Adjusted Grid Search to Find Hyper-parameters in SARIMAX Models: Efficiently Filling The Shelves in Kruidvat Stores

William Steenbergen
University of Twente
w.a.t.steenbergen@gmail.com

October 21, 2018

1 Abstract

[1] Replenishment processes, promotions and assortment choices in retail are determined by forecasting sales. Sales are affected by external factors such as seasonality, weather or promotions. This relationship can differ per store or product. Currently, researchers focus on developing a SARIMAX model for one product and store specifically to predict sales, and the novelty of this research is that it develops an algorithm that automates SARIMAX modeling to allow for model generation tailored to a store and product. The algorithm was tested with a case study at Kruidvat. It was found that the algorithm works good enough to potentially save Kruidvat more than 5 million Euros annually.

Keywords
Time series analysis, SARIMAX, sales forecasting, automated hyper-parameter optimization, operations research

2 Introduction

Nowadays, companies spend a lot of attention on incorporating data analysis into their decision-making. Research has found that companies that inject big data and analytics into their operations outperform their peers by 5 percent in productivity and 6 percent in profitability [5]. Especially in retail, characterized by large amounts of available data and complex operations, there is substantial potential in using data analytics.

One of the most prominent ways of how data analytics can be useful for retailers is by making store replenishment more efficient. Retail stores must secure enough products on their shelves while minimizing transport-, storage- and labor- costs. Especially now physical stores have to compete with online webshops, which practically have unlimited stock, having enough items on the shelf is crucial.

The most important component of optimizing store replenishment is the forecast of sales. One method that is often used to forecast sales is time series analysis. Time series models have the advantages that they have better interpretation with reasonable accuracy, and are easier to implement compared to competitive approaches [7].

One of the most commonly used time series model is the Seasonal Autoregressive Integrated Moving Average model including exogenous variables (SARIMAX)[2]. The SARIMAX model describes by a regression with SARIMA errors u_t and exogenous variables [2]:

$$\begin{gathered} Sales_t = \sum_{n=Monday}^{Saturday} D_{n,t}\phi_n + \\ \phi_{Prec}Prec_t + \phi_{MaxT}MaxT_t + u_t \\ where \\ \Phi_p(L)\Phi_P(L^s)\Delta^d\Delta_s^D u_t = \Theta_q(L)\Theta_Q(L^s)\epsilon_t \\ D_{Monday,t}\begin{cases}1 & \text{t is a Monday} \\ 0 & \text{else}\end{cases} \end{gathered} \tag{1}$$

L is the lag operator, and is defined as:

$$u_t L = u_{t-1} \tag{2}$$

Δ_s^d is called a difference operator and is defined as:

$$\Delta_s^d = (1 - L^s)^d \tag{3}$$

p, q, P and Q are the (seasonal) lag orders. They determine which lags are taken into account when estimating the sales. p determines the amount of auto-lags in the model (the AR order) and q the amount of error lags (the MA order). P and Q determine the seasonal AR and MA order. d determines how often the time series has to be differenced with lag one in order to make it stationary. ϕ and θ are the coefficients, they determine what the influence of the corresponding lag is on the predicted value. These coefficients can be estimated by maximum likelihood estimation. p, d, q, P, D and Q are also called the hyperparameters of the SARIMAX model.

[2] This summary does not cover the time series theory in depth. Please see the complete paper for more information.

ϕ_{Monday} is the coefficient for sales on Monday, and is defined as the average difference from sales on a Sunday. The other coefficients work similarly. $Prec_t$ is the amount of precipitation in 0.1mm on day t. $MaxT_t$ is the maximum temperature measured in 0.1 degrees Celsius on day t. ϕ_{Prec} and ϕ_{MaxT} are the coefficients for precipitation and maximum temperature respectively and represent the effect of 0.1mm more precipitation and 0.1 degree Celsius more. Equation 1 shows that a SARIMAX model describes sales by a regression with SARIMA errors u_t.

Most research on SARIMAX models that forecast sales in retail proposes a model that can only be applied to one specific product or store (e.g. [1] and [4]). This research aims to bring SARIMAX modeling to forecast retail sales to the next level by developing a program that allows the user to apply product and store specific SARIMAX models to many different combinations of products and stores. The forecasting program automatizes decisions about the SARIMAX model that researchers would normally make empirically. This automation comes with the cost that the resulting model might not perform as well as a thoroughly analyzed one, but it does allow for the possibility of forecasting several time series with specific models. The practicality of this is illustrated by a case study at Kruidvat, the largest health and beauty retailer in the Netherlands. Kruidvat faces the problem that their shelves are filled inefficiently. Forecasting the sales of all product and store combinations by the use of the algorithm can help solve this problem.

