Uses and Abuses of ARIMA in PPNR Modeling and Risk Management: Why Not to Fear ARIMA

Dimitri Bianco

email: Dimitri.Bianco@hotmail.com

Abstract

Many forms of the ARIMA (autoregressive integrated moving average) modeling method are used across risk management and specifically within PPNR (Pre-Provision Net Revenue) for CCAR (Comprehensive Capital Analysis and Review) and DFAST (Dodd-Frank Act Stress Testing). The ARIMA method allows for flexible modeling of PPNR and the inclusion of exogenous variables however model stability can be a concern. I argue that model instability is occurring because of improper ARIMA model development and the practice of forcing all data into the ARIMA framework. I apply a basic method of testing model stability over time and have chosen to model both Citigroup and the S&P 500 using Federal Reserve domestic data which is used in the actual CCAR and DFAST exercises. This paper aims to show common mistakes that occur throughout risk management from the perspective of model development, validation, implementation, and internal audit at major financial institutions.

1. Introduction

Through increased regulation in banking, the modeling of financial data has increased and the demand for risk managers has increased as well. PPNR and other areas of banking have been using time series models that incorporate macroeconomic variables to meet the stress testing demands of CCAR and DFAST. ARIMA[1] is a commonly used method in PPNR modeling and other areas of risk management. It is hard to determine what is the industry standard for modeling time series data and specifically PPNR data due to the proprietary nature of banking. The main issue with ARIMA is the lack of stability which is caused by misspecified models. Misspecification is caused by a lack of understanding the complex ARIMA structure. ARIMA is more complex than other methods such as OLS (ordinary least squares) and the added complexity can be hard to explain to the Federal Reserve if the model developers do not fully

[1] ARIMA will refer to the general category of models that include AR, MA, ARMA, ARIMA, and ARIMAX. Exogenous variables are always used in stress testing however the general term ARIMA is used to describe this family of models.

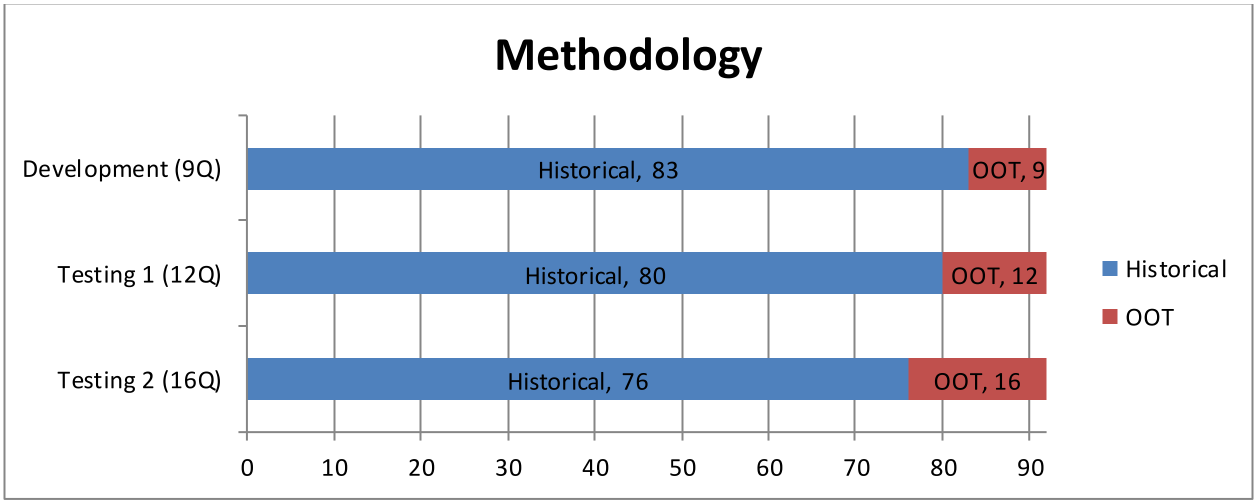
understand the benefits of using ARIMA. The focus of this paper will be on three key issues that are both misunderstood and are causing model instability. These three issues are differencing data, over-fitting, and variable selection.

2. Methodology

For CCAR, models have to be run for year-end and mid-year. As new data is added to these models, p-values of the coefficients on independent variables as well as AR (autoregressive) and MA (moving average) terms can break down and the models become unstable even after annual recalibration. Other adverse issues can arise in unstable models such as serial correlation, however this paper will focus on the coefficients and their p-values. P-values will be considered significant if they are less than 0.05, marginally significant between 0.05 and 0.1 (including 0.1), and insignificant at values greater than 0.1. If a model fails during the annual or semi-annual review, redevelopment by the bank is required.

For this paper ARIMA models were built to forecast nine quarters which is the required timeframe for CCAR. The model structures were selected on the development data set and recalibrated on the 12 and 16 quarter forecast data sets. To test model stability across time the models were tested with a nine quarter forecast, a 12 quarter forecast, and a 16 quarter forecast (see Figure 1). The end forecast date will stay constant and the last "historic" data point will be moved back in time. The reason the forecast grows back in time instead of growing forward is solely a personal preference. By growing the forecast back in time the model can be developed and calibrated on more data. Since time series have small amounts of data, especially internal PPNR data which may not have been collected too far back, maximizing the development data set has advantages and disadvantages. It is important to note that a tradeoff is being made between development data and OOT (out of time) data. It is acknowledged that the model is first fit on the maximum dataset which can cause over-fitting. Due to the small amount of observations in these time series, this method was preferred over using a smaller historic data set with a larger OOT sample and testing forward. For CCAR purposes the nine quarter OOT sample has given stable results in practice.

Figure 1: Stability Testing Method



Due to the proprietary nature of a bank's actual data, internal banking data could not be used for this paper. Two independent variables were selected from the stock market to be modeled because this data is freely available to the public. The S&P 500 ETF (SPY, referred here on as SP500) was used because it is well behaved and represents the market. Citigroup (C, referred here on as Citi) was modeled because it is a bank and should have behavior similar to other banks that are modeling internal PPNR data (see Appendix A for the dependent variable list). The data from Citi has a large drop during the financial crisis which is important to model because internal data will have these large swings during financial crises. The data for the SP500 and Citi was downloaded from Yahoo Finance (Yahoo Finance, 2016) as monthly data. The adjusted close prices were used and the data was taken on March, June, September, and December for the four quarters of every year. The day was set to the last day of every quarter for modeling in SAS purposes. To model the SP500 and Citi, the "Supervisory Historical Domestic" data was downloaded from the Federal Reserve's website under the CCAR 2016 related data (Stress Tests and Capital Planning, n.d.). All data was trimmed so that the quarterly time frame ranged from Q1 1993 through Q4 2015. The trimming was done so that it matched the same time period of the SP500 and Citi.

Variable selection will be covered in section five of this paper however for the examples in this paper only independent variables from the Federal Reserve's domestic data will be considered (see Appendix B for a list of independent variables considered). Better variables could be selected with more information about Citi and the SP500 however to keep the models simple, common economic variables are being used. For example, using international Federal Reserve variables for Citi would be reasonable because they have a global presence however this would require a deeper knowledge of Citi and their global operations. For Citi the unemployment rate (LBR) and Dow Jones Industrial Average (DJIA) are used as independent variables. The expected relationship between Citi and LBR is negative because as unemployment rises less people have money which results in less business for Citi. The expected relationship between Citi and the DJIA is positive because as the markets improve Citi should have more business and higher profits. For the SP500, the BBB Corporate Yield (BBB), and the real GDP growth (rGDPg) are used as independent variables. The expected relationship between the SP500 and BBB is negative because as rates go up, lending becomes expensive for corporations which results in less lending leading to less business expansion and less profits which will mean the SP500 should be declining. The expected relationship between the SP500 and rGDPg is positive because as GDP grows it indicates corporations are becoming more profitable which will drive up the SP500. These relationships are over simplified but will shed light on the issues in modeling data with ARIMA.

3. Differencing

During variable selection stationarity should be tested. It is an industry practice to difference data that is non-stationary to make the data stationary. I have found two camps when it comes to differencing and making data stationary. Camp one performs minimal stationarity testing or avoids ARIMA because of the stability issues. This camp is usually comprised of less experienced developers and managers who do not quite understand time series and the ARIMA methodology. For example, the author of this paper has witnessed lead developers write in CCAR model documentation that stationarity and cointegration are the same.

Camp two sees non-stationary data and automatically differences it. This seems logical however by differencing the data there are cases where the forecasts are no longer reasonable because the differencing loses information that was crucial to the model (Wang & Wang, 2010). More research needs to be conducted on information lost from differencing to see if and how this affects financial data. Developers, validators, and oversight committees need to be aware that information loss could be a cause of unstable models. This second camp is usually comprised of seasoned developers who have followed academia but have not kept current with time series analysis or do not fully understand the data they are examining. The second camp is much more common than the first camp as it is comprised of more seasoned developers who usually correct the mistakes of camp one. However the second camp is deeply rooted despite recent academic research which illustrates stability issues in ARIMA models caused by blindly differencing data. Differencing data is appropriate in many situations and can hint at issues that can be determined when testing stability over time as is mentioned in the methodology of this paper. Within banks there are usually a handful of models that are redeveloped almost every development cycle because crucial long-run information about the data is being lost. Reiss (2015) states

> *"Unfortunately, data preparation does too much and too little at the same time. Regarding only contemporaneous statistical relations, conditioning the past of variables and differencing, detrending, etc., all result in the loss of important long-run information of which a prudent statistician should make use (see for instance Hendry 1995: Sect. 7.4)."*

Dickey and Fuller (1979) state that *"the hypothesis that $\rho = 1$ is of some interest in applications because it corresponds to the hypothesis that it is appropriate to transform the time series by differencing."* Dickey and Fuller's hypothesis may hold in many cases but it does not always hold in practice. The loss of information from differencing is crucial when working with macroeconomic variables such as those required for use as independent variables by the Federal Reserve. Hoover (2003) mentions that most macroeconomic time series are I(1) meaning they are only stationary when differenced once. Industry experience has also proven that most macroeconomic variables are I(1) and this can be seen in the Federal Reserve macroeconomic variables (see Table 1).

