mathematics

MDPI

Article

On the Fundamental Diagram for Freeway Traffic: Exploring the Lower Bound of the Fitting Error and Correcting the Generalized Linear Regression Models

Yidan Shangguan [1], Xuecheng Tian [1,*], Sheng Jin [2], Kun Gao [3], Xiaosong Hu [4], Wen Yi [5], Yu Guo [1] and Shuaian Wang [1]

[1] Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University, Hung Hom, Hong Kong; 18092732117@163.com (Y.S.)
[2] College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China; jinsheng@zju.edu.cn
[3] Department of Architecture and Civil Engineering, Chalmers University of Technology, 412 96 Göteborg, Sweden
[4] State Key Laboratory of Mechanical Transmission/Automotive Collaborative Innovation Center, Chongqing University, Chongqing 400044, China
[5] Department of Building and Real Estate, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
* Correspondence: xuecheng-simon.tian@connect.polyu.hk

Abstract: In traffic flow, the relationship between speed and density exhibits decreasing monotonicity and continuity, which is characterized by various models such as the Greenshields and Greenberg models. However, some existing models, i.e., the Underwood and Northwestern models, introduce bias by incorrectly utilizing linear regression for parameter calibration. Furthermore, the lower bound of the fitting errors for all these models remains unknown. To address above issues, this study first proves the bias associated with using linear regression in handling the Underwood and Northwestern models and corrects it, resulting in a significantly lower mean squared error (MSE). Second, a quadratic programming model is developed to obtain the lower bound of the MSE for these existing models. The relative gaps between the MSEs of existing models and the lower bound indicate that the existing models still have a lot of potential for improvement.

Keywords: speed and density relationship; linear regression; quadratic programming

MSC: 90-10

check for updates

Citation: Shangguan, Y.; Tian, X.; Jin, S.; Gao, K.; Hu, X.; Yi, W.; Guo, Y.; Wang, S. On the Fundamental Diagram for Freeway Traffic: Exploring the Lower Bound of the Fitting Error and Correcting the Generalized Linear Regression Models. *Mathematics* **2023**, *11*, 3460. https://doi.org/10.3390/math11163460

Academic Editor: Chunqiao Tan

Received: 4 July 2023
Revised: 25 July 2023
Accepted: 8 August 2023
Published: 9 August 2023

1. Introduction

The traffic fundamental diagram is crucial in traffic flow theory [1–5], representing the relationship between traffic flow (vehs/h), speed (km/h), and traffic density (vehs/km). Greenshields [1] first proposed a linear model to describe the relationship between speed and density and made a pioneering work in this field. This rudimentary relationship has since been refined through the introduction of numerous models [3–18]. These studies try to define precise relationships, utilizing practical parameters to reflect the traffic flow features more accurately. This paper focuses on the four well-known models listed in Table 1, each having two parameters.

Table 1. Four speed–density models (Qu et al., 2015) [19].

Models	Function	Parameters
Greenshields [1]	$v = v_f\left(1 - \frac{k}{k_j}\right)$	v_f, k_j
Greenberg [3]	$v = v_0 ln\left(\frac{k_j}{k}\right)$	v_0, k_j

Table 1. *Cont.*

Models	Function	Parameters
Underwood [5]	$v = v_f \exp\left(-\frac{k}{k_0}\right)$	v_f, k_0
Northwestern [20]	$v = v_f \exp\left[-\frac{1}{2}\left(\frac{k}{k_0}\right)^2\right]$	v_f, k_0

Note: v denotes the speed (the dependent variable), km/h; k denotes the density (the independent variable), veh/km; v_f denotes the free-flow speed, km/h; k_j denotes the jam density, veh/km; k_0 denotes the at-capacity density, veh/km; v_0 denotes the at-capacity speed, km/h.

At the same time, a great number of calibration models have been proposed related to these well-known models. Qu et al. [19] proposed a least-squares method to calibrate the model so that the model can be applied to both in light-traffic/free-flow conditions and congested/jam conditions. Fan and Seibold [21] and Qu et al. [22] published research works using data-driven approaches to generate a percentile-based speed–density relationships for freeway traffic. Wang [23] addressed the shortcomings of data-driven stochastic fundamental maps of diagram traffic by proposing a holistic modelling framework based on the concept of mean absolute error minimization. For more related literature, please refer to Bramich et al. [24]. Nearly all existing studies employ linear regression to solve these famous models and estimate parameters [25–28]. For models that cannot be solved directly by linear regression, such as the Underwood and Northwestern models, many researchers resort to defining $y = \ln v$ and $x = k$ for the Underwood model and $y = \ln v$ and $x = k^2$ for the Northwestern model to transform them into linear models of (x, y), whose parameters can be easily estimated by linear regression. However, this transformation is fundamentally flawed, as it fails to obtain an unbiased estimate of v. The problem arises from the fact that the estimate of parameter $\ln v$ cannot accurately represent the estimate of parameter v, leading to a distorted and biased final estimate. Given this challenge, this study aims to address this issue.

In the calibration and validation of traffic flow fundamental diagrams, numerous studies use a specific dataset [13,19,22,23,29,30], which makes our comparison more consistent, as shown in Figure 1. This dataset comprises 47,815 speed-density observations collected over a year by loop detectors from 76 stations on Georgia State Route 400 (hereafter referred to as the GA400 dataset). The GA400 dataset facilitates the examination of the performance of the four models, as shown in Figure 1. Each of the four models has its own strengths when describing the characteristics of the speed and density relationship: for example, the Greenshields and Northwestern models perform better in low-density datasets, while the Underwood model performs better in medium- to high-density datasets. Despite the widespread application of the four models, a key issue—the gap between their fittings and the "ideal" lower bound of the fitting error—remains unanswered in the existing literature. To address this research gap, this paper defines the model that minimizes the MSE of the dataset among all monotonically decreasing models as an "ideal" prediction model whose optimal objective function value is thus termed the lower bound of the fitting error.

The main contributions of this paper are twofold. We first show that applying the transformation on the Underwood and Northwestern models produces biased results. In response to this finding, we correct the methodological errors involved in using linear regression for parameter estimation in these models. Second, we construct a quadratic programming model with the objective of minimizing the MSE to find the "ideal" lower bound of the fitting error for existing models. The results show that the average relative gap between the lower bound and the MSEs of existing models is about 197.322%. Therefore, there is still a lot of room for further development of existing models.

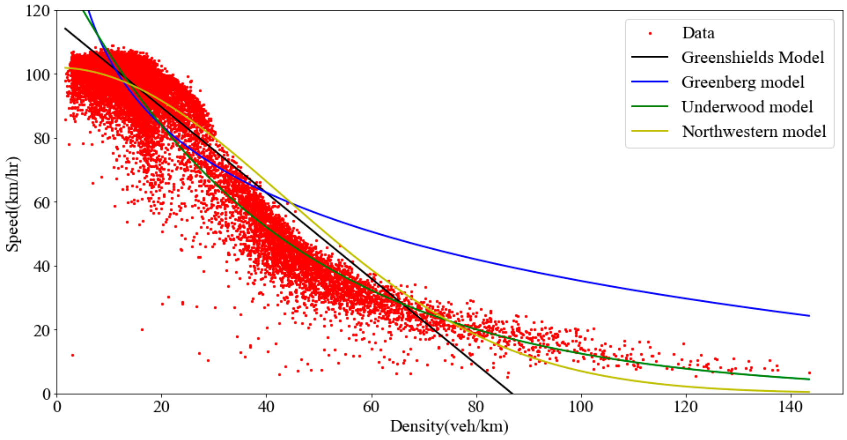


Figure 1. Performance of four models in the GA400 dataset.

