General Sales Forecast Models for Automobile Markets and their Analysis

Marco Hülsmann[1], Detlef Borscheid[2], Christoph M. Friedrich[1,3], and Dirk Reith[1,4]

[1] Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany
[2] BDW Automotive, Maybachstr. 35, 51381 Leverkusen, Germany
[3] Present address: University of Applied Science and Arts Dortmund; Department of Computer Science, Emil-Figge-Str. 42 /B.2.02, 44227 Dortmund, Germany
[4] Corresponding Author, Phone: +49-2241/142746, Fax: +49-2241/141328, E-mail: dirk.reith@scai.fraunhofer.de

Abstract. In this paper, various enhanced sales forecast methodologies and models for the automobile market are presented. The methods used deliver highly accurate predictions while maintaining the ability to explain the underlying model at the same time. The representation of the economic training data is discussed, as well as its effects on the newly registered automobiles to be predicted. The methodology mainly consists of time series analysis and classical Data Mining algorithms, whereas the data is composed of absolute and/or relative market-specific exogenous parameters on a yearly, quarterly, or monthly base. It can be concluded that the monthly forecasts were especially improved by this enhanced methodology using absolute, normalized exogenous parameters. Decision Trees are considered as the most suitable method in this case, being both accurate and explicable. The German and the US-American automobile market are presented for the evaluation of the forecast models.

Keywords: Sales Forecast, Time Series Analysis, Data Mining, Automobile Industry, Decision Trees.

1 Introduction

Strategic planning based on reliable forecasts is an essential key ingredient for a successful business management within a market-oriented company. This is especially true for the automobile industry, as it is one of the most important sectors in many countries. Reliable forecasts cannot only be based on intuitive economic guesses of the market development. Mathematical models are indispensable for the accuracy of the predictions as well as for the efficiency of their calculations, which is also supported by the increase of powerful computer resources.

The application of time series models to forecasts of the registrations of new vehicles was originally established by Lewandowski [1, 2] in the 1970s. Afterwards, a general equilibrium model for the automobile market concerning both new car sales and used car stocks was presented by Berkovec [3]. Thereby, equilibrium means that the demand equals the supply for every vehicle type. Later on, Dudenhöffer and Borscheid [4] published a very important application of time series methods to the German automobile market. However, the number of efforts undertaken in this field of research is quite small to date. Methods based on statistical learning theory [5] are powerful instruments to get insight into internal relationships within huge empirical datasets. Therefore, they are able to produce reliable and even highly accurate forecasts. However, Data Mining algorithms have become more and more complex over the last decades. In this work, the accuracy of the prediction has the same importance as the explicability of the model. Hence, only classical Data Mining methods [6] are applied here.

In a previous contribution [7], basic time series methods were used together with a trend estimation performed by *Multivariate Linear Regression (MLR)* or a *Support Vector Machine (SVM)* with a Gaussian kernel [5, 17]. The

associated models were able to produce reliable forecasts and at the same time easy to explain. However, in this work, even enhanced models are presented which increase both the accuracy and to some extent also the explicability. As in [7], the distinction between yearly, quarterly, and monthly economic data is made. Again, it turns out that quarterly data is the most suitable and stable collection of data points, although here, the focus lies on the improvement of monthly predictions. Due to the higher amount of data, the economic explicability of the model is best in the case of monthly data, which is shown in this work.
Both the German and US-American automobile market were considered. The limitations of the forecasts are mainly due to the poorness or lack of estimates for the market-specific special effects, which will be figured out as well.

2 Data and Workflow

2.1 Exogenous Parameters

Newly registered automobiles as well as exogenous indicators are considered for both the German and the US-American automobile market. In the case of the German market, all data were adopted from [7], which also holds for their units and sources. The latter were the Federal Statistical Office, the German Federal Bank, and BDW Automotive, whereas the new registrations were taken from the Federal Motor Transport Authority. Most of the published indicators were seasonally adjusted. The feature selection performed in [7] is not taken into account here, i.e. all ten exogenous parameters are considered. The reason for this is the fact that the parameter reduction consistently delivered worse results in the case of a non-linear model. For the quarterly model, all exogenous parameters were chosen to be relevant, i.e. no parameter reduction was made. As the non-linear model turned out to be superior to the linear one,

it was decided not to perform a feature selection in this work. The enhancements here are based on different approaches. However, it is not excluded that a feature selection could even improve the predictions of some of the Data Mining methods applied.

Again, the German market was chosen to be used for the assessment of the modeling algorithms. Thereby, all three data intervals, i.e. yearly, quarterly, and monthly data, were employed because the assessment also included the data representation. Also the units of the exogenous data were modified: In [7], there was a mixture of absolute parameters and relative deviations in relation to the previous period. On the first hand, this mixture makes the explicability of the model more difficult, and on the other hand, it intuitively makes more sense to use absolute values only. As an example, the gasoline prices may have a significant influence on the car sales only after having exceeded a certain threshold. This threshold may be recognized by the underlying model whenever absolute exogenous parameters are involved. Using relative deviations, this hidden information cannot be discovered at all. This heuristic consideration was the reason for a comparison between a model based on absolute values only and a model based on a mixture of absolute and relative values. Furthermore, it seemed to be interesting to study the effects of some economic indices. For the German market, both the DAX and IFO indices were taken. Their explanations are given in Table 1. Their units and data sources are given in Table 2.
In the case of the US-American market, nearly the same exogenous parameters as for the German market were taken because general economic descriptors like Gross Domestic Product, Personal Income, Unemployment and Interest Rate, Consumer and Gasoline Prices, as well as Private Consumption are also very important for the US-American market. The indices used here are the Dow

