MWSUG BL-101

EXPLORING AND CHARACTERIZING TIME SERIES DATA IN A NON-REGRESSION BASED APPROACH

Steven C. Myers
Department of Economics, College of Business Administration
The University of Akron, Akron OH

ABSTRACT

Business leaders as well as data analysts and data scientists need to have an understanding of the particularities of time series data. This paper reports on an introduction to time series as taught to students in a first business analytics course making use of data from FRED, the marvelous time series repository at the St. Louis Federal Bank. Students are cautioned not to run to advance techniques before stopping to fully explore the data and this approach is designed to instill a EDA mentality into the students while teaching them how to manipulate and characterize time series data in SAS® and thereby set the ground work for more advanced work in time-series econometrics, forecasting and predictive analytics. Also, instilled in the students is an appreciation of knowing the data generating process. SAS programming is taught through this approach focusing on SAS functions such as DIF and LAG, PROCS CORR, MEANS, TABULATE and SGPLOT. The paper concludes with a basic coverage of random walk and spurious correlation that can easily result in mistakes when one does not first investigate data stationarity.

INTRODUCTION

This paper derives from class lectures and exercise about using actual economic data from FRED (fred.stlouis.org) and that class starts out orienting the students with exploratory data analysis of the time series in a non-regression based approach and encouraging a highly visual summary of the data. One motivation is to get students to slow down and appreciate the data generating process before they rush to higher statistical and mathematical methods which may be more fun to do, but necessarily obscure the larger notions of the data. Much of this lecture process is based on the work by Silvia, et al. (2014) especially their chapter 6 “Characteristics a Time Series Using SAS Software.”

The process we are following has us concentrating on the visualization of the data as well as using natural groupings over time to depict a set of statistics that help us interpret the series in one group of time versus another. Our focus is not to claim primacy over more serious statistical efforts of analyzing time series for their movement and forecasts, but to emphasis that in any exploratory data analysis one can learn much about the data generating process. That is, how does the data arise, what does it mean, and how can we use this to better our later sophisticated modeling.

Time series analysis also requires establishing the stationarity or nonstationarity of series through the use of a type of EDA using partial autocorrelations and graphics. This process of identifying the process requires transformation of the data such as in logs or first differences or both. A macro to create useful data transformations is offered in this paper and a discussion of the consequences of not knowing the time series data generating process concludes the paper.

ACQUIRING FRED DATA INTO SAS.

The Federal Reserve Bank of St. Louis maintains a total of 589,000 US and international time series data from 87 sources. All macroeconomic series are available there in nearly real time. All data is extremely well documented and a user interface is provided that allows for real time data exploration in a graphical context. For our purposes in this paper we must download the desired time series to SAS. For the first part of the paper various labor force statistics, specifically employment-population ratios for the US, total, black, men and women are acquired. In the latter part of the paper macroeconomic series measuring the gross domestic product, the money supply, the federal debt and inflation are acquired.

There are at least four ways to get these data from FRED into SAS.

(1) IF YOU HAVE SAS/ETS YOU CAN DOWNLOAD FRED DATA BY WAY OF THE SASEFRED INTERFACE ENGINE.

This requires that you request and receive an API key from FRED. You will have to have a free account at FRED before requesting your own API key at https://research.stlouisfed.org/useraccount/apikeys. Once you have your API key then you can download the series

```
options validvarname=any;
title 'Acquire Employment Population data';
libname _all_ clear;
%let dir = d:\freddata;

libname fred sasefred "&dir"
    OUTXML=epop
    XMLMAP="&dir\exportgs.map"
                 START='2002-01-01'
                        END='2019-04-01'
/* your 32-character alphanumeric API key goes here. */
    APIKEY='xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
/* IDLIST is comma delimited with no spaces between the single quotes. */
IDLIST='EMRATIO,LNS12300006,LNS12300001,LNS12300002,UNRATE,LNS14000006,LNS14000001,LNS14
000002';
data work.export_epop;
    set fred.epop ;
    run;
proc contents data=work.export_epop; run;
proc print data=work.export_epop(obs=15); run;
```

Each download from the IDLIST creates a fredtpx.sas7bdat file in &dir where fred is the libname and x=the number order of the variable in the ID list. These files contain lots of useful information and are merged together, also in &dir, into *merged_freq.sas7bdat*. For each of the x variables this file includes a line with series name, Title, Frequency, units of measurement, whether seasonality adjusted, and date range.

An undocumented requirement is the use of START= and END=. The engine runs without them and brings in all the data available, however, if the data carry blanks at the beginning then the series delivered is character and not numeric based. So far in my trails, the use of START= and END=, seem to do the trick.[1]

A useful next step is putting the titles from *merged_freq.sas7bat* as labels into the exported data file (not shown).

(2) EXPORTING DATA TO EXCEL BY THE FRED ADD-IN FOR MICROSOFT EXCEL

The same data can be downloaded directly in Excel by using a FRED add-in. “The Federal Reserve Bank of St. Louis Economic Data (FRED) Add-In is free software that will significantly reduce the amount of time spent collecting and organizing macroeconomic data. The FRED add-in provides free access to over 580,000 data series from various sources (e.g., BEA, BLS, Census, and OECD) directly through Microsoft Excel.”[2]

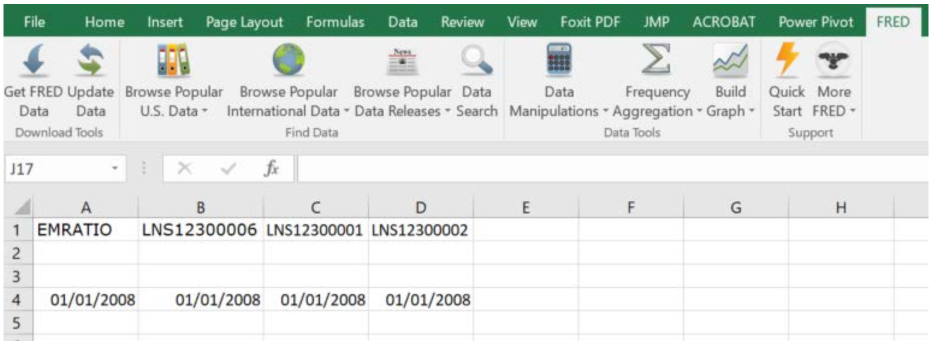
By way of demonstration, the 4 Employment Ratios can be requested by entering their names into the first 4 cells (see Figure 1). To limit the series only from 2008 one needs only to enter the start date,

[1] I am grateful to the SASEFRED developer, Kelly Fellingham, and SAS Support guru, Bari Lawhorn, for their help in resolving the character type data problem.

[2] FRED Add-In for Microsoft® Excel®, accessed at https://fred.stlouisfed.org/fred-addin/.

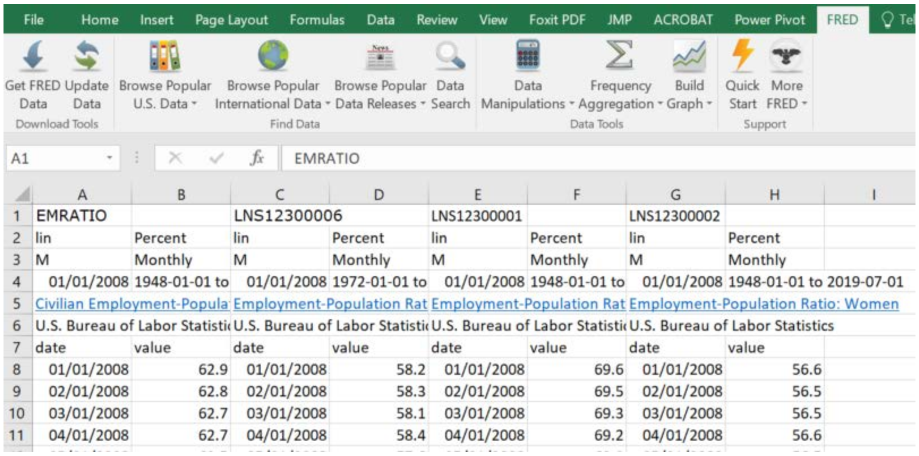
1/1/2008, in row four for each variable. Variable names can be found through searches on FRED in your browser or by use of the browse and data search buttons in Excel. Once Excel looks like Figure 1, pressing "Get FRED Data" causes the results as shown in Figure 2.

