Size and power of tests for assessing weak stationarity of time series data: an empirical investigation

Xiaoguang Luo

Leica Geosystems AG, Switzerland
E-Mail: xiaoguang.luo@leica-geosystems.com

Abstract

Whether or not a time series is weakly stationary has long been a question of major interest in the field of time series analysis. Stationary time series can be sufficiently described by means of autoregressive moving average (ARMA) processes. When modelling temporal correlations of GNSS observation noise, the applicability of ARMA processes depends on the stationarity of residual time series from GNSS data analysis. According to the property that stationary processes have homogenous variances, statistical inferences on stationarity can be made by testing for homogeneity of variance (HOV). In addition, considering a time series as a realisation of a stochastic process, stationarity can be assessed by testing for stochastic trends using unit root tests. Based on representative data simulations, this paper analyses the empirical size and power of commonly used HOV and unit root tests. The results show that the performance of the HOV test is strongly affected by serial correlations, whereas the unit root test produces high power without significant size distortions.

1 Introduction

In the preliminary stage of modelling time series data, an important question to be answered is whether a time series is stationary or not. A discrete time series $\{X_t\}$ with $t \in \mathbb{Z}$ is considered as stationary if it has the similar statistical properties to those of the "time-shifted" series $\{X_{t+h}\}$, for each $h \in \mathbb{Z}$. Strict stationarity of a time series is defined by the condition that $(X_1,\ldots,X_n)$ and $(X_{1+h},\ldots,X_{n+h})$ have the same joint distributions for all integers h and $n > 0$. A weaker form of stationarity commonly known as weak stationarity simply requires that the mean function $\mu_X(t) = \mathrm{E}(X_t)$ and the covariance function $\gamma_X(t+h,t)$ of $\{X_t\}$, i.e.,

$$\mathrm{Cov}(X_{t+h},X_t) = \mathrm{E}\{[X_{t+h}-\mu_X(t+h)]\,[X_t-\mu_X(t)]\} \tag{1.1}$$

do not vary with respect to time t for each $h \in \mathbb{Z}$, where $\mathrm{E}(\cdot)$ is the expectation operator. This indicates that

$$\mathrm{E}(X_t) = \mu_0, \tag{1.2}$$

$$\mathrm{Cov}(X_{t+h},X_t) = \mathrm{Cov}(X_h,X_0). \tag{1.3}$$

If $\{X_t\}$ is strictly stationary and $\mathrm{E}(X_t^2) < \infty$ for all t, then $\{X_t\}$ is also weakly stationary (Brockwell and Davis, 2002, p. 15). For the sake of brevity, the term stationary is used in this paper in the sense of weakly stationary.

Setting $h = 0$ in Eq. (1.3), the variance function of a stationary time series is equal to a constant:

$$\mathrm{Var}(X_t) = \mathrm{Cov}(X_t,X_t) = \mathrm{Cov}(X_0,X_0) = \mathrm{Var}(X_0). \tag{1.4}$$

This means that all random variables in the time series have the same finite variance (also known as homoscedasticity). Therefore, statistical tests for HOV,

Erschienen bei KIT Scientific Publishing
Schriftenreihe des Studiengangs
Geodäsie und Geoinformatik 2018,1

DOI Einzelbeitrag:
10.5445/KSP/1000080235

Festschrift zur Verabschiedung von
Prof. Dr.-Ing. Dr. h.c. Bernhard Heck
(Schw)Ehre, wem (Schw)Ehre gebührt

DOI Festschrift:
10.5445/KSP/1000080324

such as the two-sample β test (Teusch, 2006, p. 114), can be used to verify the necessary condition for stationarity by revising equal variances.

