Paper 1270-2017

Time Series Analysis and Forecasting in SAS® University Edition

Christopher Battiston, Women's College Hospital and Lucy D'Agostino McGowan, Vanderbilt University

ABSTRACT

Time series analysis and forecasting have always been popular as businesses realize the power and impact they can have. Getting students to learn effective and correct ways to build their models is key to having successful analyses as more graduates move into the business world. Using SAS® University Edition is a great way for students to learn analysis, and this talk focuses on the time series tasks. A brief introduction to time series is provided, as well as other important topics that are key to building strong models.

INTRODUCTION

Time series analysis is rapidly gaining in popularity in industries that have, up until now, used them sparingly or not at all. Students coming out of college or university who will be doing any sort of analytical work will, more than likely, be exposed to them, and potentially even asked to build some preliminary forecasts. Having a solid understanding of how to use SAS to perform these tasks will have students further ahead than the rest of their peers.

The aims of this paper will be to dig into key concepts for Time Series and Forecasting so that students can have a solid understanding of the topic; and exploration of the key tasks in SAS University Edition.

INTRODUCTION OF KEY CONCEPTS

There are numerous books written on the subjects of time series and forecasting; the ones that are most suited for students will be covered in the references section. To do justice to these topics would require far more time and writing than this paper will allow, however there are some basics that need to be focused in on and explained.

Time Series – Usually indicated as X_t, where X is the value of an event (temperature, surgeries, dollar value of groceries purchased) and t is the time. Note that *t* does not indicate the unit or length of time (minutes, years, etc) but simply indicates sequence. It is understood however all units of time in a dataset are the same (you can't have hourly and yearly data mixed together).

Time Series in Continuous Time –The data potentially could be viewed as a straight line in time, with values at every possible instance the event is observed. Examples include temperature, blood pressure and number of people in a mall.

Time Series in Discrete Time – These are measurements made at set points in time, whether as it's happening or historically. Some manipulation or aggregation of the data may be needed (for example if the data contains sporadic daily data and quarterly summaries are needed). It is essential that for accurate reporting, the time points (regardless of units) are equidistant apart – having data for Monday, Tuesday and Saturday won't give you a complete picture. These may be the only days available, but in order to use them in a time series analysis, the remaining days need to be filled in; SAS provides methods to do this, and they will be discussed later.

Seasonality – Although seemingly obvious, certain events happen during specific times of the year. Estimating and forecasting winter tire sales in Toronto, Canada will have fewer months than those for Anchorage, Alaska. This is seasonality, the effect that the time of the year has on the data; and there are special methods for handling this type of data. It should be noted that this comes up in rather unexpected places; for example, number of ambulatory day surgeries in a hospital – fewer people are able to travel when the roads are icy (along with increased incidences of cold/flu, travelling over Christmas, etc), and so the number of surgeries scheduled drops significantly. Taking this into account is essential for the accurate modelling of the data.

Date format – SAS Requires time series to be in a very specific format; the Time Series Preparation task, described below, does this

PRELIMINARY ASSUMPTIONS AND SUGGESTED STEPS IN PLANNING

SAS University Edition provides a number of pre-built Tasks that makes learning analysis easy. Having this out-of-the-box, "fill in the blank" functionality enables students to focus on concepts (such as which method to use for imputing their data) rather than worrying about the different options available through the syntax. As an added bonus, the University Edition environment builds the code while the options are selected, so the code can be reviewed and copied / pasted into the programming environment for further modification and testing.

Before we get started, some assumptions about the data that will be analysed and the planned analysis:

1) There is a clear plan of the requirements, either provided by the person requesting it, conversation with a professor/manager, or from some other review. These will include a well-articulated series of questions to be answered, clearly defined specifications such as how to handle outliers, types of output required, and an expected deadline.
2) A thorough preliminary data review has been done, including datatypes on the original file and on the import to SAS, detecting outliers / missing values, and a basic multi-variate review (for example, are there female subjects over 60 that answered "Yes" to "Currently pregnant", or are there temperatures in July in California that are unrealistically cold?). These may in fact be accurate, but they need to be noted and decided if they are going to be excluded.
3) Writing out the steps that will be taken, so that the analyses can be reproduced later if needed; these will need to be constantly reviewed and updated during the analytic process to ensure accuracy and completeness.

The book *Data Analysis Plans: A Blueprint for Success Using SAS* by Jablonski and Guagliardo is a great resource for students needing help in getting up and running with their reporting plans.

