Erasmus University Rotterdam
erasmus school of economics

Master thesis Econometrics and Management Science

Program: business analytics & quantitative marketing

Predicting sales peaks of weather related products by a STAR model

Author: A.M. Dijkshoorn (431936)

Supervisor: prof. dr. R. Paap

Second assessor: dr. P. Wan

Abstract

Predicting sales is important for e-commerce companies to meet the demand of products. Accurate sales forecasts can lead to less excess stock and less lost sales, which cuts the costs and increases revenue. In this thesis, it is shown how a STAR model is estimated and how forecasts on STAR are made. Sales (particularly, sales peaks) of weather related products are predicted by means of a STAR model for different forecasting steps. Adding temperature as a variable to time series models improves forecasting accuracy for weather related products. The best forecasting technique to STAR is where we include Monte-Carlo simulated error terms. Consequently, this cuts the costs of excess stock and lost sales compared when we use default AR model forecasts.

April 30, 2022

Contents

1 Introduction **1**

2 Literature **4**

2.1 Relevance 4

2.2 Methods 5

2.2.1 Modelling the effect of weather on sales 5

2.2.2 STAR models 6

3 Data **7**

3.1 Product sales time series 8

3.2 Sales and temperature 11

4 Methodology **14**

4.1 Defining a smooth transition autoregressive (STAR) model 14

4.2 Building the model 15

4.2.1 Specifying the order of the linear AR part 16

4.2.2 Testing for STAR nonlinearity 17

4.2.3 Parameter estimation 18

4.2.4 Diagnostic checks and model modification 19

4.3 Forecasting with STAR models 21

4.3.1 Point forecasts 21

4.3.2 Direct forecasting 23

4.3.3 Evaluation of forecasts 24

4.4 Extension to the STAR model 25

4.5 Excess stock and lost sales time series 26

5 Results **27**

5.1 Investigating the time series 28

5.2 Making a STAR model for fan sales 29

5.3 Economic value of the new sales model 37

5.4 Sales models for other weather-related products 40

6 Conclusion **43**

7 Limitations and further research **44**

References **46**

A Tables and figures **49**

B Theory **53**

1 Introduction

In this thesis, sales prediction of specific product groups of an e-commerce company in Europe is researched, and this company is referred to as *the e-commerce company* in this entire work. In the e-commerce sector, products are sold online. At the e-commerce company, external partners are also able to sell their products on the platform. It is important for e-commerce companies to maintain enough stock, such that customers do not miss out on buying their desired products. On the other hand, it is important not to have too much stock, because having stock costs money. This is especially important during a phenomenon called sales peaks: a high peak in the amount of sales in a period of time, deviating from other periods around the peak.

At the e-commerce company, it sometimes occurs that products in certain product groups are sold much more than in a previous period. For example, sleds whenever snow is forecast, swimming pools whenever hot weather is forecast, toys getting popular suddenly, face masks after new measures by the government, etc. These sales peaks occur frequently. At the e-commerce company, three main drivers of sales peaks are:

1. Seasonal peaks, like Thanksgiving and Christmas;
2. Climate peaks, like high temperatures and high precipitation;
3. Trend/hype peaks, like fidget toys[1] and face masks.

Even though all these types of peaks are of importance to the e-commerce company, climate peaks are not completely integrated yet in the e-commerce company's strategy. Nevertheless, obtaining an insight in when these peaks occur and in what quantities products are bought during these peaks brings major advantages to partners and logistics. It is known that when temperature increases above a certain threshold (or, e.g., when a heat wave occurs), sales of products such as swimming pools and fans will rise. However, we do not know for which exact temperatures this holds, and moreover, to what extent the sales are influenced. Does an increase in temperature of one degree Celsius lead to a sales increase of 50%, 200%, etc. If the e-commerce company or a partner knows the demand beforehand, they can anticipate on this by matching their stock to the demand. In this research, we will make an econometric model such that we can predict sales and possibly sales peaks.

[1] Fidget toys are brightly colored silicone toys that went viral on social media apps like TikTok, YouTube and Instagram. At the e-commerce company, fidget toys are sold in the chunk 'fidget'. Its sales increased heavily in the beginning of 2021.

Forecasting climate driven sales peaks is of relevance for the e-commerce company. They need these sales predictions for their own forecasts, marketing, delivery times, etc. Moreover, the e-commerce company needs this to inform their partners. Partners do not have an insight on the total sales of a product or product group, and hence they are limited to the sales of their own products. Moreover, partners do not have the capacity to make forecasts of the sales and are therefore not able to anticipate on sales peaks.

With the forecasts, we can make predictions of the amount of sales in the next days. With these predictions, the e-commerce company and its partners can order stock such that we can meet the demand of a product with the amount of stock. When predicted sales are more accurate, then we can maintain a stock that matches better with the demand. On the one hand, this leads to a reduction of excess stock, which cuts the total costs since having stock requires storage. On the other hand, it reduces the amount of "lost" sales (i.e., products that are in demand but out of stock), which means that either the products are bought at a later time or not at all occurs less, thus improving the revenue.

In this thesis I will focus on sales peaks due to climate, particularly how temperature affects sales of products and product groups that are related to the weather during a peak period. These sales peaks are modelled using previous sales and daily temperature. After making a sales model, the forecasting performance is measured, and it is investigated how this affects the total costs. The research question is therefore:

"Can sales peaks of weather related products or product groups be predicted by weather data and/or weather forecasts, and if so, how can this be modeled?"

Data that is used in this paper consists of aggregated daily sales of products in a product group, or chunk. A total of five chunks are considered here. Furthermore, as the climate impact we use daily maximum temperatures in the Netherlands. Modeling and forecasting of sales and its peaks in this thesis will be based on the smooth transition autoregressive (STAR) model, a nonlinear time series model. The STAR model consists of two separate AR series containing sales of previous days, and these AR series are then linked by a transition variable. The sales are predicted as a weighted combination of the two AR series. The weights depend on the transition variable: if this variable is below a certain threshold value then the weight on the first AR series are closer to 1 and the second weight is equal to 1 subtracted with the other weight (and vice versa). The set of observations for

which the transition variable is below $\frac{1}{2}$ is referred to as a low regime, and for higher than $\frac{1}{2}$ it is called a high regime. The transition variable can be either a combination of (differenced) lagged variables, or an external variable. In this paper, the external variable is the maximum temperature.

With the STAR model, we can make forecasts of the sales. Forecasting methods include a naïve method, Monte-Carlo, bootstrapping and direct forecasting. With the best forecasting method, we make a business model which determines the amount of excess stock and lost sales when applying such a forecasting method to the amount of stock ordered.

Usually, a series of observations in which sudden shocks (large changes in volume) occur are hard to model. The STAR model however can adequately capture these shocks by transitioning between regimes. Sales peaks showcase the same behaviour, and therefore STAR is appropriate. STAR has been researched for quite a long time. It was introduced by Luukkonen et al. (1988). Further developments are performed in Teräsvirta and Anderson (1992), Teräsvirta (1994), and Eitrheim and Teräsvirta (1996). Subsequently, van Dijk (1999) provides an extensive research on the STAR model and many extensions on it. Later on, many applications of STAR are performed (Alimi et al., 2017; Hsu & Chiang, 2011). Though, as far as my knowledge goes, no literature exists on the application of the STAR model on how retail sales are predicted by previous observations and weather data. Therefore, this thesis is a unique contribution to literature. Moreover, the economic effect of using these forecasts is captured by a unique business model.

In this paper, we find that including temperature as a variable in a sales forecasting model improves forecasting accuracy. For instance by adding temperature as a linear component to an AR model, or by using the temperature as a transition variable in STAR. Moreover, it follows that using a STAR model with the Monte-Carlo forecasting method is generally the best option to predict sales accurately. In terms of MSPE, the error is about 50% of the MSPE of an AR model, and for MAPE it is often less than 10%. Also, using STAR forecasts cuts the costs (of excess stock and lost sales) compared to AR. How much depends on how many days it takes for stock to arrive at the warehouse, and about which chunk the sales are.

The paper is organized as follows. section 2 outlines relevant academic papers related to describing sales by weather effects. In section 3 we inspect the sales and temperature data. section 4 provides theory on STAR models and describes the method used to determine the business model. The results of applying the data to the methods are given in section 5. In section 6 we make a conclusion based on the results and finally, in section 7, we discuss the limitations of this research.

2 Literature

This section outlines relevant academic papers related to describing sales by weather effects, particularly to describing sales peaks of seasonal products on temperature. Here, seasonal products are products for which its sales rely heavily on (high) temperatures and other weather effects.

2.1 Relevance

The influence of weather economic outcomes such as industrial output, energy demand, economic growth among other outcomes is researched extensively (Dell et al., 2014). Particularly, in this paper, research that models the economic effect due to the increasing threat of climate change is summarized. However, little research about weather and retail sales is found in the retail literature (Bahng & Kincade, 2012; Verstraete et al., 2019). To the best of my knowledge, there is no work on modeling sales and temperature by means of a STAR model. Therefore, this thesis will be a unique contribution to existing literature.

Even though this thesis investigates specifically the relation between temperature and sales peaks by means of the STAR model, it might inspire other related subjects to be researched as well. For instance, other sales peaks by exogenous shocks other than weather effects are also appropriately modeled by STAR. Even more generally, any peak related data may be modeled by STAR.

Sales forecasting in general is important in guiding the sales and marketing of e-commerce and warehousing (Liu et al., 2020). Moreover, sales data reflect future sales well. Forecasting future sales is a fundamental issue that guides all strategic and planning decisions in operations of retail businesses (Ramos et al., 2015). In order for retail businesses to be profitable, demand forecasts must be accurate, as it is crucial in aspects such as production, purchasing, transportation and labor force (Ramos et al., 2015). This is especially true for seasonal products which display peak behaviour, as it requires the aforementioned aspects to be extra alert due to sudden changes.

In fashion markets, selling periods of (some of) the products are likely to be short, estimated in months or weeks (Christopher et al., 2004). This holds for products that are heavily influenced by weather effects as well. Careful timing of the sales is therefore important for retailers. If products are presented or promoted too early, the product will not sell. If products are offered too late, demand remains low and retailers might need to reduce prices to keep consumers motivated to buy. If correct planning is not taken into consideration, unsold stock will remain at the end of the season (Al-Zubaidi & Tyler, 2004).

2.2 Methods

Techniques and models to forecast retail sales have been investigated extensively (Verstraete et al., 2019). Many include models based on historic observations, such as exponential smoothing models (Hyndman et al., 2002) and autoregressive integrated moving average (ARIMA) models (Permatasari et al., 2018; Ramos et al., 2015). These models are accurate for regular sales (i.e., sales are independent of external shocks). However, these models are not appropriate for nonlinear behaviour (Wang et al., 2013).

2.2.1 Modelling the effect of weather on sales

Even though literature on how retail sales are predicted using weather data is scarce, research has been done since a long time. Steele (1951) developed a multiple regression equation where sales are explained by weather variables such as precipitation, snow cover, temperature, etc. These weather variables improved the accuracy of the predictions of retail sales. Steele (1951) states that the main difficulty lies in forecasting the weather one or more days ahead. Weather forecasts have now been improved evidently compared to 1951.

Literature on forecasting energy demand and predicting retail sales are closely related, due to their directly measurable effect and to their similar aggregation level (Verstraete et al., 2019). Cui and Peng (2015) analyze the relationship between daily temperature and electricity load (i.e., energy demand) in a city. External temperature effects cause mutation structures in load data. To deal with this, an improved ARIMAX model is proposed. This model proves to be an effective method for short-term load forecasting, as it predicts more accurately than traditional time series models.

Steinker et al. (2017) performed a research on how daily weather conditions affect retail sales or demand. Their analysis involves a baseline model specifying ARIMAX models that allow to carve out the weather effect and estimate its potential to improve the short-term sales forecasting accuracy. An extended model is made by augmenting the baseline model with weather variables and weekend interaction variables. Steinker et al. (2017) discuss that other advanced techniques can help to account for nonlinearities.

Verstraete et al. (2019) provide a data-driven framework for predicting weather impact on high-volume low-margin retail products. In this paper, the proposed methodology that handles short-term weather influenced retail sales consists of selecting the best method out of six methods. The methods include machine learning techniques, Poisson regression, LASSO regression, (a combined) LASSO Poisson regression, neural networks, support vector regression and gradient boosting. Pre-

dicting is done by leave-one-period-out cross validation, and to select the best forecasting model, the metrics mean absolute error, root mean squared error and mean error are used. The LASSO Poisson regression model outperformed the other techniques both in out-of-sample forecast error and bias.

2.2.2 STAR models

Retail sales possess a time series character and therefore a time series model might be appropriate. As the objective is to model sales peaks, there is also a nonlinear behaviour involved. Teräsvirta and Anderson (1992) assume that a nonlinear time series can be adequately described by a smooth transition autoregressive (STAR) model. STAR is a model containing two different autoregressive models and a smooth transition function inbetween to switch between different 'regimes' or 'states' in a continuous manner.

(Luukkonen et al., 1988) introduced the STAR model as a special case of the self-exciting threshold autoregressive (SETAR) model (Chan & Tong, 1986). A problem that occurs with the SETAR model is that tests for linearity require simulation for each application individually (Luukkonen et al., 1988). Even though STAR is nonlinear as well, the traditional linear autoregressive model is nested within the STAR model, which makes testing for linearity suitable.

Two families of nonlinear autoregressive models, particularly two specifications of STAR, are logistic STAR (LSTAR) and exponential STAR (ESTAR), proposed by (Teräsvirta, 1994). These models specify the transition function. Teräsvirta (1994) describes a procedure on the selection of an appropriate model by first specifying the autoregression, then testing for linearity to determine the delay parameter, and lastly performing a sequence of F-tests to choose between LSTAR and ESTAR. Even though Teräsvirta (1994) makes strict assumptions (e.g., if linearity is rejected, then STAR is the only alternative), the techniques still work well, as LSTAR and ESTAR cover quite a few nonlinear time series.

Eitrheim and Teräsvirta (1996) complement the article of Teräsvirta (1994) by addressing three tests for evaluating an estimated model. First, a test of serial independence of the errors is described. Second, a test that evaluates the adequacy of the selected STAR model describing nonlinearity in the data is presented. Third, a test that checks the assumption of constant parameters is provided to detect potential misspecification.

Teräsvirta and Anderson (1992) and Teräsvirta (1994) use a STAR model without the possibility of using exogenous variables. Indeed, the transition function has a transition variable as argument,

and this variable is restricted to an endogenous lagged variable. The estimated models indicate that deep contractions are caused by exogenous shocks, and therefore an extension of STAR with exogenous variables is more appropriate for modeling sales peaks with an exogenous driver such as temperature. On the other hand, van Dijk (1999) does allow for exogenous variables in the STAR model, both linearly and in the transition function as the transition variables itself or as a transformation (i.e., a function that takes both lagged endogenous and exogenous variables).

van Dijk (1999) states that the STAR model brings two characteristic advantages. First, the regimes can be associated with certain observable variables, such as a recession or in our case a sales peak. Second, the transition between regimes are continuous instead of abrupt changes, which is more in congruence with realistic and actual data.

STAR only allows for two autoregressive regimes. This is done to characterize the different dynamics between different quantities of sales. However, there might be more than one discrepancy in the dynamics. We can capture more different dynamics by adding more regimes in an extended version of STAR: the Multiple Regime STAR (MRSTAR) model (van Dijk, 1999). In this paper it is shown that the MRSTAR model encapsulates many regime-switching mechanics, e.g. by showing that it even nests a single hidden layer artificial neural network (ANN).

Another special case of MRSTAR is the Time-Varying STAR (TVSTAR) model. TVSTAR can be used for modelling multiple nonlinearities at the same time, long-term changing observational values, and making a difference between these two features (Lundbergh et al., 2003). This could be useful for modelling sales, because there might be multiple dynamics that drive the sales simultaneously, and moreover, the average amount of sales might increase (perhaps gradually) over a long period of time.

3 Data

In this section we will inspect the sales and temperature data. Visualizing the sales data will help us understand the behaviour of this series, and to see if transformations of the time series are necessary. Visualizing or making computations of the relationship between the sales and temperature together may manifest whether temperature affects the sales significantly, thus telling us whether we should include temperature in the model.

3.1 Product sales time series

The e-commerce company's assortment is divided into categories, containing multiple layers. In descending order, the assortment is divided into cluster, shop, productgroup, productsubgroup, productsubsubgroup, brick and chunk, where chunk is the smallest and thus most specific category. Every product from the e-commerce company has a unique number, a globalid. For every globalid, content such as the volume, price, color, weight etc. is available in their database.

Moreover, data about every order is available, such as when the order was placed, if the order has been returned, a case was made, cancelled or delivered on time etc. The latter four constitute the base of how the e-commerce company measures customer satisfaction. Data from every order since 2018 is available up until 2020. Tens of millions of orders are placed every year on the e-commerce company. For every order, multiple products can be bought in various amounts, and this is displayed as the quantity ordered. Every ordered quantity contributes to the total sales of a product (or some category of that product). For this research, the total amount of sold products of a chunk on one day (i.e., the sales) is used.

Not all chunks are investigated in this research: only weather-related chunks are considered. More specifically, chunks that contain products for which its sales are heavily dependent of high temperatures. Unfortunately, this kind of information is not contained in the e-commerce company's database. However, we can extract this information by picking chunks that show relatively high increasing sales during hot weather. We pick the top forty chunks of a list of chunks with the highest percentage increase in sales of week 30 relative to week 29 of 2019, provided a minimum amount of sales in the first 30 weeks of 2019 and being contained in a list of keywords, such as: *sun, water, airco, fan, parasol* or *swim*.

