A Machine Learning Approach to Censored Bike-Sharing Demand Modeling

Gammelli, Daniele; Rodrigues, Filipe; Pacino, Dario; Kurtaran, Haci Ahmet; Pereira, Francisco Camara

Published in:
Transportation Research Board. Annual Meeting Proceedings

Publication date:
2020

Document Version
Peer reviewed version

Citation (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
A Machine Learning Approach to Censored Bike-Sharing Demand Modeling

Daniele Gammelli
PhD Student, DTU Management, Transport DTU
Danmarks Tekniske Universitet, Kgs. Lyngby, Denmark, 2800
Email: daga@dtu.dk

Filipe Rodrigues
Assistant Professor, DTU Management, Transport DTU
Danmarks Tekniske Universitet, Kgs. Lyngby, Denmark, 2800
Email: rodr@dtu.dk

Dario Pacino
Associate Professor, DTU Management, Management Science DTU
Danmarks Tekniske Universitet, Kgs. Lyngby, Denmark, 2800
Email: darpa@dtu.dk

Haci Ahmet Kurtaran
Head of BI and Finances
Donkey Republic
Copenhagen, Denmark
Email: a_kurtaran@hotmail.com

Francisco C. Pereira
Full Professor, DTU Management, Transport DTU
Danmarks Tekniske Universitet, Kgs. Lyngby, Denmark, 2800
Email: camara@dtu.dk

Word Count: 6256 words + 4 tables = 7256 words

Submitted July 26, 2019
ABSTRACT
Transportation demand is highly dependent on supply, especially for shared transport modes where the availability of the service is limited and where this limitation depends on spatio-temporal factors (e.g. consider a bike-sharing service where the availability is determined by the area and time of day in which a user is searching for a bike). More importantly, in the context of building a prediction model for demand, supply implicitly acts as an upper limit to our observed demand (i.e. observed demand will never be higher than supply), thus resulting in historical data representing a biased version of the real – i.e. underlying – demand pattern. Without making explicit these assumptions, any predictive model would, even in the best possible case, end up representing a biased model of demand, being therefore less effective in predicting user needs. To approach this problem, we first propose a general methodology by implementing a censored likelihood able to train supply-aware models. We then apply this methodology to the special case of bike-sharing demand prediction in specific areas of Copenhagen, Denmark, by implementing a Gaussian Process model for the time-series analysis of bike rentals. Experimental results on real-world datasets from Donkey Republic (bike-sharing provider in Copenhagen) show how taking into account the limiting effect of supply on demand is essential in the process of obtaining an unbiased predictive model of user demand behavior.

Keywords: Demand modeling, Censoring, Machine Learning, Gaussian Processes, Bike-Sharing
INTRODUCTION

Being able to understand and predict demand is essential in the planning and decision-making processes of any given transportation service, allowing the service provider to take decisions coherently with user behavior and needs. Having reliable models of demand is especially relevant in shared transport modes (e.g. car-sharing, bike-sharing, etc.), where the high volatility of demand and the flexibility of supply modalities (e.g. potentially infinite different collocations of a car-sharing fleet on the territory) require decisions to be made in very strict resemblance to user needs. If for instance we consider the bike-sharing scenario, the service provider finds itself in a position to deal with a great variety of complex decisions regarding the way with which to satisfy user demand. Just to name a few, concrete choices must be made for what concerns capacity positioning (i.e. where to spatially allocate supply), capacity planning (i.e. dimensioning of the bike fleet size), rebalancing (i.e. reallocate the available capacity during operations) and expansion planning (i.e. if and how to expand the reach of the bike-sharing service).

Demand modeling attempts to use statistical methods to capture the user demand behavior based on recorded historical data. However, historical data in the case of transport services, is usually highly dependent on the supply offered by the provider itself. In particular, supply represents an upper limit to our possibility of observing realizations of the real demand. To give a simple example, suppose we have a bike-sharing service with 10 bikes available in a specific moment, we might find ourselves in the situation of observing a usage (i.e. demand) of 10 bikes despite the fact that the actual demand for bikes might have been potentially higher. This leads to a situation in which our historical data is actually representing a biased, or censored, version of the underlying demand pattern in which we are truly interested. More importantly, using censored data to build demand models will, as a natural consequence, produce a biased estimate of demand and an inaccurate understanding of user needs, which will ultimately result in non-optimal operational decisions for the service provider.

To address these problems we propose a general approach to build models aware of the supply-censoring issue which are ultimately more reliable for the understanding of user behavior. We do so using a Censored likelihood in the training of our models and practically implementing a Gaussian Process model for the analysis of the rental demand of a bike-sharing service. We then apply the proposed models on real-world datasets to confirm their effectiveness in the goal of recovering the real demand of bike usage.

