Dendrochronologies: Process estimation and missing-data imputations

Time Series Analysis

2018-05-21

Abstract

A dendrochronology compiled using samples taken from wooden buildings in the Bernese Oberland is investigated using methods of time series analysis. A stationary series is obtained through a combination of scaling, log-transformation and linear trend removal. Several possible models are identified and their performance with respect to imputing missing data is investigated. We demonstrate that the transformed time series of tree ring width data is best discribed by an ARMA(1,1) model and that such a model is also superior to simple linear interpolation when applied to impute missing values.

Contents

1 INTRODUCTION

Dendrochronology, i.e. the method to date tree rings, is used in several fields of science. One relevant application of the method is in archeology where it is used to date (wooden) artifacts. In climate science, dated tree ring widths helps to learn more about the climate of the past. Tree ring chronologies also helps to backdate the radiocarbon concentration, which is used in radiocarbon calibration – another method for dating artifacts.

Usually, time series in dendrochnology span several hundred to thousands of years. Such long series are constructed by merging and overlaying shorter lasting time series from individual trees. By taking the mean of yearly ring widths, a single time series is built. Chronologies are generally constructed for defined geographical regions as the environment (i.e the local climate, soil properties and terrain) has substantial impact on the growth of trees (Baillie 2012).

From the perspective of statistics, tree ring data are of interest because they define a time series. According to Woollons and Norton (1990) and Fox, Ades, and Bi (2001) AR(1) and ARMA(1,1) are popular choices for the time series process after having removed the trend. AR(2) is seen as an appropriate choice as well (see e.g. Bunn 2008).

In this report we focus on imputing one missing value on an otherwise known time series. Our contribution to the literature is twofold. Firstly, by investigating a Swiss dendrochronology from the Bernese Oberland we give further evidence on the stochastic process of dendrochronological time series. Secondly, although dendrochronologies span hundreds or thousands of years without gaps these days, it is still important and useful to check the correctness of a chronology. One way to check the correctness is for instance to compare the observed values with what we expect from the (estimated) process of the time series. Observations that are far away from the expectation, so-called outliers, may need further attention. This paper gives a first insight how well imputations work in dendrochronology.

2 METHODS

Classical time series analysis relies on stationarity. Non-stationary time series have to be transformed such that the resulting series is stationary. In Section 2.1 we describe three common approaches proposed in the (dendrochronology) literature. We also present our method of choice that worked best on our dataset. In Section 2.2 we discuss how the best model can be found. Section 2.3 is dedicated to a short introduction about Kalman smoothing and how the method can be used to impute missing values. since we are interested on imputations on the original time series, we conclude the section by showing how to backtransform the stationary time series.

2.1 TRANSFORMATIONS

Popular methods to transform non-stationary dendrochronologies are the Box-Cox transformation, the Power transformation, and the method of Warren (1980). Additionally, we investigate an 'intuitive' approach that is not reported in the literature. The different methods are explained below.

The Box-Cox transformation aims to stabilize the variance by finding the transformation parameter λ that minimizes the coefficient of variation of the series. The time series is either transformed to $Y_t^{(\lambda)} = ((Y_t + c)^\lambda - 1)/\lambda$, for a λ unequal to 0, or $Y_t^{(\lambda)} = ln(Y_t + c)$, in case of $\lambda = 0$ (Box and Jenkins 1976). A similar approach to the Box-Cox transformation is the Power transformation proposed by Guerrero and Perera (2004) where λ is estimated in a model-independent manner, i.e. irrespective of the time series process, that is deduced from the resulting stationary series. In this method a time series is divided into $H \geq 2$ equal-sized subseries, for which

then the mean $\bar{Z}_h$ and the standard deviation S_h is computed. The optimal parameter λ is chosen such that $S_h/\bar{Z}_h^{1-\lambda} = c$, with $h = 1, ...H$ holds for a constant value $c > 0$.

The method of Warren (1980) (cited in Woollons and Norton (1990)) proposes to initially detrend the data by fitting a polynomial of the form $Y = \alpha t^{\beta} e^{\delta t}$ and then to divide the residuals by the fitted values. Taking the ln on both sides gives $ln(Y) = ln(\alpha) + \beta ln(t) + \delta t$. This approach is convenient, because the parameters can be estimated by a linear model.