Figure 1: Set up for downloading FRED data



	A	B	C	D	E	F	G	H
1	EMRATIO	LNS12300006	LNS12300001	LNS12300002				
2								
3								
4	01/01/2008	01/01/2008	01/01/2008	01/01/2008				
5								

Figure 2: Downloaded FRED Data



	A	B	C	D	E	F	G	H	I
1	EMRATIO		LNS12300006		LNS12300001		LNS12300002		
2	lin	Percent	lin	Percent	lin	Percent	lin	Percent	
3	M	Monthly	M	Monthly	M	Monthly	M	Monthly	
4	01/01/2008	1948-01-01 to	01/01/2008	1972-01-01 to	01/01/2008	1948-01-01 to	01/01/2008	1948-01-01 to 2019-07-01	
5	Civilian Employment-Popula	Employment-Population Rat		Employment-Population Rat		Employment-Population Ratio: Women			
6	U.S. Bureau of Labor Statisti	U.S. Bureau of Labor Statisti		U.S. Bureau of Labor Statisti		U.S. Bureau of Labor Statistics			
7	date	value	date	value	date	value	date	value	
8	01/01/2008	62.9	01/01/2008	58.2	01/01/2008	69.6	01/01/2008	56.6	
9	02/01/2008	62.8	02/01/2008	58.3	02/01/2008	69.5	02/01/2008	56.5	
10	03/01/2008	62.7	03/01/2008	58.1	03/01/2008	69.3	03/01/2008	56.5	
11	04/01/2008	62.7	04/01/2008	58.4	04/01/2008	69.2	04/01/2008	56.6	

Preparing Excel for SAS PROC IMPORT

Use three tabs in each Excel workbook. Name them ORIGINAL, SOURCE, and SASDATA.

To get the data of Figure 2 into SAS, a best practice is to rename the tab to "original data" or "raw data." Do not edit anything else in this original tab. Add a second tab and name it "SOURCE" including the citation to the data and each data series, include links back to FRED, perhaps you can create a data dictionary and include it in this tab. Add a third tab and name it "SASDATA." Copy all of the original tab information to SASDATA and then put SAS-legal variable names in the place of the words "value" in line 7. Verify and adjust if all dates and frequencies are not the same, then delete duplicate date columns and delete rows 1 to 6. Save the file to a directory in xlsx format. You now have an Excel Tab that you can import directly into SAS using code such as:

```
PROC IMPORT OUT= WORK.EPOP
    DATAFILE= "d:\freddata\E-POP-data.xlsx"
    DBMS=xlsx REPLACE;
    SHEET='sasdata';
    RUN;
```

(3) DOWNLOADING DATA USING THE DATA LISTS OPTION UNDER MY ACCOUNT IN FRED.

Visit https://fred.stlouis.org, go to the menu bar about half way down the page that includes "AT A GLANCE," and other choices including "TOOLS." Select Tools and then select "My Data Lists." You can also arrive at the same page by clicking on the arrow at "My Account" in the top right hand corner of every page. Look for the "+ Add New" button and choose data list. Give it a name. Find and click on that name in your list and follow the directions.

(4) DOWNLOADING DATA ASSOCIATED WITH A GRAPH YOU HAVE DESIGNED.

Visit any graph you have made or create a new graph loading it up with the series you want to download.[3] All graphs can be saved as part of your account. Look for and click the blue download button and choose CSV (data) or Excel (data). Choose Excel and a *.xls file will be downloaded to where you direct. Because PROC IMPORT uses an xlsx engine you will need to resave the downloaded data set as an *.xlsx file.

This method is only good for small data pulls and is very convenient when you have been looking at the graphs as you select the series you need. That is how the first data used for this paper were downloaded. Actually two excel files were downloaded, one with the percentage terms of the original data series and one where the series were transformed by FRED to be indexed to equal 100 at the end of the last recession (Jun 2009). The two files need to be merged after preparing the SASDATA tab in each spreadsheet and that code follows:

```
PROC IMPORT OUT= WORK.EPOP
          DATAFILE= "d:\EPOP\E-POP and U-rates FRED data.xlsx"
          DBMS=xlsx REPLACE;
          SHEET='sasdata';
          RUN;
PROC IMPORT OUT= WORK.EPOPpct
          DATAFILE= "d:\EPOP\E-POP in percent terms and U-rates FRED data.xlsx"
          DBMS=xlsx REPLACE;
          SHEET='sasdata';
          RUN;
proc sort data=epop; by date; run;
proc sort data=epoppct; by date; run;
Data work.temp;
          merge work.EPOP work.epoppct; by date;
          run;
```

EXPLORATION OF SINGLE SERIES

Each time series needs to be explored or characterized before moving to more sophisticated work including multivariate analysis. This paper does that by first looking at an economic question: How has employment recovered since the great recession? The growth of employment for each of various populations are compared to see how each group fared separately.

Since the end of the great recession that began in December 2007 and ended in June of 2009, the recovery was long and slow as compared to previous recessions.[4] Dubbed the employment recession, the number of jobs lost during the recession were not restored to prerecession levels for almost 7 years with all other modern recessions restoring jobs in 1 to 2.5 years. The only exception was the 2001 recession that was not very deep, but did last 4 years before jobs returned. When measured against the population, the civilian population ratio was 62.9 (62.9 percent of the population) at the outset of the great recession and as of August 2019 has not returned to the prerecession level. It currently stands at 60.0

[3] One of the graphs used to acquire data for the first part of the paper can be accessed here at https://fred.stlouisfed.org/graph/?g=oEKv

[4] See Cunningham (2018) and McBride (2018) for evidence of the length and depth of the Great Recession.

percent of the population. This is not necessarily a bad thing because so many things go into the decision to work, but is interesting to explore.

The Civilian Employment Population ratio from July 2009 to July 2019, adjusted for seasonality indexed such that June 2009 equals 100 is drawn for the total civilian population, EPOPTOT; for men, EPOPMEN; for women, EPOPWOMEN; and for Black, African-Americans; EPOPBLACK. For comparison the non-indexed / original data measured in percentage terms are also downloaded.

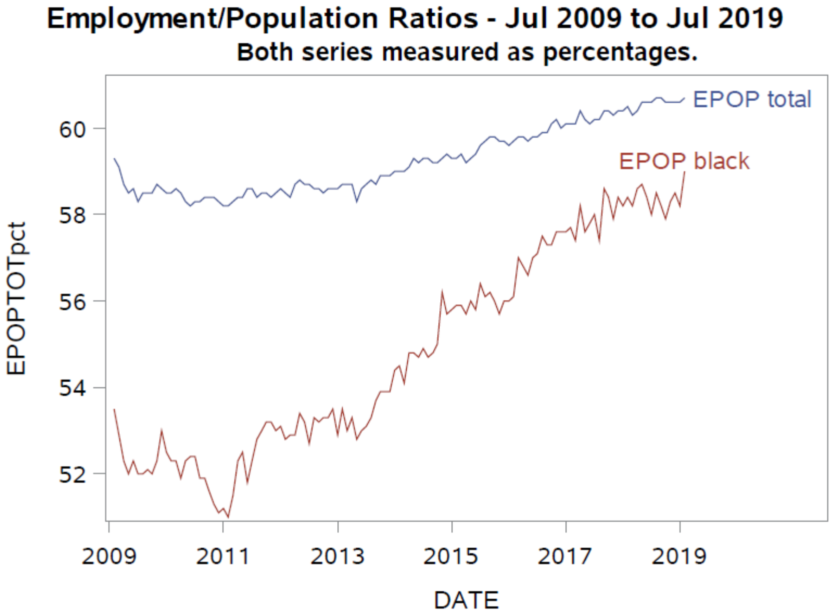
Preliminary viewing of the data on the FRED website makes one series especially interesting. The Black employment-population ratio has the highest growth coming out of the recession. The first step to broader possible questions is getting the data into SAS for deeper analysis.

