Lars Dannecker, Elena Vasilyeva, Matthias Boehm, Wolfgang Lehner, Gregor Hackenbroich

Efficient Integration of External Information into Forecast Models from the Energy Domain

Erstveröffentlichung in / First published in:

Advances on Databases and Information Systems: 16th East European Conference. Posen, 18.-21.09.2012. Springer, S. 139–152. ISBN 978-3-642-33074-2.

DOI: http://dx.doi.org/10.1007/978-3-642-33074-2_11

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

Efficient Integration of External Information into Forecast Models from the Energy Domain

Lars Dannecker[1], Elena Vasilyeva[1], Matthias Boehm[2,⋆], Wolfgang Lehner[2], and Gregor Hackenbroich[1]

[1] SAP Research Dresden, Chemnitzer Str. 48, 01187 Dresden, Germany
{lars.dannecker,gregor.hackenbroich}@sap.com
[2] Technische Universität Dresden, Database Technology Group
Nöthnitzer Str. 46, 01187 Dresden, Germany
{matthias.boehm,wolfgang.lehner}@tu-dresden.de

Abstract. Forecasting is an important analysis technique to support decisions and functionalities in many application domains. While the employed statistical models often provide a sufficient accuracy, recent developments pose new challenges to the forecasting process. Typically the available time for estimating the forecast models and providing accurate predictions is significantly decreasing. This is especially an issue in the energy domain, where forecast models often consider external influences to provide a high accuracy. As a result, these models exhibit a higher number of parameters, resulting in increased estimation efforts. Also, in the energy domain new measurements are constantly appended to the time series, requiring a continuous adaptation of the models to new developments. This typically involves a parameter re-estimation, which is often almost as expensive as the initial estimation, conflicting with the requirement for fast forecast computation. To address these challenges, we present a framework that allows a more efficient integration of external information. First, external information are handled in a separate model, because their linear and non-linear relationships are more stable and thus, they can be excluded from most forecast model adaptations. Second, we directly optimize the separate model using feature selection and dimension reduction techniques. Our evaluation shows that our approach allows an efficient integration of external information and thus, an increased forecasting accuracy, while reducing the re-estimation efforts.

Keywords: Forecasting, External Information, Efficiency.

1 Introduction

In the energy domain the availability of accurate forecasts of future electricity consumption and production is a prerequisite for the balancing of energy demand and supply and thus, for the stability and efficiency of the energy grids. Forecasting employs quantitative models—known as forecast models—that mathematically describe the historical behavior of a time series. Most forecast models

⋆ The author is currently visiting IBM Almaden Research Center, San Jose, CA, USA.

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

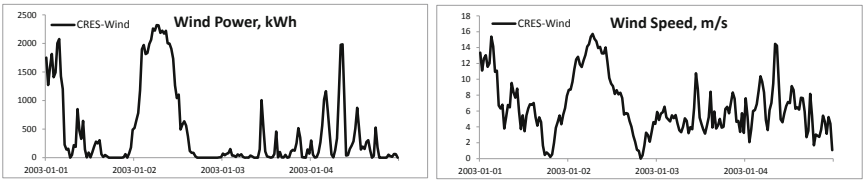


Fig. 1. Example Time Series: Wind Energy Production and Wind Speed

capture a parameterized relationship between past and future values to express different characteristics of the time series such as seasonal patterns or trends. To produce accurate forecast, the parameters are estimated on a training data set, with the goal of minimizing the forecast error. Commonly, this estimation is conducted using local (e.g., LBFGS, Nelder Mead Downhill Simplex) or global (e.g., Simulated Annealing) numerical optimization algorithms for non-linear optimization. The complexity of the parameter estimation greatly depends on the number of parameters comprised in the forecasting model, because the parameter search space increases exponentially with the number of parameters.

Current forecasting approaches for energy demand often produce accurate results, but new developments in the energy domain pose additional requirements on the calculation speed of the employed forecast models. New market dynamics in conjunction with the emerging smart grid technology and an increased integration of renewable energy sources (RES) require real-time capabilities for balancing energy demand and supply. Research projects such as MIRABEL [1] and MeRegio [2] address the issue of real-time balancing by employing new approaches like flexible demand and supply requests, dynamic price signals or demand-response systems. A fundamental prerequisite for approaches in this area is that accurate forecasts are available at any point in time. Fortunately, in the energy domain the current consumption and production can be measured constantly and thus, new values can be continuously appended to the time series. However, to enable a certain degree of accuracy, forecast models must be continuously adapted with respect to the evolving time series. This generally means the re-estimation of all model parameters. The matter becomes more involved as many forecast models used in the energy domain consider external information to increase the forecasting accuracy. The reason is that the future development of most time series does not solely rely on historical values, but is influenced by other correlated aspects [3,4,5]. The energy production of a wind power plant, e.g., greatly depends on weather conditions like the wind speed. In Figure 1 we present two time series from a real-world data set, showing (1) the energy supply of a wind park and (2) the corresponding wind speed. The positive correlation between both time series is obvious. Thus, to allow a certain degree of accuracy, it is not sufficient to predict the future production solely based on the previously produced power, but it is also necessary to consider weather information as an external influence. The resulting accuracy gain depends on the degree of dependence between the forecasted time series and the considered external influences.