An additional method is deduced on a more intuitive approach. Having raw data at hand, we are able to also take into consideration the standard deviation of yearly ring width measurements sd_t at time $t = 1, ..., T$. In particular, we proceed by:

1. Scaling the means by dividing them by the standard deviation: $B_t = \bar{r}_t/sd_t, t = 1, ..., T$,
2. Log tranform the scaled means B_t to get $C_t = ln(B_t)$,
3. Finally, fit a linear model of the form $C_t = \beta_0 + \beta_1 t$ to obtain the residuals X_t

A drawback of the last method is that it requires to estimate the standard deviation at time $t = 1, ..., T$, which is not possible if only data on aggregated yearly mean tree ring widths are at hand.

2.2 MODEL FITTING

The stochastic process of a time series can be either found visually by inspecting the ACF (autocorrelation function) and PACF (partial autocorrelation function) or computationally by comparing the AIC or a similar criterium. We use both approaches. Table 1 shows the decision matrix, which is considered to identify the form of the process visually. Automatic model selection can be done in R for instance with the `autofit` function from the `itsmr` package and the `auto.arima` function from the `forecast` package.

Table 1: Decision matrix for model identification from ACF and PACF plots.

	ACF	PACF
AR(p)	exponential decrease	p spikes
MA(q)	q spikes	exponential decrease
ARMA(p,q)	exponential decrease	exponential decrease

The candidate models are further tested by analyzing the residuals graphically. Precisely, the standardized residuals, the ACF of the resudials and the p-values of the Ljung-Box test applied on the residuals are studied.

2.3 IMPUTATIONS

As we would like to use as much information as possible, a good method to impute missing data is Kalman smoothing, since it allows to use the information of the observed values before and after the missing value. In order to apply Kalman smoothing, we have to formulate appropriate models, which we think describe the process of our time series of interest well, in state space form. Motivated from the literature and from our analysis as well, we assume AR(1), AR(2) and ARMA(1,1) as reasonable processes and compare their performance with respect to imputations.

State space models are usually applied to situations where an observed value y is an (affine) function of a state variable x and some noise. Assuming Gaussian noise for the state and the observation equation and a (prior) Gaussian distribution for the initial state as well, allows to calculate the distribution of the state at any time

t within the series. Kalman smoothing works perfectly with missing values, but a potential drawback of the method is it reliance on Gaussian distributions.[1]

For the analysis, AR(1), AR(2) and ARMA(1,1) are formulated in state space form as in Hyndman (n.d.), and the parameters are estimated with the `dlm` package in R (Petris 2009). Since an estimate of the standard deviation is required for the backtransformation of the series (see Section 2.1), an estimate for the standard deviation is required too. We will see later in the report that the standard deviation can be estimated quite well by assuming an ARMA(1,1) model. By defining X_t as the transformed time series, the original series Y_t is obtained by:

1. adding the linear trend back, i.e. $\tilde{X}_t = X_t + \alpha + \beta t$,
2. exponentiating $\tilde{X}_t$, i.e. $\hat{X}_t = exp(\tilde{X}_t)$,
3. and multiplying the standard deviation to $\hat{X}_t$, i.e. $Y = \hat{X}_t sd_t$

We also check our implementation with the `na.kalman` method of the `imputeTS` package of Moritz and Bartz-Beielstein (2017). Furthermore, all models are compared to simple linear interpolation of the original time series, which is our benchmark.

3 DATA

The dataset used for the analysis is supplied from the the archaeological service of the canton Berne through a personal contact at the dendrology laboratory in Sutz, BE, who also detailed the process of acquiring the samples (see also (Amt für Kultur / Archäologischer Dienst 2018)). The chronology comprises 234 tree ring sequences from different buildings in the Bernese Oberland that are all located above 900 meters. Typically, the wooden structure of old buildings is sampled by boring the beams. The drilling core is then analyzed by microscopy to record the sequence of tree ring widths. Spruce was a common building material and there is a wealth of houses supplying data. The data can be assumed to originate almost entirely from trees grown above 900 meters because trees were only very rarely transported uphill.

