Stockholms universitet

Count data time series models for telecommunications data

Bálint Fatér

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2015:31
Matematisk statistik
December 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2015:31**
http://www.math.su.se

Count data time series models for telecommunications data

Bálint Fatér*

December 2015

Abstract

The aim of the thesis is to try out the Poisson time series model for count data, in order to analyze telecommunications monitoring data from Ericsson AB. The amount of data a telecommunications network produces is untenable to check manually, and thus the detections of malfunctions can be both time consuming and extremely lengthy.

In this thesis we will explore the implementation of the above mentioned count data time series model on a subset of the data from a telecommunications network and the viability of such an approach, in particular, we will using identity link based generalized linear models to perform the analysis- We will also attempt to select the best model amongst several based on Akaike's information criteria. Finally we will cover how to implement said models in R and cover some of the more technical aspects of the Poisson time series model for count data.

It turns out that the model fits quite well on most of the studied time series, although, some of the data the telecommunications network produces have counts in order of magnitudes of millions, hence a better approach would probably have been to use an ordinary time series based framework using a Gaussian response model.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: balint.fater@gmail.com. Supervisor: Michael Höhle.

Acknowledgements

I would like to thank my supervisor, Michael Höhle, for the massive support and help I received with this project. His kind and some times not so kind feedback was a great help in improving the quality of this thesis, as well as, helping me staying on track. Paolo Elena and Jan Petterson both at Ericsson AB provided me with the data and readily answered any and all questions I might have had. Last but not least I would like to thank Sten Mogren for always making sure I had access to everything I needed whenever I was working at Ericsson. My most sincere thank you to you all.

Bálint Fatér
Stockholm, November 2015

Contents

1 Introduction

1.1 Aim of Thesis

In this thesis we will explore implementing the Poisson time series model for count data on time series count data provided by Ericsson. The data originated from one of their telecommunication networks, the location cannot be disclosed due to a non disclosure agreement governing the data, however the location is not important for the present analysis. In order to provide the necessary background knowledge, we will in this chapter, give a rudimentary overview of how a telecommunication network functions. In chapter two we will cover the data in detail while in chapter three go over the technical details of said Poisson time series model for count data, as well as some background calculations. I chapter four we will take look at how to implement said model into R as well as select the most suitable individual model from a set of possible candidates.

1.2 Overview of a telecommunications network

In a somewhat simplified model of a telecommunications network or radio access network (RAN) consist of one or more radio network controllers (RNC) that each handles one or more radio base stations (RBS), which in turn handles multiple cells which are geographical areas. For instance the old city center of Stockholm "Gamla stan", might in theory be handled by one RBS with four cells each spanning a quadrant as illustrated in Figure 1. In reality Gamla stan is divided into many more cells, one reason being the high roofs. There are also probably multiple networks covering the area, since I am assuming that some Swedish phone operators use separate networks.

1.2.1 RNC - Radio Network Controller

The RNC is the node that controls all RAN functions. There are two distinct roles for the RNC, to serve and to control. The Serving RNC has overall control of the phone that is connected to the RAN. It controls key and frequently used resources, such as different voice bands, internet etc. The Controlling RNC has the overall control of a particular set of cells, and their associated base stations. When a phone must use resources in a cell not controlled by its Serving RNC, the Serving RNC must ask the Controlling RNC for those resources. An RNC is in contact with both other RNCs and with all RBS under its control.

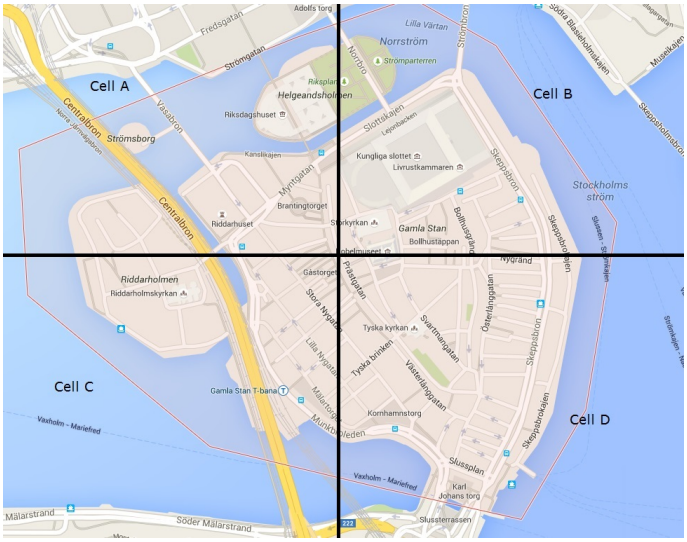

Figure 1: Google maps picture of Gamla Stan (Stockholm, Sweden) divided into 4 cells

1.2.2 RBS - Radio Base Station

The RBS the radio transmission and reception to/from the phone. It is controlled from the Radio Network Controller and its main purpose is just to shuffle signals and data between the phone and the RNC. One Radio Base Station can handle multiple cells. Figure 2 below illustrates the relation between the core network (RAN), RBS and RNCs.

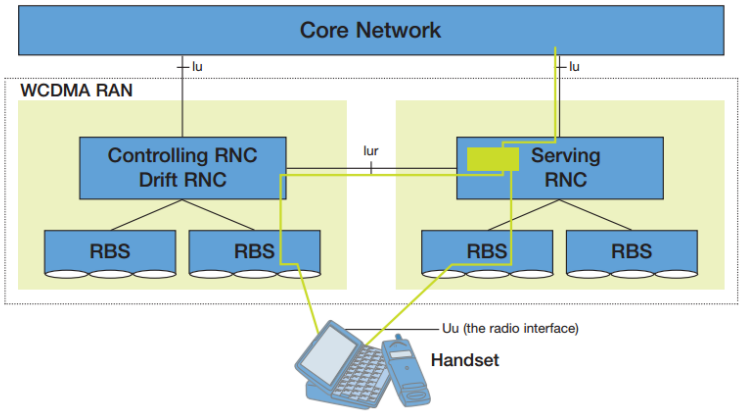

Figure 2: Figure from Ericsson Radio Systems AB (2001) illustrating basic setup of the RAN. Iu, Uu and Iur are just different protocols for communication.

2 Data Description

The data we received from Ericsson is 3 months of counter data from a full sized RAN, its location cannot be disclosed due to a confidentiality agreement with Ericsson. There are many events in such a network that needs to be tracked for maintenance and anomaly detection purposes. The way it is done here, there are counters for a set of predetermined events that count the amount of occurrences within a limited time frame. For instance, if someone were to call from this network first the counter for access would count up one, then the counter for attempts would count up one and so forth. The same way if your call would be suddenly disconnected then the counter for unexpected termination as well as the counter for drops from the network would increase by one.

Our data is divided into 15 minute intervals which gives us 96 data points for each counter each day. In total we have access to roughly 3 months worth of data, starting from 3rd of March 2015 up to 1st of June 2015, in total 8682 data points. In the present thesis we are going to focus on 4 counters, these have been chosen based on discussions with Ericsson testers for being key indicators for the health of the network as well as being within the suitable range that count data time series are useful for.

2.1 Counter 1 - A counter for number of load sharing diversions caused by high load when connecting to the network

This counter is used to count the number of times the RNC has to ask neighboring RNCs to take over some tasks due to high load. Lets take a look at the first three weeks (more would not make the resulting graph very clear) of data and see if we can discern any patterns (figure 3).

