Unraveling stochastic fundamental diagrams with empirical knowledge: Modeling, limitations, and future directions

Yuan-Zheng Lei, Yaobang Gong, Xianfeng Terry Yang *

Department of Civil and Environmental Engineering, University of Maryland, 1173 Glenn L.Martin Hall, College Park, MD 20742, United States

ARTICLE INFO

ABSTRACT

Traffic flow modeling relies heavily on fundamental diagrams. However, deterministic fundamental diagrams, such as single or multi-regime models, cannot capture the underlying uncertainty in traffic flow. To address this limitation, this study proposes a non-parametric Gaussian process model to formulate the stochastic fundamental diagram. Unlike parametric models, the non-parametric approach is insensitive to parameters, flexible, and widely applicable. The computational complexity and high memory requirements of Gaussian process regression are also mitigated by introducing sparse Gaussian process regression. This study also examines the impact of incorporating empirical knowledge into the prior of the stochastic fundamental diagram model and assesses whether such knowledge can enhance the model's robustness and accuracy. By using several well-known single-regime fundamental diagram models as priors and testing the model's performance with different sampling methods on real-world data, this study finds that empirical knowledge benefits the model only when small inducing samples are used with a relatively clean and large dataset. In other cases, a purely data-driven approach is sufficient to estimate and describe the density–speed relationship pattern.

1. Introduction

The traffic fundamental diagram is a crucial concept that provides a quantitative framework for understanding and estimating traffic behavior in different scenarios. Its application is widespread, including traffic flow theory, highway capacity analysis, and the design of intelligent transportation systems. Therefore, it plays a vital role in the planning and operating of efficient transportation networks.

The fundamental diagram is used to describe the following relationship:

$$q = \rho \times v \tag{1}$$

where q, ρ, and v represent flow, density, and space mean speed. Based on Ni (2015), it can also be regarded as an equilibrium traffic flow model since the fundamental diagram represents a steady-state or equilibrium relationship between flow, density, and speed. In this paper, the density–speed relationship is the primary focus. Given this relationship, the other two relationships can be determined using Eq. (1). Traditional density–speed fundamental diagram models can be classified into single-regime and multi-regime models, as shown in Fig. 1. Single-regime models usually describe the relationship between traffic density and speed using a linear or nonlinear function. However, single-regime models may not always fit well in all density ranges. This is why multi-regime models use piecewise functions to represent the density–speed relationship across the entire range of densities.

* Corresponding author.
  E-mail address: xtyang@umd.edu (X.T. Yang).
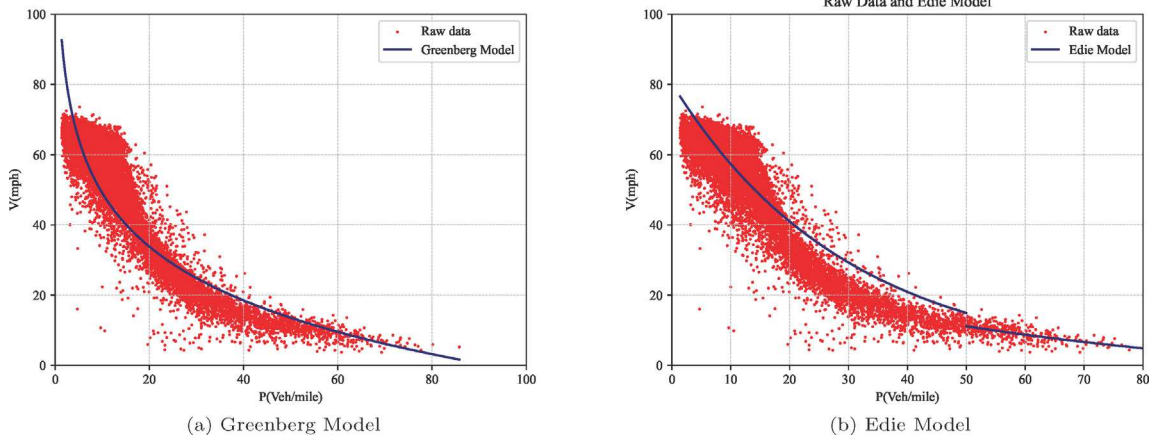
(a) Greenberg Model

(b) Edie Model

Fig. 1. Single-regime and multi-regime models.[1]

Almost all traditional single and multi-regime fundamental density–speed diagram models are shown in a deterministic form, which can be briefly represented by:

$$v = f(\rho) \tag{2}$$

In real-life traffic situations, it is difficult to find a steady-state condition. Therefore, when a traffic density ρ is given, the actual observed speed is more likely to fall within a specific range instead of a fixed value. Adopting the idea presented in Ni (2015), the stochastic fundamental diagram model operates on the principle that the density–speed relationship is not a function of density but a function of density and its distribution. This relationship is expressed mathematically in Eq. (3), where $w(\rho)$ indicates the density distribution.

$$v = f(\rho, w(\rho)) \tag{3}$$

Fig. 2 shows the stochastic fundamental diagram, where a certain speed distribution corresponds to a given density ρ. In numerous studies on network-level control, such as Gupta et al. (2023) and Li et al. (2024), the fundamental diagram is utilized to depict the traffic patterns of a route, section, or zone. Many of these studies focus on maximizing flow through signal control or route guidance strategies. In a deterministic fundamental diagram model, the target density after control is a constant value since the deterministic fundamental diagram models a curve. In the stochastic fundamental diagram model, the maximum mean flow will associate with a distribution indicating all possibilities, and it may have a wide confidence interval, indicating that it could lead to a very high maximum flow and a very low minimum flow. Therefore, the stochastic fundamental diagram model provides insight into the robust network-level control, suggesting that it is necessary to simultaneously consideration of both the mean and variance of the flow, rather than focusing solely on the mean, and it is better to steer the network toward a traffic density with a relatively high mean flow but a narrow confidence interval.

Stochastic and deterministic fundamental diagram models are essentially regression models. Assuming some data has been collected, which can be expressed as $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, a regression model is a model that used to describe the relationship between $\mathbf{y}$ and $\mathbf{x}$ that can be written as:

$$\mathbf{y} = f(\mathbf{x}, \theta) + \epsilon \tag{4}$$

where θ is a vector of k parameters, ϵ denotes an error term whose distribution may or may not be Gaussian, and $f(\cdot)$ is a function that delineates the relationship between $\mathbf{y}$ and $\mathbf{x}$.

Most existing stochastic or deterministic fundamental diagram models are parametric regression models. For a parametric regression model, the θ in Eq. (4) is a vector in finite-dimensional space, and $f(\mathbf{x}, \theta)$ need to be known functions and distribution of ϵ is known prior for the inference (Mahmoud, 2019). However, for a non-parametric regression model, the θ in Eq. (4) is a subset of an infinite-dimensional vector of vector spaces, and $f(\mathbf{x}, \theta)$ is an unknown function.

The flexibility of parametric models is limited by their fixed functional form and assumptions on error distributions, making it challenging to capture complex relationships. Non-parametric models, however, are better suited for modeling complex, nonlinear relationships. Gaussian process regression (GPR) is a powerful non-parametric model for the stochastic density–speed relationship (Yuan et al., 2021a,b). Nonetheless, handling extensive datasets presents a challenge for Gaussian Process Regression (GPR), as it demands significant computational resources. Specifically, for a dataset comprising n samples, GPR's computational complexity scales cubically, with a time complexity of $\mathcal{O}(n^3)$. Meanwhile, its space complexity, which indicates the amount of

---

[1] Both models are calibrated based on a dataset collected on the Georgia State Route 400 and weighted least square method proposed in Qu et al. (2015).
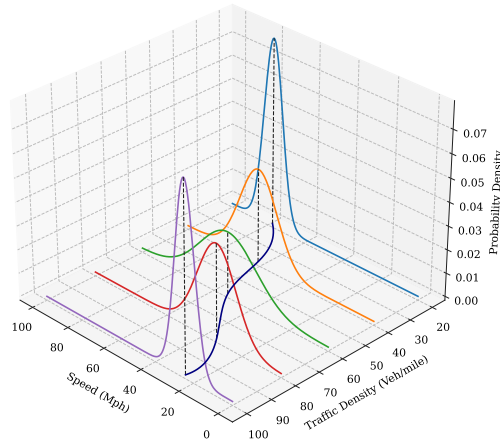
Fig. 2. Representation of the stochastic fundamental diagram.

memory needed, grows quadratically, marked by $\mathcal{O}(n^2)$. This makes GPR less feasible for very large datasets without employing variational techniques or approximation methods. To overcome this issue, a sparse Gaussian process regression (SGPR) is used to approximate the full GPR using a subset of data points, for example, $m(m \ll n)$, and the time complexity will fall into $\mathcal{O}(mn^2)$. Although the accuracy of SGPR is generally considered lower than GPR, it can still be a viable solution for deterministic fundamental diagram models. In the majority of cases, deterministic models are parametric in nature. Consequently, employing the complete dataset for their calibration does not pose significant challenges because its time complexity is generally considered as $\mathcal{O}(c^2 n)$ (Li, 1996), where c is the total number of features, $c \ll m$ in practical. This observation leads to an insightful proposition: calibrating deterministic models initially and subsequently utilizing them as priors over the unknown function may offer a viable strategy. Such an approach potentially facilitates the incorporation of empirical knowledge derived from the entire dataset without incurring additional time and space complexities associated with Sparse Gaussian Process Regression (SGPR).

The objective of this paper is twofold. Firstly, inspired by Wang et al. (2013), Qu et al. (2017), and Cheng et al. (2021), it aims to develop a non-parametric model for the density–speed relationship. This model will reflect the equilibrium density–speed relationship based on the mean function curve (the navy curve on the z = 0 plane in Fig. 2) and traffic patterns based on different speed distributions. Secondly, the paper will explore the possibility of bridging the gap between previous deterministic and stochastic models. Specifically, it will investigate the feasibility of incorporating empirical models as a prior or direct empirical model term into the loss function of the stochastic model. The paper will answer the following questions: ***Is it possible to merge empirical knowledge from deterministic models into the stochastic model? If so, how can it be done? And, is this empirical knowledge really helpful for the stochastic model?***

In this paper, our contribution can be concluded as follows:

- A sparse non-parametric regression framework is proposed to model the stochastic density–speed relationship. The mean function of the generated Gaussian process can be regarded as the density–speed relationship under the equilibrium state, and the different Gaussian distributions distributed at every single density point represent the overall traffic pattern.
- This paper describes how to model empirical knowledge into a stochastic model but also highlights the limitations caused by deterministic model parameters, which require clean data for calibration.
- In the first time, to the best of our knowledge, this paper shows that even taking an extremely well-calibrated empirical model as the prior, when the size of inducing samples is over a certain range, the advantages embedded in it may not exist.

The structure can be organized as follows: Section 2 will review deterministic fundamental diagram models and stochastic fundamental diagram models; Section 3 will introduce how deterministic models are calibrated and demonstrate the non-parametric technique that will be used for sparse regression and briefly discuss the sampling algorithms that will be used to select inducing variables for making the approximation; Section 4 will discuss the potential influence of empirical models.

2. Literature review

2.1. Deterministic fundamental diagram models

Most well-known single and multi-regime fundamental diagram models are deterministic models. The density–speed relationship is represented by a linear or nonlinear function, which can be considered to reflect the density–speed relationship under equilibrium. Given a certain density ρ, the corresponding speed is fixed. Based on the modeling methods of the density–speed relationship corresponding to different traffic conditions, the deterministic fundamental diagram models can be classified into single-regime models and multi-regime models.

Table 1
Some single-regime traffic flow model.

Model	Relationship equation	Parameters
Greenshields et al. (1935)	$v = v_f \left(1 - \dfrac{\rho}{\rho_j} \right)$	v_f, ρ_j
Greenberg (1959)	$v = v_{\text{critical}} \log \left(\dfrac{\rho_j}{\rho} \right)$	$v_{critical}, \rho_j$
Underwood (1961)	$v = v_f \exp \left(-\dfrac{\rho}{\rho_{\text{critical}}} \right)$	$v_f, \rho_{critical}$
Newell (1961)	$v = v_f \left\{ 1 - \exp \left[-\dfrac{\lambda}{v_f}(\dfrac{1}{\rho} - \dfrac{1}{\rho_j}) \right] \right\}$	v_f, ρ_j, λ
Drake et al. (1965)	$v = v_f \exp \left(-(\dfrac{\rho}{\rho_{critical}})^2 \right)$	$v_f, \rho_{critical}$
Pipes (1966)	$v = v_f \left(1 - \dfrac{\rho}{\rho_j} \right)^n$	v_f, ρ_j, n
Drew (1968)	$v = v_f \left[1 - \left(\dfrac{\rho}{\rho_j} \right)^{m_1} \right]^{m_2}$	v_f, ρ_j, m_1, m_2
Papageorgiou et al. (1989)	$v = v_f \exp \left[-\dfrac{1}{\alpha} \left(\dfrac{\rho}{\rho_j} \right)^{\alpha} \right]$	v_f, ρ_j, α
Kerner and Konhäuser (1994)	$v = v_f \left[\dfrac{1}{1 + \exp\left(\dfrac{\frac{\rho}{\rho_{critical}} - 0.25}{0.06} \right) - 372 \times 10^{-8}} \right]$	$v_f, \rho_{critical}$
Del Castillo and Benitez (1995)	$v = v_f \left\{ 1 - \exp[\dfrac{v_j}{v_f}(1 - \dfrac{\rho_j}{\rho})] \right\}$	v_f, ρ_j, v_j
Jayakrishnan et al. (1995)	$v = v_{min} + (v_f - v_{min})(1 - \dfrac{\rho}{\rho_j})$	v_f, v_{min}, ρ_j
Ardekani and Ghandehari (2008)	$v = v_{critical} \log(\dfrac{\rho_j + \rho_{min}}{\rho + \rho_{min}})$	$v_{critical}, \rho_j, \rho_{min}$
MacNicholas (2011)	$v = v_f \left(\dfrac{\rho_j^n - \rho^n}{\rho_j^n + m\rho^n} \right)$	v_f, ρ_j, n, m
Wang et al. (2011)	$v = v_{critical} + \dfrac{v_f - v_{critical}}{\left\{ 1 + \exp[\frac{\rho - \rho_{critical}}{\theta_1}] \right\}^{\theta_2}}$	$v_f, v_{critical}, \rho_{critical}, \theta_1, \theta_2$
Cheng et al. (2021)	$v = \dfrac{v_f}{[1 + (\frac{\rho}{\rho_{critical}})^m]^{\frac{2}{m}}}, 1 \leq m \leq 8.53$	$v_f, \rho_{critical}, m$

2.1.1. Single-regime models

In 1935, Greenshields et al. (1935) proposed the first fundamental traffic flow diagram. This diagram established a linear relationship between speed and density. Since then, a lot of research has been done to enhance both the accuracy and interpretability of this relationship, such as Greenberg's model (Greenberg, 1959), Underwood's model (Underwood, 1961), Newell's model (Newell, 1961), Drake's model (Drake et al., 1965) and Drew's model (Drew, 1968).

