The Dark Side of the Moon: Searching For The Other Half Of Seasonality

Gary Cornwall*
Bureau of Economic Analysis
and
Jeff Chen
Bureau of Economic Analysis

May 26, 2020

Abstract

Seasonality is among the most visible properties in time series data, yet a multitude of statistical tests devised over decades of research have only achieved limited success in its detection. In this paper we examine eight existing tests of seasonality and show that there is significant variation in how they classify a series. We then show how this variation, combined with characteristics of the time series (*e.g.* autocorrelation, frequency, skewness, kurtosis, etc.), can be exploited by a Random Forest (RF) framework to map the hypothesis test space and make more accurate predictions regarding the seasonal disposition of a series. Our proposed method reduces Type II errors by approximately sixty percentage points over the next best alternative.

Keywords: ARIMA, Simulation Study, Negative Seasonality, Random Forest, Residual Seasonality, Seasonal Adjustment

*The authors would like to acknowledge Justine Mallatt, Marina Gindelsky, Scott Wentland, Shelly Smith, Maggie Jacobson, Beau Sauley, and participants of the Midwest Econometrics Group and Seasonal Adjustment Practitioners Workshop for their helpful comments. The views expressed here are those of the authors and do not represent those of the U.S. Bureau of Economic Analysis or the U.S. Department of Commerce. First version 06/01/2019.

1 Introduction

The past several years have proven to be a volatile period for seasonal adjustment. A great deal of attention has been paid to the identification and remediation of residual seasonality, that is seasonal fluctuations which remain post-adjustment (Gilbert et al., 2015; Moulton and Cowan, 2016; Lunsford, 2017; Cowan et al., 2018; Wright, 2018). Detecting seasonality is a unique task in the data provider context since a single analyst may be responsible for identifying seasonal patterns in hundreds if not thousands of underlying series which are used to produce aggregate economic indicators. For many central statistical agencies around the world there is a preference for indirect adjustment; that is to test the underlying components of an economic indicator and adjust them individually before aggregating to a feature statistic (*e.g.* Gross Domestic Product).[1] Indirect adjustment is preferred for two primary reasons. First, it preserves the accounting relationship from the underlying components to the aggregated indicator, and second it allows individual components with wildly different seasonal patterns to be treated separately thus improving the ability to discern seasonal aspects of the data from features of interest. This makes the accuracy of the testing channel of particular importance to providing aggregate series free of seasonal features.

As with any paper covering seasonality it is important to define what constitutes seasonal features in the data. In this paper we use a rather agnostic definition of seasonality, that is if the data can be generated by a seasonal auto regressive integrated moving average (SARIMA) process we consider it to be seasonal. We show, using finite sample simulations, that many of the tests used to detect seasonality are poorly sized under the null, have low power, and often fail completely at recognizing seasonality over entire portions of the parameter space. These include tests such as, the QS test (QS), the F-test for stable seasonality (D8F), the F-test for moving seasonality (FM), the "M7" test (M7), the model based F-test (FMB), the Kruskall-Wallis test (KW), the Welch test (WE), and the

[1]The U.S. Bureau of Economic analysis has relied on indirect seasonal adjustment precisely so that users can trace the estimates of GDP and its components back to source data. See https://www.bea.gov/news/blog/2015-06-10/snapshot-seasonal-adjustment-process-gdp for a brief overview of the process. This method is also the preferred method for Eurostat (https://ec.europa.eu/eurostat/web/sector-accounts/methodology/seasonal-adjustment-key-series), Stats Canada (https://www150.statcan.gc.ca/n1/pub/12-539-x/2009001/seasonal-saisonnal-eng.htm), among others.

Friedman test (FR), many of which form the back bone of testing within the X-13 ARIMA-SEATS (hereafter X13) program provided by the U.S. Census Bureau.[2] Moreover, we show that there is significant variation in how these tests classify any single series meaning that the claim of seasonality, or lack thereof, strictly depends on which test one is willing to believe. We exploit this variation in the test statistics, and the characteristics of the time series itself (*e.g.* autocorrelation, kurtosis, frequency, etc.) to form a *pseudo-composite* test for seasonality using a Random Forest (RF) model (Breiman, 2001).[3] By projecting the hypothesis test(s) of seasonality as a classification problem we are able to reduce Type II errors by nearly sixty percentage points, and improve the overall accuracy of identifying seasonality by nearly thirty one percentage points.

In an optimal environment, when a series exhibits seasonality it would be identified and "perfectly" removed.[4] The implication is that any seasonality exhibited by aggregated indicators is purely an artifact of the aggregation, thus it can be ignored. Unfortunately, despite seasonality being among the most visible properties in time series data, our ability to detect seasonality has only achieved limited success. This is in part driven by a high degree of definitional malleability, a situation which is made clear by Gómez and Maravall (2001) who wrote "... the absence of a well-defined and generally accepted definition has fostered proliferation of procedures, and made it difficult to find common grounds for comparison." Our lack of perfection in testing for seasonality means that residual seasonality can occur through not only the aggregation channel but also through the testing channel. A failure to reject the null of no seasonality, when the null is false, means that aggregated series include seasonal features from some of the underlying components. The goal of this paper is to improve the accuracy of testing for seasonality such that we mitigate the effect the

[2]The X13 program and its predecessors are the tools most widely used for identifying and removing seasonality by practitioners around the world. Other options do exist (*e.g.* JDemetra+) though the tests used for identifying seasonality are largely consistent across platforms.

[3]We calculate these statistics using functions written by Rob J. Hyndman, see https://robjhyndman.com/hyndsight/tscharacteristics/ . These functions are based off of Wang et al. (2006) and Wang et al. (2009).

[4]Of course this is completely unrealistic as Jaynes (2003) noted that "... we do not seek to remove the trend or seasonal component from the data: that is fundamentally impossible because there is no way to know the 'true' trend or seasonal term. Any assumption about them is necessarily in some degree arbitrary, and is therefore almost certain to inject false information into the detrended or seasonally adjusted series. (p. 536)". The thought exercise is informative though in the sense that it gives us a normative view of how seasonal adjustment interacts with the aggregation of the data to provide the statistics of interest.

testing channel has on producing residual seasonality in the aggregates.

Before continuing, we would like to elucidate this concept of definitional malleability. Hopefully by doing this we will be able to contextualize why the definition we use is more agnostic and abstracts away from the arguments that have so often stymied discovery. Lovell (1963) defines seasonality by three main characteristics; orthogonality, idempotency, and symmetry. The seasonal components of a series must be independent of the non-seasonal components (*e.g.* trend or irregular), if a series is adjusted for seasonality which previously has already been adjusted there should be no changes, and if a linear model of adjustment is used, $X_t^a = AX_t$, the matrix A should be symmetric. This is a very specific definition but focuses on the outcome of a seasonal adjustment rather than what a practitioner should be looking for vis-à-vis seasonality. Meanwhile, Nerlove (1964) defines seasonality a bit more narrowly, it is present if there are spectral peaks at seasonal frequencies, where seasonal frequencies are defined through trigonometric functions. More recently Hillmer and Tiao (1982) defined seasonality as period fluctuations which are of similar intensity each year while Harvey (1990) defines seasonality as a mean-zero repetition over any one-year period. Finally, McElroy (2008) defines seasonality in a practical manner based on the behavior of the autocorrelation function (ACF); a series is seasonal if, at the seasonal periods, there are troughs or peaks in the ACF. It should be clear that many of these definitions are similar upon first glance but can produce wildly different interpretations when examined more deeply.

This definitional malleability has produced tests which are constructed in an *ad hoc* fashion, testing a variation on a theme rather than a unified approach to the problem. For example, according to the United States Bureau of Economic Analysis (BEA) and Bureau of the Census (Census), a series is considered seasonal if the F-stable statistic (D8F) is greater than seven. Alternatively, if the M7 statistic, a non-linear combination of the F-stable and F-moving statistics, is less than 1 it indicates identifiable seasonality. Finally, the QS statistic (Maravall, 2012), built on the autocorrelation function of the series, indicates seasonality if the resulting p-value is 0.01 or less (Lytras et al., 2007;

Cowan et al., 2018).[5] Consider that, enclosed within this guidance, there is no mention of fixing the Type I error rate (*alpha*), nor the distribution under the null which produced these critical values. Additionally, it is important to note that these statistical tests are only part of a broad spectrum of practices used to identify seasonal series and our goal with this paper is to evaluate only the test statistics themselves, abstracting away from the numerous interventions based on other information (*i.e.* analyst knowledge, alternative information, etc.).

We are not the first to draw a link between the inherent issues with current tests for seasonality and the potential use of Random Forests. Webel and Ollech (2017, 2018) use what's called a Random Forest of Conditional Inference Trees to choose a set of test statistics which, when used in combination with tree based decision rules, maximize the accuracy of the testing process. Their goal was to eliminate redundant information and in effect prune away tests which provide little information. We see two main issues with this approach; first, the use of p-values in the training and testing portion of the model assumes that the correct critical value under the null is used.[6] Second, and more importantly, the end result does not use the predictive power of the RF (arguably the reason to use it in the first place), but rather continues to use flawed test statistics with more complicated decision rules. We would also like to point out that, much like the decision rules outlined earlier, there is no ability to control the Type 1 error rate in this environment nor is it clear what the Type I error rate is under such a paradigm.

The remainder of this paper is structured as follows. Section 2 outlines the notation we will use throughout the remaining sections and outlines random forest modeling. Section 3 uses a simulated environment to illustrate the properties of existing tests and compares the results of RF predictions to the next best test. Section 4 contextualizes these results using two case studies. Finally, Section 5 concludes and provides additional avenues for research.

[5]While the cut-offs outlined in Lytras et al. (2007) are used by the Census, additional information can be found at https://www.census.gov/ts/papers/G18-0_v1.1_Seasonal_Adjustment.pdf which applies this more broadly to seasonal adjustment performed by the Census.

[6]In short a series is seasonal if the p-value of the QS-test is less than 0.01 or the p-value of the Kruskall-Wallis test is less than 0.002.

2 Detecting Seasonality

We begin by introducing the notation we will use hereafter. Consider the model,

$$\phi(B)\Phi(B^f)Y_t = \theta(B)\Theta(B^f)\epsilon_t, \tag{1}$$

where Y_t is the observed time series – indexed by $t = (1, \ldots, T)$ and periodicity $f \in \{4, 12\}$ – the standard and seasonal backshift operators are represented by B and B^f respectively. The terms $\phi(B)$ and $\Phi(B^f)$ are the p^{th} order autoregressive (AR) polynomials in B and B^f respectively, *e.g.* $\phi(B) = -\phi_1 B - \phi_2 B^2 - \ldots \phi_p B^p$, etc., while $\theta(B)$ and $\Theta(B)$ are the q^{th} order moving average (MA) polynomials. The noise term, ϵ_t, is an independently and identically distributed Gaussian process, $\epsilon_t \sim (0, \sigma^2)$. Going forward we will describe this model in the standard $(p\ d\ q)\ (P\ D\ Q)^f$ notation, for example $(2\ 0\ 1)\ (1\ 0\ 0)^{12}$ would indicate an AR(2) process in Y with an MA(1) process in ϵ, and an AR(1) seasonal process in Y with monthly periodicity in the observed series. The order of integration for the series is denoted by d for the series process and D for the seasonal.

A series will be considered seasonal if $\Phi \neq 0$ or $\Theta \neq 0$. Moreover, we are approaching this from the perspective of a data provider. That is, our end goal is not to estimate a model, or provide inferences regarding the development of a series. Rather, we aspire to identify candidates for seasonal adjustment. We will assume that the ARIMA portion of the process, represented by ϕ and θ are adequately modeled. More precisely, this means that our data generating process for the simulations contained herein is some variation of $(0\ 1\ 0)(P\ 0\ Q)^f$. This is akin to examining the residuals of a fitted ARIMA model which may or may not be seasonal depending on the values of P or Q.

As previously mentioned, each test for seasonality outlined was created with a slightly different definition in mind. As a result there is significant variation in the classification of series between the tests. For example, in Figures 1a to 1f we show this variation over $100,000$ simulated series with quarterly frequency for six tests as compared to the Kruskal-Wallis. Series in red are generated with a seasonal pattern while those in blue are a simple random walk process (white noise in first difference). Note that for those series in the lower left quadrant both test statistics fail to reject the null hypothesis of no seasonality;

red series in this area are the Type II errors we are looking to reduce. Those series in the upper right quadrant represent those for which both tests reject the null hypothesis, and the other two quadrants indicate disagreement on the null hypothesis between the two tests. These different patterns of variation lead us to believe that the multidimensional hypothesis space for seasonality is in fact disjoint, a feature for which decision trees and random forests are known to excel.

[Figure 1 about here.]

These trees are created by bootstrapping the observations, and randomly sub-sampling the features of interest. Unlike regression, the objective of Classification and Regression Trees (CART) is to accurately predict membership to a set of classes C, doing so by recursively splitting a sample into smaller, more homogeneous partitions known as nodes. Each split is the result of a search across all features, X, for an optimal threshold θ, first generating candidate splits, and selecting the value that minimizes Gini Impurity (G),

$$G = 1 - \sum_{c=1}^{C} p_c^2, \tag{2}$$

where p_c is the proportion of observations which belong to a class c and $G \in \{0, 1\}$. Returning to our concern of a binary outcome, if a node contains just one class then $p_c = 0$, otherwise $p_c > 0$. Lower values of G indicate greater classification accuracy. The resulting child nodes of the m^{th} split are thus defined by threshold θ_m along feature x_j,

$$r^- = \{r : x_j < \theta\}, \tag{3}$$

$$r^+ = \{r : x_j \geq \theta\}, \tag{4}$$

where r^- and r^+ are partitions of observations that fall below and above the θ, respectively.

The resulting child nodes are further partitioned until pre-defined algorithm stopping criteria are met, such as a minimum node size or the model's overall Gini Impurity reaches a threshold. All nodes in this fully grown tree are referred to as leafs, R_m, and observations within a leaf have an expected value γ_m. The predictions from this hierarchical tree

structure can be written as,

$$T(X;\Theta) = \sum_{m=1}^{M} \gamma_m I(x \in R_m), \tag{5}$$

where T is a tree defined on features X, and with parameters $\Theta = \{R_m, \gamma_m\}^m$. Thus, as CART models are trained, each observation is predicted to have a value of γ_m based on its mapping to leaf node R_m as located by nested binary criteria of X. The method offers a number of benefits, including implicit variable selection, and the ability to flexibly learn the critical value of each seasonality test under different conditions. However, the algorithm does tend to over-fit and produce noisy predictions – a less than desirable condition when detecting seasonality.

The Random Forest algorithm extends the CART machinery through bootstrap aggregation, which has the effect of producing predictions that are robust to over-fitting. The basic algorithm is simple and repetitive:

1. Construct B number of samples with replacement of n observations and randomly draw $k < K$ variables.
2. Train CART model T_b on the b^{th} sample.
3. Average the predictions from each T_b to obtain a probability of membership with class $C = c$.

We define this probability as $\hat{p}_i = \frac{1}{B}\sum_{b=1}^{B} T_b(X;\Theta)$ where $\hat{p}_i = \frac{1}{B}\sum_{b=1}^{B} R_{mb}$ and $\hat{p}_i \geq 0.5$ is flagged as a seasonal series.

In Figure 2 we outline the overall flow of the RF algorithm. Random sampling through the bootstrap is followed by a partition into training and out-of-bag (OOB) data. The OOB is used to calculate the error rate which describes the overall fit of a single decision tree. From the training set a sub sample of features is chosen at random and a decision tree is grown. This process is repeated a predetermined number of times and the collection of trees are then averaged to create the RF. Random Forests can be tuned for greater accuracy through a model validation framework, focusing on hyper-parameters such as the number of bootstrap iterations, B, and randomly drawn covariates, k. For simplicity, we fix $B = 500$, and k to $\sqrt{K}$.

[Figure 2 about here.]

3 Simulations

This section is broken into four subsections. In the first we discuss the simulation environment including the data generating process and method used to determine the appropriate parameters. In Section 3.2 we outline new facts about the existing tests for seasonality mentioned earlier. Finally, in Section 3.3 we discuss the performance of the Random Forest from an out-of-sample perspective and compare those results to the next best available test. Finally, in Section 3.4 we summarize all of the findings both about existing tests and the RF approach.

