Robust Long Term Forecasting of Seasonal Time Series

Jörg Wichard[1] and Christian Merkwirth[2]

[1] Institute of Molecular Pharmacology,
Robert-Rössle-Str. 10, D-13125 Berlin, Germany
JoergWichard@web.de,
[2] Applied Computer Science Department, Jagiellonian University
Reymonta 4, PL 30-059 Kraków, Poland
ChristianMerkwirth@web.de

Abstract. We propose the usage of a simple difference equation for predicting seasonal, trended time series with clear periodicity. By computing several forecasts for different settings of the method's single free parameter we obtain an ensemble of forecasts. These ensemble is combined to the final forecast by taking the samplewise median of those forecasts that were generated by models showing low prediction errors on left-out parts of the time-series. We show the application of this approach to the Mauna Loa atmospheric carbon dioxide concentration (ACDC) time series.

1 Introduction

Time series forecasting is a growing field of interest with applications in nearly any field of science. Since ancient times the prediction of the positions of sun, moon and the beginning of the seasons were of great importance for the early civilizations. The attempts to build proper forecasting models in Astronomy gave birth to the modern science when Newton developed his celestial mechanics and the Calculus. Nowadays, there is a huge interest to build long term market models in economics or climatic models that can give indication of global changes.

In many cases we have only the measurement of one system variable over a longer period, but no other quantitative information of variables that may influence the system of interest. In these cases we can only use the time series itself to build a predictive model. This approach was introduced by Yule, who described a linear (autoregressive) model based on a measured time series in order to predict the sunspot cycle [1]. Today the stochastic autoregressive process modelling performed by Yule belongs to the classics of linear time series analysis that is embedded in the theory of linear stochastic processes. In general the conventional linear methods for modeling and prediction fail if they are applied to time series originating from nonlinear systems. This seems to be evident because a nonlinear (deterministic) system should not be treated as a linear stochastic process.

Since the discovery of deterministic chaos, many methods for nonlinear time series modeling and prediction have been suggested and refined. A common characteristic of these models is the reconstruction of the systems state space based on the embedding theorems given by Takens [2], Sauer et al. [3] and Stark [4]. From a equally sampled time series $\{y_t\}_{t=1,\ldots,N}$ we can construct the d-dimensional state space vector

$\boldsymbol{y}_n = (y_{(n-\lambda(d-1))}, y_{(n-\lambda(d-2))}, \ldots, y_n)$ where λ denotes the time lag. We consider a model $f(\boldsymbol{y})$ for time series prediction of the form

$$\begin{aligned} f &: \mathbf{R}^d \rightarrow \mathbf{R} \\ f(\boldsymbol{y}_n) &= y_{n+\mu} \end{aligned} \tag{1}$$

where μ is the time horizon of the prediction. In the case $\mu = 1$ this is called the *one step ahead prediction* and it is the common choice in the case of iterated predictions, that means the predicted value y_{n+1} is used to construct the next state space vector $\boldsymbol{y}_{n+1}$ which is used to predict the next time series sample y_{n+2} and so on.
The difficulties of long term prediction arise from the observation, that the uncertainty increases with the horizon of prediction. This was already known before the discovery of chaotic systems, where long term predictions are in general not possible. In the case of seasonal data, we can make use of our knowledge concerning the period of the system under investigation. In the following sections, we like to introduce a simple forecasting strategy based on the periodic exponential smoothing of differences.

2 Forecasting Strategy

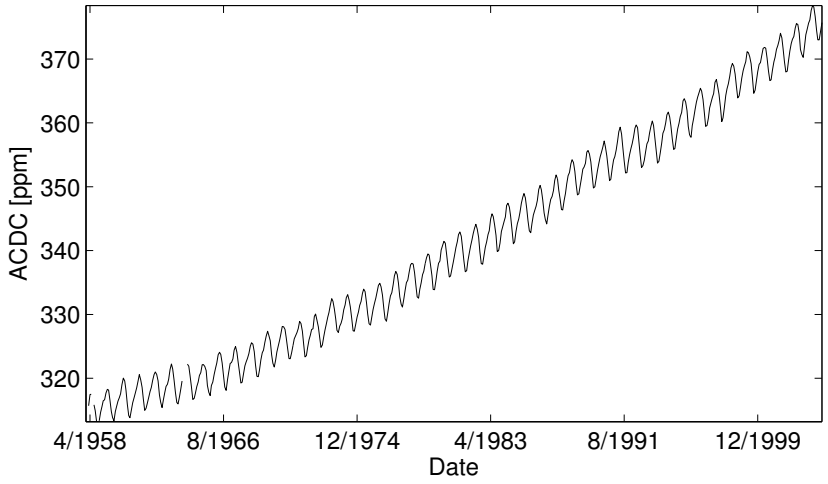
We describe a forecasting strategy that consists of two parts. The first part is a general algorithm for time-series prediction that has one or more free parameters to be optimized in order to generate a proper prediction. The second part of the strategy is a combination approach that validates the individual forecast for several parameter settings and combines these individual forecast into the final forecast, thereby relying on the validation errors to select suitable model parameters.

