Stochastic Modelling of Train Delay Time Series in Skåne, Sweden

Filip Vestin

Valentin Zulj

Bachelor’s Thesis in Statistics 15 ECTS, Fall Semester 2017

Department of Statistics, Lund University

Supervisor: Professor Krzysztof Podgórski

Abstract

The purpose of this paper is to provide a foundation for modelling train delays as multivariate time series. A pertinent issue with this kind of analysis is that the individual time series do not follow the Gaussian normal distribution. Since the normal distribution constitutes an assumption of classical time series methods, these cannot be applied blindly to the data. The solution explored by the paper is to identify the distributions of the time series and subsequently transforming the data into normal distributions.

To this end, a dataset from the Swedish Transport Administration was used containing delays for every train that either departed from the region Skåne as a first departure, or arrived in Skåne as a final destination. The dataset spans from January 1st, 2014 to December 31st, 2016. Time series were constructed for every station with a significant amount of data points by computing the mean daily delays for these stations. Using software for distribution fitting it was found that the asymmetric Laplace distribution best described the distribution of the train delays. The data was successfully transformed into a normal distribution using the empirical cumulative distribution function of the asymmetric Laplace distribution. Comparing the cross-correlation functions before and after the transformations showed mild increases in the time dependence between stations.

Table of Contents

1. Introduction

One way or another, each and every one of us has had to suffer from the consequences of delayed trains, whether it be missing a flight or being late for a meeting. With trains being such an important part of our infrastructure, especially in Skåne, it is vital to understand the way in which train delays come about, and the way in which they spread and affect one another. This thesis will focus on the latter, aiming to better understand the time dependence of delayed trains throughout Skåne, spanning from January 1st 2014 to December 31st 2016.

Interdependence, in particular over time, is best modeled using time series analysis of multivariate nature. Multivariate time series analysis allows us to study the autoregressive or moving average dependence of a time series not only upon itself, but also its dependence upon different, related series. Hence, we can combine train delay data from different stations in order to better understand the way in which they covariate. This paper will deal with complex multivariate time series data and as such, there is no convention telling us how to approach our problem. It is possible to try and model our data as is, but because train delays are unusually distributed, there are no guarantees such a model would be very fruitful. An alternative approach is transforming the data so that it becomes better suited for time series analysis, and try to model it after transformation. Hence, the scope of this thesis is initial analysis of the effects that transformations of data could have on time series and multivariate data.

As stated above, one purpose of this thesis is to study the effect of transformations, and the way in which they affect time dependence and the applicability of conventional time series analysis on train delays in the Swedish region of Skåne. That is, our focus will be preparing the data rather than modelling it, seeing as the distribution of delays is such an unusual one. This will be achieved by studying the distribution of our train delay data, attempting to make it Gaussian through transformation, and analyzing the way in which the transformations affect the time series data.

1.2 Statement of Task

- Performing initial analysis of train delay distributions.
- Determining how transformations of data affect time series and multivariate data.
- Preparing the data for multivariate time series analysis

The primary purpose of this thesis is to analyze the distributions of train delay data, as well as examining what effect transformations will have. Hence, section 3 will focus on data analysis. In section 4, the theoretical tools and justification of the methods used in section 3 are both provided and explained in further detail. Furthermore, section 4 is concluded through giving an overview of what methods can be used to continue the analysis using time series modelling. The outline of the paper may appear counter-intuitive, but a reversal of the structure would implicate that the aim of the paper is to produce a time series model, while the purpose is to create data that satisfies the Gaussianity assumption of such a model. Furthermore, the principles of the transformations are largely self-explanatory, even if the technical details are not.

2. Data

Our dataset consists of train delay data, and was provided to us by the Swedish Transport Administration (Trafikverket, STA). It contains data of planned and executed operations carried out by two of the major passenger railway operators within Skåne (Öresundstågen and Pågatågen) during the time interval spanning from January 1st 2014 to December 31st 2016. The STA records departure times of every train operating within Sweden each day of the year, hence it was able to supply data regarding every train that either departed Skåne (i.e. its first departure within Sweden took place at a railway station i Skåne) or arrived in Skåne (i.e. its last arrival at a Swedish station happened in Skåne). For each departure from an original station, and for every arrival at a final destination, our dataset shows records of planned departure time, actual departure time, planned arrival time, actual arrival time, which operator was in charge of the train, and the difference in time between the planned and executed operation (if a specific train either departed or arrived on time this, of course, means the deviation is equal to 0).

We decide to look only at data concerning arrival times at the final destinations of each train, since we personally find arriving late more of an issue than departing late, as the latter does not necessarily bring with it a negative impact on our everyday lives. For the analysis to be carried out, we filter the data so that only final destinations remain. After that, our dataset is split into several data frames, as to group arrivals by the station at which they occurred (see appendix 1). Furthermore, time series are constructed through calculating the daily mean delay of arrival at every station, enabling us to start analyzing distributions and time dependence. All calculations mentioned above are made using the statistical software R, as well as Rstudio, which is an extension of it.

Doing this, we encounter an issue concerning the length of each time series, seeing as several stations are only given as final destinations under certain conditions (for example during weekend nights, or periods of railway repairs). Because of this, the time series for some stations contain scattered data of only a few days that are hard to analyze and thus they are rendered to be of no use to the investigation, hence they are dropped from the scope of this thesis. Furthermore, there are some errors in the delays at Malmö C. At this station, there are several delays recorded as less than -30, meaning that the train arrived 30 minutes earlier than it was supposed to. This means that there is some sort of error in the time-table or in the recorded time of arrival, or potentially a train not carrying any passengers. We therefore remove all observations with delays smaller than -5 for this station. For a complete list of the set of stations, as well as the subset of stations actually analyzed, see appendix 1.

Since our data is not normally distributed, the estimators of multivariate time series coefficients are not necessarily reliable if computed based on the normal distribution paradigm. On the other hand, if we use the non-normal models the computation of the estimators may be difficult and theoretically challenging. For these reasons, transformation methods are often considered, i.e. transforming the data so that the assumption of normality cannot be rejected, and perform the analysis on the transformed data. In order to identify proper transformations, the density function and its corresponding distribution function can be utilized and fitted to the data. In our thesis, we will fit the gamma distribution and the asymmetric Laplace distribution to the data. To fit the gamma distribution, we use the method of moments, and after that we transform the data using the empirical cumulative distribution function and its inverse. In other words, we run the data through the pgamma(…) function in R, which constitutes the empirical distribution function of the gamma distribution. This gives us uniformly distributed data which is run through the inverse distribution function of the standardized normal distribution, using the qnorm(…) function in R. Similarly, we transform the data through the asymmetric Laplace distribution, using the "ald" package in R.

There is a problem with a raw application of this method in the case of the gamma distribution, that is defined for positive numbers, since some stations will have a train delay less than or equal to zero, meaning that the cumulative distribution function (c.d.f) will map to zero. If $F_X(x)$ is a cumulative distribution function, then $\lim_{a\to 0} F_X^{-1}(a) = -\infty$, and limits are impossible to handle when plotting histograms. The solution is to shift the data by a constant θ defined as:

$$\theta = \min(Y_{t,m})\left(1 + \frac{1}{n}\right)$$

Where n is the total number of observations, i.e. the length of each time series $Y_{t,m}$. This way all values will be greater than zero and qnorm(…) will not map to infinity.

The shifted time series of four stations are shown in Figure 2.1. The stations are Helsingborg central station, Kalmar central station, Lund central station, and Halmstad central station. In the main body of this thesis, the analysis of these four stations will be presented more closely, since they are representative of the time series we have dealt with while working on this paper. All in all, there are 17 stations with enough observations to make a meaningful analysis. Most of them, however, will be presented in appendices, due to the sheer amount of them.

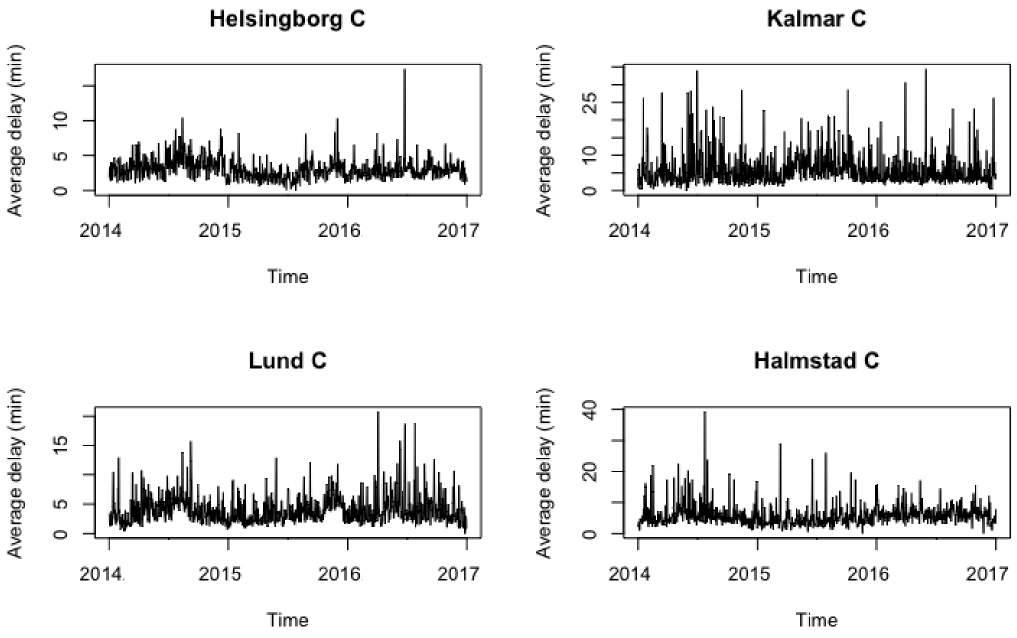


Figure 2.1: Time series showing the average daily delays for four railway stations in southern Sweden all stretching over the tree year span from 2014 to 2017

As can be seen in Figure 2.1, the θ-shift means all values in our time series are positive. Hence, it is possible to fit a gamma distribution to them.

3. Fitting Stochastic Models

3.1 Transforming the Data

This section is dedicated to transforming the data using the theory and methods described in section 4. The classical methods of time series analysis assume that the observations are normally distributed. Figure 3.1.1 shows histograms of the density of the train delays for a sample of our stations, the observations are skewed to the left and do not exhibit the symmetrical properties of a normally distributed random variable. Therefore, the inverse method given in section 4.1 is used to transform the data in order to acquire normally distributed observations.

The distribution of mean daily train delays, adjusted by the constant θ, for every station is shown in Figure 3.1.1. From these histograms, it is reasonable to suppose that the observations follow a gamma distribution. Table 3.1.1 shows the method of moments estimators of gamma parameters for each station, as well as the magnitude of θ for each station. The blue curves on the histograms are the simulated gamma distributions with the parameters estimated in Table 3.1.1.

Table 3.1.1: Estimates of gamma distribution parameters for delays at each station

Station	α = Shape	β = Rate	θ = Shift
Kalmar C	1.69	0.30	-1.75
Halmstad C	2.95	0.52	-4.00
Helsingborg C	4.53	1.49	-1.56
Hyllie	2.58	1.05	-2.05
Karlskrona C	4.85	0.58	-7.00
Kristianstad C	5.41	1.74	-2.11
Lund C	2.74	0.70	-1.10
Markaryd	2.51	1.18	-2.00
Pepparholm	3.56	1.32	-2.00
Simrishamn	3.53	2.04	-1.00
Ystad	2.99	0.79	-2.00
Malmö	0.81	0.15	-3.00
Ängelholm	1.48	0.65	-2.86
Bromölla	3.24	0.97	-3.00
Hässleholm	6.79	1.95	-1.73
Höör	1.75	0.44	-2.00
Göteborg	1.92	0.46	-3.44

The inverse-method described in section 4.1 is used to transform the gamma distributed observations into normal distributions. We use the inverse c.d.f. of the standardized normal distribution, meaning that our transformed observations should follow a $N(0,1)$ distribution assuming that the delays are indeed gamma distributed. The quantile plots of the transformed observations are shown in Figure 3.1.2. These plots describe how well the empirical distributions of the transformed observations follow the theoretical normal distribution. The observations are represented by dots, and the theoretical distribution is represented by the line. If the empirical distribution follows the normal distribution perfectly, all the dots are on the line.

