DEUTSCHE
BUNDESBANK
EUROSYSTEM

Discussion Paper

Deutsche Bundesbank
No 55/2020

A random forest-based approach to identifying the most informative seasonality tests

Daniel Ollech
Karsten Webel

Discussion Papers represent the authors' personal opinions and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

Editorial Board: Daniel Foos
Stephan Jank
Thomas Kick
Malte Knüppel
Vivien Lewis
Christoph Memmel
Panagiota Tzamourani

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de

Reproduction permitted only if source is stated.

ISBN 978–3–95729–779–2 (Printversion)
ISBN 978–3–95729–780–8 (Internetversion)

Non-technical summary

Research Question

Monitoring economic developments is often based on both seasonally adjusted and unadjusted data. Testing observed time series for the presence of seasonality is therefore an important part of economic analyses. Different seasonality tests agree in the majority of cases but also give contradictory results for a non-negligible number of time series. This raises the following questions: how can the conflicting results be settled, and how can the most reliable tests be identified?

Contribution

We treat the identification of the seasonal status of a given time series as a classification task and the outcome of the different seasonality tests as predictors. Therefore, we can use machine learning methods for this classification task and compare them to find the one which best balances high accuracy, high interpretability and availability of unbiased variable importance measures. The methods are applied to simulated data that are representative of the Bundesbank's macroeconomic time series database.

Results

Our analysis reveals tree-based methods in general and random forest variants in particular to be the most appropriate methods. The latter especially stand out due to misclassification rates which are acceptably low as well as barely affected by the time series' lengths, as opposed to some single seasonality tests. However, only one random forest variant is capable of providing unbiased variable importance measures even for correlated predictors. This variant finally identifies the modified QS and Friedman tests as the most informative seasonality tests among the considered candidates.

Nichttechnische Zusammenfassung

Fragestellung

Konjunkturbeobachtungen beruhen oft sowohl auf saisonbereinigten als auch auf unbereinigten Daten. Die Anwendung verschiedener Tests zur Erkennung saisonaler Muster in diesen Daten ist daher ein wichtiger Bestandteil volkswirtschaftlicher Analysen. Diese Saisontests liefern meistens übereinstimmende Ergebnisse, kommen aber bei einer nicht unerheblichen Zahl von Zeitreihen auch zu widersprüchlichen Einschätzungen. Dies wirft folgende Fragen auf: Wie können solche verschiedenen Testergebnisse sinnvoll zusammengefasst werden, und wie lassen sich die zuverlässigsten Saisontests identifizieren?

Beitrag

Wir betrachten die Erkennung saisonaler Muster in Zeitreihen als Klassifikationsproblem und die Ergebnisse der verschiedenen Saisontests als Prädiktoren. Wir können daher Methoden des maschinellen Lernens zur Klassifikation verwenden und vergleichen ihre Fähigkeiten hinsichtlich eines möglichst ausgewogenen Verhältnisses zwischen hoher Genauigkeit, guter Interpretierbarkeit und Verfügbarkeit unverzerrter Maße zum Informationsgehalt der Prädiktoren. Die Methoden werden dabei auf simulierte Daten angewendet, die repräsentativ für die makroökonomische Zeitreihendatenbank der Bundesbank sind.

Ergebnisse

Als geeignete Verfahren erweisen sich im Allgemeinen auf Klassifikationsbäumen aufbauende Methoden und im Speziellen Random-Forest-Varianten. Letztere zeichnen sich im Vergleich zu einzelnen Saisontests vor allem durch geringere und gegenüber der Zeitreihenlänge robustere Fehlklassifikationsraten aus. Allerdings kann nur eine Variante den Informationsgehalt auch für korrelierte Prädiktoren unverzerrt quantifizieren. Diese Variante identifiziert schließlich den modifizierten QS- und den Friedman-Test als aussagekräftigste unter allen berücksichtigten Saisontests.

Bundesbank Discussion Paper No 55/2020

A random forest-based approach to identifying the most informative seasonality tests*

Daniel Ollech
Deutsche Bundesbank

Karsten Webel
Deutsche Bundesbank

Abstract

Virtually each seasonal adjustment software includes an ensemble of seasonality tests for assessing whether a given time series is in fact a candidate for seasonal adjustment. However, such tests are certain to produce either the same result or conflicting results, raising the question if there is a method that is capable of identifying the most informative tests in order (1) to eliminate the seemingly non-informative ones in the former case and (2) to find a final decision in the more severe latter case. We argue that identifying the seasonal status of a given time series is essentially a classification problem and, thus, can be solved with machine learning methods. Using simulated seasonal and non-seasonal ARIMA processes that are representative of the Bundesbank's time series database, we compare certain popular methods with respect to accuracy, interpretability and availability of unbiased variable importance measures and find random forests of conditional inference trees to be the method which best balances these key requirements. Applying this method to the seasonality tests implemented in the seasonal adjustment software JDemetra+ finally reveals that the modified QS and Friedman tests yield by far the most informative results.

Keywords: binary classification; conditional inference trees; correlated predictors; JDemetra+; simulation study; supervised machine learning.

JEL classification: C12, C14, C22, C45, C63.

*Contact address: Daniel Ollech, Deutsche Bundesbank, Central Office, Statistics Department, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany. Phone: +49 69 9566 6250. E-mail: daniel.ollech@bundesbank.de. The authors thank Edgar Brandt, Nina Gonschorreck, Christoph Hanck, Christiane Hofer, Andreas Lorenz, Steve Matthews and seminar participants at the Bundesbank for helpful comments. Discussion Papers represent the authors' personal opinions and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

1 Motivation

To judge current economic developments and to forecast important target variables, such as quarterly gross domestic product, economists usually monitor a broad and well-established set of trustworthy key time series on a regular basis. Usually, some time series will enter this set in unadjusted form while some other series will enter it in seasonally adjusted form, depending on the seasonal status of the particular time series. Examples include the data sets used by the Bundesbank for short-term forecasting of economic activity in Germany and other countries (Götz and Hauzenberger, 2018; Pinkwart, 2018). Thus, seasonal adjustment is an integral part of many economic analyses and business processes, including in particular case-by-case decisions on whether or not observed time series are in fact seasonal and should or should not be seasonally adjusted.

A variety of statistical tests has been developed over time to answer the latter question (e.g. Busetti and Harvey, 2003; Franses, 1992; Ghysels and Osborn, 2001) and virtually each seasonal adjustment software that is currently available contains at least some of them. For example, in release version 2.2.2 of JDemetra+ (JD+), the output's diagnostics section reports the results of six different seasonality tests. However, any set of such tests is certain to give either concurring or conflicting outcomes in the sense that the tests agree or disagree, raising different questions in either case. The following example illustrates this point.

Example 1 (potential conflicts between seasonality tests)**.** Figure 1 shows sub-samples of four monthly macroeconomic time series for Germany, which, amongst other things, differ substantially in terms of their seasonal behaviour. Retail trade turnover for games and toys displays a strong seasonal pattern which primarily originates from Christmas-related spikes in December figures. In contrast, the harmonised index of consumer prices (HICP) for tobacco clearly lacks such strong (and probably any other) seasonal movements as the series is dominated by a slowly rising trend that is interrupted quite frequently by aperiodic level shifts associated with VAT increases. Finally, neither the consumer price index (CPI) for energy nor the number of persons employed in the manufacturing of wearing apparel reveal their seasonal status as immediately as the turnover and HICP series. On the one hand, either series appears to be driven mainly by trend-cyclical and irregular movements, leaving any seasonal behaviour far less distinctive compared to the turnover series. On the other hand, some recurring intra-year movements, such as the minor V-shaped troughs in the employment series in the middle of the years 2012 to 2014 and 2016, can still be eye-balled, rendering any seasonal behaviour less ignorable compared to the HICP series.

Table 1 reports the outcomes of the six JD+ seasonality tests, which will be described in detail in Section 2. As opposed to Figure 1, the tests are calculated over the entire data spans which all end in March 2020 but start in January of different years: 1991 (CPI), 1994 (turnover), 1996 (HICP), and 2009 (employment). The tests confirm the first impression gained from Figure 1 that the turnover series is seasonal and the HICP series is not. However, their outcomes conflict for the other two series, in line with visual inspection. For the CPI series, the Friedman, Kruskal-Wallis and periodogram tests as well as the F-test on seasonal dummies reject the null hypothesis of absence of seasonality at the 1% level of significance, while the modified QS test barely fails to do so, which is also supported by the test for seasonal peaks. For the employment series, the Friedman

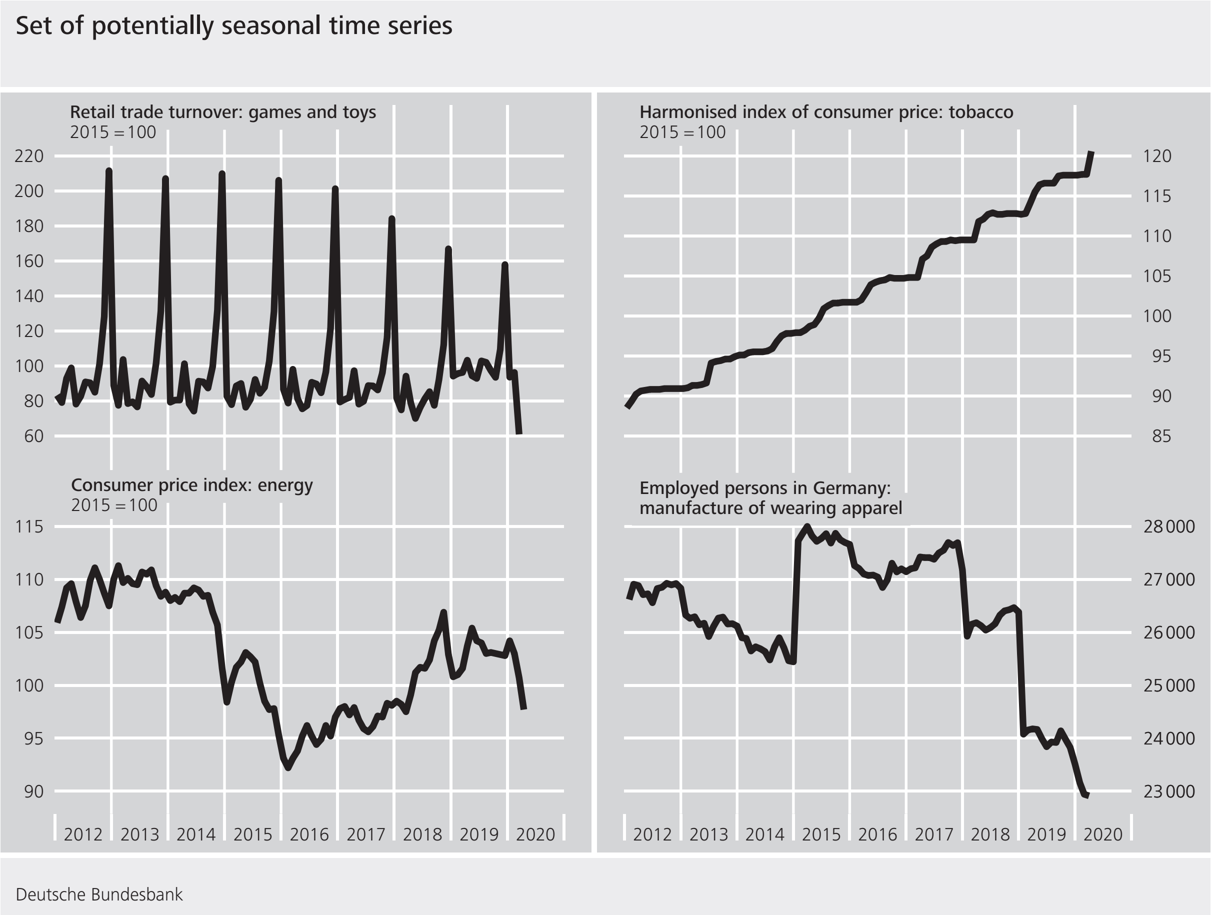

Figure 1: Four macroeconomic time series for Germany.

