Contents

1 Q1. Stationarity of Process

If E_t is a white noise process with mean 0 and variance σ^2, then it has following properties:

$$\begin{aligned} E[E_t] &= 0 \\ Var[E_t] &= \sigma^2 \\ Cov(E_t, E_{t+h}) &= E[(E_t - E[E_t])(E_{t+h} - E[E_{t+h}])] \\ &= E[E_t E_{t+h}] \\ &= E[E_t]E[E_{t+h}] \\ &= 0 \end{aligned}$$

As in white noise different elements are independent from each other, their covariance is zero.

Given the statistical properties of a stationary time series process that its expectation and variance are constant and its covariance only depends on the lag h, we can prove whether these 3 time series processes are stationary or not, as follows.

The first time series process is defined as:

$$X_t = t + E_t \qquad E_t \sim WN(0, \sigma^2)$$

It is not stationary, since:

$$E[X_t] = t$$

which means that the expectation is not constant.

The second time series process is the 1-lag differentiated process of X_t defined as:

$$\begin{aligned} Y_t &= X_t - X_{t-1} \\ &= t + E_t - (t - 1 + E_{t-1}) \\ &= 1 + E_t - E_{t-1} \end{aligned}$$

It is stationary, since:

$$\begin{aligned}
E[Y_t] &= 1 \\
Var[Y_t] &= Var[1 + E_t - E_{t-1}] \\
&= Var[E_t - E_{t-1}] \\
&= Var[E_t] + Var[E_{t-1}] - 2Cov(E_t, E_{t-1}) \\
&= Var[E_t] + Var[E_{t-1}] \\
&= 2\sigma^2 \\
Cov(Y_t, Y_{t+h}) &= Cov(1 + E_t - E_{t-1}, 1 + E_{t+h} - E_{t+h-1}) \\
&= Cov(E_t - E_{t-1}, E_{t+h} - E_{t+h-1})
\end{aligned}$$

which means that both expectation and variance of Y_t are constant and its covariance only depends on the lag h.

The third time series process is stationary, since:

$$\begin{aligned}
E[Z_t] &= E[E_t] = 0 \\
Var[Z_t] &= Var[E_t] = \sigma^2 \\
Cov(Z_t, Z_{t+h}) &= Cov(E_t, E_{t+h}) = 0
\end{aligned}$$

which means that Z_t is equal to E_t, where both expectation and variance of Z_t are constant and its covariance is not time dependency.

In conclusion, X_t is not stationary, while Y_t and Z_t are stationary.

2 Q2. SARIMA Models for Pine Year Rings

The pine data set that contains the measurements of the year rings of Douglas Fir from year 1107 to 1964 is used to conduct the time series analysis. The time series plot of the year rings from 1201 to 1500 is analysed and different Seasonal ARIMA models are fitted and compared. The chosen model is then used to predict the year rings of the next 50 years and the RMSE is computed based on the true measurements.

Time Series Plot & Data Transformation

The `pine.dat` file is imported and the R object `ts` is created for time series analysis. The `window()` function is applied to obtain the subset of time series observations from year 1201 to 1500. Time plot of the year rings in Figure 1 indicates that the time series process is clearly non-stationary. The measurements generally fluctuate and demonstrate a slight increase from year 1300 to 1400. The great variability exits in two time periods (around 1240~1250 and 1350~1400), when sudden surges and drops are identified. To stabilise the variances, the data are transformed using logarithms. As is shown in Figure 1, after log-transformation, the variability is more or less reduced but the time series process still seems non-stationary with strong seasonality.

We then apply the first differencing to the logged data, the time plot of which is shown in Figure 2 with ACF and PACF plots. According to the seasonal change approximately every 10 years in the original pine data, 10-year-lag differencing is conducted. The time plot seems stationary after differencing. Both ACF and PACF plots are examined to determine an appropriate model to use, as well as the associated orders. In ACF and PACF plots of 10-year differenced log pine data, there are significant spikes in PACF at seasonal lags 10 and 20. It may suggest a seasonal AR(2) term. Similarly, there is a significant spike in ACF at lag 10, suggesting a seasonal MA(1) term. In the non-seasonal lags, there are spikes at lag 1 to 3 and an almost significant spike at lag 8 in ACF, indicating a non-seasonal MA(3) or MA(8) term. Likewise, spikes at lag 1, 2 and 8 in PACF probably suggest a non-seasonal AR(2) or AR(8) term.