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

The inclusion of external information is typically conducted by adding the external time series as an additional component into the forecast model. While this naïve approach typically increases the forecasting accuracy, the integration also adds additional parameters to the model and thus, additional dimensions to the exponentially increasing parameter search space. As a result, the time necessary for estimating the forecast model parameters increases with the number of considered external information. Thus, highly accurate forecast models that consider external information might be unusable in a real-time environment.

To still allow highly accurate forecasts, we propose an integration approach that greatly reduces the additional efforts when adding external information. Our approach reduces the number of additional parameters and at the same time also optimizes the handling of external information. The core idea is to employ a single separate model to represent all external information and their relationship to the main dependent variable. This relationship remains relatively stable and only changes slightly over long periods of time. This leads to the assumption that the separate model is more stable and does not need to be included in most adaptation processes. Furthermore, we also apply feature selection and dimension reduction techniques to directly reduce the number of parameters in the separate model. While we describe our approach in conjunction to the energy domain, it can also be applied to time series with external influences and forecast models for other domains. The paper is organized as follows:

- First, we describe the influence of exogenous information in Section 2.
- Second, we present our integration approach for external information that especially considers the efficiency of the forecast model in Section 3.
- Fourth, we present the results of our evaluation that show significant speed-up for the parameter estimation of multi-equation models in Section 4.
- Finally, we present related work and conclude the paper in Sections 5 and 6.

2 Background of Forecasting with External Influences

The standard approach for incorporating external information into forecast models is to include additional terms that reference the values of the external time series. A typical model from the energy domain that supports external influences is the highly accurate multi-equation model EGRV [6]. This model assigns individual sub-models to each time period within a selected season (e.g., one hour per day) and considers the temperature T as external influence. For this purpose, each sub-model incorporates additional terms, where an example model for one hour is denoted as:

$$\begin{aligned} \text{Hour1} &= \alpha \cdot \text{Deterministic} + \beta \cdot \text{Temperature} + \gamma \cdot \text{Load8} + \delta \cdot \text{Lags} \\ \text{Temperature} &= \beta_1 \cdot T + \beta_2 \cdot T^2 + \beta_3 \cdot \frac{1}{N} \sum_{d=0}^{-N} \overline{T}_d + \beta_4 \cdot T_d^{max} + \beta_5 \cdot T_{d_{-1}}^{max}. \end{aligned} \quad (1)$$

Thus, the model components are deterministic variables (e.g., current day), the energy load at 8:00am of the previous day (Load8), a definable number of previous time series values (Lags) and the temperature as T plus T^2, which reflect

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

the quadratic (i.e., non-linear) dependency between energy consumption and temperature. Also, to involve the overall temperature level and the long term temperature trend, the current and past days maximal temperature T^{max} and the moving average of the past seven days mid temperature $\overline{T}_d$ are added.

Adding external information always comes with the trade-off of adding additional parameters and thus, increasing the dimensionality of the parameter search space. For multi-equation models like the introduced EGRV model, for example, the inclusion of a single influence results in five additional parameters per sub-model. This leads to an increased dimensionality of the parameter search space for each sub-model when estimating the parameters; e.g., when assuming 10 parameters in the base model, the search space dimensionality of all sub-models increases from X^{10} to X^{15}. This significantly increases the necessary time for finding a suitable parameter combination, which is especially an issue in the face of evolving time series and the resulting necessity to continuously adapt the forecast models to new situation. Thus, models considering external information might get unusable when real-time capabilities are required.

