

Article

A Kolmogorov-Smirnov Based Test for Comparing the Predictive Accuracy of Two Sets of Forecasts

Hossein Hassani ^{1,2,*} and Emmanuel Sirimal Silva ¹

¹ Statistical Research Centre, The Business School, Bournemouth University, Bournemouth, BH8 8EB, UK; E-Mail: esilva@bournemouth.ac.uk

² Institute for International Energy Studies (IIES), Tehran 1967743711, I.R., Iran

* Author to whom correspondence should be addressed; E-Mail: hhassani@bournemouth.ac.uk; Tel.: +44-1202-961916.

Academic Editor: Kerry Patterson

Received: 17 February 2015 / Accepted: 28 July 2015 / Published: 4 August 2015

Abstract: This paper introduces a complement statistical test for distinguishing between the predictive accuracy of two sets of forecasts. We propose a non-parametric test founded upon the principles of the Kolmogorov-Smirnov (KS) test, referred to as the KS Predictive Accuracy (KSPA) test. The KSPA test is able to serve two distinct purposes. Initially, the test seeks to determine whether there exists a statistically significant difference between the distribution of forecast errors, and secondly it exploits the principles of stochastic dominance to determine whether the forecasts with the lower error also reports a stochastically smaller error than forecasts from a competing model, and thereby enables distinguishing between the predictive accuracy of forecasts. We perform a simulation study for the size and power of the proposed test and report the results for different noise distributions, sample sizes and forecasting horizons. The simulation results indicate that the KSPA test is correctly sized, and robust in the face of varying forecasting horizons and sample sizes along with significant accuracy gains reported especially in the case of small sample sizes. Real world applications are also considered to illustrate the applicability of the proposed KSPA test in practice.

Keywords: forecast accuracy; Kolmogorov-Smirnov; stochastic dominance; non-parametric; size and power; predictive accuracy; KSPA; Diebold-Mariano; DM

JEL classifications: C15;C53

1. Introduction

There is a consensus that any attempt to justify the comparative superiority of forecasts from a given model is both incomplete and inadmissible if no consideration has been given to the statistical significance associated with the comparison. Tests on forecast evaluation and comparison have a long and detailed history which can be found in Chapter 3 of [1]. Few historically popular examples of such statistical tests are discussed in [2–5]. Of these, the Diebold-Mariano (DM) test [5] is one which is highly cited, and its popularity is evident via statements such as that in ([6] p. 8) according to which, “for comparing forecasts, DM is the only game in town.”

Whilst there is indeed no question regarding the popularity of the DM test, it is pertinent to note that the DM test is by no means a panacea. At present there exists other improved variants for evaluating the statistical significance between forecasts. Two sound examples would be Hansen’s [7] Superior Predictive Ability (SPA) test, and Hansen *et al.*’s [8] Model Confidence Set (MCS) which are superior to the DM test. In addition, recently there has been a renewed interest in research focussing on testing the predictive accuracy of forecasts through the work in [9–12]. Clark and McCracken [10] in particular shows that the DM test is inferior or inappropriate for use alongside nested forecasting models.

The aim of this paper is to introduce a complement statistical test (which differs from the tests noted above) for comparing between the predictive accuracy of forecasts whilst overcoming the constraints of the DM test which are identified below. Interestingly, regardless of the existence of more superior tests, the DM test continues to be cited in forecasting literature both in isolation and at times along side SPA and MCS tests, see for example [13–15]. In this paper, the DM test is used as a benchmark with the reasons being justified in what follows.

The DM test can be briefly introduced as an asymptotic z -test for the hypothesis that the loss differential is zero [6]¹. Whilst it is not the intention of this paper to ridicule any proven test currently adopted for comparing the accuracy of forecasts, we believe the need for a complement statistical test arises owing to the following reasons which relate to both theoretical and empirical issues with the DM test. Firstly, the original DM test was limited by finite sample properties [5]. Secondly, as a parametric test, the DM test requires that the loss differential has a stationary covariance [6]. The failure to meet this assumption invalidates the results and imposes a restriction on the applicability of this test. These issues were later addressed in [2] when a solution was achieved via the inclusion of a new assumption whereby all autocovariances of the mean loss differential beyond some lag length are assumed to be 0. However, according to the recent findings in [18,19] it has been proven that when the lag of a sample autocorrelation function (ACF) exceeds 1, the sum of the ACF is always equal to $-\frac{1}{2}$. In fact, according to [2] the modified DM statistic continues to be multiplied by the original DM statistic $[\hat{V}(\bar{d})]^{-0.5}\bar{d}$, where $\hat{V}(\bar{d}) \approx n^{-1}[\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k]$ and γ_k is the k th autocovariance of d_t . Then, as per recent findings [18,19] it implies that the sum of the autocovariance, $\sum_{k=1}^{h-1}\gamma_k = -\frac{1}{2}\gamma_0$ which in turn ensures that the expectation of $\hat{V}(\bar{d}) = 0$, and therefore the modified DM test statistic tends to infinity. Thus, if two models are used to forecast n data points without repeating or updating the data, then the modified DM test cannot be applied as the sum of the covariance will be zero. Thirdly, the modified DM test statistic for improved

¹ Note that the Granger and Newbold [16] assumption of forecast errors having zero mean is not essential according to [17].

small sample properties is dependent on the Student's t distribution [2] which cannot be justified unless the forecast errors are independent and normally distributed. In addition, even though [2] asserts that the modified DM test can provide efficient results when faced with small sample properties, in practice there can be instances when this assertion fails to hold. For example, in some instances where the Ratio of the Root Mean Squared Error ² (RRMSE) criterion shows that the forecasts from a particular model are for example 60% more accurate than the forecasts from another model (with a large sample size), the DM test fails to show a statistically significant difference between such forecasts. Moreover, when faced with comparing for example a small sample of $h = 12$ steps ahead forecasts there is a tendency for the modified DM test to always report a significant difference between forecasts even when the RRMSE criterion is at around 99%. Finally, according to the simulation results reported in [2] the modified DM test is not accurately sized for both small and large samples beyond the one-step ahead forecasting horizon.

The proposed test is founded upon the principles of the Kolmogorov-Smirnov (KS) test [20] and is non-parametric in nature. The choice of a non-parametric test is important as in the real world we are mostly faced with data which fails to meet the assumptions of normality and stationarity underlying parametric tests. The proposed test (referred to as the Kolmogorov-Smirnov Predictive Accuracy or KSPA test) was motivated by the work in [21,22], where cumulative distribution functions (c.d.f.s) relating to the absolute value of forecast errors are exploited to determine if one forecasting technique provides superior forecasts in comparison to another technique. The approach presented in the aforementioned papers are in fact based on the concept of stochastic dominance. However, the evidence presented relies purely on graphical representations and lacks a formal statistical test for significance which in turn leaves the final result open for debate.

The beauty of the proposed KSPA test is that it not only enables distinguishing between the distribution of forecasts from two models, but also enables to determine whether the model with the lowest error also reports the lowest stochastic error in comparison to the alternate model. Moreover, this test is not affected by the potential autocorrelation that may be present in forecast errors which is yet another advantage. The ability of exploiting the KSPA test for determining the model with the lowest stochastic error stems from the work of literature on stochastic dominance and as such deserves to be noted. Whilst the consideration of stochastic dominance in forecasting literature is novel, as noted in [23] stochastic dominance is widely used in econometric and actuarial literature and is therefore a well established and recognized concept. The use of KS tests for first and second order stochastic dominance dates back to the work in [24] where the author considered KS tests with independent samples with equal number of observations. Moreover, as the KS test compares each point of the c.d.f. [24,25] it has the potential of being a consistent test which considers all of the restrictions imposed by stochastic dominance [25].

