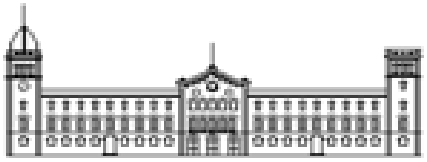

Undergraduate Final Thesis

Bachelor of Science in Mathematics
Bachelor in Business Administration and Management

Faculty of Mathematics and Computer Science
Faculty of Economics and Business
Universitat de Barcelona

Queuing Theory, Time Series and an application to Tollgate's Traffic Flow Prediction

Author: Miguel Romero Calvo

Directors:	Dr. Josep Vives Santa Eulalia
	Dra. Marién Encarnación André Romero
Realized at:	Dpt. of Mathematics and Computer Science
	Dpt. of Econometrics, Statistics and Applied Economy
Barcelona,	January 19, 2018

Greetings

I would like to greet my family and friends for their unconditional support during those bachelor degrees, specially to my mom and dad who encouraged me during those years. I would also like to thank Dr. Josep Vives Santa Eulalia and Dr. M. Encarnación Andre Ramon for guiding and helping me in this project. Last but not least, I would also like to thank Abertis Infraestructures S.A. for providing me the necessary data to implement the model.

Abstract

By introducing a methodology for forecasting the minimum amount of open gates that will be needed for a given period This paper aims to provide a useful resource to support operational management in highway companies. With this aim, it develops basic notions, results in Queuing Theory and presents and explain useful tools for a Classical Analysis of Time Series. In the practical part the model is developed, implemented in R, tested in the last trimester of 2017 and used to predict the hourly amount of minimum open gates needed for the first two months of 2018.

Contents

1 Introduction

Traffic is volatile. Rushing hours, holiday periods, accidents and even weather conditions can influence the amount and speed of cars. Industries working in this area, for example highway companies, have a thought times optimizing their resources under those conditions. This aspect is particularly important for companies managing toll gates. Those companies need to estimate, not only the traffic flow but also the ratio arriving cars to served cars, and consequently, the repercussions of investing more or less resources at each moment. Too much opened toll gates is a waste of resources for the company while too less cause bottlenecks, traffic jams and possible reductions in cars going through the same tollgate in the future.
This project aims to clarify this uncertainty and support the operational department of those companies by an early approach to the problem. The first part of the paper provides and proof the theory where the practical part is based.
Section 2.1 builds queuing system from the beginning. while providing basic notions, definitions and results, the usual methodologies and measures to analyze queuing systems are explained. However, it focuses on the development of long-run measures, the M/M/1 queue and the M/M/c queue so they constitute the necessary tools in queuing theory for the applied section.
Section 2.2 provides the basic notions to isolate, approximate and extract the trend and the cyclical effects of Time Series according to the Macroscopic or Classical approach. In this section tools such as centered and decentered moving averages, differentiation of time series or the $log-$transformation are presented and explained.
Section 3 develops and implements (6) the model for the tollgate situated at Mollet del Vallès. The model is able to forecast the number of tollgates that will be needed, by hour, in the following time-period according to a choose criteria, for example, to do not exceed certain length of queues with a desired confidence. The methodology presented to do so is based in two main steps:

1. Forecasting of the total amount of arrivals per hour.

2. Computation of the minimum number of gates needed to achieve a specific queuing measure.

The predictions of both, the number of arrivals and the minimum number of gates needed, are tested in this section to asses the accuracy of the model.

2 Theory

2.1 Queuing Theory

The starting point for queues analysis is the definition of what is considered a queuing system:

Definition 2.1 *A queuing system is defined as the composition of N queues and M servers.*

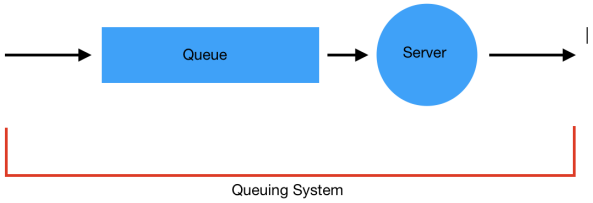

Figure 1: Queuing system with one queue and one server

Queuing Theory aims to describe the performance of a queuing system through different measures depending on the business interest. Therefore information relative to the number of servers, the amount of arrivals and the amount of departures or served costumers must be assumed to be provided somehow.

Indicators are classified into two main categories. On one hand dynamic indicators over time based on simulations provide the best way forward to analyze a queue performance and it's variance in a short period of time. On the other hand, long-run measures describe the asymptotic performance of the system when certain situations are given.

As will be developed in the upcoming sections, queuing analysis is mainly based on three objects:

1. **Distribution of the arrival times**: Main focus of this paper.
2. **Distribution of the service time**: Can be estimated for each individual toll gate.
3. **Number of services available**: Can be modified by the company depending on the obtained results.

2.1.1 Given data and notation

Kendall's Notation, $A/S/c/L$, is the standard way to describe the general features of a queuing system [3], where:

- A: The usual nomenclature is M if the inter-arrival times distribution is memoryless, M^x if each costumer brings a batch of jobs and the inter-arrival time distribution is memoryless, and G if it's any general distribution where each costumer brings a job.
- S: It can be an M if the service-time distribution is memoryless or G if the service-time distribution is any general distribution.
- c: It's the number of servers present to process jobs.
- L: In case the queue have a maximum possible length it is indicated here.

Establishing $\mathbb{N} = \{1, 2, 3, ...\}$ and $\mathbb{N}^* = \{0, 1, 2, 3, ...\}$, in this section we will deal with how information can be given and how to change from one representation to another [1]. Regarding arrivals it can be presented in terms of:

- $\{X_k\}_{k\in\mathbb{N}}$, X_k time between the arrival of costumers.
- $\{A_k\}_{k\in\mathbb{N}^*}$, A_k arrival time of the k-th costumer, $A_0 = 0$.
- $\{A(t), t \geq 0\}$ arrival process counting the number of jobs arrived in $[0, t]$.

We can transform one type of data into another as follows:

- $X_k = A_k - A_{k-1}$, $k > 0$
- $A_k = \sum_{j=1}^{k} X_j$, $k > 0$ and $A_0 = 0$
- $A(t) = \sum_{k=1}^{\infty} \mathbb{1}_{\{A_k \leq t\}} = \max\{k : A_k \leq t\}$

Regarding service times and departure times:

- $\{S_k\}_{k\in\mathbb{N}*}$, where S_k is the service time of the k-th costumer.
- $\{D_k\}_{k\in\mathbb{N}*}$, where D_k is the departure time of the k-th costumer.

- $\{D(t), t \geq 0\}$ departure process counting the number of jobs departed in $[0, t]$.

Where

- $D(t) = \sum_{k=1}^{\infty} \mathbb{1}_{\{D_k \leq t\}} = \max\{k : D_k \leq t\}$
- $D_k = \min\{t : D(t) = k\}$

2.1.2 Exponential and Poisson distributions

In this section we look at the connexion between Poisson and Exponential distribution and a queuing system. They are related via the arrival distribution. Usually, queuing system with inter-arrival times exponentially distributed have the following common features:

1. Arrivals are drawn from a large population.
2. Each of the arriving costumers decide, independently of the others, to visit the system.

The set of all inter-arrival times $\{X_i\}_{i \in \mathbb{N}}$ are independent and identically distributed in this situation.
Moreover empirical measurements in this kind of real-life queuing systems have shown that it's reasonable to model inter-arrival times with the exponential distribution.

Definition 2.2 *[2] A random variable X has a exponential distribution, $X \sim Exp(\lambda)$ if its density function is:*

$$f(x) = \begin{cases} 0 \ , \ if \ x < 0 \\ \lambda e^{-\lambda x}, \ if \ x \geq 0 \end{cases}$$

or equivalently its is distribution is:

$$F(x) = \begin{cases} 0 \ , \ if \ x < 0 \\ 1 - e^{-\lambda x}, \ if \ x \geq 0 \end{cases}$$

where $\lambda \in [0, \infty)$ is usually called the rate in this context.

Note For $X \sim Exp(\lambda)$, $P(X > t) = 1 - F(t) = e^{-\lambda t}$. It's expectation is $\mathbb{E}[X] = \frac{1}{\lambda}$ and it's variance is $\mathbb{V}ar[X] = \frac{1}{\lambda^2}$.

A fundamental property of the exponential distribution and basic for its importance in this theory is the Memoryless Property (2.3). This property translates into the theory that the time until the next arrival is independent of the time we have already been waiting.

Property 2.3 (Memoryless) *[2] Given $X \sim Exp(\lambda)$ then:*

$$P(X > t + s | X > s) = P(X > t)$$

Proof 2.3

$$P(X > t + s | X > s) = \frac{P(X > t + s)}{P(X > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t).$$

When data regarding arrivals is presented as $\{A_k\}_{k \in \mathbb{N}}$ the consideration of a counting process pops-up naturally. In our case it will be called arrival process.

Definition 2.4 *[2] We define the arrival process $A(s,t)$, $s, t \in [0, \infty)$ and $t \geq s$ as the number of arrivals in $(s,t]$.*

$$A(s,t) = \sum_{k=0}^{\infty} \mathbb{1}_{\{s < A_k \leq t\}}$$

In our case we will set $s = 0$ and notate $A(0,t) = A(t)$.

Assumption 1 makes reasonable to consider the expected arrival rate somehow proportional to the period's length:

$$\mathbb{E}[\mathrm{N}(s,t)] = \lambda(t - s), \lambda > 0$$

while assumption 2 supports that $\{A(s,t), s < t\}$ should have stationary and independent increment. Under those assumptions the arrival process is a Poisson process, i.e. Poisson distribution is used to model the number of arrivals in a period of time with independent and stationary increments. For a further justification we refer the reader to Annex 1 (5.1).

Definition 2.5 *[4] A random variable $X \sim Poiss(\lambda)$ if its probability function is $P(X = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$*

Note The Poisson's expectation and variance is $\mathbb{E}[X] = \mathbb{V}ar[X] = \lambda$.

The counting process and the inter arrival time distribution are strictly related as shown by Theorem 2.6.

