18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Long-term trend analysis of water quality in Lake Biwa

Kawasaki, Y. [1], K. Kawai [2], T. Okubo [3] and K. Kanefuji [1]

[1] Risk Analysis Research Center, The Institute of Statistical Mathematics, Research Organization of Information and Systems, Tokyo, Japan
[2] Department of Nutrition and Food Science, Beppu University, Oita, Japan
[3] Lake Biwa Environmental Research Institute, Shiga, Japan
Email: kawasaki@ism.ac.jp

Abstract: Lake Biwa is the largest freshwater lake in Japan, located in west central Japanese main land. We report the result of long-term trend analysis of time series of Chlorophyll-A in Lake Biwa. The concentration of Chlorophyll-A is a barometer of phytoplankton growth. The higher the concentration, the worse the water quality and the more probable we have the danger of red tides and blue-green algae. Data has been observed through the water quality investigations conducted by Shiga Prefecture and the Ministry of Land, Infrastructure, Transport and Tourism. Data are generally observed from April 1979 to March 2003, measured at the fixed points once in a month, though the data lengths and the observed items vary.

The first part of this presentation works on the removal of seasonal pattern from a given time series. Because the periodic pattern due to yearly climate change or some general social activities is a nuisance for trend-analysis, it should be removed in advance. For this aim, we employ unobserved component time series models where the observation is decomposed into several unobserved components. Stochastic constrains on the smoothness of the variability of the components lead to constrained least squares solution, which can be represented in a Markov form, a state space representation.

Based on the trend estimates obtained by the observation sites, we show a contour plot of the Chlorophyll-A concentration for a given time point via spatial smoothing and interpolation. Station-wise plots of the estimated trends of Chlorophyll concentration reveal that the clarification of north Lake Biwa made progress in 1980s while the south lake finally purified around 2000. This can be visually confirmed by the spatially smoothed contour plots. One of the background for the recent low concentration of Chlorophyll-A can be sought for the decrease of effluent impact, most of which could be accounted for the improvement of sewerage system of the towns around the lake.

After this preliminary stage, we obtain the trend series of measured substance by every monitoring point. In the next step, we examine the significance of the estimated trend via formal statistical tests. We report the results of two different kinds of unit root tests, ADF test and KPSS test that are standard tools in the context of econometric time series analysis. Unit root tests are originally the tools to determine whether the process is non-stationary or trend stationary. In this paper, however, we are not interested in such a distinction.

For example in ADF test, the null hypothesis is non-stationary (or the existence of stochastic trend). Even when the null is rejected, as long as the coefficient of the linear trend is significant and its sign is negative, we confirm that there is an evidence of decreasing trend in Chlorophyll concentration. On the other hand in KPSS test, the null is trend stationary. Even in case the null is not rejected, we demonstrate the existence of trend via the significant (and negative) coefficient of the linear trend.

Generally speaking, we witnessed the existence of (downward) trend at almost every site with respect to the concentration of Chlorophyll-A. If we regard the estimated trend as data, the results often support the existence of deterministic trend. On the other hand, when we use the seasonally adjusted data, stochastic trend is supported in many cases. The reason is rather obvious. If we rely on the extracted trend only, sometimes we work on very smooth data. Then data does not contain much variability, and the deterministic time trend is sufficient to describe the time series data. Seasonally adjusted series, to the contrary, contains irregular part, so it becomes more difficult to separate the irregular part from trend. In such a case, stochastic trend is so flexible that it can show better fit to the data. It is conjectured that by this reason the stochastic trend is preferred for the seasonally adjusted data.

Keywords: *Water quality, Chlorophyll concentration, seasonal adjustment, unit root test*

1. INTRODUCTION

This article reports the result of long-term trend analysis of time series of measurement of various substances in Lake Biwa. Data has been observed through the water quality investigations conducted by Shiga Prefecture and the Ministry of Land, Infrastructure, Transport and Tourism. Data are generally observed from April 1979 to March 2003, measured at the fixed points once in a month, though the data lengths and the observed items vary.