We now give some definitions and show in a more rigorous manner how the mean and standard deviation are calculated from the raw dataset. Let be $R_t := \{r_{ti}, i = 1, ...N\}$ the set of ring width measurements r for a fixed year t. Then, the cardinality for a fixed year t, also denoted the *sample depth*, is defined as $n_t := |R_t| = \sum_{i=1}^{N} \mathbb{1}_{(y_{ti} \neq \text{NA})}$. The average ring width of a fixed year t is then $\bar{r_t} := \frac{1}{n_t} \sum_{i=1}^{N} r_{ti} \mathbb{1}_{(r_i \neq \text{NA})}$, and the sample standard deviation can be computed as $sd_t := \sqrt{\frac{1}{n_t - 1} \sum_{i=1}^{N} (r_{it} - \bar{r_t})^2 \mathbb{1}_{(y_i \neq \text{NA})}}$. Finally, the time series of yearly mean tree ring widths calculated from individual trees is defined by $\bar{R} := \{\bar{r_t}, t = 1, ..., T\}$.

The series of mean tree ring widths $\{\bar{r_t}\}$ is shown in Figure 1[2]. The gray overlay represents the sample depth n_t. From the plot, a correlation between sample depth n_t and the average ring width $\bar{r_t}$ could be assumed. However, no significant relationship was found when we used the sample depth n_t as an exogeneous variable in a linear model. In contrast to that, the Spearman's correlation test is significant at $\alpha = 5\%$ with a p-value of $2.194e - 06$ and estimated $\rho = 0.1796$. Transforming the dendrochronologies with respect to the sample depth n_t could be an intresting approach for a future analysis, but is not part of this report.

Furthermore, bigger climatic trends such as the small ice age, a period of low temperatures throughout Europe from the 15th to 18th century can be seen in the data. The overall decreasing trend could be explained by population growth and an intensifying forest cultivation practice. Cutting out large amounts of old trees from a forest leaves the young trees competing for sunlight. As vertical growth is more important during the first years of a tree's life, the year rings are generally smaller.

[1] However, this is also the case when a time series model is fitted by maximum likelihood.

[2] The plot is created with the `dplR` package.

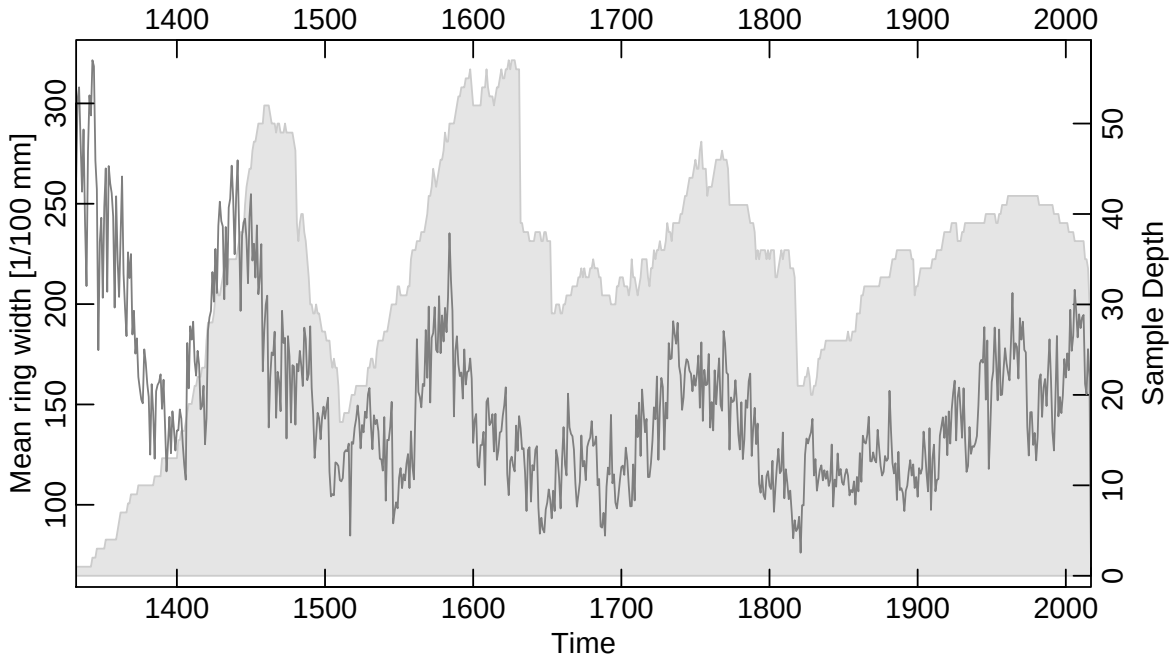

Figure 1: The timeseries of yearly mean tree ring widths. The grey overlay shows the sample depth for each year.