and Kruskal-Wallis tests provide strong evidence in favour of presence of seasonality at the 1% level of significance. The modified QS test provides mild evidence at the 5% level of significance, whereas the remaining tests do not reject the null hypothesis of absence of seasonality at this level. □

Example 1 illustrates two issues which arise in general for any set of seasonality tests. First, for some series the tests reach the same decision and we may ask whether the set is too large and if it can be reduced by eliminating seemingly non-informative tests. Second, and more importantly, for some other series the tests reach different decisions and we may ask how their outcomes can be aggregated. Of course, we could simply stick to a majority vote. In our example, this would identify the CPI series as seasonal at any conventional level of significance and classify the employment series as non-seasonal at the 1% level of significance but end in a tie at the 5% level of significance. Also, some tests may be more informative than others and, as a consequence, some tests overruled by the majority vote may still carry relevant information that should not be ignored. Thus, a weighted vote of those tests which can be identified reliably as most informative will probably provide a better aggregator of the test outcomes than a simple majority vote.

Essentially, the question whether a given time series is seasonal or non-seasonal can be viewed as a classification problem with only two classes. Since the two general issues, i.e. elimination of seemingly non-informative tests and identification of most informative

Table 1: Test statistics (TS) and p-values of the JD+ seasonality tests for the time series shown in Figure 1.

	Retail trade turnover: games and toys		HICP: tobacco		CPI: energy		Employed persons: manufacture of wearing apparel	
	TS	p-value	TS	p-value	TS	p-value	TS	p-value
QS	537.787	0.000	0.750	0.687	7.463	0.024	6.551	0.038
FT	236.337	0.000	13.031	0.291	29.804	0.002	46.864	0.000
KW	258.577	0.000	5.363	0.912	33.755	0.000	44.127	0.000
SP	AT AT AT		⋆⋆ a⋆ ⋆⋆		⋆t ⋆⋆ ⋆⋆		⋆⋆ A⋆ ⋆⋆	
	AT AT AT		⋆⋆ ⋆⋆ a⋆		⋆⋆ a⋆ ⋆⋆		⋆⋆ ⋆⋆ ⋆⋆	
PD	323.551	0.000	0.376	0.965	3.140	0.001	1.796	0.062
SD	364.898	0.000	0.359	0.970	3.031	0.001	1.747	0.071

Remarks: All series have been differenced once, the CPI series has additionally been logged, as suggested by the automatic log-level-test (Gómez and Maravall, 2001). The tests are the modified QS test (QS), Friedman test (FT), Kruskal-Wallis test (KW), test for seasonal peaks (SP), periodogram test (PD) and F-test on seasonal dummies (SD). For the SP test, lower (upper) case letters refer to visually significant peaks at the 10% (1%) level of significance in the Tukey (T) and AR(30) (A) spectra at the seasonal frequencies corresponding to the letters' positions, where visual significance is defined in Equation 7 and Equation 8, respectively. For any other test, p-values smaller than 0.0005 are indicated by 0.000.

tests, are related, they can be tackled in theory using the same classifier. To this end, low misclassification rates, high interpretability and the availability of unbiased variable importance measures are key requirements that such a classifier has to satisfy. We show by means of a large-scale simulation of representative seasonal and non-seasonal ARIMA time series that (1) several machine learning algorithms provide acceptable seasonality classifiers in terms of low error rates but (2) only random forests of conditional inference trees additionally provide unbiased variable importance measures, solving both the elimination and identification issue. In this regard, it should be noted that they are also capable of solving the aggregation issue – and we briefly demonstrate this later – but in general this topic is rather the scope of companion research (Webel and Ollech, 2018).

The remainder of this paper is organised as follows. Section 2 provides basic theory of the six seasonality tests implemented in JD+. Section 3 describes the data generation, including in particular the identification and simulation of representative ARIMA time series. Section 4 compares selected machine learning algorithms with respect to their capability of meeting the key requirements stressed above. Section 5 elaborates on the winner of this competition: random forests. Highlighting differences between forests based on unconditional and conditional inference trees, it provides basic theory of the algorithms and their respective variable importance measures. Section 6 uses both the simulated ARIMA models and the four real-world time series from Example 1 for illustration. The simulated data is also used in order to identify the most informative seasonality tests among the JD+ candidates. Finally, Section 7 concludes.

2 Seasonality tests in JDemetra+

JD+ incorporates six seasonality tests, each of which tests the null hypothesis (H_0) of absence of seasonality. We provide basic information about these tests in the order of their appearance in the JD+ output. To this end, let $\{z_t\}$ denote a weakly stationary series of length T with τ observations per year. Also, let $(pdq)(PDQ)$ abbreviate the ARIMA model

$$\phi_p(B)\,\Phi_P(B^\tau)\,\nabla_1^d\,\nabla_\tau^D x_t = \theta_q(B)\,\Theta_Q(B^\tau)\,\varepsilon_t, \tag{1}$$

where B is the backshift operator, $B^k x_t = x_{t-k}$, $\nabla_k = 1 - B^k$, d and D indicate the non-seasonal and seasonal orders of differencing, ϕ_p and Φ_P are the non-seasonal and seasonal autoregressive (AR) operators of orders p and P, θ_q and Θ_Q are the non-seasonal and seasonal moving average (MA) operators of orders q and Q, and $\{\varepsilon_t\}$ is white noise with zero mean and finite variance.

2.1 Modified QS test

The modified QS test (QS) checks the series $\{z_t\}$ for significant positive autocorrelation at seasonal lags.[1] 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,

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

where $\hat{\rho}(h)$ is the estimated lag-h autocorrelation of $\{z_t\}$. The exact null distribution of the QS-statistic (2) is unknown but can be approximated reasonably well by a χ^2-distribution with two degrees of freedom (Maravall, 2011).

2.2 Friedman test

The Friedman test (FT) checks for significant differences between the period-specific mean ranks of the values of $\{z_t\}$, being essentially a one-way ANOVA with repeated measures (Friedman, 1937). To see this, assume that each period $i \in \{1, \ldots, \tau\}$ has n observations, i.e. there are n complete years of observations.[2] 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

$$FT = \frac{\tau - 1}{\tau} \sum_{i=1}^{\tau} \frac{n\,[\bar{r}_i - (\tau+1)/2]^2}{(\tau^2 - 1)/12}, \tag{3}$$

[1] Negative autocorrelations at seasonal lags reflect alternating patterns over time and, thus, do not represent seasonal behaviour.

[2] If necessary, excess months are removed from the beginning of the series to ensure $T = n \cdot \tau$.

where $\bar{r}_i = n^{-1} \sum_{j=1}^{n} r_{ij}$. Under H_0, the FT-statistic (3) follows asymptotically a χ^2-distribution with $\tau - 1$ degrees of freedom.

2.3 Kruskal-Wallis test

The Kruskal-Wallis test (KW) follows the idea of the Friedman test up to two modifications, being essentially a one-way ANOVA without repeated measures (Kruskal and Wallis, 1952). First, period-specific numbers n_i of observations are allowed, giving $T = \sum_{i=1}^{\tau} n_i$; second, ranks are assigned over the entire observation span (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

$$KW = \frac{T-1}{T} \sum_{i=1}^{\tau} \frac{n_i \left[\bar{r}_i - (T+1)/2\right]^2}{(T^2-1)/12}. \tag{4}$$

Under H_0, the KW-statistic (4) asymptotically follows a χ^2-distribution with $\tau - 1$ degrees of freedom.

2.4 Periodogram test

The periodogram test (PD) checks if a weighted sum of the spectral density of $\{z_t\}$ evaluated at the seasonal frequencies is significantly different from zero. Let $f(\omega) = (2\pi)^{-1} \sum_h \gamma(h)\, e^{-ih\omega}$ denote the spectral density, $\omega_j^\star = 2\pi j/\tau$ the j-th seasonal frequency for $j \in \{1, \ldots, \tau/2\}$, $\mathcal{S}(\tau) = \{\omega_1^\star, \ldots, \omega_{\tau/2}^\star\}$ the set of seasonal frequencies for τ, and

$$\Sigma_{\mathcal{S}(\tau)} = 2 \sum_{j=1}^{\tau/2-1} f(\omega_j^\star) + f(\omega_{\tau/2}^\star)$$

the weighted sum of $f(\omega)$ evaluated over $\mathcal{S}(\tau)$. The null hypothesis then reads $H_0 : \Sigma_{\mathcal{S}(\tau)} = 0$, and the test statistic is given by

$$PD = \frac{T-\tau}{\tau-1} \cdot \frac{\hat{\Sigma}_{\mathcal{S}(\tau)}}{\sum_{t=1}^{T} z_t^2 - I(0) - \hat{\Sigma}_{\mathcal{S}(\tau)}}, \tag{5}$$

where

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

$\mathbb{1}_{\{\cdot\}}$ is the indicator function of the event in braces and

$$I(\omega_j) = \begin{cases} \sum_{|h| \leq T} \hat{\gamma}(h)\, e^{-ih\omega_j}, & \omega_j \neq 0 \\ T\, |\bar{z}|^2, & \omega_j = 0 \end{cases}, \tag{6}$$

is the periodogram with $\omega_j = 2\pi j/T$ being the j-th Fourier frequency for $j \in \{-\lfloor (T-1)/2 \rfloor, \ldots, \lfloor T/2 \rfloor\}$, $\lfloor x \rfloor$ the largest integer not exceeding x, and $\bar{z} = T^{-1} \sum_{t=1}^{T} z_t$. Under

H_0, the PD-statistic (5) follows an F-distribution with $\tau - 1 - \mathbb{1}_{\{T \text{ even}\}}$ and $T - \tau + \mathbb{1}_{\{T \text{ even}\}}$ degrees of freedom.[3]

2.5 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.

