Master Thesis in Statistics and Data Mining

Evaluate A/B-test in long term

- Understanding new feature adoption using Bayesian vector autoregression and Gompertz function

Marcus Bergdahl

Division of Statistics
Department of Computer and Information Science
Linköping University

Supervisor
Anders Nordgaard

Examiner
Mattias Villani

A statistician's wife had twins. He was delighted. He rang the minister who was also delighted. "Bring them to church on Sunday and we'll baptize them," said the minister. "No," replied the statistician. "Baptize one. We'll keep the other as a control."

Abstract

The thesis studies evaluates to make long term conclusions about which variant in an A/B-test is more appreciated by its users. Handling potential bias effects is also studied.

To study users' usage of Spotify, data from the *Thumbs*-test from 2013 was used. Users were exposed to different variants of the *Discover* page in which they could give feedback on music recommendation. The hypothesis was that the users would become more engaged and use Spotify more.

Bayesian *Vector Autoregression* (VAR) with non-informative priors was used to forecast users' usage of Spotify. With *Bayesian VAR* the *Markov Chain Monte Carlo* technique *Gibbs sampler* was used to make forecasts and compute prediction intervals. An approach to handling potential biases was tested, namely by using Bayesian VAR on the residuals obtained after fitting a *Gompertz function* to transformed forecasted values to the original time series. With this method, the bias would not have any negative effect when forecasting since the residuals ideally should be stationary.

The results show that using Bayesian VAR gave good forecasts both when the time series had a trend or did not have a trend in the latter part of the time series. Using Bayesian VAR on residuals obtained from fitting a Gompertz function to the time series was however a better approach when the time series was stationary in the latter part, but gave poor forecasts when the time series had a trend in the latter part.

Acknowledgments

To the people at the university, I want to especially thank my supervisor Anders Nordgaard for his excellent guidance and valuable comments in every stage of this project. I also want to thank my fast-talking opponent Uriel Chareca for his improvement suggestion and discussions about the thesis. Isak Hietala should also be thanked for proofreading the manuscript and for his help with LyX.

I would like to thank the ABBA-team at Spotify for providing the data and this interesting problem. Especially a big thank you to Danielle Jabin for guidance, proofreading the thesis and for her awesome Hadoop skills. Ali Sarrafi also deserves a big thank you for great support during the whole thesis.

Contents

1. Introduction

This chapter provides background information, the aim of the thesis and a summary of previous work in this area.

1.1. Background

At Spotify they take a data-driven approach to product development. When it is of interest to try a new feature or a new design on any of Spotify's devices, an A/B-test is performed to determine whether the feature or design is a success. With an A/B-test we perform controlled experiments where we randomly expose users to different variants (of the design or feature) of Spotify. One group, the control-group, is exposed to the standard variant which is used for comparison. In the other groups, users see other variants that the company hopes will be more appreciated by users. Those variants are called treatment groups and can be named A, B, C-group and so on, depending on how many treatment variants they choose to test. Then, after a limited time, often a few weeks, users' usage is analyzed to see how these treatments have performed relative to the control.

To determine whether a test is successful, a key metric is chosen, also called the response variable. At the end of a test, the metric is calculated for each variant relative to the control and checked for statistically significant differences between the group's metrics to determine whether the change, if any, is meaningful. There are also support metrics providing us with other information about the groups. This information is not used for evaluating which variant was the best one but is used to assess potentially side effects from a test. For each variant, the metrics are calculated from cross-selection studies each day the test is run i.e. every day during the test, new groups of users are studied. These groups are a mixture of newly registered users and previous users.

The metrics are different depending on if they measure *activation* or *retention*. Activation looks at how the user interacted with Spotify that particular date. Retention measures if the user comes back or if the user stopped using Spotify for a while after registrating an account.

1.2. Problem

There are several problems with making good conclusions about an A/B-test in the long-term when data is sampled only for a few weeks. Users' usage could be influenced by a status quo or primacy bias, which means it's possible that the reaction of users during a short period could be different from their reaction in the long run for better or for worse. Among other things, a natural resistance to change could cause previous users to initially be less enthusiastic about a new design but over time they could grow accustomed to it and ultimately like it more. Alternatively, a new feature could appear as being a great success initially, however users could simply be curious to explore the new feature and may eventually get bored of it and move on, resulting in a false positive. For new users another type of bias could be present in the sense that when they are new on Spotify, new users use it perhaps quite much in the beginning and are very enthusiastic but after a while the enthusiasm wears off. On the other hand, it could take a while for some new users to realize that they like using Spotify and start using it more often.

That metrics are from cross-selection studies is also a problem since we want to see how user usage looks during the test, and since we don't study the same group of people during the test we don't really know if changes depend on what we are testing or what group of people was studied that particular day. The mix of newly registered and previous users in a test could produce a bias in the test metrics as they are not necessarily under the same conditions. This since newly registered users has not seen any previous version of Spotify like previous users has, and that new users are potentially influenced by another type of bias then previous users are. Also, since the next date in the metrics includes a new mix of new and previous users, it is important that the percentage size of new and previous user is approximately the same for all the dates.

1.3. Objective

The objective of this thesis is to find a statistical methodology to evaluate A/B-test in the long term. To succeed with this we will also study the effect from potential bias and how it could be treated.

1.4. Previous work

We will use information about users' usage from multiple dynamic time series in which bias of various kinds can be expected to be present. Research in the area of forecasting users' usage in social media is not very developed. Moreover, there is not any previous work in forecasting users' usage in A/B-test. General methods for

multiple dynamic time series can however be found in econometrics. Examples of models that proved successful for dynamic forecasting is given, among other things Sims (1980) argued that *Vector Autoregression* (VAR) gave a credible approach for describing data and forecasting. However, several models are in its nature complex and require relatively long time series to estimate efficiently. Therefore, the Bayesian approach VAR was discussed in several articles (see e.g. Doan et al. (1984); Litterman (1986)).

In the area of psychology longitudinal studies are often used to study trends in an individual's life. Longitudinal studies are preferred over cross-sectional studies in which different individuals with the same characteristics are compared. Longitudinal studies evaluate the same group of people over a period of time. Any potential difference between people in the group is less likely to be the result of cultural differences across generations that could appear in cross-sectional studies (Carlson et al., 2009). So in this thesis, instead of studying previously used metrics to make long term conclusions it is better to study a group of users in a longitudinal study during the test. At Spotify all data on user interaction is stored which makes it possible to extract longitudinal user-level data on a daily basis.

2. Data

This chapter describes a previous test and the data for this test that will be used in this thesis.

2.1. Thumbs-test

To find the best approach and best statistical methods, it's important to see what users' usage pattern look like, therefore data from a previous test from 2013 will be used. The test was called *Thumbs*, and the hypothesis was that users wanted to give feedback on music recommendations made in the *Discover* page with the end goal that they will become more engaged useing Spotify. In Figure 2.1 the different variants are presented.

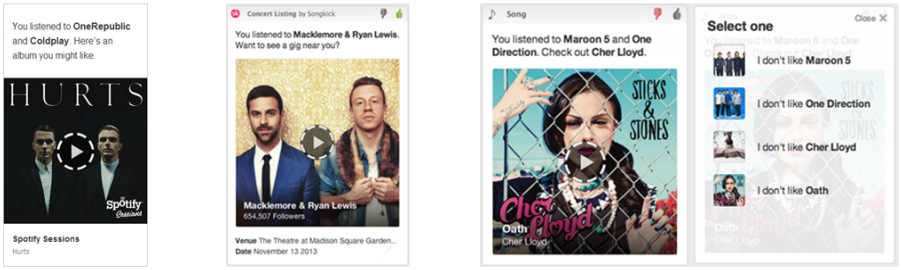

Figure 2.1.: The different variants in the Thumbs-test

The left-most picture shows the variant exposed to the control group in which a user can't give feedback on recommendations. The next picture shows the variant for the first treatment group that was exposed to the thumbs feature, for which the user could tell if they liked the recommendation or not. The third and fourth pictures from the left show the variant for the second treatment group that received the thumbs feature. With this variant you could in addition give a reason for why you didn't like the recommendation.

2.2. User-level data

Instead of mixing previous and new users, only a group of new users that registered an account on 30 July 2013 were studied in this thesis in the time period between

31 July 2013 to 31 August 2013. Since some users remove their accounts, users that did not keep their account during the whole test duration were discarded. There was 33 339 new users in the test that did not remove their accounts, among them 30 038 users was in the *control*-group, 1 694 in *A: Thumbs*-group and 1 607 in *B: Thumbs and Reason*-group. The variables used were mostly boolean variables like *Daily Active User* (DAU) which tells us if the user was online that particular day, *Weekly Active User* (WAU) which shows if the user had been online at any time the last week, *Premium User* which tells if the user is a premium user and the last variable used for this test was *Number of streams* which is a continuous variable with information about how many streams the user listened to that particular day. To summarize this data for each group, data was aggregated by *date* as well as by *test*-group with calculated variables presented in table Table 2.1.

Table 2.1.: Variables

Variable name	**Description**
Percentage of DAU	Number of DAU in the group divided by number of users in the group
Average streams per DAU	The group's total number of streams divided with the group's number of DAU
Percentage of WAU	Number of WAU in the group divided by number of users in the group
Percentage of premium users	Number of premium users in the group divided by number of users in the group

The variable *Percentage of DAU* is the response variable that will be used to evaluate the test, and the other time series will work as explanatory variables in modeling the response variable. The result from the aggregation is presented in Figure 2.2.

We can see that the time series for each group are very similar, so the thumbs feature is probably not that big of a change to change the group's user usage of Spotify. We can also see - what looks like a bias - higher values in the beginning of the test run in *Percentage of DAU*, *Average streams per DAU* and *Percentage of WAU*. Variables *Percentage of DAU* and *Average streams per DAU* has seasonal effects since users use Spotify less during the weekends. The variable *Percentage of premium users* does not increase until at the very end, probably because users receive a premium account for free the first month, making this variable extremely non-representative and useless in a model when analyzing data for new users.

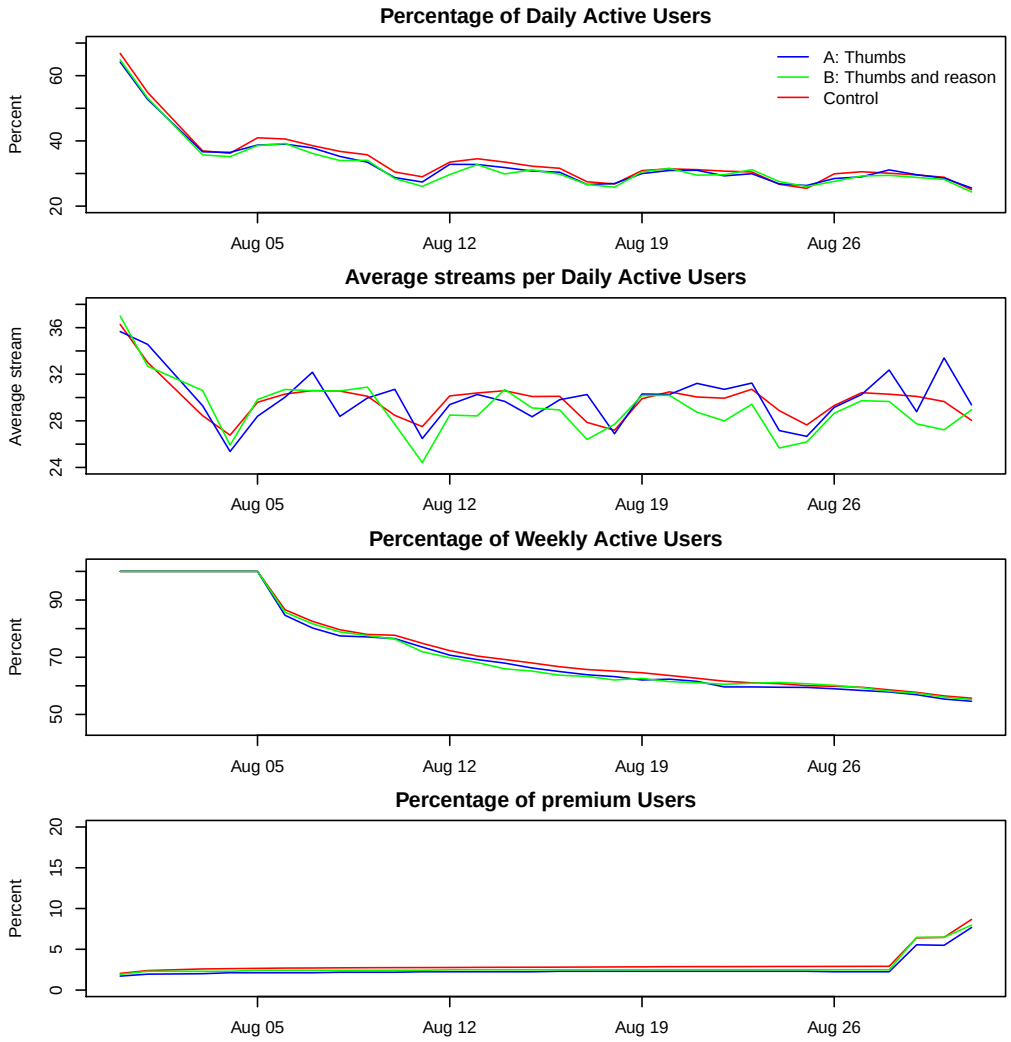

Figure 2.2.: Time series for each group in the Thumbs-test

3. Methods

This chapter will explain Bayesian inference and forecasting, Gibbs sampling, Vector Autoregressions both with the frequentistic and Bayesian approach and the Gompertz function.