Theorem 2.6 *[4] Given a counting process* $\{N(t), t \geq 0\}$ *and a set of inter-arrival times* $\{X_i, i \in I\}$ *of the same queuing system, then:*

$$N(t) \sim Pois(\lambda t), \forall t \geq 0 \iff X_i \sim exp(\lambda), i \in I$$

Proof Let N(t) be a Poisson process with rate λ, i.e. for any time $t > 0$, $N(t)$ follows a Poisson distribution. Given that the process started with zero, the probability that the count remains at zero at time t is

$$P(N(t) = 0) = e^{-\lambda t}$$

Notice that the event that the counting process remains at zero until time t, i.e. $\{N(t) = 0\}$, is equivalent to the event that the first arrival is after t, i.e. $\{A_1 > t\}$.

$$P(N(t) = 0) = P(A_1 > t) = e^{-\lambda t}$$

As $A_1 = X_1$ we have our result for the first inter-arrival time. To check the result for the rest of the inter-arrival times we need check that:

$$P(N(A_k + s) = n \mid N(A_k) = n) = \frac{P(N(A_k + s) = n, N(A_k) = n)}{P(N(A_k) = n)} \tag{1}$$

Solving the numerator

$$\begin{aligned} P(N(A_k + s) = n, N(A_k) = n) &= P(N(A_k + s) - N(A_k) = n - n, N(A_k) = n) \\ &= P(N(s) = 0)P(N(A_k) = n). \end{aligned} \tag{2}$$

Where we have used the properties of independent increments and stationarity of the arrival process in the second equality. Hence by substituting (2) in (1) the result is proved:

$$\begin{aligned} P(N(A_k + s) = n \mid N(A_k) = n) &= \frac{P(N(A_k + s) = n, N(A_k) = n)}{P(N(A_k) = n)} \\ &= \frac{P(N(s) = 0)P(N(A_k) = n)}{P(N(A_k) = n)} = \frac{(\lambda s)^0}{0!} e^{-\lambda s} \\ &= e^{-\lambda s}. \end{aligned} \tag{3}$$

2.1.3 Construction of Discrete-Time Queuing Process

In this section we will develop a simple recursion used to analyze the performance of a queue by a time discretization. Our final goal will be to obtain

the queue length over the time. E.g. By simulation it's possible to analyze the length of a queue in a highways' tollgate each 5 minutes.

The general way to proceed is:

1. Chop-up time in periods.
2. Develop recursions for the behavior of the queue from period to period.

Observe that considering discrete and fixed periods of time implies fixed inter-arrival times.
Assuming that jobs arrived at period k can't be served in this period.

$$d_k = \max\{Q_{k-1}, c_k\}$$
$$Q_k = Q_{k-1} + a_k - d_k$$

Where

- a_k is the cardinal of jobs arriving at period k.
- c_k is the maximum amount of jobs that can be served at period k.
- d_k is the cardinal of jobs departing the queue at period k.
- Q_k is the cardinal of jobs in queue at the end of period k.

2.1.4 Construction of the G/G/1 queuing process in continuous time

Suppose our initial data are the continuous time variables $\{X_k, k = 1, 2, ...\}$ and $\{S_k, k = 1, 2, ...\}$ of a $G/G/1$ queuing process.
The goal in this section is develop a recursion for the waiting time of each costumer and a general formula in continuous time for the queue length.

To begin with a recursion for its waiting time of each costumer observe that for any costumer k sojourn time in the system is composed by the sojourn time in queue and the service time as is illustrated in figure 2.

Waiting time in queue can either be zero or larger depending on the server occupation. In case servers are busy the kth costumer must wait the remaining time in queue of the $(k-1)th$ costumer plus the service time of that same costumer $(k-1)$.

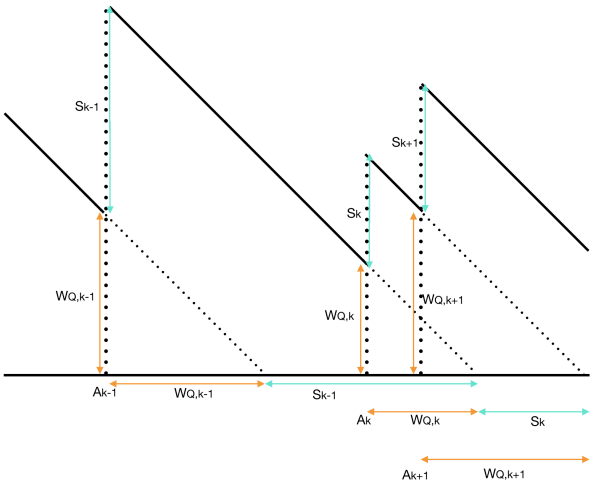

Figure 2: Waiting time in system

$$W_{Q,k} = [W_{Q,k-1} - X_k + S_{k-1}]^+ = \max\{W_{Q,k-1} - X_k + S_{k-1}, 0\}$$

Bringing the first costumer's waiting time in queue to zero all the waiting times in queue can be computed $\{W_{Q,k}, k = 0, 1, 2, ...\}$.

By this information multiple measures are immediate. The departure time from queue $D_{Q,k}$ must be $D_{Q,k} = A_k + W_{Q,k}$ what implies that costumer k leaves the system at time $D_k = D_{Q,k} + S_k$. And the sojourn time in system can then be expressed as $W_k = D_k - A_k = W_{Q,k} + S_k = D_{Q,k} + S_k - A_k$.

Secondly, to compute the queue length by a general formula we first need a new expression for the departure times. Let's define $D(\cdot) : \mathbb{R} \to \mathbb{N}$ as the number of departures in $[0, t]$.

$$D(t) = max\{k \in \mathbb{N} : D_k \leq t\} = \sum_{k=1}^{\infty} \mathbb{1}_{\{D_k \leq t\}} \tag{4}$$

And an general formulation for the queue length, $L(t) = L(0) + A(t) - D(t)$, logically follows. Furthermore observe that with this expression the number of jobs in the system as seen by the kth costumer at it's arrival can be computed substituting $t = A_k$.

Note It can be shown that the map $L(\cdot) : \mathbb{R} \to \mathbb{N}$ is right continuous.

In a queuing system jobs can be in the queue or in the service, therefore we will distinguish between:

- $L(t)$: Cardinal of jobs in the system at time t.
- $L_Q(t)$: Cardinal of jobs in the queue at time t, $L_Q(t) = L(0) + A(t) - D_Q(t)$.
- $L_S(t)$: Cardinal of jobs in the servers at time t $L_S(t) = L(t) - L_Q(t) = D_Q(t) - D(t)$.

2.1.5 Rate Stability and Utilization

In this section we will introduce long-run measurements that will help us to develop an asymptotic study of the queue.

Definition 2.7 *[3] The arrival rate of the system is defined as:*

$$\lambda = \lim_{t\to\infty} \frac{A(t)}{t} \tag{5}$$

Proposition 2.8 *[3]* $\{X_k, k = 1, 2, 3, ...\}$ *independent and identically distributed sequence of inter arrival times following a generic distribution* G_k *with a finite expectation. Then* λ *exists and is well defined.*

Proof 2.8
As $\frac{1}{n}\sum_{k=1}^{n} X_k$ is a positive increasing function over the $\mathbb{R}$ if we proof that the limit is finite then the limit must exist in $\mathbb{R}$.

$$\begin{aligned}\mathbb{E}[X] &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \lim_{n\to\infty} \frac{1}{n} A_n = \lim_{n\to\infty} \frac{A_n}{A(A_n)} \\ &= \lim_{t\to\infty} \frac{t}{A(t)} = \frac{1}{\lambda} \implies \lambda = \frac{1}{\mathbb{E}[X]} < \infty\end{aligned}$$

Definition 2.9 *[3] The departure rate of the system is defined as*

$$\gamma = \lim_{t\to\infty} \frac{D(t)}{t} \tag{6}$$

Proposition 2.10 *[3]* $\{X_k, k = 1, 2, 3, ...\}$ *independent and identically distributed sequence of inter arrival times following generic distributions* G_k *with a finite expectation. Then* γ *exists and is well defined.*

Proof 2.10
If the system is empty at time 0, then:

$$D(t) \leq A(t) \implies \gamma = \lim_{t\to\infty} \frac{D(t)}{t} \leq \lim_{t\to\infty} \frac{A(t)}{t} = \lambda < \infty$$

As it is a limit bounded in a close interval of $\mathbb{R}$ and $D(t)$ is an increasing function on t, the limits exist.

For the first approach to the service rate let us suppose a system with only one server. Total service time required by the first n jobs can be computed as a function $U_n : \mathbb{N} \to \mathbb{R}$ defined $U_n = \sum_{k=1}^{n} S_k$. On the other hand the number of jobs served until time t can be can by computed as a function $U_n : \mathbb{R} \to \mathbb{N}$ defined like $U(t) = \sup\{n : U_n \leq t\}$.

Definition 2.11 *[3] In a one service queuing system we can define the service rate as:*

$$\mu = \lim_{t\to\infty} \frac{U(t)}{t}$$

Proposition 2.12 *[3]* $\{S_k, k = 1, 2, 3, ...\}$ *independent and identically distributed sequence of service times following generic distributions* G_k *with a finite expectation. Then* μ *exists and is well defined.*

Proof 2.12
As $\frac{1}{n}\sum_{k=1}^{n} S_k$ is a positive increasing function over the $\mathbb{R}$ if we proof that the limit is finite then the limit must exist in $\mathbb{R}$.

$$\begin{aligned}\mathbb{E}[S] &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} S_k = \lim_{n\to\infty} \frac{1}{n} S_n = \lim_{n\to\infty} \frac{S_n}{S(S_n)} \\ &= \lim_{t\to\infty} \frac{t}{S(t)} = \frac{1}{\mu} \implies \mu = \frac{1}{\mathbb{E}[S]} < \infty\end{aligned}$$

Definition 2.13 *[3] A system is rate stable if and only if* $\lambda = \gamma$.

Note [3] Observe that when the arrival rate is higher than the departure rate the queue length tends to infinity:

$$\lim_{t\to\infty} L(t) = \lim_{t\to\infty} A(t) - D(t) + L(0) = \lim_{t\to\infty} \lambda t - \gamma t + L(0) = \infty.$$

Note $U(t) \leq D(t) \leq U(t) + 1$. Hence, dividing by t and taking limits we have that $\gamma = \mu$.

Corollary 2.14 *[3] In a queuing system we have:*

$$\gamma \leq \min\{\lambda, \mu\}$$

Moreover this implies that when the service rate is higher than the departure rate $\mu \geq \lambda$ from 2.1.5 and 2.14 implies that $\gamma = \lambda = \mu$.

