The Effects of Variable Stationarity in a Financial Time-Series on Artificial Neural Networks

Matthew Butler and Dimitar Kazakov
Artificial Intelligence Group,
School of Computer Science,
University of York, York, UK, YO10 5DD
mbutler,kazakov @cs.york.ac.uk

***Abstract*—This study investigates the characteristic of non-stationarity in a financial time-series and its effect on the learning process for Artificial Neural Networks (ANN). It is motivated by previous work where it was shown that non-stationarity is not static within a financial time series but quite variable in nature. Initially unit-root tests were performed to isolate segments that were stationary or non-stationary at a pre-determined significance level and then various tests were conducted based on forecasting accuracy. The hypothesis of this research is that when using the de-trended/original observations from the time series the trend/level stationary segments should produce lower error measures and when the series are differenced the difference stationary (non-stationary) segments should have lower error. The results to date reveal that the effects of variable stationarity on learning with ANNs are a function of forecasting time-horizon, strength of the linear-time trend, sample size and persistence of the stationary process.**

I. Introduction

A ubiquitous characteristic of a financial time-series is that of non-stationarity which means that the moments (i.e. mean and variance) of the system are not static in time. The consequences of modelling a non-stationary series are well documented in the statistical modelling literature [10] [12] [5], where a violation of the stationarity assumption of these techniques leads to spurious results. A common tactic for countering non-stationarity is to difference the data to some degree so that the original series $\{Y_t\}$ is transformed to $\{\Delta^P Y_t\}$, where P is the degree of differencing. A popular statistical model that utilizes this process is the autoregressive integrated moving average (ARIMA) where the degree of integration (differencing) is determined to transform the series to stationary. In financial time-series analysis specifically in computational intelligence (CI) a common pre-processing step for input attributes is converting the observations to continuously compounded log-normalized returns, where X_t = log(Y_{t+1}/Y_t) or arithmetic returns over some time period (X_t = (Y_{t+1} - Y_t)/Y_t). The advantage of using logarithmic returns is that they are symmetric and for small returns are very similar to arithmetic. There is strong evidence to support the procedure of differencing; in figure 1 we have a dual representation of 3 major stock market indices, where on the top we have a plot of the daily closing prices and at the bottom a plot of the first differences of the same series. It is quite obvious that the daily closing prices are not stationary but the first differences appear to be. This pre-processing step is routinely performed without consideration for the local characteristics of the time-series. Based on unit-root testing over a long time-horizon many of the financial markets are non-stationary [8] [6]. Considering results obtained in Butler et al. [3], the presence of a unit-root (a necessary condition for a process to be non-stationary) is not static and tends to fluctuate with time. Figure 2 is a plot of p-values obtained from the Augmented Dickey Fuller (ADF) test, a popular unit root test that has the null hypothesis that the series contains a unit-root, using a sliding window approach. We can see that the p-values fluctuate between 0.99 and 0.01 implying that the series goes through cycles where at times it contains a unit root at the 99% confidence interval and at other times can reject the presence of a unit-root at the 99% confidence interval. This result has implications for forecasting, where the optimal pre-processing step is a function of the characteristics of the time-series. If there are stationary localities within a time-series, what effect does the application of differencing the data, when the market in trend stationary, have on the robustness of the forecasting models? In the traditional time-series literature this question has already been asked, where in Diebold et al. [4] the usefulness of using a unit-root test for selecting a forecasting model was explored. The experiments were conducted with several simulations of an AR(1) process that mimicked GNP data with increasing levels of persistence. The AR(1) model used was:

$$(Y_t - a - bt) = \rho(Y_{t-1} - a - b(t-1)) + \epsilon_t \qquad (1)$$

where $\epsilon_t \sim$ N(0,σ^2), a = 7.3707 and b = 0.0065. The authors applied an ADF test to the series as a pre-test using the 5% finite-sample critical values to determine if a unit-root was present. The results strongly suggest that using a pre-test improves the forecasting accuracy of the models, which implies that the presence of a unit root is not static.

Within computation intelligence (CI) there has been a prior study which focused on the effects of non-stationarity on Artificial Neural Networks [7]. In this work the possibility of variable stationarity was not explored nor the implications of non-stationarity based on forecasting time-horizon or sample size. The general conclusion was counter-intuitive in that over-fitting the training data was a superior methodology to using a validation set, however these conclusions are based

