A Comparative Study of the KPSS and ADF Tests in terms of Size and Power

Lina Sjösten

Bachelor’s thesis in Statistics

Advisor

Yukai Yang

2022
Abstract

This thesis investigates through simulation why tests of unit root and stationarity occasionally result in different conclusions. The thesis focusses on the KPSS test and the ADF test and both review cases with and without a trend. The goal is to bring additional knowledge of whether one of the tests are more reliable in terms of size and power and when contradictory results occur. The result shows that both KPSS and ADF suffer from low power and size distortion and that the problems persists for the most common time series lengths. Problems particularly arise when the time series contains a trend or is a process with both an autoregressive and a moving average part. There is no clear evidence that one of the tests are superior to the other, it rather depends on sample size, parameter value and type of ARIMA process.

Keywords: Stationarity, unit root, time series, simulation, trend, ARIMA, parameter value, sample size.
# Contents

1. **Introduction** ................................................................. 1  

2. **Theoretical Background** .................................................. 3  
   2.1 ARIMA ............................................................................ 3  
   2.2 Augmented Dickey-Fuller Test (ADF) ................................ 4  
   2.3 Kwiatkowski Phillips Schmidt and Shin Test (KPSS) .......... 5  
   2.4 Beveridge-Nelson Decomposition ..................................... 7  
   2.5 Size Adjusted Power ....................................................... 10  

3. **Simulation Study** ............................................................. 11  
   3.1 Data Generation ............................................................ 12  
   3.2 Test Design ................................................................. 13  
   3.3 Evaluation ...................................................................... 14  

4. **Result** ............................................................................. 15  
   4.1 Size .............................................................................. 15  
   4.2 Power ........................................................................... 20  

5. **Discussion** .................................................................... 24  

6. **Conclusion** .................................................................... 26  

**References** ........................................................................ 28  

**Appendix A** ....................................................................... 30
### Notation and Symbols

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADF</td>
<td>Augmented Dickey-Fuller test</td>
</tr>
<tr>
<td>AR</td>
<td>Autoregressive process</td>
</tr>
<tr>
<td>ARMA</td>
<td>Autoregressive moving average process</td>
</tr>
<tr>
<td>KPSS</td>
<td>Kwiatkowski Phillips Schmidt and Shin test</td>
</tr>
<tr>
<td>MA</td>
<td>Moving average process</td>
</tr>
<tr>
<td>NID</td>
<td>Normally and independently distributed</td>
</tr>
<tr>
<td>iid</td>
<td>Independent and identically distributed</td>
</tr>
<tr>
<td>$Y$</td>
<td>Dependent variable</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Autoregressive parameter</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Moving average parameter</td>
</tr>
<tr>
<td>$e$</td>
<td>White noise</td>
</tr>
<tr>
<td>$L$</td>
<td>Lag polynomial</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Mean</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>Difference operator</td>
</tr>
<tr>
<td>$\delta t$</td>
<td>Linear trend</td>
</tr>
<tr>
<td>$S$</td>
<td>Sum of residuals</td>
</tr>
<tr>
<td>$T$</td>
<td>Number of observations</td>
</tr>
<tr>
<td>$u$</td>
<td>Stationary ARMA process</td>
</tr>
<tr>
<td>$\eta_t$</td>
<td>Stationary process</td>
</tr>
<tr>
<td>$Y_0 - \eta_0$</td>
<td>Initial condition</td>
</tr>
<tr>
<td>$\psi(1) \sum_{s=1}^t e_s$</td>
<td>Stochastic trend</td>
</tr>
</tbody>
</table>
1. Introduction

Stationarity is one of the most important concepts and assumptions in time series analysis. Stationary in this case is referred to covariance-stationary which means that the time series statistical properties are independent of time (Hamilton, 1994). Because of this it is possible to do analysis of the time series that is trustworthy since the behavior of the time series is known and does not change over time. If a process instead is nonstationary, this will affect the possibility to make inference on the time series without making any transformation of the data (Cryer and Chan, 2010). However, if an analysis still is performed on nonstationary time series there are potential consequences involving unreliable predictions and test statistics that follow another distribution then assumed (Levendis, 2018). To correctly recognize the order of integration is therefore of high importance.

To investigate if a time series is stationary or not there are numerous tests that can be used alongside the visual inspection of correlograms (Maddala and Kim, 1998). It has been showed that these tests at times disagree and result in different conclusions about the order of integration, meaning wrong conclusions sometimes inevitably are drawn (Caner and Kilian, 2001; Stock, 1994). No statistical test will always be correct, but the question is if the magnitude of size and power is problematic.

Previous studies have shown that many unit root tests and tests of stationarity suffer from low power and size distortion. Schwert (1989) investigated the effects of model misspecification on the size of a variety of unit root tests. He raised the problem of severe size distortion when the process is of another type than an autoregressive of order one. While DeJong et al. (1992) criticized both unit root tests and tests of stationarity for having low power after comparing the two types of tests. Even though these problems were discovered several years ago, researchers continue to use these tests in lack of better alternatives. Many tests have been suggested and some have tried to improve existing ones but there is no agreement in just one test that is superior (Stock, 1994). A couple of studies have therefore tried to give recommendations on which test to use depending on various factors. One example is Arltová and Fedorová (2016) who conducted a simulation study with the goal to give recommendations based on the length of the time series and the parameter value. They concluded that the KPSS test performs well for very small values of the autoregressive parameter whereas the ADF test is especially good
when the time series has many observations. A limitation with this study is that they only consider an autoregressive process of order one as well as only positive parameter values.

Due to the absence of an overall good test of unit root and stationarity the existing tests need to be further analyzed to bring knowledge about the problems surrounding them and their magnitude. Previous studies have not been comprehensive and that more research around this topic is needed has been pointed out by numerous researchers (Arltová and Fedorová, 2016; Caner and Kilian, 2001). To get the full picture of these tests, negative parameter values and combined autoregressive moving average processes needs to be studied as well.

