Dynamic prediction of the number of page views

Laurenz Eveleens

May 30, 2005

BMI-paper
Vrije Universiteit
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam

1 Introduction

We call each time a person enters a website a page view. Many programs exist that keep track of the number of page views. The resulting data give statistics such as the origin of the views or the total number of page views on a given time epoch. Programs like Nedstat have these features. More interesting is a prediction of the total number of page views at the end of a day. This could be done on the day before, but it would be useful to also use the information gathered on the day itself.

In this paper we discuss models which can estimate the number of page views on a day and use the number of page views that have already occurred on that day. We will first discuss the model described in [1] and derive that estimator mathematically. Then we will construct two different models which we will compare with the first one. Contrary to the models described in [1], these models use information about the number of page views which already have occurred on a day and information gathered in the past.

The statistical problem of predicting the number of occurrences over the whole day described above also applies to other areas. Take for instance a call center. A call center delivers services by telephone. In order to do so, agents are hired to give this service. An important problem is how many agents are needed to give an acceptable level of service. This needs to be known for the long term but also for each day or even each half an hour. Here it also plays an important role to have a good prediction on the number of incoming telephone calls on a day. The better the prediction the better the call center can plan their agents. If we could use information of the first few hours of the day in our prediction for the rest of the day, this may result in better service or lower costs for the call center by planning more or less agents.

For software companies it is also an issue. When software companies produce new products they will receive reports of faults in the software. The companies need employees to deal with these. In order to schedule them efficiently a good prediction is needed on the amount of reports. In [2] such a model is described in which only information of previous releases is used for prediction.

The paper is constructed in the following way. In chapter 2 we describe the mathematical assumptions we will use. Based on these assumptions the paper discusses a number of prediction methods in chapter 3. Chapter 4 describes the model which we will use to compare the estimators. Chapter 5 applies the estimators to a data set and compares the models to each other using the model of chapter 4. Chapter 6 concludes the paper with some final remarks.

2 Mathematical assumptions

We look at a single day of 24 hours and split the day in m time epochs, for instance half hours. Denote the number of page views per time epoch by the random vector $\overrightarrow{N} = (N_1, N_2, ..., N_m)$. We model the occurrence of page views as an inhomogeneous Poisson process, which means that within each time epoch the number of page views is Poisson distributed with parameter $\lambda_i (i \in m)$ and for disjunct time epochs these numbers are stochastically independent of each other. This means that

$$\mathbb{P}(N_i = n_i) = \frac{\lambda_i^{n_i}}{n_i!} e^{-\lambda_i} \quad n_i = 0, 1, 2, ...$$

The parameter vector $\overrightarrow{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)$ is unknown. We want to estimate this, at least for future time epochs. We assume that there is information on the relation between these parameters. There is some positive constant τ and a probability vector $\overrightarrow{p} = (p_1, p_2, ..., p_m)$ such that

$$\overrightarrow{\lambda} = \tau \overrightarrow{p}$$

We assume this probability vector is known. This probability vector could be derived from the data from the past. We assume to have a data set $\overrightarrow{x} = (x_1, x_2, ..., x_j)$ with j the number of data points from the past. This could for instance be 3 weeks of data where each data point is the number of page views in a half an hour.

The total number of page views T during the day has a Poisson distribution with parameter τ.

$$\mathbb{P}(T = t) = \frac{\tau^t}{t!} e^{-\tau} \quad t = 0, 1, 2, ...$$

We also have data for the first i $(i < m)$ time epochs of the day $(n_1, n_2, ..., n_i)$. Finding an estimate for $\overrightarrow{\lambda}$ is equivalent of finding an estimate for τ.

In the introduction we discussed that we wanted to predict T, the number of page views on a day. We see here however that estimating τ also results in a prediction of T. This is true because the squared error between the true value and the estimator is minimized by taking as estimator the expectation of the true value. In this case T is Poisson distributed which has an expectation of τ.

2.1 Probability vector

Is it realistic to assume there exist a probability vector which determines the course of the day? It is not strange to think there is a certain relation between the number of page views for the half hours. In the night we would expect lower amounts than during the day-time with peaks round 7 or 8 in the Netherlands. This vector should perhaps be different for the week and weekends. This probably would hold also for special days and holidays.

It would be interesting to assess the validity of the assumption that this probability vector exists. In this paper we will not address this issue however. Since we are only interested in comparing the models we do not look at differences of the vector between different weekdays either. We look only at data for one given day and compare the models based on those data.

3 Estimation models

In this chapter we discuss three models to predict the total number of page views at the end of the day. The first model is the one used by Nedstat. Then we discuss two other models we have constructed.

3.1 Model 1: Maximum likelihood estimator

The first model is based on the maximum likelihood estimator. We predict the total number of page views by finding an estimator for its expectation τ. Since we know the amount of page views for the first i epochs we can use the function $l_i(\tau) = \mathbb{P}_\tau(N_j = n_j, 1 < j < i)$ as likelihood function. Since we are only interested in the value of τ and not the value of the maximum we can take the logarithm of the function, since that does not change the location of the maximum. After differentiating and setting the derivative equal to zero, we find the maximum likelihood estimator for τ.