978-1-4244-9932-8/11/$26.00 ©2011 IEEE

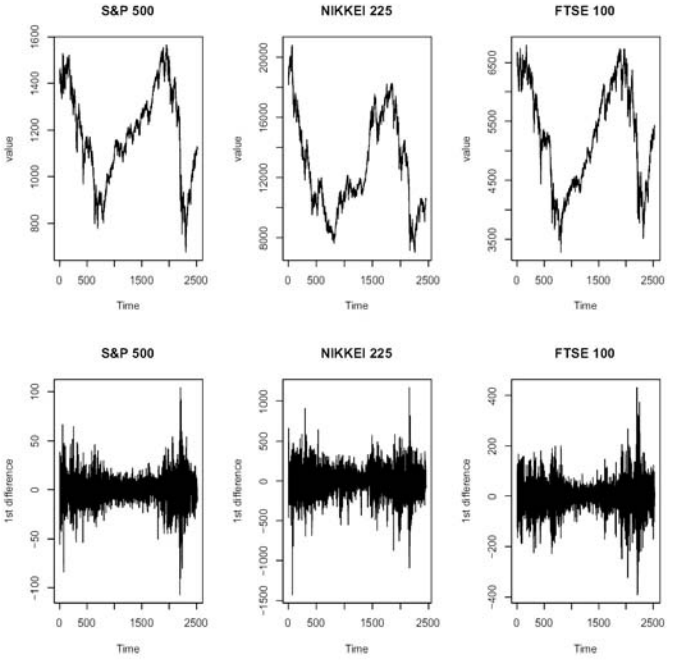

Fig. 1. (top) A plot of the S&P 500, NIKKEI 225 and FTSE 100 market indices using the daily closing prices over a 10 year period. (bottom) A plot of the first differences of the same market indices.

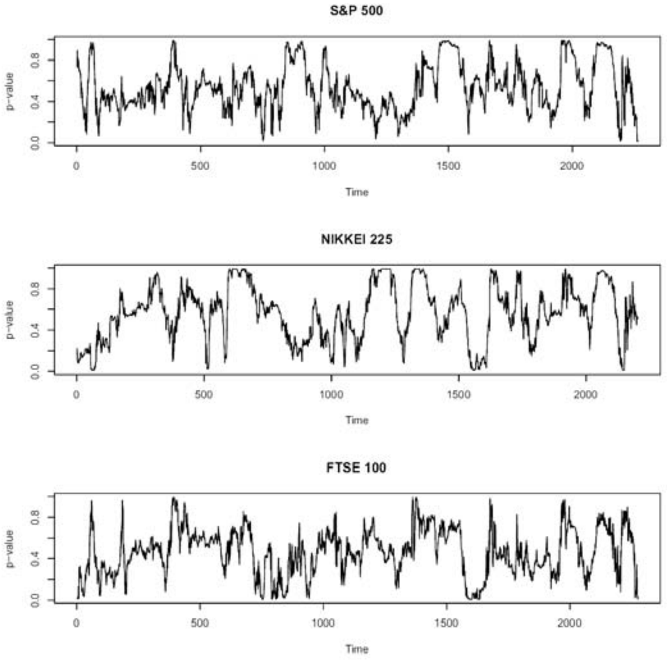

Fig. 2. Plots of the p-values generated from an ADF test using a sliding window of 250 observations for three stock market indices.

on sample auto-correlation functions (SACF) of the residuals rather than on forecasting error (i.e. MSE or RMSE). Based on this prior work we are left with the questions as to the effects of routinely differencing financial time-series data on the forecasting abilities of ANNs and whether we can improve upon these forecasts taking into account the temporal nature of stationary localities. To facilitate this exploration we will analyse the effects of stationary and non-stationary time-series on out-of-sample forecasts from ANNs using simulated and real-world time-series data over varying sample sizes and forecasting time-horizons. From [11] we know that near unit-root process for short forecasting time-horizons still benefit from differencing even though the series is technically trend or level stationary and therefore it will be meaningful to experiment with increasing forecast time-horizons.

Comparing to previous work [3], [4] and [7] on stationarity within financial time-series our contributions consist of 1) demonstrating the effects of stationarity on ANNs with simulated data, 2) applying the unit-root tests to real-world data as a filter for pre-processing steps, 3) testing the effect of varying unit-roots on ANN out-of-sample predictions, 4) providing insight into the preferred window size for the ADF test from a filtering perspective and 5) evaluating the result that over-fitting is the preferred method for training in terms of prediction error.

II. Stationarity

An important distinction between systems characterised by stationary vs. non-stationary time series is how they respond to shocks. If the system is stationary then any shocks will be temporary and will dissipate with time. However with a non-stationary time series the shocks have a more permanent effect on the characteristics (moments) of the series. Put more formally, consider a time series $\{Y_t\}$ and an AR(1) model with drift:

$$y_t = a_0 + a_1 y_{t-1} + \epsilon_t \tag{2}$$

where y_t is the observation of the series at time t, a_0 is an intercept or drift term, a_1 is a constant, and ϵ_t is white noise. The process is stationary *iff* the constant $a_1 < 1$, so as to enable the system to return to a stable trend. If $a_1 = 1$ then the equation simplifies to:

$$y_t = a_0 + y_{t-1} + \epsilon_t \tag{3}$$

which is a random walk with drift model which is non-stationary and any shocks to the system will have a permanent effect. To observe why shocks to the system have a lasting effect we can solve for y_t given an initial condition y_0, so:

$$y_t = y_0 + a_0 t + \sum_{i=1}^{t} \epsilon_i \tag{4}$$

here the value of y_t is governed by a deterministic trend $a_0 t$ which depends solely on time and a stochastic trend Σ_{ϵ_i}, that is an accumulation of all previous shocks to the system from the initial conditions until time t. The presence of a unit root has important consequences for modelling and forecasting a time series where non-stationary behaviour increases the complexity of the task and violates assumptions during a regression with OLS which assumes the data generating process (DGP) is stationary. For these reasons it is important to determine the characteristics of the time series under study in order to transform it to stationary via differencing or de-trending.