RELATED WORK

In the context of bike-sharing demand forecasting there is a long line of literature where research has mainly focused on obtaining predictions of demand based on historical data in combination with some external input (e.g. weather, temporal information, etc.). Examples of this can be found in (1), where a Random Forest model is used for rental predictions; On the same line, (2) approaches the demand forecasting problem through the use of count models based on Generalized Linear Models (GLMs), in particular Poisson and Negative-Binomial regression models; (3) applies multivariate regression using data gathered from multiple bike-sharing systems rather than just one system; (4) defines a hierarchical prediction model for both rentals and returns by implementing Gradient Boosting Regression Trees at a city-aggregation-level, and subsequently learn rent proportions for clusters of bike-stations through a multi-similarity based inference. The returns are then predicted in a consistent manner with respect to rentals by learning the transition probabilities of rented bikes (i.e. where the bikes will end once they have been rented). A different approach can be found in (5) where the proposed prediction model refers to
the *Origin-Destination (O-D)* trip demand rather than the rental/return demand in specific locations/areas. Practically, the O-D demand is modeled through a log-log regression; (6) instead proposes a deep learning-based approach to model bike in-flow and out-flow for the bike-sharing system areas.

Within this stream of research, the censoring problem discussed above is widely accepted, however, to the best of our knowledge, there has been no extensive study on how observed demand can differ from the real – underlying – demand and on how these differences could impact a predictive model. To assess this issue, common approaches regard cleaning strategies as in (2, 7–9). Here (2, 7, 9) attempt to avoid the bias induced by the censored observations by filtering out the time periods where censoring might have occurred before concretely implementing the respective models. (8) applies a data imputation technique to substitute the censored observations with the mean of the historical (non-censored) observations regarding the same period. A different, but related, approach is proposed by (10, 11) which focus on obtaining an unbiased estimate of rate of arrival process by taking into account only the time in which bikes were available, rather than the total time given by the historical data. These approaches manage, to some degree, to correct the bias characterizing the observed demand. However, they also give relevance to the fact that this problem represents an important gap requiring further study in order to obtain reliable predictions in shared transportation scenarios, and other fields. We believe there are two main reasons to why there is the need for a more structured view on the censoring problem: 1) the approaches described above might not be applicable in many real-world scenarios in which the number of censored observations is very high, leading either to an excessive elimination of useful data or to an inadequate imputation of the censored data; 2) rather than resorting to cleaning procedures as the ones above, it would be desirable to “inject” in the forecasting model some sort of awareness towards the censoring problem, giving the model the possibility to coherently adjust its predictions and at the same time make use of the entire information captured in the available observations.

**METHODOLOGY**
In the following sections we sequentially address the various building blocks in the construction of the proposed censored models. First, we introduce the idea of censored regression in a general context. We then describe the technicalities of Gaussian Processes with a particular focus on the role of kernels. Lastly, the two ideas are consequently put together by defining Censored Gaussian Processes and a specific inference procedure for the proposed model.

**Censored Regression**
In order to address the censoring problem, we will refer to the Tobit model. Initially introduced by James Tobin in (12), this framework is also known as censored regression model in econometrics and has many applications also in the field of survival analysis (13). For each observation $i$, the model assumes that there is a latent (i.e. unobservable) variable $y_i^*$ and an observable censored realization $y_i$. In the shared transport demand prediction setting, the real demand would be represented by $y_i^*$, while the available historical observations by $y_i$. In its simplest form, the Tobit model parametrizes the dependency of the latent variable on the input data $x_i$ through a linear relationship with a vector of parameters $\beta$ and the addition of an error term $\varepsilon_i$ which we assume to be normally distributed with mean 0 and variance $\sigma^2$ (just as in a linear regression model):
\[ y_i^* = \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \]  

(1)

There are multiple variations of this model depending on where and when the censoring arises. The specific model relevant for this study is known as Type I model, and deals with upper censoring. Assuming that the given upper limit is \( y_U \), the observable variable \( y_i \) is defined in the following way:

\[
y_i = \begin{cases} 
    y_i^*, & \text{if } y_i^* < y_U \\
    y_U, & \text{if } y_i^* \geq y_U,
\end{cases}
\]  

(2)

where \( y_i^* \) is as defined in equation (1).

The parameters of a Tobit Type I model are thus the linear coefficients \( \beta \), which we need to estimate before making predictions with the model. To this end, we shall use the model likelihood function, which discerns between censored and non-censored observations, as follows: 1) if the observations are non-censored, then the likelihood is

\[
\prod_{i \in N_{nc}} \left( \frac{1}{\sigma} \varphi \left( \frac{y_i - \beta x_i}{\sigma} \right) \right),
\]  

(3)

where \( N_{nc} \) represents the set of non-censored observations, \( \varphi \) is the Gaussian pdf with standard deviation \( \sigma \), and \( y_i \) together with \( \beta x_i \) indicate the observation and the model’s prediction for instance \( i \), respectively.

