Claudio Hartmann, Martin Hahmann, Wolfgang Lehner, Frank Rosenthal

Exploiting big data in time series forecasting: A cross-sectional approach

Erstveröffentlichung in / First published in:

2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris, 19.-21.10.2015. IEEE, S. 1-10. ISBN 978-1-4673-8272-4.

DOI: http://dx.doi.org/10.1109/DSAA.2015.7344786

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786

Exploiting Big Data in Time Series Forecasting: A Cross-Sectional Approach

Claudio Hartmann, Martin Hahmann,
Wolfgang Lehner
Technische Universität Dresden
Database Technology Group
01062 Dresden, Germany
Email: <firstname.lastname>@tu-dresden.de

Frank Rosenthal
GfK SE
Marketing & Data Sciences
Nordwestring 101
90419 Nürnberg, Germany
Email: Frank.Rosenthal@gfk.com

***Abstract*—Forecasting time series data is an integral component for management, planning and decision making. Following the Big Data trend, large amounts of time series data are available from many heterogeneous data sources in more and more applications domains. The highly dynamic and often fluctuating character of these domains in combination with the logistic problems of collecting such data from a variety of sources, imposes new challenges to forecasting. Traditional approaches heavily rely on extensive and complete historical data to build time series models and are thus no longer applicable if time series are short or, even more important, intermittent. In addition, large numbers of time series have to be forecasted on different aggregation levels with preferably low latency, while forecast accuracy should remain high. This is almost impossible, when keeping the traditional focus on creating one forecast model for each individual time series. In this paper we tackle these challenges by presenting a novel forecasting approach called cross-sectional forecasting. This method is especially designed for Big Data sets with a multitude of time series. Our approach breaks with existing concepts by creating only one model for a whole set of time series and requiring only a fraction of the available data to provide accurate forecasts. By utilizing available data from all time series of a data set, missing values can be compensated and accurate forecasting results can be calculated quickly on arbitrary aggregation levels.**

I. Introduction

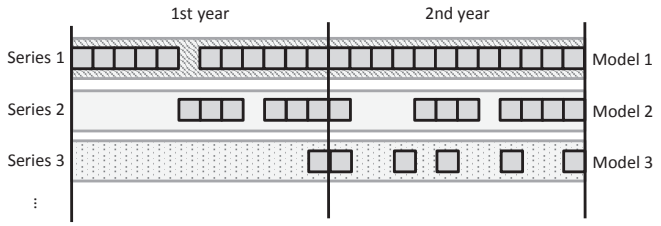
Nowadays, forecasting of time series data has become an irreplaceable tool for management, planning and decision making. Without a solid idea on the future development of certain measures or parameters it is impossible to efficiently manage inventories, organize production processes, and allocate resources in advance. Obviously, the foundation of good forecasts is the availability of sufficient data. Therefore, the ongoing Big Data trend looks like a natural contributor to forecasting, allowing more accurate forecasts and widespread application. While this might be true in some cases, Big Data also introduces new challenges. Naturally, time series have increased in volume as their granularity gets finer and recorded histories become longer. More importantly, the ongoing focus on data gathering has led to a strong increase of monitored data sources. This includes the number of types of data sources as well as the number of instances per type that are monitored. This variety of heterogeneous data sources causes two major problems regarding data availability. The first problem are missing values. While these have always been a problem due to technical factors, e.g., reliability of certain sensors, Big Data also introduces logistic missing values. As data sources can be distributed over different locations, administrated by different organizations/authorities or use different formats, errors or delays during data delivery cause missing values. The second problem is that data is often available in such large volumes that timely creation of a forecast cannot be guaranteed. In the following we use two examples to illustrate some of these problems.

Example 1: We analyzed several data sets from the sales domain and found that there are always items with stable sales over time and other items showing very strong fluctuations in their sales behavior. This leads to irregular time series which are very hard to forecast due to a lack of reproducible behavior. Furthermore, these fluctuations can lead to the temporary exclusion of individual items that are neither sold nor held in stock, and thus, are not included into an outlet's periodic data delivery. This results in missing values and causes incomplete time series histories, where properties that are essential for the creation of many forecast models, e.g., trend or seasonality, are not recognizable any more.

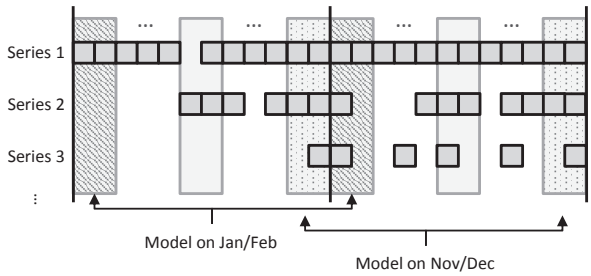
Example 2: Smart Meter Data is collected for electricity, gas, water and heating. While smart meters were originally installed for major consumers, they become more and more common in private households due to several smart grid initiatives [1], [2], [3]. In this domain, missing values occur due to different reasons than in the sales domain. We investigated smart meter data for the energy consumption of households in Ireland, that was recorded on a 30 minute granularity which leads to high-volume time series. In this scenario, missing values occur when the Smart Meter temporarily has no internet connection or it shows any other kind of technical malfunction.

These two examples illustrate the current trend in Big Data of extensive data collection with thousands of time series gathered from a multitude of heterogeneous data sources. Furthermore, they reveal some of the different reasons for missing values in time series data. These aspects are challenging for the forecasting process and make it hard or even impossible

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786



(a) Traditional time series forecasting.



(b) Cross-sectional time series forecasting.

Fig. 1. Forecasting approaches overview.

to apply traditional time series forecasting techniques. Let us consider the panel data example and assume we want to forecast the sales of individual items showing different degrees of sparseness (Figure 1a). Traditional techniques, such as Exponential Smoothing or ARIMA models [4], capture the historical behavior of a time series and model the dependency of future on past time series values. Thus, a forecast model is created for each time series, i.e., each individual item in this example. However, such forecast models require complete historical data without any missing values. For example, to learn typical seasonal patterns on data recorded with a monthly granularity, as often present in sales data, the widely used Triple Exponential Smoothing model [5], requires at least 17 historical time series values [6]. Moreover, to learn meaningful model parameters and to compensate for randomness, much more historical data should be available and used for model training [6].