2.5.1 Tukey spectrum

The Tukey spectrum is a non-parametric "lag window" estimator. To transform (6) 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 $\gamma(h)$. 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 T. A peak at any Fourier frequency ω_j is called visually significant at the α-level of significance if

$$\frac{2\hat{f}_T(\omega_j)}{\hat{f}_T(\omega_{j-1}) + \hat{f}_T(\omega_{j+1})} \geq F_{d_1, d_2, 1-\alpha}, \tag{7}$$

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).

2.5.2 AR(30) spectrum

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.

[3] To improve performance in small samples, $\mathcal{S}(\tau)$ should be a subset of the set of Fourier frequencies. To this end, T is made a multiple of τ by removing observations at the beginning of the series.

In general, the spectrum of an 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 estimated 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 (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

$$\frac{\hat{f}_{AR}(\omega_j) - \max\left\{\hat{f}_{AR}(\omega_{j-1}), \hat{f}_{AR}(\omega_{j+1})\right\}}{\max_j \hat{f}_{AR}(\omega_j) - \min_j \hat{f}_{AR}(\omega_j)} \tag{8}$$

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 universally use the critical value associated with $\omega = 0.696\,\pi$, i.e. the first trading day frequency for $\tau = 12$.

2.5.3 Decision rule

For $\tau = 12$, $\{z_t\}$ is now said to have seasonal peaks, giving $SP = 1$, if visually significant peaks show up in[4]

(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.

[4] For quarterly series, a similar but smaller set of rules applies (Maravall, 2011).

2.6 Seasonal dummies

The F-test on seasonal dummies (SD) checks if the effects of the $\tau - 1$ seasonal dummies are simultaneously zero in a time series regression on $\{z_t\}$. 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

$$D_{i,t} = \begin{cases} 1, & t \in \mathcal{P}_i \\ -1, & t \in \mathcal{P}_\tau \\ 0, & \text{otherwise} \end{cases}, \quad i \in \{1, \dots, \tau - 1\},$$

and $\mathcal{P}_k$ is time index set of the k-th period for $k \in \{1, \dots, \tau\}$. Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{\tau-1})^\top$. The null hypothesis is then specified as $H_0 : \boldsymbol{\beta} = \mathbf{0}$, and the test statistic is given by

$$SD = \frac{\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}^{-1} \hat{\boldsymbol{\beta}}}{\tau - 1} \cdot \frac{T - d - p - q - \tau - 1}{T - d - p - q}, \tag{9}$$

where $\hat{\boldsymbol{\beta}}$ is the OLS estimator of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$. Under H_0, the SD-statistics (9) follows an F-distribution with $\tau - 1$ and $T - d - p - q - \tau - 1$ degrees of freedom. Two variants of non-seasonal orders are considered: first, $(pdq) = (011)$ is used; second, (pdq) is determined via automatic model identification (Gómez and Maravall, 2001).

3 Data generation

We aim at simulating "realistic" seasonal and non-seasonal time series that portray as accurately as possible the macroeconomic monthly data analysed regularly by the Bundesbank. We therefore use the Bundesbank's time series database and ARIMA models as main building blocks of our data generating mechanism. The latter are chosen for two reasons: first, ARIMA models cover a broad range of correlation structures and perform very well in forecasting exercises; second, ARIMA models are an essential element of the two seasonal adjustment approaches available in JD+.[5]

Section 3.1 first identifies the relevant seasonal and non-seasonal ARIMA model types. Using notations from Equation 1, an ARIMA model is said to be seasonal (S) if $(PDQ) \neq (000)$ and non-seasonal (N-S) if $(PDQ) = (000)$. Section 3.2 then introduces the core simulation algorithm which is capable of replicating any given dependence structure of ARMA parameters from estimated ARIMA models. Finally, Section 3.3 applies this algorithm to

[5] JD+ implements the X-11 and ARIMA model-based (AMB) approaches to seasonal adjustment alongside their respective regARIMA and TRAMO pre-processors, which are essentially linear time series regression models with ARIMA errors.

the ARIMA model types which have been identified as relevant, using appropriate simulation weights. The simulated ARIMA processes are considered as being "realistic" if, after fitting ARIMA models to the time series in the database, they resemble as closely as possible the estimated ARIMA models in terms of model shares and the ARMA parameters' multivariate distribution for each model type.

The data generating mechanism is based on the Bundesbank's time series database as accessed in January 2017 and implemented using `R` (R Core Team, 2019) and the `{gsarima}`, `{logspline}` and `{MASS}` `R` packages (Briët, Amerasinghe, and Vounatsou, 2013; Kooperberg and Stone, 1992; Venables and Ripley, 2002). It is still generic since the feeding database can be changed easily and both ARIMA models and the core simulation algorithm are generally applicable.

3.1 Model identification

The identification of seasonal ARIMA models is based on a set of 3,308 macroeconomic time series which have been seasonally adjusted on a regular basis by the Bundesbank for many years.[6] All monthly series from this set are selected, amounting to 2,907 series in total with lengths ranging from 6 to 67 years. This data set contains information on a variety of key economic indicators, such as industrial production, orders received by industry, turnover in industry and retail trade, labour market statistics and consumer price indices, amongst others, and is referred to as set SA ("seasonally adjusted time series").

The identification of non-seasonal ARIMA models is based on a set of 10,635 macroeconomic time series which are not seasonally adjusted at all. This set is obtained after cleaning the candidate data as described in Remark 1 below. These series have lengths between 4 and 72 years, include information on national accounts, external sector and money and capital markets, amongst others, and are referred to as set NA ("not seasonally adjusted time series").

Remark 1. The Bundesbank's database stores about 12,000,000 time series which in theory qualify as candidates for non-seasonal data. In practice, however, this set contains a fair amount of time series which are inappropriate for our purpose for a variety of reasons. Due to computational restrictions, we first draw a simple random sample of size 500,000 without replacement from the set of candidate series. We then remove series which are not monthly OR have zero variance OR contain observations for periods past the year 2018 (as in January 2017 such series are usually regression variables used for calendar adjustment or artificial data) OR contain more than two zero observations OR have less than 36 observations excluding missing values and zeros (as such series are usually not considered as candidates for seasonal adjustment). □

[6] The Bundesbank has been concerned with seasonal adjustment for more than 60 years, using moving average-based methods such as X-11 and X-12-ARIMA for most of the time (Bank Deutscher Länder, 1957; Deutsche Bundesbank, 1970, 1999). The decision which series should be seasonally adjusted has been clear in most cases. Unclear cases have usually been solved by inspection of seasonality diagnostics and adjustment if at least one of them found (circumstantial) evidence of seasonality. In addition, expert knowledge, e.g. on data collection, and experience of staff economists and statisticians has also been considered in some cases.

Table 2: Shares and simulation weights of the ARIMA models identified automatically by the JD+ pre-processors TRAMO and regARIMA (REG).

$k = \text{SA}$				$k = \text{NA}$			
$\mathcal{M}(k)$	$p_{mk}^{(\text{TRAMO})}$	$p_{mk}^{(\text{REG})}$	w_{mk}	$\mathcal{M}(k)$	$p_{mk}^{(\text{TRAMO})}$	$p_{mk}^{(\text{REG})}$	w_{mk}
(011)(011)	39.60	25.64	47.5	(011)(000)	23.35	19.27	22.8
(010)(011)	6.09	5.34	8.3	(311)(000)	9.46	11.01	11.0
(311)(011)	4.45	6.21	7.7	(110)(000)	9.31	9.96	10.3
(210)(011)	2.01	7.11	6.6	(100)(000)	9.73	6.22	8.5
(110)(011)	4.26	4.48	6.4	(211)(000)	4.64	5.62	5.4
(211)(011)	2.01	4.03	4.4	(001)(000)	4.91	3.95	4.7
(012)(011)	2.41	3.13	4.0	(010)(000)	7.33	0.49	4.2
(111)(011)	2.92	2.06	3.6	(111)(000)	2.45	5.14	4.0
(011)(111)	2.62	0.87	2.5	(012)(000)	3.94	2.93	3.6
(010)(100)	1.19	1.48	2.0	(210)(000)	3.51	3.19	3.6
(010)(101)	1.68	0.87	1.9	(000)(000)	2.60	2.56	2.8
(310)(011)	0.23	1.81	1.5	(301)(000)	1.97	2.56	2.5
(112)(011)	1.63	0.27	1.4	(021)(000)	1.46	2.76	2.3
(021)(011)	1.17	0.73	1.4	(212)(000)	0.34	3.76	2.2
(110)(101)	1.02	0.14	0.8	(101)(000)	0.90	2.97	2.1
(110)(000)	4.88	6.25	0.0	(200)(000)	2.41	1.34	2.0
(011)(000)	2.38	4.18	0.0	(310)(000)	2.60	0.67	1.8
(010)(000)	0.00	2.94	0.0	(112)(000)	0.22	2.71	1.5
(111)(000)	1.75	1.09	0.0	(300)(000)	1.42	1.37	1.5
(021)(000)	1.19	0.72	0.0	(201)(000)	0.56	1.37	1.1
(210)(000)	0.69	1.12	0.0	(002)(000)	1.19	0.49	0.9
(310)(000)	0.11	1.15	0.0	(102)(000)	0.16	1.26	0.8
				(202)(000)	0.12	1.01	0.6

Remark: Shares p_{mk} and weights w_{mk} according to Equation 10 are reported as a percentage for models which have been identified by at least one pre-processor for at least one percent of all time series in set k, i.e. for models $m \in \mathcal{M}(k)$ with $p_{mk}^{(\text{TRAMO})} \geq 0.01$ or $p_{mk}^{(\text{REG})} \geq 0.01$ (or both).

For each set of series, we run the automatic ARIMA model identification routines of the TRAMO and regARIMA pre-processors as implemented in JD+, including automatic detection and modelling of outliers. Let $\mathcal{M}(k)$ denote the join of identified models in set $k \in \{\text{NA}, \text{SA}\}$ and $p_{mk}^{(\text{TRAMO})}$ and $p_{mk}^{(\text{REG})}$ the set-k shares of model $m \in \mathcal{M}(k)$ identified by the TRAMO and regARIMA pre-processors, respectively. Table 2 shows that mostly seasonal but also some non-seasonal ARIMA models have been identified for the time series in set SA, whereas only non-seasonal ARIMA models have been identified for the time series in set NA.

For each identified ARIMA model $m \in \mathcal{M}(k)$ listed in Table 2, we calculate the model's overall set-k share as $p_{mk} = \left(p_{mk}^{(\text{TRAMO})} + p_{mk}^{(\text{REG})}\right) \Big/ 2$ and the model's simulation weight[7] as

$$w_{mk} = \frac{\tilde{p}_{mk}}{\sum_j \tilde{p}_{jk}}, \tag{10}$$

rounded to 1 decimal place, with $\tilde{p}_{mk} = p_{mk} \cdot \mathbb{1}_{\left\{p_{mk}^{(\mathrm{TRAMO})} \geq 0.01 \,\mathrm{OR}\, p_{mk}^{(\mathrm{REG})} \geq 0.01\right\}}$. Thus, we intentionally exclude those models from subsequent simulations which have been identified rather rarely, i.e. for less than one percent of all time series, narrowing our intended focus on relevant models.