2) if the observations are censored, the likelihood is

\[
\prod_{i \in N_c} \left( 1 - \Phi \left( \frac{y_i - \beta x_i}{\sigma} \right) \right),
\]  

(4)

where \( N_c \) represents the set of censored observations, \( \Phi \) is the Gaussian cdf and where \( \sigma, y_i \) and \( \beta x_i \) are as defined above. Now, given equations (3) and (4), the joint likelihood of all observations can be defined as:

\[
L(\beta, \sigma) = \prod_{i \in N_{nc}} \left( \frac{1}{\sigma} \varphi \left( \frac{y_i - \beta x_i}{\sigma} \right) \right) \prod_{i \in N_c} \left( 1 - \Phi \left( \frac{y_i - \beta x_i}{\sigma} \right) \right).
\]  

(5)

**Gaussian Processes**

Gaussian Processes (GPs) (14) are an extremely powerful and flexible tool belonging to the field of probabilistic machine learning (15), and have been applied successfully to both classification and regression tasks regarding various transport related scenarios such as travel times (16, 17), congestion models (18), crowdsourced data (19, 20), traffic volumes (21), etc. In the setting of regression, to give an example, given a finite set of points, there is potentially an infinite number of functions which fit the data. Gaussian processes offer an elegant approach to this problem by assigning a specific probability to every possible function. Moreover, GPs implicitly adopt a full probabilistic approach, making possible the structured quantification of the model’s confidence, or uncertainty, on its predictions. This ease in uncertainty quantification is one of the principal reasons why we chose to use this particular framework for our demand prediction problem. A hypothetical transport service provider would not only be interested in obtaining a
highly accurate demand model, but would also, and maybe most importantly, take operational
decisions based on the measure with which the model is confident of its predictions.

Given a training set $\mathcal{D} = \{(x_i, y_i), i = 1, ..., n\}$ of $n$ input vectors $x_i$ and scalar outputs $y_i$,
the goal of GPs is to learn the underlying distribution from data defining a distribution over functions. GPs model this distribution by placing a multivariate Gaussian distribution defined by a mean function $m(x)$ and a covariance function $k(x, x')$, or kernel, between all possible pairwise combinations of input points $\{x, x'\}$ in the dataset on a latent variable $f$. In case of continuous observations $y_i$, these are typically assumed to be generated from a Gaussian distribution centered in $f_i$ and characterized by a variance $\sigma^2$. More concretely, a GP can be seen as a collection of random variables which have a joint Gaussian distribution. GPs are therefore a Bayesian approach which assumes a priori that function values behave according to:

$$p(f|x_1, ..., x_n) = \mathcal{N}(0, K).$$

where $f = [f_1, ..., f_n]^T$ is the vector of latent variables, and $K$ is the covariance matrix whose entries are defined by applying the covariance function $K_{ij} = k(x_i, x_j)$. Notice that here we assume, without loss of generality, that the joint Gaussian distribution is centered in 0.

**Kernels**

A fundamental step in GPs is the choice of the covariance matrix $K$. This will not only describe the shape of the underlying multivariate Gaussian distribution, but it will ultimately define the characteristics of the function we want to model. As stated in the previous section, the covariance matrix is determined by applying a covariance function, or kernel, between all possible pairwise combinations of points in the dataset. Intuitively, since entry $ij$ of the covariance matrix defines the correlation between the $i$-th and the $j$-th random variables, the kernel describes a measure of similarity between points, ultimately controlling the shape that any GP function can adopt.

In this study we will be using three specific kernel functions to model our data:

1) **Squared Exponential Kernel (SE)**

$$k_{SE}(x, x') = \lambda^2 \exp\left(-\frac{(x-x')^2}{2l^2}\right)$$

The SE kernel intuitively encodes the idea that nearby points should be correlated, therefore generating “smooth” functions, up to some extent. The kernel is parametrized by:
- $l$ (lengthscale): the bigger $l$, the more the model will be able to consider further away points as similar
- $\lambda^2$ (variance): measures the spread of the function from its mean. Every kernel will be defined by a variance parameter (which is acting as a scale factor).

2) **Periodic Kernel**

$$k_{per}(x, x') = \lambda^2 \exp\left(-\frac{2\sin^2\left(\frac{\pi|x-x'|}{p}\right)}{l^2}\right)$$
The Periodic kernel allows to model functions which repeat with time. It can be extremely useful in extrapolating seasonality in time-series. It is parametrized by:
- l (lengthscale): same interpretation as in the SE case
- p (period): defines the distance between repetitions
- \( \lambda \) (variance)

3) Matérn Kernel

The Matérn covariance between two points separated by distance \( d \) is given by:

\[
k_{\text{Mat}}(d) = \lambda^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left( \frac{\sqrt{2\nu} d}{l} \right)^\nu K_\nu \left( \frac{\sqrt{2\nu} d}{l} \right),
\]

(9)

where \( \Gamma \) is the Gamma function and \( K_\nu \) is the Bessel function. The Matérn kernel can be considered as a generalization of the SE kernel parametrized by positive parameters \( \nu \) and \( l \) where smaller values of \( \nu \) correspond to less smooth functions.

Typically the parameters regarding any number of kernels are selected by using Maximum Likelihood techniques. What characterizes the great flexibility of kernels is that any valid kernel can be combined (for example by addition and multiplication) with other kernels to generate more complex covariance functions which can be used to better encode domain knowledge in the regression model.