There are different types of fundamental diagrams that are classified based on the functional forms of the speed–density relationship. These diagrams can be categorized into three main groups — exponential function-related models, logarithmic function-related models, and polynomial function-related models along with their variants. Exponential function-related models include the models proposed by Newell (1961), Drake et al. (1965), Kerner and Konhäuser (1994), Del Castillo and Benitez (1995), and Wang et al. (2011). On the other hand, Ni et al. (2016) proposed logarithmic function-related models. Polynomial function-related models and their variants are proposed by Pipes (1966), Drew (1968), Jayakrishnan et al. (1995), MacNicholas (2011), and Cheng et al. (2021).

According to the Ni (2015), it has been observed that none of the early single-regime models are capable of fitting the empirical data reasonably well over the entire density range. Each model is only suitable for a specific density range. Due to this limitation, researchers started exploring the idea of fitting the data in a piecewise manner using multiple equations, which resulted in the development of multi-regime models.

2.1.2. Multi-regime models

Multi-regime models incorporate a series of piecewise functions. Common multi-regime models (Edie, 1961; Drake et al., 1965; May, 1990; Sun and Zhou, 2005) include two-regime or three-regime models. Recent studies have adopted non-parametric regression frameworks. For instance, Sun et al. (2014) adopted a B-spline non-parametric regression structure. An incomplete list of multi-regime models is shown in Table 2. The basic idea of multi-regime models is to divide density into two or three groups and then use different functions to fit corresponding empirical data located in different density ranges. Drawing on the findings of Ni (2015), it is evident that the multi-regime model exhibited limited efficacy in accurately fitting the empirical data. Moreover, these models often present issues of discontinuity and lack of differentiability.

Table 2
Some multi-regime traffic flow model[a].

Model	Free flow	Transitional	Congested
Edie model	$v = 54.9e^{-\frac{\rho}{163.9}}, \rho \le 50$	–	$v = 26.8 \ln \frac{162.5}{\rho}, \rho > 50$
Two-regime model	$v = 60.9 - 0.515\rho, \rho \le 65$	–	$v = 40 - 0.265\rho, \rho > 65$
Modified Greenberg model	$v = 48, \rho \le 35$	–	$v = 32 \ln \frac{145.5}{\rho}, \rho > 35$
Three-regime model	$v = 50 - 0.098\rho, \rho \le 40$	$v = 81.4 - 0.913\rho, 40 < \rho \le 65$	$v = 40 - 0.265\rho, \rho > 65$

[a] The speed of vehicles is measured in miles per hour, while traffic density is quantified as vehicle per mile.

There has been significant progress in improving the accuracy of both single and multi-regime models. With various attempts and exploration of both parametric and non-parametric regression structures, recent deterministic models have already shown a good fit for empirical data in all density ranges. Therefore, accuracy is no longer the greatest weakness of these models. The deterministic approach has an inherent deficiency in reflecting the density–speed relationship under different states rather than a steady state.

2.2. Stochastic fundamental diagram models

In the 1970s, Soyster and Wilson (1973) proposed an SFD model that utilized a probabilistic density distribution to describe specific flow rates based on a birth-and-death stochastic process. This model, however, relied on highly restrictive assumptions, including the notion that vehicle arrivals followed a Poisson process. Building upon this, Kerner (1998) expanded the framework by offering speed ranges for various densities through three-phase traffic flow models. Further advancements came from Alperovich and Sopasakis (2008), who developed a multi-lane stochastic traffic flow model grounded in microscopic dynamics. Nevertheless, this model did not specify an explicit SFD functional form, omitting a crucial noise or error term. Wang et al. (2013) later introduced a macroscopic stochastic method to depict the speed-flow equilibrium relationship, employing Karhunen–Loeve expansion for estimating variance and mean, yet it lacked explicit stochastic speed–density relationships.

Subsequent research efforts, including those by MacNicholas (2011), Fan and Seibold (2013), Jabari et al. (2014), Qu et al. (2017), and Wang et al. (2021), succeeded in generating distributions for speed or flow as functions of density, achieving relations based on percentiles. However, the models proposed by the first three studies were limited to specific traffic flow scenarios, which restricted their general applicability in broader traffic flow stochasticity contexts. Specifically, MacNicholas (2011) introduced a probabilistic graphical approach to the fundamental traffic diagram based on the triangular deterministic model. Fan and Seibold (2013) incorporated a novel variable, the empty-road velocity, to capture randomness within the ARZ model. Jabari et al. (2014) crafted a probabilistic stationary speed–density relation using Newell's simplified car-following model.

Further, Qu et al. (2017) proposed a percentile-based SFD calibration method, delving into the continuity, differentiability, and convexity of their model. Wang et al. (2021) extended this percentile-based SFD, proposing a comprehensive model capable of coordinating a family of speed–density curves. This model transformed a convex but nondifferentiable model into a linear programming model that was efficiently solvable with advanced solvers. Exploring other aspects, Laval and Chilukuri (2014) utilized a stochastic triangular fundamental diagram to examine kinematic wave models, while Siqueira et al. (2016) analyzed the SFD model from a mesoscopic perspective, focusing on vehicle speed transitions. Building on the triangular fundamental diagram concept, Zhang et al. (2018) established an SFD model that primarily captured the distributions of shockwave speeds.

More recent studies, such as Ni et al. (2018), investigated the SFD through a stochastic process inspired by the Maxwell–Boltzmann distribution, aiming to encapsulate the scattering features in the fundamental diagram. Zhou and Zhu (2020) applied a Gaussian mixture model to depict a triangular fundamental diagram, introducing stochasticity primarily through stochastic headways in a mixed human-driven and automated environment. Ahmed et al. (2021) developed a multimodal SFD model for heterogeneous and undisciplined traffic flows, leveraging unmanned aerial vehicle data from Pakistan. Addressing the uncertainty of speed heterogeneity and rainfall intensity, Bai et al. (2021) proposed an SFD model to uncover both observable and unobservable heterogeneities in traffic flows. Most recently, Cheng et al. (2024) tackled the limitations of deterministic models in highway traffic flow theory by introducing two stochastic fundamental diagram (SFD) models that employ lognormal and skew-normal distributions. These models aim to accurately represent traffic state variations through stochasticity in free-flow speeds and critical density speeds. Validated with real-world loop detector data, these innovative models effectively replicate empirical data scatterings, demonstrating significant potential in analyzing freeway capacities and identifying sources of stochasticity. However, a stochastic model holds both a stochastic form to reflect the overall traffic pattern and a deterministic form to represent the equilibrium state is needed.

Contrary to the majority of macroscopic traffic models, which often use first or second-order approaches and are typically formulated through partial differential equations (PDEs) — much like numerous models in physics, the deterministic models explored in this study do not adhere to a PDE-based framework. Moreover, these models are functions with physics and non-physics parameters, introducing a significant challenge: the imperative for high-quality data to calibrate these empirical models accurately. In the absence of such data, the effectiveness of these models is at risk. Although it is possible to incorporate empirical models into the loss function, doing so will fall into a modeling dilemma and will not necessarily enhance the model's robustness. A robust model can provide accurate estimations and predictions even when the input data is noisy. However, to obtain a robust model, clean data are needed to establish the entire loss function in our case. Once clean data has been identified and processed, using noisy data as input is pointless. The successful enhancement of model robustness through this integration is contingent upon the availability of

pristine data. This research proposes the empirical models into the prior as the mean function in sparse Gaussian process regression (SGPR), with the aim of improving the model's precision rather than its robustness. Most deterministic fundamental models can be directly utilized as this mean function. Further into the text, an examination is conducted on whether this inclusion effectively elevates the performance of the model and offsets the limitations of sparse approximation. The primary objective is to facilitate non-parametric models in extracting more valuable insights from the empirical prior — derived from the entire dataset, even when they are trained on only a subset of this data due to significant computational demands. This strategy seeks to enhance the accuracy of non-parametric regression models by making full use of the empirical data, which cannot be done if only rely on non-parametric regression. The primary goal of utilizing empirical models as the mean function is to enhance the precision of non-parametric models when fitting the empirical data. While this approach shares similarities with methods employed in physics-informed machine learning, it is important to note that this research does not classify as physics-informed machine learning.

3. Methodology

In this section, we will discuss how to generate the non-parametric stochastic fundamental diagram model. The basic methodology consists of three main parts: calibration of the empirical deterministic model, non-parametric stochastic modeling, and integration of empirical prior.

3.1. Empirical deterministic model

In our overall modeling framework, the empirical deterministic model plays a very important role in bridging the traditional deterministic models and the statistics model. As the prior of the statistics model, we need to ensure that all empirical deterministic models must be well-calibrated. The least square method (LSM) has been widely used to calibrate regression-based models in transportation research areas, such as Sun et al. (2003) and Meng and Qu (2013). When calibrating the single-regime fundamental diagram model, the LSM may encounter many drawbacks. The real-world data used to calibrate the single-regime fundamental diagram model is unbalanced distributed, and most density–speed pairs are located in the low-traffic density area. As a result, the fitted model based on LSM will be well fitted in high-density areas but skewed in low-density areas and may not be able to reflect the accurate real pattern of the density–speed relations. Therefore, following Qu et al. (2015), a weighted least square (WLSM) method is adopted to calibrate all the single-regime and multi-regime fundamental diagram models in this paper. Considering a density–speed pairs dataset $\mathcal{D}$, $\mathcal{D} = \{(\rho_1, v_1), (\rho_2, v_2), \ldots, (\rho_m, v_m)\}$ and a function $v = f(\rho, \theta)$, the LSM is aiming to minimize the sum of squared errors as shown in Eq. (5), due to the continuously differentiable assumption of this unconstrained optimization problem, its first-order optimality condition can be expressed as Eq. (6). Unlike LSM, WLSM minimizes Eq. (7), whose first-order optimality condition can now be written as Eq. (8).

$$\min \mathcal{L}(\theta) = \sum_{i=1}^{n} (v_i - f(\rho_i, \theta))^2 \tag{5}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} = -2 \sum_{i=1}^{n} [v_i - f(\rho_i, \theta)] \frac{\partial f(\rho_i, \theta)}{\partial \theta_j} = 0, \quad j = 1, \ldots, n \tag{6}$$

$$\min \mathcal{L}(\theta) = \sum_{i=1}^{n} \varpi_i (y_i - f(x_i, \theta))^2 \tag{7}$$

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} = -2 \sum_{i=1}^{n} \varpi_i [v_i - f(\rho_i, \theta)] \frac{\partial f(\rho_i, \theta)}{\partial \theta_j} = 0, \quad j = 1, \ldots, n \tag{8}$$

In our paper, all the single-regime fundamental diagram models will be calibrated through WLSM proposed by Qu et al. (2015), and the calibration[2] will be done by an optimizer in Virtanen et al. (2020).

Several classic and recent single-regime traffic flow models have been listed in Table 1. The calibrated parameters of each model are presented in Table 3. However, the jam density of MacNicholas's Model (MacNicholas, 2011) has lost its physical meaning. Therefore, only a portion of the calibrated models are shown in Figs. 4(a) to 4(l).

3.2. Non-parametric stochastic modeling

In the section, we will introduce the non-parametric method that we used to generate our stochastic model. Sparse Gaussian process regression (SGPR), as a powerful non-parametric regression tool, has been selected to archive stochastic modeling.

Before formally introducing the SGPR, let us quickly review the Gaussian process regression. Following Williams and Rasmussen (2006), the Gaussian process (GP) is employed to depict a distribution over functions, allowing GP to serve as a powerful analysis method from a function-space perspective. Formally:

Definition 1 (*Gaussian Process*). A Gaussian process is a collection of random variables, any finite number of which have a Gaussian distribution.

---

[2] The calibration code is shown in https://github.com/YuanzhengLei/Weighted-the-least-square-method-for-single-regime-fundamental-diagram-models.

Table 3
Parameters of single-regime traffic flow model after calibration.

Model	Parameters[a]
Greenshields et al. (1935)	$v_f = 52.12$ $\rho_j = 76.68$
Greenberg (1959)	$v_{critical} = 22.06$ $\rho_j = 92.49$
Newell (1961)	$v_f = 69.69$ $\rho_j = 25.000$ $\lambda = 1209.02$
Drake et al. (1965)	$v_f = 80.50$ $\rho_{critical} = 50.01$
Underwood (1961)	$v_f = 80.51$ $\rho_{critical} = 92.49$
Papageorgiou et al. (1989)	$v_f = 79.49$ $\rho_j = 24.83$ $\alpha = 1.02$
Kerner and Konhäuser (1994)	$v_f = 60.17$ $\rho_{critical} = 106.27$
Del Castillo and Benitez (1995)	$v_f = 69.69$ $\rho_j = 108.41$ $v_j = 11.15$
Jayakrishnan et al. (1995)	$v_f = 52.1198$ $v_{min} = 35.0052$ $\rho_j = 25.1779$
Ardekani and Ghandehari (2008)	$v_{critical} = 40.41$ $\rho_j = 56.84$ $\rho_{min} = 0.01$
MacNicholas (2011)	$v_f = 70.17$ $\rho_j = 2410.54$ $n = 2.00$ $m = 13730.07$
Pipes (1966)	$v_f = 76.05$ $\rho_j = 51.00$ $n = 1.22$
Wang et al. (2011)	$v_f = 65.23$ $v_{critical} = 6.02$ $\rho_{critical} = 9.73$ $\theta_1 = 1.53$ $\theta_2 = 0.10$
Cheng et al. (2021)	$v_f = 68.70$ $\rho_{critical} = 20.02$ $m = 2.21$

[a] The speed of vehicles is measured in miles per hour (mph), while traffic density is quantified as vehicle per mile.

For any Gaussian process, it can be specified by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. This function is commonly called the kernel of the Gaussian process (Jäkel et al., 2007), and a prevalent choice of a kernel function is the radial basis function kernel, which is defined as $k(\mathbf{x}, \mathbf{x}') = \sigma^2 e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\lambda^2}}$, σ and λ are hyper-parameters, which will decide the shape of the kernel function. The mean function and the covariance function of a real process $f(\mathbf{x})$ can be defined as:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{9}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \tag{10}$$

and the Gaussian process can be expressed as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{11}$$

Typically, the mean function is assumed to be zero. Consequently, in subsequent sections of this study, for a pure Gaussian process regression (GPR) model, its mean function $m(\rho) = 0$, for other empirical prior Gaussian process regression (EPGPR) models, because they are based some deterministic fundamental diagram model $f(\rho)$, their mean function $m(\rho) = f(\rho)$.