3.1 Simulation Structure

To study the efficacy of a composite test for seasonality produced by a random forest, we use a simulated environment to create both a training and test set of data. Each set consists of $M = 5,000,000$ simulated time series, indexed by $m = (1, \ldots, M)$, of varying length and structure. For the sake of brevity we will contextualize the simulation structure through the training set. To begin, let $\pi_s \sim U(0,1)$ be the probability of a series being seasonal or not such that,

$$m_s = \begin{cases} \phi(B)Y_t = \epsilon_t & \text{for } 0.00 \leq \pi_s < 0.50 \\ \phi(B)\Phi(B^S)Y_t = \epsilon_t & \text{for } 0.50 \geq \pi_s < 0.75 \\ \phi(B)Y_t = \Theta(B^S)\epsilon_t & \text{for } 0.75 \geq \pi_s \leq 1.00. \end{cases} \tag{6}$$

Note that in this structure we are not allowing a "mixed" seasonal structure, that is one which contains both AR and MA components as originally outlined in Equation 1. This restriction is mainly to limit the number of possible cases within the simulation study thus providing sufficient sampling within the model space, but is also a recommended guideline used by the U.S. Census Bureau in their seasonal adjustment practices.[7]

[7] See https://www.census.gov/ts/papers/G18-0_v1.1_Seasonal_Adjustment.pdf, page 2, for what specifically constitutes a mixed model.

Conditional upon a series being seasonal, $\pi_s \geq 0.50$, the dimensionality, $m_d(\Phi, \Theta) \in \{1, 2, 3\}$, is determined through probability $\pi_d \sim U(0, 1)$ such that,

$$m_d(\Phi, \Theta) = \begin{cases} 1 & \text{for } 0.00 \leq \pi_d < 0.33 \\ 2 & \text{for } 0.33 \leq \pi_d \leq 0.66 \\ 3 & \text{for } 0.66 < \pi_d \leq 1.00. \end{cases} \tag{7}$$

For example, $m_d(\theta) = 3$ implies a seasonal series of the MA type with three seasonal lags. This means that, of the 2.5 million series we generate that are seasonal, approximately $400,000$ series of each seasonal structure (sAR and sMA) and dimensionality will be generated. In all cases the value of $\phi = 1$ and $\sigma^2 = 1$, thus the AR process for each DGP is a random walk. The only difference is the presence, position, and dimensionality of the seasonal lags.

Two main considerations of the data are left to be outlined. First, frequency of the data, which we will define as $f \in \{4, 12\}$, is determined through a draw from $\pi_f \sim U(0, 1)$ such that,

$$m_f = \begin{cases} 4 & \text{for } 0.00 \leq \pi_f < 0.50 \\ 12 & \text{for } 0.50 \leq \pi_f \leq 1.00. \end{cases} \tag{8}$$

This implies a balanced sample of quarterly and monthly data sets.[8] Finally we determine the length of the series by drawing from a $\pi_l \sim U(10, 50)$ distribution. Taking the floor of each draw from this distribution, $m_l = \lfloor \pi_l \rfloor$, means that the series we generate are only complete year series between ten and fifty years in length with the final number of observations determined via $T = m_f \times m_l$. Combining the notation above each series can be characterized by $m_{\{s,d,f,l\}}$, for example a series listed as $m_{2,2,4,12}$ would indicate a series generated from a $(0\ 1\ 0)(2\ 0\ 0)^4$ process with 12 full years of data, $T = 48$. Based on this outline there are 482 unique data structures which are contained with in simulated sample space. Figure 3 outlines the process flow for the data generating process that governs the

[8]Note that, at this time we are limiting the frequency only to monthly and quarterly primarily because these are the data frequency examined by most statistical agencies. This could be expanded to include more high/low frequency data sets if appropriate.

simulations.

[Figure 3 about here.]

Once the features of each data set, m, have been determined we generate a vector of parameters subject to the constraint(s) $\sum_{d=1}^{3} |\Phi_d| < 0.95$ or $\sum_{d=1}^{3} |\Theta_d| < 0.95$ for any series which is seasonal. This vector of parameters is distributed as $\text{MVN}(0, I\sigma^2)$ where $\sigma = 0.25$. This means that 99% of all parameter draws fall in the interval of $(-0.644, 0.644)$, and the constraint implies that the seasonal portion of the series remains stationary.

3.2 New Facts, Old Tests

Each of the alternative tests are outlined in Appendix A, and as mentioned previously these include the major tests used in X13 and those found in both alternative adjustment software and literature. One of our primary concerns is the lack of knowledge around the existing tests, specifically with respect to the null distribution and the critical values a practitioner should use to control Type I errors (see Dagum (2005) & Ladiray and Quenneville (2012) among others). This section focuses on evaluating the empirical size and power for each of the eight tests of seasonality we will utilize in the RF so as to provide a basis for comparison, not only within the existing tests but between them and the proposed alternative.

To examine the distribution of each test statistic under the null we simulated $20,000$ series from a $(0\ 1\ 0)(0\ 0\ 0)^s$ SARIMA where $s \in \{4, 12\}$. In all cases we set $T = 1200$ so as to approximate long finite sample conditions that might feasibly be seen from the standpoint of a central statistical agency.[9] For each of the $20,000$ series we record each test statistic to form an empirical approximation of the distribution under the null hypothesis. From this we are able to evaluate if the recommended critical value (*e.g.* χ^2_2 for the QS statistic) is close to the appropriate quantile of the empirical distribution.

[Table 1 about here.]

[Table 2 about here.]

[9]For context, the BEA tracks Gross Domestic Product on a quarterly basis from 1947 to present; this entire series is $T < 300$.

In Tables 1 and 2 we show that the current critical values for roughly half the tests are poorly sized when considering a 5% Type I error rate. The QS, D8F, and M7 are all undersized tests in both quarterly and monthly data which indicates a false reduction in power, and a false increase in the Type II error rate relative to what it would be if the Type I rate were fixed correctly at 5%. Conversely the FM statistic is over-sized under the null – dramatically so in the monthly frequency data – meaning that it over-rejects and presents with false power and lower Type II error rates than would otherwise be the case. The remaining tests – the FMB, WE, KW, and FR – are all appropriately sized under the null distribution.

[Figure 4 about here.]

In Figure 4 we plot both the nominal and empirical distributions of the QS and FM statistics under the null to illustrate the key differences. Recall that we used $T = 1200$ to construct these test statistics and it is expected that, though this is a finite sample, the nominal and empirical distributions should be similar and/or converging. For the QS statistic, the null distribution of which is unknown but can be reasonably approximated by a χ^2_2 distribution (Maravall, 2011), the empirical distribution is far from the nominal. The resulting critical values for $\alpha = 5\%$ are roughly 40% lower than the assumed. This figure makes it clear that using the nominal critical value(s) facilitates an under-rejection of the null, which has implications for the test's power when the null is false.

The story however is reversed with respect to the FM statistic. The nominal critical values are too small, which leads to an over-rejection of the null. This makes the FM test look more powerful than it actually is. The key take away from Figure 4, and both Tables 1 and 2, is that even at $T = 1200$ the empirical null distribution is often nowhere near the assumed null distribution. This has implications on everything from the critical values under some assumed α, to the p-values calculated for the test statistic. In Appendix B we provide tables that indicate recommended critical values for each test statistic from small sample sizes (five (ten) years of monthly (quarterly) data) to the larger sample size outlined in Tables 1 and 2. It is important to note that the "settling" time for these test statistics is anywhere between five to forty years worth of data depending on its frequency.

[Figure 5 about here.]

Having fixed the size of each test, we can now turn to comparing the power of the tests, and determine if there is a uniformly more powerful test for seasonality when applied to the SARIMA DGP. The interest here is two-fold, first we want to determine what the most powerful test is so as to provide practitioners with a guide for reliability. Second, we would like to know which single test (or group of tests) to use as a comparison for our results using the RF. The power of each test is shown using the empirical critical values outlined earlier. To illustrate the power of each test we have plotted both $(0\ 1\ 0)(2\ 0\ 0)^4$ and $(0\ 1\ 0)(2\ 0\ 0)^{12}$ seasonal processes which were generated through the aforementioned simulation parameters. The AR plots found in Figure 5 are shown through a contour plot of the parameter space for Φ. Each point represents a series with coordinates (ϕ_1, ϕ_2), and for simplicity we have omitted the case where $\phi_1 = \phi_2 = 0$. This means that every series in this plot is by definition seasonal, and that a perfect test for seasonality would reject the null for all of the series. If a test correctly rejects the null of no seasonality then the point is red while if it fails to reject the null it is blue.

[Figure 6 about here.]

From these plots, and those included in Figure 6, we can see that, of the eight tests, the QS is the most powerful over the domain of the parameter space with power approaching 40%. It is interesting to note that for all tests with the exception of the FM, the coverage is limited to only the right half of the parameter space. If $\phi_1 < 0$, many of these tests will always fail to reject the null when the null is false. These results hold not only as the dimensionality changes but across seasonal structures (*e.g.* MA models).[10]

3.3 Results from the Random Forest

In order to evaluate results from the RF we used our trained model to predict an out-of-sample test set, comparing the classification of each series to its true seasonal status. With an out-of-bag (OOB) error rate of 17.05% within the training set, predictions on the $5,000,000$ series in the test set have an out-of-sample classification error rate of 5.00%. Since we have a large number of features, some of which are only minimally informative, the

[10]Though not plotted here we do discuss the accuracy of MA models when we examine the results from the RF predictions and provide results through Table 3.

training includes trees which have very little predictive power, and thus create more OOB error. On the other hand, predictions use the entire set of features, including their relative information content, and as a result our out-of-sample predictions are more accurate than the OOB.

In Table 3, we show the accuracy of each test using the empirical critical values for the traditional tests.[11] The first row outlines accuracy over the entire test space previously outlined in Figure 3. The RF predictions outperform the next best alternative by thirty-one percentage points. This top line accuracy measure includes both cases when the null is false (a seasonal series) and true (a non-seasonal series). In the subsequent blocks of each table we outline the accuracy conditional upon the null being false and dimensionality of the seasonal component. Moreover, we break down those blocks based on the domain of the parameters in question.

[Table 3 about here.]

For example, the first block outlines the accuracy of each test conditional upon the DGP being a $(0\ 1\ 0)(1\ 0\ 0)^f$, where $f \in 4, 12$. This of course implies that the null is false, or more plainly that every series in this block is seasonal by construction. Here, the RF predictions outperform the next best alternative by fifty-eight percentage points. If we further condition upon the sign of Φ we see that if $\Phi > 0$ the RF outperforms the best alternative by twenty-nine percentage points, and if $\Phi < 0$ this improvement increases to eighty-three percentage points. Note that in the last column we have included the number of series which exist within that portion of the model space (*e.g.* the $(0\ 1\ 0)(1\ 0\ 0)^f$ space includes a total of $416,529$, simulated series out of the $5,000,000$ generated). This pattern of improvement over the alternatives persists across each of the data segments, with the largest improvements coming when $\Phi_p < 0$.

For series generated with a moving average seasonal process the results are largely consistent with those from the autoregressive structure, however overall accuracy does decline in all cases. For example, when looking at accuracy conditional upon $Q > 0$ the RF predictions are sixty-one percentage points better than the alternative, but five percentage points lower than the corresponding RF accuracy for $P > 0$. Note however, that the loss in

[11] We discuss the empirical table here rather than the nominal and include the latter in Appendix C.

accuracy within the RF between MA and AR structures is smaller (5.26%) than the best alternative (14.71%). Not only is the RF more accurate compared to the alternatives, but it also produces less within variation across data structures.

[Figure 7 about here.]

While Table 3 shows a clear benefit from the RF approach, it is still important to understand what features drive these results. Figure 7 provides the top fifteen features as measured by their overall contribution to the mean reduction of Gini Impurity. The QS Statistic is not only the best performing of the traditional tests for seasonality, it is also one of the most important features in the RF.

Most interesting about this is that components of the QS statistic are as important, if not more so, to the classification of each series than the QS itself. The QS statistic is calculated as,

$$\text{QS} = T(T+2)\left(\frac{\hat{\rho}^2(f)}{T-f} + \frac{[\max\{0, \hat{\rho}^2(2f)\}]^2}{T-2f}\right), \tag{9}$$

with null hypothesis $H_0 : \gamma(m) \leq 0$ for $m \in \{f, 2f\}$. The value $\hat{\rho}(f)$ corresponds to the value of the ACF at a one-year lag, while $\hat{\rho}(2f)$ corresponds to that of a two-year lag, both of which are highly valued by the RF. This is important because the domain of a seasonal autoregressive or moving-average term does not always produce a positive value of $\gamma(m)$ for $m \in \{f, 2f\}$. In Figure 8 we show the accuracy of the RF and QS conditional upon either a $(0\ 1\ 0)(1\ 0\ 0)^{12}$ or $(0\ 1\ 0)(0\ 0\ 1)^{12}$ DGP. Note how the QS statistic has an accuracy of zero if the value of Φ or Θ is less than zero. The series are seasonal by construction, however the negative coefficient produces a value for $\gamma(12)$ which is negative, not positive. As a result the QS is zero by construction, and it fails to reject the null hypothesis even when we know the null hypothesis is false! It is also clear that, even in the portion of the domain under which the QS has positive accuracy, the RF outperforms the alternative by a considerable margin. The fact that components of the QS are as important to the RF as the QS itself means that the QS statistic leaves information on the table with respect to the ACF, even if $\gamma(m) > 0$.

[Figure 8 about here.]

In Figure 9 we extend this plot to a second dimension in the autoregressive and moving-average term. Here we plot the contour of joint distribution for Φ (row one) and Θ (row two) with each point representing a single simulated series with parameter pair (Φ_1, Φ_2) or (Θ_1, Θ_2) respectively. If the test in question rejects the null hypothesis, then the series is red, while those it fails to reject are in blue. A perfect test would reject the null for all of these series as we have omitted the simulated, non-seasonal series, and all points would appear red. However, it is expected that most of the blue series, those in which we commit a Type II error, will be close to the center of the distribution. For example, this is where Φ_1 and Φ_2 are approximately zero. The results of the RF, shown on the right, provide what is expected; as we move away from the origin the RF quickly predicts that a series will be seasonal across the entire domain. On the other hand, the QS test correctly identifies seasonal series in less than half of the domain. Recall that the QS statistic is by far the best performing of the traditional statistics meaning that these number of series in which the null is correctly rejected goes down for the alternatives.

[Figure 9 about here.]

Finally, we would like to point out that our choice of cutoff for the classification of a series as seasonal, 0.50, is relatively arbitrary and may not be the most efficient. To illustrate how sensitive to this cutoff our results are we plot the Receiver Operating Characteristic (ROC) curve in Figure 10. Here we plot the True Positive Rate (TPR), representing the series which are correctly classified as seasonal, against the False Positive Rate (FPR), or those which are incorrectly classified as seasonal. If this decision were random then it would lie upon the 45° angle and any portion of the curve above this line indicates a cutoff which is better than random. The results of this exercise show us that the choice of 0.50, while initially quite arbitrary, is nearly optimal.

[Figure 10 about here.]

3.4 Recapping the Results from Simulations

Through out this section we have uncovered several new facts about existing tests. Specifically, the most common tests used by statistical agencies such as the U.S. Census Bureau,

U.S. Bureau of Economic Analysis, United Kingdom Office of National Statistics, etc. suffer from poor statistical properties, including insufficient size under the null. Using simulations we show that the length of a time series needed to approach nominal critical values for the QS, D8F, FM, and M7 statistics exceeds that which is feasible for a central statistical agency, or for that matter any practical time series. To combat this we provide new critical values based on a Monte Carlo Study which maps the null distribution under a variety of series lengths and provide those in Appendix B.

