IADIS International Journal on Computer Science and Information Systems
Vol. 15, No. 2, pp. 58-75
ISSN: 1646-3692

PREDICTING CONSUMER GOODS PRICES – THE SHORT-, MEDIUM- AND LONG-TERM PERSPECTIVE

Anne Falkenberg[1] and Benjamin Buchwitz[2]
[1]*Catholic University of Eichstaett-Ingolstadt, Auf der Schanz 49, 85049 Ingolstadt, Bavaria, Germany.*
[2]*South Westphalia University of Applied Sciences. Lindenstraße 53, 59872 Meschede, North Rhine-Westphalia, Germany*

ABSTRACT

The large number of available prices gave rise to price comparison sites that yet lack recommendation services to support their customers in scheduling buying decisions. Using a large data set with 1.46 million daily minimum price observations for four product categories of electronic consumer goods, we outline reasons for the slow adoption of recommendation services, evaluate 6.56 billion price forecasts and show that forecasted product prices can be used to build recommendation services to advise customers on their purchase time decision. We compare 16 different methods that can act as the core of price prediction services and give detailed insights on performance as well as advice on model selection for short-, medium and long-term forecasting horizons, outline differences by category and advocate the transition towards prescriptive price analytics services.

KEYWORDS

Price Forecasting, Big Data, e-Commerce, Price Time Series, Predictive Analytics

1. INTRODUCTION

Fundamental changes in consumer behavior, competition landscape, and cost structures led to an enormous complexity growth in the retail environment over the past two decades. Consumers increasingly adopt online and mobile retail channels, which not only offer more convenient shopping experiences but also increase price transparency. Despite the universal availability of price information that allows comparing different offers, a wide range of prices exists especially for standardized and homogeneous or highly comparable products (Pan et al., 2002). A vital part and key element of modern retail business models are pricing strategies that critically shape retailers' margins and profits (Bolton et al., 2006).

1.1 Price Comparison and Recommendation Services

Price changes in traditional channels were rather rare. Now products are available online and pricing receives much more attention and price-setting behavior becomes more dynamic (Kannan et al., 2001; Kopalle et al., 2009; Levy et al., 2004). These price changes partially follow rational patterns and prices are adjusted based on demand and inventory considerations or as differentiation from other e-commerce platforms. Smaller or specialized retailers however often follow opportunistic patterns where the objectives are not transparent or price changing actions are not guided by algorithms. This uncertainty for the customer in combination with their high price sensitivity especially for homogeneous consumer goods gave rise to price aggregation platforms that act as a facilitator and allow customers to easily gain an overview of the retail landscape.

Apart from static or historical information, price comparison sites (PCS) like *idealo* or *shopbrain* recently started offering services to actively support customers in the buying process and enable them to actually use the provided information to optimize their purchase time decisions. The core of these services is usually a methodology to extrapolate prices based on historical information that allows condensing expectations about future price developments. The airfare industry discovered the usefulness of such analytics services quite early and is using them effectively. However, compared to those predictive or prescriptive analytics services, which provide active decision support based on validated predictions, the approaches observable in the consumer goods retail market are quite elementary.

1.2 Forecast Setting, Data Structure and Research Goal

While it seems surprising that there is such a large difference between the airfare and consumer goods industries, this can be at least partially explained by characteristics of the underlying data generation process, the available data and, therefore, methodologies disposable to solve the underlying forecasting problem.

Given a product or service that a customer wants to purchase, s/he is usually not interested in the price from one specific seller. If a customer is not willing to choose from a range of quality controlled and trusted sellers but sticks to one retailer, there will be no need for her/him to actually monitor the retail landscape via an aggregation platform. The entire business model of PCS, therefore, grounds on the users' flexibility about the place of purchase. The price of an individual retailer alone is no longer of primary interest to the customer. Instead, the minimum of all prices from the listed retailers constitutes the relevant reference price. The corresponding composed data source of daily minimum prices for one specific product is crucial to the customer and represents what PCS usually display and report on their webpages and in their historical pricing charts.

This forecasting setting leads to price time series with specific characteristics. Product prices of consumer goods usually deteriorate over time, thus exhibit a time-dependent level. Price adjustments occur irregularly and with varying magnitude leading to calmer and more active price changing phases as well as to entirely constant segments. Time series of minimum prices exhibit similar characteristics than the ones from individual sellers but show more extreme values and shorter, more irregular segments with constant prices. When analyzing the day-to-day price movements and therefore the sequence of price changes these constant segments lead to many observations that are zero, which is referred to as zero-inflation (Kömm

and Küsters, 2015; Winkelmann, 2008). Many traditional forecasting techniques assume a strictly continuous distribution of the forecasted data (Hyndman and Athanasopoulos, 2018). Obviously, this assumption is violated in the case of the price time series for consumer goods. However, methodological approaches that explicitly consider these shortcomings are rare, computational cumbersome and thus not scalable or directly applicable to forecast durable consumer goods prices (Buchwitz and Küsters, 2018; Rydberg and Shephard, 2003; Sucarrat and Grønneberg, 2016). The specific nature of the forecasting setting, as well as different challenges evolving from the concerning industry, customer environment and the specific time series characteristics, make forecasting for minimum price time series from multiple product categories demanding. However, it is necessary if PCS want to enhance their service to help customers optimize their purchase time decisions.

In this paper, we, therefore, aim to investigate the forecasting performance of available and scalable forecasting techniques for analytics services when applied to composed minimum price time series of consumer goods. In doing so, we evaluate and compare 16 methods for short-, medium- and long-term forecast horizons exemplary on 4 product groups: refrigerators, computer memory, graphic cards and smartphones.

We focus on short-term forecast horizons for two reasons: First, it is the most relevant for customers' buying decisions. Second, notification services like price alarms have an average waiting time up to two months for a 5% price reduction independent of the product category (idealo, 2017), showing that this time period is highly relevant for the enhancement of shopping support services. As a consequence of the business model of price comparison sites, a significant part of the user basis is highly price sensitive. Some of these users prefer to plan their purchases on the long run (which is especially true for some European countries due to cultural differences) and wait for large price drops and external shocks, like the introduction of follow-up models or introductions of competing products, so that we add medium-term forecast horizons to our study. While the short- and medium-term forecasts aim at designing data-driven support systems for customers directly, PCS also collaborate closely with the listed retailers. As part of their efforts, they firmly integrate with their sellers' IT systems and feedback data that helps to steer inventory management, ordering systems and spare parts storage. During our research, we were confronted with evidence that precise long-term price forecasts (which are interpretable as a forecast of the average residual value in the market) can help to optimize the supply chain further, which is the reason for also covering long-term forecast horizons that expand to the end of the products life cycles.