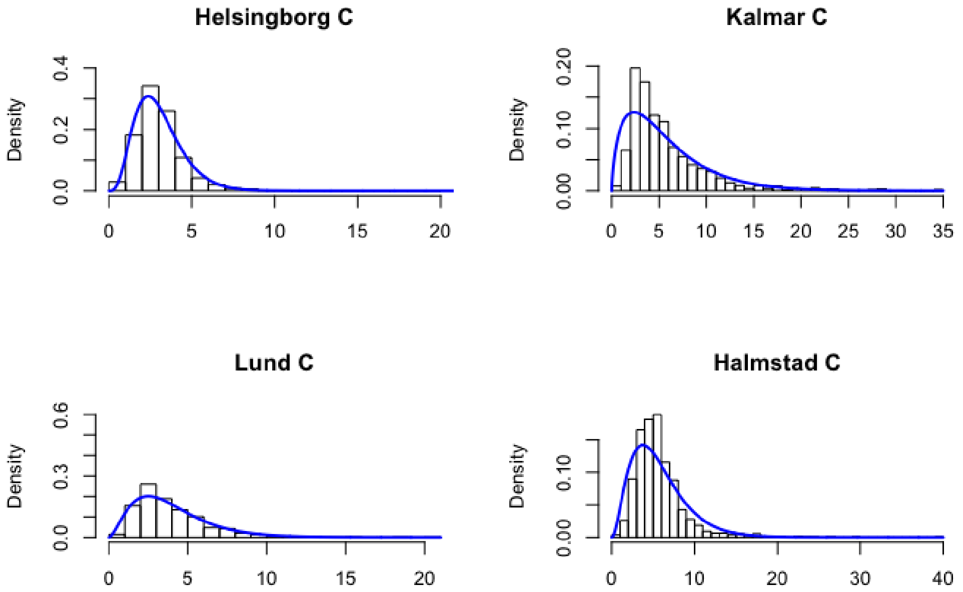


Figure 3.1.1: Histograms of shifted data, with theoretical gamma p.d.f.

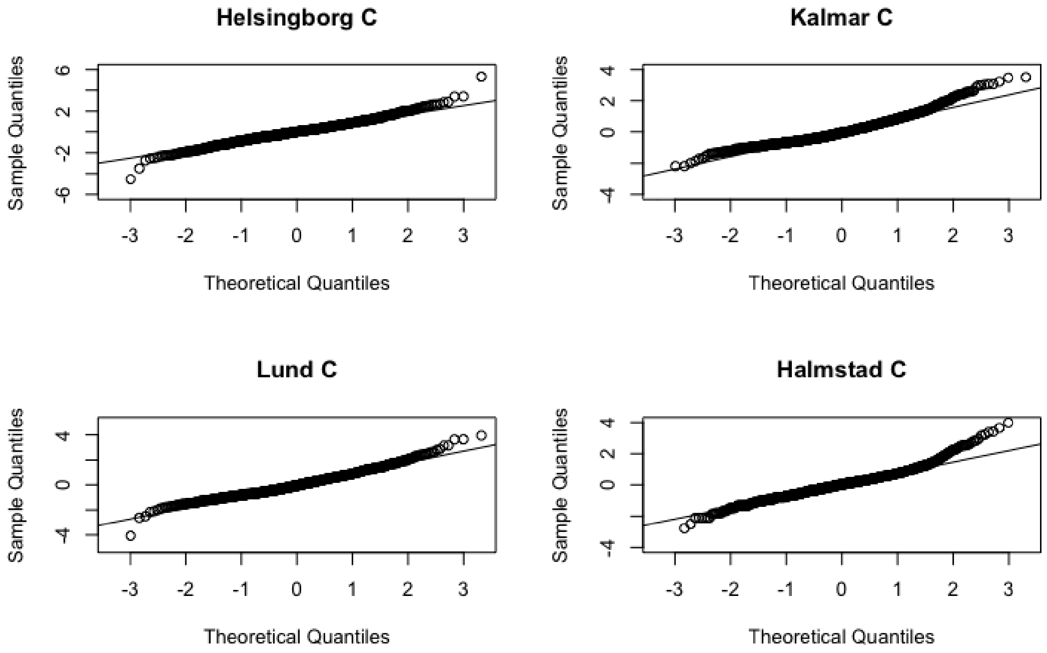


Figure 3.1.2: Normal quantile plots showing normality of transformed data

Evidently, there is a tendency of deviation from Gaussianity in the transformed data, indicated by the heavy tails and skewness in the quantile plots and in Figure 3.1.3 below. Theoretically, if the fitting of gamma parameters was right, the transformed data should follow the normal distribution. However, there are two plausible explanations as to why it does not; one being the method used to estimate parameters, and the other that transformation through the gamma distribution is not a suitable approach. When it comes to estimation, the method of moments has been used to fit gamma parameters to the delay data. An alternative is using maximum likelihood estimators instead, as they may estimate the parameters more precisely.

As for fitting the wrong distribution, it is possible that the asymmetric Laplace distribution is more appropriate for the transformation of train delay data. For the remaining stations, see appendix 3.

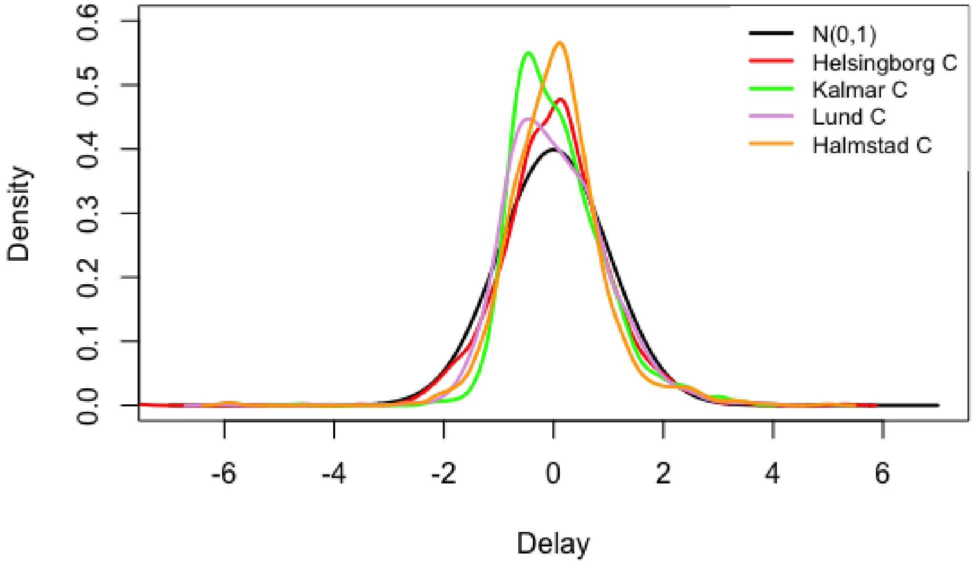


Figure 3.1.3: Theoretical density of N(0,1) plotted with empirical densities of transformed delay data

In order to gain a better fit, the asymmetric Laplace distribution is fitted to the four stations that have been presented so far. The asymmetric Laplace distribution might be more suited to our data, seeing as the gamma distribution does not manage to capture the rapid increase in density around the mode of the distributions at some stations (see Kalmar C and Halmstad C in Figure 3.1.1. in particular). The maximum likelihood method is used to fit asymmetric Laplace parameters to the delay data. In Figure 3.1.3, the estimated asymmetric Laplace probability density function (p.d.f) has been plotted over the histograms showing the densities of the data.

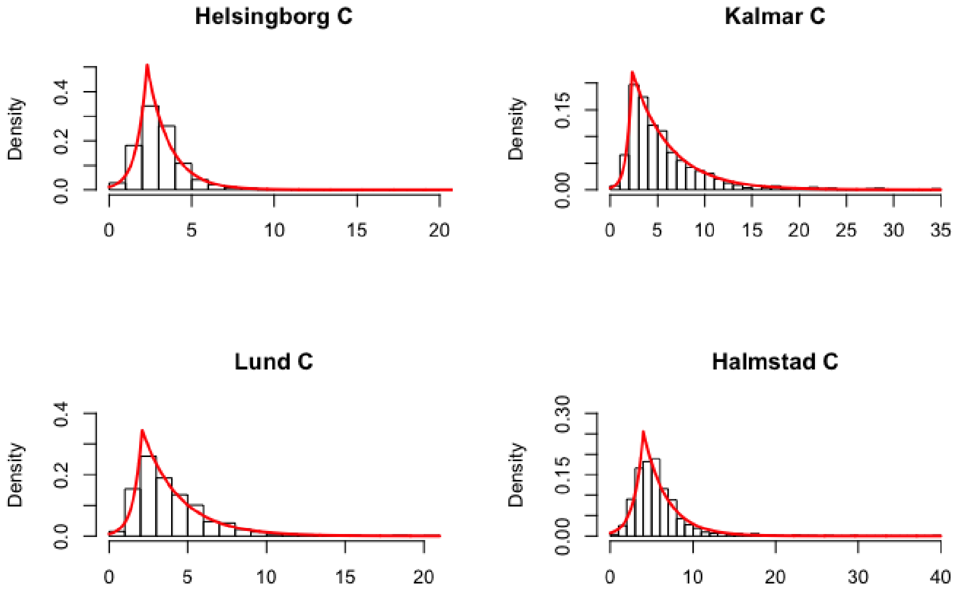


Figure 3.1.3: Fitted asymmetric Laplace p.d.f. plotted over density histograms

As is seen i Figure 3.1.3, the asymmetric Laplace p.d.f. manages to capture the rapid rise and decrease in density around the mode of the distributions better than its gamma equivalent. To carry on the Laplace analysis, the delays are transformed again, this time using the estimated Laplace distribution. Figure 3.1.4. shows the normal quantile plots of the delays after transformation using the asymmetric Laplace distribution.

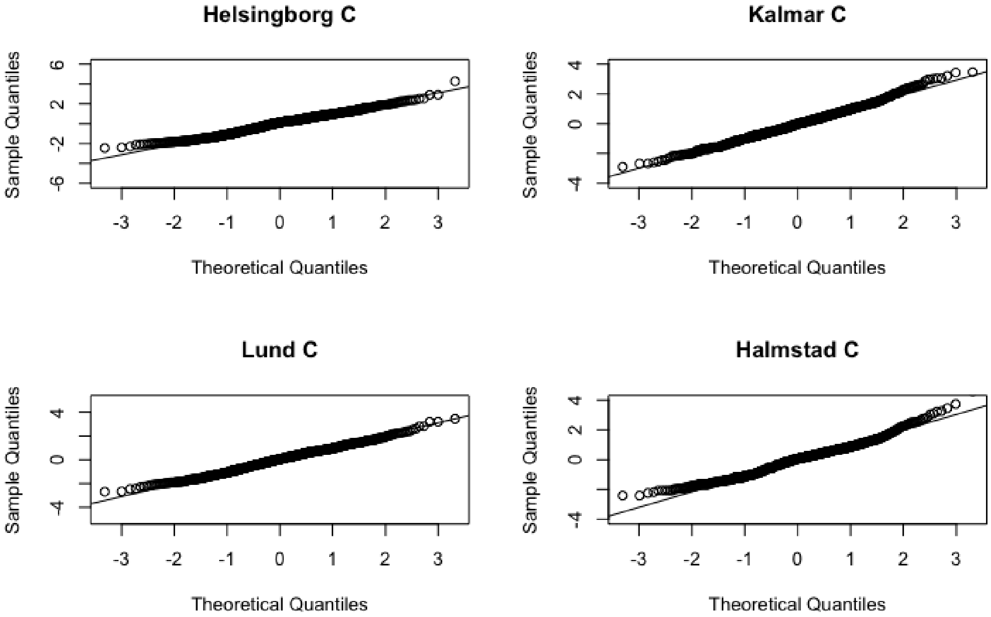


Figure 3.1.4: Normal quantile plots showing normality of transformed data

The Laplace transformation seems to handle the tails of our distributions better than the gamma transformation in general. Comparing Figure 3.1.4. to Figure 3.1.2, the deviation from Gaussianity seems to be more moderate at low and high quantiles when the Laplace distribution is used. In particular, this can be seen in the quantile plots for Halmstad C, where the tails are much better adapted to the normal distribution when Laplace is used. Figure 3.1.5, again, shows the theoretical and empirical densities plotted together.