Of course there already exist several approaches to deal with missing values. One possible approach to overcome this issue is to exploit the hierarchical structure of a data set, if there is one [7]. An aggregation function, e.g., sum, is applied on the intermittent time series to create a higher aggregation level like outlet or city on which gaps are filled due to summarization. In many domains predictions on higher aggregation levels are desired anyway, market researchers for example are interested in the market share of a brands whole portfolio, while energy providers require the total energy production and consumption in their distribution areas to balance prices on the energy trade market. However convenient this approach might seem, it does not solve the problems of traditional forecasting approaches entirely. Considering all dimensional attributes and their combinations leads to an exponential number of possible aggregation levels. Calculation of forecast values for all these aggregations with traditional forecast models is almost impossible. Furthermore, even at higher aggregation levels, it is still possible to suffer from missing values. On top of that, utilization of coarser aggregation levels prohibits the computation of forecasts on lower aggregation levels.

Another approach to dealing with missing values is filling up the gaps using interpolation. Several techniques for such calculations already exist [8], [9], [10], but the application of this approaches are domain-specific, i.e., if a non-fitting interpolation method is chosen, the calculated values will introduce errors and decrease the accuracy of the forecast. In addition certain application scenarios do not allow a clear identification of missing values. In our sales data example a gap in a time series can either mean that a measured value is actually missing or that an item was neither sold nor stocked and thus no value was reported. Obviously, in the first case the gap should be filled while in the second case it should be left unchanged. If a clear differentiation of these cases is impossible, filling up with calculated values always yields the risk of error introduction.

In conclusion, we can state that traditional forecasting approaches generally assume one model per time series as well as historical data that is complete and as extensive as possible. Big Data offers large numbers of time series, which have higher volumes of data but no higher degree of completeness. Thus, forecasting becomes more costly due to a large number of models etc., but not necessarily more accurate due to missing values. In this paper we present an approach that is able to tackle these challenges. For this, we break with traditional forecasting in two major points: First, we drop the fixation on large histories and focus on data-economical forecasting, taking only as much data as necessary in order to create accurate forecasts instead of whole histories. Second, we abandon the concept of one model per time series and focus on modeling whole sets of time series. In doing so, we exploit the fact that Big Data offers large numbers of time series that originate from the same domain. Our approach is called *cross-sectional forecasting*, whereas the term cross-sectional refers to a set of related time series, observed at the same point in time [11].

Figure 1b shows a rough sketch of our concept. Instead of training a model for every individual time series, we compute one model over a large set of time series. Thus, our approach is able to incorporate available information from all time series of the data set. This makes it resistant to missing values and also robust against outliers and high randomness of individual time series. In addition, only small time slices are used for the model creation instead of the whole available historical data. As we only pick the necessary parts out of the available historical data, we do not rely on long and complete time series histories, which makes our model creation much faster and easier to apply.

The remainder of this paper is structured as follows: We begin with a description of the data set characteristics our approach addresses and detail the current state of the art of traditional forecasting techniques in Section II. We describe our cross-sectional forecasting approach as well as possible

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786

parameters that may be used for its optimization in Section III. In Section IV we extensively evaluate our approach on two real world data sets from different domains. We finally conclude with a summary of our contributions and open topics we plan to address in the future in Section V.

II. State of the Art in Time Series Forecasting

In this section we shortly describe the structure and characteristics of the data sets used in this work. Afterwards, we give a brief overview of commonly used existing forecasting techniques [5] and highlight their drawbacks with respect to the challenges formulated earlier.

A. Data Characteristics

The data sets of our example scenarios consist of several time series $x_1, \ldots, x_n$ which are monitored over the same period of time $1, \ldots, t$. An arbitrary portion of the time series lacks a sufficient history of data points, as shown in Figure 1. Many time series are very short and sparse; especially for the example from the sales domain there are just a few items which are sold in every period. Specifically, only 8.4% of all item-level time series have a complete history for all 36 months, making them the only time series where traditional forecasting approaches can be applied without any adjustments. Additionally, 5.3% of all items are reported in one single month only, and thus, constitute unpredictable "surprise items".

B. Existing Forecasting Techniques

According to a survey of McCarthy et al. [5] judgmental forecasting, where domain experts derive predictions with the help of statistical tools, is still widely used. For Big Data sets with a high number of time series as described in the previous section this is not feasible anymore. Even in the forecasting of aggregates where the number of time series may be reduced due to aggregation there are many different influences that have to be analyzed to obtain the optimal forecast result. Therefore, we will review modeling approaches which allow the automated forecast of a high number of time series.

Univariate Modeling Techniques: The most commonly used model based forecasting techniques are ARIMA and smoothing approaches such as HoltWinters' Triple Exponential Smoothing [4]. Both model types count to the group of univariate statistical models, which means they only take one variable into account: the modeled time series itself. This type of models describes the dependency of future time series values on historical values and often follows the subsequent general equation:

$$\hat{x}_{n,t+1} = \alpha_t \cdot x_{n,t} + \cdots + \alpha_{t-m} \cdot x_{n,t-m}. \qquad (1)$$

Hereby, $x_{n,t}$ denotes the time series value of series n in period t with $t \geq 0$ and $\hat{x}_{n,t+1}$ the forecast value of the future period $t+1$. Hence, the future time series values at time $t+1$ result from a linear combination of previous time series values at time $\leq t$, where $m+1$ corresponds to the number of past values which are included in the model. The parameters $\alpha_t, \ldots, \alpha_{t-m}$ specify the contribution of each past time series value to the forecast. For example, exponential smoothing includes all previous values ($m = t$) with exponentially decaying weights. Auto-regressive models AR(p) require the explicit specification of the number of past values ($m = p-1$). Both approaches can be extended to cope with trend and seasonal patterns in time series data.

However, as discussed in the introduction, such traditional forecasting techniques require sufficient historical, as well as complete time series data to initialize and optimize the model parameters $\alpha_t, \ldots, \alpha_{t-m}$. Unfortunately, these are prerequisites which are often not met by Big Data sets which consist of information from a variety of heterogeneous data sources.