It can be shown that even when $\mu = \lambda$ if $Var[S_n] > 0$ or $Var[X_n] > 0$ the long-run average length of the queue $\lim_{t\to\infty} \frac{L(t)}{t}$ does not necessarily exists. Therefore to work with a system we are going to require a service rate strictly higher than the arrival rate $\mu > \lambda$.

Definition 2.15 *[3] When the $\{S_k, k = 1, 2, 3, ...\}$ independent and identically distributed sequence of service times following generic distributions G_k with a finite expectation. We define the load or utilization ρ as:*

$$\rho = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{1}_{\{L(s)>0\}} ds$$

Proof 2.15
Existence of the utilization rate ρ. On one hand, notice that at time t them system has been under utilization for no more than the accumulated service time of all the costumers arrived until time t. Hence:

$$\sum_{k=1}^{A(t)} S_k \geq \int_0^t \mathbb{1}_{\{L(s)>0\}} ds \tag{7}$$

On the other hand, the time the system has been under utilization until time t can't be lower than the accumulated service time of the departure costumers until time t because that time the system has been in utilization.

$$\int_0^t \mathbb{1}_{\{L(s)>0\}} ds \geq \sum_{k=1}^{D(t)} S_k \tag{8}$$

Dividing the first summation by $\frac{1}{t} > 0$ and taking the limit in t to infinity from 2.12 the limit can be solved as follows.

$$\lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{D(t)} S_k = \lim_{t\to\infty} \frac{D(t)}{t} \frac{1}{D(t)} \sum_{k=1}^{D(t)} S_k = \gamma \mathbb{E}[S]$$

We can solve the limit in the second summation by the same procedure but with $A(t)$ instead of $D(t)$. Hence, taking together equations ((7)) and ((8)), dividing by t and taking the limit in t to ∞ we have:

$$\lambda\mathbb{E}[S] \geq \rho \geq \gamma\mathbb{E}[S]$$

As $\mathbb{1}_{\{L(s)>0\}}$ is an increasing positive function of t and the limit of the expression is bounded in a close interval of $\mathbb{R}$ it exists.

Proposition 2.16 *[3] If the system is rate stable then $\rho = \lambda\mathbb{E}[S]$ and by the identities $\mathbb{E}[X] = \frac{1}{\lambda}$ and $\mathbb{E}[S] = \frac{1}{\mu}$ we have:*

$$\rho = \frac{\lambda}{\mu} = \frac{\mathbb{E}[S]}{\mathbb{E}[X]}$$

Corollary 2.17 *It's equivalent to have a rate stable system and a system where the expected service time is lower than the expected arrival time.*

Proof 2.17
Given $\mu > 0$:

$$\rho < 1 \iff \frac{\lambda}{\mu} < 1 \iff \lambda < \mu$$

2.1.6 Limits of Empirical Performance Measures

In this section we will present long-run averages over number of arrivals or over time involving two features: waiting time and queue length. Further empirical performance measures can be defined assuming a rate stable system, in this case the following limits exists.

Definition 2.18 *[2] We can define the empirical expected waiting time in the system as:*

$$\mathbb{E}[W] = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} W_k$$

Definition 2.19 *[2] We can define the empirical expected waiting time in queue as:*

$$\mathbb{E}[W_Q] = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} W_{Q,k}$$

Definition 2.20 *[2] An empirical approximation of the cumulative distribution of the waiting time can be found by counting:*

$$P\{W \leq x\} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{W_k \leq x\}}$$

Definition 2.21 *[2] The average number of jobs in the system as seen by arrivals is given by:*

$$\mathbb{E}[L] = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} L_k \ , \ L_k = L(A_k)$$

Definition 2.22 *[2] The empirical cumulative distribution of the queue length as seen by customers upon arrival, is*

$$P\{L \leq m\} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{L_k \leq m\}}$$

Definition 2.23 *[2] Supposing no initial queue. The time-average number of jobs in the system can be defined*

$$\mathcal{L}(t) = \frac{1}{t} \int_0^t L(s)ds = \frac{1}{t} \int_0^t (A(s) - D(s))ds, \ , \ t \in \mathbb{R}_{\{\geq 0\}}$$

Assuming $\exists \lim_{t\to\infty} \mathcal{L}(t)$,

$$\mathbb{E}(\mathcal{L}) = \lim_{t\to\infty} \mathcal{L}(t)$$

Definition 2.24 *[2] The time-average fraction of time the system contains at most m jobs can be defined as:*

$$P\{L \leq m\} = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{1}_{\{L(s)\leq m\}} ds$$

2.1.7 Level Crossing and Balance Equations

First of all $L(0) = 0$ can be assumed for simplicity. Queuing system can also be approached as a Markov Chain, where an arrival can be seen as an up-crossing and a departure as a down-crossing.

In this context:

- System in state n if and only if it contains n jobs.
- System crossed level n if and only if states go from state n to state $n + 1$ or vice versa.

Relative to this approach, assuming only one job can arrive or depart in each instant, new terms can be defined as follows:

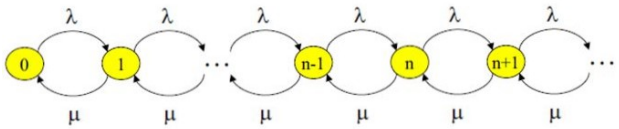

Figure 3: Waiting time in system

Definition 2.25 *[2] We can define the number of arrivals until time t that makes the costumer have $n+1$ jobs.*

$$A(n,t) = \sum_{k=1}^{\infty} \mathbb{1}_{\{A_k \leq t\}} \mathbb{1}_{\{A_k^- = n\}}$$

where A_k^- is the moment just before the $k-th$ costumer arrives.

Definition 2.26 *[2] Time that the system is in state n until t is*

$$Y(n,t) = \int_0^t \mathbb{1}_{\{L(s)=n\}} ds$$

Definition 2.27 *[2] The time that system is in state n relative to the total time until t is defined as follows.*

$$p(n,t) = \frac{1}{t} \int_0^t \mathbb{1}_{\{L(s)=n\}} ds = \frac{Y(n,t)}{t}$$

The concept of the arrival and departure rate are also generalized for every specific state as follows:

Definition 2.28 *[2] The arrival rate in state n is*

$$\lambda(n) = \lim_{t\to\infty} \frac{A(n,t)}{Y(n,t)}$$

Definition 2.29 *[2] By defining the departure rate to state n as:*

$$D(n,t) = \sum_{k=1}^{n} \mathbb{1}_{\{D_k \leq t\}} \mathbb{1}_{\{L(D_k)=n\}}$$

The departure rate from state $n+1$ is:

$$\mu(n+1) = \lim_{t\to\infty} \frac{D(n,t)}{Y(n+1,t)}$$

Theorem 2.30 *[2] In a rate stable queuing system the following equation holds:*

$$\lambda(n)p(n) = \mu(n+1)p(n+1)$$

Proof 2.30 Observe that costumers arrives and depart as single units. Therefore if we notate by T_k an arbitrary departure or arrival moment, the queue length must hold $L(T_k^-) - 1 \leq L(T_k) \leq L(T_k^-) + 1$ i.e. $\mid A(n,t) - D(n,t) \mid \leq 1$. Hence,

$$\lim_{t\to\infty} \frac{A(n,t)}{t} = \lim_{t\to\infty} \frac{A(n,t) \pm 1 \mp 1}{t} = \lim_{t\to\infty} \frac{A(n,t) \pm 1}{t} + \frac{\mp 1}{t} = \lim_{t\to\infty} \frac{D(n,t)}{t} \tag{9}$$

By solving each limit the result is proved.

$$\begin{aligned} \lim_{t\to\infty} \frac{A(n,t)}{t} &= \lim_{t\to\infty} \frac{A(n,t)}{Y(n,t)} \frac{Y(n,t)}{t} = \lambda(n)p(n) \\ \lim_{t\to\infty} \frac{D(n,t)}{t} &= \lim_{t\to\infty} \frac{D(n,t)}{Y(n+1,t)} \frac{Y(n+1,t)}{t} = \mu(n+1)p(n+1) \end{aligned} \tag{10}$$

Combining equations (9) and (10):

$$p(n+1) = \frac{\lambda(n)}{\mu(n+1)} p(n), \forall n = 0, 1, 2, ... \implies p(n+1) = \frac{\prod_{i=0}^{n} \lambda(i)}{\prod_{j=1}^{n} \mu(j)} p(0)$$

As they are frequencies they must hold $\sum_{i=0}^{\infty} p(i) = 1$ and recurrently compute the empirical time-probability of the system being in any state $n \in \mathbb{N}$.

Various performance measures can also be expressed in terms of $p(n)$:

$$\begin{aligned} L(s) &= \sum_{n=0}^{\infty} n \mathbb{1}_{\{L(s)=n\}} \\ \mathcal{L}(t) &= \frac{1}{t} \int_0^t (\sum_{n=0}^{\infty} n \mathbb{1}_{\{L(s)=n\}}) ds = \sum_{n=0}^{\infty} \frac{n}{t} \int_0^t \mathbb{1}_{\{L(s)=n\}} ds = \sum_{n=0}^{\infty} np(n,t) \\ \mathbb{E}(\mathcal{L}) &= \sum_{n=0}^{\infty} np(n) \end{aligned}$$

When level crossing can't be used we can take a different approach and focus on the stability of state n, the times the system enters in the state and the times that the same system changes from state to other states in a long-run

time-average should be infinitesimally the same. From this perspective the relation can be described as follows:

$$p(n)\lambda(n) + p(n)\mu(n) = p(n+1)\mu(n+1) + \lambda(n-1)p(n-1) \tag{11}$$

The proof is not the goal of this thesis, nonetheless the reader can find it in [1].

2.1.8 M/M/1 Queue

In this specific kind of queue the arrival distribution and the service distribution are memoryless, each costumer brings one job to the system and there is only one server. Thanks to the theory presented in section 2.1.7 and in addition to the empirical measures presented in section 2.1.6, long-run probability distributions can be described.

