Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Relazione Finale

Anomaly Detection in Time Series with the Prophet Model

Relatore
Prof. Matteo Ceccarello
Dipartimento di Ingegneria dell'Informazione

Laureando
Giovanni Zedda
Matricola N. 2082887

Anno Accademico
2024/2025

Acknowledgements

Writing this thesis means to me the crowning achievement of three years of passionate studies and hard work, but it was only made possible with the help of a lot of special people I really want to say thanks to. I start with thanking my supervisor Matteo Ceccarello, who supported me a lot with helpful tips to improve my thesis, and was also an excellent teacher. Then I must give recognition to all my other professors too, who taught me much and gave me the right inspiration to continue my studies. Of course, I cannot forget all my friends from university, because I shared with them a large part of my uni-life, by pushing each other towards our goals, but also by having fun together and making my staying in Padova exceptional. There are also many other people that have always believed in me and treated me well, therefore I am grateful to them as well. But my greatest THANK YOU goes to my family, that has always been by my side whenever I needed them and with patience and sacrifice supported me in all possible imaginable ways.

Abstract

Insofar human activities constitute a major facet among all possible aspects that describe the world, they represent a valuable source of data for statisticians. Data coming from anthropic behaviours are often influenced by temporal factors, such as the month of the year, the day of the week or the presence of holidays and thereby tend to show periodic structures, sometimes along with underlying trends. However, observations that deviate away from the expected pattern occur more frequently than one may think. In this regard, a relevant task for the analyst is to detect anomalies and possibly identify the origins of these alienated points. For this purpose, the Prophet model, a Bayesian mixed linear-nonlinear model designed for time series forecasting, comes to the rescue. We conduct a practical analysis using Prophet on a real-world dataset. Additionally, we explore several methodologies to evaluate the anomaly score.

Contents

1 Introduction

Sequences of numerical data points, each associated with a specific instant or interval of time, take the name of time series, which form a ubiquitous topic in statistics. Among the several statistical questions that arise when dealing with this kind of sequential data, one concerns the identification of irregularities in the series and possibly a measurement of their degree of aberration and the retrieval of their causes. Such deviations could be global outliers, observations that differ noticeably from all the others, as well as contextual outliers, data points that oddly diverge from their neighbours or from an estimated baseline. There exist also pattern-wise anomalies, and many other types too. This kind of task is referred in literature as *anomaly detection* and finds application in many fields, including statistical quality control, cybersecurity and fraud prevention. This thesis discusses an interesting case in the field of urban mobility. In chapter 2 introductory concepts of Bayesian statistics are presented, in particular, the foundations of the Bayesian thinking, the development of a working Bayesian model and a short overview of the underlying algorithmics are herein elaborated. Most of this chapter relies on McElreath (2020), with supplements by Kruschke (2014) and Pettigrew and Weisberg (2019), especially for the first two sections. Chapter 3 introduces the Prophet model, a Bayesian model for time series, together with an explanation of all its components and a short discussion about parameter selection. Chapter 4 is dedicated to the real world case study, where the Prophet model is applied to the number of taxi trips in New York City and then evaluated, before trying a few heuristics and more formal statistical techniques to spot anomalies hiding in the dataset. Chapter 5 concludes with a resumption and some final considerations.

2 The Bayesian Approach to Statistical Inference

2.1 An overview of the concept of probability

When tackling a statistical inference problem, much of the underlying mathematics is clearly founded on probability calculus. However, the definition of *probability* itself is far from trivial.

Definition 2.1 (Kolmogorov's axioms)**.** *Given a collection of subsets $\mathcal{A}$ of a set Ω,a probability measure $Pr(\cdot) : \mathcal{A} \to \mathbb{R}$, associates to each element of $\mathcal{A}$ a numerical value such that the following rules hold:*

- non-negativity*: $Pr(A) \geq 0$, $\forall A \in \mathcal{A}$;*
- normality*: $Pr(\Omega) = 1$;*
- σ-additivity*: if $A \cap B = \emptyset$, $A, B \in \mathcal{A}$, then $Pr(A \cup B) = Pr(A) + Pr(B)$.*

For precise mathematical reasons, the collection $\mathcal{A}$ must have specific properties: $\Omega \in \mathcal{A}$; if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$; if $A_1, A_2, \cdots \in \mathcal{A}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$. In other words, $\mathcal{A}$ is defined as a σ-algebra on Ω.

Nevertheless, the axioms outlined above do not constitute, by themselves, an operational definition of probability, as the numerical quantification of the probabilities of various events remains an open question.

Traditionally, probability is defined as the limit of a relative frequency.

Definition 2.2. *Let Ω be the sample space of all existing outcomes of a random experiment and let $E \subseteq \Omega$ be an arbitrary event. Hypothetically, the same experiment is repeated an infinite number of times. Let n_E be the number of* successes *(i.e.,*

occurrences of E) obtained in n trials. Hence, the probability of the event E can be defined as

$$Pr(E) = \lim_{n\to\infty} \frac{n_E}{n}.$$

This definition raises some issues: primarily, it is not obvious that this limit does exist; secondly, it relies on a - even if theoretical - replication of the experiment, which sounds unnatural for propositions such as "There is life on Mars" or "The birth rate in Africa will be at least halved in fifty years".

Another way of conceiving probability is based upon a subjective sense of plausibility that a certain event occurs (de Finetti, 1933).

Definition 2.3. *The probability of an event is the degree of belief assigned to its occurence by a rational agent (i.e, coherent with Kolmogorov's axioms), interpretable as a virtual bet, that is, up to how much money the same agent is inclined to stake in order to win one unit of money in the case of success.*

This alternative definition is therefore more flexible, although the implied lack of unambiguity might be instead regarded as a limitation.

It is worth mentioning that this "degree of belief" can change not only between different agents, but also within the same individual throughout time, by accumulating knowledge and experience.

2.2 Bayes's theorem

Let A and B be two events and let $Pr(A)$ and $Pr(B)$ be their respective probabilities. Finally, let $Pr(A|B)$ and $Pr(B|A)$ be defined as the probability of A *conditioned* on B and the probability of B *conditioned* on A, respectively.

Theorem 2.1 (Bayes's theorem)**.**

$$Pr(B|A) = \frac{Pr(A|B) \cdot Pr(B)}{Pr(A)}.$$

This result is a direct consequence of the definition of conditional probabilities, in particular $Pr(A|B) = Pr(A \cup B)/Pr(B)$ and $Pr(B|A) = Pr(A \cup B)/Pr(A)$, with $Pr(A)$ and $Pr(B)$ other than zero.

If $H_1, \ldots, H_n$ form a finite partition of B, then the following equivalence holds:

Theorem 2.2 (Law of total probability)**.**

$$Pr(A) = Pr(A|H_1) \cdot Pr(H_1) + \cdots + Pr(A|H_n) \cdot Pr(H_n).$$

Combining these two theorems together leads to

$$Pr(B|A) = \frac{Pr(A|B) \cdot Pr(B)}{Pr(A|H_1) \cdot Pr(H_1) + \cdots + Pr(A|H_n) \cdot Pr(H_n)}. \tag{2.1}$$

This formula offers one interesting interpretation: if A is an observable event, for example the positivity of a diagnostic test for a disease, and B is a *latent*, not immediately observable, event, such as having the actual disease, then the belief of the latter is altered by the outcome of the test, in a way that makes the *posterior probability* $Pr(B|A)$ proportional to the product of the *prior probability* $Pr(B)$ and the probability $Pr(A|B)$ of a test being positive given the affection to the disease. This procedure of recalculating probabilities in light of new evidence is also known as *Bayesian update* and can possibly be iterated, as more data become available, for instance, through a sequence of clinical trials or repeated measurements.

When it comes to random variables, a probability measure is usually characterised by a probability mass function (PMF) or a probability density function (PDF), both referred in this work as $p(\cdot)$ or $\pi(\cdot)$. It can be derived that a continuous random variable Y, with support $\mathcal{Y}$, conditioned on another random variable $X = x$, has distribution

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y) \cdot p_Y(y)}{p_X(x)} = \frac{p_{X|Y}(x|y) \cdot p_Y(y)}{\int_{\mathcal{Y}} p_{X|Y}(x|y) p_Y(y) dy}. \tag{2.2}$$

Clearly, in the case of discrete random variables, the integral will be replaced by a summation.

2.3 Two paradigms for statistical knowledge

The two operational definitions of probability introduced in section 2.1 influence the way to deal with statistical problems, which include an evaluation of a proportion or of a mean in a population, or the degree of association between a treatment and the recovery, amongst the many others. In particular, after Fisher's works carried out during the early twentieth century, the so-called *frequentist statistics* has spread. The frequentist approach requires all probabilities to be defined as frequencies in large samples, so it is based on the *repeated sampling principle*, in according to which the observed sample is just the outcome of a random variable, whose distribution relies on a number - finite or infinite - of parameters, that are unknown, but fixed - not random.