The rest of the paper is organized as follows. In Section 2, we prove that using linear regression to calibrate nonlinear relationships between k and v is biased and then correct this error using the enumeration approach. Section 3 establishes a quadratic programming model to find the "ideal" lower bound of the fitting error of existing models. Section 4 concludes this study.

2. Correcting Generalized Linear Regression Models

2.1. Analysis

In the existing studies, the parameters v_f and k_0 of Underwood and Northwestern models are estimated by linear regression. The procedures are as follows.

In the Underwood model, $v = v_f \exp\left(1 - \frac{k}{k_0}\right)$, and the parameters to be estimated are v_f and k_0. By taking the logarithm on both sides of the equation, the model is equivalent to $\ln v - \ln v_f = -\frac{k}{k_0}$. After letting $y = \ln v$ and $x = k$, the model is transformed into $y = \ln v_f - \frac{x}{k_0}$. By performing a linear regression on x and y, we obtain the equation $y = ax + b$, where a and b are the parameters derived from the regression. Consequently, the parameters v_f and k_0 can be estimated as $v_f = \exp(b)$ and $k_0 = -\frac{1}{a}$.

In the Northwestern model, $v = v_f \exp\left[-\frac{1}{2}\left(\frac{k}{k_0}\right)^2\right]$, and the parameters to be estimated are v_f and k_0. By taking the logarithm on both sides of the equation, the model is equivalent to $\ln v = \ln v_f - \frac{1}{2}\left(\frac{k}{k_0}\right)^2$. After letting $y = \ln v$ and $x = k^2$, the model is transformed into $y = \ln v_f - \frac{1}{2}\frac{x}{k_0{}^2}$. By performing a linear regression on x and y, we obtain the equation $y = cx + d$, where c and d are the parameters derived from the regression. Consequently, the parameters v_f and k_0 can be estimated as $v_f = \exp(d)$ and $k_0 = \sqrt{-\frac{1}{2c}}$.

The above procedures use the logarithm of v and then apply linear regression. In order to correctly use linear regression to estimate the parameters of the models, we should guarantee that the unbiased estimate of v is equivalent to the exponential of the unbiased estimate of y. However, this condition may not be satisfied in some cases. For example, assume v has three realizations: 3, 4, and 5. The unbiased estimate of the expectation of v is 4 (the sample mean); however, $\exp\left(\frac{\ln 3 + \ln 4 + \ln 5}{3}\right) \approx 3.915$ is not the original unbiased estimate of the expectation of v. Therefore, the exponential of the unbiased estimate of $\ln v$ results in a biased estimate of v. In the following, we discuss the unbiased and biased estimation cases under transformation.

Lemma 1. *If the transformed samples used for linear regression are strictly linearly correlated, the estimates of parameters are unbiased.*

Proof. Using the least-squares method for linear regression, $\hat{y}_i = ax_i + b$ $(i \in \{1, \ldots, n\}$, where n is the number of data samples, we minimize the sum of squares of the errors $RSS(SSE)$, which can be expressed as given:

$$RSS(SSE) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2.$$

Solving the above equation by means of derivatives, we can obtain the following:

$$b = \overline{y} - a\overline{x},$$

$$a = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

When solving the Underwood model using linear regression, let $y = \ln v$ and $x = k$, and we can obtain the following:

$$\begin{aligned} \hat{v}_i &= \exp(\hat{y}_i) \\ &= \exp\left(\hat{a}x_i + \hat{b}\right) \\ &= \exp(\hat{a}x_i + \overline{y} - \hat{a}\overline{x}) \\ &= \exp\left[\hat{a}\overline{x} + \hat{b} + \hat{a}(x_i - \overline{x})\right] \\ &= \exp\left(\hat{a}x_i + \hat{b}\right). \end{aligned}$$

If all points (x_i, y_i) are co-linear, then

$$\begin{aligned} \hat{v}_i &= \exp\left(\hat{a}x_i + \hat{b}\right) \\ &= \exp(\hat{y}_i) \\ &= \exp(y_i). \end{aligned}$$

Therefore, the estimates of the linear regression after transmission are unbiased; namely, we have the following:

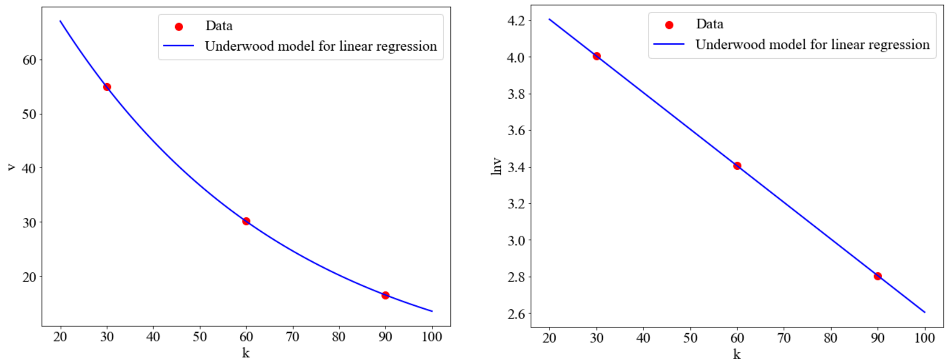
$$\mathrm{E}(\hat{v}) = \overline{v}$$

where $\hat{v}$ is the estimated v, and $\overline{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$. □

Taking the Underwood model as an example, suppose there are three given points of (k, v), which are (30, 54.881), (60, 30.119), and (90, 16.530), as shown in Figure 2a. Let $y = \ln v$ and $x = k$; the three points are transformed to (30, 4.005), (60, 3.405), and (90, 2.805). Obviously, these three points can be linked by a straight line, as shown in Figure 2b. Performing a linear regression on $(k, \ln v)$, we obtain the fitted linear expression $y = 4.6052 + (-0.02)x$. We use the MSE to express the fitting error, which is the cumulative value of the differences between actual observations and predicted values. The MSE can be computed as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (v_i - \hat{v}_i)^2 \tag{1}$$

where $\hat{v}_i$ is the value predicted by the model, v_i is the real value, and n is the number of observations in the dataset. Thus, the MSE of the fitted line to the transformed samples is zero.



(**a**) (**b**)

Figure 2. Unbiased case in the Underwood model. (**a**) The relationship between v and k. (**b**) The relationship between ln$v$ and k.

Transforming the parameters from the linear regression back into the original model, we obtain $v_f = \exp(4.6052)$ and $k_0 = -\frac{1}{-0.02}$. The original Underwood model should be $v = \exp(4.6052 + 0.02k)$, and the MSE of the fitted exponential curve to the original samples is also zero. Consequently, the density k and speed v of these samples obey the exponential relationship and strictly adhere to the Underwood model, as illustrated in Figure 2a.

