Introduction

This paper examines whether the DF-GLS test improves on the power of the Augmented Dickey-Fuller test at near stationary autoregressive coefficients when there is a small sample size. In so doing, we aim to test the claims of Stock and Watson that the DF-GLS achieves better power than the ADF test under these conditions. DF-GLS is a two-step process, in which the time series is estimated by generalized least squares in the first step before a normal Dickey-Fuller test is used to test for a unit root in the second step. Supposedly this process improves the power of a regular ADF test when the autoregressive paramater is near one.

We created a simulation of a stationary autoregressive process and count the number of times the Dickey-Fuller, DF-GLS, and Phillips-Perron tests failed to reject the null hypothesis that the series is a random walk. Having generated the time series, we know that the null hypothesis is false. Power is thus equal to one minus the percentage of times that we fail to reject the null.

We examine the differences between power levels for the ADF, DF-GLS, and Phillips-Perron tests as we vary sample size at two different values of the coefficient on the lag variable. First we hold the coefficient constant at 0.9 and vary observation length from 100 to 1000 in increments of 100. We then hold sample size constant and vary the coefficient. For a given observation length or coefficient, we ran the simulation 1000 times. In our results, Phillips-Perron typically closely follows the trends of a regular ADF test. This is because our simulation has homoskedastic errors, while the Phillips-Perron test is designed to correct for heteroskedasticity, and consequently the Phillips-Perron does not differ significantly from a Dickey-Fuller test. We are therefore not very interested in the Phillips-Perron test and will focus our analysis on the differences between ADF and DF-GLS.

We find that the DF-GLS test is significantly worse at correctly identifying near stationary processes than the ADF or Phillips-Peron tests. We offer several different possibilities for why we fail to reproduce Stock and Watson's results, primarily that their methodology is not well-documented, and that Stata's version of the DF-GLS test is unwieldy.

Data Generation

We generate data using a Monte Carlo methodology, with random errors. First, we set the time parameter, t, with t running from 0 to T. We then generate an error term, μ_t, which is normally distributed with mean zero and a standard error of one. We then generate a y_t such that $y_1 = \mu_1$ and:

$$y_t = \theta y_{t-1} + \mu_t$$

where θ is a constant such that $0 < \theta < 1$. This makes y_t a stationary autoregressive process of order one, with no trend component.

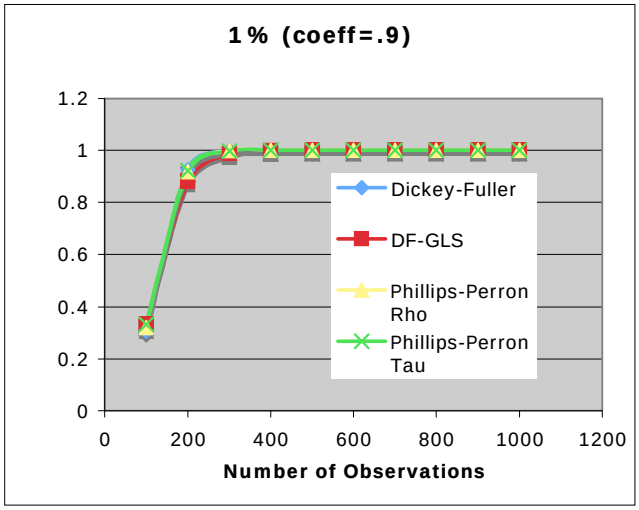
Base Case

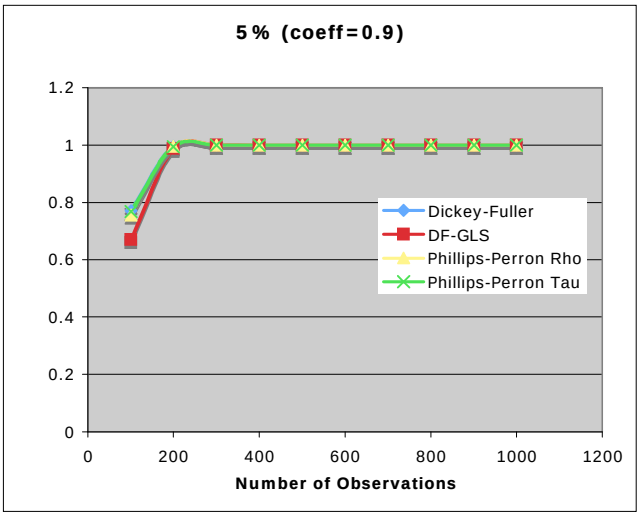
Stock and Watson (p 652) claim that the DF-GLS test is vastly superior to the ADF test in detecting near non-stationarity in small sample sizes. Specifically, they site a Monte-Carlo study they conducted with sample size of 200, and auto-regressive coefficient of 0.95, with no trend component. They claim that they find at the 5% significance level that the DF-GLS rejects the unit-root null-hypothesis 75% of the time, whereas the ADF test rejects it only 31% of the time, a huge difference in power.

We could not reproduce these results. Several possible problems may have occurred. The first is that Stock and Watson do not report how they set their lag lengths for the DF-GLS test. They also do not report on the setup of the ADF test - whether lags are included, and if so, how many. This may explain the huge difference in our results. We first tried the DF-GLS test with zero lags. Then, with one through 10 lags. Finally, we used the AIC criterion to pick the appropriate number of lags for each run of the simulation. *In no case were we able to reproduce Stock and Watson's results*. We did not experiment with the ADF test, but the results here are also incredibly different from Stock and Watson's.