3 Solution Overview: Integrating External Information

The core idea underlying our approach is to separate the modeling of external information from the actual forecast model. The idea is reasoned by the fact that the relationships between the dependent variable and the external information as well as between the external information only change very slightly. This holds especially true for external information forming a stable physical system (because physical systems typically do not change rapidly) and external information with a high correlation to the main dependent variable (rapid changes would lead to a rather low correlation). The weather information typically used as external information in the energy domain, form such a stable physical system and thus, exhibit a stable relationship to energy demand and supply. The stability of the relationships lead to the assumption that a separate model for the external information is also more stable than the forecasting model. Thus, for the forecast model adaptation in most cases it is sufficient to just re-estimate the base model's parameters and exclude the separate external information model. This means that in most cases no additional time for adapting the forecast model is needed, even when integrating external information. Furthermore, when dealing with multi-equation models, only a single external information model is necessary that can be reused for all involved sub-models. To create the external

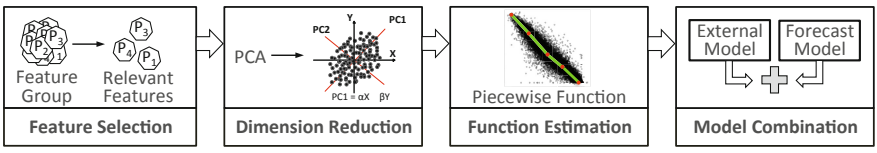


Fig. 2. Process for Integrating External Information

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

information model our framework comprises the process illustrated in Figure 2: First, we use feature selection techniques to select the most relevant external information. Second, we apply dimension reduction techniques to reduce the number of parameters involved in the external information model. Afterwards, we estimate a function that describes the output of the dimension reduction best. Finally, we combine the base forecast model with the external information model. In the following, we describe the steps of our framework in more detail.

3.1 Feature Selection

In a first step to optimize the external information model, we determine the most relevant external information. With this step, we avoid the inclusion of irrelevant factors that might have negative implications when used in conjunction with dimension reduction techniques. For this purpose, we create a correlation matrix, comprising the correlation between the dependent variable and the external factors as well as the correlation between external factors. The correlation is quantified using a correlation measure like the Pearson Correlation Coefficient (PCC) $r_{x,y}$. The choice of a suitable correlation measure depends on the type of the considered variables. While the PCC provides good results in conjunction with continuous variables, rank correlations like Spearman's Ro ρ [7] additionally support ordinal variables. To also include nominal variables more advanced techniques like statistical hypothesis testing are required. In this paper, we only focus on continuous and ordinal variables. Figure 3 illustrates an example correlation matrix using the PCC. With the help of the correlation matrix, we now select the features as follows:

1. We determine an inclusion threshold ϵ that is the median of the absolute correlations to the dependent variable. In our example $\epsilon = 0.35$, thus X_1 and X_2 (Black box in Figure 3) are selected as relevant influences.
2. To avoid redundant external factors, we evaluate the correlation between the selected influences and prune one of two influences whenever the correlation $|r|$ is larger than our defined similarity threshold ω. In our example we define $\omega = 0.9$ and thus, no influences are pruned.
3. To also cover inter-relationships between the selected influences and all available external factors, we also evaluate the cross correlation between them.

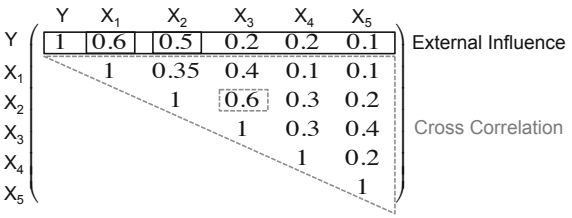


Fig. 3. Example Correlation Matrix

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

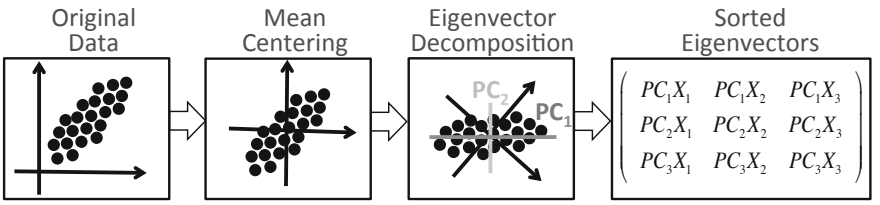


Fig. 4. Steps of the Principal Component Analysis

We then select all influences with a correlation factor $|r|$ larger than our defined cross-correlation threshold σ. In our example, $\sigma = 0.5$ and X_3 with a correlation of 0.6 to X_2 (Grey dashed box in Figure 3) is selected.

As a result, from our example the influences X_1, X_2 and X_3 are candidates for the external information model. From multiple empirical evaluations we recommend to define $\omega = 0.9$ and $\sigma = 0.5$. While our approach provides a sufficient selection for most cases, it is still possible that no influences are pruned at all. However, in this case it is at least indicated that all provided external influences will create a benefit for the forecasting process.