An apparently opposite, but actually more general procedure expects that parameters are seen as random variables, therefore a probability density function is assigned to each parameter, initially, according to a personal opinion about how likely each parameter value is. Ultimately, this belief changes in accordance with available data. Since the readjustment of the parameter distribution is obtained by means of Bayes's formula, the overall idea is known as *Bayesian statistics*.

All epistemic debates aside, there is not a unique way to face a statistical question. Indeed, no method, whatever Bayesian or frequentist, is universally valid nor sufficient by itself, rephrasing the motto «all models are wrong, but some may be useful», attributed to Box (1976). Each of the aforementioned methods brings pros and cons. For instance, in Bayesian statistics, inferential results may vary depending on different prior distributions for the parameter. Moreover, expensive computational costs are often required. On the other hand, assumptions on the generative process of the data is made forcedly explicit, while within a frequentist framework it is often possible to exploit opaque pre-constructed procedures that might mislead the statistician, unless he or she takes the due precautions. Missing check of working hypothesis (homoscedasticity, independence etc.) and *p-value fishing* are just some of the possible causes of bias in the conclusions. Generally, frequentist methods are still the most widespread, however, more efficient computer systems and algorithms have lead to a rapid grow of Bayesian techniques since the early 1990s.

2.4 Building a Bayesian model

To construct a Bayesian model, first of all it is necessary to set a parametric model for the observed data $\boldsymbol{y}$, which belongs to a family of parametric distributions $\mathcal{F} = \left\{p(\cdot|\theta) : \theta \in \Theta \subseteq \mathbb{R}^k\right\}$, where θ is the parameter and Θ is called *parameter space*. Function $p(\cdot|\theta)$ is claimed to be known, when θ is known. In Bayesian inference, θ can be handled as a random variable with distribution $\pi(\cdot)$ and support Θ. Before looking at the sample, a *prior distribution* $\pi(\theta)$ is chosen by the statisticians (further details about this step will be discussed in the next section). After observing the data $\boldsymbol{y}$, it is possible, for all θ, to calculate the function $p(\boldsymbol{y}|\theta)$, named *likelihood*, and in other contexts pointed out as a function of the parameter $\mathcal{L}(\theta|\boldsymbol{y})$. Finally, Bayes's formula updates the distribution of θ as follows:

$$\pi(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)\pi(\theta)}{\int_\Theta p(\boldsymbol{y}|\vartheta)\pi(\vartheta)d\vartheta}. \tag{2.3}$$

$\pi(\theta|\boldsymbol{y})$ is called *posterior distribution* and $\int_\Theta p(\boldsymbol{y}|\vartheta)\pi(\vartheta)d\vartheta$ is the average probability of observing $\boldsymbol{y}$, where "average" means properly the mathematical expectation $\mathbb{E}_\pi[p(\boldsymbol{y}|\theta)]$, and plays the role of a normalization constant for the posterior distribution, which must integrate (or sum) to 1.

2.5 Practices for choosing priors

The choice of the prior distribution falls on whoever conducts the analysis, though there are a few guidelines for this initial and sometimes crucial decision that are more likely to be embraced by a community, as they are more principled than other more opinionated preferences for the prior.

2.5.1 Conjugate priors.

Definition 2.4 (Conjugate prior)**.** *A prior distribution for the parameter θ is called a* conjugate *prior for the likelihood $p(\boldsymbol{y}|\theta)$, if its posterior belongs to the same distributive family.*

One example is the *Beta-conjugate* to the binomial distribution: if $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$

is a vector of the outcomes of n i.i.d. Bernoulli variables with mean θ, and if the prior $\pi(\theta)$ is a Beta-distribution with *hyperparameters* α and β, then $\theta|\boldsymbol{y}$ is still Beta-distributed, with parameters $\alpha + \sum_{i=1}^{n} y_i$ and $\beta + n - \sum_{i=1}^{n} y_i$.

There exist many other conjugate priors (see Kruschke, 2014), yet they are not able to cover less trivial likelihood assumptions. In addition, hyperparameters still have to be tuned. A *hyperparameter* is a value that rules the parameter distribution. Some *hierarchical model* specifications settle a distribution also for the hyperparameters, but there is always a bottom-level distribution that implicitly or explicitly relies on one or more hyperparameters.

2.5.2 Informative and uninformative priors. Historically, conjugate priors overcame the difficulty of the integration, though, as more powerful computational tools were developed, conjugate priors became no longer a constraint. That allowed to introduce freer procedures for the choice of the prior, most of them include heuristics premised on information theory.

Definition 2.5 (Entropy)**.** *The* entropy *of a random variable X with distribution $p(\cdot)$ and support $\mathcal{X}$ is a measure of its uncertainty defined as*

$$H(X) = \mathbb{E}_p \left[\log \left\{ \frac{1}{p(X)} \right\} \right] = - \int_{\mathcal{X}} p(x) \log p(x) dx.$$

Every base of the logarithm greater than 1 fits the definition. Here is intended the natural logarithm. So, *information* can be perceived as a reduction in uncertainty, i.e. in entropy. On this preamble, an *uninformative prior* is a prior that brings in some sense the least amount of information in the posterior. The oldest and simplest rule for selecting an "unbiased" (not in a statistical meaning) prior is the *principle of indifference* or *principle of insufficient reason*, which states that in absence of any evidence equal credibility must be assigned to each possible parameter value. In other words, it sets a uniform distribution for the prior, which also happens to be the highest entropy distribution for random variables with finite support. When the support is not finite a uniform distribution is considered an *improper prior*, because it does not integrate to one, although the posterior might. A more general approach is Jaynes's *maximum entropy principle* (Jaynes, 1957), a technique for finding a probability distribution that is most consistent with the current state of knowledge.

Such a distribution is called *maximum entropy* (maxent) because it maximizes entropy, under some constraints. For instance, the *exponential distribution* maximizes entropy in a positive range when the mean is fixed, while a *normal distribution* has maximum entropy among all distributions with the same variance, indeed both the exponential and the normal distribution are common priors in Bayesian models, but of course not the only ones.

Actually, no prior is truly uninformative, because it necessarily entails information about the parameter space, but also about its mean or variance, leaving out improper priors. One situation where less vague priors are preferred is in Bayesian regression, where the prior distribution of a β coefficient that links a covariate to the mean of the response variable is called a *regularizing prior* if it gravitates tightly around zero. When β is initially normally distributed, this is more or less equivalent to having mean zero and low variance. As a matter of fact, more conservative priors can prevent from *overfitting* and bad inference.

Also, uninformative priors are unlikely to be efficient, therefore it might be thoughtful to take domain knowledge into account, so that there cannot be almost any unreasonable deviation from the reality.

2.6 Output summaries

The object of the Bayesian inference is the entire posterior distribution of the parameters, nevertheless it might be useful to have point or interval statistics that help in visualizing and communicating the results of a Bayesian model.

2.6.1 Point estimates. When a single value for the parameter is desired, a well-reasoned approach is minimizing a task-specific loss function. The most common loss functions are the *mean squared error* (MSE) $\mathbb{E}[(\theta - \tilde{\theta})^2 \mid \boldsymbol{y}]$ and the *mean absolute error* (MAE) $\mathbb{E}[|\theta - \tilde{\theta}| \mid \boldsymbol{y}]$, which are at their minimum when $\tilde{\theta}$ is equal to the mean and to the median of the posterior distribution of the parameter, respectively. Another point estimate is the *mode*, in this context known as *maximum a posteriori* (MAP). These estimates very often converge to the same value.

2.6.2 Intervals of defined mass. A more common summary is a range of parameter values called a *credible interval* or *compatibility interval* that have together a fixed - possibly high - probability mass. Two popular credible intervals are the *percentile interval* (PI), where equal probability mass is assigned to each of the two tails of the posterior distribution, and the *highest posterior density interval* (HPDI), that contains all most plausible parameter values until the prescribed probability mass is reached. These two intervals are usually similar as well.

2.6.3 Probability within defined boundaries. Another question that can be asked is how high the probability is that the parameter lies in a specified subset Ψ of Θ. The answer is quite immediate

$$\int_{\Psi} \pi(\vartheta|\boldsymbol{y})d\vartheta \tag{2.4}$$

and allows to say how likely a parameter is to be greater or lower than a threshold, or to be contained in a range of interest.

2.7 Computational aspects

As mentioned before, the computation of the posterior distribution is a tough challenge in Bayesian statistics. However, multiple solutions have been found to solve this task. Here are presented three algorithms that serve the purpose.

The most basic algorithm is *grid approximation*. It consists in discretizing the sample space, then calculating the posterior as the product of the likelihood and the prior for each new bin of the new parameter space and finally dividing by the sum of all products in order to normalize the posterior. Unfortunately, as the dimension of the parameter space increases, this algorithm becomes computationally inapplicable.