Multivariate Modeling Techniques: In contrast to univariate forecasting models multivariate techniques take several variables into account. One of the most popular of these techniques is Vector Autoregression (VAR) [12]. This approach creates one single model to predict the future values for several time series. The model is designed such that all time series affect the predictions of each other and thus, incorporates knowledge of many time series into the training process. The influencing dependencies can be modeled using one or more periods of historical data. The following equation shows for the example of a VAR(1) model, which takes only one historical value into account, how the calculation of the forecast values is realized:

$$\begin{pmatrix} \hat{x}_{1,t+1} \\ \vdots \\ \hat{x}_{n,t+1} \end{pmatrix} = \begin{pmatrix} \alpha_{1,1} & \cdots & \alpha_{n,1} \\ \vdots & \ddots & \vdots \\ \alpha_{1,n} & \cdots & \alpha_{n,n} \end{pmatrix} \cdot \begin{pmatrix} x_{1,t} \\ \vdots \\ x_{n,t} \end{pmatrix}. \qquad (2)$$

$\hat{x}_{1,t+1}, \ldots, \hat{x}_{n,t+1}$ are the forecasted values for period $t+1$ of the time series $x_1, \ldots, x_n$. $x_{1,t}, \ldots, x_{n,t}$ are the corresponding time series values in period t. The parameters $\alpha_{1,1}, \ldots, \alpha_{n,n}$ describe the influence of the time series values to the forecast values. In particular there is one set of parameters for predicted time series, $\alpha_{1,n}, \ldots, \alpha_{n,n}$, which models the influence of all available time series values $x_{1,t}, \ldots, x_{n,t}$ in period t to the forecast $\hat{x}_{n,t+1}$ of time series n. This leads to a set of linear equations where the parameters have to be optimized to correctly model the influence of all time series to each other. With an increasing number of time series this optimization process needs an increasing number of historical data. Picking up the example of the sales of consumer electronics devices we have to handle thousands of product time series in one data set. In a specific example we may have to predict the future sales values for 100 brands. In a monthly time resolution reliable data for 8 years and 4 month would be necessary to even calculate the start parameters for this VAR(1) model without the opportunity of optimization. On a more fine grained aggregation level this problem gets even worse. Predicting sales values for 5000 electronic devices, e.g., televisions, would require available data back to the year 1599 which is clearly not possible. In addition to this, VAR still requires a complete data history without missing values, otherwise the influences between the time series can not be modeled properly.

Modeling Techniques for Incomplete Time Series: The area of forecasting intermittent time series has already been

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786

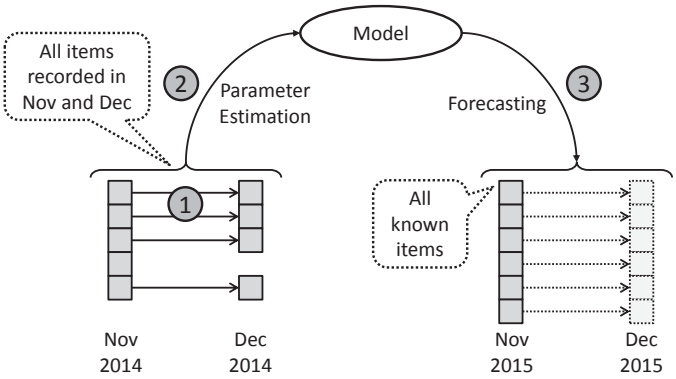


Fig. 2. Cross-sectional forecasting process.

discussed in the forecasting literature. Croston's method is a forecasting approach especially designed for time series with intermittent occurrence [13], [14]. This method models the time series using two smoothing processes:

$$Q^*_{n,j} = (1-\alpha)Q^*_{n,j-1} + \alpha Q_{n,j}, \tag{3}$$

$$x^*_{n,j} = (1-\alpha)x^*_{n,j-1} + \alpha x_{n,j}. \tag{4}$$

For a time series n, $Q^*_{n,j}$ denotes the smoothed interval from the last monitored non-zero time series value $x_{n,j-1}$ in period $j-1$ to the next non-zero value in period j. $Q_{n,j}$ is the not smoothed distance between $x_{n,j-1}$ and $x_{n,j}$. The value for j is not fixed, it denotes an arbitrary period with a non-zero time series value. $j-1$ is not the direct predecessor period to j, but the last non-zero occurrence of a time series value before period j. $x^*_{n,j}$ and $x_{n,j}$ denote the corresponding smoothed respectively not smoothed values of time series n. α is the smoothing parameter which is used to smooth the non-zero time series values as well as the intervals between them. The prediction for the next non-zero time series value is then determined by $x^*_{n,j}$ which is used as the forecasted value that has a distance of $Q^*_{n,t}$ to the last non-zero time series value $x_{n,t}$.

Actually, Croston's method is also an univarite modeling technique, since it only uses information derived from the time series itself. Nevertheless, it is listed in a separate category, since, to our knowledge, it is the only forecasting model which is capable of dealing with missing values in the time series history, which is a very important point in our application scenario. However, Croston's method may be able to calculate forecast values for time series with incomplete histories. But for complete time series this technique reduces to single exponential smoothing, which hardly ever is an appropriate choice to model time series data. Thus, this approach is not entirely suited to predict the data sets in the focus of our work, because we aim for accurate forecast values for a mix of *intermittent and complete* time series.

III. Cross-Sectional Forecasting

This section describes our new forecasting approach which is capable of working with incomplete time series histories. It incorporates knowledge from many time series into the creation of one single model, that is used to predict the future values of all time series at once. Our approach follows three simple core steps, which are shown in Figure 2 and are further detailed in the following paragraphs. In general, cross-sectional forecasting assumes that relative transitions from one period to the next remain stable over several seasons. For example, due to Christmas sales, there is an increase in the sales from November to December in every year for many product groups, which is almost stable and therefore, can support the creation of a robust forecast model.

A. Model Creation

In the first step all necessary information for the model creation process are extracted from the historical data. Our model derives its knowledge from all available time series of a data set. As these have different value ranges it is necessary to make them comparable. To achieve this, we do not take the time series values itself into account, but the relative transitions within the series, i.e., the change from one period to the next. Coming from the same data set, these are very likely to be similar, e.g., sales of consumer electronics generally increase before Christmas. We obtain these transitions by extracting two-element sub-sequences from the time series that contains the values recorded for two periods t and $t-1$ (e.g., November and December 2014) – ① in Figure 2. We translate this dependency into a $model\ function\ f$, that describes the relation of time series values in period t based on the previous period $t-1$. Moreover, we allow the inclusion of additional features besides the target measure:

$$x_{i,t} = f(x_{i,t-1}, e^1_{i,t-1}, \ldots, e^k_{i,t-1}), \tag{5}$$

$x_{i,t}$ denotes the the target measure value of time series i at time t which has to be forecasted. $x_{i,t-1}$ denotes the previous target measure value at time $t-1$, $e^1_{i,t-1}, \ldots, e^k_{i,t-1}$ are possible external influences which may be taken into account. For example, to forecast sales units x, we might also include stock units (e^1) and ordered units (e^2) at period $t-1$ of the corresponding item n. The relevant features to be used as external influences depend on the particular use case and can be determined manually or using standard feature selection techniques (e.g., correlation-based measures) from the literature [15].