3.1.1 Calculations

$$
\begin{aligned}
l_i(\tau) &= \mathbb{P}_\tau(N_j = n_j, 1 < j < i) = \prod_{j=1}^{i} \mathbb{P}_\tau(N_j = n_j) \\
\log l_i(\tau) &= \log \prod_{j=1}^{i} \mathbb{P}_\tau(N_j = n_j) \\
&= \sum_{j=1}^{i} \log \mathbb{P}_\tau(N_j = n_j) = \sum_{j=1}^{i} \log \frac{(\tau p_j)^{n_j}}{n_j!} e^{-\tau p_j} \\
&= \sum_{j=1}^{i} (n_j \log \tau p_j - \log n_j! - \tau p_j) \\
&= \sum_{j=1}^{i} (n_j \log \tau + n_j \log p_j - \log n_j! - \tau p_j)
\end{aligned}
$$

Now we try to find the value of τ which maximizes this function. Here we use the fact that we can forget terms that do not relate to τ.

$$\begin{aligned}\frac{\partial}{\partial\tau}(\sum_{j=1}^{i}(n_j\log\tau-\tau p_j)) &= 0\\ \frac{\sum_{j=1}^{i} n_j}{\tau}-\sum_{j=1}^{i} p_j &= 0\\ \frac{\sum_{j=1}^{i} n_j}{\tau} &= \sum_{j=1}^{i} p_j\\ \tau &= \frac{\sum_{j=1}^{i} n_j}{\sum_{j=1}^{i} p_j}\end{aligned}$$

Here we have found a maximum likelihood estimator for τ. This means that if we know the amount of page views which occurred during the first i time epochs we can predict the total number of page views we will get by

$$\hat{\tau}=\frac{\sum_{j=1}^{i} n_j}{\sum_{j=1}^{i} p_j} \tag{1}$$

This estimator is also described in [1].

3.2 Model 2: Gamma prior

In the second model we assume τ has some prior distribution G. We can now use the Bayes procedure to find an estimator for τ. From Bayes theory we know that if the conditional probability density of g given $\tau = t$ is $g(x|t)$, then the posterior probability density function of τ when $X = x$ is $g(t|x) \propto g(x|t)g(t)$, where $\propto$ means up to some additive or positive multiplicative constant independent of the values of the parameters.

The question remains which prior distribution to choose. Each choice of a prior distribution would result in a different posterior. As [4] points out this choice is rather arbitrary. We choose a distribution only for mathematical convenience. The advantage is of course that due to the statistical properties we can find an estimator which is relatively easy to deal with.

For this purpose we assume in this case the prior distribution to be a Gamma distribution. We choose this distribution because it is a conjugate prior. That means if the likelihood of the data ($= g(x|t)$) is Poisson and the prior distribution is Gamma, that the resulting posterior distribution is again Gamma. So we take

$$f_{\alpha,\beta}(t)=\frac{(\frac{t}{\beta})^{\alpha-1}}{\beta\Gamma(\alpha-1)}e^{-\frac{t}{\beta}}\quad t>0$$

The parameters $\alpha > 0$ and $\beta > 0$ could be estimated using the historical data. One way would be to estimate this α and β using the maximum likelihood estimator. In our data analysis we will use this method. Other methods however could also be applied.

3.2.1 Calculations

The posterior distribution we get from the Bayes procedure

$$\begin{aligned}\mathbb{P}(N_i = n_i|t) &= \frac{(tp_i)^{n_i}}{n_i!}e^{-tp_i}\\ f_{\alpha,\beta}(t|N_1 = n_1, \ldots, N_i = n_i) &\propto \mathbb{P}(N_1 = n_1, \ldots, N_i = n_i|t) f_{\alpha,\beta}(t)\\ &\propto \prod_{j=1}^{i} \mathbb{P}(N_i = n_i|t) f_{\alpha,\beta}(t)\\ &\propto \prod_{j=1}^{i} \frac{(tp_j)^{n_j}}{n_j!}e^{-(tp_j)} f_{\alpha,\beta}(t)\end{aligned}$$

Rewriting this formula so that all factors depending on t are grouped gives:

$$\begin{aligned}f_{\alpha,\beta}(t|N_1 = n_1, \ldots, N_i = n_i) &\propto \prod_{j=1}^{i} \frac{(tp_j)^{n_j}}{n_j!}e^{-(tp_j)} f_{\alpha,\beta}(t)\\ &\propto t^{\sum_{j=1}^{i} n_j} e^{-(t\sum_{j=1}^{i} p_j)} f_{\alpha,\beta}(t)\\ &\propto t^{\sum_{j=1}^{i} n_j} e^{-(t\sum_{j=1}^{i} p_j)} \frac{(\frac{t}{\beta})^{\alpha-1}}{\beta\Gamma(\alpha-1)} e^{-\frac{t}{\beta}}\\ &\propto t^{\sum_{j=1}^{i} n_j} e^{-(t\sum_{j=1}^{i} p_j)} t^{\alpha-1} \frac{(\frac{1}{\beta})^{\alpha-1}}{\beta\Gamma(\alpha-1)} e^{-\frac{t}{\beta}}\\ &\propto t^{\alpha-1+\sum_{j=1}^{i} n_j} e^{-t(\frac{1}{\beta}+\sum_{j=1}^{i} p_j)}\end{aligned}$$

Up to some factor not depending on t this is again a Gamma distribution with parameters $\alpha_i = \alpha + \sum_{j=1}^{i} n_j$ and $\frac{1}{\beta_i} = \frac{1+\beta\sum_{j=1}^{i} p_j}{\beta}$, $\beta_i = \frac{\beta}{1+\beta\sum_{j=1}^{i} p_j}$. We find the posterior probability density of τ to be Gamma(α_i, β_i).

As described in [5] we can extract a proper estimator from the posterior by looking at the squared error loss function, which is a function of the difference between the estimator and the real value of τ. As explained on page 228 we can find the minimum of this function when $\hat{\tau} = \mathbb{E}(\tau|N_1 = n_1, \ldots, N_i = n_i)$. This is the expectation of the posterior distribution we found to be Gamma. Thus

$$\hat{\tau} = \alpha_i\beta_i = \frac{(\alpha + \sum_{j=1}^{i} n_j)\beta}{1 + \beta\sum_{j=1}^{i} p_j} \tag{2}$$

If β is big and $\alpha\beta$ is small we get the same predictor as method 1.

3.3 Model 3: Conditional expectation of T

Instead of looking for an estimator for τ we now want to find a predictor for T directly. Let us assume T has some discrete distribution G. $g_k = \mathbb{P}_G(T = k)$. We can now find an estimator by looking at $E_G(T|N_1 = n_1, \ldots, N_j = n_j)$. We will see that the resulting estimator needs to be calculated numerically. Each choice of G gives us a different estimator using this method. The choice of G however is somewhat arbitrary. Based on the available data a suitable distribution could be found and used. In our data analysis we will use the empirical distribution of T. We get this by using the days of which we know the total number of page views that occurred that day.