The most efficient algorithm among these is *quadratic approximation*. Under quite general conditions (see Freedman, 1963 and Freedman, 1999), the posterior distribution around the MAP is approximately Gaussian (the logarithm is a parabola, hence the name), therefore the posterior depends only on the mean and the covariance matrix of $\theta|\boldsymbol{y}$, which can be calculated.

When a quadratic approximation is not sufficient, *Metropolis-Hastings* algorithm represents a valid alternative. *Metropolis-Hastings* belongs to a broader class of algorithms called *Markov chain Monte Carlo* (MCMC) because it generates a *Markov chain* whose elements form asymptotically a sample from the posterior distribution. The following pseudocode illustrates how the algorithm works.

Algorithm 2.1 Metropolis-Hastings algorithm

1: Let $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{B-1}) \leftarrow \mathbf{0}_B$
2: Choose an initial value θ_0
3: **for** $j = 1$ to $B - 1$ **do**
4: Sample $\theta' \sim q(\theta' \mid \theta_{j-1})$
5: Compute acceptance probability:

$$\alpha = \min\left(1, \frac{\pi(\theta'|\boldsymbol{y}) \cdot q(\theta_{j-1} \mid \theta')}{\pi(\theta_{j-1}|\boldsymbol{y}) \cdot q(\theta' \mid \theta_{j-1})}\right)$$

6: Sample $\theta_j = \begin{cases} \theta' & \text{with probability } \alpha \\ \theta_{j-1} & \text{with probability } 1 - \alpha \end{cases}$
7: **end for**

Naturally, $\pi(\theta'|\boldsymbol{y})$ and $\pi(\theta_{j-1}|\boldsymbol{y})$ are individually unknown, but their ratio is

$$\frac{p(\boldsymbol{y} \mid \theta')\pi(\theta')}{p(\boldsymbol{y} \mid \theta_{j-1})\pi(\theta_{j-1})}.$$

q is a proposal distribution where the candidate θ' is selected from, for example $\theta' \mid \theta_{j-1} \sim \mathcal{N}(\theta_{j-1}, \epsilon^2)$. In this case $q(\theta_{j-1} \mid \theta') = q(\theta' \mid \theta_{j-1})$, so the algorithm is downgraded to a simple *Metropolis algorithm*.

Once it is possible to draw a sample from the posterior, not only finding joint and marginal credible intervals for the parameter becomes immediate, but it is also possible to get the *posterior predictive distribution* (ppd) by sampling at first a vector of θ's from the posterior $\pi(\theta|\boldsymbol{y})$, then a y from $p_1(y|\theta)$ (p_1 is the likelihood for a single observation), for each θ generated. The posterior predictive distribution is important both because it displays predictive estimates with their relative uncertainty and because it shows how consistent the original data are with the current model.

3 The Prophet Model

Time series modelling is a very common task in statistics and data science, because a large variety of data are intrinsically linked to their chronological order and to the moment at which they arise, from second-to-second stock prices in financial markets to daily book sales and to the annual number of emigrants. Because of their utmost relevance in this thesis, we define time series as follows (di Fonzo & Lisi, 2005):

Definition 3.1 (Time series)**.** *A* time series *is a sequence of numerical data-points sorted according to a time index. The time series is* univariate *if each observation is a single real number. Time parameter t that defines the ordering of the data belongs to a parametric space $\mathcal{T}$, and if $\mathcal{T} = \mathbb{Z}$ the series is called* discrete-time *time series.*

From a Bayesian perspective, Prophet (Taylor and Letham, 2018) is a time series model, designed with the goal of forecasting at scale, which means that it claims to adapt to a wide diversity of contexts with specific features, while maintaining a high level interface for the analyst. Besides mere forecasts, a well-fitted Prophet model may help accomplish many other tasks as well, such as revealing trend variations or detecting the presence of anomalies in the dataset.

3.1 Model components

A Prophet model consists of a trend component $g(t)$, a seasonal component $s(t)$ and a holiday component $h(t)$ that contribute additively to the response $y(t)$, as equation 3.1 shows.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t, \qquad t = 1, \ldots, T. \tag{3.1}$$

Here ϵ_t is the idiosyncratic component, the noise. The functional forms of the other components follow precise patterns that are discussed thoroughly in the next paragraphs.

3.1.1 Trend. For the trend component $g(t)$ Prophet provides two alternatives: a nonlinear, saturating growth or a piecewise linear model. The first type of trend is basically a logistic curve, whose core is

$$g(t) = \frac{C(t)}{1 + \exp\{-k(t-m)\}} \tag{3.2}$$

with $C(t)$ the carrying capacity at the time t, k the growth rate, and m an offset parameter. If the limit of saturation does not change with time, $C(t) = C$ constant.

The Prophet model also incorporates the possibility to include manually or automatically S change points at times $s_1, \dots, s_S$ over a history of T observations. At these change points, the base rate k is allowed to shift by an amount δ_j, $j = 1, \dots, S$. Formally, the formula above updated to

$$g(t) = \frac{C(t)}{1 + \exp\left\{-(k + \boldsymbol{a}(t)^\top \boldsymbol{\delta})(t - m - \boldsymbol{a}(t)^\top \boldsymbol{\gamma})\right\}} \tag{3.3}$$

where $\boldsymbol{a}(t)$ is a vector of length S that is valued 1 on position j if $t \geq s_j$ and 0 otherwise, $\boldsymbol{\delta}$ is the vector of δ_j, and $\boldsymbol{\gamma}$ is a vector, function of $\boldsymbol{\delta}$, that gives the trend continuity at the change points.

The other alternative is a linear model, once again with change points, outlined by the expression

$$g(t) = (m + \boldsymbol{a}(t)^\top \boldsymbol{\gamma}) + (k + \boldsymbol{a}(t)^\top \boldsymbol{\delta}) \cdot t \tag{3.4}$$

where $\boldsymbol{a}(t)$ is the same as before, while the other symbols have a similar interpretation to their correspondent in formula 3.3 too.

Weakly informative priors are set for k and m, which are both normally distributed, centred in 0 and with variance 5 by default, while a more sparse (i.e., regularizing) prior is put on $\boldsymbol{\delta}$: $\delta_j \sim \text{Laplace}(0, \tau)$, $\tau > 0$. The closer to 0 is τ, the less fluctuating will be the evaluated trend, whereas a larger value for τ implies more susceptibility to trend changes.

3.1.2 Seasonality. Seasonality in the Prophet model is modelled through a Fourier approximation of a periodic function. A Fourier approximation is a truncation of a Fourier series, which is a way to write functions as a sum of sines and cosines.

Definition 3.2 (Fourier series)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous periodic function with period P. The Fourier series of f is*

$$S(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2\pi nt}{P}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2\pi nt}{P}\right)$$

where $a_0, a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ are called Fourier coefficients and can be computed exactly when $f(x)$ is completely known.

The Fourier approximation that follows definition 3.2 is then

$$s(t) = \sum_{n=1}^{N} a_n \cos\left(\frac{2\pi nt}{P}\right) + \sum_{n=1}^{N} b_n \sin\left(\frac{2\pi nt}{P}\right) \tag{3.5}$$

or in matrix notation

$$s(t) = X(t)\boldsymbol{\beta} \tag{3.6}$$

where $\boldsymbol{\beta} = (a_1, b_1, \ldots, a_N, b_N)^\top$ and

$$X(t) = \left(\cos\left(\tfrac{2\pi(1)t}{P}\right), \quad \ldots, \quad \sin\left(\tfrac{2\pi(N)t}{P}\right)\right).$$

a_0 can be left out because it belongs to the trend component. In the model, $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$ is the prior. According to Prophet's authors, $N = 10$ and $N = 3$ work well for most problems for yearly ($P = 365.25$) and weekly ($P = 7$) seasonality, respectively. Nested seasonalities can be included as well simply by adding them together.

3.1.3 Holidays and events. In many time series, special days like Christmas Day, New Year's Eve or other less regular huge events trigger shocks that simple models with only a trend and a seasonal component cannot foresee. Suppose there are L different holidays in the time series. For each holiday i, let D_i be the set of all dates for that holiday. With the help of an indicator function $\mathbf{1}_{D_i}(t)$ that is equal to 1 if $t \in D_i$ and 0 otherwise, a matrix of covariates

$$Z(t) = \left(\mathbf{1}_{D_1}(t), \quad \ldots, \quad \mathbf{1}_{D_L}(t)\right)$$

can be generated and the holiday component can be expressed as

$$h(t) = Z(t)\boldsymbol{\kappa}, \tag{3.7}$$

where $\kappa \sim \mathcal{N}(0, \nu^2)$. Both σ and τ are hyperparameters that can be selected before fitting the model.