B. Model Optimization

In the second step, we estimate the parameters of the model function over a *large group* of time series – ② in Figure 2. This approach is based on the assumption that related time series (e.g., items within the same product group) show similar behavior over time. By exploiting this, our approach becomes not only resistant to missing values, but also to outliers of individual time series.

For a more detailed description of the parameter estimation process, we express the model function (Equation 5) as a simple linear regression model. There might exist use cases that require more complex models to express the effect of

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786

the external influences on the target measure, e.g., having a multiplicative dependency instead of a linear one. However, as none of the data sets we used for evaluation shows the necessity of using such a more complex model we used the following linear regression model throughout this paper:

$$\vec{X}_t = \alpha_1 \cdot \vec{X}_{t-1} + \alpha_2 \cdot \vec{E}^1_{t-1} + \cdots + \alpha_k \cdot \vec{E}^k_{t-1} + \alpha_{k+1}. \quad (6)$$

The vector $\vec{X}_t$ denotes the target forecast measure, e.g., sales units, of all time series at time t, and $\vec{X}_{t-1}$ the corresponding measure values at the previous point in time $t-1$. The vectors $\vec{E}^1_{t-1}$ to $\vec{E}^k_{t-1}$ correspond to additional features at time $t-1$ that are included as external influences in the model. Sticking to our example we would model the influence of the stock and order units to the sales units we want to predict as $\vec{E}^1$ and $\vec{E}^2$. Other use cases may require another set of external influences ranging from a pure autoregressive model where only the target measure is available to dozens of influences, e.g., for solar energy supply forecasting. Furthermore, each feature vector only includes those time series of a data set that were actually recorded in the considered periods $t-1$ and t *and* hold a value for every feature used for modeling. The parameters $\alpha_1, \ldots, \alpha_k$ specify the influence of each feature vector and α_{k+1} is an unobserved error term. Thus, we apply a cross section over all items for the parameter estimation and train exactly one parameter value per feature. As a consequence, our model only requires a small set of parameters and it can easily be optimized with standard optimization techniques, unlike the VAR model which requires a very high number of parameters to be optimized.

C. Model Application

Finally, the trained model is applied to compute the forecast values – ③ in Figure 2. The target period is denoted by $t+p$ where p is the distance between the period used for modeling $t-1$ and the last period recorded $t+p-1$. In most cases this will correspond to the length of one season (e.g., 12 months in the case of monthly data). We tested the data sets used in the evaluation section and found the natural seasonality to be the optimal choice. This task may be automated by using the auto correlation function (acf) which analyzes the correlation of the time series with itself at an earlier stage. However, if there is no recognizable seasonality in a data set or there are many very sparse time series such that no transitions can be derived for the preferred value of p, a manual adaption is still possible.

The forecast calculation can be summed up with the following formula:

$$\hat{\vec{X}}_{t+p} = \alpha_1 \cdot \vec{X}_{t+p-1} + \alpha_2 \cdot \vec{E}^1_{t+p-1} + \cdots + \alpha_k \cdot \vec{E}^k_{t+p-1} + \alpha_{k+1} \quad (7)$$

The forecast values of all time series $\hat{\vec{X}}_{t+p}$ at time $t+p$ (e.g., December 2015) are computed based on the recorded feature values $\vec{X}$ and $\vec{E}^1, \ldots, \vec{E}^k$ at time $t+p-1$ (e.g., November 2015) and the estimated parameters $\alpha_1, \ldots, \alpha_{k+1}$ of the same transition one season ago: $t-1 \rightarrow t$ (e.g., November 2014 $\rightarrow$ December 2014). As a precondition, we require only one observation for a particular time series to be able to derive a forecast value: the recorded value at time $t+p-1$. Take the last time series on the right side of Figure 2 as an example and assume it's first value is recorded in period $t+p-1$ (e.g., November 2015). None of the forecasting approaches mentioned in Section II would be able to derive a forecast value, since there is no data for the model creation. Our approach still can compute a forecast value at time $t+p$ by using the estimated parameters over all available time series (Equation 6) and applying the model to the first monitored value of the specific time series.

TABLE I
DATA SETS USED DURING EVALUATION.

dataset	sales	electricity
#base time series	2409	1443
history length	36	25730
seasonality (p)	12	48

Now, to be able to forecast all periods within one season, a model is computed for each transition(e.g., Jan $\rightarrow$ Feb, Feb $\rightarrow$ Mar, ..., Nov $\rightarrow$ Dec, Dec $\rightarrow$ Jan). Thus, for monthly sales data, we result in a total number of 12 models. Furthermore, with the whole set of models over all monthly transitions, we also cover the modeling of seasonal effects. To achieve a higher data density and increase the robustness to outliers when there are just a few transitions, our model can be adapted to use the same transition from several previous seasons of the available historical data (e.g., Nov 2014 $\rightarrow$ Dec 2014, Nov 2013 $\rightarrow$ Dec 2013, Nov 2012 $\rightarrow$ Dec 2012, ...).

By definition, our approach calculates forecast values on the base/most fine grained aggregation level. This makes it easy to obtain forecasts on every higher aggregation level, e.g., brand or total sales amount for the sales domain, by simply executing an aggregation step after the forecast calculation. Thus, our cross-sectional approach is able to forecast *every* possible aggregation level with just one model.

IV. EVALUATION

We conduct an experimental study to evaluate the performance of our cross-sectional forecasting approach in comparison to the existing forecasting methods described in Section II-B. We begin by giving an overview of the experimental setting, including the data sets we used for evaluation. This is followed by a detailed description of the experiments and the discussion of their results.

A. Experimental Settings

Our cross-sectional forecasting approach is implemented in the statistical computing software environment R v3.1.2 [16], which provides efficient built-in functions for model parameter estimation and commonly used forecasting techniques. We build the core of our forecasting approach (i.e., Equation 6) using a multiple linear regression model. The experiments were executed on a notebook with an Intel Core i7-3630QM@2.4GHz processor and 8GB of RAM.