Consider a time series $\{X_t\}$ as a linear stochastic process and assume that it can be described by an autoregressive model of order p, i.e., AR(p)

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t, \qquad (1.5)$$

where $\{Z_t\}$ denotes a white noise (WN) process representing a sequence of uncorrelated random variables, each with zero mean and variance σ_Z^2. Such a process is indicated by the notation $\{Z_t\} \sim \mathrm{WN}(0, \sigma_Z^2)$. If $z = 1$ is a root of the associated p^{th}-degree autoregressive characteristic equation given by

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0, \qquad (1.6)$$

$\{X_t\}$ represents a unit root process and is non-stationary (Brockwell and Davis, 2002, p. 85). In comparison to a stationary process, where all roots of Eq. (1.6) lie outside the unit circle (Box et al., 2016, p. 55), a unit root process illustrates a mean-diversion behaviour and has a time-dependent variance diverging to infinity. In general, unit root processes can be rendered stationary by serially differencing, for example, by applying the first-order or lag-1 difference $Y_t = X_t - X_{t-1}$. If the differenced time series $\{Y_t\}$ can be modelled by an ARMA process, the original time series $\{X_t\}$ is called an autoregressive integrated moving average (ARIMA) process. $\{X_t\}$ is difference-stationary and has stochastic trends. If the increments of $\{X_t\}$ represent a WN process, i.e.,

$$X_t - X_{t-1} = Z_t, \qquad (1.7)$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma_Z^2)$, $\{X_t\}$ is called a random walk process. According to Eqs. (1.5) and (1.6), the autoregressive characteristic equation of a random walk process has a unit root. Therefore, a time series is non-stationary if it has random walk components. Pretesting for unit roots plays an important role not only in assessment of stationarity but also in selection of appropriate time series models. One of the most famous unit root tests is the augmented Dickey-Fuller (ADF) test (Dickey and Fuller, 1979; Said and Dickey, 1984).

The remainder of this paper is organised as follows. In Sect. 2 the mathematical backgrounds of the applied HOV and unit root tests for assessing stationarity are briefly described. Sect. 3 presents the data simulation by means of representative AR(I)MA processes. In Sect. 4 the performance of the tests is analysed based on empirical size and power values. Finally, Sect. 5 provides concluding remarks and an outlook on future research work.

2 Tests for assessing stationarity

This section summaries the core characteristics of the two-sample β test and the ADF test which are used in this study to verify homogeneity of variance and the existence of unit roots, respectively. The significance level α is the probability that the test falsely rejects the null hypothesis and commits a Type I error.

2.1 Two-sample β test

Based on the ergodic theorems that establish the relation between time and space averages (Birkhoff, 1942; Anosov, 2001), the HOV tests verify the equality of variance among individual groups longitudinally formed by subdividing a univariate time series rather than transversally built by assembling independent realisations. Let $(X_{11}, \ldots, X_{1n_1}) \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $(X_{21}, \ldots, X_{2n_2}) \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_2^2)$ be two independent samples, where iid denotes independently identically distributed. The corresponding unbiased estimators for population variances $\sigma_j^2, j \in \{1, 2\}$ are given by

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(X_{ji} - \overline{X}_{j\cdot}\right)^2, \text{ where } \overline{X}_{j\cdot} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ji}. \qquad (2.1)$$

Under the normal distribution assumption and the null hypothesis $H_0: \sigma_1^2 = \sigma_2^2$, the two-sample β test statistic T_β follows the β distribution (Abramowitz and Stegun, 1972, p. 944)

$$T_\beta := \frac{s_1^2}{s_1^2 + s_2^2} \sim \beta(a, b), \qquad (2.2)$$

where $a = \frac{n_1 - 1}{2}$ and $b = \frac{n_2 - 1}{2}$ (Teusch, 2006, p. 115). The null hypothesis of equal variances is rejected at a significance level of α if

$$T_\beta < \beta_{\frac{n_1-1}{2}, \frac{n_2-1}{2}; \frac{\alpha}{2}} \text{ or } T_\beta > \beta_{\frac{n_1-1}{2}, \frac{n_2-1}{2}; 1-\frac{\alpha}{2}}. \qquad (2.3)$$

In the case of independent and normally distributed random variables, the β test is a uniformly most powerful two-tailed test (Lehmann, 1986; Teusch, 2006, p. 129). However, it is very sensitive to deviations from the normal distribution. In addition, the assumed uncorrelatedness among the random variables within each sample and the independence between two samples are hardly fulfilled in practise. As an application example, Howind (2005) utilised the two-sample β test to evaluate the performance of an advanced stochastic model of GPS carrier-phase observations.