3.2 Increasing Estimation Efficiency by Reducing Dimensionality

In the feature selection step unnecessary influences are pruned form the external influence model to avoid negative side effects. In a second step, we directly reduce the parameters involved in the external influence model, using dimension reduction techniques. We decided to focus on the Principal Component Analysis (PCA) [8,9] that is a statistical approach to reduce the dimensionality of the data. The PCA removes redundancies in the data, but aims to preserve as much valuable information as possible. The PCA uses an orthogonal linear transformation to project a data set comprising multiple features into a new vector space with less or equal dimensions than the original space. In the new space, the original variables are represented by a set of linear combinations of them, where as part of the transformation the correlation between the linear combinations is minimized. The resulting combinations are called the principal components and can be seen as lines through the multi-dimensional space starting from the origin and minimizing the mean square error (MSE) to the data points. The principal component with the smallest MSE is called the first principal component and describes the greatest variance in the data. Each further principal component must be orthogonal to the preceding one [9]. Thus, in contrast to other dimension reduction techniques like the factor analysis, the PCA produces a result ordered by the significance of the influence on the dependent variable. This helps to decide which linear combinations should be included. Figure 4 illustrates the steps of the PCA, which is applied to our external influence model as follows:

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

1. We subtract the mean from the influences to create a mean-centered matrix.
2. From the mean-centered matrix we compute the $p \times p$ covariance matrix.
3. We conduct the eigenvalue decomposition to create the eigenvalues and corresponding eigenvectors.
4. Eigenvectors are principal components and can be sorted by their eigenvalues. The vector with the greatest eigenvalue is the first principal component.

The number of principal components beneficial for the forecasting model depends on the desired accuracy and performance. Typically, the first principal component provides sufficient accuracy and means the least number of involved parameters. However, for some use cases it is also possible to dynamically add additional principal components as long as they significantly increase the accuracy. This means that in a worst case scenario the number of principal components equals the number of original variables, resulting in a similar estimation time compared to the naïve integration approach.

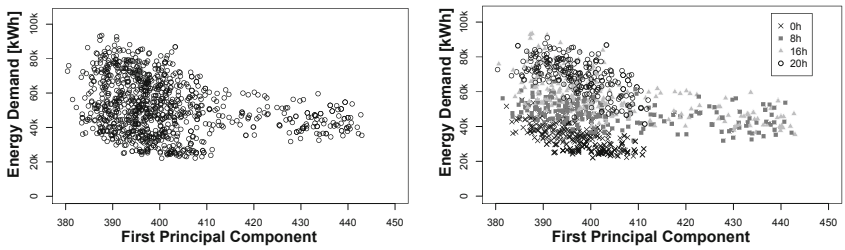
3.3 Determining the Final External Model

The selected principal components represent the entirety of all external factors that mainly influence the development of the main time series. While we could directly use them as linear combinations for our external model, from our experiments we found that relating the principal components to the dependent variable and estimating functions describing the relationship increases the final accuracy for most data sets. In particular, each function approximates the response values of the dependent variable for different values of the selected principal component. To estimate suitable functions, we combine linear regression with the concept of piecewise functions. Thus, we use different linear functions for different value ranges. For this purpose, we follow a two-step process, were we (1) determine suitable ranges and (2) then estimate a linear function for each range.

Finding the Value Ranges. To divide the data into suitable ranges, we first estimate a function that describes the entire value range of the explanatory variable. We can then use the extreme points of this function as the borders for the value ranges. However, the most suitable degree of the function is unknown in advance. Thus, we apply non-linear regression and increase the power of the polynomial step-by-step, starting with a linear function. This process converges when the accuracy benefit of increasing the degree is smaller than the configurable significance parameter θ (e.g., from our experience: 1.0%). As soon as we found a suitable polynomial, we calculate its extreme points and use these points to divide our data. If the resulting polynomial is still linear, no extreme points exist and we directly use the resulting linear function.

However, for some data this method fails, because the principal components do not sufficiently describe the dependent variable. This is illustrated in Figure 5(a), where we see the first principal component and the energy response values. This data is hard to fit and thus, it is hard to find suitable ranges for the principal component. The issue mostly occurs when the external information are

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11



(a) Without Additional Range Variable (b) With Additional Range Variable

Fig. 5. Range Splitting Using Additional Variable

not significant enough. However, for the sake of an increased accuracy it might still be worth to consider them. To address this issues, instead of only dividing the data with respect to the principal component, we additionally employ the most influencing factor of the dependent variable. For most time series this influencing factor is time. Figure 5(b) illustrates the data divided per hour. The data for a single hour is much easier to describe using a polynomial function. Thus, after dividing the data with respect to the most influencing factor, we calculate the extrema for all functions using the approach mentioned above.