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786

Data Sets: We evaluated our approach on two data sets from different domains. Table I summarizes the most important information about the data sets.

Sales The first data set is taken from the sales domain. It is provided as a private data set by a market research company. It contains 2409 base-level time series from the field of consumer electronics sold in Germany recorded in a monthly granularity. 4.6% of all time series have a complete history over the full 36 months. This data set allows us to calculate forecasts on several aggregation levels. In addition to the top aggregate, i.e., total sales amount in one period, and the base aggregation level, we can consider aggregation levels in between, e.g., the sales for any technical or descriptive attribute of the products like brand, color or energy consumption. This aggregation level is denoted as attribute level in the following experiments. The missing values in this data set may have two different causes. The most common cause was already mentioned in Section I, the sales of an item are not reported by outlets which did neither sell nor stock the specific item. This leads to missing values in item-level time series when there where no sales in any outlet at all and can even cause gaps on the brand-level for very small and exclusive brands. The second possible cause is that time series values may become missing when they are not transmitted or received correctly for the monthly report. Since the most common reason for missing values is that a product was not sold, the default strategy for handling gaps is to assume a zero value for every missing value on any aggregation level.

Electricity The second data set originates from the energy domain. It is a public data set of Smart Meter data monitored by the Commission for Energy Regulation (CER) and made available by the Irish Social Science Data Archive (ISSDA)[3]. The data set consists of 1443 time series, representing the energy consumption of individual households. Time series are monitored in half hour granularity and only 0.7% of them have a complete history of 25730 values, which equals a monitoring time of one and a half year. Missing values in this data set arise when a Smart Meter temporarily looses internet connection or suffers any other kind of technical malfunction. Filling the gaps in such a data set is not a trivial process, because missing values can't just be assumed to be zero values. To enable the comparison of our cross-sectional approach with other algorithms on this data set we interpolated the missing values according to the metering code [10]. It defines a set of rules prescribed by German authorities to interpolate incomplete Smart Meter data. Basically it contains two rules: The first rule states that gaps with a size of up to two hours have to be filled via linear interpolation. The second rule states that gaps longer than two hours are filled with substitute value calculation. To achieve this, values from a preceding week are used and scaled to fit the values before and after the gap. If this is not possible because no earlier historical data is available, values of a similar measuring site should be used. This is not possible for our example as there is no metadata available to determine similarity between individual time series. Due to the lack of an alternative, we used zero values in this case. The only other option would have been to manually compare intermittent time series to every other time series in order to find similarities and derive values for interpolation accordingly. However, we considered the manual effort for this as unreasonably high.

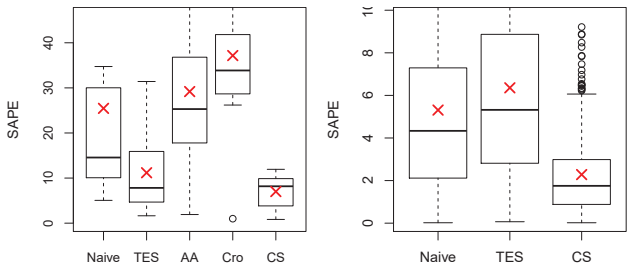
B. Forecast Accuracy

In the first series of experiments we evaluate the accuracy of our forecasting approach in comparison to the traditional forecasting techniques described in Section II, on several aggregation levels. We begin with the top aggregation level where all base time series are aggregated to the total sales per month in the sales data set, respectively, the overall energy consumption of all households in the electricity data set. We continue with the attribute aggregation level for the sales data, where all sales are aggregated to any aggregation level in between top and base-level and end with the base aggregation level where every base time series is forecasted and evaluated individually.

Top Aggregation Level: The first experiment is conducted on the highest aggregation level, where all time series are summed up only grouped by the time attribute. For the sales example we calculated the forecast value with Triple Exponential Smoothing (TES), the auto.ARIMA (AA) implementation of R's forecast package [17], Croston's method (Cro) and our cross-sectional forecasting approach. Vector Autoregression (VAR) was not evaluated because on the top aggregation level it equals the ARIMA model since there are no other time series that could be used for modeling. Additionally, the VAR model is not applicable on lower and more fine grained aggregation levels since no sufficiently long historical data is available (see Section II-B). For the electricity example we only used TES for comparison as the implementations of auto.ARIMA, Croston's method and VAR provided by R's forecasting package take a very long time for the optimization of models on long time series histories. We started the corresponding experiments but stopped the calculations after three days without a result. We considered this to be an unacceptable long time for these forecasting tasks, which will be underlined later with the results of our run time experiments in Section IV-E. Additionally, we provide the results of the naïve forecast, which assumes that every period will show exactly the same value as its predecessor. This can be seen as a baseline, should a forecasting technique perform worse than the naïve forecast, it is not an appropriate model for the specific data set, as it does not properly represent its characteristics.

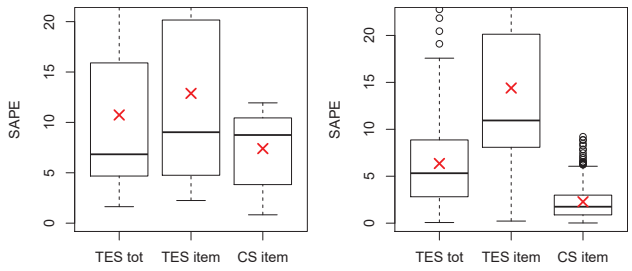
Both data sets were divided into a training and an evaluation part. For the sales data set we used the last year (12 values) for evaluation and for the electricity data the last week (336 values). All preceding data was used for the model training. For TES model training on a monthly granularity a minimum 24 training values was necessary. In both examples we applied a rolling forecast, where we create a new model for every value in the evaluation part of the time series and calculate the corresponding forecast. Then we compare the forecast values of all approaches to the corresponding real time series

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786



(a) Sales (b) Electricity

Fig. 3. Forecast error on the top aggregation level.



(a) Sales (b) Electricity

Fig. 4. Forecast error for the top aggregation level using hierarchical forecasting.

values and calculate the forecast error with the SAPE measure (Symmetric Absolute Percentage Error):

$$SAPE = \frac{|x - \hat{x}|}{(x + \hat{x})/2} \cdot 100, \tag{8}$$

where x is the real time series value and $\hat{x}$ is the corresponding forecast value of one of the evaluated techniques. We use the SAPE measure because it is easier to interpret and compare than the absolute error, which shows the absolute deviation of the forecast from the real time series value. Furthermore, this error measure can be applied even when the real time series value equals zero, where other relative error measures are not defined. If the time series value and the corresponding forecast value both equal to zero and even the SAPE is not defined anymore we have in a forecast error of zero.