2.2 Augmented Dickey-Fuller test

The augmented Dickey-Fuller (ADF) test is based on the existence and uniqueness property of an ARMA process, i.e., the p^{th}-degree autoregressive characteristic equation given by Eq. (1.6) has no unit root (Brockwell and Davis, 2002, p. 85). To understand the basic concept of autoregressive unit root tests, let $(X_1,\ldots,X_n)$ be observations from a first-order autoregressive AR(1) process

$$X_t = \phi X_{t-1} + Z_t, \tag{2.4}$$

where $Z_t \sim \text{WN}(0,\sigma_Z^2)$ and $|\phi| < 1$. It can be shown that the ordinary least squares (OLS) estimator of ϕ asymptotically follows a normal distribution (Hamilton, 1994, p. 216):

$$\hat{\phi} \overset{A}{\sim} \mathcal{N}(\phi, \frac{1-\phi^2}{n}). \tag{2.5}$$

However, in the unit root case with $\phi = 1$, the normal distribution approximation $\hat{\phi} \overset{A}{\sim} \mathcal{N}(1,0)$ is no longer applicable, which precludes its use for testing the unit root hypothesis $H_0 : \phi = 1$ against the alternative $H_1 : |\phi| < 1$. The problem is that under H_0, $\{X_t\}$ is neither stationary nor ergodic, and the usual sample moments do not converge to fixed constants. Dickey and Fuller (1979) first considered the autoregressive unit root test and derived the limiting distribution as $n \to \infty$ for the test statistic

$$t_{\phi=1} = \frac{\hat{\phi}-1}{SE(\hat{\phi})}, \tag{2.6}$$

where $SE(\hat{\phi})$ denotes the standard error of $\hat{\phi}$ resulting from the OLS evaluation. The limiting distribution of $t_{\phi=1}$ is referred to as the Dickey-Fuller distribution.

When testing for autoregressive unit roots in practise, many time series have more complicated dynamic structures which cannot be fully characterised by a simple AR(1) process as given in Eq. (2.4). Said and Dickey (1984) augmented the basic autoregressive unit root test to accommodate general ARMA processes with unknown order parameters. Assuming that the stochastic dynamics in the data can be sufficiently described by an ARMA process, the regression model of the ADF test, verifying the null hypothesis H_0 that $\{X_t\}$ is difference-stationary (non-stationary) against the alternative hypothesis H_1 that $\{X_t\}$ is trend-stationary, is formulated as

$$X_t = \mathbf{C}^T\mathbf{D_t} + \phi X_{t-1} + \sum_{j=1}^{l-1} \psi_j \Delta X_{t-j} + Z_t, \tag{2.7}$$

where $\mathbf{C} = (c,d)^T$ and $\mathbf{D_t} = (1,t)^T$ capture the deterministic trend, ϕ is the AR(1) coefficient, and the $l-1$ difference terms $\psi_j \Delta X_{t-j}$ approximate the ARMA structure of the residuals. Neglecting the deterministic trend in Eq. (2.7), the presentability of an ARMA process by the ADF regression model is mathematically proved in Luo (2013, pp. 303–305). If the truncation lag l is set to a too small value, the remaining serial correlations in regression residuals will bias the test. If l is set to a too large value, the power of the test will suffer. Ng and Perron (1995) suggested a data-based approach to optimizing the truncation lag selection. It begins with a maximum lag length $l_{\max}$ (Schwert, 1989) given by

$$l_{\max} = \left\lfloor 12 \cdot \left(\frac{n}{100}\right)^{1/4} \right\rfloor, \tag{2.8}$$

where $\lfloor x \rfloor$ denotes the integer part of x. Then, the significance of the coefficient of the last lagged difference is evaluated by means of the t-statistic. If this coefficient is statistically significant, the unit root test is carried out. Otherwise, the truncation lag l is reduced by one, and the procedure is repeated. By doing this, the lag value determined leads to a stable size (Sect. 4.1) and a minimum power loss. Based on the OLS estimates of Eq. (2.7), the ADF test statistic $t_{\phi=1}$ can be calculated using Eq. (2.6). The ADF test is a one-sided and left-tailed test, meaning that the unit root null hypothesis H_0 is rejected at a significance level of α if $t_{\phi=1} < DF_\alpha$, where DF_α denotes the α-quantile of the Dickey-Fuller distribution. Note that under H_0,

the asymptotic distribution of $t_{\phi=1}$ is influenced by the type of the deterministic terms in Eq. (2.7), but not by their parameter values.

