

Bayesian model validation metrics in retail datasets

Leevi Rönty

School of Science

Bachelor's thesis
Espoo 14.10.2020

Supervisor

Prof. Fabricio Oliveira

Advisors

DSc (Tech.) Mikko Ervasti

DSc (Tech.) Paavo Niskala



Aalto University
School of Science

Copyright © 2021 Leevi Rönty

The document can be stored and made available to the public on the open internet pages of Aalto University.
All other rights are reserved.



Author Leevi Rönty

Title Bayesian model validation metrics in retail datasets

Degree programme Engineering Physics and Mathematics

Major Mathematics and Systems Sciences

Code of major SCI3029

Teacher in charge Prof. Fabricio Oliveira

Advisors DSc (Tech.) Mikko Ervasti, DSc (Tech.) Paavo Niskala

Date 14.10.2020

Number of pages 14+2

Language English

Abstract

Your abstract in English. Keep the abstract short. The abstract explains your research topic, the methods you have used, and the results you obtained.

Keywords Bayesian models, model validation, information criterion, Facebook Prophet



Tekijä Leevi Rönty

Työn nimi Bayeslaisten mallien validointi vähittäismyynnin aineistoissa

Koulutusohjelma Teknillinen fysiikka ja matematiikka

Pääaine Matematiikka ja systeemitiheet

Pääaineen koodi SCI3029

Vastuopettaja Prof. Fabricio Oliveira

Työn ohjaajat TkT Mikko Ervasti, TkT Paavo Niskala

Päivämäärä 14.10.2020

Sivumäärä 14+2

Kieli Englanti

Tiivistelmä

Tiivistelmässä on lyhyt selvitys kirjoituksen tärkeimmistä sisällöstä: mitä ja miten on tutkittu, sekä mitä tuloksia on saatu.

Tämän opinnäytteen tiivistelmäteksti kirjoitetaan opinnäytteen luettavan osan lomakkeen lisäksi myös pdf-tiedoston metadataan \thesisabstract-makron avulla (kastoyllä). Kirjoita tähän luettavaan tiivistelmälomakkeeseen menevä teksti. Tässä saa olla erikoismerkkejä kuten kreikkalaiset kirjaimet ja rivinvaiho- ja kappaleenjako-merkit. Tämän tekstin on muuten oltava sama kuin metadatatiiivistelmän teksti.

Jos tiivistelmäsi ei sisällä erikoismerkkejä eikä kaipaa kappaleenjakoja, voit hyödyntää makroa \abstracttext luodessasi lomakkeen tiivistelmää (katso kommentti alla).

Avainsanat Bayeslaiset mallit, mallin validointi, informaatiokriteeri, Facebook Prophet

Contents

Abstract	3
Abstract (in Finnish)	4
Contents	5
1 Introduction	6
2 Background	6
2.1 Rakenne	7
3 Datasets and methods	7
4 Results	9
5 Summary	12
A Esimerkki liitteestä	15
B Toinen esimerkki liitteestä	16

1 Introduction

Businesses need forecasting to succeed. Correctly forecasting demand for goods is essential for any retail business, as shortages can cause loss of sales. As companies collect more and more data, the possibilities of forecastable subjects and possible features increase dramatically. However, one can not just include more features in the model and expect it to perform well. Also, not all models are suitable for all forecasting tasks. The amount of possible models is endless, but most of them will perform poorly. To find the useful ones, we must be able to measure the goodness of a model. Being able to objectively compare models allows for systematic approaches of model selection to be adopted.

Model validation is an integral part of a robust modelling framework. Bayesian models are a relatively novel model type, which has gained considerable popularity in recent years, as computational power of modern computers has kept rising. As a relatively new method, not too many validation metrics have been proposed with these kinds of models in mind.

The current state of model validation for bayesian models can be described as "unsatisfactory". Information criteria that try to predict out of sample fitness can have strong biases, and some may not take into consideration the nature of bayesian models and distributions of parameters. Cross validation can be extremely computationally intensive. Also, most information criteria are based on a loss function proportional to the root mean square of errors, but that may not always be the most useful function to optimize for. In some businesses the consequences of errors in forecasts can be described better with the mean absolute percentage error. All in all there does not exist a clearly better method for solving all bayesian model validation problems.

In this thesis we will model sales time series of some Walmart stores in the United States. We will be using the Facebook's Prophet modelling framework. It consists of a flexible bayesian model which can be customized to fit a wide range of possible time series modelling tasks. We will implement some information criteria for the model and compare it to more ad-hoc methods for model validation.