The nature of the proposed KSPA test is such that it evaluates the differences in the distribution of forecasting errors as opposed to relying on the mean difference in errors as is done in the DM approach. This in itself enables the KSPA test to benefit from several advantages. Firstly, relying on the distribution

² See [14] for the calculation and interpretation of the RRMSE criterion.

of errors enables the KS test to have more power than the DM test. This is because the KSPA test essentially considers an infinite number of moments whilst the DM test only tests the first moment which is popularly referred to as the mean. Secondly, the presence of outliers can severely impact the DM test as the mean is highly sensitive to outliers in data whereas the cumulative distribution function for errors are less affected. Thirdly, a test statistic which is concentrated around a mean fails to account for the variation around the data. For example, it is possible to have two populations with identical means and yet these two populations would not really be identical if the variation around the mean is not the same. By considering the distribution of the data as is done via the proposed KSPA test, we are able to study and obtain a richer understanding of the underlying characteristics which in turn enables a more efficient and accurate decision.

The remainder of this paper is organized as follows. Section 2 presents the theoretical foundation underlying the proposed statistical test for comparing between forecasting models. Section 3 is dedicated to the results from the simulation study which compares the size and power properties of both the KSPA and modified DM tests for different sample sizes and forecasting horizons. Section 4 presents empirical evidence from applications to real data where the performance of the KSPA test is compared alongside the modified DM test, and the paper concludes in Section 5.

2. Theoretical Foundation

In this section we begin by briefly introducing the theory underlying the Kolmogorov-Smirnov test which is followed by the introduction of the hypothesis for the two-sided and one-sided KS tests which are of interest to this study. Thereafter, the KSPA test is presented for distinguishing between the distribution of forecasts errors and identifying the model with the lower stochastic error. The first part of the KSPA test, which is the two-sample two-sided KSPA test, aims at identifying a statistically significant difference between the distribution of two forecast errors (and thereby comparing the predictive accuracy of forecasts). The second part, which is the two-sample one-sided KSPA test aims at ascertaining whether the forecast with the lowest error according to some loss function also has a stochastically smaller error in comparison to the competing forecast (and thereby enables the comparison of the predictive accuracy of forecasts).

2.1. The Kolmogorov-Smirnov (KS) Test

The cumulative distribution function (c.d.f.) is an integral component of the KS test. As such, let us begin by defining the c.d.f., $F(x)$ for a random variable X . The c.d.f of X is denoted as:

$$F(x) = P(X \leq x), \quad (1)$$

where x includes a set of possible values for the random variable X . In brief, the c.d.f. shows the probability of X taking on a value less than or equal to x . The next step is to obtain the empirical c.d.f. This is because the one sample KS test (which is introduced below) aims at comparing the theoretical c.d.f. with an empirical c.d.f., whereby the latter is an approximation for the former. The empirical c.d.f. can be defined as:

$$F_n(x) = P_n(X \leq x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x), \tag{2}$$

where n is the number of observations, and I is an indicator function such that I equals 1 if $X_i \leq x$ and 0 otherwise. According to [26], as implied by the law of large numbers, for any fixed point $x \in \mathbb{R}$, the proportion of the sample contained in the set $(-\infty, x]$ approximates the probability of this set as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x) \Rightarrow \mathbb{E}I(X \leq x) = F(x), \tag{3}$$

where \mathbb{E} represents the expectation.

Then, the one sample Kolmogorov–Smirnov statistic for any given $F(x)$ can be calculated as

$$D_n = \max_x |F_n(x) - F(x)|, \tag{4}$$

where \max_x denotes the maximum of the set of distances. Note that the one sample KS test in Equation (4) compares the empirical c.d.f. with a theoretical c.d.f. However, presented next is the two sample KS test statistic which is of direct relevance to the proposed KSPA test. In contrast to the one sample KS test, the two sample KS test compares the empirical c.d.f.’s of two random variables in order to find out whether both random variables share an identical distribution, or whether they come from different distributions. Assuming two random variables X and Y , the two sample KS test statistic will be

$$D_{n_1, n_2} = \max_x |F_{X, n_1}(x) - F_{Y, n_2}(x)|. \tag{5}$$

Next, we introduce the hypothesis which are relevant for the proposed KSPA test. Let us begin by presenting the hypothesis for the two-sided KS test. Let X and Y be two random variables with c.d.f.’s F_X and F_Y , respectively. Then, a two sample, two-sided KS test will test the hypothesis that both c.d.f.’s have an identical distribution, and the resulting null and alternate hypothesis can be expressed as:

$$H_0 : F_X(z) \equiv F_Y(z) \quad \forall z \in \mathbb{Z}, \tag{6}$$

$$H_1 : F_X(z) \neq F_Y(z), \text{ for some } z \in \mathbb{Z}. \tag{7}$$

In simple terms, the null hypothesis in Equation (6) states that both X and Y share an identical distribution whilst the alternate hypothesis states that X and Y do not share the same distribution.

Finally, the hypothesis for the two sample one-sided KS test which is also known as the one-sided test of stochastic dominance is presented as in [24]:

$$H_0 : F_X(z) \leq F_Y(z) \quad \forall z \in \mathbb{Z}, \tag{8}$$

$$H_1 : F_X(z) > F_Y(z), \text{ for some } z \in \mathbb{Z}. \tag{9}$$

The important point to note here is that the alternate hypothesis in Equation (9) states that the c.d.f. of X lies above and to the left of the c.d.f. of Y , which in turn means that X has a lower stochastic error than Y . Note that in our case we consider X and Y in absolute or squared terms for example.

As with all tests, the decision making process requires the calculation of the probability value. For the KS test, there are various formulas for calculating the p -value, each with its own advantages and limitations. See for example, [27–29]. Here we rely on the formulae used in [29] to calculate the p -values for both two-sided and one-sided KS tests. In what follows, we introduce the two-sided and one-sided KSPA tests which are based on the foundations of the KS test which has been concisely explained above.

2.2. Testing for Statistically Significant Differences between the Distribution of Two Sets of Forecast Errors

The aim here is to exploit the two sample two-sided KS test (which is referred to as the two-sided KSPA test hereafter) to ascertain the existence of a statistically significant difference between the distributions of two forecast errors. Let us begin by defining forecast errors. Suppose we have a real valued, non zero time series $Y_N = (y_1, \dots, y_t, \dots, y_N)$ of sufficient length N . Y_N is divided into two parts, *i.e.*, training set and test set such that $Y_1 = (y_1, \dots, y_t)$ represents the training set and $Y_2 = (y_{t+1}, \dots, y_N)$ represents the test set. The observations in Y_1 are used to model the data whilst the observations in Y_2 are set aside for evaluating the forecasting accuracy of each model. Assume we use two forecasting techniques known as m_1 and m_2 . A loss function \mathcal{L} can be used to assess and compare between the out-of-sample forecast errors. Whilst there are varied options for \mathcal{L} , here we define \mathcal{L} as:

$$\mathcal{L}(y_{i+h} - \hat{y}_{i+h}), \quad (i = t, \dots, N - h) \quad (10)$$

where $h \geq 1$ denotes the forecasting horizon, and \hat{y}_{i+h} denotes the h -step ahead forecast of Y_i . If the forecast error is denoted by ε , then we have the expression