Censored Gaussian Processes

In this study we aim at constructing a model which, on one hand, is specifically equipped to deal with a censored regression problem, while on the other, is also able to exploit the flexibility of GPs to fit highly complex patterns. We do this by defining a censoring-aware likelihood within the previously introduced GP framework, using it in place of the Gaussian likelihood from the standard GP setting. Using the same notation introduced previously, the likelihood can be expressed as follows:

\[
L = \prod_{i \in \mathcal{N}_c} \left( \frac{1}{\sigma} \phi \left( \frac{y_i - f_i}{\sigma} \right) \right) \prod_{i \in \mathcal{N}_c} \left( 1 - \phi \left( \frac{y_i - f_i}{\sigma} \right) \right).
\]

(10)

Inference

In a Bayesian setting, we are interested in computing the posterior over the latent variable vector \( f \) given the output vector \( y \) and inputs \( X \) using Bayes’ rule:

\[
p(f|y, X) = \frac{p(f)p(y|f, X)}{p(y|X)}.
\]

(11)

However, because of the way the Censored GP likelihood is defined, exact inference turns out to be intractable and for this reason, it is necessary to resort to some kind of approximate inference method. Below we describe the Expectation Propagation method and the implementation details for the specific Censored GP likelihood.

As outlined in (14), the Expectation Propagation (EP) algorithm (22) aims at overcoming the posterior intractability given by the non-normality of the likelihood, by approximating this...
same likelihood by a local likelihood approximation in the form of an un-normalized Gaussian distribution in the latent variable $f_i$:

$$p(y_i|f_i) \approx t_i(f_i|\mathcal{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) \triangleq \mathcal{Z}_i \mathcal{N}(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2).$$

where $t_i$ is the $i$-th site with site parameters $\mathcal{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2$. It follows that the product of the $n$ independent local likelihood approximations is:

$$\prod_{i=1}^{n} t_i(f_i|\mathcal{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \mathcal{N}(\tilde{\mu}, \tilde{\Sigma}) \prod_{i=1}^{n} \mathcal{Z}_i,$$

where $\tilde{\mu}$ is the vector of $\tilde{\mu}_i$ and $\tilde{\Sigma}$ is a diagonal covariance matrix with $\tilde{\Sigma}_{ii} = \tilde{\sigma}_i^2$. The posterior $p(f|y, X)$ is then approximated by $q(f|y, X)$

$$q(f|y, X) \triangleq \frac{1}{Z_{EP}} p(f|X) \prod_{i=1}^{n} t_i(f_i|\mathcal{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2) = \mathcal{N}(\mu, \Sigma)$$

with $\mu = \Sigma \tilde{\Sigma}^{-1} \tilde{\mu}$ and $\Sigma = (K^{-1} + \tilde{\Sigma}^{-1})^{-1}$, where $Z_{EP}$ is the EP approximation to the marginal likelihood.

The key idea in EP is to update the single site approximation $t_i$ sequentially by iterating the following four steps: firstly, given a current approximation of the posterior, the current $t_i$ is left out from this approximation, defining a cavity distribution:

$$q_{-i}(f_i) \propto \int p(f|X) \prod_{j \neq i} t_j(f_j|\mathcal{Z}_j, \tilde{\mu}_j, \tilde{\sigma}_j^2) df_j,$$

$$q_{-i}(f_i) \triangleq \mathcal{N}(f_i|\mu_{-i}, \sigma_{-i}^{-2}),$$

where $\mu_{-i} = \sigma_{-i}^{-2}(\sigma_i^{-2} \mu_i - \tilde{\sigma}_i^{-2} \tilde{\mu}_i)$, and $\sigma_{-i}^2 = (\sigma_i^{-2} - \tilde{\sigma}_i^{-2})^{-1}$

The second step is the computation of the target non-Gaussian marginal combining the cavity distribution with the exact likelihood $p(y_i|f_i)$:

$$q_{-i}(f_i)p(y_i|f_i).$$

Thirdly, a Gaussian approximation to the non-Gaussian marginal $\hat{q}(f_i)$ is chosen such that it best approximates the product of the cavity distribution and the exact likelihood:

$$\hat{q}(f_i) \triangleq \mathcal{Z}_i \mathcal{N}(\tilde{\mu}_i, \tilde{\sigma}_i^2) \approx q_{-i}(f_i)p(y_i|f_i).$$

To achieve this, EP uses the property that if $\hat{q}(f_i)$ is Gaussian, the distribution which minimizes the Kullback-Leibler divergence $KL(q_{-i}(f_i)p(y_i|f_i)||\hat{q}(f_i))$ (where KL is a widely used measure of distance between distributions) is the one whose first and second moments match the ones of $q_{-i}(f_i)p(y_i|f_i)$. In addition, given that $\hat{q}(f_i)$ is un-normalized, the EP algorithm imposes the condition that also the zero-th moment should match.