We note that there seem to be a strong daily and weekly cycle in data, weekends seem to produce less load then weekdays (remember data starts on a Tuesday). Lets take a look at some box plots to see if our suspicion is correct (Figure 4):

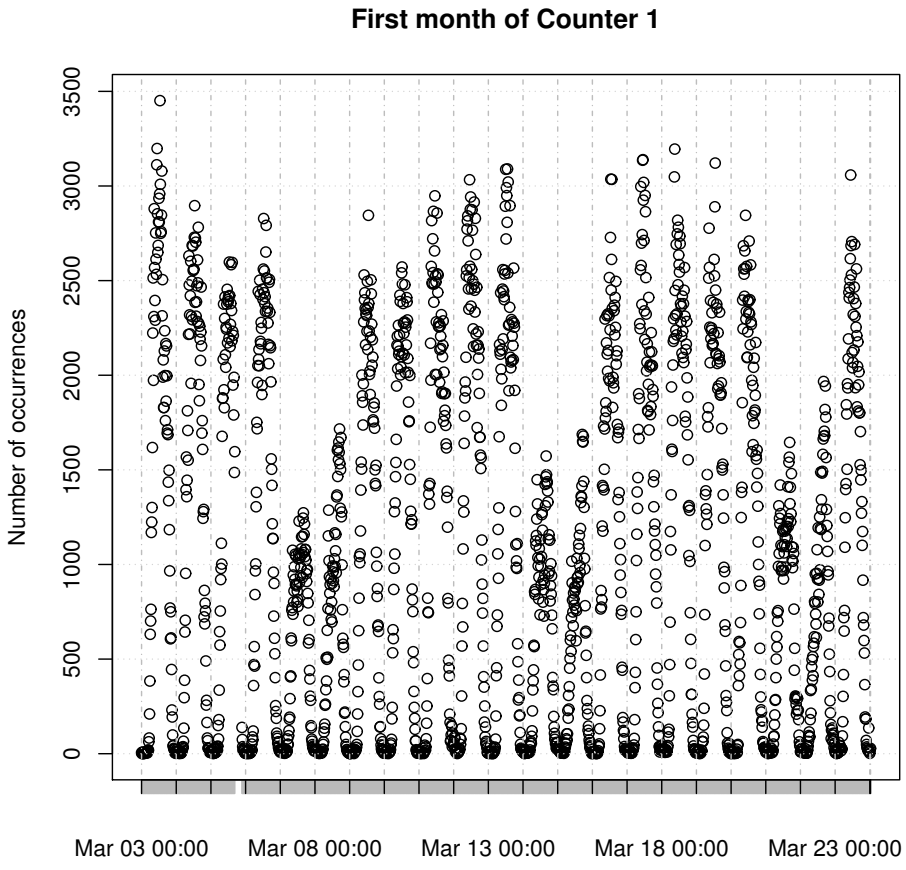

Figure 3: First 3 weeks of Counter 1

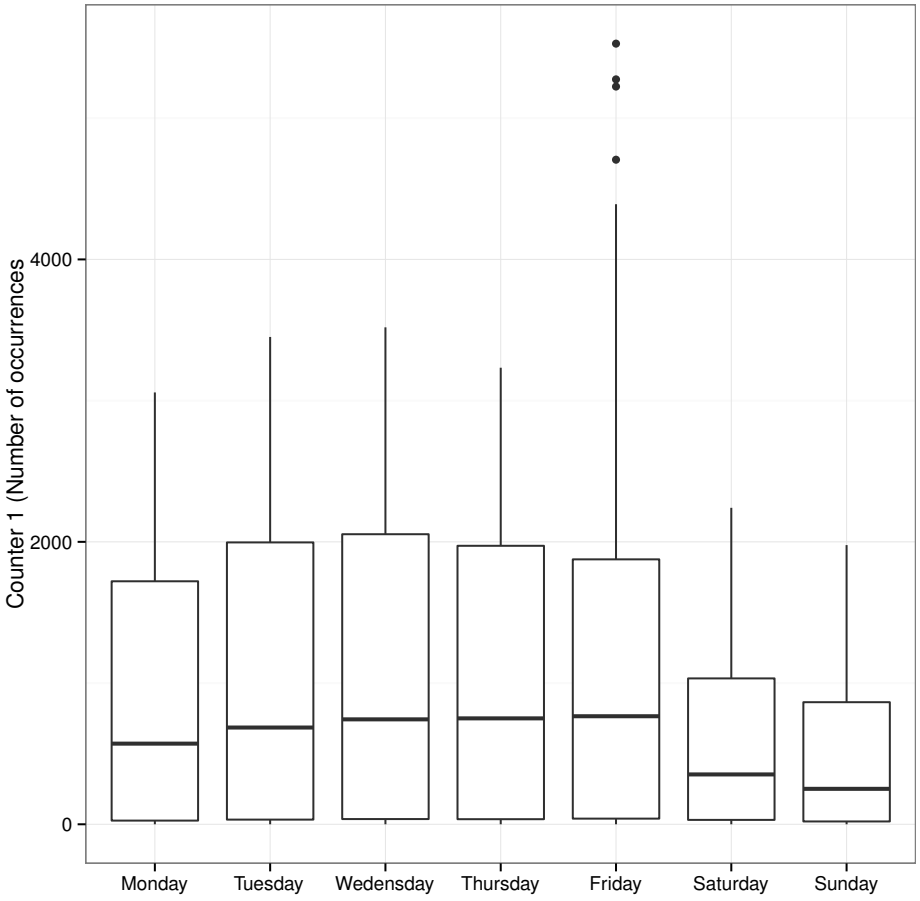

Figure 4: Box plots of daily distribution of Counter 1

The slightly lower upper hinges on Friday and Monday can be explained by that during the duration of the dataset there are two national holidays each that occur on Monday and Friday respectively,and no holidays for the rest of the weekdays. With this in mind let us take a look at box plots for the weekday (figure 5) respectively weekend pattern for said box plots.

Noticeable is the higher top on weekdays and that the top occurs on the afternoon rather then around noon on weekends.

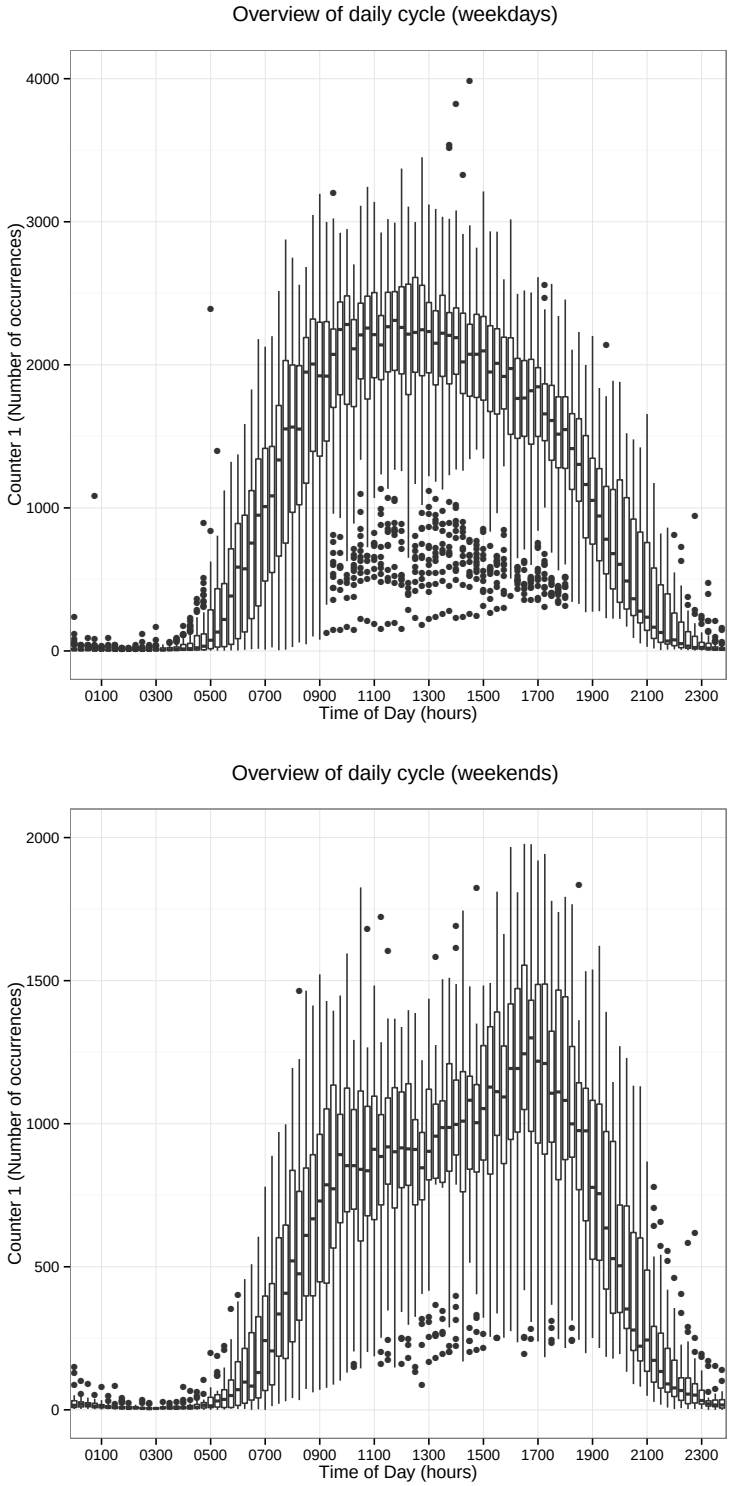

Figure 5: Box plot of daily cycle for weekday respectively weekends

2.2 Counter 2 - A counter for number of dropped calls

There are a multitude of reasons dropped calls can occur in the RAN. These include bad coverage and lack of resources during a handover between networks. Neither of which is desirable. Lets take a look at a box plot for the daily distribution of data for counter 2 (figure 6):