Our research is to the best of our knowledge the first contribution that benchmarks the forecasting performance of core methodologies for prices of consumer durable goods for such a broad range of product categories and for different forecast horizons. The obtained insights help to advance PCS towards building predictive and prescriptive services and lays the foundation for implementing and developing those services.

2. DATA SET

The analysis is based on a sample consisting of daily minimum price observations from the German e-commerce market for 2,000 electronic consumer goods ranging from home appliances over computer hardware to smartphones, equally split into four categories. To make results comparable, we focus on the length of a typical life cycle for electronic consumer

products and analyze the first two years of data resulting in 730 observations per item. In total, this yields a dataset with 1.46 million daily observations. Products with less than 730 observations, not enough price movements or obvious data errors have not been considered and were removed beforehand. All items stem from well-known and established brands and each time series represents a specific entity with completely homogeneous properties and features, meaning different product configurations, sizes or colors constitute different time series.

Table 1. Descriptive statistics by product category

Category	**Avg. Initial Price in €**	**Average Price in €**	**Average Daily Price Change in €**	**% of Constant Observations (zero-inflation)**	**N**
Computer Memory	237.47	145.32	-0.17	38 %	500
Graphic Cards	279.86	235.14	-0.05	45 %	500
Refrigerators	829.04	709.79	-0.19	72 %	500
Smartphones	496.71	333.29	-0.29	64 %	500
All	460.77	355.89	-0.18	55 %	2,000

Table 1 provides details and descriptive statistics for the nominal price level and the price changing frequency for the four product categories – refrigerators, computer memory, graphic cards, and smartphones. The products with the highest average initial price are refrigerators with 829.04€ directly followed by smartphones with 496.71€. However, the single highest initial price with roughly 3,900€ comes from an AMD Graphic Card for Professional Workstations. Interestingly, the average daily price change is not directly linked to the price level so that computer memories reach an average daily price reduction of 0.17€, compared to graphic cards that only deteriorate with a rate of 0.05€ per day-to-day price movement. The average daily price change of all products in the sample amounts to -0.18€. The most volatile products in terms of occurring price changes are memory modules, where only 38% of the observations do not change compared to the price from the previous day. Compared to this, refrigerator prices are more constant. With 72% zero-inflation in the data, the minimum price for a refrigerator changes on average about twice a week.

Figure 1 provides an aggregated overview of the utilized data for the four considered product categories. Each of the figures (1a-1d) shows the average relative price development over the first two years of the product life cycle for the respective category. Given a single time series $\{y_t\}_{t\in\mathbb{N}}$, each daily price y_t is divided by the initial price y_1 to obtain the relative price development. The solid black line displays the relative price averaged over all products of the respective category. The shaded areas around the solid black line correspond to the 50%, 75% and 90% empirical confidence intervals of the relative prices. On average, all prices decrease over time but show different intensities of the price deterioration. Besides, the longer the products are in the market, the wider the confidence intervals become. While this is nicely illustrated in figures 1c and 1d, for refrigerators the intervals around the average relative price stay almost constant independent of the products' time in the market. On average, the price level for the home appliance products (1a) drops rapidly but the steepness quickly decreases with the price level before stabilizing at around 85% of their starting price. While this is very comparable with the price development of graphic cards shown in figure 1c, the average price level here starts to slightly increase at the end of the life cycle. Graphic cards show an identifiable global minimum approximately 1.5 years after their product launch, which cannot be observed in the other three categories.

Both – computer memory modules and smartphones – show a comparably strong price deterioration. However, their respective slope also decreases with the level, but computer memory prices stabilize at around 65% of the starting price as shown in figure 1b. The settlement point for the smartphone product category is even lower with around 60%. However, while the general price dynamic of these two categories seem comparable, computer memories exhibit their own momentum. The intervals shown in figure 1b are almost three times wider than the ones shown in the remaining graphs despite representing the same quantiles. This indicates strong heterogeneity between individual items and may also be connected to the low zero-inflation within this category.

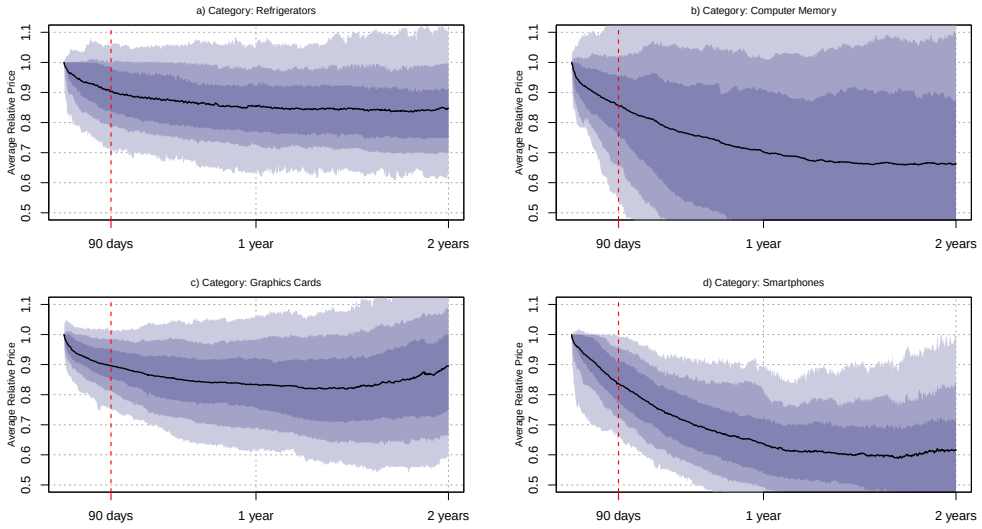

Figure 1. Relative price developments for each product category

Summing up the average price development, one can state that the overall shape is comparable for all average relative prices as each category shows deterioration, however, due to variations in slope and settlement levels, the average appearance varies. Prices that exceed the initial price are rare for refrigerators and do almost not occur for smartphones, which is coherent with the previously mentioned strong price deterioration. However, they are widely present for memory modules, which can be partially explained with the unique and difficult market dynamics in the storage industry in recent years which forced retailers to adjust prices on a regular basis.