3 Data simulation

In order to empirically investigate the size and power of the above-introduced tests for assessing stationarity, representative time series are simulated by means of low-order AR(I)MA processes. Keeping the context of GNSS stochastic modelling in mind, the parameters of the data-generating processes are specified considering the temporal correlation characteristics of GNSS observation noise presented in Wang et al. (2002), Howind (2005), Schön and Brunner (2008) and Luo et al. (2012). Following the notation of a general ARMA(p, q) process defined in Brockwell and Davis (2002, p. 83), i.e.,

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}, \qquad (3.1)$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma_Z^2)$, Table 3.1 provides the model parameters of the AR(I)MA processes used for data simulation. The model parameters of AIM are related to the lag-1 differenced process, and the random walk process (RWP) is given by Eq. (1.7). For each AR(I)MA process listed in Table 3.1, 1000 time series have been simulated for each of the data lengths $2^6, 2^7, \ldots, 2^{12}$.

Once a stationary ARMA process is uniquely defined, the associated autocorrelation function (ACF) can be derived from the model parameters (Brockwell and Davis, 2002, p. 88). Fig. 3.1 shows the theoretical ACFs of the ARMA processes used for data simulation. All model ACFs exclusively exhibit positive correlations, reflecting the general assumption that GNSS observation noise is positively correlated in time. In comparison to the AR(1) processes, the ACFs of the higher-order ARMA(3,2) processes illustrate significantly larger correlation lengths, indicating the capability of higher-order ARMA models of describing more complex temporal correlation behaviour (Luo et al., 2012).

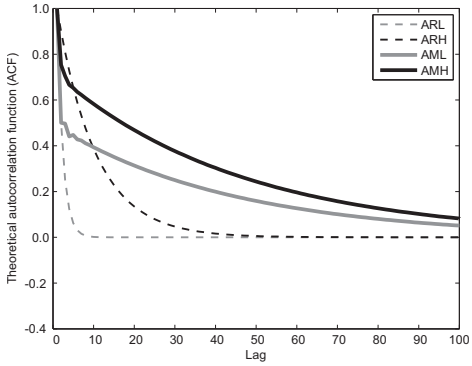

Figure 3.1: Theoretical ACFs of the stationary ARMA processes used for data simulation (cf. Table 3.1).

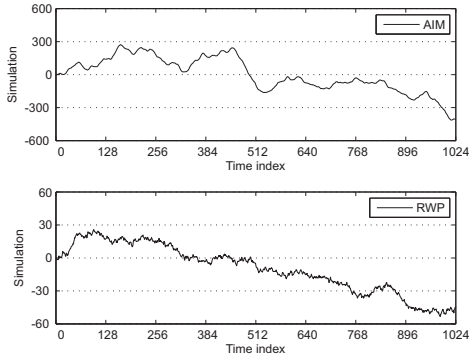

Figure 3.2: Path examples of the non-stationary ARIMA models used for data simulation (cf. Table 3.1).

Fig. 3.2 visualises two path examples of the non-stationary ARIMA processes which seem indistinguishable from trend-stationary processes. Being different from deterministic trends that are regulated by time, temporal increases or decreases due to stochastic trends are caused by cumulated shocks which have persistent effects over time. In order to effectively differentiate between stochastic and deterministic trends, hypothesis tests considering the underlying stochastic processes are more appropriate than those merely based on empirically derived statistical characteristics.

4 Empirical size and power analysis

Applying the two-sample β test and the ADF test to the simulated data, the resulting empirical size and power are analysed in this section. All tests are performed at a significance level of 5%, where different regression models and truncation lags are considered for the ADF test.

4.1 Size and power of a statistical test

The size of a hypothesis test, also known as significance level α, gives the probability of falsely reject-

Table 3.1: Model parameters of the low-order AR(I)MA processes used for data simulation. (1)-Wang et al. (2002), (2)-Luo et al. (2012), (3)-Maddala and Kim (1998, pp. 20, 76).