At this stage, the last missing piece is a distributive assumption for the vector $\boldsymbol{y}$ of observations. Again, the normal distribution is the chosen one.

3.2 Model evaluation

The approach to model evaluation illustrated in this section is independent from the forecasting method used. It is somehow similar to train-test validation, since it is based on out-of-sample predictions. Let H be a *horizon*, the number of instants that follow the last available instant T. Let $\hat{y}(t|T)$ represent the prediction for time t given all history until time T and $d(y, \hat{y})$ be an error metric, such as RMSE, MAE or MAPE. The goal is to calculate the error for future observations, that are obviously not available. A way to overcome this obstacle is via *simulated historical forecasts* (SHFs): K cutoff points in the time series are selected from which a fictional horizon of length H starts, then K models are fitted with data points that precede each cutoff and as many predictions are made for $h = 1, \ldots, H$. All computed errors for each distance h form a sample whose mean is an estimate of the prediction error at $T + h$. As a heuristic, $K \approx \frac{2T}{H}$ is a number of predictions that is neither too small to be precise, nor so large to slow down the process. Figure 3.1 shows an example of comparison between two models through SHF.

As a final consideration, Prophet includes change points that are randomly selected throughout the history, when not provided by the user. This leads to two main consequences: late change points might bias the future trend because of too few observations and the other issue is that the trend rate is constant in the future, but there is no evidence in favour of that, indeed frequent trend changes in the past might indicate the presence of as many deviations in the future. The former problem is addressed by considering only a reduced portion, e.g., the first 80%, of

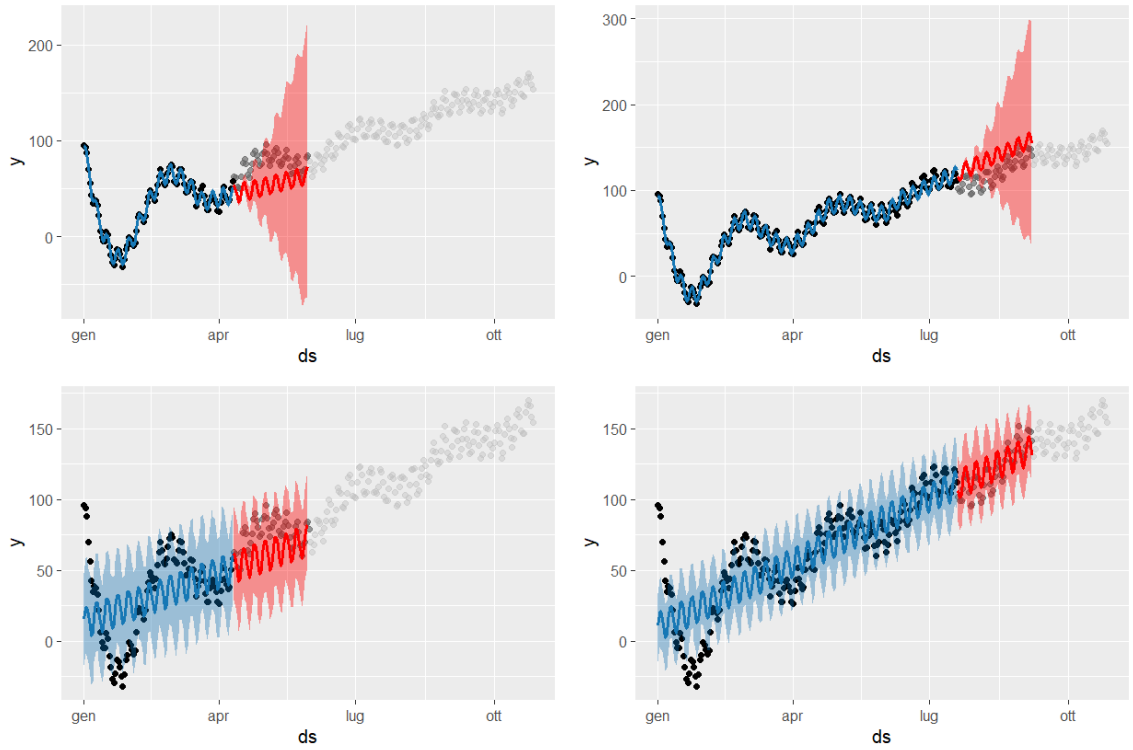


Figure 3.1. SHF for synthetic data using two different Prophet models (one more overfitting model at the top and a more regularized one at the bottom), two cutoffs and a horizon within 50 periods ahead. The second model is much more robust and has a smaller prediction error.

the history during the change point selection process, the latter by guessing future trend changes on the basis of both the number of change points in the dataset and the posterior distribution of the parameter $\boldsymbol{\delta}$, with the goal of having a more robust approximation of the uncertainty around the forecast. However, this kind of predictive issues is beyond the scope of this work.

4 A Case Study: Demand for Taxi Services in New York City

As it was discussed previously, Prophet finds different applications in several contexts. A particular scenario is given by road traffic data and related options. In this regard, the municipality of New York City releases the data relative to all registered taxi rides around the City, including pick up and drop off time and location, charges, payment type, the number of passengers and many other covariates. After data from the 1st November 2023 to 28th February 2025 were collected, they were grouped by start hour and finally counted. Figures 4.1 and 4.2 depict hourly and daily counts, respectively, in the whole interval. The large amount of data makes the graph look

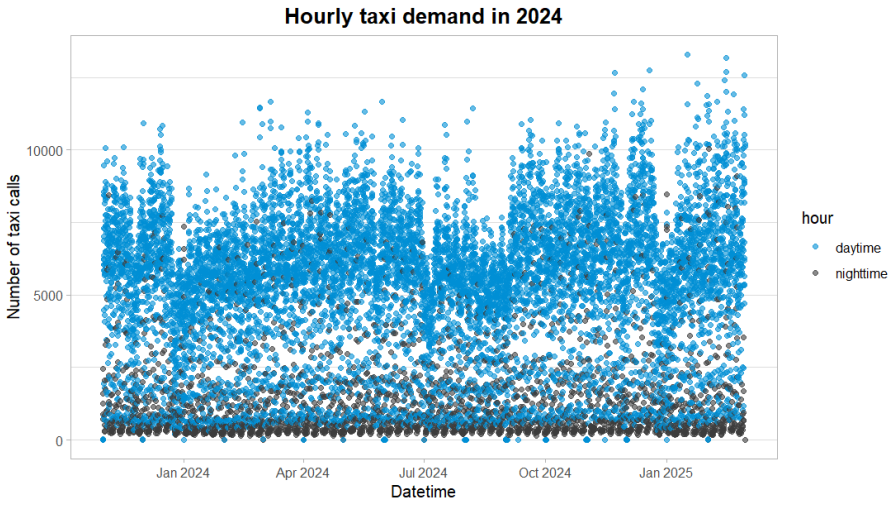


Figure 4.1. Scatterplot of hourly taxi demand in 2024 ($\pm$ 2 months).

very choppy, nevertheless it is clear that, on average, fewer taxi requests come with night hours, while there is a higher demand in the daytime. Likewise there are more calls on weekends than on weekdays and a non-linear overall trend. All of these are good hints for attempting a Prophet fit to model the data. All analyses are performed using R Statistical Software (R Core Team, 2025).

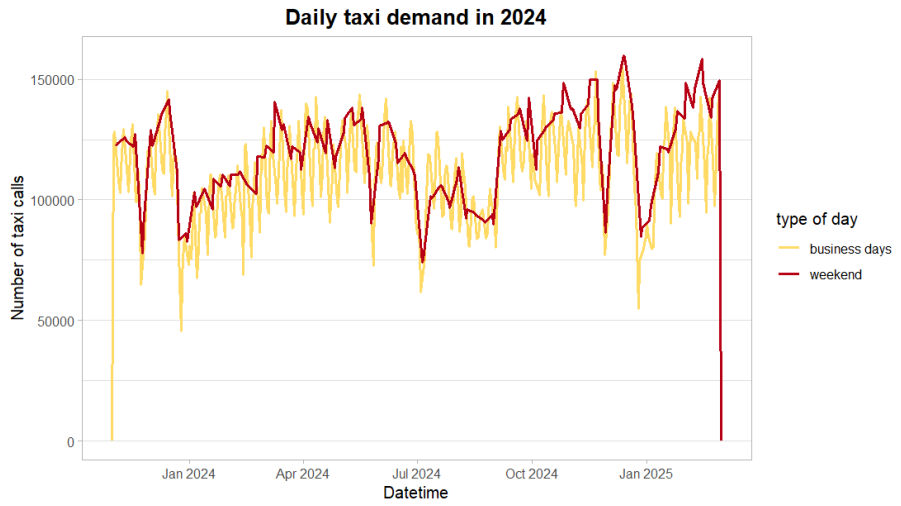


Figure 4.2. Lineplot of daily taxi demand in 2024.