However, when the data points used for linear regression do not lie on a straight line, the linear fitting is meaningless, and the estimates are biased. Therefore, we give the case where the linear transformation presents a bias against the Underwood and Northwestern models.

Lemma 2. *If the transformed samples used for linear regression are not strictly linearly correlated, the estimates of the parameters are generally biased.*

Proof. If we only have two points, they must be co-linear. We now discuss the case of three points. If the estimate is biased when the transformed three points are not co-linear, then the estimate must also be biased when more transformed points are not co-linear. Consider the three points $(x_1, y_1), (x_2, y_2),$ and(x_3, y_3) in the dataset that are not co-linear. If there are two points with equal y values, the x values are different. However, it is not possible for the y values of the three points to be equal since they would be co-linear. Hence, the relationship between the y values of these three points can be expressed as given: $y_1 < y_2 < y_3$ or $y_1 \leq y_2 < y_3$ or $y_1 < y_2 \leq y_3$. We define the following:

$$\begin{aligned} \overline{v} &= v_1 + v_2 + v_3 \\ &= \exp(y_1) + \exp(y_2) + \exp(y_3). \end{aligned}$$

Let $\hat{y}_i = y_i + \Delta_i (i = 1, 2, 3)$; then, we define the following:

$$E(\hat{v}) := \exp(y_1 + \Delta_1) + \exp(y_2 + \Delta_2) + \exp\left(y_3 + \Delta_3\right).$$

Therefore, we have the following:

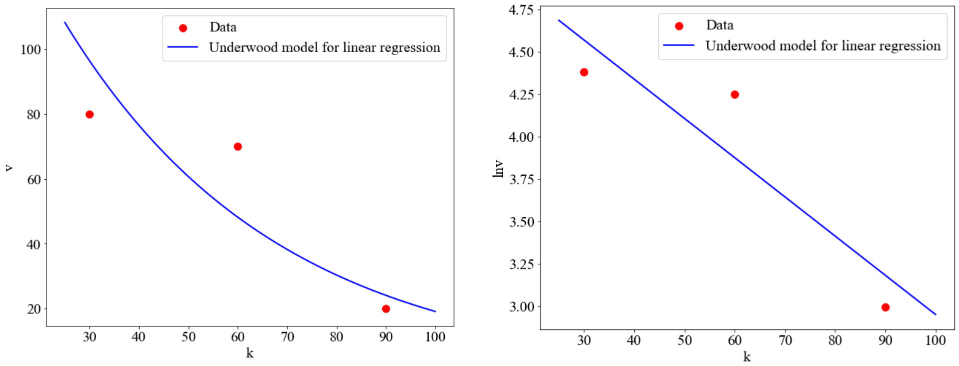
$$E(\hat{v}) - \overline{v} = \int_{y_1}^{y_1+\Delta_1} \exp(x)dx + \int_{y_2}^{y_2+\Delta_2} \exp(x)dx + \int_{y_3}^{y_3+\Delta_3} \exp(x)dx.$$

Meanwhile, in linear regression, the estimated y is unbiased; namely, we obtain $E(\hat{y}) = \overline{y}$, and $\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i + \Delta_i) = \frac{1}{n}\sum_{i=1}^{n} y_i$. Thus, $\Delta_1 + \Delta_2 + \Delta_3 = 0$. Therefore, we obtain the following:

$$(\hat{v}) - \overline{v} = \int_{y_1}^{y_1+\Delta_1} \exp(x)dx + \int_{y_2}^{y_2+\Delta_2} \exp(x)dx + \int_{y_3}^{y_3-\Delta_1-\Delta_2} \exp(x)dx.$$

In $\exp(x)$, all the different ranges of x values correspond to different function values. Therefore, to guarantee the estimates are unbiased, $E(\hat{v}) - \overline{v} = 0$ should be satisfied. To meet $E(\hat{v}) = \overline{v}$, we need $y_1 + \Delta_1 = y_2$ and $y_3 - \Delta_1 - \Delta_2 = y_1$; namely, $\hat{y}_1 = y_2$, and $\hat{y}_3 = y_1$. Obviously, this situation does not exist. Thus, in the transformed dataset, the solution using linear programming is biased as long as the three points are not co-linear. □

Taking the Underwood model as an example, suppose there are three given points of (k, v), which are (30, 80), (60,70), and (90,20), as shown in Figure 3a. Let $y = \ln v$ and $x = k$; the three points are transformed to (30, 4.382), (60, 4.249), and (90, 2.996). Clearly, these three points are not collinear, as shown in Figure 3b. Performing a linear regression on $(k, \ln v)$, we obtain the fitted linear expression $y = 5.2617 + (-0.023105)x$, whose MSE is 0.069593. However, when transforming the parameters from the linear regression back into the original model, $v_f = \exp(5.2617)$, and $k_0 = -\frac{1}{-0.023105}$, and the original Underwood model should be $v = \exp(5.2617 + 0.023105k)$, whose MSE is 253.6947. Because the transformed samples used for linear regression is not on the fitted line, it is meaningless to use linear regression to estimate the parameters of the model. Therefore, the fitted results obtained from the linear regression are not the true picture of the model, and the estimates of the model are biased.



(**a**) (**b**)

Figure 3. Biased case in the Underwood model. (**a**) The relationship between v and k. (**b**) The relationship between $\ln v$ and k.

Taking the example in the Underwood model, suppose the three given points of coincide with the above example, as in Figure 4a. Let $y = \ln v$ and $x = k^2$; the three points are transformed to (900, 4.382), (3600, 4.248), and (8100, 2.996). Obviously, these points are also not collinear, as in Figure 4b. Performing a linear regression on $\left(k^2, \ln v\right)$ gives results with an MSE of 0.03248981, and the fitted linear expression is $y = 4.7209 + (-0.0002013)x$. However, when substituting the parameters from the linear regression back into the original model, we obtain $v_f = \exp(4.7209)$ and $k_0 = \sqrt{-\frac{1}{2\times(-0.0002013)}}$, and the original Underwood model should be $v = \exp\left(4.7209 - 0.0002013k^2\right)$, whose MSE is 144.75979. Although the MSE value of the linear regression is good, this advantage cannot be reflected in the original model because the points used for linear regression are not collinear (as shown in Figure 4a). As a result, the linear regression approach is biased.

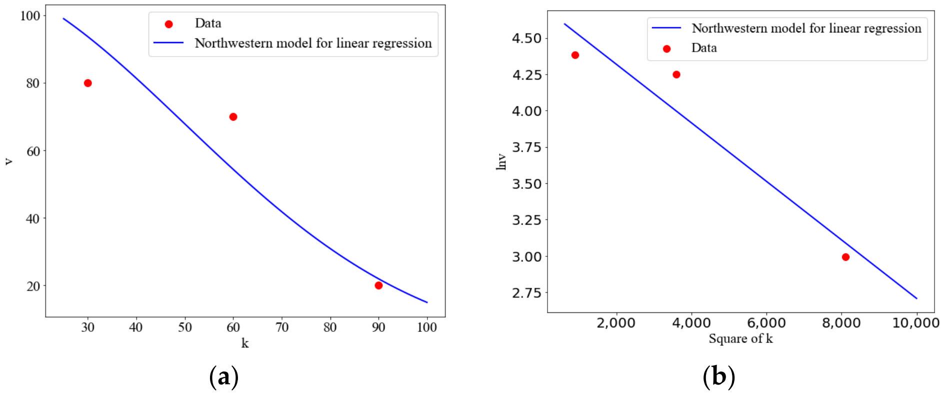


Figure 4. Biased case in the Northwestern model. (**a**) The relationship between v and k. (**b**) The relationship between ln$v$ and k^2.