III. Unit-root tests and ANNs

This section will outline the theory of the ADF test and explain the ANN utilized in this study.

A. Augmented-Dickey Fuller Test

The ADF test is an extension of the Dickey-Fuller test which was created to handle more complex time-series, where the time-series is fitted or represented by a differenced AR(P) process of one of three possible forms. The AR(p) model with p=1 and the three possible forms of the ADF test are:

$$y_t = a_1 y_{t-1} + \epsilon_i \tag{5}$$

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \epsilon_i \tag{6}$$

$$\Delta y_t = a_0 + \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \epsilon_i \tag{7}$$

$$\Delta y_t = a_0 + \gamma y_{t-1} + a_2 t + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \epsilon_i \tag{8}$$

Equation 5 is a simple AR(1) model, if we assume a unit-root process for the difference equations then equation 6 is a simple random walk, equation 7 is a random walk with drift and equation 8 is random walk with drift and a linear time trend. The process to determine if the time series contains a unit root (i.e. $a_1 = 1$) entails estimating the coefficients (a_0, a_2 and β) for one of the equations above using ordinary least squares (OLS). Let γ = a_1 - 1 and so testing $a_1 = 1$ in the AR(1) model is equivalent to testing γ = 0 in the difference equations. The estimate of γ along with its standard error (SE) will form a t-statistic which can be evaluated against the critical values in the Dickey-Fuller tables. The effectiveness of the ADF test is sensitive to the choice of P the lag parameter. There are two popular methods for choosing the lag parameter, the first is using an information criterion such as Akaike's information criterion (AIC) and the other is based on the statistical significance of the estimated coefficients, as suggested by Ng and Perron [9], denoted the NP method. In [9] the authors found that the use of an information criterion tended to underestimate the correct number of lags in the autoregressive equation and therefore the power of ADF test suffers. The NP method involves the follow steps:

1) Choose the maximum lag (P_{max}), given by equation 9 as proposed by Schwert [11] where T is the sample size or the number of observations per rolling window.

$$P_{max} = \left[12 \times \left(\frac{T}{100}\right)^{1/4} \right] \tag{9}$$

2) Set the order of the AR model lag(p) = P_{max}
3) Fit the AR model which will be used in the ADF test using p lags
4) Calculate the t-statistic for lag(p), given by equation 10

$$t_{stat} = \hat{\phi} / SE(\hat{\phi}) \tag{10}$$

where $\hat{\phi}$ is the estimated coefficient using OLS and SE($\hat{\phi}$) is its standard error.

5) If lag(P) is significant then set the number of lags in the ADF test to P, otherwise repeat steps 2-5 for p ← p - 1 and so forth until an estimated lag coefficient is significantly different from zero.

This procedure is more effective in the general → specific direction, as in the reverse the number of lags is generally underestimated. This is due to the fact that having a lag k which is significant does not imply that all lags $< k$ will also be significant.

B. Artificial Neural Networks

Since variable stationarity would most likely effect any algorithm from Artificial Intelligence (AI) in time-series prediction it seems appropriate from a CI perspective to start the analysis of this phenomenon on what is arguably the most robust modelling technique from AI for forecasting financial time-series, i.e. Artificial Neural Networks [2] [13] [1]. ANNs are considered a non-linear function approximaters which makes them attractive for complex time-series, where studies have shown that modelling non-linear data with ANNs produces more accurate models than their linear counterparts. For this study we will be experimenting with a feed-forward multi-layer perceptron with two hidden layers and the canonical back-propagation for updating connection weights. The topology of the ANN used with the real-world data is shown below in figure 3, where we have a (5,2,4,1) architecture with 5 inputs representing 5 lagged values of the time-series. For the simulated data the topology is slightly changed to (1,2,4,1) to reflect the need for only the most recent lagged value as the we know a priori what the underlying DGP is. The hidden nodes have sigmoid activation functions and the output node is linear as we require a real-valued output. The linear activation function is:

$$\hat{Y} = \sum_{m=0}^{L} X_m W_m \tag{11}$$

where $\hat{Y}$ is the output of the ANN, X_m is the output of neuron m the final hidden layer, W_m is the connection weight and L is the number of neurons in the final hidden layer.

IV. Data Description

The data considered for this study are daily observations of the adjusted closing prices from 18 market indices from around the world. The sample spans a 10 year period from the beginning of January 2000 to the end of December 2009. All observations are pre-processed prior to training based on experiment objectives. The different pre-processing steps are described below. Table I displays the 18 market indices (and their corresponding country) and the number of observations for each. Although the time span is of equal length the number of trading days for each market varies slightly.