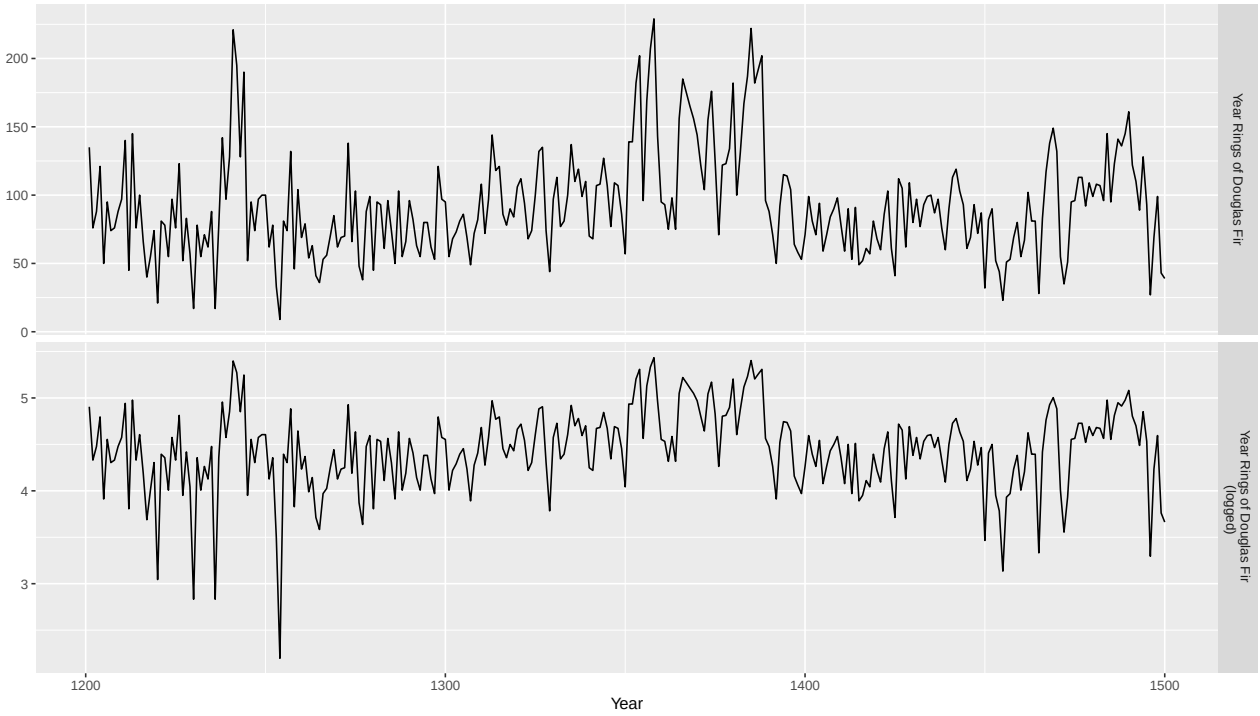

Figure 1: Time series plot of original pine data and log-transformed from 1201 to 1500.