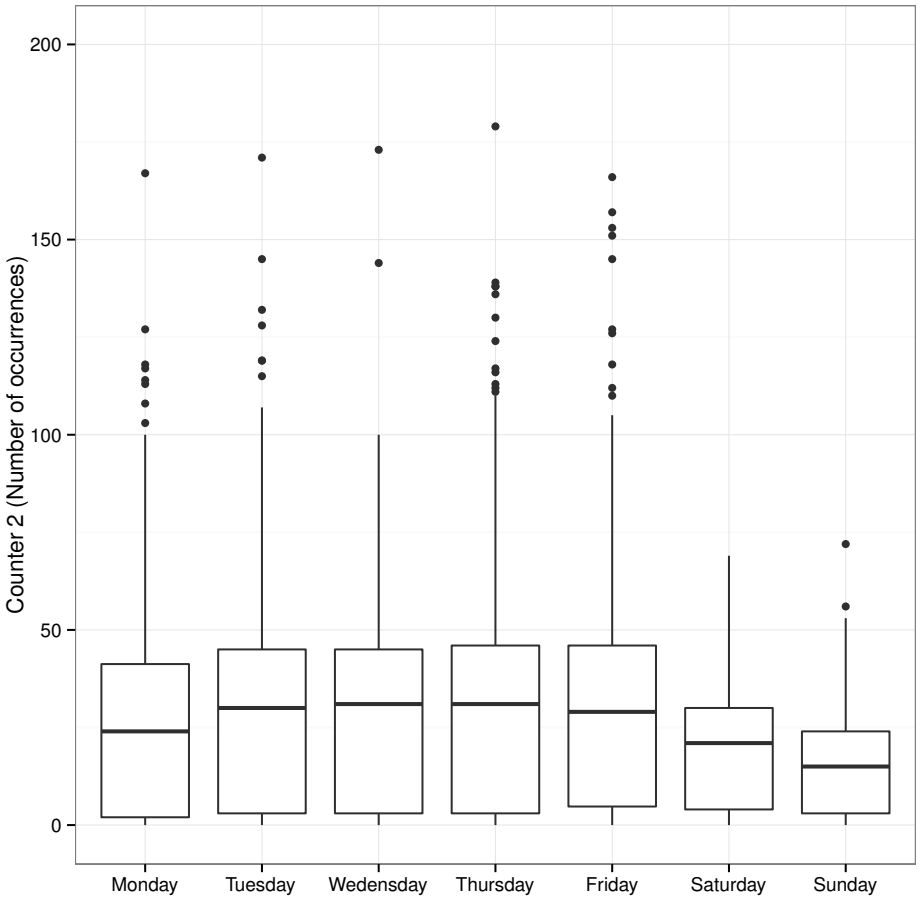

Figure 6: Box plots of daily distribution of Counter 2

Similar to counter one we see that weekends have somewhat lower load then weekdays. Lets explore this in the plots below (figure 7):

Just as counter two we see very similar behavior between weekends and weekdays with the distinction being weekdays having consistently higher load then weekdends.

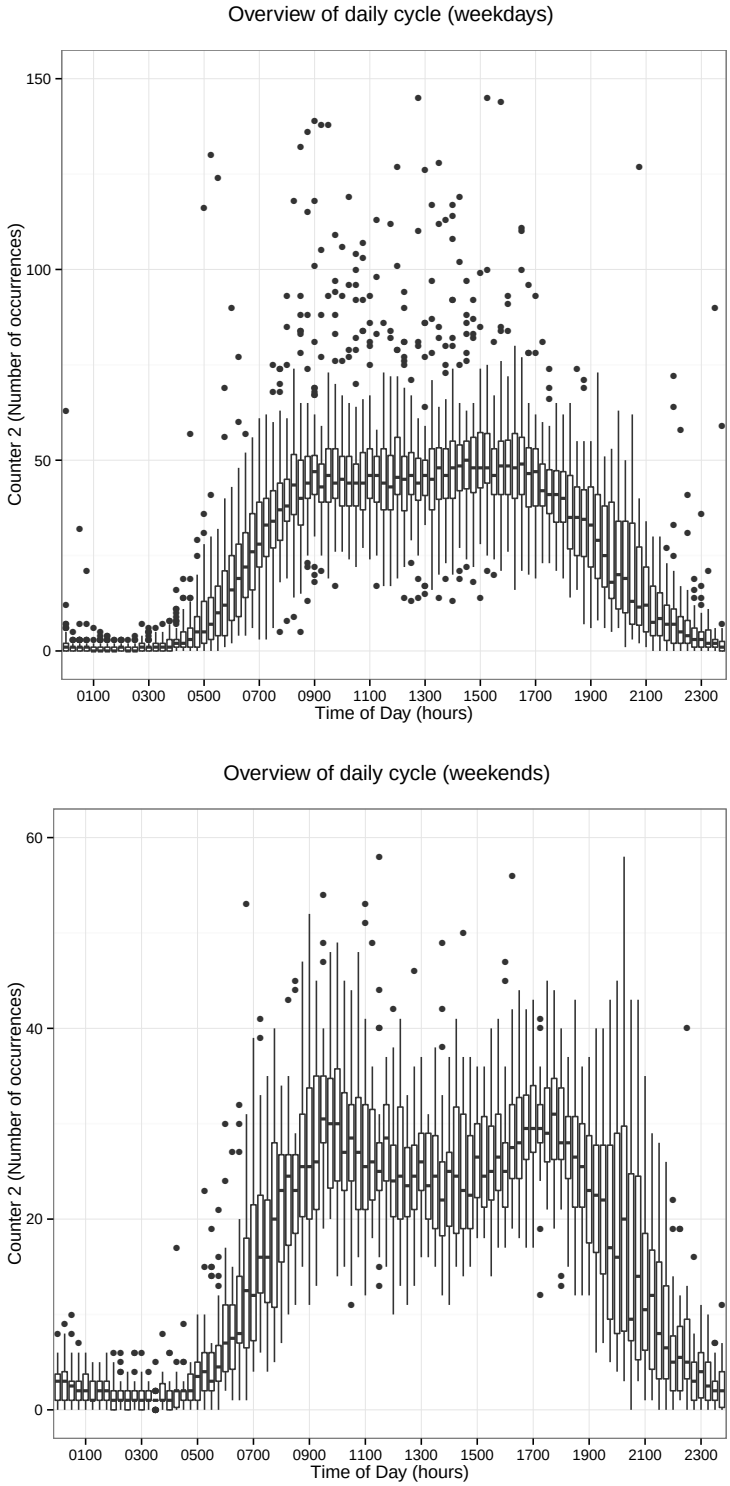

Figure 7: Box plot of daily cycle for weekday respectively weekends

2.3 Counter 3 - A counter for number of occurances when lacking rescources due to high load on the network

A counter for number of occurrences when the network is overloaded to the degree that all the resources cant be distributed to all request due to its load. The box plot for weekly pattern below (figure 8) looks very similar to the previous counters

