Assessing the Informational Content of Seasonality Tests by Random Forests of Conditional Inference Trees

Daniel Ollech* Karsten Webel†

Abstract

Virtually all seasonal adjustment programs include a variety of statistical tests for assessing whether a given time series is a candidate for seasonal adjustment. However, any ensemble of seasonality tests is certain to be either consistent or inconsistent. In the former case, the tests arrive at the same decision, raising the question if there is a method that is capable of eliminating seemingly redundant tests. In the latter case, the tests arrive at different decisions, raising the question if there is a method that is capable of identifying the most informative tests and finding a final decision. Using simulated seasonal and non-seasonal ARIMA models that are representative of the Deutsche Bundesbank's time series database, we apply random forests of conditional inference trees in order to answer the two questions. In particular, we quantify the informational content of the seasonality tests implemented in JDemetra+ and find that the modified QS- and the Friedman-test yield by far the most informative results.

Key Words: JDemetra+; simulation study; supervised machine learning; test combination

1. Motivation

When deciding whether an observed time series is seasonal or non-seasonal, a variety of statistical tests can be used. For example, in release version 2.1 of JDemetra+ (JD+), the output's diagnostics section reports the results of six different seasonality tests, among many other things. However, any set of seasonality tests is certain to be (1) consistent or (2) inconsistent. In the first case, the tests reach the same decision, raising the questions whether the set is too large and if it can be reduced by eliminating redundant tests. In the second case, the tests reach different decisions, raising the questions if the most informative tests can be identified reliably and how their outcomes can be condensed into a final decision. We show by means of a simulation study that random forests of conditional inference trees provide convenient answers to the questions in either case.

The remainder of this paper is organised as follows. Section 2 briefly describes the seasonality tests currently implemented in JD+. Section 3 provides basic theory of random forests, including the assessment of variable importance. Section 4 describes the simulation algorithm. Section 5 simulates a manifold set of seasonal and non-seasonal ARIMA processes, applies the six JD+ seasonality tests to the simulated series, and grows random forests on the tests' outcome. Furthermore, it compares the misclassification rates between the entire forest and the single seasonality tests, and uses variable importance measures to identify the most informative tests for the ensemble decision on seasonality. Finally, Section 6 draws some conclusions.

*Deutsche Bundesbank, Central Office, Directorate General Statistics, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany

†Deutsche Bundesbank, Central Office, Directorate General Statistics and Research Centre, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany

2. Seasonality tests in JD+

JD+ incorporates six seasonality tests, each of which states absence of seasonality as the null hypothesis (H_0). To provide an overview of these tests, let $\{z_t\}$ be a weakly stationary series of length T and τ the seasonal period of $\{z_t\}$, i.e. $\tau = 12$ for monthly series and $\tau = 4$ for quarterly series.

2.1 Time domain tests

JD+ implements four tests for seasonality in the time domain: the modified QS-test is a Ljung-Box-type test for significance of the seasonal autocorrelations of $\{z_t\}$; the Kruskal-Wallis- and Friedman-tests are ANOVA-type tests on the ranks of $\{z_t\}$; the F-test on seasonal dummies is a regression-based test for significant differences between the period-specific means of $\{z_t\}$.

2.1.1 Modified QS-test

The modified QS-test (QS) checks for significant positive autocorrelation at seasonal lags. Let $\gamma(h) = \mathbb{E}(z_{t+h}\, z_t) - \mathbb{E}^2(z_t)$ and $\rho(h) = \gamma(h)/\gamma(0)$ denote the lag-h autocovariance and autocorrelation, respectively, of $\{z_t\}$. Then, the null hypothesis is specified as H_0 : $\rho(k) \leq 0$ for $k \in \{\tau, 2\tau\}$, and the QS-statistic is obtained as follows: if $\hat{\rho}(\tau) \leq 0$, then $QS = 0$; otherwise,

$$QS = T(T+2)\left(\frac{\hat{\rho}^2(\tau)}{T-\tau} + \frac{\left[\max\{0, \hat{\rho}(2\tau)\}\right]^2}{T-2\tau}\right).$$

The exact null distribution of the QS-statistic is unknown but can be approximated reasonably well by a χ^2-distribution with two degrees of freedom, (Maravall 2011).

2.1.2 Friedman-test

The Friedman-test (FT) checks for significant differences between the period-specific mean ranks of the observations, being essentially a one-way ANOVA with repeated measures. To see this, assume that each period $i \in \{1, \ldots, \tau\}$ has n observations. Furthermore, let r_{ij} be the rank of the observation in the i-th period of the j-th year, where the ranks are assigned separately for each year (i.e. $1 \leq r_{ij} \leq \tau$), and $\mu_i = \mathbb{E}(r_{ij})$. The null hypothesis is then given by $H_0 : \mu_1 = \mu_2 = \cdots = \mu_\tau$, and the test statistic is defined as

$$FT = \frac{\tau - 1}{\tau} \sum_{i=1}^{\tau} \frac{n\left[\bar{r}_i - (\tau+1)/2\right]^2}{(\tau^2-1)/12},$$

where $\bar{r}_i = n^{-1} \sum_j r_{ij}$. Under H_0, the FT-statistic asymptotically follows a χ^2-distribution with $\tau - 1$ degrees of freedom.