4.1 Fitting Prophet

Initially, default options are selected, which involve 25 change points, and orders 10, 3 and 4 for the Fourier series of the yearly, weekly and daily seasonalities, respectively, and prior scales are set equal to 0.05 for trend change parameters and equal to 10 for seasonal parameters. Growth is assumed linear, not logistic. Notice that including a yearly seasonality for periods shorter than 12 months might induce the software to raise a warning: this can be serenely ignored, since the estimates are only related to the total number of observations, not to the number of periods. The only assumption that preferably must be held is that every after 365.25 days yearly variations cancel out, but this is quite credible, even without having the chance to check it out empirically with these data.

However, 16 months of observations are available, so this warning arises only during model evaluation through SHF. Figure 4.3 shows the adaptation of the model to the data, and figure 4.4 zooms in on a shorter time frame (20th October - 11th November, 2024), just for visualization purposes.

Data points follow so nicely the predictions, that a coefficient of determination $R^2 = 0.7967$ is calculated. Predictions here are to be intended as the mean of the posterior predictive density. Single model components are illustrated in figure 4.5a,

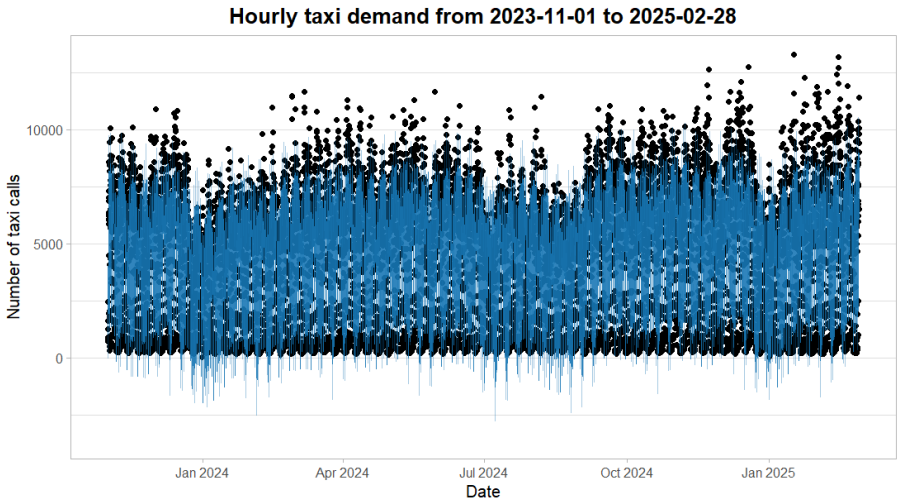


Figure 4.3. 80% prediction intervals of the fitted model.

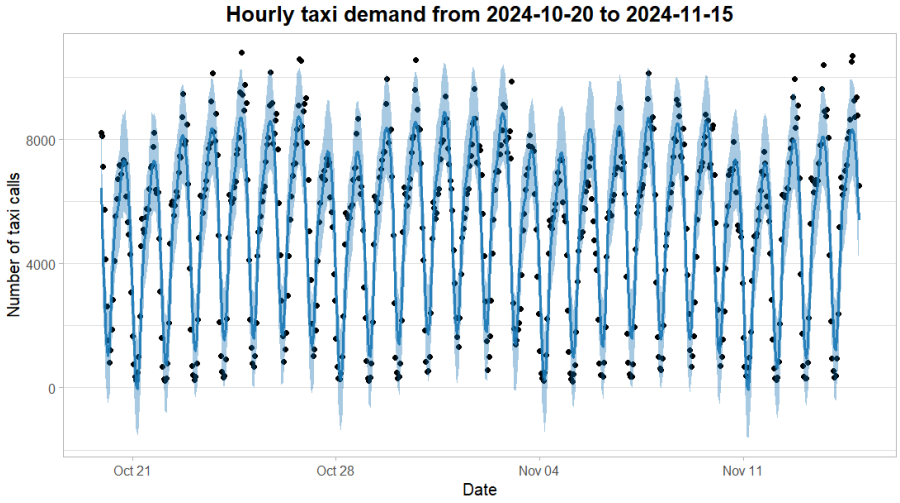
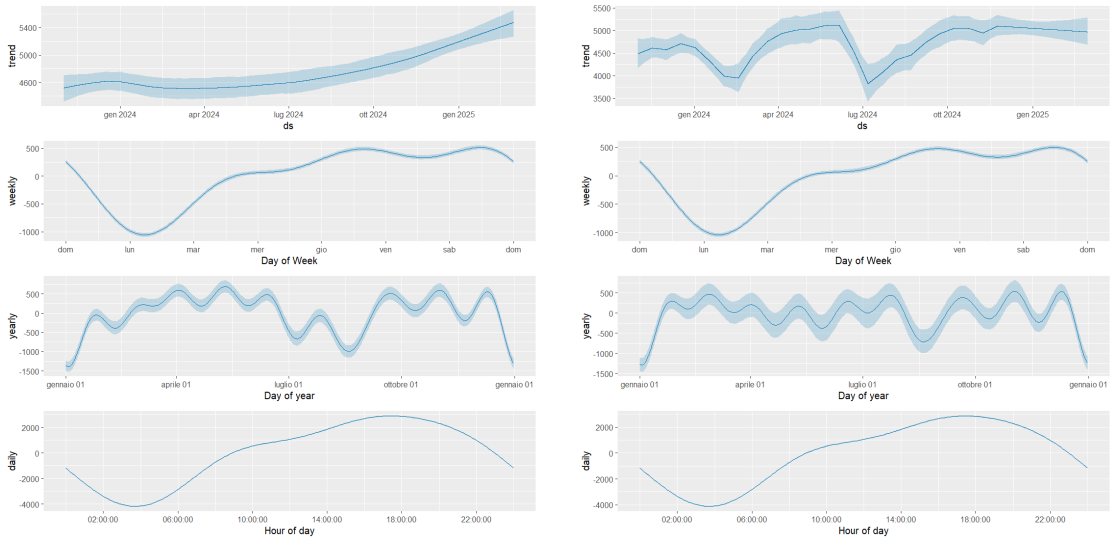


Figure 4.4. 80% prediction intervals of the fitted model on a smaller time window.



(a) Original model **(b)** Cross-validated model

Figure 4.5. Model components with an 80% uncertainty.

with their relative 80% uncertainty intervals.

Cross-validation is performed to identify the optimal parameters and hyperparameters. Prophet documentation suggests tuning the change point prior scale and the seasonality prior scale for regularization. Values 0.5 for the former and 0.01 for the latter result in the lowest RMSE and MAE at a 24-hour horizon. Here, the variability is caught by more pronounced trend changes within the year, rather than by an annual seasonality, as figure 4.5b shows.

It is also possible that a longer time interval would have lead to different conclusions.

4.2 Unsupervised anomaly detection

In unsupervised anomaly detection, there is no anomaly label over the observation, and the task of the analyst is to try to reconstruct the separation between outliers and normal observation. A quick review of state-of-art methods for unsupervised anomaly detection in time series can be found in Mejri et al. (2024). The approaches adopted in this chapter are point-based and reconstruction-based, that means that

an observation y_i is labelled as an anomaly if $|e_i| = |y_i - \hat{y}_i| > \delta$, in other words, the focus is kept on the residuals from the estimated baseline and the ones which exceed a threshold are catalogued as contextual outliers.

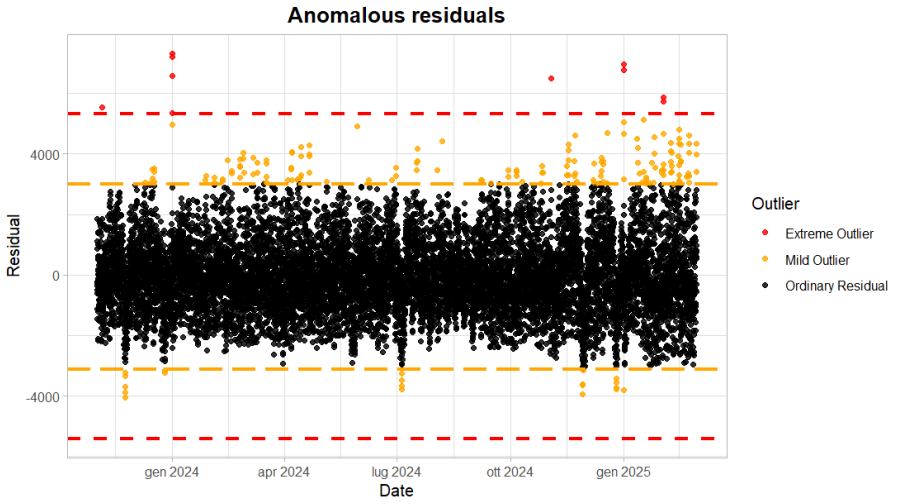


Figure 4.6. Residual vs. Date scatter plot with thresholds for outliers.