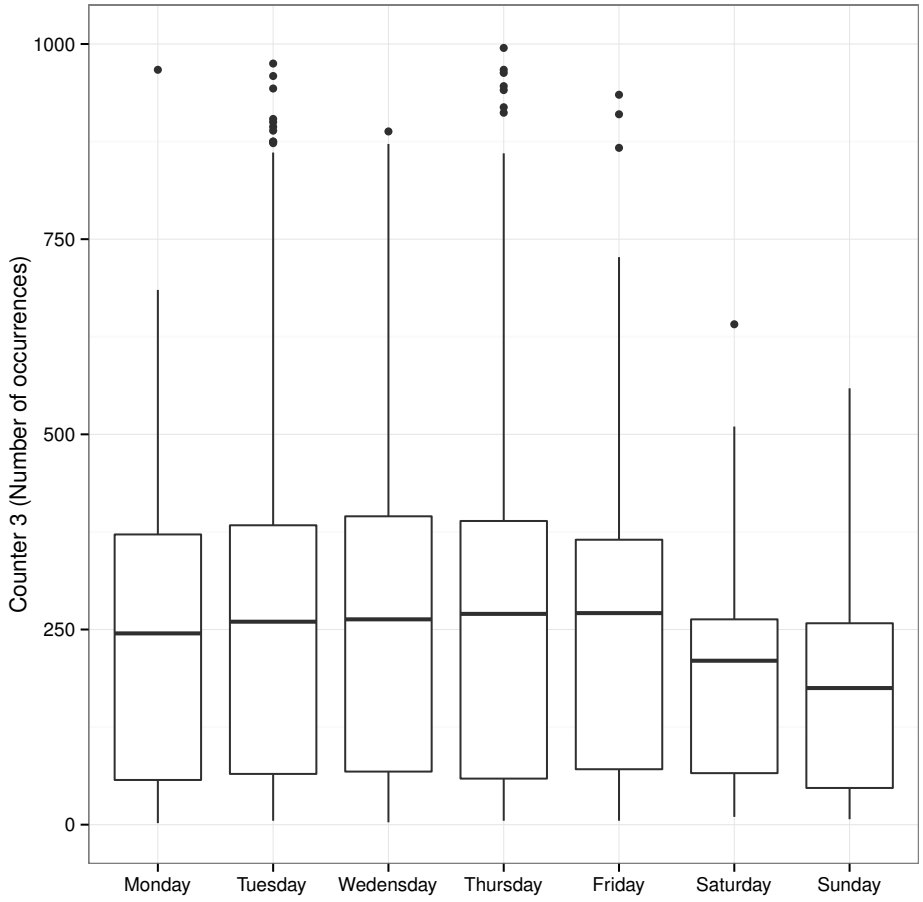

Figure 8: Box plots of daily distribution of Counter 3

Just as the previous counters weekdays seem to have a higher load then weekends. Lets look at the daily patterns (figure 9):

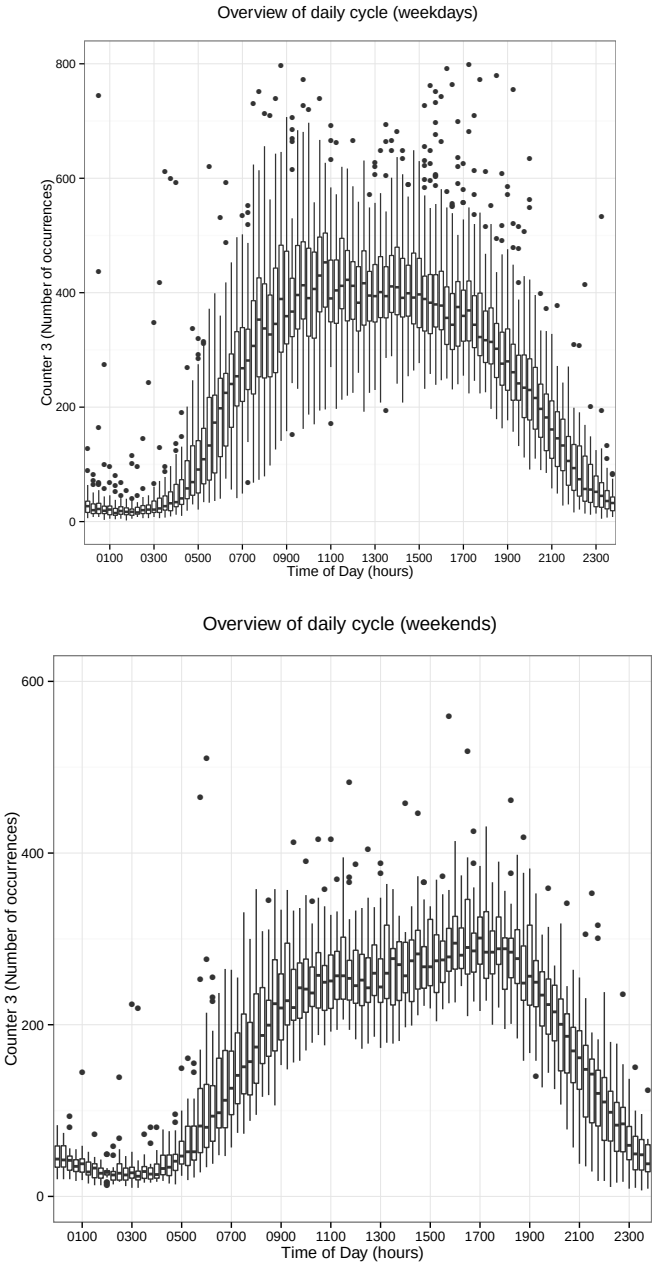

Figure 9: Box plot of daily cycle for weekday respectively weekends

Again very similar to the previous 2 counters weekends have a consistently lower load then weekdays.

2.4 Counter 4 - A counter for the sum of all CS64 rescources added to an existing call

Counter 4 is a counter that counts the amount of occurrences of CS64 resource being added to an existing call. Resources are usually added to a call if during the call one starts video chat, starts to surf etc, CS64 is one such resource. Here we see less of a weekly pattern with Mondays having similar density to Saturdays (figure 10):

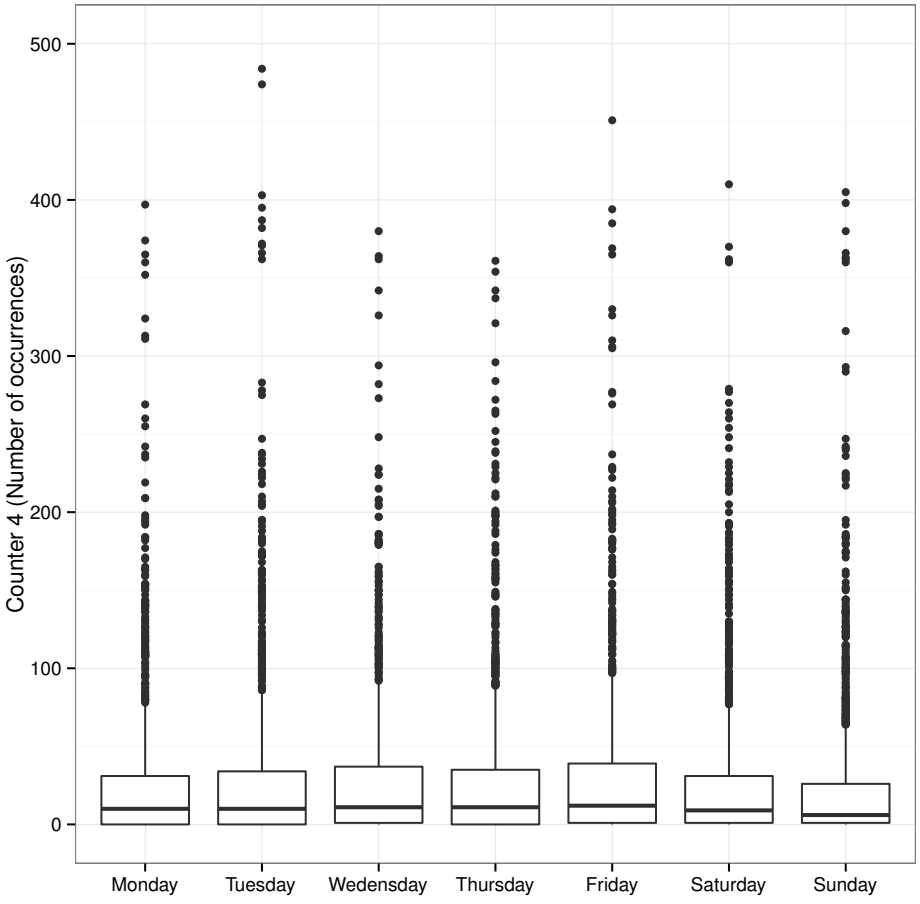

Figure 10: Boxplots of daily distribution of Counter 4

With this in mind lets explore the daily patterns for weekdays and weekends and see if they differ from the previous counters (figure 11):