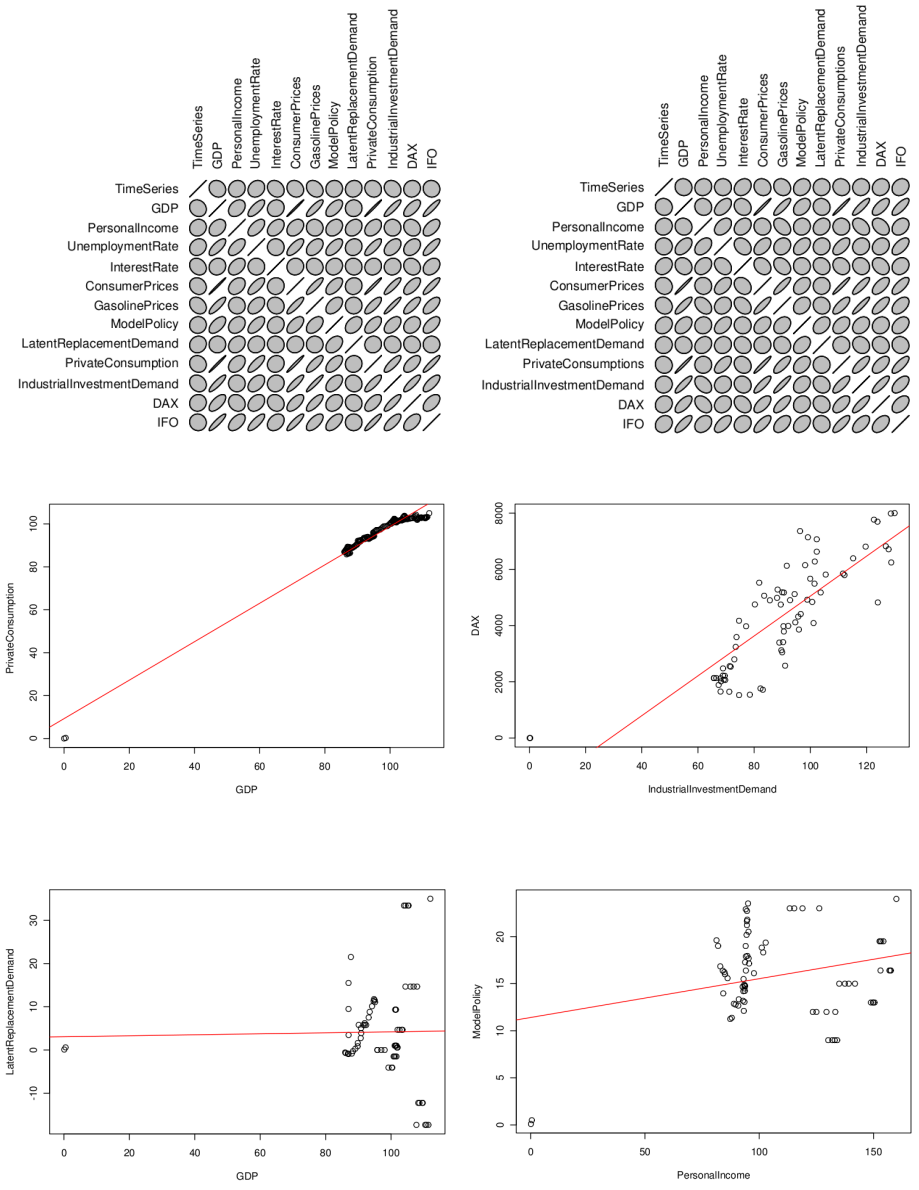

Fig. 1. Correlation analysis for the German market. The first two plots depict a correlation matrix of the exogenous parameters for quarterly and monthly data, respectively. The seasonally adjusted time series is included as well. The more a circle is transformed into an ellipse, the more correlation is present. A bias to the left indicates correlation and a bias to the right indicates anticorrelation. If the latter is present, it is only weak and not interpretable. The correlations in the case of quarterly data are stronger than in the case of monthly data but they are qualitatively equal. No correlations between the time series and the exogenous data are visible. The last four plots show some explicit examples for correlation and non-correlation (quarterly data): There is correlation between the Gross Domestic Product and the Private Consumption as well as between the Industrial Investmend Demand and the DAX. There is no correlation between the Gross Domestic Product and the Latent Replacement Demand as well as between the Personal Income and the Model Policy.

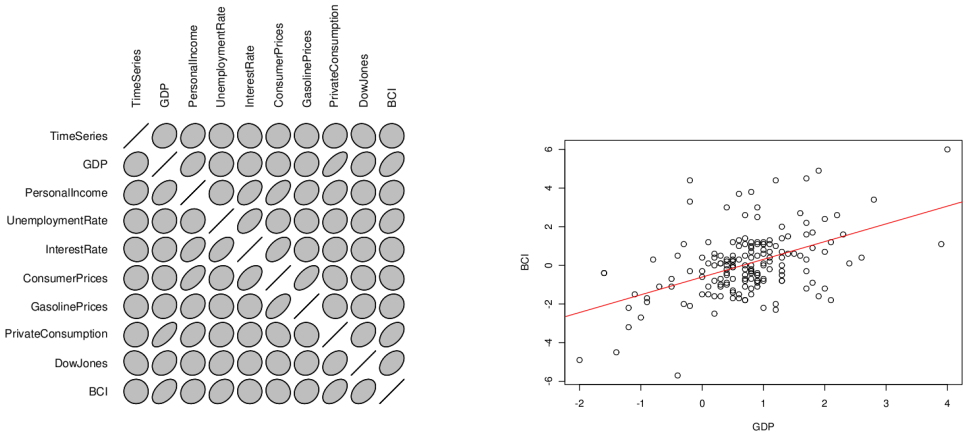

Fig. 2. Correlation analysis for the US-American market (quarterly data). Due to the lack of estimates for special effects and the presence of a mixture between absolute and relative data, only a few interpretable correlations are visible. An example is given by the correlation between the Gross Domestic Product and the Business Confidence Index (BCI).

Jones Industrial Average and the Business Confidence Index. Their units and data sources are given in Table 2.

2.2 Correlation Analysis

Figures 1 and 2 show some correlation plots for the exogenous parameters and the time series in the case of the German and the US-American market, respectively. The time series was seasonally adjusted in order to eliminate the spread within the data and hence to perform a correlation analysis on the trend component only. The more a circle within the correlation matrix is transformed into an ellipse, the more correlated are the corresponding two parameters. Thereby, a bias to the right represents correlation and a bias to the left represents anticorrelation.

It can be claimed that there are stronger correlations between quarterly data than between monthly data. In the former case, there are less data than in the latter case. Furthermore, quarterly are in most cases aggregated and smoothed monthly data. Hence, the spread within the data is much lower. However, the

correlations for quarterly and monthly data are qualitatively equal. There are only a few very weak anticorrelation but those are not interpretable, e.g. there is an ostensible anticorrelation between the Personal Income and the DAX as well as the Industrial Investment Demand. Furtheremore, no visible correlations between the time series and the exogenous data can be observed. There are only some ostensible anticorrelations, e.g. with the Gross Domestic Product and the Industrial Investmend Demand.

Moreover, Figure 1 depicts four explicit examples for quarterly data: First, the Gross Domestic Product is highly correlated with the Private Consumption and the DAX is highly correlated with the Industrial Investment Demand. Second, the Gross Domestic Product is not correlated with the Latent Replacement Demand and the Personal Income is not correlated with the Model Policy. The correlations are directly interpretable. In the case of the non-correlations, the following interpretation can be made: Latent Replacement Demand and Model Policy are specific parameters for the automobile industry. These do not have a direct influence on general indicators like the Gross Domestic Product or the DAX.

In the case of the US-American market (Figure 1), only the results for quarterly data are indicated. As described later, special random events within the market development should be estimated and eliminated from the time series. Due to the lack of such estimates for the US-American market and the fact that the exogenous data are a mixture of absolute parameters and deviation rates, there are no notable correlations to be mentioned. A correlation could be detected between the Gross Domestic Product and the Private Consumption, as for the German market, as well as between the Gross Domestic Product and the BCI. In both cases the data consists of deviation rates, so they are comparable.

2.3 Workflow

The workflow for the evaluation of the models based on the data listed above has been described in [7]. There are only three differences in this work:

1. No feature selection was performed for the reasons mentioned above.
2. The estimation of the calendar component in the case of monthly data was made before the estimation of the seasonal and trend components. The reason for this was that it seemed more reliable to estimate the seasonal component of a time series without calendar effects because otherwise, the seasonal component could be falsified. Hence, the calendar component was eliminated before.
3. No ARMA model [15] was built because it could be detected that the Data Mining algorithm used for the trend estimation had already included the ARMA component in the model. Hence, it did not make any sense to perform an additional ARMA estimation. The results were improved whenever the ARMA estimation was left out.

3 Methodology

The superior model is an additive time series model: If $x_t,\ t = 1, ..., L$, with L being the length of the observed time window used for training the model and $t \in \{1, ..., L\}$ the time period, are the new registrations of automobiles in the past, i.e. the main time series, then the equation

$$x_t = c_t + s_t + m_t + e_t,\ t = 1, ..., T,$$

holds, where c_t is the calendar component, s_t is the seasonal component, and m_t is the trend component, which have to be estimated in a reliable way. Please note that $\forall_{t=1,...,L}\ c_t = 0$ for yearly and quarterly data as well as

$\forall_{t=1,...,L}\ s_t = 0$ for yearly data. The last component e_t is the error component. The simple additive model turned out to be a good approximation of the reality [1, 2] and is easily interpretable at the same time.