$$\varepsilon_{i+h} = y_{i+h} - \hat{y}_{i+h}. \quad (11)$$

In this case the forecast errors for Y_2 , obtained using models m_1 and m_2 can be denoted by

$$\varepsilon_{i+h}^{m_1} = y_{i+h} - \hat{y}_{i+h}^{m_1}, \quad \varepsilon_{i+h}^{m_2} = y_{i+h} - \hat{y}_{i+h}^{m_2}, \quad (12)$$

where $\varepsilon_{i+h}^{m_1}$ is the h -step ahead forecast errors generated from model m_1 and $\varepsilon_{i+h}^{m_2}$ is the h -step ahead forecast errors generated from model m_2 . The most common loss functions consider errors in the form of absolute values or squared values (see for example, the Mean Absolute Percentage Error and Root Mean Squared Error). As such, we can use either the absolute value of errors or squared errors when calculating the KSPA test depending on the loss function in use. Then, the absolute values and squared values of forecast errors can be calculated as

$$\varepsilon_{i+h}^{m_1} = |y_{i+h} - \hat{y}_{i+h}^{m_1}|, \quad \varepsilon_{i+h}^{m_2} = |y_{i+h} - \hat{y}_{i+h}^{m_2}|. \quad (13)$$

$$\varepsilon_{i+h}^{m_1} = (y_{i+h} - \hat{y}_{i+h}^{m_1})^2, \quad \varepsilon_{i+h}^{m_2} = (y_{i+h} - \hat{y}_{i+h}^{m_2})^2. \quad (14)$$

The forecast errors in (13) or (14) are inputs into the KSPA test for determining the existence of a statistically significant difference in the distribution of forecasts from models m_1 and m_2 . As the requirement is to test the distribution between two samples of forecast errors, the two sample two-sided KSPA test statistic can be calculated as:

$$D_{i,i+h} = \max_x |F_{\varepsilon_{i+h}^{m_1}}(x) - F_{\varepsilon_{i+h}^{m_2}}(x)|, \tag{15}$$

where $F_{\varepsilon_{i+h}^{m_1}}(x)$ and $F_{\varepsilon_{i+h}^{m_2}}(x)$ denote the empirical c.d.f.'s for the forecast errors from two different models.

Accordingly, in terms of forecast errors, the two-sided KSPA test hypothesis can be approximately represented as follows; where $\varepsilon_{i+h}^{m_1}$ and $\varepsilon_{i+h}^{m_2}$ are the absolute or squared forecast errors from two forecasting models m_1 and m_2 with unknown continuous empirical c.d.f.'s, the two-sided KSPA test will test the hypothesis:

$$H_0 : F_{\varepsilon_{i+h}^{m_1}}(z) \equiv F_{\varepsilon_{i+h}^{m_2}}(z), \tag{16}$$

$$H_1 : F_{\varepsilon_{i+h}^{m_1}}(z) \neq F_{\varepsilon_{i+h}^{m_2}}(z). \tag{17}$$

Then, if the observed significance value of the two-sample two-sided KSPA test statistic $D_{i,i+h}$ is less than α (which is usually considered at the 1%, 5% or 10% level), we reject the null hypothesis and accept the alternate which is that the forecast errors $\varepsilon_{i+h}^{m_1}$ and $\varepsilon_{i+h}^{m_2}$ do not share the same distribution. In such circumstances we are able to conclude with $1-\alpha$ confidence that there exists a statistically significant difference between the distribution of forecasts provided by models m_1 and m_2 , and thereby conclude the existence of a statistically significant difference between the two forecasts based on the two-sided KSPA test.

2.3. Testing for the Lower Stochastic Error

The aim of the two sample one-sided KS test (referred to as the one-sided KSPA test hereafter) is to identify whether the model which reports the lowest error based on some loss function also reports a stochastically smaller error in comparison to the alternate model. The usefulness of the one-sided KSPA test in distinguishing between the predictive accuracy of forecasts is most apparent in circumstances where forecasts from two models may share an identical distribution with some degree of error (as otherwise this would mean the two forecasts are exactly the same), such that one model will clearly report a comparatively lower forecast error based on some loss function. In such instances, the two-sided KSPA test would fail to identify a statistically significant difference between the two forecasts, but the one-sided KSPA test has the ability of testing the out-of-sample forecasts further in order to identify whether the model with the lower error also reports a stochastically smaller error, and thereby test for the existence of a statistically significant difference between two forecasts.

In terms of forecast errors, the two-sample, one-sided KSPA test hypothesis can be approximately represented as follows. Once again, where $\varepsilon_{i+h}^{m_1}$ and $\varepsilon_{i+h}^{m_2}$ are the absolute or squared forecast errors from two forecasting models m_1 and m_2 with unknown continuous empirical c.d.f.'s, the two sample one-sided KSPA test will test the hypothesis:

$$H_0 : F_{\varepsilon_{i+h}^{m_1}}(z) \leq F_{\varepsilon_{i+h}^{m_2}}(z), \tag{18}$$

$$H_1 : F_{\varepsilon_{i+h}^{m_1}}(z) > F_{\varepsilon_{i+h}^{m_2}}(z). \tag{19}$$

The acceptance of the alternate hypothesis in this case translates to the c.d.f. of forecast errors from model m_1 lying towards the left and above the c.d.f. of forecast errors from model m_2 . More specifically

the acceptance of the alternate hypothesis confirms that model m_1 reports a lower stochastic error than model m_2 . Recall the relationship identified in [21] that if the c.d.f. for absolute value of forecast errors from one model lies above and hence to the left of that for the other model, the model lying above had a lower stochastic error than the other model. The one-sided KSPA test evaluates this notion and provides a statistically valid foundation which was previously lacking.

3. Simulation Results

3.1. Size of the Test

The first part of the simulation study focuses mainly on the size properties of the proposed KSPA test. The actual size of nominal 10% level tests are estimated against a two-sided alternative as in [2,5], and the simulation study itself follows the exact process as in [2]. Whilst we have considered and reported the results from errors generated via various noise distributions, as an example we explain one of the processes, *i.e.*, the process involved in the Gaussian white noise simulation. We have simulated independent standard normal white noise error series (e_{1t}, e_{2t}) , $t = 1, 2, \dots, n$, for various sample sizes n . Forecasts which cover both short and long run horizons, more specifically up to $h = 10$ were evaluated. As in [2], the information related to the simulated white noise error series were incorporated in the test statistics only in the case of $h = 1$. In order to enable comparison with the results in [2], we considered the squared errors, *i.e.*, e_{1t}^2 and e_{2t}^2 over the entire simulation study. All simulation results reported are based on 10,000 replications and were programmed in *R*.

The results for the size properties are reported in Table 1. Firstly, as noted in [2] the modified DM test remains somewhat oversized as visible in the results shown in Table 1. Yet, the authors concluded this was acceptable as the modified DM test results showed a major improvement over the previous version. Based on the results, we see that for the Gaussian white noise errors which are directly comparable with the modified DM test results in [2], the proposed KSPA test is correctly sized across all sample sizes, both large and small, and across all horizons of up to 10 steps-ahead.

Unlike the results reported in [2], we also show the outcomes from the simulation study which considered Uniform distribution, Cauchy distribution, and heavy tailed errors. The Cauchy white noise distributions are likely in time series which are affected by catastrophic events. The heavy tailed distribution is a Student's t distribution with six degrees of freedom as considered and explained in [2]. The findings from the simulation study indicates the superiority of the proposed KSPA test over the modified DM test in terms of being correctly sized across all sample sizes and all horizons even when faced with varying noise distributions.