This article consists of two parts. The first part works on the removal of seasonal pattern from a given time series. Because the periodic pattern due to yearly climate change or some general social activities is a nuisance for trend-analysis, it should be removed in advance. After this preliminary stage, we obtain the trend series of measured substance by every monitoring point. In the next step, we examine the significance of the estimated trend via formal statistical tests. We report the results of two different kinds of unit root tests that are standard tools in the context of time series analysis.

2. SEASONAL ADJUSTMENT OF ENVIRONMENTAL TIME SERIES

There are a variety of observed items available. In the sequel, partly due to the limit of the space, we only report the results on Chlorophyll-A.

2.1. Constrained Least Squares

Suppose we concentrate our interest only on Chlorophyll-A. Now we have a single time series observed at a given monitoring point. For this time series y_t $(t=1,\ldots,T)$, we assume it can be decomposed into trend component μ_t, seasonal component s_t, and irregular component ε_t in such a way as $y_t = \mu_t + s_t + \varepsilon_t$, or as $y_t = \mu_t \times s_t \times \varepsilon_t$. As a multiplicative form can be reduced to an additive form by logarithmic transformation, we only discuss an additive model here.

An additive decomposition model can be regarded as a linear regression model where ε_t is an error term, while μ_t and s_t are unknown regression coefficients. However, μ_t and s_t are not fixed constants but time varying. The number of unknown parameters are $2T+1$ including the variance of the observational noise ε_t compared to the number of observations is T. We need additional assumptions to solve this least squares problem.

Now we assume some kind of smoothness of time transition for trend and seasonal components. Intuitively, it is meant by $\mu_t \approx \mu_{t-1}$ and $s_t \approx s_{t-12}$. Let u_{1t}, u_{2t}, u_{3t} be the normally distributed random variables with mean zero and variance τ_i^2 respectively, then we specify the smoothness constraints as $\Delta^2\mu_t = \mu_t - 2\mu_{t-1} + \mu_{t-2} = u_{1t}$ and $\Delta^{12}s_t = s_t - s_{t-12} = u_{2t}$. For the seasonal components, we can put the assumption that the total of the seasonal variations within a year sums up to nearly zero, namely $\sum_{j=0}^{11} s_{t-j} = u_{3t}$.

Together with the assumption on the irregular part that ε_t follows a normal distribution with mean 0 and variance σ^2, μ_t and s_t can be obtained as the solutions to the following constrained least squares problem,

$$\sum_{t=1}^{T}(y_t - \mu_t - s_t)^2 + \lambda_1\sum_{t=3}^{T}(\Delta^2\mu_t)^2 + \lambda_2\sum_{t=13}^{T}(\Delta^{12}s_t)^2 + \lambda_3\sum_{t=12}^{T}\left(\sum_{j=0}^{11}s_{t-j}\right)^2$$

where $\lambda_i = \tau_i^2/\sigma^2$. μ_t and s_t are the parameters in some original sense, while τ_i^2 is the parameter that governs the probability distributions of the original unknown parameters. Hence it is called the hyper parameter which is usually estimated by the method of maximum likelihood based on the marginalized likelihood. One of the software implementation of the idea explained so far is BAYSEA by Akaike and Ishiguro (1980).

2.2. Use of State Space Form

Introducing a state space form is much more advantageous than a constrained least squares point of view to understand the structure of unobserved components models. To save type setting space, we consider a quarterly time series case in which we assume the second order trend model $\Delta^2 \mu_t = v_{1t}$ and the summation type seasonal component model $\sum_{j=0}^{3} s_{t-j} = v_{2t}$ to guarantee the identifiability of the components. The observed time series y_t essentially consists of two unobservable components μ_t and s_t. Note that μ_t and s_t depend on $x_{t-1} = (\mu_{t-1}, \mu_{t-2}, s_{t-1}, s_{t-2}, s_{t-3})'$. The vector x_t constitutes the essential set of information that describe the probability law of the time series, hence is often referred to as the state vector. Let $v_t = (v_{1t}, v_{2t})'$, then the assumed constraints can be rewritten as the transition equation of x_t as $x_t = Fx_{t-1} + Gv_t$, where the matrices F, G are given by

$$F = \begin{pmatrix} F_1 & O \\ O & F_2 \end{pmatrix}, G = (G_1, G_2),$$

where

$$F_1 = \begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}, F_2 = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, G_1 = (1,0,0,0,0)', G_2 = (0,1,0,0,0)'$$

respectively. This is called the system equation.