3.1. Bayesian inference and forecasting

Information about Bayesian inference and forecasting for this chapter has been collected from Hastie et al. (2009); Karlsson (2012); Petris et al. (2009). It is very unusual to have perfect information about data in the real world. Even if a deterministic model describes data in an accurate way, there are always uncertainties in the model like the effects of unobserved variables and measurement errors that are not within our control. A keystone in Bayesian statistics is to describe parameters and observation predictions in terms of probability distributions. In doing so the uncertainty is also present in the model which gives us an easier interpretation of the result.

3.1.1. Bayesian inference

For inference in a Bayesian approach, a sampling model is specified in terms of a *Likelihood*

$$L(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{3.1}$$

for the unknown parameter θ with the given data $x_1, ..., x_n$ based on $f(x_i|\theta)$ which is a density function for one observation in the sample, and the knowledge about the parameter before we see the data is specified with a prior distribution $\pi(\theta)$ for θ. A posterior distribution for θ is then calculated as

$$q(\theta|x_1, ..., x_n) = \frac{L(x_1, ..., x_n|\theta)\pi(\theta)}{\int L(x_1, ..., x_n|\theta)\pi(\theta)\, d\theta} \tag{3.2}$$

which is our updated knowledge about θ after we have seen the data. The posterior distribution is also used to construct the predictive distribution

$$p(x_{n+1}|x_1, ..., x_n) = \int f(x_{n+1}|\theta)q(\theta|x_1, ..., x_n)\, d\theta \tag{3.3}$$

for the value of a future observation. The main difference between the frequentist and Bayesian inference is that in frequentist inference the unknown parameters are considered to be fixed and the data to be random. In Bayesian inference, data is considered to be fixed and parameters to be random because they are unknowns, meaning that we describe the uncertainty for the parameters with probability distributions that update with data. In multivariate time series we often work with forecasting more than one lead which will be explored in the next chapter.

3.1.2. Bayesian forecasting

In time series the predictive distribution $p\left(y_{T+1:T+H}|Y_T\right)$ of future time points $y_{T+1:T+H}$ conditional on observed data Y_T, is the significant object in Bayesian forecasting since it includes information about the unknown future. For the actual forecast we can choose between taking the posterior mean, median or mode together with a probability interval indicating the range of likely outcomes. In this thesis the median will represent the actual forecast since it handles outliers well in the Gibbs sampling (see further below).

To construct the predictive distribution we need the specification of three distributions: First, $f\left(y_{T+1:T+H}|Y_t, \theta\right)$ which is the distribution of the future time point values conditional on observed data and unknown parameters θ. Second we need the likelihood $L\left(Y_T|\theta\right)$ for the unknown parameters with the available data. The third distribution needed is the prior knowledge about the unknown parameters θ called the prior distribution $\pi\left(\theta\right)$. The likelihood takes the form as in equation 3.4, and the distribution on future time points is of the form as described in equation 3.5. With these we can then construct the predictive distribution in equation 3.6

$$L\left(Y_T|\theta\right) = \prod_{t=1}^{T} f\left(y_t|Y_{t-1}, \theta\right) \tag{3.4}$$

$$f\left(y_{T+1:T+H}|Y_T, \theta\right) = \prod_{t=T+1}^{T+H} f\left(y_t|Y_{t-1}, \theta\right) \tag{3.5}$$

$$p\left(y_{T+1:T+H}|Y_T\right) = \frac{\int f\left(y_{T+1:T+H}|Y_T, \theta\right) L\left(Y_T|\theta\right) \pi\left(\theta\right)\, d\theta}{\int L\left(Y_T|\theta\right) \pi\left(\theta\right)\, d\theta}. \tag{3.6}$$

In practice Bayes theorem in equation 3.7 is used as an intermediate step to get the posterior distribution for the unknown parameters θ conditional on Y_T,

$$q\left(\theta|Y_T\right) = \frac{L\left(Y_T|\theta\right)\pi\left(\theta\right)}{\int L\left(Y_T|\theta\right)\pi\left(\theta\right)\, d\theta} \propto L\left(Y_T|\theta\right)\pi\left(\theta\right) \tag{3.7}$$

where $\int L\left(Y_T|\theta\right)\pi\left(\theta\right)\, d\theta$ is a constant term not affected by θ. We simply say that $q\left(\theta|Y_T\right)$ is proportional to $L\left(Y_T|\theta\right)\pi\left(\theta\right)$. Now with the posterior distribution the predictive distribution could be constructed as

$$p\left(y_{T+1:T+H}|Y_T\right) = \int f\left(y_{T+1:T+H}|Y_T,\theta\right) q\left(\theta|Y_T\right)\, d\theta. \tag{3.8}$$

In equation 3.8 we can observe that the predictive distribution takes into consideration the uncertainty in future time points with $f(y_{T+1:T+H}|Y_T,\theta)$ and the uncertainty about the true parameter value in $q\left(\theta|Y_T\right)$.

In equation 3.8 marginalizing out the parameters from the multivariate distribution $p(\theta, y_{T+1:T+H}|Y_T) = f(y_{T+1:T+H}|Y_T)q(\theta|Y_T)$ is very difficult. Therefore a simulation technique for the marginalization can be used in which we can discard the draws of θ from the multivariate distribution $(\theta, y_{T+1:T+H})$ and it will be as if they were sampled from a distribution without θ. To do that, say that we simulate θ R times from its posterior distribution $q(\theta|Y_T)$, then we can for each draw also generate a sequence of draws of $y_{T+1}, ..., y_{T+H}$ by repeatedly drawing from $f(y_t|Y_{T-1},\theta)$ and adding the draws of y_t to the conditioning set for the distribution of y_{t+1}. The sample from the predictive distribution for $y_{T+1:T+H}$ can be used to estimate $E(y_{T+1:T+H}|Y_T)$.

3.2. Gibbs sampling

Gibbs sampling presented in Algorithm 3.1 is a *Markov Chain Monte Carlo* (MCMC) method described in several books and articles (see e.g. James (2002); Nakajima and Ginkō (2011); Karlsson (2012); Doh and Connolly (2013)) and is used when generating samples from a multivariate distribution. We generate samples from a full conditional posterior distribution $q(\theta_i|Y_T,\theta_{-i})$, and it will have the same effect as generating them directly from the multivariate posterior distribution $q(\theta_1, ..., \theta_d)$.

Algorithm 3.1 Gibbs sampling

1. Set $j = 0$ and choose arbitrary initial values for parameter $\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_d^{(0)}$.
2. Generate $\theta_1^{(j+1)}$ conditionally on $\theta_2^{(j)}, \theta_3^{(j)}, ..., \theta_d^{(j)}$
   Generate $\theta_2^{(j+1)}$ conditionally on $\theta_1^{(j+1)}, \theta_3^{(j)}, ..., \theta_d^{(j)}$
   ...
   Generate $\theta_d^{(j+1)}$ conditionally on $\theta_1^{(j+1)}, \theta_2^{(j+1)}, ..., \theta_{d-1}^{(j+1)}$
3. Set $j = j + 1$ and repeat step 2 R times.

It is possible that autocorrelation between $\theta^{(j-1)}$ and $\theta^{(j)}$ is present in the result from Gibbs sampling. Therefore using the strategy to only retain every k:th draw, autocorrelation between $\theta^{(j-k)}$ and $\theta^{(j)}$ is negligible. It is then possible to estimate the posterior mode by taking the sample median of the retained values.

3.2.1. Geweke diagnostic

To test that the output from the Gibbs sampler has converged and is stationary, Geweke et al. (1991) proposed a convergence diagnostic for Markov chains called *Geweke diagnostic* that tests for equality of the means of the first 10 percent and last 50 percent of a Markov chain. If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke's statistic has an asymptotic standard normal distribution. The test statistic is a standard Z-score that should be less than ± 1.96 if the samples from Gibbs sampling has converged and is stationary if we test with a 95 percent confidence. If Z-score are higher than ± 1.96 samples from the first iterations could be discarded until only converged samples remain.

3.3. Vector autoregression

Zivot and Wang (2007) and Karlsson (2012) which describe *Vector Autoregression* (VAR) state that it's a method for dynamic multivariate time series which is a natural extension of the univariate autoregressive model and is an easy, successful and flexible model to use. VAR is proven to be useful in describing dynamic behavior and forecasting time series. Forecasting with VAR has quite flexible since the model can be made conditional on the potential future paths of specified variables in the model. The basic p-lag VAR(p) model has the form

$$\begin{aligned} y_t' &= x_t'C + A_1 y_{\mathbf{t-1}}' + ... + A_p y_{\mathbf{t-p}}' + u_{\mathbf{t}}' \\ &= z_t'\Gamma + u_t' \end{aligned} \tag{3.9}$$

where $y_t' = (y_{1t}, ..., y_{jt}, ..., y_{mt})$ is a (1 x m) vector of time series variables at time point t, $t = 1$,..., T, x_t' is a (1 x d) vector where d is the number of deterministic variables, C is a (d x m) matrix with coefficients, A_i is a (m x m) coefficient matrix for $i = 1$, ..., p and the so-called white noise $u_{\mathbf{t}}'$ is a (1 x m) vector which is normally distributed with a zero mean and covariance matrix Ψ. $z_t' = \left(x_t', y_{t-1}', ..., y_{t-p}'\right)$ is a (1 x k) vector where $k = mp + d$. Finally Γ contains the coefficients stacked in a (k x m) matrix $\left(C', A_i', ..., A_p'\right)'$.

We estimate the coefficients from equation 3.9 by using the assumption that the model for VAR(p) is covariance stationary and that there are no restrictions on the parameters. Each j:th univariate time series can be written on the form

$$Y_j = Z\Gamma_j + U_j \tag{3.10}$$

for $j = 1$, ..., m, where Y_j is a (T x 1) vector with the j:th time series observations, Z is a (T x k) matrix with its t:th row given by $z_t = \left(x_t', y_{t-1}', ..., y_{t-p}'\right)$, Γ_j is a (k x 1) coefficient vector and U_j is a (T x 1) vector of errors with covariance matrix $\sigma_j^2 I_T$. Since the VAR(p) in equation 3.10 is in the form of *seemingly unrelated regressions* (SUR), a frequentistic approach would be to estimate each equation separately by *ordinary least squares* (OLS) $\Gamma_j = (Z'Z)^{-\mathbf{1}} Z'Y_j$. For the multivariate model

$$Y = Z\Gamma + U \tag{3.11}$$

the coefficients can then be estimated by $\Gamma = (Z'Z)^{-\mathbf{1}} Z'Y$, and hence $f(y_t|Y_{t-1}, \theta)$ will be equal to $N(y_t; z_t\Gamma, \Psi)$.