3.1 Calendar Component Estimation

In the case of monthly data, the calendar component c_t is estimated as follows: Let W_t be the number of working days in a period t, $A_{i(t)}$ the average number of working days in all according periods (e.g. $i(t) \in \{1, ..., 12\}$ in the case of monthly data), and N_t the total number of days. Consider the coefficient

$$\lambda_t := \frac{W_t - A_{i(t)}}{N_t},\ t = 1, ..., L,$$

which is positive, whenever there are more working days in a period than on average, and negative, whenever there are less. Let $\bar{x}_t := s_t + m_t + e_t$ the calendar-adjusted time series. Then $c_t := \lambda_t \bar{x}_t$, and $\lambda_t > 0 \Leftrightarrow c_t > 0$. Hence,

$$\begin{aligned} x_t &= \bar{x}_t + c_t = \bar{x}_t + \lambda_t \bar{x}_t \\ \Rightarrow \bar{x}_t &= \frac{x_t}{1 + \lambda_t},\ c_t = \lambda_t \frac{x_t}{1 + \lambda_t}. \end{aligned}$$

3.2 Seasonal Component Estimation

Phase Average Method As described in [7], the phase average method [16] is a suitable way to estimate the seasonal component and at the same time easy to interpret. Thereby, as the underlying time series must be trendless, a univariate trend u_t has to be eliminated first, which is estimated by moving averages. It shall be pointed out again that the explicability of the model is of outmost interest. As it corresponds to one's intuition that periods which are situated too far away in the past or the future will not have a significant influence on the actual period, only the n nearest neighbors were included in

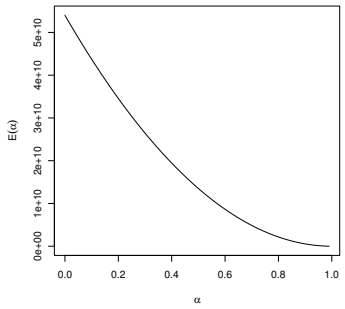

Fig. 3. Typical shape of a quadratic error function $E(\alpha)$ between a calendar-adjusted time series $\bar{x}_t$ and its univariate trend u_t estimated by exponential smoothing, as a function of the smoothing parameter α. The global minimum for $\alpha \in [0, 1]$ is reached at $\alpha = 1$ with $E(1) = 0$, since $\forall_{t=1,..,L}\ u_t = \bar{x}_t$, which is a completely overfitting univariate trend. As there is no local minimum in $(0, 1)$, the parameter α was manually adjusted in this work so that the Mean Average Percentage Error (MAPE) of the time series model applied to a test time series was as small as possible.

the average calculations. In this work, three different univariate moving averages were considered:

1. **P**ast **M**oving **A**verage (PMA), i.e. a moving average only considering periods of the past:
$$u_t^{\mathrm{PMA}} := \frac{1}{n} \sum_{i=0}^{n-1} \bar{x}_{t-i},\ n < t.$$

2. **C**lassical **M**oving **A**verage (CMA), i.e. a symmetric moving average considering both periods of the past and the future:
$$u_t^{\mathrm{CMA}} := \frac{1}{2n+1} \sum_{i=-n}^{n} \bar{x}_{t-i},\ n < \min(t, L-t+1).$$

3. **E**xponential **S**moothing **M**oving **A**verage (ESMA), i.e. a moving average based on an exponential smoothing formula only considering periods of the

past:

$$u_t^{\mathrm{ESMA}} := \alpha \sum_{i=0}^{n-2} (1-\alpha)^i \bar{x}_{t-i} + (1-\alpha)^{n-1} \bar{x}_{t-n+1},\ n < t,\ \alpha \in [0,1].$$

Actually, the smoothing parameter α is determined by minimizing the quadratic error function

$$E(\alpha) := \sum_{t=1}^{L} \left(u_t^{\mathrm{ESMA}} - \bar{x}_t\right)^2,$$

cf. [1]. Figure 3 shows a typical shape of such an error function for the present application: The global minimum is reached at $\alpha = 1$ with $E(\alpha) = 0$, which means that the trend overfits the time series completely, since $\forall_{t=1,..,L}\ u_t = \bar{x}_t$. As this is not desired and there is no local minimum in between, i.e. in the open interval $(0,1)$, the parameter α would have to be determined by cross-validation or bootstrapping so that the test error on a validation set is minimized. However, for this validation set, the real univariate trend would have to be available but it is not. Hence, all univariate trend parameters—the same holds for the size n of the time window—were adjusted manually so that the so-called **M**ean **A**verage **P**ercentage **E**rror (MAPE), which is an error estimating the quality of a prediction of time series values based on a complete time series model [7], was as small as possible.

Fourier Method If $\bar{x}_t$ is a periodic time series with period P, it can be expressed by the following discrete Fourier series:

$$\bar{x}_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j \cos(j\omega t) + \sum_{j=1}^{m} \beta_j \sin(j\omega t),$$

where $\omega = \frac{2\pi}{P}$ is the fundamental frequency of the Fourier series. The idea behind this is that the seasonal component can be expressed as a sum of cosine and sine oscillations of a certain frequency, if there is some periodicity in the time series. The $2m + 1 < L$ coefficients $\alpha_j,\ j = 0, ..., m$, and $\beta_j,\ j = 1, ..., m$, are determined by linear regression. In the case of quarterly data, $m = 2$ and $P = 4$, and in the case of monthly data, $m = 2$ and $P = 12$ are reasonable choices leading to good estimations of the seasonal component.

3.3 Trend Component Estimation

As it was assumed that the trend of the new car registrations were influenced by the exogenous parameters indicated in Tables 1 and 2, a multivariate trend model had to be created. The multivariate trend estimation was performed by Data Mining methods. The simplest ones considered here were linear models like Ordinary Least Squares (OLS) [18] and Quantile Regression (QR) [19]. However, more reliable algorithms were applied because they mostly performed significantly better without being too complex. It was decided to use a Support Vector Machine (SVM) with ϵ–regression and Gaussian kernel [5, 17], Decision Trees (DT) [20], k–Nearest Neighbor (KNN) [21], and Random Forest (RF) [22].

4 Results

4.1 Performance of Data Mining Methods

The predicted and real new car registrations of the German automobile market are plotted in Figure 4. The predictions result from the best performing Data Mining methods. The results of all Data Mining methods are indicated in Table 3. In the case of yearly data, the spread of the relative errors within the columns is the highest, when the test period was 2007 only. This is because the