3.3.1 Calculations

$$E_G(T|N_1 = n_1, \ldots, N_j = n_j) = \sum_{k=1}^{\infty} k\mathbb{P}(T = k|N_1 = n_1, \ldots, N_j = n_j)$$

$$\mathbb{P}(T = k|N_1 = n_1, \ldots, N_j = n_j) = \frac{\mathbb{P}(T = k, N_1 = n_1, \ldots, N_j = n_j)}{\mathbb{P}(N_1 = n_1, \ldots, N_j = n_j)}$$

$$= \frac{\mathbb{P}(N_1 = n_1, \ldots, N_j = n_j|T = k)\mathbb{P}_G(T = k)}{\sum_{l=0}^{\infty} \mathbb{P}(N_1 = n_1, \ldots, N_j = n_j|T = l)\mathbb{P}_G(T = l)} \qquad k \geq \sum_{j=1}^{i} n_j$$

We know that the vector N given T has a multinomial distribution with parameters $(T, p_1, \ldots, p_m)$. From [6] we know that a subset of the remaining time epochs of this vector is again multinomially distributed. This yields:

$$\mathbb{P}(N_1 = n_1, \ldots, N_j = n_j|T = k) = \frac{k!}{(k - \sum_{j=1}^{i} n_j)! \prod_{j=1}^{i} n_i!}(1 - \sum_{j=1}^{i} p_j)^{k - \sum_{j=1}^{i} n_j} \prod_{j=1}^{i} p_j^{n_j}$$

$$\mathbb{P}(T = k|N_1 = n_1, \ldots, N_j = n_j) = \frac{\frac{k! g_k}{(k - \sum_{j=1}^{i} n_j)! \prod_{j=1}^{i} n_i!}(1 - \sum_{j=1}^{i} p_j)^{k - \sum_{j=1}^{i} n_j} \prod_{j=1}^{i} p_j^{n_j}}{\sum_{l=\sum_{j=1}^{i} n_j}^{\infty} \frac{l! g_l}{(l - \sum_{j=1}^{i} n_j)! \prod_{j=1}^{i} n_i!}(1 - \sum_{j=1}^{i} p_j)^{l - \sum_{j=1}^{i} n_j}) \prod_{j=1}^{i} p_j^{n_j}}$$

$$E_G(T|N_1 = n_1, \ldots, N_j = n_j) = \frac{\sum_{k=\sum_{j=1}^{i} n_j}^{\infty} \frac{k \cdot k! g_k}{(k - \sum_{j=1}^{i} n_j)! \prod_{j=1}^{i} n_i!}(1 - \sum_{j=1}^{i} p_j)^{k - \sum_{j=1}^{i} n_j} \prod_{j=1}^{i} p_j^{n_j}}{\sum_{l=\sum_{j=1}^{i} n_j}^{\infty} \frac{l! g_l}{(l - \sum_{j=1}^{i} n_j)! \prod_{j=1}^{i} n_i!}(1 - \sum_{j=1}^{i} p_j)^{l - \sum_{j=1}^{i} n_j} \prod_{j=1}^{i} p_j^{n_j}}$$

From this last formula we can remove the terms $\prod_{j=1}^{i} n_i!$, $\prod_{j=1}^{i} p_j^{n_j}$, $(1 - \sum_{j=1}^{i} p_j)^{-\sum_{j=1}^{i} n_j}$ and $(1 - \sum_{j=1}^{i} p_j)^{-\sum_{j=1}^{i} n_j}$ since they do not depend on either k or l and occur in both the numerator and denominator. Rearranging the formula gives:

$$E_G(T|N_1 = n_1, \ldots, N_j = n_j) = \frac{\sum_{k=s}^{\infty} k \cdot \frac{k! g_k}{(k-s)!}(1 - \sum_{j=1}^{i} p_j)^k}{\sum_{l=s}^{\infty} \frac{l! g_l}{(l-s)!}(1 - \sum_{j=1}^{i} p_j)^l} \tag{3}$$

with $s = \sum_{j=1}^{i} n_j$.

We can verify this function by taking G to be the Poisson(τ) distribution. If T has a Poisson distribution the expectation of T given the first time epochs would be $\sum_{j=1}^{i} n_j + \tau(1 - \sum_{j=1}^{i} p_j)$. It can be seen that 3 gives the same expectation. Appendix: Check of model 3 shows this.

4 Model comparison

In the previous chapter we have discussed three different ways to estimate the total number of page views we expect to occur at the end of the day. In this chapter we describe the method we use to compare the second and third model to the first one.

Let $T_{ij}^{(k)}$ be the estimation of the number of page views on day i of time epoch j of model k. Let T_i be the actually observed number of page views on day i. Now let

$$e_{ij}^{(k)} = (T_{ij}^{(k)} - T_i)^2$$

$e_{ij}^{(k)}$ is the difference of the estimated number of page views and the real number of page views on a day. Looking at the values $e_{ij}^{(k)}$ we would like to find out whether the location of the distribution of $e_{ij}^{(1)}$ differs from the location of the distribution of $e_{ij}^{(2)}$ or $e_{ij}^{(3)}$. We hope to find that one of these locations is in fact smaller then the one of model 1. This would mean that the estimators of model 2 and 3 can predict the number of page views at the end of the day better then model 1.

We can test this using the Wilcoxon Rank-Sum Test. Let us call the distribution of $e_{ij}^{(1)}$ A and of $e_{ij}^{(2)}$ or $e_{ij}^{(3)}$ B. We then get the hypothesis test

$$H_0 := A \lesssim B$$
$$H_1 := A > B$$

5 Data analysis

In this chapter we apply the models we discussed in chapter 3 to the data. Then we will compare the models as described in the previous chapter.