2 Background

Predictions of sales can be used to minimize waste and ensure sufficient supply of goods. Advertising products can yield greater sales, but not all ads are equally effective. Shifting marketing investments to more impactfull marketing channels can increase sales while keeping costs the same. Both of these examples demonstrate scenarios where statistical models can be used to avoid making suboptimal decisions.

Models can be viewed as simplifications of reality. This means that no model never really matches the true data generating processes, but if we can formulate a model that works well enough, we can try to draw some conclusions from them. Most of the models we are concerned with link together some explanatory variables to the measured data. The parameters of the function are learned by fitting the data to the model. The fitted model can then be used to predict future observations if we can

know the explanatory variables beforehand. The fitted model parameters can also give insight on the behaviour of the physical world, assuming that said parameters have a sensible interpretation.

Bayesian models differ from frequentist models in two major ways. Firstly, in bayesian thinking parameters are not expressed as point values but distributions. Secondly, the distributions are affected also by the prior beliefs of the distribution. This means, that the data is not the only source of information affecting the results. These qualities allow for embedding uncertainty and business knowledge in the models. Highly informative priors can also make the models behave more robustly in case the data has some outliers.

Model selection methods can be categorized by the view they take on whether the considered models contain the actual data generating model. This data generating model would be represent the actual real world process by which some events result in the observed data. M-closed view assumes that the model is present is the considered models. M-open view instead attempts to model the data with minimal assumptions. M-open view is more often applicable than M-closed.

M-open methods for model selection can be further classified by reuse of data. M-open assumes that pseudo-random samples of future data can be obtained.

2.1 Rakenne

3 Datasets and methods

The original dataset consists of department-level weekly sales data of 45 stores. The data spans a little over two years, from 20.2.2010 to 26.10.2012. Each store location is associated with features that may help with sales prediction. The features are temperature, unemployment, price of gas, and consumer price index CPI. Also the store size and type of store is known. The original dataset contains information about markdowns, but this feature is available only from 11.11.2011 onwards. This renders the feature useless, as it is hard to compensate for the missing data.

The sales data for store groups A and B demonstrate strong seasonality. Especially during decemblers the turnover spikes due to strong holiday effect. For store group C the sales seem a lot more consistent. Temperature is the only feature that clearly behaves seasonally. Originally unemployment was expressed to equal the latest unemployment statistic. This made the feature behave rather discontinuously. To get around the discontinuity affecting the results, the unemployment for each week was calculated by interpolating unemployment from last and next changepoint. After the transformation, both unemployment and consumer price index (CPI) have a clear trend.

The original dataset contains sales data on a departement level. There are a total of 81 departements, but not all are present in every store. We are still left with 3331 sales timeseries. Not all timeseries can be modelled or fitted, thus we need to reduce the amount of data to have any hope of modelling anything in a reasonable time. The departement-level sales are first aggregated to store-level sales by summing sales in each departement in each store. This still leaves us with 45 timeseries, which is

Table 1: Different settings used in each model. Extra features are fuel price, CPI, unemployment and temereature.

	M1	M2	M3	M4	M5
Yearly seasonality coefficients	10	10	10	20	None
Extra features		X			
Trend changepoints	25	25	25	25	1
Response variable as regressor			X		X

still a bit too much. To simplify things a bit more, the stores are further aggregated based on store type (A, B or C). The features for these store groups are obtained by calculating a size-weighted average. As the aggregation process of the features can render the features less meaningfull, a small number of stores is also selected for modelling from each store group.

The prophet model is a linear model used to predict time series. The model consists of seasonality components, trend, and possible explanatory variables. The seasonalities are modelled using a fourier approximation with a limited amount of coefficients. Trend is modelled linearly between a number of pre-determined changepoints. After the last change point the model also assumes a linear trend. The prophet model handles time as a feature ranging from 0 to 1. As such, the actual time interval between datapoints does not have to be constant.

Each of the five fitted models is an prophet model with different settings and features. The models were constructed to showcase the behaviour of the metrics in situations where the model is used as one could use it in a business environment, overfitting seasonality, a case where a data-generating model can be constructed from the explanatory variables, and a case similar to the last-mentioned but with even less distracting variables. The model differences are represented in table 1.