3.3.1. Bayesian vector autoregression

The are several reasons why we should use a Bayesian approach when estimating coefficients in the VAR model. According to Zivot and Wang (2007), unrestricted estimation of the coefficients in a VAR model requires a lot of data. Karlsson (2012) states that because of rich parameterization the VAR model is flexible and fits data well, but it comes with the risk of overfitting the data which could lead to imprecise inference and a larger uncertainty when predicting future paths. The Bayesian

approach provides the optimal way of combining the data with prior beliefs for θ which gives us a sharper inference and more precise forecasts (Karlsson, 2012).

In Bayesian VAR when the data is stacked like in equation 3.11 the likelihood is of the form

$$L(Y|\Gamma,\Psi) = (2\pi)^{-mT/2} exp\left(-\frac{1}{2} tr\left[\Psi^{-1}(Y - Z\Gamma)'(Y - Z\Gamma)\right]\right). \tag{3.12}$$

Using non-information priors, a uniform distribution for Γ and an objective Jeffreys' prior for Ψ is chosen (Karlsson, 2012). We then have the joint prior distribution of the form

$$\pi(\Gamma,\Psi) = |\Psi|^{-(m+1)/2}. \tag{3.13}$$

To construct the joint posterior for Γ and Ψ we multiply the prior with the likelihood, just like in equation 3.7 and receive the form

$$\begin{aligned} q(\Gamma,\Psi|Y_T) \quad \propto \quad & |\Psi|^{-T/2} exp\left(-\frac{1}{2} tr\left[\Psi^{-1}(Y - Z\hat{\Gamma})'\left(Y - Z\hat{\Gamma}\right)\right]\right) \\ & \times exp\left(-\frac{1}{2} tr\left[\Psi^{-1}\left(\Gamma - \hat{\Gamma}\right) Z'Z\left(\Gamma - \hat{\Gamma}\right)\right]\right) |\Psi|^{-(m+1)/2} \end{aligned} \tag{3.14}$$

which is referred as a *Normal-Wishart* distribution. To estimate $E(y_{T+1:T+H}|Y_T)$ from the predictive distribution we will use a straightforward procedure by using Gibbs sampling to simulate first from the conditional distribution of the parameters and then recursively calculate y_{T+h} for $h = 1, ..., H$. To get the conditional distribution for Γ, Karlsson (2012) shows that if we focus on $tr\left[\Psi^{-1}\left(\Gamma - \hat{\Gamma}\right) Z'Z\left(\Gamma - \hat{\Gamma}\right)\right]$ from equation 3.14, we can then see that it's equal to $(\gamma - \hat{\gamma})'(\Psi^{-1} \otimes Zt Z)(\gamma - \hat{\gamma})$, where $\gamma = vec(\Gamma)$ and $\hat{\gamma} = vec(\hat{\Gamma}) = ([I_m \otimes Z'Z]^{-1} Z')\, y$. Then γ is multivariate normal distributed depending on Ψ, which gives us the conditional distribution

$$\gamma|Y_T,\Psi \sim N\left(\hat{\gamma}, \Psi \otimes [Z'Z]^{-1}\right).$$

Integrating out γ from the joint posterior in equation 3.14 we get the marginal posterior distribution for Ψ as

$$q(\Psi|Y_T) \propto |\Psi|^{-(T+m+1-k)/2} exp\left(-\frac{1}{2} tr\left[\Psi^{-1}S\right]\right)$$

where $S = \left(Y - Z\hat{\Gamma}\right)' \left(Y - Z\hat{\Gamma}\right)$. This is an *inverse Wishart* distribution $\Psi \sim iW\left(S, T-k\right)$ where T-k is the degrees of freedom. With this information and that $u_t \sim N(0, \Psi)$ we can use Algorithm 3.2 below to estimate $E(y_{T+1:T+H}|Y_T)$.

Algorithm 3.2 Simulating the predictive distribution with a Normal-Wishart distribution

For $j = 1, ..., R$

1. Generate $\Psi^{(j)}$ from the marginal distribution

$$\Psi|Y_T \sim iW\left(S, T-k\right)$$

2. Generate $\gamma^{(j)}$ from the conditional posterior

$$\gamma|Y_T, \Psi^{(j)} \sim N\left(\hat{\gamma}, \left(\Psi^{(j)} \otimes Z'Z\right)^{-1}\right)$$

3. Generate $u_{T+1}^{(j)}, ..., u_{T+H}^{(j)}$ from

$$u_t \sim N\left(0, \Psi^{(j)}\right)$$

4. Calculate for each $h = 1, ..., H$

$$y'_{T+h} = x'_{T+h}C \; + \; A_1 y'_{T+h-1} + ... + A_p y'_{T+h-p} + u'_{T+h}$$

where we use predicted values as observed value when $h > 1$.

After using Algorithm 3.2 the samples for Γ and Ψ are discarded and the calculated sample for y_{T+h} are kept. We can then easily estimate $E\left(y_{T+1:T+H}|Y_T\right)$ by taking the sample median for each y_{T+h}. A prediction interval is an estimate of an interval in which future observations will fall, with a certain probability, given what has already been observed. This will be used to check if there are any statistically significant differences between groups users' usage forecast. To receive a two-sided prediction interval with 95 percent credibility interval for $E\left(y_{T+1:T+H}|Y_T\right)$ from Gibbs sampling we order the samples for y_{T+h}, then we can obtained the lower and upper prediction interval limit by taking the value for ranks $(R/k) * 0.025$ and $(R/k) * 0.975$ respectively, where k is the k:th draw from the sample that are retained.

In *Appendix A* functions to perform Bayesian VAR are given. The function *setup* (A.2) gives the right matrices and information to be put into function *BVAR_forecast* (A.3) which returns a list with all the samples of the predicted values. In A.1 the function *weekd* is given and being used in function *setup* and *BVAR_forecast* that return the right values for the seasonal dummies.

3.4. Gompertz function

The Gompertz function (see an example of the function in Figure 3.1) is a mathematical function that lies between two asymptotes and has the properties that the growth is slowest at the start and the end of an time period and is, unlike the logistic function, non-radial symmetric (Vieira and Hoffmann, 1977)

$$y(t) = a * exp\left(b * exp\left(c * t\right)\right). \tag{3.15}$$

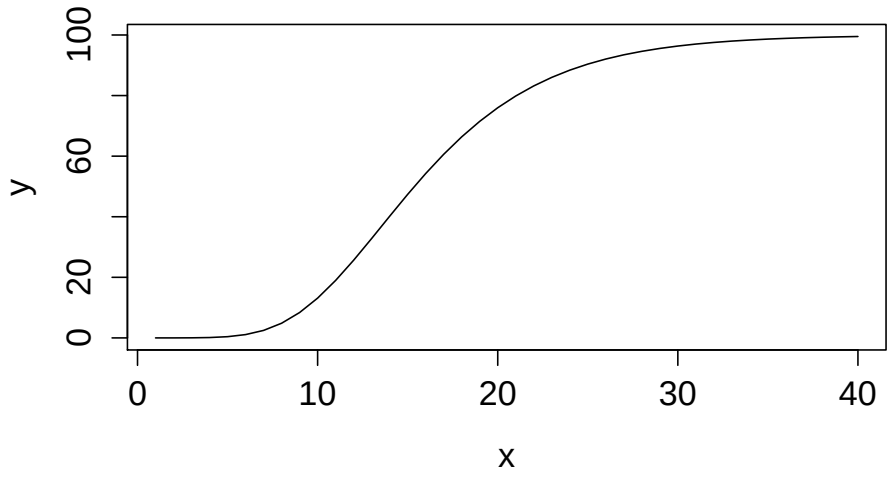

Figure 3.1.: Example of a Gompertz function

Therefore, a Gompertz function will be used to fit the mean value function of some time series (Figure 3.2, plot 1) and then calculate the residuals which will ideally be stationary (i.e. the values have approximately the same mean value), and then forecast on the residuals (Figure 3.2, plot 2). With this approach potential bias in time series will not affect the estimation of parameters, and will still capture the time series trend by adding the residual's forecasted values to the functions value at time point $T + h$ (Figure 3.2, plot 3).

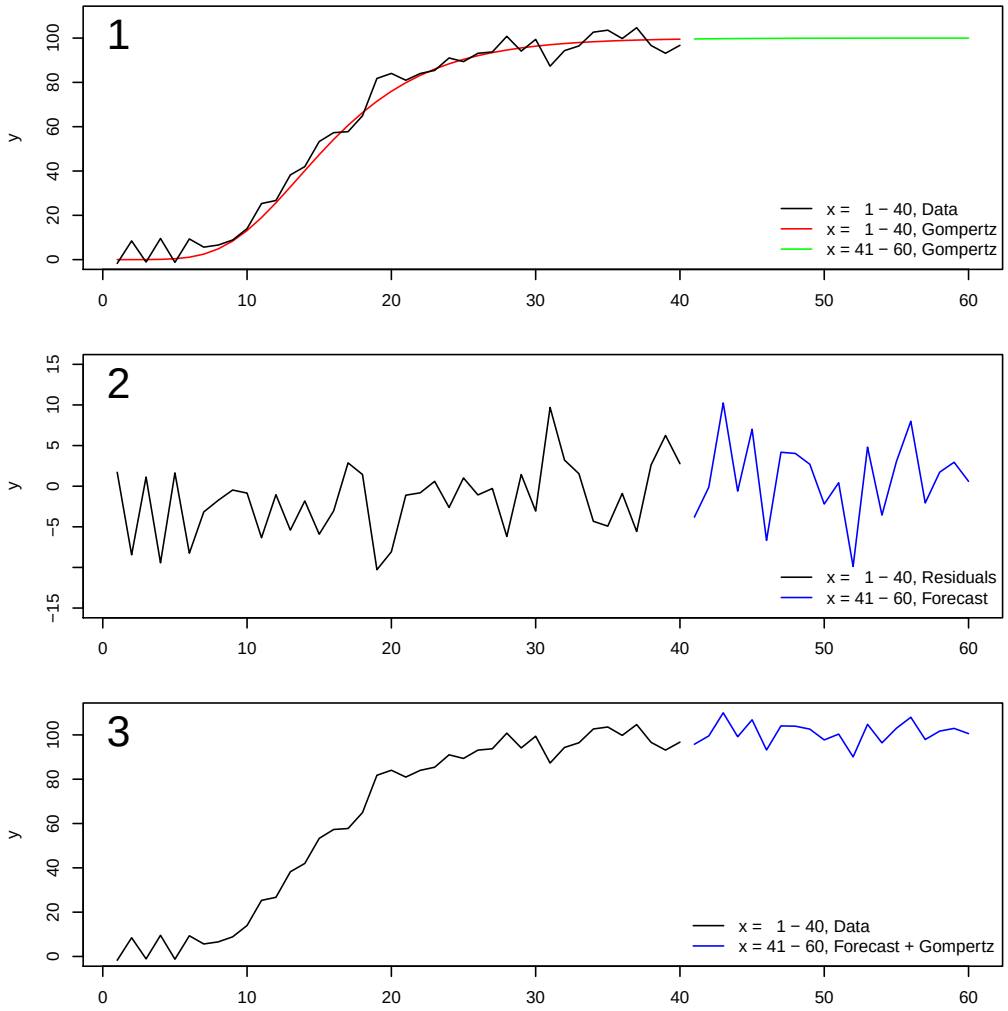

Figure 3.2.: Explanation on how Gompertz will be applied

To estimate the constant parameters a, b and c, function *nls* (nonlinear Least squares) from the *stats* package in the statistical software R is used.

4. Result

This chapter will show the result when a standard Bayesian VAR approach is used on multiple time series original values. It will also present results from using a adjusted Bayesian VAR approach to forecast residuals obtained after fitting a Gompertz function. At the end of the chapter the two approaches forecasting abilities will be evaluated by using less time points and comparing them to the time series true values.

The time series variables chosen are *Percentage of DAU*, *Percentage of WAU* and *Mean streams per DAU*. The variable *Percentage of premium users* will not be used because it is non-representative of its true value on new users since they have a free month with a premium account. The number of lags in the Bayesian VAR model was chosen to be one because of the small amount of data and after studying the time series autocorrelation plots. When using all the data to forecast, the number of forecasts will be 21 days with a caveat that the forecasting accuracy is very uncertain for such a long forecast. To forecast one week could be a better choice, but with 21 days the forecasted values trend is easier to study.