3.2 Simulation algorithm

Given a set of observed time series from the Bundesbank's database that follow the same ARIMA model, we aim at simulating versions of this model under the restriction 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 usually difficult to determine the exact family of distributions of the latter parameters, we impose the following three proxy restrictions: the ARMA parameters of the simulated time series

(P1) do not induce (additional) unit roots in the model's characteristic polynomial,

(P2) display the same correlation structure as the estimated ARMA parameters of the ARIMA models fitted to the observed time series,

(P3) should have univariate distributions with the same shape as the univariate distributions observed for the corresponding estimated ARMA parameters.

To meet these proxy restrictions, we combine the "NORmal-To-Anything" (NORTA) algorithm (Cario and Nelson, 1997) with logspline density estimation according to a particular knot addition and deletion algorithm (Stone, Hansen, Kooperberg, and Truong, 1997).

Algorithm 1 describes the technical details. Its basic idea is to initially draw sets of ARMA parameters from a multivariate Gaussian distribution such that the dependence structures of the simulated and estimated parameters are similar. The multivariate Gaussian density is then transformed into a set of univariate logspline densities. After elimination of ARMA parameters that induce additional unit roots, a random sample of admissible parameters is drawn and their dependence structure is compared to that of the ARMA parameters of the model fitted to the observed time series. In case of unacceptably high differences, the entire algorithm is restarted with an appropriately modified covariance matrix of the multivariate Gaussian distribution. Eventually, it yields simulated ARIMA time series with parameters that mirror the dependence structure of the estimated parameters as closely as possible.

3.3 Training and validation data

For each model $m \in \mathcal{M}(k)$ with $k \in \{\mathrm{NA}, \mathrm{SA}\}$ and each length $N \in \{60, 120, 240\}$, we run Algorithm 1 with $\tilde{\nu} = 100{,}000 \cdot w_{mk}$, $\nu = 100{,}000$, $\varepsilon = 0.02$ and $\alpha = 0.5$, yielding a

[7] The non-seasonal ARIMA models which have been identified for some time series that are seasonally adjusted regularly are automatically given a simulation weight of zero.

Algorithm 1 NORTA algorithm with logspline density estimation

Let n be the number of observed time series which all follow the same ARIMA model of order $(pdq)(PDQ)$. Also, let $\tilde{\nu}$ be the number of ARIMA time series to be simulated under the proxy restrictions (P1) to (P3).

```
1: Set $m = p + q + P + Q$ and let $\mathbf{X} \in \mathbb{R}^{m \times n}$ be the matrix of the estimated ARMA parameters. Calculate $\mathbf{\Sigma}_{\mathbf{X}} \in \mathbb{R}^{m \times m}$, the correlation matrix of the parameters.
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 parameter, where $j \in \{1, \ldots, m\}$.
3: Set $\mathbf{\Sigma}_{\mathbf{Y}}^{(1)} = \mathbf{\Sigma}_{\mathbf{X}}$ to initialise the simulation of ARMA parameters, where $\mathbf{Y} \in \mathbb{R}^{m \times \nu}$ denotes an empty matrix to be filled during the following loop.
4: repeat
5:   In the $i$-th loop, simulate $\nu \gg \tilde{\nu}$ independent parameter 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 $j \in \{1, \ldots, n\}$. Set $\mathbf{Y} = (\mathbf{Y}_1 \ldots \mathbf{Y}_\nu)$.
6:   Set $\mathbf{Z} = (z_{jk}) \in \mathbb{R}^{m \times \nu}$, where $z_{jk} = \hat{F}_j^{-1}[\Phi(y_{jk})]$ for all $(j,k) \in \{1, \ldots, m\} \times \{1, \ldots, \nu\}$ and $\hat{F}_j(\cdot)$ and $\Phi(\cdot)$ are the distribution functions of $\hat{f}_j(\cdot)$ and the standard normal distribution, respectively.
7:   Let $l \in \{0, \ldots, \nu\}$ be the number of columns of $\mathbf{Z}$ which contain ARMA parameters that induce additional 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 parameters.
8:   Select $\tilde{\nu}$ columns from $\tilde{\mathbf{Z}}$ according to simple random sampling without replacement, where $\tilde{\nu} \in \{1, \ldots, \nu - l\}$. Store the sampled columns in $\tilde{\mathbf{Z}}^{(\tilde{\nu})} \in \mathbb{R}^{m \times \tilde{\nu}}$.
9:   Calculate $\mathbf{\Sigma}_{\tilde{\mathbf{Z}}^{(\tilde{\nu})}} \in \mathbb{R}^{m \times m}$, the correlation matrix of the sampled admissible ARMA parameters.
10:  Calculate $\mathbf{\Delta} = |\mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}_{\tilde{\mathbf{Z}}^{(\tilde{\nu})}}| = (\delta_{jk})$ and $\mathbf{C}_{\mathbf{\Delta}} = (c_{jk})$, where $c_{jk} = \mathbb{1}_{\{\delta_{jk} > \varepsilon\}}$ for all $(j,k) \in \{1, \ldots, m\}^2$ and some $\varepsilon > 0$.
11:  if $\mathbf{C}_{\mathbf{\Delta}} \neq \mathbf{0}$ then
12:    Set $\mathbf{\Sigma}_{\mathbf{Y}}^{(i+1)} = \mathbf{\Sigma}_{\mathbf{Y}}^{(i)} + \alpha\,[\mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}_{\tilde{\mathbf{Z}}^{(\tilde{\nu})}}] \odot \mathbf{C}_{\mathbf{\Delta}}$, where $\alpha > 0$ and $\odot$ denotes the Hadamard product of two matrices, i.e. $\mathbf{A} \odot \mathbf{B} = (a_{jk} \cdot b_{jk})$.
13:  end if
14: until $\mathbf{C}_{\mathbf{\Delta}} = \mathbf{0}$, or the maximum number of iterations is reached.
15: Simulate $\tilde{\nu}$ ARIMA models of order $(pdq)(PDQ)$ with the parameters stored in the columns of $\tilde{\mathbf{Z}}^{(\tilde{\nu})}$.
```

total of 600,000 simulated ARIMA time series with Gaussian innovations. In Step 2, we use default settings for the logspline density estimation but have to restrict the density support in some cases in order to ensure simulation of appropriate ARIMA models.[8] This is mainly done because the order of the finite AR representation, used as an approximation to the infinite AR representation, of MA polynomials increases exponentially with the absolute size of the MA parameters. This potentially results in computational issues due to exceeding memory capacity. In Step 15, we let the length of the burn-in period for the simulation depend on the order of the finite AR representation of the ARMA polynomial,

[8] The density support for parameters in seasonal or non-seasonal polynomials of order 1 is $[-1; 0.975]$ for AR parameters and $[-0.99; 0.975]$ for MA parameters, whereas it is $[-2; 2]$ for each parameter in non-seasonal AR or MA polynomials of order 2 or 3.

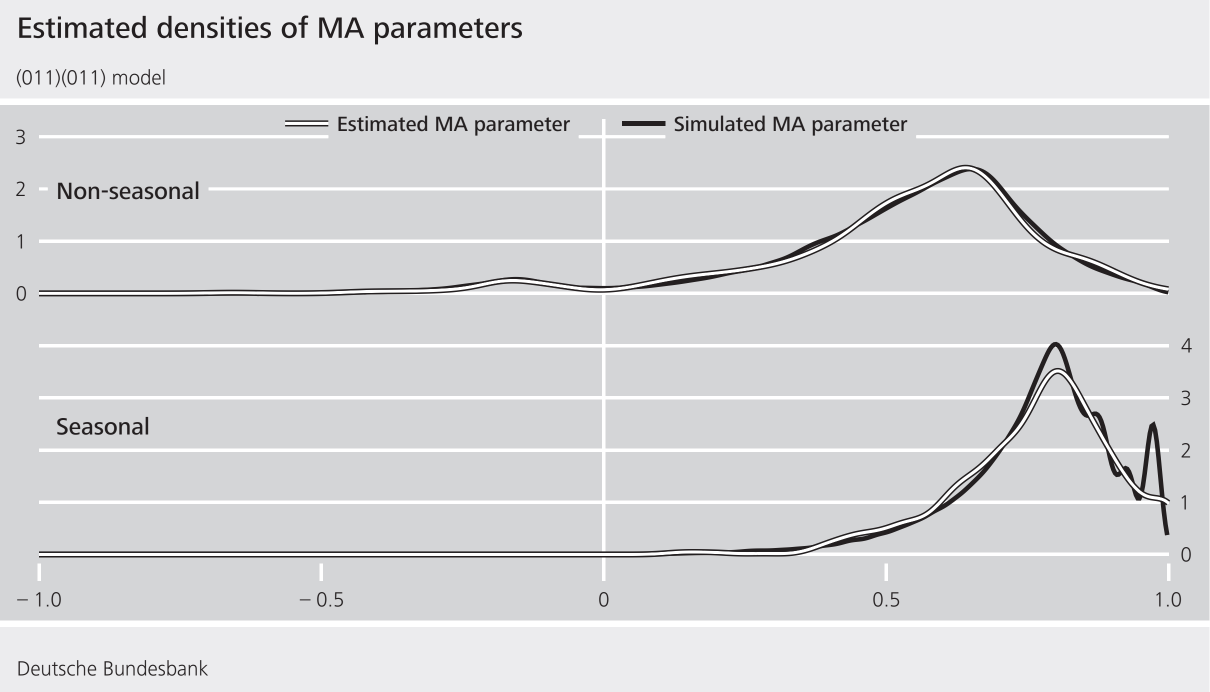

Figure 2: Kernel density estimates for the estimated MA parameters of model (11) and the corresponding simulated MA parameters obtained from Algorithm 1, using Gaussian kernels.

as in Briët et al. (2013).

Example 2 (application of Algorithm 1 for the (011)(011) model)**.** Figure 2 shows the estimated univariate densities of the estimated and simulated non-seasonal and seasonal MA parameters of the model

$$(1-B)\left(1-B^{12}\right)x_t = (1-\theta B)\left(1-\Theta B^{12}\right)\varepsilon_t, \tag{11}$$

where $(\theta, \Theta) \in (-1;1) \times (-1;1)$. The estimated densities of θ are almost indistinguishable except for slightly different shapes over the range $\theta \in [0.2; 0.6]$. The estimated densities of Θ are also shaped similarly in general. However, the density of the simulated parameter appears to be somewhat compressed in the range $\Theta \in [0.5; 1)$ and, what is more, exhibits some unwanted ripples for $\Theta > 0.8$, especially in the vicinity of 1. This is directly induced by the technical necessity of narrowing slightly the support of Θ in the simulation (see Footnote 8). Overall, Algorithm 1 still replicates very well the distributional properties of the estimated MA parameters. □

For each of the 600,000 simulated ARIMA time 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. By unchangeable default in JD+, the simulated series are differenced once before the test calculation in order to ensure stationarity. Since the test on seasonal peaks cannot be calculated for monthly time series with less than seven years of observations, we use the p-values of the other five seasonality tests and a dummy for the seasonality class as the set of predictors $\mathbf{X}$ and the categorical response $\mathbf{Y}$, respectively, of the "mother" training data $\mathcal{L} = (\mathbf{XY})$. We also randomly sample without replacement

- 50 independent "daughter" training data sets $\mathcal{L}^{(i)}$ of size 7,500 from $\mathcal{L}$ and
- 50 independent "sampled" validation data sets $\mathcal{V}_s^{(i)}$ of size 10,000 from the non-sampled data $\mathcal{V}^{(i)} = \mathcal{L} \setminus \mathcal{L}^{(i)}$, i.e. one $\mathcal{V}_s^{(i)}$ is sampled from each $\mathcal{V}^{(i)}$.