Figures 5 and 6 depict the samples after the transformation of GA400 for the Underwood and Northwestern models. It is evident that these simple, straight lines in Figures 5 and 6 cannot fully capture the underlying structure of these points. Consequently, these two linear regression models provide biased estimates in this context.

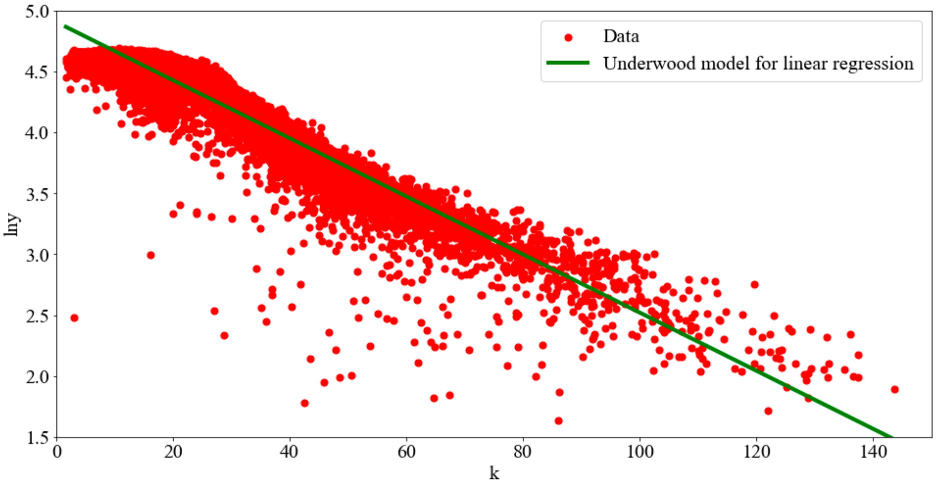


Figure 5. Sample points used for linear regression in the Underwood model.

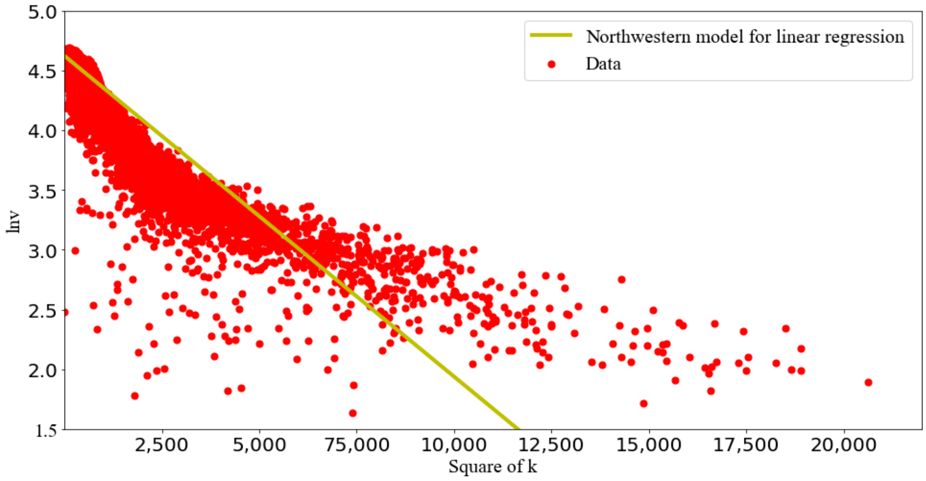


Figure 6. Sample points used for linear regression in the Northwestern model.

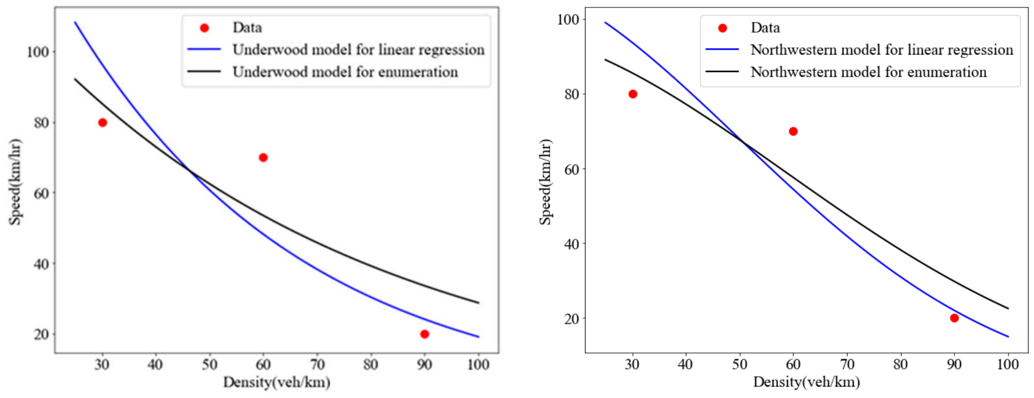
2.2. Correction

For the case where the linear regression provides biased estimates, we re-solve the model parameters using an enumeration algorithm. That is, we try to find the parameter values corresponding to the smallest MSE within the feasible ranges of the parameters, as shown in Algorithm 1. The estimated parameters obtained are unbiased for a given precision, and better estimates may exist as the precision becomes smaller. The enumeration algorithm is universal for estimated parameters that are difficult to solve by approximation or derivation methods.

Algorithm 1: An enumeration algorithm.

```
Input: A set of candidate pairs of parameters $\left\{\left(\left(v_{fi}, k_{0j}\right)\middle| i = 1,2,\ldots,M; j = 1,2,\ldots,N\right)\right\}$.
Output: The minimum MSE, the optimal values of parameters.
$MSE\left(v_{fi}, k_{0j}\right)$ denotes the MSE value of the pair of parameters $\left(v_{fi}, k_{0j}\right)$; the minimum MSE and its corresponding optimal parameters are denoted as $MSE', {v_f}', {k_0}'$.
Initialize the $MSE' = \infty, {v_f}' = 0, {k_0}' = 0$.
For $i = 1,2,\ldots,M$ do:
    For $j = 1,2,\ldots,N$ do:
        Calculate the MSE value $MSE\left(v_{fi}, k_{0j}\right)$ for the pair of parameters $\left(v_{fi}, k_{0j}\right)$.
        If $MSE\left(v_{fi}, k_{0j}\right) \leq MSE'$ do:
            ${v_f}' = v_{fi}$,
            ${k_0}' = k_{0j}$,
            $MSE' = MSE\left(v_{fi}, k_{0j}\right)$.
        End if
    End for
End for
```

The examples in Lemma 2 are solved with the enumeration algorithm, as shown in Figure 7. For the Underwood and Northwestern models, we enumerate the two parameters v_f and k_0 in functions $v = v_f\exp\left(1 - \frac{k}{k_0}\right)$ and $v = v_f\exp\left[-\frac{1}{2}\left(\frac{k}{k_0}\right)^2\right]$, both with a precision of 1 and a range of 0 to 200. The resulting optimal MSE values are 161.36348 and 93.4532, respectively. They are much better than the MSE values 253.6947 and 144.75979 obtained from the linear regression.