Based on Wang (2023), if points x_i and x_j are considered similar by the covariance function, their corresponding function outputs, $f(x_i)$ and $f(x_j)$, are expected to be similar too. The Gaussian process regression is illustrated as: given observed data $(\mathbf{x}_i, f_i | i = 1, 2, 3, \ldots, n)$ and a mean function $\mathbf{f}$ estimated from the observed data, to predict at new points $\mathbf{x}_*$ as $\mathbf{f}_*$. The joint distribution of $\mathbf{f}$ and $\mathbf{f}_*$ is actually a multivariate Gaussian/normal distribution, which can be expressed as:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{nn} & \mathbf{K}_{n*} \\ \mathbf{K}_{*n} & \mathbf{K}_{**} \end{bmatrix} \right) \tag{12}$$

where $\mathbf{K}_{nn} = \mathbf{K}(\mathbf{x}, \mathbf{x})$,[3] $\mathbf{K}_{n*} = \mathbf{K}(\mathbf{x}, \mathbf{x}^*)$,[4] $\mathbf{K}_{**} = \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*)$,[5] Eq. (13) gives the joint probability distribution of $\mathbf{f}, \mathbf{f}_* | \mathbf{x}, \mathbf{x}_*$ over $\mathbf{f}$ and $\mathbf{f}_*$. However, in a regression problem, what we really need is the conditional distribution $\mathbf{f}_* | \mathbf{f}, \mathbf{x}, \mathbf{x}_*$ over $\mathbf{f}_*$, which can be written as[6]:

$$\mathbf{f}_* | \mathbf{f}, \mathbf{x}, \mathbf{x}_* \sim \mathcal{N}(\mathbf{K}_{*n} \mathbf{K}_{nn}^{-1} \mathbf{f}, \mathbf{K}_{**} - \mathbf{K}_{*n} \mathbf{K}_{nn}^{-1} \mathbf{K}_{*n}) \tag{13}$$

However, in a more realistic modeling scenarios, $\mathbf{f}$ itself is intractable. The tractable one is the noisy version expressed as $\mathbf{y} = f(\mathbf{x}) + \epsilon$. Assuming ϵ is an additive independent and identically distributed Gaussian noise, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, then the prior on

---

[3] If there are n training points, $\mathbf{K}(\mathbf{x}, \mathbf{x}) = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & k(x_1, x_3) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & k(x_2, x_3) & \cdots & k(x_2, x_n) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ k(x_n, x_1) & k(x_n, x_2) & k(x_n, x_3) & \cdots & k(x_n, x_n) \end{bmatrix}$.

[4] If there are n testing points, $\mathbf{K}(\mathbf{x}, \mathbf{x}^*) = \begin{bmatrix} k(x_1, x_1^*) & k(x_1, x_2^*) & k(x_1, x_3^*) & \cdots & k(x_1, x_n^*) \\ k(x_2, x_1^*) & k(x_2, x_2^*) & k(x_2, x_3^*) & \cdots & k(x_2, x_n^*) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ k(x_n, x_1^*) & k(x_n, x_2^*) & k(x_n, x_3^*) & \cdots & k(x_n, x_n^*) \end{bmatrix}$.

[5] If there are n training and testing points, $\mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) = \begin{bmatrix} k(x_1^*, x_1^*) & k(x_1^*, x_2^*) & k(x_1^*, x_3^*) & \cdots & k(x_1^*, x_n^*) \\ k(x_2^*, x_1^*) & k(x_2^*, x_2^*) & k(x_2^*, x_3^*) & \cdots & k(x_2^*, x_n^*) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ k(x_n^*, x_1^*) & k(x_n^*, x_2^*) & k(x_n^*, x_3^*) & \cdots & k(x_n^*, x_n^*) \end{bmatrix}$.

[6] Details of how to obtain $\mathbf{f}_* | \mathbf{f}, \mathbf{x}, \mathbf{x}_*$ can be found in Bishop (2006) and Williams and Rasmussen (2006).

the noisy observations becomes: and Eq. (12) can be rewritten as:

$$
\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{nn} + \sigma_\epsilon^2 \mathbf{I} & \mathbf{K}_{n*} \\ \mathbf{K}_{*n} & \mathbf{K}_{**} \end{bmatrix} \right)
\tag{14}
$$

By deriving the conditional distribution, the predictive equations for Gaussian process regression will be expressed as:

$$
\mathbf{f}_* | \mathbf{y}, \mathbf{x}, \mathbf{x}_* \sim \mathcal{N}(\overline{\mathbf{f}}_*, cov(\mathbf{f}_*))
\tag{15}
$$

where $\overline{\mathbf{f}}_*$ and $cov(\mathbf{f}_*)$ can be further expressed as:

$$
\overline{\mathbf{f}}_* = \mathbf{K}_{*n}[\mathbf{K}_{nn} + \sigma_\epsilon^2 \mathbf{I}]^{-1} \mathbf{y}
\tag{16}
$$

$$
cov(\mathbf{f}_*) = \mathbf{K}_{**} - \mathbf{K}_{*n}[\mathbf{K}_{nn} + \sigma_\epsilon^2 \mathbf{I}]^{-1} \mathbf{K}_{n*}
\tag{17}
$$

Similarly, the prediction of the output $\mathbf{y}_*, \mathbf{y}_* = \mathbf{f}_* + \epsilon$ is described by (Titsias, 2009a):

$$
\mathbf{y}_* | \mathbf{y}, \mathbf{x}, \mathbf{x}_* \sim \mathcal{N}(\mathbf{y}_* | \mathbf{K}_{*n}[\mathbf{K}_{nn} + \sigma_\epsilon^2 \mathbf{I}]^{-1} \mathbf{y}, \mathbf{K}_{**} - \mathbf{K}_{*n}[\mathbf{K}_{nn} + \sigma_\epsilon^2 \mathbf{I}]^{-1} \mathbf{K}_{n*})
\tag{18}
$$

The posterior GP also depends on the values of the hyper-parameters $\Theta = (\sigma, \lambda)$ (Shown in kernel function), which are determined by maximizing the log marginal likelihood as follows:

$$
\Theta_* = \arg\max_{\Theta} \log(\mathbf{y} | \mathbf{x}, \Theta) = \arg\max_{\Theta} \log[\mathcal{N}(\mathbf{y} | m(\mathbf{x}), \mathbf{K}_{nn} + \sigma^2 \mathbf{I})]
\tag{19}
$$

However, although GP is a compelling and popular non-parametric method and has been shown to perform well on various tasks (Walder et al., 2008), the application of GP models in large datasets is still tricky (Titsias, 2009a) due to the time complexity as $\mathcal{O}(n^3)$ and the storage as $\mathcal{O}(n^2)$, where n is the size of samples. Many approximate or sparse methods (Williams and Seeger, 2000; Smola and Bartlett, 2000; Csató and Opper, 2002; Lawrence et al., 2002; Seeger et al., 2003; Schwaighofer and Tresp, 2002; Snelson and Ghahramani, 2005; Quinonero-Candela and Rasmussen, 2005; Titsias, 2009a) have been proposed to solve this dilemma. The core idea of those studies is to employ a small set with m inducing variables to form a reduction of the time complexity from $\mathcal{O}(n^3) \dashrightarrow \mathcal{O}(m^2 n)$.

In this paper, we adopted the variational framework proposed in Titsias (2009a). Suppose m inducing variables are selected to conduct the sparse approximation, which can be expressed as $\mathbf{f}_m = \{f_1^m, f_2^m, \ldots, f_m^m\}$ that are latent function values evaluated at input points $\mathbf{x}_m, \mathbf{x}_m = (x_1^m, x_2^m, \ldots, x_m^m)$. For the posterior GP, it can be written by:

$$
p(\mathbf{f}_* | \mathbf{y}) = \int p(\mathbf{f}_* | \mathbf{f}) p(\mathbf{f} | \mathbf{y}) d\mathbf{f} = \int p(\mathbf{f}_* | \mathbf{f}_m, \mathbf{f}) p(\mathbf{f} | \mathbf{f}_m, \mathbf{y}) p(\mathbf{f}_m | \mathbf{y}) d\mathbf{f} d\mathbf{f}_m
\tag{20}
$$

where $\mathbf{f}_*$ is the predict function values at new inputs $\mathbf{x}_*$. Referring to Tiao (2020) and Krasser (2020), then $p(\mathbf{f}_* | \mathbf{y})$ can be further written as an approximate posterior (For detailed derivatives, please refer to Lemma 1):

$$
q(\mathbf{f}_*) = \int p(\mathbf{f}_* | \mathbf{f}_m) \phi(\mathbf{f}_m) d\mathbf{f}_m
\tag{21}
$$

where $\phi(\mathbf{f}_m)$ is an approximation to $p(\mathbf{f}_m | \mathbf{y})$, which is intractable and can be further written as:

$$
\phi(\mathbf{f}_m) = \mathcal{N}(\mathbf{f}_m | \boldsymbol{\mu}_m, \mathbf{A}_m)
\tag{22}
$$

The mean and covariance matrix of the Gaussian approximate posterior $q(\mathbf{f}_*)$ can be further defined as:

$$
q(\mathbf{f}_*) = \mathcal{N}(\mathbf{f}_* | \boldsymbol{\mu}_*^q, \boldsymbol{\Sigma}_*^q)
\tag{23}
$$

$$
\boldsymbol{\mu}_*^q = \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1} \boldsymbol{\mu}_m
\tag{24}
$$

$$
\boldsymbol{\Sigma}_*^q = \mathbf{K}_{**} - \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1} \mathbf{K}_{m*} + \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1} \mathbf{A}_m \mathbf{K}_{mm}^{-1} \mathbf{K}_{m*}
\tag{25}
$$

Then, by minimizing the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) between the approximate $q(\mathbf{f})$ and the exact posterior $p(\mathbf{f} | \mathbf{y})$, which is relevant to minimize the KL divergence between the augmented variational posterior $q(\mathbf{f}, \mathbf{f}_m | \mathbf{y})$ and the augmented true posterior $p(\mathbf{f}, \mathbf{f}_m | \mathbf{y})$. Based on Titsias (2009a,b), the minimization is equivalently expressed as the maximization of the following variational lower bound of the true log marginal likelihood (Please check the detailed derivative in Lemma 2):

$$
F_V(\mathbf{x}_m, \phi) = \int p(\mathbf{f} | \mathbf{f}_m) \phi(\mathbf{f}_m) \log \frac{p(\mathbf{f} | \mathbf{y}) p(\mathbf{f}_m)}{\phi(\mathbf{f}_m)} d\mathbf{f} d\mathbf{f}_m
\tag{26}
$$

And we can finally have the optimal values of $\boldsymbol{\mu}_m$ and $\mathbf{A}_m$ as (Please check the detailed derivative in Lemmas 3 and 4):

$$
\boldsymbol{\mu}_m = \frac{1}{\sigma^2} \mathbf{K}_{mm} \boldsymbol{\Sigma} \mathbf{K}_{mn} \mathbf{y}
\tag{27}
$$

$$
\mathbf{A}_m = \mathbf{K}_{mm} \boldsymbol{\Sigma} \mathbf{K}_{mm}
\tag{28}
$$

where $\Sigma = (\mathbf{K}_{mm} + \sigma^{-2}\mathbf{K}_{mn}\mathbf{K}_{nm})^{-1}$, and after substituted $\boldsymbol{\mu}_m$ and $\mathbf{A}_m$ into Eqs. (25) and (26), we can compute the approximate posterior $q(\mathbf{f}_*)$ at new inputs $\mathbf{x}_*$.

As mentioned in the preceding section, sparse Gaussian Process Regression (SGPR) requires m inducing variables to approximate the exact GP model closely. Thus, in the paper, four sampling methods, including simple random sampling, systematic sampling, cluster sampling, and weighted random sampling, are used to choose inducing variables from the entire dataset to conduct comparison studies, and details of sampling methods can be found in the Appendix A.

3.3. Empirical prior integration

In this section, the methodology for integrating an empirical prior into a non-parametric model is explored. As previously delineated in Section 3, a Gaussian Process (GP) is defined by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$. Given a regression model $\mathbf{y} = f(\mathbf{x}) + \epsilon$, the mean function $m(\mathbf{x})$ encapsulates prior knowledge regarding the unknown function $f(\mathbf{x})$. In most cases, the prior is assumed to be zero, and integrating an empirical prior into a non-parametric model essentially replaces the prior with empirical functions about the speed–density relationship. After the change in the assumption on the prior, the variation distribution $\phi(\mathbf{f}_m)$ shown in Eq. (23) becomes (Please check the detailed derivative in Lemma 5):

$$\phi(\mathbf{f}_m) = \mathcal{N}(\mathbf{f}_m | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \tag{29}$$

$$\boldsymbol{\Sigma}_* = \left(\mathbf{K}_{mm}^{-1} + \frac{1}{\sigma^2}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm} \right)^{-1} \tag{30}$$

$$\boldsymbol{\mu}_* = m(\mathbf{x}_m) + \boldsymbol{\Sigma}_* \left(\frac{1}{\sigma^2}\mathbf{K}_{mm}^{-1}\mathbf{K}_{nm}(\mathbf{y} - m(\mathbf{x})) \right) \tag{31}$$

The $q(\mathbf{f}_*)$ now becomes(Please check the detailed derivative in Lemma 6):

$$q(\mathbf{f}_*) = \mathcal{N}(m(\mathbf{x}_*)) + \mathbf{K}_{*m}\mathbf{K}_{mm}^{-1}(\boldsymbol{\mu}_* - m(\mathbf{x}_m)), \mathbf{K}_{**} - \mathbf{K}_{*m}\mathbf{K}_{mm}^{-1}\mathbf{K}_{m*} + \mathbf{K}_{*m}\mathbf{K}_{mm}^{-1}\boldsymbol{\mu}_*\mathbf{K}_{mm}^{-1}\mathbf{K}_{m*}) \tag{32}$$

Predominantly, deterministic models are parametric, devoid of partial derivatives, and describe the equilibrium speed–density relationship. Consequently, these models can be seamlessly employed as the mean function within SGPR. Namely, the mean function $m(\mathbf{v}) = 0$ will be replaced by $m(\mathbf{v}) = f(\rho)$, $f(\rho)$ is a deterministic fundamental diagram model. For example, if Greenshield's model (Greenshields et al., 1935), Greenberg's model (Greenberg, 1959), and Newell's model (Newell, 1961) will be used as the prior, then the corresponding mean function of the corresponding SGPR will be set as:

$$m_1(v) = v_f \left(1 - \frac{\rho}{\rho_j} \right) \tag{33}$$

$$m_2(v) = v_{\text{critical}} \log \left(\frac{\rho_j}{\rho} \right) \tag{34}$$

$$m_3(v) = v_f \left\{ 1 - \exp \left[-\frac{\lambda}{v_f}(\frac{1}{\rho} - \frac{1}{\rho_j}) \right] \right\} \tag{35}$$

The diagram in Fig. 3 summarizes the process of integrating an empirical prior into a non-parametric model. Modules highlighted with green backgrounds utilize the entire dataset, whereas those with blue backgrounds only employ a subset of the dataset. Owing to the computationally intensive nature of Gaussian processes, a subset of the dataset is used for sparse modeling to manage computational demands. In contrast, the calibration of deterministic models exploits the full dataset. By employing calibrated empirical models as the prior, we aim to mitigate the accuracy loss inherent in the variational approximation.