The sample sizes are chosen in order to meet computational restrictions imposed by the complexity of some classification algorithms to be applied in Section 4 and Section 6.

4 Classification algorithms

Several machine learning (ML) and other – more traditional – classification methods could be used in general to group the simulated ARIMA time series into seasonal and non-seasonal models. Each of them has advantages and disadvantages, bearing in mind different salient features of different types of predictors. To identify the one that best suits our set of predictors $\mathbf{X}$, we now compare popular methods with respect to the following key requirements:

(1) **High accuracy**: The method should have low misclassification rates for both the simulated non-seasonal and seasonal ARIMA time series.

(2) **High interpretability**: The method should be intrinsically transparent and provide useful output for practitioners, such as visualisation of results and quantification of each predictor's informational content.

Remark 2. Given that the high flexibility of most ML methods sometimes comes at the cost of some opaqueness, interpretability refers to the extent to which the choices and decisions made by the ML method can be understood by users. In this sense, "intrinsic transparency" and "useful output" can be seen as quasi ex ante and ex post interpretability, respectively. □

(3) **Unbiasedness**: The method should provide measures of the predictors' informational content that are unbiased given the predictors' statistical properties.

The set of candidate ML methods consists of the following nine classifiers: classical and conditional random forests (Breiman, 2001; Hothorn, Hornik, and Zeileis, 2006), adaptive and stochastic boosting (Freund and Schapire, 1997; Friedman, 2002), support vector machines (Boser, Guyon, and Vapnik, 1992; Vapnik, 1995), feed forward neural networks (Ripley, 1996), logistic regression, weighted K-nearest neighbours (Samworth, 2012) and naive Bayes. Further details on these classifiers are also given by Breiman, Friedman, Olshen, and Stone (1984), Hastie, Tibshirani, and Friedman (2009) and James, Witten, Hastie, and Tibshirani (2013).

Each candidate ML method is trained on each "daughter" training data set $\mathcal{L}^{(i)}$ using essentially its default implementation in the respective R package (Alfaro, Gamez, and Garcia, 2018; Breiman, Cutler, Liaw, and Wiener, 2001; Culp, Johnson, and Michailidis, 2016; Hechenbichler and Lizee, 2016; Hothorn, Hornik, Strobl, and Zeileis, 2015; Meyer, Dimitriadou, Hornik, Weingessel, Leisch, Chang, and Lin, 2019; Ripley and Venables, 2016) and evaluated on each sampled validation data set $\mathcal{V}_s^{(i)}$. Table 3 summarises the key results. Distinguishing broadly between tree-based methods (top four rows) and other methods (bottom five rows), the following conclusions emerge with respect to our key requirements:

Table 3: Fulfilment of key requirements by candidate ML methods.

		High accuracy				High interpretability				Unbiasedness
		Misclassification rate in class						Measures for informational content		
		Non-seasonal		Seasonal		Graphical output				
Classifier	R package (version)	Mean*	SD**	Mean*	SD**	Availability	Type	Availability	Type	Correlated predictors
Conditional random forests	{`party`} (1.3-0)	0.77	0.17	2.13	0.24	Yes	Decision trees	Yes	Variable importance	Yes
Classical random forests	{`randomForest`} (4.6-14)	0.82	0.18	1.98	0.19	Yes	Decision trees	Yes	Variable importance	No
Stochastic boosting	{`ada`} (2.0-5)	1.03	0.22	2.03	0.21	Yes	Error plots	Yes	Variable importance	No
Adaptive boosting	{`adabag`} (4.2)	1.07	0.22	1.98	0.23	Yes	Error plots	Yes	Variable importance	No
Support vector machines	{`e1071`} (1.6-8)	2.37	0.24	2.12	0.22	Yes	Scatter plots	Yes***	Variable importance	No
Feed forward neural networks	{`nnet`} (7.3-12)	1.89	0.19	1.89	0.22	No		Yes	Connection weights	No
Logistic regression	{`stats`} (3.4.4)	6.31	0.51	1.45	0.23	Yes	Standard plots	Yes	Standardised coefficients	No
Weighted K-nearest neighbours	{`kknn`} (1.3.1)	1.17	0.18	2.13	0.19	No		No		
Naive Bayes	{`e1071`} (1.6-8)	1.98	0.49	2.67	0.21	No		No		

$\star$ As a percentage. $\star\star$ In percentage points. $\star\star\star$ No implementation in the respective R package.

Remark: Regarding accuracy, means and standard deviations (SD) of misclassification rates have been calculated over the sampled validation data sets $\mathcal{V}_s^{(i)}$, $i \in \{1, \dots, 50\}$.

(1) The tree-based methods have universally lower average misclassification rates for the non-seasonal models and competitive average misclassification rates for the seasonal models, although they perform slightly less accurately in the latter class. Support vector machines, feed forward neural networks, weighted K-nearest neighbours and to a lesser extent naive Bayes also have acceptably low average misclassification rates for both non-seasonal and seasonal models but do not universally outperform any tree-based method. Logistic regression classifies the seasonal models extremely well but fails to do so for the non-seasonal models, as it has by far the highest average misclassification rate in this class. The standard deviations of the misclassification rates are almost the same for all methods and classes, except for naive Bayes and logistic regression in the class of non-seasonal models where they are more than twice as high as in any other case. Overall, the tree-based methods seem to be slightly preferable in terms of the "high accuracy" requirement.

Remark 3. Although being derived from analysing simulated data, the last conclusion is in line with the general statement that random forests are highly competitive in a broad range of real-world prediction and classification tasks, such as forecasting stock price movements (Patel, Shah, Thakkar, and Kotecha, 2015), diagnosing diseases (Hsieh, Lu, Lee, Chiu, Hsu, and Li, 2011), and detecting fraudulent emails (Almomani, Gupta, Atawneh, Meulenberg, and Almomani, 2013). □

(2) The key principle of each tree-based method, which makes binary decisions (either simultaneously in the random forest approaches or sequentially in the boosting approaches), forms a sound basis for visualising and understanding the entire decision-making process and the final classification. Besides graphical output, it enables straightforward calculation of variable importance measures which basically evaluate each predictor's role in each binary decision and then derive its contribution to the ensemble decision. The other ML methods yield output that is noticeably less informative. Graphical output is available only for support vector machines and logistic regression as scatter/contour plots for the final classification and standard regression plots, respectively. Information on the importance of predictors is provided only for feed forward neural networks and logistic regression as "best" connection weights between input, hidden and output layers and standardised estimated regression coefficients, respectively. Overall, the tree-based methods appear to be preferable in terms of the "high interpretability" requirement.

(3) Table 4 reveals that in general the p-values of the JD+ seasonality tests exhibit high positive empirical correlations, as the latter exceed 0.81 for any pair of tests. Therefore, correlation among predictors is a serious issue which the candidate ML methods should be able to deal with, especially when quantifying variable importance. However, to the best of our knowledge, only conditional random forests are capable of providing unbiased variable importance measures in the presence of such correlations (Strobl, Boulesteix, Kneib, Augustin, and Zeileis, 2008; Strobl, Boulesteix, Zeileis, and Hothorn, 2007; Webel and Ollech, 2017). Thus, this particular tree-based method is preferable in terms of the "unbiasedness" requirement.

Given our three key requirements and weighing the candidate ML methods' respective advantages against their disadvantages, we generally favour the random forest approaches

Table 4: Spearman correlations between the p-values of the JD+ seasonality tests.

	QS	FT	KW	PD	SD	SP
QS	1.000					
FT	0.875	1.000				
KW	0.849	0.942	1.000			
PD	0.844	0.941	0.972	1.000		
SD	0.844	0.921	0.946	0.944	1.000	
SP	0.870	0.839	0.825	0.814	0.828	1.000

Remarks: The tests have been applied to the 600,000 simulated ARIMA time series (pairwise complete cases). The color code, which indicates certain clusters of tests, is explained in Remark 13. Other correlation measures, such as Bravais-Pearson coefficients, yield very similar results.

due to their competitive accuracy in classifying simulated non-seasonal and seasonal ARIMA time series and the availability of interpretable visual aids and variable importance measures. In particular, we select random forests based on conditional inference trees due to their unique capability of dealing with correlated predictors in the sense of producing unbiased variable importance measures. Nevertheless, we will also consider classical random forests in the subsequent analyses for the purpose of illustrating potential differences from the conditional approach.

Remark 4. The above comparison of the candidate ML methods intentionally focussed on those characteristics that are most relevant to the key requirements in our classification context. A related comparison in the more general context of data mining is given by Hastie et al. (2009, Chapter 10.7), who interestingly identify decision trees as preferable off-the-shelf base learners, despite the predictive inaccuracy single trees tend to suffer from. □

Remark 5. Our approach to combining several statistical tests for the same null hypothesis is based on using the tests' unadjusted p-values as predictors in a particular ML exercise. From a conceptual point of view, it is thus different from the multiple hypothesis testing procedures which are often applied in time series econometrics and usually include some sort of p-value adjustment. Prime examples are resampling-based tests for multiple structural breaks with uncertain break dates (Bergamelli, Bianchi, Khalaf, and Urga, 2019; Bernard, Idoudi, Khalaf, and Yélou, 2007; Dufour, 2006) and cointegration tests (Bayer and Hanck, 2013). □

5 Random forests

Section 5.1 briefly describes key ideas and concepts of the classical random forest (RF) approach, putting special emphasis on measuring variable importance. Section 5.2 does the same for the conditional RF approach.