The models were compared using five diferent metrics. The metrics used were Akaike information criterion (AIC), deviance information criterion (DIC), Watanabe-Aikaike information criterion (WAIC), mean average percentage error (MAPE), and 10-fold cross validation. The information criterions are calculated using only in-sample fits. They aim to estimate the out-of-sample predictive performance of the model using the likelihood of past observations. Notably only WAIC of the information criterions fully considers the bayesian nature of the model. It utilises the whole posterior distribution instead of some aggregated values. MAPE is calculated by reserving a small portion of data from the tail of the dataset to act as a test set. The percentage errors of the predictions are then averaged to obtain MAPE. 10-fold cross validation is calculated by assigning randomly 1/10 of the datapoints as test set and then calculating the likelihood of the prediction for the test set. This is done 10 times so that each datapoint belongs to the test set once. 10-fold CV is the average of the results.

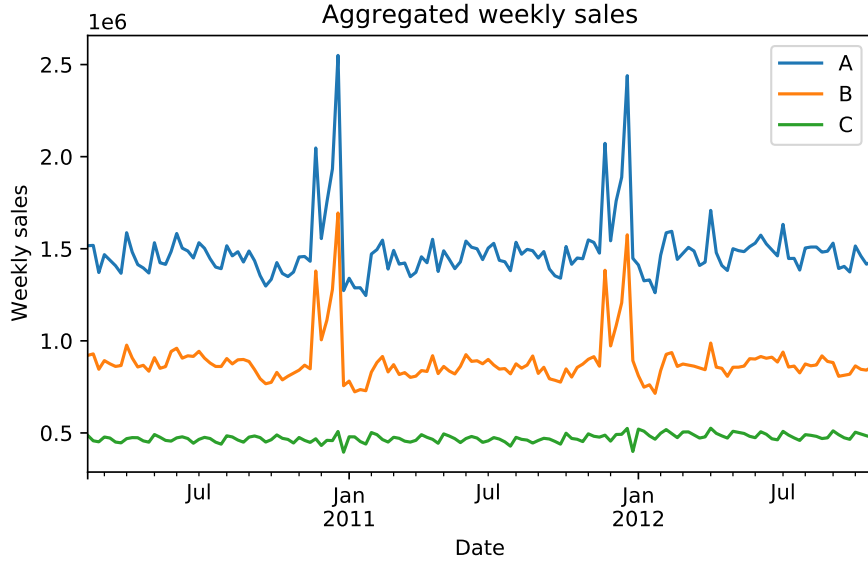


Figure 1: Aggregated weekly sales by store group. Groups A and B demonstrate holiday effects while group C remains relatively stable through each year.

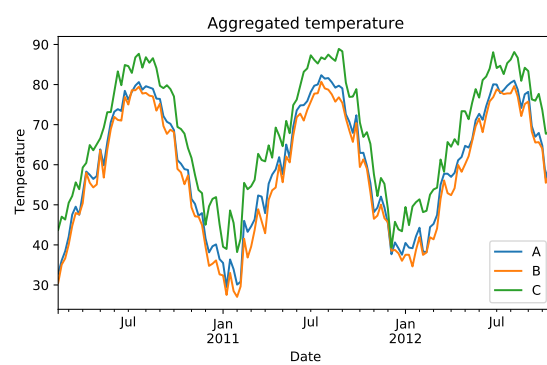
4 Results

Applying the metrics yielded promising results. The results are presented visually in figures 3, 4 and 5. The visual representation highlights the order different metrics rank the models. AIC and DIC ranked all the models similarly for all the datasets, but there were differences in the margins between the models. WAIC seemed to mostly agree with AIC and DIC, but in datasets B and C the order of models m1 and m4 was reversed.

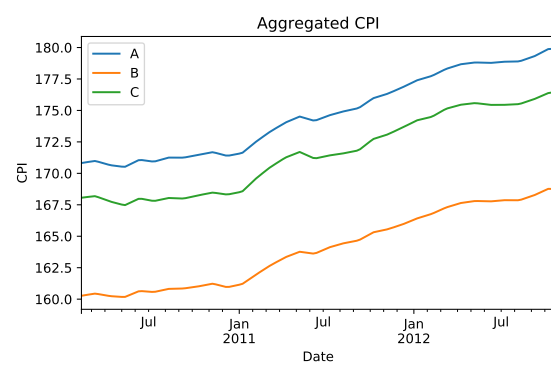
Ranking by MAPE has many similarities with the information criterion rankings. However the model m5 performed worse with all datasets, dropping behind m3 in A and C, but also behind m1 and m4 in B. In dataset B MAPE ranked m1 better than m4. This agrees with the WAIC ranking, but it is the other way around with AIC and DIC. The other notable difference in MAPE scores is that in datasets A and B the model m2 seems to perform much worse than the other models. However this phenomenon is not present in C.