(**a**) (**b**)

Figure 7. Corrected results for examples. (**a**) The Underwood model. (**b**) The Northwestern model.

Above, we have corrected two simple examples using the enumeration algorithm, and next, we will examine how this algorithm performs on the entire GA400 dataset.

In the Underwood model, for parameters v_f and k_0, we set the iteration precision to 0.1 and the range to (0, 160) and (0, 120), respectively. The optimal values of parameters obtained are $v_f = 126.5790$ and $k_0 = 52.3435$, and the corresponding MSE is 50.36096, smaller than the MSE 59.4544 obtained from linear regression. Figure 8 illustrates the curves before and after the correction.

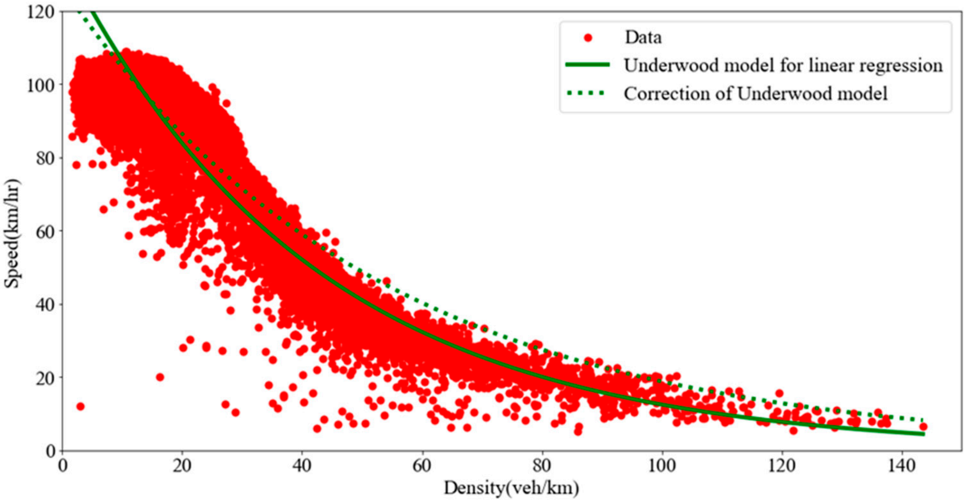


Figure 8. Correction of the Underwood model.

In the Northwestern model, for parameters v_f and k_0, we set the iteration precision to 0.1 and the range to (0, 160) and (0, 120), respectively. Then, we use the enumeration algorithm to find $v_f = 107.0668$ and $k_0 = 34.9348$. The MSE is 25.9371, much smaller than the MSE 44.3233 obtained from linear regression. The curves before and after correction are shown in Figure 9.

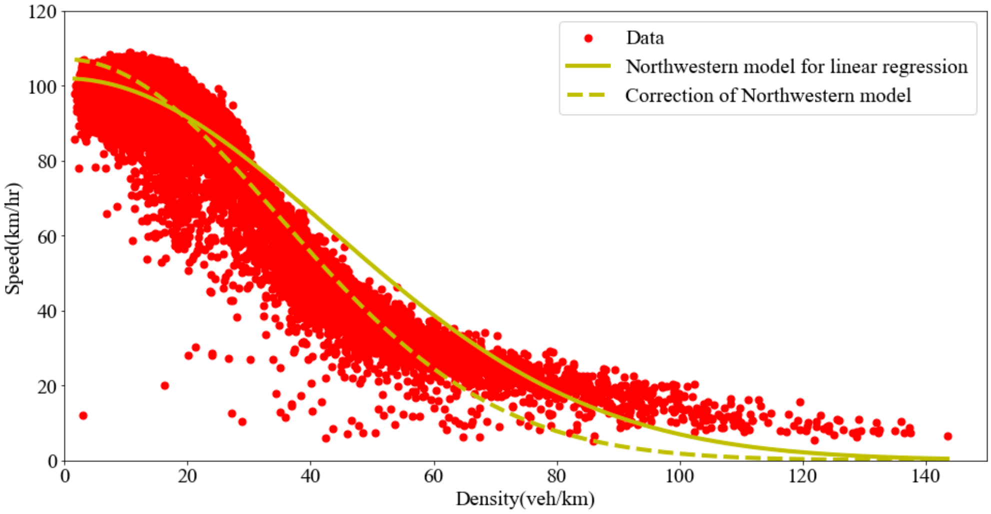


Figure 9. Correction of the Northwestern model.

From Figures 8 and 9, the corrected models appear to dominate only in the low-density range. This is because about 86% of the data points in the GA400 are concentrated in the [0, 20) range of the density. Figure 10a shows the average MSE value of the Underwood model for different density intervals, where the corrected results outperform the results solved by linear regression for densities in [0, 40) and [140, ∞), which account for 93% of all data points. Figure 10b shows the average MSE values of the Northwest model for different density intervals, and the corrected results are better than those solved by linear regression for densities [0, 60) and [140, ∞), which account for 98% of all data points. As a result, the features of a small portion of the data may be discarded in order to optimize the fit for the entire dataset.

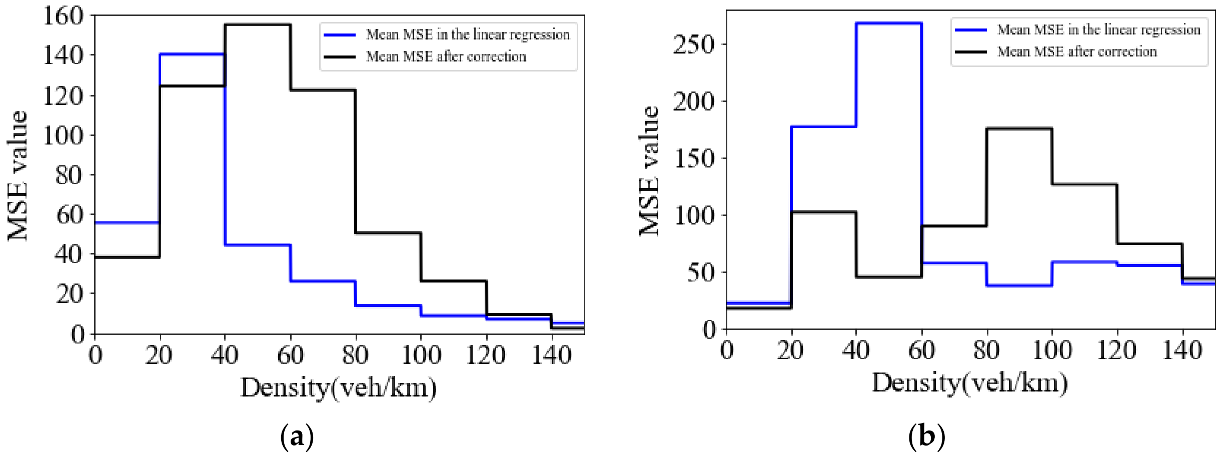


Figure 10. Average values of MSE for different density intervals. (**a**) The Underwood model. (**b**) The Northwestern model.