Algorithm 2 Classical random forest algorithm (Breiman, 2001)

Let $\mathcal{L} = (\mathbf{XY})$ be the training data, where $\mathbf{X} = (\mathbf{X}_1 \ldots \mathbf{X}_p)$ is a set of p predictors with $\mathbf{X}_j = (x_{1j}, \ldots, x_{Nj})^\top$ for all $j \in \{1, \ldots, p\}$ and $\mathbf{Y} = (y_1, \ldots, y_N)^\top$ is a vector of categorical responses with $y_i \in \{1, \ldots, K\}$ for all $i \in \{1, \ldots, N\}$. A random forest of B classification trees is grown on $\mathcal{L}$ as follows.

```
for $b \in \{1, \ldots, B\}$ do
    Draw a bootstrap sample $\mathcal{L}_b$ of size $N$ with replacement from $\mathcal{L}$.
    Grow a binary classification tree $\mathcal{T}_b$ on $\mathcal{L}_b$. To this end, initialise the following loop
    with a single root node, i.e. $|\mathcal{M}_1(\mathcal{T}_b)| = 1$, where $\mathcal{M}_i(\mathcal{T}_b)$ is the set of terminal
    nodes, alias split candidates, at the $i$-th loop.
    repeat
        for each $m \in \mathcal{M}_i(\mathcal{T}_b)$ do
            Draw a random sample $\tilde{\mathbf{X}}$ of size $\tilde{p} < p$ without replacement from $\mathbf{X}$.
            for each $\tilde{\mathbf{X}}_j$ in $\tilde{\mathbf{X}}$ do
                Determine the split which minimises the impurity of $m$ among all possible
                splits of $m$.
            end for
            Find $\tilde{\mathbf{X}}_{j*}$ that generates the split which minimises the impurity of $m$ among
            all $\tilde{\mathbf{X}}_j$.
            Use $\tilde{\mathbf{X}}_{j*}$ to create a binary split of $m$.
        end for
    until $\mathcal{M}_{i+1}(\mathcal{T}_b) = \emptyset$, i.e. each terminal node at this stage has an irreducible impu-
    rity, or contains a pre-specified minimum number of observations, $n_{\min}$.
end for
Take the unweighted majority vote of the tree classifications as the forest classification.
```

5.1 Classical approach

The classical RF algorithm is an ensemble ML method that has been developed by Breiman (2001). It is based on bootstrap aggregation (bagging) developed by Breiman (1996a) and further analysed by Bühlmann and Yu (2002) and is applicable equally well to regression and classification problems. Here, we restrict ourselves to the latter case.

5.1.1 Algorithm

The basic idea of the classical RF algorithm is to grow a large and diverse, i.e. decorrelated, set of unpruned binary classification trees built upon bootstrap samples of the training data. The classifications made by the single trees are then aggregated in order to smooth the hard cut decisions of the binary splits, which usually results in an improvement in classification accuracy. During tree growing, a fresh sample from the set of available predictors is considered for each node splitting. This is done to prevent strong predictors in the entire set from dominating all other predictors, which in turn increases the diversity of the single trees compared to bagging, where all predictors are considered at each split. Algorithm 2 formalises this basic principle.

Remark 6. Measuring node impurity, which in general can be done in a variety of ways, is key to Algorithm 2. A popular measure is the Gini index. Let $\hat{q}_{mk} = N_m^{-1} \sum_{\mathbf{x}_i \in R_m} \mathbb{1}_{\{y_i=k\}}$

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 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ and R_m denote the i-th observation of the p predictors and the classification region corresponding to node m, respectively. Then, the Gini index is given by

$$Q_m(\mathcal{T}_b) = \sum_k \hat{q}_{mk} \left(1 - \hat{q}_{mk}\right). \tag{12}$$

The cross-entropy and misclassification error provide alternative impurity measures which are also based on $\hat{q}_{mk}$ (Hastie et al., 2009). □

Remark 7. A celebrated advantage of RFs is the possibility of using subsets of the training data for validation purposes. Let $\mathcal{O}_b = \mathcal{L} \setminus \mathcal{L}_b$ be the "out-of-bag" (OOB) data of the b-th bootstrap sample, i.e. 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. □

5.1.2 Variable importance

For a single classification tree, the importance of a given predictor $\mathbf{X}_j$ is determined directly by its position in the tree. However, this concept does not translate to the classical RF algorithm in a straightforward way. Therefore, two types of variable importance measures have been suggested which essentially quantify the mean decrease in node impurity and prediction accuracy caused by the predictor and its absence, respectively.

Mean decrease in node impurity. The basic idea of binary node splitting is to separate a set of observations into two sets that are less heterogenous than the original one. Since strong splitting variables should give greater reduction in heterogeneity, the importance of any predictor can be quantified by its average contribution to the decrease in node impurity. Let $\mathcal{M}(\mathcal{T}_b, \mathbf{X}_j)$ be the set of internal 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 internal nodes in the forest. Measuring node impurity by the Gini index (12), 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. Essentially, this measure is the average difference between the impurities of the nodes split by $\mathbf{X}_j$ and the weighted sum of the impurities of their descendant nodes. Its effectiveness has been studied by Archer and Kimes (2008) by means of a large-scale simulation study tailored to the empirical characteristics of microarray gene expression data.

Mean decrease in prediction accuracy. Alternatively, variable importance can be measured by the mean decrease in prediction accuracy after randomly permuting the values of $\mathbf{X}_j$ 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

$$\mathrm{VI}^P(\mathbf{X}_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i\in\mathcal{O}_b}\left[\frac{\mathbb{1}_{\{y_i\neq\hat{y}_i(\mathcal{T}_b,\mathbf{X}_{\pi(j)})\}}}{|\mathcal{O}_b|} - \frac{\mathbb{1}_{\{y_i\neq\hat{y}_i(\mathcal{T}_b,\mathbf{X}_j)\}}}{|\mathcal{O}_b|}\right]. \tag{13}$$

Sometimes, this measure is normalised using the standard deviation of the differences between the misclassification rates in the OOB samples. However, Díaz-Uriarte and de Andrés (2006) find that the unscaled version (13) is preferable as it allows to compare outcomes obtained from different parameter settings, especially the number of trees in each forest and the number of predictors sampled at each split.

5.2 Conditional inference trees

RFs based on conditional inference trees deviate from the classical RF approach in two respects: variable selection and variable importance measures.

5.2.1 Variable selection

In classical RFs, variable selection tends to be biased towards predictors with larger measurement scales, higher numbers of categories and, sometimes, missing values (Hothorn et al., 2006; Strobl et al., 2007). 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; Webel and Ollech, 2017), which is a consequence of the linkage between variable selection and node splitting. More precisely, in the absence of a truly influential predictor in a sample $\tilde{\mathbf{X}}$ of candidate predictors, a predictor that is highly correlated with the truly influential predictor but not with the response may be selected as splitting variable. In this case, the substitute only appears to be an effective splitting variable as the true dependencies between the predictors are not taken into account appropriately. As a consequence, its influence on the response is likely to be overestimated, regardless of the considered variable importance measure.[9]

To overcome this drawback, Hothorn et al. (2006) develop a conditional inference framework for decision trees which is based on the theory of permutation tests developed by Strasser and Weber (1999) and avoids potential biases by untangling variable selection and node splitting. The rationale of this separation is an ex ante exclusion of those predictors which are not strongly related to the response. Using notation from Algorithm 2, let each split candidate $m \in \mathcal{M}_i(\mathcal{T}_b)$ be represented by a N-dimensional vector $\mathbf{w}_m = (w_{m,1}\dots w_{m,N})^\top$ of integer case weights, where $w_{m,i}$ is positive if the i-th observation $(\mathbf{x}_i, y_i)$ is an element of node m and zero otherwise. Then, they propose the following generic 2-step algorithm which essentially replaces lines 7 to 11 of Algorithm 2:

1. **Selection step**: Test the global null hypothesis of no association between $\mathbf{Y}$ and each $\tilde{\mathbf{X}}_j$ in the sample $\tilde{\mathbf{X}}$ given the node's case weights $\mathbf{w}_m$. Stop if this hypoth-

[9] This effect tends to be higher/lower for lower/higher values of $\tilde{p}$, i.e. the number of candidate predictors in the sample (Grömping, 2009; Strobl et al., 2008). However, increasing $\tilde{p}$ in order to remedy the bias issue may again lead to an unwanted dominance of strong predictors.

esis cannot be rejected. Otherwise, identify the predictor $\tilde{\mathbf{X}}_{j*}$ with the strongest association to $\mathbf{Y}$.

2. **Split step**: Take $\tilde{\mathbf{X}}_{j*}$ as splitting variable. Find the optimal binary split of node m using a pre-specified splitting criterion. Calculate the case weights $\mathbf{w}_{m_L}$ and $\mathbf{w}_{m_R}$ of the left and right descendent nodes of m.

Remark 8. In Step 1, the global null hypothesis of no association between $\mathbf{Y}$ and each $\tilde{\mathbf{X}}_j$ in node m can be written formally as

$$H_0^{(m)} = \bigcap_{j=1}^{\tilde{p}} H_0^{(m,j)} \quad \text{with} \quad H_0^{(m,j)} : \mathcal{D}(\mathbf{Y}|\tilde{\mathbf{X}}_j, \mathbf{w}_m) = \mathcal{D}(\mathbf{Y}|\mathbf{w}_m),$$

where $\mathcal{D}(\mathbf{Z})$ denotes the distribution of $\mathbf{Z}$. This hypothesis is rejected if the minimum of the (adjusted) p-values for rejecting the local null hypotheses $H_0^{(m,j)}$ is smaller than a pre-specified nominal level α. In this case, $\tilde{\mathbf{X}}_{j*}$ can be identified from standardised linear statistics within the permutation test framework or from the local null hypothesis $H_0^{(m,j*)}$ which has been rejected at the smallest (adjusted) p-value. □

Remark 9. 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. □

Remark 10. The key idea of Hothorn et al. (2006), i.e. separating variable and split point selection, has also been considered in others approaches in order to deal with the selection bias of classical RFs. Prime example are the QUEST and CRUISE methods suggested by Loh and Shih (1997) and Kim and Loh (2001), respectively. Utilising conditional independence tests, Lee and Shih (2006) extend the selection schemes of these methods to classification trees with multivariate responses. □

5.2.2 Variable importance

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 such 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 13 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

$$\mathrm{VI}^{CP}(\mathbf{X}_j) = \frac{1}{B}\sum_{b=1}^{B}\sum_{i\in\mathcal{O}_b}\left[\frac{\mathbb{1}_{\{y_i\neq\hat{y}_i(\mathcal{T}_b,\mathbf{X}_{\pi(j)|\mathbf{x}_j^c})\}}}{|\mathcal{O}_b|} - \frac{\mathbb{1}_{\{y_i\neq\hat{y}_i(\mathcal{T}_b,\mathbf{X}_j)\}}}{|\mathcal{O}_b|}\right], \tag{14}$$

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.

Remark 11. Conditioning on $\mathbf{X}_j^c$, that is on all other predictors, in the permutation scheme might be seen as a very cautious strategy. A slightly more incautious scheme could consider only those predictors in $\mathbf{X}_j^c$ whose correlation with $\mathbf{X}_j$ exceeds a certain threshold (Strobl et al., 2008). In this case, 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 used. Either way, it should also be kept in mind that the differences between classical and conditional RFs primarily concern variable importance measures and tend to be negligible in terms of misclassification rates (Hothorn et al., 2006; Webel and Ollech, 2017). □

6 Application

Section 6.1 applies the classical and conditional RF approaches to the training data that was generated in Section 3 from the set of 600,000 simulated ARIMA time series and identifies the most informative seasonality tests from the latter approach.[10] Section 6.2 applies the conditional RF approach to the four time series from Example 1, disentangling especially the conflicting test results for the CPI and employment series.