Lastly, the $t_i$ which makes the posterior have the marginal from step three is chosen.
Censored Likelihood Moments

To implement the previous steps we used GPy (23), an open source Gaussian Processes package in Python, where we implemented the Censored likelihood and defined the following moments for the EP inference procedure:

\[ \hat{Z}_i = \mathbb{E}_q(f_i) \]
\[ \hat{\mu}_i = \mathbb{E}_q[f_i] \]
\[ \hat{\sigma}_i^2 = \mathbb{E}_q[(f_i - \mathbb{E}_q[f_i])^2] \]

where in the specific case of the Censored likelihood the moments can be defined analytically (14), depending on the censoring upper limit \( y_U \), as follows:

If \( y_l < y_U \):

\[ \hat{Z}_i = \frac{1}{\sqrt{2\pi(\sigma^2+\sigma_{-i}^2)}} \exp\left(-\frac{1}{2} \frac{(y_l-\mu_{-i})^2}{\sigma^2+\sigma_{-i}^2}\right) \]
\[ \hat{\mu}_i = \mu_{-i} + \sigma_{-i}^2 \frac{(y_l-\mu_{-i})}{\sigma^2+\sigma_{-i}^2} \]
\[ \hat{\sigma}_i^2 = \sigma_{-i}^2 - \sigma_{-i}^4 \left(\frac{1}{\sigma^2 + \sigma_{-i}^2}\right) \]

If \( y_l = y_U \)

\[ \hat{Z}_i = \Phi(z_l^U) \]
\[ \hat{\mu}_i = \mu_{-i} + \frac{\sigma_{-i}^2 \mathcal{N}(z_l^U)}{\phi(z_l^U)\sqrt{\sigma^2 + \sigma_{-i}^2}} \]
\[ \hat{\sigma}_i^2 = \sigma_{-i}^2 - \frac{\sigma_{-i}^4 \mathcal{N}(z_l^U)}{\phi(z_l^U)(\sigma^2 + \sigma_{-i}^2)} \left( z_l^U + \frac{\mathcal{N}(z_l^U)}{\Phi(z_l^U)} \right) \]

where \( z_l^U = \frac{\mu_{-i}-y_U}{\sqrt{\sigma^2+\sigma_{-i}^2}} \) and \( \Phi \) is the Gaussian cdf.

EXPERIMENTS

In what follows we will be dealing with the problem of building a demand prediction model for a bike-sharing system. In what follows we will be considering real-world data provided by Donkey Republic, a bike-sharing provider in the city of Copenhagen, Denmark. Within the bike-sharing scenario, Donkey Republic can be considered a hub-based service, meaning that the user of the service is not free to pick up or drop-off a bike in any location, but is restricted to a certain number of virtual hubs around the city of Copenhagen. In this context, our objective is to determine a model for predicting daily rental demand in the current hub network. The data which was used for experimentation regards records of bike rentals indicating time, date and hub location for both pickup and drop-off for 32 hubs belonging to the Donkey Republic network. In practice, these hubs were aggregated in three super-hubs by selecting three main service areas (such as the central stations and tourist attractions) and considering a 200m radius around these points of interest (Figure 1a). The choice of the radius is justified by business insights which show that 200m can be considered a threshold for user conversion rates. The data at our disposal allowed us to retrieve the time-series of daily rental pickups regarding the three super-hubs which will represent the target of our prediction model (Figure 1b).
Figure 1 - a) Map showing the location and division of hubs in the three super-hubs (in red, green and yellow). In blue other hubs belonging to the Donkey Republic network (here not considered). b) Time series of daily rentals for each of the three super-hubs.
In practice, the historical data upon which our prediction model will rely on is composed of daily counts of bike rentals for a period going from 1 March 2018 until 14 March 2019, defining for each of the super-hubs under consideration a time-series of demand observations regarding 379 days of Donkey Republic usage. It is important to remark that the aggregation of single hubs into super-hubs or areas is an important modeling step in building the prediction model. This is true for two main reasons: 1) the rental time-series belonging to single hubs showed excessive fluctuations in the rental behavior. This would have likely ended up in hiding most of the valuable regularities hidden in the data and ultimately exposing the predictive model to an excessive amount of noise; 2) as it is possible to notice in Figure 1a, the individual hubs are very close one to the other, especially in central areas of the city. It is therefore reasonable to assume that the demand patterns between neighboring hubs are extremely correlated, and that for a bike-sharing provider it is actually more important to understand the demand of the overall area rather than of a single hub (for example, consider a user willing to walk 50m or 100m in order to rent a bike, making the demand conceptually belonging to an entire area rather than an individual location).

In the context of understanding bike-sharing demand, weather surely plays a central role. For this reason, we paired the rental time-series with relevant weather measurements collected in the Copenhagen area by (24). For the purpose of this study we used data regarding irradiance, wind, temperature, rain fall (duration and intensity) and humidity. Concretely, the features used during the training of the models are summarized in the following table:

<table>
<thead>
<tr>
<th>Global shortwave solar irradiance [W/m²]</th>
<th>Wind speed maximum [m/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short wave horizontal solar irradiance [W/m²]</td>
<td>Air temperature [°C]</td>
</tr>
<tr>
<td>Short wave direct normal solar irradiance [W/m²]</td>
<td>Relative humidity [%]</td>
</tr>
<tr>
<td>Downwelling horizontal longwave irradiance [W/m²]</td>
<td>Air pressure [hPa]</td>
</tr>
<tr>
<td>Wind direction minimum [°]</td>
<td>Rain fall [mm]</td>
</tr>
<tr>
<td>Wind direction average [°]</td>
<td>Rain duration [s]</td>
</tr>
<tr>
<td>Wind direction maximum [°]</td>
<td>Rain intensity [mm/h]</td>
</tr>
<tr>
<td>Wind speed minimum [m/s]</td>
<td>Temporal index</td>
</tr>
<tr>
<td>Wind speed average [m/s]</td>
<td></td>
</tr>
</tbody>
</table>

Lastly, in order to define a measure of supply, we use recorded data of bike availability represented by counts of available bikes, regarding every individual hub, during the period under analysis.

