Transportation Research Part C 95 (2018) 636–651

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Heteroscedastic Gaussian processes for uncertainty modeling in large-scale crowdsourced traffic data

Filipe Rodrigues[a,*], Francisco C. Pereira[a,b]

[a] Technical University of Denmark (DTU), Bygning 116B, 2800 Kgs. Lyngby, Denmark
[b] Massachusetts Institute of Technology (MIT), 77 Mass. Ave., 02139 Cambridge, MA, USA

ARTICLE INFO

Keywords:
Gaussian processes
Heteroscedastic models
Traffic data
Crowdsourcing
Uncertainty modeling
Forecasting
Imputation
Floating car data

ABSTRACT

Accurately modeling traffic speeds is a fundamental part of efficient intelligent transportation systems. Nowadays, with the widespread deployment of GPS-enabled devices, it has become possible to crowdsource the collection of speed information to road users (e.g. through mobile applications or dedicated in-vehicle devices). Despite its rather wide spatial coverage, crowdsourced speed data also brings very important challenges, such as the highly variable measurement noise in the data due to a variety of driving behaviors and sample sizes. When not properly accounted for, this noise can severely compromise any application that relies on accurate traffic data. In this article, we propose the use of heteroscedastic Gaussian processes (HGP) to model the time-varying uncertainty in large-scale crowdsourced traffic data. Furthermore, we develop a HGP conditioned on sample size and traffic regime (SSRC-HGP), which makes use of sample size information (probe vehicles per minute) as well as previous observed speeds, in order to more accurately model the uncertainty in observed speeds. Using 6 months of crowdsourced traffic data from Copenhagen, we empirically show that the proposed heteroscedastic models produce significantly better predictive distributions when compared to current state-of-the-art methods for both speed imputation and short-term forecasting tasks.

1. Introduction

Modeling traffic speeds is an essential task for developing intelligent transportation systems, because it provides real-time and anticipatory information about the performance of the network. This information is not only essential for traffic managers, since it allows them to properly allocate resources (e.g. control traffic lights) and identify problematic situations, but it also helps users to make better travel decisions by providing them with a complete "picture" of the traffic status throughout the city (e.g., suggest to take an alternative route or delay/advance the departure) (Liu et al., 2013). The role of accurate traffic speed modeling is even more significant when we consider innovative car-sharing, autonomous vehicles and connected vehicles technologies (Tajalli and Hajbabaie, 2018), where inappropriate routing of vehicles and poor system-wide optimization and coordination can have severe adverse effects in the behavior of the road network (e.g., congestion and poor quality of service) and, ultimately, it can be decisive to the adoption of these technologies.

There are two main sources of traffic speed data: static traffic sensors located at fixed location and GPS sensors from floating vehicles. While traditional speed modeling approaches tend to rely solely on static traffic sensors, which are accurate but expensive to

* Corresponding author.
E-mail addresses: rodr@dtu.dk (F. Rodrigues), camara@dtu.dk (F.C. Pereira).
URL: http://www.fprodrigues.com (F. Rodrigues).

https://doi.org/10.1016/j.trc.2018.08.007
Received 16 January 2018; Received in revised form 17 August 2018; Accepted 20 August 2018

deploy and maintain, nowadays, with the development and widespread deployment of GPS-enabled devices, it has become possible to achieve a much better sensing coverage of the entire road network. In fact, the development of crowdsourcing technologies, where individual users contribute with their own GPS data from their mobile devices, further provides a unique potential for obtaining rather accurate, inexpensive and complete measurements of the speed conditions throughout the network. Hence, it is not surprising that this type of traffic data is becoming increasing popular among traffic managers, operators and local authorities, with many of these acquiring traffic data consisting of aggregated speed measurements from probe vehicles from providers such as INRIX[1] or HERE.[2] However, despite its potential, this data also brings many interesting challenges.

A key fundamental challenge for using crowdsourced speed data in practice is accurately modeling the uncertainty associated with it. Since this data typically consists of aggregated speeds based on individual GPS measurements from a heterogeneous fleet of contributing vehicles and devices (probe vehicles/devices), the resultant speed information can be extremely noisy. This can be due to several reasons, such as low number of samples (probe devices), accuracy of the GPS-enabled devices (as studied by Guido et al. (2014)), different drivers' behavior, etc. As a consequence, in some situations, the overall picture of the traffic conditions that this data provides can be significantly blurred, causing applications that rely on it to fail. For example, anomaly detection algorithms can be misled to believe that there is something wrong in a certain road segment, when the problem is simply due to momentarily poor data quality. Similarly, forecasting algorithms can be led to produce erroneous predictions by not accounting for the noise in the speed data when modeling it.