5.1 Website data

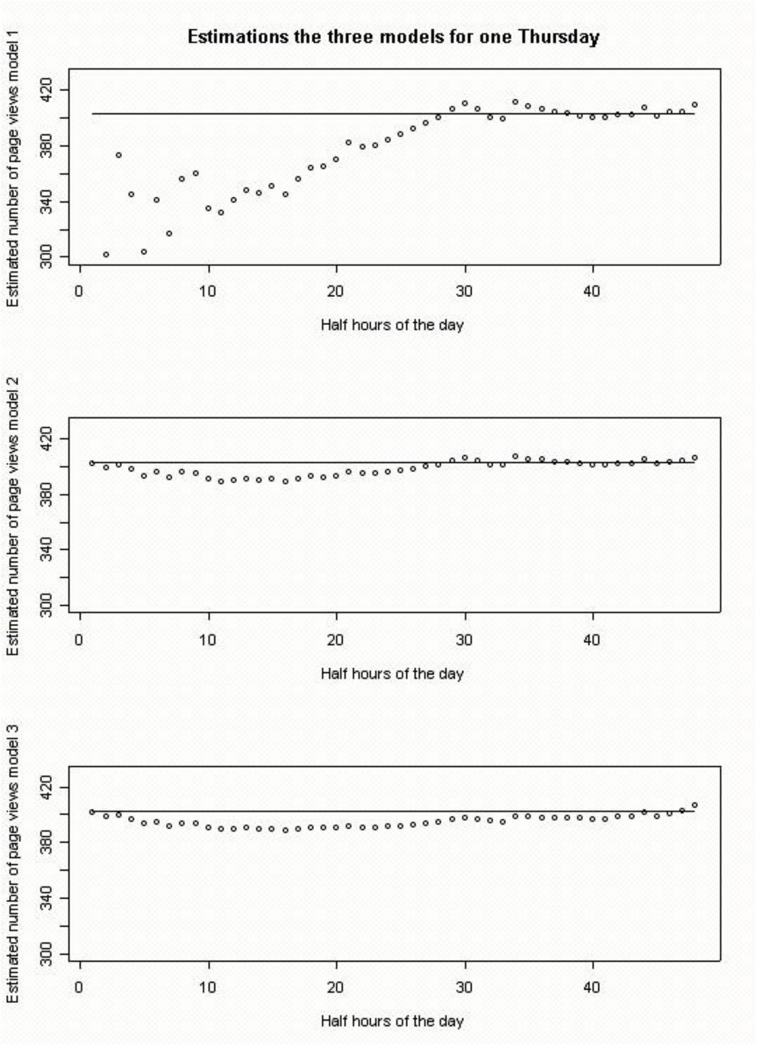
In this section we analyze the models using data gathered from the site 'www.cs.vu.nl'. For one and a half month we collected the number of page views that occurred at each half hour per day. We restrict attention to the data obtained on Thursdays because of reasons discussed in chapter 2. This data consists of 6 days with each $m = 48$ time epochs. In the table below a sample of the data is shown along with the total number of page views on these Thursdays.

half hour	23-09-2004	30-09-2004
12:30-13:00	14	9
13:00-13:30	15	14
13:30-14:00	13	12
14:00-14:30	16	11
....	....	....
day total	403	390

With this data we constructed the probability vector p by taking the sum of the page views for each half hour over these 6 days and divide it by the sum of all half hours. Since we only took Thursdays we do not have to take into account possible differences of the probability vector between days. We also ignore possible trends within our data. We can do this since we are only interested in comparing the estimators. If the model is applied in practice it is advisable to pay more attention to the choice of the probability vector.

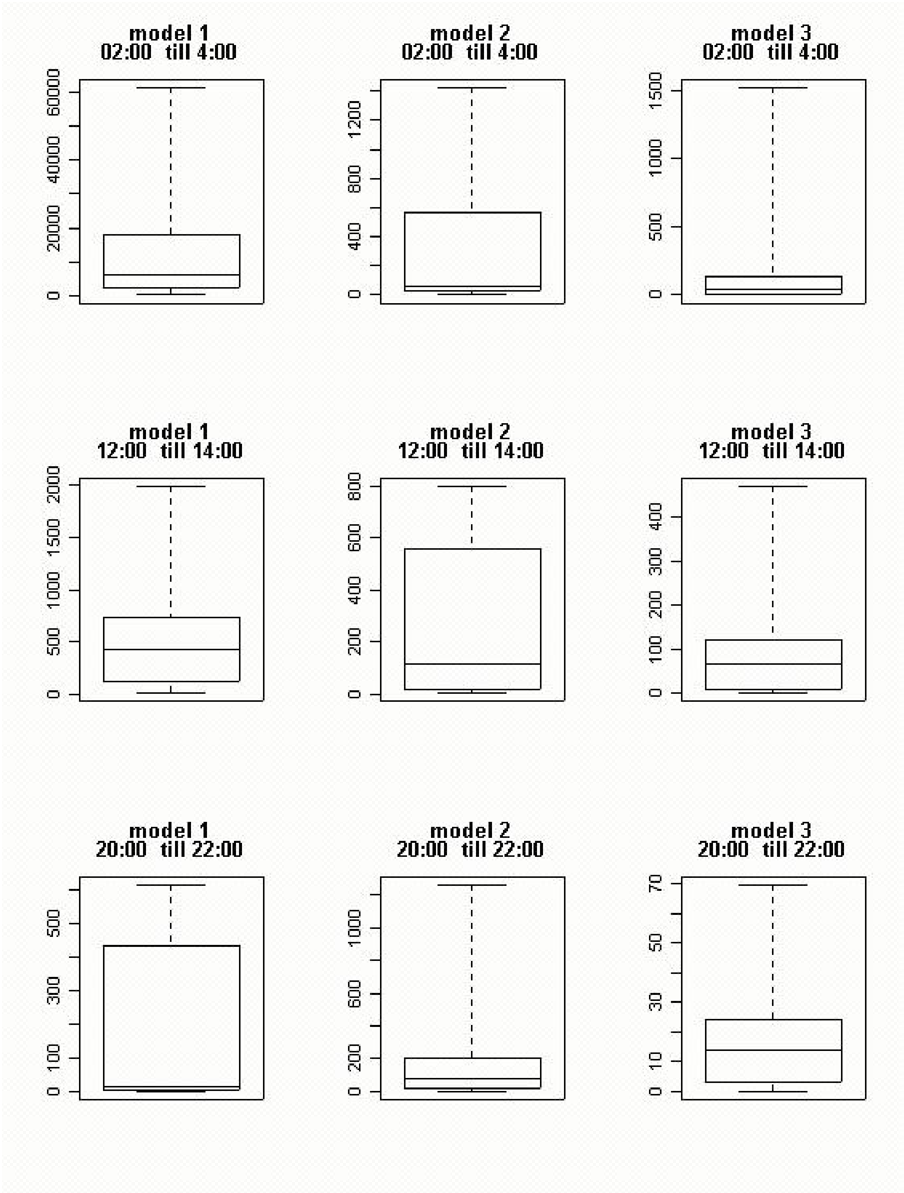
In appendix *Models programmed in R* the functions for calculating the estimators can be found written in the statistical package R. For the second model we used the maximum likelihood estimator to estimate α and β. For our third model we used the empirical distribution for g_k. Note that we have only few data points in this model. For this reason we modified g_k so that each data point ±5 has a probability mass. This way g_k consists of 66 points instead of 6. In the next chapter we will look at a larger set of data.