Arrivals and departures follow the Poisson processes $N_\lambda(t)$ and $N_\mu(t)$, respectively, and level-crossing equation can be used.

$$p(n+1) = \frac{\lambda}{\mu}p(n) = \rho p(n) \implies p(n+1) = \rho p(n) = \rho^n p(0).$$

Imposing a rate stable system and consequently taking into account that $\rho < 1$:

$$1 = \sum_{n=0}^{\infty} p(n) = p(0)\sum_{n=0}^{\infty}\rho^n = \frac{p(0)}{1-\rho} \implies p(0) = 1-\rho.$$

Hence $p(n) = \rho^n(1-\rho)$, $\forall n \in \mathbb{N}$

Proposition 2.31 *[3] In this case the expected long-run timer average and probability of the queue being greater or equal than a threshold are described as follows:*

$$\mathbb{E}[\mathcal{L}] = \frac{\rho}{1-\rho}$$
$$P(L \geq n) = \rho^n$$

Proof The expected long-run average length is:

$$\mathbb{E}[\mathcal{L}] = \sum_{n=0}^{\infty} np(n) = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = (1-\rho)\sum_{n=0}^{\infty} n\rho^n$$
$$= (1-\rho)\frac{\rho}{(1-\rho)^2} = \frac{\rho}{(1-\rho)}$$

The long-run length distribution:

$$P(\mathcal{L} \geq n) = \sum_{i=n}^{\infty} p(i) = \sum_{i=n}^{\infty} \rho^i(1-\rho) = (1-\rho)\sum_{i=n}^{\infty} \rho^i = (1-\rho)\frac{\rho^n}{1-\rho} = \rho^n$$

Those expressions that thinking about the empirical expected queue length as a function over ρ it has an asymptotic discontinuity in $\rho = 1$, or equivalently when the arrival and the departure rate are equal.

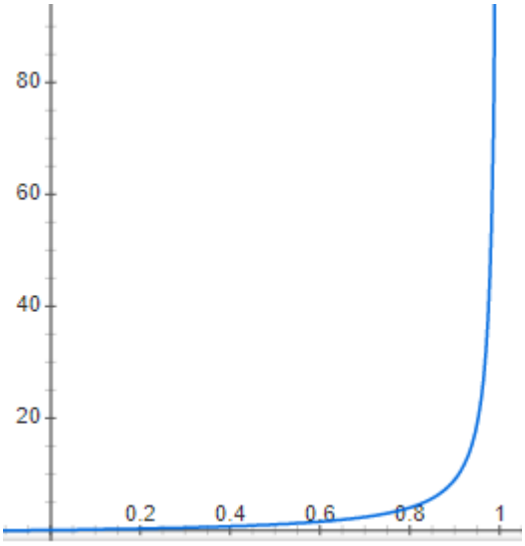

Figure 4: Empirical expected queue length against the load ρ

Another fact is that, implicit from $P(\mathcal{L} \geq n)$, the probability that queue length exceeds some threshold decreases exponentially.

2.1.9 M/M/c Queue

Also known as the Erlang C-formula, In this model the arrival rate stays as in the M/M/1, the only variance is in the number of servers, instead of one there are $c \in \{2, 3, 4, ...\}$ with the same service rate. As can be seen in figure 5 the model suppose one queue and c servers.

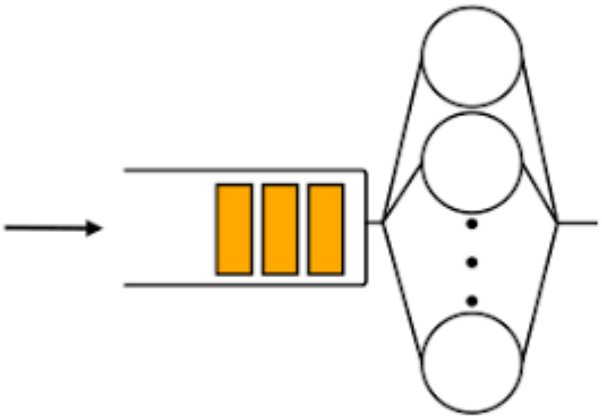

Figure 5: Diagram of the M/M/c queue

Observe that in this queuing system the arrival rate and the service rate will be as follows:

$$\lambda(n) = \lambda$$

$$\mu(n) = \begin{cases} n\mu \ , \ n = 0, ..., c-1 \\ c\mu \ , \ n = c, c+1, ... \end{cases}$$

To illustrate this imagine the Markov Chain associated with the $M/M/3$ queue such as seen in Figure 6.

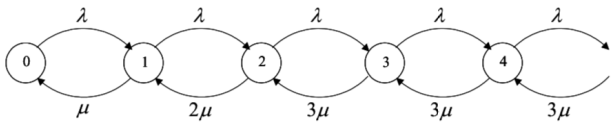

Figure 6: State-transition diagram of the M/M/3 queue

Moreover the system load is, then, defined as $\rho = \frac{\lambda}{c\mu}$. Hence by applying the general equations (11) for this case we obtain the normalizing factor (Equation (12)).

$$G = \sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{(1-\rho)c!} \tag{12}$$

and the long-run probabilities (13):

$$p(n) = \begin{cases} \frac{1}{G}\frac{(cp)^n}{n!} \ , \ n = 0, ..., c-1 \\ \frac{1}{G}\frac{c^c\rho^n}{c!} \ , \ n = c, c+1, ... \end{cases} \tag{13}$$

At this point the expectation for the queue length and the servers occupation can be computed. Results can be seen in Equations (14) and (15).

$$\mathbb{E}[L_Q] = \sum_{n=c}^{\infty}(n-c)p(n) = \frac{(c\rho)^c}{c!G}\frac{\rho}{(1-\rho)^2} \tag{14}$$

$$\mathbb{E}[L_S] = \sum_{n=0}^{c-1} np(n) = \frac{\lambda}{\mu} \tag{15}$$

Assuming people tends to go to the shortest queue.The length of each of the server's queues is the length of the queue of the model divided between c servers.

Moreover the probability of the queue being greater than a certain threshold can also be immediately computed by the Equations exposed in (13). Suppose we have c servers and we aim to know the probability of the queue being greater or equal than a length l.

On one hand, if c is greater or equal than L the cumulative probability can be calculated by Equation (16):

$$P(L_Q \geq l) = 1 - P(L_Q < l) = 1 - \sum_{n=0}^{l-1}\frac{1}{G}\frac{(cp)^n}{n!} \tag{16}$$

On the other hand, if c is strictly smaller than L the cumulative probability can be calculated by Equation (17):

$$P(L_Q \geq l) = 1 - P(L_Q < l) = 1 - \sum_{n=0}^{c-1}\frac{1}{G}\frac{(cp)^n}{n!} - \sum_{n=c}^{l-1}\frac{1}{G}\frac{c^c\rho^n}{c!} \tag{17}$$

2.2 Time Series

A time series is a set of discrete data with a given order where either the data and it's order provides information.

Times series can be identified in any variable that evolves in time such a

company's supplies of a specific item or the amount of traffic in a tollgate. It's analysis usually aims to predict it's evolution for the upcoming periods.

Time series have a critical difference from classic statistics and it's study is carried out by stochastic analysis or Markov chains. On one hand, if we were to consider time series from a classical statistician point of view we would be likely to assume the each observation as a sample of a given random variable which would usually be independent and identically distributed. Thus, it neglects the correlation of almost any evolutionary process. On the other hand, from the perspective of a stochastic process or a Markov chain, the observations would come from a random variables which are not independent but have a common evolutionary line and, hence, no independence.

A clear separation between time series can be made by differentiating between:

- A continuous or discrete time
- A continuous or discrete sample space of the variables in the series

As we will focus on the cyclic influence of the week day in the traffic flow problem we will focus on time series with discrete time and a continuous sample space.

$$\{x_i, i = 1, 2, ...\} \text{ and } \{X_i, i = 1, 2, ...\} \tag{18}$$

Where x_i and X_i are, respectively, the observation and the variable at time i. This kind of series are currently studied as Stochastic Process whereas those with a discrete sample space are usually studied as Markov Chains. However, as our goal is only to extract the cyclic influence of the week day in the traffic flow we will be make use of the classical theory.

2.2.1 Classical or Macroscopic analysis of time series

This perspectives assumes that a time series is composed by a trend, some cycles and a stationary residual. Hence, for a clear understanding of the series we need to extract and identify each of those components.

Assume, for instance, $\{\theta_i, i = 1, 2, 3, ...\}$ is the frequency of cars arriving at the toll gate at day i. It is not difficult to imagine that, in a specific toll gate, the data is going to vary influenced by both the annual increase or decrease of use of cars and the month day and the weekly day. If the toll gate is to enter to an industrial area there is likely to be a significant difference between

a Monday in May and a Sunday in August.

Hence (reference [5]), data is going to be:

$$X_i = f(T_i, C_i^{(1)}, ..., C_i^{(k)}, R_i) \tag{19}$$

where

- T_i: The trend of the data, describing the overall trend of the data. For instance, the trend of pollution has been increasing in the last years. This effect does not depend on any cycle.
- $C_i^{(1)}, ..., C_i^{(k)}$: The different cycles that can be identified in the data describing seasonality effects where the studied variable increase or decrease depending on the specific moment.
- R_i: The stationary residual. Noise that was not described by neither trending or cyclical effects.

This function is usually going to be linear or can be converted to linear by a log transformations as it's the case of the prices. Log transformation is mainly used for three reasons:

- To convert the sample space from positive real values to real whole real set.
- To convert an exponential evolution to a linear one. It's useful to simplify the model.
- To convert products in sums, decreasing then, the computational error.

After a visual analysis of the time series the following techniques can be used to extract the trend and cycles. For the following part we will suppose $x_i, i = 0, ..., N$ to be the initial data.

Trend Identification

Fitting
A polynomial approximation of the data (independent variable) as function of time (dependent variable) can describe relatively well the trend of the data.

$$x_i = f(i) + e_i$$

with

$$f(i) = a_0 + a_1 i + a_2 i^2 + \cdots + a_m i^m$$

The coefficients $a_0, a_1, ..., a_m$ can be found by minimum squares.

Filtering
A filter is defined as a linear combination of variables such as:

$$A_q(x_n) = \sum_{r=-q}^{q} a_r x_{n+r} \tag{20}$$

Usually assumed symmetric, $a_r = a_{-r}$, and such that $\sum_{r=-q}^{q} a_r = 1$ due to time influence of one observations. The resulting time series will be $\{A_q(x_i), i = q, ..., N - q\}$ Observe that with this method we are summarizing the initial time series in a shorter one were each term represents a weighted average.
This method, when suitable used, can isolate the trend by smothering the cyclic effects.
The simples case of a filter is the case of moving averages. In this case the coefficients equally distribute the influence of the observations between the $2q + 1$ observations, $a_r = \frac{1}{2q+1}$.