In addition to these new facts regarding existing tests, we show how a RF can be used to exploit the variation in classification by these tests to more accurately predict if a series is seasonal. We trained this random forest using $5,000,000$ series generated from a SARIMA structure with varying dimensionality and parameter values. The gains in accuracy are substantial across the board regardless of the seasonality structure (moving average vs. autoregressive), or size of the seasonal effect. We show how existing tests, as compared to the RF, not only fail to perform well over the domain of parameters governing the size of the seasonal effect, but also that they fail to extract maximal information from the inputs upon which they are derived.

4 Empirical Examples

4.1 U.S. Retail Sales: Shoe Stores

To contextualize the use of RF in determining the seasonality classification for an individual series we turn to McElroy (2008) as a guideline, and look at U.S. Retail Sales of Shoe Stores, both seasonally and not seasonally adjusted. Recall that the purpose of this paper is not to adjust any single series, nor offer guidance on when an adjustment is adequate, but rather to identify if a series should be classified as seasonal, or not, for the purposes of identifying candidates for adjustment. Our choice of U.S. Retail Sales of Shoe Stores is important because in the aforementioned paper the author noted that the seasonally adjusted series is residually seasonal based on the concept of "negative seasonality", that is seasonality which creates troughs at the seasonal frequencies in the ACF. Recall that our objective is to reduce residual seasonality through the testing channel by reducing the Type II error rate.

In this instance we have a series which is known to be seasonal, and post adjustment is still considered to be seasonal. Though the author states that the seasonally adjusted series is still seasonal, this is done through a visual inspection of the ACF, spectrum, and other data elements which are not reproducible to scale. Our main focus then is to determine if the RF agrees with the expert opinion regarding the seasonal classification of this series both pre and post adjustment.

[Figure 11 about here.]

Figure 11 plots, on the left, raw data for monthly sales as reported by the U.S. Census Bureau (black), and seasonally adjusted sales (red) from 1984 to 1998.[12] In the middle we have plotted the autocorrelation function for the first-difference log of the not seasonally adjusted data ($\Delta ln(Shoes)$), and to the right the first-difference log of the seasonally adjusted data ($\Delta ln(Shoes^*)$). The positive autocorrelation at lags 12, 24, and 36 in Figure 11b show a clear twelve month seasonal pattern, which is supported by the appropriate test statistics found in Table 4. All seven tests are in agreement, the autocorrelation function of the first-difference log, and a visual inspection of the series confirm that the not seasonally adjusted data are a clear candidate for seasonal adjustment. As expected, the RF agrees with the overwhelming evidence indicating that the null hypothesis of no seasonality should be rejected. Figure 12d outlines the relative contribution each of the included features had in this decision with the final prognosis representing the probability that the null should be rejected.

[Table 4 about here.]

Recall that, in the aforementioned paper, it was of the expert's opinion that the seasonally adjusted series exhibited residual seasonality based on the concept of negative seasonality. In Figure 11c we plot the autocorrelation function of the first-difference log for the seasonally adjusted series. While the clear spikes at 12, 24, and 36 have been removed there is now a clear downward spike in the ACF at lags 24 and 36 with a lesser drop at

[12] It is important to note that we did not seasonally adjust this data ourselves but rather are relying on the published seasonally adjusted numbers put out by the U.S. Census Bureau. In this way our example differs from McElroy (2008).

lag 12. Returning to Table 4 we see that there is agreement among the tests that there is insufficient evidence to reject the null of no seasonality. Recall however that over the previous sections we have shown that these tests exhibit poor coverage over the parameter space domain in a SARIMA framework, and what coverage they do have is plagued by poor size and power. The RF however uses the variation in the test statistics, combined with the other features, and predicts that this series, the published seasonally adjusted figures, is in all likelihood seasonal. In Figure 12e we show the contribution of top features to this decision. We caution the reader in generalizing the breakdown of contributions found in Figures 12d and 12e; while the inferences regarding the feature contribution to the decision are valid for this particular example, they should not be generalized across other series as those series may lay elsewhere in the sample space.[13]

[Figure 12 about here.]

4.2 U.S. Import/Export Series

One important facet of the RF predictions is the scalability of the testing procedure. For example, a single analyst producing a single aggregated data series may be responsible for synthesizing information from hundreds, if not thousands, of underlying component series. Obtaining predictions for many series from the RF requires little additional computational or time burden when compared to a single series. Here we illustrate this by examining import/export data from the Census both from the standpoint of classification by traditional tests and that of the RF. This data consists of 255 monthly series outlining imports to the United States and 254 monthly series of exports.[14] The series vary in length from a minimum of eight years to a maximum of thirty-nine years.

[Figure 13 about here.]

In Figure 13 we two dimensions of variation found using the aforementioned Import/Export data. For example, in Figure 13a, each point represents a single series from this data with

[13] We calculated these feature contributions using the "breakDown" package in R version 3.6.1 (2019-07-05) – "Action of the Toes". See Staniak and Biecek (2018) for more information.

[14] We obtained this information from the U.S. Census Bureau Foreign Trade, https://www.census.gov/foreign-trade/index.html. The raw data that formed the basis for this analysis can be obtained through https://www.census.gov/foreign-trade/balance/country.xlsx (accessed 10/01/2019).

coordinate pair ($x = \text{KW}, y = \text{QS}$). The dashed, black lines represent the appropriate nominal critical values for the statistics in question while those in gray represent the empirical values presented herein. Similar to Figures 1a to 1f from the simulated environment, a series which appears in the bottom left quadrant represents one in which both tests agree on a failure to reject the null hypothesis. Those series in the upper right quadrant represent series for which both tests fail to reject the null hypothesis. The remaining quadrants represent those series in which there is disagreement regarding the seasonal disposition of the series. Export series in each figure are those which appear as circles (red) and imports are represented by triangles (blue). It is important to note that there does not appear to be a pattern of disagreement, *e.g.* the series in which there is disagreement is not dominated by imports alone.

[Table 5 about here.]

In Table 5 we outline the fraction of series for which each test rejects the null hypothesis of no seasonality. This table is constructed using the empirical critical values for each test. Note that for the vast majority of the tests roughly half of the series in question reject the null hypothesis though the M7, D8F, and FM tests stand out as outliers in the respect. The RF predicts that nearly every one of the series, both imports and exports, is seasonal and should be considered a candidate for adjustment. A result that closely aligns with the FM test. Again, it is important to note that we are not commenting on whether these series have been, or will be, seasonally adjusted. Rather, our focus is on identifying series which exhibit seasonal characteristics in an effort to define the scope of what might need to be adjusted in an effort to minimize the possibility of residual seasonality should these series be aggregated. Given the evidence presented in Section 3 regarding the accuracy of tests it would appear that many of these tests fail to recognize the seasonal nature of these series either because the seasonal component is relatively weak, or – if framed in SARIMA terms – the parameter governing the seasonal component is negative.

5 Conclusion

Residual seasonality has recently proven to be a thorn in the side of statistical agencies, and spawned criticism for a number of official series produced by those agencies (see McElroy (2008); Lunsford (2017); Cowan et al. (2018); Wright (2018) for examples). In this paper we show that inconsistencies in the definition of seasonality have led to tests which have poor statistical properties in the face of a basic seasonal autoregressive integrated moving average (SARIMA) data generating process. In addition to their own idiosyncratic weaknesses (*e.g.* strict reliance upon stationarity, violation of distributional assumptions, etc.) the tests produce poor accuracy over the domain of parameters in both single and multidimensional SARIMA structures.

Since a failure to reject the null hypothesis of no seasonality when the null is false can lead to residual seasonality in aggregates, any increases in accuracy from the testing channel will reduce the likelihood of finding residual seasonality. To increase this accuracy we map the hypothesis space using a Random Forest where our features include the test statistics themselves and other characteristics of the time series. We show that the predictions from this random forest are better than any single test at identifying series which are seasonal across a number of different SARIMA specifications. In our simulated environment we show that the trained RF model is thirty-one percentage points more accurate than the next best alternative. Depending on the stratification, this improvement ranges from eleven to eighty-eight percentage points with larger gains being realized when the SARIMA parameters are negative. In all cases the RF is strictly more accurate than the alternative options. Moreover, since the RF is scalable in its predictions, predicting many is not significantly more resource intensive than predicting a single series. Thus, this tool is useful in the resource constrained environment many data agencies face.

We then apply the trained RF to two case studies, the first dealing with a single series, monthly retail trade of shoe stores, shown to have residual seasonality, and the second examining foreign trade series (McElroy, 2008). In both cases the data are publicly available, and provided by the U.S. Census Bureau. When examining the former, the primary goal was to determine if the published seasonally adjusted series should be considered a candidate for seasonal adjustment thus exhibiting residual seasonality. In this case, the

trained RF agreed with the aforementioned paper – the seasonally adjusted series is likely seasonal – with the primary driver of this decision being the autocorrelation function. In the second application we show that the trained RF indicates that nearly all of the series included are seasonal, while the alternatives range from 12.3% to 94.9%. The RF identifies a total of six series out of the 505 examined as seasonal for which no other test agrees.

Recall that the focus of this paper is not to determine if a series has been or ever will be adjusted, and that a great deal of information beyond the test statistics is used in this determination. The RF is used primarily as an effort to synthesize the information available through multiple test statistics with disparate assumptions and conclusions regarding the same series, and should be taken as but a piece of evidence regarding the seasonal disposition of a series rather than the definitive authority on the matter. Finally, we make no claims about the optimal method for seasonal adjustment and recognize that some filters may leave fragmented seasonality in the adjusted series which will often go undetected. This is an avenue for further study with respect to the use of ML methods in the current seasonal adjustment paradigm.

A Tests for Seasonality

In this appendix we would like to take a moment to outline the eight tests for seasonality that are of primary interest. The first four are statistics used in the X13 program employed by many statistical agencies, and the remaining five are other tests found throughout the literature. It is important to note that this list is not exhaustive and our comments about these tests are relevant only so far as the tests themselves pertain to our chosen domain of interest.

The QS Statistic: The QS statistic is formed by examining the autocorrelation function at the appropriate seasonal lags (Maravall, 2012). As mentioned earlier, let S be the seasonal frequency, *e.g.* $f = 4$ for quarterly data with a yearly seasonal pattern, and T be the length of the series. Furthermore, let $\gamma(g) = \mathbb{E}[y_{t+g}y_t] - \mathbb{E}^2[y_t]$ represent the autocovariance and $\rho(g) = \gamma(g)/\gamma(0)$ the autocorrelation of Y. The QS statistic can be written as,

$$\text{QS} = T(T+2)\left(\frac{\hat{\rho}^2(f)}{T-f} + \frac{[\max\{0, \hat{\rho}^2(2f)\}]^2}{T-2f}\right). \tag{10}$$

Note that the null hypothesis is $H_0 : \gamma(m) \leq 0$ for $m \in \{f, 2f\}$. This naturally leads to a restriction on the domain of $\gamma(m)$ such that if $\hat{\gamma}(m) \leq 0$, then QS $= 0$. This is a notable restriction since, if there is a trough in the ACF at the seasonal frequency then, by definition the QS statistic is zero and the null hypothesis of no seasonality fails to be rejected. By construction then the QS cannot see the aforementioned negative seasonality. It is important to note that we know very little about the distribution of the QS statistic under the null and that we generally use critical values from a χ^2 distribution with 2 degrees of freedom as an approximation Maravall (2011). We will present evidence in Section 3.2 that this is in fact a poor approximation which leads to an undersized test.

The F-stable Statistic: Let k be the number of periods in a yearly cycle and n_j be the number of observations in the j^{th} period of k. The F-stable test is a one way analysis of variance test written as:

$$\text{FS} = \frac{\sum_{j=1}^{k} n_j(\bar{y}_{.j} - \bar{y}_{..})^2/(k-1)}{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2/(n-k)}, \tag{11}$$

where the null hypothesis is $H_0 : \mathbb{E}[y_1] = \ldots = \mathbb{E}[y_k]$ and alternative $H_0 : \mathbb{E}[y_p] \neq \mathbb{E}[y_q]$ for at least a single pair (p, q). Theoretically this test statistic should follow a an F-distribution with $(k-1)$ and $(n-k)$ degrees of freedom. However, in practice the F-stable statistic, FS (also known as the D8F), has a nominal critical value of 7. This is because "...several of the basic assumptions in the F test are probably violated..." (Dagum, 2005; Ladiray and Quenneville, 2012). A failure to reject the null using this test does not necessarily imply the series is not seasonal. Rather, the null is stating that there is no difference in the means of each period over the time span being reviewed. It could be that the seasonality is affecting the variance in these periods over years, not the mean. It is for this reason that many practitioners will tell you the D8F should be a small part of the evidentiary profile when determining if a series is seasonal.

There is one additional point that must be considered, the D8F is computed based on a transformed series, not the original. Why does this matter? Consider the following example, suppose we are interested in determining if a series is stationary by using an augmented Dickey-Fuller test. We transform the original series by taking the difference and calculate the test statistic which leads us to reject the null that the differenced series is stationary. This of course implies that the original series was also not stationary since taking the difference is supposed to move us closer to an I(0) process. Now let's consider the alternative, suppose that we fail to reject the null, that the differenced series is stationary. Unlike the case where we rejected the null, this is not backward compatible with our original series. In levels the original series could be stationary or not, and our failure to reject the null on the transformed series provides us no information. In short, our failure to reject the null of equal means when evaluating the D8F tells us little to nothing about whether the original series is seasonal or not, rather it is only by rejecting the null that we can safely claim anything about the original series.

Of course, testing for stationarity is not testing for seasonality so how does this apply to the topic du jour? The transformation of the series for the D8F includes a decomposition and testing on what is thought to be the seasonal and irregular pattern. While this is trivial in cases where the standard trend, seasonal, irregular decomposition holds it is not clear that this is a benign transformation. The decision of whether the time series

is additive or multiplicative, and thus determining the type of decomposition, is far from inconsequential and commonly is decided through information criterion such as AIC, AICc, BIC, etc. While it is beyond the scope of the current work to prove that this transformation is inconsequential it is important to note that the D8F and M7 statistic, the latter of which is a non-linear function of the former, are the worst performing of the tests evaluated herein. While it may not be causal, the fact that these two are the only tests which rely upon a transformation of the original series should at least put some skepticism in their use.

The F-moving Statistic: In a complementary fashion to the D8F, a second F-test can be used to evaluate moving seasonality (Higginson, 1975). This is a two-way analysis of variance test looking at both the month (or quarter if applicable) and the year. Again, let k be the number of periods in a yearly cycle and let N represent the number of complete years in the data. Following notation by Ladiray and Quenneville (2012), the F-moving statistic is based off of the model,

$$|\text{SI}_{ij} - \bar{y}| = y_{ij} = b_i + m_j + e_{ij}, \tag{12}$$

where b_i is the object of interest and refers to the annual effect. Note again that this test statistic is performed on the transformed time series, much like the D8F. The F-moving statistic can then be written as:

$$\text{FM} = \frac{k \sum_{i=1}^{N} (\sum_{j=1}^{k} y_{ij}/k - \bar{y}_{..})^2/(N-1)}{\sum_{i=1}^{j} \sum_{j=1}^{k} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2}, \tag{13}$$

with the null hypothesis $H_0 = b_1 = \ldots = b_N$. This statistic is distributed as an F-distribution with $(k-1)$ and $(k-1)(N-1)$ degrees of freedom under the null.

The M7 Statistic: The M7 statistic, outlined by Equation 14, is a non-linear combination of the F-stable and F-moving statistics, Equations 11 and 13 respectively (Lothian and Morry, 1978).

$$\text{M7} = \sqrt{\frac{1}{2}\left(\frac{7}{\text{FS}} + \frac{3\text{FM}}{\text{FS}}\right)}. \tag{14}$$

The M7 statistic uses a base critical value of 1 and is a left-tailed test. This means the null hypothesis of no seasonality is rejected if M7 < 1 though originally it was designed with a decision tree mind and not for use in isolation (see Ladiray and Quenneville (2012) pp. 138 for the diagram).[15]

The Model Based F-test: The Model Based F-test, hereafter FMB, is a regression based approach which combined an ARIMA model with seasonal dummies to estimate the presence of a seasonal effect. The base of this statistic is an χ^2-test with the null hypothesis that the coefficients on seasonal dummies are all equal to zero. In finite samples, which is the case with all tests for seasonality, the variance can greatly influence the test statistic and as a result a correction is used. This corrected test can be written as,

$$\text{SD} = \frac{\hat{\beta}'\Sigma_{\hat{\beta}}\hat{\beta}}{f-1}\left(\frac{T-d-f}{T-d}\right), \tag{15}$$

which follows an F-distribution with $(f-1)$ and $(T-f-d)$ degrees of freedom. This statistic is shown to have favorable properties in terms of size and power when compared to some alternatives Lytras et al. (2007).