To examine the growth of employment-population rates for blacks versus all workers we make use of PROC SGPLOT:

```
Title1 'Employment/Population Ratios - Jul 2009 to Jul 2019';
title2 'Series measured as percentages of the respective population. ';
proc sgplot data=work.temp;
    series x=date y=EPOPTOTpct / curvelabel='EPOP total';
    series x=date y=EPOPblackpct  /curvelabel='EPOP black';
    format date year4. ;
    xaxis values=('1jun09'd to '1jul19'd by year);
    run;
```

Which yields the results shown in Figure 3.

Figure 3: Employment Population Ratios in levels (percent of the respective population).



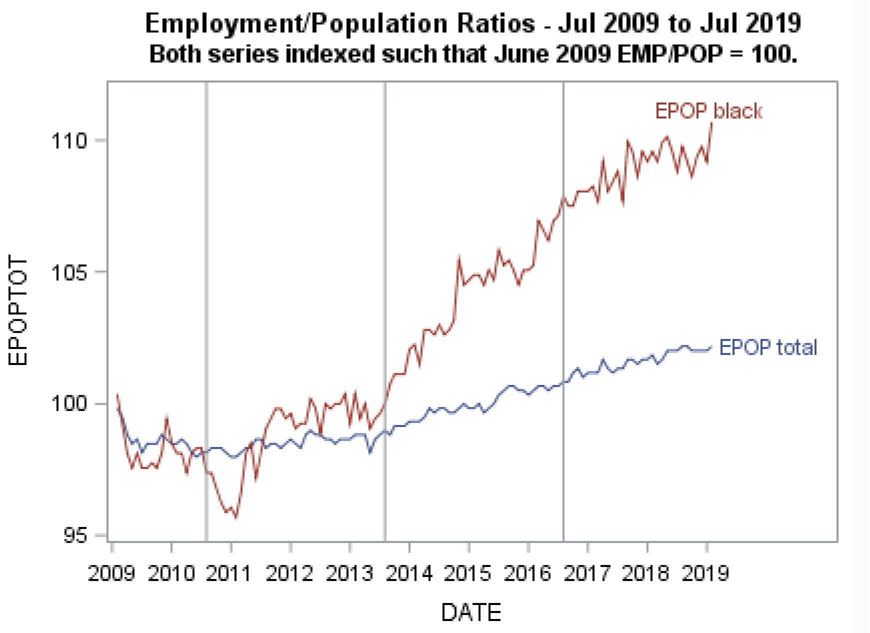
Total employment-population ratio (EPOP) has risen by 1.3 percentage points, but the Black EPOP ratio have dramatically risen more. Although black/African Americans start at the lowest level of all other measured groups, the growth appears the fastest.

Another way to show this is to rerun the code with the measures of EPOP being indexed so that July 2009 equals 100 and to overlay them again on a plot using similar code but substituting EPOPTOTpct for EPOPTOT and EPOPBLACKPct for EPOPBLACK. Also added to the code is a set of reference lines (refline) that visibly divide the visual into 4 time-slices.

```
title2 'Both series indexed such that June 2009 EMP/POP = 100. ';
proc sgplot data=work.temp;
    series x=date y=EPOPTOT / curvelabel='EPOP total';
    series x=date y=EPOPblack  /curvelabel='EPOP black';
    refline '1jan11'd '1jan14'd '1jan17'd /axis=x;
    format date year4. ;
    xaxis values=('1jun09'd to '1jul19'd by year);
    run;
```

The results are in Figure 4 and show more dramatic growth based on the two series starting out at the same point (100 in July 2019). Moreover the differences are more dramatic starting in 2014.

Figure 4: Relative Employment Population growth.



USE OF DESCRIPTIVE STATISTICS OVER PORTION OF THE TIME SERIES

Each series can be characterized by its (1) central tendency, (2) volatility and (3) stability on both the entire length of the data we have, and by various time-slices of the data.

The mean, $\bar{X}_i$, of a series on different portions of time (i) indicates the central tendency of the series, such as the average value over each decade or between each trough and peak of the business cycle. The standard deviation, s, of a series on different portions of time indicates a relative measure of volatility of that series during each time-slice. The coefficient of variation, $s/\bar{X}_i$, gives a measure of the stability of the series over each time-slice.

To illustrate this a simple employment measure is used. Hence our problem can be expressed as:

How do the two series behave over various time-slices of the post-recession months?

The following code creates a variable, group, to divide the time-series into various time slices. In this case, some partially arbitrary points were chosen to yield 4 sets of years that corresponds to the reference lines added in the last graph:

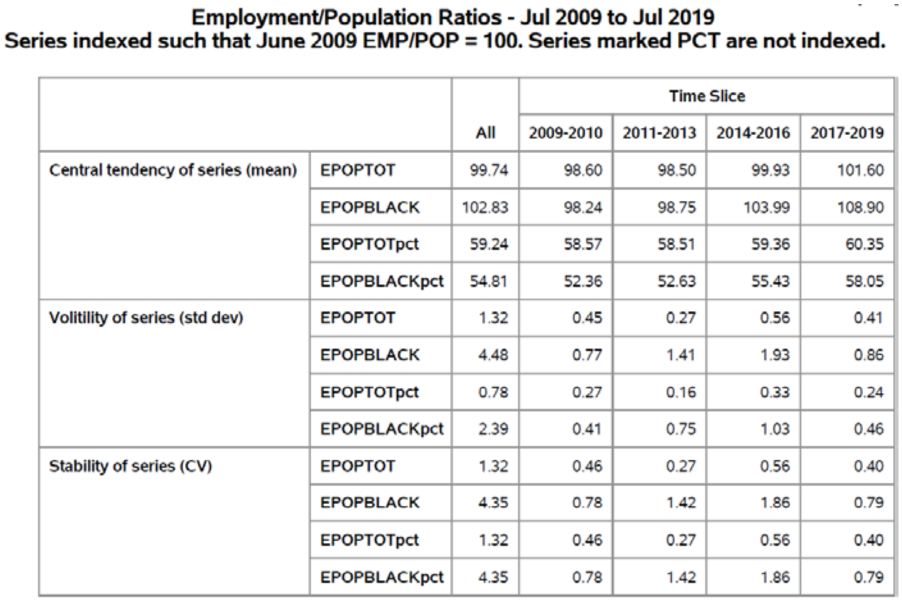
```
length group $9;
if date ge '1jan09'd and date lt '1jan11'd then group='2009-2010';
if date ge '1jan11'd and date lt '1jan14'd then group='2011-2013';
if date ge '1jan14'd and date lt '1jan17'd then group='2014-2016';
if date ge '1jan17'd and date lt '1jan20'd then group='2017-2019';
```

We find summary statistics on the four time-slices by running PROC TABULATE to calculate three statistics for our four different variables in each of the four time periods. Employment-population ratios for EPOPTOT and EPOPBLACK are indexed to be 100 in July 2009 and the variables EPOPTOTpct and EPOPBLACKpct are the original variables expressed in percentage of the population. The statistics are calculated by the following code:

```
title2 'Series indexed such that June 2009 EMP/POP = 100. Series marked PCT are not
indexed.';
Proc tabulate data=temp;
    class group;
    var epoptot epopblack  epoptotpct epopblackpct;
    Table  ( mean='Central tendency of series (mean)'
             stddev='Volitility of series (std dev)'
             cv='Stability of series (CV)')*
           ( epoptot epopblack epoptotpct epopblackpct)
           , ( group='Time Slice');
           run;
```

And the output results are shown in Table 1.