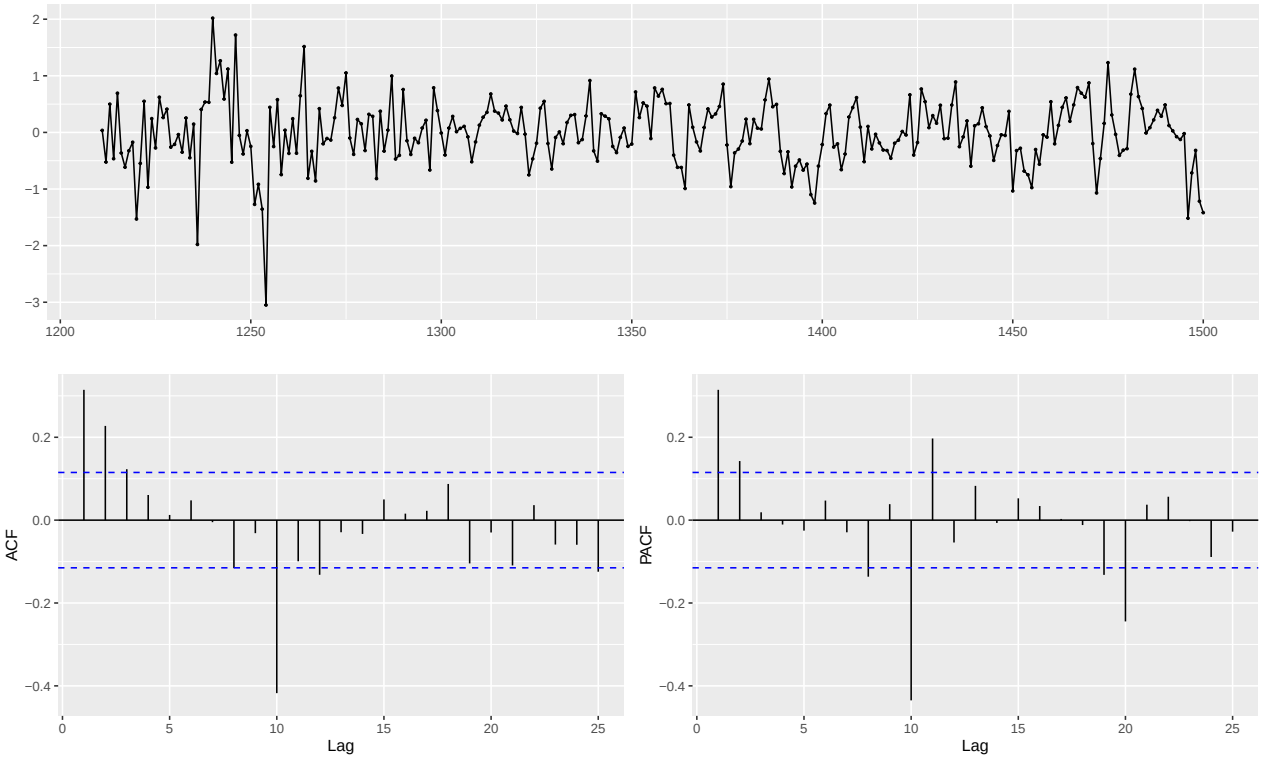

Figure 2: 10-year differenced log time series with associated autocorrelations and partial autocorrelations.

Model Comparison

Consequently, this initial analysis suggests that a $SARIMA(2,0,3)(2,1,1)^{10}$ model might be appropriate to fit on the log-transformed pine data. Several models with variations on the non-seasonal and seasonal orders are fitted, and AIC values are used to search for a better model. The fitted models and the associated AIC values are shown in Table 1. The $SARIMA(3,0,3)(2,1,1)^{10}$ model is selected with the smallest AIC of 328.943.

In Figure 3, the residuals of the chosen model are assessed by using the time plot and ACF of the residuals and comparing with the normal distribution. The ACF plot of the residuals shows that all the autocorrelations are within the threshold limits and the histogram plot of the residuals suggests that the residuals are close to normal distribution, indicating that the residuals look like a white noise series. The p-value of Ljung-Box test is calculated as 0.270 (>0.05), which means that we do not reject the null hypothesis. Hence, we conclude that the residuals are not distinguishable from white noise and the chosen SARIMA model does not exhibit significant lack of fit.