The purpose of this thesis is therefore to compare two tests that can be used for evaluating if a time series is stationary or not and to bring additional knowledge about them by also studying models with both an autoregressive and a moving average part as well as negative parameter values. The tests that are being compared are Augmented Dickey-Fuller test (ADF) and Kwiatkowski Phillips Schmidt and Shin test (KPSS). These are chosen due to their popularity in applied research. Since it has been showed that these tests can result in different conclusions, this thesis will evaluate by simulation why this is the case and if one of them are more reliable. The research questions that will be answered are:

- Why does Augmented Dickey-Fuller test (ADF) and Kwiatkowski Phillips Schmidt and Shin test (KPSS) sometimes result in contradictory conclusions?
- When does Augmented Dickey-Fuller test (ADF) and Kwiatkowski Phillips Schmidt and Shin test (KPSS) result in contradictory conclusions?
- Are one of Augmented Dickey-Fuller test (ADF) and Kwiatkowski Phillips Schmidt and Shin test (KPSS) more reliable?

The remainder of this thesis is outlined as follows. In Section 2 are ARIMA models, KPSS and ADF test, Beveridge-Nelson decomposition and size adjusted power described. Section 3 presents the data generation, test design and evaluation of the simulation study. The results are reviewed in Section 4 and are thereafter discussed in Section 5. The thesis finishes with a conclusion of the study in Section 6.
2. Theoretical Background

To fully grasp the results provided by this thesis some concepts need to be further explained. One of them are ARIMA models which are the type of models considered in the simulation study. Also, the two tests used, KPSS and ADF, need a deeper explanation as well as the Beveridge-Nelson decomposition. The reasons and theory behind the need to size adjust power when evaluating hypothesis tests are also important to clarify.

2.1 ARIMA

One type of univariate time series models is ARIMA\( (p,d,q) \) models which stands for autoregressive integrated moving average models. The letters \( p \) and \( q \) express the order of the autoregressive and moving average part respectively. The order of integration, \( d \), represents the number of unit roots in the stochastic process. So, when there is no deterministic trend, the stochastic trend can be removed by differentiating the time series \( d \) times. A time series integrated of order zero is a stationary process and can simply be written as an ARMA\( (p,q) \) process (Cryer and Chan, 2010).

A process that only consist of the autoregressive part can be expressed as AR\( (p) \) and written as

\[
Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t. \tag{1}
\]

This is a process where the current value of the time series is a linear combination of previous values of the time series and a white noise term, \( e_t \). The white noise term is a random shock that consists of everything that is not explained by the past values of the time series (Cryer and Chan, 2010). White noise further means that the process has a zero mean, a constant variance and is uncorrelated across time (Hamilton, 1994).

If a process only consists of a moving average part, it can be expressed as MA\( (q) \). A MA process of order \( q \) can then be written as

\[
Y_t = e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}. \tag{2}
\]

This is a process where the current value of the time series depends on a weighted sum of a white noise process (Hamilton, 1994).
The process can also be a combination of an autoregressive process and a moving average process, ARMA\((p,q)\), and is then expressed as

\[
Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}.
\] (3)

This is a time series where the current value of the time series depends on both a number of previous values and on a weighted sum of a white noise process. This can also be written in lag polynomials,

\[
\phi(L)Y_t = \theta(L)e_t.
\] (4)

The first part, \(\phi(L)\), is the one that is checked for a unit root (Hamilton, 1994). In the next two sections two test that can be used for testing whether the time series contains a unit root is presented.

2.2 Augmented Dickey-Fuller Test (ADF)

The Augmented Dickey-Fuller test (ADF) was created by Dickey and Fuller (1981) and is one of many unit root tests. The authors derived the test from considering a higher order auto-regressive process, see equation 1. The white noise term, \(e_t\), has a zero mean and variance \(\sigma_e^2\) whereas \(\phi_i\) are fixed coefficients. By using the lag operator, it is possible to write the process as

\[
\phi(L)Y_t = e_t,
\]

where

\[
\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p.
\]

If the polynomial

\[
\phi(1) = 1 - \phi_1 - \cdots - \phi_p = 0,
\]

then the process has a unit root. Therefore, the hypothesis of interest is if \(\phi(1) = 0\). To be able to test this the following equation is used and estimated by OLS (Lütkepohl et al., 2004)

\[
\Delta Y_t = a Y_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta Y_{t-i} + e_t.
\] (5)
The equation can also include a linear trend which then looks like

\[ \Delta Y_t = \delta + aY_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta Y_{t-i} + e_t, \quad (6) \]

where \( \sum_{i=1}^{p-1} \phi_i \Delta Y_{t-i} + e_t \) is a stationary progress and \( e_t \) are NID(0,\( \sigma^2 \)) (Dickey and Fuller, 1981). Further, \( a = -\phi(1) \) and \( \phi_i = -\left(\phi_{i+1} + \cdots + \phi_p\right) \). This results in the following hypotheses for the ADF test (Lütkepohl et al., 2004):

- \( H_0 \): \( a = 0 \) meaning the time series contains a unit root
- \( H_1 \): \( a < 0 \) meaning the time series does not contain a unit root

The test follows an asymmetric t-distribution where most of the values will be negative (Maddala and Kim, 1998). After the t-value is calculated it is compared to the simulated critical values provided by Fuller (1976). The null hypothesis is rejected for small values which can be seen from the hypotheses where the alternative is that \( a \) is less than zero.

For a process where \( p \) is one the Dickey-Fuller test is a special case of the Augmented Dickey-Fuller test. Since it is only of order one the lagged differenced terms of previous periods in equation 5 and 6, \( \sum_{i=1}^{p-1} \phi_i \Delta Y_{t-i} \), will disappear. The equations then result in

\[ \Delta Y_t = aY_{t-1} + e_t \]

and

\[ \Delta Y_t = \delta + aY_{t-1} + e_t. \]

These are the regression models suggested by Dickey and Fuller (1979) for the Dickey-Fuller test.