	Power		
Test	**1%**	**5%**	**10%**
DFGLS Lags			
0	0.305	0.693	0.874
1	0.271	0.635	0.823
2	0.261	0.587	0.82
3	0.262	0.58	0.775
4	0.231	0.549	0.76
5	0.23	0.543	0.761
6	0.205	0.516	0.73
7	0.204	0.545	0.751
8	0.186	0.513	0.732
9	0.165	0.512	0.724
10	0.183	0.526	0.716
AIC	0.194	0.511	0.742
ADF	0.215	0.7	0.466

Varying *T*





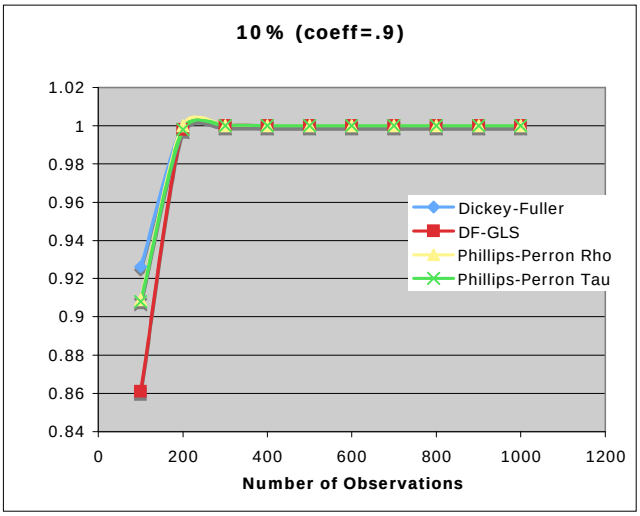


Figure 1. Observed power of the Dickey-Fuller, DF-GLS, and Phillips-Perron test for sample sizes varying between 100 and 1000. The lag coefficient of 0.9.

With a coefficient of 0.9, the results from each test are very similar. The biggest differences lie generally below a sample size of 300. Above that sample size all tests seemingly reject the null hypothesis 100% of the time. One notices that in the region where the tests differ, DF-GLS actually has less power than both Dickey-Fuller and Phillips-Perron, which contradicts our original expectation.

Figure 2 displays the difference between Dickey-Fuller and DF-GLS at all three significance levels. A positive value indicates that Dickey-Fuller does better than DF-GLS. At virtually all lag lengths and significance levels Dickey-Fuller does at least as well as DF-GLS.

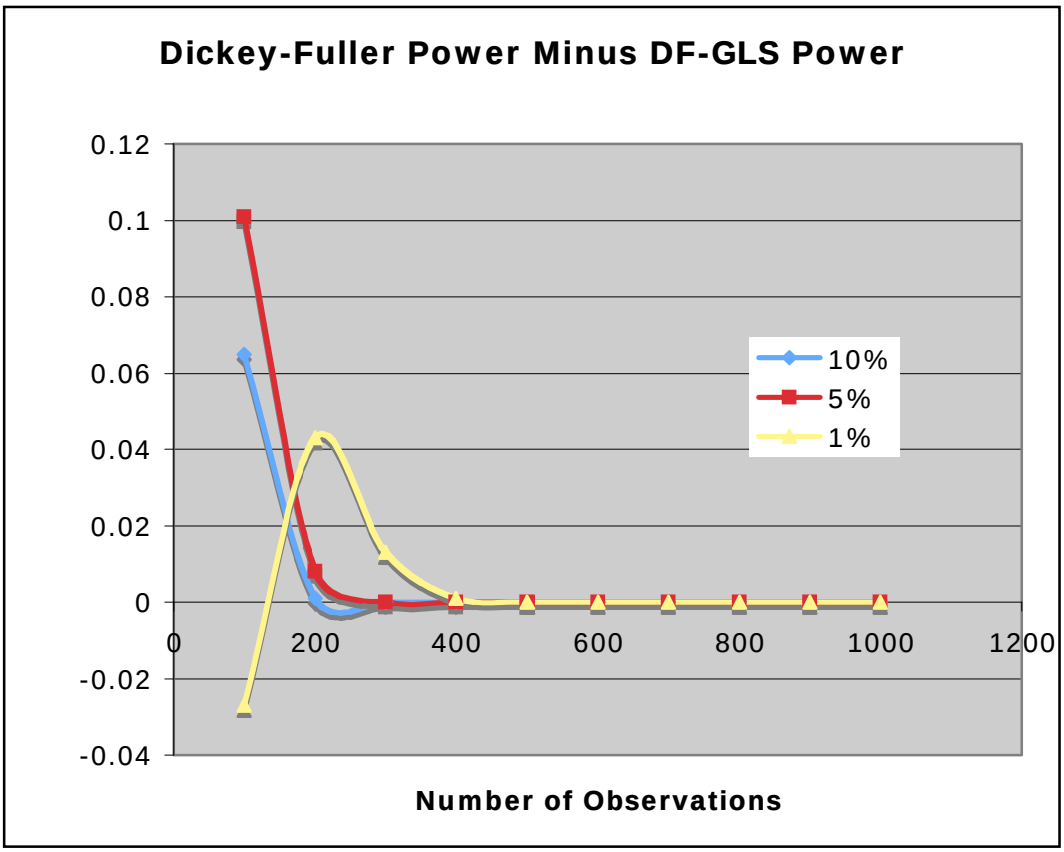
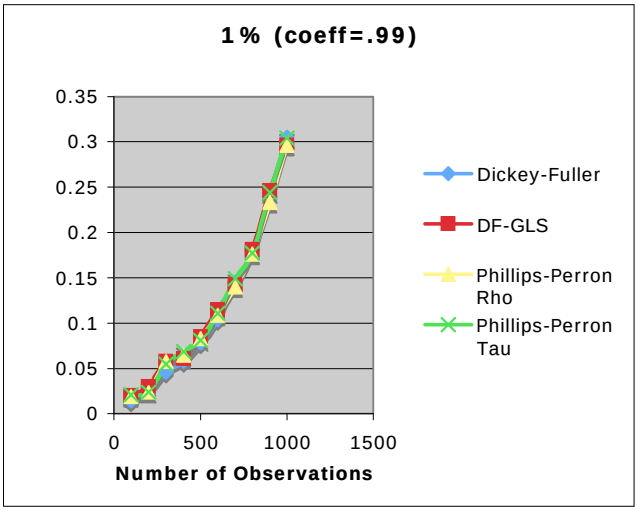
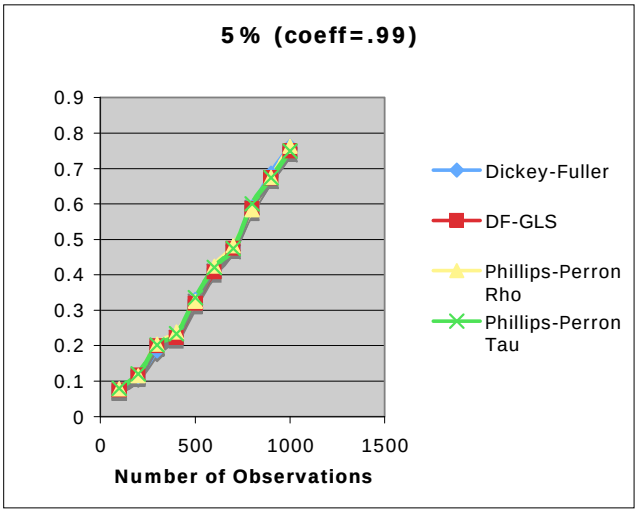


Figure 2 Difference in power between Dickey-Fuller and DF-GLS for observations between 100 and 1000. The lag coefficient is 0.9.