10-fold cross validation ranks the models m5 and m3 the best in all datasets, just like the information criteria do. However the ranking of the rest varies. In dataset C the models m1, m2 and m4 are ranked similarly as by AIC and DIC. In both A and B the model m4 takes the last place, while m2 and m1 get a score similar to each other. M2 was ranked better than m1 for dataset A and vice versa for B, but the margin was rather slim.

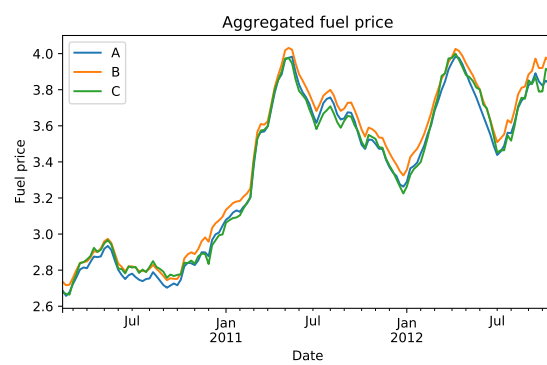
Overall it seems, that the metrics correctly recognize the "data generating models" from other models. Also M5 mostly ranks better than M3, as it has less non-informative regressors. When it comes to the ranking of models M1, M2 and M4, the results are not as clear. In store groups A and B M2 performs much worse



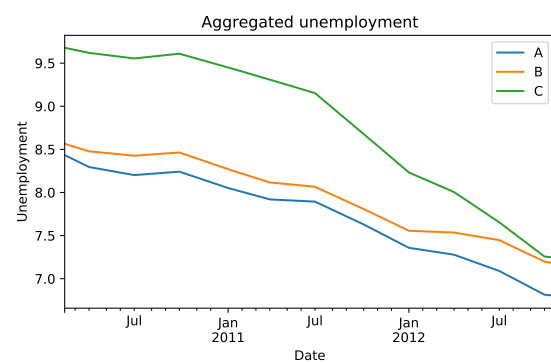
(a) Temperature by store group



(b) Consumer price index by store group



(c) Fuel price by store group



(d) Interpolated unemployment rate by store group

than the other models when evaluating with MAPE. This likely indicates, that the model overestimates the importance of the external regressors. As the regressors have trends, the prediction drifts off the actual sales. In 10-fold cv this is not observed, as the datapoints in test set are picked randomly. The test points have similar-valued training points next to them, thus the trend does not have time to cause the predictions to drift off.

10-fold cv appears to rank M4 worse than the other metrics. This may be due to overfitting the seasonality with too little training data. As the model is evaluated with 10-fold cv, 10% of the data is reserved for testing. This reduction of training data can be the cause for the model to overfit the seasonality, as it had a greater number of coefficients to fit than the other models.

One thing to keep in mind when examining the results is the difference of loss functions used in the metrics, in particular the difference in MAPE and the others. Other metrics are based on estimating the log-likelihood of out-of-sample predictions, while MAPE evaluates the mean absolute percentage error. Log-likelihood is proportional to MSE as the model consist an normally distributed error term with a constant variance. In a perfect world the loss function would be representing the actual losses due to errors in predictions. However, it is really difficult to estimate those accurately.

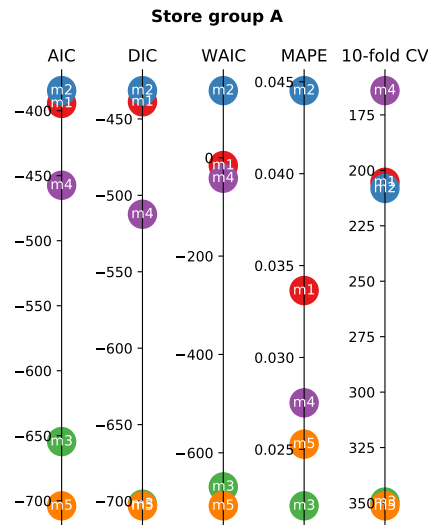


Figure 3: Visualization of model validation metric results applied on store group A. Arranged such that visually lower is better. Each axis is scaled so that extreme values are at the ends. Please note the inverted axis of 10-fold cross validation metric. For precise values please refer to Table 2.