When running the Gibbs sampler, the number of iterations (R) will be 10 000. Since the samples from each iteration are independent from each other, which can be seen in Algorithm 3.2 where the parameters S and $\hat{\gamma}$ are never updated, all the samples will be used and the strategy to only retain every k:th draw discussed in chapter Gibbs sampling will not be applied.

4.1. Forecasting non-stationary time series variables

In this section a standard approach is used to perform a forecast with Bayesian VAR on the selected time series from Figure 2.2.

The samples for each y_{T+h} for all three groups are normally distributed with some heavy outliers for the later time points. Performing a Geweke diagnostic it was found that for some of the later time points the Z-score was higher than ± 1.96. Studying the iterations in a time series shows that the samples are stationary through the whole series and there are some heavy outliers. The outliers affect the means in Gewekes diagnostic enough to get the result that the samples are not converged and stationary. Since $E(y_{T+1:T+H}|Y_T)$ and prediction intervals are estimated by the median and percentiles the outliers are too few to affect the estimates. When applying

Bayesian VAR on the original non-stationary time series the $E(y_{T+1:T+H}|Y_T)$ can be studied in Figure 4.1.

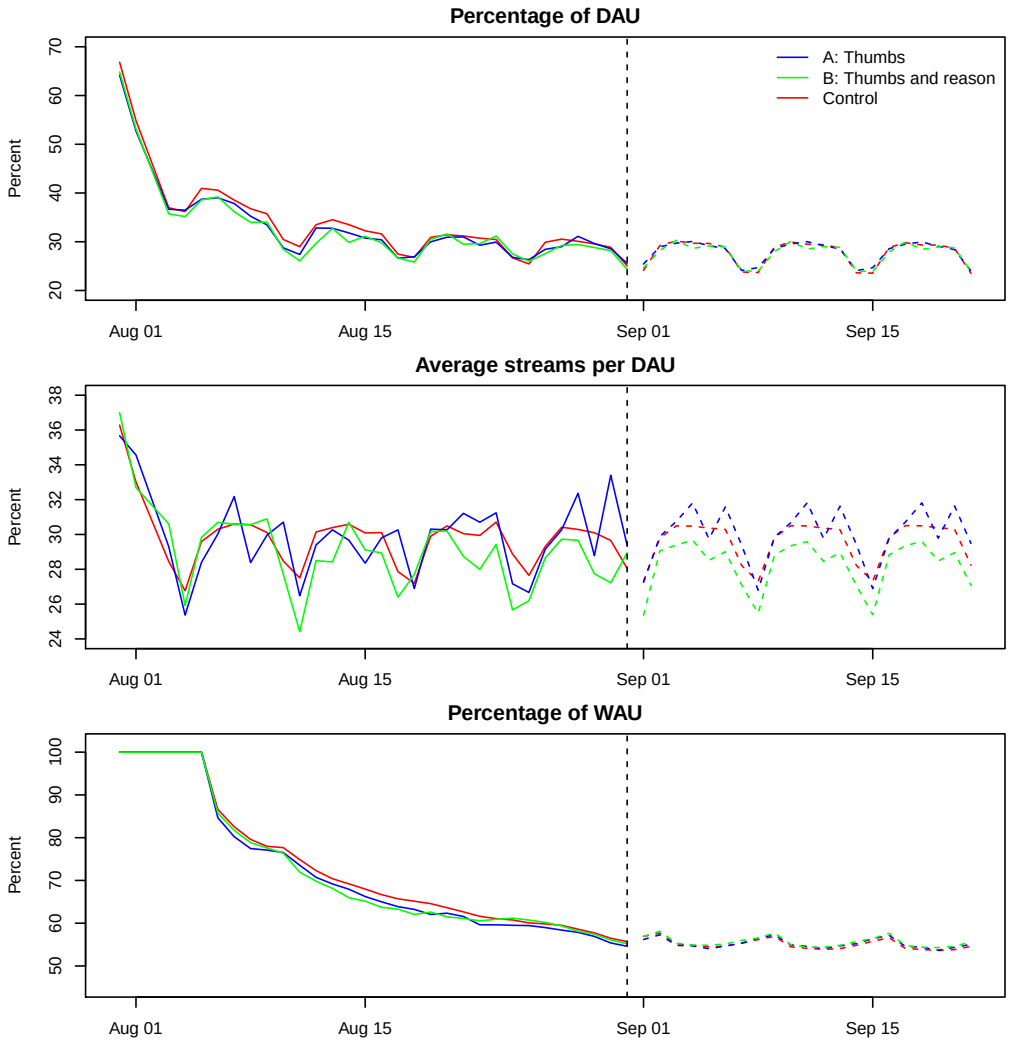

Figure 4.1.: 21 days forecast when using all data

In Figure 4.1 we see that the test groups time series behave very similar to each other for this test. Perhaps group *B: Thumbs and Reason* listen to less streams according to the forecasts in *Average streams per DAU*. To study if there are any differences between predictions for each group we study prediction intervals in Figure 4.2 for each group and time series. We also notice that the seasonal dummies in the model have an effect when forecasting *Percentage of WAU*.

It is not a big surprise that the prediction intervals for the groups cover each other in all the time series in Figure 4.2 since they are so alike each other in Figure 4.1. Because there are no differences between the groups looking at *Percentage of DAU* which we evaluate the test on, the treatment variants are neither less or more ap-

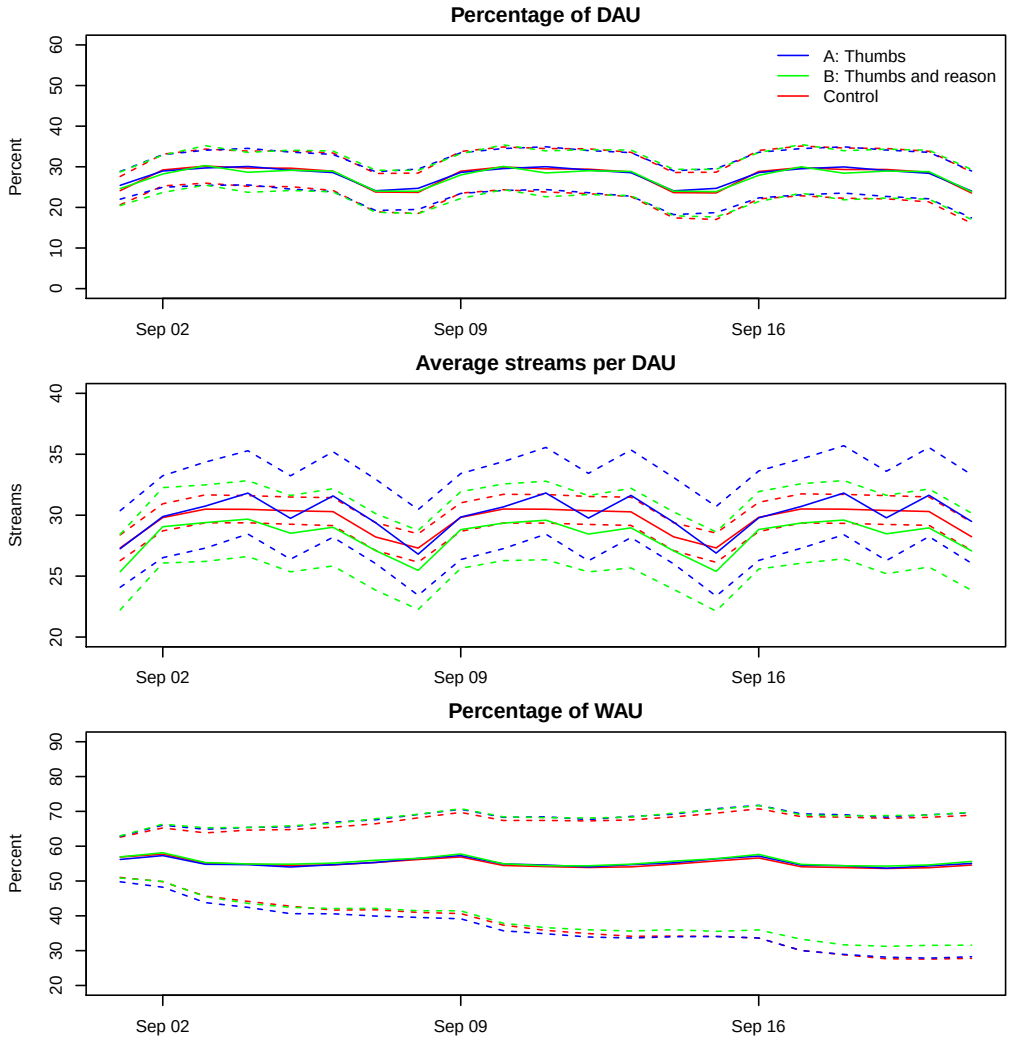

Figure 4.2.: Prediction intervals for the forecasted values when using all data

preciated by the new users. There is neither any difference between the groups in any of the other time series forecast.

4.2. Forecasting stationary time series variables

In the previous chapter we tried to make a forecast without handling any bias. In this chapter we will use an adjusted Bayesian VAR approach, in which we first apply a Gompertz function to each time series so residuals will hopefully have a mean value close to zero (see Figure 3.2). Since seasonal patterns are of interest no test for stationary and independence will be used.

To make use of the Gompertz function property that it is non-radial symmetric as the curve is increasing, some time series must be transformed since they are decreasing

(see Figure 2.2). The property that it is non-radial symmetric is important to use since the time series, and especially *Percentage of WAU*, has a slow start in the beginning which is not radial symmetric with the slow finish of the time series. The Gompertz function could be decreasing but then the properties that its growth is slower in the beginning and at the end of the time series would not work well. The function would then instead have a lower decay in the beginning and at the end when the function is decreasing.

Studying Figure 2.2 we see that time series *Percentage of DAU* and *Percentage of WAU* looks to follow a Gompertz function, but decreasing instead of increasing as Gompertz function does. Therefore these time series (y) will be transformed (x) by taking $x = 100 - y$ since they are in percentage form. After the actual forecast and estimations these time series will will be transformed back by $y = 100 - x$. The time series *Average streams per DAU* is not that much affected by the bias and a decreasing Gompertz function will be fitted to the time series. This since we do not need the property that the Gompertz function is slow in the beginning and in the finish for this specific time series. Plots for the fitted Gompertz curve and the residuals are presented in Appendix B. The forecasts for the residuals with Bayesian VAR can be seen in Figure 4.3.

At the end of *Percentage of WAU* there is an increase, but the forecasted values keep being around level zero. We note that an incorrect seasonal pattern is present in the forecast for *Percentage of WAU*, since *Percentage of WAU* checks if the users been online at any point the last seven days and should therefor have no seasonal trend. To transform the residual forecast to the original time series, we add the Gompertz function to the forecast at time point $T + h$. For *Percentage of DAU* and *Percentage of WAU* we transform them back to their decreasing trend upon adding the Gompertz function values. In Figure 4.4 the forecasts for the original time series are shown when they were transformed back.

In Figure 4.4 the time series for each group behave very similarly to each other, just like in Figure 4.1. Comparing these two plots to each other, we see that in Figure 4.1 the time series are slightly decreasing, which they are not in Figure 4.4. Also for this forecast we will study if there are any difference between the groups in the forecasts, so we take a look at the prediction intervals in Figure 4.5 for each group and time series. However, since the prediction intervals here are based on the forecasts on the residuals, we cannot compare these intervals with those in Figure 4.2.

The predictions intervals overlay in *Percentage of DAU* in which we evaluate the test from, meaning that the treatment variants are neither less or more appreciated by the new users. There is neither any difference between the groups in the other two time series.

