 574–581.
13. Williams, J. C., H. Mahmassani, and R. Herman. Analysis of Traffic Network Flow Relations and Two-Fluid Model Parameter Sensitivity. *Transportation Research Record: Journal of the Transportation Research Board*, 1985. 1005: 95–106.
14. Ardekani, S. A., J. C. Williams, and S. Bhat. Influence of Urban Network Features on Quality of Traffic Service. *Transportation Research Record: Journal of the Transportation Research Board*, 1992. 1358: 6–12.
15. Dixit, V. V., A. Pande, M. Abdel-Aty, A. Das, and E. Radwan. Quality of Traffic Flow on Urban Arterial Streets and its Relationship With Safety. *Accident Analysis & Prevention*, Vol. 43, No. 5, 2011, pp. 1610–1616.
16. Dixit, V. V. Behavioural Foundations of Two-Fluid Model for Urban Traffic. *Transportation Research Part C: Emerging Technologies*, Vol. 35, 2013, pp. 115–126. <https://doi.org/10.1016/j.trc.2013.06.009>.
17. Ambühl, L., A. Loder, L. Leclercq, and M. Menendez. Disentangling the City Traffic Rhythms: A Longitudinal Analysis of MFD Patterns Over a Year. *Transportation Research Part C: Emerging Technologies*, Vol. 126, 2021, p. 103065. <https://doi.org/10.1016/j.trc.2021.103065>.
18. Lopez, C., L. Leclercq, P. Krishnakumari, N. Chiabaut, and H. van Lint. Revealing the Day-to-Day Regularity of Urban Congestion Patterns With 3D Speed Maps. *Scientific Reports*, Vol. 7, 2017, p. 14029. <https://doi.org/10.1038/s41598-017-14237-8>.
19. Hintz, K. S., K. O’Boyle, S. L. Dance, S. Al-Ali, I. Ansper, D. Blaauboer, M. Clark, et al. Collecting and Utilising Crowdsourced Data for Numerical Weather Prediction: Propositions From the Meeting Held in Copenhagen, 4–5 December 2018. *Atmospheric Science Letters*, Vol. 20, No. 7, 2019, p. e921.
20. Kumar, N., and M. Raubal. Applications of Deep Learning in Congestion Detection, Prediction and Alleviation: A Survey. *Transportation Research Part C: Emerging Technologies*, Vol. 133, 2021, p. 103432. <https://doi.org/10.1016/j.trc.2021.103432>.
21. Didi Chuxing. *About Us*. Didi, 2021. <https://www.didiglobal.com/about-didi/about-us>. Accessed July 31, 2021.
22. Didi Chuxing. DiDi Research Outreach Initiative. Didi, 2022. <https://outreach.didichuxing.com/en/>. Accessed October 13, 2022.
23. Zhang, J., and Y. Sun. An Automatic Data Cleaning Method for GPS Trajectory Data on Didi Chuxing GAIA Open Dataset Using Machine Learning Algorithms. *Proc., 6th International Conference on Systems and Informatics (ICSAI)*, Shanghai, China, IEEE, New York, 2019, pp. 1522–1526.
24. Wang, C., Y. Hou, and M. Barth. Data-Driven Multi-Step Demand Prediction for Ride-Hailing Services Using Convolutional Neural Network. In *Proc., Advances in Computer Vision: Science and Information Conference* (K. Arai, and S. Kapoor, eds.), Las Vegas, NV, April 25–26, 2019, Springer, Cham, pp. 11–22.
25. Huang, Z., G. Huang, Z. Chen, C. Wu, X. Ma, and H. Wang. Multi-Regional Online Car-Hailing Order Quantity Forecasting Based on the Convolutional Neural Network. *Information*, Vol. 10, No. 6, 2019, p. 193.
26. Niu, K., C. Wang, X. Zhou, and T. Zhou. Predicting Ride-Hailing Service Demand via RPA-LSTM. *IEEE Transactions on Vehicular Technology*, Vol. 68, No. 5, 2019, pp. 4213–4222.
27. Kuang, L., X. Yan, X. Tan, S. Li, and X. Yang. Predicting Taxi Demand Based on 3D Convolutional Neural Network and Multi-Task Learning. *Remote Sensing*, Vol. 11, No. 11, 2019, p. 1265.
28. Liang, X. *Applied Deep Learning in Intelligent Transportation Systems and Embedding Exploration*. PhD thesis. New Jersey Institute of Technology, 2019.
29. Khezerlou, A. V., X. Zhou, L. Tong, Y. Li, and J. Luo. Forecasting Gathering Events Through Trajectory Destination Prediction: A Dynamic Hybrid Model. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, No. 3, 2019, pp. 991–1004.
30. Shi, D., X. Li, M. Li, J. Wang, P. Li, and M. Pan. Optimal Transportation Network Company Vehicle Dispatching via Deep Deterministic Policy Gradient. In *Proc., Wireless Algorithms, Systems, and Applications: International Conference on Wireless Algorithms, Systems, and Applications* (E. Biagioni, Y. Zheng, and S. Cheng, eds.), Honolulu, HI, June 24–26, 2019, Springer, Cham, pp. 297–309.
31. He, S., and K. G. Shin. Spatio-Temporal Capsule-Based Reinforcement Learning for Mobility-on-Demand Network Coordination. *Proc., The World Wide Web Conference*, San Francisco, CA, 2019, pp. 2806–2813.
32. Li, W., Z. Pu, Y. Li, and X. J. Ban. Characterization of Ridesplitting Based on Observed Data: A Case Study of Chengdu, China. *Transportation Research Part C: Emerging Technologies*, Vol. 100, 2019, pp. 330–353.
33. Zhang, Y., T. Cheng, and Y. Ren. A Graph Deep Learning Method for Short-Term Traffic Forecasting on Large Road Networks. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 34, No. 10, 2019, pp. 877–896.
34. Luo, X., B. Liu, P. J. Jin, Y. Cao, and W. Hu. Arterial Traffic Flow Estimation Based on Vehicle-to-Cloud Vehicle Trajectory Data Considering Multi-Intersection Interaction and Coordination. *Transportation Research Record:*

- Journal of the Transportation Research Board*, 2019. 2673: 68–83.
35. Gao, R., X. Guo, F. Sun, L. Dai, J. Zhu, C. Hu, and H. Li. Aggressive Driving Saves More Time? Multi-Task Learning for Customized Travel Time Estimation. *Proc., 28th International Joint Conference on Artificial Intelligence, IJCAI*, Macao, China, 2019, pp. 1689–1696.
 36. Zhang, X., L. Xie, Z. Wang, and J. Zhou. Boosted Trajectory Calibration for Traffic State Estimation. *Proc., 2019 IEEE International Conference on Data Mining (ICDM)*, Beijing, China, IEEE, New York, 2019, pp. 866–875.
 37. OpenStreetMap. OpenStreetMap. 2021. <https://openstreetmap.org>. Accessed July 31, 2021.
 38. Lee, G., et al. *eviltransform*. googollee, 2015. <https://github.com/googollee/eviltransform>.
 39. Kaiser, R., and A. Maravall Herrero. *Short-Term and Long-Term Trends, Seasonal Adjustment, and the Business Cycle*. Banco de España, Servicio de Estudios, Madrid, Spain, 1999.
 40. Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin. Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure Are we That Economic Time Series Have a Unit Root? *Journal of Econometrics*, Vol. 54, No. 1–3, 1992, pp. 159–178.