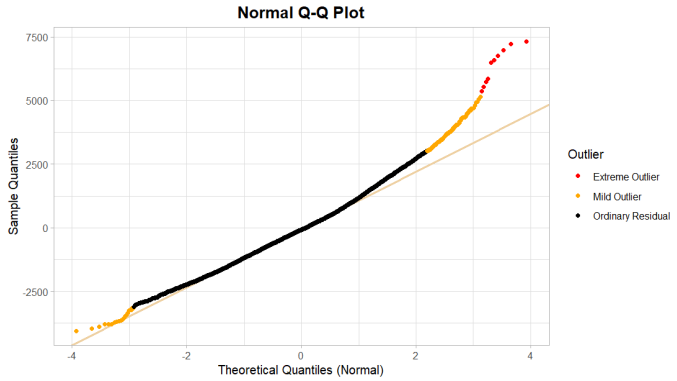


Figure 4.7. Q-Q plot of residuals vs. normal distribution.

4.2.1 Tagging through interquartile range (IQR). The first method is a classic quantile-based strategy that catalogues anomalies with the following rule:

$$Anomalous(y_i) = \begin{cases} 0 & \text{if } q_1 - k \cdot IQR \leq y_i \leq q_3 + k \cdot IQR \\ 1 & \text{otherwise} \end{cases} \tag{4.1}$$

with q_1 and q_3 the first and the third empirical quartiles of the residuals, respectively, and $IQR = q_3 - q_1$ is the interquartile range. Common choices for k are $k = 1.5$

Date and hour	y	$\hat{y}$	$\lvert y - \hat{y} \rvert$
2024-01-01 01:00	7355	33.6	7321
2024-01-01 02:00	6220	−989	7209
2025-01-01 01:00	8468	1494	6974
2025-01-01 02:00	7257	493	6764
2024-01-01 03:00	4936	−1639	6575
2024-11-03 01:00	9869	3376	6493
2025-02-01 23:00	11347	5488	5859
2025-02-02 00:00	10023	4294	5729
2023-11-05 01:00	8435	2888	5547
2024-01-01 00:00	6596	1232	5364

Table 4.1. All 10 extreme outliers detected.

(mild outliers) or $k = 3$ (extreme outliers). For normally distributed samples this method is roughly equivalent to label as anomalies observations whose absolute value is more than $0.674 + 1.349k$ standard deviations away from the mean. However, in this case, the sample distribution of the residuals seems to have a heavier right tail than a normal distribution, as Q-Q plot 4.7 suggests. Out of $n = 11643$ total counts, 186 (1.59%) are marked as anomalies, of which 176 (1.51%) are mild outliers and 10 (0.08%) are extreme outliers. From table 4.1 stands out that six of the ten "extreme anomalies" occur on 2024 and 2025 New Year's Eve. Interestingly, 5th November 2023 and on 3rd November 2024 were the dates of the 52nd and 53rd edition New York City Marathon, respectively. What happened on the night between 1 and 2 February 2025 is yet to be investigated, nevertheless there seems to be enough evidence to believe that these biased results are likely a consequence of not including the holiday component.

4.2.2 Histogram-based outlier score (HBOS). Another method introduced by Goldstein and Dengel (2012) is based on the height of each bin in the histogram of the observations. Each observation - a residual of the model - is placed in a bin of the histogram and an anomaly score inversely proportional to the bin height is assigned to it.

$$HBOS(y_i) = -\sum_{j=1}^{d} \log(hist_j(y_i)) \tag{4.2}$$

where $hist_j(\cdot)$ is the height of the histogram of feature j, normalized to be between 0 and 1, and d is the number of features, $d = 1$ in this case of univariate time series. The authors of this technique argue that both static fixed-width bins and dynamic bin-width histograms can be used, with the due attention to the shape of the distribution. The former settles k break points over the value range, the latter fills each bin with $\left\lfloor \frac{n}{k} \right\rfloor$ observations. As a rule of thumb, $k \approx \sqrt{n}$, although other common choices are Sturges's rule of $k = 1 + \lfloor \log_2(n) \rfloor$ or Friedman-Diaconis' rule of Bin width $= 2\frac{IQR}{\sqrt[3]{n}}$. Sticking to $k = \lfloor \sqrt{n} \rfloor = 107$, figure 4.8 shows HBOS

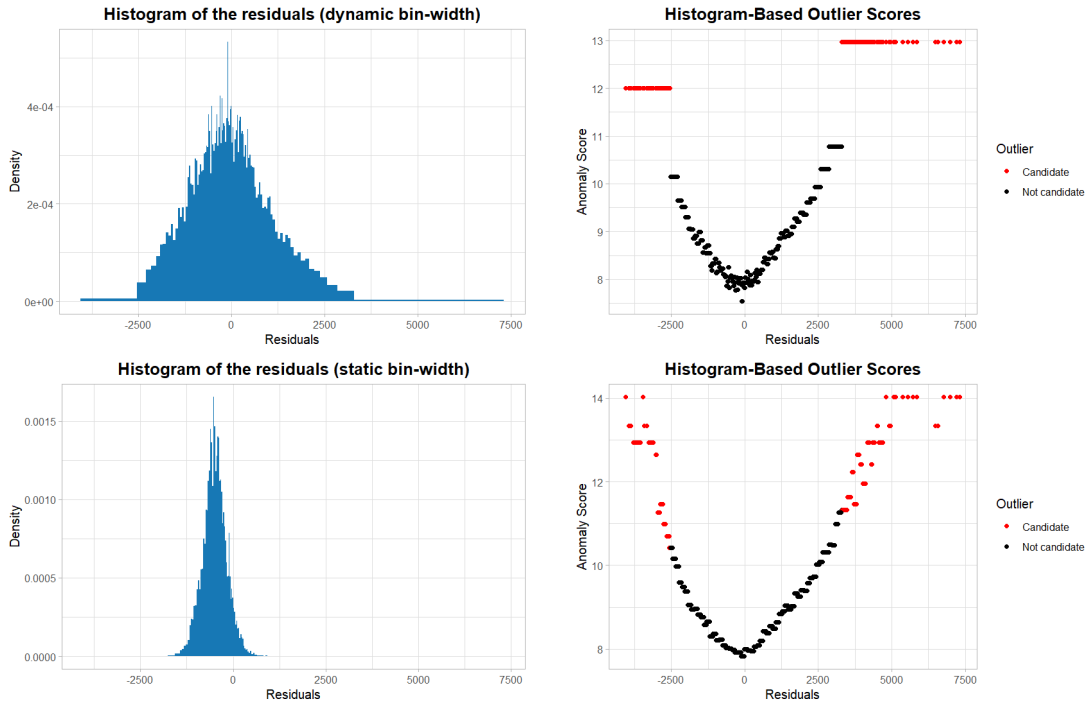


Figure 4.8. Histograms and HBOSs of the residuals with a dynamic (top) and a static (bottom) binning method. Residuals that exceed a HBOS threshold of 11.5 with the first histogram are coloured in red. The same colouring is kept in the static bin-width histogram.

measured with two different histogram binning strategies. Whereas the dynamic type of histogram always treats equally chunks of 109 residuals, the constant bin-width is less sensitive to extreme values, when they are aggregated into tight clusters of two or more points.

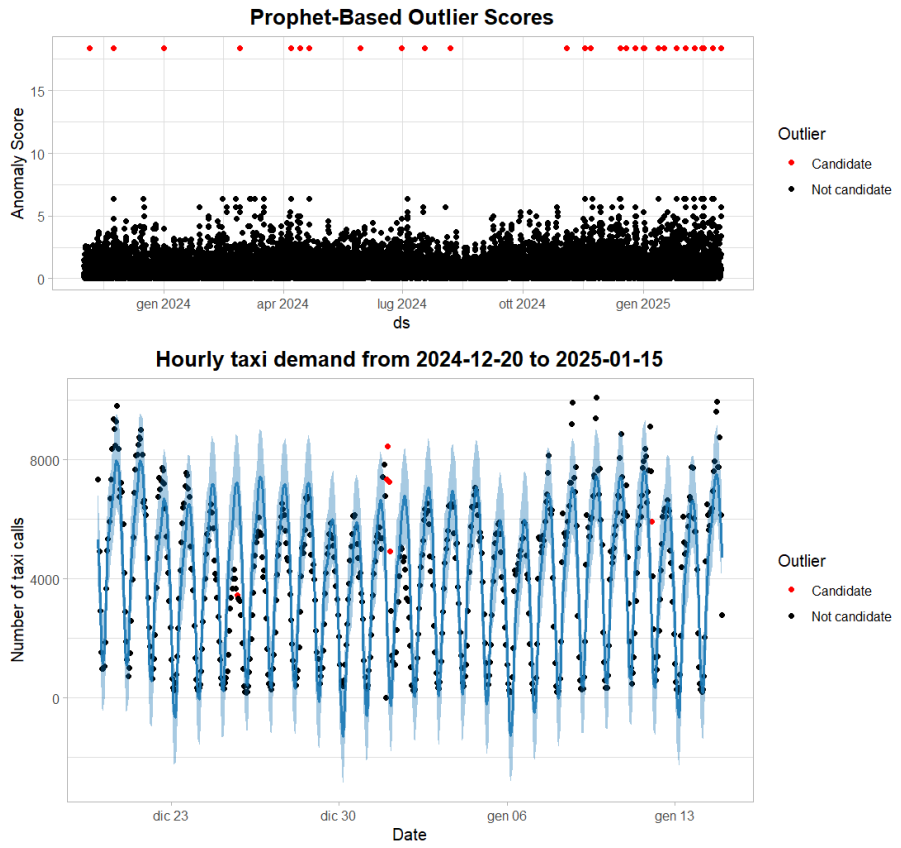


Figure 4.9. Anomaly scores for the whole time series (top) and anomalous values reported within a time window (bottom).