2.1.3 Kruskal-Wallis-test

The Kruskal-Wallis-test (KW) basically follows the same idea as the Friedman-test but comes up with two modifications, being essentially a one-way ANOVA without repeated measures. First, period-specific numbers n_i of observations are allowed; second, ranks are

assigned over the entire observation period (i.e. $1 \leq r_{ij} \leq T$). The null hypothesis again reads $H_0 : \mu_1 = \mu_2 = \cdots = \mu_\tau$, and, assuming absence of ties, the test statistic is given by

$$KW = \frac{T-1}{T} \sum_{i=1}^{\tau} \frac{n_i \left[\bar{r}_i - (T+1)/2\right]^2}{(T^2-1)/12}.$$

Under H_0, the KW-statistic asymptotically follows a χ^2-distribution with $\tau - 1$ degrees of freedom.

2.1.4 Seasonal dummies

The F-test on seasonal dummies (SD) checks if the effects of the $\tau - 1$ seasonal dummies are simultaneously zero. Dropping the stationarity assumption on $\{z_t\}$ and assuming absence of additional regression variables, the $(pdq)(000)$ regARIMA model

$$\phi_p(B)(1-B)^d \left(z_t - \sum_{i=1}^{\tau-1} \beta_i D_{i,t} \right) = \mu + \theta_q(B)\, \varepsilon_t$$

is considered, where $D_{i,t} = 1$ if $t = i$, $D_{i,t} = -1$ if $t = \tau$, and $D_{i,t} = 0$ otherwise. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{\tau-1})^\top$. The null hypothesis is then specified as $H_0 : \boldsymbol{\beta} = \mathbf{0}$, and the test statistic is given by

$$SD = \frac{\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}^{-1} \hat{\boldsymbol{\beta}}}{\tau - 1} \cdot \frac{T - d - p - q - \tau - 1}{T - d - p - q},$$

where $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$. Under H_0, the SD-statistics follows an F-distribution with $\tau - 1$ and $T - d - p - q - \tau - 1$ degrees of freedom. Two variants of non-seasonal orders are included in JD+ first, $(pdq) = (011)$ is used; second, (pdq) is determined via automatic model identification.

2.2 Frequency domain tests

JD+ implements two tests for seasonality in the frequency domain: the periodogram-test evaluates a weighted sum of the periodogram of $\{z_t\}$ at the seasonal frequencies; the dummy-type test for seasonal peaks is based on visually significant peaks at the seasonal frequencies in the Tukey and AR(30) spectra of $\{z_t\}$.

2.2.1 Periodogram-test

We follow Brockwell and Davis (1991) and define the discrete periodogram in terms of the discrete Fourier transform of the observed time series. To this end, let $\omega_j = 2\pi j/T$ be the j-th Fourier frequency satisfying $-\pi < \omega_j \leq \pi$, i.e. $j \in \{\lfloor (T-1)/2 \rfloor, \ldots, \lfloor T/2 \rfloor\}$, where $\lfloor x \rfloor$ denotes the integer part of x. Then, the discrete periodogram at frequency ω_j is defined as

$$I(\omega_j) = \frac{1}{T} \left| \sum_{t=1}^{T} z_t e^{-it\omega_j} \right|^2, \tag{1}$$

which is closely related to the empirical autocorrelation function $\{\hat{\gamma}(h)\}$. More precisely, equation 1 is equivalent to the following representation:

$$I(\omega_j) = \begin{cases} \sum_{|h| \leq T} \hat{\gamma}(h) e^{-ih\omega_j}, & \omega_j \neq 0 \\ T\,|\bar{z}|^2, & \omega_j = 0 \end{cases}, \tag{2}$$

where $\bar{z} = T^{-1}\sum_{t=1}^{T} z_t$. Equation 2 can be used to construct the periodogram-based seasonality test. To this end, let $\omega_j^\star = 2\pi j/\tau$ denote the j-th seasonal frequency for $j \in \{1,\dots,\tau/2\}$, $\mathcal{S}(\tau) = \{\omega_1^\star,\dots,\omega_{\tau/2}^\star\}$ the set of seasonal frequencies for any given τ, and

$$\Sigma_{\mathcal{S}(\tau)} = 2\sum_{j=1}^{\tau/2-1} I(\omega_j^\star) + I(\omega_{\tau/2}^\star)\cdot \mathbb{1}_{\{T \text{ even}\}}$$

the weighted sum of the periodogram evaluated at the seasonal frequencies, where $\mathbb{1}$ is the indicator function of the event in braces. The null hypothesis of the periodogram-test then reads $H_0 : \Sigma_{\mathcal{S}(\tau)} = 0$, and the test statistic is given by

$$PD = \frac{T-\tau}{\tau-1}\cdot\frac{\Sigma_{\mathcal{S}(\tau)}}{\sum_{t=1}^{T} z_t^2 - I(0) - \Sigma_{\mathcal{S}(\tau)}}.$$

Under H_0, the PD-statistic follows an F-distribution with $\tau - 1 - \mathbb{1}_{\{T \text{ even}\}}$ and $T - \tau + \mathbb{1}_{\{T \text{ even}\}}$ degrees of freedom.