Furthermore, kernel functions play a crucial role in SGPR. After evaluating various kernel functions such as the Squared Exponential Kernel, Rational Quadratic Kernel, Matérn kernels, and the Exponential Kernel, it was determined that the Exponential Kernel demonstrated the highest accuracy across all sampling methodologies. Therefore, this kernel function has been selected for implementation, as shown below:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 e^{-\frac{\|\mathbf{x} - \mathbf{x}'\|}{2}} \tag{36}$$

4. Case study

4.1. Data description

In this case study, data collected from loop detectors across 76 stations on the Georgia State Route 400 serve as the empirical dataset (Denotes as GA400). This data has been extensively utilized within the research community for developing fundamental diagram models, as evident by studies such as Wang et al. (2011, 2013) and Qu et al. (2015, 2017). The GA400 dataset comprises 44,787 pairs of density–speed observations, indicating a dataset size of $n = 44,787$. The choice of $m = 288$, where $288 \ll 44,787$, is justified by the data collection frequency: with a pair recorded every 5 min, a single day yields $\frac{24 \times 60}{5} = 288$ data pairs. This parameterization comprehensively represents daily traffic patterns within a manageable computational framework.
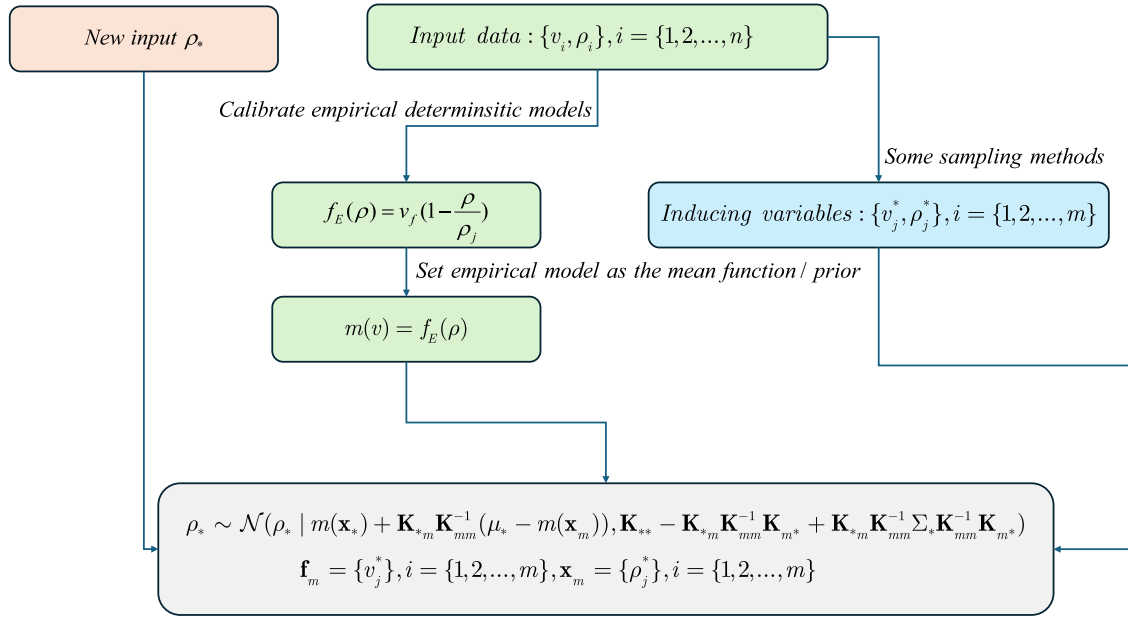
Fig. 3. Empirical prior integration.

4.2. Model validation

Figs. 5(a) through 5(m) illustrate the speed–density relationships in a two-dimensional depiction based on prior assumptions. The estimation curve represents the posterior mean function curve, which embodies the speed–density relationship with the maximum likelihood. This means that for any given density value, the speed value situated along this curve has the highest likelihood of occurrence in empirical scenarios.

The stochastic fundamental diagram model proposed in this study is essentially a Gaussian process. Its domain of definition includes an infinite one-dimensional Gaussian distribution looking towards the y-axis direction. The confidence intervals of 90%, 95%, and 99% have been specified, and within the 95% confidence interval, for any given density, the stochastic model believes the corresponding speed has a 95% likelihood of falling within this specified range.

Furthermore, Figs. 6(a) through 6(m) show the corresponding three-dimensional depiction of the speed–density relationships, informed by various prior assumptions, referring to Fig. 2. These figures illustrate the different probability densities of traffic speed given any traffic density. Therefore in this paper, the distribution of traffic speed, denoted as $w(v)$, is considered as a function of traffic density, denoted as $f(\rho)$. This is different from the conventional deterministic models, where the traffic speed is solely a function of traffic density (i.e., $v = f(\rho)$). Here, we take into account the fact that the distribution of traffic speed is a function of traffic density (i.e., $w(v) = f(\rho)$). $w(v) = f(\rho)$ depicts the overall traffic pattern of the density–speed relationship under all situations, and the generated posterior mean function curve $v = m(\rho)$ can be regarded as a deterministic representation of the equilibrium density–speed relationship under a steady state.

Two metrics, Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE), are employed to evaluate the accuracy of fitting models to empirical data. The formulas for these metrics are as follows:

$$\textbf{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i^o - v_i^e)^2} \tag{37}$$

$$\textbf{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{v_i^o - v_i^e}{v_i^o}\right| \times 100\% \tag{38}$$

where v_i^o denotes the observed speed corresponding to the observation density ρ_i^o, and v_i^e represents the estimated speed for the same observation density. These metrics provide insights into the comparative performance of different models.

Table 4 shows the comparison of the results of different empirical models. Table 5 compares the estimation accuracy of various empirical prior Gaussian regression models with the pure Gaussian process regression model across different sampling techniques. Table 6 shows the percentage of the empirical data points that fall within the 95% confidence interval. Here, $\mathcal{RS}$, $\mathcal{SS}$, $\mathcal{CS}$, and $\mathcal{WRS}$ denote simple random sampling, systematic sampling, cluster sampling, and weighted random sampling, respectively. We also used the I-80 detector dataset from NGSIM for testing, and all results are shown in the appendix from Tables 7 to 9.

First, it is observed that among empirical models, the most precise one is Cheng's model (Cheng et al., 2021). However, the Full Gaussian Process (GP) Model surpasses these empirical models in terms of accuracy. The Pure GP Model's superior performance is
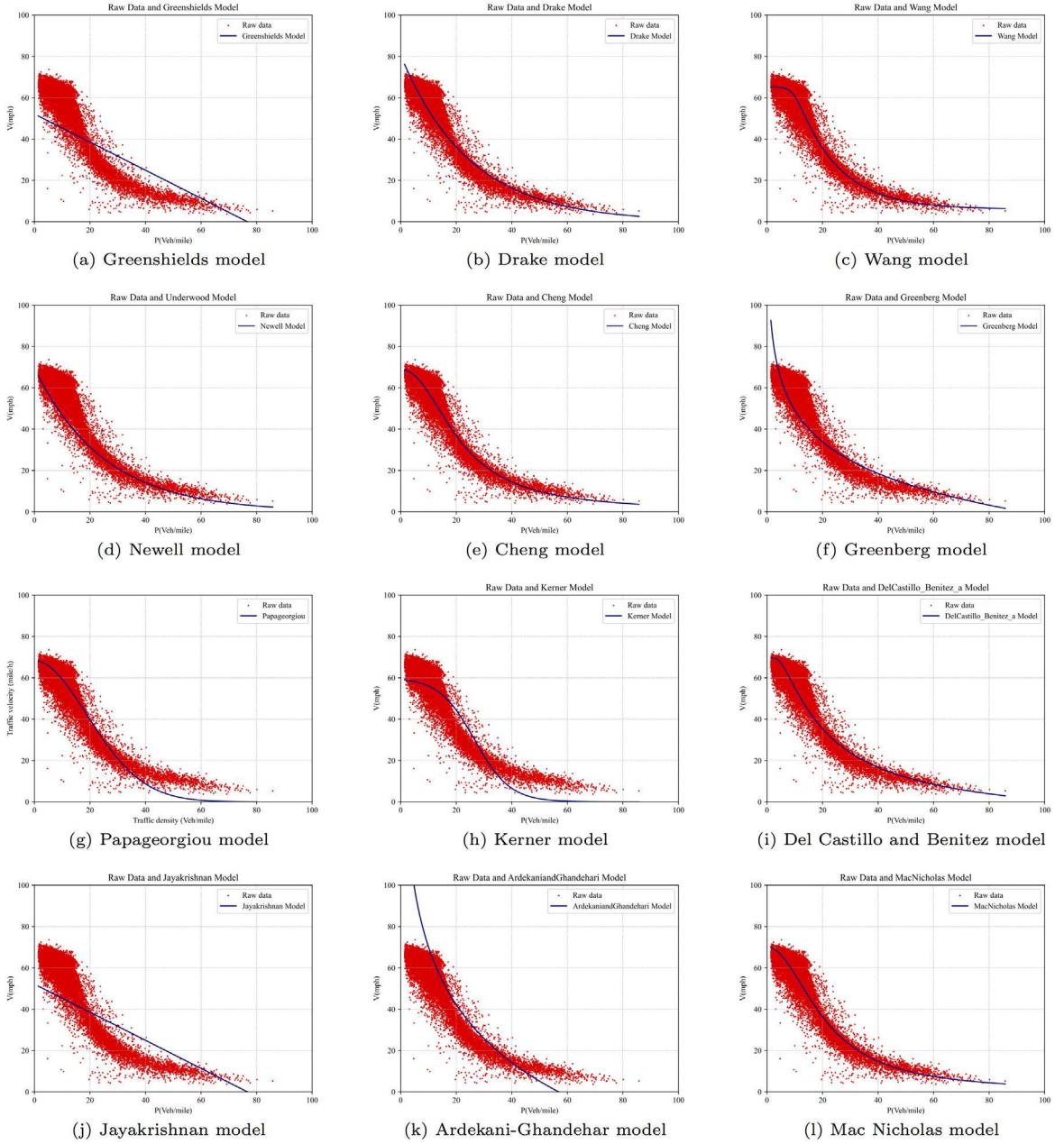
Fig. 4. Single-regime traffic flow models after calibration.

achieved with only 288 data pairs as opposed to the 44,787 utilized by the empirical models; also, from Table 6, all models are well-fitted because they all capture the distribution of the empirical data under sparse approximation. Also, we should notice that model performances are nearly the same under all sampling methods for all models, which indicates that all sampling methods in our paper are suitable choices for the SGPR modeling.

Secondly, to figure out whether the sparse regression successfully captures the stochastic of the density–speed relationship. The percentage density–speed points that fall within the 95% confidence interval (PWCI) have been tested. The percentage of points that fall within the 95% confidence interval (PWCI) of all models is higher than 94%. The superior performance of sparse Gaussian process regression methods in accurately modeling the stochastic density–speed relationship is convincingly illustrated by the above facts.

Thirdly, the results shown in Table 5 unequivocally indicate that the Pure GP Model is more accurate compared to other models across all sampling methods employed. This leads to the inference that an accurate prior mean function may not sufficiently
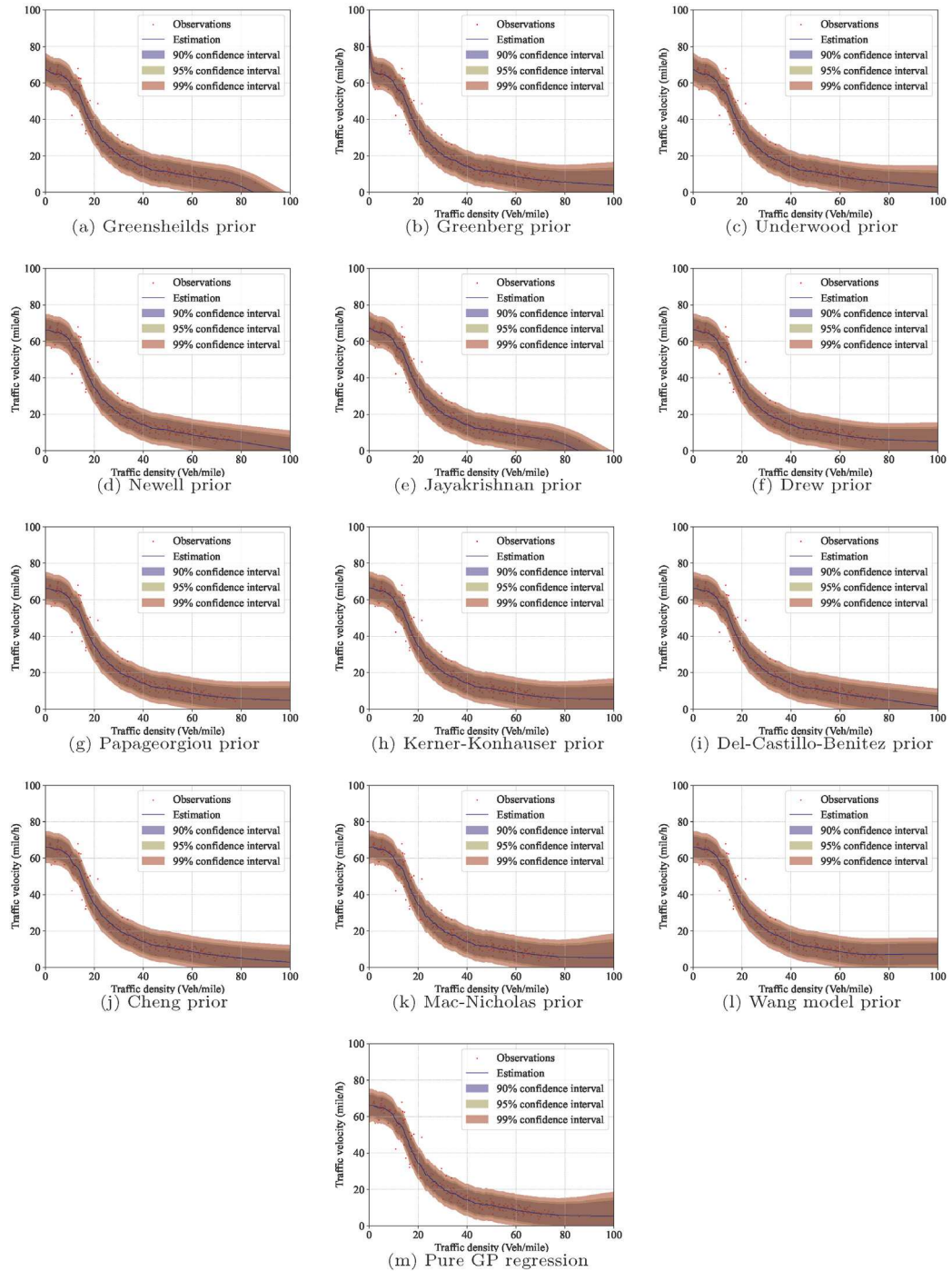
Fig. 5. 2D Visualization of speed–density relationships based on different priors.