Once the value of the state x_t realized, the time series y_t is observed as the sum of the part of state variables and the present irregular variation ε_t. Namely, the way we obtain the observation can be expressed as $y_t = Hx_t + \varepsilon_t$ by the matrix $H = (1,0,1,0,0)$. This is referred to as the observational equation. We say we have a state space form of a given model when both the system and the observational equations become available.

Once we obtain a (linear Gaussian) state space form, given the value of hyperparameters, we can estimate the state x_t by using the recursive algorithm called the Kalman filter. The values of the hyperparameters can be calibrated by MLE, defining the likelihood by the accumulation of one-step ahead prediction error $y_t - Hx_t$. See Kitagawa (1993) for example. Kitagawa's DECOMP is one of the software implementation based on the idea stated above. (See Kitagawa (1981).) What is actually used in the empirical analysis of this paper is E-DECOMP which is implemented as the add-on macro works on Microsoft Excel.

2.3. Preliminary Data Analysis

Figure 1 shows the plots of the estimated trend values of Chlorophyll-A by the 46 observation locations, and also by a fixed time. The blue line corresponds to April 1980, red to April 1990, and green to April 2000, respectively. Note that the horizontal axis is not the time but just the nominal number of the measurement sites. So these are not time series plots. Having said so, left in the horizontal axis generally corresponds to the north of the lake, and right to the south vice versa. Change in time is reflected on the shift of the curves; they are shifted downwardedly as time goes, which shows that the concentration of Chlorophyll-A is recently

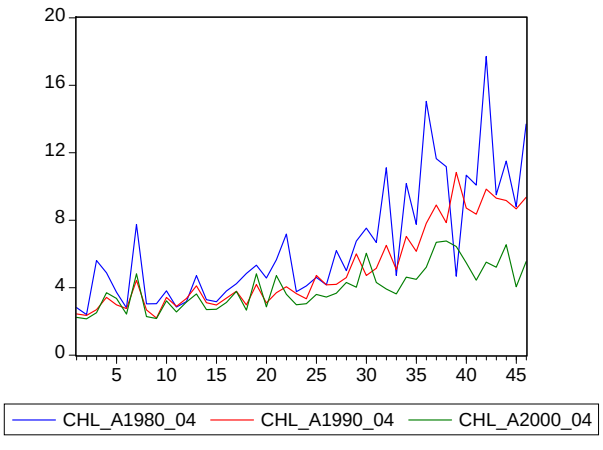

Figure 1. Secular change of the concentration of Chlorophyll-A.

gradually diminishing.

The concentration of Chlorophyll-A is a barometer of phytoplankton growth. The higher the concentration, the worse the water quality and the more probable we have the danger of red tides (*Akashio*) and blue-green algae (*Aoko*). One of the background for the recent low concentration of Chlorophyll-A can be sought for the decrease of effluent impact, most of which could be accounted for the improvement of sewerage system of the towns around the lake.

Based on the trend estimates obtained by the observation sites, we can draw a contour plot of the Chlorophyll- A concentration for a given time point, by performing spatial smoothing and interpolation. Contour plots in Figure 2 are, so to say, made by embedding the data drawn in Figure 1 in to a map (and by smoothing).

Blue colored area shows that there we have relatively good quality of water. Looking at these three panels give us an impression that in the last two decades there has been a substantial improvement of the water quality in Lake Biwa. Comparing April 1980 and April 1990, we visually confirm the remediation in north area of the lake, while the south still had some room for improvement. Setting April 2000 against April 1990, we see the last decade witnessed the water quality improvement in south area, too.

Three contour plot panels in Figure 2 are drawn by using GMT, The General Mapping Tools. (http://gmt.soest.hawaii.edu/) Generally speaking, spatial smoothing and interpolations heavily depend on the choice of basis functions and tuning parameters. Admittedly, doubts should be casted to the colors of area where the data is sparse. However, the aim of these drawings here is a sort of rough visual check of Chlorophyll-A concentration in every 10 years up to 2000. We just tried 2D visualization of 1D plots in Figure 1, and do not mean to assert these show a sort of 'optimal' interpolations.