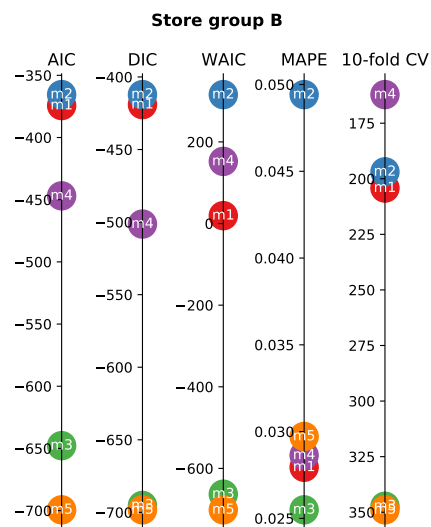


Figure 4: Caption this

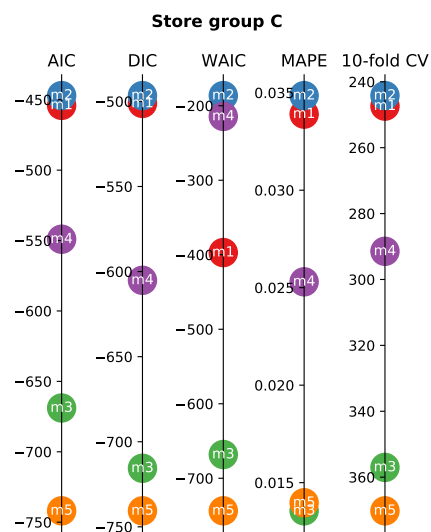


Figure 5: Caption this

5 Summary

References

Table 2: Results of validation metrics applied on store group A

Model	AIC	DIC	WAIC	MAPE	10-fold CV
m1	-394.4	-438.8	-15.9	0.0336	205.4
m2	-384.6	-431.5	136.5	0.0445	208.2
m3	-654.5	-702.1	-669.3	0.0219	349.6
m4	-457.4	-512.4	-41.6	0.0275	164.1
m5	-703.9	-703.2	-707.8	0.0253	351.3

Table 3: Results of validation metrics applied on store group B

Model	AIC	DIC	WAIC	MAPE	10-fold CV
m1	-374.3	-419.1	19.3	0.0279	204.2
m2	-365.5	-412.2	315.9	0.0494	196.9
m3	-647.8	-695.3	-663.2	0.0255	347.3
m4	-446.9	-501.4	152.2	0.0286	162.1
m5	-699.4	-698.3	-701.5	0.0297	348.9

Table 4: Results of validation metrics applied on store group C

Model	AIC	DIC	WAIC	MAPE	10-fold CV
m1	-454.0	-501.2	-397.1	0.0339	247.4
m2	-447.1	-496.6	-186.4	0.0349	244.2
m3	-668.9	-715.6	-668.1	0.0136	357.0
m4	-549.1	-605.1	-214.3	0.0253	291.4
m5	-741.9	-740.6	-743.7	0.014	370.2