compensate for the limitations inherent in sparse approximations when seeking to enhance estimation accuracy. A purely data-driven approach may yield more accurate results in situations with a relatively small amount of quality data. Notably, the same results are also shown in the I-80 dataset, shown in Tables 7 to 9. Such findings imply that even a training dataset with a modest size can contain sufficient information for effective model training. Consequently, these insights prompt further questions: *"What is the minimal size of a dataset that remains effective? How much data is necessary to counterbalance the advantages of employing an empirical prior?"*
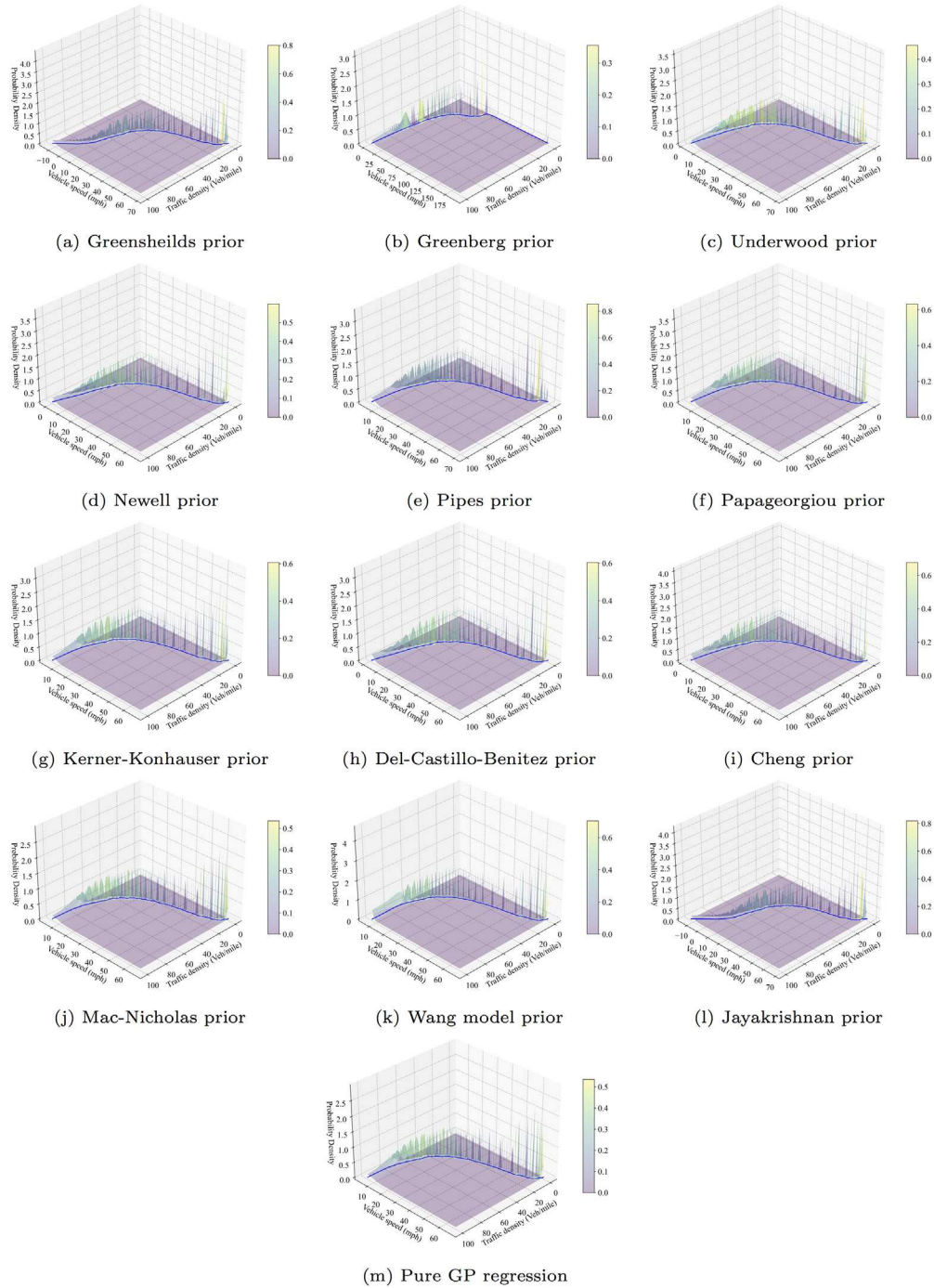
(a) Greensheilds prior

(b) Greenberg prior

(c) Underwood prior

(d) Newell prior

(e) Pipes prior

(f) Papageorgiou prior

(g) Kerner-Konhauser prior

(h) Del-Castillo-Benitez prior

(i) Cheng prior

(j) Mac-Nicholas prior

(k) Wang model prior

(l) Jayakrishnan prior

(m) Pure GP regression

Fig. 6. 3D Visualization of speed–density relationships based on different priors.

Therefore, a series of experiments have been conducted to find the fitting accuracy of both empirical prior Sparse Gaussian process regression model (EPGPR) and pure Sparse Gaussian process regression model under increasing training dataset size. As shown in Figs. 7(a) to 7(k) and 8(a) to 8(k), a comparison has been made to discover both RSME and MAPE between the pure GPR model and different EPGPR models with the increase of the training dataset size. Our findings indicated no significant improvement in accuracy after the training dataset size exceeded 300 for all the models. It is worth noting that all results shown in Figs. 7(a) to 8(k) are based on weighted random sampling. Experimental results based on other sampling methods are shown in the supplementary materials. The warm color area indicates the ESME or MPAE of the relevant proposed EPGPR model exceeds that of the pure

Table 4
Comparison of the result of different empirical models.

Model	Speed RMSE (mph)	Speed MAPE
Original Greenshields' Model	15.43	26.80%
Original Greenberg's Model	68.59	116.52%
Original Underwood's Model	15.58	34.82%
Original Newell's Model	14.23	29.33%
Original Pipes' Model	4.25	7.41%
Original Papageorgiou's Model	5.52	8.77%
Original Kerner–Konhause's Model	6.73	12.49%
Original Del Castillo–Benitez's Model	12.14	18.97%
Original Jayakrishnan's Model	15.43	26.80%
Original Cheng's Model	3.81	5.64%
Original Wang's Model	3.39	4.48%

Table 5
Comparison of the result of different EPNR models with Exponential kernel function (m = 288).

EPNR model	Speed RMSE (mph)				Speed MAPE			
	$\mathcal{RS}$	$\mathcal{SS}$	$\mathcal{CS}$	$\mathcal{WRS}$	$\mathcal{RS}$	$\mathcal{SS}$	$\mathcal{CS}$	$\mathcal{WRS}$
Pure GP	3.31	3.32	3.31	3.31	4.43%	4.43%	4.43%	4.43%
Cheng's	3.32	3.32	3.32	3.32	4.43%	4.43%	4.43%	4.43%
Wang's	3.32	3.32	3.32	3.32	4.44%	4.44%	4.44%	4.43%
Greenshields's	3.32	3.32	3.32	3.32	4.44%	4.44%	4.43%	4.43%
Greenberg's	3.31	3.32	3.32	3.32	4.43%	4.43%	4.43%	4.43%
Underwood's	3.32	3.32	3.32	3.32	4.43%	4.43%	4.43%	4.43%
Newell's	3.32	3.32	3.32	3.32	4.43%	4.43%	4.43%	4.43%
Pipes's	3.32	3.32	3.32	3.32	4.43%	4.43%	4.43%	4.43%
Papageorgiou's	3.32	3.32	3.32	3.32	4.43%	4.43%	4.43%	4.43%
Kerner's	3.32	3.32	3.32	3.32	4.43%	4.43%	4.43%	4.43%
Del-Castillo's	3.32	3.32	3.32	3.32	4.44%	4.44%	4.43%	4.43%
Jayakrishnan's	3.32	3.32	3.32	3.32	4.44%	4.44%	4.43%	4.43%

Table 6
Comparison of the result of percentages of points falling within the 95% confidence interval.

EPNR model	PWCI			
	$\mathcal{RS}$	$\mathcal{SS}$	$\mathcal{CS}$	$\mathcal{WRS}$
Pure GP	94.27%	94.27%	94.29%	94.26%
Cheng's	94.22%	94.21%	94.23%	94.21%
Wang's	94.22%	94.22%	94.22%	94.22%
Greenshields's	94.26%	94.25%	94.25%	94.26%
Greenberg's	94.26%	94.26%	94.25%	94.26%
Underwood's	94.25%	94.26%	94.25%	94.26%
Newell's	94.23%	94.23%	94.22%	94.23%
Pipes's	94.24%	94.23%	94.23%	94.24%
Papageorgiou's	94.21%	94.21%	94.22%	94.23%
Kerner's	94.23%	94.23%	94.24%	94.24%
Del-Castillo's	94.22%	94.23%	94.23%	94.22%
Jayakrishnan's	94.25%	94.25%	94.25%	94.26%

SGPR model. Conversely, the cold color area indicates the ESME or MPAE of the relevant proposed EPGPR model is lower than the pure SGPR model. Drawing insights from the above pictures, only when well-fitted empirical models are utilized as the mean function, and notably, this improvement is largely confined to scenarios involving small training datasets. Despite the potential benefits offered by well-calibrated priors, such as those proposed in Wang's model (Wang et al., 2011), Cheng's model (Cheng et al., 2021), Papageorgiou's model (Papageorgiou et al., 1989), Kerner's model (Kerner and Konhäuser, 1994), and Del-Castillo's model (Del Castillo and Benitez, 1995), the empirical advantage diminishes as the size of the training dataset expands. Considering RSME and MAPE simultaneously, only the prior based on Cheng's and Kerner's models can benefit the EPGPR under a very small training dataset. Employing a judicious sampling strategy for choosing inducing variables, a training dataset comprising merely 30 density–speed pairs suffices for a purely data-driven model to outperform a model incorporating a well-fitted empirical prior for the GA 400 dataset. The findings of this study suggest that the speed–density relationship may not exhibit a complex nonlinear nature. Applying highly specific prior may lead to overfitting of the entire model. Consequently, selecting a suitable sampling method, coupled with the advantage of relatively clean and ample datasets, indicates that data-driven modeling could emerge as a more effective approach for developing stochastic fundamental diagram models.
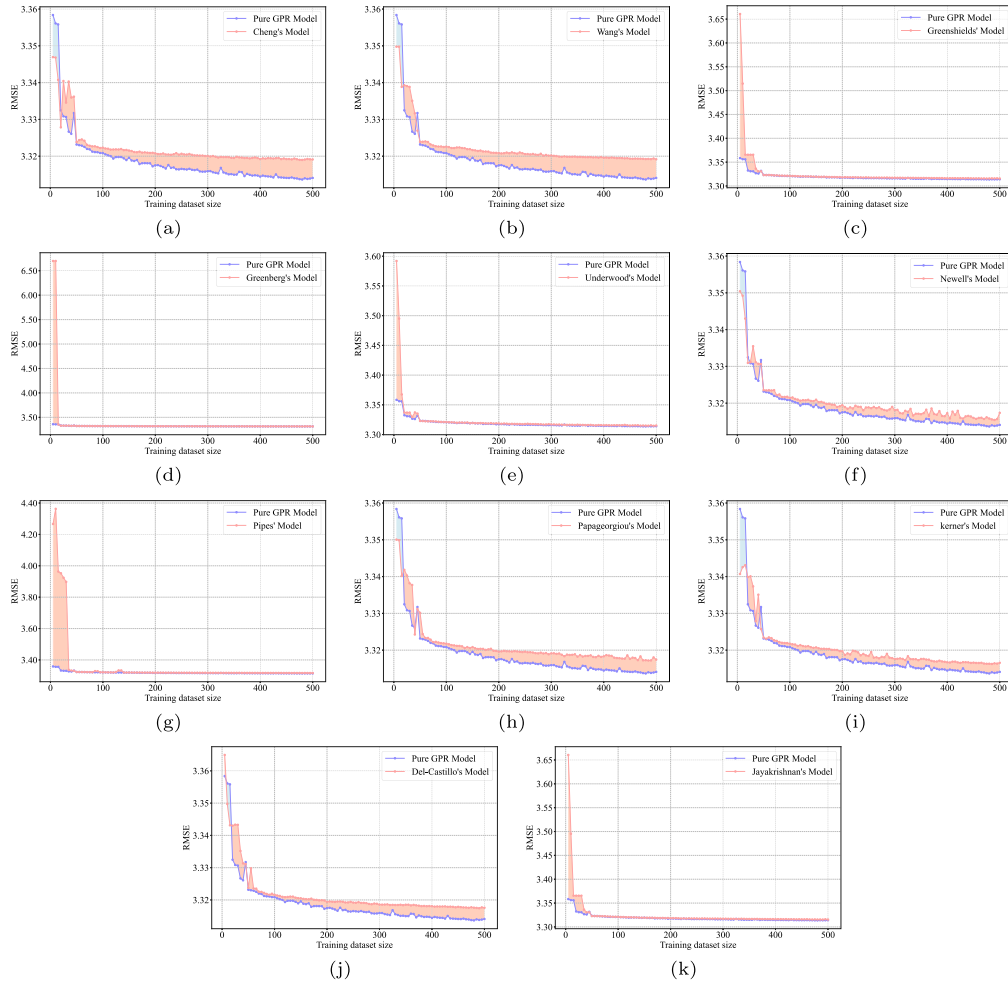
Fig. 7. RSME comparison between pure GPR and EPGPR model under different training dataset size.

5. Conclusion

This research employs a sparse Gaussian process regression framework to analyze the stochastic fundamental diagram model. SGPR, a non-parametric regression technique, counters its intrinsic high computational requirements with a sparse approximation strategy, thereby enhancing fitting accuracy beyond that of all empirical models discussed herein. The Gaussian process distinguishes itself from other non-parametric regressions by its inherent stochastic nature, which is well-suited for modeling the speed distribution at a given traffic density.

The integration of the empirical model into the prior demonstrates that empirical models, derived and calibrated from comprehensive datasets, improve fitting performance over fully data-driven non-parametric regression models only when applied to smaller training datasets. It is noteworthy that the precision of an empirical model does not invariably translate to superior performance in fitting empirical data. The findings suggest that a moderately sized training dataset of relatively clean data suffices for the pure GPR model to represent the stochastic relationship between speed and density accurately.

Looking ahead, future investigations could incorporate methodologies from information theory and statistics to enhance the selection of the input-inducing variables that convey greater informational content about the data. Instead of solely encoding empirical knowledge into the mean function, as explored in this study, it may prove beneficial to embed such knowledge within the kernel function, potentially improving the precision and robustness of non-parametric regression models.