This article proposes the use of heteroscedastic Gaussian processes in order to produce models that account for the non-constant variance of speeds through time. Gaussian processes (GPs) are flexible non-parametric Bayesian models that are widely used for modeling complex time-series. Indeed, GPs have been successfully applied to model and predict with state-of-the-art results various traffic related phenomena such as traffic congestion (Liu et al., 2013), travel times (Idé and Kato, 2009), pedestrian and public transport flows (Neumann et al., 2009; Rodrigues et al., 2016), traffic volumes (Xie et al., 2010), driver velocity profiles (Armand et al., 2013), etc. The fully Bayesian non-parametric formulation of GPs makes them particularly well suited for modeling uncertainty and noise in the observations. Heteroscedastic approaches using GPs to model complex noisy time-series additionally extend the capabilities of GPs to capture the uncertainty in the data by allowing the latter to vary between different time periods. In this article, we take these approaches one step further by proposing a heteroscedastic Gaussian process in which the speed variance is conditioned on the observed sample size, i.e. the number of vehicles/devices per minute (which can also be regarded as a noisy proxy for traffic flow), and on the current traffic regime. The intuition is that the uncertainty associated with a speed observation varies with the number of samples (vehicles) that were used to produce that observation, with more samples producing more accurate speed measurements, as well as the traffic regime. As it turns out, conditioning the observation uncertainty on the number of samples per time interval leads to significantly more accurate predictive distributions.

Using a large-scale dataset of crowdsourced speeds provided by Google for Copenhagen, we consider two major tasks: speed imputation and short-term forecasting. Speed imputation refers to the post hoc problem of predicting the speeds that were not observed due to the absence of sensing devices traveling along a road segment or due to any other data collection issue. On the other hand, short-term forecasting refers to the problem of predicting the speeds in a road network for short periods ahead of time (typically 5–15 min). While short-term forecasting is a fundamental part of any intelligent transportation system, speed imputation can be critical for the success of any application that makes use of that type of data, especially when the missing observation rate is higher, as it is common with crowdsourced data.

By applying the proposed heteroscedastic GP model to crowdsourced speed data, we are effectively able to quantify the uncertainty in the observed speeds and obtain significantly more accurate predictive distributions for traffic speed imputation and short-term forecasting. This, in turn, allows us to produce precise prediction intervals, which are of vital importance for many real-world applications. In fact, the value of accurate prediction intervals is often neglected in the transportation literature. However, for various tasks, it is often more important to be able to estimate prediction intervals than single point estimates. For example, when planning a trip, it is common the case where we need to guarantee with some level of confidence (e.g., 95%) that the users will arrive on time to the destination. Similarly, when modeling travel demand, such as for public transport or autonomous vehicles, it is essential to ensure that the allocated resources are enough to accommodate the demand. Therefore, rather than planning and allocating resources in accordance to the estimated mean demand, it is better to rely on a different quantile of the predictive distribution in order to guarantee quality of service and avoid travelers' dissatisfaction and frustration. However, if the prediction intervals are not accurate enough, we risk either underestimating our uncertainty, causing problems to the users, or overestimating it, thus wasting valuable resources. By providing a heteroscedastic treatment of the speed data, we are able to produce accurate prediction intervals, while also reducing the error of the mean predictions. Furthermore, by using an approximate inference technique based on variational inference, we are also able to scale the proposed approach to relatively large datasets such as the one used in our experiments.

The rest of this article is organized as follows. First, Section 2 reviews related works. Section 3 introduces GPs and discusses how to use them for modeling time-series data. The proposed sample-size-and-regime-conditioned heteroscedastic GP (SSRC-HGP) is presented in Section 4. A thorough experimental evaluation of the proposed methodology in comparison with other state-of-the-art approaches is presented in Section 5. Finally, we conclude in Section 6.

[1] http://inrix.com.

[2] http://here.com.