Theoretically the variables should not be stationary because of economic shocks in the market and differing rates of information coming to the market. When financial crises occur, the variables tend to have large moves in one direction over a short period of time. If these series are differenced crucial information about the crises can be lost. The point of CCAR and DFAST is to have enough capital on hand when the next crisis occurs and to stress the portfolios and accounts to see what could happen to the banks when a crisis occurs. If differencing is not appropriate this crisis information will be diminished or lost and the purpose of the stress testing is lost.

Table 1: Stationarity of Fed Macroeconomic Variables Using ADF Test

Not Stationary	**I(0)**	**I(1)**
HPI	rGDPg	LBR
	nGDPg	ir3m
	rDIg	ir5yr
	sDIg	ir10yr
	CPI	BBB
		Mort
		Prime
		DJIA
		CRE
		Vol
		sp500
		Citi

There are many models currently being used in practice that are ARIMA type models and are very stable and reliable in stress testing. The issue arises when careless differencing occurs by developers. When a model is redeveloped frequently, analysis should be taken to understand why this is happening. It often occurs that banks redevelop models that fail without analyzing why the model failed. This lack of analysis can come from careless behavior or most likely due to the time constraints imposed by CCAR and other regulations. The industry seems scared of cointegration and VECM (vector error correction models) in general which seems reasonable as most banks and developers either abuse ARIMA by applying it to everything or fear ARIMA from a lack of understanding. Some portfolios and accounts cannot be modeled with OLS or ARIMA type models. The industry needs to move on to the next step which is a cointegration analysis of the variables in the model. It has been seen that developers will try to select different variables or over-fit models to solve the constant redevelopment issue but are refusing to realize there is a problem which could be solved with a model that involves cointegrated relationships. Cointegration is a potential solution because it addresses stationarity in a different way and can incorporate nonlinear dynamics between variables which could exist (Nesmith & Jones, 2008). It is also important to note that not all accounts and portfolios can be modeled with data provided from the Federal Reserve for stress testing. This attitude seems to be supported by the Federal Reserve (2013) in the statement, *"BHC[2]s should not use weak models just for the sake of using a modeled approach to PPNR."*

An example of how differencing is not black and white is nGDPg (nominal GDP growth) and HPI (house price index). When analyzing nGDPg it looks to be single mean and has a p-value at lag three of 0.0588 which means it is not significant at the 5% threshold when using the Tau p-values. When nGDPg is differenced the p-values all drop to <.0001 which many developers would accept as an improvement. There are a few issues that should be considered. The bars on the ACF (autocorrelation function) plot oscillate between positive and negative, the standard deviation increases from 2.731045 to 2.864685, and the value for Rho and Rho's p-value for I(1) at lag four are very different from both the I(0) values and the Tau values (see Figures 2 and 3). These three issues indicate there could be an issue with differencing nGDPg. The above reasoning to difference nGDPg seem reasonable for many developers however the main flaw is that in the industry many developers only use one set of p-values either Rho or Tau. Dickey and Fuller (1979) show that the Rho test should be used when Rho is less than one and the Tau test should be used when Rho is greater than one. With this information from Dickey and Fuller it is clear that nGDPg is stationary when using the p-values of Rho instead of Tau under the single mean (see Figure 2). In Figures 2 through 6 the "*variable(#)*" indicates the number of differencing done to that variable.

[2] Bank Holding Company

Figure 2: ADF Stationarity Test for nGDPg(0)

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-13.6138	0.0088	-2.71	0.0072		
	1	-5.7862	0.0950	-1.72	0.0804		
	2	-4.3208	0.1504	-1.47	0.1313		
	3	-3.4666	0.1983	-1.39	0.1515		
	4	-3.7697	0.1795	-1.44	0.1402		
Single Mean	0	-49.6830	0.0009	-5.76	<.0001	16.56	0.0010
	1	-28.7642	0.0009	-3.69	0.0057	6.83	0.0010
	2	-25.8934	0.0016	-3.26	0.0200	5.31	0.0325
	3	-21.1910	0.0059	-2.83	0.0588	4.04	0.0896
	4	-30.2943	0.0008	-3.04	0.0353	4.66	0.0528
Trend	0	-56.9341	0.0003	-6.38	<.0001	20.40	0.0010
	1	-36.4818	0.0008	-4.21	0.0065	8.89	0.0010
	2	-36.2211	0.0008	-3.81	0.0206	7.27	0.0294
	3	-30.5082	0.0040	-3.25	0.0809	5.29	0.1344
	4	-49.9278	0.0003	-3.52	0.0437	6.19	0.0649

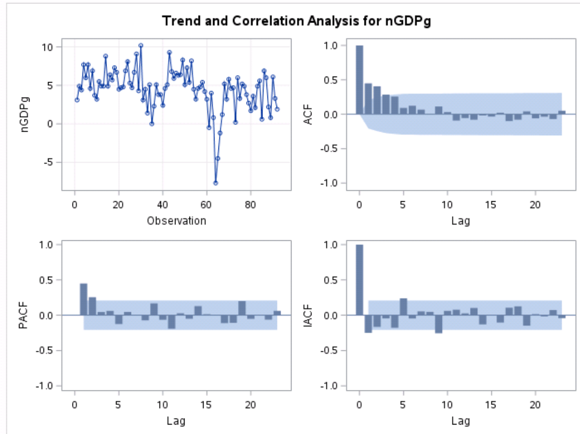
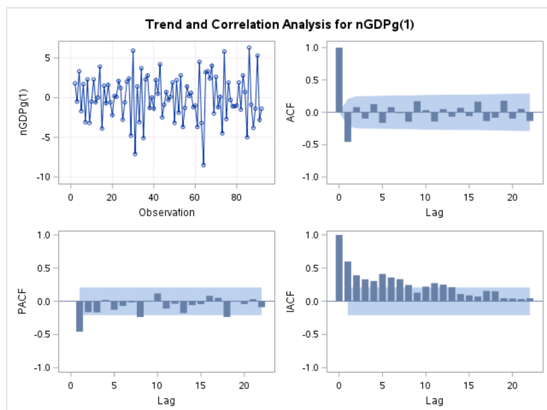


Figure 3: ADF Stationarity Test for nGDPg(1)

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-131.303	0.0001	-15.51	<.0001		
	1	-184.021	0.0001	-9.39	<.0001		
	2	-471.895	0.0001	-7.90	<.0001		
	3	-365.100	0.0001	-5.81	<.0001		
	4	404.6149	0.9999	-5.64	<.0001		
Single Mean	0	-131.302	0.0001	-15.42	<.0001	118.92	0.0010
	1	-183.981	0.0001	-9.33	<.0001	43.55	0.0010
	2	-472.665	0.0001	-7.86	<.0001	30.92	0.0010
	3	-368.492	0.0001	-5.79	<.0001	16.74	0.0010
	4	399.8812	0.9999	-5.62	<.0001	15.80	0.0010
Trend	0	-131.337	0.0001	-15.34	<.0001	117.67	0.0010
	1	-184.007	0.0001	-9.28	<.0001	43.08	0.0010
	2	-473.050	0.0001	-7.81	<.0001	30.52	0.0010
	3	-368.373	0.0001	-5.75	<.0001	16.53	0.0010
	4	401.3444	0.9999	-5.58	<.0001	15.60	0.0010



HPI is not stationary by itself or by differencing once or twice (see Figures 4, 5, and 6). The non-stationarity should be expected because of the volatile rise and fall that occurred leading up to and during the Financial Crisis from 2007-2009. HPI should not be used in ARIMA because it is non-stationary however it could be used in another model type such as VECM if a proper cointegrated relationship with a dependent variable was found.