Process	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	σ_Z	Notation	Reference
AR(1)	0.50	–	–	–	–	1.00	ARL	(1)
AR(1)	0.90	–	–	–	–	1.00	ARH	(1)
ARMA(3,2)	0.62	0.43	−0.09	−0.42	−0.34	0.24	AML	(2)
ARMA(3,2)	0.73	0.38	−0.14	−0.33	−0.35	0.29	AMH	(2)
ARIMA(1,1,1)	0.90	–	–	0.5	–	1.00	AIM	(3)
ARIMA(0,1,0)	–	–	–	–	–	1.00	RWP	(3)

ing the null hypothesis H_0 (Type I error). In order to protect H_0 and to prevent the investigator from inadvertently making false claims, the size of a hypothesis test should be kept as small as possible. A significance level of 5% is normally used in the practise of hypothesis testing (Stigler, 2008). The empirical size is defined as the rejection rate of H_0 tested based on data for which H_0 is actually true.

The power of a hypothesis test measures the test's ability to reject H_0 when it is actually false. In other words, the power of a test is the probability of not committing a Type II error which means failing to reject H_0 when it is in fact false. The maximum power of a statistical test is 1 and ideally a test is desirable to possess high power close to 1. The empirical power is determined by calculating the rejection rate of H_0 tested using data for which the alternative hypothesis H_1 is true. Note that decreasing the size of a test raises the probability of Type II errors and reduces the test power.

Table 4.1 gives an overview of the AR(I)MA data for the empirical size and power analysis (cf. Table 3.1). The size is investigated based on the data for which H_0 is true, whereas the power is evaluated using the data for which H_1 is true.

Table 4.1: AR(I)MA data for empirical size and power analysis.

Measure	Two-sample β test	ADF test
Size	ARL, ARH, AML, AMH	AIM, RWP
Power	AIM, RWP	ARL, ARH, AML, AMH

4.2 Two-sample β test

The two-sample β test is applied to the simulated AR(I)MA time series of different lengths, the empirical size and power values are presented in Table 4.2. It can be seen that in most cases both the size and power rise with an increasing data length. This means for a larger n, it is more likely to falsely reject the null hypothesis of HOV, and thus stationarity. On the other hand, in terms of power, the probability of committing a Type II error becomes smaller as n grows.

The empirical sizes of ARL, which illustrates the shortest zero-crossing correlation length in Fig. 3.1, are closest to the nominal level of 5%. Moreover, the size increases with correlation length, which can be seen by comparing the results between ARL (AML) and ARH (AMH). For a sufficient data volume of $n \geq 1024$ that is approximately 10 times the correlation length of AMH (Luo et al., 2011), the HOV null hypothesis is rejected for more than 50% of the data with stronger serial correlations, i.e., ARH, AML and AMH. Such high rejection rates are due to the deviation from the iid assumption of the two-sample β test. For $n = 64$, the empirical sizes of AML and AMH are also close to the nominal level of 5%. Such low sizes are due to the small data length which is insufficient to reflect the stochastic properties of the data-generating processes. Regarding $n = 1024$, almost 90% of the simulated non-stationary data can be correctly rejected by the two-sample β test, showing favourable rejection rates with high empirical power values.

For $n = 1024$, Fig. 4.1 illustrates the empirical cumulative distribution functions (CDFs) of the two-sample β test statistic T_β, along with the theoretical CDF of the β distribution. Depending on the degree of serial correlations, the empirical CDFs of ARL and AMH show the smallest and largest deviations from the theoretical CDF curve, respectively.

Table 4.2: Empirical size and power values of the two-sample β test ($\alpha = 5\%$).