We conclude that a coefficient of 0.9 was not near enough to 1 for DF-GLS to have an advantage over the ADF test. At 0.9, with a large-enough sample size, all three tests were able to distinguish the near-unit root from a unit-root. Consequently, we decided to look at the results when the coefficient is 0.99. Figure 3 shows the results when we replicate the process used above with a coefficient of 0.99.





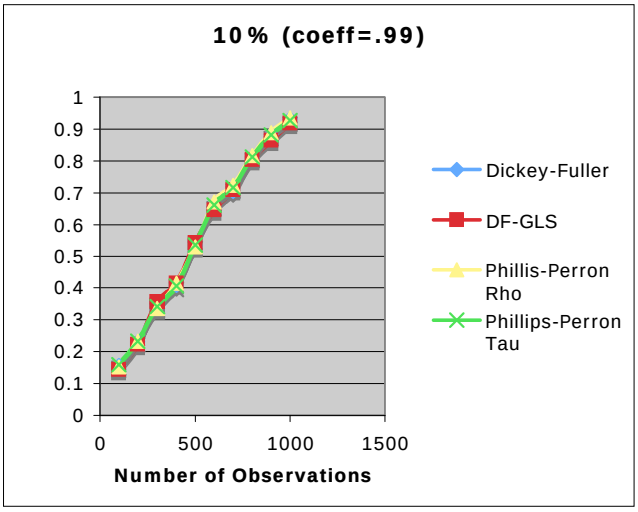


Figure 3 Observed power of the Dickey-Fuller, DF-GLS, and Phillips-Perron test for sample sizes varying between 100 and 1000. The lag coefficient of 0.9.

The graph of the 1% level shows the power appears to improve exponentially as we increase the number of observations. In contrast, the graph of the 10% level appears to curve slightly in the opposite direction. The power for all three tests is lower across sample size at the 1% level than it is at the 10% level, which follows intuitively because it is harder to reject a null at the 1% level than at the 10% level. The shape of the curves, however, is puzzling.

We find that for each test, the 1% critical values gradually get closer to 0 as sample size increases. The 10% critical values get farther from 0 as sample size increases. Though the differences are slight, this might help explain the different graph shapes that we observe. Table 1 lists the critical values for each sample size and test.

Table 1.

stat_obs	DF 1%	DF 5%	DF 10%	DFGLS 1%	DFGLS 5%	DFGLS 10%
100	-2.601	-1.95	-1.61	-2.6	-2.119	-1.811

200	-2.587	-1.95	-1.617	-2.587	-2.035	-1.72
300	-2.58	-1.95	-1.62	-2.58	-2.007	-1.689
400	-2.58	-1.95	-1.62	-2.58	-1.992	-1.673
500	-2.58	-1.95	-1.62	-2.58	-1.983	-1.663
600	-2.58	-1.95	-1.62	-2.58	-1.978	-1.657
700	-2.58	-1.95	-1.62	-2.58	-1.973	-1.652
800	-2.58	-1.95	-1.62	-2.58	-1.97	-1.649
900	-2.58	-1.95	-1.62	-2.58	-1.968	-1.646
1000	-2.58	-1.95	-1.62	-2.58	-1.966	-1.644

	Tau 1%	Tau 5%	Tau 10%
100	-2.601	-1.95	-1.61
200	-2.587	-1.95	-1.617
300	-2.58	-1.95	-1.62
400	-2.58	-1.95	-1.62
500	-2.58	-1.95	-1.62
600	-2.58	-1.95	-1.62
700	-2.58	-1.95	-1.62
800	-2.58	-1.95	-1.62
900	-2.58	-1.95	-1.62
1000	-2.58	-1.95	-1.62

For the most part the critical values stop changing after a sample size of 300. We therefore expect power to increase slowly at low sample sizes for the 1% level, and to increase faster-than-normal at small sample sizes for the 10% level. This behavior is roughly reflected in the graphs.

Figure 4 shows the difference between Dickey-Fuller and DF-GLS with a 0.99 coefficient for each significance level.