The results shown in Figure 3.1.5 support the fact that the asymmetric Laplace distribution is better suited to train delays than the gamma distribution. Hence, we seem to have found a transformation that makes it possible to fit multivariate time series models to our train delay data. Maximum likelihood estimates of asymmetric Laplace parameters are given in Table 3.1.2.

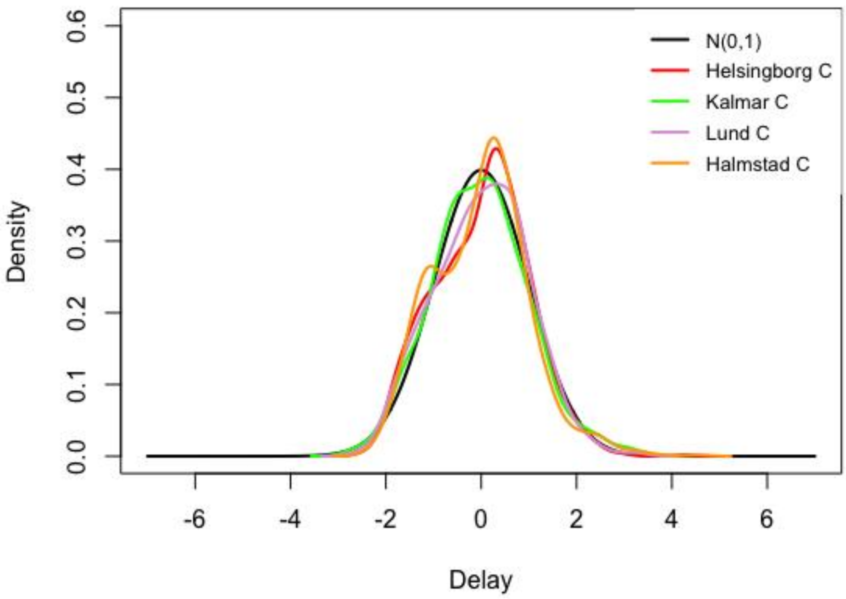


Figure 3.1.5: Theoretical p.d.f. of the N(0,1) distribution plotted with empirical densities of train delay data

Table 3.1.2: Estimates of asymmetric Laplace distribution parameters for delays at each station

Station	θ	σ	κ
Bromölla	3.00	0.67	0.44
Göteborg C	2.48	0.62	0.25
Halmstad C	4.00	0.80	0.29
Helsingborg C	2.29	0.42	0.31
Hyllie	1.34	0.24	0.17
Hässleholm C	2.88	0.38	0.33
Höör	2.93	0.75	0.35
Karlskrona C	5.51	0.66	0.18
Kalmar C	2.32	0.49	0.12
Kristianstad C	2.30	0.34	0.27
Lund C	2.10	0.43	0.18
Malmö C	2.50	0.62	0.17
Markaryd	1.37	0.18	0.18
Pepparholm	1.69	1.69	1.69
Simrishamn	1.11	0.21	0.24
Ystad	2.22	0.41	0.20
Ängelholm	1.00	0.38	0.16

3.2 Time Series Modelling

Table 3.2.1 shows the cross-correlation matrix for the stations Ystad, Lund, Kristianstad and Helsingborg for the original and transformed data up to three lags. We use the ccm(…) function in the R-package MTS compute the correlation matrices. In general, the effect of the gamma transformation on time dependence is relatively mild, although the cross-correlations grow stronger. Also, the cross-correlations look to be significant only at lag 0 and at lag 1, which is reasonable since the mean delay of one day is not likely to be affected by the mean delay two or more days before. Consequently, the transformation seems to reveal stronger correlations, and therefore time-dependence, between stations. For data transformed through the asymmetric Laplace distribution, se appendix 5.

Univariate autocorrelation functions are given in Table 3.2.2, and graphically illustrated in appendix 2. Table 3.2.2. shows that, again, the transformations amplify the magnitude of time dependence at most lags. Also, the autocorrelations in appendix 2 indicate that there is a seasonal trend in the data, meaning that this will have to be considered before preforming any time series analysis. This, however, is thought to be outside the scope of our paper. Consequently, the transformation reveals stronger relations in the data through time and between stations.

Table 3.2.1: Cross-correlation matrices of original and transformed data for station, Ystad, Lund, Kristianstad and Helsingborg at lags 0, 1, 2 and 3

	Original Data	Transformed Data (Gamma)
$P(0)$	$\begin{pmatrix} 1 & 0.385 & 0.278 & 0.531 \\ 0.385 & 1 & 0.315 & 0.334 \\ 0.278 & 0.315 & 1 & 0.296 \\ 0.531 & 0.334 & 0.296 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.397 & 0.253 & 0.479 \\ 0.397 & 1 & 0.288 & 0.339 \\ 0.253 & 0.288 & 1 & 0.300 \\ 0.479 & 0.339 & 0.300 & 1 \end{pmatrix}$
$P(1)$	$\begin{pmatrix} 0.232 & 0.180 & 0.150 & 0.214 \\ 0.261 & 0.388 & 0.212 & 0.285 \\ 0.205 & 0.193 & 0.415 & 0.224 \\ 0.162 & 0.164 & 0.220 & 0.392 \end{pmatrix}$	$\begin{pmatrix} 0.260 & 0.221 & 0.154 & 0.207 \\ 0.275 & 0.434 & 0.200 & 0.275 \\ 0.170 & 0.186 & 0.430 & 0.217 \\ 0.172 & 0.208 & 0.226 & 0.443 \end{pmatrix}$
$P(2)$	$\begin{pmatrix} 0.1427 & 0.0702 & 0.104 & 0.0743 \\ 0.1403 & 0.1615 & 0.156 & 0.1443 \\ 0.1905 & 0.1741 & 0.292 & 0.2077 \\ 0.0944 & 0.0680 & 0.175 & 0.2501 \end{pmatrix}$	$\begin{pmatrix} 0.1682 & 0.0924 & 0.125 & 0.0663 \\ 0.1529 & 0.2302 & 0.158 & 0.1542 \\ 0.1464 & 0.1780 & 0.339 & 0.1942 \\ 0.0926 & 0.1003 & 0.205 & 0.2692 \end{pmatrix}$
$P(3)$	$\begin{pmatrix} 0.0949 & 0.0629 & 0.0700 & 0.0876 \\ 0.0408 & 0.0757 & 0.0826 & 0.0845 \\ 0.1161 & 0.0943 & 0.2485 & 0.1891 \\ 0.0433 & 0.0639 & 0.1570 & 0.2305 \end{pmatrix}$	$\begin{pmatrix} 0.1260 & 0.0844 & 0.0646 & 0.0829 \\ 0.0564 & 0.1292 & 0.0930 & 0.0944 \\ 0.1079 & 0.1023 & 0.2855 & 0.1881 \\ 0.0461 & 0.0904 & 0.1766 & 0.2510 \end{pmatrix}$

Table 3.2.2. Autocorrelations of the four given stations, all stretching 10 lags back in time

	Original				**Transformed (Gamma)**				**Transformed (Laplace)**			
Lag	Hels.	Kalm.	Lund	Halm.	Hels.	Kalm.	Lund	Halm.	Hels.	Kalm.	Lund	Halm.
1	0.392	0.092	0.389	0.153	0.433	0.163	0.435	0.234	0.462	0.193	0.431	0.273
2	0.250	0.084	0.162	0.082	0.269	0.098	0.230	0.149	0.270	0.099	0.236	0.186
3	0.231	0.095	0.076	0.132	0.251	0.122	0.129	0.170	0.252	0.119	0.146	0.214
4	0.238	0.063	0.067	0.118	0.250	0.077	0.125	0.162	0.253	0.086	0.137	0.211
5	0.220	0.019	0.115	0.104	0.243	0.020	0.197	0.174	0.243	0.019	0.221	0.211
6	0.294	0.024	0.173	0.105	0.340	0.082	0.248	0.139	0.358	0.096	0.273	0.189
7	0.419	0.144	0.209	0.104	0.479	0.210	0.263	0.161	0.519	0.236	0.282	0.218
8	0.263	0.036	0.186	0.111	0.306	0.075	0.210	0.138	0.327	0.107	0.212	0.169
9	0.158	0.066	0.086	0.017	0.174	0.076	0.135	0.061	0.167	0.067	0.153	0.095
10	0.162	0.027	0.035	0.063	0.186	0.067	0.074	0.110	0.175	0.085	0.102	0.144

4. Theory and Methods

4.1 Probability models

In the following section, the relevant probability models are presented and the method of transforming random variables is discussed. Given that the classical methods of estimating the parameters of ARMA (p, q) and VARMA (p, q) processes assume that the time series are samples of normally distributed random variables, this poses a significant problem. The problem can be circumvented using some elementary probability theory.

Suppose that $X \sim N(\mu, \sigma)$ is a normally distributed random variable of the delay of a given train at a station. Then the probability density function (p.d.f.) of X is given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

For the normal distribution, $E[X] = \mu$ and $Var[X] = \sigma^2$. Furthermore, the probability density function of the normal distribution is symmetrical around μ, this can be seen in Figure 4.1.1 (Hogg and Tanis). However, the normal distribution is not an accurate model of train delays for our random variable X. This was shown by the distribution of the data in section 3, but intuitively it is unlikely since technical problems or extreme weather produce very large delays which skew the distribution of train delays.

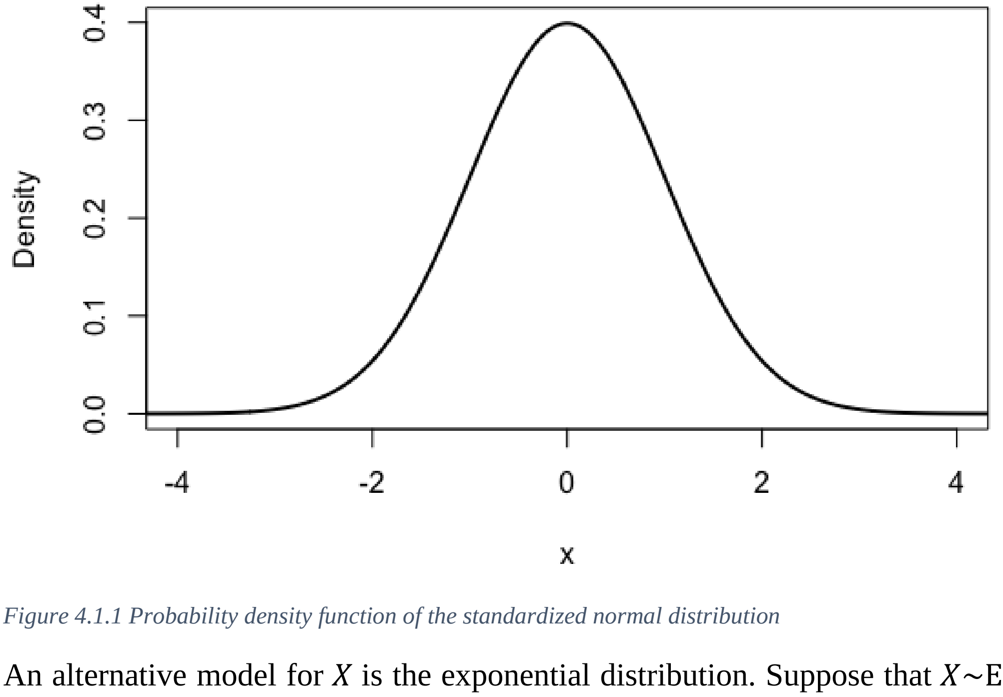


Figure 4.1.1 Probability density function of the standardized normal distribution

An alternative model for X is the exponential distribution. Suppose that $X \sim \text{Exp}(\theta)$ with $E(X) = \frac{1}{\theta}$ and $Var(X) = \frac{1}{\theta^2}$, then the probability density function is given by:

$$f_X(x) = \begin{cases} \theta e^{-\theta x}, & 0 < x < \infty \\ 0, & elsewhere \end{cases}$$

In this model the density of train delays declines exponentially as the magnitude of delays increases so that the probability of the train arriving in proximity to the time-table is very high while the probability of large delays is very low. Perhaps it is charitable towards the Swedish railroad system to assume that X follows an exponential distribution but since we transform X to produce our time series $Y_{t,m}$ so that $Y_{t,m}$ follows a gamma distribution, it will not matter whether we assume X to be an exponential or gamma distribution. Besides, a previous study on the distribution of train delays in Sweden suggests the exponential distribution is an accurate model of train delays (Bergström and Kruger).