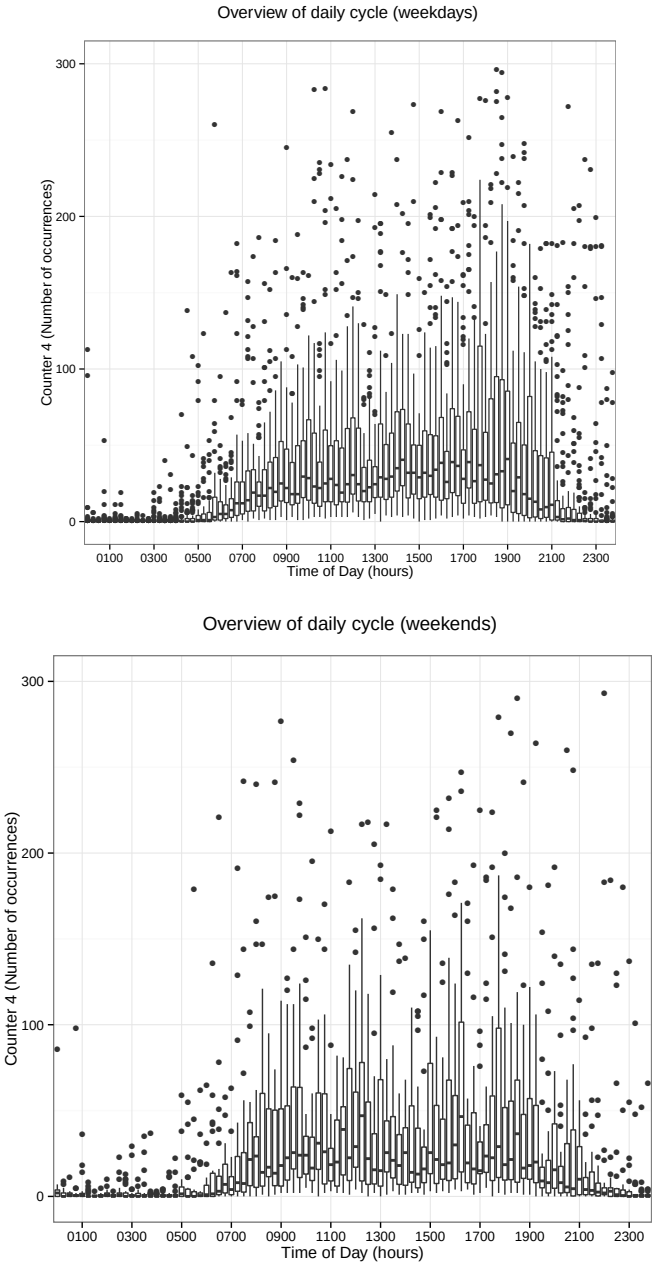

Figure 11: Boxplot of daily cycle for weekday resepectivly weekends

Here we also see a much more erratic behavior over the course of the day, with weekdays being slightly higher in load compared to weekends.

2.5 Summary

One thing that is clear from the previous descriptive analysis is that all four of our studied counters have a very strong daily and weekly cycles. The daily cycle is 24 hours and hence 96 observations long. Its noteworthy that activity during nightly hours are much lower across all 4 counters. The weekly cycle is 674 data points long and here we note that weekends tend to have a lower "top" then weekdays. In addition we can note that national holidays very similar to weekends. Counter 4 is somewhat a special case, while the patterns from the first three counters are present, the differences are much more minute, with the daily pattern being much less clear as well.

Now that we are more familiar with the behavior of the data lets move on to the theory behind the model we are going to implement.

3 Short introduction to generalized linear models

In this section we will take a look at the generalized linear model, more precisely the Poisson response model for time series that we will be using in this thesis. We will also look at parameter estimation and model selection for this type of models.

Consider a nonnegative integer valued time series $\{Y_t\}$ called the *response* time series of said process. Furthermore let $\mathcal{F}_{t-1}$ denote all available information on the process up to time t, this could also include covariates that are known at time t, in our case, current date and time are two examples. According to Kedem and Fokianos (2002, p. 140) a most natural candidate distribution for the response process is the Poisson. Then the conditional distribution of $\{Y_t\}$ is specified by assuming that the conditional density of the response given the past is Poisson with mean μ_t.

$$f(y_t; \mu_t|\mathcal{F}_{t-1}) = \frac{\exp(-\mu_t)\mu_t^{y_t}}{y_t!}, \quad t = 1, ..., N. \tag{1}$$

Furthermore we will define $\mathbf{Z}_{t-1}, t = 1, ..., N$ as a p- dimensional *covariate vector* that includes past values of the response up to time $t-1$ and possibly any other supplementary information available up to time t.

In order to model the process $\{Y_t\}$ we will use a generalized linear model, which is an extension of the general linear model:

$$y_t = \mathbf{x}_t\boldsymbol{\beta} + \epsilon_t. \tag{2}$$

The differences being the relaxation of the assumption that y is independently normally distributed with constant variance, instead the generalized linear model permits any distribution of the response that belongs to the exponential family, in our case the Poisson, and instead of modeling $\mu_t = E[y_t|x_t]$ directly as a function of the *linear predictor* $\eta_t = \mathbf{x}_t\boldsymbol{\beta}$, we model some function $g(\mu_t)$ of μ_t. Thus, the generalized linear model takes the form

$$g(\mu_t) = \eta_t = \mathbf{x}_t\boldsymbol{\beta}, \tag{3}$$

with $g(\cdot)$ being called the *link function*. With the above in mind our model then becomes, using the *inverse link function* $h(\cdot) = g^{-1}(\cdot)$, above mentioned covariate vector $\mathbf{Z}_{t-1}$ and the p-dimensional parameter vector for the model β:

$$\mu_t = h(\mathbf{Z}_{t-1}\boldsymbol{\beta}), \quad t = 1, ..., N. \tag{4}$$

We will however concern ourselves with a Poisson model with identity link, since it fits the aggregated nature of our data very well and the fact that

the traditional link function being a log link have the unfortunate tendency that for unbounded covariates the proccess tends to grow at an exponatial rate (for more details see Kedem and Fokianos (2002, p. 143)). In this case $g(\mu_t) = \mu_t$ and as a consequence $h(\mathbf{Z}_{t-1}\boldsymbol{\beta}) = \mathbf{Z}_{t-1}\boldsymbol{\beta}$.

3.1 Parameter estimation

Now that we have formulated our time series Poisson model, we need to estimate the parameter vector β. The likelihood function is not always easily obtainable, we will therefore use the partial likelihood function to estimate the $\boldsymbol{\beta}$ vector. Kedem and Fokianos (2002, p. 2) defines partial likelihood as follows:

Definition 1 *Denote the density of Y_t, given the information $\mathcal{F}_{t-1}$, by $f_t(y_t;\theta)$ where $\theta \in \mathbb{R}^p$ is a fixed parameter vector. The partial likelihood (PL) function relative to $\theta, \mathcal{F}_t$, and the data $Y_1, Y_2, ..., Y_N$, is given by the product*

$$\mathrm{PL}(\theta; y_1, y_2, ..., y_N) = \prod_{t=1}^{N} f_t(y_t;\theta). \tag{5}$$

Conditional likelihood takes only into account what is known to the observer up to time t. The partial likelihood function of β for the Poisson model is given by

$$\mathrm{PL}(\boldsymbol{\beta}) = \prod_{t=1}^{N} f(y_t;\boldsymbol{\beta}|\mathcal{F}_{t-1}) = \prod_{t=1}^{N} \frac{\exp(-\mu_t(\beta))\mu_t(\beta)^{y_t}}{y_t!} \tag{6}$$

Then the partial log -likelihood is (using equation (4)):

$$\ell(\beta) = \log \mathrm{PL}(\boldsymbol{\beta}) = \sum_{t=1}^{N} y_t \log(\eta_t) - \sum_{t=1}^{N} \eta_t - \sum_{t=1}^{N} \log(y_t!) \tag{7}$$

By differentiation, we obtain the partial score function

$$\mathbf{S}_N(\boldsymbol{\beta}) = \nabla \ell(\boldsymbol{\beta}) = \left(\frac{\partial \ell(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \ell(\beta)}{\partial \beta_p} \right) \tag{8}$$

Calculation of which is done by using the chain rule (analogous to Kedem and Fokianos (2002, p.11)):

$$\frac{\partial \ell_t}{\partial \beta_j} = \frac{\partial \ell_t}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \beta_j}, \quad j = 1, \ldots, p$$

Where

$$\frac{\partial \ell_t}{\partial \mu_t} = \frac{y_t}{\mu_t} - 1$$

Furthermore since $\eta_t = \mu_t$ when using identity link and $\eta_t = \sum_{j=1}^{p} z_{(t-1)j}\beta_j$ we get $\frac{\partial \eta_t}{\partial \beta_t} = z_{(t-1)j}$. Combining all parts of the chain rule and inserting them into (8) we obtain the *partial score function* as a p-dimensional vector:

$$\mathbf{S}_N(\beta) = \sum_{t=1}^{N} (\mathbf{Z}_{t-1} \frac{y_t}{\mu_t(\beta)} - 1) \tag{9}$$

The solution $\hat{\beta}$ such that $S_N(\hat{\beta}) = 0$ constitutes the partial maximum likelihood estimator, i.e.