**Design**

The goal of this experiment is, as stated previously, to assess how well different predictive models are able to recover the real demand given a set of censored historical observations. However, in the given experiment, the historical data at our disposal naturally represents only the censored version of the demand, making the evaluation of any given model a non-trivial task. Moreover, given the aggregation in super-hubs, the same concept of supply (and consequently of censorship), changes. To address this issue, in what follows we will assume that a generic user, in the process of requesting a bike from a specific hub, in case there are no bikes available, is willing to walk to any other hub within the same super-hub in order rent a bike. This allows to define $a_k(t)$, the availability in super-hub $K$ at time $t$, as the sum of availabilities of all hubs $j$ belonging to that same super-hub at time $t$. More formally:
\[ a_K(t) = \sum_{j \in K} a_j(t). \]  

Once a measure of supply for the super-hub has been introduced, it is possible to define the concept of censored observation of daily demand in order to implement the Censored likelihood described in the previous section. The reasoning is the following: if the number of available bikes in a given super-hub is zero at any moment during the day, this means that potentially on that day there could have been additional demand which the service was not able to satisfy (and therefore not recorded in the historical data at our disposal). Concretely, this enables us to divide each of the 379 days under analysis between the sets of censored \((N_c)\) and non-censored observations \((N_{nc})\).

In what follows, we will (i) assume our historical data to be representing the real demand (which is what ideally we would like to predict) (ii) given the set of censored observations \(N_c\) defined above, apply a known censoring ratio to the above real demand in order to simulate a censored version of the demand (iii) train the models on this last censored demand and evaluate them on the real observations. This can be observed more concretely in the example of Figure 2. Here the time-series representing the real demand (in blue) is superposed to the curve where, in those days corresponding to censored observations (green markers), the demand has been decreased by a certain censoring ratio, or intensity, \(p=50\%\) (in yellow).

![Real vs Censored Observations in Zone 2](image)

**Figure 2** - Time-Series showing experimental sample. In blue the real demand, in yellow the simulated censored counterpart in which the demand is described by a 50\% censoring intensity.

As for any real-world instance, the target for the implemented models will be to recover the real demand after being trained on its censored counterpart.

Concretely, the main goal of our experiments is to analyze how well different models are able to recover the real demand pattern after being trained on a censored version of the same demand. With this intention in mind, it is important to explicitly define two experimental design choices:

1. The experiments compare results regarding two distinct versions of the same model. In particular, these are represented by a GP model with Censored Likelihood (in what follows Censored GP) and a GP with standard Gaussian Likelihood (Non-Censored GP). Both models are implemented by using a combination of SE and Periodic kernels on the temporal index feature, together with a Matérn kernel on weather data. More formally, let \(x_{i,w}\) and \(x_{i,t}\) be the vector of weather data and temporal index for sample \(x_i\) respectively, the covariance function \(k(x_i, x_j)\) between sample \(i\) and sample \(j\) is defined as follows:

\[ k(x_i, x_j) = k_{SE}(x_{i,t}, x_{j,t}) + k_{Per}(x_{i,t}, x_{j,t}) + k_{Mat}(x_{i,w}, x_{j,w}) \]  

(22)
2. Not only is it interesting to investigate to what degree censored models are able to recover the underlying demand pattern compared to standard regression models, but also how the comparison evolves with different censoring intensities. Meaning: the more severe is the censoring, the more any predictive model will struggle to recover the real demand pattern. For this reason the experiment measures the performances of the two models for different censoring intensities ranging from absence of censoring (*full observability* of the real demand) to the extreme case of complete censoring (no demand observed in historical data).

**Results**

This section presents results for the predictive models implemented on each of the three time-series independently after a *10-fold Cross Validation*. Let us initially concentrate on the results obtained for the demand prediction of Zone 1. The plots presented in Figure 3 and Figure 4 can be interpreted as a visual representation of Table 2 and compare the performances of the Censored GP against the Non-Censored GP for different censoring intensities evaluated on both the training set and after the cross-validation technique respectively.

Practically, these results represent two regression evaluation metrics (Mean Absolute Error and $R^2$) for models trained on data with different degrees of censoring (i.e. from 0% to 100% by increments of 10%) and evaluated on the non-censored (i.e. real) demand. Considering that a predictive model is better the more its MAE is to close 0 and the more its $R^2$ is close to 1, the following plots clearly show how, for data characterized by small degrees of censoring, the two models seem comparable. However, the more the data is affected by censoring, the more the Censored GP outperforms its Non-Censored counterpart proving of being able to recover the underlying demand in a substantially more consistent way.

![Figure 3 - Censoring performance analysis on Training set: Mean Absolute Error (left) and $R^2$ (right)](image-url)
More importantly, the Non-Censored regression model ends up being almost completely biased by the censored observations, while the Censored GP doesn’t. As outlined in previous sections, the non-parametric censored nature of the Censored GP allows the model to effectively exploit the concept of censoring, therefore not allowing the censored observations to bias the entire demand model, but somehow treating them differently.