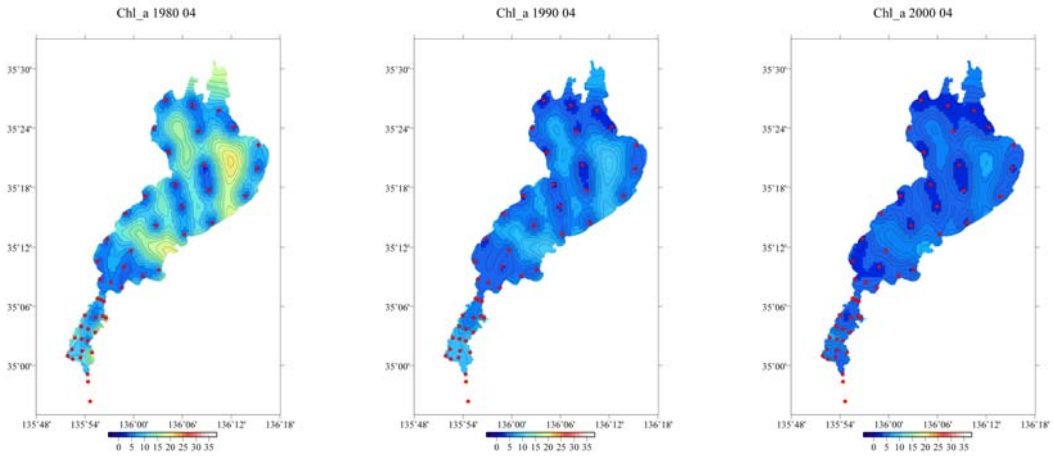

Figure 2. Contour plots of the concentration of Chlorophyll-A (based on trend estimates).

Though such a graphical representation like this helps our understandings a little, we cannot adopt them as a scientific evidence of the existing of trend as it depends many tuning parameters and factors. To make things more formal, we have to resort to some formal testing procedure to validate the existence of trends. In the next section, we report the results of a couple of unit root tests applied for the trend series extracted from Chlorophyll-A data.

3. UNIT ROOT TESTS

3.1. Review of Procedures

Unit root tests are originally developed in the context of econometric analysis of time series. main concern lies in the following question; trend in economic time series, is it a deterministic function of time or is it

essentially a stochastic trend characterized by, for example, a random walk model? We want to decide based on data, which necessitates formal statistical tests on the characteristics of time series.

In other words, a deterministic trend model has a trend that increase or decrease at a fixed increment. If it is true, the trend part of the series can be predictable. On the other hand, if the trend is stochastic, it is impossible to tell to which direction the trend moves. As the prediction horizon becomes longer, the larger the prediction error bound becomes.

At least in an economic sense, these two kinds of trends have different implication. However, we do not have to care the difference. Whether deterministic or stochastic, we are more interested in the existence of trends than the nature of trends.

3.2. Test Procedure

In fact, we include a constant term and the linear time trend in a regression equation, and performed the following two kinds of tests.

- Augmented Dickey-Fuller (ADF) Test
- Kwiatkowski, Phillips, Schmidt, and Shin (KPSS) Test

We are going to make some brief remarks on the test procedure used here.

The name ‘unit root test’ comes from the simple fact seen subsequently. Suppose that the time series y_t is truly generated by the following model, $y_t = \rho y_{t-1} + \varepsilon_t$ where ε_t is i.i.d. $N(0, \sigma^2)$ random variable. By introducing lag operator L defined by $Ly_t = y_{t-1}$, the above model has an alternative expression $(1-L)y_t = \varepsilon_t$. Generally we can put the dependence structure of a time series into the corresponding polynomial of lag operator L. We investigate the behavior of time series model through the analysis of the polynomial associated with y_t (The expression for this example might be too simple for us to understood its meaningfulness).

This associated polynomial is sometimes called the characteristic polynomial. If the solution to the equation setting the characteristic polynomial equal to zero lies just on the unit circle, the time series is said to follow a unit root process. Repeating simple backward substitutions, we obtain

$$y_t = \sum_{j=0}^{\infty} \varepsilon_{t-j} .$$

Apparently, the time series y_t is expressed as the accumulation of independent noise. The term ‘stochastic trend’ comes from the right hand side of this expression.