Estimating the Range Functions. In the second step, we apply linear regression for each identified range to calculate a linear function that describes this data portion best. The result of our external influence model creation step is a set of linear functions that forms in its entirety the separate external influence model.

3.4 Creating a Combined Forecast Model

In the next step we need to combine the external information model with the original forecast model. We suggest two options: First, the *Indirect Integration*, where both models are combined using a weighted linear combination. The efforts for this additional estimation are low, because the linear combination consists of only two parameters. Second, the *Hybrid Integration*, which is an enhancement of the indirect integration. There, in addition to the external information model, we include the most important external factor directly (and without using PCA) into the forecast model. Thus, we combine the efficiency of the indirect integration with the possible accuracy gain of directly integrating external information. Finally, with the external influence model in place, in most cases only the parameters of the base forecast model must be re-estimated and thus, adding additional external information does only have a minimal performance impact. Plus, even in the rare case where all parameters must be estimated (e.g., for the initialization), the estimation of the external information model can run in parallel. In addition, when dealing with multi-equation models, instead of creating a separate external influence model for each employed sub-model, a single external information model can be reused for all sub-models.

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

4 Experimental Evaluation

In our evaluation, we substantiate the claims of our integration framework and discuss options for implementing it. The evaluation shows that our approach significantly reduces the additional efforts when integrating external information. In addition, in some cases it even provides a better accuracy than the direct integration. We used the introduced EGRV [6] model and the double seasonal exponential smoothing (DSESM) model [10]. The employed datasets are:

- Wind energy supply from CRES [11]: Local wind park; January 1st 2003 to December 31st 2003; 30 min granularity; External Information: windspeed, wind direction, number of wind turbines, min/avg/max energy production.
- Energy demand from the MeRegio project [2]: Households (Selected 7, 40), November 1st 2009 to June 30th 2012; 1 h granularity; External Information: Weather data from the Deutsche Wetterdienst [12]; air temperature, ground temperature, cloud cover, sun duration, wind speed, humidity, pressure.

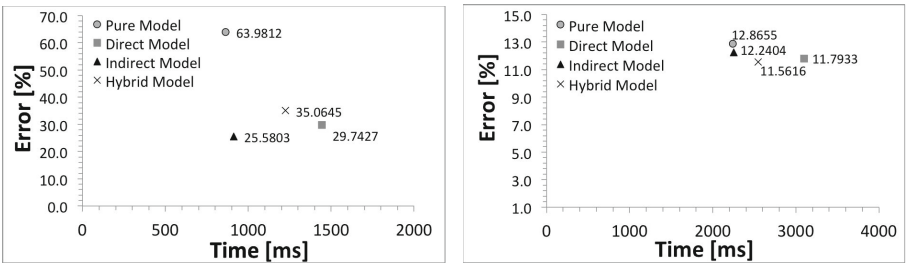
For the parameter estimation, we used the Nelder Mead Downhill Simplex algorithm [13]. The re-estimation process comprises the estimation of the forecast model, the estimation of the separate external information model (EGRV: one model for all sub-models) and the estimation of the final combination. For our experiments we assumed exact predictions for the external information, because uncertain time series are an orthogonal issue that we handle in future work.

As test system we used: Intel Core i7 2635QM (2.0 GHz), 4GB RAM, Mac OSX 10.6.8, C++ (GCC 4.2.1). All results are the average of 20 subsequent runs.

4.1 Time vs. Accuracy

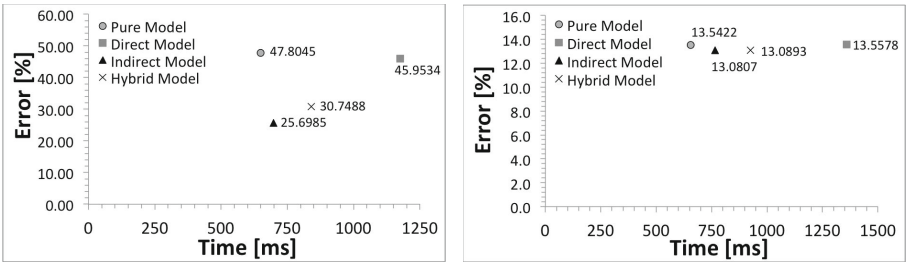
In the first experiment we evaluated the runtime and final accuracy of the parameter re-estimation. Overall, we compared our separate external information model (*Indirect Model*) with the models when adding no external factors (*Pure Model*), a naïve direct integration of the external information without using a separate model (*Direct Model*) and the hybrid solution (*Hybrid Model*) (compare Section 3.4). Selected results are illustrated in Figure 6 (EGRV model) and Figure 7 (DSESM model). We observed very high forecast errors for the CRES supply data set (Figures 6(a) and 7(a)) when not considering external information (*Pure Model*). Thus, only including external information enables accurate predictions for energy supply. For both forecast models and both data sets, the *Indirect Model* showed the best runtime of all solutions considering external information. In addition, our approach even showed the best accuracy for the CRES supply dataset, even though we used PCA. Focusing on the EGRV forecast model: Using our *Indirect Model* approach we limited the additional efforts for adding external information to only 48.64 ms and increased the accuracy by 38.40 %. In contrast, using the naïve *Direct Model* the additional effort was 580.77 ms and the accuracy increase was only 34.24 %. The *Hybrid Model* did not provide better results. Originally we assumed that our method exhibits the lowest runtime, while the *Direct Model* provides the best accuracy. We account the