Differencing
When our goal is to simply erase the trend instead of isolating it. The idea behind this method is that the increase behind the data will eventually be stationary, i.e. have the same statistical mean, reflecting neither a positive nor negative trend. To do so the following operators are introduced:

$$Bx_i := x_{i-1} \text{ for } i = 1, ..., N$$

and

$$\nabla x_i := (Id - B)x_i \text{ for } i = 1, ..., N$$

This last operation can be recursively defined as

$$\nabla^j x_i := (Id - B)^j x_i \text{ for } i = 1, ..., N$$

The resulting series $\{y_i = \nabla^j x_i, i = j, ..., N\}$ is call the series of the jth-differences of the series $\{x_i, i = 0, ..., N\}$.
Differences are usually applied until no trend is seen in the visual analysis of the last differenced series.

Identification of cycles

The seasonal or cyclical component of a time series $\{s_i, i = 0, .., N\}$ needs to hold some theoretical assumptions.

- The seasonal component need to be the same every period, equivalently, $s_i = s_{i+p}$, $i = 0, ..., N - p$.
- The cyclic influence in the trend must be null, equivalently, $\sum_{i=1}^{p} s_i = 0$

Where p is the period of observations. In our case, for instance, if we are computing week days the period would be of 7 days. *Direct identification* Our goal with this method is to identify $\{s_i, i = 0, .., N\}$ or equivalently $\{s_i, i = 1, .., P\}$. Assume $m > 0$ periods in the observed data. To find the cyclic component means for each component of the cycle need to be computed:

$$e_i := \frac{1}{m} \sum_{j=0}^{m-1} x_{i+pj},\ i = 1, ..., p \tag{21}$$

In order to make the component centered to assure the second assumption the mean of period means is defined:

$$\overline{e} := \frac{1}{p} \sum_{i=1}^{p} e_i \tag{22}$$

obtaining the seasonal component as:

$$s_i = e_i - \overline{e}$$

Filtering
This method is used to erase the cyclic effect. Any symmetric filter will eliminate a cycle e with an odd period p. The simplest one is the moving averages. Setting $2q + 1 = p$, the series defined by:

$$A_q(x_i) := \frac{1}{2q+1} \sum_{j=-q}^{q} x_{i+j}$$

Constitutes the series $\{A_q(x_i),\ i = q, \ldots, N-q\}$ which has erased the cyclical effect. Equivalently, when the cycle has an even period as it is the case of months in a year or the hours in a day, the filter might be adapted as follows:

$$A_q(x_i) := \frac{1}{2q}(\frac{1}{2}x_{i-q} + x_{i-q+1} + \cdots + x_{i+q-1} + \frac{1}{2}x_{i-q})$$

Differencing
An alternative method to erase the cyclical effect is by developing a series with the following operator. Suppose the period of the cycle to be p, hence,

$$\nabla_p x_i := (Id - B^p)x_i \text{ for } i = p, \ldots, N$$

Where

$$B^p x_i = x_{i-p} \text{ for } i = 1, \ldots, N$$

By doing so the series will just carry out information about increase from a stages of the period to the following one, erasing the initial difference that this stage may have. In a yearly cycle where the different months are the different stages of the periods and $p = 12$ the operator transform the initial data of each month to the difference from the data in one month with the data in the same month but in the latest year. This usually eliminates the base difference between months.

3 Application

This section aim to asses the error of the model and predict hourly queuing measures for *January* and *February* of 2018 in the tollgate situated at Mollet del Vallès by using the developed theories.

The available data is the number of arrivals by hour from $00:00:00$ of $1/1/2013$ until $23:00:00$ of $31/1/2017$, with this information arrivals forecasting will be tested for *October*, *November* and *December*. Afterwards, the predicted total number of arrivals per hour will constitute the base for the queuing analysis using the theory presented in Sections 2.1.8 and 2.1.9. In this last part the goal is not to compute the queue measures but to provide the company with the minimum amount of gates needed to reach different queuing criteria.

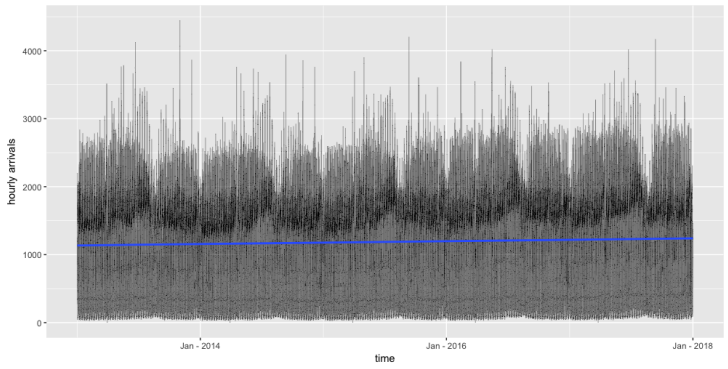

Figure 7: Initial Time Series of the arrivals.

The following theoretical assumptions are required to implement the model:

1. *Inter arrivals times are exponentially distributed* or, equivalently, arrivals follow a Poisson process for periods smaller or equal than an hour.

2. *Service times are exponentially distributed* or, equivalently, the amount of cars served follow a Poisson process for periods smaller or equal than an hour.

3. *Inter arrival per hour follow the same exponential distribution.* Thus, there is a certain homogeneity of the inter arrival times during this

period. For example, a lot of cars arrive in the first 10 minutes and non in the last 10 minutes of an hour.

4. *The hourly arrival-rate is a good approximation of the expected arrival-rate in the period.*

All the assumptions can be assessed individually for each tollgate by performing a specific test or a visual analysis. We will mention but not develop some procedures to asses them and assume their validity for our case as it is not the goal of this thesis.

Provided that the company collects the specific arrival times and the service times of the system a way to proceed could be as follows:

- Assumptions **1** and **2** can be evaluated by plotting the inter-arrival and service times respectively for an hour-periods and check if its reasonable to suppose an exponential distribution.
- Assumption **3** can be assessed by visually representing the moments when the arrivals are produces, group them and test if they are better represented by two or more exponential distributions
- Assumption **4** can be assessed by plotting the evolution of the limit as the time goes on in an hour and check if in an hour period it is already stable enough.

3.1 Analysis of the time series

The first step is to analyze the amount of arrivals by hour ,$A(60)_d$ following the notation of Section 2.2.1, with t in minutes, with classical time series theory. The goal is then to describe the time series by a function f as presented in Equation 23 as presented in section 2.2.1.

$$A(60)_d = f(T_d, C_d^{(1)}, C_d^{(2)}, ..., C_d^{(N)}, R_d) \tag{23}$$

In order to do a post-validation and assess its accuracy we start by splitting the data set between the training and testing sets where the test is composed by the observations in *October*, *November* and *December* of 2017 and the training set by the remaining part of the data.

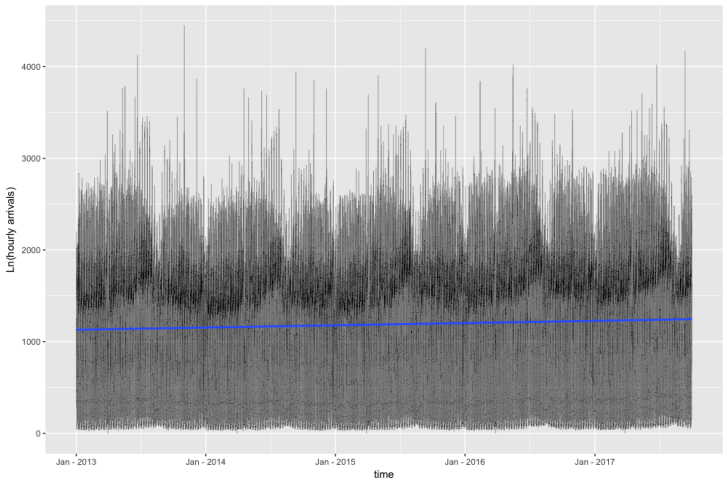

Figure 8: Training time series

As one can appreciate by Figure 8 number of arriving cars per hors seems to remain stable rather than increase or decrease over the years. As the data is highly skewed and the arrivals are positives or equal to zero we can use a $log-$transformation to reassure the linear assumption in the trend. Technically, we do so by first assigning one arrival to those hours with no arrivals so the log can be applied. Pragmatically, in the model, this modification of the data is not really important for the analysis as we are increasing the number of arriving cars only by one in two specific hours our of the total number of hours in 5 years.

After applying the $log-$transformation, as presented in Figure 9, the presence of five outliers is clear. By filtering the data frame for $log(arrival)$ values smaller than 2 and extracting various chronological features we obtain Table 1.

Day	Month	Hour	Weekday
31	3	2	Sunday
30	3	2	Sunday
29	3	2	Sunday
27	3	2	Sunday
26	3	2	Sunday

Table 1: Outliers of the time series

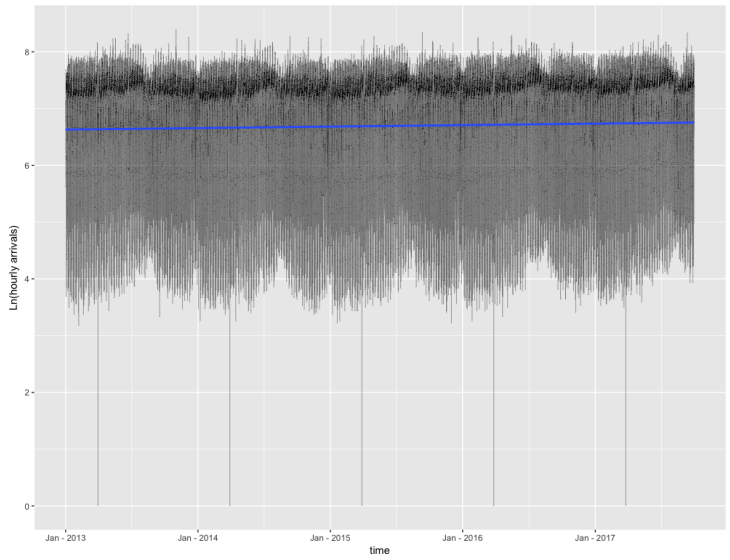

Figure 9: Log of the training time series.

Oddly, those 5 outliers correspond to the 2 a.m. of last Sunday of March of every year. Extracting this pattern as a remark, as it is an isolated event, and normalizing the observations, the influence of this event in the rest of the analysis can be reduced. For normalization the initial value will be substituted by the average of the rest of the Mondays in March of the same year. The result is the time series represented in Figure 10.