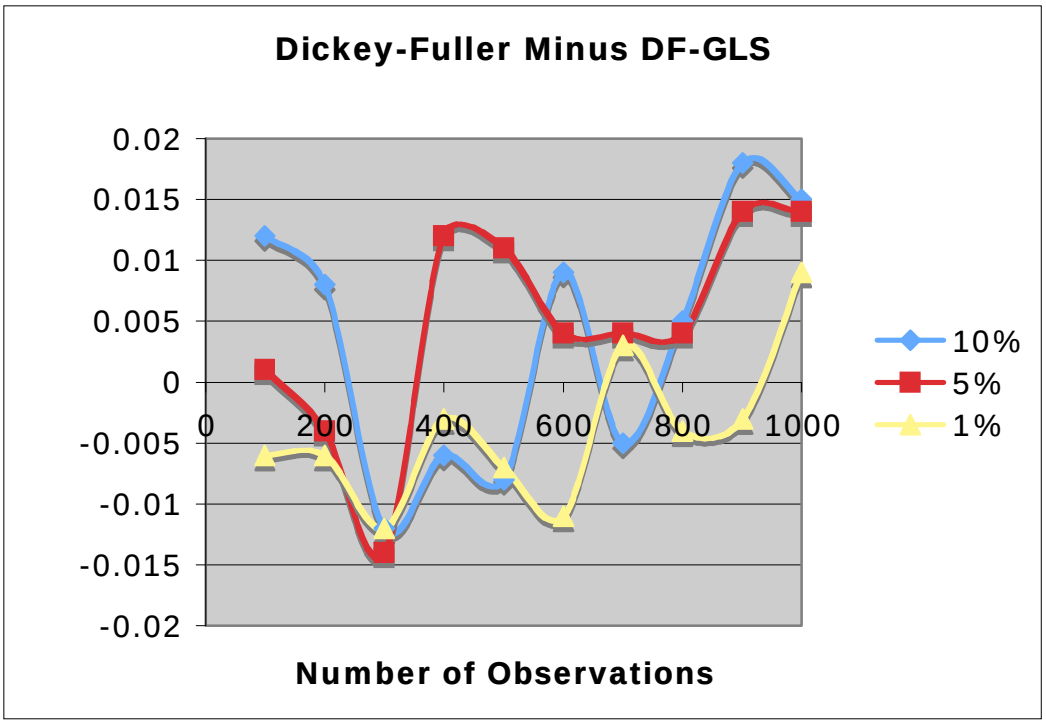


Figure 4 Differences between the observed power for the Dickey-Fuller and DF-GLS unit-root tests. The lag coefficient is 0.99

For the most part, the differences fluctuate and don't seem to have a clear trend. Also, at most sample sizes, DF-GLS has higher power than Dickey-Fuller at the 1% level, and lower power than Dickey-Fuller at the 10% level. However, around a sample size of 300, DF-GLS has higher power than Dickey-Fuller at all significance levels. At a sample size of 300, every significance level reports a higher power for DF-GLS than from Dickey-Fuller, with a difference of about -0.013. Although the connection may be coincidental, it is interesting that the agreement between tests occurs at the point where the critical values stop changing after a sample size of 300. The improvement in DF-GLS over Dickey-Fuller near a sample size of 300 could potentially reflect the prediction that DF-GLS has the best advantage in power at small sample sizes. However, due to a lack of data points at small sample sizes, we decided to investigate this region before making conclusions.

Figure 5 shows the power of the three tests for each significance level with small sample sizes ranging from 50 to 200 by increments of 25. In retrospect we realized that the range around 300 is also interesting. Unfortunately, we didn't simulate observation lengths between 200 and 300. We have included the previous observations for a sample size of 300 to approximate the trends.

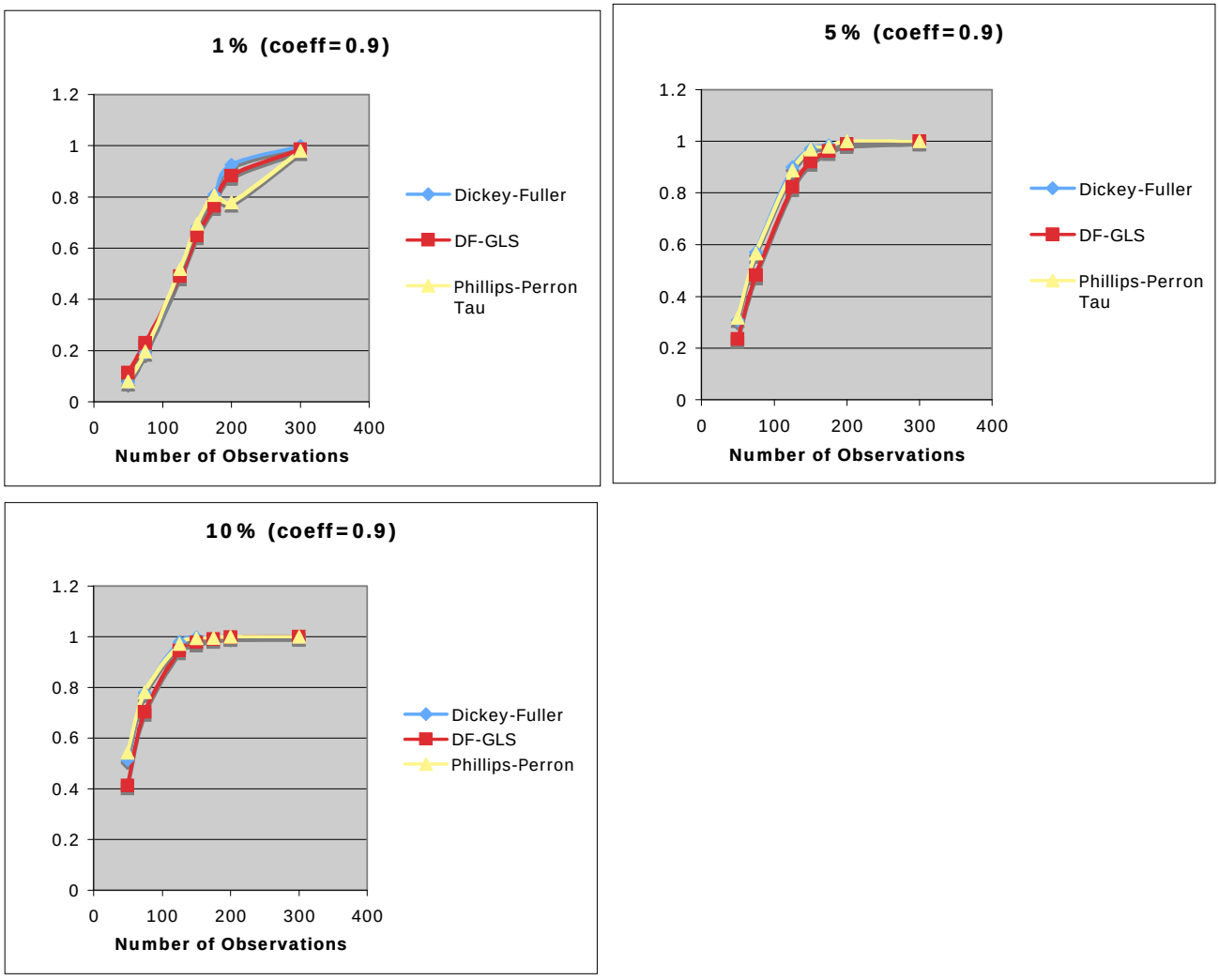


Figure 5. Power at small-sample sizes with a coefficient of 0.9 for the Dickey-Fuller, DF-GLS and Phillips-Perron tests.