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}(\mathrm{PL}(\boldsymbol{\beta}))$$

This, however, is a system of nonlinear equations and is usually solved by the Fischer scoring method. For more information on how $\hat{\beta}$ is obtained see Kedem and Fokianos (2002, p. 12)

3.2 Model Validation

In order to evaluate how well the model fit data we will be looking at Pearson Residuals which is defined in Olsson (2002, p.56) as follows:

$$r_t^p = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{\mu}_t}} \tag{10}$$

In chapter 4 we will use these residuals to gauge model performance and to identify obvious outliers.

3.3 Model Selection

In order to select the most suitable model we will primarily be looking at Akaike's Information Criteron (AIC) and the Bayesian Information Criterion (BIC).

Accoridng to (Kedem and Fokianos 2002, p. 25), *Akaike's Information Criteron* is defined as a function of the number of independent model paramaters,

$$\mathrm{AIC}(p) = -2\log(\mathrm{PL}(\hat{\boldsymbol{\beta}})) + 2p \tag{11}$$

where $\hat{\beta}$ is the maximum partial likelihood estimator of β and $p = dim(\beta)$. This criterion evidently penalises models with many parameters.

The Bayesian Information Criterion is defined as follows, with N being the length of the time series

$$\mathrm{BIC}(p) = -2\log(\mathrm{PL}(\hat{\boldsymbol{\beta}})) + p\log(N) \tag{12}$$

Notable is that BIC penalises models with many parameters even further than AIC. We will choose the model that for a given p will minimise BIC and AIC, since BIC is harsher we will prefer BIC.

4 Implementation on telecom data

Now that we are familiar with both data and the intended model, lets take a look at the specific models we will be comparing.

Model 1 *The first model we will be looking at will be a very simplistic model, with the assumption that* $Y_t \overset{indep.}{\sim} \text{Po}$*, namely*

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} \tag{13}$$

Here we only incorporate information from the last observation.

Model 1 is a branching process with immigration (see Leonhard Held, Michael Höhle and Mathias Hofmann (2005, p.189) equation (1.2) for details) and as such is stationary when $0 \leq \beta_1 \leq 1$. Lets expand this model, and take the weekly cycle into account. This would give us the following model (the assumption that $Y_t \overset{indep.}{\sim} \text{Po}(\lambda_t)$ holds for this and all subsequent models)

Model 2 *In this second model we will add the following variable*

$$I_t = \begin{cases} 1, \textit{if day} \in \{\textit{Saturday, Sunday}\} \\ 0, \textit{otherwise} \end{cases}$$

thus our model thus becomes

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_t \tag{14}$$

Lets now take those national holidays in account:

Model 3 *In model 3 we will introduce the following variable*

$$H_t = \begin{cases} 1, \textit{if day} \in \{\textit{weekday is national holiday}\} \\ 0, \textit{otherwise} \end{cases}$$

thus our model thus becomes

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_t + \beta_3 H_t \tag{15}$$

Remember those daily cycles, we will attempt to incorporate them into the model using a Fourier series expansion (see Hansen (2002, p. 187)), which is given by the relation

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

with P being the period of the series. Here, β_0 in combination with the terms $\beta_4 \sin\left(\frac{2\pi t}{96}\right) + \beta_5 \cos\left(\frac{2\pi t}{96}\right)$ constitute the first three terms in such a series.

Model 4 *Here we will add a* $\cos(\cdot)$ *and* $\sin(\cdot)$ *function to take our daily cycle of 96 observations into account:*

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_t + \beta_3 H_t + \beta_4 \sin\left(\frac{2\pi t}{96}\right) + \beta_5 \cos\left(\frac{2\pi t}{96}\right) \tag{16}$$

In models 5 and 6 we will remove the indicator for weekend and holiday respectivly. While in model 7 we will remove both:

Model 5

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_t + \beta_4 \sin\left(\frac{2\pi t}{96}\right) + \beta_5 \cos\left(\frac{2\pi t}{96}\right) \tag{17}$$

Model 6

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_3 H_t + \beta_4 \sin\left(\frac{2\pi t}{96}\right) + \beta_5 \cos\left(\frac{2\pi t}{96}\right) \tag{18}$$

Model 7

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_4 \sin\left(\frac{2\pi t}{96}\right) + \beta_5 \cos\left(\frac{2\pi t}{96}\right) \tag{19}$$

The following R code is used to fit Model 4 for the Counter 4 Data:

```
> mdl4 <- glm(Counter4[2:numrow] ~ 1 + Counter4[1:trow] +
    nWeekend[2:numrow] + nHoliday[2:numrow] + sin(2/96 * pi *
    nTime[2:numrow]) + cos(2/96 * pi * nTime[2:numrow]), data =
    xData, family = poisson(link = "identity"), start = c
    (1,0,0,0,0,0))
> summary(mdl4)

Call:
glm(formula = Counter4[2:numrow] ~ 1 + Counter4[1:trow] +
    nWeekend[2:numrow] + nHoliday[2:numrow] + sin(2/96 * pi *
    nTime[2:numrow]) + cos(2/96 * pi * nTime[2:numrow]), family
        = poisson(link = "identity"),
    data = xData, start = c(1, 0, 0, 0, 0, 0))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-20.405    -5.017    -3.491     0.923    47.293

Coefficients:
                                      Estimate Std. Error z value
                                         Pr(>|z|)
(Intercept)                          12.719191   0.062515 203.457
    < 2e-16 ***
Counter4[1:trow]                     0.571631    0.001935 295.347  < 2e
    -16 ***
nWeekend[2:numrow]                   -0.863481   0.102203  -8.449
    < 2e-16 ***
nHoliday[2:numrow]                   -1.425307   0.219904  -6.481
    9.08e-11 ***
sin(2/96 * pi * nTime[2:numrow])     -0.837293   0.065788 -12.727
    < 2e-16 ***
cos(2/96 * pi * nTime[2:numrow])     -0.727017   0.066031 -11.010
    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 487540  on 8680  degrees of freedom
Residual deviance: 364620  on 8675  degrees of freedom
AIC: 395132

Number of Fisher Scoring iterations: 8
```

Start value of $(1,0,0,0,0,0)$ was provided out of necessity due to the default start value of $(0,0,0,0,0,0)$ would result in the log likelihood for the starting value becoming infinite due to having a poisson process with zero mean but non zero response.

The table below contains our selection criteria for each model (for counter 1):

model	LogLik	p	AIC	BIC
Model 1	-166152	2	332308	332323
Model 2	-166094	3	332194	332215
Model 3	-166088	4	332184	332212
Model 4	-166032	6	**332075**	**332118**
Model 5	-166040	5	332089	332124
Model 6	-166091	5	332193	332228
Model 7	-166094	4	332196	332225

One can see that the model for Counter 1 model 4 has the lowest AIC, however since the model that doesnt take national holidays into account is pretty close in AIC, one might consider to settle with model 5 and allow the modeling to be more robust in terms of which network its implemented on.

Lets take a look at counter 2

model	LogLik	p	AIC	BIC
Model 1	-45366	2	90736	90750
Model 2	-45354	3	90713	90734
Model 3	-45341	4	90690	90719
Model 4	-45324	6	**90660**	**90702**
Model 5	-45335	5	90679	90714
Model 6	-45338	5	90687	90722
Model 7	-45345	4	90699	90727

Just like Counter 1 we see that model 4 is still the most suitable model for Counter 2. One can attribute the overall higher AIC for this counter in part to systematic downshift last few weeks on Counter 2. Lets take a look at Counter 3:

model	LogLik	p	AIC	BIC
Model 1	-214857	2	429717	429731
Model 2	-214834	3	429673	429694
Model 3	-214801	4	429609	429637
Model 4	-214774	6	**429560**	**429603**
Model 5	-214807	5	429625	429660
Model 6	-214804	5	429618	429653
Model 7	-214829	4	429667	429695

Again we see that Model 4 is the superior model for Counter 3, just like it was for Counter 1 and Counter 2. Finally lets take a look at Counter 4:

model	LogLik	p	AIC	BIC
Model 1	-197722	2	395448	395462
Model 2	-197697	3	395401	395422
Model 3	-197685	4	395378	395406
Model 4	-197560	6	**395132**	**395174**
Model 5	-197577	5	395165	395200
Model 6	-197592	5	395194	395229
Model 7	-197603	4	395215	395243

Unsuprisingly Model 4 is once more proving to be the best model out of the 7 at our disposal.