Suppose we introduce a new random variable which follows a gamma distribution with a parameter determining the shape of the p.d.f, β, and a parameter determining the rate, α. Call this random variable $Z\sim\Gamma(\alpha,\beta)$ with a p.d.f given by:

$$f_Z(z)=\begin{cases}\dfrac{\beta^{\alpha}}{\Gamma(\alpha)}z^{\alpha-1}e^{-\beta z}, & z,\alpha,\beta>0\\ 0, & elsewhere\end{cases}$$

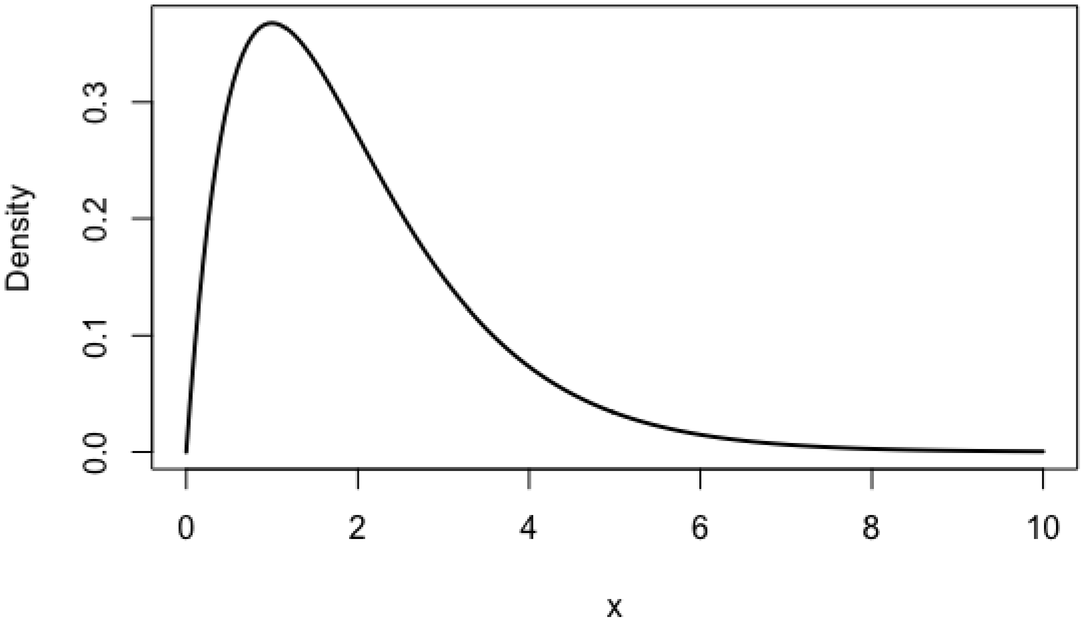


Figure 4.1.2 Probability density function of gamma distribution with parameters $\alpha=2$ and $\beta=1$

The density function of a $\Gamma(1,2)$ is plotted in 4.1.2 (Hogg and Tanis). Our time series $Y_{t,m}$ is the daily average delay at the station m. If the train delays are exponentially distributed we can define and interpret the sample of train delays of size n as a sample from independent and identically distributed random variables $(X_1,\ldots,X_n)$. Suppose that Z is an observation in $Y_{t,m}$, then $Z=\frac{\sum_{i=1}^{n}X_i}{n}$, where n is the number of train arrivals in a day at the station. The distribution of Z is derived using the moment-generating function of the exponential distribution defined as:

$$M_X(t) = E(e^{Xt}) = \int_0^{\infty} e^{Xt}\theta e^{-\theta x} = \frac{\theta}{\theta - t} \qquad 4.1.1$$

$$M_Z(t) = E(e^{Zt}) = M_X(t) = E\left(e^{\frac{\sum_{i=1}^{n} X_i}{n}t}\right) = \prod_{i}^{n} e^{\frac{t}{n}X_i} = \left(M_X\left(\frac{t}{n}\right)\right)^n = \left(\frac{1}{1-\frac{t\theta}{n}}\right)^n \qquad 4.1.2$$

The moment-generating function of $X \sim \Gamma(\alpha, \beta)$ in turn is given by:

$$M_x(t) = \left(\frac{\beta}{\beta - t}\right)^{\alpha}$$

Since the moment-generating function is unique for every given probability density function, it follows from equation 4.1.2 that $Z \sim \Gamma(n, \frac{\theta}{n})$ (Hogg and Tanis). Note that if we assume that $X \sim \Gamma(\alpha, \beta)$ we get the same result. An example of a gamma p.d.f. is showed in Figure 4.1.2.

In many cases the assumption of independence is reasonable considering that our data is only for final stations on a route, meaning that the time between observations are long enough, so that a previous delay will not affect the timeliness of the next train. It is therefore reasonable to assume that the observations in $Y_{t,m}$ approximately follow a gamma distribution. However, the fact that the gamma distribution did not fit some stations convincingly suggests that the data for some stations violate this assumption.

In section 3.1 we found that the gamma distribution did not fit the data very well. This could be due to the method of moment estimators or the fact that the mean delays do not follow a gamma distribution. An alternative model that could be used is the asymmetric Laplace distribution. Assuming that $X \sim A\text{Laplace}(\theta, \kappa, \sigma)$ then the p.d.f. is given by:

$$f_{\theta,\kappa,\sigma}(x) = \frac{\sqrt{2}}{\sigma}\frac{\kappa}{1+\kappa^2}\begin{cases} \exp\left(\frac{-\sqrt{2}\,\kappa}{\sigma}|x-\theta|\right), & if\ x \geq \theta \\ \exp\left(\frac{-\sqrt{2}}{\sigma}|x-\theta|\right), & if\ x < \theta \end{cases}$$

Figure 4.1.3 shows the p.d.f of the asymmetric Laplace distribution for different parameter values. The Laplace distribution increases and decreases sharply before and after the mode value of X. Looking at the histograms in Figure 3.1.1, the mode of the gamma distribution is consistently lower than the mode of the data. This suggests that the asymmetric Laplace distribution may provide a better fit to the data (Kotz, Kozubowski and Podgorski).

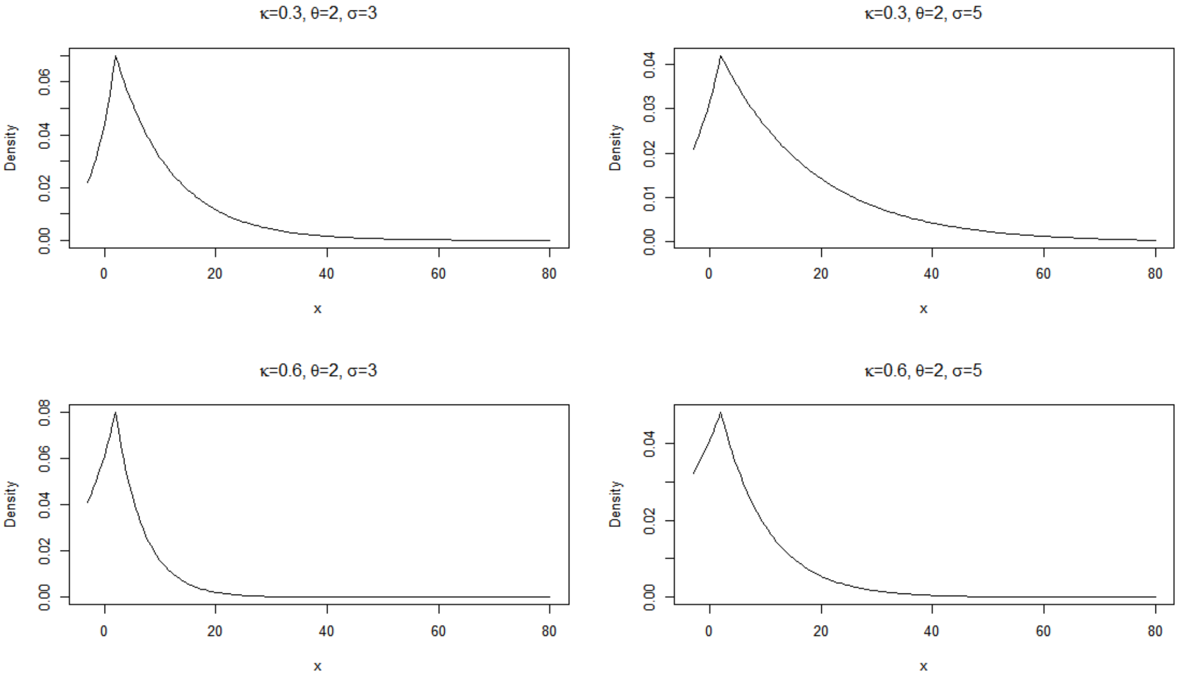


Figure 4.1.3: Asymmetric Laplace distribution probability distribution function

To estimate the parameters of the gamma distribution we use the method of moments, which means we equate the sample moments of our data with the theoretical moments of the gamma distribution. The kth moment is defined as $E[X^k]$ meaning that the first and second moments of the gamma distribution are given by:

$$E[X] = \int_0^\infty x \cdot f(x)dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^\alpha e^{-\beta x} dx = \frac{\alpha}{\beta} \tag{4.1.3}$$

$$E[X^2] = \int_0^\infty x^2 \cdot f(x)dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha+1} e^{-\beta x} dx = \frac{\alpha(\alpha-1)}{\beta^2} \tag{4.1.4}$$

Notice that the right-hand sides of equations (4.1.3) and (4.1.4) constitute a system of equations. Solving for α and β gives the method of moments estimators:

$$\alpha = \frac{E[X]^2}{E[X^2] - E[X]^2} \tag{4.1.5}$$

$$\beta = \frac{E[X]}{E[X^2] - E[X]^2} \tag{4.1.6}$$

Where α corresponds to the shape and β to the rate of the distribution. We can now use the estimators for $E[X]$ and $E[X^2] - E[X]^2$ to estimate α and β:

$$\hat{\alpha}_{MM} = \frac{\bar{X}^2}{\hat{\sigma}^2}$$

$$\hat{\beta}_{MM} = \frac{\bar{X}}{\hat{\sigma}^2}$$

The method of moments estimators are both functions of $\bar{X}$, meaning that very large values will have large effects on the estimators. This poses a potential problem for data that follows a gamma distribution since it is a characterized by a concentration of small values and a small amount of very large values. The method of moment estimator is consistent, but if an unusual amount of large values are generated, this will have a large effect on the estimators.