The discrete periodogram can be extended to the continuous periodogram by defining $I(\omega) = T^{-1}\left|\sum_{t=1}^{T} z_t e^{-it\omega}\right|^2$ for all $\omega \in [-\pi,\pi]$. However, the discrete periodogram has refined statistical properties at the Fourier frequencies, which do not necessarily apply to other frequencies. In either case, a periodogram estimator of the spectral density $f(\omega) = (2\pi)^{-1}\sum_h \gamma(h)\, e^{-ih\omega}$ is given by

$$\hat{f}_{Per}(\omega) = \frac{I(\omega)}{2\pi},$$

which is asymptotically unbiased for all $\omega \in [-\pi,\pi]$ if $\mathbb{E}(z_t) = 0$ and for all $\omega \neq 0$ if $\mathbb{E}(z_t) \neq 0$ but not consistent regardless of $\mathbb{E}(z_t)$ (Brockwell and Davis 1991).

2.2.2 Seasonal peaks

Since the test for seasonal peaks (SP) combines information from the Tukey and AR(30) spectra of $\{z_t\}$, we first introduce the two estimators of $f(\omega)$ as well as respective criteria for calling a spectral peak visually significant.

The Tukey spectrum is a non-parametric "lag window" estimator. To transform equation 2 into a consistent estimator of $f(\omega)$, the general idea of "lag window" estimators is to put relatively more weight on smaller lags of $\gamma(h)$, which are considered to be more reliable, and relatively less weight on higher lags of $\gamma(h)$, which are considered to be less reliable. For that purpose, an even and piecewise continuous window function $w(\cdot)$ is introduced which satisfies the following three conditions: (1) $w(0) = 1$, (2) $|\,w(x)\,| \leq 1$ for all $x \in \mathbb{R}$, and (3) $w(x) = 0$ for $|\,x\,| > 1$. The Tukey spectrum is then defined as

$$\hat{f}_T(\omega) = \frac{1}{2\pi}\sum_{|h|\leq H} w_a(h/H)\,\hat{\gamma}(h)\,e^{-ih\omega},$$

where $w_a(\cdot)$ is the Blackman-Tukey window given by

$$w_a(x) = \begin{cases} 1 - 2a + 2a\cos(\pi x), & |\,x\,| \leq 1 \\ 0, & |\,x\,| > 1 \end{cases}$$

with $a \in [0, 0.25]$ and H is any truncation lag, not necessarily the number of observations available, T. A peak at any Fourier frequency ω_j is called visually significant at the α-level of significance if

$$\frac{2\hat{f}_T(\omega_j)}{\hat{f}_T(\omega_{j-1}) + \hat{f}_T(\omega_{j+1})} \geq F_{d_1,d_2,1-\alpha},$$

where $F_{d_1,d_2,1-\alpha}$ is the critical value of the F-distribution with d_1 and d_2 degrees of freedom, which are determined empirically via simulations described by Maravall (2011).

The AR(30) spectrum is a parametric "plug in" estimator. The basic idea of this class of estimators is to choose a particular time series model for $\{z_t\}$, derive its theoretical spectrum $f(\omega)$, and replace the unknown parameters in $f(\omega)$ with well-established estimators. In general, the spectrum of an autoregressive (AR) process of order $p > 0$ is given by

$$f(\omega) = \frac{\sigma_\varepsilon^2}{2\pi} \left| 1 - \sum_{h=1}^{p} \phi_h e^{-ih\omega} \right|^{-2},$$

where σ_ε^2 is the variance of the white noise process driving the AR process. The AR(30) spectrum is then given by

$$\hat{f}_{AR}(\omega) = \frac{\hat{\sigma}_\varepsilon^2}{2\pi} \left| 1 - \sum_{h=1}^{30} \hat{\phi}_h e^{-ih\omega} \right|^{-2},$$

where $\hat{\sigma}_\varepsilon^2$ and $\hat{\phi}_h$ are some estimators of the white noise's variance and the AR coefficients, respectively (Priestley 1981). The choice of 30 as the truncation lag is justified pragmatically by Soukup and Findley (1999) who argue that "this choice [...] can potentially produce the largest number of peaks possible, i.e. 30, in a plot with 61 frequencies. Thus, it has the greatest resolving power." A peak at any Fourier frequency ω_j is called visually significant if (1) $\hat{f}_{AR}(\omega_j)$ is larger than the median AR spectrum of all Fourier frequencies and (2) the quantity

$$\frac{\hat{f}_{AR}(\omega_j) - \max\left\{\hat{f}_{AR}(\omega_{j-1}), \hat{f}_{AR}(\omega_{j+1})\right\}}{\max_j \hat{f}_{AR}(\omega_j) - \min_j \hat{f}_{AR}(\omega_j)}$$

is larger than some critical value which may be set to $6/52$ for all frequencies (i.e. the X-12-ARIMA default) or be chosen individually for each frequency ω_j. As a compromise, Maravall (2011) provides critical values based on a large-scale simulation of random walk processes and suggests to use the critical value associated with the first TD frequency for all frequencies.

For $\tau = 12$, $\{z_t\}$ is now said to have seasonal peaks, giving $SP = 1$, if visually significant peaks show up in[1]

(1) $\hat{f}_T(\omega)$ OR $\hat{f}_{AR}(\omega)$ at four or more frequencies $\omega_j^\star$,

(2) $\hat{f}_T(\omega)$ OR $\hat{f}_{AR}(\omega)$ at three frequencies $\omega_j^\star$ PLUS in $\hat{f}_T(\omega)$ AND $\hat{f}_{AR}(\omega)$ at one or more frequency $\omega_j^\star$,

(3) $\hat{f}_T(\omega)$ OR $\hat{f}_{AR}(\omega)$ at three frequencies $\omega_j^\star$ PLUS there is no peak at $\omega_6^\star$,

(4) $\hat{f}_T(\omega)$ AND $\hat{f}_{AR}(\omega)$ at $\omega_6^\star$ and another frequency $\omega_j^\star$,

(5) $\hat{f}_T(\omega)$ OR $\hat{f}_{AR}(\omega)$ at two or more frequencies $\omega_j^\star$ INCLUDING in $\hat{f}_T(\omega)$ AND $\hat{f}_{AR}(\omega)$ at one frequency $\omega_j^\star$ PLUS there is no peak at $\omega_6^\star$.