It is noteworthy that the results reported in Table 1 represents a subset of results obtained from an extensive simulation study. Following the simulation study in [2], we also considered (1) contemporaneously correlated forecast errors with contemporaneous correlations of 0.5 and 0.9, and (2) autocorrelated forecast errors. We noticed the results were similar to what is reported for the other distributions in Table 1 as it continued to illustrate that the KSPA test is indeed correctly sized across all sample sizes and forecasting horizons. As such in order to save space we do not report these results here.

In summary, it is evident that in comparison to the modified DM test, the KSPA test shows major improvements not only across different forecasting horizons, but more importantly over small sample sizes. As noted in [2] it is the performance over small sample sizes that is of utmost importance to practitioners as in reality very large number of forecasts are not often available for comparison purposes, and the proposed KSPA test has proven to be comparatively more accurate in this case with far better results.

Table 1. Percentage of rejections of the true null hypothesis of equal prediction mean squared errors for the Diebold-Mariano test and equal distribution of squared prediction errors for the Kolmogorov-Smirnov Predictive Accuracy (KSPA) test at nominal 10% level.

| <i>h</i> | Error Distribution | Test | <i>n</i> = 8 | <i>n</i> = 16 | <i>n</i> = 32 | <i>n</i> = 64 | <i>n</i> = 128 | <i>n</i> = 256 | <i>n</i> = 512 |
|----------|--------------------|------|--------------|---------------|---------------|---------------|----------------|----------------|----------------|
| 1 | Gaussian | DM | 8.4 | 9.6 | 9.7 | 10.1 | 9.9 | 10.4 | 10.6 |
| | Gaussian | KSPA | 8.6 | 9.4 | 8.9 | 9.6 | 8.4 | 9.4 | 8.6 |
| | Uniform | KSPA | 9.1 | 8.9 | 8.6 | 9.4 | 8.9 | 8.9 | 8.5 |
| | Cauchy | KSPA | 9.0 | 9.1 | 8.4 | 9.2 | 8.5 | 8.9 | 8.6 |
| | Student's <i>t</i> | KSPA | 8.5 | 9.4 | 9.3 | 9.5 | 9.0 | 8.7 | 8.6 |
| 2 | Gaussian | DM | 16.4 | 14.2 | 12.2 | 11.2 | 10.8 | 10.5 | 10.3 |
| | Gaussian | KSPA | 9.0 | 9.5 | 8.5 | 9.2 | 8.6 | 9.1 | 8.4 |
| | Uniform | KSPA | 9.1 | 9.4 | 8.9 | 9.8 | 8.8 | 9.2 | 8.8 |
| | Cauchy | KSPA | 9.3 | 9.5 | 9.0 | 9.3 | 8.8 | 9.4 | 9.0 |
| | Student's <i>t</i> | KSPA | 8.7 | 9.3 | 9.1 | 9.1 | 8.4 | 9.7 | 8.9 |
| 3 | Gaussian | DM | 18.1 | 18.5 | 14.3 | 12.2 | 10.7 | 10.8 | 10.9 |
| | Gaussian | KSPA | 8.6 | 9.6 | 8.7 | 9.2 | 8.7 | 9.1 | 9.1 |
| | Uniform | KSPA | 8.7 | 9.8 | 9.0 | 9.2 | 8.6 | 9.4 | 8.7 |
| | Cauchy | KSPA | 8.4 | 9.4 | 9.3 | 9.7 | 8.7 | 9.5 | 8.7 |
| | Student's <i>t</i> | KSPA | 8.2 | 9.7 | 8.8 | 9.5 | 8.9 | 9.1 | 8.6 |
| 4 | Gaussian | DM | 16.3 | 19.8 | 16.1 | 13.4 | 11.5 | 10.9 | 11.0 |
| | Gaussian | KSPA | 8.5 | 9.4 | 8.3 | 8.9 | 8.6 | 9.2 | 9.0 |
| | Uniform | KSPA | 8.7 | 9.6 | 8.6 | 9.2 | 9.4 | 9.6 | 9.1 |
| | Cauchy | KSPA | 8.4 | 9.4 | 9.0 | 9.4 | 9.6 | 9.7 | 8.7 |
| | Student's <i>t</i> | KSPA | 8.7 | 9.1 | 8.8 | 9.9 | 8.7 | 9.7 | 8.8 |
| 5 | Gaussian | DM | 12.9 | 19.9 | 17.8 | 14.9 | 12.2 | 11.1 | 11.0 |
| | Gaussian | KSPA | 8.4 | 9.4 | 8.9 | 9.4 | 8.3 | 9.7 | 8.3 |
| | Uniform | KSPA | 8.2 | 9.2 | 8.7 | 9.1 | 8.4 | 9.3 | 8.9 |
| | Cauchy | KSPA | 8.8 | 9.6 | 8.5 | 9.5 | 9.0 | 8.8 | 8.9 |
| | Student's <i>t</i> | KSPA | 8.4 | 9.3 | 9.1 | 9.9 | 9.1 | 9.6 | 8.6 |

Table 1. Cont.

| <i>h</i> | Error Distribution | Test | <i>n</i> = 8 | <i>n</i> = 16 | <i>n</i> = 32 | <i>n</i> = 64 | <i>n</i> = 128 | <i>n</i> = 256 | <i>n</i> = 512 |
|----------|--------------------|------|--------------|---------------|---------------|---------------|----------------|----------------|----------------|
| 6 | Gaussian | DM | 10.6 | 19.8 | 18.8 | 16.0 | 12.9 | 11.4 | 11.2 |
| | Gaussian | KSPA | 8.6 | 9.5 | 8.9 | 9.5 | 8.6 | 9.1 | 9.0 |
| | Uniform | KSPA | 8.7 | 9.4 | 8.8 | 9.1 | 8.4 | 9.2 | 8.3 |
| | Cauchy | KSPA | 8.9 | 9.8 | 9.1 | 9.9 | 8.5 | 9.2 | 8.6 |
| | Student's <i>t</i> | KSPA | 8.7 | 9.3 | 8.8 | 9.4 | 9.0 | 9.8 | 9.1 |
| 7 | Gaussian | DM | 9.9 | 18.2 | 19.5 | 16.8 | 13.6 | 11.6 | 11.4 |
| | Gaussian | KSPA | 8.6 | 9.5 | 9.3 | 8.9 | 8.8 | 9.3 | 9.0 |
| | Uniform | KSPA | 8.4 | 9.0 | 8.7 | 9.9 | 9.0 | 9.1 | 8.7 |
| | Cauchy | KSPA | 8.5 | 9.2 | 8.7 | 9.1 | 9.0 | 9.4 | 8.9 |
| | Student's <i>t</i> | KSPA | 8.8 | 9.1 | 9.0 | 9.0 | 8.6 | 8.8 | 9.2 |
| 8 | Gaussian | DM | - | 17.4 | 20.2 | 18.0 | 13.8 | 11.9 | 11.4 |
| | Gaussian | KSPA | - | 9.3 | 8.6 | 9.1 | 8.5 | 9.5 | 8.7 |
| | Uniform | KSPA | - | 9.5 | 8.7 | 9.8 | 9.0 | 9.7 | 8.7 |
| | Cauchy | KSPA | - | 9.5 | 8.3 | 9.2 | 8.8 | 8.9 | 8.9 |
| | Student's <i>t</i> | KSPA | - | 9.7 | 8.3 | 9.6 | 8.6 | 9.1 | 9.1 |
| 9 | Gaussian | DM | - | 15.1 | 20.2 | 19.0 | 14.7 | 12.4 | 11.6 |
| | Gaussian | KSPA | - | 9.5 | 8.6 | 9.2 | 8.5 | 9.4 | 8.8 |
| | Uniform | KSPA | - | 9.4 | 9.0 | 9.7 | 8.0 | 9.5 | 8.9 |
| | Cauchy | KSPA | - | 9.8 | 8.6 | 8.9 | 8.6 | 9.4 | 8.8 |
| | Student's <i>t</i> | KSPA | - | 9.1 | 8.6 | 9.2 | 8.9 | 9.6 | 9.0 |
| 10 | Gaussian | DM | - | 14.0 | 20.2 | 19.1 | 15.1 | 12.6 | 11.8 |
| | Gaussian | KSPA | - | 9.2 | 8.9 | 9.3 | 8.7 | 9.7 | 9.0 |
| | Uniform | KSPA | - | 9.2 | 8.7 | 9.8 | 8.7 | 9.1 | 9.4 |
| | Cauchy | KSPA | - | 9.2 | 8.8 | 9.7 | 9.1 | 9.5 | 9.3 |
| | Student's <i>t</i> | KSPA | - | 9.3 | 8.8 | 9.0 | 8.7 | 9.1 | 8.6 |