### 2.3 Kwiatkowski Phillips Schmidt and Shin Test (KPSS)

The Kwiatkowski Phillips Schmidt and Shin test (KPSS) was first founded by Kwiatkowski, Phillips, Schmidt and Shin (1992). This in contrary to the ADF test is a test of stationarity which was first thought of as a complement to tests of unit root. The test assumes that the time series can be divided into a deterministic trend, a random walk and a stationary error. This means that
the time series can be illustrated in the following way

\[ Y_t = \delta t + r_t + e_t, \tag{7} \]

where \( \delta t \) is the deterministic trend, \( r_t \) is the random walk and \( e_t \) is the stationary error with zero mean. The random walk can be displayed as

\[ r_t = r_{t-1} + v_t, \]

where \( v_t \) is iid\((0, \sigma_v^2)\) and \( r_0 \) is the equations intercept. The null hypothesis implies that \( \sigma_v^2 = 0 \). This results in the time series, \( Y_t \), being trend stationary. The KPSS then tests if there is a unit root in \( r_t \) when \( \delta \) is nonzero. The hypotheses for the test is then:

\[ H_0: \text{The time series is trend stationary} \]
\[ H_1: \text{The time series contains a unit root} \]

The KPSS test can also be used for testing for level stationary. In this case \( \delta = 0 \) in equation 7 and the hypotheses changes to:

\[ H_0: \text{The time series is level stationary} \]
\[ H_1: \text{The time series contains a unit root} \]

The test statistic that the test uses is a one-sided Lagrange Multiplier and is given by

\[ LM = \frac{\sum_{t=1}^{T} S_t^2}{\hat{\sigma}_e^2}, \tag{8} \]

where \( S_t \) is the sum of the residuals and \( \hat{\sigma}_e^2 \) is the estimated error variance from equation 7. This is for the trend stationary hypothesis. For the level stationary null hypothesis, the only difference is that the residuals are from the equation excluding the time trend. Otherwise, the test statistic is the same. The test statistic is derived under the assumption that the error of the time series, \( e_t \), is iid which is typically not met in reality since the time series often are dependent over time. Therefore, the following test statistic for the trend stationary (\( \tau \)) and level stationary (\( l \)) null hypothesis is used instead.

\[ \eta_{\tau/l} = T^{-2} \frac{\sum_{t=1}^{T} S_t^2}{s^2(l)} \tag{9} \]

The only difference is again that the residuals is based on equation 7 for the trend stationary hypothesis versus the same equation excluding the time trend for the level stationary hypothesis.
In equation 9, \( s^2(l) \) is a consistent estimator of the long-run variance of the errors, \( \sigma^2 \). The estimator is created from the residuals \( e_t \). This is an appropriate denominator of the equation instead of using an estimate of \( \sigma^2_e \) since the errors are not iid which is used in equation 8. The equation is normalized by \( T^{-2} \) where \( T \) is the number of observations (Kwiatkowski et al., 1992). The null hypothesis is rejected for large values of the test statistic (Müller, 2005).

### 2.4 Beveridge-Nelson Decomposition

KPSS and ADF tests the same thing but with different hypotheses meaning they do the same thing but in different ways. This makes it possible to compare the two tests. To help understand the hypotheses of the tests and the underlying process the Beveridge-Nelson decomposition proposed by Beveridge and Nelson (1981) can be used. Hamilton (1994) explained it as when a process follows a random walk without drift given by

\[
Y_t = Y_{t-1} + u_t,
\]

where the first difference fulfills

\[
u_t = \psi(L)e_t = \sum_{j=0}^{\infty} \psi_j e_{t-j}\]

and

\[
E(e_t) = 0
\]

\[
E(e_t e_{\tau}) = \begin{cases} 
\sigma^2 & \text{for } t = \tau \\
0 & \text{otherwise}
\end{cases}
\]

\[
\sum_{j=0}^{\infty} |\psi_j| < \infty
\]

the process can be written as the sum of a stochastic trend, initial condition, a linear trend and a stationary process. The initial condition can be seen as the equations intercept. This is what is called the Beveridge-Nelson decomposition. To further illustrate this, consider a process integrated of order one given by

\[
\Delta Y_t = \delta + u_t,
\]

\( (10) \)
where \( u_t = \psi(L)e_t \) and \( \delta = E(\Delta Y_t) \). The connection between the lag polynomial for the ARMA\((p,q)\) in equation 4 and equation 10 can be illustrated in the following way by assuming non-zero unconditional expectation and trend in equation 4.

\[
\phi(L)(Y_t - \mu - \delta t) = \theta(L)e_t
\]

If there is a unit root, \( \phi(L) \) can be written as

\[
(1 - L)\phi^*(L),
\]

where \( (1 - L) \) is the difference operator \((\Delta)\) and \( \phi^*(L) \) is the lag polynomial without unit root which means that this part can be inverted. Then equation 4 becomes

\[
(1 - L)(Y_t - \mu - \delta t) = (\phi^*)^{-1}(L)\theta(L)e_t = \psi(L)e_t = u_t.
\]

By applying the difference operator, the result is

\[
(1 - L)Y_t = \delta + u_t,
\]

since \((1 - L)\mu = 0\) and \((1 - L)\delta t = \delta\). This shows that equation 4 and 10 are equivalent.