3. Lower Bound of the Fitting Error of Existing Models

3.1. MSE Values of Existing Models

Table 2 illustrates the MSE values of the four models based on the GA400 dataset. Since linear regression is biased, the Underwood and Northwestern models make the differences in MSE values before and after the correction. The results show that the Northwestern model performs the best. Nevertheless, the MSE value of the Northwestern model is still high, motivating us to explore the lower bound of the fitting error for existing models.

Table 2. MSE values of the four models for the GA400 dataset.

Models	Function	Transformation	Original MSE	Corrected MSE
Greenshields (Greenshields et al., 1935) [1]	$v = v_f\left(1 - \frac{k}{k_j}\right)$	$v = y, k = x$	46.727	46.727
Greenberg (1959) [3]	$v = v_0 \ln\left(\frac{k_j}{k}\right)$	$v = y, \ln k = x$	107.948	107.948
Underwood (1961) [5]	$v = v_f \exp\left(1 - \frac{k}{k_0}\right)$	$\ln v = y, k = x$	59.4544	50.3609
Northwestern (Drake et al., 1967) [20]	$v = v_f \exp\left[-\frac{1}{2}\left(\frac{k}{k_0}\right)^2\right]$	$\ln v = y, k^2 = x$	44.3233	25.9371

We use an example to illustrate how to compute the “ideal” lower bound of the fitting error. Given a dataset containing the three points (30,80), (60,78), and (90,40), Table 3 presents the MSEs for each of the four models, with the corresponding fitted curves displayed in Figure 11. Due to the models’ structure, none of them could be adjusted to achieve an MSE of zero, as evidenced by their inability to pass through all three points simultaneously. Consistent with the monotonicity and decreasing characters of the traffic flow, the speeds corresponding to each density value to achieve the minimum MSE are found and simply connected to form a piecewise linear function. This value is the lower bound of the model’s fitting error. Therefore, assessing the differences between the existing models’ MSEs and the lower bound exposes potential areas for improvement.

Table 3. MSE values of the four models based on the three data points.

Models	Corrected MSE
Greenshields (Greenshields et al., 1935) [1]	72.0000
Greenberg (1959) [3]	117.3113
Underwood (1961) [5]	95.7534
Northwestern (Drake et al., 1967) [20]	57.0006

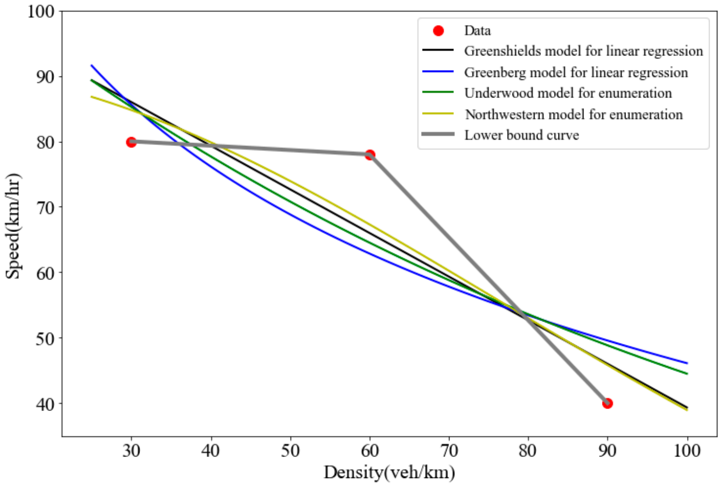


Figure 11. Non-optimal solution cases for the four models.

3.2. Quadratic Programming Model

Considering the monotonically decreasing and continuous characteristics of traffic flow, the prediction model, denoted by $f(k)$, with the minimum MSE should be selected among all possible monotonically decreasing continuous functions. This means that for two given densities, i.e., $k_1 < k_2$, we should have $f(k_1) \geq f(k_2)$ and that for each density, there is only one speed output. We use the following two cases to illustrate this model.

Case 1: As shown in Figure 12a, actual speed may increase with increasing density, contrary to the general relationship where speed decreases as density increases. However, to capture the overall characteristic of traffic flow, any fitted model should exhibit both continuity and a monotonically decreasing trend. This allows the model to accommodate the unique cases while reflecting the general behavior of traffic flow.

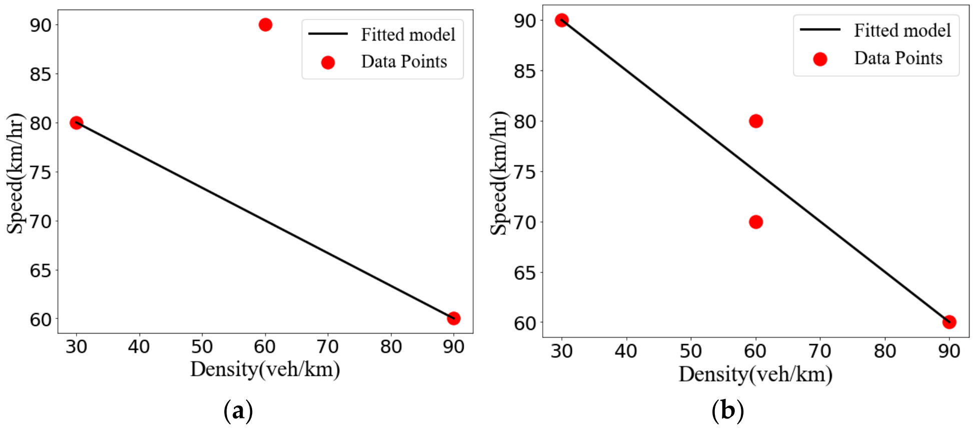


Figure 12. The case where the limit of MSE is not zero. (**a**) Case 1. (**b**) Case 2.

Case 2: As shown in Figure 12b, different speeds can exist at the same density. However, the estimated speed in the model can only be a single value, which should ideally be the average of these speeds.

Considering these factors, we developed a quadratic programming model that defines the lower bound of the fitting error. The optimal objective function value of this model corresponds to the lower bound, providing a quantifiable measure of the fitting error. The model is shown as follows:

$$min\frac{1}{m}\sum_{i=1}^{m}\left(v_{ij}-\hat{v}_i\right)^2 \tag{2}$$

Subject to

$$\hat{v}_i - \hat{v}_{i+1} \geq 0,\ \forall i = 1, \ldots, m. \tag{3}$$

Here, $\hat{v}_i = f(k_i)$ denotes the decision variable, representing the estimated speed at the i-th density $f(k_i)$, and m is the number of all different densities. Considering that a same speed may correspond to multiple densities, we denote v_{ij} as the j-th real speed value at the i-th density. Equation (2) is the objective function of the model that minimizes the MSE value. Constraint (3) requires that the estimated speeds should satisfy the characteristics of monotonically decreasing continuity in traffic flow.