2.1 Periodic exponential smoothing of differences

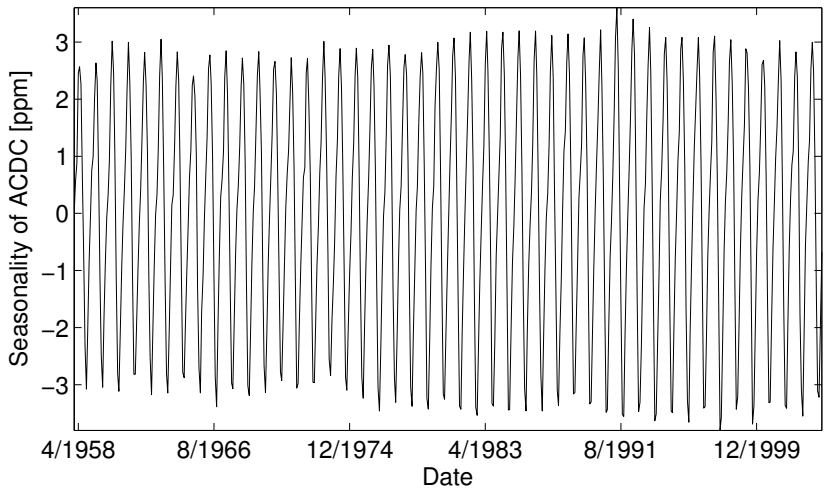
We propose the usage of a simple periodic exponential smoothing difference equation to compute predictions based on the observed time series y_t:

$$y_{t+1} = y_t + \frac{1}{\sum_{i=1} \alpha^{i-1}} \sum_{i=1} \alpha^{i-1}(y_{t+1-ip} - y_{t-ip}) \tag{2}$$

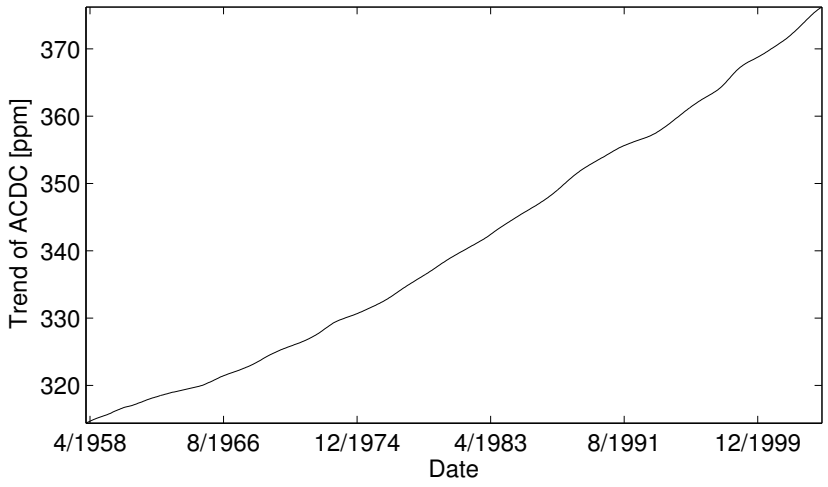
The equation determines the difference between the actual sample y_t and the one step ahead forecast y_{t+1} by computing the weighted sum over all differences between pairs of consecutive samples $p, 2p, 3p, \ldots$ from the past. The period length p is determined by the type and the seasonality of the time series, for monthly sampled time series p is usually 12. The weighting factors α^{i-1} are exponentially decreasing, thus giving recent observations relatively more weight than the older observations. Please note that after performing p such one step ahead prediction steps, Equ. 2 sums up already forecasted sample values and thus can be seen as an iterated forecasting method. Missing values and the corresponding weighting factor are simply omitted from above sum, making this forecasting model applicable to time series with a few missing values. Though the method is only applicable to data with a clear periodicity, it has the advantage of being robust and has only a single *topological* parameter $\alpha \in]0, 1]$ that has be to be externally



(a) Monthly levels of Atmospheric Carbon Dioxide Concentration in ppm



(b) Seasonal part of the above ACDC time series



(c) Trend part of the above ACDC time series

Fig. 1. Monthly atmospheric carbon dioxide concentration in ppm as measured by Keeling, Bacastow and Whorf on the Mauna Loa. The time series covers 46 consecutive years from January 1958 to December 2003. A few missing values occur close to the beginning of the measurement period. The time series was decomposed into the seasonal and the trend part by means of non-stationary harmonic regression with a residual RMS of 0.1625. Both seasonal and trend part are not stationary. The seasonal part shows variations of the amplitude of the yearly oscillations while the slope of the trend part varies and is in average higher at the end of the time period than in the beginning.

specified. The next section details a method that reliefs the user from the necessity to specify a single, optimal α. Instead, a range of putative values for α has to be given which is usually much easier to infer from the data set of interest.