The results for both data sets are shown in Figure 3. The left diagram (Figure 3a) shows the forecast errors for the sales data as a Box-Whisker-Plot. The y-axis denotes the SAPE forecast error. Each box represents one evaluated forecasting technique, that shows the distribution of the errors over the full evaluation part. The red cross × denotes the corresponding average error. The right diagram (3b) shows the results for the electricity data set. It is clearly visible that our approach achieves the lowest forecast errors, and therefore the highest accuracy, out of all evaluated techniques. Since both data sets show a well recognizable seasonal pattern Triple Exponential Smoothing also achieves good results. In contrast, auto.ARIMA and Croston's method are not capable of modeling seasonal effects and, hence, achieve the lowest accuracy. The naïve forecast achieves good results for the electricity data set because this data set is recorded in a very fine grained time granularity and therefore shows only minor changes between two successive time series values.

In a second experiment we demonstrate that the high accuracy of our approach is not the result of much better forecastable time series on the base aggregation level, which is an effect often exploited in hierarchical forecasting [7]. To obtain the optimal forecast, the aggregation level on which the forecast values are calculated is chosen by an optimization process before the forecasting process takes place and then the forecasts are aggregated or disaggregated to receive the demanded target aggregation level. In this experiment we execute TES on the base aggregation level for both data sets and aggregate the results to receive a prediction for the top aggregation level. The results of this experiment are shown in Figure 4 next to the results for TES and our cross-sectional approach from the first experiment. The right box in both diagrams shows the results of TES applied on the top aggregation level. The second box shows the results of TES applied on the base aggregation level with subsequent aggregation and the third box shows the application of our cross-sectional approach applied on the base aggregation level as described in Section III.

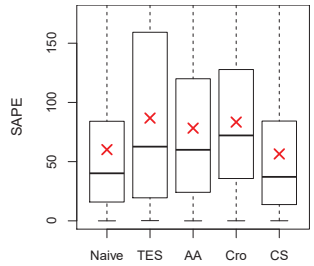
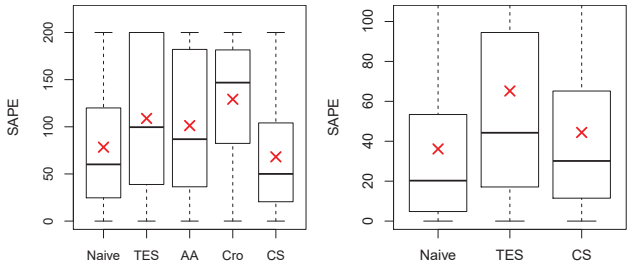


Fig. 5. Forecast error on the attribute aggregation level.

The results show that the aggregation of the base forecasts calculated with TES does not lead to an increased forecast accuracy. The base time series show much more randomness in their behavior than the time series on higher aggregation levels. Therefore, it is much harder to calculate individual forecasts for every time series based only on its own history. Our approach on the other hand uses the historical data of all available base-level time series and can compensate for unpredictable behavior of individual time series. Other forecasting techniques may achieve better results than TES, but the choice of an appropriate modeling technique states a different optimization problem and it is still questionable if it will lead to a higher accuracy than our cross-sectional forecasting approach.

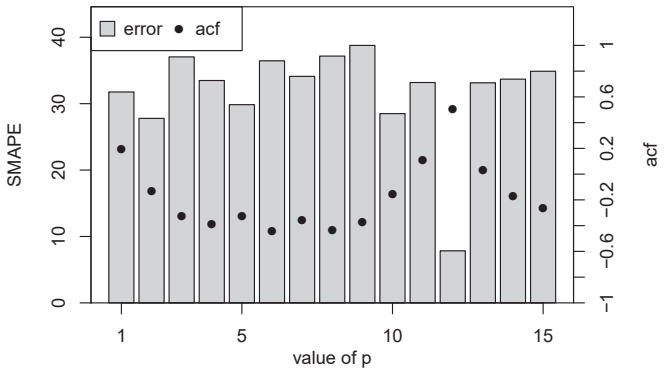
Attribute Aggregation Level: The third experiment focuses on the attribute aggregation level. Since only the sales data set features a hierarchical structure with more than top and base aggregation levels, we only use this data set for evaluation. The examined aggregation level represents any describing attribute, e.g., a technical feature, the brand or the color, of the monitored consumer electronics devices. We conduct the same

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786

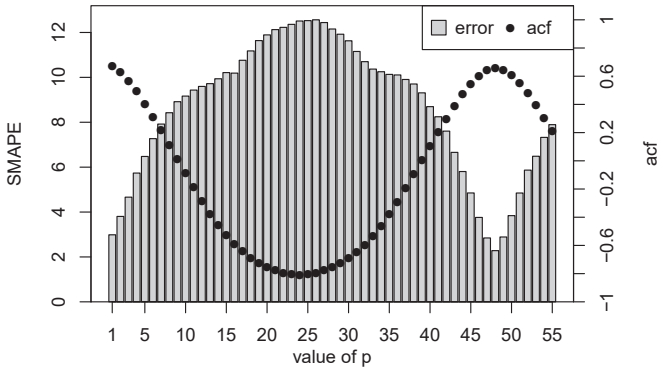


(a) Sales (b) Electricity

Fig. 6. Forecast error on the base aggregation level.



(a) Sales



(b) Electricity

Fig. 7. Influence of the Parameter p on the forecast error.

experiment as on the total aggregation level in the previous section. First, we calculate a forecast for all distinct attribute values, then the deviation to the corresponding real time series value is computed. TES, AA and Cro are applied directly on the aggregated data and our cross-sectional approach is again applied on the base aggregation level and the forecasts are aggregated afterwards.

The obtained results are shown in Figure 5. Again, there is one box for every evaluated forecasting technique, including the naïve forecast. As in the previous experiment our approach reaches the highest accuracy. The naïve forecast achieves the second best result, because on the attribute aggregation level there are already some time series which show such a high degree of randomness that the other three techniques are not able to identify any systematic behavior in the historical data. Please note, that the general increase in the average forecast error is caused by the high number of attribute level time series which show significantly stronger and less predictable fluctuations than the top-level aggregate time series.