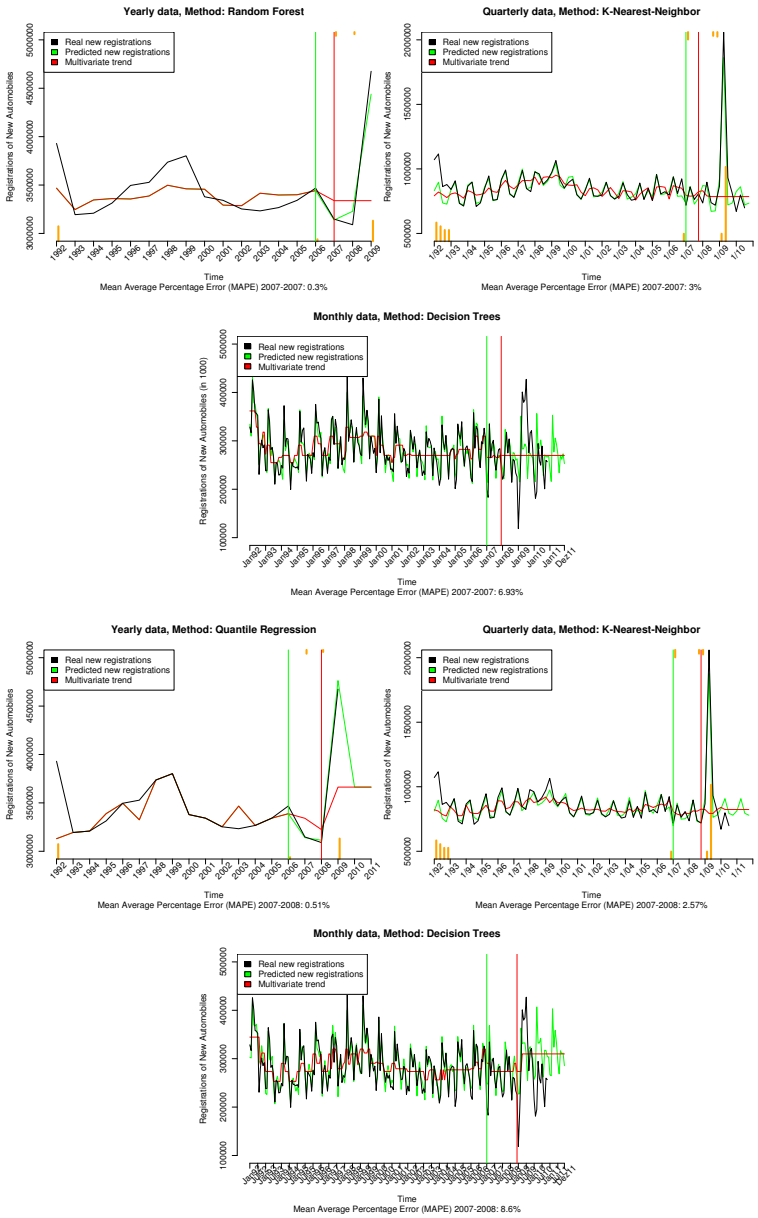

Fig. 4. Predictions for the German automobile market in comparison to real data using the same exogenous parameters as in [7], i.e. without DAX and IFO. In each plot, the results of the best performing Data Mining method are indicated (cf. Table 3) for yearly, quarterly, and monthly data. In the case of monthly data, only the results of DT are plotted, as this method turned out to be the most robust and explicable one for this kind of data. In all cases, the training period was 1992–2006. In the first three plots, the test set was 2007 only, and in the last three, it was 2007–2008. The test period is the interval enclosed by the green and red bars. The orange rugs indicate the amount of special effects. Rugs on the bottom stand for positive and rugs on the top stand for negative special effects. Please note that the first three models differ from the last three, as the exogenous parameters were updated by the FSO in 2008.

yearly MAPEs can be considered as completely random results, as the test data only consisted of one data point. For the last two test sets, QR turned out to be the best method. However, this was only the case for $\tau = 0.55$, q_τ being the τth quantile of the response values, i.e. the new car registrations in the training set. For $\tau \neq 0.55$, the results were much worse in comparison to the other methods. The yearly results of all applications and the quarterly results in the case of the first test period (2007, only four test points) can be considered as random results as well. From the other applications, it can be seen that the quarterly spreads are always lower than the monthly spreads within the columns, which indicates that quarterly data are the most stable data interval. This could already be concluded in [7] as well. In that publication, it was also discussed that the best results can be achieved in the case of yearly data (<1%), followed by quarterly data (2–3%) and monthly data (<10%). This can be confirmed again in this work.

The most suitable and robust Data Mining algorithms are SVM, DT, KNN, and RF, whereas OLS and QR mostly deliver poor results. This is because their underlying models are linear, which is not reliable for the present application [7]. It is natural that QR always performs better than OLS because there always exists a $\tau \in [0, 1]$ for which the τth quantile leads to a model with a smaller test error than the mean. One of the methods DT, KNN, and RF mostly outperformed the SVM, which was the only nonlinear method used in [7]. In the case of monthly data, DT turned out to be the most suitable method for two reasons: First, it delivered a MAPE which was significantly lower than 10%, except in the case of the third test period (2007–2009). Second, its application led to very reliable decision trees, which makes DT an exceedingly explicable method in the case of monthly data. The explicability of the algorithms will be discussed later.

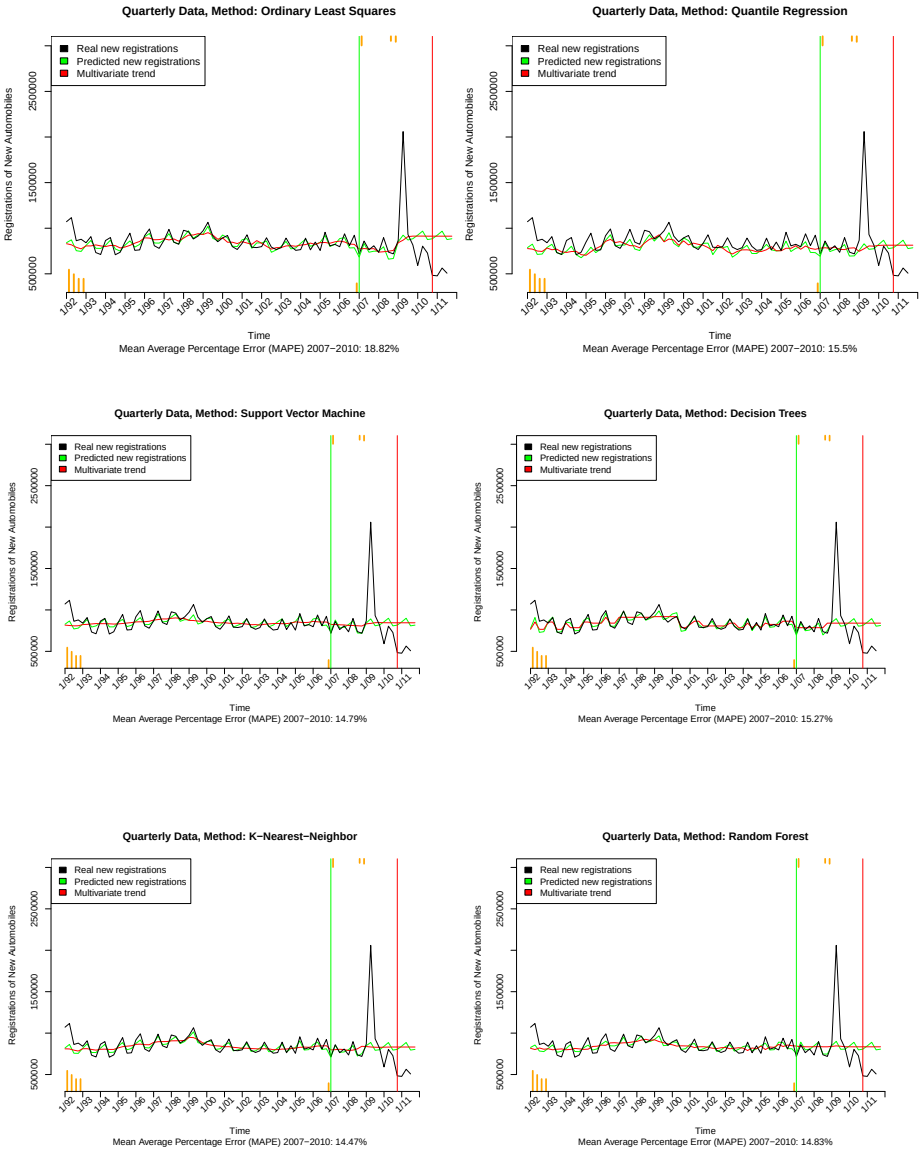

Fig. 5. Quarterly predictions for the German automobile market in comparison to real data using absolute exogenous parameters, including DAX and IFO. Each plot corresponds to a different Data Mining method. The test set was 2007–2010. None of the methods was able to predict the unusual behavior of the new car registrations after 2008.