2.2 Combining forecasts

In order to generate the final forecast, we propose the following strategy:

1. Leave out a certain fraction L of samples from the end of the time series and keep those as a test set.
2. Compute forecasts $\hat{y}_{T+1}(\alpha_j), \ldots, \hat{y}_{T+M}(\alpha_j)$ for a range of different parameter settings $\alpha_j, j = 1, \ldots, J$.
3. Compute the forecasting error $E_L(\alpha_j)$ on the left-out test set for each setting α_j.
4. Repeat this for different numbers of left out samples $L_1, L_2, \ldots, L_N$.
5. Average the forecasting error for each setting of α_j over all settings of L_i:

$$E(\alpha_j) = 1/N \sum_{i=1}^{N} E_{L_i}(\alpha_j). \tag{3}$$

6. Forecast the full time series using the parameter settings α_j.
7. Compute the combined forecast $\hat{y}_{T+1}, \ldots, \hat{y}_{T+M}$ by taking the samplewise median over all forecasts whose errors $E(\alpha_j)$ are lower than the median of all errors $\bar{E} = \text{median}_{j=1,\ldots,J} E(\alpha_j)$

$$\hat{y}_{T+m} = \text{median}_{j:E(\alpha_j)\leq\bar{E}}\ \hat{y}_{T+m}(\alpha_j). \tag{4}$$

The choice of the median value instead of the arithmetic average increases the robustness of the approach. Combining the output of several forecasts has the additional advantage of easily deriving lower and upper bounds by computing the samplewise lower and upper quartile of the forecasts.

3 Application

3.1 Atmospheric carbon dioxide concentration time series (ACDC)

The level of ACDC has been closely measured and documented for more than 45 years. In this study we use the famous Mauna Loa[3] monthly time series, displayed in Figure 1. The details regarding the measurement of the ACDC level can be found in Keeling et al. [5]. Atmospheric carbon dioxide concentration (ACDC) is a crucial variable for many environmental simulation models. Whether or not rising atmospheric carbon dioxide levels have an impact on the global climatic system is still under investigation and beyond the scope of this study. The time series is an ideal test case for long term forecasting. Despite its apparent simple structure neither the trend is linear nor is the seasonality strictly periodic. An advantage of this time series besides the sufficient number of samples is that new measurements are available on a regular basis, thus allowing to truly validate forecasting results.

[3] The time series data with additional information and updates are available under http://cdiac.esd.ornl.gov/trends/co2/sio-mlo.htm .

3.2 Results

We produced forecasts for all possible combinations of the 50 settings of parameter $\alpha_j = 1-10^{-8+0.1592(j-1)},\ j = 1,\ldots,50$ and the number of left out samples L chosen out of {6,12,28,56,138}. The the range for the values of $(1-\lambda)$ is exponentially spaced between $1e^{-8}$ and $1e^{-0.2}$ in order to cover a broad range of possible settings for λ. However, the proposed ensembling method selects those parameters that perform worse than the median out, which makes the choice of the primary range for the parameter λ highly uncritical. By varying the number of left-out samples from 6 to 138, we intend to access the method's short, mid and long term forecasting performance. Forecasting errors $E_L(\alpha_j)$ were computed as described in the previous section. For several settings of the number of left-out samples, we computed forecasts covering only half of the length of the left-out part in order to avoid giving the samples at the end of the time series a too high influence on the computed errors (see Table 1).

Despite of the simplicity of the method, the forecasting errors displayed for periods up to 11.5 years do not exceed 2.0 in case of the RMS and 0.5% in case of the MAPE[4]. However, since the errors of the individual forecasts on the left-out part were used to determine which models were taken into the combined forecast, the validation errors displayed in Table 1 might be to optimistic. We validated this by applying the proposed method on the ACDC time series from which the last five years (1999-2003) where omitted right at the beginning of the whole procedure. Both the RMS error of the combined forecast with 0.80 and the MAPE of 0.18% agree readily with the results displayed in Table 1. These validated results are printed there in bold face.