Data length	Empirical size				Empirical power	
n	ARL	ARH	AML	AMH	AIM	RWP
$2^6 = 64$	0.11	0.38	0.07	0.13	0.69	0.54
$2^7 = 128$	0.11	0.45	0.11	0.27	0.75	0.66
$2^8 = 256$	0.12	0.47	0.23	0.44	0.82	0.75
$2^9 = 512$	0.13	0.51	0.38	0.56	0.85	0.83
$2^{10} = 1024$	0.10	0.52	0.43	0.61	0.89	0.88
$2^{11} = 2048$	0.13	0.54	0.54	0.68	0.92	0.92
$2^{12} = 4096$	0.14	0.50	0.56	0.68	0.95	0.93

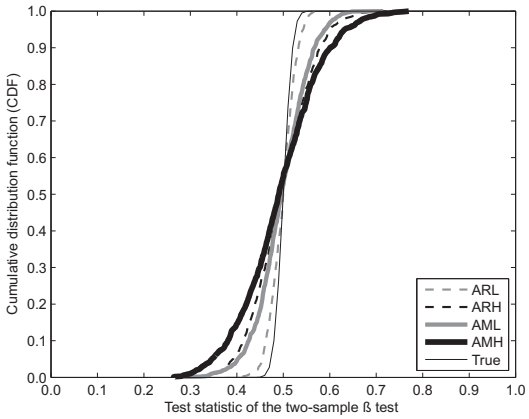

Figure 4.1: Empirical and true CDFs of the two-sample β test statistic T_β ($n = 1024$); see Eq. (2.2).

4.3 Augmented Dickey-Fuller test

When performing the ADF test, it is important to properly specify the regression model and the truncation lag. The regression model should be able to describe the trend behaviour of the data, whereas the truncation lag directly impacts upon the test performance. Table 4.3 provides the parameter settings for the ADF test. Regarding the path examples of AIM and RWP shown in Fig. 3.2, it is reasonable to consider a non-zero constant and a time-dependent trend within the regression model. According to Schwert (1989) and Kwiatkowski et al. (1992), three lag values $l = 0$, l_{short} and l_{long} are used, where l_{short} and l_{long} are given by

$$l_{\text{short}} = \left\lfloor 4 \cdot \left(\frac{n}{100}\right)^{1/4} \right\rfloor, \tag{4.1}$$

$$l_{\text{long}} = \left\lfloor 12 \cdot \left(\frac{n}{100}\right)^{1/4} \right\rfloor. \tag{4.2}$$

Applying the ADF test to the non-stationary AIM and RWP time series, the empirical sizes are provided in Table 4.4. The alternative hypothesis is a stationary ARMA process around a constant mean (a time-dependent trend) for the regression model C0 (CD). It can be seen that the empirical sizes of the ADF test are fairly close to the nominal level of 5%, even for moderate data lengths. Furthermore, including a trend in the regression model improves the size value (C0 vs. CD), which is particularly visible for L0. When changing the lag parameter from L4 to L12, only slight size improvements are observed. This indicates that the truncation lag given by Eq. (4.1) seems sufficient to characterise the serial correlations in the regression residuals.

Table 4.3: Parameter settings for the ADF test; see Eq. (2.7).

Parameter	Setting	Notation
Regression	$c \neq 0, d = 0$	C0
model $c + d \cdot t$	$c \neq 0, d \neq 0$	CD
Truncation	$l = 0$	L0
lag l	$l = l_{\text{short}}$; Eq. (4.1)	L4
	$l = l_{\text{long}}$; Eq. (4.2)	L12

Tables 4.5 and 4.6 show the empirical power of the ADF test against the alternative stationary processes ARL, ARH, AML and AMH. Regarding the same data length and the same regression mode, the empirical power decreases with increasing serial correlations (ARL vs. ARH, AML vs. AMH). In addition, considering a time-dependent trend in the regression model reduces the power of the ADF test (e.g., L4: C0 vs. CD). As mentioned in Sect. 2.2, the power of the ADF test will suffer if the truncation lag is set to a too large value. This can be observed by comparing the power results between L4 and L12. Taking both the empirical size and power into account, the regression model CD together with the truncation lag L4 seems to be an appropriate choice for $n \geq 1024$. Using this parame-

Table 4.4: Empirical sizes of the augmented Dickey-Fuller (ADF) test ($\alpha = 5\%$).