Note: The DM test results relate to the modified DM test and were extracted from Table 1 in [2].

3.2. Power of the Test

Next, we consider the power of the modified DM test and the proposed KSPA test. For this purpose we generate forecast errors from different combinations of distributions which will certainly result in significantly different forecast errors so that we can obtain an accurate evaluation of the power of the tests in relation to small and large sample sizes. The details of the combinations evaluated are explained in the footnotes of Table 2 which also reports the results. We evaluate the power of the two tests over the one-step horizon because power calculations are only valid if a given test is correctly sized, and we know that the modified DM test suffers from problems of being oversized especially beyond $h = 1$ [2].

Once again, we report a subset of our results as the general conclusion remains similar. Firstly, it is evident that the KSPA test is more powerful than the modified DM test for both small and large sample

sizes. Secondly, we notice the KSPA test converging towards a power of 100% faster than the modified DM test in all cases whereas on most instances our results (including those not reported here) showed that the modified DM test fails to converge to 100% over these sample sizes. The only exception being in the case of autocorrelated errors as in Case 3, skewed errors as in Case 4 or where we compared forecast errors generated from a MA(1) process against those generated from an AR(1) process.

In summary, the simulation study has shown that the proposed KSPA test is correctly sized across all sample sizes and forecasting horizons, and that it is more powerful than the DM test, and thereby proving its practicality and suitability as a complement statistical test for distinguishing between the predictive accuracy of forecasts.

Table 2. Percentage of rejections of the false null hypothesis of equal one-step prediction mean squared errors for the Diebold-Mariano test and equal one-step distribution of squared prediction errors for the KSPA test at nominal 10% level.

| Combinations | Test | $n = 8$ | $n = 16$ | $n = 32$ | $n = 64$ | $n = 128$ | $n = 256$ | $n = 512$ |
|--------------|------|---------|----------|----------|----------|-----------|-----------|-----------|
| Case 1 | DM | 7.3 | 17.5 | 31.9 | 37.3 | 39.3 | 40.3 | 40.9 |
| | KSPA | 19.6 | 35.8 | 61.0 | 91.7 | 99.9 | 100.0 | 100.0 |
| Case 2 | DM | 5.2 | 13.4 | 26.5 | 35.4 | 39.5 | 41.0 | 40.8 |
| | KSPA | 15.9 | 25.8 | 42.0 | 75.3 | 97.6 | 100.0 | 100.0 |
| Case 3 | DM | 59.3 | 96.0 | 99.7 | 100.0 | 100.0 | 100.0 | 100.0 |
| | KSPA | 65.1 | 92.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Case 4 | DM | 91.6 | 99.7 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| | KSPA | 97.3 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Note: Case 1: Compares errors from a Cauchy distribution with mean 0 and standard deviation 1 against errors from a $N(0, 1)$ distribution. Case 2: Compares errors from a Student's t distribution with 6 d.f. against errors from a Cauchy distribution. Case 3: Compares errors from $N(0, 1)$ against autocorrelated errors. Case 4: Compares errors from a skewed χ^2 distribution with 3 d.f. against errors from a χ^2 distribution with 10 d.f.

4. Empirical Evidence

Following the simulation study which illustrated the superiority of the proposed KSPA test in terms of being correctly sized and more powerful than the modified DM test, in this section we consider the KSPA test for several real world applications. Note that all applications here use the RMSE as the loss function, and therefore the KSPA test like the DM test relies on squared errors in all instances. These real world applications have been carefully selected to illustrate that: (i) The KSPA test can accurately perform the same task as the modified DM test in practice when faced with real data. (ii) Both two-sided and one-sided KSPA tests can be of benefit in practice. (iii) The KSPA test is applicable where the modified DM test cannot be applied. (iv) The KSPA test can handle both small and large sample sizes. (v) The KSPA test is suitable across different forecasting horizons. (vi) The KSPA test is not affected by the generation of forecast errors from either parametric or non-parametric models.

Note that the R code for the KSPA test is provided in the Appendix, and the data relevant for the scenario's discussed below are also provided as Supplementary Materials to enable the reader to replicate the results shown here.

4.1. Scenario 1: Tourism Series

The rationale for this application is to mainly show that the KSPA test can perform the same task as the modified DM test in practice in addition to showing its performance when faced with a relatively large number of out-of-sample forecast errors. We consider testing forecasts from two models, *i.e.*, Singular Spectrum Analysis (SSA) which is non-parametric and ARIMA (parametric) for a statistically significant difference in terms of providing $h = 1$ step ahead forecasts for total U.S. tourist arrivals³. This data set was recently used in [14] and the related forecasts of $n = 69$ observations were extracted from that study. Figure 1 shows the out-of-sample forecasts, distribution of errors and the empirical c.d.f. for U.S. tourist arrivals obtained via SSA and ARIMA models. Based on the forecasts figure alone one is not able to determine whether there exists a statistically significant difference between the forecasts from SSA and ARIMA. As such, we then look to the distribution of the squared forecast errors from ARIMA and SSA which can be seen in Figure 1 (middle). However, without a formal statistical test it is not possible to determine whether there exists a statistically significant difference between the distribution of these errors. Thirdly, we look at the empirical c.d.f.'s shown in Figure 1 (right) to identify if one model does indeed provide a lower stochastic error than the other model as suggested in [21]. In this case it is clear that based on the empirical c.d.f., it appears that the out-of-sample forecasts from SSA provide a lower stochastic error than the out-of-sample forecasts from ARIMA. However, as mentioned in the introduction this conclusion is open to debate as it lacks a mandatory statistical test.