Chan and McAleer [6] present a study in which they employ variations of the generalized autoregressive conditional heteroscedasticity (GARCH) model that account for the dynamics of the conditional variance to forecast the ACDC time series. Their two best performing models achieve a RMS of 0.680 and 0.458 and a MAPE of 0.135% and 0.101% on the left-out period from January 2002 to December 2002. The method described in this article achieves for the same left-out period a RMS of 0.482 and a MAPE of 0.102 without taking into account time-series values from January 2003 to December 2004 which were not available at the time Chan et al. conducted their study.

The forecasted atmospheric carbon dioxide levels are detailed for the next 4 years in Table 2. No data for the year 2004 was available at the time this study was conducted. We thing it is a good way of allowing the reader to objectively validate the quality of the methods by publishing forecasts for forthcoming periods. New time series values will be released yearly, each update covers the 12 month of the previous calendar year. This allows the reader to compute successively the respective errors for a forecast length of 12, 24, 36 and 48 month. We expect the first update covering the values of 2004 to be released before the conference takes place.

According to the long term prediction as displayed in Figure 2, an atmospheric carbon dioxide level of 400 ppm should not be exceeded within the forecast period of 138 months starting from January 2004. However, accelerated economic growth of large parts of Asia could lead to an earlier crossing of this threshold. On the other hand

[4] The **M**ean **A**bsolute **P**ercentage **E**rror is a common error measure in the field of econometrics. It is defined as MAPE$= 1/T \sum_{t=1}^{T} \frac{|\hat{y}_t - y_t|}{y_t}$ and can only be applied to strictly positive time series.

Left out samples	6	12	12	28	28	56	56	138	**60**
Forecast length	6	6	12	12	28	28	56	138	**60**
RMS	0.18	0.23	0.32	0.72	1.63	0.77	0.83	1.94	**0.8**
MAPE [%]	0.04	0.05	0.07	0.18	0.39	0.19	0.20	0.42	**0.18**

Table 1. RMS and MAPE of the combined forecast for different numbers of left-out samples and forecast lengths. The values displayed in all but the last row are slightly optimistic since the forecasts used to compute the final, combined forecast where selected according to their error on the left-out part. The errors given in the last column were obtained by leaving out 60 samples (years 1999 to 2003) from the end of the time series right at the beginning of the whole procedure and can serve thus as validated errors. They agree seamlessly with the other errors in this Table.

the amount of fossil fuels conveyed at present seems to be close to an upper limit of the capacity which makes a sudden strong increase of the anthropogenic CO_2 emission in the near future less likely. Another source of forecasting error could be a decreasing capacity of natural carbon dioxide sinks, however, the prediction of these influences is surely out of the scope of the proposed method.

Conclusion and Outlook

The proposed method delivers excellent forecasting performance when validating it by leaving out several month or even years from the end of the time series. The results are comparable with more sophisticated forecasting approaches. We consider the low number of free parameters of the proposed method not as a disadvantage, instead we believe this to be beneficial for the purpose of long term forecasting. A main drawback of the method is the need for a clear periodicity. However, the method consists of two parts, a simple forecasting algorithm and a method for combining several forecasts to a final forecast. We are optimistic that the latter method can be applied successfully to different forecasting methods that allow the treatment of more complex time series. The method of combined forecasts could also be applied to forecasts generated by different forecasting algorithms.

References

1. Yule, U.: On a method of investigating periodicities in disturbed series with special reference to wolfer's sunspot numbers. Philos. Trans. Roy. Soc. Series A **226** (1927) 267–298
2. Takens, F.: Detecting strange attractors in turbulence. In: Dynamical Systems and Turbulence. Lect. Notes Math. Sringer-Verlag, Berlin (1981)
3. Sauer, T., Yorke, J., Casdagli, M.: Embedology. J.Stat.Phys. **65** (1991) 579–618
4. Stark, J., Broomhead, D., Davies, M., Huke, J.: Takens embedding theorems for forced and stochastic systems. Nonlinear Analysis **30** (1997) 5303–5314
5. Keeling, C., Bacastow, R., Whorf, T.: Measurements of the concentration of carbon dioxide at mauna loa observatory, hawaii. In: Carbon dioxide review. Oxford University Press, New York (1982) 377–385