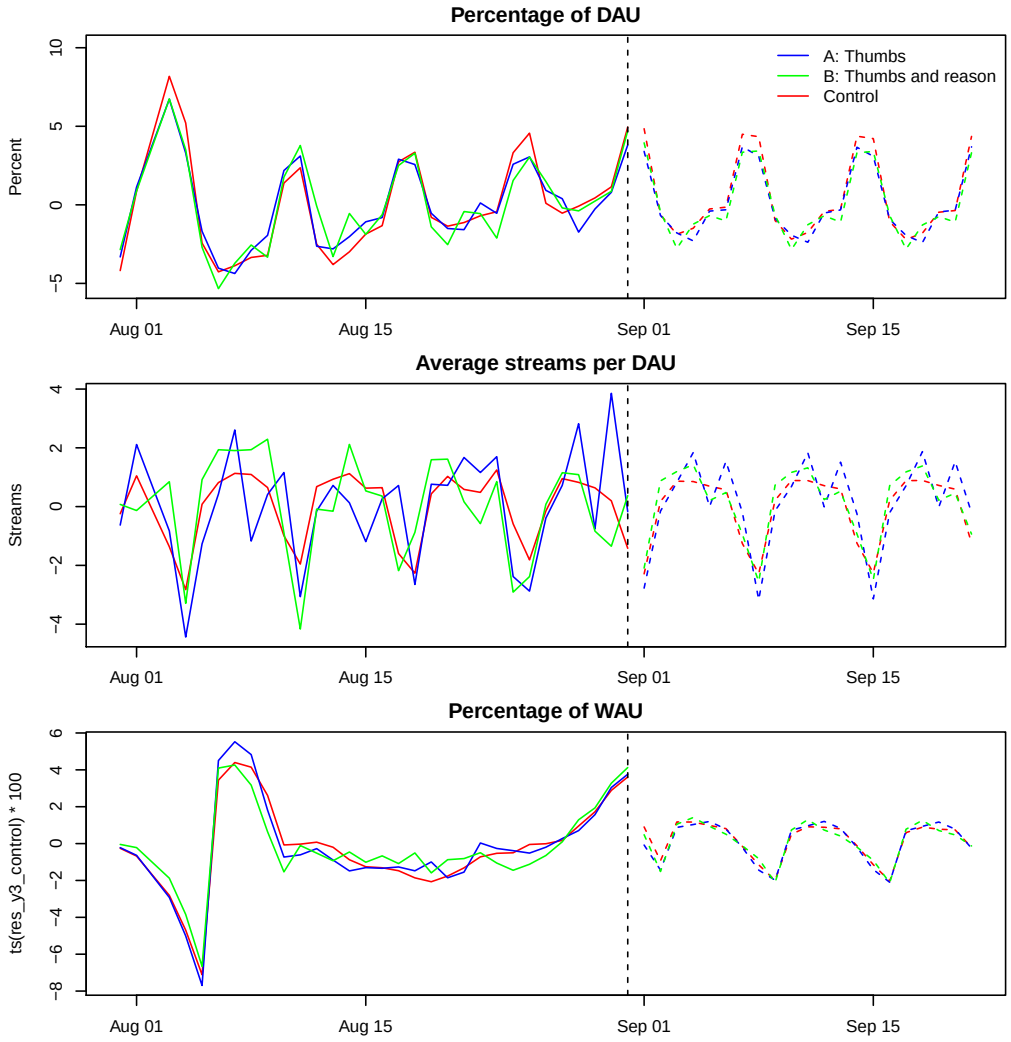

Figure 4.3.: 21 days residuals forecasts when using all data

4.3. Compare approaches on shorter time series

Since an A/B-test could be performed in less than one month, this chapter will evaluate which approach is the better one when we have less data, either the standard or the adjusted Bayesian VAR approach. Here, the first 15 and 25 days time points in the time series will be used, so the forecast can be compared with the rest of the actual true values. For simplicity, only the time series from the *control*-group will be studied since the groups are so similar to each other.

First we study the first 25 days, i.e. data from the time period 31 July 2013 to 24 August 2013. Thus, the forecasts will be for the next seven days, 25 August to 31 August, and will be compared against the true values in the same period. The values of $E(y_{T+1:T+H}|Y_T)$ for both approaches are presented in Figure 4.6. In Table 4.1 we give the calculated *Mean Squared Prediction Error* (MSPE) for the approaches for

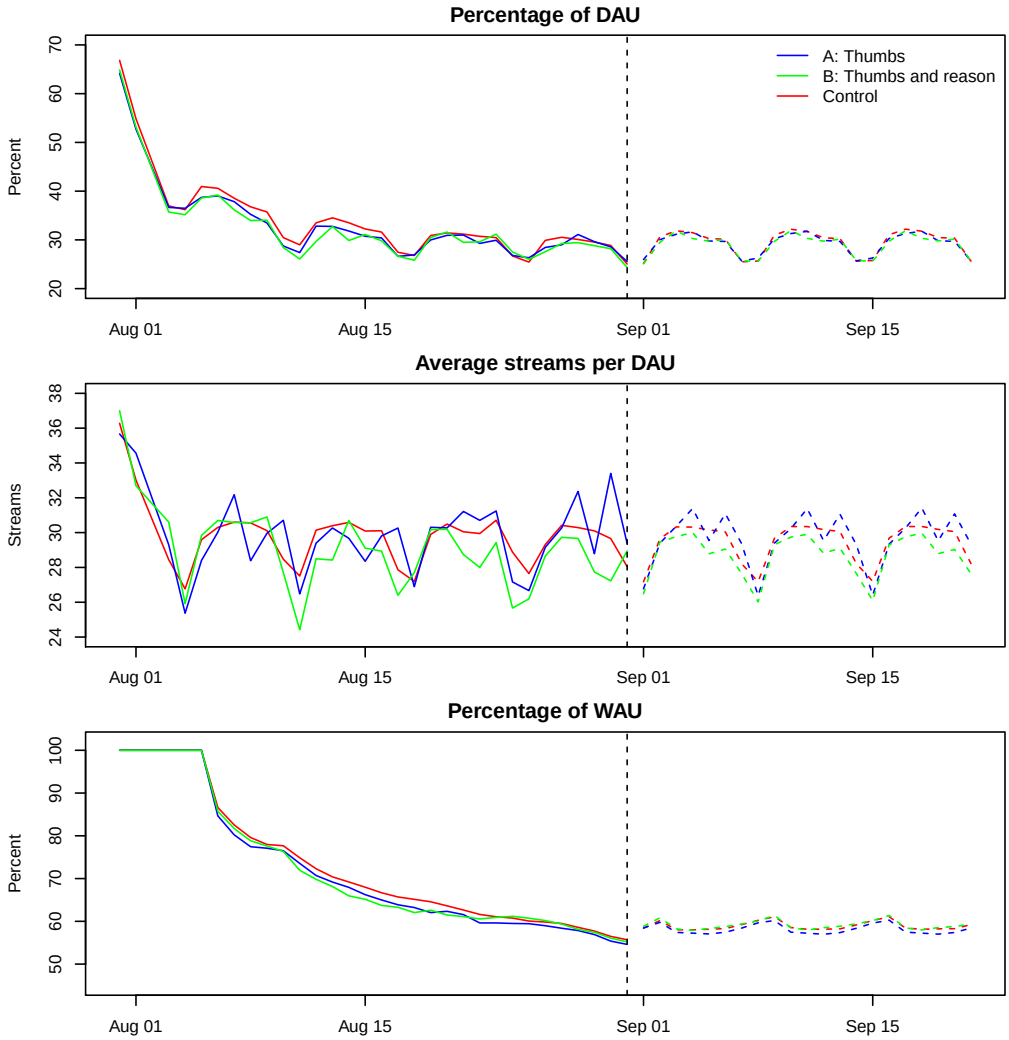

Figure 4.4.: 21 days forecast when using all data

each time series. The MSPE measures the expected mean squared distance between forecasts and true values with the formula $\frac{1}{H}\sum_{h=1}^{H}\left(\hat{y}_{T+h} - y_{T+h}\right)^2$.

As we see in Figure 4.6 the forecast seems to follow the trend of the original series for *Percentage of DAU* and *Average streams per DAU*, but for *Percentage of WAU* it seems to be harder to fit the decreasing trend, especially for the adjusted Bayesian VAR approach.

Now let us see the result when we use only the first 15 days, i.e. data from the time period 31 July 2013 to 14 August 2013. Hence the forecast will be for the next 17 days, 15 August to 31 August, and will be compared against the true values in the same period. The result for both approaches are presented in Figure 4.7 and the MSPE when using 15 days are also in Table 4.1.

Again we see that the forecast seems to follow the original series *Percentage of*

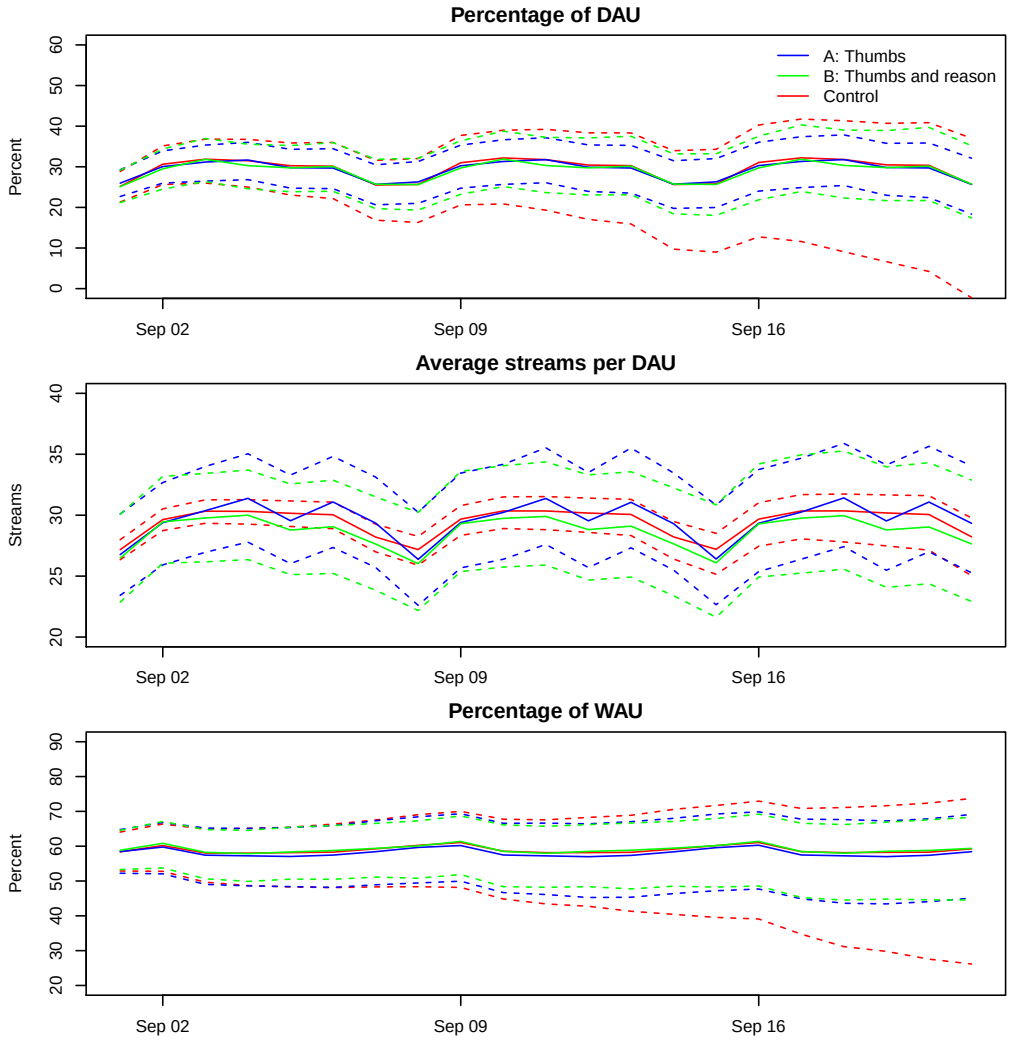

Figure 4.5.: Prediction intervals for the forecasted values when using all data

DAU and *Average streams per DAU* in Figure 4.7. Here it is also clearer that for the adjusted Bayesian VAR approach, the fit for *Percentage of WAU* gives a really poor forecast. While forecasting with the standard Bayesian VAR approach, the decreasing trend is captured. An undesirable seasonal pattern in shown with both approaches in *Percentage of WAU*.