Process	Data length	Regression model C0			Regression model CD		
	n	L0	L4	L12	L0	L4	L12
AIM	64	0.11	0.07	0.06	0.06	0.10	0.06
	128	0.15	0.05	0.06	0.06	0.04	0.06
	256	0.16	0.06	0.06	0.05	0.05	0.05
	512	0.13	0.04	0.04	0.06	0.05	0.04
	1024	0.15	0.05	0.05	0.06	0.06	0.04
	2048	0.14	0.05	0.05	0.07	0.06	0.05
	4096	0.15	0.05	0.06	0.07	0.05	0.05
RWP	64	0.05	0.04	0.05	0.04	0.05	0.04
	128	0.04	0.05	0.04	0.04	0.04	0.04
	256	0.05	0.05	0.05	0.05	0.04	0.03
	512	0.05	0.05	0.05	0.05	0.04	0.04
	1024	0.05	0.05	0.05	0.05	0.04	0.05
	2048	0.05	0.05	0.05	0.03	0.04	0.04
	4096	0.06	0.06	0.06	0.05	0.04	0.05

Table 4.5: Empirical power of the augmented Dickey-Fuller (ADF) test ($\alpha = 5\%$).

Process	Data length	Regression model C0			Regression model CD		
	n	L0	L4	L12	L0	L4	L12
ARL	64	1.00	0.73	0.12	0.99	0.49	0.07
	128	1.00	0.99	0.49	1.00	0.94	0.28
	256	1.00	1.00	0.94	1.00	1.00	0.78
	512	1.00	1.00	1.00	1.00	1.00	1.00
	1024	1.00	1.00	1.00	1.00	1.00	1.00
	2048	1.00	1.00	1.00	1.00	1.00	1.00
	4096	1.00	1.00	1.00	1.00	1.00	1.00
ARH	64	0.16	0.11	0.05	0.09	0.08	0.04
	128	0.48	0.32	0.16	0.30	0.21	0.10
	256	0.98	0.84	0.49	0.84	0.60	0.28
	512	1.00	1.00	0.96	1.00	1.00	0.83
	1024	1.00	1.00	1.00	1.00	1.00	1.00
	2048	1.00	1.00	1.00	1.00	1.00	1.00
	4096	1.00	1.00	1.00	1.00	1.00	1.00

ter combination, the unit root hypothesis is correctly rejected at high power levels above 95% without significant size distortions.

5 Conclusions and outlook

This paper presents an empirical investigation of the size and power of hypothesis tests for assessing weak stationarity of time series data. Based on representative AR(I)MA simulations, statistical inferences on stationarity are made by testing for homogeneity of variance (HOV) and for the existence of autoregressive unit roots. The two-sample β test is used to verify the HOV null hypothesis, whereas the augmented Dickey-Fuller (ADF) test is applied to the detection of autoregressive unit roots.

In spite of the high sensitivity to non-stationary alternatives, the two-sample β test overrejects the assumption of stationarity in this study, which becomes more severe with increasing data length and serial correlations. In comparison to the two-sample β test, the empirical sizes of the ADF test are significantly closer to the specified nominal level. Including a time-dependent trend in the regression model and utilising larger truncation lags improve the empirical size of the ADF test, but lead to power loss.

Table 4.6: Empirical power of the ADF test ($\alpha = 5\%$; continuation of Table 4.5).

Process	Data length	Regression model C0			Regression model CD		
	n	L0	L4	L12	L0	L4	L12
AML	64	1.00	0.49	0.06	1.00	0.39	0.05
	128	1.00	0.64	0.10	1.00	0.59	0.08
	256	1.00	0.78	0.16	1.00	0.75	0.13
	512	1.00	0.95	0.33	1.00	0.90	0.22
	1024	1.00	1.00	0.87	1.00	1.00	0.65
	2048	1.00	1.00	1.00	1.00	1.00	1.00
	4096	1.00	1.00	1.00	1.00	1.00	1.00
AMH	64	0.98	0.28	0.04	0.98	0.24	0.04
	128	1.00	0.35	0.06	1.00	0.30	0.07
	256	1.00	0.47	0.09	1.00	0.40	0.07
	512	1.00	0.75	0.29	1.00	0.63	0.17
	1024	1.00	0.98	0.78	1.00	0.95	0.55
	2048	1.00	1.00	1.00	1.00	1.00	0.97
	4096	1.00	1.00	1.00	1.00	1.00	1.00

The essential limitation of the HOV test applied in this study is the assumption of independent samples. To test for variance homogeneity of correlated variables, the robust large-sample methods proposed by Harris (1985) will be considered in future studies. Moreover, for a moderate data length, the ADF test exhibits low test power against the alternatives which are close to the unit root null hypothesis. To overcome this deficiency, Elliott et al. (1996) suggested the efficient unit root tests, which are recommended for future research.