Table 1: Means, Volatility and Stability of time series by time-slice



Employment/Population Ratios - Jul 2009 to Jul 2019
Series indexed such that June 2009 EMP/POP = 100. Series marked PCT are not indexed.

			Time Slice			
		All	2009-2010	2011-2013	2014-2016	2017-2019
Central tendency of series (mean)	EPOPTOT	99.74	98.60	98.50	99.93	101.60
	EPOPBLACK	102.83	98.24	98.75	103.99	108.90
	EPOPTOTpct	59.24	58.57	58.51	59.36	60.35
	EPOPBLACKpct	54.81	52.36	52.63	55.43	58.05
Volitility of series (std dev)	EPOPTOT	1.32	0.45	0.27	0.56	0.41
	EPOPBLACK	4.48	0.77	1.41	1.93	0.86
	EPOPTOTpct	0.78	0.27	0.16	0.33	0.24
	EPOPBLACKpct	2.39	0.41	0.75	1.03	0.46
Stability of series (CV)	EPOPTOT	1.32	0.46	0.27	0.56	0.40
	EPOPBLACK	4.35	0.78	1.42	1.86	0.79
	EPOPTOTpct	1.32	0.46	0.27	0.56	0.40
	EPOPBLACKpct	4.35	0.78	1.42	1.86	0.79

By the numbers we can see that relative to the total employed, those who were black began to pull away by 2014 to higher levels. Blacks showed far more volatility and less stability than the total population.

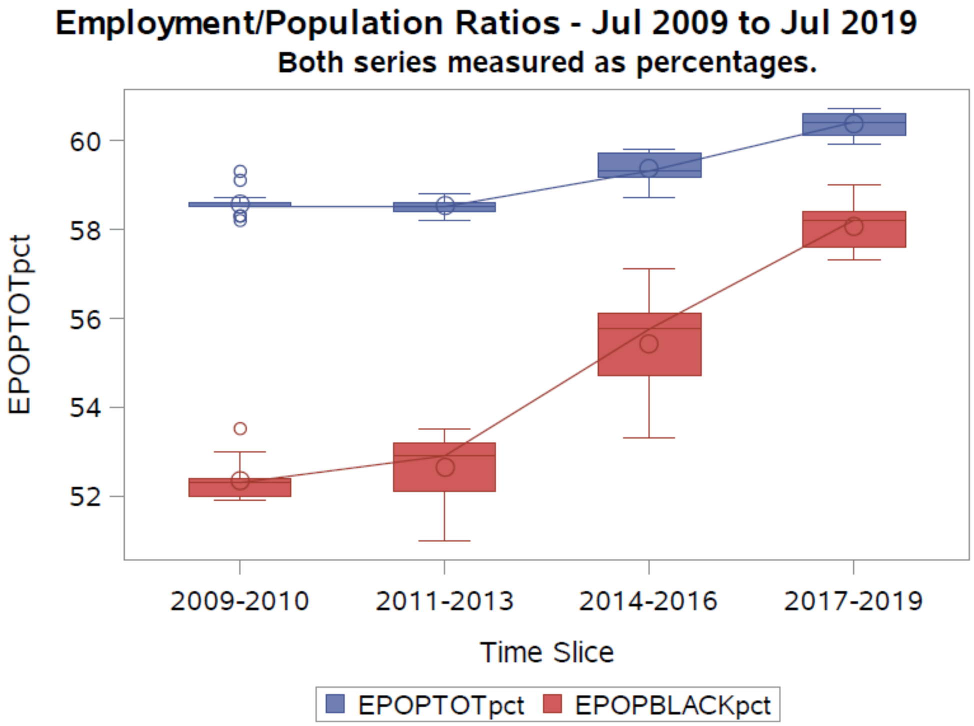
While it is useful to have the statistics, can visually display the pattern by using box plots. Using PROC SGPLOT we add the following code.

```
title2 'Both series measured as percentages. ';
proc sgplot data=temp ; label group='Time Slice';
    vbox EPOPTOTpct  / category = group boxwidth=0.50 connect=median ;
    vbox Epopblackpct/ category = group boxwidth=0.50 connect=median ;
    run;

title2 'Both series indexed such that June 2009 EMP/POP = 100.';
proc sgplot data=temp; label group='Time Slice';
    vbox EPOPTOT / category=group boxwidth=0.40 discreteoffset=-0.25 connect=median;
    vbox Epopblack/ category=group boxwidth=0.40 discreteoffset= 0.25 connect=median;
    run;
```

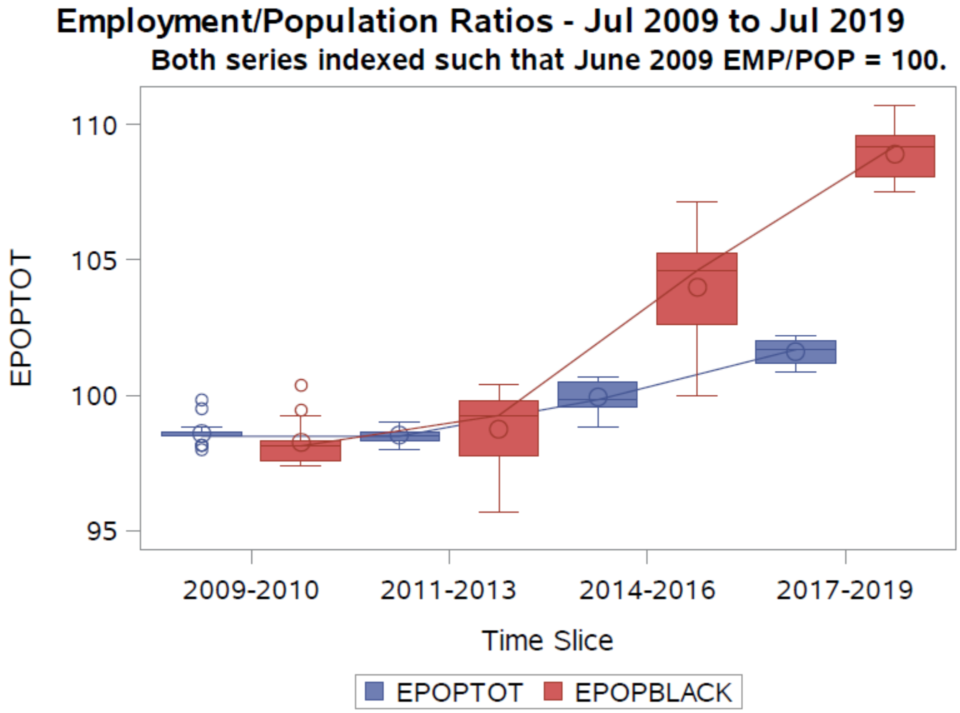
Which produces the following two graphics. Figure 5 shows the box plots based on the actual percentage data and Figure 6 is based on the indexed data.

Figure 5: Employment-Population Ratios in levels



Again, we can see the dramatic rise of the employment of blacks relative to the total population. However the next graph shows it even more dramatically by having the data indexed to start out at the same point. Using the option discreteoffset, we can show the two series side-by-side and avoid visual data collisions.

Figure 6: Employment-Population Ratios: Relative growth from end of Great Recession (indexed 6/2009=100)



DATA TRANSFORMATIONS – THE DIF MACRO

When working with time-series data there are many times that the researcher will want to LAG the data or LOG the data or both. The researcher may want many and varied transformations, perhaps in first differences, perhaps in log difference forms. [5]

FRED allows the transformation of variables on their website. Table 2 shows all of the transformations that you can apply in FRED to visualize your series with the mathematics and the SAS code to do the same transformation in SAS. The DIF macro shown below when called with a levels variable specified will create the 11 new SAS variables and add to the data set which includes 7 different growth rates.