In the range of sample sizes between 50 and 300, DF-GLS does not usually perform better than Dickey-Fuller or Phillips-Perron. The tests differ most below 200, but even in that range the differences are not very large. At a sample size of 300, most of the tests have a power of 1 at all significance levels. Figure 6 displays the differences between Dickey-Fuller and DF-GLS at each significance level in order to better evaluate the magnitude of differences.

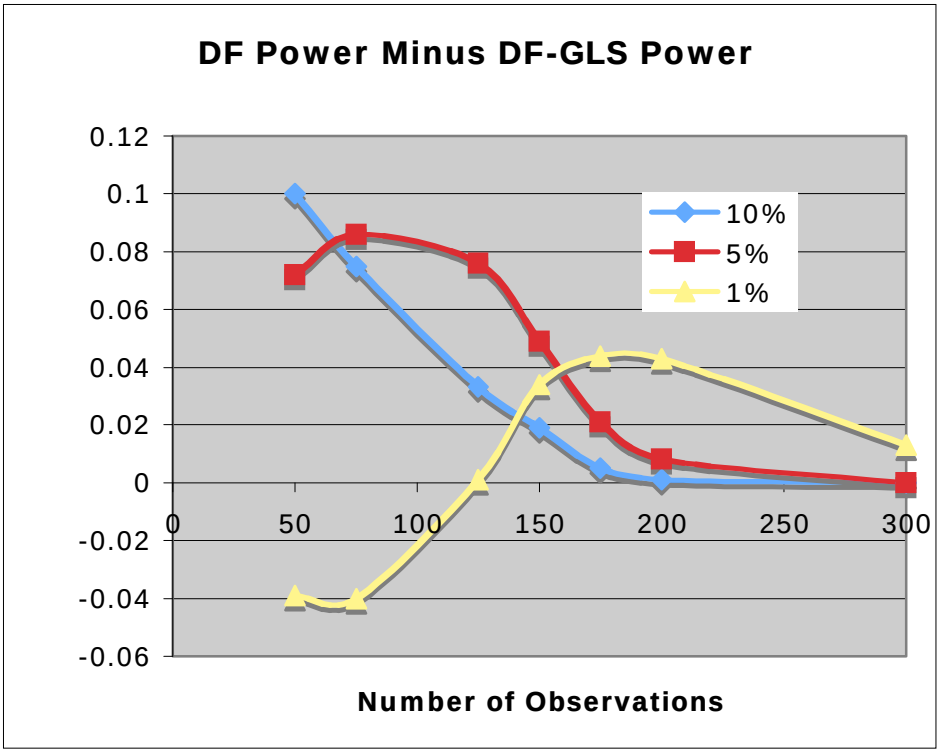


Figure 6 Difference between Dickey-Fuller and DF-GLS power, with coefficient value 0.9.

Except at the 1% level for sample sizes below 125, Dickey-Fuller always performs better than DF-GLS. This again contradicts our expectations

Throughout our analysis of the effect of sample size on the power of unit-root tests, we consistently find that DF-GLS does not perform as well as a standard Dickey-Fuller or Phillips-Perron. One very likely explanation for the poor performance of the DF-GLS test in relation to other tests is that we used a maxlags(0) specification. Although our results are valid for this specification, one should not conclude from these results that DF-GLS is worse than Dickey-Fuller for other lag-length specifications. This was not the focus of our analysis, but it would be worthwhile to repeat our procedure with other lag-lengths. Unfortunately, this process is hampered because of limitations within STATA, i.e. the program does not record critical values for the DF-GLS test.

Varying θ

Supposedly, the major advantage of the DF-GLS test over both the Phillips-Peron and ADF tests is with θ's close to one, i.e. when the series closely resembles a non-stationary autoregressive process, in this case a random walk. We begin by keeping the sample size constant ($T = 100$), and comparing the Phillips-Peron, ADF, and DF-GLS tests. The Phillips-Peron, ADF and DF-GLS tests were originally set with no trend component, and zero lags. Results are reported below:

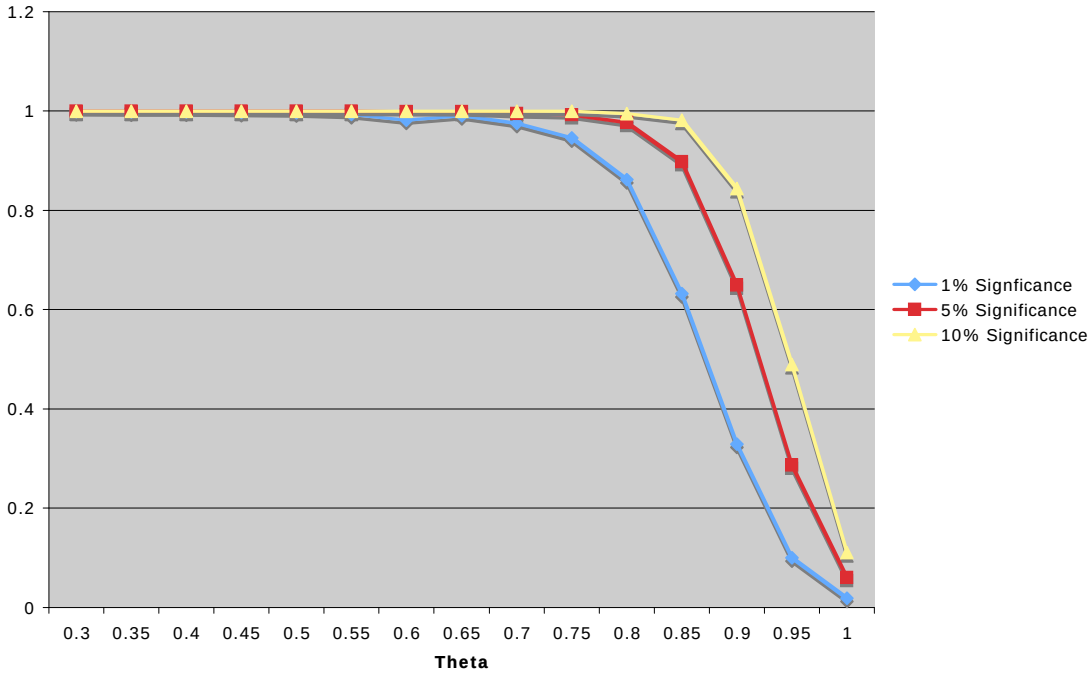
	Augmented Dickey-Fuller		
Theta	1%	5%	10%
0.3	1	1	1
0.35	1	1	1
0.4	1	1	1
0.45	1	1	1
0.5	1	1	1
0.55	1	1	1
0.6	1	1	1
0.65	1	1	1
0.7	0.999	1	1
0.75	0.992	0.999	1
0.8	0.916	0.999	1
0.85	0.68	0.971	0.996
0.9	0.309	0.744	0.917
0.95	0.072	0.3	0.534
1	0.012	0.052	0.103