Seeing as the moments estimators did not turn out a great fit, maximum likelihood estimators might be used to improve the accuracy. The estimators can be obtained by solving an optimization problem along the following relations for the gamma likelihood:

$$L(\underline{x}, \alpha, \beta) = \prod_{i}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} {x_i}^{\alpha-1} e^{-\beta x_i} \qquad (4.1.7)$$

$$\Rightarrow \log L\left(\underline{x}, \alpha, \beta\right) = (n\alpha)\log(\beta) - n\log\big(\Gamma(\alpha)\big) + (\alpha - 1)\sum_{i=1}^{n} \log x_i - \beta \sum_{i=1}^{n} x_i \qquad (4.1.8)$$

$$\Rightarrow \frac{\partial L}{\partial \alpha} = n \log \beta - n\frac{d}{d\alpha}\log\big(\Gamma(\alpha)\big) + \sum_{i=1}^{n} \log x_i \qquad (4.1.9)$$

$$\Rightarrow \frac{\partial L}{\partial \beta} = n\frac{\alpha}{\beta} - \sum_{i=1}^{n} x_i \qquad (4.1.10)$$

(4.1.9) and (4.1.10) then constitute a system of equations that can be solved numerically for α and β (Choi and Wette). The maximum-likelihood estimators may produce better fits to the data.

The parameters of the asymmetric Laplace distribution can be estimated by maximizing the likelihood function. The full derivation of the estimators is given by Kotz et al., 2002 (Kotz, Kozubowski and Podgorski). In practice, this method of estimation is included in the “ald” package in R.

Knowing the distribution of $Y_{t,m}$ allows us to transform the data to a normal distribution. Suppose that the random variable Z follows a gamma distribution with p.d.f $f_Z(z)$ and c.d.f $F_Z(z)$, and the random variable X follows a normal distribution with p.d.f. $f_X(x)$ and c.d.f. $F_X(x)$, and we define a uniform distribution between zero and one as $U \sim [0,1]$ then, using some elementary probability theory, we get the following:

$$P(Z \leq z) = P(F_Z^{-1}(U) \leq z) = P\big(F_Z\big(F_Z^{-1}(U)\big) \leq F_Z(z)\big) = P\big(U \leq F_Z(z)\big)$$

Showing that $F_Z(z)$ follows a uniform distribution. Since any c.d.f. $F_Z(z)$ will map to a set A defined on the interval [0,1] we can transform $f_Z(z)$ to the normal p.d.f. $f_X(x)$ using $F_X^{-1}(x)$:

$$F_Z(z) \to A$$

$$F_X^{-1}(a) \to x,$$

where $a \epsilon A$. Consequently, if we plot $F_X^{-1}(a)$ for all elements in A, we get the normally distributed variable X. This method is inspired by the inverse method for random number generation, but since we use the empirical cumulative distribution function to compute the values in A, the generated values from $F_X^{-1}(a)$ are random only in so far as the empirical distribution function is generated from a random variable (Devroye).

After successfully fitting a distribution and using it to transform the data into normality, the theory presented in section 4.2 can be used to build time series models with regards to the delay series.

4.2 Time series analysis

A time series is a series of values generated by a stochastic process over time, for example GDP, interest rates, and, of course, train delays. Time series analysis mainly focuses on trying to determine the characteristics of the stochastic process in question. Often, time series are believed to consist of completely random terms (white noise) called e_t, and autoregressive terms Y_{t-k}, where t is a given point in time and k determines the order of the lag of autoregression. These can be combined into different types of time series models, where the value at a point in time Y_t, is explained by recursions on either random terms, or by autoregression.

The simplest form of time series is one made up entirely of white noise terms, commonly known as moving average (MA) processes of order q. An MA(q) process is defined as follows:

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

where θ_i are unknown coefficients determining the recursive power of each white noise term (Cryer and Chan). The MA(q) process is often complemented by a time series model of autoregressive nature, meaning that the value at a certain time t is calculated as the sum of weighted earlier terms. The autoregressive process, AR(p), of order p is defined as below:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$

where e_t is a white noise error term and ϕ_i are unknown coefficients measuring the strength of the recursion. Typically the two are combined, creating a mixed autoregressive moving average (ARMA) model. The ARMA (p,q) process is defined as

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q}$$

where, again, ϕ_i and θ_i are unknown.

A multivariate time series $\boldsymbol{Y_t} = (Y_{t,1}, Y_{t,2}, \dots, Y_{t,k})$ is a row vector of univariate time series $Y_{t,m}$ where t denotes the point in time and $1 \leq m \leq k$ denotes the time series. In our data m denotes the station that generated the time series. If our multivariate time series $\boldsymbol{Y_t}$ is a vector with dimension $1 \times K$, and we assume that every univariate time series $Y_{t,m}$ with length k is dependent on $Y_{t,j}$ for $j \neq m$, through an autoregressive process of order p and a moving average process of order q, then $\boldsymbol{Y_t}$ can be represented as a vector mixed autoregressive moving average process of orders p and q, VARMA(p, q), which is defined as follows:

$$\boldsymbol{Y_t'} = \sum_{i=1}^{p} \boldsymbol{\Phi_i Y'_{t-i}} + \sum_{i=1}^{q} \boldsymbol{\Theta_i \varepsilon'_{t-i}} + \boldsymbol{\varepsilon'_t}$$

Where $\boldsymbol{Y_t'}$ is the transpose of our multivariate time series $\boldsymbol{Y_t}$, $\boldsymbol{Y_{t-i}} = (Y_{1,t-i}, Y_{2,t-i}, \dots, Y_{k,t-i})$ is a row vector of the values of the time series $Y_{j,t-i}$ at lag i. The matrices $\boldsymbol{\Phi_i}$ and $\boldsymbol{\Theta_i}$ are the AR and MA components respectively of order i. Finally $\boldsymbol{\varepsilon_t}$ is a column vector of the white noise components.

For the vector ARMA$(1,1)$ process we have:

$$\boldsymbol{Y_t'} = \boldsymbol{\Phi_1 Y'_{t-i}} + \boldsymbol{\Theta_1 \varepsilon'_{t-i}} + \boldsymbol{\varepsilon'_t} \tag{4.2.1}$$

$$\begin{pmatrix} \phi_{1,1} & \cdots & \phi_{1,k} \\ \vdots & \ddots & \vdots \\ \phi_{k,1} & \cdots & \phi_{k,k} \end{pmatrix} \begin{pmatrix} Y_{1,t-1} \\ \vdots \\ Y_{k,t-1} \end{pmatrix} + \begin{pmatrix} \theta_{1,1} & \cdots & \theta_{1,k} \\ \vdots & \ddots & \vdots \\ \theta_{k,1} & \cdots & \theta_{k,k} \end{pmatrix} \begin{pmatrix} \varepsilon_{1,t-1} \\ \vdots \\ \varepsilon_{k,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \vdots \\ \varepsilon_{k,t} \end{pmatrix} \tag{4.2.2}$$

$$\begin{cases} Y_{1,t} = \ \phi_{1,1}Y_{1,t-1} + \phi_{1,2}Y_{2,t-1} + \dots + \phi_{1,k}Y_{k,t-1} + \theta_{1,1}\varepsilon_{1,t-1} + \dots + \theta_{1,k}\varepsilon_{k,t-1} \dots + \varepsilon_{1,t} \\ \vdots \\ Y_{k,t} = \ \phi_{k,1}Y_{1,t-1} + \phi_{k,2}Y_{2,t-1} + \dots + \phi_{k,k}Y_{k,t-1} + \theta_{k,1}\varepsilon_{1,t-1} + \dots + \theta_{k,k}\varepsilon_{k,t-1} \dots + \varepsilon_{k,t} \end{cases} \tag{4.2.3}$$

From the system of equations in (4.2.3), we clearly see the codependence structure of the time series in $\boldsymbol{Y_t'}$. To understand the concept of stationarity it is useful define the row vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$ which is a vector of the expected values of our time-series in $\boldsymbol{Y_t}$.

To estimate the order and magnitude of the coefficients of a VARMA(p, q) process it is necessary to compute the covariance of our time-series with each other at different time-lags l. This is called the cross-covariance of $Y_{t,m}$ and $Y_{t,j}$ at lag l and is given by:

$$\gamma(l)_{m,j} = E[(Y_{t,m} - \mu_m)(Y_{t+l,j} - \mu_j)]$$

The cross-covariance of all $Y_{t,m}$ and $Y_{t,j}$ for $1 \leq m, j \leq k$ at lag l can be represented by the cross-covariance $K \times K$ matrix given by:

$$\boldsymbol{\Gamma}(l) = \begin{pmatrix} \gamma_{11}(l) & \cdots & \gamma_{1k}(l) \\ \vdots & \ddots & \vdots \\ \gamma_{k1}(l) & \cdots & \gamma_{kk}(1) \end{pmatrix}$$

The diagonal of $\mathbf{\Gamma}(l)$ represents the auto-covariances of the time-series $Y_{t,m}$ at lag l and the elements correspond to $\gamma(l)_{m,j}$ (Reinsel). The cross-correlation function between m and j is given by:

$$\rho(l)_{m,j} = \frac{\gamma(l)_{m,j}}{\sqrt{\left[\gamma(0)_{m,m} \cdot \gamma(0)_{j,j}\right]}}$$

Which gives the cross-correlation matrix at lag l denoted by (l) :

$$\mathbf{P}(l) = \begin{pmatrix} \rho(l)_{1,1} & \cdots & \rho(l)_{1,k} \\ \vdots & \ddots & \vdots \\ \rho(l)_{k,1} & \cdots & \rho(l)_{k,k} \end{pmatrix}$$

In order for statistical inference to be possible, simplifications and assumptions have to be made. When it comes to time series analysis, the vital assumption is that of stationarity. Stationarity comes in two forms; it can be both strict and weak. Throughout this thesis, whenever stationarity is mentioned, it will be in reference to the weaker (or second order) kind. A multivariate time series $\boldsymbol{Y_t}$ is stationary if $\boldsymbol{\mu}_t$ and $\mathbf{\Gamma}(t+l,t), l = 0, \pm 1, \ldots$, are both independent of t (Brockwell and Davis).

A common method of estimating the parameters of the VARMA(p,q) model, assuming that the order of p and q is known, is maximum-likelihood estimation. Assuming that we have a sample of size n from a stochastic variable X with a single parameter θ and probability density given by $f_X(x,\theta)$, then the maximum likelihood function for our random sample is given by $L(\underline{x},\theta) = \prod_{i=1}^{n} f_X(x_i)$. The maximum-likelihood estimator (MLE) of θ is then given by:

$$\hat{\theta} = argmax_\theta \left[\prod_{i=1}^{n} f_X(x_i,\theta)\right]$$

To keep the exposition concise we consider only the MLE of the VAR(1) process which can easily be extended to the VARMA(p,q) case. A VAR(1) is then given by $\boldsymbol{Y}'_{\boldsymbol{t}} = \boldsymbol{\Phi_1 Y'_{t-1}} + \boldsymbol{\varepsilon'_t}$**.** Assuming that our observations in the $1 \times K$ vector $Y_{t,m}$ are normally distributed for all m, the maximum likelihood function of our estimators in the $K \times K$ matrix $\boldsymbol{\Phi_1}$ is:

$$L(\boldsymbol{Y'_t}, \boldsymbol{\Phi}) = [2\pi * \det(\boldsymbol{\Sigma})]^{-k/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{K}[\boldsymbol{Y_t} - \boldsymbol{\Phi_1 Y_{t-1}}]^T \boldsymbol{\Sigma^{-1}} [\boldsymbol{Y_t} - \boldsymbol{\Phi_1 Y_{t-1}}]\right)$$

Where $\boldsymbol{\Sigma}$ is the variance of $\boldsymbol{\varepsilon_t}$ (Jakobsson)**.** Maximizing this likelihood function with respect to the parameters is not a problem. If non-Gaussianity is assumed, the maximum-likelihood function becomes non-linear producing several stationary points. In the multivariate case this becomes very difficult to solve (Lehr and Lii). This is why the data needs to be transformed to a normal distribution.