Figure 4: ADF Stationarity Test for HPI(0)

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	0.7123	0.8544	3.29	0.9997		
	1	0.0711	0.6970	0.05	0.6969		
	2	0.3657	0.7693	0.39	0.7943		
	3	0.3089	0.7548	0.30	0.7697		
	4	0.4679	0.7954	0.65	0.8549		
Single Mean	0	-0.5285	0.9233	-0.70	0.8418	6.98	0.0010
	1	-8.7738	0.1697	-2.19	0.2105	2.67	0.3974
	2	-4.8870	0.4370	-1.61	0.4717	1.71	0.6396
	3	-6.1255	0.3269	-1.77	0.3916	1.93	0.5845
	4	-3.7093	0.5648	-1.47	0.5431	1.71	0.6394
Trend	0	-0.8376	0.9892	-0.63	0.9744	0.28	0.9900
	1	-27.5992	0.0086	-3.65	0.0308	6.69	0.0446
	2	-16.0575	0.1288	-2.64	0.2645	3.50	0.4849
	3	-24.0497	0.0206	-2.97	0.1453	4.45	0.3000
	4	-12.0328	0.2873	-2.14	0.5153	2.38	0.7052

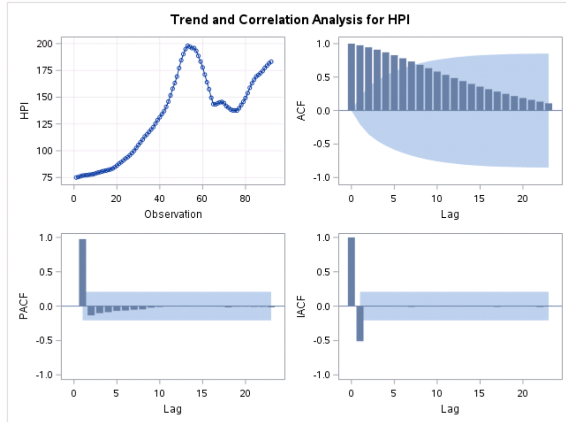


Figure 5: ADF Stationarity Test for HPI(1)

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-5.2322	0.1129	-1.62	0.0992		
	1	-10.4773	0.0224	-2.26	0.0238		
	2	-8.6289	0.0393	-1.98	0.0457		
	3	-18.4269	0.0020	-2.68	0.0079		
	4	-12.8670	0.0109	-2.21	0.0266		
Single Mean	0	-6.1339	0.3264	-1.77	0.3948	1.56	0.6756
	1	-12.1752	0.0705	-2.43	0.1376	2.94	0.3302
	2	-10.2193	0.1171	-2.15	0.2263	2.31	0.4889
	3	-22.5136	0.0040	-2.90	0.0492	4.21	0.0792
	4	-16.1141	0.0241	-2.42	0.1379	2.94	0.3304
Trend	0	-6.1607	0.7210	-1.76	0.7151	1.56	0.8652
	1	-12.2319	0.2774	-2.42	0.3672	2.93	0.5973
	2	-10.2889	0.3931	-2.15	0.5137	2.30	0.7192
	3	-22.8726	0.0272	-2.91	0.1650	4.25	0.3391
	4	-16.4560	0.1176	-2.44	0.3572	3.00	0.5833

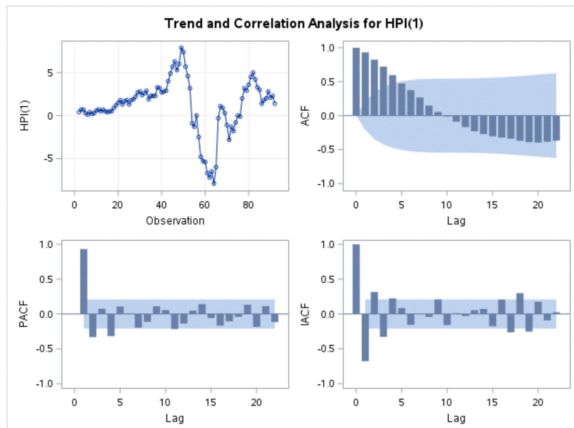
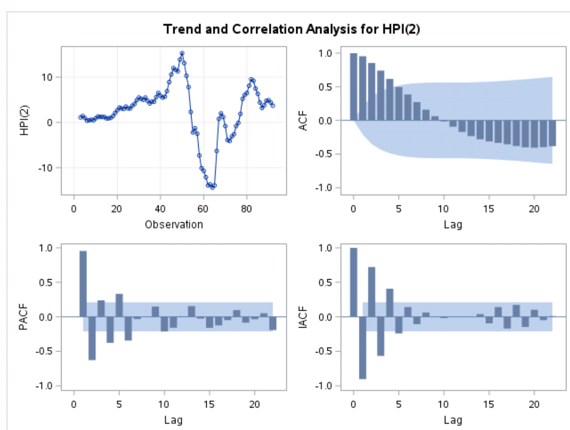


Figure 6: ADF Stationarity Test for HPI(2)

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-3.3428	0.2067	-1.27	0.1868		
	1	-15.2272	0.0054	-2.72	0.0071		
	2	-8.2943	0.0435	-1.93	0.0514		
	3	-21.8437	0.0006	-2.85	0.0048		
	4	-7.8895	0.0492	-1.79	0.0705		
Single Mean	0	-4.0160	0.5296	-1.42	0.5710	1.02	0.8134
	1	-17.7719	0.0153	-2.93	0.0462	4.29	0.0747
	2	-10.0286	0.1229	-2.13	0.2330	2.28	0.4968
	3	-27.4687	0.0010	-3.14	0.0276	4.93	0.0437
	4	-9.9433	0.1253	-1.99	0.2914	1.98	0.5714
Trend	0	-4.0264	0.8806	-1.41	0.8510	0.99	0.9674
	1	-17.8552	0.0871	-2.92	0.1618	4.26	0.3369
	2	-10.1155	0.4046	-2.13	0.5227	2.27	0.7263
	3	-27.9147	0.0077	-3.14	0.1038	4.94	0.2043
	4	-10.1158	0.4040	-1.99	0.5963	1.99	0.7798



4. Over-fitting

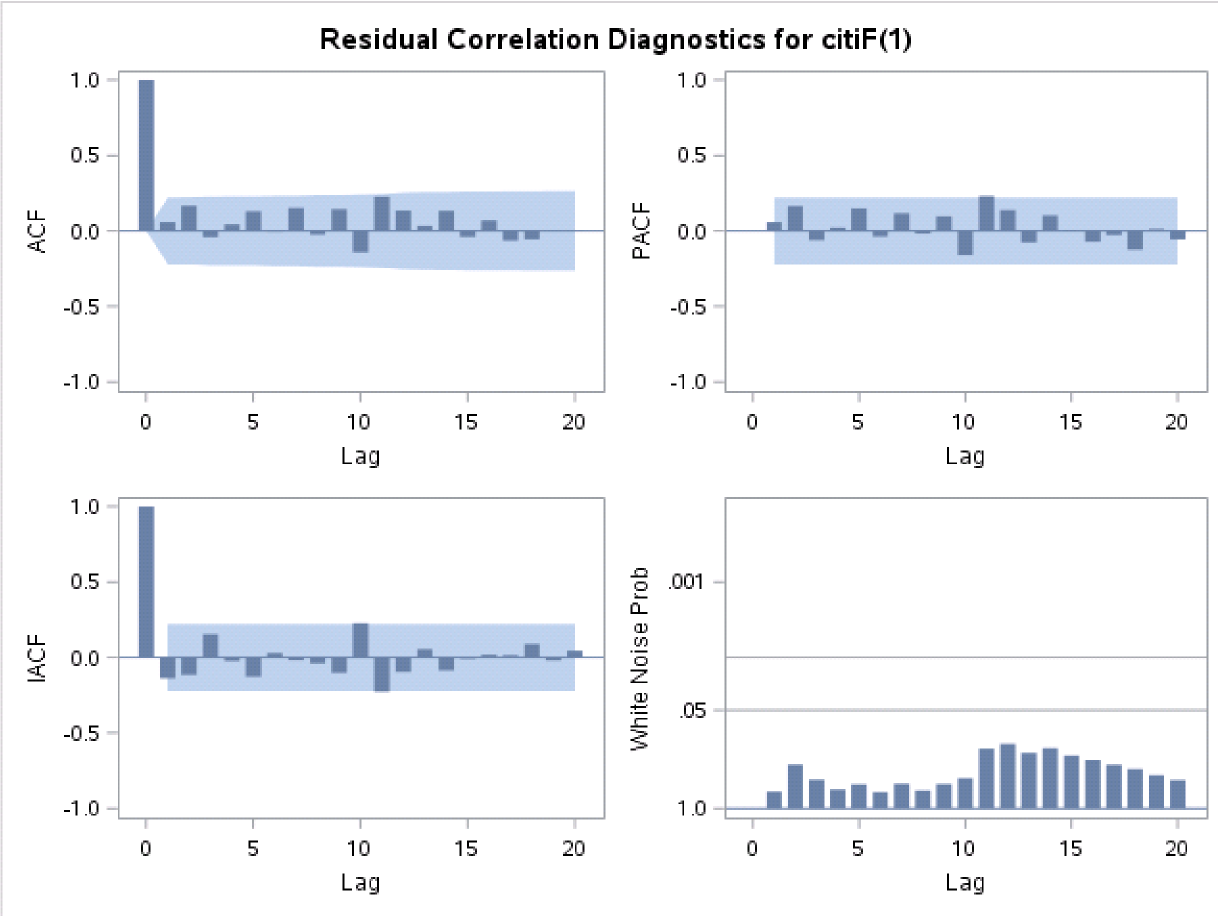
Over-fitting occurs when too many independent variables, AR (Autoregressive), and/or MA (Moving Average) terms have been added to the model. This paper will focus on the over-fitting associated with too many AR and MA terms as this is a more common pitfall in ARIMA modeling. AR and MA terms can come up as significant (low p-values) however these terms will be competing with other parts of the model when over-fitting occurs. If a developer, validator, or oversight committee does not recognize these signs, the model will most likely need redevelopment more frequently than a model with a proper fit. An example of over-fitting can be seen while fitting a model to Citi. Citi is non-stationary and has a significant change in the data structure when looking at the data before the crisis compared to the data after the crisis. The Stock Watson test was conducted to look for common trends and the Johansen cointegration test was conducted. All tests were conducted on a single independent variable model with Citi as the dependent variable. No trends or cointegration were found with LBR or DJIA and so LBR and DJIA were differenced to make them stationary. The data for Citi was differenced in this model and two versions were created where one version has a decent fit and the other is over-fitted. Model (1) shows no signs of serial correlation, white noise in the residuals, correct sign on the coefficient, and a significant p-value for the coefficient (see Figures 7 and 8).