One chunk is highlighted in particular for visualization purposes: fans. It is the highest selling chunk of the aforementioned list of chunks. A plot of the total unit fan sales time series is displayed in fig. 1. During the entire period from 1 January 2018 up to and including 31 December 2020, there are 1096 observations and the sales range from a minimum of 12 units to a maximum of 26244 units on a day.

Observations are most often near a constant mean, with the exception of the periods around the high peaks, at the late spring and the entire summer. During the entire period, the mean equals 794.2, while the median is 106. This indicates a right skewed distribution. In fact, most data is near the median, and the the high peaks are relatively far from the median, pulling the mean upwards. The skewness equals 5.51, so there is indeed a positive (right) skew. The kurtosis is $39.46 > 3$ and

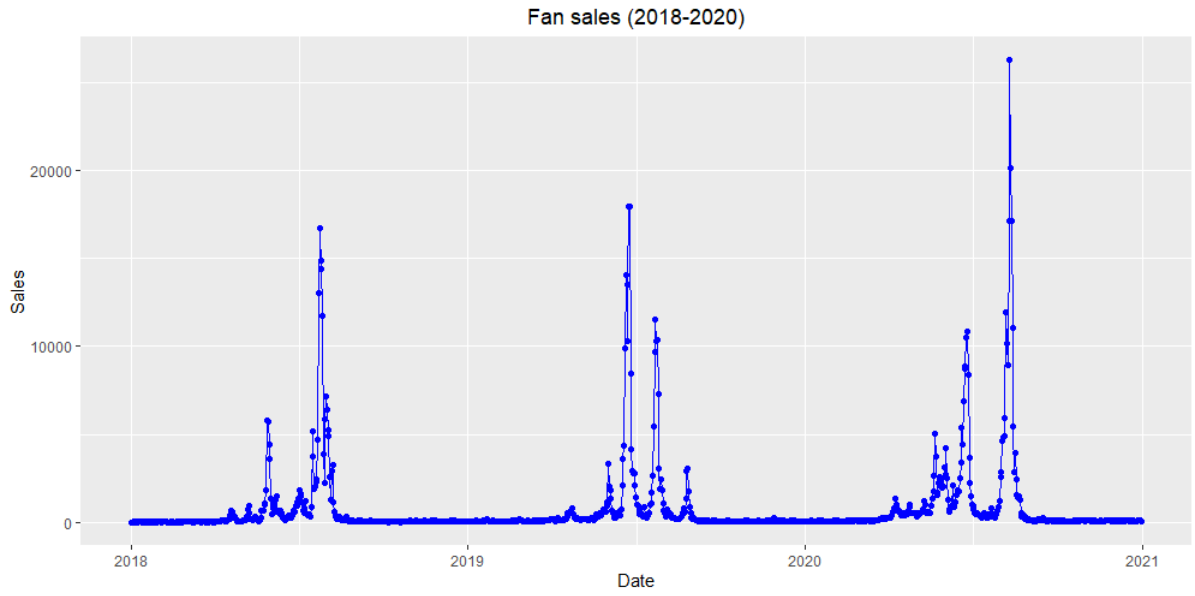

Figure 1: Time series of the e-commerce company's total unit fan sales starting in 2018 and ending in 2020.

hence the distribution is leptokurtic, indicating a thin peak.

Let us first consider some of the characteristics of the time series. We will look into the first two years of the sales observations in our sample, as this will be our estimation sample. The last year of our sample is used for forecasting purposes. To find out if a trend is present, we regress the sales y_t on a constant and the trend t, the number of days past the start of the series. The coefficient of the trend is 0.035 with p-value 0.337, which means that the coefficient is not significant and therefore we conclude that a linear trend is not present in the series.

The series are stable (in the sense that the sales volume fluctuate around a constant mean) throughout the fall and winter. During the late spring, the series start to show some peaks, and during the summer more peaks follow. This pattern is shown every year, where in the late spring and summer the mean and variance of the series is much higher than in the rest of the year, clearly indicating the seasonal character of the sales volume.

Furthermore, we want to know if aberrant observations are present in the series. Due to the relatively large difference between regular sales and high peaks, these peak observations are potential outliers. However, these peaks occur every summer and are therefore not rare or unexpected. Moreover, they are assumed to be caused by the (somewhat) predictable exogenous variable temperature. Therefore, the high peaks in the late spring and summer are not considered to be aberrant observations. Instead, they can be seen as contributions to the nonlinear pattern in the series and will be modeled by the STAR model in section 4.

We want to test whether the series is stationary, or in other words if the time series has no unit root. This can be determined by the Augmented Dickey-Fuller (ADF) test. It is the extended version of the Dickey-Fuller test, where we regress the first differences of the series on a constant, a trend and the first lag of the series (Cheung & Lai, 1995). The ADF test follows the same procedure, with the addition of lagged variables to remove autocorrelation from the series. The regression equation is

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{t-p+1} \Delta y_{t-p+1} + \varepsilon_t,$$

where p is the lag order of the autoregressive process. A trend is not present in the fan sales time series, hence we have $\beta = 0$ in our case. The test statistic for the ADF test is the t-value of the coefficient of the lagged value y_{t-1}, i.e., $\tau = \frac{\hat{\gamma}}{\mathrm{SE}(\hat{\gamma})}$. This statistic has their own critical values, known as the Dickey-Fuller table. For the ADF test with a constant, without a trend, and a sample size of 730, the critical value is -2.86 on a 95% confidence level. The τ-value for the fan sales time series equals -6.20. Since $-6.20 < -2.86$, we reject the null hypothesis of a unit root, and therefore the series is stationary. Thus, the series do not need a differenced transformation. We might however need a log transformation, and this is investigated in section 5.

We can quantify the relationship (or correlation) between observations and their lags by autocorrelation function (ACF) and a partial autocorrelation function (PACF). The ACF are the correlations between an observation and a set of its lags. The PACF are similar to the ACF, except that the correlation between shorter lag is removed. In other words, a correlation of lag p in the ACF might be partially explained by shorter lags such as lag $p-1, p-2, \ldots, 1$, whereas in the PACF the correlation is 'pure'. The ACF and PACF of the fan sales time series are plotted in fig. 2.

From the ACF plot, we see a high correlation for low lags. The correlation is decreasing immediately from lag 1, indicating less correlation with higher lags. From lag 10 onward, the correlations are not significant anymore. Furthermore, we see an increase of the correlation around lag 30, indicating a monthly seasonal pattern. The PACF has, unlike the ACF, an alternating pattern. The first lag has a high correlation of 0.914. The lags hereafter have smaller correlations but are still significant up to lag 8. We conclude that only the first 8 lags are interesting to include in a time series model, at least in a linear time series.

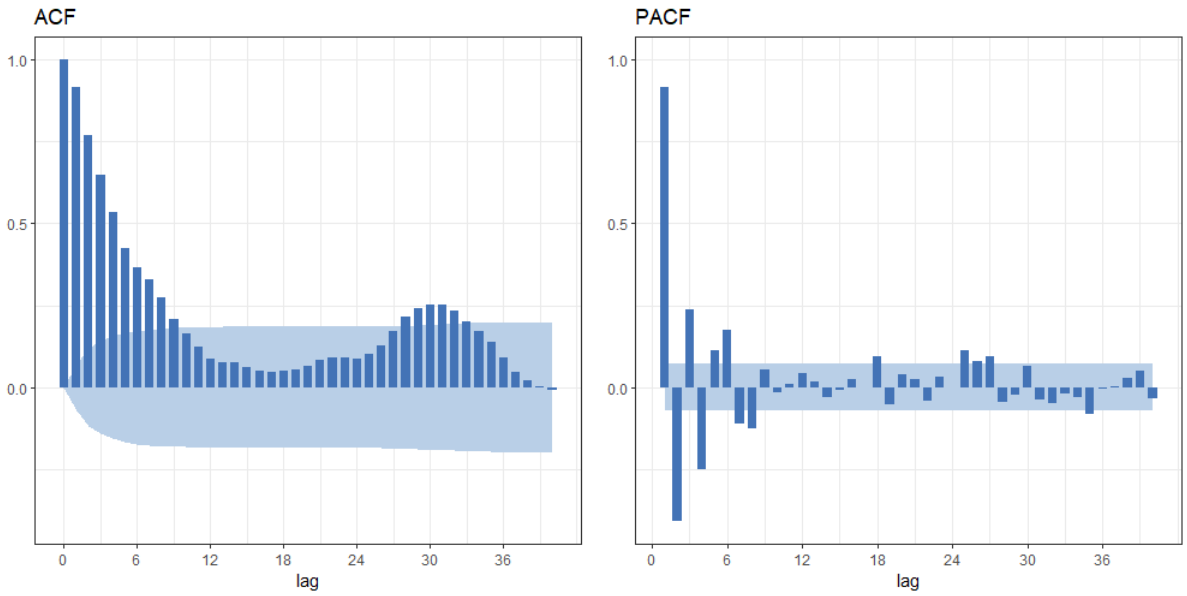

Figure 2: ACF (left) and PACF (right) plots of the fan sales time series. The lightly shaded blue area displays the significance area: bars that fall outside this region are significant correlations.

3.2 Sales and temperature

Besides using lagged sales to predict sales, we want to model the effect, if any, of temperature on the sales. For trying to predict sales peaks using weather forecasts, the Dutch national weather service Royal Netherlands Meteorological Institute (KNMI) is consulted. They provide an open to public database containing all daily weather information. Maximum temperature is measured in 0.1 degrees Celsius each day and is measured since 1 January 1901.

Using current temperatures might not be enough to accurately predict sales. Rather, temperature forecasts are required. Forecasts are however not always completely equal to the actual outcomes. It is therefore interesting to know the accuracy of these weather forecasts. Haiden et al. (2021) investigate the development of the quality of weather forecasts in Europe, see fig. A1. Here, the number of days for which forecasts are still useful and accurate is given. The 12-month average shows that this number is between 6 and 7 since 2006, and nearly 7 since 2018. We conclude that predicting up to 7 days is accurate enough.

To investigate the effect of temperature on sales, we first look into a combined plot of the sales and temperature time series, see fig. 3. This plot contains the again fan sales and the daily maximum temperature of only 2019. The reason for excluding the other years for which we also have sales data (i.e., 2018 and 2020), is that it is easier to visualize similarities or relations with just one year. And, moreover, each year shows more or less the same characteristics.

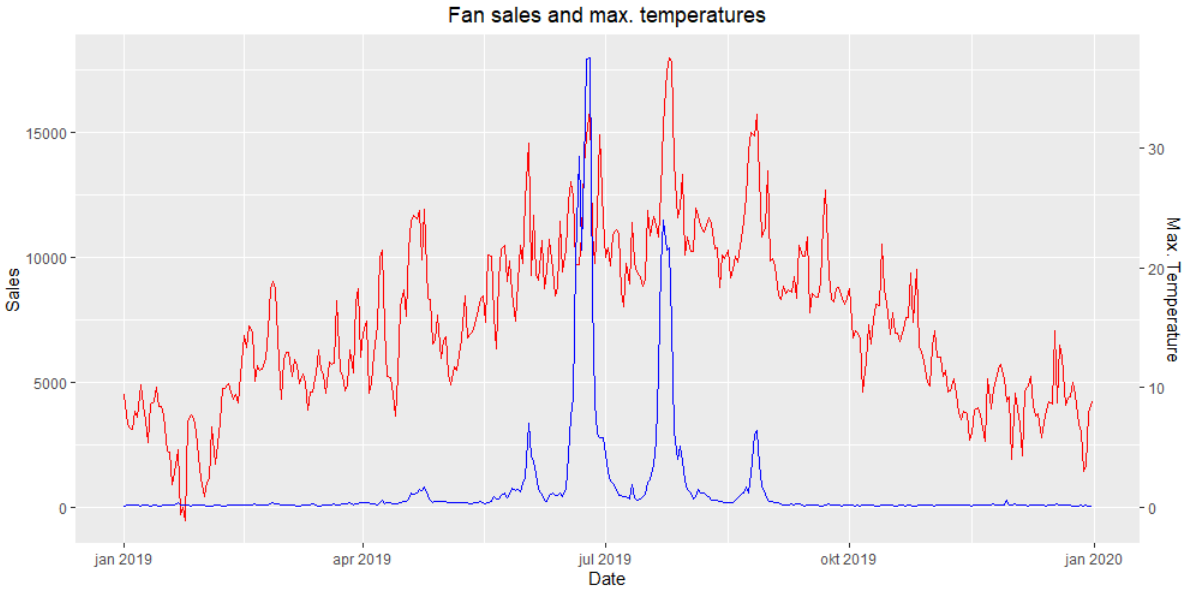

Figure 3: Time series of daily total fan sales (blue) and maximum temperature (red) of 2019. The left axis portray the sales, while the right axis portrays the temperature in degrees Celsius.

The first thing to notice about the relationship between temperature and sales is that the peaks of both time series coincide. There are four sales peaks, see table 1. Whenever one of these four sales peak occurs, we also see a peak in temperature. This indicates their relationship with each other. In fact, we see that temperature peaks are followed by sales peaks. In other words, sales peaks anticipate temperature peaks. This consumer behaviour is explained by the fact that consumers purchase products to fulfill their needs and desire. Consumers know that hot weather is coming beforehand due to weather forecasts, and therefore they anticipate on this by purchasing the products they want to utilize later in that period of high temperatures. Thus, high temperatures have a direct effect on sales, however, it is the weather forecast that is decisive.

Date	**Sales**	**Temperature**
2 June	3349	30.4
25 June	17966	33.2
23 July	11495	31.9
27 August	3062	32.8

Table 1: Total fan sales and maximum temperatures (in degrees Celsius) on the fan sales peaks in 2019.

Let us look into the period starting in the late spring and ending at the beginning of the autumn. Sales peaks occur only in this period. In the period before the midsummer, temperature

has a smaller effect on sales than in the midsummer, even though the temperature peak is not much smaller than later peaks. This indicates an exponential effect of temperature on sales. Another explanation is that consumers know that high temperatures occur less often before the summer, and therefore weather related products are used less than during the summer, and therefore these products are less desired. A similar explanation is used for the fact that the sales peaks are lower after the summer. At that time, consumers know that they will probably not need weather related products anymore in that year. Another explanation for this behaviour is that these products are already purchased often in that year, and hence consumers do not need the products anymore as they have already bought the products before. In conclusion, despite higher temperatures, the summer has a boosting effect on sales.

Outside the aforementioned period around the summer, the sales are stable and low. The temperature does however show much variation. This volatility must have a marginal effect on the sales during that period, compared to the summer. This gives rise to not only and exponential effect of temperature on sales, but also to use some sort of threshold value on temperature. For instance, we can subtract a threshold value of 25 on every temperature, such that for positive values, changes in temperature have a large effect on sales, and for negative values, changes in temperature have a small effect on sales.

Finally, we calculate the Pearson correlation coefficient, a measure of the linear correlation between two time series. We can compute this correlation between the sales and temperature of the same day: 0.473. Furthermore, since the temperature is known beforehand by weather forecasts, we can also look into the correlation between the sales and a j-day lead temperature. The highest correlation is attained for $j = 1$ (for $j \leq 7$), with a correlation of 0.458. Moreover, we can also calculate the sales and the largest maximum temperature of the next j days (like a rolling window). The highest correlation is attained for $j = 3$ (for $j \leq 7$), with a correlation of 0.496. It follows that if we use the highest maximum temperature of the upcoming 3 days, we attain the highest correlation with the sales.

It should be noted that previous conclusions that are drawn hold for the fan sales of 2019. These conclusions do not have to hold for other time series. Thus, these features should not be fixed in a model. Rather, they should be optional, for instance using parameters that are estimated.

4 Methodology

The (most) discussed model in section 2 is the STAR model, and this model will be explained in this section, as the STAR model is the method that will be applied in this research. The model, techniques and tests are prominently based on Teräsvirta (1994), van Dijk (1999), and van Dijk et al. (2002).

4.1 Defining a smooth transition autoregressive (STAR) model

The smooth transition autoregressive (STAR) model is a nonlinear extension of the standard linear autoregressive (AR) time series model. It describes the time series process of univariate variable y_t observed at $t = 1-p, -p, \ldots, 0, 1, \ldots, T$:

$$y_t = \phi_1' x_t (1 - G(s_t; \gamma, c)) + \phi_2' x_t G(s_t; \gamma, c) + \varepsilon_t, \quad t = 1, \ldots, T, \tag{1}$$

Here, $x_t = (1, y_{t-1}, \ldots, y_{t-p}, z_{1t}, \ldots, z_{kt})'$ is a vector of the endogenous lagged variable y_t and an exogenous variable z_t, and $\phi_i' = (\phi_{i,0}, \ldots, \phi_{i,m})$ are the fixed parameters, for $i = 1, 2$ and $m = p + k$. ε_t describes a martingale difference sequence, i.e., its expectation with respect to its past and the exogenous variables is zero: $\mathrm{E}(\varepsilon_t | x_t) = 0$.