5. Conclusion

This paper provides the groundwork for the possibility of modelling train delays as multivariate time series. The use of the gamma distribution to transform the train delay data proved mildly successful, meaning that the results show some signs of deviation from the normal distribution. The transformed data, however, are better suited to analysis using multivariate time series than the original data. The gamma transformation revealed an increase in the magnitude of univariate autocorrelations and multivariate cross correlations. Seeing as the gamma transformation did not turn out to be as fruitful as hoped, the asymmetrical Laplace distribution was used in order to make new transformations. This resulted in new time series that are distributionally very close to Gaussianity. Again, the transformed data show signs of a slight increase in time dependence, meaning that autocorrelations and cross correlations were mildly amplified. Hence, the thesis has successfully paved the way to how one could initiate multivariate time series analysis of train delays in Skåne, Sweden.

The transformations seem to have a strengthening effect on time dependence in our time series. Autocorrelations and cross correlations seem to indicate a first order multivariate autoregressive model could be used to model delays. This is believed to be a reasonable conclusion, seeing as time dependence is believed to be unlikely to stretch further back in time than one day. Cross correlations at lags greater than 1 are weak, and any significance is thought of as doubtful. Looking at the univariate time series, there are clear periodical effects. In order to fully prepare the data for multivariate time series analysis, these trends need to be taken into account. The diagonals of the cross-correlation matrices suggest that delays on the previous day might be used as a predictor of train delays today. For the purpose of prediction, it may be sufficient to analyze the univariate time-series.

As the Laplace transformation yielded relatively Gaussian data, the time series theory given in section 4.1 of this paper could be used to study the way in which train delays can be modelled using moving average and autoregressive components, both individually in univariate models and together in multivariate models. Of course, the above mentioned periodicity needs to be addressed before any such model fitting is attempted. Multivariate time series analysis could be used to determine whether there is any sort of interdependence between the stations when it comes to train delays, and perhaps a model as such could be of use in order draw up measures aimed to limit the spreading of delays between heavily linked cities. Furthermore, it could be used as a guide to decide how to distribute resources aimed at prevention of delays, since the dependence might be stronger on some routes, while other routes provide a more natural clearing of delays.

Finally, due to some irregularities in the Swedish Transport Administration’s data, it may be prudent to request data from the train operators themselves.

6. Acknowledgements

Lastly, we would like to extend our sincerest gratitude to Professor Krzysztof Podgórski at the Department of Statistics for his advising us and helping us in reaching a conclusion to this paper. Also, we would like to thank the Swedish Transport Administration for providing us with data, allowing us to perform this analysis.

7. References

Bergström, A., & Kruger, N. A. (2013). *Modeling Passenger Train Delay Distributions.* Stockholm: Centre for Transport Studies.

Brockwell, P. J., & Davis, R. A. (1991). *Time Series: Theory and Methods.* New York: Springer-Verlag.

Choi, S., & Wette, R. (1969). Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their. *Technometrics, Vol. 11, No. 4*, 683-690.

Cryer, J. D., & Chang, K.-S. (2010). *Time Series Analysis With Applications in R.* New York: Springer Science.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* New York: Springer-Verlag New York.

Hogg, R. V., & Tanis, E. A. (2010). *Probability and Statistical Inference, 8th edition.* Upper Saddle River, New Jersey: Pearson Prentice Hall.

Jakobsson, A. (2013). *An Introduction to Time Series Modelling.* Lund: Studentlitteratur AB.

Kotz, S., Kozubowski, T. J., & Podgorski, K. (2001). *The Laplace distribution and generalizations. A revisitwith applications to Communications, Economics, Engineering and Finance.* Boston: Birkhauser.

Kotz, S., Kozubowski, T., & Podgorski, K. (2002). Maximum Likelihood Estimation of Asymmetric Laplace Parameters. *Annals of the Institute of Statistical Mathematics, vol. 54*, 816-826.

Lehr, M., & Lii, K.-S. (1998). Maximum Likelihood Estimates of Non-Gaussian ARMA Models. University of California Riverside.

Reinsel, G. C. (1997). *Elements of Multivariate Time Series Analysis .* New York: Springer-Verlag New York, Inc.

Appendix 1

Table: Summary of data and indication of which stations were used in analysis

Final Destination	Frequency	Final Destination	Frequency
Alvesta	979	Laholm Västra	71
Arlöv	2	Landskrona Östra	18
Båstad Norra	5	Lockarp	227
Billberga	95	**Lund C**	**21314**
Bjuv	3465	Mörrum	1
Bräkne-Hoby	1	**Malmö C**	**16378**
Bromölla	**5972**	**Markaryd**	**10310**
Eslöv	77	**Peberholm**	**71772**
Förslöv	7975	Perstorp	1162
Göteborg C	**23736**	Rydsgård	2
Grantofta	1	Sölvesborg	1
Grevie	2	**Simrishamn**	**19096**
Hässleholm	**39696**	Skurup	5
Höör	**11251**	Stenhag	2
Halmstad C	**11927**	Svågertorp	68
Helsingborg C	**105640**	Teckomatorp	409
Helsingborg godsbangård	51	Tomelilla	2
Hyllie	**46101**	Trelleborg	14044
Jordholmen	1	Triangeln	2
Kävlinge	3872	Växjö	5212
Kalmar C	**12009**	Varberg	78
Karlshamn	4947	**Ystad**	**21860**
Karlskrona C	**16611**	Åkarps Norra	3
Klippan	4	Åstorp	5655
Kristianstad C	**41077**	Älmhult	619
Kungsbacka	1	**Ängelholm**	**17784**
Kvidinge	1	Ödåkra	1

The above table summarizes all stations which were listed as final destination of at least one train during the time of the investigation, as well as the amount of trains that did indeed arrive at each station. Stations which have been used in our models are listed in bold characters.

Appendix 2

In appendix 2 we present autocorrelation functions for all stations that were part of the analysis, adding to the four that were presented in the main body of the paper. Autocorrelations of both original and transformed data are given.