Observing the linearity in of the overall trend in Figure 10 and by the log properties we try to use an additive model, hence f will describe the function observed in Equation 24.

$$f(T_d, C_d^{(1)}, C_d^{(2)}, ..., C_d^{(N)}, R_d) = T_d + C_d^{(1)} + C_d^{(2)} + \cdot + C_d^{(N)} + R_d \qquad (24)$$

Compute moving-averages for different sliding windows, as can be observed in Figure 11, we can appreciate that the seasonal yearly effects are erased as the windows get wider, leading to a positive trend where is reasonable to do a linear approximation.

The time series after computing and extracting the linear approximation from the in represented in Figure 12.

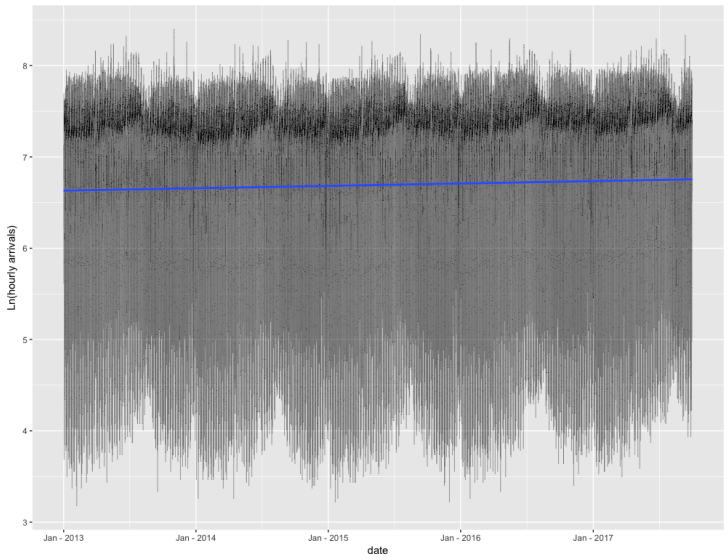

Figure 10: Time series after modifying the outliers. The blue line represents the line resulting of computing linear regression by minimum squares

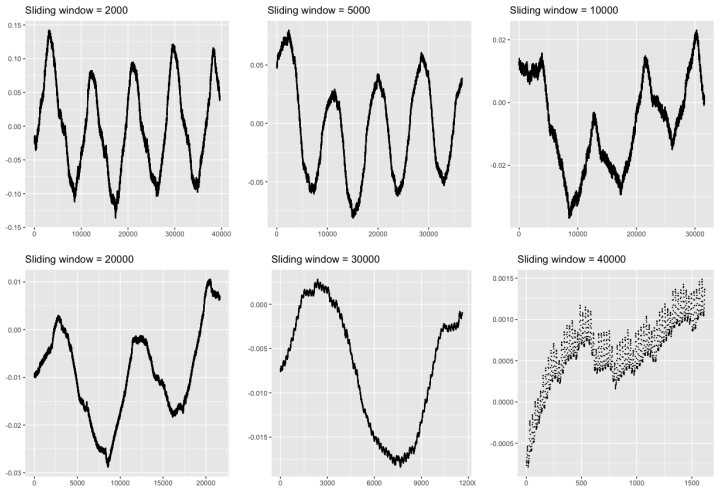

Figure 11: Moving Averages with different sliding windows.

Let's now proceed to compute the seasonality effects. To do so the theory must be applied as follows:

First:

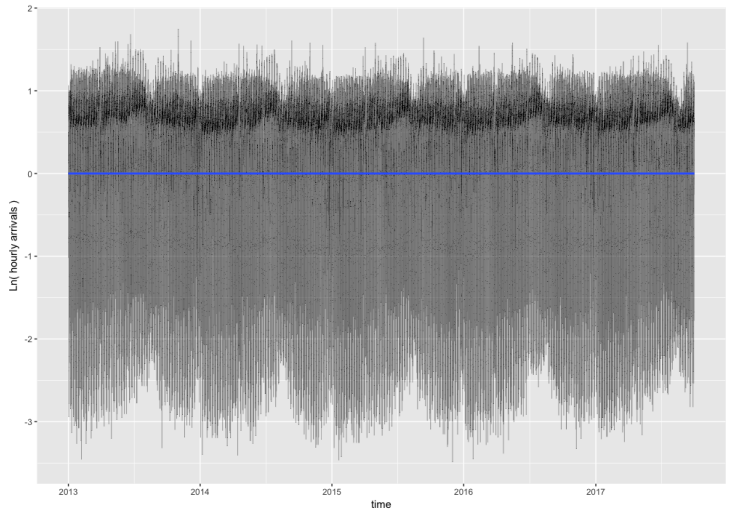

Figure 12: Time series after extracting the linear approximation of the trend.

$$A(60)_d - T_d = C_d^{(1)} + R_d \tag{25}$$

So the trend's effect is subtracted from the time series and not considered as part of the cyclical effect.

Secondly:

$$A(60)_d - T_d - C_d^{(1)} = C_d^{(2)} + R_d \tag{26}$$

So the trend and the first cycle's effect is subtracted from the time series and not considered as part of the second cyclical effect. This procedure continues until no more cyclical effect can be seen in the time series.

Remark 3.1 *In contrast to the provided theory we are also considering cycles with non fixed periods. The implemented code takes it in account and use the difference amount of elements in each cycle to weight the averages. Thus, instead of Equation* (21) *we use Equation* **??**.

$$e_i := \frac{1}{m_i} \sum_{x_j \in C_i} x_j,\ i = 1, ..., p \tag{27}$$

where C_i corresponds to the group of elements in cycle i and m_i is the cardinal of elements in C_i. For example, all the 12th of Decembers in the data

set. Consequently the computation of (22) *takes it in account by changing it by a weighted average of the* $\{e_i\}_{i\in P}$ *as shown by Equation* **??**.

$$e := \sum_{i=1}^{p} \frac{m_i}{m} x_i, \text{ where } m = \sum_{i=1}^{p} m_i \tag{28}$$

Going back to our time seres an additional problem arises, the time series is too skewed and large to do any visual identification of cycles. Thus, the cycles, at least at the beginning, will be identified by considering the possible situations that may effect the number of cars arriving to a tollgate. To assure the influence of those possible situations we will first the isolated cycle, decide if it may have a real meaning or it is just describing noise, and also plot the data after subtracting the cyclical effect to check if it helps reducing the time series only to noise.

The first seasonal effect to be considering, $C_d^{(1)}$, is the one produced by the effect of peak and off-peak hours depending on the week day. To illustrate the idea behind considering this cycle, imagine, for example, the difference between 9 a.m. of a Saturday, 12 p.m. of a Saturday and 9 a.m. of a Monday so it distinguishes by week days and hours. To do so we take a fixed 168h-cycle. This cycle can be appreciated in Figure 13.

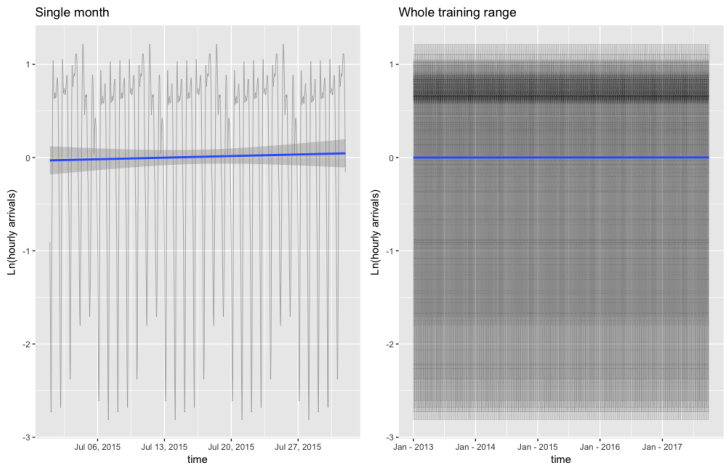

Figure 13: Seasonal effect depending on the week day and hour.

As the cycle represented in Figure 13 seems to have a clear e meaningful effect we now proceed to remove it's effect from the time series. The time

series after removing this effect is represented in Figure 14.

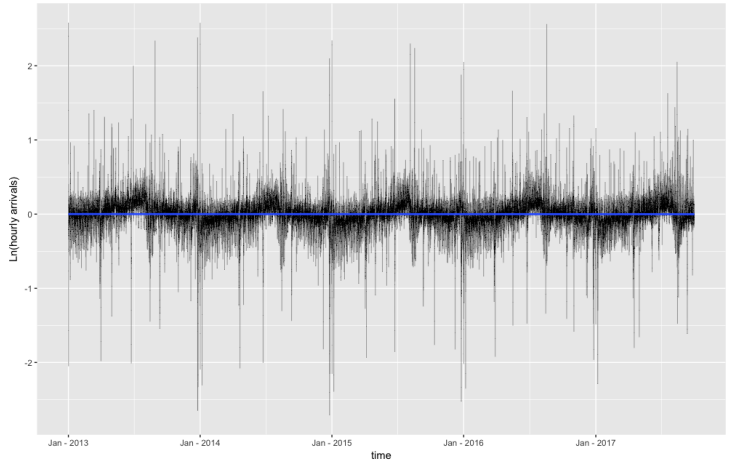

Figure 14: Time series after extracting the the first seasonal effect.

At this point it is already possible to identify some cycles. For instance a decreasing trend in the time series can be appreciated in August or in other weeks during the year. Hence, we consider $C_d^{(2)}$ to describe those differences. To do so we take the number of the week in the year and the week day. This cycle won't have a fixed period, nonetheless, we take it in account in the code as mentioned in Remark 3.1.This stationary effect can be appreciated in Figure 15.

Even though, it's effect is not highly influential to the series as their values vary, roughly, from -0.3 to 0.3, the cycle seems to have a meaningful effect, describing, for example, the decrease in traffic of August. The time series after the extraction of this cycle can be observed inf Figure 16.

Indeed, the noise in the data has been reduced. However we can still appreciate some periodic, significant, increase and decrease in traffic. Thus we define $C_d^{(3)}$ grouping the elements depending on the month, the day of the month and the hour. Observe that, when defining the cycle in this way, it encompasses the effect of festivities such as Christmas or Easter, between others. Notice that $C_d^{(3)}$ won't have a fixed period, nonetheless, we take it in account in the code as mentioned in Remark 3.1.This cycle can be appreciated in Figure 17.