3. METHODOLOGY

As shown by the descriptive statistics and the aggregated graphical representations of the price time series in figure 1, forecasting approaches have to deal with a variety of paradigms that are present in the data. Besides the already mentioned non-transparent data generation process and its implications, this includes products with rapidly changing dynamics, potentially time-varying mean structures and local up- or downward movements that overlay the global

price development. Due to the variety and variability of the time series, it is unrealistic to expect a single method to deliver the best forecasts independent of these specific characteristics and independent of the forecast horizon. The approaches in the chosen methodological corpus, therefore, possess different functionalities that can deal with and dynamically adapt to the features in the price data.

3.1 Forecasting Models

In total, we estimate and benchmark over 40 different methods and configurations for our analysis. Shown results focus on methods that are computationally robust and specialized for time series forecasting applications. While modern statistical methods like Deep Neural Networks or Artificial Intelligence deliver very good results in many situations, they perform surprisingly bad on time series data (Makridakis et al., 2018). The few exceptions (e.g. Smyl et al., 2018) consist of highly engineered frameworks that incorporate excessive amounts of domain-specific knowledge, which cannot be relied on in the case of minimum price time series. However, we comment more on the usage of modern methods throughout the paper but focus on a more conservative methodological corpus for now. As several of the chosen approaches belong to the same model families, represent simple or special cases of modified models or can be differentiated by configuration or pre- and post-processing options, the approaches can be grouped into the following seven classes. Due to the restricted length of this article, we only briefly outline the main ideas or working principles behind the methods but restrain from completely describing or deriving the respective forecasting functions. However, the technical details are cited accordingly. For each class, we focus on two to four configuration options, which leads to a selection of 16 methods, that we elaborate exclusively on in the upcoming results section. Additionally, a more in-depth description and breakdown of the results can be obtained from the authors upon request.

Global and Local Trend: Given the fact that the majority of electronic consumer product prices deteriorate over time, models that assume and estimate a global trend, present an intuitive benchmark for more complex approaches. We, therefore, include linear time series regression models (Makridakis et al., 1997) that are estimated using ordinary least squares as a benchmark in our comparative study. While the simplest form models a global linear trend (*Trend*), we also generate forecasts using an exponential trend (*Trend exp.*) and employ a configuration that contains a quadratic term (*Trend sq.*) to account for different shapes of price deterioration. As the data contains local dynamics as well, it may be plausible to restrict the amount of data used for the trend extrapolation to make the model more adaptive to local changes. For this reason and to be comparable to the aggregation platform *Bidvoy* – one of the few commercial platforms that actually communicates extrapolations to users – we generate forecasts based on a quarterly sliding window (*Trend 90*). However, while this localization allows the trend to react (partially) to recent developments in the series, generated forecasts are not tied to the price level of the last observation. Trend curves can lie significantly below or above the local price level, which is why we expect poor performance for small forecast horizons when extrapolating global trends. Following the same logic, we expect the localized trends to be outperformed when large forecast horizons are of interest.

Mean Models formulate expectations for future price developments based on the arithmetic mean and, therefore, assume that deviations from the mean are based on random perturbations. As seen in figure 1, electronic consumer products show strong price deterioration, thus this

assumption may not hold. For this reason and due to the missing alignment to the last observation, mean-based forecasts are expected to perform poorly in general and especially poor on short forecast horizons. Following the approach applied by *eBay* to signal price tendencies to their users, we include a global mean model (*Mean*) and the arithmetic mean based on the last 90 observations (*Mean 90*) in our comparison. Both trend and mean models together represent the forecasting approaches known to be currently in place and observable in practice and are, therefore, used mainly as a benchmark for more complex approaches.

Random Walk: Assuming the difference between two consecutive price observations is just white noise, pure random walk models (*RW*) use the last available observation as the forecast for all future horizons. While mean and random walk models implement two divergent principles, the mean forecast converges towards a random walk prediction with increasing zero-inflation. Given that there are multiple constant segments in the price time series and that prices show strong price deterioration, we enhance the random walk with a drift term (*RW Drift*) (Hyndman and Athanasopoulos, 2018). This allows the model to incorporate a global tendency while eradicating some shortcomings of the presented trend approaches. Forecasts generated by random walk models that include a drift term are equal to a trendline that crosses the first and the last price observations of the time series. Generated predictions are closely linked to the last price observation, which is why we expect both random walk models to deliver good performances for small- to medium-sized forecast horizons.

ARIMA: Autoregressive Integrated Moving Average models (ARIMA) depict the inherent dynamics in the price time series and therefore represent a more adaptive group of forecasting approaches. While such models are often described as short-term oriented, this is especially true for types that assume a stationary time series (Box et al., 1994). However, ARIMA models can also be applied to time series originating from an instationary data generation process and include coefficients to incorporate longer-term tendencies in the data. The fitted slope reflects the inherent direction of the price development after accounting for the partial autocorrelation (moving average) and autoregressive structure in the data. To estimate models, a concrete specification has to be determined, for which we revert to a procedure that automatically selects the model order that fits best to the price data (*AutoARIMA*) (Hyndman and Khandakar, 2008). This setting considers stationary as well as instationary model configurations and includes an appropriate number of autoregressive and moving average coefficients to capture the price dynamics. In addition to the fully automated selection procedure, we also employ forecasts that originated from models assuming instationary data generation processes, so that only the autoregressive and moving average orders can be automatically selected (*AutoARIMA inst.*). Due to the weak but present memory structure in the price data, we expect ARIMA models to perform reasonably on short-term forecast horizons. Additionally, due to adaptive drift coefficients, while simultaneously accounting for serial correlation, we also expect robust performance on medium- to long-term horizons.

Exponential Smoothing: While the already discussed methods strictly assume global trend or drift dynamics, exponential smoothing models (EXS) allow for time-varying trends. The inherent modeling idea is that forecasts are adjusted based on the available level, trend and/or error estimates, while the speed of the adjustment is controlled via smoothing constants. Some model configurations allow for additive or multiplicative components in each case with or without damped trends (Holt, 2004) or combine these approaches with additive or multiplicative seasonality indices. An overview of the existing and considered configurations is given by Pegels (1969) and Gardner (2006, 1985). To estimate and forecast using exponential smoothing models, we refer to their state-space representation that also enables automatic model selection

(*Auto EXS*) (Hyndman et al., 2008). However, due to the good forecasting results in forecasting competitions (Makridakis and Hibon, 2000), we include some pure models in addition to the automated model selection approach – most importantly an exponential smoothing model with additive, damped trend (*EXS dHolt*). This allows us to derive conclusions based on the different components considered by the models. Due to the versatility of the models, we expect a reasonable performance throughout all forecast horizons for the automatic procedure as well as for the model that incorporates a damped trend.