As recommended by the supplier of the data, we ignore data after 1800 due to the prevalence of young trees distorting average ring widths $\bar{r_t}$ in that time range. Additionally, we discard data before 1400 because of a very small sample depth in the years before. The unit of the untransformed ring width is 1/100 mm throughout the paper.

4 RESULTS

4.1 STATIONARITY

On our specific dataset, a combination of scaling, log-transforming and subtracting a linear trend – i.e. what we refer to as the intuitive approach in Section 2.1 – produces in our view the best result, yielding a times series that looks reasonably (weakly) stationary as the Figure 2 visually assures.[3] The series is mean-centered and shows a relatively constant variance. However, two outliers, which were already evident in the original data, remain. A possible explanation in general for extremly small growth rates could be volcanic activity, which has been shown to impact tree growth (Sigl et al. 2015). However no major events are known to us for that period.

The ACF and PACF for the stationary series (see Figure 3) give further evidence for the plausibility of the transformation. The tansformation we propose results in a time series with a drecreasing ACF.[4] The ACF of the transformed time series is discussed in more details in Section 4.2.

[3] Stationarity is only assessed graphically. Unit root tests such as the Augemented Dickey-Fuller (ADF) or Phillips-Perron (PP) test, or other tests for stationarity like Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test have usually very limitied explanatory power due to their sensitivity on assumptions and are thus not performed.

[4] The other three approaches discussed in Section 2.1 show a slow ACF decrease, which let us whether the resulting time series are indeed stationary, since such patterns are seen with random walks, which are non-stationary.

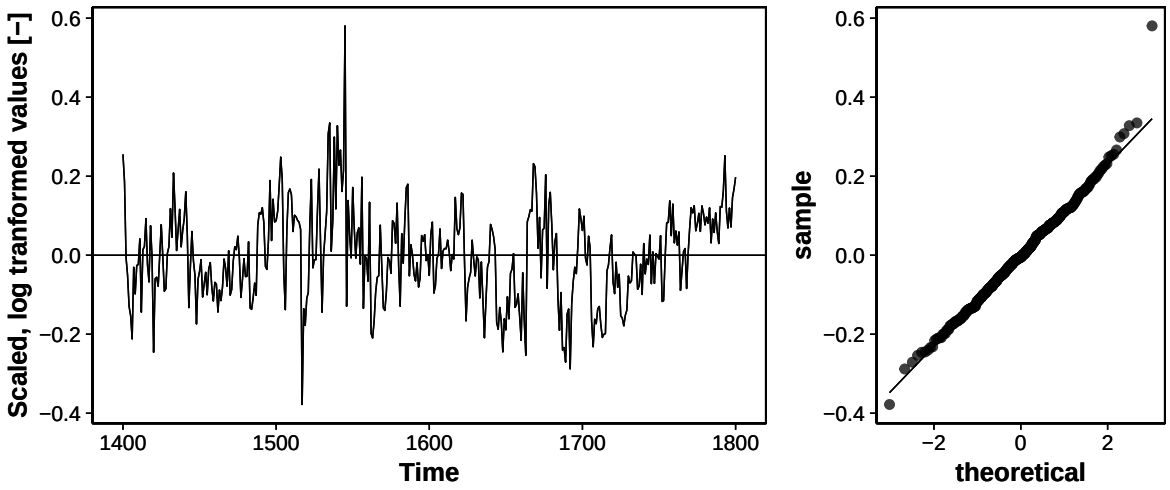

Figure 2: The stationary series obtained throgh scaling, log-transformation and order-1 trend removal is displayed on the left. On the right, the qq-plot of the series is shown. The series seems to be normally distributed, except for the two outliers at around 1520 and 1550.

4.2 MODEL SELECTION

As explained in Section 2.2, the patterns of the ACF and PACF can be used to estimate the parameters p and q of an ARMA(p,q) process (see Table 1). Figure 3 gives some basis to argue for either an AR(1) or ARMA(1,1) process.

The visual guess that the process could be ARMA(1,1) is confirmed by automatic model selection methods provided by the `autofit` method from the `itsmr` package[5] and the `auto.arima` method from the `forecast` package[6].

Despite the fact that ARMA(1,1) is the best model in terms of AIC, we decided to include AR(1), AR(2) and ARMA(1,1) in our analysis too, because all of them are frequently found in the literature to fit dendrochronological data well (Woollons and Norton 1990). Table 2 lists values of the AIC for the different models. ARMA(1,1) shows the best fit to the data, followed by the AR(2) model. The performance of the AR(1) model is the weakest. In Table 3 the estimated parameters for all models are presented. We also report the standard deviations of the estimates.