The Friedman Test: The Friedman test (FR) is a non-parametric testing method which evaluates if samples are drawn from a population with equal medians. This test uses ranking of the observations rather than relying on distributional assumptions. Following notation from Webel and Ollech (2018), let r_{ij} be the rank of i^{th} observation in the j^{th} year and $\mu_i = \mathbb{E}(r_{ij})$. The test statistic can then be written as,

$$\text{FR} = \frac{\tau-1}{\tau}\sum_{\tau}^{i=1}\frac{n[\bar{r}_i-(\tau+1)/2]^2}{(\tau^2-1)/12}, \tag{16}$$

where $\bar{r}_i = n^{-1}\sum_j r_{ij}$, n represents the number of observations in each period i, indexed by $i \in \{1,\ldots,\tau\}$. The null hypothesis is $H_0 : \mu_1 = \ldots = \mu_\tau$ and follows a χ^2 distribution with $\tau - 1$ degrees of freedom under the null.

[15] We would like to point out that this varies somewhat. For example, the Office of National Statistics in the United Kingdom uses a critical value of 1.250 and 1.050 for monthly and quarterly series respectively while the IMF adheres to the flat critical value of 1.

The Kruskal-Wallis Test: The Kruskal-Wallis test (KW) is similar to the Friedman test, in fact it shares the same null hypothesis $H_0 : \mu_1, \ldots, \mu_\tau$, but calculates the test statistic as,

$$\text{KW} = \frac{T-1}{T} \sum_{i=1}^{\tau} \frac{n_i[\bar{r}_i - (T+1)/2]^2}{(T^2-1)/12} \tag{17}$$

where ranks are produced over the entire sample, T, rather than each period year. Like the Friedman test, the Kurskall-Wallis test follows a χ^2 distribution with $\tau - 1$ degrees of freedom under the null.

The Welch Test: Again, borrowing notation from Webel and Ollech (2017) let $\mu_i = \mathbb{E}(z_{ij})$ and $\bar{z}_i = n_i^{-1} \sum_{j=1}^{n_i} z_{ij}$. The Welch test (WE) (Welch, 1951) can be written as,

$$\text{WE} = \frac{(\tau-1)^{-1} \sum_{i=1}^{\tau} w_i(\bar{z}_i - w^{-1} \sum_{i=1}^{\tau} w_i \bar{z}_i)^2}{1 + 2(\tau+2)(\tau^2-1)^{-1} \sum_{i=1}^{\tau} (n_i-1)^{-1}(1-w^{-1}w_i)^2}, \tag{18}$$

where $w_i = n_i/(n_i-1)^{-1} \sum_{j=1}^{n_i}(z_{ij} - \bar{z}_i)^2$ and $w = \sum_{i=1}^{\tau} w_i$. WE follows an F distribution under the null, $H_0 : \mu_1, \ldots, \mu_2$, with $(\tau - 1)$ and $\left(3(\tau^2-1)^{-1} \sum_{i=1}^{\tau} (n_i-1)^{-1}(1-\frac{w_i}{w})^2\right)^{-1}$ degrees of freedom.[16]

[16] A special thanks to the authors of Webel and Ollech (2017) and Webel and Ollech (2018) for making many of these seasonal tests available through replication code of their work in the "seastests" R package.

B Critical Values at Common Values of α

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

[Table 10 about here.]

[Table 11 about here.]

C Nominal Accuracy for RF

[Table 12 about here.]

References

1. Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32. 3

2. Cowan, B., S. Smith, and S. Thompson (2018). Seasonal adjustment in the national income and product accounts. *Survey of Current Business 98*. 2, 5, 21

3. Dagum, E. (2005). *Essays Collection of Estela Bee Dagum in Statistical Sciences*, Chapter New Developments in the X-11-ARIMA, pp. 977–1002. Naples: Liguori Editore. 11, 24

4. Gilbert, C., N. J. Morin, A. D. Paciorek, C. R. Sahm, et al. (2015). Residual seasonality in GDP. Technical report, Board of Governors of the Federal Reserve System (US). 2

5. Gómez, V. and A. Maravall (2001). Seasonal adjustment and signal extraction in economic time series. *A Course in Time Series Analysis*, 202–247. 3

6. Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge university press. 4

7. Higginson, J. (1975). *An F Test for the presence of moving seasonality when using census method II-X-11 variant*. Statistics Canada. 25

8. Hillmer, S. C. and G. C. Tiao (1982). An arima-model-based approach to seasonal adjustment. *Journal of the American Statistical Association 77*(377), 63–70. 4

9. Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press. 3

10. Ladiray, D. and B. Quenneville (2012). *Seasonal adjustment with the X-11 method*, Volume 158. Springer Science & Business Media. 11, 24, 25, 26

11. Lothian, J. R. and M. Morry (1978). *A test for the presence of identifiable seasonality when using the X-11-ARIMA program*. Seasonal Adjustment and Time Series Analysis Staff, Statistics Canada. 25

12. Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association 58*(304), 993–1010. 4

13. Lunsford, K. G. (2017). Lingering residual seasonality in GDP growth. *Economic Commentary* (2017-06). 2, 21

14. Lytras, D. P., R. M. Feldpausch, and W. R. Bell (2007). Determining seasonality: a comparison of diagnostics from x-12-arima. *US Census Bureau*. 4, 5, 26

15. Maravall, A. (2011). Seasonality Tests and Automatic Model Identification in TRAMO-SEATS. *Mimeo, Bank of Spain*. 12, 23

16. Maravall, A. (2012). Update of Seasonality Tests and Automatic Model Identification in TRAMO-SEATS. *Bank of Spain*. 4, 23

17. McElroy, T. (2008). A modified model-based seasonal adjustment that reduces spectral troughs and negative seasonal correlation. *Research Report Series 2008-12*. 4, 17, 18, 21

18. Moulton, B. R. and B. D. Cowan (2016). Residual seasonality in GDP and GDI: Findings and next steps. *Survey of Current Business 96*(7), 1–6. 2

19. Nerlove, M. (1964). Spectral analysis of seasonal adjustment procedures. *Econometrica: Journal of the Econometric Society*, 241–286. 4

20. Staniak, M. and P. Biecek (2018). Explanations of model predictions with live and breakdown packages. *arXiv preprint arXiv:1804.01955*. 19

21. Wang, X., K. Smith, and R. Hyndman (2006). Characteristic-based clustering for time series data. *Data mining and knowledge Discovery 13*(3), 335–364. 3

22. Wang, X., K. Smith-Miles, and R. Hyndman (2009). Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. *Neurocomputing 72*(10-12), 2581–2594. 3

23. Webel, K. and D. Ollech (2017, September). An overall seasonality test based on recursive feature elimination in conditional random forests. In O. Valenzuela, F. Rojas, H. Pomares, and I. Rojas (Eds.), *Proceedings of the 61st ISIS World Statistics Congress*, pp. 21–31. Godel Impresiones Digitales S.L. 5, 27

24. Webel, K. and D. Ollech (2018). An overall seasonality test based on recursive feature elimination in conditional random forests. *Proceedings of the 5th International Conference on Time Series and Forecasting*, 20–31. 5, 26, 27

25. Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika 38*(3-4), 330–336. 27

26. Wright, J. H. (2018). Seasonal Adjustment of NIPA Data. Technical report, National Bureau of Economic Research. 2, 21

Table 1: Critical Values and Size: Frequency = 12 Years = 100

	Current CV	Current Size	Suggested CV	New Size
QS	5.991	0.016	3.668	0.050
D8F	7.000	0.000	2.471	0.055
FM	1.792	0.824	8.956	0.049
M7	1.000	0.000	1.712	0.060
FMB	1.797	0.059	1.791	0.060
WE	1.809	0.066	1.802	0.067
KW	19.675	0.048	19.571	0.050
FR	19.675	0.048	19.617	0.049

Table 2: Critical Values and Size: Frequency = 4 Years = 300

	Current CV	Current Size	Suggested CV	New Size
QS	5.991	0.023	3.602	0.059
D8F	7.000	0.002	3.638	0.049
FM	2.615	0.124	3.578	0.050
M7	1.000	0.008	1.259	0.051
FMB	2.612	0.053	2.660	0.050
WE	2.618	0.052	2.642	0.051
KW	7.815	0.052	7.877	0.047
FR	7.815	0.051	7.844	0.048

Table 3: Empirical Accuracy Table: Test Data

	RF	QS	M7	D8F	FM	FMB	WE	KW	FR	N. Series
Accuracy	0.95	0.64	0.52	0.54	0.55	0.53	0.53	0.54	0.53	5,000,000
$\text{ACC}\|P=1$	0.88	0.30	0.14	0.10	0.03	0.19	0.19	0.13	0.13	416,529
$\Phi_1>0$	0.88	0.59	0.26	0.19	0.03	0.34	0.34	0.24	0.23	208,181
$\Phi_1<0$	0.88	0.00	0.03	0.01	0.03	0.04	0.05	0.01	0.02	208,348
$\text{ACC}\|P=2$	0.95	0.34	0.14	0.15	0.19	0.22	0.22	0.18	0.17	416,795
$\Phi_1,\Phi_2>0$	0.96	0.81	0.40	0.50	0.22	0.63	0.62	0.55	0.49	103,721
$\Phi_2,\Phi_2<0$	0.95	0.00	0.01	0.00	0.15	0.01	0.02	0.00	0.01	103,933
$\sum_{p=1}^{2}\Phi_p>0$	0.96	0.61	0.26	0.30	0.19	0.42	0.42	0.35	0.32	208,371
$\sum_{p=1}^{2}\Phi_p<0$	0.95	0.08	0.01	0.01	0.18	0.02	0.03	0.01	0.01	208,424
$\text{ACC}\|P=3$	0.98	0.37	0.13	0.19	0.40	0.24	0.24	0.21	0.20	417,260
$\Phi_1,\Phi_2,\Phi_3>0$	0.99	0.91	0.61	0.78	0.46	0.85	0.84	0.80	0.74	52,098
$\Phi_2,\Phi_2,\Phi_3<0$	0.98	0.10	0.00	0.00	0.35	0.01	0.01	0.00	0.00	52,023
$\sum_{p=1}^{3}\Phi_p>0$	0.98	0.55	0.26	0.37	0.38	0.46	0.46	0.42	0.39	208,401
$\sum_{p=1}^{3}\Phi_p<0$	0.98	0.18	0.00	0.00	0.41	0.01	0.01	0.00	0.01	208,859
$\text{ACC}\|P>0$	0.94	0.34	0.14	0.15	0.21	0.22	0.22	0.17	0.16	1,250,584
$\text{ACC}\|Q=1$	0.87	0.28	0.13	0.06	0.00	0.15	0.15	0.08	0.09	416,757
$\Theta_1>0$	0.87	0.56	0.23	0.12	0.00	0.26	0.27	0.16	0.16	208,479
$\Theta_1<0$	0.87	0.00	0.03	0.01	0.00	0.04	0.04	0.01	0.02	208,278
$\text{ACC}\|Q=2$	0.93	0.30	0.15	0.08	0.00	0.16	0.17	0.10	0.10	416,714
$\Theta_1,\Theta_2>0$	0.93	0.74	0.38	0.23	0.00	0.40	0.40	0.28	0.27	104,512
$\Theta_2,\Theta_2<0$	0.92	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.01	104,652
$\sum_{p=1}^{2}\Theta_p>0$	0.93	0.56	0.28	0.16	0.00	0.30	0.31	0.20	0.20	208,435
$\sum_{p=1}^{2}\Theta_p<0$	0.93	0.05	0.02	0.01	0.00	0.03	0.03	0.01	0.01	208,279
$\text{ACC}\|Q=3$	0.92	0.30	0.15	0.09	0.00	0.17	0.17	0.11	0.11	416,462
$\Theta_1,\Theta_2,\Theta_3>0$	0.92	0.75	0.48	0.34	0.00	0.50	0.50	0.39	0.36	52,076
$\Theta_2,\Theta_2,\Theta_3<0$	0.91	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	52,426
$\sum_{p=1}^{3}\Theta_p>0$	0.91	0.48	0.29	0.18	0.00	0.32	0.32	0.22	0.21	207,990
$\sum_{p=1}^{3}\Theta_p<0$	0.92	0.11	0.01	0.00	0.00	0.02	0.02	0.01	0.01	208,472
$\text{ACC}\|Q>0$	0.90	0.29	0.14	0.08	0.00	0.16	0.17	0.10	0.10	1,249,933
$\text{ACC}\|P,Q=0$	0.98	0.95	0.91	0.96	1.00	0.88	0.87	0.95	0.93	2,499,483

Here we calculate the accuracy based on the empirical critical values for each test and compare them to the out-of-sample prediction accuracy of the Random Forest. The breakdowns are to show accuracy when the null is false (equivalent to power) under specific seasonal dimensionality conditions. Note that there is no constraints on the parameter space of Φ or Θ beyond those which are standard.

Table 4: Retail Sales: Shoe Stores Seasonality Test Results

	QS	M7	D8F	FM	FMB	WE	KW	FR	Obs.
Shoes - NSA	294.188	0.150	200.255	0.659	170.860	225.663	145.312	135.253	179
Shoes - SA	0.000	4.953	0.234	1.491	0.093	0.091	1.361	0.462	179

Table 5: Percentage of Import/Export Series Flagged as Seasonal

Type	RF	QS	M7	D8F	FM	FMB	WE	KW	FR	N
Exports	1.000	0.424	0.267	0.329	0.078	0.443	0.435	0.537	0.490	255
Imports	0.996	0.516	0.335	0.433	0.114	0.524	0.512	0.587	0.504	254

Table 6: Empirical Critical Values: Quarterly Data, $\alpha = 10\%$

Time (Years)	QS	D8F	FM	M7	FMB	WE	KW	FR
10	7.823	3.035	2.437	1.322	2.441	2.425	4.953	6.067
20	7.083	2.895	2.266	1.369	2.204	2.231	5.901	6.158
30	2.811	2.822	2.285	1.402	2.145	2.146	6.053	6.145
40	2.241	2.870	2.318	1.403	2.158	2.147	6.227	6.231
50	2.177	2.849	2.371	1.405	2.148	2.141	6.219	6.233
60	2.173	2.843	2.441	1.414	2.119	2.125	6.204	6.214
70	2.205	2.858	2.461	1.412	2.114	2.131	6.250	6.200
80	2.214	2.821	2.526	1.423	2.081	2.093	6.162	6.144
90	2.237	2.829	2.538	1.425	2.110	2.121	6.243	6.222
100	2.225	2.837	2.581	1.422	2.093	2.099	6.225	6.176
110	2.288	2.827	2.593	1.425	2.102	2.104	6.185	6.171
120	2.280	2.820	2.628	1.430	2.083	2.085	6.194	6.156
130	2.256	2.825	2.658	1.429	2.092	2.099	6.218	6.163
140	2.282	2.834	2.654	1.424	2.089	2.092	6.261	6.230
150	2.267	2.839	2.685	1.423	2.107	2.106	6.204	6.221
160	2.275	2.843	2.716	1.424	2.113	2.108	6.267	6.260
170	2.266	2.848	2.737	1.424	2.094	2.103	6.215	6.224
180	2.298	2.835	2.753	1.426	2.105	2.112	6.241	6.231
190	2.302	2.820	2.749	1.428	2.087	2.092	6.207	6.225
200	2.269	2.784	2.758	1.438	2.072	2.073	6.187	6.238
210	2.268	2.810	2.779	1.437	2.088	2.095	6.224	6.262
220	2.276	2.812	2.794	1.436	2.101	2.101	6.285	6.260
230	2.269	2.829	2.785	1.428	2.098	2.103	6.235	6.228
240	2.301	2.830	2.799	1.430	2.098	2.095	6.237	6.219
250	2.296	2.847	2.823	1.426	2.095	2.093	6.262	6.181
260	2.298	2.817	2.823	1.428	2.096	2.097	6.240	6.211
270	2.318	2.815	2.820	1.431	2.095	2.093	6.246	6.239
280	2.334	2.826	2.824	1.426	2.109	2.114	6.295	6.239
290	2.304	2.825	2.825	1.424	2.099	2.100	6.291	6.252
300	2.310	2.838	2.834	1.426	2.099	2.100	6.280	6.267

Note: For each time set we generated $20,000$ series where the null hypothesis is true, there is no seasonality present. We then calculated the test statistic for each of the $20,000$ series and calculated the value at the 90^{th} quantile to provide an empirical critical value under those finite sample conditions. This table allows us to review the convergence of the critical values which varies substantially between tests.