To get an impression of the estimators we depict below the estimations per half hour for one Thursday (see next page). The straight line depicts the real amount of page views on that day. As we can see in the graph, the first model has difficulty getting close to the real number of page views at the beginning of the day. As the day progresses the estimation of this model gets better and better. Model 2 and 3 however seem, also at the beginning of the day, able to estimate the total number of page views accurately.



5.1.1 Model comparison

In order to compare the models we calculate the values of $e_{ij}^{(k)}$. We chose to combine 4 time epochs of a half an hour for the $e_{ij}^{(k)}$ giving $j = 1, .., 12$. For each Wilcoxon test we now have $4 * 6 = 24$ data points. To given an idea of the values of the differences we show box plots for the epochs $j = 2, 6, 11$.



Again we see that the estimations of model 2 and 3 are closer to the real number of page views in the beginning of the day. Around noon they still have less difference from the real value than the first model, but it is less distinct. At the end of the day it is more difficult to tell the difference between the models. It does however seem that at the end of the day model 3 performs the best.

If we preform the Wilcoxon test we get the following p-values.

half hour	Model 1 vs Model 2 p-value	Model 1 vs Model 3 p-value
0:00-2:00	0.0000	0.0000
2:00-4:00	0.0000	0.0000
4:00-6:00	0.0000	0.0000
6:00-8:00	0.0000	0.0000
8:00-10:00	0.0000	0.0000
10:00-12:00	0.0000	0.0000
12:00-14:00	0.0103	0.0000
14:00-16:00	0.3914	0.1034
16:00-18:00	0.2466	0.0039
18:00-20:00	0.4553	0.1504
20:00-22:00	0.9089	0.1430
22:00-24:00	0.9909	0.2664

If we take a rejection value of 0.05 ($\alpha = 5\%$) we see that we may reject our hypothesis till 14.00 for model 2. This means that till 14.00 the estimation of model 2 is closer to the real value than the estimation of model 1. After 14.00 however we cannot reject the null hypothesis anymore. From that point on, we can no longer state that model 2 preforms any better than model 1.

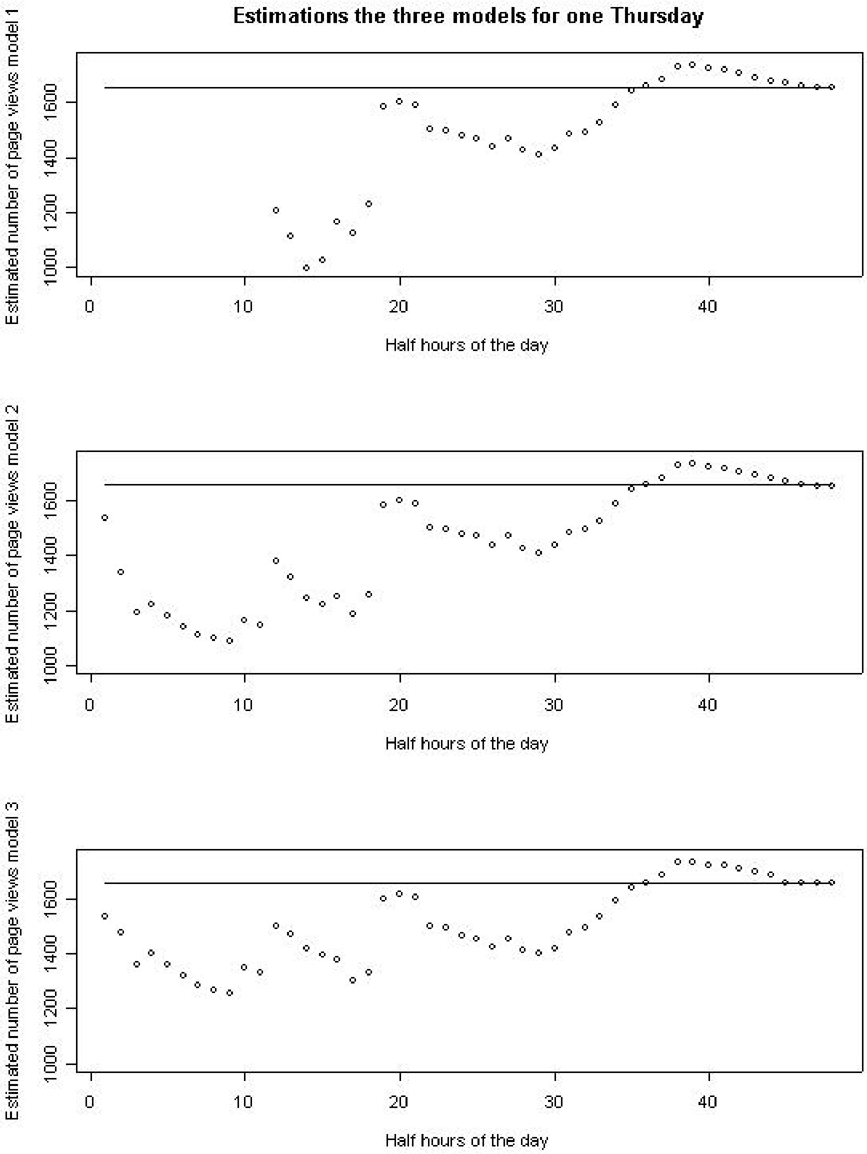
For model 3 we see the same thing as with model 2. Up till 14.00 we can reject the null hypothesis. After that the estimations of model 3 and model 1 can not be said to be any different. We see the results from the Wilcoxon test also in the graphs. Till 14.00 the estimation of model 1 is further from the real value then the other two models. After 14.00 it gets harder to distinguish the difference between the models.