Theta: The Theta method is a decomposition approach where the segmented components of a time series are called Theta-lines (Assimakopoulos and Nikolopoulos, 2000). While the original derivation is algebraically complex, Hyndman and Billah (2003) show that the generated forecasts are numerically equivalent to a single exponential smoothing model combined with a drift term. This drift coefficient is equal to half of the slope coefficient when fitting a linear regression model to the data. For this reason, the Theta method is known to generate conservative predictions that show no extreme up- or downward tendencies. We employ an adapted version of the Theta method (*FourTheta*) (Nikolopoulos and Assimakopoulos, 2004) that is more tailored to the data in addition to the original version that we present in combination with a logarithmic transformation of the prices (*logTheta*) (Assimakopoulos and Nikolopoulos, 2000; Box and Cox, 1964). Due to its relation to the exponential smoothing models, we expect reasonable performance for the Theta models for all forecast horizons as well.

Combination Methods are well known to produce reliable results as they have the potential to eradicate shortcomings from the individual methods through calculating the (weighted) average of the derived point forecasts (Clemen, 1989). Obviously, the resulting performance depends on the incorporated method. We combine multiple of the proposed approaches and consider static and equally weighted combinations for the automatic selection procedures of EXS and ARIMA models (*Combi EA*) as well as for a group of exponential smoothing models consisting of the simple exponential smoothing and exponential smoothing with and without damped trend (*Combi SHD*) that was also used in the M4 forecasting competition (M4 Competition, 2018). However, the success of forecast combinations depends on multiple factors including the correlation of the generated forecasts. Expressing general performance expectations – other than the fact that combinations usually show strong performance – is therefore difficult.

3.2 Forecasting Models

To evaluate the forecasting accuracy of the presented methods, we use genuine out-of-sample forecasts. Each of the time series considered in the study contributes 730 observations to the sample. Because a sufficient amount of observations is needed to actually estimate an initial model, we reserve the first 90 time points as an introductory training set, while the remaining 640 observations are allocated to the initial test set. The forecast generation and evaluation scheme is illustrated in figure 2, where the black blocks represent observations that are incorporated in the model estimation, while grey and white circles in the yellow blocks represent the generated forecasts. All forecasts that belong to one forecast horizon are diagonally arranged and marked with the same color. The one-step-ahead forecasts are exemplarily highlighted in dark grey. In total, this cross-validation scheme yields 205,120 evaluable forecasts per method and time series. Given the 2,000 electronic consumer products and 16 considered model

variants, we extract our findings from a basis of 6.56 billion evaluated predictions. Even though our initial model portfolio contained additional methods, the presented 16 methods deliver the most interesting results. However, in case additional comprehensive insights were generated, we briefly highlight special model variants. Note that with increasing forecast horizon, the number of evaluable predictions declines and thus results become more volatile and less informative. The upcoming results chapter, therefore, mainly focuses on short-, medium- and long-term sections that not only contain the most robust results but were specially chosen to cover the majority of practically relevant applications.

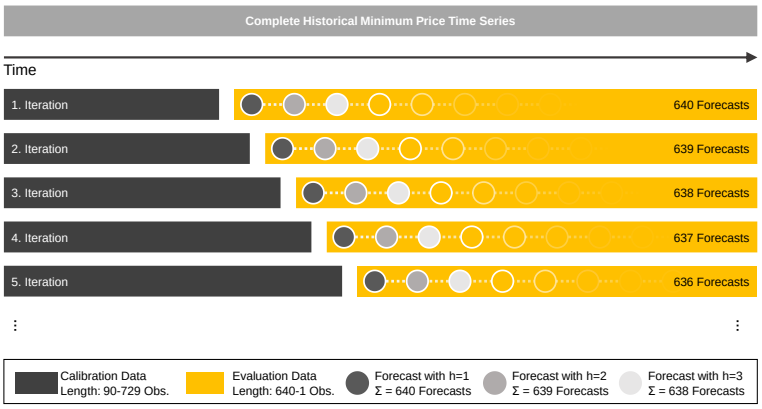

Figure 2. Rolling forecast evaluation scheme.

Because we forecast prices, the generated predictions and their accuracy are related to the level of the respective time series. Thus, the need to make results comparable over all time points, forecast horizons, and products in the dataset arises. For the evaluation, we, therefore, refer to the mean absolute percentage error (MAPE) per product as the main performance indicator. Armstrong (1985) points out that the MAPE is not a symmetric evaluation measure, which is correct when the evaluated data are not on a percentage scale. Resulting from this, an adapted version of the MAPE, the symmetric MAPE (sMAPE) has been developed, which is sometimes favored over the traditional one. We also calculated the sMAPE beside some other evaluation measures and found our results to be consistent with the findings presented below. We focus on presenting our findings for the traditional MAPE, which is not only more widely used but also allows comparability with other studies (Clements, 2005). For each method, the evaluation measure is calculated over all forecasts that have been generated for one horizon (e.g. for $h = 1$ over the grey circles in figure 2). The MAPE for a specific forecast horizon h conditional on a specific product and method is given by

$$MAPE_h = \frac{1}{640 - h + 1} \sum_{k=1}^{640-h+1} \left| \frac{y_{90+h+k-1} - \hat{y}_{90+h+k-1}}{y_{90+h+k-1}} \right|$$

By being an aggregated measure, the MAPE allows gaining an understanding of the overall performance for the specific product and for different forecast horizons. Additional to the MAPE, we also analyze the mean percentage error (MPE) that is calculated in the same way as the MAPE but without using the absolute value of the relative forecasting error. Therefore, positive and negative relative errors cancel out and the MPE gives an impression of whether a method systematically over- or underestimates future price developments.