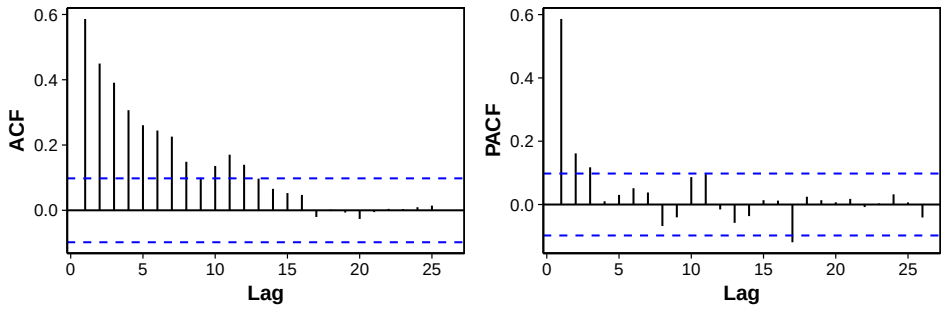

Figure 3: ACF (left) and PACF (right) of the stationary time series X_t.

[5] The bounds for the parameters p and q are set to be 0 and 3, respectively.

[6] The bounds for the parameters p and q are set to be 0 and 5, respectively. These are the default values and will be later used in the `na.kalman` method of the `imputeTS` package.

Table 2: Comparison of AIC of the fitted models.

Model	AIC
ARMA(1,1)	-765.8969
AR(1)	-752.1661
AR(2)	-761.6195

Table 3: Model parameters for ARMA(1,1) and AR(1).

Model	Parameter name	Parameter value	standard error
ARMA(1,1)	ϕ	0.8135	0.0481
	θ	-0.3540	0.0791
	$\sigma^2_{ARMA(1,1)}$	0.0085	-
AR(1)	ϕ	0.5960	0.0405
	$\sigma^2_{AR(1)}$	0.0089	-
AR(2)	ϕ_1	0.4972	0.0493
	ϕ_2	0.1691	0.0496
	$\sigma^2_{AR(2)}$	0.0086	-

The validity of all models is further assessed graphically by plotting the standardized residuals and their ACFs. Moreover, Ljung-Box tests are perfomed for different lags to check the independence assumption. The plots are shown in Figure 4. Some insights can be gainded from the results of the Ljung-Box tests. The p-values of the ARMA(1,1) model are all greater than 5%, whereas one p-value of the AR(2) model and all of the AR(1) are smaller than 5%. Small p-values indicate that the residuals might not be independent, which would violate the model assumption of independent residuals. This further confirms that the ARMA(1,1) is the best model, followed by the AR(2) model. The residuals of the AR(1) show the most prominent temporal dependencies, which is an indication that the model is too simple.