5.2 Call center data

In this section we analyze the models using data acquired from [7]. The data describes telephone calls recorded over 12 months from 1 January 1999 to 31 December 1999 at a call center of a bank in Israel. (For more information about the data see [7]). Again we looked only at the data from the Thursdays. This data consists of 52 days. We summed the data to get a data set with $m = 48$ time epochs with the number of calls that occurred during that time epoch. In the table below a sample of that data set is shown along with the total number of page views on these Thursdays.

half hour	07-01-1999	14-01-2004
12:30-13:00	41	59
13:00-13:30	63	53
13:30-14:00	33	54
14:00-14:30	42	42
....	....	....
day total	1655	1657

With this data we constructed the probability vector p by taking the sum of the calls for each half hour over these 52 days and divide it by the sum of all half hours. We use the same functions as we did with the website data. To get an impression of the estimators we depict below the estimations per half hour for one Thursday. The straight line depicts the real amount of page views on that day.



As we can see in this graph the model 1 has difficulty getting close to the real number of page views at the beginning of the day. As the day progresses the estimation of this model gets somewhat closer to the real value. Model 2 and 3 estimate the real value at the beginning of the day more accurately. At the end of the day however we cannot see much difference in this graph.

5.2.1 Model comparison

In order to compare the models we again calculate the values of $e_{ij}^{(k)}$. We chose to combine 4 time epochs of a half an hour for the $e_{ij}^{(k)}$ giving $j = 1, .., 12$. For each Wilcoxon test we now have $4 * 52 = 208$ data points. If we perform the test we get the following p-values.

half hour	Model 1 vs Model 2 p-value	Model 1 vs Model 3 p-value
0:00-2:00	0.0000	0.0000
2:00-4:00	0.0000	0.0000
4:00-6:00	0.0000	0.0000
6:00-8:00	0.0002	0.0000
8:00-10:00	0.1592	0.0054
10:00-12:00	0.3850	0.1020
12:00-14:00	0.3686	0.0996
14:00-16:00	0.4140	0.0726
16:00-18:00	0.4735	0.0660
18:00-20:00	0.4573	0.0094
20:00-22:00	0.4108	0.0000
22:00-24:00	0.4940	0.0000

The comparison between model 1 and model 2 shows the same results as we had with the data from the website. At the beginning of the day model 2 estimates the real number of incoming calls more accurately than model 1. After in this case 8.00 we no longer can be certain the estimation of model 2 is better, we cannot reject our null hypothesis from that point onward.

The comparison between model 1 and model 3 however shows a different picture. Only between $10 : 00$ and $18 : 00$ we can not reject our null hypothesis. For all other time epoch we know that the location of the distribution of $e_{ij}^{(3)}$ is smaller then the location of the distribution of $e_{ij}^{(1)}$.

6 Conclusions

In this paper we have discussed different estimators to estimate the total number of page views at the end of a day T. For these estimations we used the number of page views N_i that already have occurred till time i. We assume this T to be Poisson distributed with parameter τ. We also assume there exists a probability vector p such that $\overrightarrow{\lambda} = \tau \overrightarrow{p}$.

Our first model used a maximum likelihood estimator for τ:

$$\hat{\tau} = \frac{\sum_{j=1}^{i} n_j}{\sum_{j=1}^{i} p_j} \tag{4}$$

This is the estimator that is used by Nedstat. The second model used Bayes theory and a Gamma prior to find the estimator:

$$\hat{\tau} = \alpha_i \beta_i = \frac{(\alpha + \sum_{j=1}^{i} n_j)\beta}{1 + \beta \sum_{j=1}^{i} p_j} \tag{5}$$

The last model used the expectation of T given the time units that have already passed to find:

$$E_G(T|N_1 = n_1, \dots, N_j = n_j) = \frac{\sum_{k=s}^{\infty} k \cdot \frac{k! g_k}{(k-s)!} (1 - \sum_{j=1}^{i} p_j)^k}{\sum_{l=s}^{\infty} \frac{l! g_l}{(l-s)!} (1 - \sum_{j=1}^{i} p_j)^l} \tag{6}$$

This last model needs to be calculated numerically to find the estimation.

Data analysis was preformed on two data sets. One set was data gathered from the site "www.cs.vu.nl". For 1,5 month the total number of page views per half hour per day was collected. The second data set was data acquired from a call center of a bank in Israel. This data set contained 12 months of data of incoming telephone calls.

To compare the models only the data from the Thursdays was used for both data sets. The models were compared using a Wilcoxon test on the values

$$e_{ij}^{(k)} = (T_{ij}^{(k)} - T_i)^2$$

$T_{ij}^{(k)}$ is the prediction of the number of page views (calls) on day i of time epoch j of model k and T_i is the real number of page views (calls) on day i.