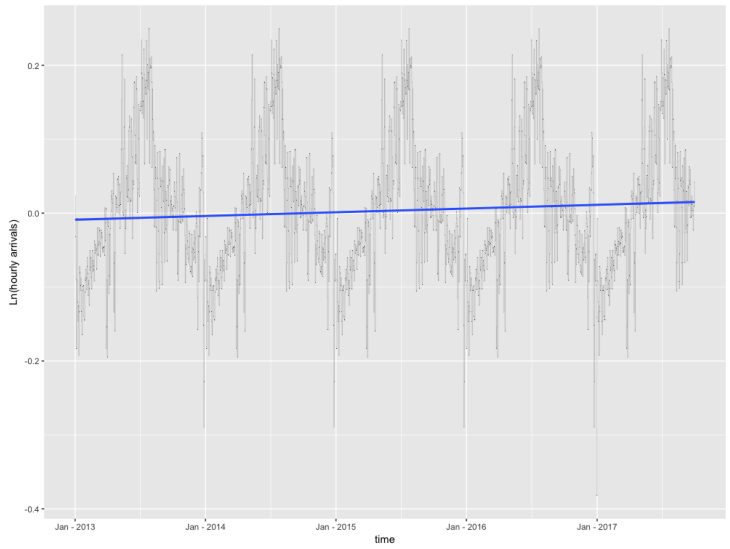

Figure 15: Seasonal effect depending on the week of the year and the week day.

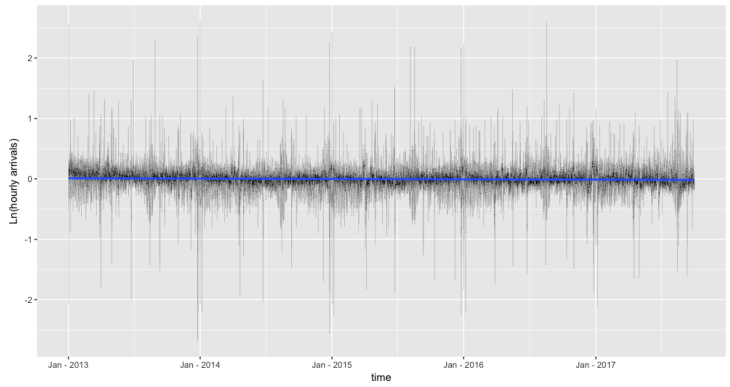

Figure 16: Time series after extracting the the second seasonal effect.

Observe that this cycle (17) is highly similar to 16 what indicate a good explanation in the noise. Subtracting this effect from the time series we obtain the output represented in Figure 18.

Indeed, with this cycle a reduction in the range of the noise have been achieved (reduction in scale). Observe now that the data does not show any regular pattern except for few yearly off-peaks around March. To study this phenomena we will filter the data set for values smaller than -1.1 and

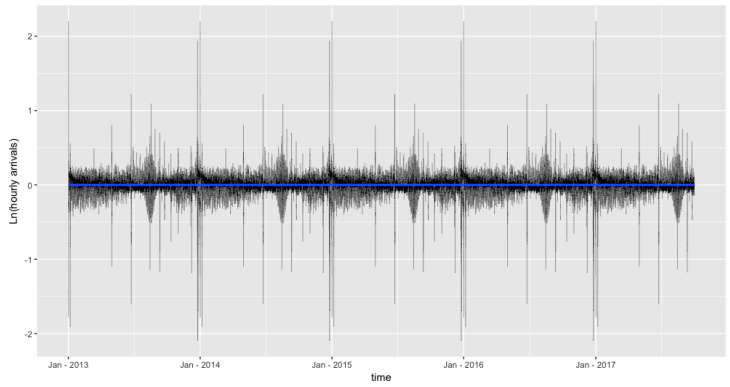

Figure 17: Seasonal effect depending on the month, day of the month, and hour.

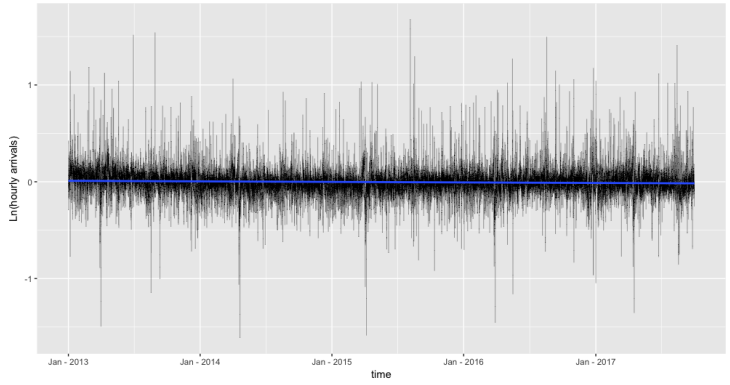

Figure 18: Time series after extracting the the third seasonal effect.

extract some chronological features as can be appreciated in Table 2.

A temporal relation between the observations can be appreciated, the recursive off-peak happens at 7 and 8 in the morning of a specific Monday each year (excluding the observation at $2016-05-1607:00:00$) in May or April. However, there is not a clear pattern and thus a we can not be sure when will happen the following off-peak. Even thought it can not be treated directly using cyclic analysis we may trust experienced workers about festivities, weather conditions or any possible circumstance causing this situation and then reflecting it when doing the prediction just as any other seasonal effect. Apart from those off-peaks no other clear time pattern is observed in Figure 18. Thus, the first step in the analysis is concluded and Figure 18 is

Date and Time	Weekday	Week
2013-04-01 07:00:00	Monday	13
2013-04-01 08:00:00	Monday	13
2014-04-21 07:00:00	Monday	16
2014-04-21 08:00:00	Monday	16
2015-04-06 07:00:00	Monday	14
2015-04-06 08:00:00	Monday	14
2016-03-28 07:00:00	Monday	13
2016-03-28 08:00:00	Monday	13
2016-05-16 07:00:00	Monday	20
2017-04-17 07:00:00	Monday	16
2017-04-17 08:00:00	Monday	16

Table 2: Off-peaks from time series represented in Figure 18.

describing the noise R_d.

The following step in the project is testing the analysis on *October*, *November* and *December* of 2017. The construction of the hourly predictions are constructed as follows:

1. Add-up the corresponding cyclic effects $C_d^{(1)}$, $C_d^{(2)}$ and $C_d^{(3)}$ and trend T_d.
2. Take the exponential of the values computed in (1).

A visual comparison between the predicted and real hourly arrivals is presented in Figure 19.

The performance of the forecasting is assessed by the *Normalized Root-Mean-Squared Deviation* which is computed following Equations (29) and (30).

$$RMSD = \sqrt{\frac{\sum_{i=0}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{29}$$

$$CV(RMSD) = \frac{RMSD}{\bar{y}} \tag{30}$$

Where $\{\hat{y}_i\}_{i\in T}$ are the predicted values of the test set, $\{y_i\}_{i\in T}$ are the real values of the test set, T is the index of the test set and n is the cardinal of

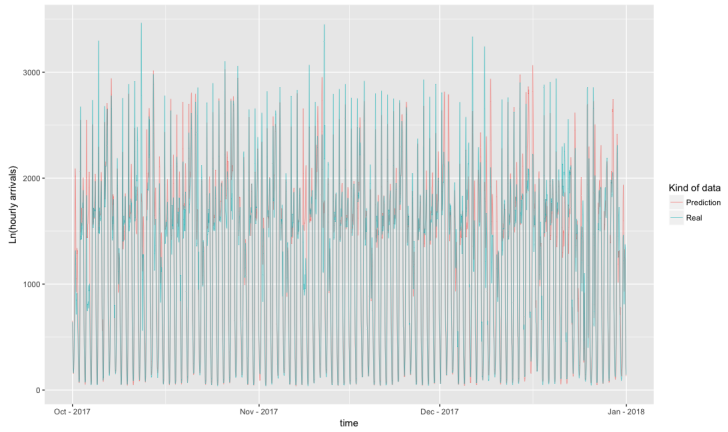

Figure 19: A comparison between the predicted real values of the testing period.

T.

In our case the normalized root-mean-squared deviation is 0.2064687 and the standard deviation of 245.2296, what indicate us that we are likely to expect a deviation between real and predicted estimations of 20.72% on average, relative to the real values or a deviation of the order of 245 cars on average.

3.2 Queuing Measures

Now, with the predicted arrivals and using the assumptions mentioned in theory, the arrival rate per minute can be computed. Supposing or estimating a common service rate (expected service rate) per minute of the servers in a specific tollgate, the company can variate the number of open gates (servers) for every hour depending on their criteria and goals. The forecasting of the number of gates is done using the theory developed in Sections 2.1.8 and 2.1.9. For this application two criteria will be used:

1. *Criteria 1: Expected queue length.* The expected queue length to don't be greater than a desired *threshold*.

2. *Criteria 2: Confidence in a queue smaller than a threshold.* The probability of the queue of being greater than a specific number of cars ($threshold_1$) to be smaller than a specific $threshold_2$.

For testing and illustrating purposes algorithms for *Criteria 1* and *Criteria 2* will be implemented. The first step in both cases is to compute the expected service time of the gates. It is assumed in theory that the gates have the same service time distribution. Thus, if we have different kind of gates, for example, those operated by machines and those operated by humans they must be studied separately. We suppose a expected service rate of 2 cars per minute.

Criteria 1

For this criteria we will start by specifying the desired expected length of the queue, for example, $L = 2$. Now with the R function $nGates.EL$ in Annex 2 (6) we start by computing the utilization rate ρ. Starting with $c = 1$ server it checks if the utilization rate is smaller than 1 and the queue expectation per server is smaller than the threshold L. If the amount of servers don't meet the conditions the algorithm re-computes the utilization rate ρ and checks the conditions for $c + 1$ servers.

The comparison between the number of servers using the real number of arrivals and the predicted number of cars can bee observed in Figure 20.

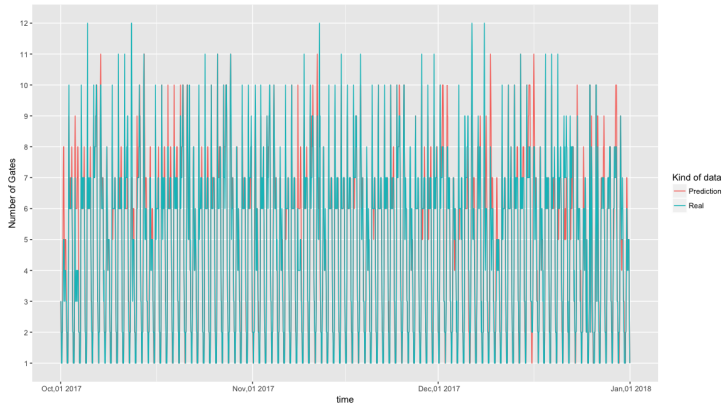

Figure 20: Comparison between the predicted and real minimum amount of open gates needed each hour, according *Criteria 1*, to have a expected queue smaller than 2 cars per line for *October*, *November* and *December* of 2017.