To determine whether or not time series has the stochastic trend usually rely on the estimates of the autoregressive parameters in a linear model. The tool to be employed for this aim is unit root test. Common specification is $y_t = \alpha + \delta t + \rho y_{t-1} + \varepsilon_t$, and we test the null hypothesis of $\rho = 1$. This is so-called Dickey-Fuller test (DF test hereafter), and an extended version where the serial correlation in the innovation process is allowed is referred to as the Augmented Dickey-Fuller test, ADF test hereafter. See Dickey and Fuller (1979, 1981). In our analysis, we confirm the existence of trend as long as δ is significant even when ρ is judged to be less than unity. Once again we do not care about the characteristic of trends. When the unit root hypothesis is accepted and at the same time δ is significant, we have a deterministic trend plus a stochastic trend.

In the ADF test, the null hypothesis is ‘unit root.’ To validate the test from another side, we also perform the test where the null of ‘no unit root’. One of such tests is so-called KPSS test proposed by Kwiatkowski, Phillips, Schmidt, and Shin (1992). The essential idea of KPSS test is as follows; if there is no stochastic trend, the innovation process of the stochastic trend must be degenerated. If we write this in a model, y_t has the following data generating structure,

$$y_t = \alpha + \delta t + \eta_t + u_t ,$$

$$u_t = u_{t-1} + \xi_t,$$

where η_t stands for a stationary process. It is apparent that the unobserved component u_t corresponds to the stochastic trend generated by the innovation process ξ_t. We test whether or not $\sigma_\xi^2 = 0$ where σ_ξ^2 stands for the variance of ξ_t. When it is the case, the second equation results in a difference equation, and u_t is eventually absorbed into the constant term α.

Once again, we do not care whether $\sigma_\xi^2 = 0$ or $\sigma_\xi^2 > 0$, even accept both. If $\sigma_\xi^2 > 0$, we confirmed the existence of a stochastic trend. Even if $\sigma_\xi^2 = 0$, we could validate the existence of a deterministic trend as long as the estimated δ is significantly different from zero.

We will close this subsection by mentioning a few remarks on the specification of tests. The lag length in ADF test is determined by AIC. The choice of kernel function in the estimation of spectral density in KPSS test is Bartlett kernel. This is the default setting in EViews. The bandwidth is chosen according to the Newey-West automatic variable bandwidth selection, which is again the default in EViews.

3.3. Results of Data Analysis

In this subsection, we report the results on Chlorophyll-A series (Chl_a), a monthly time series observed from April 1979 to Mach 2003. We cannot determine in advance whether we should use the estimated trend only or to work on the seasonally adjusted series. In this paper, we report both results.

Test results are reported only for 19 out of 46 monitoring locations. 13 sites are chosen from the north lake, the rest from the south lake. (See Figure 3.) Every monitoring station is marked as P followed by a two digits number. To list the selected locations, P04, P05, P06, P07, P11, P12, P13, P17, P18, P19, P23, P24, P25 in the north, P36, P37, P38, P44, P45, P46 in the south.

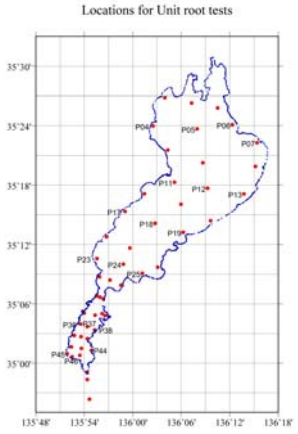

Figure 3. Location of monitoring sites.

Table 1 shows the test results on trend estimates, while Table 2 for the seasonally adjusted series. In the table, the columns of ADF and KPSS report the values of test statistic, and their levels of marginal significance are indicated by asterisk(s): *** (1%), ** (5%), and * (10%). The column of 'Trend' reports the t-statistic of the estimated coefficient of linear trend.

Generally speaking, we witnessed the existence of trend at almost every site. If we regard the estimated trend as data, the results often support the existence of deterministic trend. On the other hand, when we use the seasonally adjusted data, stochastic trend is supported in many cases.