The MAPEs in the case of the second test period (2007–2008) are similar to or higher than the ones in the case of the first test period (2007). This is because of the special effects in 2008 due to the financial crisis. In the last half of 2008, the number of new car registrations was decreased, which can only be predicted if the special effects are estimated and considered within the prediction. For quarterly data, the decrease in 2008 could be predicted by the most suitable methods DT, KNN, and RF. For KNN and DT, cf. Figure 4. For monthly data, there are no special effects as they are difficult to estimate on a monthly basis, and for yearly data, only the balance of quarterly special effects is taken into account. Hence, the decrease in 2008 can only be predicted at random in the case of yearly or monthly data.

The last three rows of Table 3 show the univariate trends used for the estimation of the seasonal component together with their specific parameters. It is interesting to see that in most cases, exactly the data of one quarter or one year were taken into account in order to calculate the moving averages. It was desisted from setting $n > 12$ for monthly and $n > 4$ for quarterly data because taking more data into account could lead to overfitting and reduce the explicability of the models. Please note that the phase average method was consistently used in all applications of this work, as the Fourier method did not deliver any noticeable improvements. For the comparison, the parameters for both methods were manually adjusted so that the MAPEs were as small as possible.

In 2009, the car-scrap bonus in Germany led to an enormous increase of new registrations of automobiles. This cannot be predicted by any forecast method. The results in the case of the third test period (2007–2010) are much worse than in the case of the first two. Figure 5 shows the limitations of such

economic forecasts for quarterly data: In the first quarter of 2009, a huge peak due to the car-scrap bonus was observed. Then, the number of newly registered cars decreased again because of the slow recovery after the financial crisis in 2008. The year 2010 was detected to be the worst year after the German reunification. In contrast, the years 2011 and 2012 are expected to feature an increase of new automobile registrations. The reasons for this are the high model policy and the recent positive economic development in 2011. However, none of the Data Mining methods applied could predict the unusual market behavior after 2008. All MAPE values were much greater than 10%. An accurate forecast would only be possible with reliable estimates for the respective special effects but these can only be calculated in retrospect. Pre-estimating them would equal reading the future from a crystal ball.

4.2 Stability Analysis

The box plots in Figure 6 show the MAPEs resulting from 50 statistically independent bootstrap replicates: Thereby, the training and test periods were merged into one data set. Then, this data set was divided randomly into a new training and a new test set. This procedure was repeated 50 times for two test periods, 2007 and 2007–2008. Mostly, the results indicated in the table lie within the lower whisker domain or are outliers showing that the time information is of very high importance here. The models must always learn from the past and cannot be based on random data points corresponding to random time periods. The stability analysis is divided into three parts, i.e. the *vertical*, the *horizontal*, and the *data set* analysis.

The vertical analysis considers the widths of the box plots, i.e. it compares the methods with respect to their robustness: The SVM turned out to be the most robust method followed by KNN, RF, and DT. Moreover, the box plots show

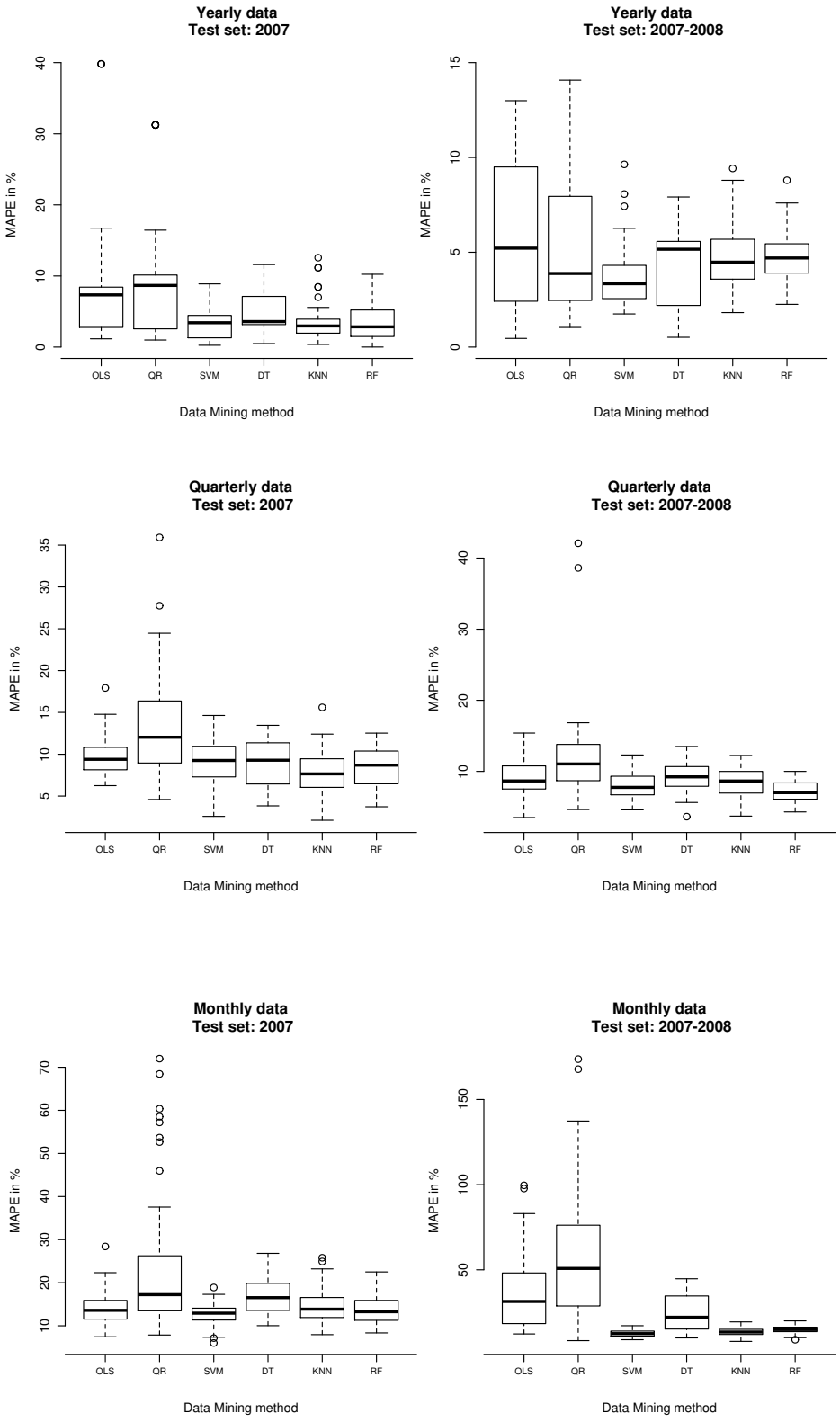

Fig. 6. Box Plots for each method resulting from 50 statistically independent bootstrap replicates. The training period was 1992–2006. The left and the right column shows box plots for the test period 2007 and 2007–2008, respectively. The most robust method is the SVM, followed by KNN and RF. Quarterly data are the most stable data interval, and in the case of the second test set (2007–2008), the results were less stable than in the case of the first test set due to the presence of special effects in 2008.

again that OLS and QR are not reliable for the present problem. Especially QR exhibits a large spread and a high number of outliers. The reason for the robustness of the SVM with respect to outliers and noise lies in its soft margin approach.