Accordingly, the null hypothesis is specified as $H_0 : SP = 0$. When assessing visual significance of spectral peaks, $\alpha = 0.1$ is always used for the Tukey and AR(30) spectra.

[1] For quarterly series, a similar but smaller set of rules applies which is not discussed here.

3. Random forests

The growing literature on statistical and machine learning has provided a vast amount of forecasting methods and algorithms. While several of these could be used to combine the results of a set of seasonality tests, random forest has some favourable properties:

(1) Random forest is among the best performing algorithms for a broad range of prediction and classification tasks, such as forecasting stock price movements (Patel et al. 2015), diagnosis of diseases (Hsieh et al. 2011), and detection of phishing emails (Almomani et al. 2013).

(2) Measures to quantify the variable importance are readily available (Archer and Kimes 2008).

(3) Using random forests based on conditional classification trees (Strobl et al. 2007, 2008), we can take into account high correlations among predictors that otherwise would bias the estimation of variable importance.

(4) Random forests consists of a large set of classification or regression trees. Their output can be depicted as decision trees, which is an ideal basis for combining seasonality tests conditional on their p-values in a non-linear, interpretable fashion.

3.1 Original approach

Random forest is an ensemble method that has been developed by Breiman (2001). It is based on bootstrap aggregation (bagging) developed by Breiman (1996) and further analysed by Bühlmann and Yu (2002) and applicable equally well to regression and classification problems. Restricting ourselves to the latter case, the basic idea is to combine a large and diverse set of unpruned binary classification trees built upon bootstrap samples of the training data. Thereby, aggregation smoothes the hard cut decisions of the binary splits in a single tree and, usually, results in an improvement in classification accuracy. Further information on classification trees and statistical/machine learning is provided by Breiman et al. (1984) and Hastie et al. (2009).

Let $\mathcal{L} = (\mathbf{XY})$ be the training or learning data, where $\mathbf{X} = (\mathbf{X}_1 \dots \mathbf{X}_p)$ is a set of p predictors with $\mathbf{X}_j = (x_{1j}, \dots, x_{Nj})^\top$ for all $j \in \{1, \dots, p\}$, and $\mathbf{Y} = (y_1, \dots, y_N)^\top$ is a vector of categorical responses with $y_i \in \{1, \dots, K\}$ for all $i \in \{1, \dots, N\}$. For $b \in \{1, \dots, B\}$, a bootstrap sample $\mathcal{L}_b$ is drawn with replacement from $\mathcal{L}$, and an unpruned classification tree $\mathcal{T}_b$ is grown for the sample. To this end, assume that $\mathcal{T}_b$ currently has M terminal nodes corresponding to M classification regions R_m. To create a binary split of any terminal node $m \in \{1, \dots, M\}$, a random sample is drawn from the set of p predictors. Here, the intention of random sampling is to prevent strong predictors from dominating all other predictors, which in turn increases the diversity among the single trees compared to bagging, where all predictors are considered at each split. For each sampled predictor, the best split of node m among all possible splits is determined, and the predictor that generates the best split in the sample is eventually chosen as splitting variable for node m. Thereby, best splits are identified by means of minimum node impurity. Let $\hat{p}_{mk} = N_m^{-1} \sum_{\mathbf{x}_i \in R_m} \mathbb{1}_{\{y_i = k\}}$ be the proportion of training data in node m from class k, where $N_m = \sum_i \mathbb{1}_{\{\mathbf{x}_i \in R_m\}}$ and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ denotes the i-th observation of the p predictors. A popular impurity measure is given by the Gini index $Q_m(\mathcal{T}_b) = \sum_k \hat{p}_{mk} \, (1 - \hat{p}_{mk})$. Tree growing is stopped if the minimum number of observations in terminal nodes is reached or the node impurity cannot be reduced further with

the given predictors. The forest classification is finally given by the unweighted majority vote of the tree classifications.

A celebrated advantage of random forests is the possibility of using subsets of the training data for validation purposes. Let $\mathcal{O}_b = \mathcal{L} \backslash \mathcal{L}_b$ be the "out-of-bag" (OOB) data of the b-th bootstrap sample, that is the training data not selected in $\mathcal{L}_b$. The forest's performance can then be judged by means of misclassification rates in the OOB samples. Alternatively, external validation (VAL) data can be considered as usual.