Table 7: Empirical Critical Values: Monthly Data, $\alpha = 10\%$

Time (Years)	QS	D8F	FM	M7	FMB	WE	KW	FR
5	2.018	2.245	4.408	1.613	1.787	2.293	15.832	16.500
10	2.236	2.079	4.566	1.783	1.665	1.791	16.882	16.915
15	2.220	2.025	4.887	1.833	1.625	1.696	17.047	16.989
20	2.248	2.029	5.249	1.866	1.616	1.659	17.122	17.040
25	2.233	2.035	5.530	1.879	1.602	1.640	17.126	17.174
30	2.277	2.028	5.753	1.892	1.601	1.628	17.175	17.260
35	2.300	2.023	5.914	1.903	1.594	1.613	17.210	17.231
40	2.288	2.011	6.053	1.909	1.585	1.607	17.199	17.249
45	2.299	2.009	6.186	1.911	1.591	1.610	17.227	17.241
50	2.346	2.013	6.296	1.913	1.595	1.604	17.243	17.220
55	2.318	2.011	6.391	1.905	1.591	1.595	17.227	17.237
60	2.338	2.007	6.491	1.912	1.584	1.590	17.242	17.266
65	2.345	2.001	6.574	1.917	1.577	1.582	17.216	17.267
70	2.337	2.002	6.636	1.919	1.578	1.583	17.205	17.228
75	2.379	1.997	6.652	1.930	1.569	1.580	17.237	17.192
80	2.380	2.013	6.724	1.922	1.567	1.577	17.117	17.206
85	2.366	1.996	6.722	1.920	1.569	1.581	17.166	17.172
90	2.405	2.010	6.742	1.912	1.565	1.572	17.150	17.099
95	2.412	1.999	6.811	1.920	1.564	1.572	17.153	17.167
100	2.357	1.994	6.853	1.916	1.574	1.578	17.209	17.204

Note: For each time set we generated $20,000$ series where the null hypothesis is true, there is no seasonality present. We then calculated the test statistic for each of the $20,000$ series and calculated the value at the 90^{th} quantile to provide an empirical critical value under those finite sample conditions.

Table 8: Empirical Critical Values: Quarterly Data $\alpha = 5\%$

Time (Years)	QS	D8F	FM	M7	FMB	WE	KW	FR
10	11.205	3.893	3.083	1.158	3.219	3.156	6.420	7.533
20	18.449	3.712	2.850	1.205	2.801	2.831	7.453	7.737
30	5.472	3.551	2.880	1.239	2.681	2.704	7.585	7.634
40	3.707	3.602	2.947	1.235	2.700	2.707	7.728	7.708
50	3.458	3.603	3.045	1.246	2.667	2.706	7.722	7.751
60	3.397	3.583	3.109	1.254	2.646	2.653	7.759	7.759
70	3.428	3.598	3.132	1.258	2.665	2.670	7.727	7.661
80	3.485	3.544	3.199	1.264	2.608	2.625	7.714	7.724
90	3.496	3.617	3.215	1.258	2.634	2.637	7.755	7.800
100	3.537	3.579	3.272	1.262	2.630	2.634	7.777	7.788
110	3.568	3.545	3.322	1.269	2.613	2.613	7.776	7.668
120	3.559	3.580	3.334	1.267	2.624	2.621	7.750	7.760
130	3.563	3.602	3.368	1.261	2.610	2.615	7.735	7.716
140	3.582	3.581	3.380	1.263	2.634	2.629	7.799	7.722
150	3.504	3.649	3.423	1.256	2.622	2.635	7.862	7.752
160	3.522	3.604	3.467	1.259	2.629	2.641	7.817	7.732
170	3.571	3.597	3.474	1.256	2.626	2.627	7.818	7.758
180	3.606	3.605	3.519	1.259	2.617	2.623	7.813	7.780
190	3.615	3.562	3.525	1.263	2.608	2.605	7.769	7.775
200	3.569	3.598	3.513	1.271	2.608	2.614	7.854	7.830
210	3.606	3.636	3.526	1.263	2.623	2.634	7.865	7.817
220	3.570	3.613	3.529	1.259	2.637	2.642	7.899	7.795
230	3.537	3.604	3.546	1.262	2.639	2.638	7.853	7.810
240	3.556	3.572	3.564	1.265	2.628	2.621	7.781	7.715
250	3.561	3.553	3.603	1.269	2.638	2.633	7.797	7.733
260	3.616	3.569	3.614	1.271	2.626	2.628	7.829	7.740
270	3.604	3.611	3.604	1.263	2.639	2.643	7.827	7.764
280	3.634	3.618	3.599	1.255	2.643	2.642	7.896	7.804
290	3.597	3.651	3.570	1.253	2.677	2.670	7.870	7.767
300	3.602	3.638	3.578	1.259	2.660	2.642	7.877	7.844

Note: For each time set we generated $20,000$ series where the null hypothesis is true, there is no seasonality present. We then calculated the test statistic for each of the $20,000$ series and calculated the value at the 95^{th} quantile to provide an empirical critical value under those finite sample conditions. This table allows us to review the convergence of the critical values which varies substantially between tests (e.g. the QS doesn't stabalize until approximately 40 years of quarterly data is available while the Friedman converges relatively quickly at 20 years).

Table 9: Empirical Critical Values: Monthly Data $\alpha = 5\%$

Time (Years)	QS	D8F	FM	M7	FMB	WE	KW	FR
5	3.356	2.804	5.697	1.422	2.126	2.895	17.801	18.269
10	3.664	2.599	5.877	1.583	1.921	2.094	19.011	19.000
15	3.563	2.512	6.405	1.641	1.865	1.976	19.288	19.330
20	3.566	2.505	6.861	1.659	1.844	1.916	19.382	19.324
25	3.543	2.486	7.222	1.683	1.824	1.878	19.425	19.379
30	3.565	2.507	7.553	1.684	1.825	1.865	19.484	19.547
35	3.608	2.494	7.757	1.701	1.826	1.855	19.500	19.516
40	3.623	2.491	7.944	1.708	1.806	1.843	19.471	19.560
45	3.571	2.478	8.083	1.704	1.812	1.837	19.521	19.626
50	3.617	2.495	8.190	1.696	1.811	1.831	19.555	19.581
55	3.640	2.488	8.305	1.705	1.807	1.827	19.576	19.601
60	3.604	2.464	8.440	1.700	1.803	1.823	19.492	19.616
65	3.630	2.467	8.454	1.714	1.803	1.820	19.575	19.616
70	3.669	2.463	8.609	1.714	1.807	1.828	19.611	19.670
75	3.640	2.469	8.673	1.718	1.802	1.815	19.555	19.561
80	3.685	2.472	8.750	1.713	1.793	1.801	19.542	19.501
85	3.635	2.454	8.835	1.723	1.796	1.813	19.549	19.500
90	3.731	2.451	8.875	1.719	1.788	1.804	19.522	19.382
95	3.694	2.466	8.955	1.718	1.792	1.812	19.564	19.442
100	3.668	2.471	8.956	1.712	1.791	1.802	19.571	19.617

Note: For each time set we generated $20,000$ series where the null hypothesis is true, there is no seasonality present. We then calculated the test statistic for each of the $20,000$ series and calculated the value at the 95^{th} quantile to provide an empirical critical value under those finite sample conditions.

Table 10: Empirical Critical Values: Quarterly Data, $\alpha = 1\%$

Time (Years)	QS	D8F	FM	M7	FMB	WE	KW	FR
10	18.140	6.194	5.054	0.912	5.479	4.981	9.482	10.335
20	36.330	5.543	4.713	0.970	4.147	4.266	10.826	10.705
30	41.174	5.340	4.817	1.013	4.042	4.136	10.828	10.945
40	10.110	5.370	4.832	1.018	4.010	4.055	11.103	11.215
50	7.091	5.359	5.035	1.015	3.927	3.957	11.109	11.033
60	6.707	5.458	5.159	1.016	3.972	3.988	11.272	11.156
70	6.648	5.375	5.171	1.012	3.919	3.902	11.278	11.035
80	6.631	5.332	5.244	1.020	3.832	3.913	11.090	11.248
90	6.578	5.294	5.179	1.032	3.872	3.861	11.149	11.171
100	6.557	5.295	5.179	1.032	3.833	3.823	11.077	11.218
110	6.837	5.333	5.227	1.023	3.819	3.857	11.031	11.246
120	6.778	5.337	5.148	1.029	3.810	3.844	11.268	11.188
130	6.665	5.370	5.230	1.019	3.874	3.874	11.232	11.279
140	6.747	5.335	5.215	1.019	3.862	3.891	11.224	11.340
150	6.609	5.400	5.205	1.019	3.888	3.919	11.498	11.279
160	6.674	5.325	5.332	1.028	3.825	3.825	11.221	11.279
170	6.631	5.329	5.317	1.024	3.817	3.867	11.357	11.230
180	6.613	5.321	5.297	1.027	3.790	3.816	11.210	11.253
190	6.603	5.307	5.282	1.025	3.797	3.832	11.145	11.229
200	6.615	5.328	5.301	1.025	3.857	3.846	11.355	11.394
210	6.682	5.373	5.294	1.014	3.907	3.872	11.340	11.469
220	6.636	5.349	5.242	1.024	3.881	3.853	11.387	11.296
230	6.699	5.292	5.261	1.025	3.849	3.876	11.306	11.201
240	6.679	5.329	5.236	1.020	3.818	3.848	11.346	11.220
250	6.643	5.382	5.307	1.021	3.815	3.819	11.371	11.202
260	6.711	5.328	5.283	1.022	3.783	3.822	11.328	11.261
270	6.721	5.358	5.312	1.028	3.812	3.817	11.415	11.387
280	6.694	5.341	5.314	1.023	3.803	3.811	11.306	11.245
290	6.737	5.342	5.302	1.019	3.884	3.893	11.402	11.246
300	6.667	5.359	5.252	1.021	3.866	3.882	11.527	11.256

Note: For each time set we generated $20,000$ series where the null hypothesis is true, there is no seasonality present. We then calculated the test statistic for each of the $20,000$ series and calculated the value at the 99^{th} quantile to provide an empirical critical value under those finite sample conditions. This table allows us to review the convergence of the critical values which varies substantially between tests.

Table 11: Empirical Critical Values: Monthly Data, $\alpha = 1\%$

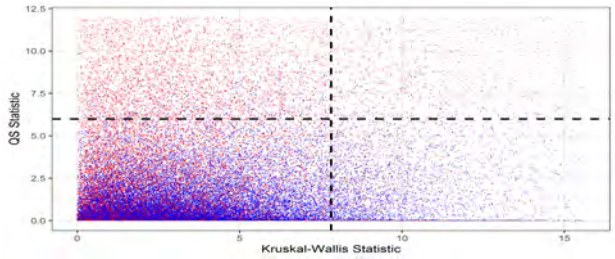
Time (Years)	QS	D8F	FM	M7	FMB	WE	KW	FR
5	6.622	4.401	9.349	1.124	2.892	4.796	21.722	21.808
10	7.591	3.814	9.587	1.283	2.488	2.854	23.244	23.496
15	6.986	3.620	10.533	1.348	2.384	2.582	23.907	23.945
20	6.792	3.667	10.930	1.371	2.347	2.495	24.041	23.745
25	6.939	3.649	11.230	1.374	2.310	2.393	24.175	24.013
30	6.906	3.548	11.575	1.394	2.301	2.404	24.265	24.279
35	7.039	3.548	11.727	1.397	2.282	2.374	24.270	24.299
40	7.089	3.595	12.032	1.396	2.285	2.347	24.211	24.282
45	6.924	3.580	12.271	1.400	2.271	2.335	24.089	24.381
50	6.889	3.564	12.380	1.388	2.270	2.316	24.137	24.551
55	6.932	3.593	12.484	1.397	2.257	2.299	24.231	24.316
60	6.808	3.553	12.435	1.399	2.279	2.317	24.590	24.450
65	6.944	3.500	12.679	1.398	2.282	2.329	24.760	24.416
70	6.829	3.536	12.768	1.413	2.293	2.299	24.565	24.650
75	6.808	3.522	12.758	1.405	2.273	2.300	24.594	24.626
80	6.801	3.494	12.994	1.419	2.278	2.283	24.243	24.595
85	6.887	3.511	13.013	1.406	2.262	2.272	24.317	24.522
90	6.911	3.504	13.084	1.415	2.248	2.275	24.303	24.544
95	6.935	3.490	13.117	1.414	2.249	2.270	24.316	24.272
100	6.835	3.599	13.207	1.396	2.244	2.263	24.360	24.382

Note: For each time set we generated $20,000$ series where the null hypothesis is true, there is no seasonality present. We then calculated the test statistic for each of the $20,000$ series and calculated the value at the 99^{th} quantile to provide an empirical critical value under those finite sample conditions.

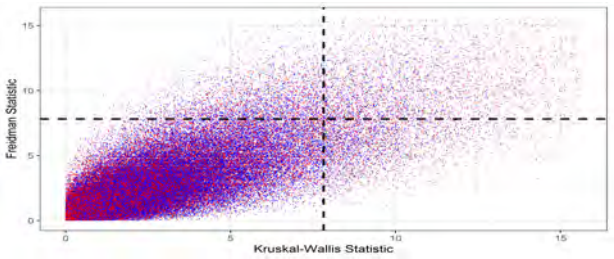
Table 12: Nominal Accuracy Table: Test Data