4.1 Improving on model 4 for counter 2

Model 4 performed uniformly best for all four of our chosen counters, lets take a look at what happens if we expand the fourier series in model 4 for counter 2. To this end we will compare the following, more elaborate models:

Model 8

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_t + \beta_3 H_t + \beta_4 \sin\left(\frac{2\pi t}{96}\right) + \beta_5 \cos\left(\frac{2\pi t}{96}\right) + \beta_6 \sin\left(\frac{4\pi t}{96}\right) + \beta_7 \cos\left(\frac{4\pi t}{96}\right) \tag{20}$$

Model 9

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_t + \beta_3 H_t + \beta_4 \sin\left(\frac{2\pi t}{96}\right) + \beta_5 \cos\left(\frac{2\pi t}{96}\right) + \beta_6 \sin\left(\frac{4\pi t}{96}\right) + \beta_7 \cos\left(\frac{4\pi t}{96}\right) + \beta_8 \sin\left(\frac{6\pi t}{96}\right) + \beta_9 \cos\left(\frac{6\pi t}{96}\right) \tag{21}$$

Model 10

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_t + \beta_3 H_t + \beta_4 \sin\left(\frac{2\pi t}{96}\right) + \beta_5 \cos\left(\frac{2\pi t}{96}\right) + \beta_6 \sin\left(\frac{4\pi t}{96}\right) + \beta_7 \cos\left(\frac{4\pi t}{96}\right) + \beta_8 \sin\left(\frac{6\pi t}{96}\right) + \beta_9 \cos\left(\frac{6\pi t}{96}\right) + \beta_{10} \sin\left(\frac{6\pi t}{96}\right) + \beta_{11} \cos\left(\frac{6\pi t}{96}\right) \tag{22}$$

model	LogLik	p	AIC	BIC
Model 4	-45324	6	90660	90702
Model 8	-45313	8	90641	90699
Model 9	-45298	10	90616	**90687**
Model 10	-45293	12	**90610**	90695

We can see here that of the 4 modesl, model 9 has the lowest BIC, while model 10 has the lowest AIC, however since BIC is a harsher criteria for large datasets such as ours, model 9 is the model we will settle for. Taking a look at the the R - summary of said model:

```
> summary(mdl9)

Call:
glm(formula = Counter2[2:numrow] ~ 1 + Counter2[1:trow] +
    nWeekend[2:numrow] + nHoliday[2:numrow] + sin(2/96 * pi *
    nTime[2:numrow]) + cos(2/96 * pi * nTime[2:numrow]) + sin(4/
        96 *
    pi * nTime[2:numrow]) + cos(4/96 * pi * nTime[2:numrow]) +
    sin(6/96 * pi * nTime[2:numrow]) + cos(6/96 * pi * nTime[2:
        numrow]),
    family = poisson(link = "identity"), data = xData, start = c
        (1,
```

```
         0, 0, 0, 0, 0, 0, 0, 0, 0))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-32.141   -1.405   -0.303    0.855   70.143

Coefficients:
                                    Estimate Std. Error z value
                                        Pr(>|z|)
(Intercept)                         1.327473   0.034532  38.442
    < 2e-16 ***
Counter2[1:trow]                    0.951151   0.002338 406.826  < 2e-16
    ***
nWeekend[2:numrow]                 -0.311960   0.056404  -5.531
    3.19e-08 ***
nHoliday[2:numrow]                 -0.482402   0.104231  -4.628
    3.69e-06 ***
sin(2/96 * pi * nTime[2:numrow])   -0.006681   0.036467  -0.183
       0.855
cos(2/96 * pi * nTime[2:numrow])   -0.218734   0.036448  -6.001
    1.96e-09 ***
sin(4/96 * pi * nTime[2:numrow])   -0.174287   0.036328  -4.798
    1.61e-06 ***
cos(4/96 * pi * nTime[2:numrow])   -0.035751   0.036507  -0.979
       0.327
sin(6/96 * pi * nTime[2:numrow])   -0.032317   0.036370  -0.889
       0.374
cos(6/96 * pi * nTime[2:numrow])    0.197641   0.036351   5.437
    5.42e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 196825  on 8680  degrees of freedom
Residual deviance:  54067  on 8671  degrees of freedom
AIC: 90616

Number of Fisher Scoring iterations: 8
```

We observe that both weekends (Estimate: -0.311960, p-value: 3.19e-08) and national holidays have (Estimate: -0.482402, 3.69e-06) have a negative effect on mu_t this is well in line with what we observed in chapter 2. Conversely the coefficients for some of the Fourier series part of the model, namely $\sin\left(\frac{2\pi t}{96}\right)$, $\cos\left(\frac{4\pi t}{96}\right)$ and $\sin\left(\frac{6\pi t}{96}\right)$, have non-significant p-values, for the test that the corresponding $\beta_j \neq 0$, (0.855, 0.327, 0.374) suggesting that changes in their value is not associated with a corresponding change in the response mu_t. We will therefore omit these from the final model:

Model 11

$$\mu_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 I_t + \beta_3 H_t + \beta_5 \cos\left(\frac{2\pi t}{96}\right) + \beta_6 \sin\left(\frac{4\pi t}{96}\right) + \beta_9 \cos\left(\frac{6\pi t}{96}\right) \tag{23}$$

Implemented in R it becomes:

```
> summary(mdl11)

Call:
glm(formula = Counter2[2:numrow] ~ 1 + Counter2[1:trow] +
    nWeekend[2:numrow] + nHoliday[2:numrow] + cos(2/96 * pi *
    nTime[2:numrow]) + sin(4/96 * pi * nTime[2:numrow]) + cos(6/
        96 *
    pi * nTime[2:numrow]), family = poisson(link = "identity"),
    data = xData, start = c(1, 0, 0, 0, 0, 0, 0))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-32.141    -1.404    -0.297     0.855    70.157

Coefficients:
                                      Estimate Std. Error z value
                                          Pr(>|z|)
(Intercept)                           1.327506   0.034527  38.448
    < 2e-16 ***
Counter2[1:trow]                   0.951151   0.002338 406.840  < 2e-16
    ***
nWeekend[2:numrow]                   -0.311726   0.056424  -5.525
    3.30e-08 ***
nHoliday[2:numrow]                   -0.484994   0.104404  -4.645
    3.39e-06 ***
cos(2/96 * pi * nTime[2:numrow])     -0.221211   0.036545  -6.053
    1.42e-09 ***
sin(4/96 * pi * nTime[2:numrow])     -0.177735   0.036079  -4.926
    8.38e-07 ***
cos(6/96 * pi * nTime[2:numrow])      0.195996   0.036255   5.406
    6.44e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 196825  on 8680  degrees of freedom
Residual deviance:  54068  on 8674  degrees of freedom
AIC: 90612

Number of Fisher Scoring iterations: 8

> AIC(mdl11)
```

```
[1] 90611.71
> BIC(mdl11)
[1] 90661.2
> logLik(mdl11)
'log Lik.' -45298.86 (df=7)
```

Noticeable is that model 11 has both a lower AIC (90611.71) and BIC (90661.2) as well as all coefficients being significant with a p value lower then 0.001. This will be the model we settle for in this thesis, as a final step lets take a look at how model 11 are able to predict μ_t for counter 2:

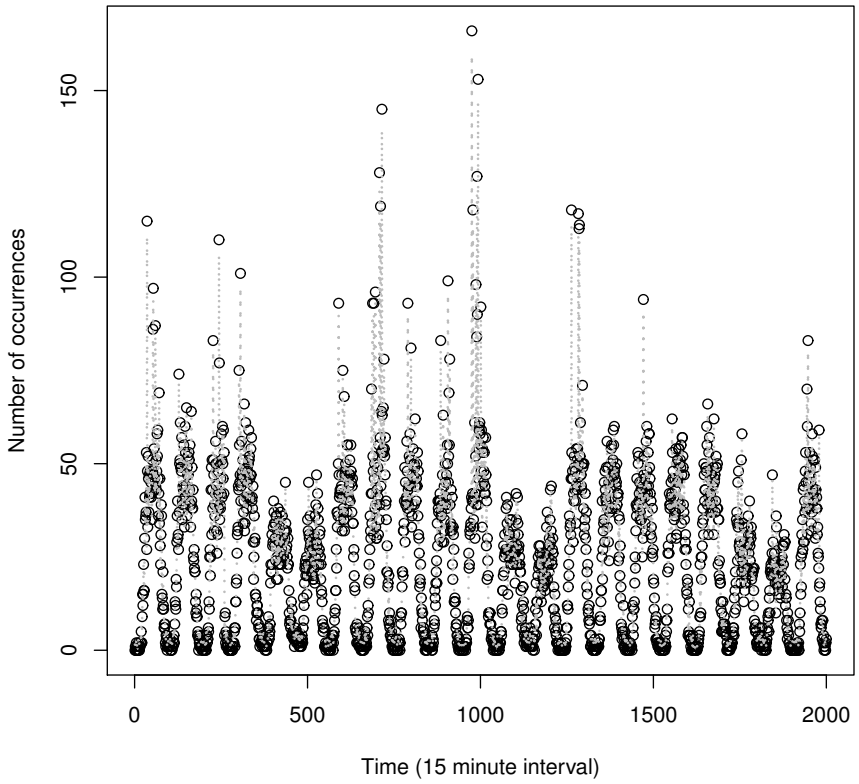

Figure 12: Dotted gray is predicted value, black circle is measured value at time t