The beauty of this macro is you can customize it to meet your individual, but frequent, needs. Since time series data are in the realm of small data the extra storage used is a good option for the convenience. The transformations follow a similar naming convention: Y_t is the original variable measured at time period t, n equals the number of observations in a year (n_obs_per_year) which is the frequency of the series. Regardless of what named series Y is used, the prefix shown is concatenated to create the new variable name, as can be seen in the table and below in the discussion of the macro call.

[5] The LAG and DIF functions are documented at https://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_tsdata_sect048.htm. The DIF function is not to be confused with the DIF macro created in this paper.

Table 2: What formulas are used to calculate growth rates and other transformations?[6]

Variable	Prefix	Math representation	SAS CODE
level	LEV_	Y_t	LEV_&id = &var;
One period lag	LAG1	Y_{t-1}	LAG1&id = lag(&var);
One year ago lag	LAGn	Y_{t-n}	LAG&n_obs_per_year&id =lag&n_obs_per_year(&var);
Change from one period ago	D1	$Y_t – Y_{t-1}$	D1&id=dif1(&var);
Change from 1 Year ago	Dn	$Y_t – Y_{t-n}$	D&n_obs_per_year&id =dif&n_obs_per_year(&var);
Percent change in variable, period over period	PCTD1	$((Y_t/Y_{t-1})-1)*100$	PCTD1&id=((&var/LAG1&id)-1)*100;
Percent change from one year ago, year over year	PCTDn	$((Y_t/Y_{t-n})-1)*100$	PCTD&n_obs_per_year&id =((&var / LAG&n_obs_per_year&id)-1)*100;
Compounded Annual Rate of Change	CARCn	$(((Y_t/Y_{t-1})^n)-1)*100$	CARC&n_obs_per_year&id =(((&var/LAG1&id)**&n_obs_per_year) 1)*100;
Log of variable	LOG	lnY_t	LOG&id=log(&var);
Lag of log variable	LAGLOG	lnY_{t-1}	LAGLOG&id=lag(LOG&id);
Continually Compounded Rate of Change	CCR	$(lnY_t - lnY_{t-1})*100$	CCR&id=(LOG&id-LAGLOG&id)*100;
Continually Compounded Annual Rate of Change	CCAR	$(lnY_t - lnY_{t-1})*100*n$	CCAR&id =(LOG&id-LAGLOG&id)*100*&n_obs_per_year;

The code for the DIF macro is:

```
%macro DIF(var,n_obs_per_year, id);

    /* id = short abreviation for the variable name */
    /* n_obs_per_year = frequency of the series */
    /* Annual = 1, Monthly = 12, Quarterly =4, Bi-weekly = 26, Weekly=52 */
    /* var =variable name of variable to be manipulated, Y in the examples below */

    /* level variable */
```

[6] Formulas are duplicative of the transformations that can be done in FRED. See Federal Reserve Bank of St. Louis. What Formulas are Used to Calculate Growth Rates accessed at https://fredhelp.stlouisfed.org/fred/data/understanding-the-data/formulas-calculate-growth-rates/

```
      LEV_&id = &var;

      /* lag variable */
      LAG1&id=lag(&var);
      /* if Y = variable on the RHS, this gives y(t-1) */

      /* lag n_obs_per_year variable */
      LAG&n_obs_per_year&id=lag&n_obs_per_year(&var);
      /* if Y = variable on the RHS, this gives y(t-1) */

      /* Change from one period ago */
      D1&id=dif1(&var);
      /* if y = variable on RHS, this gives y(t)-y(t-1) */

      /* Change from 1 Year ago */
      D&n_obs_per_year&id=dif&n_obs_per_year(&var);
      /* if y = variable on RHS, this gives y(t)-y(t-1) */

      /* percent change in variable, period over period */
      PCTD1&id= ((&var / LAG1&id)-1)*100;
      /* if y = variable on RHS, this gives [y(t)/y(t-1)-1]*100 */

      /* percent change from one year ago, year over year */
      PCTD&n_obs_per_year&id= ((&var / LAG&n_obs_per_year&id)-1)*100;
      /* if y = variable on RHS, this gives [y(t)/y(t-1)-1]*100 */

      /* Compounded Annual Rate of Change */
      CARC&n_obs_per_year&id=(((&var/LAG&id)**&n_obs_per_year )-1)*100;
      /* if y = variable on RHS, this gives [((y(t)/y(t-1)^n)-1]*100 */

      /* Log of variable*/
      LOG&id=log(&var);
      /* if y = variable on RHS, this gives ln(y) */

      /* Lag of log variable*/
      LAGLOG&id=lag(LOG&id);
      /* if y = variable on RHS, this gives ln(y) */

      /* Continually Compounded Rate of Change */
      CCR&id=(LOG&id-LAGLOG&id)*100;
      /* if y = variable on RHS, this gives lny(t) - lny(t-1) */

      /* Continually Compounded Annual Rate of Change */
      CCAR&id=(LOG&id-LAGLOG&id)*100*&n_obs_per_year;
      /* if y = variable on RHS, this gives lny(t) - lny(t-1) */

  %mend;
```

The macro is called by %DIF(var, n, id) where var is the SAS variable name of variable to be transformed, n is the frequency of the series in terms of number of observations in a year, and id is a unique short abbreviation of the variable. So by calling %DIF(EPOPTOT,12,EPT) the SAS time-series variable

EPOPTOT which is measured monthly will cause all transformations to be made to EPOPTOT and each of the created variables will have the prefix as shown in Table 2 concatenated to the short name, EPT.

We run the macro before the data step and further modify the data step by making calls to the DIF macro for each of the employment-population ratios of interest:

```
Data work.temp;
    merge work.EPOP work.epoppct; by date;
    length group $9;

    T=_N_;

    group='         ';
    if date ge '1jan09'd and date lt '1jan11'd then group='2009-2010';
    if date ge '1jan11'd and date lt '1jan14'd then group='2011-2013';
    if date ge '1jan14'd and date lt '1jan17'd then group='2013-2016';
    if date ge '1jan17'd and date lt '1jan20'd then group='2017-2019';

    %dif(EPOPTOT,12,EPT);
    %dif(EPOPmen,12,EPM);
    %dif(EPOPwomen,12,EPW);
    %dif(EPOPblack,12,EPB);

    run;
```

The first macro call, %dif(EPOPTOT,12,EPT), passes the named variable, EPOPTOT, to the macro and instructs the macro to use EPT as the id token that is combined in the macro with the appropriate prefix to name the transformed variables. The prefixes can be seen in the Table 2.

SERIES DESCRIPTIVE STATISTICS – THE EXPLORE MACRO

Volumes of graphs can be produced by reproducing the nearly same SGPLOT statements over all of the original and transformed variables. A second macro, named explore, produces for each variable (1) a table with descriptive statistics (mean, volatility and stability). (2) A time plot with an overlaid LOESS[7] (local regression) line overlaid and (3) a graph of vertical box plots by the defined time-slices in the variable group.

The only output that is created is for the variable specified outside of the %explore macro call and the transformations listed by prefix inside the macro. The number of pages presented by the %explore macro is equal to 3 times the number of prefixes coded in %var_list times the number of time-series listed in %name_list. EXPLORE[8] requires that the DIF macro has been executed on all variables of interest and that the variable group is appropriate named.

The macro Explore has one parameter, n, which is the frequency of the series being measured:

```
%macro explore(n);

%let var_list = LEV_ D&n PCTD&n CARC&n CCAR;
%local i next_name;
%local j next_var;
```

[7] LOESS is a nonparametric regression that can trace out the best regression through a scatter of points and is useful to explore patterns in data like here. For a good paper on LOESS see Bilenas (2014).