4. RESULTS

The following chapter presents the results of our conducted empirical study. We show the majority of our findings and present performance measures by forecast horizon as well as by product category for the different forecasting methods. However, for brevity and clarity, we concentrate on displaying the performance of the 16 previously explained methods. In cases where models perform surprisingly well or where configurations, which we expected to significantly improve performance, stay behind expectations, we explicitly include them in the discussion. As all performance differences between methods are highly significant due to the extremely large sample size, the following section does not present evidence and results from statistical tests for mean differences or predictive performance in general. In order to be able to visualize the results, the product-dependent evaluation measures have been condensed by averaging over all products in the sample or over the respective product category. The performance indicators for each method, therefore, represent 1.28 million in total respectively 320,000 forecasts per category for the first horizon. We illustrate our results by grouping the forecast horizons into a short-, medium- and long-term perspective.

4.1 Overall Forecasting Results

Table 2 shows the resulting MAPE and MPE performance measures in percent for the 16 selected methods. While the rows indicate the respective method, the columns refer to the short-, medium- or long-term perspectives as well as the overall view. Of course, the definition of these horizons highly depends on the application context, for our purposes, we decided to have non-connecting time periods spanning two months to make results comparable. We defined short-term as the first 56 days after the forecast origin, corresponding to the first 8 weeks and roughly two months. Medium and long-term describe the time from day 91 to 146 respectively 331 to 386. As expected, the presented MPE and MAPE values increase with growing forecast horizons due to the rising uncertainty. This behavior is consistent throughout all methods. The MPE measures the bias of its associated methods, thus, a negative MPE indicates that the corresponding method produces forecasts that are too high.

On average, almost no bias for short- as well as medium-term horizons is found for the combination methods *Combi EA* and *Combi SHD* as well as *logTheta*. For longer forecast horizons, the difference between the methods becomes increasingly clear. Overall, by far the lowest absolute MPE occurs when using the *logTheta* (2.03%) and *Combi SHD* (-2.23%) model; followed by *EXS* (5.42%) and the exponential trend (-6.07%). The original Theta method also has notable low systematic disturbances with 6.11%. The complete opposite applies for forecasts generated by the *Mean* method, which has the most noticeable bias of all methods for short- and medium-term horizons (17-22%). However, over all horizons, the worst MPE results from the quadratic trend (84.37%) and the random walk with drift (-47.88%).

Table 2. Combined results for different forecast horizons in %

Method	Short-term		Medium-term		Long-term		Overall	
	MAPE	**MPE**	**MAPE**	**MPE**	**MAPE**	**MPE**	**MAPE**	**MPE**
Trend	13.23	-7.61	24.25	-15.03	69.78	-45.57	63.80	-41.33
Trend sq.	11.61	2.45	35.75	8.43	249.63	61.89	311.02	84.37
Trend exp.	9.80	-3.44	16.44	-5.69	39.18	-8.98	36.07	-6.07
Trend 90	7.14	-0.63	18.49	-3.89	62.36	-27.56	57.74	-26.46
Mean	20.77	16.91	26.53	22.00	41.21	35.73	37.62	32.42
Mean 90	8.35	3.96	15.25	8.94	32.73	25.26	28.92	22.14
RW	5.01	1.40	12.17	5.69	28.15	19.77	24.60	17.11
RW Drift	6.31	-2.60	19.46	-12.76	70.69	-53.16	64.18	-47.88
AutoARIMA	6.06	-0.91	17.82	-5.25	58.49	-24.05	51.59	-20.09
AutoARIMA inst.	5.98	-1.54	17.77	-7.63	58.88	-29.32	51.47	-24.81
EXS	5.21	0.77	13.58	2.58	37.58	6.40	33.31	5.42
EXS dHolt	4.99	1.24	11.99	5.25	27.69	18.83	24.21	16.24
FourTheta	4.98	1.18	12.05	4.71	28.13	16.91	24.51	14.65
logTheta	4.93	0.23	11.40	0.52	26.85	1.85	24.12	2.03
Combi EA	5.37	-0.07	14.22	-1.33	42.20	-8.82	37.81	-7.34
Combi SHD	4.99	0.25	11.81	0.36	29.48	-2.06	27.45	-2.23

While table 2 only shows the results for the standard configuration of the fully automated *EXS* selection procedure, changing the optimization criterion for the model selection results in a strong decrease of the MPE especially for medium- to long-term horizons. The default selection criterion of the automated *EXS* is the one-step-ahead mean squared error. Prolonging this to the average of the one- to 14-step-ahead (or 28-step-ahead) mean squared error leads to a decrease of the MPE from 5.42% to 0.98% (-0.11%). Considering the average relative price development presented in figure 1, it is not surprising that methods that are configured to produce a non-linear sequence of forecasts tend to have smaller biases. Comparing the accuracy of the exponential and linear trend makes it obvious that the MPE grows much slower when non-linear forecasts are produced, which bend in the same fashion as the price deterioration. This leads to an overall offset of only -6.07% instead of -41.33%. With few exceptions, most of our considered methods have a notable bias, which seems to be rooted in the special statistical properties of the data. While a bias is not a favorable property, the direction of it, however, is important when designing decision recommendation services. Albeit the specific risk functions of the end-users, that the prediction is presented to, are unknown, it is unlikely that a customer will weight predictions, which are too high or too low, equally. Therefore, there might be an implicit preference for the sign of the bias depending on the application. Conditional on the context, it may be better to implement a methodology that consistently ‘promises’ lower prices than actually occur or vice versa. Interestingly, no sign change of the MPE can be observed over the horizons. This means that the sign of the method-specific bias for larger horizons can already be identified in cases where only short-term forecasts can be generated or are evaluable.

When interpreting the MAPE, it is noticeable that the short- and medium-term perspectives are quite comparable. Here, random walk, *AutoARIMA*, and *EXS* perform expectedly good, however, the best performance is noted by the *logTheta*, *FourTheta*, *EXS dHolt*, and *Combi SHD* methods. The worst performances expectedly come from models that are not tied to the last observations of the series, so *Mean*, as well as the linear trend, fail to deliver convincing results. Both methods cannot recover in the long run so that their extrapolations also do not align with the long-term price deterioration (long-term MAPE between 41-70%). Following the same principle, good methods continue to deliver good results. So, the best long term MAPE is generated by *logTheta* (26.85%), directly followed by the Holt model with a damped trend and the random walk method. Exceptions from these findings are the two ARIMA models. While the distance to the best models in the short and medium perspective was rather narrow, the results strongly deviate in the long run. Besides, it is noticeable that *Combi SHD*, which performed nicely for short- and medium-term horizons, is later outpaced by the *logTheta* and random walk methods. Finally, it can be concluded that the most consistent and continuously stable results are generated by *logTheta* (24.12%) and Holt exponential smoothing method with a damped trend (24.21%) and that the quadratic trend (311.02%) produces by far the worst results.