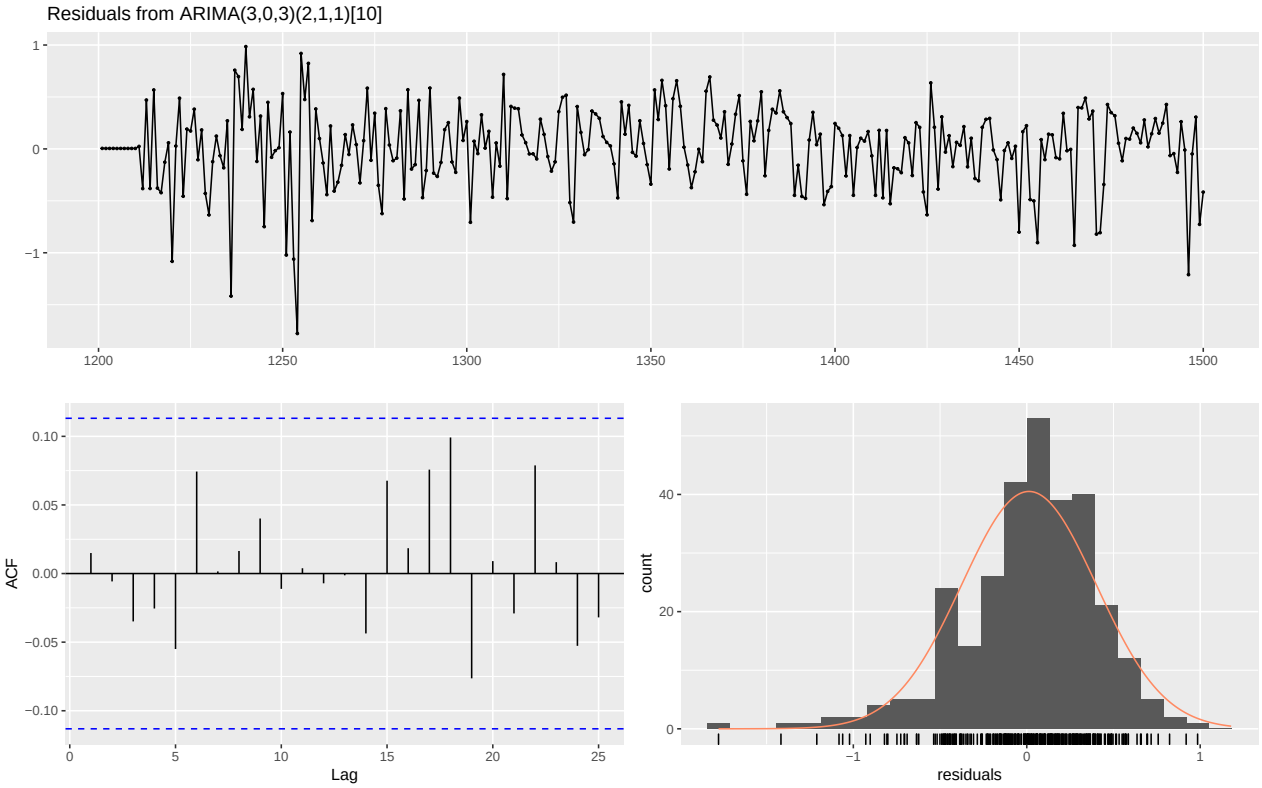

Figure 3: Residual analysis of SARIMA(3,0,3)(2,1,1)[10] model.

Table 1: Fitted SARIMA models with AIC values.

Model	AIC
$SARIMA(2,0,3)(2,1,1)^{10}$	336.438
$SARIMA(2,0,2)(2,1,1)^{10}$	335.368
$SARIMA(2,0,8)(2,1,1)^{10}$	342.678
$SARIMA(1,0,3)(2,1,1)^{10}$	334.077
$SARIMA(3,0,3)(2,1,1)^{10}$	328.943
$SARIMA(8,0,3)(2,1,1)^{10}$	334.525
$SARIMA(3,0,3)(2,1,0)^{10}$	386.034
$SARIMA(3,0,3)(2,1,2)^{10}$	333.992
$SARIMA(3,0,3)(1,1,1)^{10}$	332.930

Forecasting

Forecasting based on the chosen $SARIMA(3,0,3)(2,1,1)^{10}$ model is carried out. Forecasts of the year rings for the following 50 years from year 1500 are shown in Figure 4. It presents the fitted 50 logged year rings from year 1501 to 1550; dark blue and light blue areas indicate 80% and 95% confidence values associated with the prediction intervals. The fitted values fluctuate and demonstrate generally a slight increasing trend.