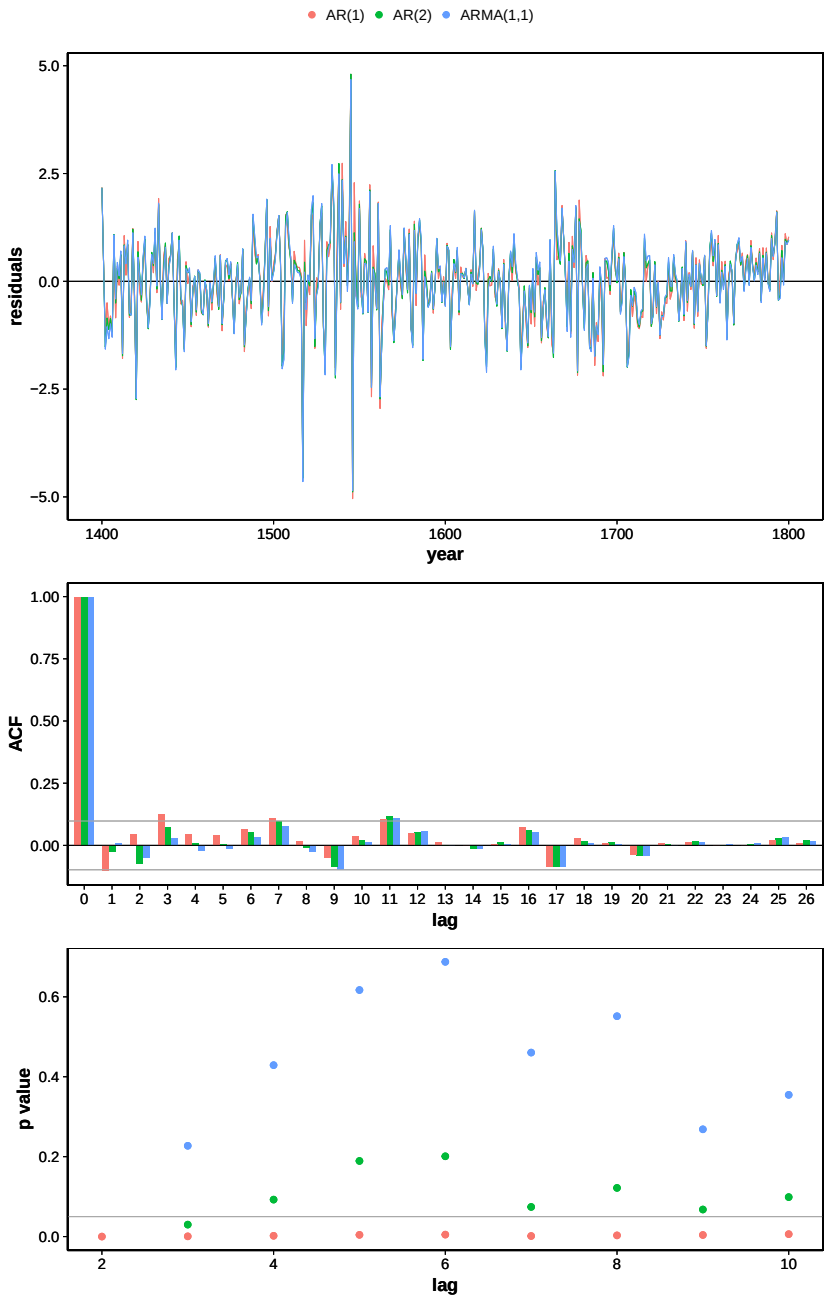

Figure 4: Diagnostic plots for ARMA(1,1), AR(1) and AR(2). Top: Standardized residuals, middle: ACF, bottom: Ljung-Box test

4.3 IMPUTATIONS

Section 3.2 gives evidence that the ARMA(1,1) model describes the stochastic process best. We expect to see the same pattern when we compare the models with respect to their imputation performance. The missing values are restricted to be spread equally over the time series by defining 40 blocks, the first beeing from the year 1401 to 1409, the second from year 1411 to 1419, and so forth.[7] We run 40 simulations in total. For the first simulation, block number 1 is chosen and one observation of block number 1 is randomly defined as missing. All other values remain known. We then try to impute the missing value. For the other 39 blocks this procedure is repeated. The unknown standard deviation, needed for the backtransformation, is either estimated (i) with an ARMA(1,1) (model selection) process or (ii) linearly interpolated. We also compare the results to the case where we know the standard deviation (see Table 4) and to the case where the imputed value is obtained by linear interpolation on the original time series. The coefficients of the linear model needed for the backtransformation are estimated each time without knowing the the missing value.

Table 4 shows the MSE (mean squared errors) for the stationary series denoted as “transformed”, and the back-transformed series denoted as “original”.Assuming an ARMA(1,1) process for the transformed tree rings and for the standard deviation outperforms the other approaches. Not surprisingly, the best results are obtained when the standard deviation is known. Simple linear interpolation on the original time series performs poorly with an MSE of 302.4. The results of the imputations are plotted in Figure 5.

Table 4: MSE results of the simulation study.

Approach	AR(1)	AR(2)	ARMA(1,1)	Linear Interpolation
Transformed	5.796e-03	5.582e-03	5.366e-03	5.661e-03
Original with sd as ARMA(1,1)	2.694e+02	2.646e+02	2.583e+02	2.910e+02
Original with sd lin. interpolated	2.735e+02	2.783e+02	2.757e+02	2.859e+02
Original with true sd	1.758e+02	1.703e+02	1.633e+02	1.691e+02

The `na.kalman` method of the `imputeTS` package produces very similar results. In case of assuming an ARMA(1,1) process for the transformed times series and for the standard deviation – both assessed by automatic model selection within the package – results in a MSE of 256.1. In case of linearly interpolate, a MSE of 275.7 is obtained.

[7] We separate the blocks by one observation to avoid having consecutive missing missing values. This will later allow us to plot results in one plot.