Studying the MSPE in Table 4.1 for all three time series for both approaches, the standard approach outperforms forecasting with the adjusted approach. When forecasting *Percentage of WAU* the adjusted approach performs really poorly with MSPE at 7.54 and 2.75, compared to MSPE at 0.39 and 0.68 when forecasting on the original time series using 15 and 25 days long time series respectively. There is also a difference between the approaches when forecasting *Percentage of DAU* with a 25 days long time series, where the adjusted approach gives an MSPE at 1.02 compared to forecasting with the standard approach which gave a much lower

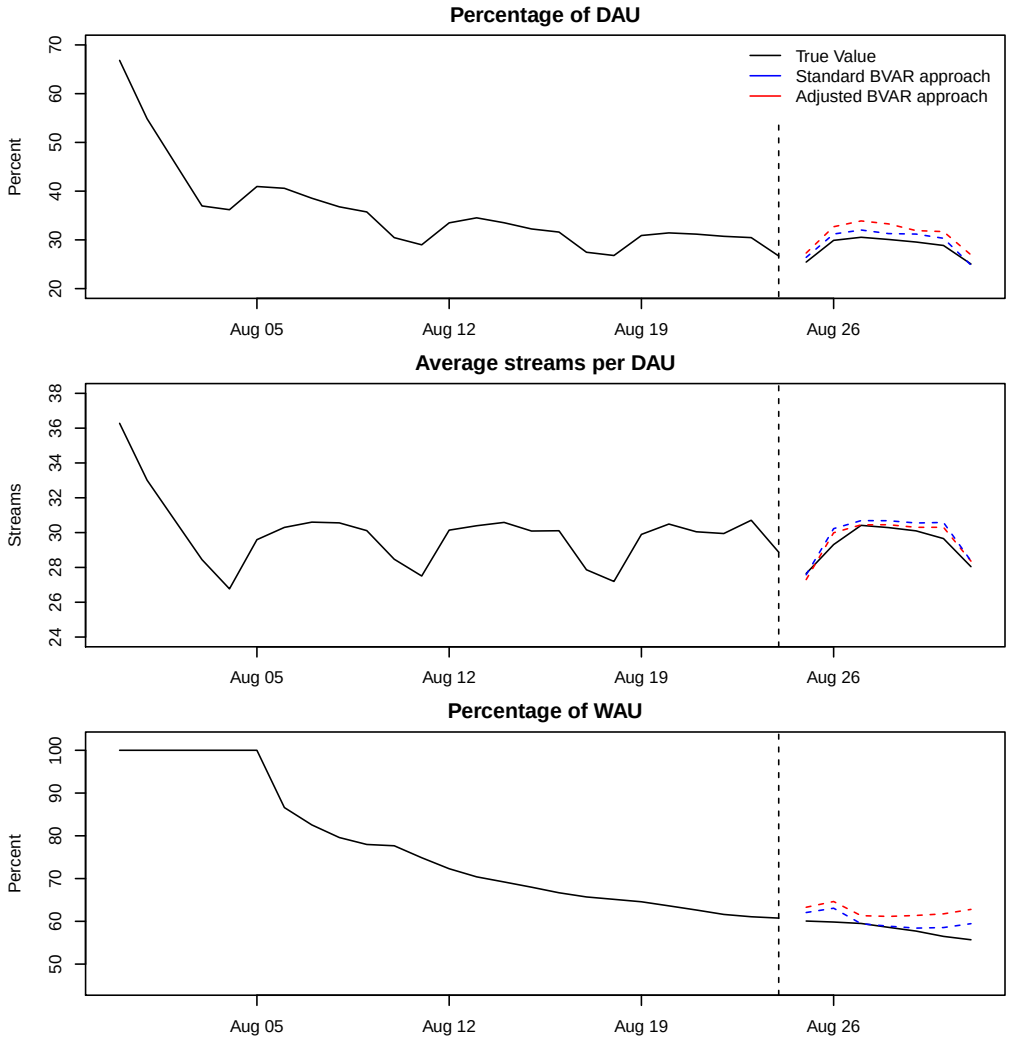

Figure 4.6.: 7 days forecast when using 25 days data

MSPE at 0.18. Forecasting *Average streams per DAU* with the adjusted approach gave a lower MSPE when using both 15 and 25 days.

The final result from both approaches is that there is no difference between the users in the three groups *A: Thumbs*, *B: Thumbs and reason* and *control* predicted future usage of Spotify. This is based on the three variables *Percentage of DAU*, *Average streams per DAU* and *Percentage of WAU*. The feature seems to have too been small of a difference to engage the users to spend more time on Spotify.

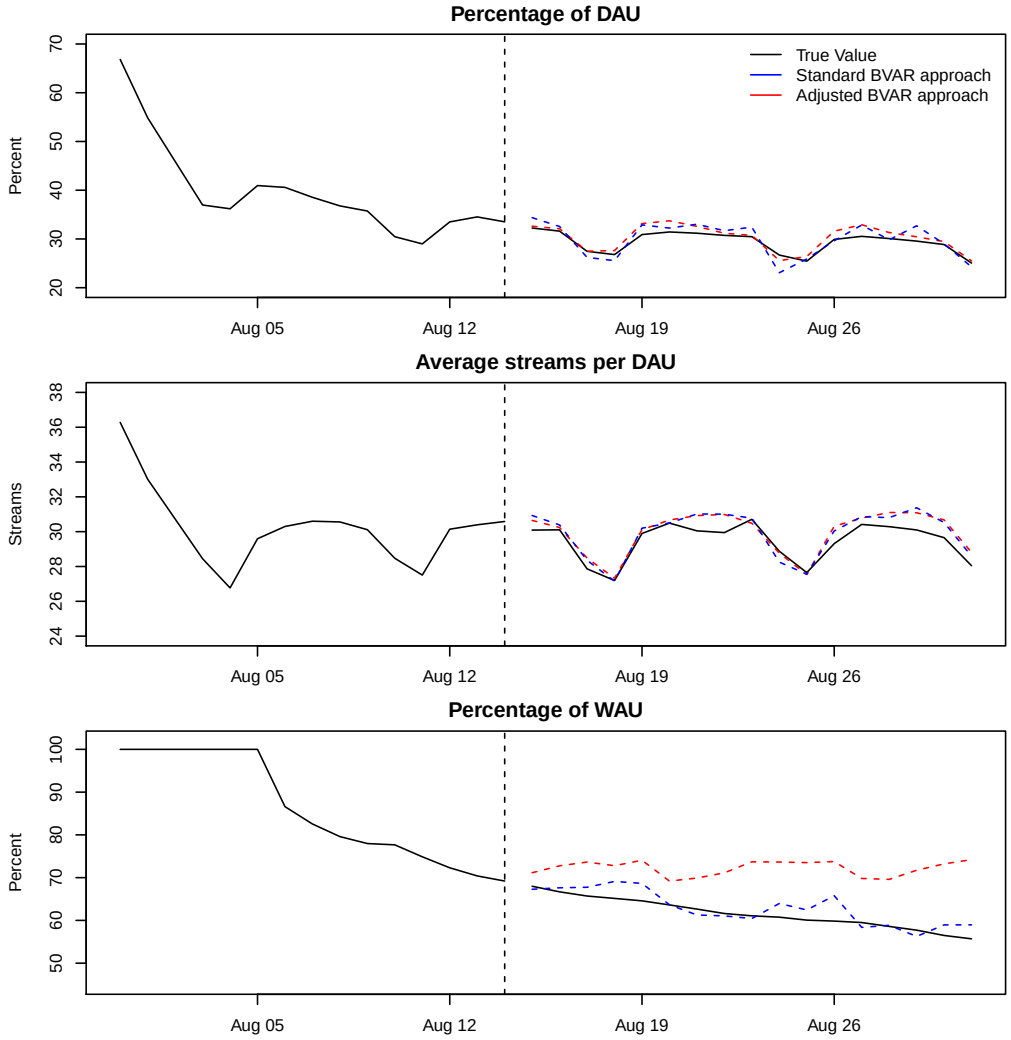

Figure 4.7.: 17 days forecast when using 15 days data

Table 4.1.: MSPE for the standard and adjusted Bayesian VAR approach

	Standard		**Adjusted**	
	15 days	25 days	15 days	25 days
Percentage of DAU	0.18	0.23	0.09	1.02
Average streams per DAU	0.03	0.05	0.02	0.02
Percentage of WAU	0.39	0.68	7.54	2.75

5. Discussion

With both approaches we saw that there were no differences between the groups' future usage of Spotify. There were differences between the two approaches' estimations of $E(y_{T+1:T+H}|Y_T)$. With the standard Bayesian VAR approach, the forecasts in Figure 4.1 were decreasing slowly in *Percentage of DAU* and *Percentage of WAU*, while with the bias adjusted Bayesian VAR approach, the forecasts were more stationary. To evaluate which approach was better, a reduced amount of data was used to compare $E(y_{T+1:T+H}|Y_T)$ to the true values from the time series. The conclusion that could be made, by studying Figure 4.6 and especially Figure 4.7, is that the forecast from the adjusted approach performed poorly since it could not follow the trend of the original series. Forecasting with the adjusted approach depends on when we add the Gompertz value to each forecast estimate. If the Gompertz curve has already leveled off, there is no trend, meaning that we are adding the same value to each forecast estimate at time point $T + h$. This leads to a problem following a decreasing or increasing trend. Furthermore, the residuals from the bias adjustment are stationary around zero, creating a stationary forecast for each time series. This is why the adjusted approach gives poor results when the latter part of the time series is non-stationary. In comparison, the standard approach can capture the time series' trend and does not seem to be to affect much by the bias. However, this is not always the case. When the time series is stationary in the latter part, for example in *Average streams per DAU*, the adjusted approach performs slightly better than the standard approach. Unfortunately the difference between the two approaches is very small, so this result could very well fall into the statistical margin of error.

The seasonal pattern was captured well with both approaches thanks to the seasonal dummies in the Bayesian VAR model. However, the model encounters problems when the seasonal effect is zero, seen in *Percentage of WAU*. Because the users are online the first day, the first six days in this time series will have the value 100 percent. Since the seasonal dummies cannot be removed from individual time series, another type of prior should be chosen with a strong belief that the seasonal dummies for this specific time series should be zero.

To check for statistical significance between the groups, two-sided prediction interval with 95 percent credibility interval were computed. The intervals' role was to check for a potential statistically significanct difference between the groups' forecasted values. A problem with this is that when a feature or a new design has a small impact on the users' usage of Spotify, the predictions intervals will probably be too wide to make the conclusion that there are any differences between the groups.

In this thesis only new users' usage of Spotify has been analyzed since only the data for new users was extracted from Spotify databases. In this paragraph we will discuss both previous and new users based on the studied data for new users and theories for previous users. Studying the new users time series in Figure 2.2 we can see a clear bias effect, especially in *Percentage of DAU* and *Percentage of WAU* which has a negative trend during these 32 days, and is still decreasing in the latter part of the time series. This makes it hard, nearly impossible, to forecast many time points with few time points because the bias in *Percentage of DAU* and *Percentage of WAU* decreases so slowly we do not know when it will level off. Making a long forecast based on data for a month will then have a low reliability because the model will keep forecasting predicted values with a decreasing trend and never level off. However, in Figure 4.7 we saw a really good performance of the forecast when forecasting with the standard Bayesian VAR approach. It could maybe be a coincidence for this data or it could be a really good approach to forecast another two weeks of the time series with 15 days of sampled data since the bias has not subsided in the first month. It would be good to test the method on other test data where we can compare the forecasts with true values.

A user should be classified as a previous user when they have been registered at Spotify long enough for the bias that new users have to wear off. Kohavi et al. (2012) says that the status quo and primacy bias effect usually have subsided after one week, meaning that the status quo and primacy bias is not as long as the bias for new users**.** We would then not need to sample much more data than a week to see a difference between the test groups. The status quo and primacy bias could also depend on how big of a change is being investigated. Bigger changes should probably have a longer status quo or primacy bias, meaning that data should be sampled for a longer time such as two to three weeks. If a forecast would be applied, data should be sampled for at least two or three weeks to estimate a trustworthy model. In the case that time series bias has leveled off and is stationary in the observed time series' latter part, the adjusted Bayesian VAR approach could be performed. If there still is a trend in the time series latter part, forecasting with the standard Bayesian VAR approach could be performed, but since we have not study the data for previous users, a forecast more than one week should not been performed since the forecasting accuracy would then be unreliable for the forecast's subsequent values.