2.1 Case study at Kruidvat

The applicability of the sales forecast is that one can use the sales forecast to forecast when to fill the shelves in the stores. If this is known, employees can save a lot of time since they will not waste time trying to fill products when it is not needed. On average, employees spent 59% of their time filling shelves trying to fill shelves that are already full (measured by going to the store and timing the actions of 5 different employees). It should be noted that this is a rough estimation, and should be further tested to give more reliable results.

Since the main disadvantage of the proposal is errors in the forecast, the performance of the proposal heavily depends on the performance of the forecasting program. The program is meant to forecast sales, while Kruidvat needs the forecast of when to fill shelves. Therefore, this thesis also developed a tool that incorporates the algorithm to forecast the shelf replenishment moments, and can communicate it to the store manager.

The tool was tested in a store, by questioning employees that have the task to fill shelves on the usability of the tool.

3 The program

The program[3] is built around SARIMAX models that include the weather, day of the week, history of sales and promotions to forecast the sales for the upcoming 7 days. To take into account the weather, historical precipitation and maximum temperature per day was used (extracted from [3]). The day of the week is modeled using dummy variables as shown in section 2. Products can be in multiple promotions during one day. To include promotions in the model, a variable $Prom_t$ was created that represents the number of promotions a products is in during day t. This leads to the model in equation 1.

The program must work generically so that it can estimate the best lag orders and best parameters for any combination of product and store. This means that it must work on different sets of data that do not necessarily have the same characteristics like stationarity and normality. Sales are forecasted by day, which means that lags are defined as days in the past (e.g. lag 7 is the sales of 7 days back).

The program works as following. First, the data must be made stationary. To ensure stationarity, an iterative algorithm was developed that uses the Dickey-Fuller test to determine whether the data must be (seasonally) differenced or the logarithm must be used. After the data is made stationary, it must be determined what lags and exogenous variables to take into account for the specific product and store combination. This is done by testing all kinds of different models in an intelligent way. Insignificant variables are taken out of the model if necessary, and the AIC criterion is used to determine what model scores best. After a model was chosen, a depth first search was used to remove insignificant endogenous variables without harming the AIC score of the model too much.

When the sales are predicted with the model as specified above, the timing of refilling the shelf needs to be forecasted. To forecast when to fill a shelf, shelf capacity, the sales prediction, theft, breakage and loss data and presentation quantity was taken into account. The presentation quantity (Cap_{pres}) is specified by Kruidvat and determines at what percentage of shelf the total shelf capacity the shelf should be filled. For example, if the presentation quantity is 30%, the shelf should be refilled when 70% of the shelf is sold. To reduce the risk associated with forecasting errors, a security input was installed in the algorithm. This security input lets the user determine how much risk you accept to have a shelf lower than the presentation quantity.

[3]The code can be found at https://gist.github.com/WilliamSteenbergen/754c9b6ff05ba8484189332474696b31

4 Results and testing

The algorithm has been tested on 20 products (that have been carefully selected to represent most products) in 5 different stores, giving a test sample of 100 cases. The test instances used half a year of data and makes predictions for dates that have a known true value. Different versions of the program have been tested, using the AIC, BIC and RSS [2] criteria to compare models with different lags. The performance of the different versions are measured by summing the absolute value of the 7 forecast errors and dividing by the actual number of times that should be filled (see table 1):

$$\frac{\frac{1}{n}\sum_{p}^{n}\sum_{t=1}^{7}|Forecast_{t,p} - Actual_{t,p}|}{\frac{1}{n}\sum_{p}^{n}\sum_{t=1}^{7}Actual_{t,p}} \quad (4)$$

With n being the number of product-store combination, p being a product-store combination and t being a day in the forecast range. $Forecast_{t,p}$ is the forecasted number of shelf refills of product-store combination p on day t. $Actual_{t,p}$ is the number of shelf refills if future sales would be perfectly known. To give insight in what the risk is for Kruidvat to have an empty shelf, it was also measured how often the algorithm advices employees to fill too often (table 2).

	Security = 50	**Security = 65**
RSS	21%	25%
	22%	24%
AIC	18%	22%
	21%	24%
BIC	21%	25%
	21%	24%

Table 1: Performance score for different versions of the algorithm. All versions are also tested on two different security levels. Cap_{pres} was chosen to be $0.3 \cdot Cap_{max}$. The two values indicate the forecast error by using the best first search to remove insignificant variables and not using it respectively.

	Security = 50	**Security = 65**
RSS	61%	75%
	62%	76%
AIC	61%	76%
	61%	75%
BIC	61%	77%
	61%	76%

Table 2: Percentage of the errors where the algorithm proposes to fill more than is actually needed.

All versions tested barely show significant correlation between sales and precipitation and sales and the maximum temperature. Most models did have significant correlations with the day of the week and promotions. On average, the algorithm took 417 seconds to make a prediction about one product at one store when removing insignificant endogenous variables. When not removing insignificant endogenous variables, the algorithm took 219 seconds on average.