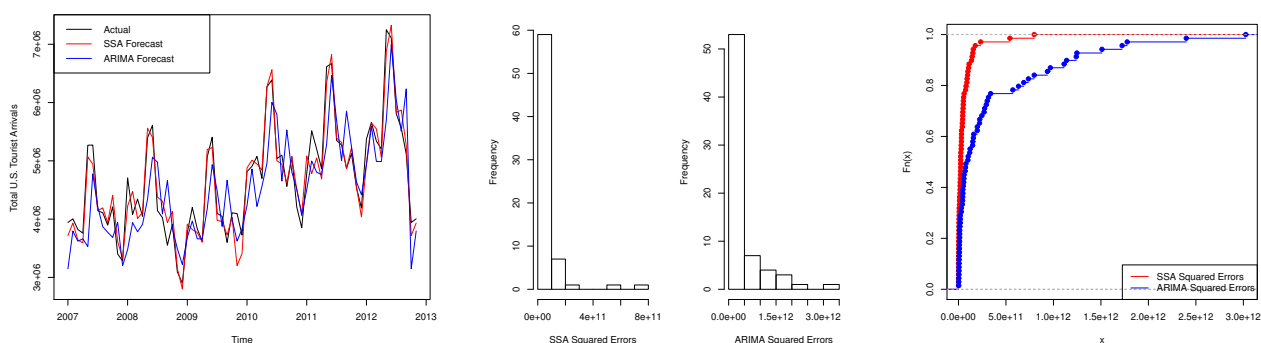


Figure 1. U.S. Tourist arrivals forecast, distribution of errors and empirical cumulative distribution functions (c.d.f.) of errors.

When we calculate the RRMSE statistic, it shows that the forecasts obtained from the SSA model are 60% better than the forecasts obtained via the ARIMA model. Accordingly one would expect a statistically significant difference between the forecasts of SSA and ARIMA. We apply both modified

³ Data source: <http://travel.trade.gov/research/monthly/arrivals/>.

DM and KSPA tests. The results are reported in Table 3. In this case, the modified DM test correctly identifies that there exists a statistically significant difference between the forecasts from SSA and ARIMA. In terms of the newly proposed KSPA test, firstly the two-sided KSPA test confirms that there is indeed a statistically significant difference between the distribution of forecast errors from SSA and ARIMA for U.S. tourist arrivals at $h = 1$ step ahead, and thereby confirms the existence of a statistically significant difference between the two forecasts. Next, we apply the one-sided KSPA test to find out whether SSA forecasts (which has the lower RMSE) reports a lower stochastic error than ARIMA forecasts. The one-sided KSPA test confirms that SSA does in fact provide forecasts which report a lower stochastic error than the ARIMA model as suggested by the empirical c.d.f.'s in Figure 1 (right), and provides supplementary evidence to the conclusion from the two-sided KSPA test for the existence of a statistically significant difference between the two forecasts. The results from the modified DM test and KSPA tests are significant at a 95% confidence level.

Table 3. Evaluating $h = 1$ step ahead forecasts for U.S. tourist arrivals.

| Test | Two-Sided (p -Value) | One-Sided (p -Value) |
|-------------|-------------------------|-------------------------|
| Modified DM | <0.01 * | N/A |
| KSPA | <0.01 * | <0.01 * |

Note: * indicates results are statistically significant based on a p -value of 0.05.

4.2. Scenario 2: Accidental Deaths Series

The main reason to present this next application is to show how the KSPA test can overcome a limitation of the modified DM test. The U.S. death series is well known and has been used widely in previous time series analysis and forecasting applications (see for example, [30–33]). We follow a similar forecasting approach to that reported in [31]. In our case we consider forecasting the last 12 points of the death series such that the first forecast point represents the horizon of $h = 1$, the second forecast point represents $h = 2$ and so on, up until the final forecast point which represents the $h = 12$ steps ahead forecast. As explained in Section 1, the modified DM test cannot be used in such scenarios where the out-of-sample forecast errors relate to various horizons within a single forecasting exercise as the sum of the covariance will equate to zero. However, it is possible to rely on the original DM test (*i.e.*, without considering the covariance effect) in such scenarios, but it is not advisable owing to the many limitations of the original DM test as identified in [2]. The forecasts are obtained via the parametric ARIMA model and a non-parametric Neural Networks (NN) model, and the ARIMA forecasts report a lower RMSE.

Figure 2 shows the out-of-sample forecasts, distribution of errors and the empirical c.d.f. for the U.S. death series obtained via ARIMA and NN models. In this case based on the empirical c.d.f. we are able to state that the ARIMA forecasts report a stochastically smaller error than the NN forecasts. The two-sided KSPA test can be used to test for statistically significant differences between the two forecasts whilst the one-sided KSPA test can be exploited to provide statistical evidence for the claim based on [21]. The resulting output from the KSPA and original DM tests are reported in Table 4. Initially, the two-sided KSPA test confirms that there is indeed a statistically significant difference between the distribution of

forecast errors from ARIMA and NN at a 95% confidence level. Secondly, the one-sided KSPA test confirms that ARIMA does in fact provide forecasts which report a lower stochastic error than forecasts from the NN model as suggested by the empirical c.d.f.'s in Figure 2 (right). Note that whilst the original DM test too proves the existence of a statistically significant difference between the two forecasts, the two-sided KSPA test reports a lower p -value than the original DM test.

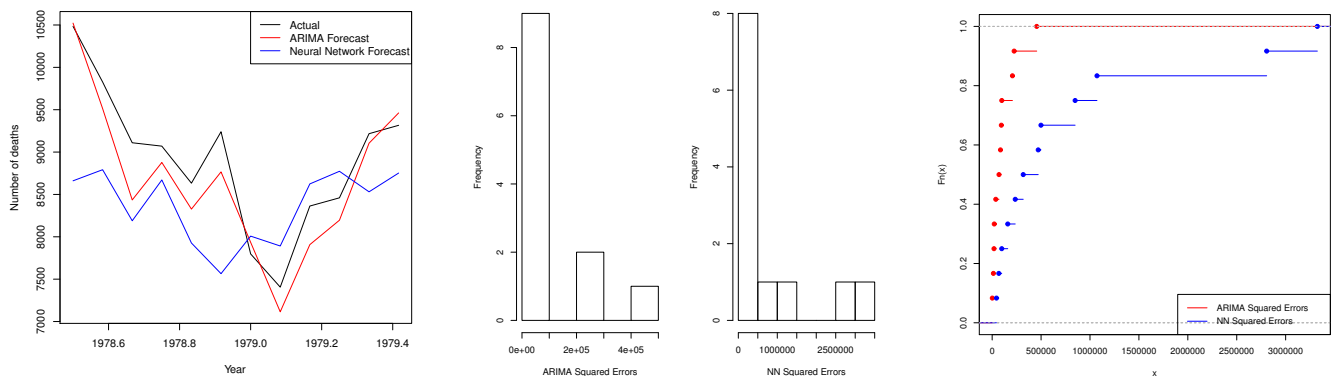


Figure 2. U.S. death series forecast, distribution of errors and empirical c.d.f. of errors.

Table 4. Evaluating $h = 1, \dots, h = 12$ step ahead forecasts for the U.S. death series.

| Test | Two-Sided (p -Value) | Greater (p -Value) |
|-------------|-------------------------|-----------------------|
| DM | 0.04 * | N/A |
| Modified DM | N/A | N/A |
| KSPA | 0.03 * | 0.02 * |

Note: * indicates results are statistically significant based on a p -value of 0.05.