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11



(a) CRES Wind Supply (b) MeRegio Single Customer (7)

Fig. 6. EGRV Model: Different Integration Approaches



(a) CRES Wind Supply (b) MeRegio Single Customer (7)

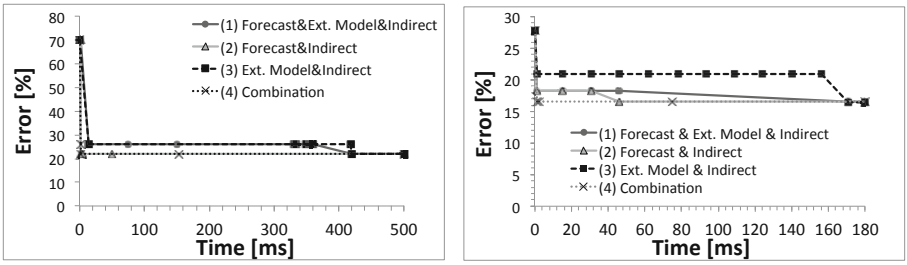
Fig. 7. DSESM: Different Integration Approaches

accuracy advantage of our *Indirect Model* to a possible over fitting when directly including external information. The results for the DSESM model are similar. For the MeRegio single customer energy demand dataset (smaller dependency on external information) the results are more diverse (Figures 6(b) and 7(b)). Using our *Indirect Model* approach we still increased the accuracy with small additional effort of only 12.40 ms, but as expected the accuracy gain was less significant (only 0.63 %). For the EGRV model the *Direct Model* approach provided a better final accuracy (increase of 1.07 %), however, the additional effort is much higher (857.24 ms). A suitable alternative in this case is the *Hybrid Model*, providing a good balance between additional effort (305.78 ms) and accuracy gain (1.30 %). Overall, our framework provides an efficient way of integrating external information and increasing the forecasting accuracy.

4.2 Different Re-estimation Strategies

In this experiment, we evaluated different re-estimation strategies. We compared (1) the re-estimation of all models, (2) the re-estimation of the forecast model and the combined model, (3) the re-estimation of the external influence model and the combined model and (4) the re-estimation of the combined model (Combination) only. Figure 8 illustrates the results. For the CRES supply data set (Figure

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11



(a) EGRV Equation 23: CRES Supply (b) DSESM: MeRegio Customer 40

Fig. 8. Comparison of Re-Estimation Strategies

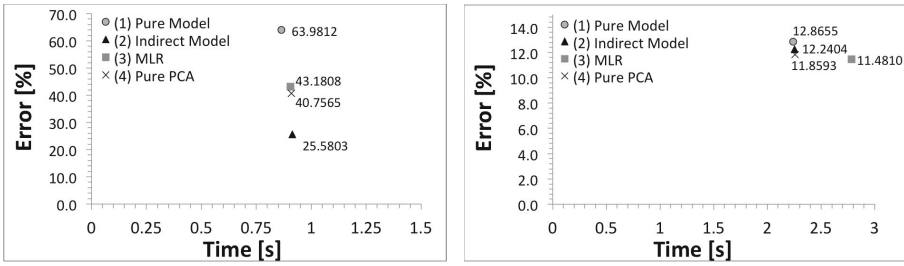
8(a)) all estimation methods finally reached the same accuracy. However, the runtimes of strategies 1 and 3 were longer. When repeating the experiment at different points in time, we observed that strategy 4 did not always reach the best accuracy. Strategy 2 produced the maximal accuracy at all times. The results are similar for the DSESM model. Using the MeRegio demand data set (Figure 8(b)), still all strategies ultimately reach the same accuracy, but their speed is different. We observed the shortest runtime using strategy 4 followed by strategy 2. Strategy 1 and 3 are slower. Strategy 3 exhibits the longest runtime. However, strategy 4 again did not reach the maximal accuracy at all points in time, which finally renders method 2 as most appropriate for both data sets.