3.2 Variable importance

Another advantage of random forests is the capability to provide information about the predictors' importance with respect to the ensemble decision. For an individual classification tree, the importance of a given predictor is determined directly by the predictor's position in the tree. However, this concept does not apply to random forests in a straightforward way. Therefore, variable importance measures have been suggested to identify strong predictors. A popular measure is given by the forest's mean decrease of node impurity caused by the predictor. Let $\mathcal{M}(\mathcal{T}_b, \mathbf{X}_j)$ be the set of nodes in $\mathcal{T}_b$ that were split by $\mathbf{X}_j$ and $M_j = \sum_b |\mathcal{M}(\mathcal{T}_b, \mathbf{X}_j)|$ the respective total number of nodes in the forest. Measuring node impurity by the Gini index, the variable importance of $\mathbf{X}_j$ is given by

$$\mathrm{VI}^G(\mathbf{X}_j) = \frac{1}{M_j} \sum_{b=1}^{B} \sum_{m \in \mathcal{M}(\mathcal{T}_b, \mathbf{X}_j)} \left\{ Q_m(\mathcal{T}_b) - \left[\frac{N_{m_L}}{N_m} Q_{m_L}(\mathcal{T}_b) + \frac{N_{m_R}}{N_m} Q_{m_R}(\mathcal{T}_b) \right] \right\},$$

where m_L and m_R are the left and right descendent nodes of m.

Alternatively, variable importance can be measured by the mean decrease in prediction accuracy after randomly permutating the values of the predictor in the OOB samples. The rationale of this approach is that random permutation mimics absence of the predictor. Let $\hat{y}_i(\mathcal{T}_b, \mathbf{X}_j)$ and $\hat{y}_i(\mathcal{T}_b, \mathbf{X}_{\pi(j)})$ denote the predicted classes of y_i obtained from $\mathcal{T}_b$ before and after random permutation of the values of $\mathbf{X}_j$ in $\mathcal{O}_b$. The permutation-based variable importance of $\mathbf{X}_j$ is then given by

$$\mathrm{VI}^P(\mathbf{X}_j) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i \in \mathcal{O}_b} \left[\frac{\mathbb{1}_{\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{X}_{\pi(j)})\}}}{|\mathcal{O}_b|} - \frac{\mathbb{1}_{\{y_i \neq \hat{y}_i(\mathcal{T}_b, \mathbf{X}_j)\}}}{|\mathcal{O}_b|} \right]. \tag{3}$$

Sometimes, this measure is normalised using the standard deviation of the differences between the misclassification rates in the OOB samples, though the unscaled version usually should be preferred (Diaz-Uriarte and de Andrés 2006).

3.3 Conditional inference trees

Random forest based on conditional inference trees (or conditional random forests) deviate from the original approach in two respects: variable selection and variable importance measures. In classical random forests, variable selection tends to be biased towards predictors with larger measurement scales, a higher number of categories and, sometimes, missing values (Hothorn et al. 2006, Strobl et al. 2009). Variable importance measures are likely to be biased in the same cases as well as in the presence of correlated predictors (Strobl et al. 2007, 2008).

Regarding variable selection, Hothorn et al. (2006) develop a conditional inference framework for decision trees which avoids potential biases by separating variable selection and node splitting. The rationale of this separation is an ex ante exclusion of those predictors $\mathbf{X}_j$ which are not strongly related to the response $\mathbf{Y}$. For any terminal node $m \in \{1, \ldots, M\}$, they propose the following generic algorithm:

1. Test the global null hypothesis of independence between $\mathbf{Y}$ and any $\mathbf{X}_j$. Stop if this hypothesis cannot be rejected. Otherwise, identify the predictor $\mathbf{X}_{j*}$ with the strongest association to $\mathbf{Y}$.

2. Take $\mathbf{X}_{j*}$ as splitting variable. Find the optimal binary split of node m using a prespecified splitting criterion.

In Step 1, the association between $\mathbf{Y}$ and any $\mathbf{X}_j$ is measured by means of standardised linear statistics within the permutation test framework developed by Strasser and Weber (1999). In Step 2, any splitting criterion can be considered in principle. However, Hothorn et al. (2006) suggest using two-sample linear statistics which are in line with the criteria applied in Step 1 of the generic algorithm.

Regarding variable importance measures, Strobl et al. (2008) develop a conditional permutation scheme which avoids potential biases by taking the correlation structure among the predictors into account. The aim of this scheme is to prevent ex ante the overestimation of seemingly influential predictors $\mathbf{X}_j$ that in fact are not strongly associated with $\mathbf{Y}$ but appear as though due to a high correlation with a truly influential predictor, such as $\mathbf{X}_{j*}$. To this end, the original permutation scheme $\pi(\cdot)$ which underlies equation 3 is applied to the values of $\mathbf{X}_j$ only within subgroups of observations of $\mathbf{X}_j^c = (\mathbf{X}_1 \dots \mathbf{X}_{j-1}\mathbf{X}_{j+1} \dots \mathbf{X}_p)$, resulting in the conditional permutation scheme $\pi(\cdot)|\mathbf{X}_\cdot^c$. The respective conditional permutation-based variable importance measure is given by

$$\text{VI}^{CP}(\mathbf{X}_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i\in\mathcal{O}_b}\left[\frac{\mathbb{1}_{\{y_i\neq\hat{y}_i(\mathcal{T}_b,\mathbf{X}_{\pi(j)|\mathbf{X}_j^c})\}}}{|\mathcal{O}_b|} - \frac{\mathbb{1}_{\{y_i\neq\hat{y}_i(\mathcal{T}_b,\mathbf{X}_j)\}}}{|\mathcal{O}_b|}\right], \tag{4}$$

where for each tree $\mathcal{T}_b$ the permutation grid for $\mathbf{X}_j$ is defined by the cut-points of $\mathbf{X}_j^c$ in $\mathcal{T}_b$. Thus, the conditional variable importance measure is feasible for both categorical and continuous predictors.