Returning to equation 10 and substitution \( u_t = \psi(L)e_t \) the equation becomes

\[
\Delta Y_t = \delta + \psi(L)e_t.
\]  \( \text{(11)} \)

\( u_t \) represents the stationary ARMA process where \( e_t \) is white noise. By writing equation 10 without the difference operator and using substitution the result is equation 12.

\[
Y_t = Y_{t-1} + \delta + u_t
\]

\[
Y_t = Y_{t-2} + 2\delta + u_t + u_{t-1}
\]

\[
Y_t = Y_{t-3} + 3\delta + u_t + u_{t-1} + u_{t-2}
\]

\[
\vdots
\]

\[
Y_t = Y_0 + \delta t + \sum_{s=1}^{t} u_s
\]  \( \text{(12)} \)

By summing the stationary ARMA processes the result, \( \sum_{s=1}^{t} u_s \), is a random walk (Hamilton, 1994).
The first component of the stationary process in equation 1 can be written as \( \psi(L) = \psi(1) + \Delta \alpha(L) \) where \( \psi(1) = \sum_{j=0}^{\infty} \psi_j \) and \( \alpha(L) = \sum_{j=0}^{\infty} \alpha_j L^j \). Further, \( \alpha_j = -(\psi_{j+1} + \psi_{j+2} + \cdots) \) for \( j = 0, 1, \ldots \). By using this it is possible to write

\[
u_t = \psi(1) e_t + \eta_t - \eta_{t-1}, \tag{13}\]

where \( \eta_t = \alpha(L) e_t \) is a zero-mean process integrated of order zero. By substituting equation 13 into equation 12 the result is

\[
Y_t = \delta t + \psi(1) \sum_{s=1}^{t} e_s + \eta_t + Y_0 - \eta_0. \tag{14}\]

Which is the equation for the Beveridge-Nelson decomposition. The equation consists of \( \delta t \) which is the linear trend, a stochastic trend \( \psi(1) \sum_{s=1}^{t} e_s \), a stationary process \( \eta_t \) and an initial condition \( Y_0 - \eta_0 \) (Hamilton, 1994). To see that the equation for the ADF test can be written in this way consider equation 6 which is the equation with a linear trend. By comparing it to the Beveridge-Nelson decomposition in equation 14 it becomes clear that \( \delta \) represents the linear trend since by removing the difference operator and using substitution this will become \( \delta t \). Whereas \( \alpha Y_{t-1} \) is the stochastic trend since it is assumed that the process is integrated of order one. \( \sum_{i=1}^{h} \phi_i \Delta Y_{t-i} + e_t \) is the stationary process. In this case the initial condition is zero.

As stated before, the KPSS assumes that the process can be divided into a deterministic trend, a random walk and a stationary error (Kwiatkowski et al., 1992). Where the intercept is in the random walk component. This makes it clear that the process in equation 7 can be written in the same way as proposed by the Beveridge-Nelson decomposition. This is therefore proof that when the process contains a unit root, the equations for the KPSS and ADF can be decomposed into the same parts. If the process is instead stationary, \( \psi(1) \sum_{s=1}^{t} e_s \) is zero in equation 14. For the ADF test \( a \) is less than one and \( \alpha Y_{t-1} \) is no longer a stochastic trend but a part of the stationary process. For the KPSS test the random walk component is instead a stationary component. This inevitably means that the two tests can be decomposed into the same parts regardless of whether the process has a unit root or is a stationary process.
Figure 1. Illustration of the Beveridge-Nelson decomposition. The blue dotted line represents the endpoints of the confidence interval for future predicted values when the process is stationary. The red dotted line represents the endpoints of the confidence interval for future predicted values when the process is nonstationary. The angle between the horizontal black dotted line and the solid black line is the arctangent of $\delta$.

The Beveridge-Nelson decomposition can also be used to illustrate why an uncorrected nonstationary time series is a problem when doing predictions. This origin from the fact that the variance of a nonstationary time series grows linearly with time (Box-Steffensmeier et al., 2014). This is illustrated in Figure 1 and is due to the stochastic trend in equation 14. Even though the process is nonstationary it may be possible to make an accurate prediction for one time period ahead. The problem particularly arises when predictions is made for more than one time period ahead since the variance contribute to making the confidence interval for the predictions very wide. For the stationary case the variance is constant and independent of time which means that the confidence interval will have the same width for all time periods ahead (Hamilton, 1994). This is instead due to the stationary process in equation 14 since this assures that the variance is constant.

2.5 Size Adjusted Power

When evaluating the performance of hypothesis tests, size and power properties are usually considered (Lloyd, 2005). This is because of its close connection to the two possible errors of hypothesis testing. The first error is Type I error which is the probability to reject the null hypothesis when it is true. This is the same as the size of the test. The other type of error is Type
II error which is defined as: *Type II error = 1 - Power*. Type II error is the probability to not reject the null hypothesis when it is false which then evidently means that the power is the probability to reject the null hypothesis when it is false (Devore et al., 2021). Therefore, it is desirable that a hypothesis test has high power and low size.

One important aspect to be aware of when evaluating the power and size of a test is that there is a trade-off between size and power where the power can be increased by lowering the critical value. Though, this will result in an increase in size. So, if one test has higher actual size than the other test the power will also be higher in the first test which makes it hard to know if the higher power is a result of the higher size or if it actually has higher power. It is therefore important that the actual size of the test is the same when comparing the power of the tests, preferably the same as the significance level. The problem when the actual size differs from the nominal size is called size distortion (Lloyd, 2005). The reason why this problem arise is because of the critical values for the ADF test and KPSS test being valid asymptotically (Dickey and Fuller, 1979; Kwiatkowski et al., 1992). Meaning that in finite samples the critical values for the size to be equal to the significance level might be different from when the sample size is infinite. By adjusting the critical value so the test rejects the null hypothesis by a rate equal to the chosen significance level, it is possible to make a fair comparison of hypothesis tests (Lloyd, 2005).

### 3. Simulation Study

To evaluate the two tests, KPSS and ADF, a Monte Carlo simulation study is conducted in R Studio¹. A Monte Carlo simulation can briefly be explained as a method for generating a large number of random samples from a mathematical model. The key concept is that the procedure is iterated multiple times which result in outcomes that are purely random making it possible to generalize the results (Thomopoulos, 2013). This is also the reasons why a simulation study is performed and not for example real data are considered. It may not be possible to replicate real data which would make it hard to generalize the result and draw accurate conclusions. There would also be a fundamental problem when evaluating if the test is correct or not since there would take a lot of effort to make sure that the true order of integration is known in advance.