4.3. Scenario 3: Trade Series

Finally, the purpose of this empirical example is to show the superiority of the KSPA test over the modified DM test, and also to show how the one-sided KSPA test is useful when the two-sided KSPA test fails at showing a statistically significant difference between two forecasts. In this application we consider forecasts for U.S. imports⁴ between March 2011-December 2011 at $h = 3$ steps ahead using ETS and SSA which are both non-parametric techniques. This data set was recently used in [13] and the forecasts considered here are those generated in that study. This is another example of a scenario with a small sample size, *i.e.*, $n = 10$. Figure 3 shows the out-of-sample forecasts, distribution of errors and the empirical c.d.f. of errors obtained via ETS and SSA. Here, unlike on previous occasions, based on this forecast figure alone one is able to see that there exists a significant difference between the forecasts from both models. However, it cannot be verified in the absence of statistical evidence. The distribution

⁴ Data source: <http://www.bea.gov/international/index.htm>.

of out-of-sample forecast errors are shown in Figure 3 (middle) and the resulting empirical c.d.f. for the squared forecast errors are also presented (right). Based on the empirical c.d.f. in Figure 3 we can see that except for three points, at every other observation, the forecasts from SSA appear to report a smaller stochastic error than the forecasts from the ETS model (according to the inference in [21]). Once again, relying solely on this empirical c.d.f. in Figure 3 (right) will only result in conclusions which are debatable. The RRMSE criterion shows that forecasts from the SSA model are 54% better than those from the ETS model. The expectation would be that such a significant gain reported through the RRMSE will appear as statistically significant. In order to confirm the expectations we apply both modified DM and KSPA tests.

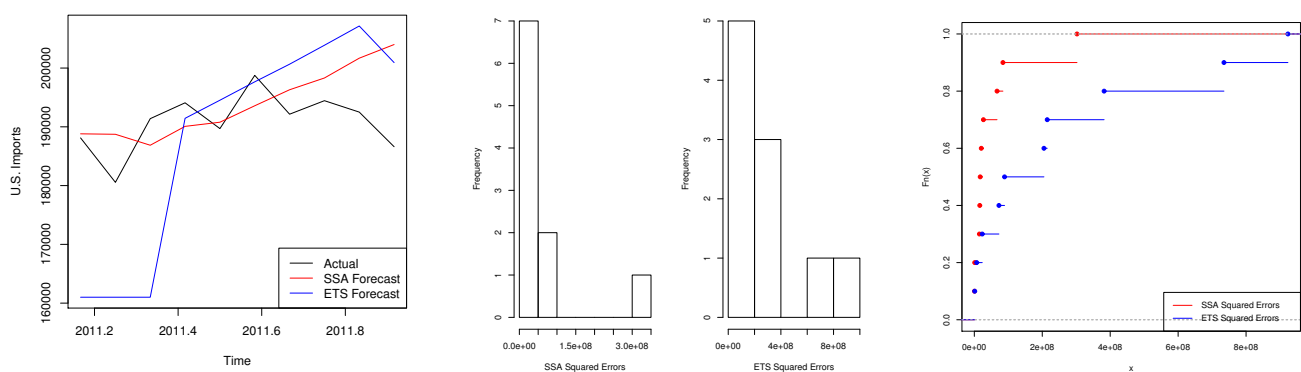


Figure 3. U.S. imports forecast, distribution of errors and empirical c.d.f. of errors.

The results from the two tests are reported in Table 5. Based on the modified DM test we are inclined to conclude that there exists no statistically significant difference between the forecasts obtained via ETS and SSA. The modified DM test statistic reports a p -value which exceeds 10% thus leading to the acceptance of the null hypothesis. Likewise, the two-sided KSPA test suggests that forecast errors obtained via ETS and SSA share an identical distribution. Accordingly, there is no sufficient evidence based on the modified DM test and the two-sided KSPA test for a statistically significant difference between the two forecasts. However, this is where the one-sided KSPA test becomes exceedingly useful. We know based on the RRMSE criterion that forecasts from SSA report a lower RMSE than forecasts from ETS. As such, we can use the one-sided KSPA test to find out whether the SSA forecast which reports a lower error based on the RMSE loss function also reports the lowest stochastic error in comparison to the ETS forecast. We find sufficient evidence based on the one-sided KSPA test at the 10% significance level to conclude that SSA forecasts report a lower stochastic error than forecasts from ETS. Thereby, we are able to confirm the existence of a statistically significant difference between the two forecasts which is expected given that forecasts from SSA are 54% better than forecasts from ETS according to the RRMSE criterion.

Table 5. Evaluating $h = 3$ step ahead forecasts for U.S. imports.

| Test | Two-Sided (p -Value) | Greater (p -Value) |
|-------------|-------------------------|-----------------------|
| Modified DM | 0.30 | N/A |
| KSPA | 0.17 | 0.08 * |

Note: * indicates results are statistically significant based on a p -value of 0.10.

The results from this case not only show the advantage of the one-sided KSPA test, but also proves that the graph of one c.d.f. need not lie “strictly” above the graph of another c.d.f. as suggested in [21] in order for one model to report a stochastically smaller error than the alternate model. It is clear that if a larger proportion of the c.d.f. of errors from one model lies above the c.d.f. of errors from another model, then the KSPA test is able to accurately show that one model reports a lower stochastic error than the other, and thereby pick up a statistically significant difference between the forecasts from two models.

5. Conclusions

Developing on the ideas presented in [21,22] with respect to using an empirical c.d.f. for determining whether the forecast errors from one model are stochastically smaller than those obtained from a competing model, we introduce a complement statistical test for distinguishing between the predictive accuracy of forecasts. The proposed non-parametric Kolmogorov-Smirnov Predictive Accuracy (KSPA) test serves two purposes via the two-sided KSPA test and the one-sided KSPA test. A simulation study is called upon to evaluate the efficiency and robustness of the KSPA test which is followed by an application to real data. The need for the KSPA test is further evidenced by limitations of the DM test in relation to issues in sample size or inherent assumptions which have been left invalidated in the face of recent findings.

Through the simulation study, the KSPA test is directly compared with the widely accepted modified DM test. In order to enable a meaningful comparison, we consider distributions as used in [2] for their simulation study. The simulation results provide a clear indication that the proposed KSPA test is more robust than the DM test especially when the number of out-of-sample forecast errors available for comparison purposes are considerably small.

We also consider applications to real data which capture forecasts from different cases in real world applications for validating the proposed KSPA test, and compare the results against those obtained via the modified DM test. As expected, we observed that when the number of observations are small the KSPA test is able to accurately identify a statistically significant difference between forecasts whilst the modified DM test fails. Furthermore, through another scenario in real world applications we show that the KSPA test can be applied in forecasting exercises where the modified DM test is not applicable. In addition, another scenario is used to show that the two variations of the KSPA test can be extremely useful in practice.

Another advantage in the proposed KSPA test is that given its nature, which is to compare the empirical c.d.f. of errors from two forecasting models, we are able to compare both parametrically estimated model-based forecasts and survey-based forecasts with no restrictions on whether these models are nested or non-nested. This is because, regardless of the model used, a forecast error would always be

calculated as the actual value minus the predicted value, and the proposed KSPA test will compare the distribution of these errors to differentiate between them. In addition, as the KSPA test is non-parametric it is not dependent on any assumptions relating to the properties of the underlying errors which is also advantageous in practice.

In conclusion, the KSPA test has shown promising results in comparison to the modified DM test and is presented as a viable alternative for comparing between the predictive accuracy of forecasts. The non-parametric nature of the test enables one to overcome issues with the assumptions underlying the DM test which have recently been proven void (see for example, [18,19]). Additionally, in the process we have provided statistical validity to the ideas presented in [21,22] whilst showing the relevance and applicability of the KSPA test via simulations and applications to real data. Our research now continues to ascertain whether there is a possibility of extending the use of the KSPA test to enable comparisons between more than two forecasts as this would add more value to its practical use.