As a final remark, conditioning on $\mathbf{X}_j^c$ in the permutation scheme might be seen as a very conservative strategy. As an alternative, only those predictors in $\mathbf{X}_j^c$ whose correlation with $\mathbf{X}_j$ exceeds a certain threshold could be used. To this end, the association measures calculated in Step 1 of the generic algorithm developed by Hothorn et al. (2006) may give an intuition about which predictors could be considered. Either way, it should also be kept in mind that the differences between classical and conditional random forests primarily concern variable importance measures and tend to be negligible in terms of misclassification rates (Hothorn et al. 2006, Webel and Ollech 2017).

4. Simulation algorithm

We aim at simulating ARIMA processes that are representative of the non-seasonal and seasonal economic data of the Deutsche Bundesbank's time series database. More precisely, given n observed time series that follow the same ARIMA model of order $(pdq)(PDQ)$, we wish to simulate $\tilde{\nu}$ versions of this model under the restrictions that the ARMA parameters of the simulated models should have the same multivariate distribution as the estimated ARMA parameters of the models fitted to the observed data. As it is difficult to determine the exact family of distributions of the latter parameters, we impose the proxy restrictions that the former ARMA parameters

(1) do not induce (additional) unit roots in the characteristic polynomial of the simulated ARIMA model,

(2) display the same correlation structure as the estimated ARMA parameters of the ARIMA models fitted to the observed time series,

(3) follow the same univariate distribution.

To meet these proxy restrictions, we assorted the following procedure which combines the "NORmal-To-Anything" (NORTA) algorithm of Cario and Nelson (1997) with logspline density estimation as described by Stone et al. (1997):

1. Set $m = p + q + P + Q$ and let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be the matrix of the estimated ARMA coefficients. Calculate $\mathbf{\Sigma_X} \in \mathbb{R}^{m \times m}$, the correlation matrix of the coefficients.

2. Apply logspline density estimation to each row of $\mathbf{X}$ to obtain a non-parametric estimate $\hat{f}_j(\cdot)$ of the density of the j-th ARMA coefficient, where $j \in \{1, \dots, m\}$.

3. Set $\mathbf{\Sigma}_\mathbf{Y}^{(1)} = \mathbf{\Sigma_X}$ to initialise the simulation of ARMA coefficients, where $\mathbf{Y} \in \mathbb{R}^{m \times \nu}$ denotes an empty matrix to be filled during the following recursion.

4. Start of recursion. In the i-th loop, simulate $\nu \gg \tilde{\nu}$ independent coefficient vectors $\mathbf{Y}_j \in \mathbb{R}^m$, where $\mathbf{Y}_j \sim \mathcal{N}\left(\mathbf{0}_m, \mathbf{\Sigma}_\mathbf{Y}^{(i)}\right)$ for each $j \in \{1, \dots, n\}$. Set $\mathbf{Y} = (\mathbf{Y}_1 \dots \mathbf{Y}_\nu)$.

5. Define $\mathbf{Z} = (z_{jk}) \in \mathbb{R}^{m \times \nu}$, where $z_{jk} = \hat{F}_j^{-1}[\Phi(y_{jk})]$ for all $(j, k) \in \{1, \dots, m\} \times \{1, \dots, \nu\}$ and $\hat{F}_j(\cdot)$ and $\Phi(\cdot)$ are the distribution functions of $\hat{f}_j(\cdot)$ and the standard normal distribution, respectively.

6. Let $l \in \{0, \dots, \nu\}$ be the number of columns of $\mathbf{Z}$ which contain ARMA coefficients that induce unit roots. Remove the l columns from $\mathbf{Z}$ to obtain $\tilde{\mathbf{Z}} \in \mathbb{R}^{m \times (\nu - l)}$, the matrix of admissible ARMA coefficients.

7. Select $\tilde{\nu}$ columns from $\tilde{\mathbf{Z}}$ according to simple random sampling without replacement, where $\tilde{\nu} \in \{1, \dots, \nu - l\}$. Store the sampled columns in $\tilde{\mathbf{Z}}^{(\tilde{\nu})} \in \mathbb{R}^{m \times \tilde{\nu}}$.

8. Calculate $\mathbf{\Sigma}_{\tilde{\mathbf{Z}}^{(\tilde{\nu})}} \in \mathbb{R}^{m \times m}$, the correlation matrix of the sampled admissible ARMA coefficients.

9. Calculate $\mathbf{\Delta} = \left|\mathbf{\Sigma_X} - \mathbf{\Sigma}_{\tilde{\mathbf{Z}}^{(\tilde{\nu})}}\right| = (\delta_{jk})$. Define $\mathbf{C_\Delta} = (c_{jk})$, where $c_{jk} = \mathbb{1}_{\{\delta_{jk} > \varepsilon\}}$ for all $(j, k) \in \{1, \dots, m\}^2$ and some $\varepsilon > 0$. If $\mathbf{C_\Delta} \neq \mathbf{0}$, restart Step 4 with $\mathbf{\Sigma}_\mathbf{Y}^{(i+1)} = \mathbf{\Sigma}_\mathbf{Y}^{(i)} + \alpha \left[\mathbf{\Sigma_X} - \mathbf{\Sigma}_{\tilde{\mathbf{Z}}^{(\tilde{\nu})}}\right] \odot \mathbf{C_\Delta}$, where $\alpha > 0$ and $\odot$ denotes the Hadamard product of two matrices, i.e. $\mathbf{A} \odot \mathbf{B} = (a_{jk} \cdot b_{jk})$. If $\mathbf{C_\Delta} = \mathbf{0}$, or the maximum number of iterations is reached, end of recursion.