Let $F(x_t; \Omega) = \phi_1' x_t (1 - G(s_t; \gamma, c)) + \phi_2' x_t G(s_t; \gamma, c)$ denote the fitted value of eq. (1), where $\Omega = \{\phi_1, \phi_2, \gamma, c\}$ is the parameter space. From here, we assume that the exogenous variable z_t is excluded from x_t. The model is compactly written as $y_t = F(x_t; \omega) + \varepsilon_t$

Furthermore, G is the transition function: it takes as argument the transition variable s_t, has parameters γ and c, and is bounded between 0 and 1. The transition function is continuous. If $G(s_t; \gamma, c) = 0$, then eq. (1) reduces to $y_t = \phi_1' x_t + \varepsilon_t$, an autoregressive model in past observations and exogenous variables, and we refer to this as the base time series. If $G(s_t; \gamma, c) = 1$, then eq. (1) reduces to $y_t = \phi_2' x_t + \varepsilon_t$, a similar autoregressive model with different parameter values, and we refer to this as the alternative time series. The STAR model ensures that there is a smooth transition between the base and alternative time series. These time series can also be seen as regimes, as the complete model consists of the two different time series and the continuous transition inbetween. For practical data, transitions usually develop fluently instead of with discrete steps, and hence STAR captures practical data adequately.

For the transition variable s_t, there are different choices. The transition variable can be equal to an endogenous lagged variable $s_t = y_{t-d}$ for some integer $d > 0$, indicating that the time series has a different character for distinct values of the endogenous variable. The transition variable

can also be equal to the exogenous variable $s_t = z_t$, indicating that transition between regimes are initiated by exogenous variables. In this research, this would be suitable to capture the influence of temperature as an exogenous variable on the sales. We can also choose some general transformation of the variables, i.e., $s_t = h(x_t)$, where h is a function. The following four variables are proposed as transition variable:

- $s_{1,t} = z_{t+j}$, i.e., the j-th lead on the exogenous variable, or the temperature forecast in j days;
- $s_{2,t} = z_{t+j}^2$, which is $s_{1,t}$ squared, in order to let the model blow up the sales at an even higher rate to capture peaks more accurately;
- $s_{3,t} = \max(z_t, z_{t+1}, \ldots, z_{t+j})$ is the highest maximum temperature of the coming j days using the weather forecast;
- $s_{4,t} = \max(z_t, z_{t+1}, \ldots, z_{t+j})^2$, which is $s_{3,t}$ squared, blowing up the sales more rapidly similar to $s_{2,t}$;
- $s_{5,t} = \max(z_{t+1}, \ldots, z_{t+j})$, similar to $s_{3,t}$, but now excluding the current temperature z_t;
- $s_{6,t} = \max(z_{t+1}, \ldots, z_{t+j})^2$, which is $s_{5,t}$ squared, blowing up the sales more rapidly similar to $s_{2,t}$;

As stated in section 2, there are two (amongst others) specifications of STAR. In this paper, we only consider LSTAR, and defines G as

$$G(s_t; \gamma, c) = \frac{1}{1 + \exp\{-\gamma(s_t - c)\}}, \quad \gamma > 0.$$

Parameter c can be interpreted as the threshold value between the two regimes. The transition variable s_t is the only variable that alters the transition function G during time. Therefore, switching between regimes is fully dependent of this variable. The point between the two regimes is located at $s_t = c$. For large γ, transitioning occurs here. The parameter γ is interpreted as the rate in which the time series switches between regimes. Indeed, it is the slope of the transition function at $s_t = c$: $\frac{\partial G}{\partial s_t}(s_t = c) = \gamma$.

4.2 Building the model

The STAR model can be built with a specific-to-general approach. A brief description of setting up an adequate model is as follows. First, the order p is determined by specifying a linear AR(p) model

of the time series. Then, linearity is tested against STAR nonlinearity by performing an LM-type test. If linearity is rejected, then transition variables are selected by testing different predetermined variables against each other and after that, the transition function is selected by testing a series of null hypotheses extracted from the auxiliary regressions from the logistic and exponential transition functions. Then, estimation of parameters is done by nonlinear least squares (NLS). Finally, the valuation of the model is done by a series of diagnostics tests: the adequacy of STAR should be tested by null hypotheses such as no residual autocorrelation and no remaining nonlinearity. The model is modified if necessary.

4.2.1 Specifying the order of the linear AR part

Building a STAR model starts with choosing an appropriate lag order for the linear AR model. An AR(p) model is given by

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \quad t = 1, \ldots, T. \tag{2}$$

Given p, one can estimate the parameters $\phi_0, \ldots, \phi_p$ by using OLS. From the estimated AR(p) models, the sum of squared estimated residuals $\hat{\sigma}^2 = \sum_{t=1}^{T} \hat{\varepsilon}_t^2$ and the number of parameters $p+1$ are used to select the appropriate order. Criteria to select a model are the Akaike Information Criterion (AIC):

$$\text{AIC} = T \log \hat{\sigma}^2 + 2(p+1),$$

or the Schwarz Information Criterion (BIC):

$$\text{BIC} = T \log \hat{\sigma}^2 + (p+1) \log T.$$

The value for p that minimizes these criteria should be selected as order for the linear AR model. Moreover, the residuals should approximately be white noise to prevent residual autocorrelation later on in building the STAR model. Therefore, we test for residual autocorrelation by means of the Ljung-Box test. The null hypothesis H_0 is that there is no autocorrelation in the residuals of the AR(p) model at lags 1 to $m > p$. Any large value of m should be sufficient, as they often lead to the same conclusions. Let $r_k(\hat{\varepsilon}) = \sum_{t=k+1}^{T} \varepsilon_t \varepsilon_{t-k} / \sum_{t=1}^{T} \varepsilon_t^2$ be the $k-$th autocorrelation of the residuals, then the Ljung-Box statistic equals

$$\text{LB}(m) = T(T+2) \sum_{k=1}^{m} \frac{r_k^2(\hat{\varepsilon})}{T-k}.$$

Under H_0, we have that $\text{LB}(m) \sim \chi^2_{m-p}$, and we reject the null hypothesis for large values of LB. In case of rejection, the model should not be picked, due to remaining residual autocorrelation. Instead, the smallest order p for which there is no rejection, p_{LB}, should be selected. Especially whenever AIC and BIC are not minimized by the same order p, the LB-statistic can be used to make a decision.

The decision of the order $\hat{p}$ should be the p that minimizes AIC and BIC, provided that $p \geq p_{LB}$. In case of rejection in the LB test for order p, we should pick $p = p_{\text{LB}}$. In other words, $\hat{p} = \max\{\min\{p_{\text{AIC}}, p_{\text{BIC}}\}, p_{\text{LB}}\}$.

4.2.2 Testing for STAR nonlinearity

The next step in building the structure of the STAR model is testing STAR against linearity. For this, we rearrange terms in eq. (1):

$$y_t = \phi_1' x_t + (\phi_2' - \phi_1') x_t G(s_t; \gamma, c) + \varepsilon_t, \quad t = 1, \ldots, T. \tag{3}$$

Looking at eq. (3), it follows that when we restrict the parameters to $\phi_1 = \phi_2$, eq. (3) reduces to the linear AR model given in eq. (2). However, Luukkonen et al. (1988) argue that due to the presence of unidentified nuisance parameters (i.e., parameters that are not restricted by the null hypothesis such as c and γ), conventional statistical theory does not hold anymore. Instead, they propose to approximate the transition function G by a (third-order) Taylor series (around $\gamma = 0$):

$$\begin{aligned} G(s_t; \gamma, c) &= G(s_t; 0, c) + \gamma \frac{\partial G}{\partial \gamma}\bigg|_{\gamma=0} + \frac{1}{2}\gamma^2 \frac{\partial^2 G}{\partial \gamma^2}\bigg|_{\gamma=0} + \frac{1}{6}\gamma^3 \frac{\partial^3 G}{\partial \gamma^3}\bigg|_{\gamma=0} + R_3(s_t; \gamma, c) \\ &= \frac{1}{2} + \frac{1}{4}\gamma(s_t - c) - \frac{1}{48}\gamma^3 (s_t - c)^3 + R_3(s_t; \gamma, c), \end{aligned}$$

where $R_3(s_t; \gamma, c)$ is the remainder term. For computations of the partial derivatives, see appendix B.1. If we neglect the remainder term and substitute this in eq. (3), we get

$$\begin{aligned} y_t &= \phi_1' x_t + (\phi_2' - \phi_1') x_t \left(\frac{1}{2} + \frac{1}{4}\gamma(s_t - c) - \frac{1}{48}\gamma^3 (s_t - c)^3 \right) + e_t \\ &= \beta_0' x_t + \beta_1' x_t s_t + \beta_2' x_t s_t^2 + \beta_3' x_t s_t^3 + e_t, \end{aligned} \tag{4}$$

with $e_t = \varepsilon_t + (\phi_2' - \phi_1') x_t R_3(s_t; \gamma, c)$. Hence, linearity is obtained if we restrict the parameters to $\beta_1 = \beta_2 = \beta_3 = 0$.

In order to test STAR against linearity, we can perform an F-test on the obtained auxiliary regression in eq. (4) and the nested linear model under the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

Let SSR_0 denote the sum of squared residuals of the regression under H_0, and let SSR_1 be the sum of squared residuals of the full regression. The test statistic is

$$\text{LM} = \frac{\left(\text{SSR}_0 - \text{SSR}_1\right)/3(m+1)}{\text{SSR}_1/(T-4(m+1))}.$$

Under H_0, we have that $\text{LM} \sim F\left(3(m+1), T-4(m+1)\right)$, and we reject the null hypothesis of linearity for large values of LM. For all candidates s_t, LM should be computed. The candidate transition variable with the largest value of the LM statistic should be chosen as transition variable.

4.2.3 Parameter estimation

After specifying the order of the linear part of the model and choosing the transition variable, the parameters $\omega = (\phi_1', \phi_2', \gamma, c)'$ are estimated. Using this parameter, eq. (1) is compactly written as $y_t = F(x_t; \omega) + \varepsilon_t$, where F is nonlinear in x_t. Therefore, to estimate ω we can apply nonlinear least squares (NLS). That is, the sum of squared residuals is minimized:

$$\hat{\omega} = \arg\min_{\omega} \sum_{t=1}^{T} \left(y_t - F(x_t; \omega)\right)^2$$

The dimension of ω, and hence the number of parameters to be estimated by NLS, is $2p + 4$. Leybourne et al. (1998) suggest that we can reduce this dimensionality by splitting the NLS estimation problem into two parts. Note that the STAR model is linear in ϕ_1 and ϕ_2. Therefore, conditional on γ and c, we can estimate ϕ_1 and ϕ_2 by OLS. In other words, the NLS sum of squared residuals are concentrated with respect to ϕ_1 and ϕ_2, and the NLS estimation problem reduces to minimization with respect to only two parameters: γ and c.

For known γ and c, STAR transforms into a linear regression model with parameter vector $\phi = (\phi_1', \phi_2')'$ and explanatory variable vector $x_t(\gamma, c) = \left(x_t' \left(1 - G\left(s_t; \gamma, c\right)\right), x_t' G\left(s_t; \gamma, c\right)\right)'$. The parameter vector (conditional on γ and c) is estimated by concentrated OLS:

$$\hat{\phi}(\gamma, c) = \left(\sum_{t=1}^{T} x_t(\gamma, c) x_t(\gamma, c)'\right)^{-1} \left(\sum_{t=1}^{T} x_t(\gamma, c) y_t\right).$$

The problem is now reduced to a two-dimensional minimization problem, with minimization function $Q(\gamma, c)$:

$$Q(\gamma, c) = \sum_{t=1}^{T} \left(y_t - \phi(\gamma, c)' x_t(\gamma, c)\right)^2.$$

We start the estimation problem by finding suitable starting values for γ and c. One straightforward way is to perform a two-dimensional grid search over γ and c, applying OLS on the STAR

model conditional on γ and c, and calculating the sum of squared residuals. The combination (γ, c) that minimizes this are selected as starting values.

If we choose (grid) values for c outside the range of s_t, then the transition function will be smaller (or larger) than $\frac{1}{2}$ for all s_t. Then the time series will only be in one regime, and STAR does not make sense anymore. Therefore, grid values for c should be in the range of s_t, hence we choose N_c equidistant numbers between the α% and $(100 - \alpha)$% quantiles of s_t, e.g., $N_c = 100$ numbers between the $\alpha = 10$% and 90% quantiles of s_t.

The parameter γ is interpreted as the rate in which the time series switches between regimes. To find a good starting value for γ, we want to cover as much values of the transition function as possible between 0 and 1. Because γ is multiplied with $s_t - c$, we can divide γ by the standard deviation σ_{s_t} of s_t to make it scale-free. Hence, we choose a set of numbers 1 to 50 and divide this by σ_{s_t}

Once suitable starting values are found, we can start estimating all parameters. The problem can be solved by conventional nonlinear optimization methods. Here, we choose to optimize using the BFGS method, which is a Quasi-Newton method to solve unconstrained nonlinear optimization problems, see Wright and Nocedal (1999). The procedure of the BFGS algorithm is described in appendix B.2. Besides starting values and the minimization function, we need the gradient of the minimization function $\nabla Q(\gamma, c)$. Here, we treat the autoregressive parameters as known, i.e., $\phi(\gamma, c) = \phi$. We then have that

$$\nabla Q(\gamma, c) = \sum_{t=1}^{T} 2\left(y_t - \phi' x_t(\gamma, c)\right) \nabla(-\phi' x_t(\gamma, c)) = -2\sum_{t=1}^{T} e_t(\phi_1' - \phi_2') x_t \nabla G(\gamma, c),$$

where $e_t = y_t - \phi' x_t(\gamma, c)$ is the error term and

$$\nabla G(\gamma, c) = \left((s_t - c)\frac{e^{-\gamma(s_t - c)}}{\left(e^{-\gamma(s_t - c)} + 1\right)^2}, -\gamma \frac{e^{-\gamma(s_t - c)}}{\left(e^{-\gamma(s_t - c)} + 1\right)^2} \right)'. \tag{5}$$

4.2.4 Diagnostic checks and model modification

After setting the structure of the model and estimating the parameters, we have a fitted model and we can evaluate the adequacy of this model by performing misspecification tests. Such tests include testing hypotheses of no serial dependence and no remaining nonlinearity (van Dijk, 1999).

Serial independence

First, we test for the hypothesis of no residual autocorrelation, or serial independence. Residual autocorrelation is present when residuals and a lagged version of themselves are dependent. This

is a violation of the standard assumptions. The usual LM type test for serial correlation, the Breusch-Godfrey test, needs modification to be applicable on the STAR model. We test the null hypothesis of no residual autocorrelation up to lag order q, by performing an LM test for q-th order serial dependence in ε_t. Let $\hat{\omega}$ be the parameter estimate of the fitted STAR model under the null hypothesis, with residual estimates $\hat{\varepsilon}_t$. We regress $\hat{\varepsilon}_t$ on its q lagged residuals $\hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-q}$ and the gradient of the model fit $\nabla F(x_t; \hat{\omega})$, which equals

$$\nabla F(x_t; \hat{\omega}) = \left(x_t(1 - G(s_t; \gamma, c)), x_t G(s_t; \gamma, c), (\phi_2' - \phi_1')x_t \nabla G(\gamma, c)'\right)', \tag{6}$$

where $\nabla G(\gamma, c)$ is given in eq. (5). From this regression, we obtain the coefficient of determination R^2. The test statistic is $\mathrm{LM}(q) = nR^2$, with $n = T - q$. This test statistic is asymptotically χ^2 distributed with q degrees of freedom. We reject the null hypothesis of serial independence for large values of LM, i.e., for $\mathrm{LM}(q) > \chi^2_q$. In case of rejection we can add an extra lagged variable and repeat the process of building the model.

No remaining nonlinearity

Second, we test for the hypothesis of no remaining nonlinearity, to see if we have captured a sufficient amount of the nonlinearity with the STAR model. A natural way to do this is to extend the STAR model as given in eq. (3) with an additional regime G_2:

$$y_t = \phi_1' x_t + (\phi_2' - \phi_1')x_t G_1(s_t; \gamma_1, c_1) + (\phi_3' - \phi_1')x_t G_2(s_t; \gamma_2, c_2) + \varepsilon_t. \tag{7}$$

We can test for no remaining nonlinearity by testing the alternative of eq. (7) against the null of eq. (3). This comes down to the absence of the term $(\phi_3' - \phi_1')x_t G_2(s_t; \gamma_2, c_2)$, and can be tested by means of $H_0 : \phi_3 = \phi_1$ or $H_0 : \gamma_2 = 0$. This however gives rise to an identification problem similar to testing for STAR nonlinearity against linearity in section 4.2.2. Similar to that, we can estimate G_2 with a Taylor series approximation (around $\gamma_2 = 0$). The new auxiliary model is then given by

$$y_t = \beta_0' x_t + \beta_1' x_t G_1(s_t; \gamma_1, c_1) + \beta_2' x_t s_t + \beta_3' x_t s_t^2 + \beta_4' x_t s_t^3 + \mathrm{e}_t.$$

The new null hypothesis is $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ and the identification problem is solved. Again, an LM test is performed: we regress the residuals $\hat{\varepsilon}_t$ obtained under the null hypothesis against the partial derivatives of the regression function with respect to the parameters in the base STAR model of eq. (1) (i.e., $\phi_1, \phi_2, \gamma_1, c_1$), and the remaining terms $x_t s_t^j$ for $j = 2, 3, 4$. Put differently, we regress $\hat{\varepsilon}_t$ on $(\nabla F(x_t; \hat{\omega})', x_t s_t, x_t s_t^2, x_t s_t^3)'$, where hats indicate estimation under the null hypothesis, and ∇F is given in eq. (6). From this regression we obtain the coefficient of determination R^2, and

the test statistic equals LM = nR^2. Asymptotically, LM is χ^2-distributed with $3(m+1)$ degrees of freedom, and we reject for large values of LM. In case of rejection, we can estimate an extended MRSTAR model instead, see section 4.4.