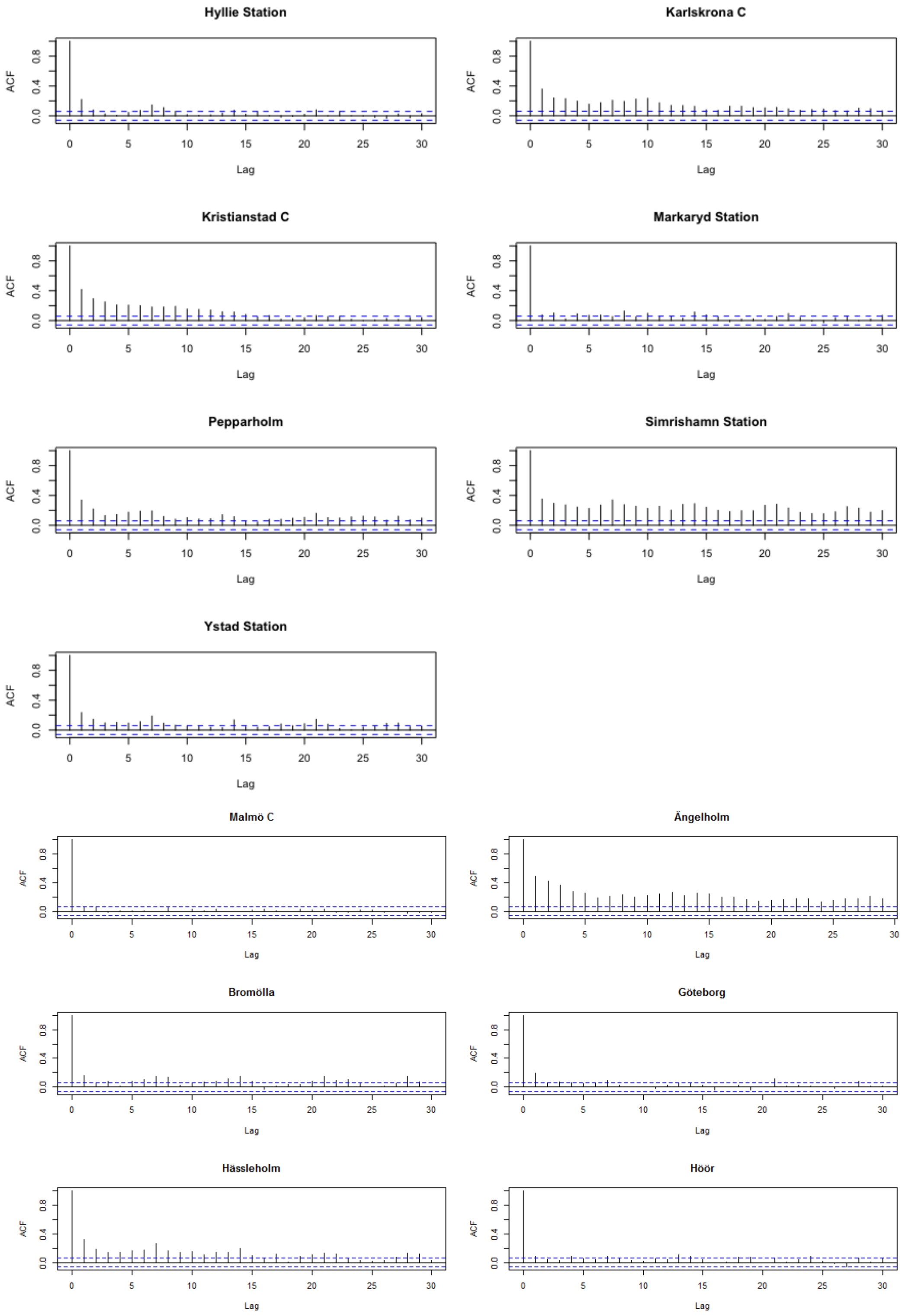


Figure 1: Autocorrelations of original data

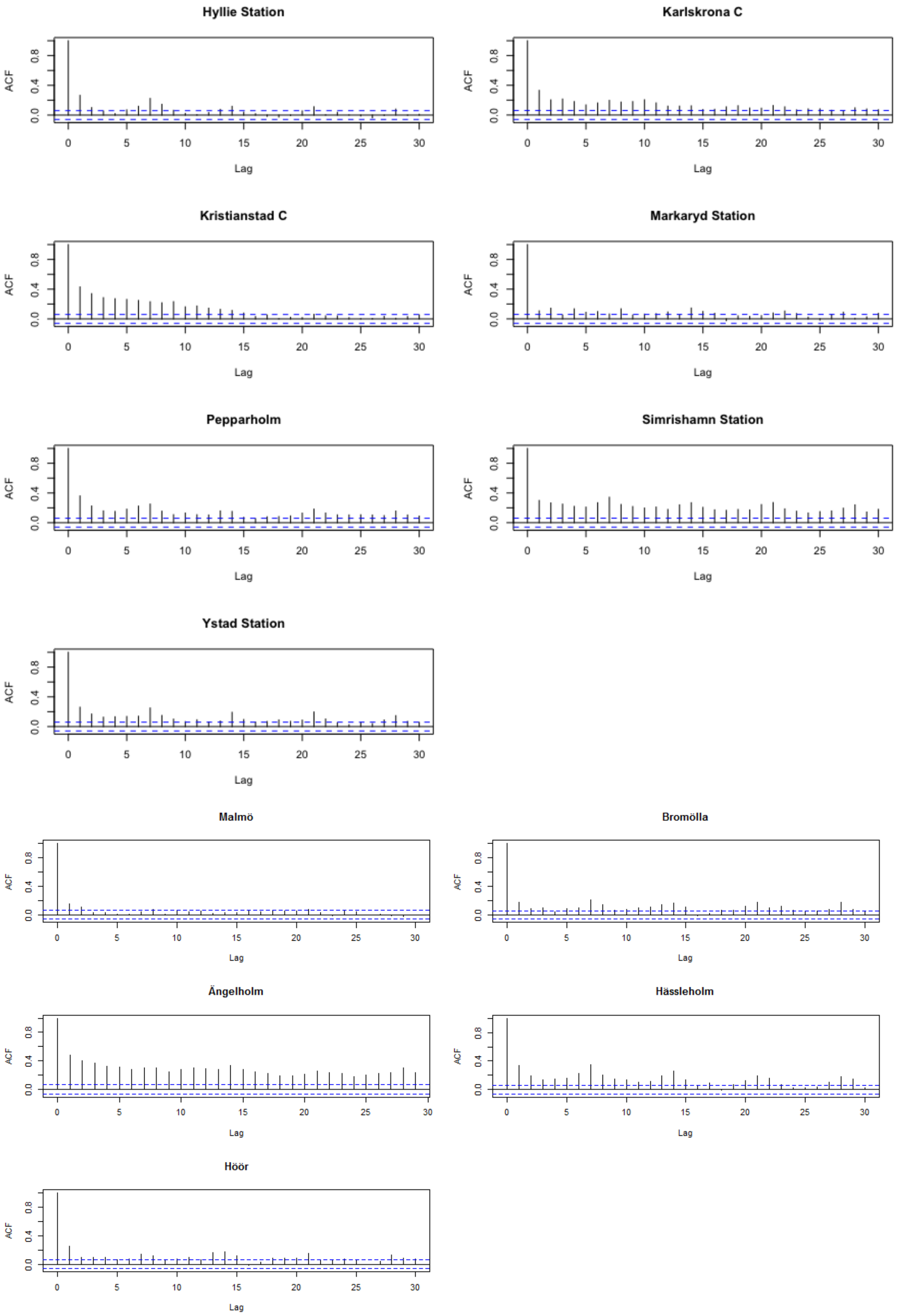
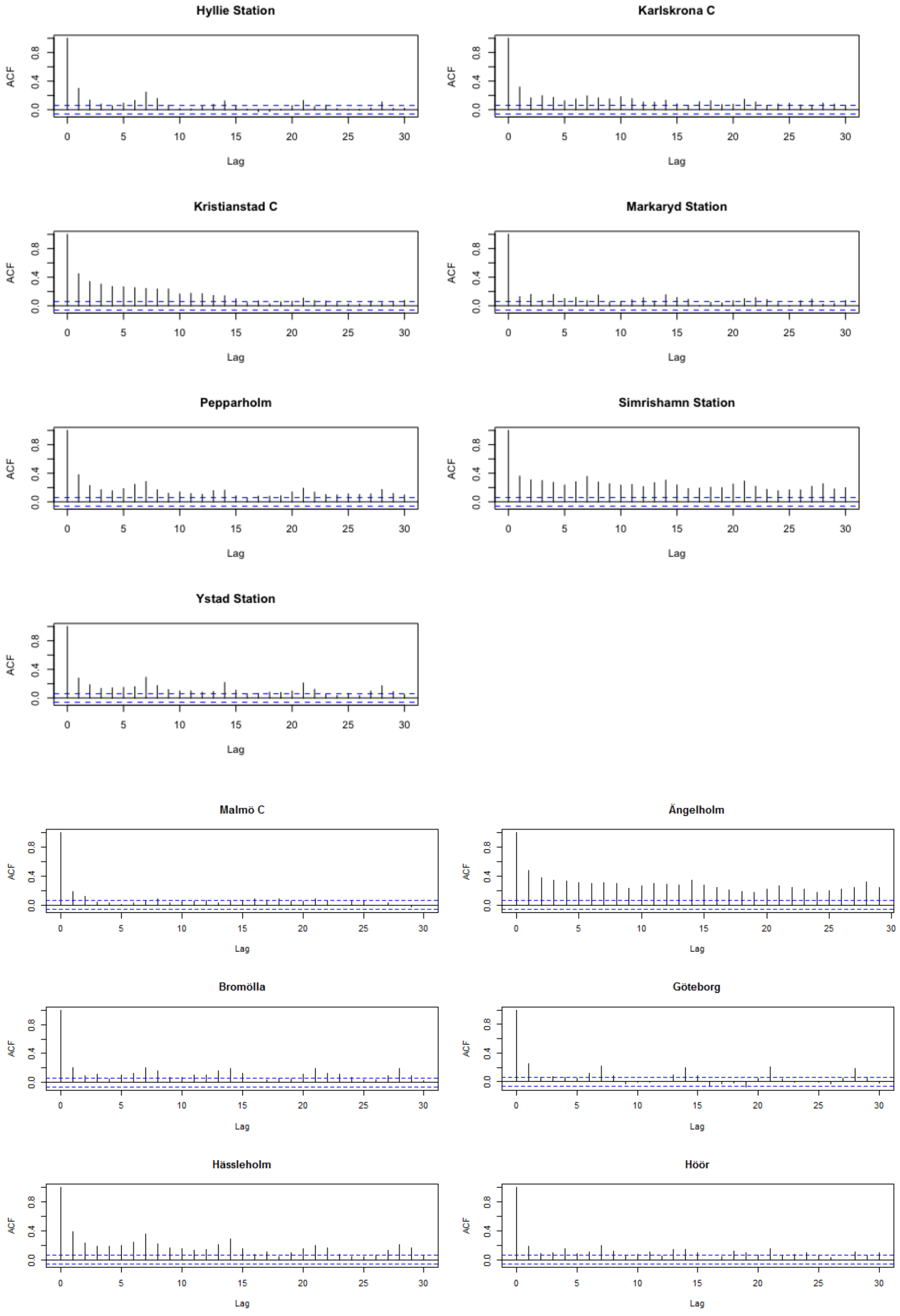


Figure 2: Autocorrelations of data transformed through gamma distribution



Hyllie Station

Karlskrona C

Kristianstad C

Markaryd Station

Pepparholm

Simrishamn Station

Ystad Station

Malmö C

Ängelholm

Bromölla

Göteborg

Hässleholm

Höör

Figure 3: Autocorrelations of data transformed using the asymmetric Laplace distribution

Appendix 3

In appendix 3 density histograms of the shifted train delays will be given, with probability functions of the estimated gamma and asymmetric Laplace distributions plotted on top of them.