In consultation with Kruidvat, it was assumed that the average employee spends 10 hours per week on filling shelves (not including driving the carts around). 59% of this is wasted due to the problems as described in section 2.1. Assuming that the method costs 1 hour extra per week of redistributing products over the carts, and furthermore 75% of the problem, it saves an average employee $((0.59 * 0.75) * 10) - 1 = 3.5 hours/week$.

5 Conclusion

This research developed an algorithm that determines the best SARIMAX model given a certain store and product to forecast sales in retail. The algorithm was used in a case study at Kruidvat to test for its applicability and accuracy of forecasting. Found was that the algorithm gives the best forecast when using the AIC to choose a model and removing insignificant endogenous variables afterwards. It can solve the problem at Kruidvat with decent accuracy, having an error score of around 18%. Given that safety measures can be incorporated to decrease the risks that are inherent to errors in the forecast, the algorithm seems to be appropriate for practical use.

If implemented in all stores in the Netherlands, this could save more than 5 million Euros annually (approximately: 50 weeks $*$ 3.5 hours $*$ 5000 employees * 6 euros per hour). On top of that, it could decrease the probability of a shelf being empty or looking messy. Finally, the algorithm can give Kruidvat insight and choice in the probability of shelves being empty, and gives them a better understanding of the relations between external variables like the weather, promotions and seasons and their sales.

6 Discussion

First of all, the program could be improved by decreasing the computing time. Most time is spent in the grid search and the depth first search. Decreasing the time of the grid search is difficult without decreasing the number of models the algorithm tries, since most time is spent running the extensively researched and optimized .fit function [6]. The best first search could be programmed more efficiently by for example cutting of nodes that have a low potential sooner. One possibility is to decrease the 5% bottleneck or decrease the depth from 3 to 2. The risk of doing this is that cutting off nodes too early

can lead to missing good solutions. More research has to be done in the decision between speed and search depth to optimize this part of the algorithm. As can be seen in table 1, not doing the depth first search at all decreases the forecasting accuracy, but also decreases the computing time with on average 198 seconds per product. On top of this, the program could be coded more efficiently, which would greatly decrease computing time.

The program is now only tested in a case study at Kruidvat. To see if it also works for other situations, it should be tested in other retail companies. When testing with other retailers it can be promising to include other exogenous variables that are more applicable to those retailers. Moreover, this research tested the program by using it in a tool that forecasts days to refill shelves, but this is not the only applicability of the algorithm. Examples of these applicabilities could be to optimize replenishment from distribution centers, optimize promotion timing or improve the planogram.

Although there might be some downsides to the solution proposed in this research, there is enough evidence that it performs better than the current situation. Forecasting allows for more efficient time planning, it decreases the risks of shelves being empty, and it gives Kruidvat insights about why products sell.

7 Role of the student

The complete program and tool as described above was developed by myself. I worked together with Kruidvat employees to gather data. I brainstormed with senior management and employees in the store to come up with a new procedure to refill shelves. This does not only contain a mathematical model, but also a practical procedure of how employees can refill shelves more efficiently (this is not so much elaborated on in this summary). I came with the idea of automatizing SARIMAX modeling, and also with the idea of restructuring the way shelves are refilled. My supervisors from the University of Twente helped me writing my report and understanding theory about SARIMAX modeling. I did this thesis as part of my bachelor University College Twente, majoring in econometrics and operations research. I was awarded an 'excellence' for this thesis, the highest possible grade within University College Twente.

References

[1] C. P. Da Veiga, C. R. P. Da Veiga, A. Catapan, U. Tortato, and W. V. Da Silva. Demand Forecasting in Food Retail: A Comparison Between the Holt-Winters and ARIMA models. *WSEAS Transactions on Business and Economics*, 11(1):608–614, 2014.

[2] P. H. Franses. *Time series models for business and economic forecasting*. Cambridge university press, 1998.

[3] KNMI. Daggegevens van het Weer in Nederland. https://www.knmi.nl/nederland-nu/klimatologie/daggegevens, 2018. Accessed: 2018-04-28.

[4] J. B. Marin, E. T. Orozco, and E. Velilla. Forecasting Electricity Price in Colombia: A Comparison Between Neural Network, ARMA Process and Hybrid Models. *International Journal of Energy Economics and Policy*, 8(3):97–106, 2018.

[5] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton. Big Data: the Management Revolution. *Harvard Business Review*, 90(10):60–68, 2012.

[6] S. Seabold and J. Perktold. Statsmodels: Econometric and Statistical Modeling with Python. In *9th Python in Science Conference-https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax. SARIMAX.html*, 2017.

[7] M. Shukla and S. Jharkharia. Applicability of ARIMA Models in Wholesale Vegetable Market: an Investigation. *International Journal of Information Systems and Supply Chain Management*, 6(3):105–119, 2013.