4.3 Forecasting with STAR models

After specifying the model, estimating the parameters, evaluating the model with diagnostic tests and adjust if necessary, we have built a model that can be used for forecasting. We can implement data (i.e., sales of a chunk) from a certain year for some period into the model to specify the parameters. With this model, we can make sales forecasts of that same chunk in a new year, just as far ahead as the weather forecast is available.

4.3.1 Point forecasts

Recall that eq. (1) is given by $y_t = F(x_t;\Omega) + \varepsilon_t$. Forecasts of this model can be either point forecasts or prediction intervals. First, define the h step ahead forecast as $\hat{y}_{t+h|t}$, a forecast of y_{t+h} made at time t, with prediction error $e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$. Forecasting is done with a quadratic loss function, i.e., $\hat{y}_{t+h|t}$ minimizes the squared prediction error

$$E\left(e^2_{t+h|t}\right) = E\left((y_{t+h} - \hat{y}_{t+h|t})^2|X_t\right).$$

Here X_t represents the set of all lagged (sales) variables and exogenous variables up until time t. The squared prediction error is minimized whenever we assign the forecast the conditional expectation of y_{t+h} at time t (van Dijk, 1999), i.e.,

$$\hat{y}_{t+h|t} = E(y_{t+h}|X_t). \tag{8}$$

For the 1 step ahead forecast where $h = 1$, eq. (8) is easily computed, as y_{t+1} can be expressed in a deterministic output (and an error term with zero mean): $\hat{y}_{t+1|t} = E(y_{t+1}|X_t) = E(F(x_{t+1};\Omega) + \varepsilon_{t+1}|X_t) = F(x_{t+1};\Omega)$. Recall that x_{t+1} consists of lags and not current variables, and it is therefore known.

The 2 step ahead forecast $\hat{y}_{t+2|t}$ is more complicated to compute. We have

$$\hat{y}_{t+2|t} = E(y_{t+2}|X_t) = E(F(x_{t+2};\Omega)|X_t).$$

Firstly, the input in the expectation contains y_{t+1} which is unknown, hence this term cannot be extracted from the expectation. Moreover, as F is nonlinear, the expectation and F cannot be swapped.

We can of course use the previous 1 step ahead forecast in the 2 step ahead forecast. For this, let $\hat{x}_{t+2}^{(N)} = (1, \hat{y}_{t+1|t}, y_t, \ldots, y_{t-(p-2)})'$, where we ignored the error term ε_{t+1}. A 2 step ahead forecast with $\hat{x}_{t+2}^{(N)}$ becomes

$$\hat{y}_{t+2|t}^{(N)} = E(F(\hat{x}_{t+2}^{(N)}; \Omega)|X_t) = F(\hat{x}_{t+2}^{(N)}; \Omega).$$

This is called a naïve forecast, as it comes down to neglecting the error term. These types of forecasts are however biased (van Dijk, 1999).

Another approach is to use a Monte Carlo method. The 2 step ahead Monte Carlo forecast is given by

$$\hat{y}_{t+2|t}^{(M)} = \frac{1}{P} \sum_{k=1}^{P} F(\hat{x}_{t+2,k}^{(M)}; \Omega),$$

for some large integer P, where $\hat{x}_{t+2,k}^{(M)} = (1, \hat{y}_{t+1|t} + \varepsilon_{t+1,k}^{(M)}, y_t, \ldots, y_{t-(p-2)})'$. Here, we draw P values of $\varepsilon_{t+1,k}^{(M)}$ independently from the distribution of the error term. We draw error terms from a normal distribution with zero mean and variance equal to the sample variance of the training data model residuals. As $P \to \infty$, the forecast becomes unbiased due to the weak law of large numbers (Lundbergh & Teräsvirta, 2002).

An alternative approach, somewhat similar to the Monte Carlo method, is the bootstrapping method. The only difference is the way we choose the error terms. The 2 step ahead bootstrap forecast is given by

$$\hat{y}_{t+2|t}^{(B)} = \frac{1}{P} \sum_{i=1}^{P} F(\hat{x}_{t+2,i}^{(B)}; \Omega),$$

where $\hat{x}_{t+2,i}^{(B)} = (1, \hat{y}_{t+1|t} + \varepsilon_{t+1,i}^{(B)}, y_t, \ldots, y_{t-(p-2)})'$ for some $i \in \{1, \ldots, P\}$. We can pick $P = T$, the size of training data set. Here, the $\varepsilon_{t+1,i}^{(B)}$ are drawn with replacement T times from $\{\hat{\varepsilon}_t : t = 1, \ldots, T\}$, the set of residuals from the estimated model. The advantage of the bootstrapping method compared to the Monte Carlo method is that we do not have to make assumptions on the distribution of the error terms.

Multiple step ahead forecasts, say $h > 2$, are computed analogously to the above described methods of computing the 2 step ahead forecasts. Every forecasting step is evaluated using the previous forecasts. We have, with bootstrapping for instance, that

$$\hat{y}_{t+h|t}^{(B)} = \frac{1}{T} \sum_{i=1}^{T} F(\hat{x}_{t+h,i}^{(B)}; \Omega),$$

where

$$\hat{x}_{t+h,i}^{(B)} = (1, \hat{y}_{t+(h-1)|t} + \varepsilon_{t+(h-1),i}^{(B)}, \hat{y}_{t+(h-2)|t}^{(B)}, \ldots, \hat{y}_{t+1|t}^{(B)}, y_t, \ldots, y_{t-(p-h)})'.$$

In order to calculate the forecast on y_{t+h}, we use the model fit function F. This function contains the transition variable of the current observation (i.e., s_{t+h}). The problem is that we do not know s_{t+h} at time t, and hence we can replace it by its predicted value $\hat{s}_{t+h}$. This predicted value is generally not known. However, in our case where we use maximum temperature as transition variable, we have already discussed in section 3 that weather forecasts are currently very precise up to a week. We therefore assume that $\hat{s}_{t+h|t} = s_{t+h}$ for $h = 1, \ldots, 7$ and all t.

Error terms

A problem that may arise in forecasting with Monte Carlo or bootstrapping methods is that the drawn error terms do not match the observations to be forecast. This is due to a potential violation of the assumption that the error terms are normally distributed with constant variance. Particularly, the variance of error terms can be significantly different between the regimes, or more generally, the variance may increase whenever the sales increase. This potential heteroskedasticity can be detected by plotting the error terms against the time series observations: if the deviation of the error terms increases while the observations increase, then this indicates heteroskedasticity. Another (more exact) way of detecting heteroskedasticity is performing the Breusch-Pagan test.

If heteroskedasticity is present, then this can be accounted for by drawing regime specific error terms. That is, if we want to forecast $h \geq 2$ steps ahead, we draw error terms out of the set of error terms corresponding to the same regime as the forecast h steps ahead. The problem is that the regime of the h steps ahead is not known, since s_{t+h} is not known at day t (and therefore, G is unknown). As discussed before, based on precise weather forecasts, we assume that $\hat{s}_{t+h|t} = s_{t+h}$ for $h = 1, \ldots, 7$.

4.3.2 Direct forecasting

Instead of making point forecasts, we can also make direct forecasts of the time series. For this, we alter the model in such a way that we fit the series directly on a future observation. That is, we substitute the dependent variable of eq. (1) with future observation y_{t+h}:

$$y_{t+h} = \phi_1' x_t (1 - G(s_t; \gamma, c)) + \phi_2' x_t G(s_t; \gamma, c) + \varepsilon_t, \quad t = 1, \ldots, T - h.$$

This model is called the h step ahead STAR model. Model estimation is similar to the estimation process described in section 4.2, except that we use h less observations. One drawback is that for every forecasting step h, a different model needs to be estimated. Nevertheless, we can use

a different transition variable for each model. In that case, not all forecasts have to depend on the same temperature forecast as with point forecasts, which makes the forecasts more flexible. A natural decision is to choose the h-th day forecast of the maximum temperature is used, i.e., $s_t = z_{t+h}$. Due to temperature forecasts being accurate until the 7th day (see section 3.2), we limit the forecasting step to $h \leq 7$. The h step ahead direct forecast is therefore given by

$$\hat{y}_{t+h}^{(D)} = \hat{\phi}_1' x_t (1 - G(s_t; \hat{\gamma}, \hat{c})) + \hat{\phi}_2' x_t G(s_t; \hat{\gamma}, \hat{c}).$$

4.3.3 Evaluation of forecasts

As there are four different ways (mentioned) to execute the forecasting, we need an evaluation technique to compare these forecasts. One approach is to use the mean squared prediction error (MSPE) for h step ahead forecasts. We can compute the evaluation of the forecasts of the entire time series by:

$$\text{MSPE}_h = \frac{1}{T-h} \sum_{t=1}^{T-h} (\hat{y}_{t+h|t} - y_{t+h})^2,$$

where h is the forecasting step. Another metric is the median absolute prediction error (MAPE), which is appropriate whenever an absolute (instead of a quadratic) loss function is used:

$$\text{MAPE}_h = \text{median}\{|\hat{y}_{t+h|t} - y_{t+h}|; t = 1, \ldots, T-h\}.$$

Of course, we choose the forecasting technique (naïve, Monte Carlo or bootstrapping) with the smallest MSPE and/or MAPE.

To test whether two forecasting methods differ significantly in MSPE, we use the Diebold-Mariano (DM) test (Diebold & Mariano, 2002). Let $\hat{y}_{i,t+h|t}$ be the h step ahead forecasts of model i, with corresponding error terms $e_{i,t+h|t} = y_{t+h} - \hat{y}_{i,t+h|t}$. The loss differential between models i and j for the h-th forecasting step is defined as $d_{t+h} = e_{i,t+h|t}^2 - e_{j,t+h|t}^2$ for the MSPE measure, and $d_{t+h} = |e_{i,t+h|t}| - |e_{j,t+h|t}|$ for the MAPE measure. Testing for no significant difference between the forecasting techniques comes down to the loss differential having zero mean. The null hypothesis is therefore $H_0 : E(d_{t+h}) = 0$, with forecasting step $h \leq 7$. For large samples, d_{t+h} is approximately normally distributed with mean $\bar{d}_h$ and variance $\frac{2\pi f_d(0)}{P}$ (Diebold & Mariano, 2002). Here, P is the number of forecasts and $f_d(0)$ is the spectral density at frequency zero. Therefore, the test statistic is given by

$$\text{DM}_h = \frac{\bar{d}_h}{\sqrt{2\pi \hat{f}_d(0)/P}}.$$

A consistent estimate of $2\pi f_d(0)$ is obtained by taking a sum of the available sample autocovariances (i.e., up to $h-1$):

$$2\pi \hat{f}_d(0) = \sum_{\tau=-(h-1)}^{h-1} \hat{\gamma}_d(\tau).$$

The DM statistic is asymptotically standard normally distributed. We reject the null hypothesis for large values of the test statistic. In case of rejection, the forecasting techniques differ significantly.

4.4 Extension to the STAR model

The STAR model is appropriate for nonlinear time series whenever one transition variable is the driver for switching between two regimes. However, when there are multiple transition variables that affect switching regimes, or when the time series show more than just two different regimes, STAR is rather limited. In that case, it can be extended to the Multiple Regime STAR (MRSTAR) model. The MRSTAR model extends the number of regimes of STAR and (optionally) the number of transition variables. If the null hypothesis of no remaining nonlinearity, as explained in section 4.2.4, is rejected, one can extend STAR by including additional regime terms like in eq. (7):

$$y_t = \phi_1' x_t + (\phi_2' - \phi_1')x_t G(s_t; \gamma_1, c_1) + \cdots + (\phi_q' - \phi_{q-1}')x_t G(s_t; \gamma_{q-1}, c_{q-1}) + \varepsilon_t, \quad t = 1, \ldots, T.$$

This is a MRSTAR model with m regimes and one transition variable. Let $G_j(s_t) := G(s_t; \gamma_j, c_j)$ and assume $c_1 < c_2$ without loss of generality. For a q-regime model, each time s_t surpasses c_j, the j-th regime G_j changes smoothly from 0 to 1. More regimes might reduce bias in the estimates, because it allows for more constant autoregressive values (the ϕ_{j1}'s).

We can also include more than one transition variable. A 4-regime MRSTAR model with two transition variables is given by van Dijk (1999). It can be constructed by replacing the $\phi_j' x_t$ in the base model (as in eq. (1)) with another star model:

$$\begin{aligned} y_t = &\left(\phi_1' x_t (1 - G_1(s_{1t}; \gamma_1, c_1)) + \phi_2' x_t G_1(s_{1t}; \gamma_1, c_1)\right)\left[1 - G_2(s_{2t}; \gamma_2, c_2)\right] + \\ &\left(\phi_3' x_t (1 - G_1(s_{1t}; \gamma_1, c_1)) + \phi_4' x_t G_1(s_{1t}; \gamma_1, c_1)\right) G_2(s_{2t}; \gamma_2, c_2) + \varepsilon_t. \end{aligned} \tag{9}$$

With 4 regimes in eq. (9), each regime is a combination of the two transition functions $G_1 = 0, 1$ and $G_2 = 0, 1$. Both transition functions incorporate a different transition variable and different parameters, s_{1t} with parameters γ_1, c_1 for the first transition function G_1 and s_{2t} with parameters γ_2, c_2 for the second transition function G_2.

The process of building a MRSTAR model begins with testing the null hypothesis of no remaining nonlinearity after a STAR model as been estimated. If the null hypothesis is rejected, then there is

evidence of remaining nonlinearity and we add another regime to the model. Building the new model begins with estimation immediately, as the first two steps (specifying the order of the linear AR part and testing for STAR nonlinearity) are redundant. Estimation is similar to the case with 2 regimes. Now there are more parameters. For the 2-regime STAR model, there were $2m+4$ parameters. For a q-regime STAR model the number of parameters is $q(m+3)-2$. After estimation, we repeat the test of no remaining nonlinearity, now adding another extra regime. The same process is repeated until there is no evidence of more remaining nonlinearity. When the final model is made, forecasts are made, also in a similar way. We finally compare these forecasts with the STAR forecasts. MSPE and MAPE will showcase if the forecasting accuracy has improved.

4.5 Excess stock and lost sales time series

With the forecasts, we can make predictions of the amount of sales in the next days. With these predictions, the e-commerce company can order stock such that we can meet the demand of a product with the amount of stock. Of course we do not want to maintain too much stock. Therefore, there is an optimal amount of products to be ordered each day. When predicted sales are more accurate, then we can maintain a stock that matches better with the demand. On the one hand, this leads to a reduction of excess stock, which cuts the total costs since having stock requires storage. On the other hand, it reduces the amount of "lost" sales (i.e., products that are in demand but out of stock), which means that either the products are bought at a later time or not at all occurs less, thus improving the revenue.

We consider a simplified representation of the process of ordering stock and selling the stock. The daily timeline on day t is as follows. First, stock that was ordered by the e-commerce company i days prior to day t arrives with an amount of $b_{t|t-i}$. Second, the products are being sold during the day with a total amount of y_t. After that, the stock v_t is measured. And finally, new stock is ordered for, say, i days ahead and is represented as $b_{t+i|t}$.

Assume that ordering stock takes just one day, then we have $i = 1$. We want to figure out the optimal order amount $b_{t+1|t}$. If no stock is left on day t (i.e., $v_t = 0$), we ideally want to order a quantity equal to the amount of sales on the next day, $b_{t+1|t} = y_{t+1}$. In that case, we perfectly fit the demand. However, on day t we do not know y_{t+1} yet. Therefore, we use the forecast value of the sales of tomorrow, so we have $b_{t+i|t} = \hat{y}_{t+1|t}$. If there is still stock left, however, then we should subtract this with the quantity of stock left, provided that this number is non-negative. Thus, we have that $b_{t+1|t} = \max(0, \hat{y}_{t+1|t} - v_t)$.

Consider the case $i = 2$, where it takes two days for an order to arrive. The optimal quantity to be ordered is equal to the expected sales on day $t + 2$ subtracted by the expected stock on day $t + 1$. That is, $b_{t+2|t} = \max(0, \hat{y}_{t+2|t} - \hat{v}_{t+1|t})$. We can decompose $\hat{v}_{t+1|t}$ into the stock on day t and all events in between (i.e., the sales and order on day $t + 1$): $\hat{v}_{t+1|t} = \max(0, v_t + b_{t+1|t-1} - \hat{y}_{t+1|t})$. Let us now consider the general case where it takes i days for an order to arrive. Analogously, it holds that

$$b_{t+i|t} = \max(0, \hat{y}_{t+i|t} - \hat{v}_{t+i-1|t}) \quad \text{and} \quad \hat{v}_{t+j|t} = \max(0, \hat{v}_{t+j-1|t} + b_{t+j|t+j-i} - \hat{y}_{t+j|t}).$$

The latter formula can recursively be applied for $j = i-1, \ldots, 1$. For $j = 1$ we have that $\hat{v}_{t+j-1|t} = \hat{v}_{t|t} = v_t$ and this is known on day t, and hence the optimal quantity ordered can be calculated in this recursive manner.