Date	Month	Forecast	Lower bound	Upper bound	Date	Month	Forecast	Lower bound	Upper bound
2004-01	Jan	376.8	376.8	376.8	2006-01	Jan	379.7	379.7	380.3
2004-02	Feb	377.6	377.6	377.6	2006-02	Feb	380.5	380.5	381.1
2004-03	Mar	378.5	378.4	378.5	2006-03	Mar	381.4	381.4	381.9
2004-04	Apr	379.8	379.7	379.8	2006-04	Apr	382.7	382.7	383.2
2004-05	May	380.4	380.3	380.4	2006-05	May	383.3	383.3	383.8
2004-06	Jun	379.8	379.8	379.8	2006-06	Jun	382.8	382.7	383.3
2004-07	Jul	378.3	378.3	378.3	2006-07	Jul	381.3	381.3	381.8
2004-08	Aug	376.3	376.2	376.3	2006-08	Aug	379.2	379.2	379.7
2004-09	Sep	374.5	374.5	374.5	2006-09	Sep	377.4	377.4	377.9
2004-10	Oct	374.4	374.4	374.5	2006-10	Oct	377.4	377.4	378
2004-11	Nov	375.8	375.8	376	2006-11	Nov	378.8	378.7	379.5
2004-12	Dec	377.2	377.2	377.4	2006-12	Dec	380.1	380.1	380.9
2005-01	Jan	378.2	378.2	378.5	2007-01	Jan	381.2	381.1	382
2005-02	Feb	379	379	379.3	2007-02	Feb	382	381.9	382.8
2005-03	Mar	379.9	379.9	380.1	2007-03	Mar	382.9	382.8	383.7
2005-04	Apr	381.2	381.2	381.5	2007-04	Apr	384.2	384.2	385
2005-05	May	381.8	381.8	382.1	2007-05	May	384.8	384.8	385.6
2005-06	Jun	381.3	381.3	381.6	2007-06	Jun	384.2	384.2	385.1
2005-07	Jul	379.8	379.8	380	2007-07	Jul	382.7	382.7	383.6
2005-08	Aug	377.7	377.7	377.9	2007-08	Aug	380.7	380.7	381.4
2005-09	Sep	376	376	376.2	2007-09	Sep	378.9	378.9	379.7
2005-10	Oct	375.9	375.9	376.2	2007-10	Oct	378.8	378.8	379.7
2005-11	Nov	377.3	377.3	377.7	2007-11	Nov	380.2	380.2	381.2
2005-12	Dec	378.6	378.6	379.2	2007-12	Dec	381.5	381.5	382.7

Table 2. Combined forecast for the ACDC for the forthcoming 48 month. The time series values for the years 2004 to 2007 were not available at the time this study was conducted.

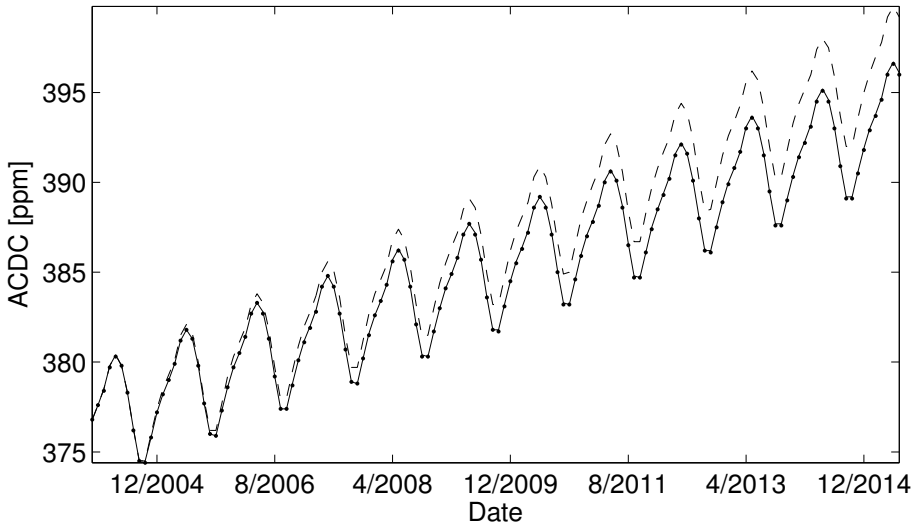


Fig. 2. Combined prediction of the ACDC for the forthcoming 138 month. While the forecast is displayed a solid line, lower and upper bounds are depicted as dotted resp. dashed lines. The forecast keeps close to its lower bound. Lower and upper bounds were simply computed as the lower and upper quartile of the range of forecasts and should therefore not be considered to be defined in a strict mathematical framework.

6. Chan, F., McAleer, M.: Analysing trends and volatility in atmospheric carbon dioxide concentration levels. In Claudia Pahl-Wostl, Sonja Schmidt, A.E.R., Jakeman, A.J., eds.: Complexity and Integrated Resources Management. Volume 3., iEMSs (2004) 1455–1461