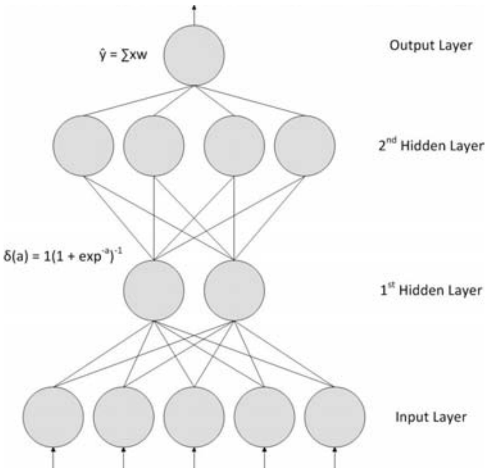

Fig. 3. The feed-forward MLP topology used in this study, displaying the sigmoid activation function in the hidden nodes and the linear activation function in the output node.

TABLE I
THE 18 MARKET INDICES CONSIDERED IN THE STUDY AND THE NUMBER OF OBSERVATIONS FOR EACH. THE SAMPLE PERIOD IS FROM JANUARY 2000 TO DECEMBER 2009.

Country	Market	obser.	Country	Market	obser.
US	S&P 500	2515	Indonesia	Jakarta	2411
Korea	KOSPI	2464	Malaysia	KLSE	2464
Taiwan	TSEC	2465	Argentina	MerVal	2466
Japan	Nikkei 225	2454	China	Shanghi	2579
Singapore	Strt. Time.	2509	UK	FTSE 100	2526
Hong Kong	Hang Seng	2489	France	CAC 40	2553
Brazil	Bovespa	2472	Germany	DAX	2543
Mexico	IPC	2506	Canada	TSX/S&P	2515
India	BSE 30	2474	Australia	ASX	2538

V. EXPERIMENT DESIGN

The main objective of this study is to examine the effects of stationary and non-stationary time-series samples on out-of-sample forecasting of an ANN. To facilitate this goal the ANNs will be trained with input data of three forms:

- First differenced (DD),
- De-trended by removing a linear time trend (DT), and
- No pre-processing (OB).

These three forms represent the preferred pre-processing method for input data when a time-series is difference, trend and level stationary respectively. The experiments will be conducted on data sets of increasing size and time-horizon. The samples from the real-world data will be collected with a sliding window approach which moves in increments of 1 day at a time. For a time series $\{Y_t\}$ and a window of size d an initial sub-sample is created consisting of observations $\{Y_{1,2..,d}\}$. Next the appropriate tests are run and then the window shifts by one day to cover $\{Y_{2,3..,d+1}\}$ and so forth until the end of the sample. If a sample is found to generate a statistically significant p-value at some specified limit then the sample is stored and a linear-time trend is evaluated. The algorithm for determining stationary and non-stationary windows is as follows:

Algorithm 1 Identify Stationary/Non-stationary Windows

```
input: Data set D
input: $P_{max}$
input: windowSize
input: upper/lower significance levels(sigLevU, sigLevL)
for i← 1:length(D)-windowSize do
    sample(S) ← i:i+windowSize
    apply NG method to S
    return Optimal lag L
    apply $ADF_{test}$(L) to D
    return p-value($p_{val}$)
    if $p_{val}$ < sigLevL then
        store S to SW
        Trend ← regress S on time
        store Trend to ST
    end if
    if $p_{val}$ > sigLevU then
        store S to NS
        Trend ← regress S on time
        store Trend to NST
    end if
end for
return SW, NS, ST, NST
```

SW and NS are matrices of stationary and non-stationary data samples respectively and ST and NST are vectors of linear time trends for the samples stored in SW and NS respectively.

The above process is only applied to the real-world data as the simulated data will be generated to adhere to the intended time-series characteristics. Once the relevant data has been collected, it is split into two equal sets for training and testing. The evaluation of the results will be performed with four commonly used error measures: MSE, RMSE, MAPE, and MAE. The simulated data will be generated with equation 1 with increasing levels of persistence, with ρ being equal to $\{0.5,\ 0.8,\ 0.99,\ \text{and}\ 1\}$. Each realization will contain 2000 observations with a 50/50 split for training and testing, re-arranging equation 1 to have only Y_t occupy the left hand side yields:

$$Y_t = (a - a\rho + b\rho) + (b - b\rho)t + \rho Y_{t-1} + \epsilon_t \qquad (12)$$

VI. SIMULATED DATA RESULTS

This section will detail the results from training and testing on the simulated data. In all of the tables PP_{step} stands for pre-processing step, OBS, DET and DIF stand for observation, de-trended and differenced data respectively. All simulations are generated with a = 7.3707 and b = 0.065. Only MSE is reported for brevity as the results were generally consistent across all four error measures, however full results are available upon request. Figure 4 displays 4 plots of 2000 realizations from equation 1 with a ρ = $\{0.50,\ 0.80,\ 0.99,\ 1.00\}$.

A. Stationary Process with ρ = {0.5, 0.8}

Table II displays the testing results from the three different pre-processing steps for each time-horizon. As expected with a trend stationary process with low persistence the error from de-trending the data is the lowest.