With the formula for the optimal quantity ordered, we can calculate the stock v_t on every day. Since this is measured after all sales are done on a day, the excess stock function (ES) is simply equal to v_t. The lost sales function (LS) is equal to the surplus of the sales compared to the previous stock and quantity ordered i days earlier combined: $\text{LS}_t = |\min(0, v_{t-1} + b_{t|t-i} - y_t)|$. We can calculate ES and LS for every forecasting technique. These are representative for the effect of a forecasting technique on inventory management. If we compare ES and LS among different techniques, it becomes clear which technique cuts the costs the most. Thus, this is also a way to compare forecasting performance, besides MSPE and MAPE. We can also pick the best (STAR) forecasting technique in the sense of MSPE and/or MAPE, and compare this with AR forecasts by ES and LS to see how STAR improves upon AR in terms of costs.

5 Results

In this section, we apply the methods discussed in section 4 to the data that was already explored in section 3. That is, we build a STAR model to weather-related product daily sales using daily maximum temperature data. First, in section 5.1 we start by investigating the sales observations to see if a log transformation is necessary, and whether adding temperature as a linear component to an AR model improves the default AR model. Secondly, in section 5.2 we apply the cycle to the highest selling chunk category: fans. We give economic value (e.g., cost reduction or increasing revenue) to this model in section 5.3. Finally, we repeat the process for other series of products and summarize these in section 5.4.

5.1 Investigating the time series

For the entire process of building and evaluating the model, we will make a division of the data into a train and a test set. The train set consists of the first two years of data (2018 and 2019), whereas the test data consists of only one year (2020), such that there is a 2/3 split. The train and test set contain respectively 730 and 366 observations, summing up to a total of 1096. Estimating the model is done with the train set, and for evaluation the test set is used.

Before the modelling cycle is started, we determine whether a log transformation of the series is beneficial. It generally depends on the data and model. Generally, when heteroskedasticity is present for a regular series, a log transform might reduce this. In that case, when the log transformed series is used, statistical theory is more accurate as it is often built on the assumption of normality of errors. Lütkepohl and Xu (2012) conclude that it is useful to apply a log transform to a time series whenever the transformation makes the variance of the series more homogeneous. For the sales time series, we have already seen that there is exponential growth during the summer (sales peaks). However, the series return to a constant mean and the variance stabilizes. Moreover, when we estimate a STAR model (with up to 8 lags, as concluded in section 3) and look into the test set residuals histogram (see fig. A2), we see that there is a bias relative to 0 and there are outliers present. The distribution is however symmetric. Thus, we do not have enough evidence for or against a log transformation, and therefore we apply both series to STAR and compare their sum of squared residuals (SSR) of the test set. The SSR of the log estimated STAR model is 2.827 times that of the regular estimated model, and therefore we conclude that a log transform does not improve the model.

We have found an optimal lag order for an AR model in section 4.2.1, and with this we can estimate an AR model. The linear AR model is considered the default model against which we compare other models in forecasting performance. We start by adding the exogenous variable maximum temperature as a linear component to the AR model, to see if including temperatures enhances the forecasts. This model is called the ARX model (where the X represents the exogenous variable). In predicting sales on day t, we estimate this based on its m lagged values and the maximum temperature on that same day. Determining the optimal lag order for ARX is analogous to AR, and we find the same optimal order of 8. Estimating both models, we obtain estimates of all coefficients. The intercept of AR equals 679.45 (unit sales), and -13.76 for ARX. The temperature coefficient estimate in ARX is 44.31. The difference in intercepts is therefore countered by the temperature coefficient: if we consider a temperature of 15.64 (degrees Celsius) as default, then the

intercepts coincide. After estimation, we can make forecasts with steps $h = 1, \ldots, 7$. We calculate MSPE and MAPE for both models. Results are in table 2, where we display MSPE and MAPE of the ARX model as proportion of the AR model. The MSPE's of ARX are smaller than in AR, and moreover, they decrease with respect to the AR model with increasing forecasting step. Furthermore, DM statistics are given for every forecasting step. For no forecasting step do the forecasts differ significantly according to the DM statistic (in the sense of MSPE). For MAPE, we see that for $h = 1, 2$ the forecasts differ significantly, but not anymore for $h \geq 3$. We conclude that including the maximum temperature as a linear component to AR (i.e., the ARX model) improves the AR model, even though the forecasts are not (always) significantly different.

h	ARX/AR model		DM statistic	
	MSPE	MAPE	MSPE	MAPE
1	0.997	0.833	0.079	3.831
2	0.988	0.842	0.160	2.314
3	0.978	0.840	0.232	1.638
4	0.971	0.857	0.286	1.378
5	0.962	0.860	0.338	1.241
6	0.954	0.871	0.393	1.076
7	0.945	0.868	0.460	1.103

Table 2: MSPE and MAPE of AR models including exogenous variable temperature as proportions of linear (AR) models for forecasting steps $h = 1, \ldots, 7$, with Diebold-Mariano (DM) statistics for every step.

5.2 Making a STAR model for fan sales

The first part of the modelling cycle is determining the optimal lag order of the linear AR part of the model as described in section 4.2.1. In the data exploration, we saw that autocorrelations of the series were significant up to lag 8. Determining the optimal lag order in a more exact manner is done by calculating the AIC, BIC and LB statistic for every model up to lag order 10. These are summarized in table 3, where the optimal lag order per measure is put in bold. From this table, it follows that AIC is minimized for a linear model with lag order 9. For BIC we find an optimal lag order of 8, and with the LB tests, the smallest order which results in a p-value above 0.05 is also 8. Therefore, the optimal lag order is $\hat{m} = \max(\min(p_{\text{AIC}}, p_{\text{BIC}}), p_{\text{LB}}) = \max(\min(9, 8), 8) = 8$. This matches the conclusion of the data exploration.

AR order	AIC	BIC	LB	p-value
1	11911.19	11924.97	212.09	0.00
2	11780.11	11798.48	226.01	0.00
3	11739.46	11762.43	155.96	0.00
4	11693.68	11721.24	66.51	0.00
5	11686.46	11718.61	51.77	0.00
6	11666.08	11702.82	31.90	0.00
7	11658.88	11700.22	21.28	0.01
8	11649.50	**11695.43**	**8.77**	0.27
9	**11649.47**	11700.00	6.92	0.33
10	11651.26	11706.38	6.26	0.28

Table 3: AIC, BIC, and LB with corresponding p-values for 10 different AR models of the training data. Numbers in bold are the optimal lag order for the criterion measure of the corresponding column.

We can test a linear model against STAR nonlinearity, to find out if adding nonlinearity to the model improves the linear model significantly. We do this by performing an LM test as described in section 4.2.2 for different transition variables. The candidate transition variables are $s_{x,t}$ for $x = 1, \ldots, 6$. Each of the six candidates is associated to a certain lead observation j, see section 4.1. As the temperature forecasts are accurate up to seven days ahead, we test for $j = 1, \ldots, 7$. Thus, there are a total of 42 candidate transition variables. We have performed the LM test for each of these variables, and the resulting F-values are summarized in table 4. The 5% critical value for an F-distribution with $3(p+1) = 27$ and $T_{\text{train}} - 4(p+1) = 694$ degrees of freedom equals 1.50. All of the F-values of the candidate transition variable exceed this value, and hence all candidates indicate that nonlinearity improve the linear model significantly. The candidate with the largest F-value is $s_{2,t} = z^2_{t+j}$ for $j = 3$ (indicated bold in table 4) with $F = 25.15$ (and p-value $1.92 \cdot 10^{-84}$), and we choose this as our transition variable.

With a lag order of the AR part and STAR nonlinearity, the structure of the model is provided. Following the modelling cycle, the parameters are to be estimated. Here, we limit ourselves to a 2-regime STAR model. For each regime, there are $m = 8$ lagged orders and 9 parameters (including the constant). The transition function contains two extra parameters γ and c. The estimated model

j	$s_{1,t}$	$s_{2,t}$	$s_{3,t}$	$s_{4,t}$	$s_{5,t}$	$s_{6,t}$
1	15.68	21.96	12.68	18.35	15.68	21.96
2	14.76	18.83	14.73	17.52	19.80	24.30
3	23.95	**25.15**	15.23	17.61	17.07	19.70
4	19.18	20.76	17.70	20.53	18.60	21.28
5	12.96	15.09	16.23	20.30	17.56	22.01
6	13.21	15.14	14.70	17.92	16.93	20.05
7	16.70	19.78	13.51	15.73	15.66	18.57

Table 4: F-values for testing STAR nonlinearity, with 6 different transition variables $s_{x,t}$ and j the number of included lead values per transition variable.

is given in eq. (10).

$$\begin{aligned} y_t = \; & [43.55 + 1.38y_{t-1} - 0.72y_{t-2} + 0.51y_{t-3} - 0.38y_{t-4} + 0.05y_{t-5} + 0.20y_{t-6} - 0.06y_{t-7} \\ & -0.06y_{t-8}] \cdot \Big(1 - G(z^2_{t+3})\Big) + [932.19 + 0.60y_{t-1} - 1.93y_{t-2} - 0.06y_{t-3} + 2.98y_{t-4} \\ & -1.93y_{t-5} + 1.34y_{t-6} + 0.23y_{t-7} - 0.72y_{t-8}] \cdot G(z^2_{t+3}) + e_t \end{aligned} \tag{10}$$

The estimated logistic transition function is given in eq. (11). The transition function variables are $\hat{\gamma} = 100$ and $\hat{c} = 30.97^2 = 959.39$. Switching between regimes occurs at a maximum temperature of 30.97 (squared), with a rate of 100.00. This transition function is plotted in fig. A3. Due to a high transition rate combined with squared temperatures, transitioning happens rapidly, and the transition function approximates a step function. That is, the transition function is roughly 0 for all temperatures below the threshold value of 30.97 and 1 for all temperatures above this threshold value. In the train set, 715 observations are in the low regime and 15 in the high regime. In the train set, 356 observations are in the low regime and 10 in the high regime. The percentage of high regime observations is 2.10% for the train data and 2.73% for the test data.

$$G(z_{t+3}; \gamma, c) = \frac{1}{1 + \exp\{-100.00(z_{t+3} - 30.97^2)\}} \tag{11}$$

We have estimated the model using train data. Now, we evaluate the model by performing diagnostic tests and forecasting methods using the test data. First, the test data sales are applied to the estimated model, to see how well the model estimates new observations, and this is plotted in fig. 4. Predicted values are generally close to actual values. In the high sales peak around 10 August, there is more deviation in the predictions compared to actual values. The peak of 10 August

had 26244 sales, while this is overestimated as 40628. On 8 August, the sales were 8910, while the prediction was underestimated as 1455.

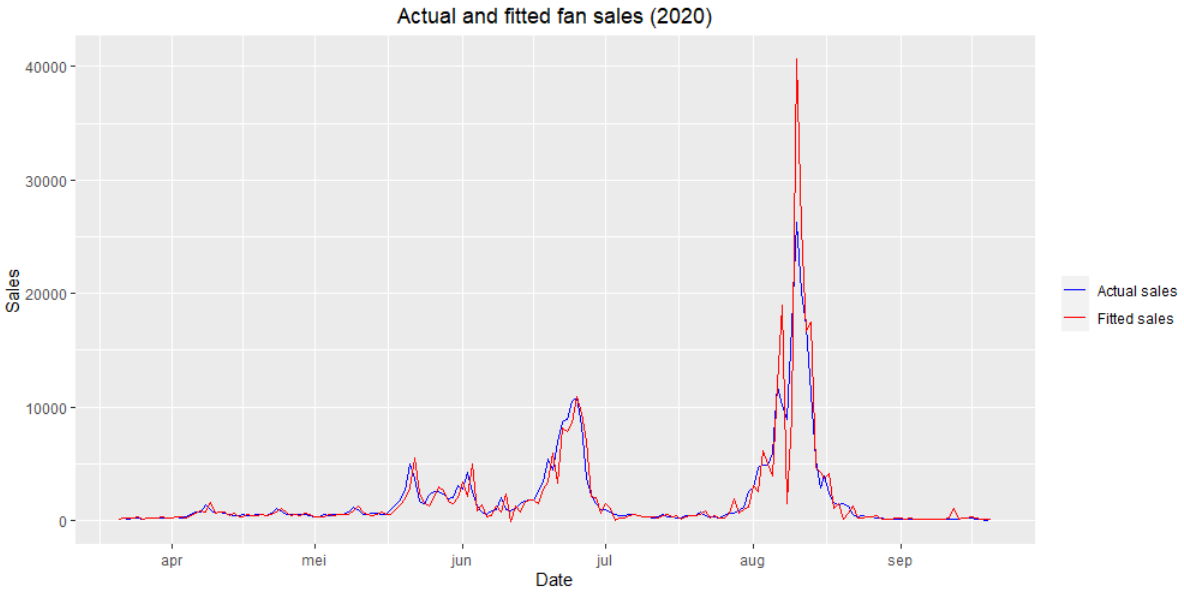

Figure 4: Spring and summer 2020 fan sales (blue) and corresponding model estimated sales (red).

From applying the test set to the model, we obtain 358 residuals, from which 10 correspond to high regime observations. A histogram of these residuals is given in fig. A2. The median of the set of residuals is -27.9, indicating a negative bias, i.e., the actual values are more often overestimated. To test normality of the residuals, we use the Jarque-Bera (JB) test. The test statistic equals $57334 > 5.99$ for the full set of residuals, and hence we reject the null hypothesis of normality. The set of high regime residuals is too small to make accurate decisions about normality. The median of the absolute residuals is 49.7, while for high regime observations it is 1603.4, indicating more deviation for high regime observations, and hence heteroskedasticity. To test for this, we regress the squared residuals on high and low regime variables $G(s_t)$ and $1 - G(s_t)$ respectively. The coefficient of the low regime variable is 461123 with p-value 0.33, and for the high regime variable we have a coefficient of 41151717 with p-value < 0.001. This indicates that the variance during high regimes is about $\frac{41151717}{461123} \approx 89$ times the variance during low regimes, and hence heteroskedasticity is present. In forecasting by the Monte-Carlo or bootstrapping methods, we can account for this by drawing regime specific error terms.

The residuals might be correlated with lagged residuals. As this is a violation of the assumptions, we want to test whether this is true. We do this by means of the serial independence LM test as described in section 4.2.4. We perform this test using all lagged residuals up to lag q, and we do this

for $q = 1, \ldots, 8$, see table 5 for the results including critical value per number of included lags. For every test, the null hypothesis is rejected, and therefore we conclude that residual autocorrelations are present up to lag 8. A solution to this is increasing the maximum lag order in the linear AR part of the model. However, we have seen that the residuals are biased. Particularly, in low regime the predictions often overestimate the actual values, and therefore consecutive residuals are often negative for a long period. Reducing this bias might help removing the serial dependence. We can account for this by expanding the number of regimes. We will do this later on.

q	LM(q)	Critical value
1	307.97	3.84
2	307.23	5.99
3	308.55	7.81
4	307.82	9.49
5	306.99	11.07
6	307.43	12.59
7	308.06	14.07
8	307.21	15.51

Table 5: q serial independence tests of the residuals of the model fitted test data.

With the estimated model, we can make forecasts. The naïve, Monte-Carlo and bootstrapping methods are compared in MSPE and MAPE. For each method, we make 1 to 7 step ahead forecasts. On top of that, for the Monte-Carlo and bootstrapping methods we also make a distinction between no regime specific variance (nsv) and regime specific variance (rsv) for the error terms. The results are summarized in table 6. Here, MSPE and MAPE of the entire data set for all five forecasting techniques are given as proportions of the MSPE and MAPE (resp.) of an estimated AR forecasting model. This is done to provide an insight of how the forecasting techniques compare to a standard AR model.

Evaluating the forecasts by MSPE of the entire set, we see in table 6 that they do not differ too much between different forecasting techniques. For almost all forecasting steps and techniques, the MSPE values are about 50 to 75% of that of the AR model MSPE. The lowest MSPE is attained for the Monte-Carlo and bootstrapping methods only. For five steps, bootstrapping works best and for two steps, Monte-Carlo works best. Moreover, we see that no regime specific variances work best five times (of which four are bootstrapping), compared to regime specific being the best method

only twice. We conclude that bootstrapping with no regime specific error terms work best (when comparing techniques by MSPE). When we look at the absolute MSPE values (that is, without dividing by the AR model MSPE), we see that the MSPE increases whenever the forecast horizon h increases, with the exception of $h = 3$. The exception is due to the choice of the transition variable, where we picked the 3-day lead temperature.