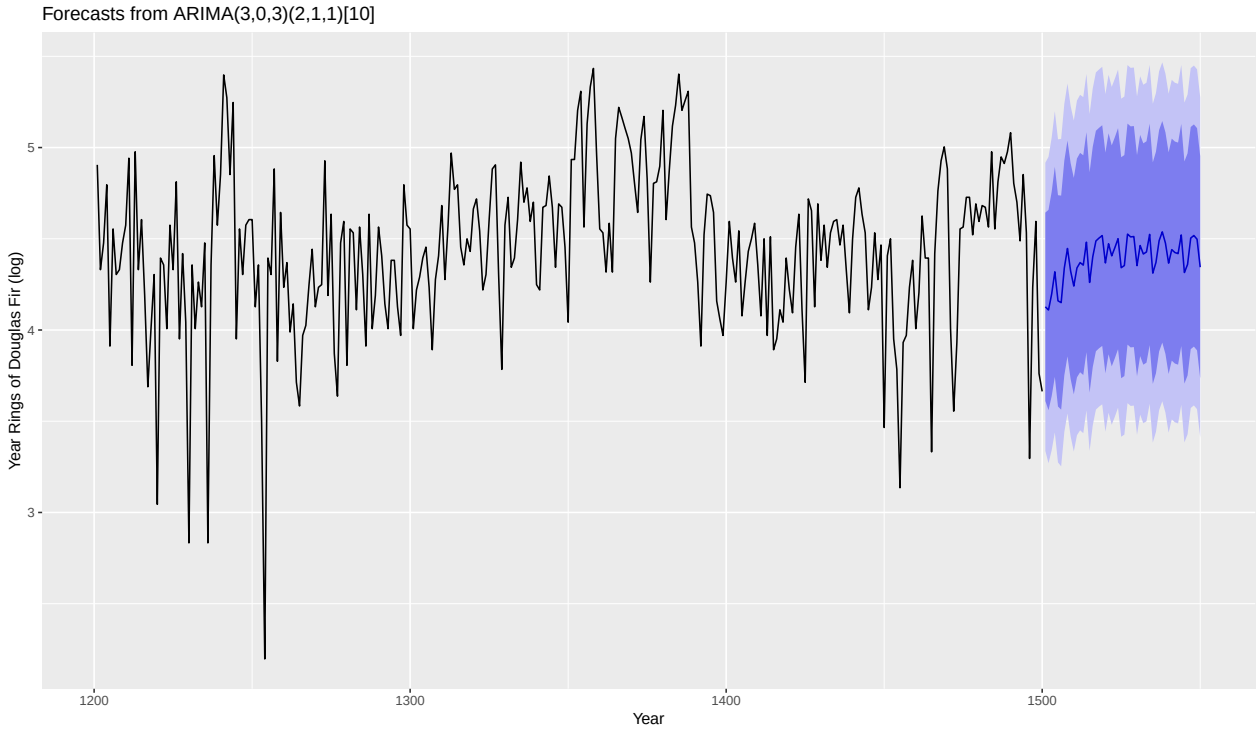

Figure 4: Forecasts for year 1501 to 1550 (logged year rings).

As the true measurements of the year rings from year 1501 to 1550 are in the pine data, we can exponentiate the fitted values first and then compare them with the true year rings. The root mean squared errors (RMSE) between the ground truth and the prediction are computed and shown in Table 2 for various SARIMA models. As indicated, $SARIMA(3,0,3)(2,1,1)^{10}$ model, which has the smallest AIC in the training process, achieves the smallest RMSE of 37.506.

Table 2: Fitted SARIMA models with RMSE values on the test set.