[8] This is a bit of programming that has its inspiration from Puryear (2015).

```
    %do i=1 %to %sysfunc(countw(&name_list));
          %let next_name = %scan(&name_list, &i);

    %do j=1 %to %sysfunc(countw(&var_list));
          %let next_var = %scan(&var_list, &j);

    %** DO whatever needs to be done for &NEXT_NAME;

    Title "Variable &next_var&next_name";
          proc means data=temp n mean std cv;
          var &next_var&next_name;
          by group;
          run;
    proc sgplot data=work.temp;
          series x=date y=&next_var&next_name ;
          loess x=date y=&next_var&next_name ;
          format date year4. ;
          xaxis values=('1jun09'd to '1jul19'd by year);
          run;

    proc sgplot data=work.temp;
          vbox &next_var&next_name / category = group boxwidth=0.50;
          run;
%end; %end;
%mend;
```

This macro is called by (1) specifying the list of the ids used in the call of the DIF macro and then (2) submitting the name, %explore(n), with n equal to the frequency of the series. This creates the data transformations on the men and women, black and total employment-population ratios.

```
%let name_list = EPT EPB;
%explore(12);
```

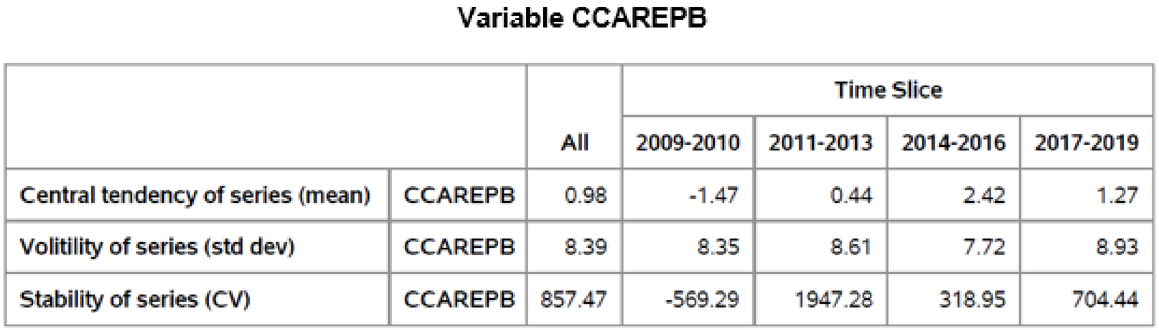
There are no parameters to pass as the EXPLORE macro includes complete DATA and PROC steps which will create a lot of graphs including the appropriate proc tabulate to give the appropriate statistics by predefined time-slice. This can be modified to create certain outcomes and expanded to run identification and stationary steps as discussed in the second part of this paper. An experienced time series analyst may benefit from the amount of and varied output which will quickly inform more sophisticated analysis.

Because of the amount of results, the entire production is written to a pdf file using the output delivery system (ODS). All of the production code is written between the following code:

```
ODS pdf file='d:\EPOP\timeseries_results.pdf';
ODS GRAPHICS on / ATTRPRIORITY=color noborder width=4.5in;
┬   <all production code>
ODS pdf close;
```

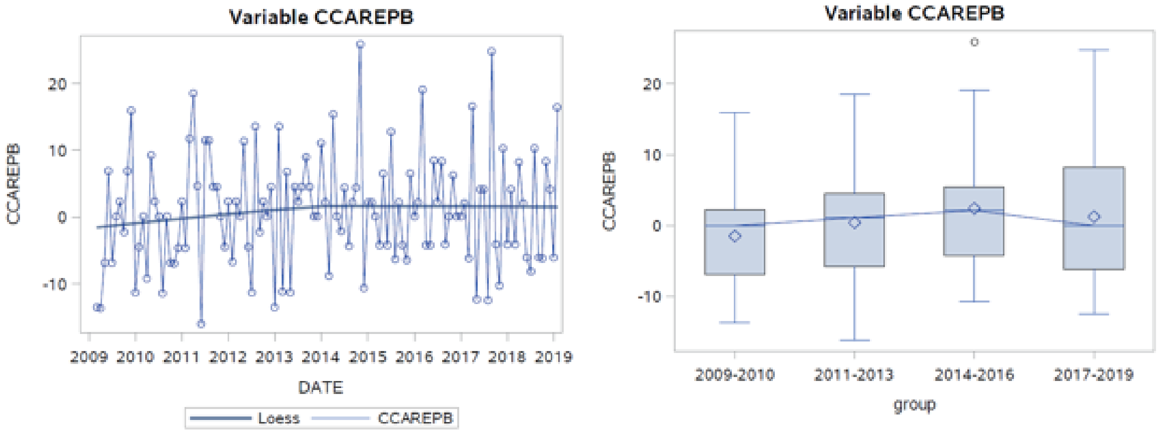
Explore produces 30 pages as specified here, but for illustration only the EPOPBLACK results for the continuously compounded annual rate of change is shown in Figure 7.

Figure 7: Sample output for each variable and transformation from Explore Macro



Variable CCAREPB

			Time Slice			
		All	2009-2010	2011-2013	2014-2016	2017-2019
Central tendency of series (mean)	CCAREPB	0.98	-1.47	0.44	2.42	1.27
Volitility of series (std dev)	CCAREPB	8.39	8.35	8.61	7.72	8.93
Stability of series (CV)	CCAREPB	857.47	-569.29	1947.28	318.95	704.44



As you review the many pages of output on your series, consider Silvia, et al. (2014) visualization questions while inspecting the data. They are:

1. Do you see any extreme values representing unusual or occasional events or possible mistakes in the data? Can you determine the difference of an unusual event from a mistake in the data?
2. Does the series have an explicit time trend? Does it appear linear or nonlinear? Is there a cyclical pattern that repeats? Do increases and decreases in value seem to occur at the same rate?
3. Is there evidence of one or more structural breaks indicating a major change in behavior? Do these happen at known instances such as related to the onset of a recession or its aftermath, or at a time of a historical event such as a disaster or other such event?

WHAT IS A RANDOM WALK AND WHY SHOULD I CARE?[9]

Many macroeconomic series are random walks and using them in analysis takes special care. An in-depth discussion is beyond this brief session, but it is important to know the special forms of a random walk and their implications. Nonstationary series contain a unit root and values do not fluctuate around a long-run mean and the variance is not constant over time. A stationary series will return to its long-run mean (mean reversion) and a series that is non stationary will not return to that mean. Further the variance may be dependent on time. You can see the lack of stationarity (non-zero mean and changing variance) in the descriptions of Table 1 because of the use of the time-slices.

[9] Maradiaga, et al. (2103) diagnose non-stationary properties of data and should be read if the reader is interested in doing a deeper dive into the types of time series structures encountered.

First, Equation 1 shows that a random walk takes the form of

1 $\boldsymbol{y_t = y_{t-1} + u_t}$

which says that the expected value of the y variable is equal to the immediate past value plus a random error term. The problem here can be seen by subtracting the RHS variable from each side yielding a new horrifying equation that any growth observed is purely random error as shown in Equation 2.

2 $\boldsymbol{\Delta y_t = y_t - y_{t-1} = u_t.}$

Since you cannot write an equation to predict random error, it stands to reason that you cannot predict current or forecast future changes in the variable of interest. So a test of the random walk hypothesis with no trend or drift is $H_0{:}\, \rho = 1\ vs.\, H_1\ |p| < 1$ in the model shown in Equation 3 where u_t is white noise (zero mean and not serially correlated). The Augmented Dickey-Fuller (ADF) and Phillips-Perone (PP) are tests of this null hypothesis.[10]

Second, random walks take three typical forms. Each allows a test of stationarity (no Random Walk) under different assumptions (zero mean, with no trend, as in Equation 3, a non-zero mean, with no trend, as in Equation 4) and a non-zero and a trend as in Equation 5).