	RF	QS	M7	D8F	FM	FMB	WE	KW	FR	N. Series
Accuracy	0.95	0.63	0.51	0.51	0.60	0.54	0.54	0.54	0.53	5,000,000
$\text{ACC}\|P=1$	0.88	0.25	0.02	0.01	0.31	0.18	0.18	0.13	0.12	416,529
$\Phi_1>0$	0.88	0.49	0.04	0.03	0.29	0.33	0.32	0.24	0.22	208,181
$\Phi_1<0$	0.88	0.00	0.00	0.00	0.32	0.04	0.04	0.01	0.02	208,348
$\text{ACC}\|P=2$	0.95	0.30	0.04	0.05	0.62	0.22	0.21	0.18	0.16	416,795
$\Phi_1,\Phi_2>0$	0.96	0.75	0.13	0.20	0.67	0.62	0.60	0.55	0.49	103,721
$\Phi_2,\Phi_2<0$	0.95	0.00	0.00	0.00	0.59	0.01	0.01	0.00	0.01	103,933
$\sum_{p=1}^{2}\Phi_p>0$	0.96	0.55	0.07	0.11	0.62	0.41	0.40	0.35	0.32	208,371
$\sum_{p=1}^{2}\Phi_p<0$	0.95	0.05	0.00	0.00	0.63	0.02	0.02	0.01	0.01	208,424
$\text{ACC}\|P=3$	0.98	0.33	0.06	0.09	0.85	0.23	0.23	0.21	0.20	417,260
$\Phi_1,\Phi_2,\Phi_3>0$	0.99	0.88	0.32	0.51	0.89	0.85	0.83	0.80	0.74	52,098
$\Phi_2,\Phi_2,\Phi_3<0$	0.98	0.06	0.00	0.00	0.82	0.00	0.00	0.00	0.00	52,023
$\sum_{p=1}^{3}\Phi_p>0$	0.98	0.51	0.11	0.18	0.84	0.46	0.45	0.42	0.38	208,401
$\sum_{p=1}^{3}\Phi_p<0$	0.98	0.15	0.00	0.00	0.86	0.01	0.01	0.00	0.01	208,859
$\text{ACC}\|P>0$	0.94	0.29	0.04	0.05	0.59	0.21	0.21	0.17	0.16	1,250,584
$\text{ACC}\|Q=1$	0.87	0.23	0.01	0.00	0.19	0.14	0.14	0.08	0.09	416,757
$\Theta_1>0$	0.87	0.46	0.02	0.01	0.18	0.25	0.24	0.16	0.16	208,479
$\Theta_1<0$	0.87	0.00	0.00	0.00	0.20	0.04	0.04	0.01	0.02	208,278
$\text{ACC}\|Q=2$	0.93	0.26	0.02	0.01	0.19	0.16	0.16	0.10	0.10	416,714
$\Theta_1,\Theta_2>0$	0.93	0.66	0.05	0.02	0.18	0.39	0.38	0.28	0.26	104,512
$\Theta_2,\Theta_2<0$	0.92	0.00	0.00	0.00	0.20	0.01	0.01	0.00	0.01	104,652
$\sum_{p=1}^{2}\Theta_p>0$	0.93	0.49	0.04	0.01	0.18	0.29	0.29	0.20	0.19	208,435
$\sum_{p=1}^{2}\Theta_p<0$	0.93	0.02	0.00	0.00	0.20	0.03	0.03	0.01	0.01	208,279
$\text{ACC}\|Q=3$	0.92	0.24	0.02	0.01	0.19	0.16	0.16	0.11	0.11	416,462
$\Theta_1,\Theta_2,\Theta_3>0$	0.92	0.67	0.08	0.04	0.18	0.49	0.48	0.39	0.35	52,076
$\Theta_2,\Theta_2,\Theta_3<0$	0.91	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	52,426
$\sum_{p=1}^{3}\Theta_p>0$	0.91	0.42	0.04	0.02	0.18	0.31	0.30	0.21	0.21	207,990
$\sum_{p=1}^{3}\Theta_p<0$	0.92	0.07	0.00	0.00	0.20	0.02	0.02	0.01	0.01	208,472
$\text{ACC}\|Q>0$	0.90	0.24	0.02	0.01	0.19	0.16	0.15	0.10	0.10	1,249,933
$\text{ACC}\|P,Q=0$	0.98	0.99	0.99	1.00	0.81	0.89	0.89	0.95	0.94	2,499,483

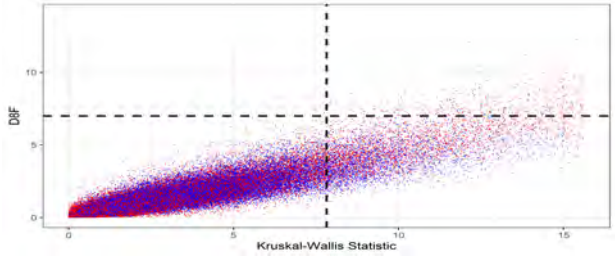
Here we calculate the accuracy based on the nominal critical values for each test and compare them to the out-of-sample prediction accuracy of the Random Forest. The breakdowns are to show accuracy when the null is false (equivalent to power) under specific seasonal dimensionality conditions. Note that there is no constraints on the parameter space of Φ or Θ beyond those which are standard.



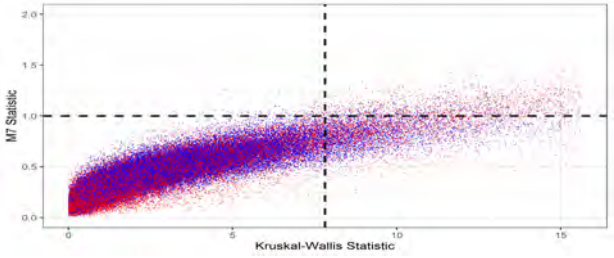
(a) Here we have generated 100,000 data sets to illustrate the variation in classification from the QS Statistic and Kruskal-Wallis Statistic. Here the QS Statistic has a nominal critical value of 5.991 (χ^2 with $df = 2$) and the axis has been scaled for both test statistics to be twice the appropriate critical value.Those series in red (blue) are seasonal (non-seasonal) by construction.



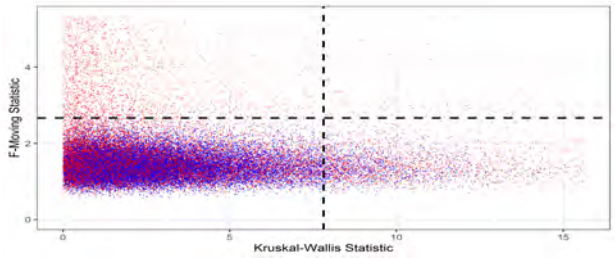
(b) Here we have generated 100,000 data sets to illustrate the variation in classification from the Freidman and Kruskal-Wallis Statistic. Both of these statistics have identical critical values (χ^2 with $df = S - 1$) and the axis has been scaled for both test statistics to be twice the appropriate critical value. Those series in red (blue) are seasonal (non-seasonal) by construction.



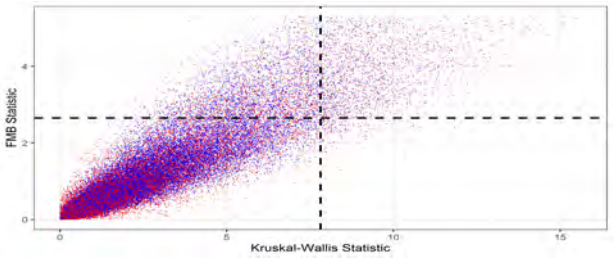
(c) Here we have generated 100,000 data sets to illustrate the variation in classification from the D8F and Kruskal-Wallis Statistic. In this case we use the suggested critical value for the D8F (7), and the axis has been scaled for both test statistics to be twice the appropriate critical value. Those series in red (blue) are seasonal (non-seasonal) by construction.



(d) Here we have generated 100,000 data sets to illustrate the variation in classification from the M7 Statistic and Kruskal-Wallis Statistic. In this case we use the suggested critical value for the M7 (1), and the axis has been scaled for both test statistics to be twice the appropriate critical value. For consistency in quadrant interpretation we have used the inverse of the M7 Statistic.



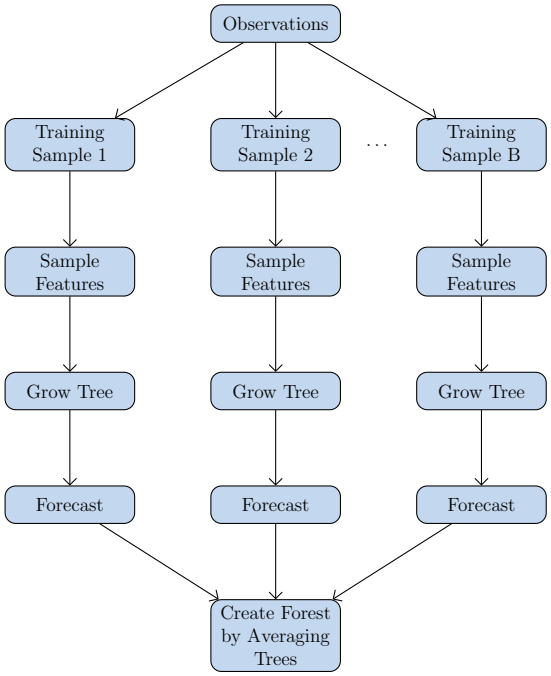
(e) Here we have generated $\approx 62,000$ data sets to illustrate the variation in classification from the F-Moving and Kruskal-Wallis Statistic. In this case we fix $T = 196, f = 4$ so as to keep the critical value for the FM constant and the axis has been scaled for both test statistics to be twice the appropriate critical value. Those series in red (blue) are seasonal (non-seasonal) by construction.



(f) Here we have generated $\approx 62,000$ data sets to illustrate the variation in classification from the F-Model Based and Kruskal-Wallis Statistic. In this case we fix $T = 196, f = 4$ so as to keep the critical value for the FMB constant and the axis has been scaled for both test statistics to be twice the appropriate critical value.

Figure 1: Variation in Classification

Figure 2: RF Algorithm



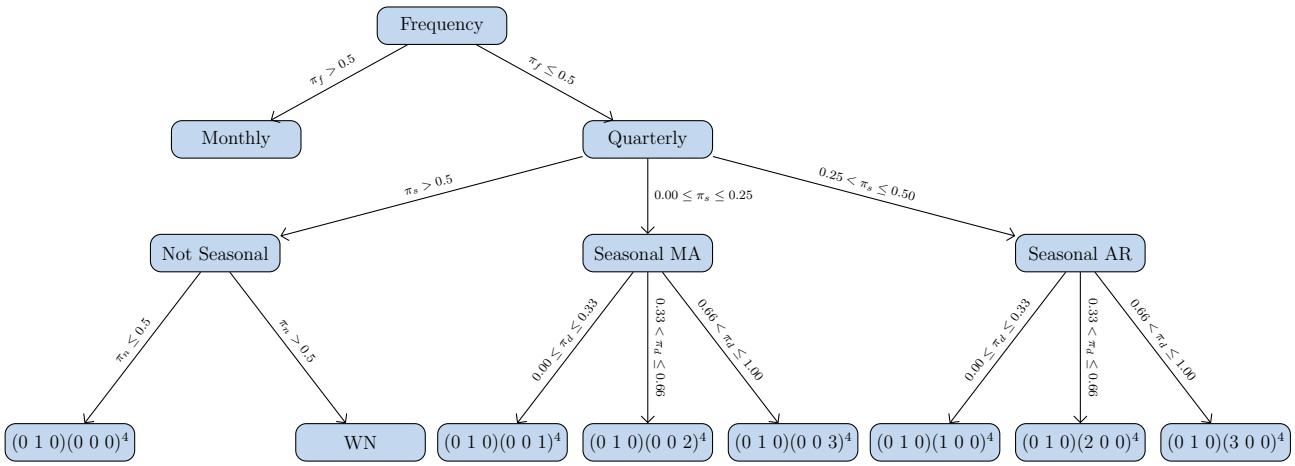
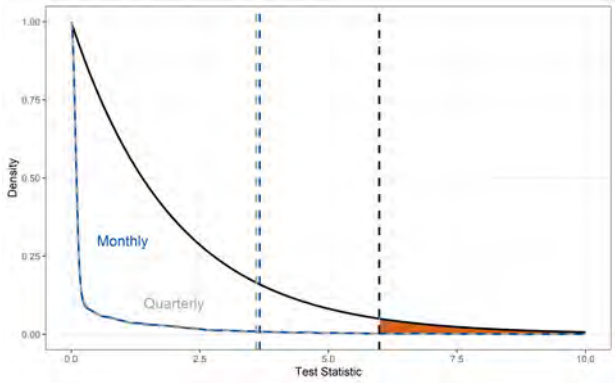
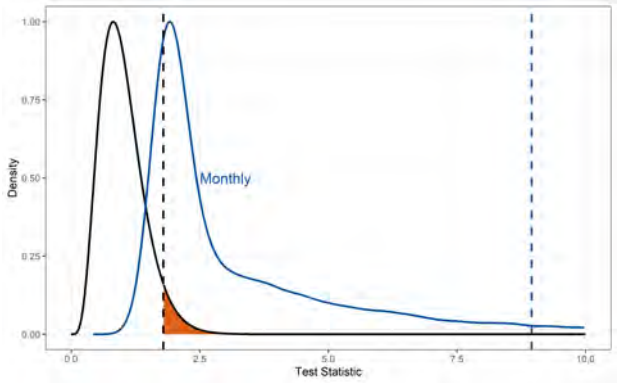


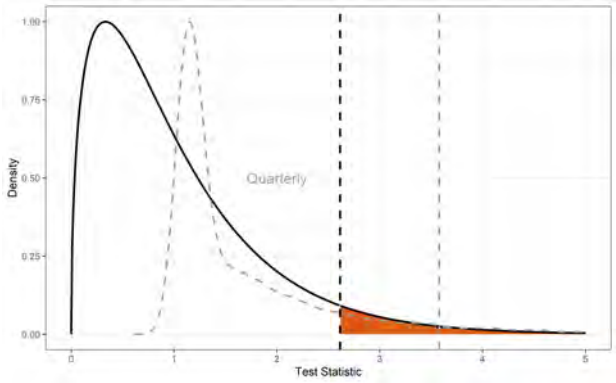
Figure 3: Simulation Model Space



(a) In this figure it is pretty clear that the empirical distribution of the QS test statistic under the null is no where near converging to the hypothesized distribution, χ^2_2. The critical value for $\alpha = 5\%$ under such sample sizes is ≈ 3.7 as compared to the nominal value of ≈ 5.9. Note that, unlike many of the other tests, the critical value for different frequency data is roughly the same for the QS statistic.

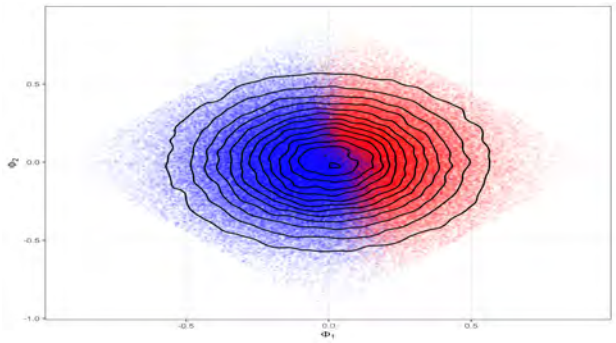


(b) This figure shows the empirical distribution (blue) and nominal distribution (black) under the null. In this case the empirical distribution is shifted to the right substantially which indicates that the corresponding critical values for any test should be much higher than are assumed. The empirical distribution implies a critical value of ≈ 8.9 compared to the nominal value of ≈ 1.79 when $\alpha = 5\%$.

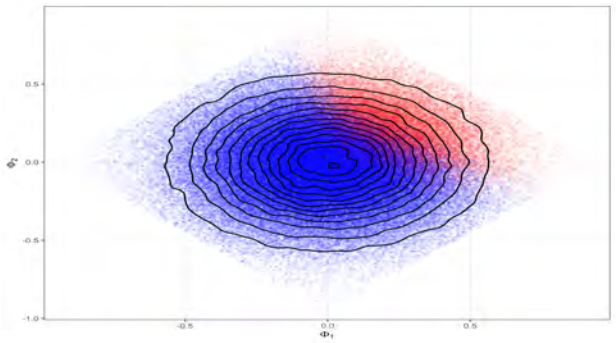


(c) Much like panel (b), this figure shows the distribution of the FM statistic under the null for quarterly data both in nominal (black solid line) and empirical (grey dashed line) terms. Here the entire distribution is shifted over indicating that the nominal critical value is too small under the null. While this is similar to the monthly data it is important to note that the test is not quite as oversized in the quarterly case.