Model	RMSE
$SARIMA(2,0,3)(2,1,1)^{10}$	38.253
$SARIMA(2,0,2)(2,1,1)^{10}$	37.570
$SARIMA(2,0,8)(2,1,1)^{10}$	39.063
$SARIMA(1,0,3)(2,1,1)^{10}$	38.834
$SARIMA(3,0,3)(2,1,1)^{10}$	37.506
$SARIMA(8,0,3)(2,1,1)^{10}$	38.824
$SARIMA(3,0,3)(2,1,0)^{10}$	41.884
$SARIMA(3,0,3)(2,1,2)^{10}$	38.138
$SARIMA(3,0,3)(1,1,1)^{10}$	37.690

Overall, the time series analysis is conducted on the measurements of a Douglas Fir's year rings from year 1201 to 1500. The non-stationary time series data are log-transformed and differenced, and ACF/PACF plots are used to choose an appropriate SARIMA model. The SARIMA models with various seasonal and non-seasonal orders are fitted and compared. The $SARIMA(3,0,3)(2,1,1)^{10}$ model is finally chosen based on AIC and residual analysis to predict the next 50 years' measurements of the year rings.

3 Q3. Time Series Plots & Correlograms

Autocorrelations measure the linear relationship between lagged values of time series. The formula of ACF is indicated in (1):

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \tag{1}$$

For a time series, the partial autocorrelation between y_t and y_{t-k} is defined as the conditional correlation between y_t and y_{t-k}, conditional on $y_{t-k+1}, ..., t_{t-1}$, the set of observations that come between the time points t and t-k.

Firstly, we match ACF and PACF plots. As is known, partial autocorrelation should be the same as autocorrelation when the lag is equal to 1 for time series process, since both measure the correlation between data points at time t with those at time t-1. From the plots, ACF (I) and PACF (A) have positive autocorrelations almost within threshold limits at lag 1. ACF (II) and PACF (C) have positive significant spikes at lag 1, whereas ACF (III) and PACF (B) have negative autocorrelations close to 0 at lag 1.

As a result, ACF (II) and PACF (C) are a match. ACF (II) shows an exponential decay while PACF (C) has a cut-off at lag 2. It suggests that an AR(2) model might be appropriate for the corresponding time series process. A closer look at the significant spike at lag 1 indicates that the time series should have positive autocorrelations at small lags. **Time series 1** is distinguished from the other two, as data points within small lags seem to be positively correlated. Cited the formula of ACF in (1), the data points at small lags should be generally above the mean or below the mean at the same time. The other two time series demonstrate greater variability in small lags, suggesting that their autocorrelations at small lags might be negative.

ACF (I) matches with PACF (A). In both plots, there are generally no significant spikes, as all the spikes except at lag 0 are more or less within the threshold limits. The corresponding time series process should behave like a white noise process, where the mean and variance are constant and the correlation between data points at lags is 0. Compared to time series 2,

of which the observations look more random, uncorrelated and not predictable, time series 3 seems to have more regular patterns and demonstrate seasonality. Additionally, the peaks happen more shortly after the neighbour troughs in time series 3, resulting in negative autocorrelations at small lags, while times series 2 seems to have positive autocorrelations at small lags. As a result, we choose **time series 2** as a match to ACF (I) and PACF (A). Thus ACF (III) and PACF (B) should correspond to **time series 3**. Moreover, ACF (III) with a cutoff at lag 2 and PACF (B) with a cutoff at lag 4 show that we might use an ARMA(4,2) model to fit time series 3.

In conclusion, **time series 1** should be corresponding to ACF (II) and PACF (C); **time series 2** might match with ACF (I) and PACF (A); and **time series 3** should be consistent with ACF (III) and PACF (B).

4 Q4. Summary of Model-based Clustering of Baltic Sea-level

In this study, researchers applied time series clustering based on forecast densities to describe regional sea-level variability in the Baltic Sea in terms of future relative heights. Data were from long (>30 years) monthly records of **relative sea-level heights** from Baltic tide gauges, and a total of 14 stations with records extending at least up to 2004 were collected, where clustering based on forecast densities had time series data ending at the same time, 2004.