Next to MSPE, the other measure is MAPE. In table 6, we see the MAPE values of all methods for the entire dataset as proportion of the AR model MAPE values. The best forecasting technique is Monte-Carlo with no regime specific variance for the error terms, and this holds for all forecasting steps. Contrary to MSPE, the MAPE values of the forecasting techniques are much smaller compared to the AR model MAPE values. Here, the MAPE values are about 5-10% of that of the AR model MAPE values. This is considerably smaller than 50-70% with MSPE. This difference is explained as follows. For an AR model, there is not enough power to capture both peak observations and more regular low observations. In estimating a model, a trade-off has to be made between these types of observations. Therefore, it predicts regular low observations with a large positive bias, and these cause a large MAPE. The outliers are responsible for a high MSPE. In the case of a STAR model, the transition function makes sure that peak observations and regular observations are distinguished more properly. In that way, regular observations are predicted much better and show less bias, keeping the MAPE low. However, since peak observations still show much volatility, these are still hard to predict and can still be considered outliers in prediction errors, and hence MSPE is still high.

h	Naïve		Monte-Carlo (nsv)		Monte-Carlo (rsv)		Bootstrap (nsv)		Bootstrap (rsv)	
	MSPE	MAPE	MSPE	MAPE	MSPE	MAPE	MSPE	MAPE	MSPE	MAPE
1	0.536	0.098	0.538	**0.075**	0.536	0.093	0.535	0.089	**0.535**	0.092
2	0.728	0.074	0.729	**0.063**	0.731	0.070	**0.728**	0.069	0.738	0.071
3	0.490	0.067	0.493	**0.058**	**0.489**	0.060	0.489	0.062	0.531	0.063
4	0.652	0.062	0.655	**0.056**	0.654	0.058	**0.650**	0.058	0.690	0.063
5	0.566	0.061	0.569	**0.055**	0.571	0.059	**0.565**	0.060	0.605	0.059
6	0.536	0.058	0.538	**0.055**	0.536	0.059	**0.535**	0.059	0.569	0.059
7	0.496	0.060	**0.496**	**0.058**	0.499	0.061	0.497	0.060	0.494	0.059

Table 6: MSPE and MAPE for the Naïve, Monte-Carlo and Bootstrapping forecasting methods as proportions of the MSPE/MAPE from the estimated AR models. The latter two forecasting techniques are divided into no regime specific variance (nsv) and regime specific variance (rsv). The forecasting steps h range from 1 to 7.

We have seen that the Monte-Carlo method with no regime specific variance for the error terms was most often chosen to be the best forecasting method. Thus, we choose this method as the optimal forecasting technique. In fig. 5 the test data series is plotted together with 3 step ahead forecasts of the Monte-Carlo method. We have only included spring and summer of 2020 for better visualization purposes. For low sales, the forecasts are close to the actual values. Daily sales that are somewhat higher (mid sales), such as the period at the end of May and June, we see a delay in the forecasts relative to the actual values. The forecasts are more driven by past sales rather than predicting the right amount of sales taking into account high upcoming temperatures. Because we have a high transition rate and a high threshold value, transitioning between regimes does not run smoothly and does not occur often. Therefore, for temperatures below the threshold value, the model is just a linear AR model, predicting sales based on previous values. In that case, 3 step ahead forecasts are mainly based on previous sales that were between 4 and 11 days prior to the forecasting date. For high sales, such as the period of the first half of August, the forecasts show even more volatility. However, since this period contains more high-regime observations, the forecasts do not show such a delay and are sometimes accurate.

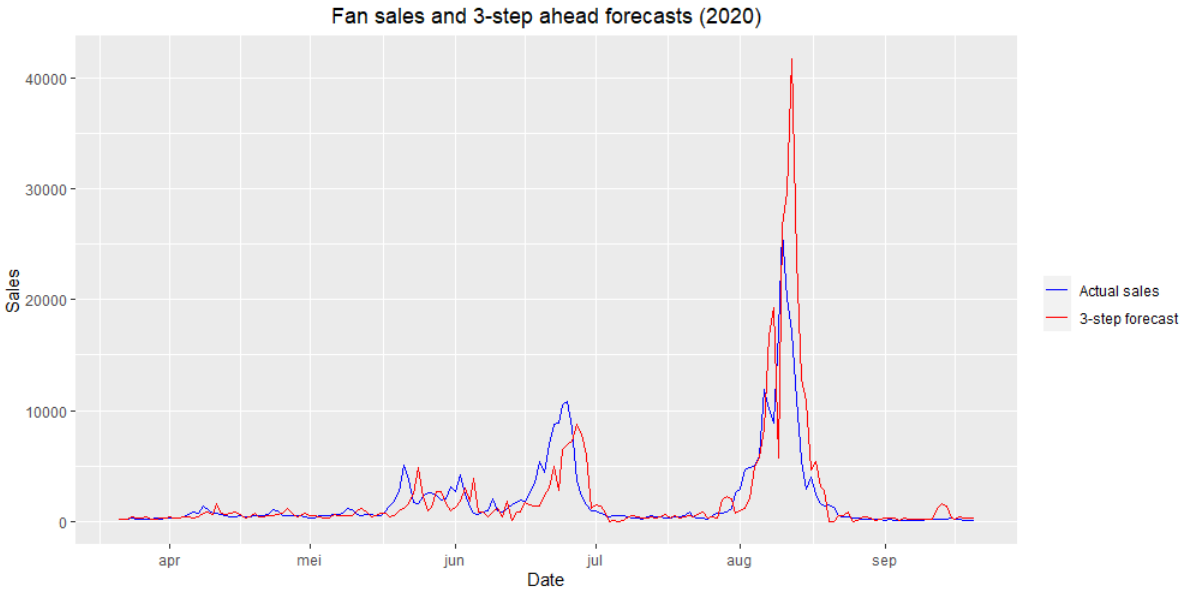

Figure 5: Spring and summer 2020 fan sales (blue) and corresponding 3 step ahead forecasts (red). Forecasting was done by the Monte-Carlo method.

For the naïve, Monte-Carlo and bootstrap forecasting methods, only one model is estimated, and with that model all forecasts are made. The other mentioned forecasting method, direct forecasting, estimates a new model for every forecasting step. The fit of a model is then the forecast. As this

method requires multiple model estimations, this method is more computer intensive. However, we might receive more accuracy in our forecasts. Moreover, since we estimate a new model for every step, we can also implement a new transition variable each step. For forecasting step h, we use transition variable $s_{2,t} = z_{t+h}^2$. The MSPE and MAPE for every forecasting step are given in table 7 (again as proportions of a linear AR model). Comparing the MSPE of direct forecasting with the MSPE of Monte-Carlo forecasting, we see that in the first two steps Monte-Carlo outperforms direct forecasting, and from forecasting step 3 and later direct forecasting works best. For MAPE, this holds already from step 2. Furthermore, both the MSPE and MAPE in direct forecasting strictly decrease in ratio against AR forecasting with increasing forecasting step (with the exception of $h = 7$). We conclude that direct forecasting gives the best results out of all four forecasting techniques.

To determine whether a STAR model significantly improves forecasting results, we compare the forecasting errors of the best STAR forecasting method with forecasting errors of an AR model by means of the Diebold-Mariano (DM) test. The DM statistic is calculated for both MSPE and MAPE, and are summarized in table 7. The critical value of this test is 1.96 on 5% significance, and we reject for larger values of the test statistic. for the first two forecasting steps, the test statistic (for MSPE) is lower than the critical value, and therefore the forecasting methods do not differ significantly. For forecasting horizon $h \geq 3$, direct forecasting improves MSPE significantly. The DM statistics for MAPE show significant improvement of direct forecasting upon linear forecasting for all forecasting steps.

h	Direct/AR model		DM statistic	
	MSPE	MAPE	MSPE	MAPE
1	0.784	0.182	1.01	6.72
2	0.850	0.068	1.03	10.40
3	0.367	0.035	4.89	13.38
4	0.270	0.032	8.80	14.40
5	0.180	0.023	8.90	14.67
6	0.142	0.021	10.17	14.52
7	0.168	0.047	11.74	13.86

Table 7: MSPE and MAPE of the best nonlinear forecasting method (Direct) as proportions of linear (AR) models for forecasting steps $h = 1, \ldots, 7$, with Diebold-Mariano (DM) statistics for every step.

Due to the serial dependence of the residuals and lack of flexibility of the model for middle to high temperatures, we propose to not limit the model to only two regimes. Therefore, we estimate a MRSTAR model. We estimate MRSTAR models and make forecasts for $h = 1, \ldots, 7$ using the best forecasting technique for the 2-regime STAR model, which is direct forecasting. The same transition variables are used for each forecasting step, and only 3-regime (MR)STAR models are estimated. The results are in table 8. For all models, except for the 1 step ahead direct model, we accept the alternative of adding an extra regime from testing no remaining nonlinearity (as described in section 4.2.4. MSPE of MRSTAR increases relative to STAR forecasts for increasing forecasting step, and are always higher. The MAPE proportions are considerably smaller than MSPE, but are still larger for MRSTAR than MAPE (with the exception of $h = 7$. We conclude that adding an extra regime to direct STAR models does not improve the forecasting accuracy.

h	p-value of testing an extra regime	$\hat{c}_1$	$\hat{c}_2$	MSPE ratio 3-regime vs. 2-regime STAR	MAPE ratio 3-regime vs. 2-regime STAR
2	$1.37 \cdot 10^{-38}$	811.32	844.76	1.115	1.119
3	$4.01 \cdot 10^{-19}$	862.03	941.16	2.148	1.430
4	$6.42 \cdot 10^{-12}$	885.21	951.31	7.895	1.289
5	$1.53 \cdot 10^{-10}$	914.98	949.29	2.560	1.658
6	$1.77 \cdot 10^{-13}$	914.29	946.61	3.024	1.720
7	$2.10 \cdot 10^{-33}$	798.43	982.15	12.875	0.951

Table 8: p-values of testing adding an extra regime to a 2-regime direct STAR model, with estimated threshold values $\hat{c}_1$ and $\hat{c}_2$, and MSPE/MAPE proportions of MRSTAR against STAR forecasts. Note that for $h = 1$, adding an extra regime was rejected.

5.3 Economic value of the new sales model

A STAR model is built to predict the sales of a product. We know which past sales are used to determine the next sales, which transition variable works best to estimate the model (and thus how many days ahead temperature forecasts works best), and at which temperature the model starts transitioning into a new regime. Several forecasting techniques are applied to the model, and we determined which technique works best (direct forecasting), and how much better (in terms of the forecasting error) it works compared to a standard AR model. In addition to this, we want to express the improvement of the forecasting technique in economic/business terms. We want to know what we can do with the more accurate predictions, what the benefits are of having more

accurate sales predictions and how do they improve compared to a normal forecasting model.

For both the AR and the STAR (direct) forecasting methods, we apply the process of determining the excess stock and lost sales as described in section 4.5. We do this for the scenarios where it takes $h = 1, \ldots, 7$ days for an order of products to arrive. The results are summarized in table 9.

h	Excess stock		Lost sales		Break-even
	AR	STAR	AR	STAR	ratio $p_{\text{LS}}/p_{\text{ES}}$
1	707	319	108	335	1.71
2	2095	713	92	410	4.34
3	4248	1725	96	505	6.17
4	7035	2196	117	488	13.04
5	10457	2252	139	528	21.06
6	14250	2394	152	514	32.71
7	18291	2629	152	561	38.36

Table 9: Mean excess stock and lost sales for AR and STAR model for forecasting steps $h = 1, \ldots, 7$. Includes the break-even proportion of the cost price of having lost sales against excess stock at which AR and STAR have equal total costs.

For all forecasting steps, the excess stock is higher in the AR model than the STAR model, whereas the lost sales of the STAR model always exceed those of the AR model. The excess stock are about $2-7\times$ larger in AR models than in STAR models. The lost sales are about $3-5\times$ larger in STAR models than AR models. The number of lost sales are however much smaller than the excess stock. If we assume that the e-commerce company values having minimal excess stock and lost sales equally, then we conclude that STAR outperforms a regular AR model.

If we do not assume that the e-commerce company values excess stock and lost sales equally, then we can calculate the ratio between the costs of excess stock and lost sales at which AR and STAR have the same total performance (equal total costs). Let a be the proportion of the cost price of having lost sales against excess stock, i.e., $p_{\text{LS}} = a \cdot p_{\text{ES}}$. We want to know (for all h) for which a we have that the costs of excess stock and lost sales is equal when using an AR or STAR model. For instance, for $h = 1$ the costs of AR are $855p_{\text{ES}} + 174p_{\text{LS}} = (855 + 174a)p_{\text{ES}}$, and the costs of STAR are $(319 + 335a)p_{\text{ES}}$. These are equal for $a = \frac{855-319}{335-174} \approx 1.71$. Thus, if lost sales are considered to cost less than $1.71\times$ excess stock, then using STAR forecasts results in less costs. The results are displayed in table 9, and we see that the proportion rises with increasing forecasting step.

Within the AR models, the excess stock are much higher than the lost sales, and their proportion

increases with forecasting step. We have discussed before (in section 5.2), that AR forecasting models contain major positive bias for regular low sales. Because of this, forecasts are often much higher than their actual values. Hence, high stocks are maintained to match the forecasts and thus excess stock is high for AR models. As there is a lot of excess stock, the consequence is that the amount of lost sales are often zero. These phenomena can be seen in fig. 6, where we consider the case $h = 3$. Moreover, we see that when a peak is predicted too late and sales are already declining, we are left with a lot of excess stock, and it decreases only slowly as there are not enough daily sales during the period after the peak. This occurs at the peak of the end of June and at the peak of mid August.

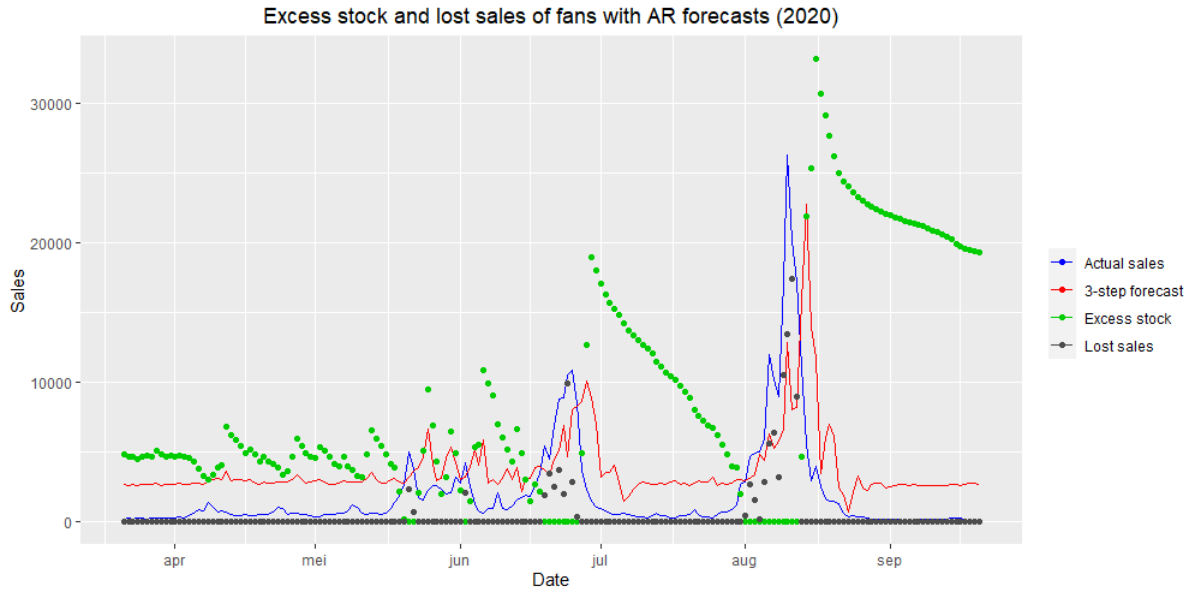

Figure 6: Spring and summer 2020 fan sales (blue) and corresponding 3 step ahead AR forecasts (red). Excess stock is represented by green dots, lost sales by gray dots.

Within the STAR models, the excess stock are smaller than the lost sales, with the exception of the first step. STAR forecasts contain less bias than AR forecasts, and therefore excess stock is generally low. There is, however, somewhat more negative bias present in the STAR forecasts compared to the actual sales, and therefore there are more lost sales here than with AR. This can be seen in fig. 6, where we again visualize excess stock and lost sales for the case $h = 3$, now for STAR forecasts. Like with AR, we see again that whenever sales peaks are predicted with a delay and the sales are already declining, we are left with excess stock in the period after the peak. Notably, the peak at the end of September is predicted very high due to high temperatures. However, since the summer is nearly over, demand of fans remains low and we are left with excess stock for the period

thereafter.

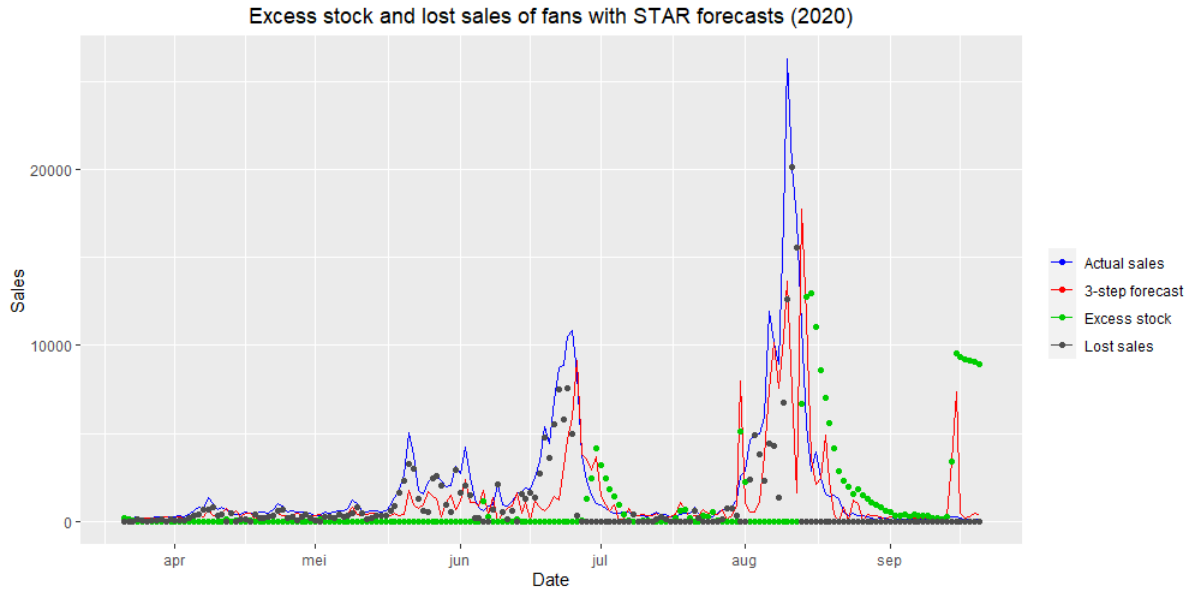

Figure 7: Spring and summer 2020 fan sales (blue) and corresponding 3 step ahead STAR forecasts (red). Excess stock is represented by green dots, lost sales by gray dots.