The comparison between the models showed that model 2 could estimate the real value more accurately than model 1 only for the first half of the day for both the data from the website as for the data from the call center. For the second half of the day, we could no longer reject our null hypothesis that the location of the distributions was different.

The comparison between model 1 and 3 show some different results for each data set. With both data sets model 3 performs better in the first half of the day. In the data from the website we can not claim model 3 is better on the second half of the day, much like the comparison between model 1 and 2. For the data from the call center however we see that model 3 for the hours 18 : 00 to 24 : 00 does perform better than model 1.

7 Acknowledgement

I thank Geurt Jongbloed for his help in creating this paper, both for the mathematical input as well as his remarks while writing this.

8 References

[1] Van Gelder, P., Beijer, G., Berger, M., 2003, Statistical analysis of page views on website, *Proceedings of the third international conference on data mining methods and databases, Management Information systems, vol 6, p. 979-988*

[2] Jongbloed, G., Verbaken, T., 2003, Modelling the process of incoming problem reports on released software products, *Applied Stochastic Models in Business and Industry, Vol 20, p. 131-142*

[3] Bhat, B., Kulkarni, N., 1966, On Efficient Multinomial Estimation. *Journal of the Royal Statistical Society, Series B, Vol 28, No 1*

[4] Koole, G., Jongbloed, G.: Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry* **17**, p. 307-318

[5] DeGroot, M., 1970, Optical statistical decisions

[6] Johnson, N., Kotz, S., 1969, Discrete distributions

[7] I. Guedj, A. Mandelbaum, 1999, *http://iew3.technion.ac.il/serveng/*, this data set is used for instance in V.d. Post, R., 2004, Risicobeheersing bij call center planning

9 Appendix: Check of model 3

We want to check whether 3 is correct. We can do this by taking G to be the Poisson(τ) distribution. This means that $\mathbb{E}(T|N_1 = n_1, \ldots, N_i = n_i)$ should be equal to $\sum_{j=1}^{i} n_j + \tau(1 - \sum_{j=1}^{i} p_j)$.

We can see that this expectation should be equal to $\sum_{j=1}^{i} n_j + \tau(1 - \sum_{j=1}^{i} p_j)$ using the expectation of the multinomial distribution. We get $\mathbb{E}(T|N_1 = n_1, \ldots, N_i = n_i) = \sum_{j=1}^{i} n_j + \mathbb{E}(N_{i+1} = n_{i+1} + \ldots + N_m = n_m|N_1 = n_1, \ldots, N_i = n_i) = \sum_{j=1}^{i} n_j + \mathbb{E}(T)\sum_{j=i+1}^{m} p_j = \sum_{j=1}^{i} n_j + \mathbb{E}(T)(1 - \sum_{j=1}^{i} p_j)$. If T is Poisson(τ) distributed the $\mathbb{E}(T) = \tau$.

We try to find the same result using 3

$$g_k = \frac{\tau^k}{k!}e^{-\tau}$$

$$E_G(T|N_1 = n_1, \ldots, N_j = n_j) = \frac{\sum_{k=\sum_{j=1}^{i} n_j}^{\infty} k \cdot k! \frac{\tau^k}{k!} e^{-\tau}(1 - \sum_{j=1}^{i} p_j)^k}{(k - \sum_{j=1}^{i} n_j)! \sum_{l=\sum_{j=1}^{i} n_j}^{\infty} \frac{l! \frac{\tau^l}{l!} e^{-\tau}}{(l - \sum_{j=1}^{i} n_j)!}(1 - \sum_{j=1}^{i} p_j)^l}$$

$$E_G(T|N_1 = n_1, \ldots, N_j = n_j) = \frac{\sum_{k=\sum_{j=1}^{i} n_j}^{\infty} k\tau^k(1 - \sum_{j=1}^{i} p_j)^k}{(k - \sum_{j=1}^{i} n_j)! \sum_{l=\sum_{j=1}^{i} n_j}^{\infty} \frac{\tau^l}{(l - \sum_{j=1}^{i} n_j)!}(1 - \sum_{j=1}^{i} p_j)^l}$$

$$= \frac{S_k}{S_l}$$

In the following calculations we use the fact that:

$$\sum_{k=0}^{\infty} \frac{u^k}{k!} = e^u$$

$$\sum_{k=1}^{\infty} \frac{ku^k}{k!} = \sum_{k=1}^{\infty} \frac{u^k}{(k-1)!} = u\sum_{k=0}^{\infty} \frac{u^k}{k!} = ue^u$$

If we look at S_k and S_l more closely we see that:

$$\begin{aligned}
S_l &= \sum_{l=\sum_{j=1}^{i} n_j}^{\infty} \frac{\tau^l}{(l-\sum_{j=1}^{i} n_j)!}(1-\sum_{j=1}^{i} p_j)^l \\
&= \sum_{l=\sum_{j=1}^{i} n_j}^{\infty} \frac{(\tau-\tau\sum_{j=1}^{i} p_j)^l}{(l-\sum_{j=1}^{i} n_j)!} \\
&= \sum_{l=0}^{\infty} \frac{(\tau-\tau\sum_{j=1}^{i} p_j)^l}{l!} \cdot (\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j} \\
&= e^{(\tau-\tau\sum_{j=1}^{i} p_j)} \cdot (\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j}
\end{aligned}$$

$$\begin{aligned}
S_k &= \sum_{k=\sum_{j=1}^{i} n_j}^{\infty} \frac{k\tau^k(1-\sum_{j=1}^{i} p_j)^k}{(k-\sum_{j=1}^{i} n_j)!} \\
&= \sum_{k=0}^{\infty} \frac{(k+\sum_{j=1}^{i} n_j)(\tau-\tau\sum_{j=1}^{i} p_j)^k}{k!} \cdot (\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j} \\
&= \sum_{k=1}^{\infty} \frac{k(\tau-\tau\sum_{j=1}^{i} p_j)^k}{k!} \cdot (\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j} + (\sum_{j=1}^{i} n_j)\sum_{k=0}^{\infty} \frac{(\tau-\tau\sum_{j=1}^{i} p_j)^k}{k!} \cdot (\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j} \\
&= \sum_{k=1}^{\infty} \frac{k(\tau-\tau\sum_{j=1}^{i} p_j)^k}{k!} \cdot (\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j} + (\sum_{j=1}^{i} n_j)e^{(\tau-\tau\sum_{j=1}^{i} p_j)}(\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j} \\
&= (\tau-\tau\sum_{j=1}^{i} p_j)e^{(\tau-\tau\sum_{j=1}^{i} p_j)} \cdot (\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j} + (\sum_{j=1}^{i} n_j)e^{(\tau-\tau\sum_{j=1}^{i} p_j)}(\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j}
\end{aligned}$$