4.2.3 Prophet based approach. The last approach is based on the prediction intervals that a fitted Prophet model provides, that makes it perhaps the most consistent with the whole set up. The core idea is to consider the posterior predictive density and see how extreme an observation is under a specified model.

Given $p_t = 2 \cdot \min\{Pr(Y_t \leq y_t),\ Pr(Y_t \geq y_t)\}$, a metric of *surprise* is proposed as anomaly score:

$$Anomaly\ Score(y_t) = -\log(p_t)\,. \tag{4.3}$$

For each observation, a sample of 1200 simulations was drawn from the full posterior predictive density, and an approximation of the probability of seeing an at least as extreme value was computed as twice the minimum between the number of simulations greater than the real observation and the number of simulated values that

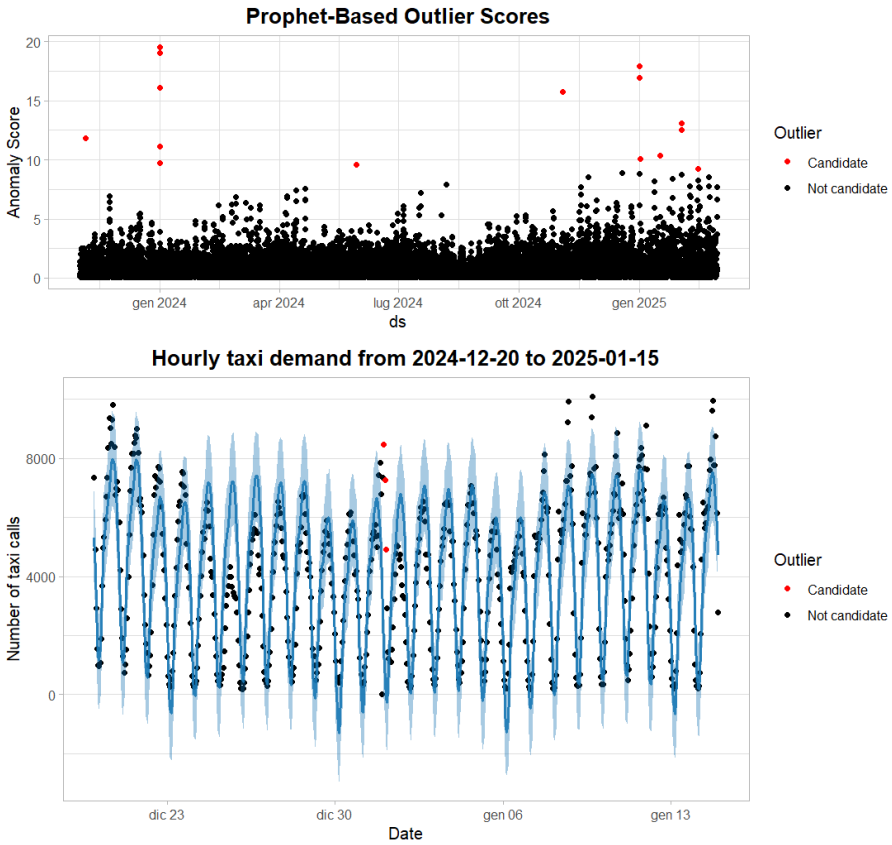


Figure 4.10. Anomaly scores fore the whole time series (top) and anomalous values reported within a time window (bottom), calculated with a mixed theoretical and computational approach.

were smaller, divided by 1200.

Although this approach is substantially the most formal way to proceed in a Bayesian framework, it is not very accurate for estimating small probabilities, since it is very likely to obtain 0 as an answer, when the true probability is much smaller than 1/1200, or whatever the simulated sample size is. Every time this happened, a 10^{-8} continuity correction was added.

Nevertheless, observations are declared to be normally distributed, hence this information can be used to compute a more precise estimate of p_t. The exact form of a posterior predictive density has the form

$$ppd(y_t) = \int_{\mathbb{R}\times\mathbb{R}^+} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_t-\mu_t}{\sigma}\right)^2} \pi(\mu_t, \sigma \mid \boldsymbol{y})\, \mathrm{d}\mu_t \mathrm{d}\sigma \tag{4.4}$$

where $\mu_t = g(t) + s(t)$ from equation 3.1, and σ is the scale of the error term. $\pi(\mu_t, \sigma \mid \boldsymbol{y})$ has no close form, but a sample from the posterior is available, and can be exploited to help building a combination of normal distributions, whose cumulative density function is relatively easy to calculate. Given

$$p_t \mid \mu_t, \sigma = 2\left[1 - \Phi\left(\left|\frac{y_t - \mu_t}{\sigma}\right|\right)\right] \tag{4.5}$$

it immediately follows

$$p_t = \mathbb{E}\left[p_t \mid \mu_t, \sigma\right] \approx \frac{2}{B}\sum_{j=1}^{B}\left[1 - \Phi\left(\left|\frac{y_t - \tilde{\boldsymbol{\mu}}_{tj}}{\tilde{\boldsymbol{\sigma}}_j}\right|\right)\right] \tag{4.6}$$

where B is the size of the sample from the posterior, in this case $B = 600$ and $\tilde{\boldsymbol{\mu}}_t$ and $\tilde{\boldsymbol{\sigma}}$ are the samples themselves. The cutoff for a candidate was arbitrarily set on $p_t < 1/10000$, although in a this long time series, and this specific threshold, one outlier is to be expected.

Figures 4.9 and 4.10 report mostly the same anomalous points, though, because of the above-mentioned reasons, the second method looks more robust to randomness. The first ten anomalies are reported in table 4.2. For clarity, all anomalies are also displayed more closely in figure 4.11.

As it could be expected, these results are coherent with table 4.1, indeed they are