5.4 Sales models for other weather-related products

A STAR model for fan sales is constructed and forecasts are made with this model in section 5.2. Their economic/business values are expressed in terms of excess stock and lost sales in section 5.3 and this is compared to a standard AR model. Now we want to apply this process to other product chunks, to find out if the forecasting techniques have similar performances, and if the business models are affected differently.

We will investigate a total of five product chunks and summarize all results in tables. Each table contains results of the prior exploration of the time series, results of estimating the model, results of the STAR forecasts and a fragment of the business model. Particularly for the forecasts and the business model, not all results are given. For instance, we limit ourselves to 3- and 6-day ahead forecasts and orders. Although this is not complete, it is representative as the other steps are similar.

The second (after the already investigated fans) product chunk is referred to as chunk2. Results of investigating this series and building the STAR model are given in table 10. The highest correlation between sales and (a shifted version of) the maximum temperature is attained for $s_{3,t} = \max\{z_t, z_{t+1}, \ldots, z_{t+7}\}$. No first differences or a log transformation is applied, and adding

temperature as a linear component to an AR model improves this model in terms of MSPE. The optimal lag order is 8 and $s_{2,t} = z_{t+3}^2$ is chosen as the candidate transition variable, just like for the fan sales. $\hat{\gamma} = 3.08$ and $\hat{c} = 515.73$, both lower than for the fan sales. Due to a lower threshold value, there are more high regime observations (56 in total). Monte-Carlo with no regime specific error variance is chosen as the best forecasting method, and these differ significantly from AR forecasts. These forecasts are about 2 times better in terms of MSPE and about 9 times better in terms of MAPE. MRSTAR again does not improve forecasting results. Average excess stock and lost sales are about $3 - 4\times$ smaller by using STAR forecasts compared to AR. The business model results are displayed in fig. 8. The sales observations of chunk2 show more volatility during Spring and Summer than the fan sales. We can see this behaviour in the forecasts too, making it easier for the STAR model to forecast medium high observations (say, between 1000 and 3000 units sales), although the high peaks are still hard to capture by forecasts. Therefore, the daily sum of excess stock and lost sales are more often distanced from 0 than in the fan sales business model.

Series summary	Min	Median	Mean	Max
	27	250	571.5	10174
Correlation temperature and sales	Highest correlation $\rho = 0.515$ for $s_{3,t} = \max\{z_t, z_{t+1}, \ldots, z_{t+7}\}$,			
First differences	No, $\tau_{\text{ADF}} = -4.09 < -2.86$			
Log transformation	No, STAR nonlinearity rejected			
ARX vs. AR	Improves in MSPE for $h = 4, \ldots, 7$. Improves in MAPE for all h			
Optimal AR lag order	$\hat{m} = 8$			
STAR nonlinearity vs. linearity	Highest F-value attained for $s_{2,t} = z_{t+3}^2$			
Est. parameters of transition function	$\hat{\gamma} = 3.08$ and $\hat{c} = 22.71^2 = 515.73$			
Size of high regime observations set	56 out of 366			
Best forecasting method	Monte Carlo with no regime specific variance			
MSPE and MAPE proportions of best STAR vs. AR	MSPE, $h = 3$	MAPE, $h = 3$	MSPE, $h = 6$	MAPE, $h = 6$
	0.570	0.128	0.503	0.121
Second best forecasting method	Naïve method			
DM tests best STAR vs. AR	For all h, reject H_0: significantly different forecasts			
MRSTAR vs. best STAR	STAR outperforms MRSTAR except for $h = 5$ in terms of MAPE			
Average excess stock + lost sales	3-day orders, AR	3-day orders, STAR	6-day orders, AR	6-day orders, STAR
	1995	592	4549	1445
Break-even price best STAR vs. AR	$p_{ES} = 12.36 p_{LS}$ for $h = 3$ and $p_{ES} = 17.25 p_{LS}$ for $h = 6$			

Table 10: Summarizing table of the chunk2 time series: exploration, STAR model and business model.

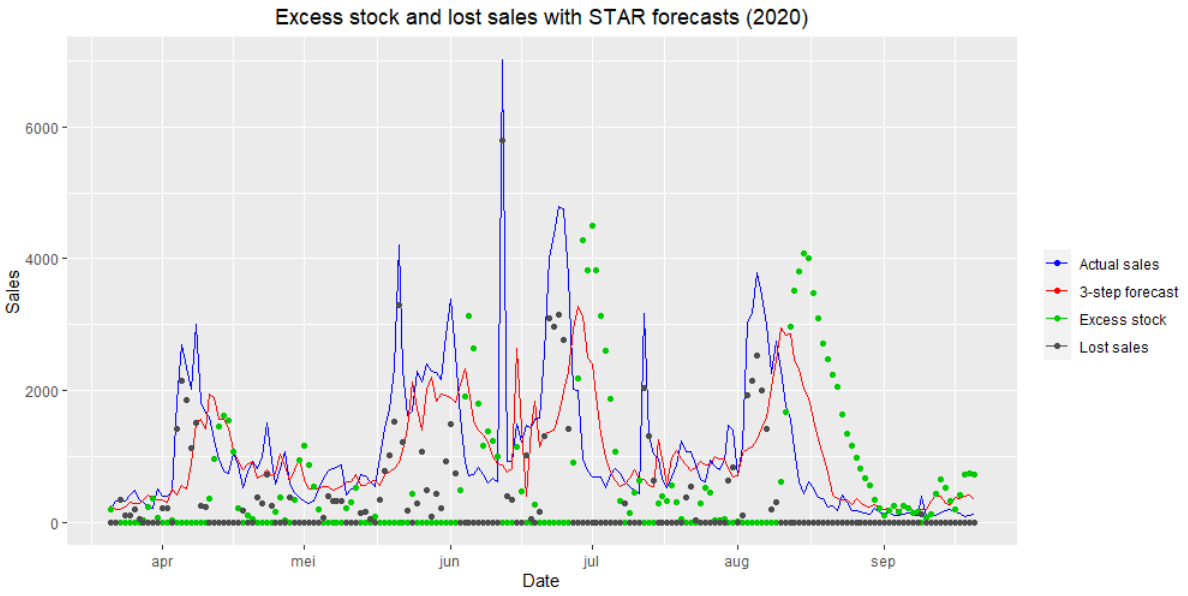

Figure 8: Spring and summer 2020 sales (blue) and corresponding 3 step ahead STAR forecasts (red) of chunk 2. Excess stock is represented by green dots, lost sales by gray dots.

The last three chunks are not presented here. The summarizing tables of chunk3, chunk4 and chunk5 can be found in table A1, table A2 and table A3 respectively. We summarize the results of all five mentioned chunks below.

Testing for first difference always leads to rejection. A log transformation is never adopted, either due to rejection of STAR nonlinearity or because MSPE does not improve compared to a regular (without a log transformation) STAR model. Adding temperature as a linear component to an AR model (i.e., an ARX model) nearly almost improves the AR model. The optimal lag orders are between 6 and 10. Different candidate transition variables are chosen by testing against STAR nonlinearity, but always a squared version and three times it included the third lead temperature at max. Monte-Carlo (either with or without regime specific errors) is always chosen as either the best or second best forecasting technique. Both bootstrap with regime specific errors and direct forecasting are chosen once as the best and once as the second best forecasting method. MSPE is improved by STAR forecasts (compared to AR) by about $1 - 2\times$. MAPE however is improved a lot more by about $10 - 20\times$. MRSTAR rarely ever improves STAR. The exception is chunk5, where MRSTAR improves STAR for mostly the higher forecasting steps (between 5 and 7). Average excess stock is always higher when using AR forecasts, while average lost sales are lower. However, total excess stock always exceeds the total lost sales. Therefore, if excess stock and lost sales are penalized equally, the costs of using an AR model are always higher than of using a STAR model.

6 Conclusion

In this research, we have applied econometric methods on time series of the unit sales of weather related product chunks to answer the research question: "Can sales peaks of weather related products or product groups be predicted by weather data and/or weather forecasts, and if so, how can this be modeled?".

Using accurate forecasting techniques to predict sales is relevant to the e-commerce company and e-commerce companies. For the e-commerce company, sales peaks forecasts are not yet implemented, and it is yet unclear if the demand of products is anticipated on by their stock. With accurate forecasts, it is easier to meet product demand of customers: the amount of excess stock and lost sales decreases, and therefore it cuts the total costs.

To find out whether adding maximum temperature to a time series model improves forecasting results, five product chunks are investigated. For each product chunk, different models are applied and forecasts varying from 1 to 7 steps are made. The default model is the AR model, and we compute forecasting results of other models relative to this AR model. First, adding the maximum temperature linearly to AR into an ARX model always improved MAPE and almost always improved MSPE. Furthermore, we have estimated STAR models with maximum temperature as transition variable, and made forecasts with these models. Not all forecasting techniques were always better than AR, but there was always at least one that outperformed AR in terms of MSPE and MAPE. Thus, adding temperature to a time series model improves forecasting results.

For the STAR models, transformed versions (in the sense of differencing) of the maximum temperature are used. By means of LM tests, the exact temperature lead is chosen. For three chunks, it was the maximum temperature (up to) 3 days ahead, and 2 and 5 for the other two chunks. Therefore, we conclude that sales are best estimated using three day ahead forecasts of the maximum temperature.

Four forecasting techniques are examined on STAR: naïve, Monte-Carlo, bootstrapping and direct forecasting. For Monte-Carlo and bootstrapping, we have also made a distinction by using regime specific error terms or not. Based on MSPE and MAPE, Monte-Carlo without regime specific error terms was the best method. Direct and bootstrap with regime specific error terms were second, although bootstrapping is preferred, since can be computer intensive as for every forecasting step a new model needs to be estimated. Generally, the MSPE of STAR was about 50% and the MAPE often less then 10% of an AR model. This was due to the fact that sales peaks are hard to estimate

for both models. The residuals are therefore high in both models. This has a relatively large contribution to MSPE, but not to MAPE, as the latter metric is more robust to outliers. Finally, MRSTAR forecasts almost never improved STAR forecasts, with the exception of one chunk for the higher forecasting steps (4 to 7).

Finally, we have determined the economic impact of using different forecasts by making a business model. This business model consisted of excess stock and lost sales, both desired to be minimized. For AR models there was more excess stock present but less lost sales. This is because AR models tend to overestimate observations more than STAR. Nevertheless, the amount of excess stock always exceeded the amount of lost sales. Therefore, assuming that the price of having excess stock and lost sales is equal, STAR cuts the total costs compared to AR.

We conclude that sales (particularly, sales peaks) of weather related product groups (chunks) can be predicted by using previous sales and weather forecasts. It is best to model this using a STAR model, particularly using Monte-Carlo forecasts without drawing regime specific error terms. It is however still hard for a STAR model to predict a sales peak, but if sales peaks occur more gradually, then STAR is still able to capture these peaks. STAR reduces the amount of excess stock vastly, but increases the amount of lost sales slightly.

We recommend the e-commerce company to utilize STAR models for weather related products. The stock should be adjusted to the STAR forecasts. Moreover, the forecasts should be communicated to partners of the e-commerce company, adjusted to the total output of each partner, such that they can adjust their stock too. In this manner, we improve the inventory management and we reduce the total amount of costs.

7 Limitations and further research

In this section, we discuss limitations of this paper and provide potential topics for future research.

In predicting sales, we have used weather forecasts. Unfortunately, these forecasts were unavailable, and hence we have assumed that the temperature forecasts were equal to the actual temperatures. However, we have substantiated this by noting that research of the ECMWF states that current temperature forecasts are accurate (in the sense of a maximum deviation of one degree Celsius) up to 7 days ahead. Nevertheless, it might be better to use actual forecasts, such that the temperature forecasts are not necessarily equal to their true value.

We have seen that observations of low unit sales are often overestimated, and observations of high

unit sales are often underestimated. For an AR model, there is not enough power to capture both peak observations and more regular low observations. For STAR, this also holds to a certain extent. Generally, in estimating a model, a trade-off has to be made between these types of observations. We might be able to counter this by applying other values of γ, because transitioning between regimes is then much smoother (or less). Even though this value is determined by the estimation procedure, it does not mean that this value is optimal for different forecasting steps. We can for instance try different γ's and compute forecasting results, and comparing these might lead to a different choice of γ.

STAR makes it possible for the time series to blow up easily when temperature is high. This works well during the late spring and summer. However, outside this period there might still be high temperatures, but little demand. This occurs for instance in the fan sales at the end of September. Due to hot upcoming weather, a sales peak is forecast. However, because it is that late in the season, the demand is not rising. The consequence is a high amount of excess stock. We might prevent this by including seasonal dummy variables in the model. They can be 1 for the second half of the spring and the entire summer (and 0 elsewhere) for instance. Another idea is to use ESTAR as transition function in combination with a time transition variable. This function is low (near zero) around a threshold value, and gradually (and symmetrically) goes to one. Thus, this indicates a high regime near a certain date in the year, and the low regime consists of dates far away from that date.

In this paper we have treated overestimation and underestimation of forecasts with equal penalty (particularly, in STAR). Consequently, the excess stock and lost sales are treated equally as well. However, it might be that the e-commerce company values having low lost sales more (not entirely) than having low excess stock. Future research might therefore include estimating STAR models where underestimation of forecasts are penalized heavier than overestimation.

Other future research might be to include more chunks. In this paper, only five chunks are considered. This is not enough to make general (in the sense of all weather related products) conclusions about which forecasting technique outperforms the others for instance. Furthermore, we have seen serial correlation in the error terms for every chunk. We can account for this by using GARCH error terms. Indeed, GARCH is appropriate for when error terms are serially correlated, or if heteroskedasticity is present. Finally, it might be interesting to look into promotion of products, as this might cause sales peaks too. This can be modeled as a promotion dummy variable, capturing the discrepancy of the promotional period.

References

Alimi, M., Rhif, A., & Rebai, A. (2017). Nonlinear dynamic of the renewable energy cycle transition in tunisia: Evidence from smooth transition autoregressive models. *International Journal of Hydrogen Energy*, *42*(13), 8670–8679.

Al-Zubaidi, H., & Tyler, D. (2004). A simulation model of quick response replenishment of seasonal clothing. *International Journal of Retail & Distribution Management*, *32*(6), 320–327.

Bahng, Y., & Kincade, D. (2012). The relationship between temperature and sales: Sales data analysis of a retailer of branded women's business wear. *International Journal of Retail & Distribution Management*, *40*.

Chan, K. S., & Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, *7*(3), 179–190.

Cheung, Y.-W., & Lai, K. S. (1995). Lag order and critical values of the Augmented Dickey–Fuller test. *Journal of Business & Economic Statistics*, *13*(3), 277–280.

Christopher, M., Lowson, R., & Peck, H. (2004). Creating agile supply chains in the fashion industry. *International Journal of Retail & Distribution Management*, *32*(8), 367–376.

Cui, H., & Peng, X. (2015). Short-Term City Electric Load Forecasting with Considering Temperature Effects: An Improved ARIMAX Model. *Mathematical Problems in Engineering*, *2015*, 1–10.

Dell, M., Jones, B. F., & Olken, B. A. (2014). What do we learn from the weather? The new climate-economy literature. *Journal of Economic Literature*, *52*(3), 740–98.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.

Eitrheim, Ø., & Teräsvirta, T. (1996). Testing the adequacy of smooth transition autoregressive models. *Journal of Econometrics*, *74*(1), 59–75.

Haiden, T., Janousek, M., Vitart, F., Ben-Bouallegue, Z., Ferranti, L., & Prates, F. (2021). Evaluation of ECMWF forecasts, including the 2021 upgrade. (884).

Hsu, K.-C., & Chiang, H.-C. (2011). Nonlinear effects of monetary policy on stock returns in a smooth transition autoregressive model. *The Quarterly Review of Economics and Finance*, *51*(4), 339–349.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439–454.

Leybourne, S., Newbold, P., & Vougas, D. (1998). Unit roots and smooth transitions. *Journal of Time Series Analysis*, *19*(1), 211–228.

Liu, J., Liu, C., Zhang, L., & Xu, Y. (2020). Research on sales information prediction system of e-commerce enterprises based on time series model. *Information Systems and e-Business Management*, *18*(4), 823–836.

Lundbergh, S., & Teräsvirta, T. (2002). Forecasting with smooth transition autoregressive models. *A companion to economic forecasting*, 485–509.

Lundbergh, S., Teräsvirta, T., & Van Dijk, D. (2003). Time-varying smooth transition autoregressive models. *Journal of Business & Economic Statistics*, *21*(1), 104–121.

Lütkepohl, H., & Xu, F. (2012). The role of the log transformation in forecasting economic variables. *Empirical Economics*, *42*(3), 619–638.