6.1 Simulated ARIMA time series

We run Algorithm 2 with $B = 100$, which was determined by cross-validation, $\tilde{p} = \lfloor\sqrt{p}\rfloor = 2$ and $n_{\min} = 1$ in order to grow classical RFs. Essentially, we use the same parameters in the conditional RF-modified algorithm (Remark 8 and Remark 9) alongside $\alpha = 0.05$. Employing univariate p-values, we thereby identify predictors with the strongest association to the response according to the "smallest p-value"-approach. These choices are in line with the suggestions of Díaz-Uriarte and de Andrés (2006) and Strobl et al. (2007).

6.1.1 Misclassification rates

Table 5 reports the misclassification rates of the candidate seasonality tests and of the classical and conditional RFs. The key findings are that (1) each test shows non-ignorable upward size distortions, especially for longer series, and (2) the performance of either RF approach is, by construction, completely independent of any pre-specified level of significance and barely affected by the length of the series.

Seasonality tests. The modified QS test has the lowest misclassification rates for seasonal series regardless of the series' length, whereas the Friedman test performs generally well at correctly classifying non-seasonal series, especially at the 1% level of significance. Interestingly, the misclassification rates of the modified QS, Friedman and Kruskal-Wallis tests increase more or less noticeably with the length for non-seasonal series, while the same is true for the misclassification rates of the periodogram test and F-test on seasonal dummies for seasonal series.

[10] Recall that each training data set contains the p-values of five JD+ seasonality tests as predictors and a seasonality dummy as categorical response. Also, recall that these predictors display high positive empirical correlations (Table 4) and, thus, justify the consideration of conditional inference trees.

Remark 12. Although the modified QS test has the lowest misclassification rate for seasonal series among all competitors, it misclassifies non-seasonal series relatively often compared to the other tests, especially for longer series. This discrepancy between the "type I" and "type II" misclassification rates can in part be explained by the fact that the modified QS test is probably most sensitive to under-differencing. For example, it falsely classifies 64.2% of the (021)(000) models as seasonal at the 1% level of significance, which accounts for 1.47 percentage points of the overall misclassification rate for non-seasonal series. In contrast, the misclassification rates of the other tests range between 0.1% and 6.3% for this particular model class. □

Remark 13. The test for seasonal peaks combines two spectrum-based diagnostics and,

Table 5: Misclassification rates of the JD+ seasonality tests and the classical and conditional RF approaches.

			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 RF	OOB	Mean*	0.76	2.08	0.63	2.38	0.83	1.99	0.82	1.87
		SD**	0.16	0.17	0.25	0.38	0.25	0.37	0.21	0.29
	VAL	Mean*	0.77	2.13	0.66	2.35	0.82	2.05	0.82	1.98
		SD**	0.17	0.24	0.22	0.33	0.24	0.39	0.28	0.42
Classical RF	OOB	Mean*	0.84	1.91	0.75	2.26	0.88	1.82	0.89	1.66
		SD**	0.16	0.16	0.27	0.35	0.22	0.34	0.25	0.29
	VAL	Mean*	0.82	1.98	0.74	2.25	0.85	1.90	0.88	1.77
		SD**	0.18	0.19	0.24	0.32	0.24	0.34	0.30	0.36
Seasonality tests	$\alpha = 0.01$	QS	4.84	1.49	2.48	1.75	4.99	1.41	7.07	1.29
		FT	2.08	2.09	1.46	2.25	2.25	1.95	2.52	2.07
		KW	2.37	3.78	1.86	3.90	2.56	3.68	2.70	3.75
		PD	3.21	3.65	3.18	3.42	3.30	3.60	3.16	3.91
		SD	4.04	2.70	4.35	2.50	4.09	2.69	3.67	2.91
		SP	∗	∗	∗	∗	6.63	1.74	5.04	1.95
	$\alpha = 0.05$	QS	7.42	1.22	4.92	1.39	7.52	1.11	9.82	1.15
		FT	6.64	1.58	5.63	1.62	6.96	1.50	7.33	1.64
		KW	6.91	3.15	6.24	3.13	7.12	3.07	7.36	3.24
		PD	8.14	3.16	8.31	2.84	8.15	3.11	7.95	3.51
		SD	9.11	2.24	9.68	2.03	9.19	2.22	8.46	2.48
		SP	∗	∗	∗	∗	6.63	1.74	5.04	1.95

$\star$ As a percentage. $\star\star$ In percentage points.
Remarks: N-S and S denote the classes of non-seasonal and seasonal series. Seasonality tests have been applied to the "mother" training data $\mathcal{L}$, whereas RF approaches have been applied to the "daughter" training data sets $\mathcal{L}^{(i)}$, $i \in \{1, \ldots, 50\}$. Means and standard deviations (SD) of the misclassification rates have been calculated over the respective OOB samples and sampled validation (VAL) data sets $\mathcal{V}_s^{(i)}$. The color code, which indicates certain clusters of tests, is explained in Remark 13.

thus, does not obey any traditional null distribution, including p-values. As a consequence, its misclassification rates do not follow a pattern similar to any of the other tests. For instance, the misclassification rates are not affected by the user's choice of a significance level. This distinctive feature is in line with grouping all seasonality tests into the following three clusters, which are highlighted in Table 4 and Table 5 in different shades of grey:

- The dark grey cluster consists only of the modified QS test. It is able to handle moving seasonality as well as volatile seasonal patterns. This results in particularly low misclassification rates for seasonal time series.
- The grey cluster contains the Friedman, Kruskal-Wallis and periodogram tests and the F-test on seasonal dummies. This set is designed to identify stable seasonality, e.g. by checking differences between period-specific means. They tend to have lower misclassification rates for non-seasonal series and slightly higher misclassification rates for seasonal series compared to the modified QS test.
- The light grey cluster consists only of the test for seasonal peaks that directly classifies a series as seasonal or non-seasonal instead of yielding p-values. As opposed to the modified QS test, the overall misclassification rate slightly improves with the time series length so that the test has particularly low misclassification rates for longer seasonal series. □

Classical RF. Based on the OOB data, the average misclassification rates are universally lower than the respective rates of the seasonality tests for the non-seasonal series and slightly higher than the rates of the best tests for seasonal series. Compared to this data, the average misclassification rates on the sampled VAL data are even marginally smaller for the non-seasonal series but slightly higher for the longer seasonal series. Also, the respective standard deviations are slightly lower for shorter series and slightly higher for longer series, regardless of seasonality class.

Conditional RF. Compared to classical RFs, the average misclassification rates on both the OOB and sampled VAL data are slightly smaller for the non-seasonal series and visibly higher for the seasonal series, especially the longer ones. In contrast, the standard deviations of the misclassification rates are almost the same for either method. The only exception are the longer seasonal series in the sampled VAL data for which a slight increase in the standard deviations can be observed for the conditional RF approach compared to classical RFs.

6.1.2 Variable importance measures

Figure 3 shows boxplots of the mean decrease in accuracy (Equation 13 and Equation 14) for the five JD+ seasonality tests whose p-values have been involved in tree growing. The left panel reveals that in the conditional RF setup the modified QS test eventually turns out to win the race for the most informative test by a narrow margin over the Friedman test. The F-test on seasonal dummies and the Kruskal-Wallis and periodogram tests finish far behind at third, fourth and fifth place, although the dispersion of the mean decrease in accuracy is remarkably lower for these tests. The right panel shows that the

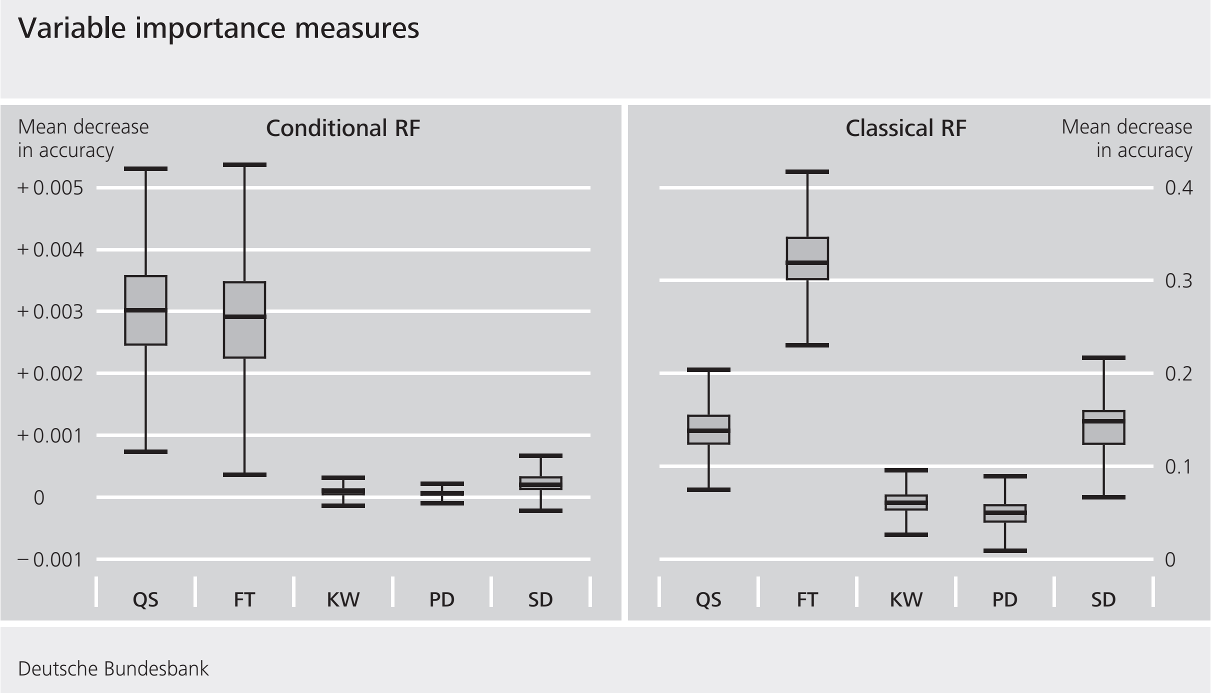

Figure 3: Variable importance measures for five JD+ seasonality tests based on the conditional and classical RFs grown on the 50 independent "daughter" training data sets.

Friedman test and the F-test on seasonal dummies appear much more important in the classical RF setup, which does not properly take into account the p-values' correlations.

Remark 14. The identification of the modified QS and Friedman tests as the most informative JD+ seasonality tests may be explained intuitively by the fact that the two tests complement each other well with respect to the different types of seasonal pattern they are designed to capture. While the Friedman test mainly covers stable seasonality, the modified QS test can also cope with seasonal patterns that change gradually over time. This is in line with the test clusters discussed in Remark 13. □

Remark 15. The informational content of the Friedman and Kruskal-Wallis tests is noticeably different, although their test statistics are structurally quite similar. This suggests that intra-year ranks seem to be more informative than intra-span ranks. The reason could be that intra-year ranks are more robust against non-monotonous trend-like behaviour that may still remain in the input series even after first-order differencing. □

6.2 Real-world time series

To demonstrate the benefits of the conditional RF approach, we return to the four time series discussed in the introductory Example 1, i.e. retail trade turnover for games and toys, the HICP for tobacco, the CPI for energy, and the number of persons employed in the manufacture of wearing apparel. Table 6 recalls that the seasonality tests in JD+ unanimously classify the first series as seasonal and the second series as non-seasonal, whereas they disagree for the other two series.