	Dickey-Fuller/GLS (no lags)		
Theta	1%	5%	10%
0.3	0.999	1	1
0.35	0.998	1	1
0.4	0.998	1	1
0.45	0.997	1	1
0.5	0.996	1	1
0.55	0.993	1	1
0.6	0.982	0.999	1
0.65	0.991	0.999	1
0.7	0.975	0.995	1
0.75	0.946	0.992	1
0.8	0.862	0.977	0.995
0.85	0.632	0.898	0.982
0.9	0.329	0.649	0.844
0.95	0.1	0.287	0.49
1	0.019	0.06	0.111

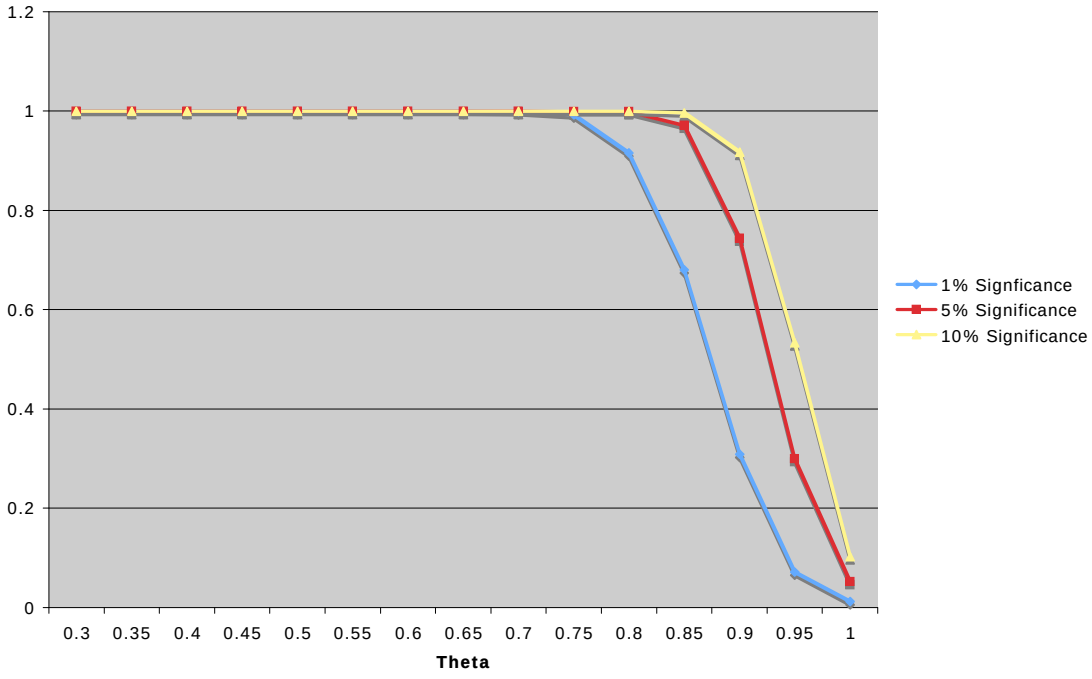
	Philips-Peron		
Theta	1%	5%	10%
0.3	1	1	1
0.35	1	1	1
0.4	1	1	1
0.45	1	1	1
0.5	1	1	1
0.55	1	1	1
0.6	1	1	1
0.65	1	1	1
0.7	1	1	1
0.75	0.989	0.998	1

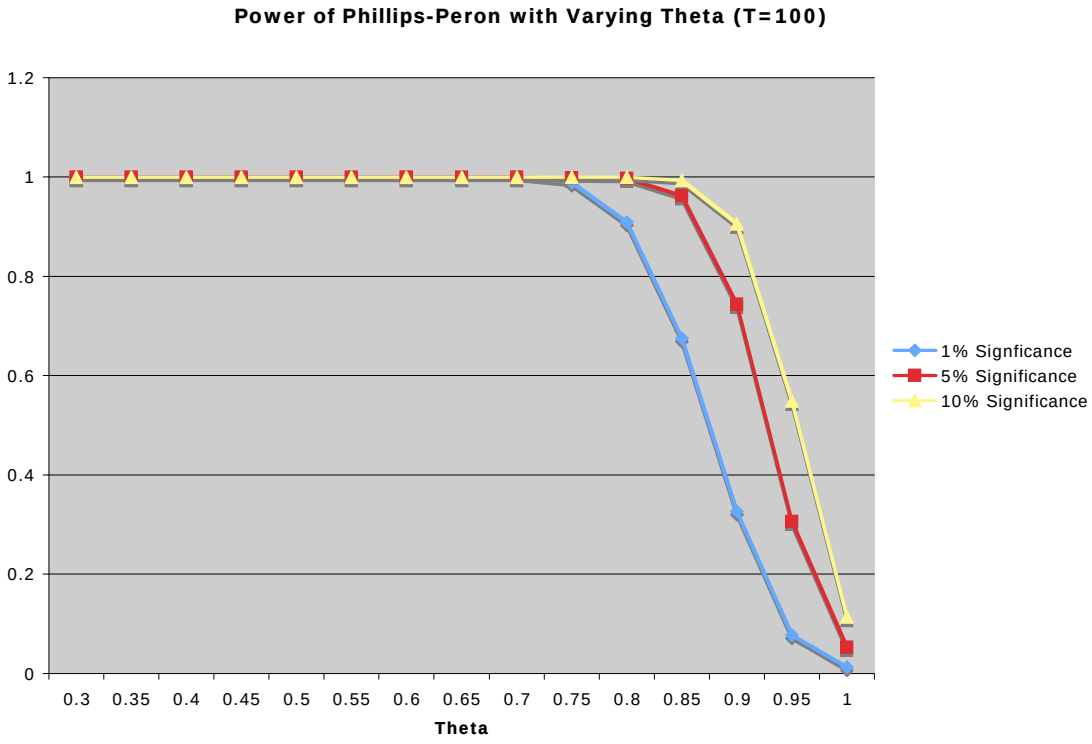
0.8	0.909	0.997	1
0.85	0.675	0.962	0.993
0.9	0.327	0.744	0.906
0.95	0.077	0.306	0.549
1	0.013	0.053	0.114