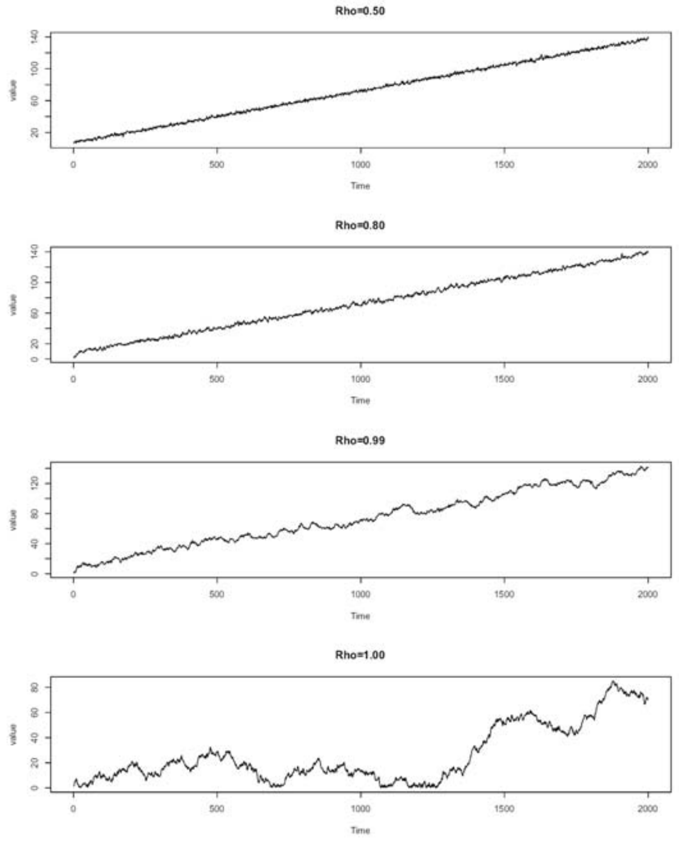

Fig. 4. A plot of 1 realization for each of the various levels of persistence generated from Y_t = (a - aρ + bρ) + (b - bρ)t + ρ Y_{t-1} + ϵ_t.

TABLE II
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE SIMULATED RESULTS GENERATED FROM EQUATION 1, WHERE $\epsilon_t \sim$ N(0,σ^2), A = 7.3707, B = 0.065 AND ρ = 0.50.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	1548	1555	1564	1578	1626	1712
DET	1.06	1.36	1.36	1.36	1.38	1.38
DIF	1.35	2.41	2.81	3.84	3.83	13.40

In table III we have the testing results from 2000 realizations of the simulated data with ρ = 0.80, so this series represents a more persistent stationary process with a trend, however the trend is not as strong as a result of the equation construction. We have similar results, with the only departure coming from a time-horizon of 1 day, where differencing the data yielded a smaller testing error.

TABLE III
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE SIMULATED RESULTS GENERATED FROM EQUATION 1, WHERE $\epsilon_t \sim$ N(0,σ^2), A = 7.3707, B = 0.065 AND ρ = 0.80.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	1547	1558	1572	1583	1631	1713
DET	2.08	2.85	2.93	3.02	3.15	2.92
DIF	1.26	5.23	5.72	6.02	6.60	8.44

B. Stationary Process with ρ = 0.99

This series will be generated twice (i) first using the equation above and (ii) secondly with the trend term slightly altered to allow for a stronger trend. A stronger linear time trend is achieved by replacing *(b - ρb)t with* bt in equation 12. The second scenario was chosen to explore the effects of the strength of the trend, in [11] the effects of pre-processing were not explored for stronger trends in near unit-root processes. The results from scenario (i) are displayed in table IV and once again we see that for shorter time-horizons the pre-processing step of differencing produces the superior results. As the time-series becomes more persistent the time-horizons where de-trending is beneficial get pushed further out.

TABLE IV
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE SIMULATED RESULTS GENERATED FROM EQUATION 1, WHERE $\epsilon_t \sim$ N(0,σ^2), A = 7.3707, B = 0.065 AND ρ = 0.99.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	1731	1741	1764	1785	1831	1925
DET	2.76	16.71	11.77	20.90	28.98	34.26
DIF	1.13	5.99	22.31	30.04	68.95	150.68

TABLE V
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE SIMULATED RESULTS GENERATED FROM EQUATION 1 SUBSTITUTING IN A STRONGER TIME TREND, WHERE $\epsilon_t \sim$ N(0,σ^2), A = 7.3707, B = 0.065 AND ρ = 0.99.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
DET	299	419	415	364	628	1999
DIF	23	692	2563	6964	4.7x10^4	2.34x10^5

In table V we have the results from scenario (ii), the testing errors from using the observations with non pre-processing have been omitted due to the poor nature of the results (they are available upon request). Using a stronger trend reveals that it does effect the performance of the ANN with respect to the de-trending or differencing the data. When a stronger trend is present, the benefits of de-trending become apparent with shorter time-horizons. However, for 1-day ahead predictions, differencing is still the preferred option.

C. Non-Stationary Process with ρ = 1

The results from testing in the out-of-sample data from a non-stationary time-series is provided in table VI.