To substantiate the results, we simulated an evolving time series, repeating the experiment at several points in time. To do so, we only used 3/4 of the CRES data set (9 month) and added the remaining 1/4 of the data (3 month) successively. After each added value we evaluated the accuracy. We triggered a re-estimation, whenever the average error for the last 10 values was 10 % higher than the last estimation error. We observed similar results at all occasions. However, when only re-estimating the forecast model or the model combination the number of re-estimations increased to 887 compared to 607 when re-estimating all models. Thus, while in most cases it is sufficient to re-estimate the forecast model only, from time to time all models should be re-estimated. However, the forecast model and the external information model can be estimated in parallel, which means an only slight increase of the runtime for most re-estimations.

4.3 Comparing Different Types of External Information Models

In this experiment, we evaluated design alternatives for the external influence model. We compared our approach to alternatives that employ multiple linear regression (MLR) or that directly use the eigenvectors of the PCA. Figure 9 illustrates the results for the EGRV model. For the CRES wind supply data set (Figure 9(a)), all approaches need almost the same time to converge (∼910 ms), but our framework provides the best final accuracy (improved by 15 %). For the MeRegio demand dataset (Figure 9(b)) the pure PCA method and the MLR provide slightly better results (∼0.75 %) compared to our approach. While the

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11



(a) EGRV: CRES Supply (b) EGRV: MeRegio Customer 7

Fig. 9. Comparison of Different External Model Variants

MLR in return needs more time, the pure PCA method finishes in similar time. We repeated the experiment for several points in time and observed similar results, but with changing final accuracies. Hence, the pure PCA method seems to be better suited for the MeRegio data set. We account the smaller dependence of the time series on the external influences to be the reason. This results in the need for a less complex model, where a single principal component is sufficient. However, our approach provides more reliable results on both demand and supply data sets. Furthermore, the difference on the demand data set is rather small, while the advantage of our approach on the supply data set is significant. Thus, our approach is more general and suits demand and supply data sets equally.

5 Related Work

Efficient statistical algorithms directly integrated into database systems gain more and more attention in research and industry. Akdere et al. [14] and Parisi et al. [15] describe several aspects of a predictive database system such as the creation and integration of special forecasting operators. Similarly Duan and Babu describe how to efficiently process forecasting queries [16]. Concerning an increased forecasting efficiency Ge and Zdonik present a skip-list approach to vary data granularity depending on the forecast requirements (e.g., horizon) [17]. In previous work we optimized the forecasting process by introducing our context-aware forecast model repository [18]. There, we store forecast models in conjunction with their time series context (e.g., time series characteristic, external information), to quickly retrieve them as soon as similar contexts appear.

With respect to integrating external information into forecast models, some approaches already exist. Besides forecast models directly supporting external information like the ARIMAX [19] or the introduced EGRV [6] models, other solutions describe more efficient ways. One prominent solution is to directly quantify the influence of the external information on the dependent time series by using a special factor that modifies the model output. Bruhns et al. [20] use such a factor to describe the temperature influence in their regression model. Similarly, Young-Min use an exponential smoothing model that is adjusted to the

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

current temperature with a weather adjustment factor [21]. While the solutions provide an efficient way of integrating external information, they often are taylor-made for a specific forecast model, because they directly describe the deviation of that specific model when facing certain weather conditions. Thus, they cannot be easily applied to other forecast models or external information.

To increase the efficiency of multi-equation models, Taylor proposes to reduce the number of sub-models by also applying Principal Component Analysis (PCA) [22,23]. With his approach, Taylor only targets the relationship between different times of the day and the similarities between sub-models. Thus, components involved in the PCA describe the same variable. In contrast, we target the relationship between the main time series and multiple external influences. Thus, we refer to different influencing aspects and multiple variables. As a result, Taylor's approach is orthogonal to our solution and does not target the integration of external factors. Thus, we can combine both solutions to (1) optimize the base forecast models and (2) to optimize the integration of external information. This would lead to a very efficient and accurate forecasting solution.

Overall, our presented solution provides a universal framework to efficiently add external information to forecast models. With the help of our approach, even complex relationships involving a large number of external information can be considered in environments with special efficiency constraints.