From a methodological perspective, the standard Non-Censored model can be considered a special case of the Censored model. This can be justified by analyzing the two likelihoods shown below:

\[ L_C = \prod_{i \in N_{nc}} \left( \frac{1}{\sigma} \phi \left( \frac{y_i - f_i}{\sigma} \right) \right) \prod_{i \in N_c} \left( 1 - \phi \left( \frac{y_i - f_i}{\sigma} \right) \right), \]
\[ L_{NC} = \prod_{i \in N} \left( \frac{1}{\sigma} \phi \left( \frac{y_i - f_i}{\sigma} \right) \right). \]

where \( N_{nc} \cup N_c = N \). From the equations above it is easy to notice that in the case of absence of censoring (i.e. \( N_c = \emptyset \)), the Censored likelihood is essentially equivalent to the standard Non-Censored case. Censored models are therefore capable of activating their censoring-awareness depending only on data. The difference of results between the Censored and Non-Censored models in our experiments for the 0% censoring case however, is completely justified given the experimental design: as outlined in Section Design, we are evaluating the models on a demand which, in reality, is intrinsically censored (the real demand will never be available from historical data). In practice this means that the models are never exposed to the case an absence of censored observations (i.e. \( N_c \neq \emptyset \)), but only to different intensities of censoring. In a real-world scenario, the 0% censoring case will naturally correspond to an absence of censored observations and, as proven above, with the two models being equivalent.

The following tables show the complete picture of results given by the models for all three zones on both the training and CV evaluations. In particular, results regard Mean Absolute Error and \( R^2 \) score, where we will refer to the Non-Censored and Censored GP models as \( NC \) and \( N \) respectively:
Table 2 - Censoring analysis for Demand model on Zone 1 (Red)

<table>
<thead>
<tr>
<th>Censoring Intensity</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE</td>
<td>NC</td>
<td>4.14</td>
<td>4.22</td>
<td>4.47</td>
<td>4.86</td>
<td>5.35</td>
<td>5.94</td>
<td>6.55</td>
<td>7.2</td>
<td>7.87</td>
<td>8.57</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>4.62</td>
<td>4.54</td>
<td>4.59</td>
<td>4.64</td>
<td>4.73</td>
<td>4.8</td>
<td>4.89</td>
<td>4.93</td>
<td>4.96</td>
<td>4.98</td>
</tr>
<tr>
<td>R^2</td>
<td>NC</td>
<td>0.86</td>
<td>0.85</td>
<td>0.82</td>
<td>0.79</td>
<td>0.74</td>
<td>0.68</td>
<td>0.61</td>
<td>0.52</td>
<td>0.43</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.84</td>
<td>0.84</td>
<td>0.83</td>
<td>0.82</td>
<td>0.81</td>
<td>0.8</td>
<td>0.79</td>
<td>0.78</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td><strong>10-fold CV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE</td>
<td>NC</td>
<td>6.23</td>
<td>6.21</td>
<td>6.3</td>
<td>6.56</td>
<td>7.02</td>
<td>7.48</td>
<td>8.05</td>
<td>8.66</td>
<td>9.3</td>
<td>9.99</td>
</tr>
<tr>
<td>R^2</td>
<td>NC</td>
<td>0.69</td>
<td>0.68</td>
<td>0.66</td>
<td>0.63</td>
<td>0.59</td>
<td>0.54</td>
<td>0.48</td>
<td>0.4</td>
<td>0.32</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.65</td>
<td>0.68</td>
<td>0.68</td>
<td>0.69</td>
<td>0.69</td>
<td>0.68</td>
<td>0.67</td>
<td>0.66</td>
<td>0.65</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 3 - Censoring analysis for Demand model on Zone 2 (Green)

<table>
<thead>
<tr>
<th>Censoring Intensity</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE</td>
<td>NC</td>
<td>3.1</td>
<td>3.19</td>
<td>3.43</td>
<td>3.65</td>
<td>4</td>
<td>4.39</td>
<td>4.82</td>
<td>5.3</td>
<td>5.79</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>3.33</td>
<td>3.28</td>
<td>3.29</td>
<td>3.32</td>
<td>3.35</td>
<td>3.36</td>
<td>3.4</td>
<td>3.42</td>
<td>3.41</td>
<td>3.45</td>
</tr>
<tr>
<td>R^2</td>
<td>NC</td>
<td>0.84</td>
<td>0.83</td>
<td>0.81</td>
<td>0.78</td>
<td>0.73</td>
<td>0.67</td>
<td>0.6</td>
<td>0.51</td>
<td>0.41</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.8</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
<td>0.8</td>
<td>0.8</td>
<td>0.79</td>
<td>0.79</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>10-fold CV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE</td>
<td>NC</td>
<td>4.65</td>
<td>4.7</td>
<td>4.82</td>
<td>5.01</td>
<td>5.29</td>
<td>5.63</td>
<td>6.02</td>
<td>6.45</td>
<td>6.91</td>
<td>7.41</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>5.09</td>
<td>4.94</td>
<td>4.83</td>
<td>4.85</td>
<td>4.93</td>
<td>5.01</td>
<td>5.11</td>
<td>5.26</td>
<td>5.36</td>
<td>5.51</td>
</tr>
<tr>
<td>R^2</td>
<td>NC</td>
<td>0.65</td>
<td>0.65</td>
<td>0.63</td>
<td>0.6</td>
<td>0.56</td>
<td>0.5</td>
<td>0.43</td>
<td>0.35</td>
<td>0.26</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.59</td>
<td>0.61</td>
<td>0.62</td>
<td>0.62</td>
<td>0.61</td>
<td>0.61</td>
<td>0.59</td>
<td>0.57</td>
<td>0.55</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 4 - Censoring analysis for Demand model on Zone 3 (Yellow)