From the results we can see that when forecasting a difference stationary process, differencing the data leads to the

TABLE VI
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE SIMULATED RESULTS GENERATED FROM EQUATION 1, WHERE $\epsilon_t \sim$ N(0,σ^2), A = 7.3707, B = 0.065 AND ρ = 1.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	3.24	67.60	47.58	148.24	408.69	1569.34
DET	110.59	136.97	126.20	167.33	219.46	308.99
DIF	1.15	6.13	17.20	45.07	155.11	519.84

optimal forecasting results for ANNs up to the 50-day time-horizon. An interesting result is that even though a non-zero linear time-trend was estimated (0.05253) it was not actually present and using it to de-trend the data lead to superior results at the 100-day time-horizon.

This concludes the experimentation on the simulated data, we have shown that stationarity, or lack there of, effects the performance of the ANN and that proper consideration for the characteristics of the time series leads to improved forecasting performance. The choice of the proper pre-processing step is a function of stationarity, forecast-time horizon, persistence and the presence and strength of a linear time trend.

VII. Stock Market Data

This section will detail the results from applying the above procedure to sub-samples of actual stock market index data.

A. *P-Values = {0.99, 0.01}*

The first step is to isolate the opposite ends of the stationarity spectrum and apply the ANN to these samples. In table VII we have the results from collecting stationary samples (p-value $\leq$ 0.01) based on an ADF test being applied with a sliding window of 100 observations. We are reporting for each sample size and time-horizon experiment pair the percentage of windows for each pre-processing step that produces superior results for the ANN (therefore each column adds up to 1.0). The forecasts only go up to a 20-day time-horizon as all predictions are made within the sample and therefore longer time-horizons could not be accommodated.

TABLE VII
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE STATIONARY SAMPLES DETERMINED BY AN ADF TEST USING A 100 OBSERVATION SLIDING WINDOW.

PP_{step}	Time Horizon			
	1	5	10	20
OBS	0.0	0.016	0.040	0.060
DET	0.0	0.379	0.504	0.488
DIF	1.0	0.604	0.456	0.452

The results from the "stationary" data (stationary in shown in quotes as the actual function of the DGP is unknown and the ADF test is only an estimation) show some expected results where for shorter time-horizons (1-day) the differencing pre-processing step dominates the other two, but as the time-horizon is lengthened the benefits of de-trending are realized. However, only about half of the samples benefit from de-trending at the 20 day time-horizon. In table VIII we have the results obtained with a 100 day sliding window isolating non-stationary windows.

TABLE VIII
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE NON-STATIONARY SAMPLES DETERMINED BY AN ADF TEST (WITH P-VALUES $\geq$ 0.99) USING A 100 OBSERVATION SLIDING WINDOW.

PP_{step}	Time Horizon			
	1	5	10	20
OBS	0.0	0.127	0.306	0.361
DET	0.0	0.052	0.347	0.609
DIF	1.0	0.821	0.346	0.029

The results for short-time horizons are similar to those observed before, however as the time-horizon is increased the benefits of differencing decrease and at the longest time-horizon (20 days) we observe that the de-trending was the preferred method the majority of the time. These results are not indicative of what was achieved with the simulated data. This further suggests that the ADF test is not as effective with a 100-day observation window. The next experiments were performed in the same manner but for a 250 day observation window, these results are reported in tables IX and X.

TABLE IX
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE STATIONARY SAMPLES DETERMINED BY AN ADF TEST (WITH P-VALUES $\leq$ 0.01) USING A 250 OBSERVATION SLIDING WINDOW.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	0.0	0.0	0.009	0.027	0.027	0.023
DET	0.0	0.091	0.356	0.598	0.703	0.831
DIF	1.0	0.909	0.635	0.374	0.270	0.146

TABLE X
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE NON-STATIONARY SAMPLES DETERMINED BY AN ADF TEST (WITH P-VALUES $\geq$ 0.99) USING A 250 OBSERVATION SLIDING WINDOW.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	0.0	0.031	0.041	0.160	0.137	0.388
DET	0.0	0.0	0.027	0.182	0.662	0.598
DIF	1.0	0.968	0.932	0.658	0.201	0.014

The results from the stationary/non-stationary samples obtained from a 250 observation sliding window are more indicative of the simulated data. The results from the stationary samples reveal that at longer time-horizons the majority of the samples where more accurately predicted by de-trending and for the non-stationary windows the differencing pre-processing step dominated all the others up to time-horizons $\leq$ 20 days and for long-time horizons the de-trended data produced the superior models. The final test considers a sliding window of 500 observations, those results are reported in tables XI and XII.