Base Aggregation Level: Finally, we also evaluate the accuracy of our forecasting approach on the base aggregation level. For this, we calculate forecasts for every single base-level time series and calculate the forecast error. The results of this experiment are shown in Figure 6. The first thing to observe is that the overall forecast error increases in comparison to the total and attribute aggregation levels, because time series on the item-level have a less predictable behavior and show higher randomness than aggregated time series. Our cross-sectional forecasting approach still reaches the highest accuracy of all compared approaches for the sales data set. In contrast to the traditional forecasting techniques TES, AA and Cro, that try to predict the future values of a time series only based on it's own historical values, our approach utilizes information from other base time series and benefits from it. The base time series in the electricity data have such a high portion of unpredictable behavior, that even our approach is not capable to properly model the data sets characteristics anymore.

C. Influence of Parameter p

In the next experiment we examined the influence of the parameter p introduced in Section III-C. This parameter denotes the time distance between the data values used for model creation and the forecasted value. We conduct the following experiment to show that 1) setting p to the natural seasonality of a data set always leads to the best forecasting result and 2) there is a high correlation between the auto correlation function and the forecast accuracy. For both example data sets we execute our cross-sectional forecasting approach with different values for p to receive forecast values for the top aggregation level. Starting with $p = 1$ we gradually increase the value for p and stop a few values after the natural seasonality of each data set. Additionally, we measured the result of the auto correlation function (acf), which shifts a time series by p periods and then calculates the correlation between the shifted and the original series. The evaluation part of the time series was the same as in the previous experiments. Results for this experiment are presented in Figure 7. The x-axis shows the values of p, the left y-axis denotes the forecast error and the right y-axis shows the corresponding result of the acf. The forecast error is measured with the SMAPE measure which is the mean of all SAPE error values calculated on the evaluation part of the time series. Thus, the gray bars, which show the forecast error of our approach, represent the same error values as the $\times$ in the previous diagrams. The black dots $\bullet$ represent the corresponding results of the acf with a time shift of p periods.

The first thing to notice is that the highest accuracy is achieved when p is set to the natural seasonality of the data sets. These are $p = 12$ for the monthly sales data and

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786

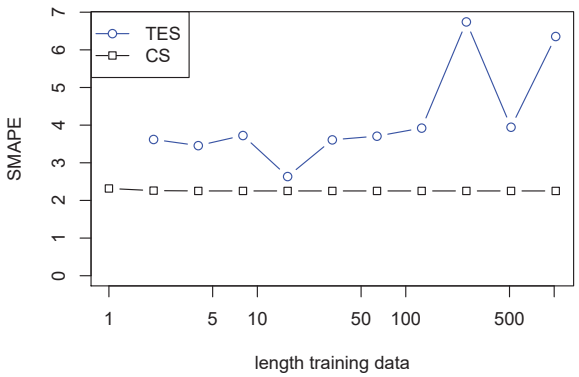


Fig. 8. Training data length

$p = 48$ for the electricity data set with a 30 minute granularity. Second, in both cases there is a strong negative correlation between the forecast error and the result of the acf, -0.76 for the sales data set and -0.96 for the electricity data set. The higher the value of the acf, the more similar the shifted and the original time series are. As a consequence our model becomes more accurate with a high value of the acf, since it assumes that time series of the same data set show a similar behavior after p periods.

D. Training Data

In this experiment we show that our approach already reaches a very high accuracy with a minimum of training data and is significantly more robust to fluctuations in the time series data than traditional approaches. We use a similar setup as in the first series of experiments: calculate forecast values for the evaluation part of the data sets and compare them to the corresponding real time series values by evaluating the SAPE error measure. In this series of experiments we vary the amount of available training data for the model training process, starting by one period and doubling the length of training data with every step. For this evaluation we choose the electricity data set since its history is long enough to vary the amount of training data in an appropriate range. Our cross-sectional approach always uses all time series of the training data to extract the corresponding time slices from every available period. TES is used as the comparison technique and always uses the whole amount of available data for model training. The results of the experiment is shown in the diagram in Figure 8. The log-scaled x-axis shows the number of periods used as training data, the y-axis shows the corresponding forecast errors to the different number of periods of training data. The black line shows the results for our cross-sectional forecasting approach and the blue line shows the forecast errors for TES.

Our approach again reaches the lowest forecast error regardless of the amount of available training data. We can observe a small increase in accuracy when moving from one period of training data to two periods. However, after that a further increase of the amount of training data does not lead to a significant change in the forecast accuracy in any direction. TES on the other hand is massively influenced by the available amount of training data. One period of training data is not enough at all for this technique, since it requires more information to even initialize the model. Furthermore, the gradually increase in the available training data does not lead to a systematic increase in the forecast accuracy, as one might expect. Actually, it leads to strong fluctuations. At 16 periods of training data the accuracy of TES shortly approaches the accuracy of our algorithm. While a more fine grained analysis and selection of the training data for TES could make this technique competitive again, it would also create a completely new optimization problem, i.e., identifying the optimal amount of training data. Solving this problem would require even more historical data to ensure reliable results. This shows that more data does not necessarily leads to more accurate forecasts. The strong fluctuations in the forecast error of TES in the last three measurements, i.e., 256, 512 and 1024 periods of training data, may be explained by the high increase of the available training data. With more available data there is also a risk of using parts of a time series which have different characteristics. This has a negative influence on the forecast accuracy of models like TES which derive forecast values only based on the history of one single time series, and therefore, can not compensate the randomness. In summary, compared to TES our approach is literally data-frugal, achieving high accuracy even on a minimum amount of training data. In addition, its pooled parameter estimation makes it very robust against fluctuations in the time series data.

E. Execution Time

In the last experiment we will show that our approach is also fast and can calculate forecast values for every aggregation level in nearly constant time. This is a property most traditional techniques severely lack, which means they do not scale well when applied on lower aggregation levels. Using the setup from the previous experiments, we calculate forecast values for all three different aggregation levels of the sales data set and use TES as comparison method. We execute both forecasting models ten times over the whole evaluation part of the data set and use the average runtime for evaluation.