Model source code

The source code of the proposed model is available for readers to download via the following link: https://github.com/YuanzhengLei/Unraveling-stochastic-fundamental-diagrams-considering-empirical-knowledge.
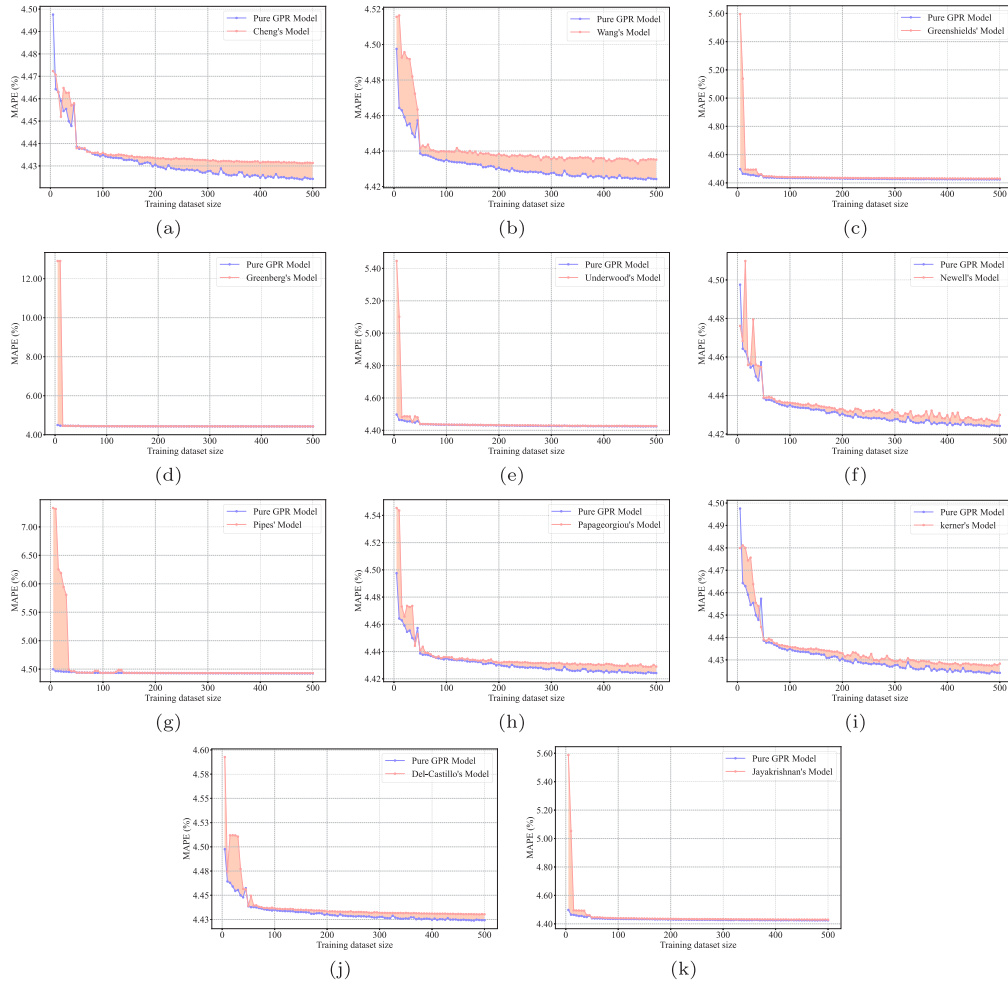
Fig. 8. MAPE comparison between pure GPR and EPGPR model under different training dataset size.

CRediT authorship contribution statement

Yuan-Zheng Lei: Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Yaobang Gong:** Writing – review & editing, Writing – original draft, Validation, Supervision. **Xianfeng Terry Yang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Data availability

I have shared the link in the manuscript and more details are in the supplemental material.

Acknowledgments

Appendix A

Lemma 1.

$$q(\mathbf{f}_*) \approx p(\mathbf{f}_*|\mathbf{y}), q(\mathbf{f}_*) = \int p(\mathbf{f}_*|\mathbf{f}_m)\phi(\mathbf{f}_m)d\mathbf{f}_m \tag{39}$$

Table 7
Parameters of single-regime traffic flow model after calibration[a].

Model	Parameters
Greenshields et al. (1935)	~~$v_f = 18.22$ $\rho_j = -111232.99$~~
Greenberg (1959)	~~$v_{critical} = -29357.73$ $\rho_j = -29083.32$~~
Newell (1961)	$v_f = 44.66$ $\rho_j = 107.05$ $\lambda = 3063.25$
Underwood (1961)	~~$v_f = 49.35$ $\rho_{critical} = -29083.32$~~
Drake et al. (1965)	$v_f = 49.35$ $\rho_{critical} = 214.11$
Papageorgiou et al. (1989)	$v_f = 47.92$ $\rho_j = 102.13$ $\alpha = 1.09$
Kerner and Konhäuser (1994)	$v_f = 41.34$ $\rho_{critical} = 391.92$
Del Castillo and Benitez (1995)	$v_f = 44.66$ $\rho_j = 485.62$ $v_j = 6.31$
Jayakrishnan et al. (1995)	$v_f = 33.72$ $v_{min} = 37.35$ $\rho_j = 27.30$
Ardekani and Ghandehari (2008)	$v_{critical} = 46.64$ $\rho_j = 52.24$ $\rho_{min} = 0.38$
MacNicholas (2011)	~~$v_f = 45.22$ $\rho_j = 10544.64$ $n = 1.95$ $m = 12412.20$~~
Pipes (1966)	$v_f = 46.89$ $\rho_j = 132.87$ $n = 0.69$
Wang et al. (2011)	$v_f = 42.72$ $v_{critical} = 6.64$ $\rho_{critical} = 40.97$ $\theta_1 = 3.50$ $\theta_2 = 0.08$
Cheng et al. (2021)	$v_f = 44.42$ $\rho_{critical} = 80.09$ $m = 2.21$

[a] Parameters are calibrated based on I-80 detector data (Station id:1).

Table 8
Comparison of the result of different EPNR models with Exponential kernel function (I-80).

EPNR model	Speed RMSE (mph)				Speed MAPE			
	$\mathcal{RS}$	$\mathcal{SS}$	$\mathcal{CS}$	$\mathcal{WRS}$	$\mathcal{RS}$	$\mathcal{SS}$	$\mathcal{CS}$	$\mathcal{WRS}$
Pure GP	5.27	5.27	5.26	5.27	12.74%	12.74%	12.73%	12.74%
Cheng's	5.27	5.27	5.27	5.27	12.76%	12.76%	12.75%	12.75%
Wang's	5.27	5.27	5.27	5.27	12.75%	12.75%	12.75%	12.75%
Newell's	5.27	5.27	5.27	5.27	12.76%	12.76%	12.74%	12.75%
Papageorgiou's	5.27	5.26	5.27	5.27	12.76%	12.76%	12.75%	12.75%
Kerner's	5.27	5.27	5.27	5.27	12.75%	12.75%	12.75%	12.75%
Del-Castillo's	5.27	5.27	5.27	5.27	12.76%	12.75%	12.75%	12.75%
Jayakrishnan's	5.27	5.27	5.27	5.27	12.76%	12.76%	12.75%	12.75%

Table 9
Comparison of the result of percentages of points falling within the 95% confidence interval.

EPNR model	PWCI			
	$\mathcal{RS}$	$\mathcal{SS}$	$\mathcal{CS}$	$\mathcal{WRS}$
Pure GP	95.79%	95.78%	95.77%	95.80%
Cheng's	95.80%	95.80%	95.78%	95.80%
Wang's	95.79%	95.79%	95.79%	95.79%
Newell's	95.79%	95.79%	95.78%	95.80%
Papageorgiou's	95.78%	86.55%	95.78%	95.79%
Kerner's	95.79%	95.79%	95.80%	95.79%
Del-Castillo's	95.79%	95.80%	95.79%	95.80%
Jayakrishnan's	95.78%	95.78%	95.77%	95.78%

Proof.

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{f})p(\mathbf{f}|\mathbf{y})d\mathbf{f}$$

$$= \int p(\mathbf{f}_*|\mathbf{f}_m, \mathbf{f})p(\mathbf{f}|\mathbf{f}_m, \mathbf{y})p(\mathbf{f}_m|\mathbf{y})d\mathbf{f}d\mathbf{f}_m$$

Since $\mathbf{f}_m$ is independent from $\mathbf{f}$ and $\mathbf{y} = \mathbf{f} + \sigma$, so $\mathbf{f}_m$ is also independent from $\mathbf{y}$, thus, $p(\mathbf{f}_*|\mathbf{f}_m) = p(\mathbf{f}_*|\mathbf{f}_m, \mathbf{f})$, $p(\mathbf{f}|\mathbf{f}_m) = p(\mathbf{f}|\mathbf{f}_m, \mathbf{y})$, thus, we can have:

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{f}_m)p(\mathbf{f}|\mathbf{f}_m)p(\mathbf{f}_m|\mathbf{y})d\mathbf{f}d\mathbf{f}_m$$

And since $\int p(\mathbf{f}|\mathbf{f}_m)d\mathbf{f} = 1$, $\phi(\mathbf{f}_m) = p(\mathbf{f}_m|\mathbf{y})$, thus:

$$p(\mathbf{f}_*|\mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{f}_m)p(\mathbf{f}_m|\mathbf{y})d\mathbf{f}_m$$

$$= \int p(\mathbf{f}_*|\mathbf{f}_m)\phi(\mathbf{f}_m)d\mathbf{f}_m$$

$$= q(\mathbf{f}_*) \quad \square$$

Lemma 2. *Minimize $KL(q(\mathbf{f}, \mathbf{f}_m)\|p(\mathbf{f}, \mathbf{f}_m|\mathbf{y}))$ is equivalent to maximize:*

$$F_V(\mathbf{x}_m, \phi) = \int p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m) \log \frac{p(\mathbf{f}|\mathbf{y})p(\mathbf{f}_m)}{\phi(\mathbf{f}_m)} d\mathbf{f} d\mathbf{f}_m \tag{40}$$

Proof.

$$
\begin{aligned}
KL(q(\mathbf{f}, \mathbf{f}_m)\|p(\mathbf{f}, \mathbf{f}_m|\mathbf{y})) &= \mathbb{E}_{q_{\phi(\mathbf{f}, \mathbf{f}_m)}}[\log \frac{q_{\phi(\mathbf{f}, \mathbf{f}_m)}}{p(\mathbf{f}, \mathbf{f}_m|\mathbf{y})}] \\
&= \mathbb{E}_{q_{\phi(\mathbf{f}, \mathbf{f}_m)}}[\log \frac{q_{\phi(\mathbf{f}, \mathbf{f}_m)}p(\mathbf{y})}{p(\mathbf{f}, \mathbf{f}_m|\mathbf{y})p(\mathbf{y})}] \\
&= \mathbb{E}_{q_{\phi(\mathbf{f}, \mathbf{f}_m)}}[\log \frac{q_{\phi(\mathbf{f}, \mathbf{f}_m)}p(\mathbf{y})}{p(\mathbf{f}, \mathbf{f}_m, \mathbf{y})}] \\
&= \log p(\mathbf{y}) + \mathbb{E}_{q_{\phi(\mathbf{f}, \mathbf{f}_m)}}[\log \frac{q_{\phi(\mathbf{f}, \mathbf{f}_m)}}{p(\mathbf{f}, \mathbf{f}_m, \mathbf{y})}] \\
&= \log p(\mathbf{y}) - \mathbb{E}_{q_{\phi(\mathbf{f}, \mathbf{f}_m)}}[\log \frac{p(\mathbf{f}, \mathbf{f}_m, \mathbf{y})}{q_{\phi(\mathbf{f}, \mathbf{f}_m)}}] \\
&= \log p(\mathbf{y}) - \int q(\mathbf{f}, \mathbf{f}_m) \log \left(\frac{p(\mathbf{f}, \mathbf{f}_m, \mathbf{y})}{q(\mathbf{f}, \mathbf{f}_m)} \right) d\mathbf{f} \, d\mathbf{f}_m \\
&= \log p(\mathbf{y}) - \int p(\mathbf{f}|\mathbf{f}_m)p(\mathbf{f}_m) \log \left(\frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{f}_m)p(\mathbf{f}_m)}{q(\mathbf{f}, \mathbf{f}_m)} \right) d\mathbf{f} \, d\mathbf{f}_m
\end{aligned}
$$

From Eqs. (20) and (21), $q(\mathbf{f}_*)$ is an approximation to $p(\mathbf{f}_m|\mathbf{y})$, i.e. $q(\mathbf{f}_*) = p(\mathbf{f}_*|\mathbf{y})$, also, $\phi(\mathbf{f}_m) = p(\mathbf{f}_m|\mathbf{y})$, therefore, we know that $q(\mathbf{f}, \mathbf{f}_m) = p(\mathbf{f}, \mathbf{f}_m|\mathbf{y}) = p(\mathbf{f}|\mathbf{f}_m)p(\mathbf{f}_m|\mathbf{y}) = p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m)$, therefore:

$$
\begin{aligned}
KL(q(\mathbf{f}, \mathbf{f}_m)\|p(\mathbf{f}, \mathbf{f}_m|\mathbf{y})) &= \log p(\mathbf{y}) - \int p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m) \log \left(\frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{f}_m)p(\mathbf{f}_m)}{p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m)} \right) d\mathbf{f} \, d\mathbf{f}_m \\
&= \log p(\mathbf{y}) - \int p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m) \log \frac{p(\mathbf{f}|\mathbf{y})p(\mathbf{f}_m)}{\phi(\mathbf{f}_m)} d\mathbf{f} d\mathbf{f}_m
\end{aligned}
$$

And since KL divergence ≥ 0, the lemma proved. $\square$

Lemma 3. *The variation lower bound after taking the optimal choice of the variation distribution ϕ_* is:*

$$F_V(\mathbf{x}_m, \phi_*) = \log[\mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{Q}_{nn})] - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn}) \tag{41}$$

where $\mathbf{Q}_{nn} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$.