<table>
<thead>
<tr>
<th>Censoring Intensity</th>
<th>0.0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE</td>
<td>NC</td>
<td>3.78</td>
<td>3.83</td>
<td>3.95</td>
<td>4.11</td>
<td>4.37</td>
<td>4.67</td>
<td>5.02</td>
<td>5.4</td>
<td>5.8</td>
<td>6.21</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>3.78</td>
<td>3.74</td>
<td>3.73</td>
<td>3.78</td>
<td>3.86</td>
<td>3.87</td>
<td>3.94</td>
<td>4</td>
<td>4.01</td>
<td>4.05</td>
</tr>
<tr>
<td>R^2</td>
<td>NC</td>
<td>0.78</td>
<td>0.78</td>
<td>0.75</td>
<td>0.73</td>
<td>0.68</td>
<td>0.63</td>
<td>0.56</td>
<td>0.49</td>
<td>0.4</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.78</td>
<td>0.78</td>
<td>0.77</td>
<td>0.76</td>
<td>0.75</td>
<td>0.74</td>
<td>0.73</td>
<td>0.72</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>10-fold CV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAE</td>
<td>NC</td>
<td>5.6</td>
<td>5.61</td>
<td>5.66</td>
<td>5.76</td>
<td>5.93</td>
<td>6.14</td>
<td>6.42</td>
<td>6.74</td>
<td>7.1</td>
<td>7.49</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>5.8</td>
<td>5.63</td>
<td>5.59</td>
<td>5.53</td>
<td>5.54</td>
<td>5.57</td>
<td>5.57</td>
<td>5.6</td>
<td>5.62</td>
<td>5.67</td>
</tr>
<tr>
<td>R^2</td>
<td>NC</td>
<td>0.54</td>
<td>0.53</td>
<td>0.51</td>
<td>0.49</td>
<td>0.46</td>
<td>0.41</td>
<td>0.36</td>
<td>0.29</td>
<td>0.22</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.52</td>
<td>0.54</td>
<td>0.55</td>
<td>0.55</td>
<td>0.54</td>
<td>0.53</td>
<td>0.52</td>
<td>0.52</td>
<td>0.51</td>
<td>0.49</td>
</tr>
</tbody>
</table>
DISCUSSION AND FUTURE WORK

Building a model for demand prediction naturally relies on extrapolating knowledge from historical data. This is usually done by implementing different types of regression models in order to explain the past demand behavior and that are able to compute reliable predictions for the future, which is a fundamental building block for a great number decision making processes. However, we have shown how a reliable predictive model must take into consideration the censoring problem, especially in those cases in which demand is implicitly limited by the supply. More importantly, we stressed the fact that, in the context of shared transport demand modeling, there is a need for models able to deal with censoring in a meaningful way rather than resorting to different data cleaning techniques. We therefore showed how ideas from censored regression models can be applied to demand modeling. In particular, we constructed a Gaussian Process model where a Censored likelihood was used opposed to a standard Gaussian likelihood in order to deal with the censoring problem. Moreover, GPs represent an elegant framework naturally able to work in a fully probabilistic way and to deal with uncertainty in predictions. Experimental studies on real-world instances in the context of bike-sharing demand prediction highlighted how standard regression models are prone to returning a biased model of demand in the case of censored observations, while the implemented Censored GP exhibited a strong consistency in its predictions, even in severe censoring conditions. As shown in Section Experiments, the experimental results seem to confirm the importance of censoring in demand modeling, especially in the transportation scenario where demand and supply are naturally interdependent. More generally, our results defend the idea of building more knowledgeable models rather than using black-box models after accurate data cleaning techniques. This can be done by feeding the demand models insights on how the demand patterns actually behave and letting these adjust automatically to the available data.

In the context of shared transport demand modeling, the present study does not aim at exploiting spatial correlations between different geographical areas. These correlations are most likely relevant especially in a context like the one of bike-sharing demand prediction, where nearby areas are most likely to exhibit similar demand patterns or where there might be regular commutes between groups of areas. Guided by this domain knowledge, we want to implement models which are able to predict demand jointly for all the areas of interest rather than treating them individually. Further research also includes the analysis of how the censored models implemented in this study interact with a multi-output setting as the one outlined above. Such considerations can further improve the construction of reliable spatio-temporal demand models and ultimately creating reliable tools for decision-making.
REFERENCES


22. Minka, T. P. *Expectation Propagation for Approximate Bayesian Inference*.