The reason is rather obvious. If we rely on the extracted trend only, sometimes we work on very smooth data. Then data does not contain much variability, and the deterministic time trend is sufficient to describe the time series data. Seasonally adjusted series, to the contrary, contains irregular part, so it becomes more difficult to separate the irregular part from trend. In such a case, stochastic trend is so flexible that it can show better fit to the data. It is conjectured that by this reason the stochastic trend is preferred for the seasonally adjusted data. What remains to be seen is which trend leads to the natural interpretation from the view point of environmental science.

Some monitoring sites show the contradiction between the results of ADF test and KPSS test. P44 for the test based on trend (See Table 1), while P07, P18, P19, P23, P38, P44 for the test based on seasonally adjusted data (See Table 2). This is probably due to the effect of outliers, and it is difficult to demonstrate the existence of trend by just applying simple form of unit root test. This issue is also a future challenge.

Table 1: Trend series

	ADF	Trend	KPSS	Trend
P4	-2.17	1.31	0.22***	-5.99***
P5	-1.64	0.75	0.21**	-4.43***
P6	-3.22*	-3.81***	0.08	-18.42***
P7	-2.10	-0.90	0.36***	-12.11***
P11	-3.18*	-2.43**	0.08	-10.31***
P12	-2.64	1.77*	0.18**	7.08***
P13	-2.21	-0.54	0.17**	-5.17***
P17	-2.24	1.05	0.19**	0.73
P18	-2.49	-0.96	0.21**	-14.30***
P19	-2.37	2.04**	0.33***	0.27
P23	-2.72	-1.05	0.19**	-5.02***
P24	-3.21*	-0.32	0.22***	-9.77***
P25	-2.11	-0.92	0.17**	-4.85***
P36	-1.95	-2.13**	0.21**	-12.44**
P37	-1.82	-1.05	0.23***	-15.78***
P38	-3.18*	-1.85*	0.28***	-31.50***
P44	-4.29***	-3.01***	0.18**	-22.83***
P45	-1.98	-1.84*	0.36***	-19.75***
P46	-1.82	-1.61	0.22***	-28.75***

Table 2: Seasonally adjusted series

	ADF	Trend	KPSS	Trend
P4	-15.84***	-1.66*	0.04	-1.65*
P5	-13.72***	-1.04	0.08	-1.18
P6	-7.84***	-1.94*	0.04	-2.42**
P7	-10.45***	-2.65***	0.24***	-4.16***
P11	-15.10***	-2.22**	0.05	-2.34**
P12	-7.14***	0.60	0.06	0.04
P13	-14.74***	-1.12	0.07	-1.24
P17	-13.57***	-0.74	0.05	-0.80
P18	-12.94***	-2.38**	0.17**	-3.10***
P19	-15.19***	-0.57	0.19**	-0.62
P23	-16.35	-1.69*	0.06	-1.71*
P24	-14.04	-1.34	0.04	-1.75*
P25	-14.89***	-1.24	0.09	-1.48
P36	-15.91***	-3.19***	0.08	-3.46***
P37	-15.48***	-2.43**	0.12	-2.56**
P38	-14.36***	-5.62***	0.23***	-7.27***
P44	-11.64***	-2.71***	0.15**	-4.16***
P45	-11.01***	-3.19	0.10	-5.22***
P46	-9.23***	-3.08***	0.05	-4.31***

ACKNOWLEDGMENTS

This research was partially supported by Research Organization of Information and Systems, Transdisciplinary Research Center, 'Function and Induction' Research Project. It was also supported by JSPS Grant-in-Aid for Scientific Research (C), No. 19500248.

REFERENCES

Akaike, H. and M. Ishiguro (1980), *BAYSEA, A Bayesian seasonal adjustment program, Computer Science Monographs No.13*, The Institute of Statistical Mathematics, Tokyo.

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, 427-431.

Dickey, D. A. and Fuller, W. A. (1981), Likelihood ratio statistics for autoregressive time series with a unit root, *Econometrica*, 49, 1057-1072.

Kitagawa, G. (1981), A nonstationary time series model and its fitting by a recursive filter, *Journal of Time Series Analysis*, 2, 103-116.

Kitagawa, G. (1993), *FORTRAN 77 Programming for Time Series Analysis (in Japanese)*, Iwanami Shoten, Tokyo.

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, *Journal of Econometrics*, 54, 159-178.