Table 6 also shows the aggregated classification of each series based on the conditional RF approach. This first confirms the two concurrent decisions of the JD+ tests, i.e. the

Table 6: Test statistics (TS) and p-values of the JD+ seasonality tests and conditional RF classification for the time series shown in Figure 1.

	Retail trade turnover: games and toys		HICP: tobacco		CPI: energy		Employed persons: manufacture of wearing apparel	
	TS	p-value	TS	p-value	TS	p-value	TS	p-value
QS	537.787	0.000	0.750	0.687	7.463	0.024	6.551	0.038
FT	236.337	0.000	13.031	0.291	29.804	0.002	46.864	0.000
KW	258.577	0.000	5.363	0.912	33.755	0.000	44.127	0.000
SP	AT AT AT AT AT AT		⋆⋆ a⋆ ⋆⋆ ⋆⋆ ⋆⋆ a⋆		⋆t ⋆⋆ ⋆⋆ ⋆⋆ a⋆ ⋆⋆		⋆⋆ A⋆ ⋆⋆ ⋆⋆ ⋆⋆ ⋆⋆	
PD	323.551	0.000	0.376	0.965	3.140	0.001	1.796	0.062
SD	364.898	0.000	0.359	0.970	3.031	0.001	1.747	0.071
RF	Seasonal		Non-seasonal		Non-seasonal		Seasonal	

Remark: See Table 1 for further details and note that the conditional RFs yield an unanimous classification of each series.

turnover series is classified as seasonal and the HICP series as non-seasonal. Regarding the conflicting results, the CPI series is classified as non-seasonal, in contrast to the majority vote of the tests, and the employment series as seasonal, in line with the Friedman and Kruskal-Wallis test results. The decision in favour of absence of identifiable seasonality in total energy prices seems reasonable given that slightly more than 60% of the subcomponents, such as fuels and lubricants for personal transport equipment, heat energy and liquid fuels, are classified as non-seasonal by all JD+ seasonality tests.

7 Summary

We argued that the identification of the seasonal status of observed data is essentially a classification task and can thus be performed by machine learning (ML) methods, using the outcomes of single seasonality tests as predictors that are potentially subject to positive correlations. Hence, ML methods may also help to eliminate seemingly redundant tests and to identify the most informative tests within a given set. Working with the seasonality tests implemented in JD+ and asking for high accuracy, high interpretability and availability of unbiased variable importance measures in the presence of correlated predictors, we compared selected ML methods in terms of their capability to balance these key requirements and identified random forests (RF) of conditional inference trees as the best method in that sense. Thereby, each method was trained and evaluated on a large set of simulated monthly seasonal and non-seasonal ARIMA time series which was obtained from combining the "NORmal-To-Anything" (NORTA) algorithm with logspline density estimation in order to be as representative of the Bundesbank's macroeconomic time series database as possible. Utilising RFs of conditional inference trees, we finally identified the modified QS and Friedman tests as the most informative seasonality tests among the ones implemented in JD+. An intuitive explanation may be that the two tests together cover a broad range of seasonal patterns as the Friedman test mainly captures stable seasonality while the modified QS test allows seasonality to gradually change over time.

Our RF-based approach to assessing the informational content of seasonality tests can be extended in a variety of ways. For example, the representative training data could be improved further by (1) becoming also unbalanced with respect to seasonality classes and time series lengths, (2) adding quarterly, outlier-prone and/or borderline seasonal series, or (3) using models other than ARIMA and/or algorithms and methods other than NORTA, such as copulas, in the simulation process. Future research could also consider seasonality tests currently not implemented in JD+ as additional candidate predictors and other RF-based methods as additional classifiers. Since obtaining results from conditional RFs is still time-consuming, drawing also random samples of split points during tree growing as in extremely randomised trees (Geurts, Ernst, and Wehenkel, 2006) could help to save computing time. Also, combining classifications of different (not necessarily RF-based) methods trained on the same data as in stacking (Breiman, 1996b; Wolpert, 1992) could be beneficial, especially in terms of accuracy. In general, considering further methods could provide a starting point for transferring our approach from supervised to unsupervised learning, working entirely with real-world data. Apart from that, changing our approach's attention to tests for residual seasonality, which is sometimes even more difficult to detect (Findley, Lytras, and McElroy, 2017), may also be worthwhile.

Putting some of the above ideas into practice, we are currently elaborating on the work of Webel and Ollech (2018) as we are using the conditional RF-based approach as a main building block for deriving an overall seasonality test by repeatedly eliminating "weak" (alias less informative) predictors from a larger initial set of candidate seasonality tests.

References

Alfaro, E., M. Gamez, and N. Garcia (2018). adabag: Applies Multiclass AdaBoost.M1, SAMME and Bagging. R Package Version 4.2.

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

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

Bank Deutscher Länder (1957, March). Eliminating Seasonal Movements from Series of Economic Data. *Monthly Report 9*(3), 38–47.

Bayer, C. and C. Hanck (2013, January). Combining non-cointegration tests. *Journal of Time Series Analysis 34*(1), 83–95.

Bergamelli, M., A. Bianchi, L. Khalaf, and G. Urga (2019, August). Combining p-values to test for multiple structural breaks in cointegrated regressions. *Journal of Econometrics 211*(2), 461–482.

Bernard, J.-T., N. Idoudi, L. Khalaf, and C. Yélou (2007, December). Finite Sample Inference Methods for Dynamic Energy Demand Models. *Journal of Applied Econometrics 22*(7), 1211–1226.

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152.

Breiman, L. (1996a, August). Bagging Predictors. *Machine Learning 24*(2), 123–140.

Breiman, L. (1996b, July). Stacked Regressions. *Machine Learning 24*(1), 49–64.

Breiman, L. (2001, October). Random Forests. *Machine Learning 45*(1), 5–32.

Breiman, L., A. Cutler, A. Liaw, and M. Wiener (2001). randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R Package Version 4.6-14.

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

Briët, O. J., P. H. Amerasinghe, and P. Vounatsou (2013). Generalized Seasonal Autoregressive Integrated Moving Average Models for Count Data with Application to Malaria Time Series with Low Case Numbers. *PloS ONE 8*(6), 1–9.

Bühlmann, P. and B. Yu (2002). Analyzing Bagging. *The Annals of Statistics 30*(4), 927–961.

Busetti, F. and A. C. Harvey (2003, July). Seasonality Tests. *Journal of Business & Economic Statistics 21*(3), 420–436.

Cario, M. C. and B. L. Nelson (1997). Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix. Technical report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Culp, M., K. Johnson, and G. Michailidis (2016). ada: The R Package Ada for Stochastic Boosting. R Package Version 2.0-5.

Deutsche Bundesbank (1970, March). Seasonal adjustment by the Census Method. *Monthly Report 22*(3), 37–41.

Deutsche Bundesbank (1999, September). The changeover from the seasonal adjustment method Census X-11 to Census X-12-ARIMA. *Monthly Report 51*(9), 39–50.

Díaz-Uriarte, R. and S. A. de Andrés (2006, January). Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinformatics 7*, Article 3.

Dufour, J.-M. (2006, August). Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics 133*(2), 443–477.

Findley, D. F., D. P. Lytras, and T. S. McElroy (2017). Detecting Seasonality in Seasonally Adjusted Monthly Time Series. Research Report 2017-03, U.S. Census Bureau.

Franses, P. H. (1992, March). Testing for Seasonality. *Economics Letters 38*(3), 259–262.

Freund, Y. and R. E. Schapire (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences 55*(1), 119–139.

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis 38*(4), 367–378.

Friedman, M. (1937, December). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association 32*(200), 675–701.

Geurts, P., D. Ernst, and L. Wehenkel (2006, April). Extremely randomized trees. *Machine Learning 63*(1), 3–42.

Ghysels, E. and D. R. Osborn (2001). *The Econometric Analysis of Seasonal Time Series.* Cambridge: Cambridge University Press.

Gómez, V. and A. Maravall (2001). Automatic Modeling Methods for Univariate Series. In D. Peña, G. C. Tiao, and R. S. Tsay (Eds.), *A Course in Time Series Analysis*, pp. 171–201. New York: Wiley.

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

Götz, T. B. and K. Hauzenberger (2018). Large mixed-frequency VARs with a parsimonious time-varying parameter structure. Discussion Paper No 40/2018, Deutsche Bundesbank, Frankfurt.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning – Data Mining, Inference, and Prediction* (Second ed.). Heidelberg: Springer.

Hechenbichler, K. and A. Lizee (2016). kknn: Weighted k-Nearest Neighbors. R Package Version 1.3.1.

Hothorn, T., K. Hornik, C. Strobl, and A. Zeileis (2015). Party: A Laboratory for Recursive Partytioning. R Package Version 1.3-0.

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

Hsieh, C.-H., R.-H. Lu, N.-H. Lee, W.-T. Chiu, M.-H. Hsu, and Y.-C. J. Li (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.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning*. New York: Springer.

Kim, H. and W.-Y. Loh (2001, June). Classification Trees With Unbiased Multiway Splits. *Journal of the American Statistical Association 96*(454), 589–604.

Kooperberg, C. and C. J. Stone (1992, December). Logspline Density Estimation for Censored Data. *Journal of Computational and Graphical Statistics 1*(4), 301–328.

Kruskal, W. H. and W. A. Wallis (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association 47*(260), 583–621.

Lee, T.-H. and Y.-S. Shih (2006, November). Unbiased variable selection for classification trees with multivariate responses. *Computational Statistics & Data Analysis 51*(2), 659–667.

Loh, W.-Y. and Y.-S. Shih (1997). Split Selection Methods for Classification Trees. *Statistica Sinica 7*(4), 815–840.

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

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, and C.-C. Lin (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group, TU Wien. R Package Version 1.6-8.

Patel, J., S. Shah, P. Thakkar, and K. Kotecha (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.

Pinkwart, N. (2018). Short-term forecasting economic activity in Germany: a supply and demand side system of bridge equations. Discussion Paper No 36/2018, Deutsche Bundesbank, Frankfurt.

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

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ripley, B. and M. Venables (2016). nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models. R Package Version 7.3-12.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Samworth, R. J. (2012, October). Optimal Weighted Nearest Neighbour Classifiers. *The Annals of Statistics 40*(5), 2733–2763.

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

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

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

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

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

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

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

Webel, K. and D. Ollech (2018). An overall seasonality test based on recursive feature elimination in conditional random forests. In *Proceedings of the 5th International Conference on Time Series and Forecasting*, pp. 20–31. Granada.

Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks 5*(2), 241–259.