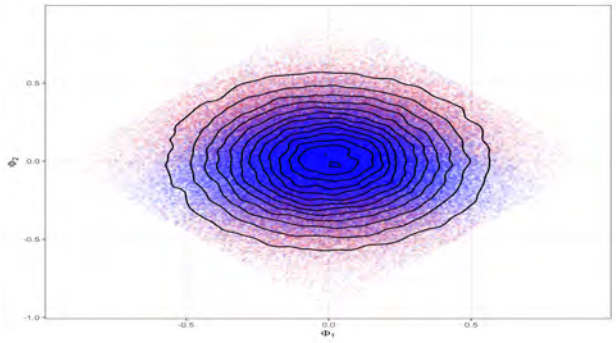
Figure 4: Empirical Vs. Nominal Distributions Under the Null



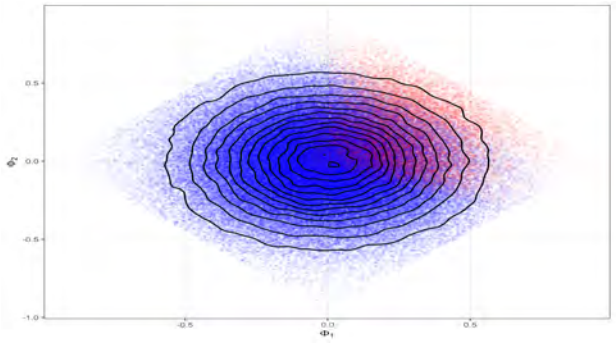
(a) QS Empirical CV: Power ≈ 38.9%



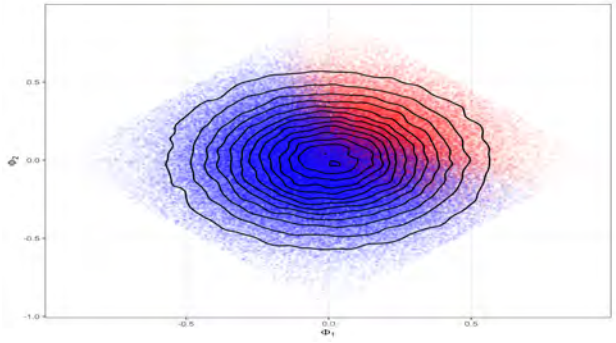
(b) D8F Empirical CV: Power ≈ 15.9%



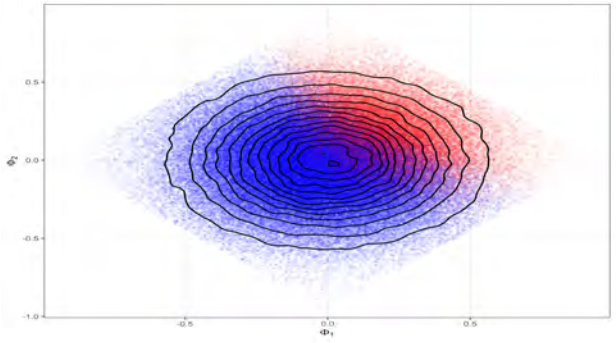
(c) FM Empirical CV: Power ≈ 18.2%



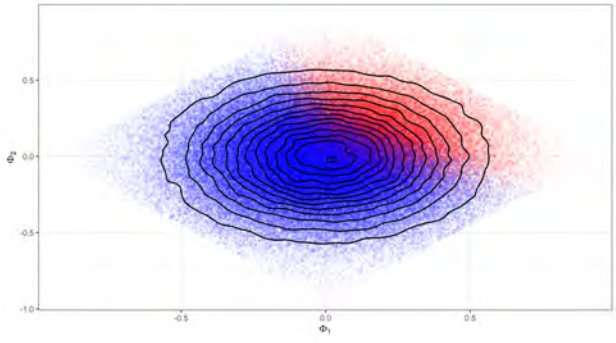
(d) M7 Empirical CV: Power ≈ 14.2%



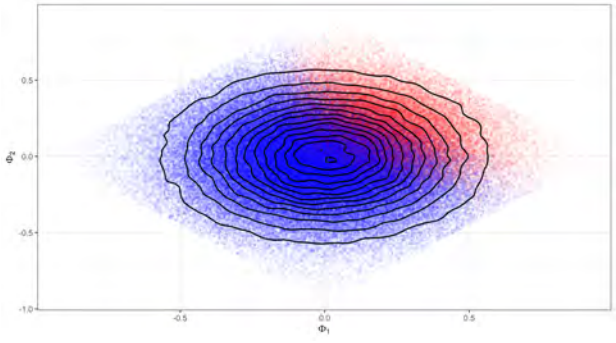
(e) FMB Empirical CV: Power ≈ 24.3%



(f) WE Empirical CV: Power ≈ 24.7%

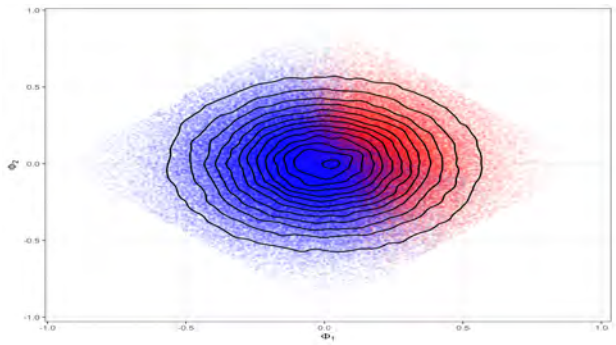


(g) KW Empirical CV: Power ≈ 21.3%

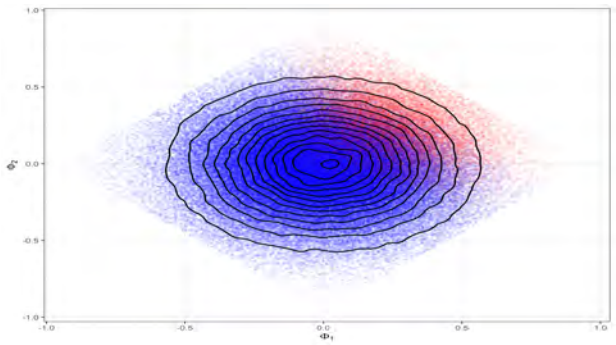


(h) FR Empirical CV: Power ≈ 19.9%

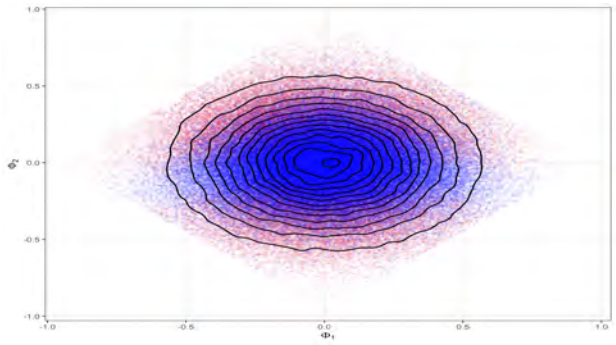
Figure 5: Power Contour Plots: Monthly Frequency



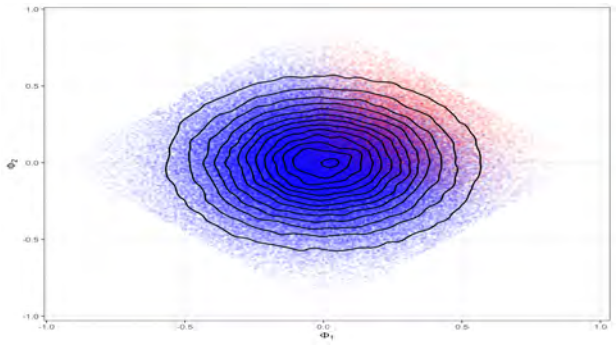
(a) QS Empirical CV: Power ≈ 29.5%



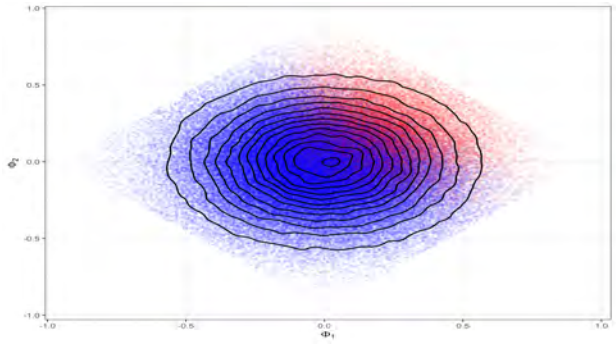
(b) D8F Empirical CV: Power ≈ 14.6%



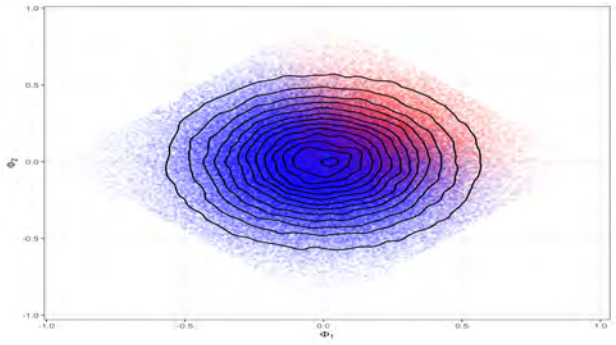
(c) FM Empirical CV: Power ≈ 19.4%



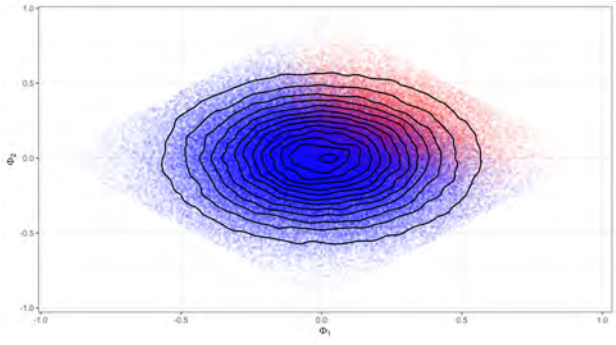
(d) M7 Empirical CV: Power ≈ 12.5%



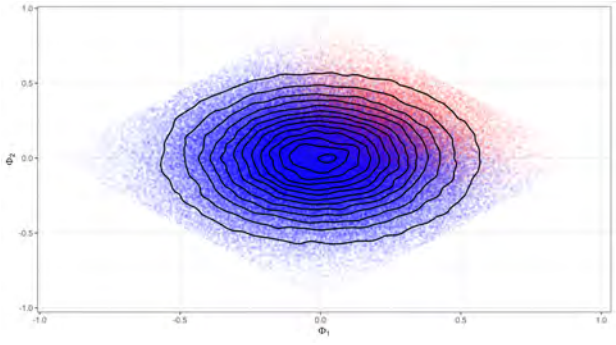
(e) FMB Empirical CV: Power ≈ 20.4%



(f) WE Empirical CV: Power ≈ 20.2%



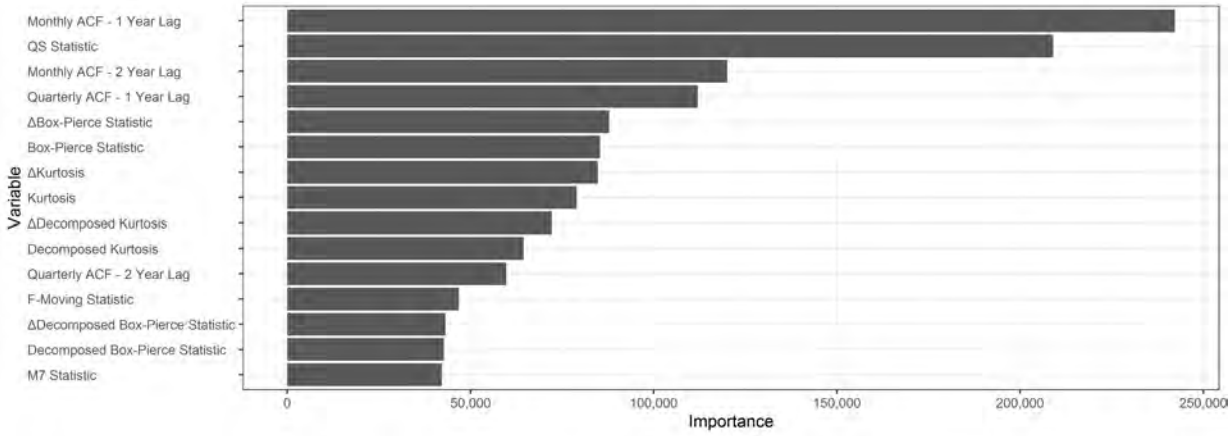
(g) KW Empirical CV: Power ≈ 14.4%



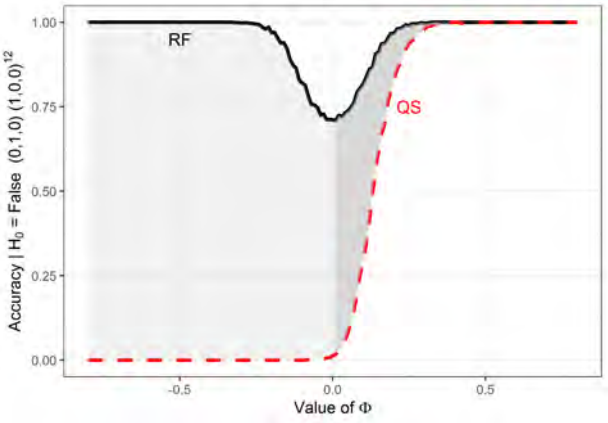
(h) FR Empirical CV: Power ≈ 13.2%

Figure 6: Power Contour Plots: Quarterly Frequency

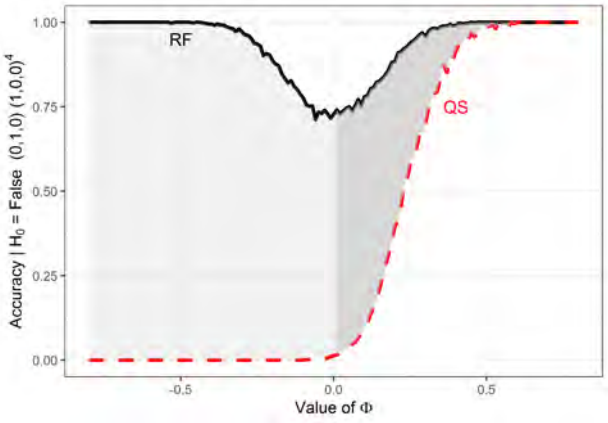
Figure 7: Top 15 Important Features



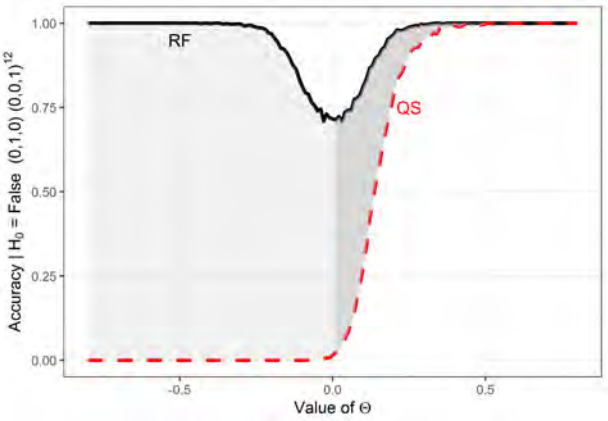
Note: Here we are providing the top fifteen features by mean reduction in Gini Impurity. We denote a first difference series by Δ and a level series with no additional qualifier (*e.g.* Δ Kurtosis refers to the Kurtosis of the first difference series. The term "Decomposed" refers to an adjusted series which removes the trend and seasonal component using an STL decomposition.



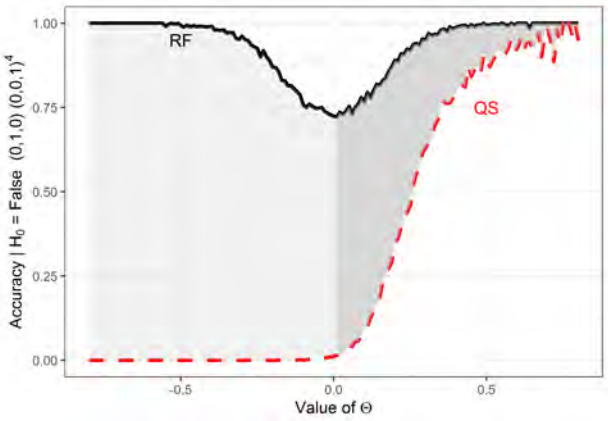
(a) Here we are plotting the accuracy curves for the RF and QS statistic over a $(0\ 1\ 0)(1\ 0\ 0)^{12}$ model structure. We have omitted the case where $\Phi_1 = 0$ since we discussed the size of the test earlier. Note how in the negative portion of the domain the QS statistic is "blind" and has zero power. This is consistent across the range of tests we examined.



(b) Here we are plotting the accuracy curves for the RF and QS statistic over a $(0\,1\,0)(1\,0\,0)^4$ model structure. We have omitted the case where $\Phi_1 = 0$ since we discussed the size of the test earlier. For quarterly data there is less data points for an equivalent number of years in monthly data and as a result the QS experiences a loss of power.