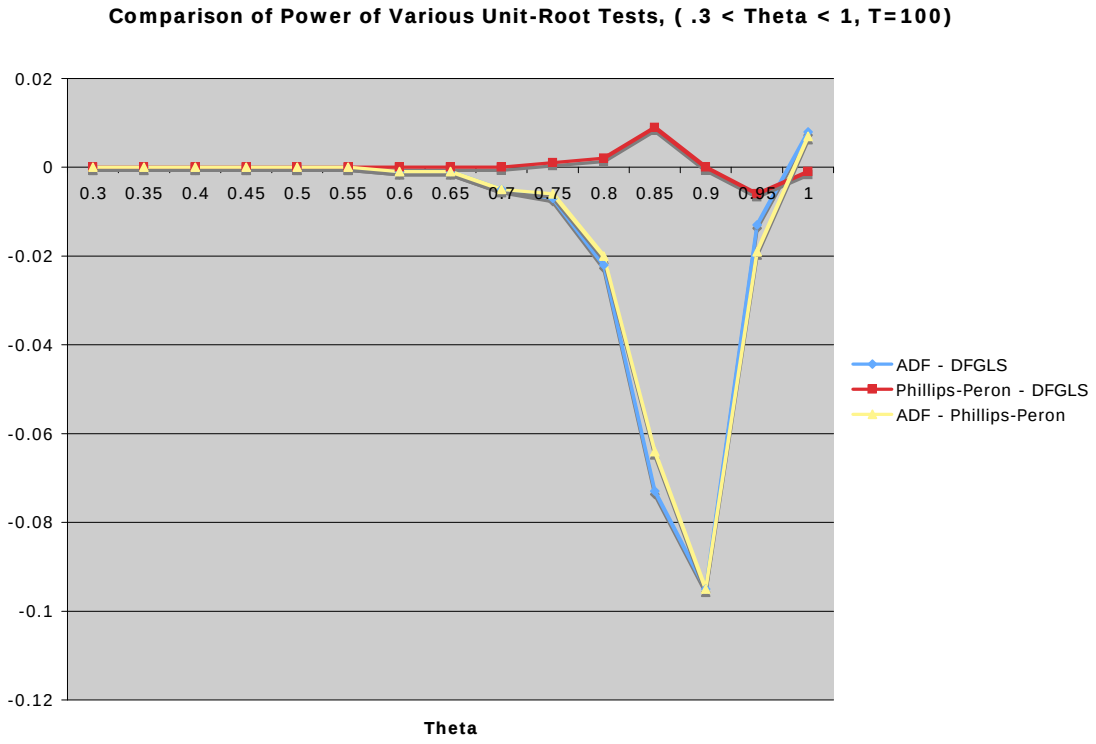
Power of DFGLS Test (no lags) with Varying Theta (T=100)



Power of ADF Test with Varying Theta (T=100)



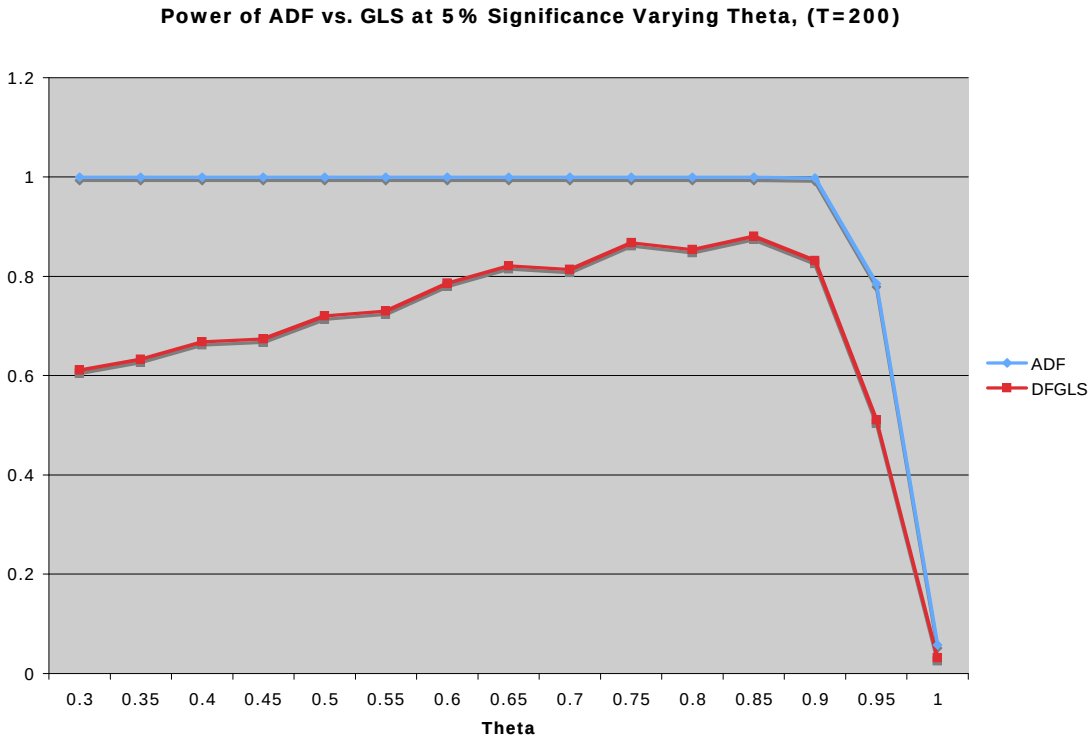

Power of Phillips-Peron with Varying Theta (T=100)
1.2
1
0.8
0.6
0.4
0.2
0
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Theta
1% Signficance
5% Significance
10% Significance