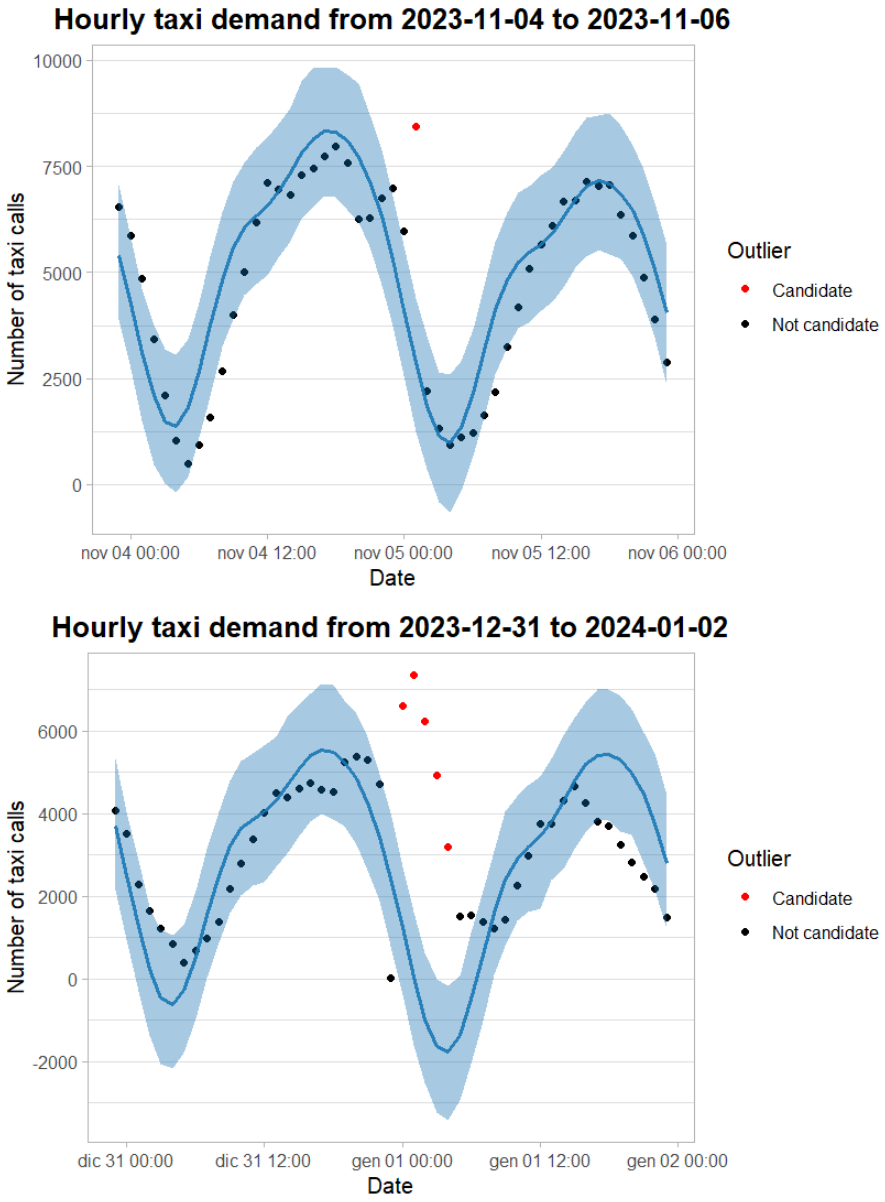


Figure 4.11. Zoom-in on the context of some of the anomalies.

Date and hour	y	$\hat{y}$	$y - \hat{y}$	p	Score
2024-01-01 01:00	7355	33.6	7321	3.17e-9	19.6
2024-01-01 02:00	6220	-989.0	7209	5.46e-9	19.0
2025-01-01 01:00	8468	1494.0	6974	1.66e-8	17.9
2025-01-01 02:00	7257	493.0	6764	4.39e-8	16.9
2024-01-01 03:00	4936	-1639.0	6575	1.04e-7	16.1
2024-11-03 01:00	9869	3376.0	6493	1.48e-7	15.7
2025-02-01 23:00	11347	5488.0	5859	2.11e-6	13.1
2025-02-02 00:00	10023	4294.0	5729	3.53e-6	12.6
2023-11-05 01:00	8435	2888.0	5547	7.24e-6	11.8
2024-01-01 00:00	6596	1232.0	5364	1.42e-5	11.2

Table 4.2. Top 10 anomalous observations with residual and score.

exactly the same here, but unlike the quantile-based approach, the Bayesian way is more quantitative and has a clearer explanation to what an outlier is defined, with an explicit probabilistic interpretation. Moreover, it takes the additional parameter uncertainty into consideration, which may usually vary over time, correctly making the anomaly score not depending only on the absolute value of the residual.

5 Discussion and conclusions

Prophet is a flexible approach to time series modelling that catches non-linear trends and nested seasonalities and has a wide range of application. The authors of Prophet designed it for "forecasting at scale", however, predictions are in some contexts not as interesting as a better understanding of the phenomenon. Luckily, Prophet handles this aspect at least as good as any classic linear model: the software provides methods for exploring visually and quantitatively trend, trend changes and all types of seasonalities, so that little statistical knowledge is required. Actually, a bold claim supported by Cotton (2021) is that Prophet works better for this kind of descriptive tasks rather than for out-of-sample predictive purposes. In any case, having a good base model, that is "robust to missing data and shifts in trend and handles outliers well" (Meta, 2023), is an excellent starting point to carry out an anomaly detection investigation. For instance, in the dataset of the taxi hires, Prophet seems to perform well, and all anomaly detection methods presented previously supply valuable information about the City of New York: events like New Year's Eve and New York Marathon boost the demand for taxi cabs. In particular, the Prophet-based approach to outlier detection seems the most promising, at least methodologically; using MAP estimates for point prediction and using residuals to tag anomalies basing on empirical quantiles is instead more user-friendly and computationally much lighter, while still looking nearly identical in the results to the first technique; histogram-based anomaly scores are perhaps not very suitable for this case study, although they can be applied to multivariate data sets and also have a natural evolution in kernel density estimation (KDE) based algorithms (see Blázquez-García et al., 2021, or Frehner et al., 2024, for example).

Of course, some criticisms raised against Prophet cannot be missing. It was already mentioned that a Prophet fit might have predictive issues, especially if the last estimated trend slope is very steep and therefore it under-estimates or over-estimates unreasonably future observations. This might be addressed by using a logistic growth, a "flat" trend (i.e., no trend), or by regularizing trend change parameters, or by setting past and future change points manually, with the support of domain knowledge. The second concern is about speed: as with most Bayesian

models, Prophet is also utterly slow, especially when a MCMC sample is drawn. A little workaround is to store the fitted model in memory as a file, though, as soon as data is streaming, the model must be entirely re-fitted each time, unlike with a classic linear model, which provides fast analytical solutions to this problem. The third major issue concerns distributive assumptions. Prophet was developed by researchers at Facebook, and the example reported in the original paper was about the number of events on the renowned social network. The case study presented in this thesis is somehow analogous, because it represents a business time series generated by human actions as well. As a matter of fact, the model is quite adequate, nonetheless other data with very similar features have poorer performance measures - not completely off target though - due to too relatively small values and "zero-inflation". Even in the taxi time series there is a low request during nighttime hours, yielding a number of negative predictions, which are meaningless in a context where the outcome is known to be a non-negative integer. A Poisson generalized linear model seems a principled and reasonable advancement to base Prophet: the mean could actually be the inverse link function of the linear combination of the trend, seasonality and holiday components. Another subtler aspect about the likelihood is that in a standard Prophet model error terms are independent one from another, but this hypothesis might not hold in the new york taxi case study. It can be seen from figure 4.6, where there are lots of residuals that are consecutively above and below x-axis. Correlated errors imply auto-correlated residuals, anomaly scores and anomaly tags, so it's surely not by chance that in table 4.1 four outliers happen at four consecutive hours.

Unfortunately, Prophet maintainers are not prone to make any large changes to the underlying model, however, the project was released as open-source in 2017 and since then any forks by the community are welcomed. Related to the topic, all data and R code used for the analysis are made available at https://github.com/GioZd/nyc-taxi for reproducibilty.

Bibliography

Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data [Online Real-Time Learning Strategies for Data Streams]. *Neurocomputing*, *262*, 134–147. https://doi.org/https://doi.org/10.1016/j.neucom.2017.04.070

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM computing surveys (CSUR)*, *54*(3), 1–33.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Cotton, P. (2021). Is Facebook's "Prophet" the time-series Messiah, or just a very naughty boy? - DataScienceCentral.com — datasciencecentral.com [Accessed 22-05-2025].

de Finetti, B. (1933). Sul concetto di probabilità. *Rivista italiana di statistica, economia e finanza*, *5*, 723–747.

di Fonzo, T., & Lisi, F. (2005). *Serie storiche economiche: Analisi statistiche e applicazioni*. Carocci editore.

Duong, C. (2023). Facebook/Prophet in 2023 and beyond [Accessed 22-05-2025].

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, *34*(4), 1386–1403.

Freedman, D. A. (1999). *On the Bernstein-von Mises theorem with infinite dimensional parameters* [From 1998 Wald Lectures].

Frehner, R., Wu, K., Sim, A., Kim, J., & Stockinger, K. (2024). Detecting anomalies in time series using kernel density approaches. *IEEE Access*, *12*, 33420–33439. https://doi.org/10.1109/access.2024.3371891

Goldstein, M., & Dengel, A. (2012). Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, *106*(4), 620–630.

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in R and Stan* (2nd edition). CRC Press.

Mejri, N., Lopez-Fuentes, L., Roy, K., Chernakov, P., Ghorbel, E., & Aouada, D. (2024). Unsupervised anomaly detection in time-series: An extensive evaluation and analysis of state-of-the-art methods. *Expert Systems with Applications*, *256*, 124922. https://doi.org/https://doi.org/10.1016/j.eswa.2024.124922

Meta. (2023). Prophet — facebook.github.io [Accessed 22-05-2025].

Pettigrew, R., & Weisberg, J. (Eds.). (2019). *The open handbook of formal epistemology* [Open-access book]. PhilPapers Foundation.

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramachandran, K. M., & Tsokos, C. P. (2015). Chapter 1 - descriptive statistics. In *Mathematical statistics with applications in R (second edition)* (Second Edition, pp. 1–52). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-417113-8.00001-1

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

TLC Trip Record Data - TLC — nyc.gov [Accessed 13-05-2025]. (2025).

Von Mises, R. (1941). On the foundations of probability and statistics. *The Annals of Mathematical Statistics*, *12*(2), 191–205.