References

Abramowitz, M. and Stegun, I. A. (1972): Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. 10th ed. *U.S. National Bureau of Standards: Applied Mathematics Series, No. 55*. U. S. Government Printing Office, Washington, D. C.

Anosov, D. V. (2001): Ergodic theory. In: Hazewinkel, M. *Encyclopaedia of Mathematics*. Dordrecht: Kluwer Academic Publishers.

Birkhoff, G. D. (1942): What is the ergodic theorem? *American Mathematical Monthly* 49(4):222–226.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2016): Time Series Analysis: Forecasting and Control. 5th ed. Wiley & Sons, Hoboken.

Brockwell, P. J. and Davis, R. A. (2002): Introduction to Time Series and Forecasting. 2nd ed. Springer-Verlag, New York.

Dickey, D. A. and Fuller, W. A. (1979): Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74(366):427–431.

Elliott, G., Rothenberg, T. J., and Stock, J. H. (1996): Efficient tests for an autoregressive unit root. *Econometrica* 64(4):813–836.

Hamilton, J. (1994): Time Series Analysis. Princeton University Press, Princeton.

Harris, P. (1985): Testing for variance homogeneity of correlated variables. *Biometrika* 72(1):103–107.

Howind, J. (2005): Analyse des stochastischen Modells von GPS-Trägerphasenbeobachtungen. *Deutsche Geodätische Kommission, Reihe C*, no. 584. Verlag der Bayerischen Akademie der Wissenschaften in Kommission beim Verlag C. H. Beck, Munich, Germany.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y. (1992): Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Economics* 54(1-3):159–178.

Lehmann, E. L. (1986): Testing Statistical Hypotheses. 2nd ed. Springer-Verlag, New York.

Luo, X. (2013): GPS Stochastic Modelling: Signal Quality Measures and ARMA Processes. *Springer Theses*. Springer-Verlag, Berlin Heidelberg.

Luo, X., Mayer, M., and Heck, B. (2011): Verification of ARMA identification for modelling temporal correlations of GNSS observations using the ARMASA toolbox. *Studia Geophysica et Geodaetica* 55(3):537–556.

Luo, X., Mayer, M., and Heck, B. (2012): Analysing time series of GNSS residuals by means of AR(I)MA processes. In: Sneeuw, N., Novák, P., Crespi, M., and Sansò, F. (eds.) Proceedings of the VII Hotine-Marussi Symposium on Mathematical Geodesy, Rome, July 6–10, 2009. Berlin: Springer, pp. 129–134.

Maddala, G. S. and Kim, I.-M. (1998): Unit Roots, Cointegration, and Structural Change. Cambridge University Press, Cambridge.

Ng, S. and Perron, P. (1995): Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* 90(429):268–281.

Said, S. E. and Dickey, D. A. (1984): Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika* 71(3):599–607.

Schön, S. and Brunner, F. K. (2008): A proposal for modelling physical correlations of GPS phase observation. *Journal of Geodesy* 82(10):601–612.

Schwert, W. (1989): Tests for unit roots: a Monte Carlo investigation. *Journal of Business and Economic Statistics* 7(2):147–159.

Stigler, S. (2008): Fisher and the 5% level. *Chance* 21(4):12.

Teusch, A. (2006): Einführung in die Spektral- und Zeitreihenanalyse mit Beispielen aus der Geodäsie. *Deutsche Geodätische Kommission, Reihe A*, no. 120. Verlag der Bayerischen Akademie der Wissenschaften in Kommission beim Verlag C. H. Beck, Munich, Germany.

Wang, J., Satirapod, C., and Rizos, C. (2002): Stochastic assessment of GPS carrier phase measurements for precise static relative positioning. *Journal of Geodesy* 76(2):95–104.