Plugging S_l and S_k in the formula gives:

$$\begin{aligned}
E_G &= \frac{S_k}{S_l} \\
&= \frac{(\tau-\tau\sum_{j=1}^{i} p_j)e^{(\tau-\tau\sum_{j=1}^{i} p_j)} \cdot (\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j} + (\sum_{j=1}^{i} n_j)e^{(\tau-\tau\sum_{j=1}^{i} p_j)}(\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j}}{e^{(\tau-\tau\sum_{j=1}^{i} p_j)} \cdot (\tau-\tau\sum_{j=1}^{i} p_j)^{\sum_{j=1}^{i} n_j}} \\
&= (\tau-\tau\sum_{j=1}^{i} p_j) + (\sum_{j=1}^{i} n_j) \\
&= (\sum_{j=1}^{i} n_j) + \tau(1-\sum_{j=1}^{i} p_j)
\end{aligned}$$

So we see that if we take G to be Poisson we get the estimation we expected. This shows that our estimator is in fact correct.

10 Appendix: Models programmed in R

10.1 Model 1

```
model1 = function(N,P){
   if(sum(P)==0){
      0
   } else {
      sum(N)/sum(P)
   }
}
```

10.2 Model 2

```
model2 = function(N,P,alpha,beta){
   (alpha+sum(N))*beta/(1+beta*sum(P))
}
```

10.3 Model 3

The goal was to solve 3 numerically. Implementing the formula directly however is not an option, since most programs (including R) cannot store high values such as 300!. In order to avoid this problem we define $u_k = \frac{k!g_k}{(k-s)!}(1 - \sum_{j=1}^{i} p_j)$. 3 now becomes $\frac{\sum_{k=s}^{\infty} ku_k}{\sum_{l=s}^{\infty} u_l}$. We then divide both the nominator and denominator by the maximum value of u_k say u_{k0}. This gives $\frac{\sum_{k=s}^{\infty} k\frac{u_k}{u_{k0}}}{\sum_{l=s}^{\infty} \frac{u_l}{u_{k0}}}$. Now these fractions in the nominator and denominator lie between 0 and 1. We could now also take the e power of the log of this fraction. Doing this gets rid of the high valued factorials. Doing all this yields the following formula which can be implemented directly:

$$\frac{\sum_{k=s}^{\infty} ke^{\sum_{i=k-s}^{k} log(i) - \sum_{i=k0-s}^{k0} log(i) + log(g_k) - log(g_{k0}) + (k-k0) log(1-\sum_{j=0}^{i} p_j)}}{e^{\sum_{i=l-s}^{l} log(i) - \sum_{i=k_0-s}^{k_0} log(i) + log(g_l) - log(g_{k_0}) + (l-k_0) log(1-\sum_{j=0}^{i} p_j)}} \tag{7}$$

```
model3 = function(N,P,gk=c("Poisson","Empirisch"), empVector, epsilon=1e-10){
   sumN <- sum(N)
   sumP <- sum(P)
```

```
sumDenom <- 0
sumNom <- 0
#start from 1 if sumN=0 or else from sumN
k <- max(1,sumN)

#function to calculate log(u_k)
logu = function(k){
   if(k == sumN){
      logu <- sum(log(1:k)) + log(g(k)) +k* log(1-sumP)
   } else{
      logu <- sum(log(1:k)) + log(g(k)) - sum(log((1:(k-sumN))))+k* log(1-sumP)
   }
}

#function to find k0
findkmax = function(stopconditie){
   maxlogu = -1e200;

   #calculate log(u_k) for all relevant k, ie the k values for which g_k has value
   repeat {
      log = logu(k)
      if(log > maxlogu){
         maxlogu <- log
         findkmax <- k
      }

      if(k >= stopconditie){
         if(log - maxlogu < (-1e10)){
            k <- max(1,sumN)
            break
         }
      }
      k <- k+1
   }
   findkmax
}

if(gk == "Poisson") {
   tau = mean(empVector)
   stopcondition <- tau
   g =function(k){
      dpois(k, tau)
      }
      kmax <-findkmax(stopcondition)
   }
   if(gk == "Empirisch"){
      stopcondition <- max(empVector)

      g =function(k){
         length(empVector[empVector==k])/length(empVector)
      }
      kmax <- findkmax(stopcondition)
   }

   repeat {
      sumk = sum(log(1:k))
      sumkmax = sum(log(1:kmax))
      if(k-sumN==0) {
         sumks =0
      } else {
         sumks = sum(log((1:(k-sumN))))
      }
      if(kmax-sumN<=0){
         sumkmaxs = 0
      } else {
         sumkmaxs = sum(log(1:(kmax-sumN)))
      }
      lastterm <- exp(sumk - sumkmax - (sumks - sumkmaxs)
      + log(g(k)) - log(g(kmax))+(k-kmax)*log(1-sumP))

      sumNom <- sumNom + k*lastterm
      sumDenom <- sumDenom + lastterm
      if(lastterm < epsilon)if(k>stopcondition) break
      k <- k+1
   }
   output <- sumNom/sumDenom
   output
}
```