Unsurprisingly, the power of all tests goes down as θ nears one, and the actual coefficient becomes more and more difficult to distinguish from a unit root. Strangely, we find that DF-GLS has lower power relative to Phillip-Peron and ADF as θ gets closer to .9, and then becomes about as accurate as the other tests as θ approaches 1, in that it is not at all accurate. This is exactly the opposite result that Stock and Watson report. This may be because we set the lags to zero for the DF-GLS test, which eliminates much of the intention of the test. Therefore we repeated the results, using the AIC lag criterion to determine lag length, and setting the sample size to 200.

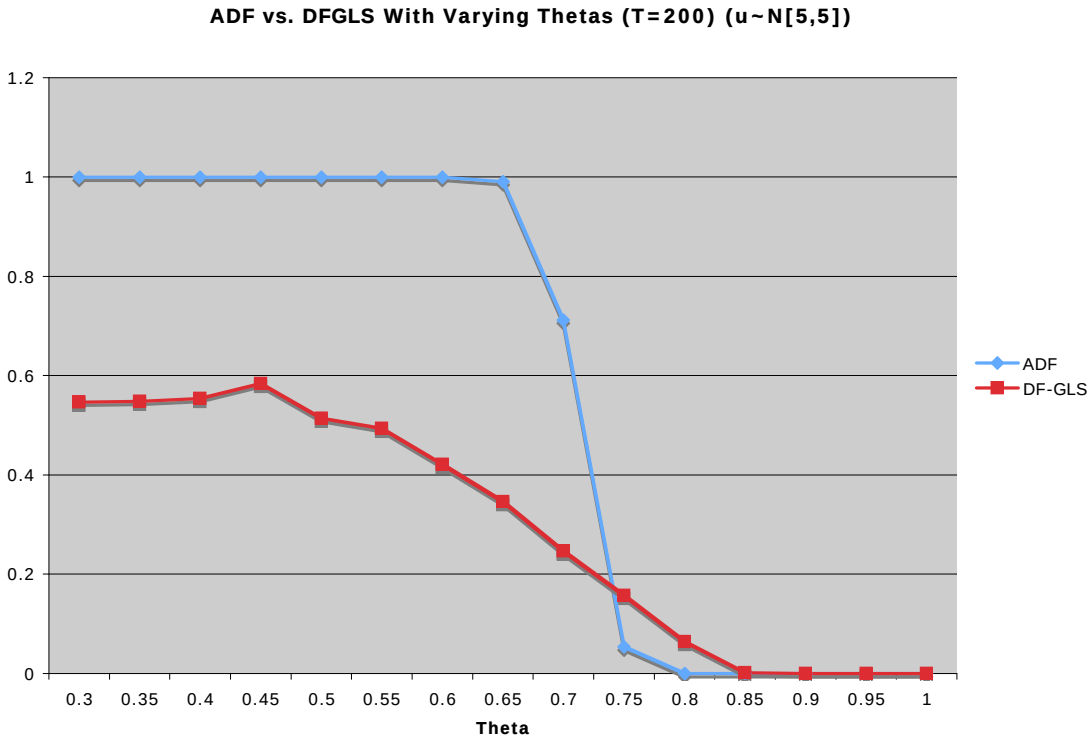
	Augmented Dickey-Fuller		
Theta	1%	5%	10%
0.3	1	1	1
0.35	1	1	1
0.4	1	1	1
0.45	1	1	1
0.5	1	1	1
0.55	1	1	1
0.6	1	1	1
0.65	1	1	1
0.7	1	1	1
0.75	1	0.999	1
0.8	1	0.999	1
0.85	1	0.971	0.996
0.9	0.998	0.744	0.917
0.95	0.785	0.3	0.534
1	0.058	0.052	0.103

	Dicke-Fuller/GLS (AIC lags)		
Theta	1%	5%	10%
0.3	0.438	0.611	0.72
0.35	0.453	0.633	0.741
0.4	0.485	0.668	0.76
0.45	0.484	0.674	0.773
0.5	0.534	0.72	0.806
0.55	0.55	0.73	0.824
0.6	0.586	0.786	0.879
0.65	0.602	0.821	0.91
0.7	0.625	0.814	0.908
0.75	0.653	0.868	0.93
0.8	0.673	0.854	0.935
0.85	0.67	0.881	0.947
0.9	0.557	0.832	0.943
0.95	0.194	0.511	0.742
1	0.009	0.032	0.07



This test was actually less successful. The DF-GLS test with AIC lags is clearly still less powerful than the ADF in the region of $\theta = 1$, and we are still left without a clue as to Stock and Watson's methodology. However, at low levels of Theta, the power of the DF-GLS test is ridiculously low, possibly because the AIC criterion doesn't have enough to work with. In any case, we still cannot find that the DF-GLS test works better at low sample sizes or thetas close to 1.

Trying to see how the DF-GLS test would be effected by a change in the distribution of the error term, we also attempted an error distributed N(5,5), on advice of the lag-length criterion Monte-Carlo group. The results, which do in fact show more power for the DF-GLS test, probably because of increased power of the AIC, and worse problems with the ADF, are shown below:



Conclusions

We can find no good evidence for the high power DF-GLS that Stock and Watson claim exists at low sample sizes and thetas close to 1. This could be caused by several different problems – probably a misspecification on our part because of Stock and Watson's vague methodology, or the even more vague nature of the Stata DF-GLS command. Clearly, if Stock and Watson are to be believed, the DF-GLS is a sensitive test, and needs to be better documented if it is to be used effectively. Without a better understanding of how it should be employed, it may be wiser to rely on the less-touchy ADF or Phillips-Peron tests.