While the presented nominal evaluation measures are interesting and already provide a good overview of the performance for the model classes, some properties and results only emerge when explicitly visualizing the performance individually per horizon. Figure 3, therefore, illustrates relative accuracy measures that result from comparing each individual method with the global linear trend and thus normalizes the performance of the linear trend to unity. Forecast accuracy values lower than one, therefore, present a relative reduction in MAPE compared to the benchmark and thus indicate better performance than the trend. While each accuracy measure is capped by the perfect forecast and cannot be smaller than zero, all obtainable values have no upper boundary. However, values higher than one indicate that the respective method performs worse than the benchmark.

Figure 3 illustrates two findings clearly: First, the MAPE difference between *Trend* and most methods rapidly decreases in the short-term perspective but stabilizes in the long run with a constant gap of 20% to 60% depending on the method. Second, there are only a few crossing points between models, meaning intersections, where it is favorable to switch from one method to another. In most cases, the performance order of the methods stays constant, showing that a superior method for short horizons, often also performs well in the long-term perspective.

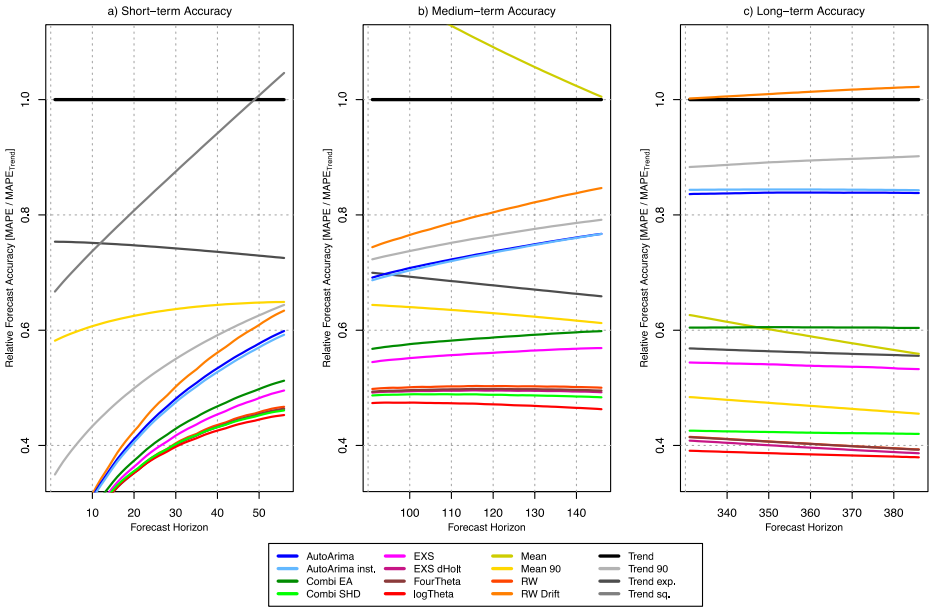

Figure 3. Forecast accuracy relative to the linear trend

Figure 3 shows the relative forecast accuracy over the already discussed short-, medium- and long-term forecast intervals for all methods included in table 2. While for the short-term horizon the global mean (above the trend, outside the plot) is, without a doubt, the worst method to choose. The horizon-dependent forecasting accuracy of *Mean* aligns with the accuracy of the linear trend in the medium-term interval and surpasses it in the long term. Contrary to this, the quadratic trend shows the inverse behavior outperforming the linear trend in the short-term and worsens for larger horizons. However, in the long run, almost all forecasting methods deliver results superior to the ones of the linear trend. Furthermore, figure 3 reveals that the short-term forecast horizons exhibit especially high heterogeneity between the forecasting methods. The best performing approach is *logTheta* that scores an improvement of 62.74% compared to the linear trend. Thereby, estimating the Theta method on a logarithmically transformed price series improves performance by 0.3% to 1.8% relative to the accuracy of the Theta model estimated on the nominal price values. Additionally, *EXS dHolt* performs well throughout the shown 56 forecast horizons. However, most of the time it is on par with the *FourTheta*, *Combi SHD* and *RW* methods that improve the accuracy of the linear trend by over 61%. While the random walk with drift and the ARIMA approaches deliver similar results for very short forecast horizons, ARIMA produces superior results when horizons grow. *Combi EA* and *EXS* produce good results for shorter horizons, considerably better than those of *Mean 90*. Surprisingly, in the long-term perspective, *Mean 90* significantly outperforms both, more complex methods and ends up as the sixth-best method in this interval. With this behavior, *Mean 90* is one of the few methods improving its placement by six places in the accuracy ranking. Methods that considerably lose places over time are *AutoArima inst.* (from 8th to 12th place) and *RW Drift* (from 10th to 15th place) with the latter one performing worse than the benchmark in the long run.

Interesting insights can also be found, when comparing models within their class as can be seen in figure 4. Trend models (4a) show very heterogeneous results, especially for forecast horizons bigger than 100. While the exponential trend performs expectedly well as it shows good adaptation to the price-wise life cycle and the price deterioration, the linear and quadratic trend are again not recommendable for any horizon and thus also fail in the overall view. Only in the short-term perspective is the localized trend with 90 observations superior to the exponential trend but is quickly overtaken by *Trend exp.* for medium and large forecast horizons.