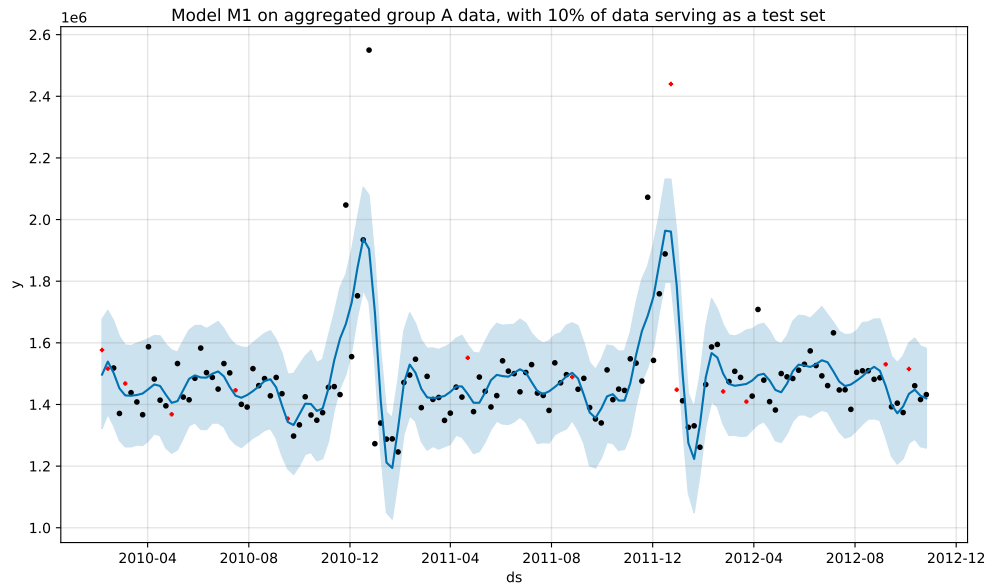


Figure 6: Demonstrating the cv metric calculation

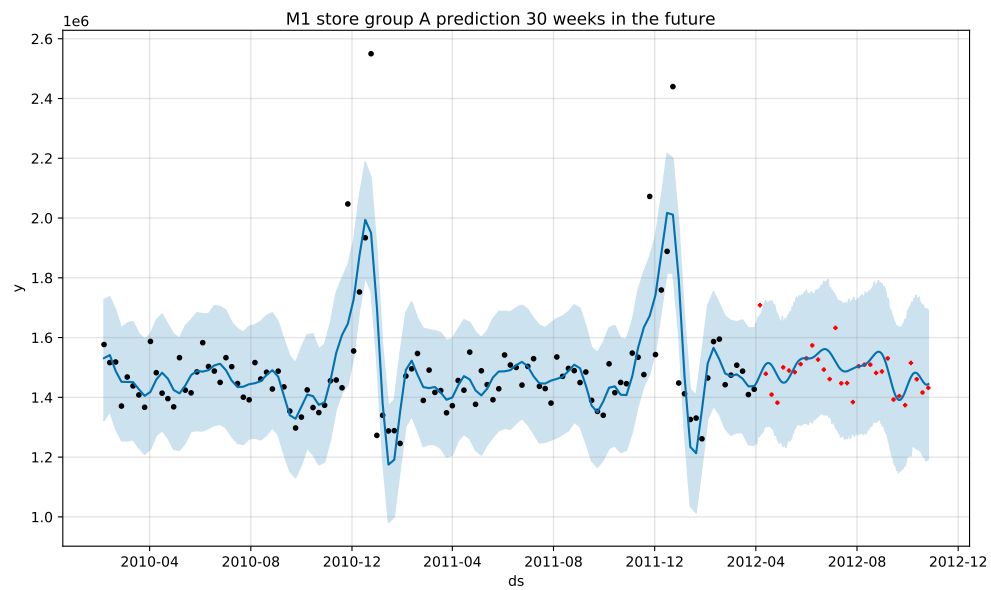


Figure 7: A prediction 30 weeks in the future. Training data in black, test data in red.

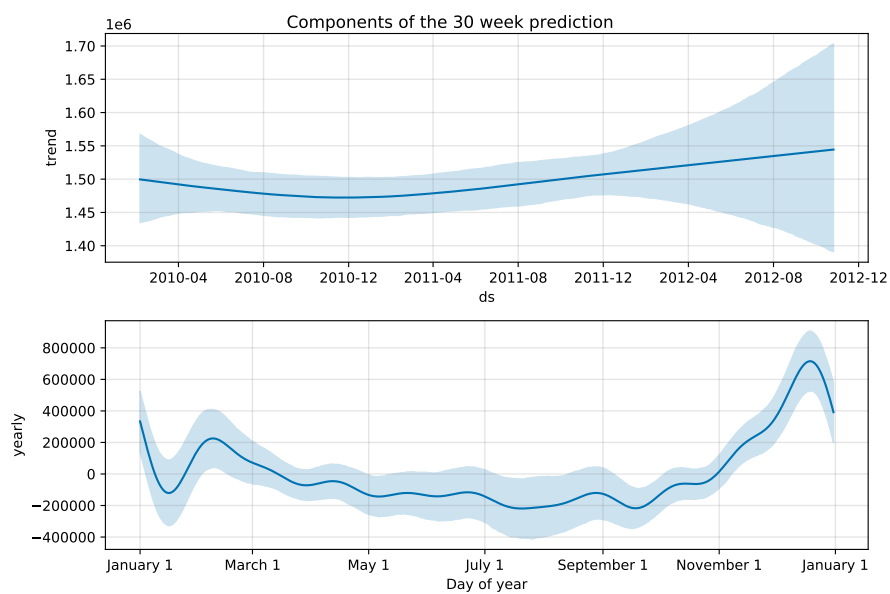


Figure 8: Components of the fitted model used in figure 7

A Esimerkki liitteestä

B Toinen esimerkki liitteestä