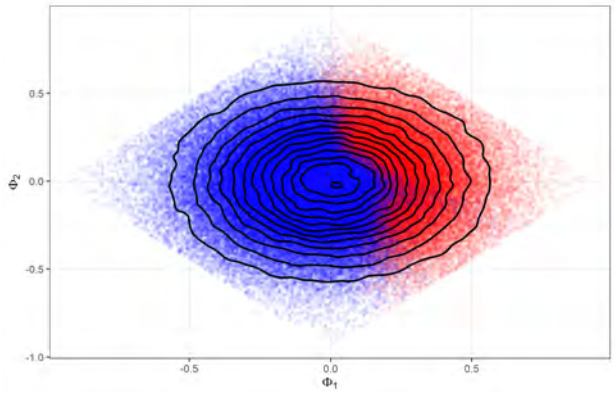


(c) Here we are plotting the accuracy curves for the RF and QS statistic over a $(0\ 1\ 0)(0\ 0\ 1)^{12}$ model structure. We have omitted the case where $\Phi_1 = 0$ since we discussed the size of the test earlier. As would be expected, a switch to an MA structure from an AR structure produces a loss in power for both the RF predictions and the QS test.

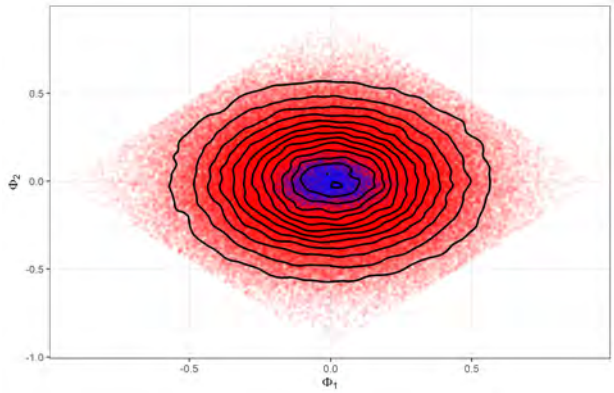


(d) Here we are plotting the accuracy curves for the RF and QS statistic over a $(0\ 1\ 0)(0\ 0\ 1)^4$ model structure. We have omitted the case where $\Phi_1 = 0$ since we discussed the size of the test earlier. The degredation continues as the number of observations shrinks due to the fruency change.

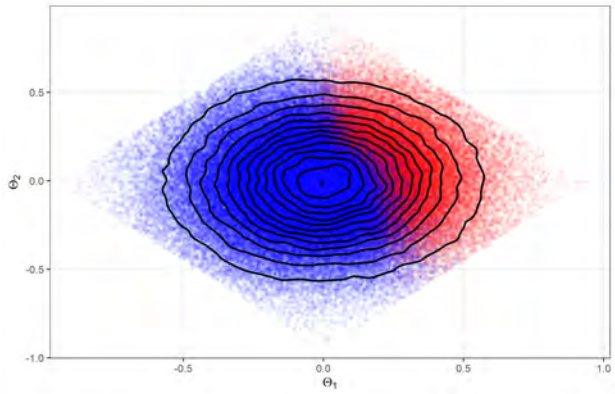
Figure 8: Accuracy over Single Dimension Seasonality



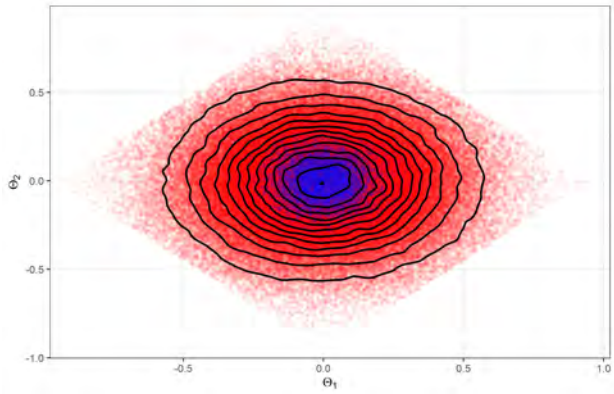
(a) Here we are plotting the contour plot of the parameters Φ from (0 1 0)(2 0 0) series. These are mixed frequency and each point represents a simulated series. If the QS statistic rejects the null hypothesis of no seasonality then the points are empty circles and red. If the QS statistic fails to reject the null they are filled circles and blue.



(b) Here we are plotting the contour plot of the parameters Φ from (0 1 0)(2 0 0) series. These are mixed frequency and each point represents a simulated series. If the RF predicts seasonality, then the points are empty circles and red. If the RF predicts no seasonality then they are filled circles and blue.



(c) Here we are plotting the contour plot of the parameters Θ from (0 1 0)(0 0 2) series. These are mixed frequency and each point represents a simulated series. If the QS statistic rejects the null hypothesis of no seasonality then the points are empty circles and red. If the QS statistic fails to reject the null they are filled circles and blue.



(d) Here we are plotting the contour plot of the parameters Θ from (0 1 0)(0 0 2) series. These are mixed frequency and each point represents a simulated series. If the RF predicts seasonality, then the points are empty circles and red. If the RF predicts no seasonality then they are filled circles and blue.

Figure 9: Accuracy over Two Dimensional Seasonality

Figure 10: Random Forest ROC: Training Data

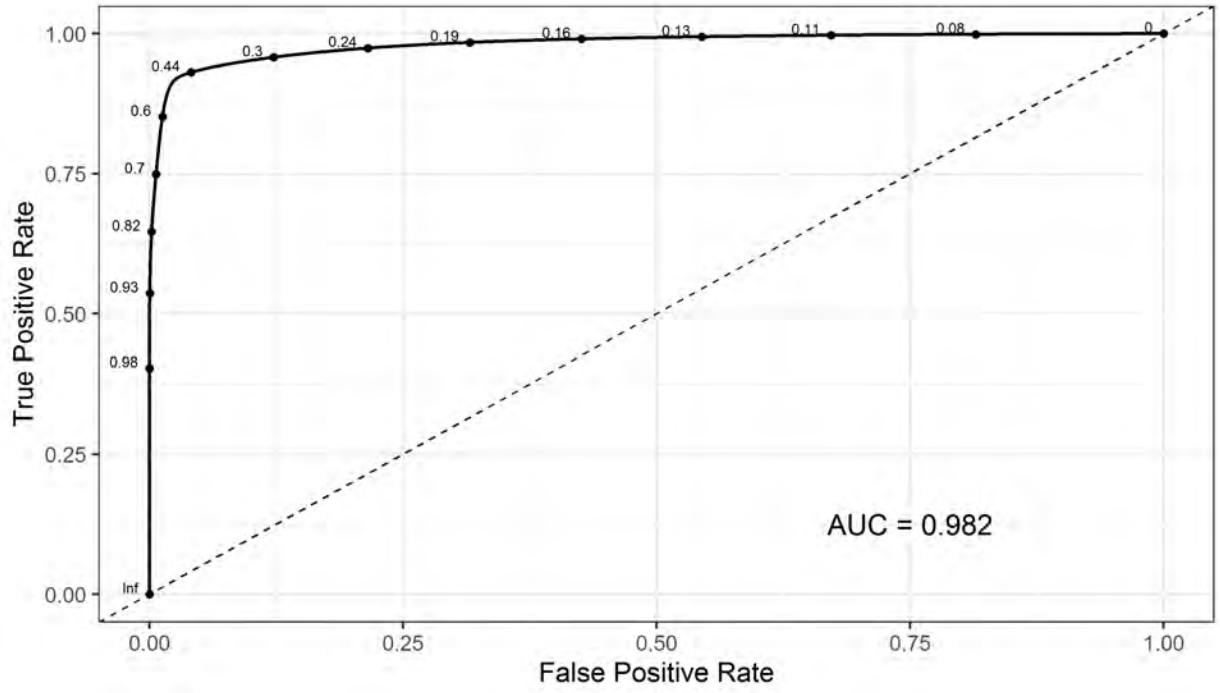
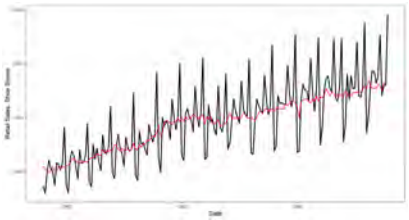
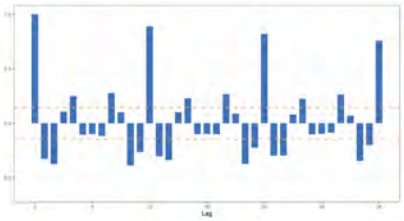


Figure 11: Retail Sales: Shoe Stores

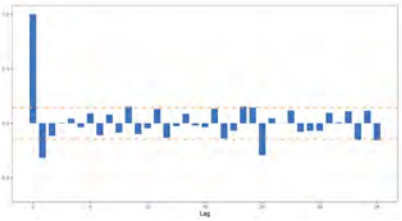
(a) Not seasonally adjusted data (black), and seasonally adjusted data (red).

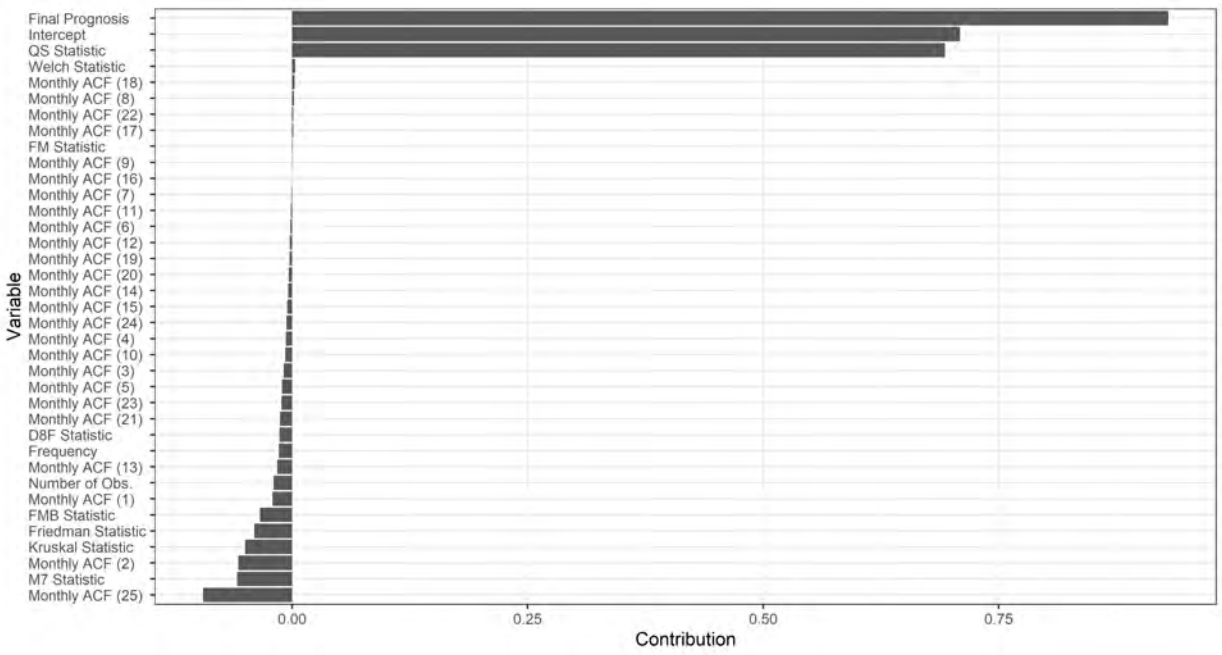


(b) Autocorrelation function of $\Delta ln(Shoes)$.

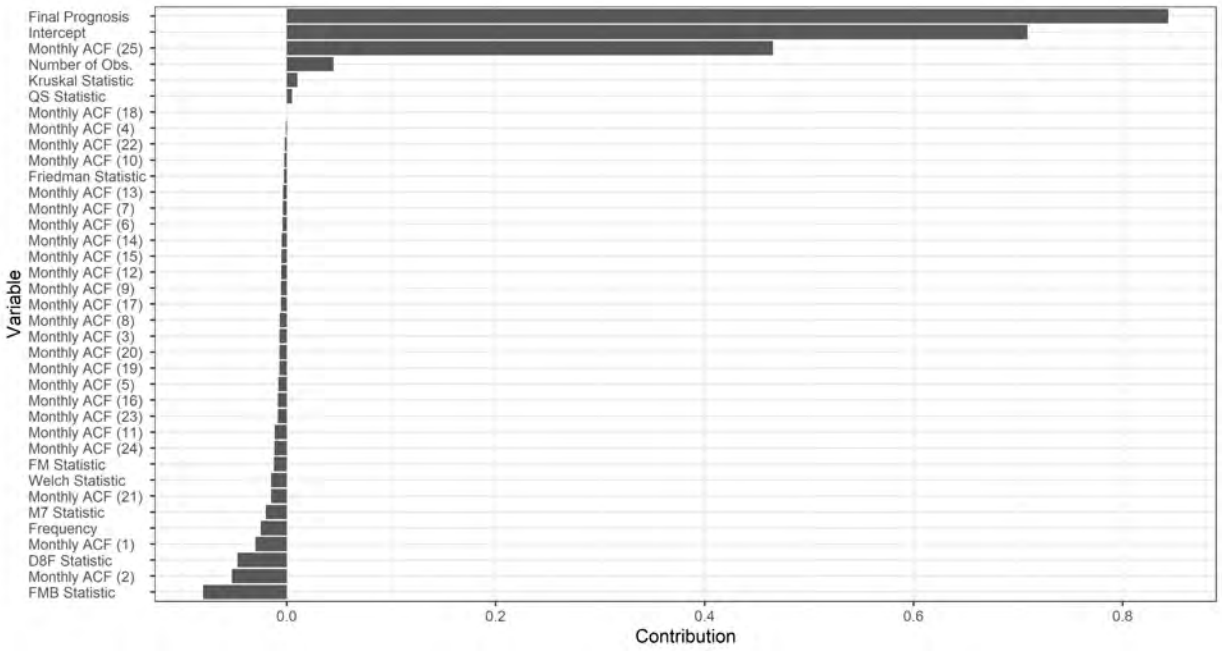


(c) Autocorrelation function of $\Delta ln(Shoes^*)$.





(d) Here we are plotting the variable contribution to predicting if the not seasonally adjusted data are seasonal and should be considered as a candidate for adjustment.



(e) Here we are plotting the variable contribution to predicting if the seasonally adjusted data are residually seasonal.

Figure 12: Retail Sales Shoe Stores: Prediction Contributions

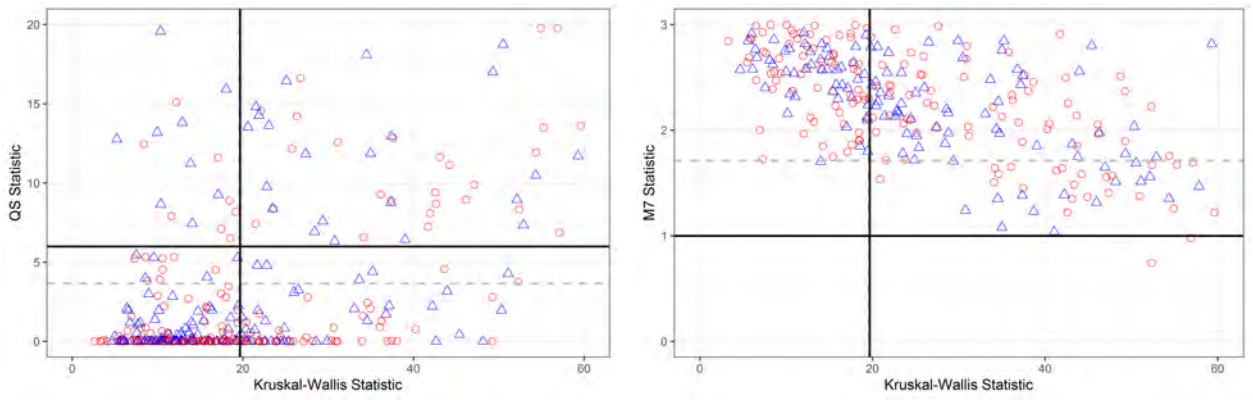


Figure 13: Variation in Test Statistics: U.S. Import/Export