Computing the error presented in Equations (29) and (30) for the number of needed servers we obtain a normalized root-mean-squared deviation of 0.1870278 and a standard deviation of 0.8971069 for the gate's forecasting.

An interpretation of those values is that reality is likely to differ, in average, from predictions in the 18.72% of the real number of gates, or, approximately, in 1 gate in average.

Criteria 2

For this criteria we start by specifying the desired length of the queue that we don't want to exceed and an upper threshold for the probability of this event to happen, for example, a length of $L = 3$ and an upper threshold of 0.15 for the probability. Furthermore, for computational reasons, we set a limitation in the number of servers that the algorithm checks. For example, if we set the limit to 50, the algorithm won't return more than 50 servers as a result. Then, with the R function $nGates.PL$ in Annex 2 (6) we start by computing the utilization rate ρ for one server. Starting with $c = 1$ servers the program checks if the utilization rate is smaller than 1 for this amount of servers and if the desired conditions are achieved. If the amount of servers don't meet the conditions, the algorithm re-computes the utilization rate ρ and checks the conditions for $c + 1$ servers until obtaining the desired results or it arrives at 50 servers.

The visual comparison between the minimum number of gates needed using the real amount of arrivals and the the minimum number of gates needed using the the predicted amount of arrivals is presented in Figure 20.

As done for *Criteria 1*, the computation of the error presented in Equations (29) and (30) for the predicted minimum number of servers needed produces a normalized root-mean-squared deviation of 0.18258556 and a standard deviation of 0.9269027. An interpretation of those values is that the minimum number of gates needed based in real arrivals is likely to differ, in average, from the minimum number of gates needed based in predictions, in the 18.26% of the minimum number of gates needed based in real arrival, or in approximately 1 gate in average.

3.3 Forecasting

Following the same procedure that has been done for testing, but using the whole data set, forecasting can be done for *January* and *February* of 2018. Predicted arrivals to the toll gate can be appreciated in Figure 22.

After obtaining the total hourly arrivals, we can also estimate the minimum required number of gates by Criteria 1, with a expected queue length of

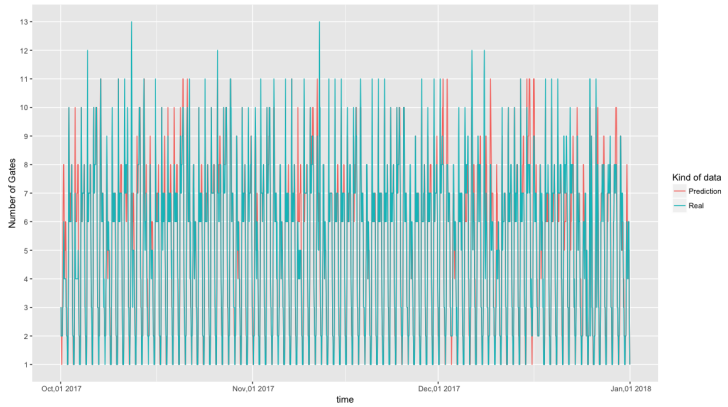

Figure 21: Comparison between the predicted and real minimum amount of open gates needed, according to *Criteria 2*, to have a probability of the queues being greater than 3 smaller than 0.15 for *October*, *November* and *December* of 2017

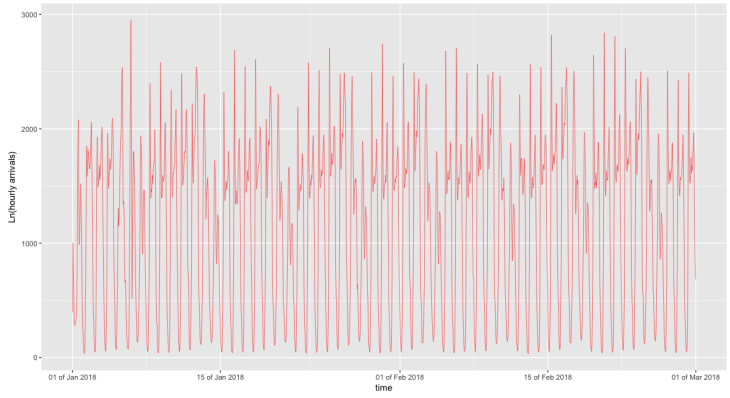

Figure 22: Expected total arrivals per hour for *January* and *February* of 2018

$L = 2$, and by Criteria 2, with a queue length of $L = 3$ and a probability threshold of 0.15 and limit 50. The results are represented in Figures 23 and 24, respectively.

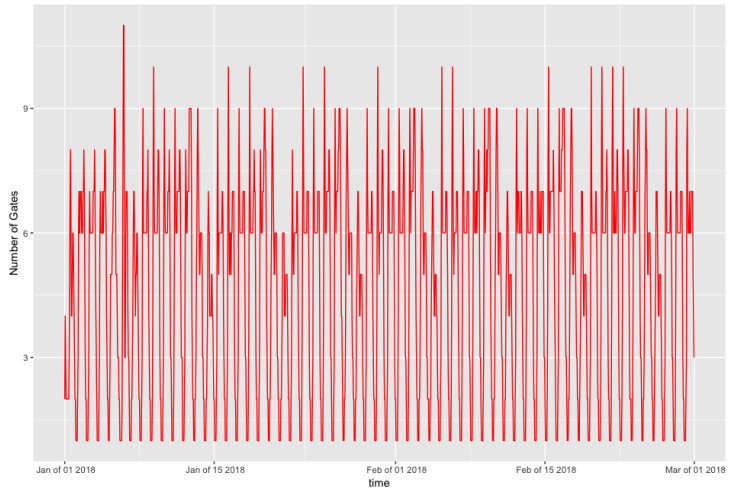

Figure 23: Predicted minimum number of gates needed each hour to have a expected queue of 2 cars per line *Criteria 1* for *January* and *February* of 2018.

Figure 24: Predicted minimum number of gates needed each hour to have a probability of the queues being greater than 3 smaller than 0.15 *Criteria 2*for *January* and *February* of 2018.

4 Conclusions and Limitations

At the beginning of this paper we aimed to clarify some of the uncertainty when managing resources related with traffic in an upcoming period of time.

While exploring how queues were classified or how a general queue is analyzed we encounter two kind of queues, the M/M/1 and the M/M/c, and the long-run time measures which are useful for our purposes. Working on their bases such as the Poisson Processes or Balance Equations, between others, we were able construct those concepts from the ground. Afterwards, we studied the classical approach of time series, as well as some of its tools and the steps to isolate the various components of the series.

Once the basics of those theories were set we boosted their use for our goal by combining them. Time Series analysis was used to forecast the basic information needed for Queuing theory, the total number of arrivals per hours. More specifically we applied a $log-$transformation and identified and extracted the trend and the cycles. The forecasting was used to compute the minimum number of gates needed to reach, under those circumstances by using the measures exposed in Sections 2.1.8 and 2.1.9.

Our gaol was achieved, an early model and a program implemented in R were developed. We obtained a standard deviation of the 20% of the expected real data by isolating the time patterns of the Time Series. Nonetheless, the model still have room for improvement.

> *A model should be as simpler as it can be but not simpler.*
>
> Albert Einstein.

The provided script in Annex 2 (6) can be automatized and productized and may be useful depending on the goal and situation. However, after testing and plotting the time series of the necessary number of gates computed by the real and predicted number of arrivals we observe that even when the averaged root-squared-error of 1 may be acceptable, reality needs 5 more gates than what was predicted. For some applications the range of errors might need to be reduced and therefore a more complex model might be needed. A possible way to develop the model from this point is to describe the remaining noise of the total arrivals per hour by other non temporal factors. For example, using the noise as dependent variable and multiple weather conditions as dependent variables in an analysis using machine learning algorithms such as XGBoosting.

References

[1] Randolph W. Hall, *Queueing Methods For Services And Manufacturing*, University of Southern California, 2013.

[2] Nicky D. Van Foreest, *Analysis of Queuing Systems with Sample Paths and Simulation*, University of Groningen, 2017.

[3] W.H.M. Zijm, *Manufacturing and Logistic Systems Analysis, Planning and Control*, 2003.

[4] *R. Gallager, MIT* `http://www.rle.mit.edu/rgallager/documents/6.262lateweb2.pdf` Consulted October the 30th of 2017.

[5] Brockwell, Peter J., Davis and Richard A., *Introduction to Time Series and Forecasting*, 2002

5 Annex 1

5.1 Poisson process

Definition 5.1 *Given a counting process $N(s,t)$, $s,t \in \mathbb{R}_{\geq 0}$ it has the stationary increment property if and only if $\forall t_0, t_1, t_2, t_3 \in \mathbb{R}_{\geq 0}$, $t_1 \geq t_0$ and $t_3 \geq t_2$:*

$$t_1 - t_0 = t_3 - t_2 \implies N(t_2, t_3) \text{ and } N(t_0, t_1) \text{ have the same CDF}$$

Besides matching the intuitive conditions for a simple queuing model this fact can be explained if we think about the Poisson distribution as a Binomial distribution with a really large number of trials:

Imagine a finite time interval $(0, T]$ chopped up in smaller intervals that small that only one occurrence could happen. In this case it's modeled by a Binomial distribution $Bin(n, p)$ such that $\lambda = np$.

$$\begin{aligned} P(N_n(t) = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n!}{k!(n-k)!} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n!}{(n-k)!} \frac{1}{n^k} (1 - \frac{\lambda}{n})^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^x} (1 - \frac{\lambda}{n})^{n-x} \end{aligned}$$

By applying the limit when n tends to infinity:

$$\begin{aligned} \lim_{n\to\infty} P(N_n(t) = k) &= \lim_{n\to\infty} \frac{\lambda^k}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^x} (1 - \frac{\lambda}{n})^n (1 - \frac{\lambda}{n})^{-x} \\ &= \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned}$$

6 Annex 2 : R code for the application

The *R* code that has been used for the application can be found in the attached Annex 2 as a R-Script (.R) file.