3 $\mathbf{y_t = \rho y_{t-1} + \varepsilon_t}$ **Zero mean**

4 $\mathbf{y_t = \mu + \rho y_{t-1} + \varepsilon_t}$ **Single mean**

5 $\mathbf{y_t = \mu + \rho y_{t-1} + \delta t + \varepsilon_t}$ **Trend**

Third, if non-stationary series are used in correlation and regression this will lead to spurious results which we take up in the next section.

TESTING FOR UNIT ROOTS

While the tests may seem to be straight forward, tests for unit roots are sensitive to the lag length, the degree of autocorrelation, in the series. Our equations assumed no such autocorrelation. There are two procedures that test for unit root in SAS/ETS, PROC ARIMA and PROC AUTOREG.

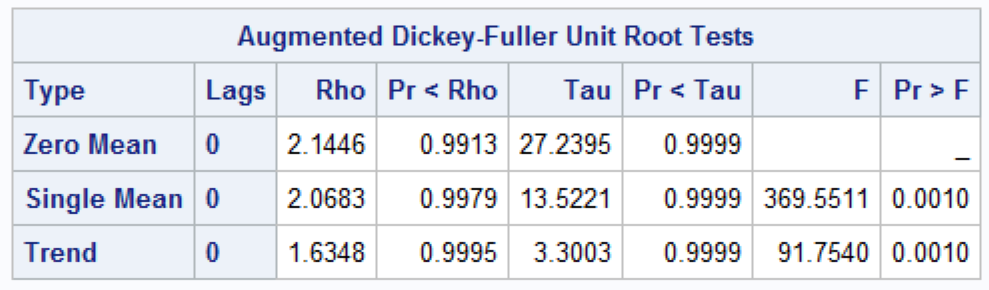
To run an ADF unit-root test in PROC AUTOREG we can use the code:

```
proc autoreg data=gdpm2;
    model m2=           /stationarity=(adf=0);
    model PCTD4CPI=     /stationarity=(adf=0);
    run;
```

where m2 is the money supply called M2 and has been aggregated to a quarterly frequency. CPI is the Consumer Price Index and also aggregated to a quarterly frequency and PCTD4CPI is the year over year percentage change in the CPI and is our measure of inflation. The separate results are in Table 3 and Table 4 show a clear non-stationary M2 and a stationary inflation measure in PCTD4CPI based on the Tau test and assuming there is no autoregressive terms (adf=0).

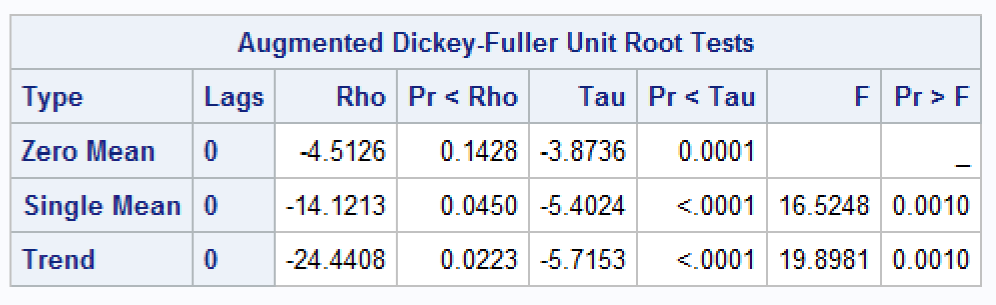
[10] For details see SAS/ETS. The ARIMA Procedure. Stationarity Tests accessed at https://documentation.sas.com/?docsetId=etsug&docsetTarget=etsug_arima_details07.htm&docsetVersion=15.1&locale=en and SAS/ETS. The AUTOREG Procedure. Testing for Stationarity accessed at https://documentation.sas.com/?docsetId=etsug&docsetTarget=etsug_autoreg_details27.htm&docsetVersion=15.1&locale=en and Dickey (2016).

Table 3: ADF Tests of a unit-root in the money supply, (M2)



Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	2.1446	0.9913	27.2395	0.9999		_
Single Mean	0	2.0683	0.9979	13.5221	0.9999	369.5511	0.0010
Trend	0	1.6348	0.9995	3.3003	0.9999	91.7540	0.0010

Table 4: ADF Tests of a unit-root in Inflation (PCTD4CPI)



Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-4.5126	0.1428	-3.8736	0.0001		_
Single Mean	0	-14.1213	0.0450	-5.4024	<.0001	16.5248	0.0010
Trend	0	-24.4408	0.0223	-5.7153	<.0001	19.8981	0.0010

EXPLORATION OF THE RELATIONSHIP BETWEEN TWO SERIES

Back in Figure 3 two series are shown, but one is really an arithmetic function of the others. What if we look at two series that we believe may have a causal relationship with each other? Spurious Correlation (2014) shows how a false correlation may be picked up using the same variables we use below, M2 and the federal debt (GFDEBTN). It shows three visualizations and we repeat the first two here using a slightly altered time dimension. The point, the strong relationship between M2 and DEBT (as in Equation 1) vanishes when the changes in M2 and changes in DEBT are compared (as in Equation 2).

Here is a near replication of Spurious Correlation (2014).[11]

The code below, uses the SASEFRED engine to access FRED and bring down a few macroeconomic data series that ought to have some economic relationship between them. The series are loaded from the beginning of 1981 and continues through the latest data.

```
options validvarname=any;
title 'Acquire M2 GDP and Federal Debt Data';
libname _all_ clear;
%let dir = d:\freddata;

libname fred sasefred "&dir"
    OUTXML=gdpm2
    XMLMAP="&dir\gdpm2.map"
    FREQ='q'
    START='1981-01-01'
    END='2019-06-01'
/*  Your 32-character alphanumeric API key goes here. */
    APIKEY='xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
/*  IDLIST is comma delimited with no spaces between the single quotes.  */
```

[11] The data in the FRED blog example are from 1966Q1 to 2014Q1. See Federal Reserve Bank of St. Louis (2014).

```
    IDLIST='gdp,m2,GFDEBTN,CPILFENS';
data work.gdpm2;
    set fred.gdpm2 ;
    %dif(gfdebtn,4,DBT);
    %dif(M2,4,M2);
    %dif(GDP,4,GDP);
    %dif(CPILFENS,4,CPI);
    run;
```

SPURIOUS CORRELATION & REGRESSION

When we regress two random walks we will find an R^2 close to 1.0. Normally, this would indicate that the series are extremely related to each other. An observer, may recognize this pattern and wonder not only if there is high correlation between them, but whether there is a causal relationship as well. Table 5 shows correlations, p-values and the R^2 from running a linear regression of one of the variables on the other.

Table 5: Correlations, p-values and R-square for Debt, M2 and GDP

		Levels		Change from a Year Ago		Percentage Change from a Year Ago	
		DEBT	**M2**	**DEBT**	**M2**	**DEBT**	**M2**
M2	Correlation	.993		.617		.257	
	p-value	<.0001		<.0001		.0015	
	R^2	.986		.381		.066	
GDP	Correlation	.963	.976	.097	.307	.059	-.047
	p-value	<.0001	<.0001	.2384	.0001	.4745	.5627
	R^2	.928	.953	.009	.095	.003	.002

The R^2, sometimes referred to a goodness of fit statistic, is the percentage of total variation in the LHS (dependent) variable that is explained by the RHS (independent) variable. We would run this regression only to make a point as there is much to consider before trusting these results. Our only interest here is what the R^2 is. As it turns out we do not have to run the regressions at all, because when there is only a single independent variable and a single dependent variable in the model, the R^2 is the square of the correlation coefficient.

We are told a spurious regression exists when the R^2 of the regression in levels is close to 1 and the R^2 of the regression in differences is low, even close, to zero. In Table 5 we see that M2, DEBT and GDP are all random walks.