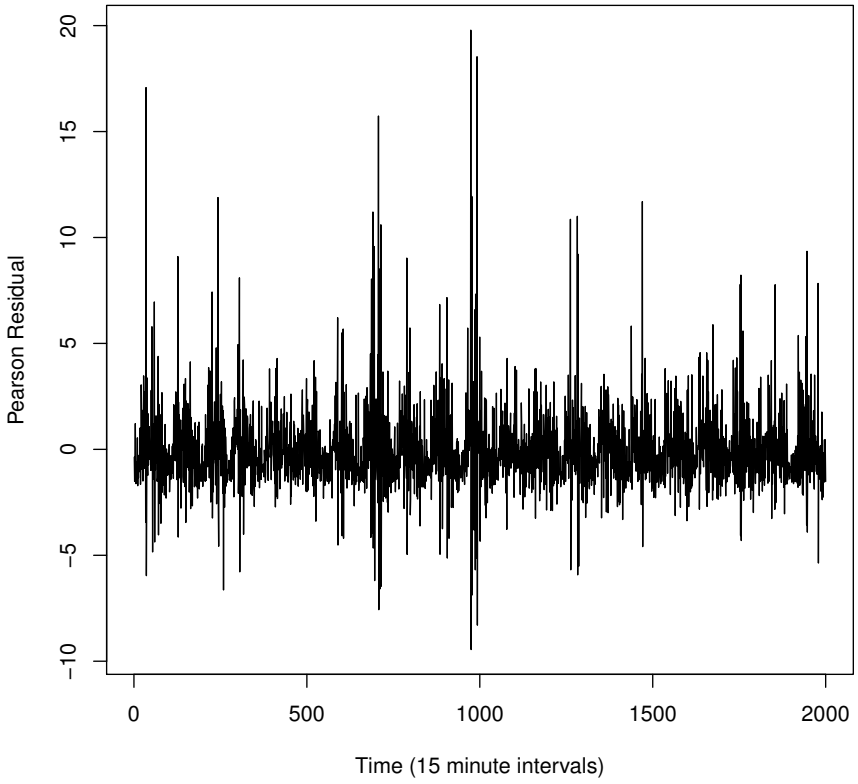

Figure 13: Pearson Residuals

As seen in figure 13, the our model does a really good job predicting the behavior of the counter in question. With the residuals only jumping at values that are outside of the main pattern and are usually classified as anomalies at Ericsson. Looking at a 95% (Figure 14) prediction interval we see that the model does a fine job of detecting anomalies. Noteworthy however is that the autoregression results in a "normal" observation following an anomaly falling outside the prediction interval as well.

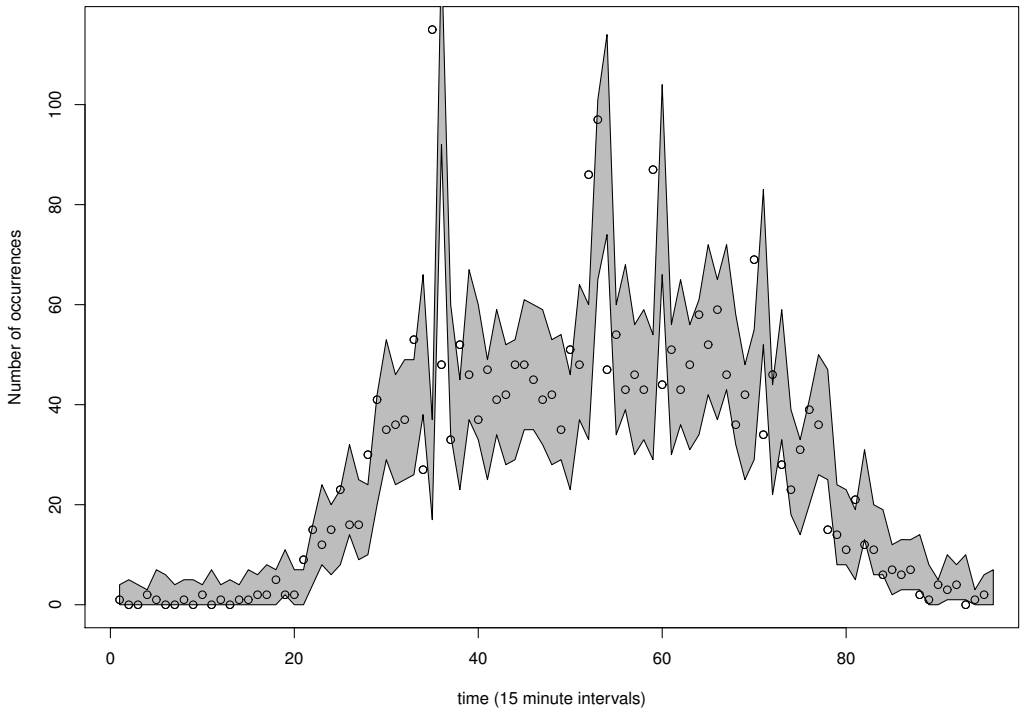

Figure 14: Gray area is the bounds for a 95% prediction interval for the first day of Counter 2, circles are actual observations of counter 2

5 Conclusion

We started out with data for four counters. After a short introduction in chapter 1 to how the network functions we then delved into data in detail trying to display both the uniqueness and similarities of each counter vis-á-vis the others.

In chapter 2 we then covered generalized linear models and in particular the Poisson model for count data that we intended to use for modeling. We also covered Parameter estimation and criteria we then went on to use to select the best possible model. We then choose between 7 models, finding the most complete one (model 4) to be the best. Then we expanded model 4 into 3 more models attempting to model the daily cycle better by adding more sin/cos terms. Choosing model 9 we then noticed that a few coefficients where non significant we omitted these eventually settling on model 11. We then took a look at how well model 11 would predict the first 3 weeks worth of data for counter 2. Judging from figure 13, the models ability of predicting μ_t was excellent.

The problem with the Poisson model for count data, is overdispersion, meaning that the variability in data will be greater then expected given our model. One way to deal with this is to use an alternative, more complicated, model like the Negative-Binomial model or the Zeger-Qaqish model both of which are discussed in Kedem and Fokianos (2002) and implemented in the R package tscount.

Possible future expansion of this work would be to consider if the counters actually covariate, i.e. if a model for one counter using another counter as a covariate would yield even better results. Furthermore considering not just additive covariates, but multiplicative ones as well and past predictions are both viable expansion of the current model (model 11) in this thesis. Both of which I suspect would improve the fit.

From Ericssons point of view, implementing this work into an automated anomaly detection scheme would prove very fruitful, at least based on our experiments in chapter 4. In this case comparing the ability of the above mentioned models ability to predict anomalies would again prove another interesting extension of this thesis.

References

Ericsson Radio Systems AB (2001). *Basic Concepts of WCDMA Radio Access Network*. URL: http://www.cs.ucsb.edu/~almeroth/classes/W03.595N/papers/wcdma-concepts.pdf.

Kedem, Benjamin and Konstantinos Fokianos (2002). *Regression Models for Time Series Analysis*. John Wiley & Sons, Inc. ISBN: 0-471-36355-3.

Olsson, Ulf (2002). *Generalized Linear Models - An Applied Approach*. Ulf Olsson and Studentlitteratur 2002. ISBN: 978-91-44-03141-5.

Leonhard Held, Michael Höhle and Mathias Hofmann (2005). "A statistical framework for the analysis of multivariate infectious disease surveillance counts". In: *Statistical Modelling* 5, pp. 187–191.

Hansen, Eric W. (2002). *Fourier Transforms - Principles and Applications*. John Wiley & Sons, Inc. ISBN: 978-1-118-47914-8.