10. Simulate $\tilde{\nu}$ ARIMA models of order $(pdq)(PDQ)$ with the coefficients stored in the columns of $\tilde{\mathbf{Z}}^{(\tilde{\nu})}$.

5. Application

To simulate ARIMA models that are representative of the Deutsche Bundesbank's time series database, we draw a random sample of almost $14{,}000$ seasonal and non-seasonal monthly time series without replacement from the database and identify the ARIMA model of each sampled series using the automatic identification routine of JD+. For each identified model m of "seasonality class" $k \in \{\text{N-S}, \text{S}\}$, where N-S and S denote the class of non-seasonal and seasonal models, respectively, we calculate the model's share p_{mk} among

Table 1: Misclassification rates as a percentage (N-S = non-seasonal series, S = seasonal series).

			Simulated ARIMA series							
			All lengths		5-year		10-year		20-year	
Classifier			N-S	S	N-S	S	N-S	S	N-S	S
Conditional random forests		OOB	0.8	2.2	0.7	2.5	0.9	2.0	0.9	2.0
		VAL	0.8	2.2	0.7	2.4	0.9	2.1	0.9	2.1
Seasonality test	$\alpha = 0.01$	QS	4.8	1.5	2.5	1.7	5.0	1.4	7.1	1.3
		FT	2.1	2.1	1.5	2.2	2.2	2.0	2.5	2.1
		KW	2.4	3.8	1.9	3.9	2.6	3.7	2.7	3.7
		PD	3.2	3.6	3.2	3.4	3.3	3.6	3.2	3.9
		SD	4.0	2.7	4.4	2.5	4.1	2.7	3.7	2.9
		SP	$*$	$*$	$*$	$*$	6.6	1.7	5.0	2.0
	$\alpha = 0.05$	QS	7.4	1.2	4.9	1.4	7.5	1.1	9.8	1.2
		FT	6.6	1.6	5.6	1.6	7.0	1.5	7.3	1.6
		KW	6.9	3.1	6.2	3.1	7.1	3.1	7.4	3.2
		PD	8.1	3.2	8.3	2.8	8.2	3.1	8.0	3.5
		SD	9.1	2.2	9.7	2.0	9.2	2.2	8.5	2.5
		SP	$*$	$*$	$*$	$*$	6.6	1.7	5.0	2.0

all models in the same class and the simulation weights $w_{mk} = \tilde{p}_{mk} / \sum_m \tilde{p}_{mk}$, where $\tilde{p}_{mk} = p_{mk} \cdot \mathbb{1}_{\{p_{mk} \geq 0.01\}}$. For each model m of class k and each length $N \in \{60, 120, 240\}$, we run the NORTA algorithm with $\tilde{\nu} = 100{,}000 \cdot w_{mk}$, $\nu = 100{,}000$, $\varepsilon = 0.02$ and $\alpha = 0.5$, yielding 600,000 simulated ARIMA time series in total.

For each simulated series, we calculate the six seasonality tests in JD+, where we restrict ourselves to the version $(pdq) = (011)$ of the F-test on seasonal dummies. The p-values of the seasonality tests are then used as predictors for the random forests.[2] More precisely, we randomly draw 50 independent training data sets of size 7,500 from the 600,000 simulated ARIMA models, keeping the respective non-sampled models for validation purposes. As the empirical correlation between any two p-values is larger than 0.70, we grow a random forest of conditional inference trees for each training data set. Each random forest consists of $B = 100$ trees.[3] For each single tree, $\lfloor \sqrt{p} \rfloor = 2$ predictors are considered at each split[4], and the minimum size of terminal nodes is set to one.

Table 1 reports the misclassification rates of the candidate seasonality tests[5] and the conditional random forests, where the OOB and VAL misclassification rates are averaged over the respective 50 forests. In general, the modified QS-test seems to be the best test for identifying seasonal series, whereas the Friedman-test performs particularly well at detecting non-seasonal series. Interestingly, the misclassification rates of the modified QS-, the Friedman- and the Kruskal-Wallis-test increase with the length for non-seasonal series,

[2] Note that in JD+ the test on seasonal peaks cannot be calculated for monthly time series with less than seven years of observations. For that reason, we consider only the lengths $N \in \{120, 240\}$ for this test.

[3] We also grew larger forests but did not observe a significant decrease of the OOB misclassification rates.

[4] Grömping (2009) notes for regression problems that increasing the number of candidate predictors at each split will make the variable importance measures more conditional, i.e. the dependence structure among the predictors is taken more strongly into account. In our case, even including all predictors at each split did barely change the calculated variable importance measures.

[5] Note that in practice the estimation of the sample autocorrelations is sensitive to the order of differencing. Correspondingly, the QS-statistic is biased in the case of under-differencing. In JD+, first differences are taken once and the series is mean-adjusted, whereas in X-13ARIMA-SEATS and TRAMO/SEATS the default order of differencing is $\max\{1, \min\{d + D, 2\}\}$ (U.S. Census Bureau 2016, Maravall 2011).