TABLE II
COMPARISON OF EXECUTION TIMES.

	total	attribute	base
TES	0.15s	21.87s	564.92s
CS	0.73s	0.73s	0.69s

The results are shown in Table II. The first column shows the run time for the total aggregation level for TES (upper row) and our cross-sectional approach (lower row). On this aggregation level TES is four to five times faster than our approach. However, both techniques have a run time of under one second and therefore the measured difference is negligible. The results on the attribute aggregation level (column two) show that, even if the creation of a single model is faster,

Final edited form was published in "2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris 2015 2009", S. 1-10,
ISBN 978-1-4673-8272-4
http://dx.doi.org/10.1109/DSAA.2015.7344786

it takes TES 30 times more time to create all the models and calculate the forecasts than it takes for our cross-sectional approach. On the base-level (column three) TES has to create and optimize more than 2400 models for every period in the evaluation part of each time series and, hence, needs significantly more time than our approach. Cross-sectional forecasting even gets slightly faster on the base-level than on higher aggregation levels, because it already calculates the forecasts on the lowest aggregation level and, therefore, can omit the aggregation step.

V. Conclusions

In this paper, we introduced a new forecasting approach that addresses the challenges of data incompleteness and volume, brought forth by Big Data. For this, we break with some of the main principles of traditional forecasting techniques: create one model for each time series and the more historical data the higher the forecast accuracy. Our cross-sectional forecasting focuses on sets of time series, for which it creates and optimizes one single model to calculate forecast values for every time series. Instead of using complete sequences of historical values for model construction, our approach only uses small time slices, that represent the transition between two consecutive time periods. As these time slices are taken from all time series of a data set in a cross section, our approach can handle missing values without additional processing and also becomes robust against random fluctuations of individual time series. Our experimental evaluation shows, that our cross-sectional forecasting approach achieves a higher accuracy on any aggregation level than traditional forecasting approaches, while only requiring a minimum of training data and a much shorter runtime.

With our cross-sectional forecasting we proposed an approach that is clean, simple and lightweight at its core. Besides high accuracy and fast runtime, it can natively handle missing values and allows the inclusion of additional external factors into the forecasting process. This is why we see a lot of potential and many interesting future research directions for our approach. Some of them are described in the following.

Feature Selection: Currently, we have to select the features (e.g., sales units and order units for the sales data set) of the cross-sectional forecasting approach manually. We plan to investigate and expand our approach with an automatic feature selection mechanism, that will identify and include the most useful features into the forecasting process, in order to further improve the accuracy.

Long Range Forecasting: In this work we focused on the calculation and evaluation of one-step ahead forecasts. However, there are many application scenarios, where forecasts for more than only one period ahead are necessary to properly plan for future developments. This is why the extension of our cross-sectional forecasting approach to long range forecasting will be a major goal of our future research.

Advanced Model Creation: Currently our model creation is pretty straight forward and utilizes a simple cross- section of all transitions from one preceding season. In the future, we aim to introduce more functionality in order to increase the accuracy. This involves mechanisms for training data selection, e.g., inclusion of additional cross sections if the current one is too sparse, as well as methods for model calculation, e.g., weighting the transitions of a cross-section.

Acknowledgment

The work presented in this paper has been funded by the GfK-Nürnberg e.V. Further, we would like to thank Marcel Spranger for supporting our work.

References

[1] "Arrowhead Framework," http://www.arrowhead.eu/, 28.04.2015.
[2] "The MIRABEL Project," http://www.mirabel-project.eu/, 28.04.2015.
[3] Irish Social Science Data Archive (ISSDA), *CER Smart Metering Project*, The Commission for Energy Regulation (CER), 28.04.2015, www.ucd.ie/issda.
[4] J. S. Armstrong, *Principles of forecasting: A handbook for researchers and practitioners*. Norwell: Kluwer Academic Publishers, 2001.
[5] T. M. McCarthy, D. F. Davis, S. L. Golicic, and J. T. Mentzer, "The Evolution of Sales Forecasting Management: A 20-Year Longitudinal Study of Forecasting Practices," *Journal of Forecasting*, vol. 25, no. 5, pp. 303–324, Aug. 2006. [Online]. Available: http://doi.wiley.com/10.1002/for.989
[6] R. J. Hyndman and A. V. Kostenko, "Minimum sample size requirements for seasonal forecasting models," *Foresight*, no. 6, pp. 12–15, 2007. [Online]. Available: http://www.robjhyndman.com/papers/shortseasonal.pdf
[7] G. Fliedner, "Hierarchical forecasting: issues and use guidelines," *Industrial Management & Data Systems*, vol. 101, no. 1, pp. 5–12, 2001.
[8] R.-S. Jeng, C.-Y. Kuo, Y.-H. Ho, M.-F. Lee, L.-W. Tseng, C.-L. Fu, P.-F. Liang, and L.-J. Chen, "Missing data handling for meter data management system," *Proceedings of the the fourth international conference on Future energy systems - e-Energy '13*, no. 2, p. 275, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2487166.2487204
[9] S. Van Buuren and K. Groothuis-Oudshoorn, "Multivariate Imputation by Chained Equations," *Journal Of Statistical Software*, vol. 45, pp. 1–67, 2011. [Online]. Available: http://igitur-archive.library.uu.nl/fss/2010-0608-200146/UUindex.html
[10] VDE Verband der Elektrotechnik Elektronik Informationstechnik e.V., "Messwesen Strom (Metering Code); VDE-AR-N 4400," 2011.
[11] J. M. Wooldridge, *Introductory econometrics: a modern approach*, fifth edit ed. Mason: South-Western, Cengage Learning, 2013.
[12] T. Riise and D. Tjozstheim, "Theory and practice of multivariate arma forecasting," *Journal of Forecasting*, vol. 3, no. 3, pp. 309–317, Jul. 1984. [Online]. Available: http://doi.wiley.com/10.1002/for.3980030308
[13] J. D. Croston, "Forecasting and Stock Control for Intermittent Demands," pp. 289–303, 1972.
[14] L. Shenstone and R. J. Hyndman, "Stochastic models underlying Croston's method for intermittent demand forecasting," *Journal of Forecasting*, vol. 24, no. 6, pp. 389–402, 2005. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/for.963/abstract
[15] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the Twentieth International Conference on Machine Learning*, 2003, pp. 856–863.
[16] R Development Core Team, *R: A Language and Environment for Statistical Computing, Reference Index Version 3.1.1*, R Foundation for Statistical Computing, 2014, http://www.r-project.org.
[17] R. J. Hyndman and Y. Khandakar, "Automatic Time Series for Forecasting: The Forecast Package for R," *Journal of Statistical Software*, vol. 27, no. 3, 2008. [Online]. Available: http://webdoc.sub.gwdg.de/ebook/serien/e/monash_univ/wp6-07.pdf