What this looks like graphically is shown in Figure 8 and Figure 9. A strong pattern exists in levels, but practically vanishes when examined in changes. The code to create the graph in Figure 8 is:

```
Title 'Federal Public Debt and the money supply (M2)';
proc sgplot data=work.gdpm2;
    series x=date y=lev_dbt / curvelabel='Federal Debt'
           CURVELABELPOS=end lineattrs=(thickness=3px color=red);
    series x=date y=lev_m2  / curvelabel='Money supply'
           CURVELABELPOS=max y2axis lineattrs=(thickness=3px color=blue);
    format date year4. ;
    xaxis values=('1jan81'd to '1jul19'd by year);
    run;
```

The code to create the graph in Figure 9 makes use of both series and loess statement and is:

```
ODS GRAPHICS / ATTRPRIORITY=NONE;
Title 'Percent Change from a year ago: Federal Public Debt and the money Supply (M2)';
proc sgplot data=work.gdpm2;
    styleattrs  datacontrastcolors=(grey brown) datasymbols=(circle circleFilled );
    series x=date y=pctd4dbt;
    loess x=date y=pctd4dbt / curvelabel='LR Federal Debt' CURVELABELPOS=max
           lineattrs=(thickness=5px color=red pattern=1) legendlabel="Loess debt";
    series x=date y=pctd4M2;
    loess x=date y=pctd4M2 / curvelabel='LRMoney supply' CURVELABELPOS=start
           lineattrs=(thickness=5px color=blue pattern=1) legendlabel="Loess M2";
    format date year4.;
    xaxis values=('1jan81'd to '1jul19'd by year);
    run;
```

Figure 8: Debt and Money Supply track each other tightly

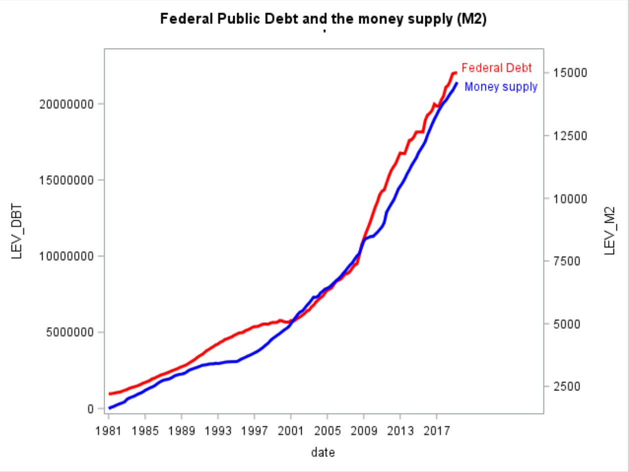
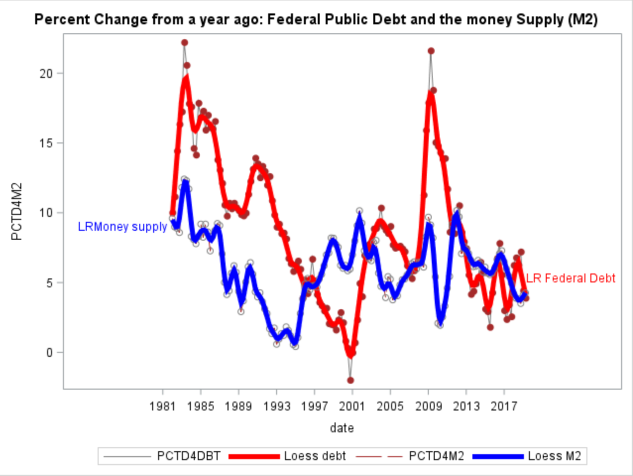


Figure 9: Differenced Debt and Money Supply do not track each other as tightly



CONCLUSION

Data from FRED seems mostly clean and quickly useful. All of the series are as reported from the original provider or owner of that data. The quality would seem to be pretty good. However, this paper shows exploratory methods to examine each series, identify outliers, and visually check for structural breaks or other data anomalies. A way to characterize the data by time-slice is suggested as well as extensive visualizations so that the analyst may fully understand the data generating process.

A macro of data transformations is shown and a second macro that allows for “tons” of pages of graphics and basic statistics. While not shown here, dozens to hundreds of visualizations can be read by an experienced eye very quickly and anomalies, perturbations and possible breaks can be seen just as fast. The experienced analyst can hone on quickly on series that need another look.

Finally, a brief look at relationships between economic variables start with a warning that nonstationary series can lead an analyst to ruin while suggesting a found relationship that is actually spurious. While the paper is a non-regression exploration, it seemed prudent to remind any explorer that there is much more to consider in order to establish good results. Moreover, all of this needs to be done before any inference about the economy or reasonable attempt at forecasting should be undertaken. In time series analysis, the hasty regression approach is almost always wrong.

REFERENCES

Bilenas, Jonas V. 2014. Scatter Plot smoothing using PROC LOESS and Restricted Cubic Splines, Paper 1503-2014. Accessed at https://support.sas.com/resources/papers/proceedings14/1503-2014.pdf.

Cunningham, Evan. 2018. "Great Recession, great recovery? Trends from the Current Population Survey," Monthly Labor Review, U.S. Bureau of Labor Statistics, April 2018, https://doi.org/10.21916/mlr.2018.10.

Dickey, David A. 2016. What’s the Difference? Paper 7080-2016, accessed at https://support.sas.com/resources/papers/proceedings16/7080-2016.pdf

Federal Reserve Bank of St. Louis. 2014. Spurious Correlation. The Fred Blog, July 28, 2014 accessed at https://fredblog.stlouisfed.org/2014/07/spurious-correlation/

Horstman, Joshua SESUG Paper 205-2018 Doing More with the SGPLOT Procedure Joshua M. Horstman, Nested Loop Consulting https://www.lexjansen.com/sesug/2018/SESUG2018_Paper-205_Final_PDF.pdf

McBride, Bill. 2018. Update: Scariest Jobs Chart Ever. Calculated Risk blog. February 2, 2018 accessed at https://www.calculatedriskblog.com/2018/02/update-scariest-jobs-chart-ever.html

Maradiaga, David, Aude Pujula, Hector Zapata. 2013. Exploring Time Series Data Properties in SAS®, Paper 456-2013, SAS Global Forum, accessed at https://support.sas.com/resources/papers/proceedings13/456-2013.pdf

Puryear, Cindy. 2015. Getting the macro language to perform a %DO loop over a list of values, SAS Learning Post blog. Jan 30, accessed at https://blogs.sas.com/content/sastraining/2015/01/30/sas-authors-tip-getting-the-macro-language-to-perform-a-do-loop-over-a-list-of-values/

SAS. The BOXPLOT Procedure. Accessed at https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.3&docsetId=statug&docsetTarget=statug_boxplot_details09.htm&locale=en

Silvia, J. E., Iqbal, A., Swankoski, K., Watt, S., & Bullard, S. (2014). *Economic and business forecasting: Analyzing and interpreting econometric results*. John Wiley & Sons. https://amzn.to/2RrK4qA

ACKNOWLEDGMENTS

I wish to thank the section chair, Carl Nord, for accepting my paper and for his patience in my completion of it. And my thanks to Jessica Chen, the Academic Chair, for her support as well. I am grateful to Kirk Paul Lafler and Josh Horstman for their encouragement to begin presenting at SAS conferences in general and MWSUG in particular.

RECOMMENDED READING

Dickey, David A. 2011. SAS/ETS® and the Nobel Prize , Paper 328-2011 accessed at https://support.sas.com/resources/papers/proceedings11/328-2011.pdf

Dickey, David A. 2005. Stationarity Issues in Time Series Models, Paper 192-30 accessed at https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/192-30.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Steven C. Myers
myers@uakron.edu
https://www.linkedin.com/in/stevencmyers/
https://econdatascience.com