6 Conclusion

In this paper, we introduced a framework for an efficient integration of external information into forecast models. We excluded the modeling of external information from the main forecast model and proposed to create a separate model that represents them. Due to the stable relationship between the external information and the main time series, it is possible to exclude the separate model from most model adaptations. In addition, for multi-equation models only a single external information model is needed for all involved sub-models. We also increased the efficiency of the external information model by applying feature selection and dimension reduction techniques. The proposed framework can be configured very flexibly regarding necessary accuracy and desired performance. Our experiments showed that with the help of our approach, the time for re-estimating a forecast model, while considering external influences, can be significantly reduced. At the same time, for some data sets we even improved the accuracy compared to the naïve direct integration. As a result, our approach enables a broad use of external influences in the face of efficiency constraints as well as evolving time series and thus, increases the accuracy when forecasting in real-time environments.

Acknowledgment. The work presented in this paper has been carried out in the MIRABEL project funded by the EU under the grant agreement number 248195.

Final edited form was published in "Advances on Databases and Information Systems: 16th East European Conference. Posen 2012", S. 139–152, ISBN 978-3-642-33074-2.
http://dx.doi.org/10.1007/978-3-642-33074-2_11

References

1. MIRABEL Project (2011), http://www.mirabel-project.eu
2. MeRegio Project (2011), http://www.meregio.de/en/
3. Chateau, B., Lapillonne, B.: Long-term energy demand forecasting a new approach. Energy Policy 6(2), 140–157 (1978)
4. Hor, C.L., Watson, S., Majithia, S.: Analyzing the impact of weather variables on monthly electricity demands. Power Systems 20(4), 2078–2085 (2005)
5. Ruźić, S., Vuckovic, A., Nikolic, N.: Weather sensitive method for short term load forecasting in electricpower utility of serbia. Power Systems 18(4), 1581–1586 (2003)
6. Ramanathan, R., Engle, R., Granger, C.W., Vahid-Araghi, F., Brace, C.: Short-run forecasts of electricity loads and peaks. International Journal of Forecasting 13(2), 161–174 (1997)
7. Spearman, C.: The proff and measurement of association between two thins. American Journal of Psychology 15, 72–101 (1904)
8. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2(6), 559–572 (1901)
9. Jolliffe, I.: Principal Component Analysis, 2nd edn. Springer Series in Statistics. Springer Verlag Inc. (2002)
10. Taylor, J.W.: Triple seasonal methods for short-term electricity demand forecasting. European Journal of Operational Research 204, 139–152 (2009)
11. Center for Renewable Energy Sources and Saving (2012), http://www.cres.gr/kape/index_eng.htm
12. Deutscher Wetterdienst (2012), http://www.dwd.de
13. Nelder, J., Mead, R.: A simplex method for function minimization. The Computer Journal 7(4), 308–313 (1965)
14. Akdere, M., Cetintemel, U., Upfal, E.: Database-support for continuous prediction queries over streaming data. In: VLDB 2010 (2010)
15. Parisi, F., Sliva, A., Subrahmanian, V.S.: Embedding Forecast Operators in Databases. In: Benferhat, S., Grant, J. (eds.) SUM 2011. LNCS, vol. 6929, pp. 373–386. Springer, Heidelberg (2011)
16. Duan, S., Babu, S.: Processing forecasting queries. In: VLDB 2007 (2007)
17. Ge, T., Zdonik, S.: A skip-list approach for efficiently processing forecasting queries. In: Proceeding of the VLDB 2008 (2008)
18. Dannecker, L., Schulze, R., Böhm, M., Lehner, W., Hackenbroich, G.: Context-Aware Parameter Estimation for Forecast Models in the Energy Domain. In: Bayard Cushing, J., French, J., Bowers, S. (eds.) SSDBM 2011. LNCS, vol. 6809, pp. 491–508. Springer, Heidelberg (2011)
19. Box, G.E.P., Jenkins, G.M., Reinsel, G.C.: Time Series Analysis: Forecasting and Control. John Wiley & Sons Inc. (1970)
20. Bruhns, A., Deurveilher, G., Roy, J.S.: A non-linear regression model for mid term load forecasting and improvements in seasonality. In: Proceedings of the 15th PSCC 2005 (2005)
21. Wi, Y.M., Kim, J.H., Sung-Kwan Joo, J.B.P., Oh, J.C.: Customer baseline load (cbl) calculation using exponential smoothingmodel with weather adjustment. In: Proceedings of the 2009 TDCE (2009)
22. Taylor, J.W., de Menezes, L.M., McSharry, P.E.: A comparison of univariate methods for forecasting electricity demand up to a day ahead. International Journal of Forecasting 22, 1–16 (2006)
23. Taylor, J.W., McSharry, P.E.: Short-term load forecasting methods: An evaluation based on european data. Power Systems 22, 2213–2219 (2007)