---

¹ R code is available upon request.
3.1 Data Generation

The simulation study is based on four different time series processes: stationary with trend, stationary without trend, nonstationary with trend and nonstationary without trend. Nonstationary in this case refers to a unit root process. The trend component is a linear time trend which means that the process is trend stationary for the case where the process is stationary with trend. Autoregressive, moving average and combined autoregressive moving average processes are all reviewed with the limitation of only first order processes. In addition, both parameter value and sample size are varied. The parameter values that are used are 0.3, 0.7, 0.9, -0.3, -0.7, -0.9. They are chosen with the aim of having a wide variety of values which are both high and low as well as for complementing previous studies by using negative parameter values. A value close to one means the time series is more persistent compared to a time series with smaller parameter values. To be able to generalize the results it is therefore important to have parameter values that is both high and low. For the combined autoregressive moving average process the parameter values for the autoregressive part and the moving average part are restricted to be the same since the simulation would be too comprehensive otherwise. This means that the first model for example will be a model where $\phi = 0.3$ and $\theta = 0.3$ resulting in the following model

$$Y_t = 0.3Y_{t-1} + e_t + 0.3e_{t-1}.$$ 

The sample sizes are chosen to reflect the common lengths of real time series and are therefore set to be 20, 50, 100 and 500. All model combinations are replicated 7 500 times which according to Mundfrom et. al. (2011) are enough to produce stable results in Monte Carlo studies. The reason for having different sample sizes and parameter values is to see if there is any difference among them as well as for making the study more applicable to real time series settings. This is also the reason for not only considering one type of process but instead simulate all three; autoregressive, moving average and combined autoregressive moving average processes.

The simulation of all four cases is based on equation 10 from the Beveridge-Nelson decomposition presented in Section 2.4. To simulate the different ARIMA models the function arima.sim from the package stats is used. The part produced by arima.sim is represented by $u_t$ in equation 10. For the nonstationary cases the unit root component is located in the left-hand side of equation 10 which means that all the arima.sim functions are specified as integrated of
order zero regardless of whether the process is stationary or not. For the cases containing a linear trend component the value of \( \delta \) is set to be 0.5.

### 3.2 Test Design

The two tests are designed as follows. The function kpss.test from the package tseries is used to perform the KPSS test. When the simulated time series has a trend, the null hypothesis is set to be trend stationary whereas when there is no trend the null hypothesis is set to be level stationary. This means that the function will estimate a regression as in equation 12 either with or without a trend. To perform the ADF test the function ur.df from the package urca is used. In this function the test type is set to be drift when the time series has a trend since this means that the test regression will include an intercept. In the case where there is no trend the test type is set to be none since this means that the test regression does not have an intercept. This intuitively comes from the fact that the test estimates a test regression in the same form as equation 10 and not as the one in equation 12 as the KPSS test does. By considering equation 6 for the ADF test it becomes clear that this is in the same form as equation 10 where the difference operator is on the left-hand side and that the trend component is just an intercept before removing the difference operator using substitution. This is the reason for the function ur.df to include an intercept when the time series has a trend. Independent of whether the time series has a trend or not the lag selection is set to be AIC. The critical values used for the ADF and KPSS tests are presented in Table 1.

The input for both tests are \( Y_t \) in equation 12. The equation differs depending on which of the four time series cases considered. For the case where the time series is nonstationary with trend the equation which the tests are applied to are exactly the one in equation 12. For the case where

<table>
<thead>
<tr>
<th>Sample size</th>
<th>KPSS With trend</th>
<th>KPSS Without trend</th>
<th>ADF With trend</th>
<th>ADF Without trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.146</td>
<td>0.463</td>
<td>-3.000</td>
<td>-1.950</td>
</tr>
<tr>
<td>50</td>
<td>0.146</td>
<td>0.463</td>
<td>-2.930</td>
<td>-1.950</td>
</tr>
<tr>
<td>100</td>
<td>0.146</td>
<td>0.463</td>
<td>-2.890</td>
<td>-1.950</td>
</tr>
<tr>
<td>500</td>
<td>0.146</td>
<td>0.463</td>
<td>-2.870</td>
<td>-1.950</td>
</tr>
</tbody>
</table>
the time series is nonstationary without trend the value of $\delta$ is zero since there is no trend. Equation 12 are then $Y_t = Y_0 + \sum_{s=1}^{t} u_s$. The third case is where the time series is stationary with trend. This means that equation 12 reduces to $Y_t = \delta t + u_t$ instead since this process does not contain a unit root. For the case where the time series is stationary without trend the trend component is zero so equation 12 is just the stationary ARMA process, $Y_t = u_t$.

### 3.3 Evaluation

To evaluate how well the KPSS and ADF test performs the size and power properties of the tests are studied. Since the two tests have different null hypothesis, it is misleading to compare each of the four cases respectively. This is because the size property can be evaluated by simulating cases where the null hypothesis is true and then investigate the rejection rate. Whereas the power property can be investigated by simulating cases where the null hypothesis is false and then evaluate the rejection rate (Martin et al., 2013). This evidently means that the null hypothesis needs to be the same for the tests which is not fulfilled for the KPSS and ADF test. Therefore, the cases where the null hypothesis is true, nonstationary for the ADF test and stationary for the KPSS test, are compared when evaluating size. Whereas stationary for the ADF test and nonstationary for the KPSS are compared when investigating power. The degree of reliability among the tests is evaluated by considering a more reliable test as the one who has higher power and lower size.