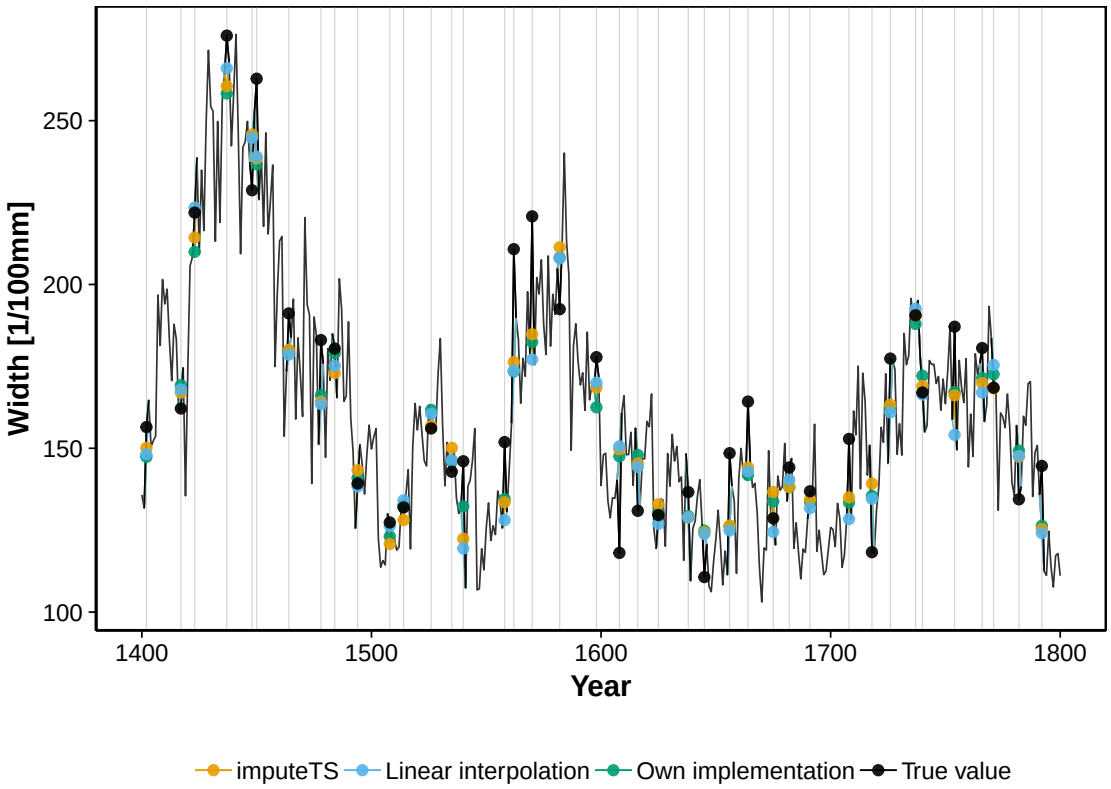

Figure 5: Comparison of imputed values obtained through different methods. Vertical lines represent location of missing values.

5 DISCUSSION AND CONCLUSION

We achieve stationarity by a – to our knowledge – not yet described transformation in the literature with very satisfying results on our dataset. Our results support the assumption that a transformed tree ring time series can be reasonalbly described by an ARMA(1,1) process. Furthermore, we show that an ARMA(1,1) process can also be helpful to impute missing data. In particular, we give some evidence that the approach based on an ARMA(1,1) process outperfoms simple linear interpolation on the original time series.

In our analysis, the shape of the trend, which we use to achieve stationarity, is not explained by exogenous variables. However, the growth of trees is likely influenced by many factors as for example the climate, soil, slope, exposition, forest cultivation, sun activity (e.g. 11-year cycle) and others. In case the exogoneous variables are known at a specific time, but tree ring data are missing, the knowledge of such relationships might be helpful to improve the imputations. A preliminary analysis based on an ARMA model with exogeneous variables[8] from the U.S. National Oceanographic and Athmosperic Administration (NOAA) for Arizona and Arizonian tree ring data from R-package `dplr` has shown the potential usefulness of such an approach. An unsolvable problem is however the very limited data availability since climatic data were not (systematically) collected before 1895.

6 REFERENCES

Amt für Kultur / Archäologischer Dienst. 2018. "Dendrochronologische Reihen für Fichten und Weisstannen." *Erziehungsdirektion Des Kantons Bern*. Sutz, BE: Departement für Bildung und Kultur Bern.