Acknowledgments

The authors would like to thank Kerry Patterson and three anonymous referees for their invaluable comments, feedback and suggestions which helped improve the quality of this manuscript. Our heartfelt gratitude to the referees' whose comments were very helpful in contributing to a major improvement in the revised version. In addition, we also wish to thank Mansi Ghodsi (Bournemouth University) for her critical feedback, comments and guidance right throughout the production of this manuscript, and Dick Van Dijk of the Erasmus University (Rotterdam) and Rob Hyndman of Monash University (Australia) for the guidance and comments given for this research during its initial stages.

Author Contributions

The authors contributed equally to this work.

Conflicts of Interest

The authors declare no conflict of interest.

Appendix: R Code for the KSPA Test

The Kolmogorov-Smirnov Predictive Accuracy test

```
# Install and load the "stats" package in R.
```

```
install.packages("stats")
```

```
library(stats)
```

```
# Input the forecast errors from two models. Let Error1 show errors from the model with the lower error based on some loss function.
```

```
Error1<-scan()
Error2<-scan()

# Convert the raw forecast errors into absolute values or squared values
depending on the loss function.

abs1<-abs(Error1)
abs2<-abs(Error2)

sqe1<-(Error1)^2
sqe2<-(Error2)^2

# Perform the KSPA tests for distinguishing between
the predictive accuracy of forecasts from the two models*.

# Two-sided KSPA test:

ks.test(abs1,abs2)

# One-sided KSPA test:

ks.test(abs1,abs2, alternative = c("greater"))

OPTIONAL GRAPHS FOR MORE INFORMATION

# Draw histograms for the forecast errors from each model.

par(mfrow=c(1,2))
hist(abs1, xlab="Model 1 Absolute Errors", main="")
hist(abs2, xlab="Model 2 Absolute Errors",main="")

# Plot the cdf of forecast errors from each model*.

plot(ecdf(abs1),do.points=T,col="red",xlim=range(abs1,abs2),main="")
  plot(ecdf(abs2),do.points=T,col="blue",add=TRUE, main="")
legend("bottomright",legend=c("Model 1 Absolute Errors","Model 2
Absolute Errors"), lty=1, col=c("red","blue"))

#NOTE: *Replace abs1 and abs2 with sqe1 and sqe2 as appropriate.
```

References

1. Elliot, G.; Timmermann, A. *Handbook of Economic Forecasting*; North Holland: Amsterdam, Netherlands, 2013.
2. Harvey, D.I.; Leybourne, S.J.; Newbold, P. Testing the equality of prediction mean squared errors. *Int. J. Forecast.* **1997**, *13*, 281–291.
3. Meese, R.; Rogoff, K. Was it real? The exchange rate-interest rate differential relation over the modern floating-rate period. *J. Finance* **1988**, *43*, 933–948.
4. Christiano, L.J. p^* : Not the inflation forecasters holy grail. *Fed. Reserve Bank Minneap. Q. Rev.* **1989**, *13*, 3–18.
5. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263.
6. Diebold, F.X. *Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests*; Department of Economics, University of Pennsylvania: Philadelphia, PA, USA, 2013; pp. 1–22.
7. Hansen, P.R. A test for superior predictive ability. *J. Bus. Econ. Stat.* **2005**, *23*, 365–380.
8. Hansen, P.R.; Lunde, A.; Nason, J.M. Model confidence set. *Econometrica* **2011**, *79*, 453–497.
9. Clark, T.E.; McCracken, M.W. In-Sample Tests of Predictive Ability: A New Approach. *J. Econom.* **2012**, *170*, 1–14.
10. Clark, T.E.; McCracken, M.W. Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy. *J. Econom.* **2015**, *186*, 160–177.
11. Gilleland, E.; Roux, G. A new approach to testing forecast predictive accuracy. *Meteorol. Appl.* **2015**, *22*, 534–543.
12. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378.
13. Silva, E.S.; Hassani, H. On the use of Singular Spectrum Analysis for Forecasting U.S. Trade before, during and after the 2008 Recession. *Int. Econ.* **2015**, *141*, 34–49.
14. Hassani, H.; Webster, A.; Silva, E.S.; Heravi, S. Forecasting U.S. Tourist arrivals using optimal Singular Spectrum Analysis. *Tour. Manag.* **2015**, *46*, 322–335.
15. Hassani, H.; Silva, E.S.; Gupta, R.; Segnon, M.K. Forecasting the price of gold. *Appl. Econ.* **2015**, *47*, 4141–4152.
16. Granger, C.W.J.; Newbold, P. *Forecasting Economic Time Series*; Academic Press: New York, NY, USA, 1977.
17. Morgan, W.A. A test for significance of the difference between two variances in a sample from a normal bivariate population. *Biometrika* **1939**, *31*, 13–19.
18. Hassani, H. A note on the sum of the sample autocorrelation function. *Phys. A Stat. Mech. Its Appl.* **2010**, *389*, 1601–1606.
19. Hassani, H.; Leonenko, N.; Patterson, K. The sample autocorrelation function and the detection of Long-Memory processes. *Phys. A Stat. Mech. Its Appl.* **2012**, *391*, 6367–6379.
20. Kolmogorov, A.N. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituto Italiano degli Attuari* **1933**, *4*, 1–11.

21. Hassani, H.; Heravi, S.; Zhigljavsky, A. Forecasting European industrial production with Singular Spectrum Analysis. *Int. J. Forecast.* **2009**, *25*, 103–118.
22. Hassani, H.; Heravi, S.; Zhigljavsky, A. Forecasting UK industrial production with multivariate Singular Spectrum Analysis. *J. Forecast.* **2013**, *32*, 395–408.
23. Horváth, L.; Kokoszka, P.; Zitikis, R. Testing for stochastic dominance using the weighted McFadden-type statistic. *J. Econom.* **2006**, *133*, 191–205.
24. McFadden, D. Testing for Stochastic Dominance. In *Studies in the Economics of Uncertainty: In Honor of Josef Hadar*; Fomby, T.B., Seo, T.K., Eds.; Springer: New York, NY, USA; Berlin, Germany; London, UK; Tokyo, Japan, 1989.
25. Barrett, G.F.; Donald, S.G. Consistent tests for stochastic dominance. *Econometrica* **2003**, *71*, 71–104.
26. DeGroot, M.H.; Schervish, M.J. *Probability and Statistics*, 4th ed.; Addison-Wesley: Boston, MA, US, 2012.
27. Marsaglia, G.; Tsang, W.W.; Wang, J. Evaluating Kolmogorov's distribution. *J. Stat. Softw.* **2003**, *8*, 1–4.
28. Birnbaum, Z.W.; Tingey, F.H. One-sided confidence contours for probability distribution functions. *Ann. Math. Stat.* **1951**, *22*, 592–596.
29. Simard, R.; L'Ecuyer, P. Computing the Two-Sided Kolmogorov-Smirnov Distribution. *J. Stat. Softw.* **2011**, *39*, 1–18.
30. Brockwell, P.J.; Davis, R.A. *Introduction to Time Series and Forecasting*; Springer: New York, NY, US, 2002.
31. Hassani, H. Singular Spectrum Analysis: Methodology and Comparison. *J. Data Sci.* **2007**, *5*, 239–257.
32. Hassani, H.; Mahmoudvand, R.; Omer, H.N.; Silva, E.S. A Preliminary Investigation into the Effect of Outlier(s) on Singular Spectrum Analysis. *Fluct. Noise Lett.* **2014**, *13*, doi:10.1142/S0219477514500291.
33. Sanei, S.; Hassani, H. *Singular Spectrum Analysis of Biomedical Signals*; CRC Press: Boca Raton, FL, US, 2015.