To get the size adjusted power, the Monte Carlo simulation conducted for the size experiment is used. For the KPSS test the two cases of stationarity is considered and for the ADF test the two cases of nonstationarity is considered. For each sample size, parameter value, process and process per sample size the calculated values for the test statistic are sorted from smallest to largest. Thereafter the value where exactly five percent of the values are smaller, the fifth percentile, are chosen to be the critical value for the ADF test since it rejects for small values. This is where the rejection frequency of the test are five percent. Whereas the value where exactly five percent of the values are larger, the 95th percentile, are to be the critical value for the KPSS test since this test instead rejects for large values. The rejection frequency is then evaluate based on these critical values which gives the size adjusted power of the test.
4. Result

The results are reported for size and power separately and visualized per sample size, parameter value and process for all four types of time series. In addition, the size and power per sample size are also divided by process and then graphically displayed for each of the four types of time series separately. The power presented is the size adjusted power where the size is adjusted to be five percent. The critical values used are presented in Table 2-8 in Appendix A.

4.1 Size

Figure 2 represents the size evaluated per sample size. All four cases have an actual size higher than the nominal size of five percent. Both cases for the ADF test have a higher size than the cases for the KPSS test. The actual size closest to the nominal size is when the sample size is 20 independent of which case is considered. For the other sample sizes the size are rather constant except for the ADF where the true process is nonstationary without trend. Here the size slightly increases as the sample size grows.

Figure 2. Size per sample size. The black dotted line represents the significance level of five percent.
Figure 3. Size per parameter value. The black dotted line represents the significance level of five percent.

Figure 3 illustrates the size for the different parameter values used in the simulation. The performance of the two tests is rather different besides when the parameter value is 0.3 or -0.3. Then the size is around the nominal level for both tests. The ADF has a very large size when the parameter value is -0.7 or -0.9. When the process is nonstationary with trend the size for the ADF is more than 60 percent. Whereas KPSS has problems recognizing the process is stationary when the parameter value is 0.9. It also has a somewhat higher size for a parameter value of 0.7.

The size for the three different processes: autoregressive, moving average and autoregressive moving average is presented in Figure 4. The size for the ADF test is almost equal to the nominal size when the process is autoregressive regardless of whether the time series exhibits a trend or not. The test has a much higher size when the process is a moving average or an autoregressive moving average. For the KPSS test the process that has an actual size close to the significance level is instead the moving average. The size is somewhat higher when the process is an autoregressive or autoregressive moving average but the KPSS overall has a much more constant size compared to the ADF.
Figure 4. Size per process. The black dotted line represents the significance level of five percent.

Figure 5 displays the size for the ADF test per sample size and process when the true time series is nonstationary with trend. The size is close to the significance level when the process is autoregressive but much higher when the process is either a moving average or autoregressive moving average. The size is closest to the nominal when the sample size is 20 for the moving average and autoregressive moving average. For the autoregressive process the size is rather constant around five percent.

Figure 5. Size per sample size and process for nonstationary time series with trend for the ADF test. The black dotted line represents the significance level of five percent.
Figure 6. Size per sample size and process for nonstationary time series without trend for the ADF test. The black dotted line represents the significance level of five percent.

The size for the ADF test per sample size and process when the time series is nonstationary without trend is illustrated in Figure 6. As for the case in Figure 5, when the time series exhibits a trend, the size for the autoregressive process is close to five percent. Whereas the size for the moving average and autoregressive moving average processes are well above five percent. They both exhibit an increase in size as the sample size becomes larger and have the smallest size when the sample size is 20.

Figure 7 displays the size for the KPSS test per sample size and process when the time series is stationary with trend. The size when the process is a moving average is somewhat below five percent independent of sample size. Instead looking at a process that is an autoregressive or autoregressive moving average the size is above five percent. Though, it is closest to five percent when the sample size is 20. Thereafter the size increase to a level around 15 percent for all other sample sizes considered.
Figure 7. Size per sample size and process for stationary time series with trend for the KPSS test. The black dotted line represents the significance level of five percent.

Figure 8 represents the size for the KPSS test per sample size and process when the true time series is stationary without trend. As when the time series exhibits a trend the KPSS has a size closest to the nominal size when the process is a moving average. For the four sample sizes considered, the size is rather constant around 12 percent when the process is an autoregressive or autoregressive moving average.

Figure 8. Size per sample size and process for stationary time series without trend for the KPSS test. The black dotted line represents the significance level of five percent.
4.2 Power

Figure 9 displays the power for each sample size. There is a clear pattern that for both KPSS and ADF, no matter if the process has a trend or not, the power increases as the sample size increases. For a sample size of 500 both time series for the KPSS test have a power close to unity. The increase in power is somewhat steeper for the KPSS compared to the ADF. The ADF test generally has a high power when the time series is without a trend whereas the power is around 15 to 30 percentage point lower when the time series has a trend compared to when it does not have a trend.

The power of the tests for each parameter value is displayed in Figure 10. The general pattern for both tests is that when the parameter value is close to unity or negative one the power is lower. There is one exception which is when the time series is without trend for the KPSS test, the power is then still pretty high when the parameter value is -0.9. The ADF test generally has the highest power when the time series is without trend of all four cases compared except when the parameter value is -0.9.

Figure 9. Power per sample size.
Figure 10. Power per parameter value.

Figure 11 represents the power for each of the three processes. The ADF has the highest power for the autoregressive process. Whereas the KPSS has the highest power for the moving average process. Both tests have the lowest power for the autoregressive moving average regardless of whether the process contains a trend or not. The ADF test has the highest power for all types of processes when the time series contains a trend compared to the other three cases.

Figure 11. Power per process.
Figure 12. Power per sample size and process for nonstationary time series with trend for the KPSS test.

Figure 12 displays the power for the KPSS test for each sample size and process when the time series is nonstationary with trend. The KPSS test has highest power when the process is a moving average and lowest when the process is a combined autoregressive moving average independent of sample size. When the sample size is 500 the power is almost unity for all types of processes. The power for the KPSS test for each sample size and process when the time series is nonstationary without trend are displayed in Figure 13. The pattern is the same as in Figure 12 but with slightly higher values.