For the actual *Thumbs* test, we could not find any difference between the test groups when studying the prediction intervals for both approaches. If there were any difference between the groups in another unobserved variable, e.g. time spent on the Discover page, it did not have any impact on the variables in this study that the test were evaluated from. However, it could be interesting to study other types of variables when smaller changes like the thumbs feature are tested. Another approach could have been to only analyze users that have been using the Discover page, or only users that have been listening to music recommendations. This is because of whether the majority of the users do not use the Discover page, the groups will be

similar to each other since for the majority does not see the change. Therefore a potential difference between the groups will not be shown when analyzing an A/B-test since the majority of the variants will be exactly the same.

6. Conclusions and further work

- Longitudinal studies are preferred to cross-sectional studies to avoid potential difference in test groups because of cultural differences across generations (e.g. perhaps one day a lot of users from Europe were randomly selected, and the next day a lot of users from Asia were randomly selected. If there is a change between how Europeans and Asians things about the change, a difference in a group will depend on cultural differences).
- Newly registered users and previous users are not under the same conditions when performing A/B-test and should therefore be analyzed separately.
- The use of a Bayesian approach is preferred to a frequentist approach when having less data. It also gives an easier and better inference.
- When a bias has not subsided for new users in time series and there is still a decreasing trend for the later part of time series, the standard Bayesian VAR approach is preferred to the adjusted Bayesian VAR approach which then gives a poor forecast. In the case that the bias has subsided so the time series are stationary for the latter part of time series, the adjusted approach is slightly better than the standard approach.
- When analyzing newly registered users, the bias subsides slowly so we do not know with 32 days of sample data when it is possible to discover a potential difference between test groups. However, it was shown in this thesis that when a test has been running for 15 days, it is possible to perform a credible forecast the next 17 days.
- For previous users, the status quo and primacy biases level off in an early stage so the data most likely does not need to be sampled for more than two weeks to discover a potential difference between test groups. Since the status quo and primacy biases are less extreme than the bias for newly registered users, the standard and adjusted Bayesian VAR approach could be applied to forecast previous users' usage of Spotify under the same conditions as when forecasting new users' usage.

6.1. further work

There are some other approaches to handling potential status quo and primacy biases when studying a previous user's reaction. One suggestion according to Kohavi et al.

(2012) is to exclude the first week of sampled data since the bias effect usually has subside by then. If we only have sampled data for a short while, though, that is probably not feasible. Another approach could be to model for any potential bias, both for new and previous users. The approach suggested by Primiceri (2005) with time-varying coefficients to the Bayesian VAR could be a suitable method for this problem since time series in econometrics are also often non-stationary. In the paper Primiceri models three time series in the US economy: inflation rate, unemployment rate and short-term nominal interest rate. He then uses time-varying coefficients in the VAR model since there is strong evidence that US unemployment and inflation were higher and more volatile in the period between 1965 and 1980 than in the last twenty years. This could be an interesting approach for this study were bias could have an effect in the beginning of the test run while not in the rest of the test run.

Evaluating a test on a single time series could be a ***bit robust***. It could be interesting to evaluate tests with an *Overall Evaluation Criterion* (OEC). An OEC would in this case be a single vector time series calculated by taking information from other time series. An easy example of OEC, not from a time series perspective, is when a student has multiple exams in school, and the results from all of them are used to give one final grade. Perhaps one exam is more important than other exams, and then we can compute weighted results from all the exams. This could be applied by someone who knows how their OEC should be constructed for their specific test when evaluating their A/B-test. A univariate time series method should then be applied.

In this thesis time series were transformed to have an increasing trend instead of an original decreasing trend so they would suit the Gompertz function $a * exp(b * exp(c * t))$. Another approach to fitting a Gompertz function to decreasing data while keeping its properties of being slower in the beginning and at the end is to rewrite the function to $a * (1 - exp(b * exp(c * t)))$ which means that transforming is not necessary. The problem is then to estimate the coefficients since the function *nls* in *R* has problems with the rewritten Gompertz functions derivative. Trying optimization techniques to minimize the *Sum of Squared Error* (SSE) result is a poor estimation of the coefficients.

Bibliography

Carlson, N. R., Heth, D., Miller, H., Donahoe, J., and Martin, G. N. (2009). *Psychology: the science of behavior.* Pearson.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100.

Doh, T. and Connolly, M. (2013). *The state space representation and estimation of a time-varying parameter VAR with stochastic volatility.* Springer.

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning, volume 2.* Number 1. Springer.

James, E. (2002). *Gentle. Elements of Computational Statistics.* Statistics and Computing. Springer.

Karlsson, S. (2012). Forecasting with bayesian vector autoregressions. Technical report, Orebro university.

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., and Xu, Y. (2012). Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 786–794. ACM.

Litterman, R. B. (1986). Forecasting with bayesian vector autoregressions-five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38.

Nakajima, J. and Ginkō, N. (2011). Time-varying parameter var model with stochastic volatility: An overview of methodology and empirical applications. Technical report, Institute for Monetary and Economic Studies, Bank of Japan.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic linear models with R.* Springer.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Sims, C. (1980). Macroeconomics and reality. *Econometrica*, 48:1–48.

Vieira, S. and Hoffmann, R. (1977). Comparison of the logistic and the gompertz growth functions considering additive and multiplicative error terms. *Applied Statistics*, 18:143–148.

Zivot, E. and Wang, J. (2007). *Modeling Financial Time Series with S-PLUS®*, volume 191. Springer.

A. R-code

In this appendix code used in R will be presented.

A.1

```
# This function gives me the right values on the
# dummies depending on which day it is.
# This function will be used in function 'setup' and 'BVAR_forecast'.
weekd <- function(date){
  day <- weekdays(as.Date(date))
  a <- matrix(diag(6), 6)
  if(day == 'Monday'){
    vec <- a[1,]
  }else if(day == 'Saturday'){
    vec <- a[2,]
  }else if(day == 'Sunday'){
    vec <- a[3,]
  }else if(day == 'Thursday'){
    vec <- a[4,]
  }else if(day == 'Tuesday'){
    vec <- a[5,]
  }else if(day == 'Wednesday'){
    vec <- a[6,]
  }else{
    vec <- rep(0, 6)
  }
  return(matrix(vec, 1))
}
```

A.2

```
# This function should give me the right matrices that will be used
# when forecasting with BVAR.

#### INPUT
# matrix = matrix with time series (T x m)
# p = number of lags you want
# date = sequence of dates for when the data was sampled (only if seasonal = TRU
# seasonal = if you want seasonal dummies or not

#### OUTPUT
# Z (big) = intercept, lags and dummies for t = 1, ..., T-p
# Y = m time series for t = p+1, ..., T
# T = length of time series
# k = number of coefficients
# p = number of lags
# d = number of deterministics variables
# m = number of time series
# z (small) = explanatory variables for time point T+1,
#             will be the values for the first forecast in BVAR

setup <- function(matrix, p=1, date=NULL, seasonal=FALSE){
  m <- ncol(matrix) # Tells how many variables there are
  T <- nrow(matrix) # How many time points are observed

  lags_var <- list()
  Z <- cbind(rep(1, T-p), matrix[p:(T-(1)),]) # Z with intercept and lag 1
  if(p > 1){ # add other lags to Z
    for(i in 1:(p-1)){
      lags_var[[i]] <- matrix[(p-i):(T-(i+1)),]
      Z <- cbind(Z, lags_var[[i]]) # lag > 1
    }
  }

  d <- 0
  if(seasonal){ # add seasonal dummies to Z
    dummies <- matrix(0, ncol=6, nrow=(T-p))
    wd <- date[(p+1):length(date)]
    for(i in (p):(length(date)-p)){
      dummies[i,] <- weekd(wd[i])
    }
    d <- 6
    Z <- cbind(Z, dummies)
  }
```

```
  Y <- matrix[(p+1):nrow(matrix),] # Fix Y depending on how many lags there are
  colnames(Y) <- rownames(Y) <- colnames(Z) <- rownames(Z) <- NULL

  d <- 1+d # intercept + dummies
  k <- m*p+d # number of coefficients
  p <- p # number of lags

  # Fix z (small) that will be the values for the first forecast in BVAR
  o <- matrix(Y[nrow(Y):(nrow(Y)-p+1),], ncol=m)
  r <- o[1,]
  if(p > 1){
    for(i in 2:p){
      r <- cbind(r, o[i,])
    }
  }
  if(seasonal){ # add seasonal dummies to z
    z <- matrix(c(1, as.vector(r), weekd(wd[i]+1)), nrow=1)
  }else{ # if no seasonal dummies
    z <- matrix(c(1, as.vector(r)), nrow=1)
  }
  return(list(Z = Z, Y = Y, T = T, k = k, p = p, d = d, m = m, z=z))
}
```

A.3

```
## Forecast model with seasonal dummies

#### INPUT
# H = Number of time points you want to forecast
# R = number of iterations for Gibbs sampler
# Y = time series (output from function 'setup')
# Z (big) (output from function 'setup')
# z (small) = (output from function 'setup')
# p = number of lags (output from function 'setup')
# date = The date for when t = T

#### OUTPUT
# predY = A list, each folder in the list is a (R x H) matrix

BVAR_forecast <- function(H, R, Y, Z, z, p, date = as.Date('2013-08-31')){
```

```
m <- ncol(Y)
k <- ncol(Z)
t <- nrow(Z)
T <- t+p
# Parameters
Gamma_hat <- solve(t(Z) %*% Z) %*% t(Z) %*% Y
S <- t(Y - Z %*% Gamma_hat) %*% (Y - Z %*% Gamma_hat)
P <- solve(t(Z) %*% Z)

# Set up
Gamma <- matrix(0, ncol=R, nrow=(ncol(Gamma_hat)*nrow(Gamma_hat)))
psi <- matrix(0, ncol=R, nrow=m*m)
mat <- rbind(Z, matrix(0, ncol=k, nrow=H))
pred <- matrix(0, ncol=m, nrow=H)
predY <- list()
for(o in 1:m){
  predY[[o]] <- matrix(0, ncol=H, nrow=R)
}

## GIBBS
IW <- matrix(0, m, m)
for(j in 1:R){
  # Sample Psi
  IW <- riwish(T-k, S)
  psi[,j] <- matrix(IW, ncol=1)

  Q <- IW
  # Sample Gamma
  Gamma[,j] <- matrix(mvrnorm(1, as.vector(Gamma_hat), kronecker(Q, P)), ncol=1

  # Predict Y_T+1 : Y_T+H
  for(h in 1:H){
    # Sample u for every m:th time series at time point t
    u <- as.vector(mvrnorm(1, rep(0, m), IW))
    if(h == 1){
      # predict Y_T+1
      pred[h,] <- z %*% matrix(Gamma[,j], ncol=m) + u
    }else{
      # predict Y_T+h
      pred[h,] <- matrix(mat[t+h-1,],1) %*% matrix(Gamma[,j], ncol=m) + u
    }
    for(i in 1:m){
      # put the predicted values in the right list
```

```
        # (one folder in the list for each time series)
        predY[[i]][j,h] <- pred[h,i]
      }
      # add predicted values to Z so it's possible to predict next value.
      # these values will be replaced for every iteration in Gibbs sampler.
      if(p > 1){
        mat[t+h,] <- cbind(1, matrix(pred[h,], 1), matrix(mat[t+h-1, 1:(m*(p-1))+1],
      }else if(p == 1){
        mat[t+h,] <- cbind(1, matrix(pred[h,], 1), weekd(date+h))
      }
    }
  }
  return(predY)
}
```