The horizontal analysis considers the position of the box plots, i.e. it compares the data intervals with respect to their stability regarding the behavior of the methods (except for QR for the reason mentioned above): Quarterly data turned out to be the most stable data interval for both test periods. All medians were situated approximately on the same level in this case, and the fewest outliers were detected. The reason for this is the fact that quarterly data are a compromise between a sufficiently high amount of data and a small spread within the data.

Finally, the data set analysis compares the results of the two test periods: Due to the different scales, a clear comparison cannot be derived from Figure 6. However, after a more detailed analysis, it could be recognized that the box plots were narrower in the case of the first period, leading to the interpretation that the second test set (2007–2008) is less stable due to the existence of special effects in 2008 and the absence of accurate estimates.

4.3 Absolute and Relative Exogenous Parameters

Table 4 shows the results using absolute exogenous parameters only instead of a mix of absolute and relative parameters. This time, all exogenous data indicated in Table 2 was taken in order to study the influence of the two German indices DAX and IFO. As the range of the absolute values of the indices differed exceedingly from the range of the other parameters, the data had to be scaled. The test period was 2007–2008. In all three applications, the SVM was the best method. In the case of yearly and quarterly data, no significant improvement compared to the results in Table 4 could be detected.

Using monthly data, all Data Mining methods delivered a MAPE smaller than 10%. Hence, the absolute data sets were easier to model for the algorithms, which is also explicable because of the motivation given in the section 2. The improvements were not caused by the incorporation of DAX and IFO, which only had a low impact on the predictions. The reason for this is the fact that they are highly correlated with the GDP and the Consumer Prices. Furthermore, the DAX is correlated with the Industrial Investment Demand and the IFO is correlated with the Private Consumption. However, they were taken because they both appeared in the decision trees in Figure 7.

4.4 Explicability of the Results

The algorithms used for the present application are standard Data Mining methods and hence do not hurt the requirement of explicability. The underlying models are understandable and descriptive. The most explicable methods is by far DT as besides delivering predictions for test data, the method also analyzes the training data and draws trees depicting the impact of the most important exogenous parameters. Figure 7 shows two of them, one for quarterly and one for monthly data. Thereby, the training set was 1992–2006 and the exogenous parameters were normalized absolute values including DAX and IFO. The root nodes are labeled with the most important parameters determined by the algorithm. In the case of monthly data, it is *Consumer Prices*. The tree indicates that the new registrations decrease with increasing consumer prices, which is meaningful. Most of the leaf nodes are explicable as well: The new registrations increase with decreasing gasoline prices, with increasing latent replacement demand, and with increasing GDP. In comparison to this, the decision tree for quarterly data is less explicable. The root note indicates that the highest number of new car registrations is

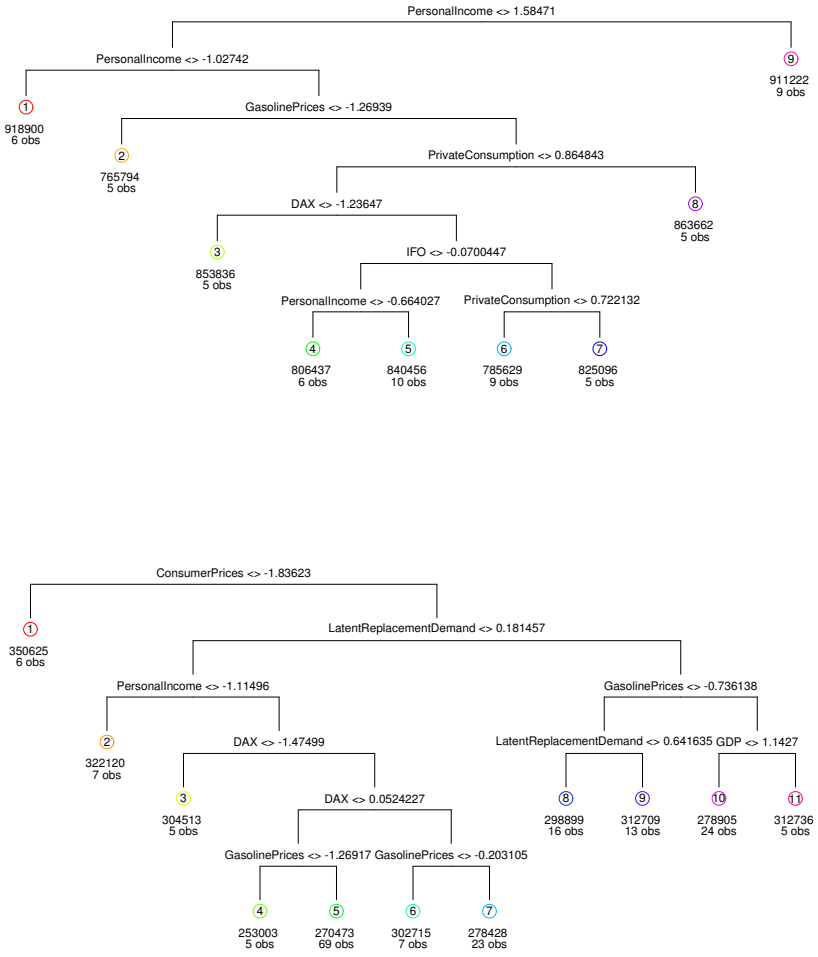

Fig. 7. Decision trees for the training set 1992–2006 using normalized absolute exogenous parameters including DAX and IFO for quarterly and monthly data. In the case of monthly data, the decision trees are more explicable than in the case of quarterly data. The leaves are labeled with the number of the class, the mean value of newly registrated automobiles of this class, and the number of observables (obs.) in the training set belonging to it.

achieved, when the personal income has a very low value, and when its value is higher, the number of new car registrations is lower. This does not make any sense. As motivated above, the usage of absolute parameters increases the explicability. In the case of monthly data, it also leads to meaningful decision trees. Furthermore, the amount of data is much lower for quarterly data, which leads to the fact that only little reasonable information can be extracted from the data. Hence, it can be concluded that the method DT together with normalized absolute exogenous parameters on a monthly data basis is the most reasonable choice in order to get explicable results. Please note that the numbers in Figure 7 are normalized values. They can easily be inverted so that interpretable thresholds can be achieved.