$$\widehat{C_t} = \beta_1(DJIA_t - DJIA_{t-1}) + C_{t-1} + \varepsilon_t \tag{1}$$

Figure 7: Model 1 Estimates

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
NUM1	0.02665	0.0034052	7.83	<.0001	0	DJIA	0

Figure 8: Model 1 Residual Analysis



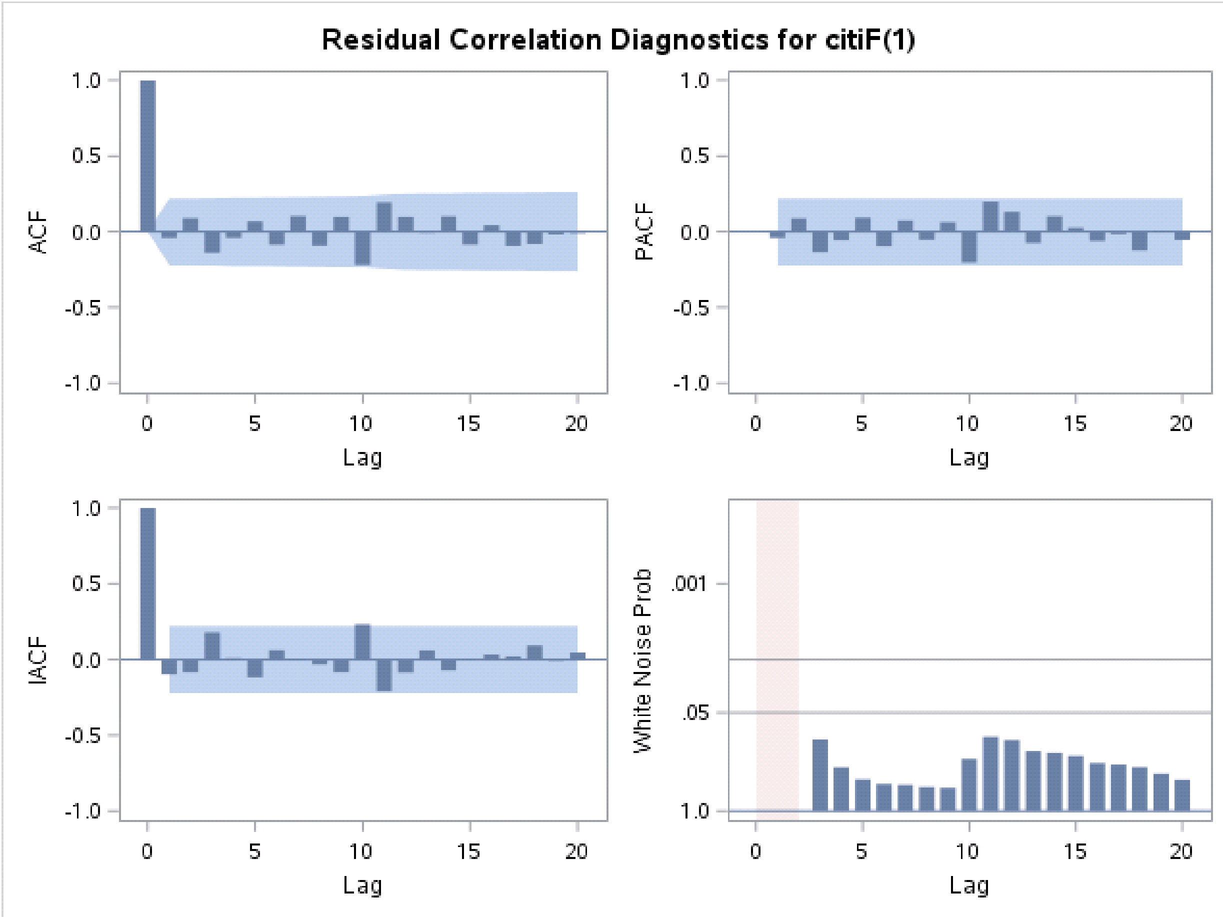
The model can be refitted using AR and MA terms to get Model (2). This model also shows white noise in the residuals, correct sign on the coefficients, significant p-values, and no serial correlation (see Figures 9 and 10).

$$\widehat{C_t} = \beta_1(DJIA_t - DJIA_{t-1}) + \beta_2(C_{t-1} - C_{t-2}) + \beta_3(\varepsilon_{t-1}) + C_{t-1} + \varepsilon_t \qquad \textbf{(2)}$$

Figure 9: Model 2 Estimates

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.88047	0.15509	5.68	<.0001	1	citiF	0
AR1,1	0.95188	0.11071	8.60	<.0001	1	citiF	0
NUM1	0.02716	0.0034064	7.97	<.0001	0	DJIA	0

Figure 10: Model 2 Residual Analysis



Both of these models remain stable over time however the second model is over-fitted which may not be detected by an inexperienced developer, validator, or oversight committee. The issue is that the effects from the AR and MA terms are canceling each other out which is seen in the correlation between the AR and MA terms being over 0.90 in all three time periods (see Figure 11).

Figure 11: Model 2 Correlation Analysis for 9Q, 12Q, 16Q

Correlations of Parameter Estimates

Variable	Parameter	citiF MA1,1	citiF AR1,1	DJIA NUM1
citiF	MA1,1	1.000	0.937	0.146
citiF	AR1,1	0.937	1.000	0.154
DJIA	NUM1	0.146	0.154	1.000

Correlations of Parameter Estimates

Variable	Parameter	citiF MA1,1	citiF AR1,1	DJIA NUM1
citiF	MA1,1	1.000	0.958	0.111
citiF	AR1,1	0.958	1.000	0.108
DJIA	NUM1	0.111	0.108	1.000

Correlations of Parameter Estimates

Variable	Parameter	citiF MA1,1	citiF AR1,1	DJIA NUM1
citiF	MA1,1	1.000	0.965	0.081
citiF	AR1,1	0.965	1.000	0.073
DJIA	NUM1	0.081	0.073	1.000

In general, a model that uses only AR or only MA terms should be used over a mixed model of AR and MA terms when possible. Simple models are easier to understand and with regards to ARIMA less AR and MA terms will result in a more stable model in general. In Model (2) it is well behaved meaning the p-values stay significant and the coefficients do not change signs. A common symptom of over-fitting is a model that is unstable in the sense that the p-values of the coefficients become insignificant over time. If the correlations were not so high the over-fitting and multicollinearity could have been missed. A way to test for this over-fitting when the multicollinearity is not clearly present is trying to build a simpler model with no AR and MA terms or build a model with only AR or only MA terms. Having a pure AR or MA model helps ensure that AR and MA terms are not canceling each other out. A pure AR or MA model also follows the academic and industry concept that complexity should only be added when absolutely necessary. The over-fitting and specifically the correlation (multicollinearity) between exogenous variables or other terms in the model is a serious problem and can be easily over looked. Perfect multicollinearity is rarely seen and making the judgement call on what is considered multicollinear will be negatively influenced by banks trying to create models for the sole purpose of passing CCAR or other regulations. Some tests such as the VIF (variance inflation factor) are being used but these are arbitrary as well. Farrar and Glauber (1967) point out that "*Multicollinearity constitutes a threat – and often a very serious threat – both to the proper specification and the effective estimation of the type of structural relationship commonly sought through the use of regression techniques.*" Their paper is in regards to least squares regression however their point is even more important with regards to ARIMA due to the added complexities.