Proof.

$$
\begin{aligned}
F_V(\mathbf{x}_m, \phi) &= \int p(\mathbf{f}|\mathbf{f}_m)\phi(\mathbf{f}_m) \log \frac{p(\mathbf{f}|\mathbf{y})p(\mathbf{f}_m)}{\phi(\mathbf{f}_m)} d\mathbf{f} d\mathbf{f}_m \\
&= \int \phi(\mathbf{f}_m) \left\{ \int p(\mathbf{f}|\mathbf{f}_m) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{f} + \log \frac{p(\mathbf{f}_m)}{\phi(\mathbf{f}_m)} \right\} d\mathbf{f}_m
\end{aligned}
$$

The integral in the brace has two parts, only the first part involving $\mathbf{f}$, which can be computed as:

$$\log G(\mathbf{f}_m, \mathbf{y}) = \int p(\mathbf{f}|\mathbf{f}_m) \log p(\mathbf{y}|\mathbf{f}) d\mathbf{f} \tag{42}$$

Since $\mathbf{y} = \mathbf{f} + \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, therefore, $p(\mathbf{y}|\mathbf{f}) \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$. $\Rightarrow$

$$
\begin{aligned}
\log G(\mathbf{f}_m, \mathbf{y}) &= \int p(\mathbf{f}|\mathbf{f}_m) \left\{ -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) \right\} d\mathbf{f} \\
&= \int p(\mathbf{f}|\mathbf{f}_m) \left\{ -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{f} + \mathbf{f}^T\mathbf{f}) \right\} d\mathbf{f} \\
&= \int p(\mathbf{f}|\mathbf{f}_m) \left\{ -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}Tr(\mathbf{y}\mathbf{y}^T - 2\mathbf{y}\mathbf{f}^T + \mathbf{f}\mathbf{f}^T) \right\} d\mathbf{f}
\end{aligned}
$$

Since $\int p(\mathbf{f}|\mathbf{f}_m)d\mathbf{f} = 1$, and $\mathbf{y}\mathbf{y}^T$ does not involve $\mathbf{f}$, therefore, let us focus on the last two integrals.

$$-\int p(\mathbf{f}|\mathbf{f}_m)2\mathbf{y}\mathbf{f}^T d\mathbf{f} = -\mathbb{E}[2\mathbf{y}\mathbf{f}|\mathbf{f}_m] = -2\mathbf{y}\mathbb{E}[\mathbf{f}|\mathbf{f}_m] \tag{43}$$

$$\int p(\mathbf{f}|\mathbf{f}_m)\mathbf{f}\mathbf{f}^T d\mathbf{f} = \mathbb{E}[\mathbf{f}\mathbf{f}^T|\mathbf{f}_m] = Cov(\mathbf{f}|\mathbf{f}_m) + \mathbb{E}[\mathbf{f}|\mathbf{f}_m]\mathbb{E}[\mathbf{f}|\mathbf{f}_m]^T \tag{44}$$

From Williams and Rasmussen (2006), we know that, given:

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{bmatrix} \right) \tag{45}$$

The conditional expectation and covariance are calculated by:

$$\mathbb{E}[\mathbf{X}|\mathbf{Y}] = \mu_{\mathbf{X}} + \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-1}(\mathbf{Y} - \mu_{\mathbf{Y}}) \tag{46}$$

$$Cov[\mathbf{X}|\mathbf{Y}] = \Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}} \Sigma_{\mathbf{YY}}^{-1} \Sigma_{\mathbf{YX}} \tag{47}$$

In our case:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_m \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{nn} & \mathbf{K}_{nm} \\ \mathbf{K}_{mn} & \mathbf{K}_{mm} \end{bmatrix} \right) \tag{48}$$

Therefore:

$$-2\mathbf{y}\mathbb{E}[\mathbf{f}|\mathbf{f}_m] = -2\mathbf{y}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m \tag{49}$$

$$\mathbb{E}[\mathbf{f}\mathbf{f}^T|\mathbf{f}_m] = Cov(\mathbf{f}|\mathbf{f}_m) + \mathbb{E}[\mathbf{f}|\mathbf{f}_m]\mathbb{E}[\mathbf{f}|\mathbf{f}_m]^T = \mathbf{K}_{nn} - \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn} + \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m \left(\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m \right)^T \tag{50}$$

Therefore, $\log G(\mathbf{f}_m, \mathbf{y})$ can be written as:

$$\log G(\mathbf{f}_m, \mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}Tr(\mathbf{y}\mathbf{y}^T - 2\mathbf{y}\boldsymbol{\alpha}^T + \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

$$= \log[\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2\mathbf{I})] - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

where $\boldsymbol{\alpha} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m$, $\mathbf{Q}_{nn} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$. Therefore, $F_V(\mathbf{x}_m, \phi)$ can be written as:

$$F_V(\mathbf{x}_m, \phi) = \int \phi(\mathbf{f}_m) \log \frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2\mathbf{I})}{\phi(\mathbf{f}_m)} d\mathbf{f}_m - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn}) \tag{51}$$

Based on Jensen's inequality (Jensen, 1906), $\varphi(\mathbb{E}[\mathbf{X}]) = \mathbb{E}[\varphi(\mathbf{X})]$ (φ is a convex function), taking the log function outside the above integral, we can have:

$$F_V(\mathbf{x}_m, \phi) = \log \int \phi(\mathbf{f}_m) \frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2\mathbf{I})}{\phi(\mathbf{f}_m)} d\mathbf{f}_m - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

$$= \log \int \mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2\mathbf{I}) d\mathbf{f}_m - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

$$= \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{Q}_{nn}) - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn}) \quad \square$$

Lemma 4.

$$\phi(\mathbf{f}_m) = \mathcal{N}(\mathbf{f}_m | \boldsymbol{\mu}_m, \mathbf{A}_m) \tag{52}$$

where:

$$\boldsymbol{\mu}_m = \frac{1}{\sigma^2}\mathbf{K}_{mm}\boldsymbol{\Sigma}\mathbf{K}_{mn}\mathbf{y} \tag{53}$$

$$\mathbf{A}_m = \mathbf{K}_{mm}\boldsymbol{\Sigma}\mathbf{K}_{mm} \tag{54}$$

$$\boldsymbol{\Sigma} = (\mathbf{K}_{mm} + \sigma^{-2}\mathbf{K}_{mn}\mathbf{K}_{nm})^{-1} \tag{55}$$

Proof. Based on Lemma 2 and Titsias (2009b), achieving the optimal bound occurs when equality holds in Jensen's inequality. i.e. :

$$\phi^*(\mathbf{f}_m) \propto \mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) p(\mathbf{f}_m)$$

$$\propto p(\mathbf{f}|\mathbf{f}_m) p(\mathbf{f}_m)$$

$$\propto \exp(\log p(\mathbf{f}|\mathbf{f}_m) + \log p(\mathbf{f}_m))$$

Since $p(\mathbf{f}|\mathbf{f}_m) = \mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$, $p(\mathbf{f}_m) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_{mm})$, $\phi^*(\mathbf{f}_m)$ can be further written as:

$$\phi^*(\mathbf{f}_m) = \exp \left\{ -\log \left((2\pi)^{\frac{n}{2}} |\sigma^2|^{\frac{1}{2}} \right) - \frac{1}{2}(\mathbf{y} - \boldsymbol{\alpha})^T \frac{1}{\sigma^2}(\mathbf{y} - \boldsymbol{\alpha}) - \log \left((2\pi)^{\frac{n}{2}} |\mathbf{K}_{mm}|^{\frac{1}{2}} \right) - \frac{1}{2}\mathbf{f}_m^T \mathbf{K}_{mm}^{-1} \mathbf{f}_m \right\}$$

$$= c \exp \left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m)^T(\mathbf{y} - \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m) - \frac{1}{2}\mathbf{f}_m^T \mathbf{K}_{mm}^{-1} \mathbf{f}_m \right\}$$

$$= c \exp\left\{ -\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y} - \frac{1}{2}\mathbf{f}_m^T(\mathbf{K}_{mm}^{-1} + \frac{1}{\sigma^2}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1})\mathbf{f}_m + \frac{1}{\sigma^2}\mathbf{y}^T\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{f}_m \right\}$$

$$= \mathcal{N}(\mathbf{f}_m | \left(\mathbf{K}_{mm}^{-1} + \frac{1}{\sigma^2}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\right)^{-1}\frac{1}{\sigma^2}\mathbf{K}_{mn}\mathbf{y}, \left(\mathbf{K}_{mm}^{-1} + \frac{1}{\sigma^2}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\right)^{-1})$$

Based on the Sherman–Morrison formula (Sherman and Morrison, 1950), we can have:

$$\left(\mathbf{K}_{mm}^{-1} + \frac{1}{\sigma^2}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\right)^{-1} = \mathbf{K}_{mm}\left(\mathbf{K}_{mm} + \frac{1}{\sigma^2}\mathbf{K}_{mn}\mathbf{K}_{nm}\right)^{-1}\mathbf{K}_{mm} \tag{56}$$

Therefore:

$$\phi^*(\mathbf{f}_m) = \mathcal{N}(\mathbf{f}_m | \sigma^{-2}\mathbf{K}_{mm}\mathbf{\Sigma}\mathbf{K}_{mn}\mathbf{y}, \mathbf{K}_{mm}\mathbf{\Sigma}\mathbf{K}_{mm})$$

where $\mathbf{\Sigma} = (\mathbf{K}_{mm} + \sigma^{-2}\mathbf{K}_{mn}\mathbf{K}_{nm})^{-1}$, c is a constant. □

Lemma 5. *When the prior of $\mathbf{f}$ and $\mathbf{f}_m$ is not zero, then:*

$$\phi(\mathbf{f}_m) = \mathcal{N}(\mathbf{f}_m | \boldsymbol{\mu}_*, \mathbf{\Sigma}_*) \tag{57}$$

$$\mathbf{\Sigma}_* = (\mathbf{K}_{mm}^{-1} + \frac{1}{2\sigma^2}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1})^{-1} \tag{58}$$

$$\boldsymbol{\mu}_* = \mathbf{\Sigma}_* \left(\frac{1}{\sigma^2}\mathbf{K}_{mm}^{-1}\mathbf{K}_{nm}(\mathbf{y} - m(\mathbf{x}))\right) \tag{59}$$

Proof. Since the prior of $\mathbf{f}$ and $\mathbf{f}_m$ is not zero, then the joint distribution changed from:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_m \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{nn} & \mathbf{K}_{nm} \\ \mathbf{K}_{mn} & \mathbf{K}_{mm} \end{bmatrix}\right) \tag{60}$$

to:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_m \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{x}) \\ m(\mathbf{x}_m) \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{nn} & \mathbf{K}_{nm} \\ \mathbf{K}_{mn} & \mathbf{K}_{mm} \end{bmatrix}\right) \tag{61}$$

Then Eqs. (49) and (50) will be written as:

$$-2\mathbf{y}\mathbb{E}[\mathbf{f}|\mathbf{f}_m] = -2\mathbf{y}\left\{m(\mathbf{x}) + \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}(\mathbf{f}_m - m(\mathbf{x}_m))\right\} \tag{62}$$

$$\mathbb{E}[\mathbf{f}\mathbf{f}^T|\mathbf{f}_m] = \mathbf{K}_{nn} - \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn} + \left\{m(\mathbf{x}) + \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}(\mathbf{f}_m - m(\mathbf{x}_m))\right\}\left\{m(\mathbf{x}) + \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}(\mathbf{f}_m - m(\mathbf{x}_m))\right\}^T \tag{63}$$

Therefore, $\log G(\mathbf{f}_m, \mathbf{y})$ can be written as:

$$\log G(\mathbf{f}_m, \mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}Tr(\mathbf{y}\mathbf{y}^T - 2\mathbf{y}\boldsymbol{\alpha}^T + \boldsymbol{\alpha}\boldsymbol{\alpha}^T + \mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

$$= \log[\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2\mathbf{I})] - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

where $\boldsymbol{\alpha} = m(\mathbf{x}) + \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}(\mathbf{f}_m - m(\mathbf{x}_m))$, $\mathbf{Q}_{nn} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$. Therefore, $F_V(\mathbf{x}_m, \phi)$ can be written as:

$$F_V(\mathbf{x}_m, \phi) = \int \phi(\mathbf{f}_m)\log\frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2\mathbf{I})}{\phi(\mathbf{f}_m)}d\mathbf{f}_m - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn}) \tag{64}$$

Similarly, based on Jensen's inequality, taking the log function outside the above integral, we can have:

$$F_V(\mathbf{x}_m, \phi) = \log\int \phi(\mathbf{f}_m)\frac{\mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2\mathbf{I})}{\phi(\mathbf{f}_m)}d\mathbf{f}_m - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

$$= \log\int \mathcal{N}(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2\mathbf{I})d\mathbf{f}_m - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

$$= \log\mathcal{N}(\mathbf{y}|\mathbf{0}, \sigma^2\mathbf{I} + \mathbf{Q}_{nn}) - \frac{1}{2\sigma^2}Tr(\mathbf{K}_{nn} - \mathbf{Q}_{nn})$$

which has the same structure as when the prior is zero, but $\boldsymbol{\alpha}$ is different. Also, based on Lemma 4, we can recalculate $\phi(\mathbf{f}_m)$ as:

$$\phi^*(\mathbf{f}_m) = \exp\left\{-\log\left((2\pi)^{\frac{n}{2}}|\sigma^2|^{\frac{1}{2}}\right) - \frac{1}{2}(\mathbf{y} - \boldsymbol{\alpha})^T\frac{1}{\sigma^2}(\mathbf{y} - \boldsymbol{\alpha}) - \log\left((2\pi)^{\frac{n}{2}}|\mathbf{K}_{mm}|^{\frac{1}{2}}\right) - \frac{1}{2}(\mathbf{f}_m - m(\mathbf{x}_m))^T\mathbf{K}_{mm}^{-1}(\mathbf{f}_m - m(\mathbf{x}_m))\right\}$$

$$= c\exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\alpha})^T\frac{1}{\sigma^2}(\mathbf{y} - \boldsymbol{\alpha}) - \frac{1}{2}(\mathbf{f}_m - m(\mathbf{x}_m))^T\mathbf{K}_{mm}^{-1}(\mathbf{f}_m - m(\mathbf{x}_m))\right\}$$

$$= c\exp\{\frac{1}{2\sigma^2}(\mathbf{y} - m(\mathbf{x}))^T(\mathbf{y} - m(\mathbf{x})) - \frac{1}{\sigma^2}(\mathbf{y} - m(\mathbf{x}))^T\mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}(\mathbf{f}_m - m(\mathbf{x}_m))$$

$$- \frac{1}{2}(\mathbf{f}_m - m(\mathbf{x}_m))^T \left(\mathbf{K}_{mm}^{-1} + \frac{1}{\sigma^2} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mn} \mathbf{K}_{nm} \mathbf{K}_{mm}^{-1} \right) (\mathbf{f}_m - m(\mathbf{x}_m)) \}$$

$$= \mathcal{N}(\mathbf{f}_m | \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$$

where $\boldsymbol{\Sigma}_* = \left(\mathbf{K}_{mm}^{-1} + \frac{1}{\sigma^2} \mathbf{K}_{mm}^{-1} \mathbf{K}_{mn} \mathbf{K}_{nm} \mathbf{K}_{mm} \right)^{-1}$, $\boldsymbol{\mu}_* = m(\mathbf{x}_m) + \boldsymbol{\Sigma}_* \left(\frac{1}{\sigma^2} \mathbf{K}_{mm}^{-1} \mathbf{K}_{nm}(\mathbf{y} - m(\mathbf{x})) \right)$, c is a constant. $\square$

Lemma 6. *When the prior of $\mathbf{f}_*$ and $\mathbf{f}_m$ is not zero, then:*

$$q(\mathbf{f}_*) = \mathcal{N}(m(\mathbf{x}_*)) + \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1}(\boldsymbol{\mu}_* - m(\mathbf{x}_m)), \mathbf{K}_{**} - \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1} \mathbf{K}_{m*} + \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1} \boldsymbol{\mu}_* \mathbf{K}_{mm}^{-1} \mathbf{K}_{m*}) \tag{65}$$

Proof. Since the joint distribution is now written as:

$$\begin{bmatrix} \mathbf{f}_* \\ \mathbf{f}_m \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m(\mathbf{x}_*) \\ m(\mathbf{x}_m) \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{**} & \mathbf{K}_{*m} \\ \mathbf{K}_{m*} & \mathbf{K}_{mm} \end{bmatrix} \right) \tag{66}$$

Based on conditional expectation and covariance formulation shown in Eqs. (47) and (48), and $q(\mathbf{f}_*) = \int p(\mathbf{f}_* | \mathbf{f}_m) \phi(\mathbf{f}_m) d\mathbf{f}_m$

$$\mathbf{f}_* | \mathbf{f}_m \sim \mathcal{N}(m(\mathbf{x}_*)) + \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1}(\boldsymbol{\mu}_* - m(\mathbf{x}_m)), \mathbf{K}_{**} - \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1} \mathbf{K}_{m*}) \tag{67}$$

$$q(\mathbf{f}_*) = \mathcal{N}(m(\mathbf{x}_*)) + \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1}(\boldsymbol{\mu}_* - m(\mathbf{x}_m)), \mathbf{K}_{**} - \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1} \mathbf{K}_{m*} + \mathbf{K}_{*m} \mathbf{K}_{mm}^{-1} \boldsymbol{\mu}_* \mathbf{K}_{mm}^{-1} \mathbf{K}_{m*}) \quad \square \tag{68}$$

Definition 2 (*Simple Random Sampling*). Simple random sampling is the basic sampling technique where we select a group of subjects for study from a larger group (Easton and McColl, 1997). Each individual is chosen entirely by chance, and each member of the population has an equal chance of being included in the sample. Every possible sample of a given size has the same chance of selection.