3.3. Results

The above model capturing the lower bound of the fitting error can be viewed as a piece-wise linear function that links the optimal speed at each density. We utilize GUROBI to solve the model on the GA400 dataset, which achieves a minimum MSE of 19.360. This fitting error is significantly lower than the results obtained by the four models, as demonstrated in Figure 13. In the GA400 dataset, more than 80% of the data points are concentrated within the 0–20 density range. As a result, models tend to primarily focus on these points. However, our model optimizes the lower bound across all density intervals, making it applicable in all cases of density distribution. Furthermore, in different models, the free-flow speed depends on the form of the model. However, the lower bound is derived from the dataset following the monotonicity and continuity characteristics of the traffic flow. Therefore, the free-flow speed of the lower bound depends on the speed when the density of the dataset is extremely small. This result ignores factors such as length and width of the road, and vehicle type and is ideal for observing the situation on the road.

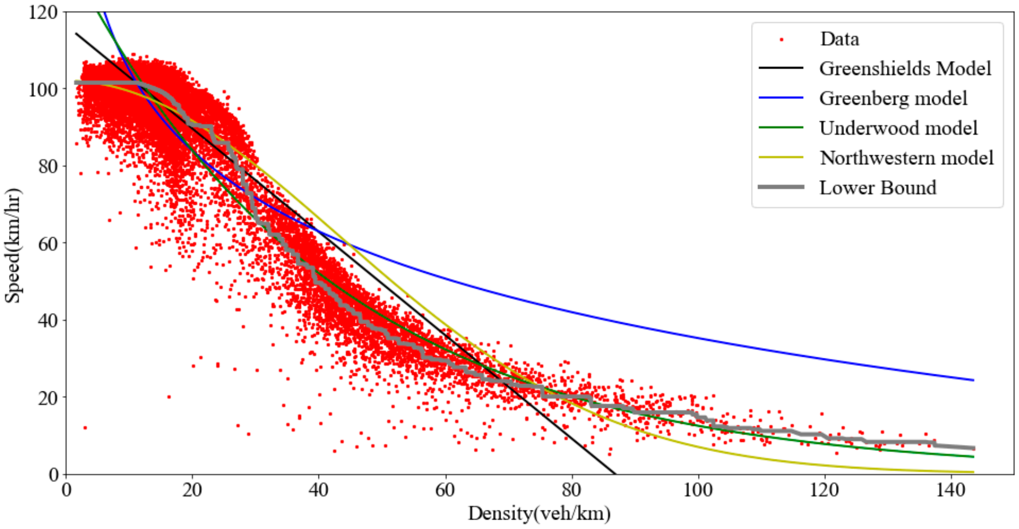


Figure 13. Lower bound of models.

The fitting results vary across models, but the lower bound is unique for the same dataset. In order to measure the effectiveness of each model and the room for improvement in a more standardized way, we define the "relative gap", $\frac{|MSE_s - MSE_L|}{MSE_L} \times 100\%$, which represents the gap between one existing model and the "ideal" lower bound. MSE_L is the MSE value of the "ideal" lower bound, and MSE_s is the MSE of any other model (i.e., Greenshields, Greenberg, Underwood, and Northwestern models). The relative gaps of the four models are shown in Table 4, where the Northwestern model performs the best but still has a 33.973% relative gap. Therefore, there is significant room for improvement for existing models to achieve a better fit of the dataset and reduce the MSE closer to the "ideal" lower bound.

Table 4. MSE and relative gap of four models based on the GA400 dataset.

Models	MSE	Relative Gap
Greenshields [1]	46.7270	137.603%
Greenberg [3]	107.9480	457.583%
Underwood [5]	50.3609	160.129%
Northwestern [20]	25.9371	33.973%
Average value	57.8053	197.322%

To further validate the correction method and the method of exploring the lower limit of fitting error, we sample datasets of different sizes from the GA400 dataset, shown in Tables 5 and 6. It can be noticed that, for different sizes, the MSE values obtained by the correction method are smaller and much closer to the lower bounds. At the same time, the lower bound always represents the limit of fitting error.

Table 5. Results of different sample sizes of the Underwood model.

	Sample Size	MSE Values for Linear Regression	MSE Values after Correction	MSE Values for Lower Bound	Relative Gap for Linear Regression	Relative Gap for Corrected Results
1	100	79.266	67.860	24.084	229.126%	181.766%
2	100	44.656	43.402	10.827	312.444%	300.863%
3	500	62.533	50.326	14.262	338.467%	252.871%
4	500	68.246	58.360	18.631	266.316%	213.249%
5	1000	57.880	48.574	16.318	254.691%	197.667%
6	1000	53.204	44.782	12.246	334.443%	265.675%
7	5000	61.323	51.584	19.455	215.196%	165.137%
8	5000	58.771	50.546	18.529	217.175%	172.787%
9	10,000	59.292	50.461	19.010	211.899%	165.446%
10	10,000	60.492	51.296	19.657	207.732%	160.950%
11	30,000	59.321	49.987	18.852	214.672%	165.157%
12	30,000	59.220	50.062	19.505	203.620%	156.668%

Note: Relative gap for linear regression = $\frac{|\text{MSE value for linear regression} - \text{MSE value for lower bound}|}{\text{MSE value for lower bound}}$; relative gap after correction = $\frac{|\text{MSE value after correction} - \text{MSE value for lower bound}|}{\text{MSE value for lower bound}}$.

Table 6. Results of different sample sizes of the Northwestern model.

	Sample Size	MSE Values for Linear Regression	MSE Values after Correction	MSE Values for Lower Bound	Relative Gap for Linear Regression	Relative Gap for Corrected Results
1	100	26.520	26.288	15.640	69.562%	68.082%
2	100	99.182	65.021	24.821	299.583%	161.955%
3	500	29.822	24.064	16.680	78.787%	44.271%
4	500	43.881	29.784	17.6181	149.064%	69.054%
5	1000	38.354	23.915	15.793	142.859%	51.433%
6	1000	49.473	23.244	14.825	233.723%	56.790%
7	5000	52.530	28.479	20.810	152.427%	36.852%
8	5000	49.058	27.132	19.101	156.829%	42.045%
9	10,000	40.869	24.552	17.541	132.990%	39.969%
10	10,000	46.855	25.510	18.658	151.124%	36.722%
11	30,000	43.821	26.410	19.684	122.624%	34.172%
12	30,000	43.286	26.440	19.655	120.226%	34.521%

Note: Relative gap for linear regression = $\frac{|\text{MSE value for linear regression} - \text{MSE value for lowerbound}|}{\text{MSE value for lower bound}}$; relative gap after correction = $\frac{|\text{MSE value after correction} - \text{MSE value for lower bound}|}{\text{MSE value for lower bound}}$.

4. Conclusions

In this study, we conducted a comprehensive analysis of the errors associated with the generalized linear regression models on the fundamental diagram, focusing on the bias introduced when linear regression is improperly applied for parameter estimation in the Underwood and Northwestern models. To address this issue, we employed an enumeration algorithm to resolve these models, resulting in significant decreases in MSE values and improving the model fits. Moreover, we developed a quadratic programming model that takes advantage of the inherent properties of monotonicity and continuity in traffic flow. This enabled us to determine the lower bound of the fitting error for existing models. Our presented model demonstrates robust performance across various density intervals, achieving a minimum MSE of 19.360. This indicates a relative gap of 33.973% between the lower bound and the best result obtained by other models. The substantial gap highlights the potential for further refinements and advancements in model performance.