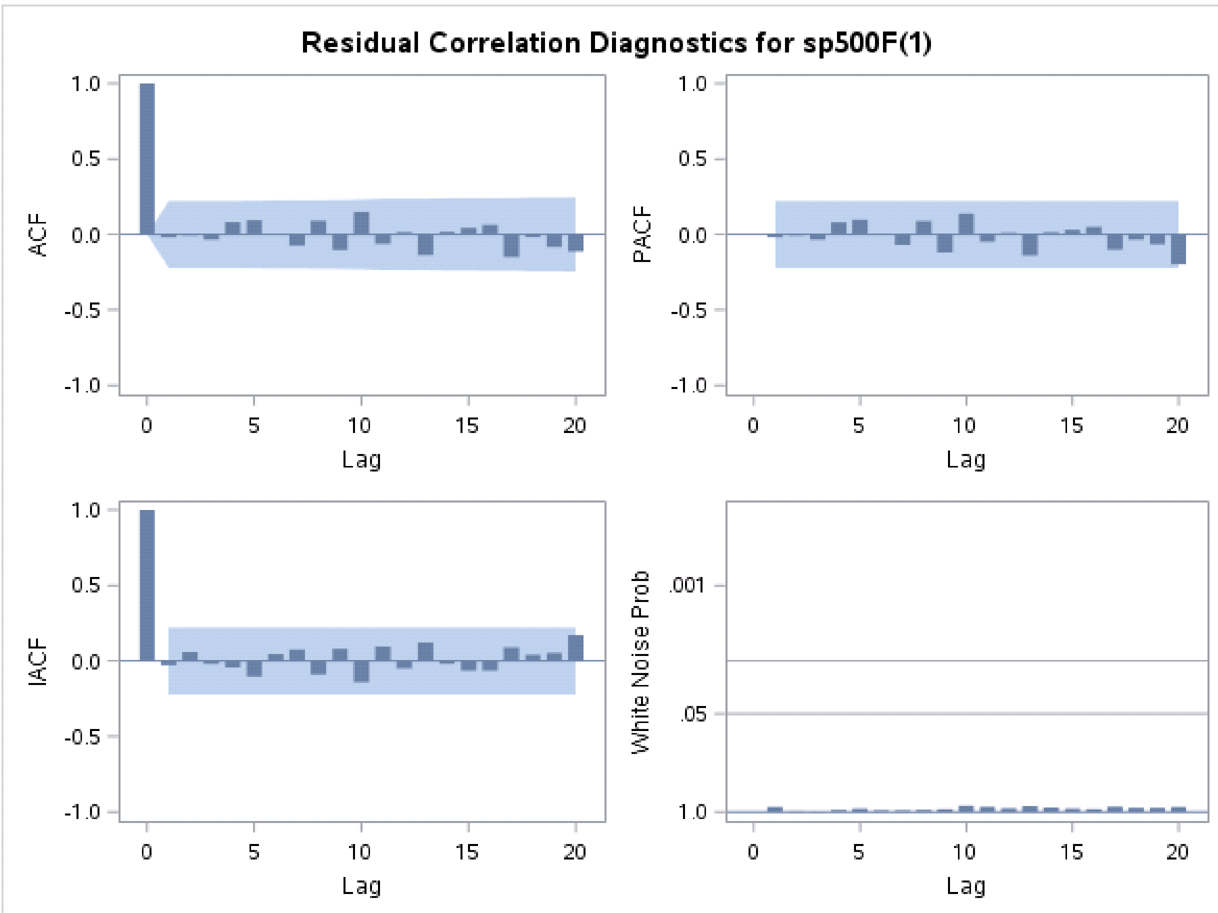
An example of a model becoming unstable over time is an SP500 model with the same issue of an AR and MA term canceling the effects of each other which is an over-fitting issue. The SP500 and BBB are differenced to make them stationary however rGDPg is not differenced because it is already stationary and a measure of change. The first version of the SP500 Model (3) again has white noise in the residuals, correct sign on the coefficients, significant p-values over all three time testing periods, and no serial correlation (see Figures 12 and 13).

$$\widehat{SP500}_t = \beta_1(BBB_t - BBB_{t-1}) + \beta_2(rGDPg_t) + SP500_{t-1} + \varepsilon_t \qquad \textbf{(3)}$$

Figure 12: Model 3 Estimates

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
NUM1	-3.83694	1.55826	-2.46	0.0160	0	BBB	0
NUM2	0.88518	0.20402	4.34	<.0001	0	rGDPg	0

Figure 13: Model 3 Residual Analysis



When an AR1 and MA1 are added to get Model (4) the model also produces white noise in the residuals, correct sign on the coefficients, significant p-values in a 9 quarter forecast period, and no serial correlation (see Figures 14, 17, 18 and 19). However when a 12 quarter forecast period is used the p-values of the AR term and MA term become 0.8859 and 0.8999 which indicates there is a stability problem over time (see Table 2). Interesting enough is that a forecast period of 16 quarters has p-values that are significant however the coefficients on the AR and MA terms switch signs (see Figure 16). Coefficient sign changes are another indicator of an unstable model. In all three forecast periods the correlation between the AR and MA terms remain high and during the 12 quarter forecast they become 0.999 which is almost perfect correlation (see Figure 14). Also of interest is that the cross correlation at lag one of the BBB looks to be slightly significant (see Figure 15). From this analysis the model structure with an AR and MA term is deteriorating. It can be seen that a model without the over-fitting of AR and MA terms results in a stable model.

$$\widehat{SP500}_t = \beta_1(BBB_t - BBB_{t-1}) + \beta_2(rGDPg_t) + \beta_3(SP500_{t-1} - SP500_{t-2}) + \beta_4(\varepsilon_{t-1}) + SP500_{t-1} + \varepsilon_t \quad \textbf{(4)}$$

Table 2: Over-fitted SP500 Model (4) P-values

	9Q	**12Q**	**16Q**
MA1,1	<.0001	0.8859	<.0001
AR1,1	<.0001	0.8999	<.0001
BBB(1)	0.0162	0.0064	0.0064
rGDPg	<.0001	<.0001	0.0003

Figure 14: Model 4 Estimates

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.92127	0.10613	8.68	<.0001	1	sp500F	0
AR1,1	0.97150	0.06223	15.61	<.0001	1	sp500F	0
NUM1	-3.80745	1.54944	-2.46	0.0162	0	BBB	0
NUM2	1.16690	0.27673	4.22	<.0001	0	rGDPg	0

Figure 15: Model 4 Estimates 12Q

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.35082	2.43574	0.14	0.8859	1	sp500F	0
AR1,1	0.31202	2.47188	0.13	0.8999	1	sp500F	0
NUM1	-4.39646	1.56574	-2.81	0.0064	0	BBB	0
NUM2	0.81716	0.19678	4.15	<.0001	0	rGDPg	0

Figure 16: Model 4 Estimates 16Q

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	-0.96099	0.10265	-9.36	<.0001	1	sp500F	0
AR1,1	-0.98681	0.06998	-14.10	<.0001	1	sp500F	0
NUM1	-4.41985	1.57241	-2.81	0.0064	0	BBB	0
NUM2	0.77015	0.20307	3.79	0.0003	0	rGDPg	0