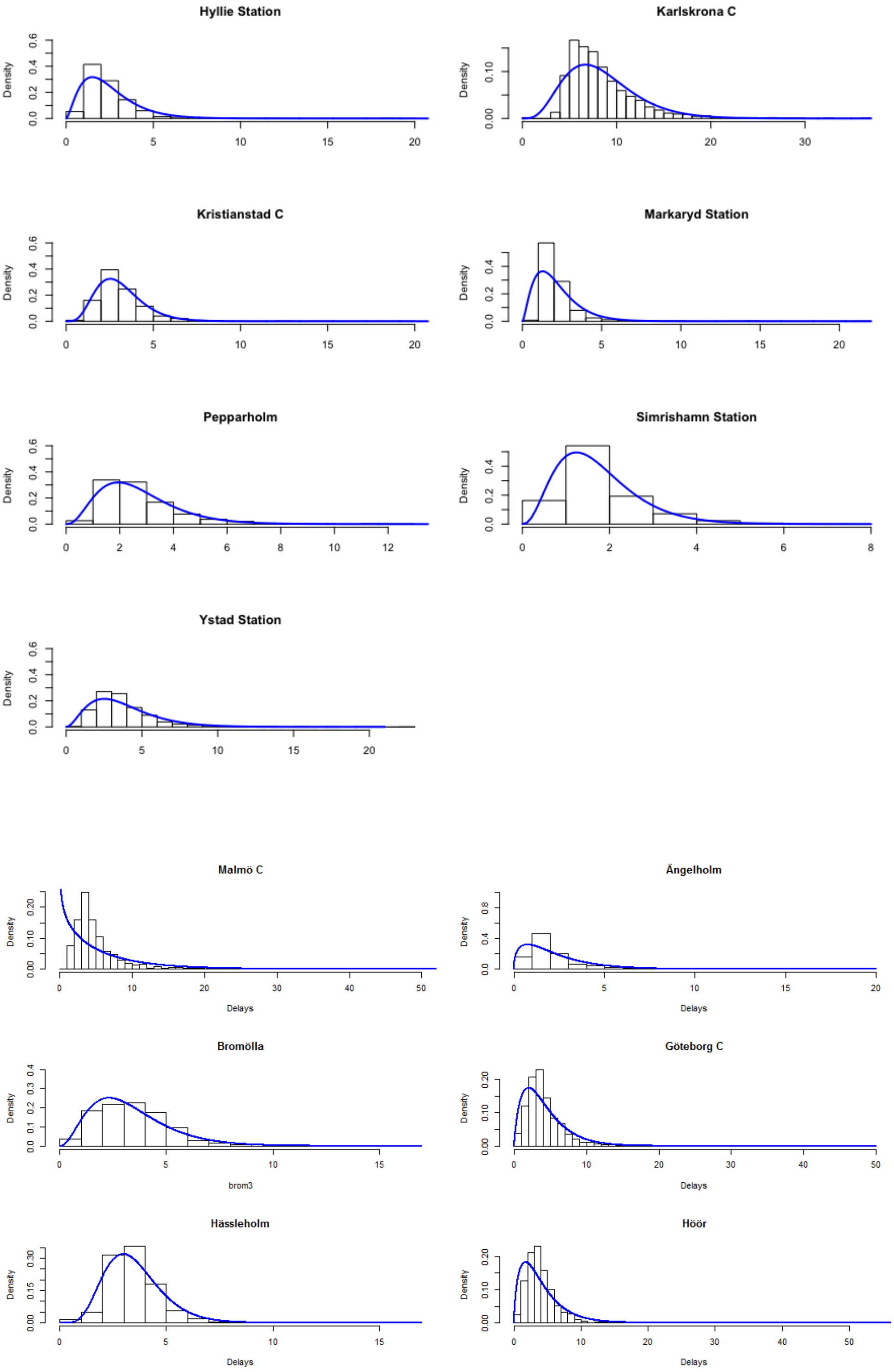


Figure 1: Histograms and gamma p.d.f.s

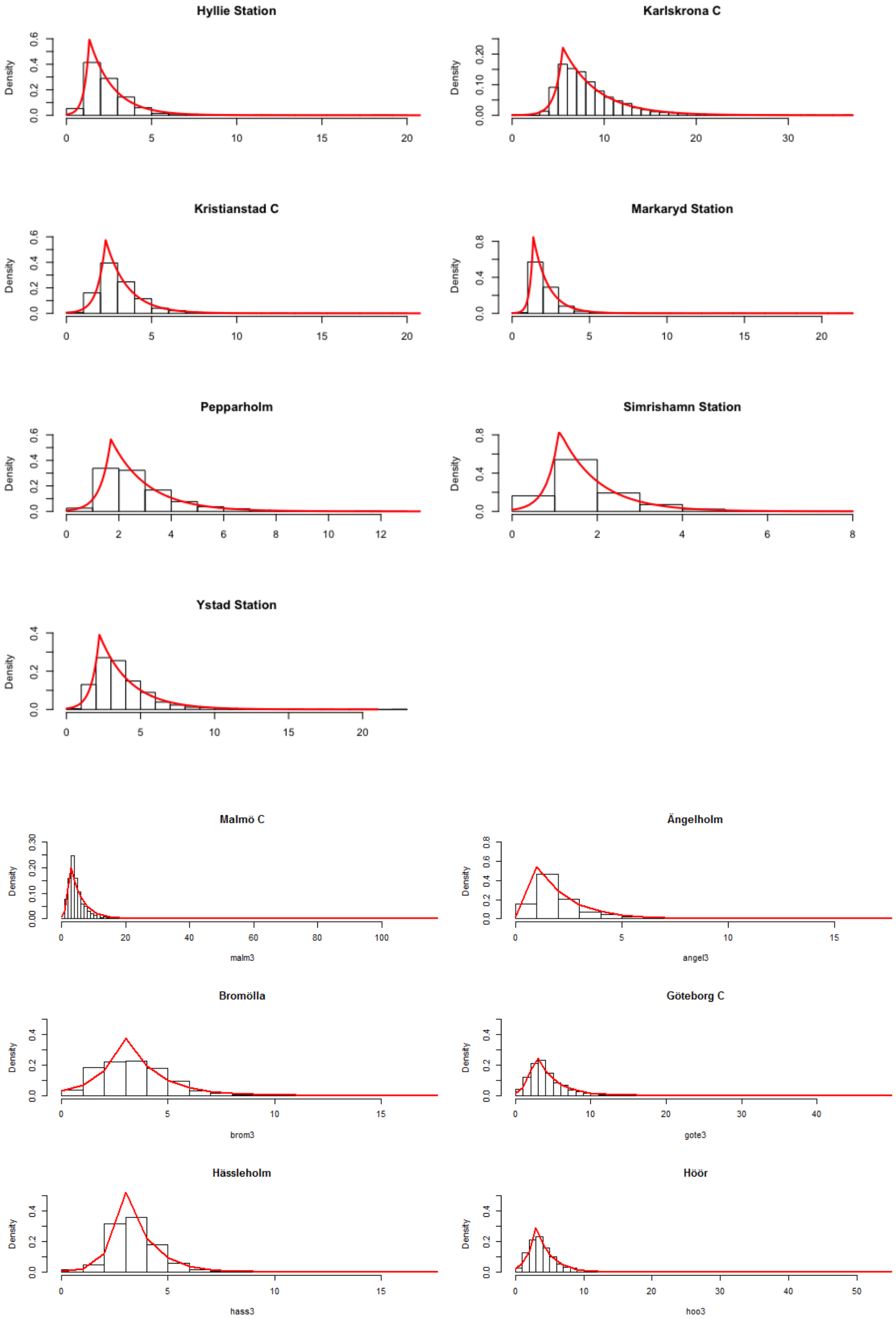
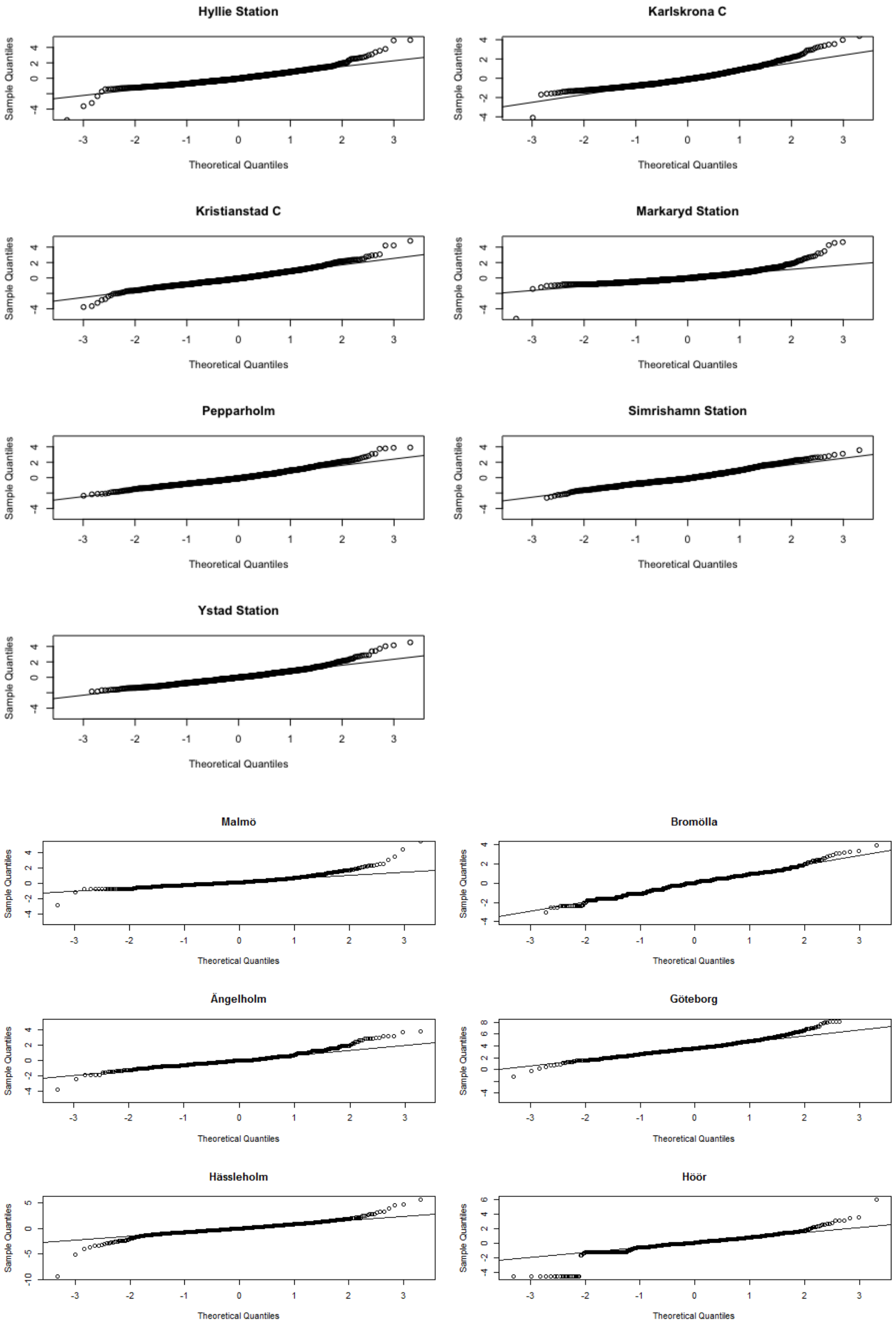


Figure 2: Histograms and asymmetric Laplace p.d.f.s

Appendix 4

Below, quantile plots showing whether the transformed data is normally distributed or not. Plots are given for data transformed using both gamma and asymmetric Laplace distributions.



Hyllie Station

Karlskrona C

Kristianstad C

Markaryd Station

Pepparholm

Simrishamn Station

Ystad Station

Malmö

Bromölla

Ängelholm

Göteborg

Hässleholm

Höör

Figure 1: Quantile plots showing normality of delays transformed through gamma distribution

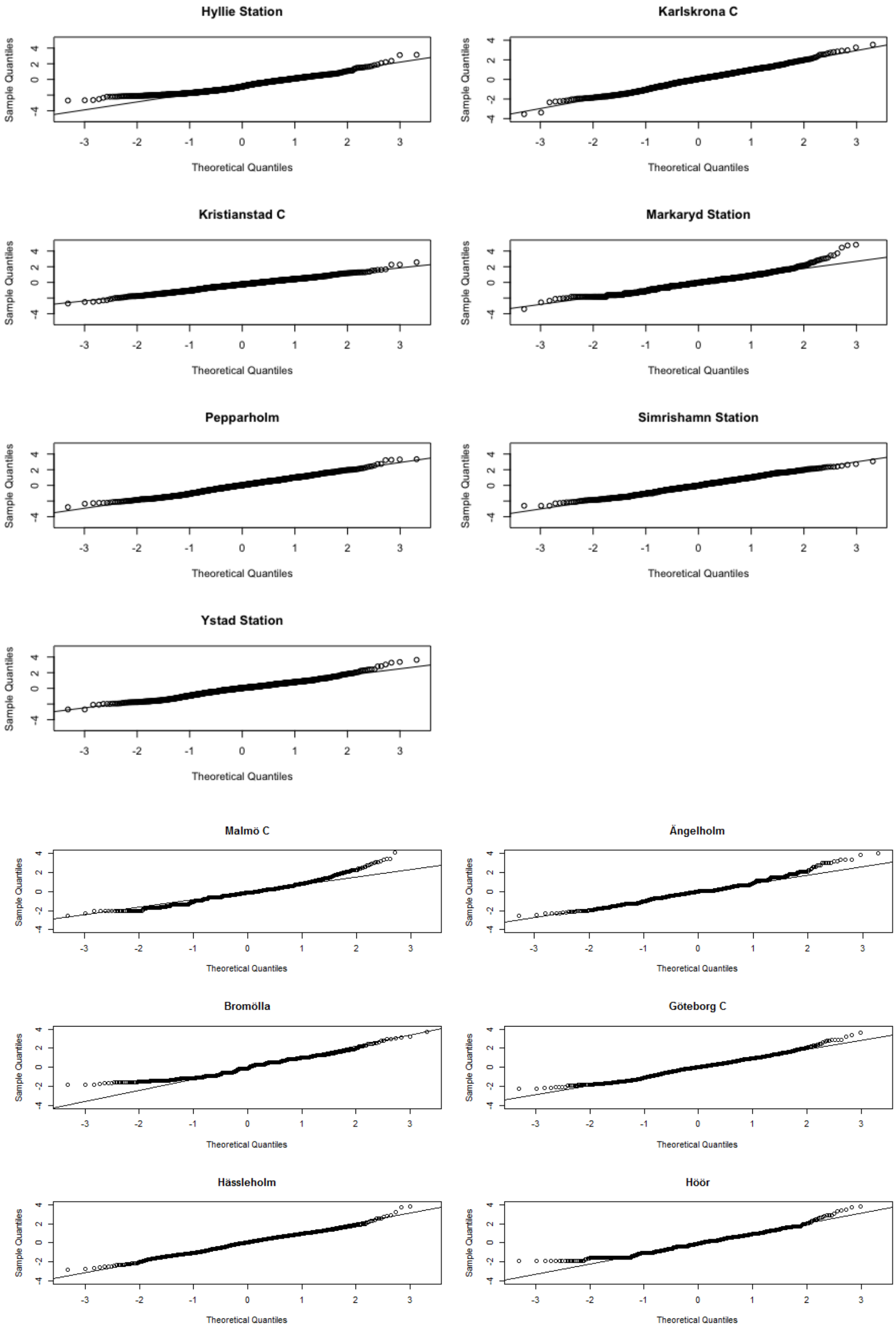


Figure 2: Quantile plots showing normality of delays transformed through the asymmetric Laplace distribution

Appendix 5

In this appendix, cross correlations for the data transformed using the asymmetric Laplace distribution is shown.

Table 3.2.1: Cross-correlation matrices of original and transformed data for station, Ystad, Lund, Kristianstad and Helsingborg at lags 0, 1, 2 and 3

	Transformed Data (Asymmetric Laplace)
$P(0)$	$\begin{pmatrix} 1 & 0.413 & 0.281 & 0.495 \\ 0.413 & 1 & 0.298 & 0.343 \\ 0.281 & 0.298 & 1 & 0.335 \\ 0.495 & 0.343 & 0.335 & 1 \end{pmatrix}$
$P(1)$	$\begin{pmatrix} 0.276 & 0.235 & 0.163 & 0.236 \\ 0.282 & 0.431 & 0.209 & 0.278 \\ 0.177 & 0.205 & 0.444 & 0.250 \\ 0.198 & 0.226 & 0.246 & 0.462 \end{pmatrix}$
$P(2)$	$\begin{pmatrix} 0.1842 & 0.112 & 0.140 & 0.0919 \\ 0.1630 & 0.236 & 0.166 & 0.1464 \\ 0.1464 & 0.193 & 0.337 & 0.2204 \\ 0.0846 & 0.106 & 0.227 & 0.2694 \end{pmatrix}$
$P(3)$	$\begin{pmatrix} 0.1341 & 0.0997 & 0.0824 & 0.101 \\ 0.0775 & 0.1456 & 0.1099 & 0.102 \\ 0.1091 & 0.1112 & 0.3016 & 0.212 \\ 0.0517 & 0.0789 & 0.1991 & 0.252 \end{pmatrix}$