Time series clustering based on forecast densities was used for each one of the observed time series. Implementations are as below: (1) A model for each one of the time series was defined (it was found that an autoregressive model of order one (AR(1)) was adequate). (2) Data transformation was carried out to remove the mean and the trend and seasonal components, prior to fitting a model. (3) To obtain dissimilarity matrix, B copies of h-step-ahead predicted values were calculated by bootstrapping. (4) Once dissimilarity matrix was constructed, three clustering techniques were applied, namely agglomerative hierarchical method with nearest distance (single linkage), average linkage and Ward's method, where a dendrogram was built for different clusters formed by the models. (5) Bootstrap predictions were computed for each time series. Clusters of sea-level observations based on forecasts at a specific future time were therefore obtained.

Based on forecast densities of future total sea-levels, including trends and seasonal variability, results show a fairly spatial coherency in terms of grouping together stations from the same sub-basin area for 3-month forecasts, while showing a high degree of similarity between most stations at a 6-month horizon but a clear separation from the stations closer to the entrance in the southwestern Baltic. To conclude, as a general pattern, three distinct groups are the northern stations in the Bothnian Sea and Gulf of Finland, the stations in the Baltic Proper, and the southern stations in the western Baltic, regardless of the diversity and complexity of factors influencing sea-level variations at each tide gauge locations. When considering detrended records, reflecting mainly the seasonal cycle, the clustering results are more homogeneous and indicate a clear response of coastal sea-level in spring and in summer.

Comments on the analysis are given as below:

- **Transformation of multivariate time series data**: The preliminary transformation of data is used to reduce the number of parameters to be estimated, thus requiring no relatively restrictive assumption on data and noise distributions (Zhou and Chan, 2014). It allows estimated residuals to be modelled as a stationary time series which admits an autoregression representation. The transforming smooths out the data to reduce noise and therefore improve accuracy, speeds up the clustering process and presents more understandable results of clustering.

- **Seasonal-Trend decomposition procedure**: Seasoned-Trend decomposition procedure based on Loess (STL) is performed to estimate the trend and the seasonal component, yielding a decomposition robust to extreme observations. With the Loess smoother, STL is a filtering procedure for decomposing a time series into trend, seasonal and reminder components, where the simplicity allows fast computation and analysis of the procedure properties (Rojo et al., 2017). It is flexible to specify the amounts of variation in the trend and seasonal components, and capable to decompose series with missing values (Robert et al., 1990). Another feature is specification of the number of observations per cycle of the seasonal components to any integer (>1). Robust trend and seasonal components are not distorted by transient and not acceptable behaviour in the data. Therefore, it makes sense to process long time sea-level series, using STL method.

- **Time series clustering based on forecast densities**: A forecast density is an estimate of probability distribution of possible future values of the process, by a resampling method combined with a nonparametric kernel estimator (Alonso et al., 2006). However, in this paper, researchers do not give reasons why forecast density method is selected. Based on extra study, it makes sense that this method is more informative about the likelihood of the target time series being met. In fact, due to an estimation based on the past observed data, the outcomes only are of value to the extent that the forecast probabilities accurately capture the true probabilities, which exactly is limited by an interval that is a band of plus/minus a fixed times of standard errors.

- **H-step-ahead prediction model**: The h-step-ahead prediction model is performed, where the horizon h is decided by the user. It trains directly an h-step-ahead model rather than a one-step-ahead prediction model to forecast one-by-one into the future for h steps. As iterated forecast, h-step-ahead method is more efficient, if correctly specified, and improves relative performance with forecast horizon (McElroy and Wildi, 2013). However, as direct forecast, one-step-ahead method is more robust to model misspecification. This paper does not give explanation why h-step-ahead model is applied. Although theoretically h-step-ahead model outperforms one-step-ahead model if models can select long lag specifications, one-step-ahead prediction model is worth carrying out as an empirical matter.