Figure 17: Model 4 Residual Analysis

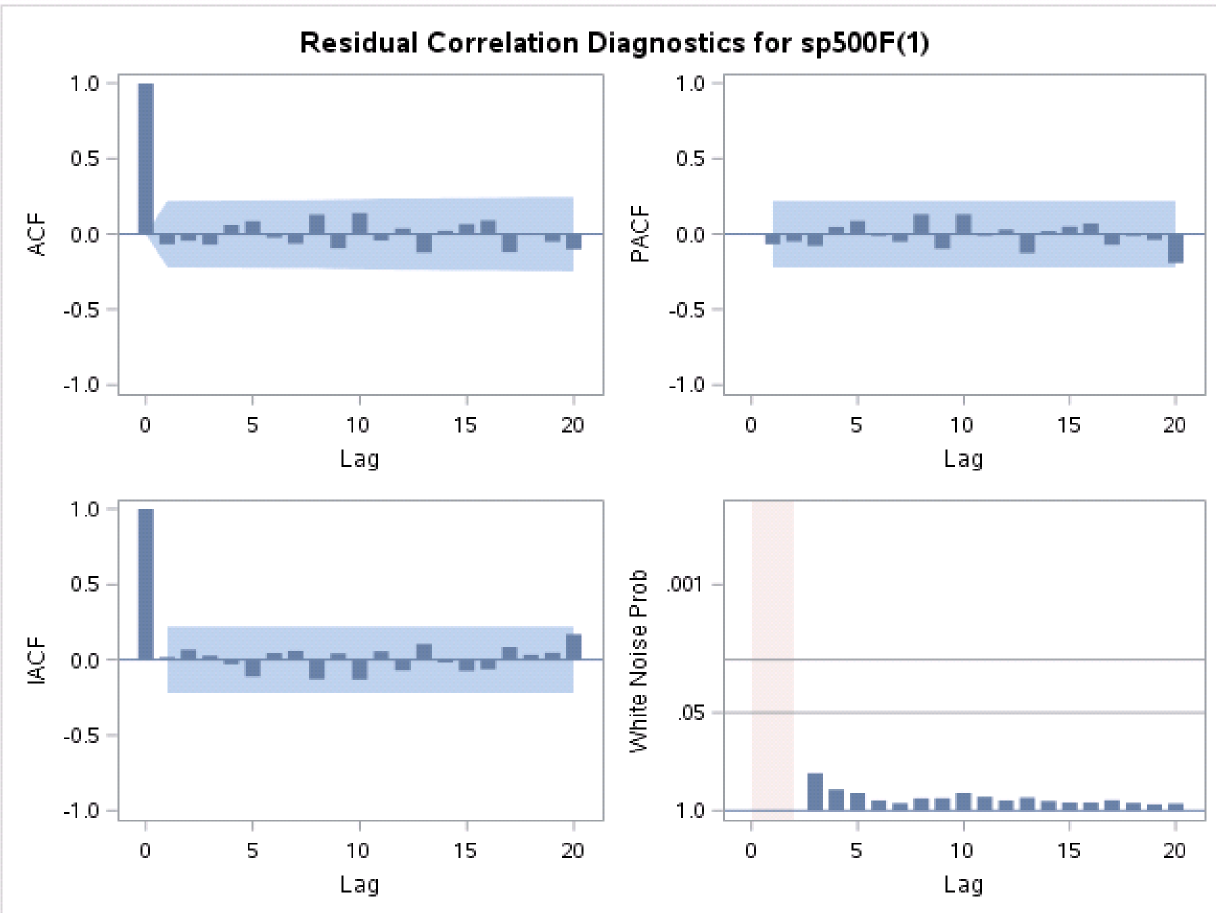


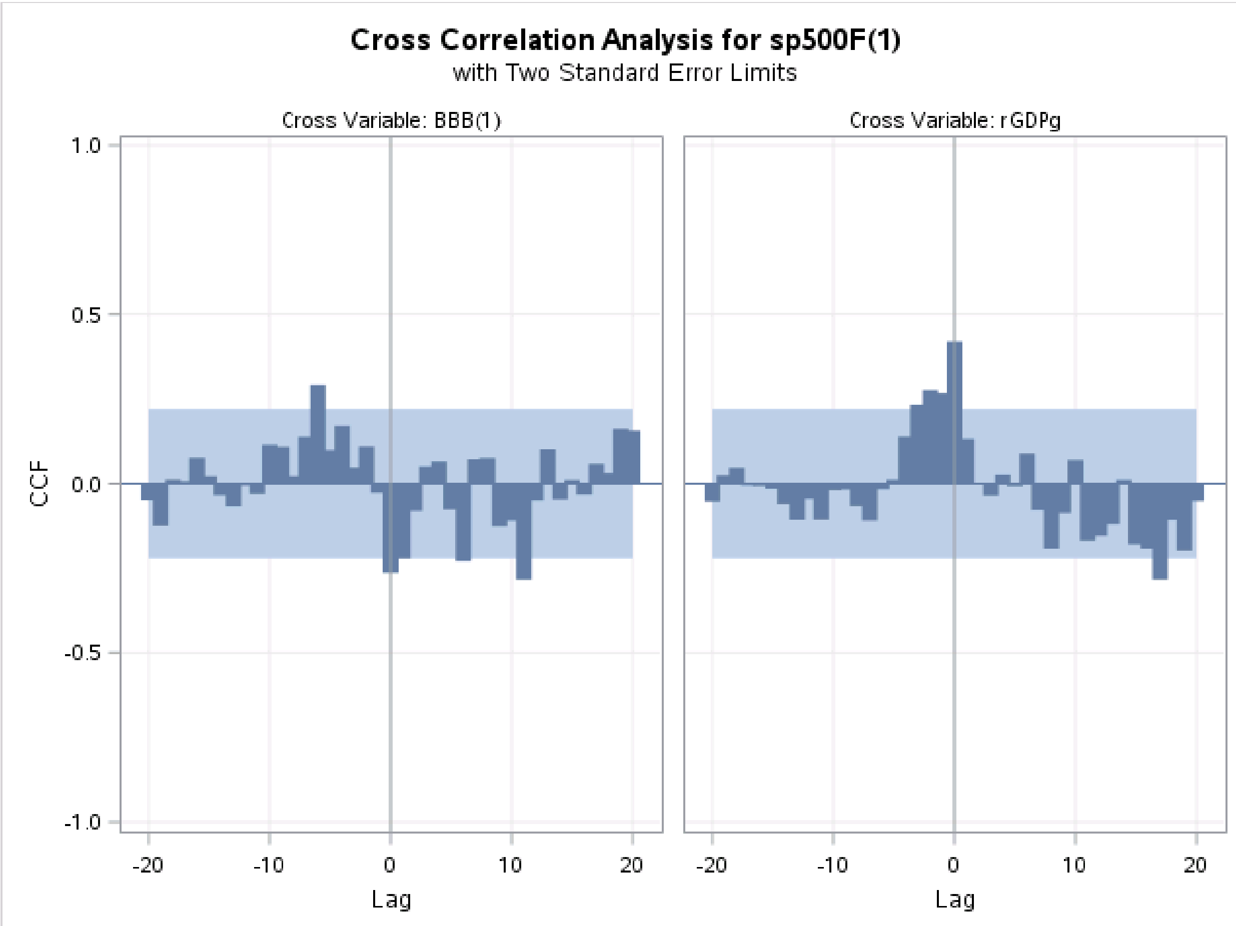
Figure 18: Model 4 Correlation Analysis for 9Q, 12Q, 16Q

Correlations of Parameter Estimates					
Variable Parameter		sp500F MA1,1	sp500F AR1,1	BBB NUM1	rGDPg NUM2
sp500F	MA1,1	1.000	0.881	0.078	-0.004
sp500F	AR1,1	0.881	1.000	0.036	0.118
BBB	NUM1	0.078	0.036	1.000	0.041
rGDPg	NUM2	-0.004	0.118	0.041	1.000

Correlations of Parameter Estimates					
Variable Parameter		sp500F MA1,1	sp500F AR1,1	BBB NUM1	rGDPg NUM2
sp500F	MA1,1	1.000	0.999	0.030	-0.059
sp500F	AR1,1	0.999	1.000	0.029	-0.058
BBB	NUM1	0.030	0.029	1.000	0.103
rGDPg	NUM2	-0.059	-0.058	0.103	1.000

Correlations of Parameter Estimates					
Variable Parameter		sp500F MA1,1	sp500F AR1,1	BBB NUM1	rGDPg NUM2
sp500F	MA1,1	1.000	0.945	0.050	0.089
sp500F	AR1,1	0.945	1.000	0.073	0.090
BBB	NUM1	0.050	0.073	1.000	0.111
rGDPg	NUM2	0.089	0.090	0.111	1.000

Figure 19: Cross Correlation between SP500 and BBB



Over-fitting is common and sometimes difficult to avoid when the data and ARIMA methodology is not understood very well. In practice I have seen models used for CCAR where the model had one independent variable and six other terms that were either AR or MA terms. This example is somewhat extreme however it sheds light onto a serious problem within PPNR modeling. As predicted that model failed (p-values became insignificant) after one use and had to be redeveloped. A model with the same variable was used however another variable was added and only one AR term was kept. This new model remained stable at the next CCAR test and redevelopment was not needed. This example indicates that a developer over-fitted which resulted in an unstable model. The other possible outcome would have been no other stable ARIMA model could be found. The solution would have been to re-analyze the data and make sure the right model structure was selected. For example, if serial correlation was not present then an OLS model might have been a better choice. Or perhaps there were variables that were cointegrated and the developer needed to move on to another method such as VECM. It is also possible that with the available data no model could be built. Again as the Federal Reserve pointed out, *"BHCs should not use weak models just for the sake of using a modeled approach to PPNR.* (Capital Planning at Large Bank Holding Companies: Supervisory Expectations and Range of Current Practice, 2013)*"* BHC stands for bank holding company.

5. Variable Selection

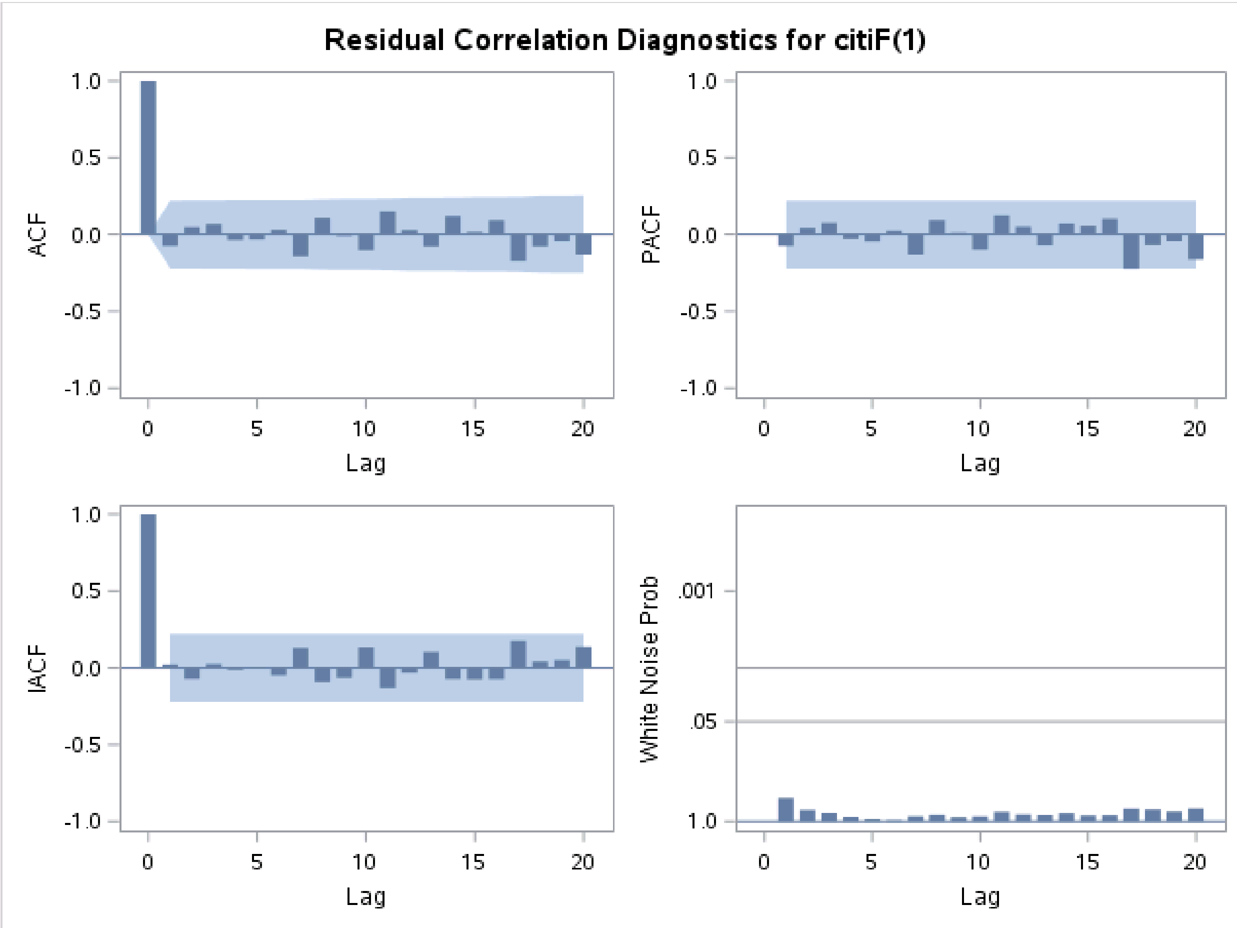
The variable selection step is one of the most important steps of model development. One of the most common mistakes in PPNR modeling is the issue of selecting a variable from a class of variables. An example of a variable class would be interest rates. When modeling accounts that are tied to loans it is reasonable to use interest rates as an independent variable. When modeling Citi with interest rates it was difficult to build a model with more than one variable including other non-interest rate variables. The 5 year, 3 month, and prime rates all created fairly well behaved models however these models are solely dependent on one variable. The best two variable model found was the use of the 5 year and rGDPg where the p-value was marginally significant for rGDPg at 0.0532 and significant for the 5 year at 0.0367. Besides the marginal p-values there was white noise in the residuals, correct sign on the coefficients, and no serial correlation (see Figures 20 and 21). After testing the model over the 9Q, 12Q, and 16Q periods the p-values only changed slightly and remained in their original ranges (i.e. marginal and significant). Many developers, validators, and oversight boards would hear out the reasoning of having a robust model with two variables, model stability, and good forecasts as long as rGDPg made sense.