The results from the 500 observation samples are similar to the 250 day samples, where the de-trended data produced the most accurate models 84.0% of the time with a forecast

TABLE XI
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE STATIONARY SAMPLES DETERMINED BY AN ADF TEST (WITH P-VALUES ≤ 0.01) USING A 500 OBSERVATION SLIDING WINDOW.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	0.0	0.0	0.0	0.0	0.0	0.0
DET	0.0	0.0	0.0	0.24	0.680	0.840
DIF	1.0	1.0	0.1	0.76	0.320	0.160

TABLE XII
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE NON-STATIONARY SAMPLES DETERMINED BY AN ADF TEST (WITH P-VALUES ≥ 0.99) USING A 500 OBSERVATION SLIDING WINDOW.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	0.0	0.005	0.019	0.025	0.184	0.230
DET	0.0	0.0	0.0	0.012	0.269	0.587
DIF	1.0	0.995	0.981	0.963	0.547	0.113

horizon of 100 days in a trend stationary sample. This compared to the 58.7% of the time in the non-stationary samples. The observation data never once produced a superior model in any time-horizon for the trend-stationary samples and in the non-stationary samples it was only the preferred method in 23.0% of the samples at the 100 day time horizon. We also observe that the difference models maintain their dominance in the non-stationary samples up to the 20-day time horizon where they were the preferred method in at least 96.3% of the samples. Additionally, at the 50-day time-horizon the difference models are preferred in 54.7% of the samples for non-stationary data but only 32.0% for stationary.

B. Summary of results for p-values = {0.01 and 0.99}

The results from the real-world data are akin to those obtained with the simulated data. The behaviour of the market is not well represented by AR(1) process and this added complexity makes the benefits of pre-processing less definitive. The following observations can be made about the real-world data results:

1) The results from using the 100 observation sliding window were counter-intuitive to the what would be expected with non-stationary samples. This leads to the conclusion that the ADF test is not effective with only 100 samples (for the markets considered).
2) Using a 250 observation sliding window produced more consistent results w.r.t the simulated data. Where in the non-stationary samples the differencing pre-processing step was dominant at the 1, 5 and 10 time horizons and still the majority top performing model at the 20-day time horizon. For the stationary windows the de-trending pre-processing step became useful at the 10-day time-horizon and was the dominant model at all time-horizons beyond that point.
3) The 500 observation window produced similar results to the 250 day observation window where the benefits of de-trending were realized at the 20 and 50-day time-horizon in the stationary data. Additionally, with non-stationary segments the benefits of differencing were realized until the 50-day time-horizon, which is consistent with the simulated results.
4) In general the results imply that at short time-horizons {1-10 days} differencing the data as a pre-processing step will produce the optimal results in out-of-sample testing. At longer time-horizons the preferred pre-processing step will depend on the stationarity characteristic but ultimately the confidence in the forecast diminishes. Regardless if the time-series is exhibiting stationary on non-stationary qualities, at long time-horizon forecasts, de-trending appears to be the optimal pre-processing step w.r.t differencing or simply using the actual observations.

C. P-Values = {0.95, 0.05}

This section will perform the same analysis as above but to samples obtained using p-values at the 5% and 95% confidence intervals.

TABLE XIII
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE STATIONARY SAMPLES DETERMINED BY AN ADF TEST USING A 100 OBSERVATION SLIDING WINDOW AND 5% SAMPLE CRITICAL VALUES.

PP_{step}	Time Horizon			
	1	5	10	20
OBS	0.008	0.036	0.069	0.056
DET	0.0	0.306	0.516	0.524
DIF	0.992	0.657	0.415	0.419

TABLE XIV
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE NON-STATIONARY SAMPLES DETERMINED BY AN ADF TEST USING A 100 OBSERVATION SLIDING WINDOW AND 5% SAMPLE CRITICAL VALUES.

PP_{step}	Time Horizon			
	1	5	10	20
OBS	0.0	0.141	0.371	0.460
DET	0.0	0.024	0.282	0.484
DIF	1.0	0.835	0.347	0.056

TABLE XV
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE STATIONARY SAMPLES DETERMINED BY AN ADF TEST USING A 250 OBSERVATION SLIDING WINDOW AND 5% SAMPLE CRITICAL VALUES.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	0.0	0.0	0.0	0.009	0.050	0.091
DET	0.0	0.046	0.160	0.475	0.740	0.758
DIF	1.0	0.954	0.840	0.516	0.210	0.151

TABLE XVI
THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE NON-STATIONARY SAMPLES DETERMINED BY AN ADF TEST USING A 250 OBSERVATION SLIDING WINDOW AND 5% SAMPLE CRITICAL VALUES.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	0.0	0.009	0.073	0.265	0.123	0.525
DET	0.0	0.0	0.005	0.192	0.498	0.447
DIF	1.0	0.991	0.922	0.543	0.123	0.027

TABLE XVII

THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE STATIONARY SAMPLES DETERMINED BY AN ADF TEST USING A 500 OBSERVATION SLIDING WINDOW AND 5% SAMPLE CRITICAL VALUES.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	0.0	0.0	0.004	0.068	0.634	0.100
DET	0.0	0.0	0.022	0.265	0.100	0.773
DIF	1.0	1.0	0.974	0.667	0.265	0.127

TABLE XVIII

THE RESULTS FROM OUT-OF-SAMPLE TESTING ON THE NON-STATIONARY SAMPLES DETERMINED BY AN ADF TEST USING A 500 OBSERVATION SLIDING WINDOW AND 5% SAMPLE CRITICAL VALUES.