The simple random sampling without replacement that is adopted to do the comparison study is the famous reservoir sampling (Vitter, 1985) shown as follows:

Algorithm 1: Reservoir sampling

    Input: Input: $\mathcal{P}[1,2,..,n], S[1,2,...,k]$
1 **for** i $\in$ **Range** (1 **to** k+1) **do**
2 | $S[i] = \mathcal{P}[i]$
3 **for** i $\in$ **Range** (k+1 **to** n+1) **do**
4 | $j = \text{RandomInteger}(1, i)$
5 | **if** $j \leq k$ **then**
6 | | $S[j] = \mathcal{P}[i]$

Definition 3 (*Systematic Sampling*). Systematic sampling is a statistical method involving the selection of elements from an ordered sampling frame. The most common form of systematic sampling is an equiprobability method, which is shown as follows: where

---

Algorithm 2: Systematic sampling

    Input: Input: $\mathcal{P}[1,2,..,N], S = \varnothing, n$
1 $k = \frac{N}{n}$
2 Randomly choose an index I, $I \in [[1,k]]$
3 $S[1] = \mathcal{P}[I]$
4 **for** i $\in$ **Range** (2 **to** k+1) **do**
5 | $\mathcal{R}[i] = S[\lceil I + k * (i - 1) \rceil]$

N is the population size, n is the sample size. To avoid the risk that the last house chosen does not exist ($I + k * (n - 1) > N$), the initial index I must be located within $[1, k]$ and should be rounded up to the next integer if it is not an integer.

Definition 4 (*Cluster Sampling*). Cluster sampling is a method used when mutually homogeneous yet internally heterogeneous groupings are evident in a statistical population. Based on k-mean algorithm (MacQueen et al., 1967), the cluster sampling can be expressed as:

Definition 5 (*Weighted Random Sampling*). This sampling method is motivated by Qu et al. (2015), which aims to avoid selecting high percentage samples with a relatively low traffic density. The core idea is that first sort the $S = \{(\rho_1, v_1), (\rho_2, v_2), \ldots, (\rho_n, v_n)\}$ with a density ascending order. And then compute the weight w for each density–speed pair based on Qu et al. (2015). Then for each density–speed pair (ρ_i, v_i), it will has a probability $\frac{w_i}{\sum_{j=1}^{n} w_j}$ to be selected.

Algorithm 3: Cluster sampling algorithm

Input: Input: $\mathcal{P} = \{x_1, x_2, ..., x_n\}, C_i = \varnothing, i = 1, 2, ..., k, S = \varnothing, \mathbf{M}$

1   Randomly selecting k random observations $m_i^{(1)}, i = 1, 2, ..., k$ from $\mathcal{P}$

2   **for** t **Range**(1, **M**+1) **do**

3   **for** x_p **in** $\mathcal{P}$ **do**

4   **for** i = **Range**(1,k+1) **do**

5   **if** $\|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \ \forall 1 \leq j \leq k$ **then**

6   Add x_p to $C_i^{(t)}$ and remove x_p from $\mathcal{P}$

7   **for** i = **Range**(1,k+1) **do**

8         $m_i^{(t+1)} = \frac{1}{|C_i^t|} \sum_{x \in C_i^t} x$

9   **for** i = **Range**(1,k+1) **do**

10      $S[i] = x^*, \{x^*|\mathbf{min}_{x^*} : \|x - m_i^{\mathbf{M}}\|^2, \forall x \in C_i^{\mathbf{M}}\}$

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.trc.2024.104851.

References

Ahmed, A., Ngoduy, D., Adnan, M., Baig, M.A.U., 2021. On the fundamental diagram and driving behavior modeling of heterogeneous traffic flow using UAV-based data. Transp. Res. A 148, 100–115.

Alperovich, T., Sopasakis, A., 2008. Stochastic description of traffic flow. J. Stat. Phys. 133, 1083–1105.

Ardekani, S., Ghandehari, M., 2008. A Modified Greenberg Speed-Flow Traffic Model. Technical Report, University of Texas at Arlington.

Bai, L., Wong, S., Xu, P., Chow, A.H., Lam, W.H., 2021. Calibration of stochastic link-based fundamental diagram with explicit consideration of speed heterogeneity. Transp. Res. B 150, 524–539.

Bishop, C., 2006. Pattern recognition and machine learning. In: Springer Google Schola. Vol. 2, Springer, pp. 5–43.

Cheng, Q., Lin, Y., Zhou, X.S., Liu, Z., 2024. Analytical formulation for explaining the variations in traffic states: A fundamental diagram modeling perspective with stochastic parameters. European J. Oper. Res. 312 (1), 182–197.

Cheng, Q., Liu, Z., Lin, Y., Zhou, X.S., 2021. An s-shaped three-parameter (S3) traffic stream model with consistent car following relationship. Transp. Res. B 153, 246–271.

Csató, L., Opper, M., 2002. Sparse on-line Gaussian processes. Neural Comput. 14 (3), 641–668.

Del Castillo, J., Benitez, F., 1995. On the functional form of the speed-density relationship—II: Empirical investigation. Transp. Res. B 29 (5), 391–406.

Drake, J., Schofer, J., May, A., 1965. A statistical analysis of speed-density hypotheses. Traffic Flow Transp..

Drew, D.R., 1968. Traffic Flow Theory and Control. Technical Report.

Easton, V.J., McColl, J.H., 1997. Statistics glossary v1. 1.

Edie, L.C., 1961. Car-following and steady-state theory for noncongested traffic. Oper. Res. 9 (1), 66–76.

Fan, S., Seibold, B., 2013. Data-fitted first-order traffic models and their second-order generalizations: Comparison by trajectory and sensor data. Transp. Res. Rec. 2391 (1), 32–43.

Greenberg, H., 1959. An analysis of traffic flow. Oper. Res. 7 (1), 79–85.

Greenshields, B.D., Bibbins, J., Channing, W., Miller, H., 1935. A study of traffic capacity. In: Highway Research Board Proceedings. Vol. 14, Washington, DC, pp. 448–477.

Gupta, N., Patil, G.R., Vu, H.L., 2023. Simple abstract models to study stability of urban networks with decentralized signal control. Transp. Res. B 172, 93–116.

Jabari, S.E., Zheng, J., Liu, H.X., 2014. A probabilistic stationary speed–density relation based on Newell's simplified car-following model. Transp. Res. B 68, 205–223.

Jäkel, F., Schölkopf, B., Wichmann, F.A., 2007. A tutorial on kernel methods for categorization. J. Math. Psych. 51 (6), 343–358.

Jayakrishnan, R., Tsai, W.K., Chen, A., 1995. A dynamic traffic assignment model with traffic-flow relationships. Transp. Res. C 3 (1), 51–72.

Jensen, J.L.W.V., 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Math. 30 (1), 175–193.

Kerner, B.S., 1998. Experimental features of self-organization in traffic flow. Phys. Rev. Lett. 81 (17), 3797.

Kerner, B.S., Konhäuser, P., 1994. Structure and parameters of clusters in traffic flow. Phys. Rev. E 50 (1), 54.

Krasser, M., 2020. Gaussian processes: Sparse approximations. URL: https://krasserm.github.io/2020/12/12/gaussian-processes-sparse/.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86.

Laval, J.A., Chilukuri, B.R., 2014. The distribution of congestion on a class of stochastic kinematic wave models. Transp. Sci. 48 (2), 217–224.

Lawrence, N., Seeger, M., Herbrich, R., 2002. Fast sparse Gaussian process methods: The informative vector machine. Adv. Neural Inf. Process. Syst. 15.

Li, J., 1996. A new complexity bound for the least-squares problem. Comput. Math. Appl. 31 (12), 15–16.

Li, X., Zhang, X., Qian, X., Zhao, C., Guo, Y., Peeta, S., 2024. Beyond centralization: Non-cooperative perimeter control with extended mean-field reinforcement learning in urban road networks. Transp. Res. B 186, 103016.

MacNicholas, M.J., 2011. A simple and pragmatic representation of traffic flow. In: Symposium on the Fundamental Diagram. Vol. 75, pp. 161–177.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1, Oakland, CA, USA, pp. 281–297.

Mahmoud, H.F., 2019. Parametric versus semi and nonparametric regression models. arXiv preprint arXiv:1906.10221.

May, A.D., 1990. Traffic flow fundamentals.

Meng, Q., Qu, X., 2013. Bus dwell time estimation at bus bays: A probabilistic approach. Transp. Res. C 36, 61–71.

Newell, G.F., 1961. Nonlinear effects in the dynamics of car following. Oper. Res. 9 (2), 209–229.

Ni, D., 2015. Traffic Flow Theory: Characteristics, Experimental Methods, and Numerical Techniques. Butterworth-Heinemann.

Ni, D., Hsieh, H.K., Jiang, T., 2018. Modeling phase diagrams as stochastic processes with application in vehicular traffic flow. Appl. Math. Model. 53, 106–117.

Ni, D., Leonard, J.D., Jia, C., Wang, J., 2016. Vehicle longitudinal control and traffic stream modeling. Transp. Sci. 50 (3), 1016–1031.

Papageorgiou, M., Blosseville, J.-M., Hadj-Salem, H., 1989. Macroscopic modelling of traffic flow on the Boulevard Périphérique in Paris. Transp. Res. B 23 (1), 29–47.

Pipes, L.A., 1966. Car following models and the fundamental diagram of road traffic. Transp. Res./UK/.

Qu, X., Wang, S., Zhang, J., 2015. On the fundamental diagram for freeway traffic: A novel calibration approach for single-regime models. Transp. Res. B 73, 91–102.

Qu, X., Zhang, J., Wang, S., 2017. On the stochastic fundamental diagram for freeway traffic: model development, analytical properties, validation, and extensive applications. Transp. Res. B 104, 256–271.

Quinonero-Candela, J., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. 6, 1939–1959.

Schwaighofer, A., Tresp, V., 2002. Transductive and inductive methods for approximate Gaussian process regression. Adv. Neural Inf. Process. Syst. 15.

Seeger, M.W., Williams, C.K., Lawrence, N.D., 2003. Fast forward selection to speed up sparse Gaussian process regression. In: International Workshop on Artificial Intelligence and Statistics. PMLR, pp. 254–261.

Sherman, J., Morrison, W.J., 1950. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. Ann. Math. Stat. 21 (1), 124–127.

Siqueira, A.F., Peixoto, C.J., Wu, C., Qian, W.-L., 2016. Effect of stochastic transition in the fundamental diagram of traffic flow. Transp. Res. B 87, 1–13.

Smola, A., Bartlett, P., 2000. Sparse greedy Gaussian process regression. Adv. Neural Inf. Process. Syst. 13.

Snelson, E., Ghahramani, Z., 2005. Sparse Gaussian processes using pseudo-inputs. Adv. Neural Inf. Process. Syst. 18.

Soyster, A., Wilson, G., 1973. A stochastic model of flow versus concentration applied to traffic on hills. Highw. Res. Board 456, 28–39.

Sun, H., Liu, H.X., Xiao, H., He, R.R., Ran, B., 2003. Use of local linear regression model for short-term traffic forecasting. Transp. Res. Rec. 1836 (1), 143–150.

Sun, L., Pan, Y., Gu, W., 2014. Data mining using regularized adaptive B-splines regression with penalization for multi-regime traffic stream models. J. Adv. Transp. 48 (7), 876–890.

Sun, L., Zhou, J., 2005. Development of multiregime speed–density relationships by cluster analysis. Transp. Res. Rec. 1934 (1), 64–71.

Tiao, L.C., 2020. A handbook for sparse variational Gaussian processes. tiao.io. URL: https://tiao.io/post/sparse-variational-gaussian-processes/.

Titsias, M., 2009a. Variational learning of inducing variables in sparse Gaussian processes. In: Artificial Intelligence and Statistics. PMLR, pp. 567–574.

Titsias, M.K., 2009b. Variational Model Selection for Sparse Gaussian Process Regression. Report, University of Manchester, UK.

Underwood, R., 1961. Speed, volume, and density relationship: Quality and theory of traffic flow, Yale Bureau of highway traffic. pp. 141–188.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al., 2020. SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods 17 (3), 261–272.

Vitter, J.S., 1985. Random sampling with a reservoir. ACM Trans. Math. Softw. 11 (1), 37–57.

Walder, C., Kim, K.I., Schölkopf, B., 2008. Sparse multiscale Gaussian process regression. In: Proceedings of the 25th International Conference on Machine Learning. pp. 1112–1119.

Wang, J., 2023. An intuitive tutorial to Gaussian processes regression. Comput. Sci. Eng..

Wang, S., Chen, X., Qu, X., 2021. Model on empirically calibrating stochastic traffic flow fundamental diagram. Commun. Transp. Res. 1, 100015.

Wang, H., Li, J., Chen, Q.-Y., Ni, D., 2011. Logistic modeling of the equilibrium speed–density relationship. Transp. Res. A 45 (6), 554–566.

Wang, H., Ni, D., Chen, Q.-Y., Li, J., 2013. Stochastic modeling of the equilibrium speed–density relationship.

Williams, C.K., Rasmussen, C.E., 2006. Gaussian Processes for Machine Learning. Vol. 2, MIT Press Cambridge, MA.

Williams, C., Seeger, M., 2000. Using the Nyström method to speed up kernel machines. Adv. Neural Inf. Process. Syst. 13.

Yuan, Y., Wang, Q., Yang, X.T., 2021a. Traffic flow modeling with gradual physics regularized learning. IEEE Trans. Intell. Transp. Syst. 23 (9), 14649–14660.

Yuan, Y., Zhang, Z., Yang, X.T., Zhe, S., 2021b. Macroscopic traffic flow modeling with physics regularized Gaussian process: A new insight into machine learning applications in transportation. Transp. Res. B 146, 88–110.

Zhang, N., Yang, X., Ma, W., et al., 2018. Empirical approximation for the stochastic fundamental diagram of traffic flow on signalized intersection. J. Adv. Transp. 2018.

Zhou, J., Zhu, F., 2020. Modeling the fundamental diagram of mixed human-driven and connected automated vehicles. Transp. Res. C 115, 102614.