Figure 13. Power per sample size and process for nonstationary time series without trend for the KPSS test.
Figure 14. Power per sample size and process for stationary time series with trend for the ADF test.

Figure 14 illustrates the power for the ADF test per sample size and process when the time series is stationary with trend. The autoregressive moving average distinguishes from the other processes as its power is under 50 percent for all sample sizes. Both processes of autoregressive and moving average has a power of unity when the sample size is 500.

The power for the ADF test per sample size and process when the time series is stationary without trend is displayed in Figure 15. When the time series is a moving average, the power is unity already at a sample size of 100. The autoregressive moving average has the lowest power for all sample sizes.

Figure 15. Power per sample size and process for stationary time series without trend for the ADF test.
5. Discussion

The result shows that the size for both ADF and KPSS sometimes is unacceptably large even for a sample size of 500 meaning the tests are not performing as well as one would wish for. Asymptotic theory about both tests says that the size should converge to the significance level of five percent as the sample size goes to infinity. There is no sign in the simulation study that this is happening, instead the size is either constant or increasing which is contradictory to theory. It could be that a sample size of 500 simply is not enough for the asymptotic properties of the tests to be valid indicating that the tests have poor size properties for a finite sample and suffer from size distortion. These findings are alarming since time series in reality rarely has more than 500 observations.

When dividing the size per process it becomes clear that the size distortion mainly origin from the insufficient performance of the ADF test when the process is of another type than autoregressive which is in line with findings by Schwert (1989). The size for the autoregressive process is close to five percent independent of sample size meaning the asymptotic properties are valid already at a sample size of 20. This may originate from the derivation of the ADF test where the creators considered an autoregressive process, and the simulated critical values are therefore possibly not adjusted for other processes than an autoregressive. For the KPSS the only type of process where the test performs adequately is moving average. There is no clear information in the original paper by Kwiatkowski et al. (1992) regarding what type of processes were used when simulating the critical values. Though, they evaluated the size properties of the test on an autoregressive process of order one. They stated that the test rejects too often when the parameter value is higher than zero and too rarely when the parameter value is less than zero. This is in line with findings in this thesis which show that the KPSS clearly have problems diagnosing the correct order of integration when the parameter value is close to unity. For the ADF test the problem is instead when the process has a parameter value close to negative one. Then the size varies between seven and 62 percent meaning the test occasionally rejects the null hypothesis far too often. This indicates that the value of the parameter can be especially problematic for the ADF test.

A somewhat surprising result is that the size of the two tests frequently is smaller when the sample size is 20 compared to larger sample sizes. This is contradictory to theory and should be further evaluated to find possible reasons and exclude potential errors. It is also unexpected
that the size sometimes grows as the sample size becomes larger. One plausible explanation for the increase in size could be that the size exhibits some fluctuations, which also can be seen in Kwiatkowski et al. (1992), meaning that it is possibly not an increase but rather a fluctuation. If more sample sizes were considered the fluctuation should then be seen given that there are no errors.

Both tests performs best when the time series is without a trend. The reason why it worsens when the time series contains a trend may possibly be due to the tests having problems determining whether the behavior of the time series is the result of a trend or a unit root. Remembering equation 14 for the Beveridge-Nelson decomposition the value of $\delta$ will make the mean of the time series to either increase or decrease with time. The same applies to the unit root component in the equation which can result in a changing mean. The similar characteristics in terms of a growing or decreasing mean may cause the tests to mistakenly regard the trend as a unit root for example.

The power of KPSS and ADF generally varies a lot but it is clear that both tests at times have problems with low power. It is especially low for the tests when the process is an autoregressive moving average. The simulation shows that for the studied sample sizes the power can be as low as 31 percent and never higher than 50 percent for the ADF when the time series has a trend. This indicates that the tests have problems managing a process that both have an autoregressive and a moving average part. This can also be seen when considering the size of the tests. Both sample size and parameter value affect the power of the tests greatly. The power increases as the sample size grows which is also expected since there generally is a positive relationship between sample size and power. In most cases the power is one or at least approaching one when the sample size is 500. Considering parameter value, the simulation shows that the power is smallest for values close to unity or negative one with the exception for KPSS when the time series is without trend. It is reasonable to think that this origins from the tests having problems distinguishing between a stationary time series with a parameter value close to one and a unit root process. When it comes to parameter values close to negative one, there is no clear explanation for why the tests have problems but due to its severity both in terms of power and size this needs further evaluation. Comparing the results from this study to the recommendations by Arltová and Federová (2016) the main conclusion about the ADF and KPSS differ. The KPSS performs better than the ADF considering 500 observations and the ADF has higher power for small positive values of the parameter. This is the opposite to the
results by Arltová and Federová (2016) who recommended using KPSS for small parameter values and the ADF when there is a lot of observations. This could be due to this study being more comprehensive in terms of processes and parameter values.

The simulation study shows that there is not one test that is clearly more reliable than the other which makes it hard to determine if any of the tests is superior. This is because there are numerous factors involved such as sample size, parameter value and type of process. The two tests seem to act differently depending on these settings. This is also a potential explanation for why the two tests at times result in different conclusions. When the size is higher and the power is lower the test will more frequently be wrong. When one of the tests are wrong the problem with contradictory results arise and by bringing knowledge about these settings one could possibly determine whether one of the tests are more likely to be correct. This in turn may help researchers in diagnosing the correct order of integration.

It is occasionally proposed that the two tests are used together which also was the original idea for the KPSS proposed by Kwiatkowski et al. (1992). Though this has been criticized by for example Caner and Kilian (2001) who argued that this will result in favoring the unit root hypothesis. Besides, using the tests together will not fully solve the problem with size distortion and low power. There will still be cases where the two tests result in contradictory conclusions.