PP_{step}	Time Horizon					
	1	5	10	20	50	100
OBS	0.0	0.0	0.0	0.022	0.168	0.494
DET	0.0	0.0	0.0	0.012	0.236	0.295
DIF	1.0	1.0	0.999	0.966	0.600	0.210

D. Summary of results for p-values = {0.05 and 0.95}

The results for each window size are similar where the benefits of using the observations or de-trended observations are realized as the time-horizon is extended. The benefit being that the forecasters can be more confident that they are using the optimal pre-processing method based on the local characteristics of the time-series. When the time-series is exhibiting trend-stationary characteristics there is evidence to suggest that the most accurate predictions can be obtained from de-trending when the desired time-horizon is of suitable length. Conversely, when the series is non-stationary at a significant critical value then the forecaster will lose confidence as to which method is optimal. At short time-horizons differencing will also be preferred but at longer time-horizons the preferred method is not discernible from the information obtained from a unit-root test alone. The proportion of trend stationary samples that benefit from de-trending (at any time-horizon) diminishes as the allowable confidence interval in the ADF test is widened.

VIII. CONCLUSIONS

The results from the simulated data confirm that the effects of stationary/non-stationary time-series on ANNs are similar to those of AR(p) models that were obtained in [4]. Where de-trending was beneficial in lowly persistent trend stationary processes at any time-horizon and in highly persistent trend-stationary processes, the advantages to using de-trending were a function of the time-horizon and strength of the trend. The results from the real-world data are not as clear as the simulated but in general given an appropriate window size for calculating the ADF test we observe results that roughly adhere to their simulated counter-parts. In the trend-stationary samples de-trending was the dominant pre-processing step at time-horizons of 50 days and above. Additionally the inferior performance on the real-world data sets could be a result of Type II errors, where some of the null acceptances could have been from structural breaks in the time-series. This could be overcome by using a unit-root test which allows for structural breaks, but these tests have less power.

The results from the real-world data also present conclusive evidence that overfitting one's training data is not the preferred method for training ANNs for financial time-series forecasting. Using just the observations (which was the method used in [7]) was commonly the inferior model for the non-stationary data w.r.t. differencing at forecasting time-horizons of 50-days or less. When the use of the observations was superior to differencing (at the 100-day time-horizon) there was not a conclusive method for producing optimal results. As a consequence, the use of differencing in non-stationary financial data will produce the most accurate models in the majority of the cases up to forecasting time-horizons of at most 50-days. This conclusion is supported by the statistical properties on stationarity and its implications for modelling a time-series.

As for the preferred window size for applying the ADF test, the 250-day and 500-day windows both produced results indicative of the simulated results and both would appear to be effective, however the 250-day window produced more samples that were of a trend-stationary nature and therefore maybe more useful as more opportunities will be presented to minimize forecasting error.

REFERENCES

[1] Bruce Vanstone Bjoern Krollner and Gavin Finnie. Financial time series forecasting with machine learning techniques: A survey. In *European symposium on artificial neural networks: Computational and machine learning. Bruges, Belgium.Apr. 2010*, 2010.

[2] Matthew Butler and Dimitar Kazakov. Modeling the behavior of the stock market with an artificial immune system. In *IEEE Congress on Evolutionary Computation*, pages 1–8, 2010.

[3] Matthew Butler and Dimitar Kazakov. Variable stationarity in financial time series: A comparison of international markets. 2010.

[4] Francis X. Diebold and Lutz Kilian. Unit-root tests are useful for selecting forecasting models. *Journal of Business and Economic Statistics*, 18(3):265–273, 2000.

[5] Gwilym M. Jenkins George Box and Gregory Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, third edition, 1994.

[6] Ricardo Gimeno, Benjamn Manchado, and Romn Mnguez. Stationarity tests for financial time series. *Physica A: Statistical Mechanics and its Applications*, 269(1):72 – 78, 1999.

[7] Tae Yoon Kim, Kyong Joo Oh, Chiho Kim, and Jong Doo Do. Artificial neural networks for non-stationary time series. *Neurocomputing*, 61:439 – 447, 2004. Hybrid Neurocomputing: Selected Papers from the 2nd International Conference on Hybrid Intelligent Systems.

[8] Denis Kwiatkowski, Peter C. B. Phillips, Peter Schmidt, and Yongcheol Shin. 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 Econometrics*, 54(1-3):159 – 178, 1992.

[9] Serena Ng and Pierre Perron. Lag length selection and the construction of unit root tests with good size and power. *Econometrica*, 69(6):1519–1554, 2001.

[10] Robert Pindyck and Daniel Rubinfeld. *Microeconomics*. Prentice Hall, seventh edition, 2008.

[11] G William Schwert. Tests for unit roots: A monte carlo investigation. *Journal of Business and Economic Statistics*, 7(2):147–59, April 1989.

[12] Walter A. Shewhart and Samuel S. Wilks. *Applied Econometric Time Series*. Wiley, 1995.

[13] Paul D. Yoo, Maria H. Kim, and Tony Jan. Machine learning techniques and use of event information for stock market prediction: A survey and evaluation. In *CIMCA-IAWTIC'06*, pages 835–841, Washington, DC, USA, 2005. IEEE Computer Society.