$$\widehat{C}_t = \beta_1(ir5yr_t - ir5yr_{t-1}) + \beta_2(rGDPg_t) + C_{t-1} + \varepsilon_t \tag{5}$$

Figure 20: Model 5 Estimates

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
NUM1	18.82721	9.59384	1.96	0.0532	0	ir5yr	0
NUM2	2.45793	1.15680	2.12	0.0367	0	rGDPg	0

Figure 21: Model 5 Residual Analysis



The alternative would be to select a model with only one variable that has significant p-values and is stable over time. In the industry the problem usually comes down to developers selecting the variable with the lowest p-value. This practice is inappropriate and should not be done. Discussions with business managers or knowledge of the account should be taken into consideration when selecting the interest rate variable. No information is available on the inner workings of Citi and the best single interest rate variable would not be able to be selected unless an expert from Citi could shed some light on their accounts. A great example is modeling mortgage accounts. Many banks are involved in originating and holding mortgages. It is important to pull information about the maturity of the mortgages. Information should be available to show the distribution of loans between 10, 15, 20, and 30 year mortgages and the type of loan such as floating vs. fixed. If Citi's only business was US mortgages and the majority of the mortgages had maturities of fixed 30 years then it would make the most sense to use a 30 year interest rate model. However if Citi was more specialized in floating rate mortgages it would make more sense to use a shorter interest rate variable such as the 3 month interest rate. By selecting the wrong interest rate variable, especially in a single variable model, the model can deteriorate quickly as the spread between rates widen. The spread rate itself can give a bank a lot of information about current and future credit markets. The importance of spreads and specifically the risk premium on corporate bonds can be useful for banks to manage risk (Elton, Gruber, Agrawal, & Mann, 2001). A 10 year spread was created with the difference between the BBB and ir10yr (10-year Treasury yield). What is interesting is that the 10 year yield was not significant by itself or in combination with other variables. However a robust two variable model using the 10 year spread and CRE (Commercial Real Estate Price Index [Level]) was created. This model outperformed the other two variable model (5yr and rGDPg) by having two variables with significant p-values, white noise in the residuals, correct signs on the coefficients, no serial correlation, stability over the 9Q, 12Q, and 16Q tests, and a lower RMSE (see Figures 22 and 23 and Table 3).

$$\widehat{C}_t = \beta_1(Spread_t - Spread_{t-1}) + \beta_2(CRE_t - CRE_{t-1}) + C_{t-1} + \varepsilon_t \tag{6}$$

Figure 22: Model 6 Estimates

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
NUM1	-43.17599	9.01969	-4.79	<.0001	0	spread	0
NUM2	1.24348	0.51705	2.40	0.0185	0	CRE	0

Figure 23: Model 6 Residual Analysis

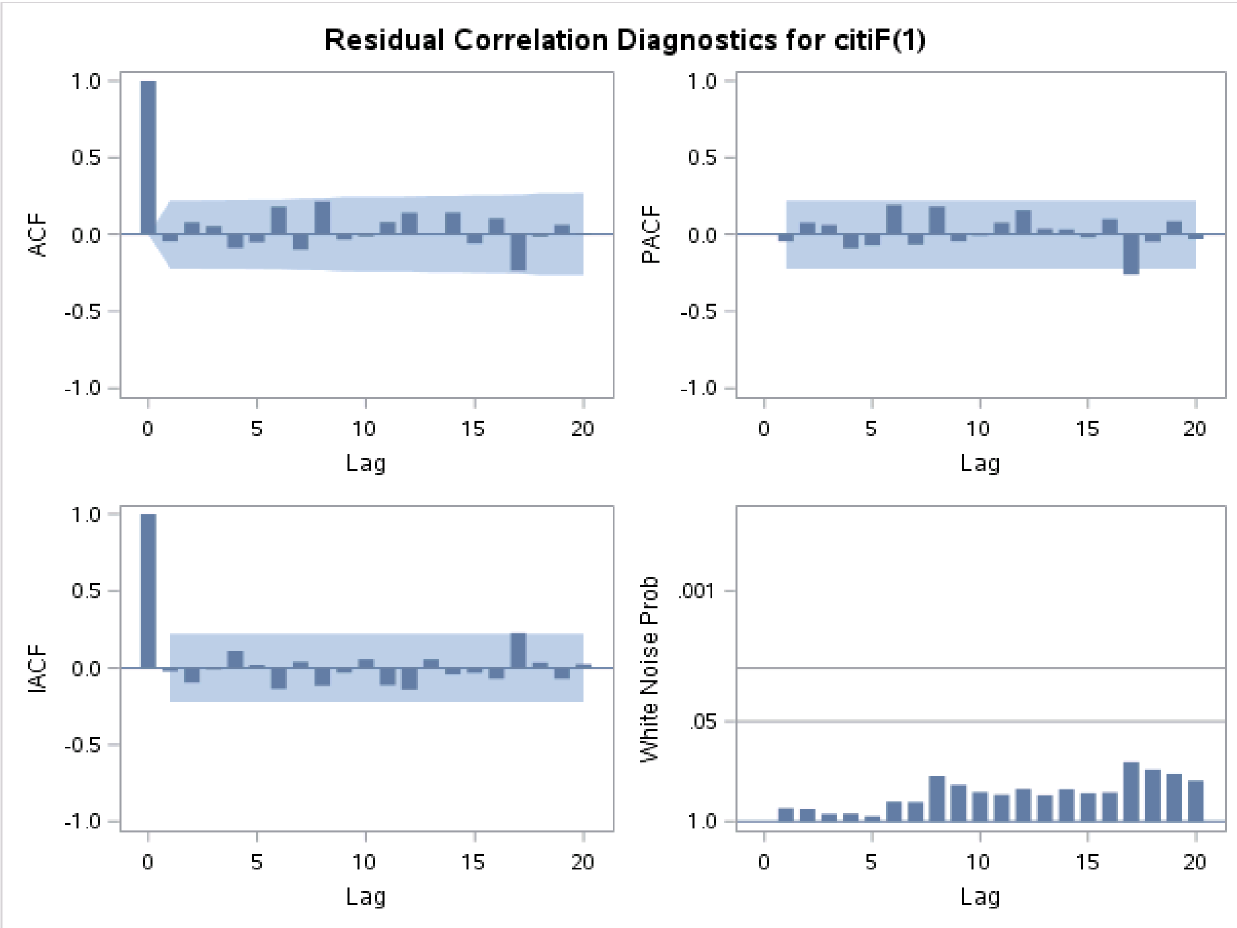


Table 3: RMSE (root mean square error) for Models 5 and 6

	Model 5	**Model 6**
RMSE	1,319.88	1,228.32

It can be seen that selecting a variable from a class of variables and variable selection in general needs business insight and consideration. Without the knowledge of spreads and an understanding of an account or business the variable selection can seem reasonable from a statistical point of view but can become unstable over time. It should also be seen that the forecasts from these two models have very different shapes and could result in very different forecasts (see Figures 24 and 25). Again business insight and specialized knowledge would be needed to select the most appropriate model. Many banks have included business insight in the variable selection process however the disconnect between how a model developer thinks and a business expert thinks creates substandard models. Both sides need to spend more time learning about the other's expertise. Simply approving models on p-values and/or a correct looking forecast is irresponsible. Model development should be part science, part art, and based on theory, not solely based on a basic list of criteria that must be met.