B. Gompertz plots and residuals

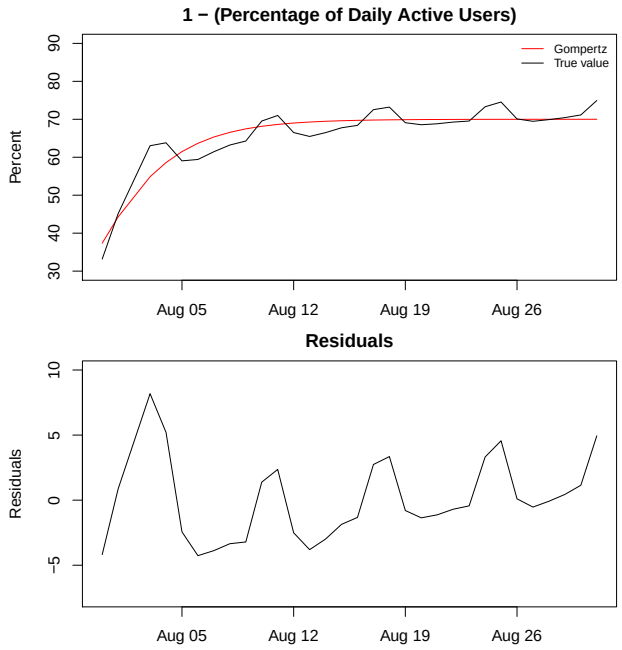

Figure B.1.: Gompertz function and residuals, Group: Control, Variable: DAU

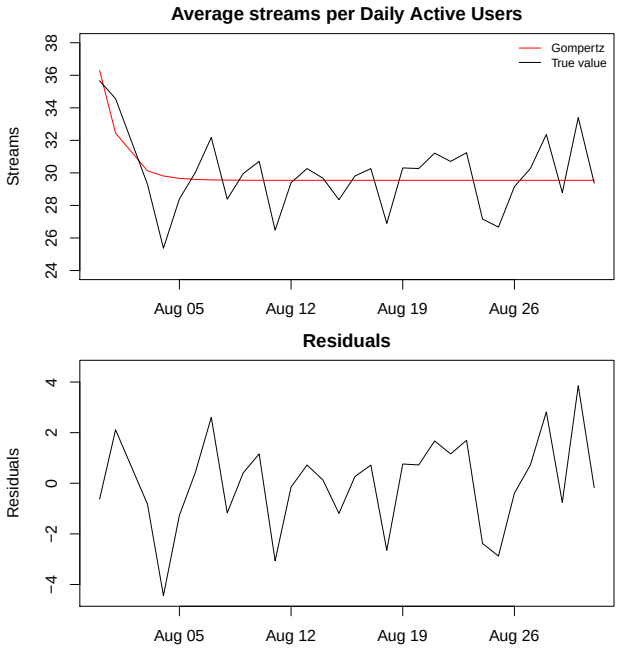

Figure B.2.: Gompertz function and residuals, Group: Control, Variable: Streams

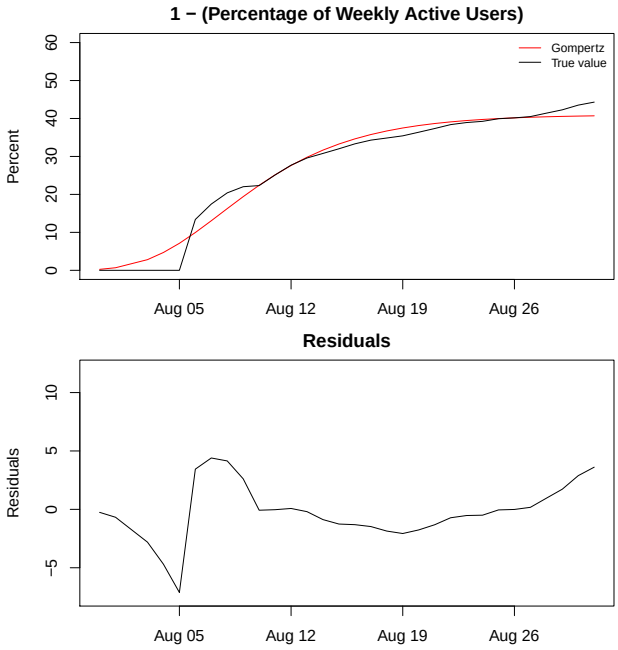

Figure B.3.: Gompertz function and residuals, Group: Control, Variable: WAU

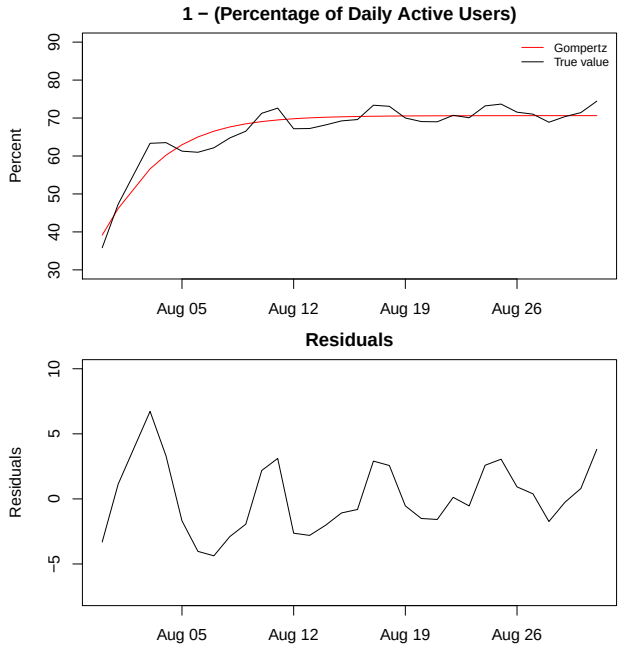

Figure B.4.: Gompertz function and residuals, Group: A: Thumbs, Variable: DAU

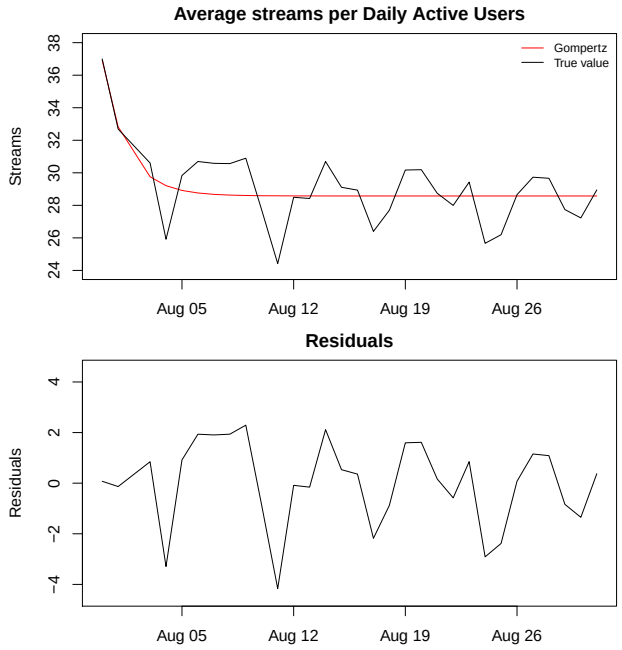

Figure B.5.: Gompertz function and residuals, Group: A: Thumbs, Variable: Streams

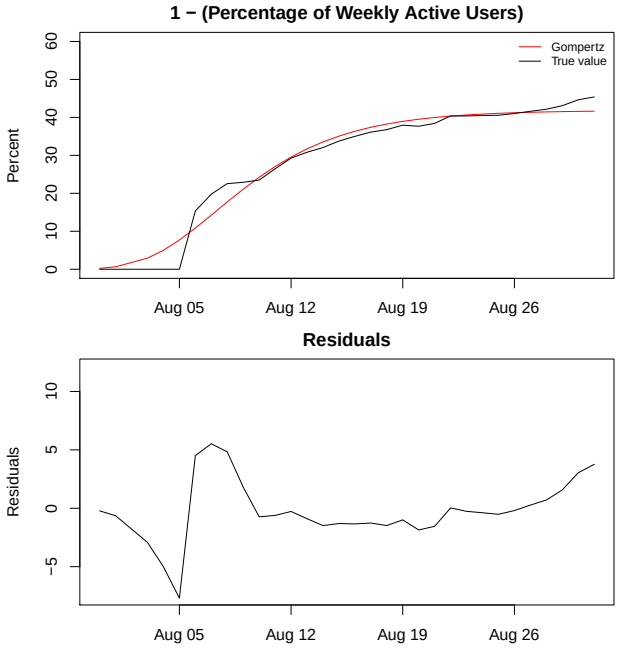

Figure B.6.: Gompertz function and residuals, Group: A: Thumbs, Variable: WAU

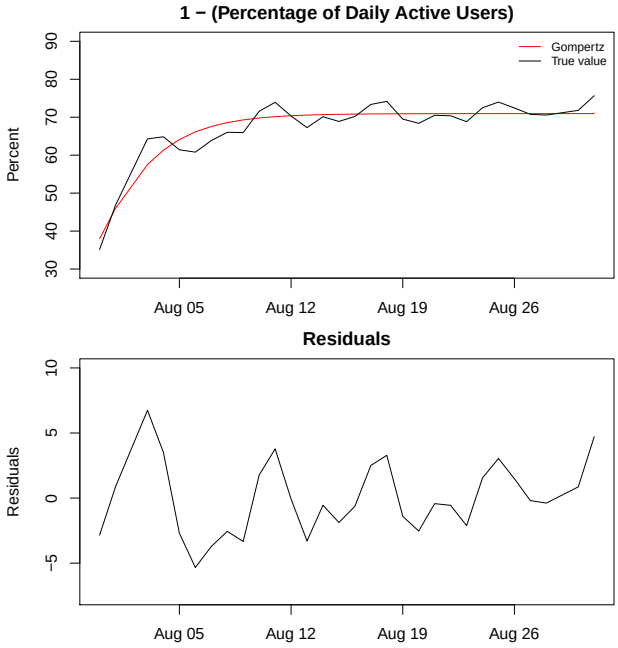

Figure B.7.: Gompertz function and residuals, Group: B: Thumbs and reason, Variable: DAU

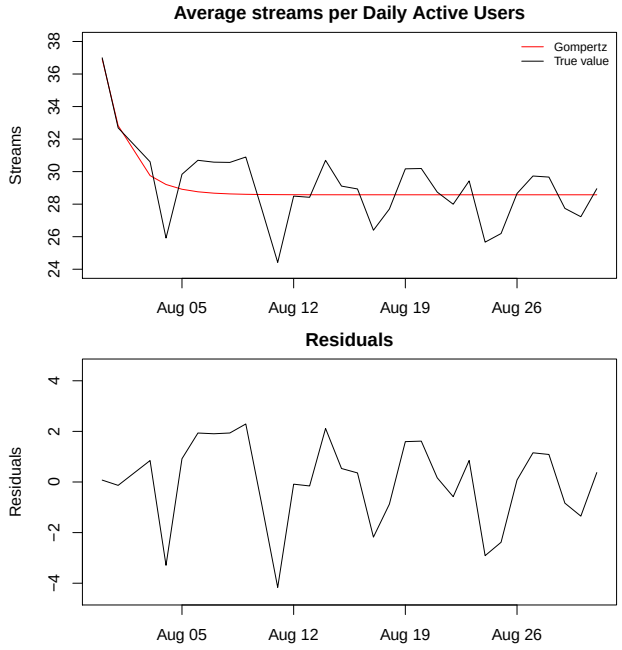

Figure B.8.: Gompertz function and residuals, Group: B: Thumbs and reason, Variable: Streams

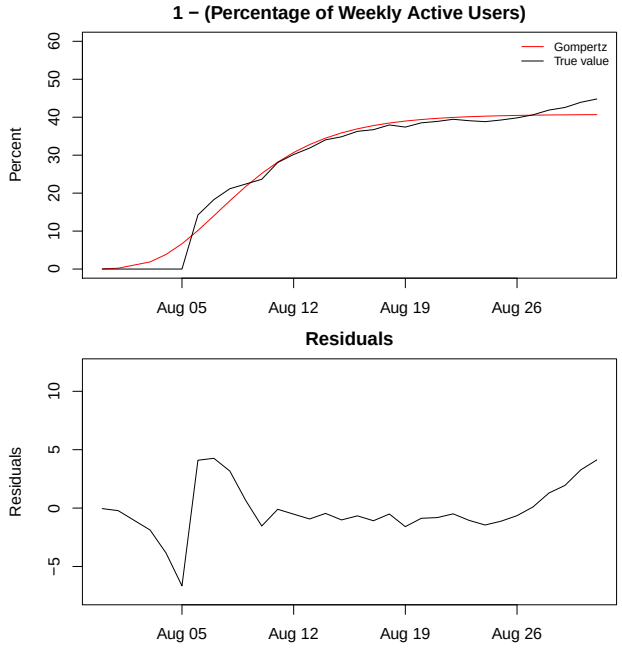

Figure B.9.: Gompertz function and residuals, Group: B: Thumbs and reason, Variable: WAU

???? ISRN ?????