4.5 Application to the US-American Automobile Market

The forecast workflow was additionally applied to the US-American automobile market, where meaningful data were available for a longer training period than for the German market. For reasons of brevity, only quarterly data were taken here. The training set was 1970–2005, whereas the test set was 2006–2008. Unfortunately, no special effect estimates could be obtained, which made the modeling procedure much more difficult. The last quarter of 2008 was not included in the test set because of the financial crisis, whose occurrence and impact could not be foreseen. However, the principal difficulties to build reliable models were due to the lack of estimates for the special effects in the past, like the Vietnam War lasting until the early 1970s, the oil crisis in 1973, the economic booms in 1972/73 and from 1977 to 1979, the energy crisis of 1979, the internet bubble burst in 2000, the aftermath of September 11th, the financial crisis in 2008, as well as the US-American scrappage program *Car*

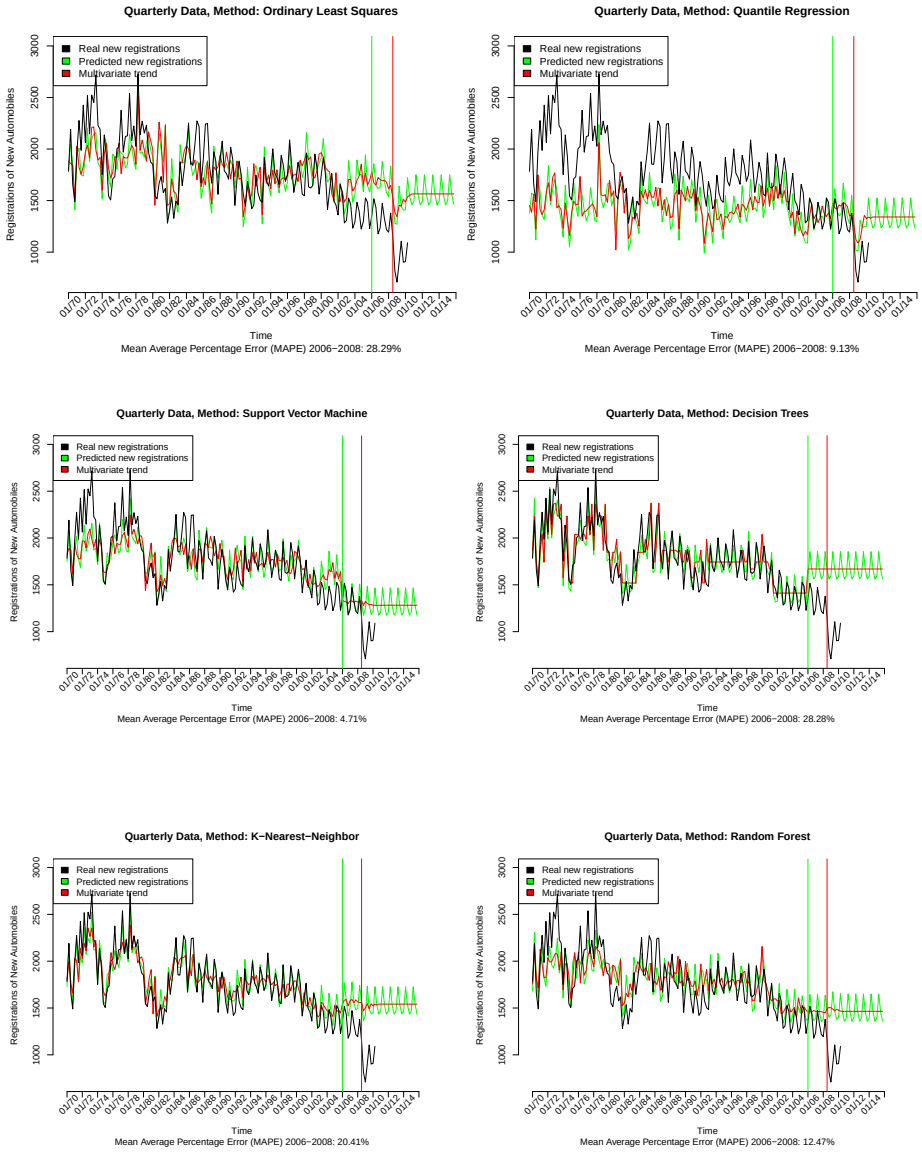

Fig. 8. Predictions for the US-American automobile market in comparison to real data using the exogenous parameters indicated in Table 2 for quarterly data. The training period was 1970–2005 and the test period was 2006–2008, where the last quarter of 2008 was omitted. Only the SVM could reproduce the collective multivariate trend of the time series in a proper way. All other methods predicted an increasing trend after 2002 remaining until 2008. The univariate average for the seasonal component was PMA with $n = 4$.

Allowance Rebate System (CARS) after July 2009. Figure 8 shows the results of two methods applied to the US-American market. The predictions were based on the exogenous data indicated in Table 2. Only the SVM was capable to detect the special effects mentioned above as outliers, which can be seen from the course of the multivariate trend. Intuitively, the trend should go up after 2002, following the shape of the time series from 1970 to 2005, which was predicted by all methods, also by the SVM. Owing to the robustness of the SVM with respect to outliers, the collective decreasing trend of the time series could be reproduced correctly leading to the low MAPE of 4.71% for the test period. All other methods overfitted the training data except Quantile Regression:

Interestingly, it was the only method which could detect the decreasing sales due to the last crisis and the increasing sales due to the subsequent boom in 2010. This could be achieved $\tau = 0.05$ corresponding to the 5% quantile of the training data. However, the training error was very high and the model built over the 1970–2005 was absolutely bad because of this choice of τ. Hence, the small validation error was only achieved at random. This example shows again that good estimates for the special effects are indispensable for reliable time series models. Forecasting the actual market situation of the United States would even be more problematic for any mathematical modeling algorithm: Recently, there were huge attacks on the US-American market by German automobile companies like Volkswagen and Audi. Positive effects occur due to the high replacement demand in 2011 and negative effects due to the recent debt crisis.

5 Conclusions

In this work, the performance and limitations of general sales forecast models for automobile markets based on time series analysis and Data Mining techniques were presented. The models were applied to the German and the US-American automobile markets. As in a recent work [7], the Support Vector Machine turned out to be a very reliable method due to its non-linearity. In contrast, linear methods like Ordinary Least Squares or Quantile Regression are not suitable for the present forecasting workflow. Owing to some modifications concerning the time series analysis procedure including the estimation of the calendar and seasonal components, the results of [7] could even be improved. However, other Data Mining methods like Decision Trees, K-Nearest-Neighbor, and Random Forest were considered leading to similar and in some cases even better results. Using absolute exogenous data instead of a mixture of absolute and relative data in the case of monthly data, the prediction errors of all suitable Data Mining methods were less than 10%, which was another enhancement. The most explicable method was the Decision Trees, which delivered meaningful models using absolute monthly exogenous parameters. In the case of monthly data, this method turned out to be the most reliable and explicable one. As in [7], quarterly data were the most stable ones. As expected, the Support Vector Machine was the most robust method, also with respect to outliers, i.e. special effects. However, useful and accurate predictions for the future cannot be achieved without reliable estimates of special effects, which could particularly be detected in the case of the German car-scrap bonus and the irregular behavior of the US-American market. Generally, it would be possible to use methods for outlier and noise detection in order to get reliable estimates for the special effects in the past. However, in

most cases, special effects occuring in the future cannot be predicted at all so that the quality of the forecasts is always limited.

References

1. Lewandowski, R.: Prognose- und Informationssysteme und ihre Anwendungen. de Gruyter, Berlin – New York (1974)
2. Lewandowski, R.: Prognose- und Informationssysteme und ihre Anwendungen, Band II. de Gruyter, Berlin – New York (1980)
3. Berkovec, J.: New Car Sales and Used Car Stocks: A Model for the Automobile Market, The RAND Journal of Economics **26** (1985) 195–214
4. Dudenhöffer, F., Borscheid, D.: Automobilmarkt-Prognosen: Modelle und Methoden. In: Ebel, B., Hofer, M. B., Al-Sibai, J. (eds.): Automotive Management – Strategie und Marketing in der Automobilwirtschaft, Springer Berlin – Heidelberg (2004) 192–202
5. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Berlin – Heidelberg – New York (1995)
6. Witten, I. H., Frank, E.: Data Mining. Morgan Kaufmann Publishers, San Francisco (2005)
7. Brühl, B., Hülsmann, M., Borscheid, D., Friedrich, C. M., Reith, D.: A Sales Forecast Model for the German Automobile Market Based on Time Series Analysis and Data Mining Methods. In: Perner, P. (ed.): Advances in Data Mining: Applications and Theoretical aspects, Springer, Berlin (2009) 146–160 (Lecture Notes in Computer Science 5633: Lecture Notes in Artificial Intelligence), Proceedings of the 9th Industrial Conference on Data Mining (ICDM) 2009, Leipzig, Germany
8. The CESifo GmbH, http://www.cesifo-group.de
9. Organisation for Economic Cooperation and Development (OECD), http://stats.oecd.org
10. Organisation for Economic Cooperation and Development (OECD), Reference Series Vol. 2010, ESDS International, (Mimas) University of Manchester.