The proposed correction method in this study is universally applicable, particularly for models where parameter estimation through derivation or approximation is not feasible. Additionally, the quadratic programming model can serve as a measure of model quality for any traffic flow dataset. Furthermore, it is important to consider the influence of heterogeneous traffic flow data on the fitting process. Therefore, future studies should investigate the effects of multiple factors on the fitting process, enhancing the comprehensiveness and credibility of the research.

Author Contributions: Conceptualization, Y.S. and S.W.; methodology, Y.S., X.T., S.J., K.G., X.H., W.Y. and Y.G.; software, Y.S.; validation, Y.S., X.T. and S.W.; formal analysis, S.J.; investigation, K.G.; resources, X.H.; data curation, W.Y.; writing—original draft preparation, Y.G.; writing—review and editing, Y.S.; visualization, Y.S.; supervision, X.T.; project administration, S.W.; funding acquisition, S.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (Grant No. 72361137006) and JPI Urban Europe and Energimyndigheten (P2023-00029, e-MATS).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Greenshields, B.D.; Bibbins, J.R.; Channing, W.S.; Miller, H.H. A study of traffic capacity. *Highw. Res. Board Proc.* **1935**, *14*, 448–477.
2. Haight, F.A. *Mathematical Theories of Traffic Flow*; Academic Press: London, UK, 1963.
3. Greenberg, H. An analysis of traffic flow. *Oper. Res.* **1959**, *7*, 255–275. [CrossRef]
4. Edie, L.C. Car-following and steady state theory for non-congested traffic. *Oper. Res.* **1961**, *9*, 66–76. [CrossRef]
5. Underwood, R.T. *Speed, Volume, and Density Relationship: Quality and Theory of Traffic Flow*; Yale Bureau of Highway Traffic: New Haven, CT, USA, 1961; pp. 141–188.
6. Newell, G.F. Nonlinear effects in the dynamics of car following. *Oper. Res.* **1961**, *9*, 209–229. [CrossRef]
7. Kerner, B.S.; Konhäuser, P. Structure and parameters of clusters in traffic flow. *Phys. Rev.* **1994**, *50*, 54–83. [CrossRef]
8. Del Castillo, J.M.; Benítez, F.G. On the functional form of the speed-density relationship—I: General theory. *Transp. Res. Part B Methodol.* **1995**, *29*, 373–389. [CrossRef]
9. Del Castillo, J.M.; Benítez, F.G. On the functional form of the speed-density relationship—II: Empirical investigation. *Transp. Res. Part B Methodol.* **1995**, *29*, 391–406. [CrossRef]
10. Li, J.; Zhang, H.M. Fundamental diagram of traffic flow: New identification scheme and further evidence from empirical data. *Transp. Res. Rec.* **2001**, *2011*, 50–59. [CrossRef]
11. Wu, N. A new approach for modelling of fundamental diagrams. *Transp. Res. Part A Policy Pract.* **2002**, *36*, 867–884. [CrossRef]
12. MacNicholas, M.J. A simple and pragmatic representation of traffic flow. In *Symposium on The Fundamental Diagram: 75 Years*; Transportation Introduction Research Board: Woods Hole, MA, USA, 2008.
13. Wang, H.; Li, H.; Chen, Q.; Ni, D. Logistic modeling of the equilibrium speed–density relationship. *Transp. Res. Part A Policy Pract.* **2011**, *45*, 554–566. [CrossRef]
14. Wu, X.; Liu, H.X.; Geroliminis, N. An empirical analysis on the arterial fundamental diagram. *Transp. Res. Part B Methodol.* **2011**, *45*, 255–266. [CrossRef]
15. Dervisoglu, G. *Automatic Calibration of Freeway Models with Model-Based Sensor Fault Detection*; University of California: Berkeley, CA, USA, 2012.
16. Keyvan-Ekbatani, M.; Kouvelas, A.; Papamichail, I.; Papageorgiou, M. Exploiting the fundamental diagram of urban networks for feedback-based gating. *Transp. Res. Part B Methodol.* **2012**, *46*, 1393–1403. [CrossRef]
17. Keyvan-Ekbatani, M.; Papageorgiou, M.; Papamichail, I. Urban congestion gating control based on reduced operational network fundamental diagrams. *Transp. Res. Part C Emerg. Technol.* **2013**, *33*, 74–87. [CrossRef]
18. Keyvan-Ekbatani, M.; Papageorgiou, M.; Knoop, V.L. Controller design for gating traffic control in presence of time-delay in urban road networks. *Transp. Res. Part C Emerg. Technol.* **2015**, *59*, 308–322. [CrossRef]
19. Qu, X.; Wang, S.; Zhang, J. On the fundamental diagram for freeway traffic: A novel calibration approach for single-regime models. *Transp. Res. Part B Methodol.* **2015**, *73*, 91–102. [CrossRef]

20. Drake, J.S.; Schofer, J.L.; May, A.D. A statistical analysis of speed–density hypotheses. *Highway Res. Rec.* **1967**, *154*, 112–117.
21. Fan, S.; Seibold, B. Data-fitted first-order traffic models and their second-order generalizations. *Transport. Res. Rec.* **2013**, *2391*, 32–43. [CrossRef]
22. Qu, X.; Zhang, J.; Wang, S. On the stochastic fundamental diagram for freeway traffic: Model development, analytical properties, validation, and extensive applications. *Transp. Res. Part B Methodol.* **2017**, *104*, 256–271. [CrossRef]
23. Wang, S.; Chen, X.; Qu, X. Model on empirically calibrating stochastic traffic flow fundamental diagram. *Commun. Transp. Res.* **2021**, *1*, 100015. [CrossRef]
24. Bramich, D.M.; Menéndez, M.; Ambühl, L. Fitting empirical fundamental diagrams of road traffic: A comprehensive review and comparison of models using an extensive data set. *IET Intell. Transp. Syst.* **2022**, *23*, 14104–14127. [CrossRef]
25. Jabeena, M. Comparative study of traffic flow models and data retrieval methods from video graphs. *Int. J. Eng. Res. Appl.* **2013**, *3*, 1087–1093.
26. Li, Y.; Lu, H.; Bian, C.; Sui, Y.G. Traffic speed-flow model for the mix traffic flow on Beijing urban expressway. In Proceedings of the 2009 International Conference on Measuring Technology and Mechatronics Automation, Zhangjiajie, China, 11–12 April 2009; Volume 3, pp. 641–644.
27. Banos, A.; Corson, N.; Lang, C.; Marilleau, N.; Taillandier, P. Multiscale modeling: Application to traffic flow. *Agent-Based Spat. Simul. NetLogo* **2017**, *2*, 37–62.
28. Anuar, K.; Habtemichael, F.; Cetin, M. Estimating traffic flow rate on freeways from probe vehicle data and fundamental diagram. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain, 15–18 September 2015; pp. 2921–2926.
29. Wang, H.; Ni, D.; Chen, Q.Y.; Li, J. Stochastic modelling of the equilibrium speed-density relationship. *J. Adv. Transp.* **2013**, *47*, 126–150. [CrossRef]
30. Zhang, J.; Qu, X.; Wang, S. Reproducible generation of experimental data sample for calibrating traffic flow fundamental diagram. *Transport. Res. Part A Policy Pract.* **2018**, *111*, 41–52. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.