6. Conclusion

The aim of this thesis was to compare the size and power properties of a test of stationarity with a test of unit root by conducting a simulation study. The goal was to find answers to why and when the two tests sometimes result in contradictory conclusions as well as to investigate if one of them are more reliable. The result shows that the performance of the two tests is highly influenced by factors such as sample size, parameter value and type of process. This makes it hard to conclude if one test is more reliable than the other. There are also signs of size distortion as well as low power. In some situations, the tests perform exceptional as when the process is an autoregressive for the ADF or a moving average for the KPSS but the big difference in performance makes the tests fragile and hard to trust. There are also indications that the asymptotic properties are not valid for the most common time series lengths which makes one question how useful the tests are in practice. The effect could be that researcher put too much
trust in these tests resulting in wrong conclusions. It is therefore of great importance to be aware of the problem with low power, size distortion and its severity.

The result of the study is limited to the two used tests, KPSS and ADF, and it may not be possible to generalize the result to other tests of stationarity and unit root. Though similar results have been found for other tests. The study is also limited in only considering first order ARIMA models. It is possible that the result from the simulation study is effected by this. The result and conclusion could differ from a study which is more comprehensive and considers for example second and third order processes as well. This makes the result less generalizable. It would therefore be interesting to also examine autoregressive, moving average as well as combined autoregressive moving average processes with higher order. Due to the inconclusive research around negative parameter values and combined processes, further research could perform a simulation study to investigate this in combination with higher order processes. The most important future research would however be to find a way to improve the size and power properties of the tests or invent a more accurate way to investigate the order of integration of a time series. It would also be interesting to in more detail illustrate what the consequences would be when one wrongly assumes stationarity and how this for example affects future predictions or the possibility to make inference. Preferably by examining real data.
References


Appendix A

Table 2. Critical values used for calculating size adjusted power per sample size. Size adjusted to five percent.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>KPSS</th>
<th>ADF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonstationary with trend</td>
<td>Nonstationary without trend</td>
</tr>
<tr>
<td>20</td>
<td>0.157</td>
<td>0.534</td>
</tr>
<tr>
<td>50</td>
<td>0.192</td>
<td>0.623</td>
</tr>
<tr>
<td>100</td>
<td>0.206</td>
<td>0.631</td>
</tr>
<tr>
<td>500</td>
<td>0.236</td>
<td>0.653</td>
</tr>
</tbody>
</table>

Table 3. Critical values used for calculating size adjusted power per parameter value. Size adjusted to five percent.

<table>
<thead>
<tr>
<th>Parameter value</th>
<th>KPSS</th>
<th>ADF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonstationary with trend</td>
<td>Nonstationary without trend</td>
</tr>
<tr>
<td>0.3</td>
<td>0.157</td>
<td>0.495</td>
</tr>
<tr>
<td>0.7</td>
<td>0.196</td>
<td>0.636</td>
</tr>
<tr>
<td>0.9</td>
<td>0.313</td>
<td>1.037</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.124</td>
<td>0.356</td>
</tr>
<tr>
<td>-0.7</td>
<td>0.101</td>
<td>0.226</td>
</tr>
<tr>
<td>-0.9</td>
<td>0.117</td>
<td>0.201</td>
</tr>
</tbody>
</table>

Table 4. Critical values used for calculating size adjusted power per process. Size adjusted to five percent.

<table>
<thead>
<tr>
<th>Process</th>
<th>KPSS</th>
<th>ADF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nonstationary with trend</td>
<td>Nonstationary without trend</td>
</tr>
<tr>
<td>AR</td>
<td>0.212</td>
<td>0.673</td>
</tr>
<tr>
<td>MA</td>
<td>0.134</td>
<td>0.395</td>
</tr>
<tr>
<td>ARMA</td>
<td>0.217</td>
<td>0.675</td>
</tr>
</tbody>
</table>
Table 5. Critical values used for calculating size adjusted power per sample size and process when the time series is nonstationary with trend. Size adjusted to five percent.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>KPSS</th>
<th>AR</th>
<th>MA</th>
<th>ARMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.162</td>
<td>0.138</td>
<td>0.166</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.206</td>
<td>0.134</td>
<td>0.215</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.230</td>
<td>0.134</td>
<td>0.237</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.276</td>
<td>0.133</td>
<td>0.281</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Critical values used for calculating size adjusted power per sample size and process when the time series is nonstationary without trend. Size adjusted to five percent.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>KPSS</th>
<th>AR</th>
<th>MA</th>
<th>ARMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.568</td>
<td>0.399</td>
<td>0.583</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.715</td>
<td>0.400</td>
<td>0.716</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.743</td>
<td>0.397</td>
<td>0.750</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.789</td>
<td>0.397</td>
<td>0.771</td>
<td></td>
</tr>
</tbody>
</table>

Table 7. Critical values used for calculating size adjusted power per sample size and process when the time series is stationary with trend. Size adjusted to five percent.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>ADF</th>
<th>AR</th>
<th>MA</th>
<th>ARMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>-3.132</td>
<td>-3.958</td>
<td>-4.781</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>-2.950</td>
<td>-5.000</td>
<td>-6.591</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>-2.915</td>
<td>-6.422</td>
<td>-8.826</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>-2.869</td>
<td>-10.465</td>
<td>-15.758</td>
<td></td>
</tr>
</tbody>
</table>

Table 8. Critical values used for calculating size adjusted power per sample size and process when the time series is stationary without trend. Size adjusted to five percent.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>ADF</th>
<th>AR</th>
<th>MA</th>
<th>ARMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>-2.001</td>
<td>-2.911</td>
<td>-3.118</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>-1.951</td>
<td>-3.814</td>
<td>-3.857</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>-1.951</td>
<td>-4.792</td>
<td>-4.739</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>-1.947</td>
<td>-7.481</td>
<td>-7.622</td>
<td></td>
</tr>
</tbody>
</table>