Baillie, M.G.L. 2012. *A Slice Through Time: Dendrochronology and Precision Dating*. Routledge, New York.

Box, G.E.P., and G.M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control, Revised Ed*. Wiley & Sons, New Jersey.

Bunn, A.G. 2008. "A Dendrochronology Program Library in R (dplR)." *Dendrochronologia* 26 (2):115–24.

Fox, J.C., P.K. Ades, and Huiquan Bi. 2001. "Stochastic Structure and Individual-Tree Growth Models." *Forest Ecology and Management* 154 (1):261–76.

Guerrero, Victor M., and Rafa Perera. 2004. "Variance Stabilizing Power Transformation for Time Series." *Journal of Modern Applied Statistical Methods* 3 (2). http://digitalcommons.wayne.edu/jmasm/vol3/iss2/9.

Hyndman, R. n.d. "State space models. 3: ARIMA and RegARMA models, and dlm." Monash University.

Moritz, S., and T. Bartz-Beielstein. 2017. "imputeTS: Time Series Missing Value Imputation in R." *The R Journal* 9 (1):207–18.

Petris, G. 2009. "Dlm: An R Package for Bayesian Analysis of Dynamic Linear Models." *University of Arkansas*.

Sigl, M., M. Winstrup, J.R. McConnell, K.C. Welten, G. Plunkett, F. Ludlow, U. Büntgen, et al. 2015. "Timing and Climate Forcing of Volcanic Eruptions for the Past 2,500 Years." *Nature* 523 (7562):543.

Woollons, R.C., and D.A. Norton. 1990. "Time-Series Analyses Applied to Sequences of Nothofagus Growth-Ring Measurements." *New Zealand Journal of Ecology* 13 (1):9–15.

[8] Such models are also called ARMAX.

7 APPENDIX

7.1 R SESSION INFO

The following output details the exact environment used for the calculations and result display of this report.

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] bindrcpp_0.2.2     matrixcalc_1.0-3   dlm_1.1-4
##  [4] imputeTS_2.6       forecast_8.3       tidyr_0.8.0
##  [7] MASS_7.3-50        itsmr_1.8          dplR_1.6.7
## [10] stringr_1.3.1      readtext_0.71      ggpubr_0.1.6.999
## [13] magrittr_1.5       egg_0.2.0          ggplot2_2.2.1.9000
## [16] kableExtra_0.8.0   knitr_1.20         gridExtra_2.3
## [19] dplyr_0.7.4
##
## loaded via a namespace (and not attached):
##  [1] tseries_0.10-44    httr_1.3.1         viridisLite_0.3.0
##  [4] uroot_2.0-9        R.utils_2.6.0      assertthat_0.2.0
##  [7] TTR_0.23-3         animation_2.5      yaml_2.1.19
## [10] pillar_1.2.2       backports_1.1.2    lattice_0.20-35
## [13] glue_1.2.0         quadprog_1.5-5     digest_0.6.15
## [16] rvest_0.3.2        colorspace_1.3-2   htmltools_0.3.6
## [19] Matrix_1.2-14      R.oo_1.22.0        plyr_1.8.4
## [22] timeDate_3043.102  XML_3.98-1.11      pkgconfig_2.0.1
## [25] purrr_0.2.4        scales_0.5.0       tibble_1.4.2
## [28] withr_2.1.2        urca_1.3-0         nnet_7.3-12
## [31] lazyeval_0.2.1     quantmod_0.4-13    evaluate_0.10.1
## [34] R.methodsS3_1.7.1  nlme_3.1-137       xts_0.10-2
## [37] xml2_1.2.0         ggthemes_3.5.0     tools_3.5.0
## [40] data.table_1.11.2  hms_0.4.2          matrixStats_0.53.1
## [43] munsell_0.4.3      compiler_3.5.0     stinepack_1.3
## [46] rlang_0.2.0        rstudioapi_0.7     labeling_0.3
## [49] rmarkdown_1.9      gtable_0.2.0       fracdiff_1.4-2
## [52] curl_3.2           R6_2.2.2           zoo_1.8-1
## [55] bindr_0.1.1        rprojroot_1.3-2    readr_1.1.1
## [58] stringi_1.2.2      parallel_3.5.0     Rcpp_0.12.16
## [61] png_0.1-7          tidyselect_0.2.4   lmtest_0.9-36
```