Figure 24: Model 5 Forecast

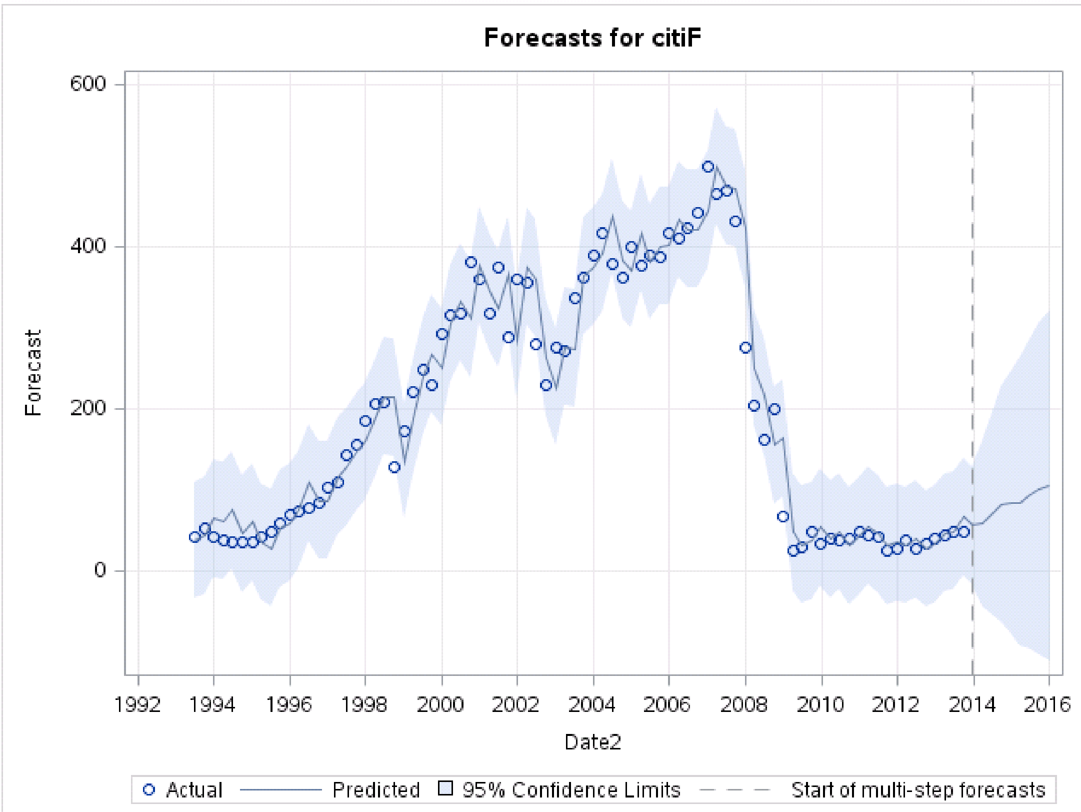
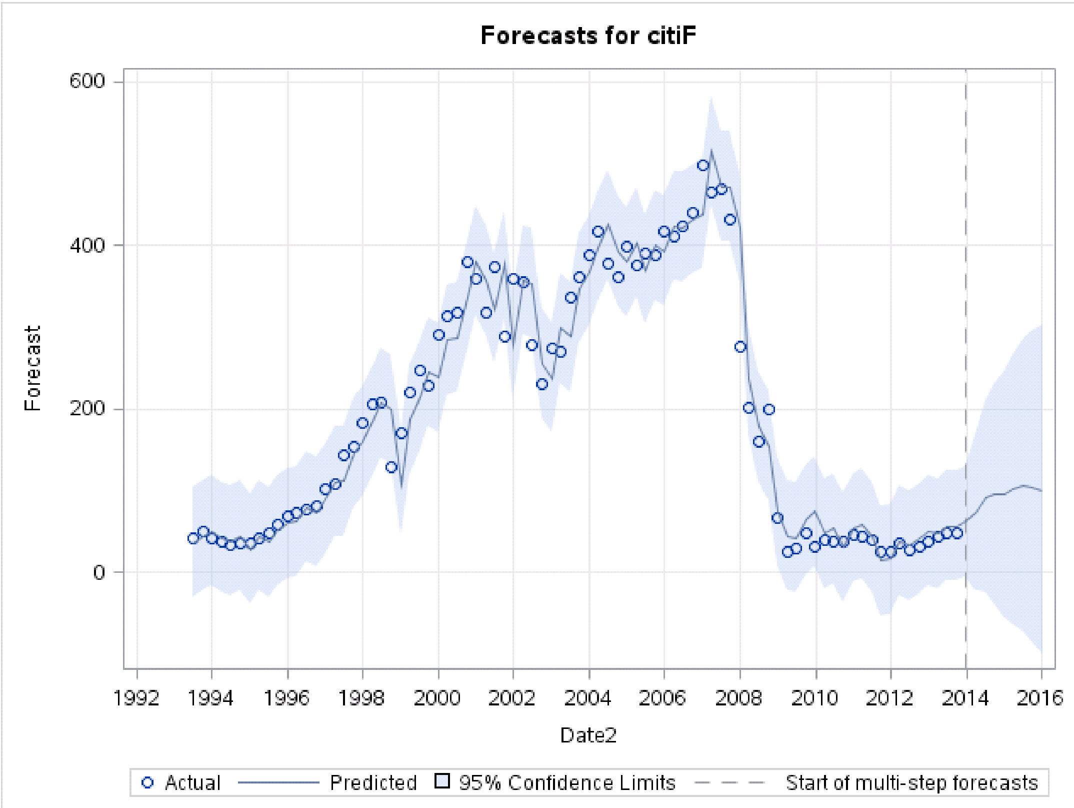


Figure 25: Model 6 Forecast



6. Conclusion

Using the ARIMA model structure has many pitfalls and stability issues if the models are not built and specified correctly. As it can be seen great attention to differencing, over-fitting AR and MA terms, and variable selection is needed to create stable models for stress testing. From an industry perspective I have seen PPNR developers at some banks refusing to use ARIMA due to the instability, some banks using ARIMA regardless of the data structures and issues, and very few instances of ARIMA being used in an appropriate setting. This is a cause for concern across risk management and banking as a whole. The main cause seems to be a lack of experts in ARIMA and the business. This lack of true expertise seems prevalent in other areas of risk management as well due to the increased requirements of risk management worldwide. Statistics is as much of an art as it is a science. The ARIMA model structure has great stability and predictive power especially in stress testing exercises when developers, validators, and board oversight understand the intricacies of the ARIMA structure.

For those refusing to use ARIMA, problems arise with finding solutions to correct for serial correlation. Some methods such as the Newey-West estimator have been used to correct for serial correlation as well as heteroskedasticity. This seems to be a good solution in some situations especially since OLS is well understood by many including those without a solid statistics background. The downfall is that the lack of AR and MA terms hides the understanding of the processes within the data. Some types of data theoretically and in practice have an autoregressive (AR) process which can be easily modeled through an ARIMA structure. Having the AR term with a coefficient shows how much impact the AR process has on the dependent variable. This can also be said for moving average (MA) processes.

For those abusing the ARIMA structure by applying it to every situation, many important intricacies within the data can be missed. Many of these banks will be those that are redeveloping models yearly. This is costly to banks as many more employees are needed to keep up with the regulatory requirements. As seen above, stability issues arise when the models are over or under differenced, data is differenced blindly without analysis of the data beforehand, over-fitting of AR and MA terms occurs, and poor variable selection is conducted. Other methods that should be considered would be GARCH (generalized autoregressive conditional heteroskedasticity) to deal with heteroskedasticity and VECM (Vector Error Corrective Models) to address cointegrated relationships when appropriate. Banks seem leery to use either GARCH or VECM in modeling PPNR due to the added complexities. This fear seems reasonable especially if issues arise in using an ARIMA structure, however model developers should be hired who have specialties in these areas.

As a concluding remark it should be noted that those who lack understanding of the intricacies of ARIMA or those who abuse its structure are causing fear in the group who is avoiding its use like the plague. This paper uses data that behaves very well for modeling purposes and it should be noted that internal PPNR is much more difficult to model. ARIMA as well as other modeling methods are being used by banks but due to the proprietary nature of banking many of their failures and success are not known across the industry.

References

Capital Planning at Large Bank Holding Companies: Supervisory Expectations and Range of Current Practice. (2013, August 01). Retrieved June 20, 2016, from http://www.federalreserve.gov/bankinforeg/bcreg20130819a1.pdf

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association, 74*(366a), 427-431. doi:10.1080/01621459.1979.10482531

Elton, E. J., Gruber, M. J., Agrawal, D., & Mann, C. (2001). Explaining the Rate Spread on Corporate Bonds. *The Journal of Finance, 56*(1), 247-277. doi:10.1111/0022-1082.00324

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics, 49*(1), 92. doi:10.2307/1937887

Hoover, K. D. (2003). Nonstationary Time Series, Cointegration, and the Principle of the Common Cause. *The British Journal for the Philosophy of Science, 54*(4), 527-551. doi:10.1093/bjps/54.4.527

Nesmith, T. D., & Jones, B. E. (2008). Linear Cointegration of Nonlinear Time Series with an Application to Interest Rate Dynamics. *Studies in Nonlinear Dynamics & Econometrics, 12*(1). doi:10.2202/1558-3708.1468

Reiss, J. (2015). *Causation, evidence, and inference* (1st ed.). New York, NY: Routledge.

Stress Tests and Capital Planning. (n.d.). Retrieved June 20, 2016, from http://www.federalreserve.gov/bankinforeg/ccar.htm

Wang, J., & Wang, S. (2010). *Business intelligence in economic forecasting: Technologies and techniques*. Hershey, PA: Information Science Reference.

Yahoo Finance - Business Finance, Stock Market, Quotes, News. (n.d.). Retrieved June 20, 2016, from http://www.yahoo.com/finance

Appendix A

Dependent Variables Chosen

Abbreviation	**Description**
C	Citigroup
SP500	Standard and Poor's 500 Index

Appendix B

Domestic Variables: Federal Reserve

Abbreviation	Description
rGDPg	Real GDP growth
nGDPg	Nominal GDP growth
rDIg	Real disposable income growth
nDIg	Nominal disposable income growth
LBR	Unemployment rate
CPI	CPI inflation rate
ir3m	3-month Treasury rate
ir5yr	5-year Treasury yield
ir10yr	10-year Treasury yield
BBB	BBB corporate yield
Mort	Mortgage rate
Prime	Prime rate
DJIA	Dow Jones Total Stock Market Index (Level)
HPI	House Price Index (Level)
CRE	Commercial Real Estate Price Index (Level)
Vol	Market Volatility Index (Level)