11. Organisation for Economic Cooperation and Development (OECD), Key Economic Indicators (KEI) Vol. 2010, ESDS International, (Mimas) University of Manchester.
12. Organisation for Economic Cooperation and Development (OECD), Financial Indicators, subset of Main Economic Indicators (MEI) Vol. 2010, ESDS International, (Mimas) University of Manchester.
13. Yahoo! Finance, http://finance.yahoo.com
14. Organisation for Economic Cooperation and Development (OECD), The Crisis and Beyond Vol. 2010, ESDS International, (Mimas) University of Manchester.
15. Box, G. E. P., Jenkins, G. M.: Time Series Analysis Forecasting and Control. Holden–Day, San Francisco (1976)
16. Leiner, B., Einführung in die Zeitreihenanalyse, R. Oldenbourg Verlag, München–Wien (1982)
17. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge (2002)
18. Chambers, J. M., Hastie, T. J.: Statistical Models in S, CRC Press, Boca Raton (1991)
19. Koenker, R. W., Quantile Regression, Cambridge University Press, Cambridge (2005)
20. Breiman L., Friedman J. H., Olshen R. A., Stone, C. J.: Classification and Regression Trees, CRC Press, Boca Raton (1984)
21. Hechenbichler K., Schliep K. P.: Weighted k-Nearest-Neighbor Techniques and Ordinal Classification, Discussion Paper 399, SFB 386, Ludwig–Maximilians University Munich (2004)
22. Breiman, L.: Random Forests. In: Machine Learning **45** (2001) 5–32

Table 1. Explanation of the economic indices used as exogenous data for the models in this work. In the case of the German automobile market, the DAX and IFO indices were chosen, and in the case of the US-American market, the Dow Jones and BCI indices were taken.

Country	**Index**	**Explanation**
Germany	DAX	most important German stock index reflecting the development of the 30 biggest and top-selling companies listed at the Stock Exchange in Frankfurt (so-called *blue chips*), published as performance or exchange rate index; in this work, the performance index was taken meaning that all dividends and bonuses of the stocks are directly reinvested; the abbreviation DAX comes from the German name **D***eutscher* **A***ktieninde***X**
	IFO	business climate index published monthly by the German Institute for Economic Research (IFO), known as an early indicator for the economic development in Germany
USA	Dow Jones	actually *Dow Jones Industrial Average (DJIA)*, known as *Dow Jones Index* in Europe, created by the Wall Street Journal and the company of Charles Dow and Edward Jones, most important US-American stock index reflecting the development of the 30 biggest and top-selling companies listed at the New York Stock Exchange (NYSE), analog to the German DAX
	BCI	**B**usiness **C**onfidence **I**ndex measuring the level of optimism that people who run companies have about the performance of the economy and how they feel about the prospects of their organizations, comparable to the German IFO

Table 2. Data units and sources for all exogenous parameters used in this work. The units and sources for the German exogenous parameters (except for DAX and IFO) are listed in [7]. The three data sources were the Federal Statistical Office (FSO), the German Federal Bank (GFB), and BDW Automotive. Please note that in the case of the US-American market, only quarterly data were taken. Here, the main data sources were the Bureau of Economic Analysis (BEA), the Bureau of Labor Statistics (BLS), and the Organisation for Economic Cooperation and Development (OECD) database [9]. If only the term *deviation rates* is indicated, this refers to the previous quarter. Title and ownership of the data remain with OECD.

Country	**Parameter**	**Data Unit and Source**
Germany	DAX	monthly: indices (1987=1000), dataset from the GFB quarterly: deviation rates of monthly averages yearly: deviation rates of monthly averages
	IFO	monthly: indices (2000=100), dataset from the CESifo GmbH [8] quarterly: deviation rates of monthly averages yearly: deviation rates of monthly averages
USA	New Car Registrations	in thousands, dataset from the BEA
	Gross Domestic Product	deviation rates, OECD [10]
	Personal Income	billions of chained 2000 dollars, dataset from the BEA
	Unemployment Rate	in % of the total population, OECD [11]
	Interest Rate	in %, OECD [12],
	Consumer Prices	deviation rates of monthly averages (price indices), dataset from the BLS
	Gasoline Prices	deviation rates of monthly averages (price indices), dataset from the BLS
	Private Consumption	deviation rates, OECD [11]
	Dow Jones	deviation rates of monthly averages (index points), dataset from Yahoo! Finance [13]
	BCI	deviation rates of monthly averages (indices, 1985=100), OECD [14]

Table 3. Yearly, quarterly, and monthly MAPEs in % between the predicted and real new car registrations of the German automobile market for all Data Mining methods and test periods. The results of the best methods are plotted in Figure 4 for each of the first six applications. In the case of RF, the average values of ten statistically independent replicates are indicated, with the standard deviations in parentheses. In the last rows, the univariate trends for the seasonal component estimation together with their specific parameters are shown.

Test period			2007			2007–2008			2007–2009
Method/Data	**Y**	**Q**	**M**	**Y**	**Q**	**M**	**Y**	**Q**	**M**
OLS	16.73	17.81	8.41	8.12	8.23	9.85	7.93	10.88	12.66
QR	16.45	9.07	7.74	**0.51**	6.08	**7.95**	**0.96**	7.12	**11.40**
SVM	1.75	3.66	7.33	1.86	3.60	12.84	3.15	5.04	16.72
DT	4.5	3.25	**6.93**	2.89	3.56	**8.60**	4.32	**4.83**	**13.22**
KNN	0.37	**3.00**	8.36	1.65	**2.57**	18.18	2.70	**4.83**	20.70
RF	**0.23**	3.82	12.70	2.50	2.99	17.80	2.74	4.77	20.94
	(0.15)	(0.09)	(1.56)	(0.26)	(0.08)	(0.93)	(0.20)	(0.04)	(0.76)
Univariate	–	PMA	ESMA	–	ESMA	CMA	–	ESMA	CMA
Trend		$n = 3$	$n = 12$		$n = 4$	$n = 4$		$n = 4$	$n = 4$
			$\alpha = 0.1$		$\alpha = 0.3$			$\alpha = 0.3$	

Table 4. Yearly, quarterly, and monthly MAPEs in % between the predicted and real new car registrations of the German automobile market for all Data Mining methods using absolute exogenous parameters. The test period was 2007–2008. In the case of RF, the average values of ten statistically independent replicates are indicated, with the standard deviations in parentheses. In the last rows, the univariate trends for the seasonal component estimation together with their specific parameters are shown. The SVM was the best method in all three applications.

Method/Data	**Y**	**Q**	**M**
OLS	6.15	6.16	9.37
QR	**1.82**	**3.37**	**6.73**
SVM	4.99	4.61	7.65
DT	8.08	4.75	8.66
KNN	1.95	3.71	9.7
RF	2.94	4.02	8.64
	(0.14)	(0.12)	(0.33)
Univariate Trend	–	ESMA $n = 4$ $\alpha = 0.5$	PMA $n = 12$