Luukkonen, R., Saikkonen, P., & Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika*, *75*(3), 491–499.

Permatasari, C. I., Sutopo, W., & Hisjam, M. (2018). Sales forecasting newspaper with ARIMA: A case study. *AIP Conference Proceedings*, *1931*(1), 030017.

Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*, *34*, 151–163.

Steele, A. T. (1951). Weather's effect on the sales of a department store. *Journal of Marketing*, *15*(4), 436–443.

Steinker, S., Hoberg, K., & Thonemann, U. W. (2017). The value of weather information for e-commerce operations. *Production and Operations Management*, *26*(10), 1854–1874.

Teräsvirta, T., & Anderson, H. M. (1992). Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics*, *7*(S1), S119–S136.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, *89*(425), 208–218.

van Dijk, D. (1999). *Smooth transition models: Extensions and outlier robust inference* (tech. rep. No. 200).

van Dijk, D., Teräsvirta, T., & Franses, P. H. (2002). Smooth transition autoregressive models — a survey of recent developments. *Econometric Reviews*, *21*(1), 1–47.

Verstraete, G., Aghezzaf, E.-H., & Desmet, B. (2019). A data-driven framework for predicting weather impact on high-volume low-margin retail products. *Journal of Retailing and Consumer Services*, *48*, 169–177.

Wang, L., Zou, H., Su, J., Li, L., & Chaudhry, S. (2013). An ARIMA-ANN Hybrid Model for Time Series Forecasting. *Systems Research and Behavioral Science*, *30*(3), 244–259.

Wright, S., & Nocedal, J. (1999). Numerical optimization. *Springer Science*, *35*(67-68), 7.

A Tables and figures

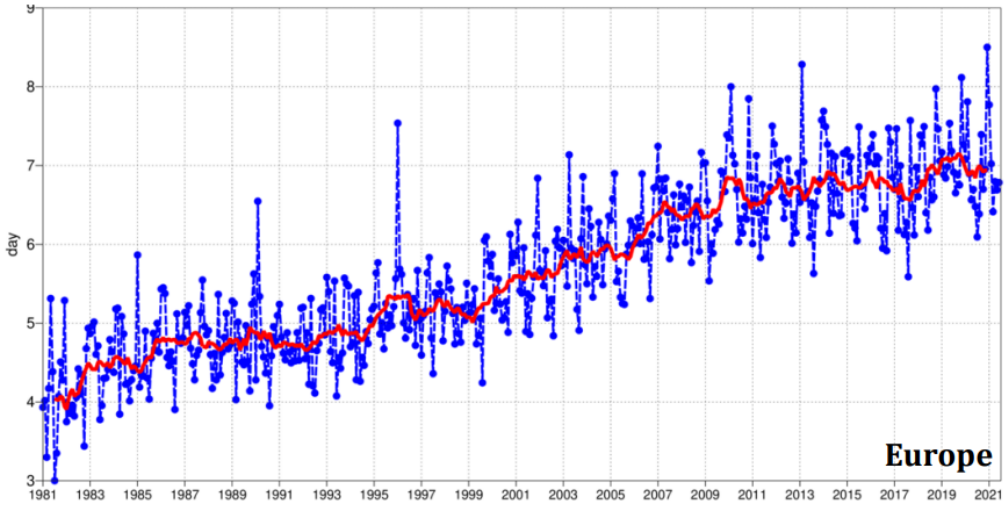

Figure A1: Development of the quality of weather forecasts in Europe between 1981 and 2021. The period in days for which the location of high- and low-pressure areas were accurately predicted over that month is given in blue, with the 12 month moving average in red. Source: European Centre for Medium-Range Weather Forecasts (ECMWF), see Haiden et al. (2021).

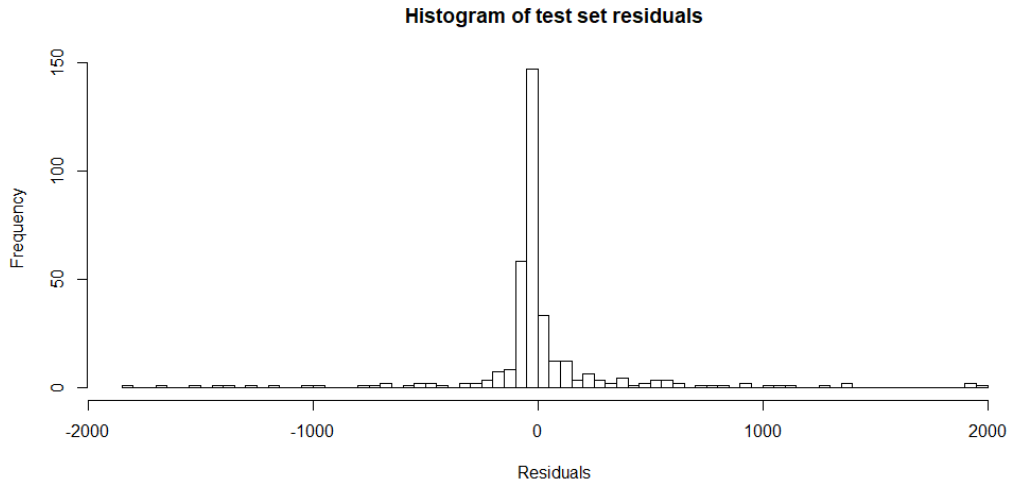

Figure A2: Histogram of the test set STAR model residuals. Due to visualization purposes, residuals having an absolute value higher than 2000 units are omitted (12 in total).

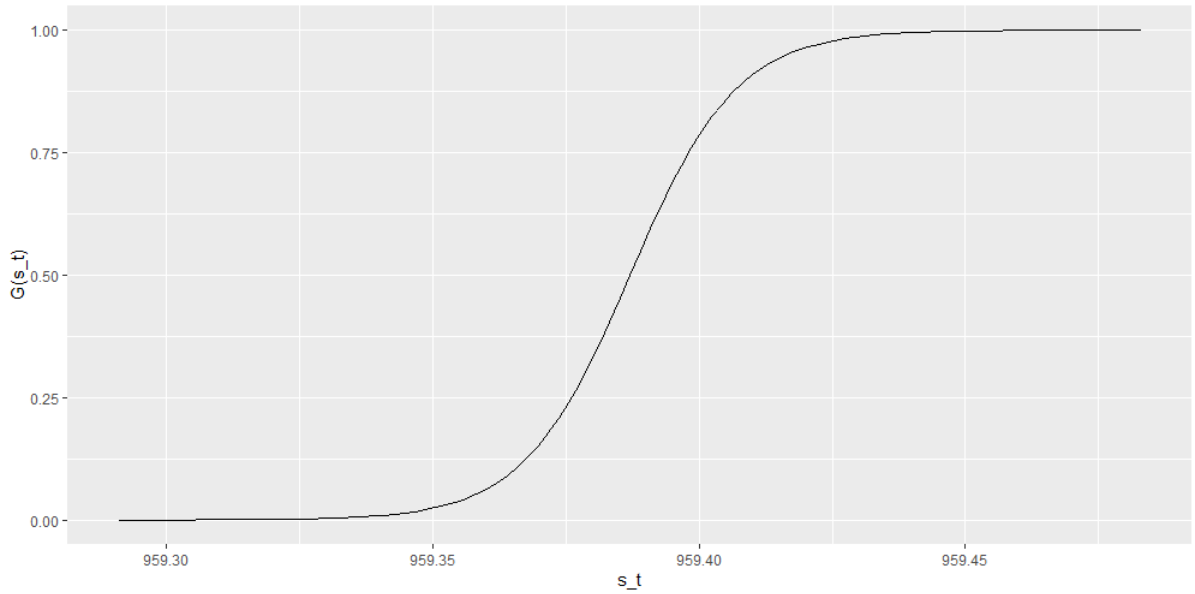

Figure A3: Logistic transition function G, with transition variable $s_t = z_{t+3}$, and parameters $\gamma = 100.00$ and $c = 30.97^2$.

Series summary	Min	Median	Mean	Max
	5	43	369.7	9758
Correlation temperature and sales	Highest correlation $\rho = 0.451$ for $s_{3,t} = \max\{z_t, z_{t+1}, \ldots, z_{t+3}\}$			
First differences	No, $\tau_{\text{ADF}} = -6.15 < -2.86$			
Log transformation	No, MSPE of log transformed STAR is 1.015 times MSPE of regular STAR			
ARX vs. AR	Improves in MSPE for $h = 2, \ldots, 7$. Improves in MAPE for all h			
Optimal AR lag order	$\hat{m} = 6$			
STAR nonlinearity vs. linearity	Highest F-value attained for $s_{4,t} = \max\{z_t, z_{t+1}, \ldots, z_{t+3}\}^2$			
Est. parameters of transition function	$\hat{\gamma} = 1.43$ and $\hat{c} = 26.61^2 = 708.08$			
Size of high regime observations set	44 out of 366			
Best forecasting method	Monte Carlo with no regime-specific error variance			
MSPE and MAPE proportions of	MSPE, $h = 3$	MAPE, $h = 3$	MSPE, $h = 6$	MAPE, $h = 6$
best STAR vs. AR	0.979	0.058	0.639	0.041
Second best forecasting method	Bootstrap with regime-specific error terms			
DM tests best STAR vs. AR	In terms of MSPE, forecasts do not differ. In terms of MAPE, forecasts differ.			
MRSTAR vs. best STAR	STAR outperforms MRSTAR except for $h = 3$ in terms of MAPE			
Average excess stock + lost sales	3-day orders, AR	3-day orders, STAR	6-day orders, AR	6-day orders, STAR
	1342	677	4111	1967
Break-even price best STAR vs. AR	$p_{ES} = 4.00 p_{LS}$ for $h = 3$ and $p_{ES} = 8.32 p_{LS}$ for $h = 6$			

Table A1: Summarizing table of the chunk3 time series: exploration, STAR model and business model.

Series summary	Min	Median	Mean	Max
	1	43	204.1	3310
Correlation temperature and sales	Highest correlation $\rho = 0.609$ for $s_{3,t} = \max\{z_t, z_{t+1}, \ldots, z_{t+3}\}$			
First differences	No, $\tau_{\text{ADF}} = -4.96 < -2.86$			
Log transformation	No, MSPE of log transformed STAR is 1.346 times MSPE of regular STAR			
ARX vs. AR	Improves in MSPE and MAPE for all h			
Optimal AR lag order	$\hat{m} = 9$			
STAR nonlinearity vs. linearity	Highest F-value attained for $s_{2,t} = z_{t+5}^2$			
Est. parameters of transition function	$\hat{\gamma} = 31.46$ and $\hat{c} = 24.08^2 = 579.75$			
Size of high regime observations set	40 out of 366			
Best forecasting method	Bootstrap with regime-specific error terms			
MSPE and MAPE proportions of best STAR vs. AR	MSPE, $h = 3$	MAPE, $h = 3$	MSPE, $h = 6$	MAPE, $h = 6$
	0.964	0.089	0.933	0.056
Second best forecasting method	Monte Carlo with no regime-specific error variance			
DM tests best STAR vs. AR	In terms of MSPE, forecasts do not differ. In terms of MAPE, forecasts differ.			
MRSTAR vs. best STAR	STAR outperforms MRSTAR for all h			
Average excess stock + lost sales	3-day orders, AR	3-day orders, STAR	6-day orders, AR	6-day orders, STAR
	527	232	1490	803
Break-even price best STAR vs. AR	$p_{ES} = 4.18 p_{LS}$ for $h = 3$ and $p_{ES} = 6.02 p_{LS}$ for $h = 6$			

Table A2: Summarizing table of the chunk4 time series: exploration, STAR model and business model.

Series summary	Min	Median	Mean	Max
	0	21	121.4	1455
Correlation temperature and sales	Highest correlation $\rho = 0.656$ for $s_{3,t} = \max\{z_t, z_{t+1}, \ldots, z_{t+3}\}$			
First differences	No, $\tau_{\text{ADF}} = -5.03 < -2.86$			
Log transformation	No, STAR nonlinearity is rejected			
ARX vs. AR	Improves in MSPE and MAPE for all h			
Optimal AR lag order	$\hat{m} = 10$			
STAR nonlinearity vs. linearity	Highest F-value attained for $s_{2,t} = z_{t+2}^2$			
Est. parameters of transition function	$\hat{\gamma} = 2.70$ and $\hat{c} = 25.21^2 = 635.70$			
Size of high regime observations set	30 out of 366			
Best forecasting method	Monte-Carlo with regime-specific error terms			
MSPE and MAPE proportions of best STAR vs. AR	MSPE, $h = 3$	MAPE, $h = 3$	MSPE, $h = 6$	MAPE, $h = 6$
	0.592	0.068	0.418	0.054
Second best forecasting method	Direct forecasting			
DM tests best STAR vs. AR	For all h, forecasts differ significantly			
MRSTAR vs. best STAR	MRSTAR outperforms STAR for $h = 5, 6, 7$ (MSPE) and $h = 4, 6$ (MAPE)			
Average excess stock + lost sales	3-day orders, AR	3-day orders, STAR	6-day orders, AR	6-day orders, STAR
	461	121	1409	217
Break-even price best STAR vs. AR	$p_{ES} = 5.94 p_{LS}$ for $h = 3$ and $p_{ES} = 12.98 p_{LS}$ for $h = 6$			

Table A3: Summarizing table of the chunk5 time series: exploration, STAR model and business model.

B Theory

B.1 Partial derivatives of the logistic function

The logistic transition function G is given by $G(s_t;\gamma,c)=\frac{1}{1+e^{-\gamma(s_t-c)}}$. Define $a:=s_t-c$, then the transition function is represented as $G(\gamma)=\frac{1}{1+e^{-a\gamma}}$. Its first three partial derivatives with respect to γ are:

$$\begin{aligned}
\frac{\partial G}{\partial \gamma} &= \frac{ae^{-a\gamma}}{\left(e^{-a\gamma}+1\right)^2};\\
\frac{\partial^2 G}{\partial \gamma^2} &= \frac{\left(e^{-a\gamma}+1\right)^2\cdot -a^2e^{-a\gamma}-ae^{-a\gamma}\cdot 2\left(e^{-a\gamma}+1\right)\cdot -ae^{-a\gamma}}{\left(e^{-a\gamma}+1\right)^4}\\
&= -a^2\frac{e^{-a\gamma}-e^{-2a\gamma}}{\left(e^{-a\gamma}+1\right)^3};\\
\frac{\partial^3 G}{\partial \gamma^3} &= -a^2\frac{\left(e^{-a\gamma}+1\right)^3\cdot\left(-ae^{-a\gamma}+2ae^{-2a\gamma}\right)-\left(e^{-a\gamma}-e^{-2a\gamma}\right)\cdot 3(e^{-a\gamma}+1)^2\cdot -ae^{-a\gamma}}{\left(e^{-a\gamma}+1\right)^6}\\
&= -a^3\frac{-e^{-2a\gamma}+2e^{-3a\gamma}-e^{-a\gamma}+2e^{-2a\gamma}+3e^{-2a\gamma}-3e^{-3a\gamma}}{\left(e^{-a\gamma}+1\right)^4}\\
&= a^3\frac{e^{-a\gamma}-4e^{-2a\gamma}+e^{-3a\gamma}}{\left(e^{-a\gamma}+1\right)^4}.
\end{aligned}$$

If we substitute $a=s_t-c$ back, it follows that

$$\left.\frac{\partial G}{\partial \gamma}\right|_{\gamma=0}=\frac{1}{4}(s_t-c),\quad \left.\frac{\partial^2 G}{\partial \gamma^2}\right|_{\gamma=0}=0,\ \text{and}\ \left.\frac{\partial^3 G}{\partial \gamma^3}\right|_{\gamma=0}=-\frac{1}{8}(s_t-c)^3.$$

B.2 The BFGS algorithm

The BFGS algorithm is a Quasi-Newton method to solve unconstrained nonlinear optimization. It is named after its discoverers Broyden, Fletcher, Goldfarb and Shannon. Suppose we want to minimize $f(x)$, with $x\in\mathbb{R}^n$ and f a differentiable scalar function. The idea is to start from an initial point x_0 and iteratively obtain better estimates by performing a line search based on the previous estimate, the gradient and an approximate Hessian. The BFGS algorithm is given in Algorithm 1, see (Wright & Nocedal, 1999). For initial inverse Hessian H_0, we can simply choose the identity matrix. The constants in the Wolfe conditions are often chosen to be $c_1=10^{-4}$ and $c_2=0.9$.

Algorithm 1: BFGS

Input: Initial value x_0;

Function f with gradient ∇f;

An inverse approximate Hessian H_0;

Convergence tolerance $\varepsilon > 0$.

$k \leftarrow 0$

while $\|\nabla f_k\| > \varepsilon$ **do**

Compute the search direction $p_k = -H_k \nabla f_k$.

Set $x_{k+1} = x_k + \alpha_k p_k$, where α_k is obtained by line search to satisfy the Wolfe conditions:

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k' p_k,$$

$$\nabla f(x_k + \alpha_k p_k)' p_k \geq c_2 \nabla f_k' p_k,$$

with $0 < c_1 < c_2 < 1$;

Define $s_k = x_{k+1} - x_k$ and $y_k = \nabla f_{k+1} - \nabla f_k$;

Compute the next inverse approximate Hessian H_{k+1} by:

$$H_{k+1} = (I - p_k s_k y_k') H_k (I - p_k y_k s_k') + p_k s_k s_k';$$

$k \leftarrow k + 1$;

end