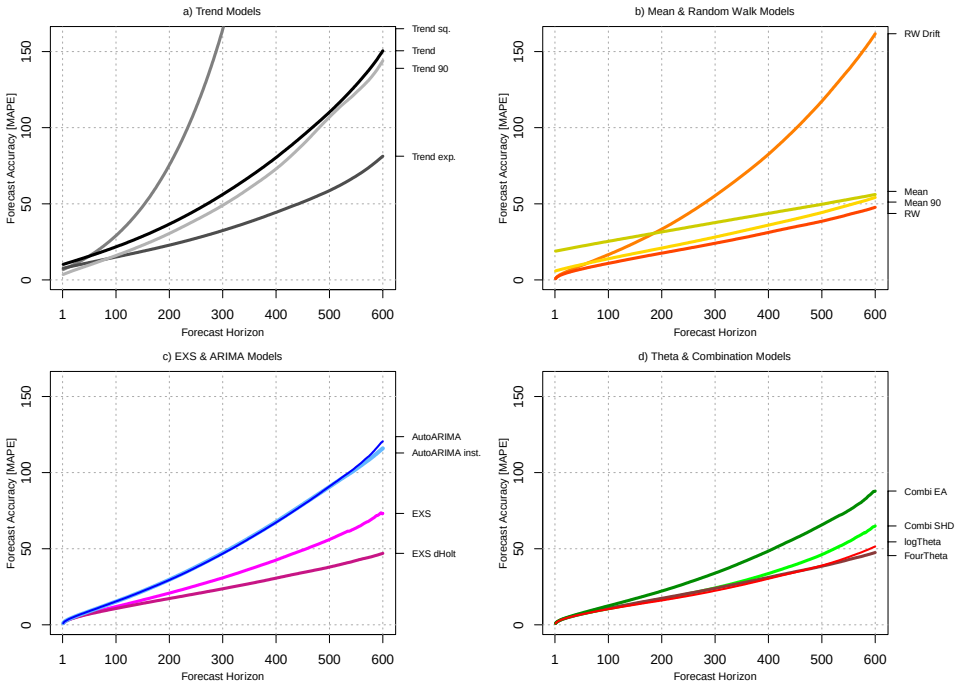

Figure 4. Result comparisons within model classes.

Figure 4b shows that even though *Mean 90* is at no point and for no forecast horizon the best performing method, it performs stable and comparably good throughout the life cycle, thus outperforming its other class member. For random walk models, there is only a small difference between the *RW* and *RW Drift* in the short run. However, this quickly changes with increasing forecast horizons leading to highly deviating results – the random walk with drift being one of the worst methods, while the normal random walk shows suitable performance for the price data set. This is partially surprising because we only expected the good performance on short horizons. However, the satisfying results, especially on longer extrapolations, may be explainable through the varying zero-inflation in the sample.

As described before, ARIMA models (4c) perform reasonably on rather short horizons in general. For the short and medium perspective, the model group that treats each time series as strictly instationary delivers slightly better performance than the fully automatic procedure that allows for stationary models. However, as figure 4c shows both *EXS* models outperform the

ARIMA class. The very good performance of the automatic exponential smoothing procedure, especially for short and medium horizons, can be easily overlooked when compared to configuration options within its model class. This is because the Holt exponential smoothing method with a damped trend delivers the second-best results in the analysis, significantly outperforming almost all other methods. As table 2 shows *FourTheta* is inferior to the logarithmized configuration of the original Theta model. However, figure 4d indicates that with increasing horizons this difference vanishes and at $h = 600$ *FourTheta* delivers the best results in its class. Finally, *Combi SHD* consistently outperforms the combination of the automatic selection procedures of EXS and ARIMA.

4.2 Category-Specific Results

When discussing category-specific results two perspectives need to be taken into account. First, it is important to analyze which method delivers the best results for a given category. Second, it must be noted that the performance of each model differs by category, leading to heterogeneous results and varying magnitudes of evaluation measures. Thus, one method might perform well for one category but fails for another. For the second perspective, it is necessary to calculate the MAPE relative to the best model for each category. Thereby, it becomes visible can see that particularly random walk and *EXS dHolt* models have a good relative performance for all categories except for smartphones, where the MAPE is in both cases over 40% higher compared to the best performing method. Besides, we could find that especially mean and trend models perform rather heterogeneously between the categories.

Considering the first perspective, for short and medium horizons of refrigerators – from the 16 described methods – the *FourTheta* method delivers the best results directly followed by the random walk. Interestingly, the method with the lowest error over all horizons is the single exponential smoothing model, which is the simplest form of exponential smoothing including only a weighted moving average and no components for trend or seasonality. For graphic cards, the pure random walk model works best for all 600 horizons. However, the *RW* is narrowly outperformed by the *FourTheta* method in the medium-term. The *Combi SHD* model delivers very good results for the smartphone product category. But pre- and postprocessing of the data in terms of applying the methods to logarithmic prices improves results significantly so that the *logTheta* method consistently outperforms all other methods independent of the forecast horizon. For the product category that is hardest to forecast – computer memory - the Holt exponential smoothing model with a damped trend works best overall, resulting mainly from the good performance for larger forecast horizons. In the short- and medium-term perspective, the *Combi SHD* model is superior to *EXS dHolt*. Across all product categories, the worst performing methods are the quadratic trend followed by the random walk with drift. For computer memories, those models are complemented by the linear trend, which also delivers consistently poor results and should be avoided for practical applications.

5. CONCLUSION

In this paper, we empirically evaluated the forecasting performance of 16 methods for the usage of customer- and business partner-centric applications of price prediction services based on a large sample of product price time series from the German consumer electronic goods e-commerce market.

Overall, we showed that univariate time series forecasting methods deliver reasonable performance over short- and medium-term horizons and that these methods are a viable option when planning and designing services, that support customers while buying consumer goods. Our results recommend different methods depending on the application context and provide a comprehensive reference for price comparison sites and developers of prescriptive analytics services in the area of digital commerce. Due to the large sample size and the variability of products and forecast horizons, the performance of particular forecasting configurations can be derived easily from our analysis.

When designing recommendation services that should be applicable to all products and decision horizons while delivering baseline performance, it is advisable to refer to the Theta method or to the closely related exponential smoothing models. Both deliver solid performance throughout all categories and most horizons. However, an ideal one-fits-all solution does not exist. This is especially true when giving long-term price-wise recommendations for household goods. Here, due to the comparably low price movement – even over very large horizons – service engines should be rather based on random walk principles. Additionally, it often seems to be advisable to transform the underlying data before generating predictions. Transformations like the exemplarily discussed logarithmic one, do not guarantee better performance but sometimes improve results and rarely lead to significantly worse outcomes. The more complex approaches in our study could not fully leverage the specific data characteristics, which indicates that custom-designed approaches should be based on robust and rather simple methods and expand them to exploit price time series characteristics like zero-inflation or asymmetric price changes. As most modern machine learning and artificial intelligence methods show generally poor performance in time series contexts, we recommend exploring hybrid methods, where modern methods parameterize simpler methods like ARIMA or exponential smoothing models, instead of using them directly for generating price forecasts.