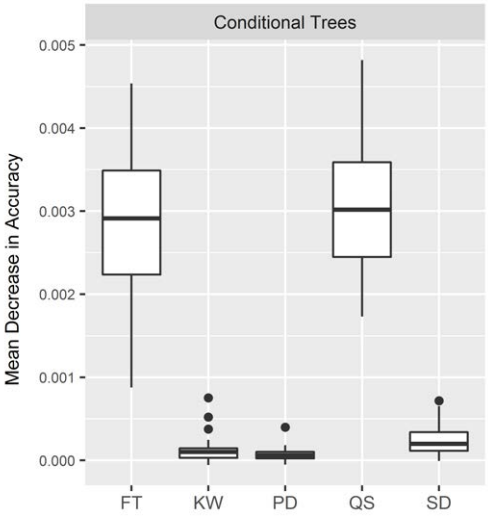

Figure 1: Boxplots of the seasonality tests' importance based on sets of 50 conditional random forests.

while the same is true for the misclassification rates of the periodogram test and the F-test on seasonal dummies for seasonal series. The average misclassification rates of the conditional random forests are universally lower for non-seasonal series than the rates of any single seasonality test. For seasonal series, they are slightly above the best candidate test. Overall, the performance of the conditional random forests is less dependent on the length of the simulated series than for the seasonality tests.

Using the 50 training data sets, figure 1 shows boxplots of the seasonality tests' mean decrease in accuracy as defined in equation 4. The modified QS-test slightly outperforms the Friedman-test in the battle for the most informative seasonality test, while the other three tests are less important by far. The test for seasonal peaks is not considered in this exercise since all lengths of simulated time series are used to obtain the mean decrease in accuracy.

6. Summary

We showed by means of a large-scale simulation study that random forests conveniently solve the problem of combining seasonality tests as they improve prediction accuracy and provide insights into the informational content of the candidate tests. In particular, we highlighted that unbiased variable importance measures, which can be obtained by utilising random forests of conditional inference trees, identify the modified QS- and Friedman-tests as the most reliable seasonality tests. An intuitive explanation may be that the Friedman-test mainly covers stable seasonality, while the modified QS-test allows for a higher degree of flexibility in the seasonal component. Future research could use these findings to construct an overall seasonality test based on variable selection within the conditional inference framework.

Acknowledgement

We thank Nina Gonschorreck and Christiane Hofer for technical support.

REFERENCES

Almomani, A., Gupta, B., Atawneh, S., Meulenberg, A. & Almomani, E. A. (2013). Survey of Phishing Email Filtering Techniques. IEEE Communications Surveys & Tutorials 15 (4), 2070-2090.

Archer, K. J. & Kimes, R. V. (2008). Empirical Characterization of Random Forest Variable Importance Measures. Computational Statistics and Data Analysis 52 (4), 2249-2260.

Breiman, L. (1996). Bagging Predictors. Machine Learning 24 (2), 123-140.

Breiman, L. (2001). Random Forests. Machine Learning 45 (1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Chapman & Hall.

Brockwell, P. J. & Davis, R. A. (1991). Time Series. Theory and Methods. Springer.

Bühlmann, P. & Yu, B. (2002). Analyzing bagging. The Annals of Statistics 30 (4), 927-961.

Cario, M. C. & Nelson, B. L. (1997). Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. Technical report, Northwestern University.

Diaz-Uriarte, R. & de Andrés, S. A. (2006). Gene Selection and Classification of Microarray Data using Random Forest. BMC Bioinformatics 7, Article 3.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician 63 (4), 308-319.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition. Springer.

Hothorn, T., Hornik, K. & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics 15 (3), 651-674.

Hsieh, C.-H., Lu, R.-H., Lee, N.-H., Chiu, W.-T., Hsu, M.-H. & Li, Y.-C. J. (2011). Novel Solutions for an Old Disease: Diagnosis of Acute Appendicitis with Random Forest, Support Vector Machines, and Artificial Neural Networks. Surgery 149 (1), 87-93.

Maravall, A. (2011). Seasonality Tests and Automatic Model Identification in TRAMO-SEATS. Mimeo, Bank of Spain.

Patel, J., Shah, S., Thakkar, P. & Kotecha, K. (2015). Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques. Expert Systems with Applications 42 (1), 259-268.

Priestley, M. (1981). Spectral Analysis and Time Series. Academic Press.

Soukup, R. J. & Findley, D. F. (1999). On the Spectrum Diagnostics Used by X-12-ARIMA to Indicate the Presence of Trading Day Effects After Modeling or Adjustment. Proceedings of the American Statistical Association, Business and Economic Statistics Section, 144-149.

Stone, C. J., Hansen, M. H., Kooperberg, C. & Truong, Y. K. (1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling. The Annals of Statistics 25 (4), 1371-1470.

Strasser, H. & Weber, C. (1999). On the Asymptotic Theory of Permutation Statistics. Mathematical Methods of Statistics 8, 220-250.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. & Zeileis, A. (2008). Conditional Variable Importance for Random Forests. BMC Bioinformatics 9, Article 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A. & Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics 8, Article 25.

Strobl, C., Malley, J. & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. Psychological Methods 14 (4), 323-348.

U.S. Census Bureau (2016). X-13ARIMA-SEATS Reference Manual Version 1.1.

Webel, K. & Ollech, D. (2017). Condensing Information from Multiple Seasonality Tests with Random Forests. Proceedings of the 61st ISI World Statistics Congress, forthcoming.