- **Clustering**: Researchers apply single and average linkage, and Ward's method. However, this paper does not explain why single linkage is chosen, where distance defined as the distance of the two closest members is a local property that is not affected by merging (Ward, 1963). Therefore, we recommend application of complete linkage which utilises distance denoted as the diameter of a cluster, where the criterion is non-local and compact clusters with small diameters are created rather than long, straggly clusters (Vijaya et al., 2019). Even though complete linkage has sensitivity to outliers, comparison can be considered with average linkage which is less dependent upon extreme values. Additionally, Ward's method evaluates distances between clusters using analysis of variance approach, but prior to Ward's method, it should be confirmed that sample coordinates are transformed to avoid the error sum of squares biased to variables with high variance (Wishart, 1969). Ward's method tends to create clusters of small size. As tide gauges have distinctive seasonal variability, small-sized clusters help detect some subtle patterns. Nevertheless, the method is computationally intensive.

- **Alternative time series clustering**: Prediction-based approach to time series clustering is applied in this paper, where each time series is created by some known models. An alternative feature-based clustering time series is best used when domain knowledge about a problem is available (Zhou and Chan, 2014). It provides interpretability for time series analysis. It clusters time series using just a set of sea-level statistical fea-

tures derived from the shapes of time-series subsequence, where it directly reduces the dimensionality of original time series rather than introduces data transformation, showing less sensitivity to missing values, and it can handle different lengths of time series (Fulcher, 2021). Hence, compared to model-based clustering, feature-based clustering is worth carrying out as well.

- **Robustness & Generality**: Prediction-based time series clustering method is implemented. Before fitting a model, researchers carry out data transformation to obtain stationary time series process. Residual analysis is performed with no significant deviation from Gaussian white noise sequence for the residuals detected, and correlation and normality between residuals are verified. AIC is used to select models. Therefore, the analysis and results are robust. However, it is stressed that all the results are based on forecast densities of future total sea-level. The interpretation of differences between individual stations within the three major sub-basins is hindered by the diverse and complex factors influencing sea-level variations in the Baltic Sea. The value at a specific point on the coast is a complex function of a range of conditions, including coastal topography and morphology, non-linear shallow water tides and currents. Thus, it is proposed that more complements and applications are needed to serve as the spatial resolution to identify such localized behaviour.

5 References

Alonso A.M., Berrendero J.R., Hernández A., Justel A., (2006), "Time series clustering based on forecast densities", Computational Statistics & Data Analysis, Volume 51, Issue 2, Pages 762-776, ISSN 0167-9473, https://doi.org/10.1016/j.csda.2006.04.035.

Fulcher, B., (2021), "Feature-based time-series analysis", [online] Arxiv-vanity.com. Available at: https://www.arxiv-vanity.com/papers/1709.08055/ [Accessed 13 April 2021].

McElroy T. and Wildi M., (2013), "Multi-step-ahead estimation of time series models", International Journal of Forecasting, 29(3), pp.378-394.

Robert B.C., William S.C., Jean E.M. and Irma T., (1990), "STL: A Seasonal-Trend Decomposition Procedure Based on Loess", Journal of Official Statistics, Stockholm, Vol. 6, Iss. 1, 3.

Rojo, J., Rivero, R., Romero-Morte, J. et al., (2017), "Modeling pollen time series using seasonal-trend decomposition procedure based on LOESS smoothing", Int J Biometeorol 61, 335–348. https://doi.org/10.1007/s00484-016-1215-y

Vijaya, Sharma S. and Batra N., (2019), "Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, pp. 568-573. doi: 10.1109/COMITCon.2019.8862232

Ward, J., (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58(301), pp.236-244.

Wishart, D., (1969), "256. Note: An Algorithm for Hierarchical Classifications." Biometrics, vol. 25, no. 1, pp. 165–170. JSTOR, www.jstor.org/stable/2528688. Accessed 12 Apr. 2021.

Zhou PY. and Chan K.C.C., (2014), "A Model-Based Multivariate Time Series Clustering Algorithm", In: Peng WC. et al. (eds) Trends and Applications in Knowledge Discovery and Data Mining. PAKDD 2014. Lecture Notes in Computer Science, vol 8643. Springer, Cham. https://doi.org/10.1007/978-3-319-13186-3_72