While the presented results in this paper are promising and show which methods can be used for what business purpose directly, the findings raise the need for future research. For practical applications, it may be especially interesting to develop a selection approach that automatically suggests a suitable core methodology based on product or historic price movement characteristics. From what we learned in this study, we believe that it should be possible to anticipate the forecasting performance for individual items or at least groups of products, by modeling the extent of which data features contribute to the forecast accuracy. This is also a promising application for modern statistical learning algorithms and the illustrated statistics can hereby serve as a starting point. Additionally, the stage in the product life cycle, general market tendencies or multivariate knowledge are possible enhancements that can further improve performance and should be investigated.

This study shows that there is no need to further restrain from using univariate time series forecasting methods for price forecasting applications in the context of price comparison sites. It became clear, that it is necessary to incorporate the recommended methods into platforms and services so that customers and business partners of PCS can actually benefit from the generated insights and help to make more informed, less costly and more efficient buying and business decisions.

REFERENCES

Armstrong, J.S., 1985. *Long-Range Forecasting: From Crystal Ball to Computer.*, 2nd ed. John Wiley & Sons, New York, NY. https://doi.org/10.5465/AMR.1979.4289149

Assimakopoulos, V., Nikolopoulos, K., 2000. The theta model: A decomposition approach to forecasting. *Int. J. Forecast.* 16, 521–530. https://doi.org/10.1016/S0169-2070(00)00066-2

Bolton, R.N. et al., 2006. Recent Trends and Emerging Practices in Retailer Pricing, in: Krafft, M., Mantrala, M.K. (Eds.), Retailing in the 21st Century - Current and Future Trends. Springer, Berlin, Germany, pp. 255–269. https://doi.org/10.1007/978-3-540-72003-4

Box, G.E.P. et al., 1994. *Time Series Analysis: Forecasting & Control*. Wiley, Hoboken, NJ.

Box, G.E.P., Cox, D., 1964. An analysis of transformations. *J. R. Stat. Soc. - Ser. B* 26, 211–252.

Buchwitz, B., Küsters, U., 2018. Should I buy my new iPhone now ? Predictive Event Forecasting for Zero-Inflated Consumer Goods Prices. *Proc. 39th Int. Conf. Inf. Syst.* 1–16.

Clemen, R.T., 1989. Combining forecasts: A review and annotated bibliography. *Int. J. Forecast.* 5, 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Clements, M.P., 2005. *Evaluating Econometric Forecasts of Economic and Financial Variables*. Palgrave Macmillan, New York, NY. https://doi.org/10.1057/9780230596146

Gardner, E.S., 2006. Exponential Smoothing: The State of the Art - Part II. *Int. J. Forecast.* 22, 637–666. https://doi.org/10.1016/j.ijforecast.2006.03.005

Gardner, E.S., 1985. Exponential Smoothing: The State of the Art. *J. Forecast.* 4, 1–28.

Holt, C.C., 2004. Forecasting seasonals and trends by exponentially weighted moving averages. *Int. J. Forecast.* 20, 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hyndman, R.J. et al., 2008. *Forecasting with Exponential Smoothing - The State Space Approach*. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-540-71918-2

Hyndman, R.J., Athanasopoulos, G., 2018. *Forecasting: Principles and Practice*, 2nd ed. OTexts.

Hyndman, R.J., Billah, B., 2003. Unmasking the Theta method. *Int. J. Forecast.* 19, 287–290. https://doi.org/10.1016/S0169-2070(01)00143-1

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: The forecast package for R. *J. Stat. Softw.* 27, 1–22. https://doi.org/10.18637/jss.v027.i03

idealo, 2017. Bestseller und ihr Preisverfall - Wann klingelt der Preiswecker [Bestsellers and their Price Deterioration - When does the Price Alarm Ring].

Kannan, P.K. et al., 2001. Dynamic Pricing on the Internet : Importance and Implications for Consumer Behavior. *Int. J. Electron. Commer.* 5, 63–83.

Kömm, H., Küsters, U., 2015. Forecasting zero-inflated price changes with a Markov switching mixture model for autoregressive and heteroscedastic time series. *Int. J. Forecast.* 31, 598–608. https://doi.org/10.1016/j.ijforecast.2014.10.008

Kopalle, P. et al., 2009. Retailer Pricing and Competitive Effects. *J. Retail.* 85, 56–70. https://doi.org/10.1016/j.jretai.2008.11.005

Levy, M. et al., 2004. Emerging trends in retail pricing practice: Implications for research. *J. Retail.* 80, XIII–XXI. https://doi.org/10.1016/j.jretai.2004.08.003

M4 Competition, 2018. Competitor's Guide.

Makridakis, S. et al., 2018. The M4 Competition: Results, findings, conclusion and way forward. *Int. J. Forecast.* 34, 802–808. https://doi.org/10.1016/j.ijforecast.2018.06.001

Makridakis, S. et al., 1997. *Forecasting Methods and Applications*, 3rd ed. Wiley, Hoboken, NJ.

Makridakis, S., Hibon, M., 2000. The M3-Competition : results, conclusions and implications. *Int. J. Forecast.* 16, 451–476.

Nikolopoulos, K., Assimakopoulos, V., 2004. Generalizing the Theta model for automatic forecasting. *Int. Symp. Forecast. ISF 2004, Sydney, Aust.*

Pan, X. et al., 2002. Why Aren't the Prices of the Same Item the Same at Me.Com and You.Com?: Drivers of Price Dispersion Among E-Tailers. *SSRN 328820*. https://doi.org/10.2139/ssrn.328820

Pegels, C.C., 1969. Exponential Forecasting : Some New Variations. *Manage. Sci.* 15, 311–315.

Rydberg, T.H., Shephard, N., 2003. Dynamics of Trade-by-Trade Price Movements: Decomposition and Models. *J. Financ. Econom.* 1, 2–25. https://doi.org/10.1093/jjfinec/nbg002

Smyl, S. et al., 2018. M4 Forecasting Competition: Introducing a New Hybrid ES-RNN Model [WWW Document]. *Uber Eng.* URL https://eng.uber.com/m4-forecasting-competition/ (accessed 4.21.19).

Sucarrat, G., Grønneberg, S., 2016. Models of Financial Return with Time-Varying Zero Probability. *MPRA Work. Pap.* 1–25.

Winkelmann, R., 2008. *Econometric Analysis of Count Data*, 5th ed. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-540-78389-3