The other basic assumption for this paper is that the student has had some exposure to SAS University Edition. If the student is still new to the environment and looking for resources, Ron Cody's book *An Introduction to SAS University Edition* is the perfect place to start. Also, *Essential Statistics Using SAS University Edition* by Geoff Der and Brian Everett does a fantastic job covering important statistical concepts through the University Edition.

TIME SERIES TASKS IN SAS UNIVERSITY EDITION

There are three main tasks that will be covered – *Time Series Data Preparation, Time Series Exploration,* and *Modelling and Forecasting.* Each of these comes with a variety of options and decisions, all of which could have serious impacts on your analysis. It is suggested that these be part of the discussion with either the person requesting the analysis or someone who has experience in Time Series analysis.

The dataset used will be from the City of Toronto's Open Data, and is the full list of reported water main breaks in the city from 1990 to 2015. The dataset is available here and is ideal for this type of project because it's simple, has a huge amount of data (34,600 lines) and has missing data which can be used to demonstrate imputation / data expansion.

For those unfamiliar with Toronto weather, there are days during the winter that the temperature can go below -30C (-22F). During these times, the pipes carrying water can break because of the expanding / freezing of the metal, and these can break through the roads and cause major delays in transportation, possible safety risks, and costs the city huge amounts of money every year. Other times, the water mains may break because of construction, old pipes, or other reasons. Having a way to predict and analyse these events would save the city time, frustration and money.

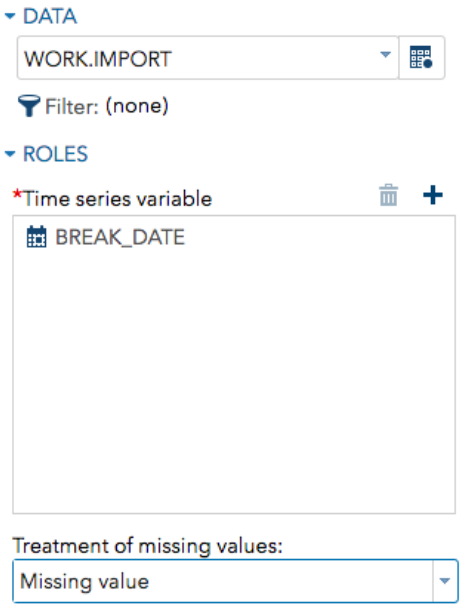
The IMPORT DATA task was used to bring the data into the SAS University Edition environment. Once in, the usual data checks are done and nothing that was of concern was raised. This means that the actual time series exploration can begin.

Time Series Data Preparation

Preparing the time series data is one of the most critical steps in the process, as a simple omission here can cause your analysis and forecast to appear to make sense, but in fact be completely incorrect. It is highly recommended that, as mentioned above, a step-by-step plan is laid out and is reviewed with either the person requesting the report (if they are knowledgeable about this type of analysis) or finding a professor or someone else with experience you can work with.

The first step is specifying the variables you want; this first example is keeping it simple, so nothing else other than the Time ID variable needs to be specified.

Figure 1. The Time Series Preparation Task



When we run the task, we find out that this won't necessarily help us – the data is set up so that each watermain break is a separate row, which is good for analyses like geospatial analysis, but we want a summary.

Figure 2 A Sample Output

Total rows: 34606 Total columns: 2

	TIME	BREAK_D...
1	1	01/01/1990
2	2	01/01/1990
3	3	01/01/1990
4	4	01/01/1990
5	5	01/01/1990
6	6	01/01/1990
7	7	01/01/1990
8	8	01/01/1990
9	9	01/01/1990
10	10	01/01/1990

In order to do this, a summary by day needs to be set up. Using SQL is a fast and efficient way to do this:

```
PROC SQL;
    Create table work.breaks as
    Select break_date, break_year, count(*) as Vol
    From work.import
    Group by break_date, break_year
    Order by break_date, break_year;
QUIT;
```

When the table opens, it becomes immediately apparent that something doesn't quite look right – we have a total of 9 cases where there is not break date available.

Figure 3. Sample Output of the BREAKS table

	BREAK_D...	BREAK_YEAR	Vol
1	.	2006	1
2	.	2014	8
3	01/01/1990	1990	29
4	01/02/1990	1990	16
5	01/03/1990	1990	19
6	01/04/1990	1990	12

Having a clear plan for handling this type of situation (which is a missing case, but it appears as though it was partially entered) is very important – having to go back and have that conversation later means potentially starting the whole analysis over again.

For the purposes of this paper, these cases have been dropped; this is done by a simple Where statement in the SQL query.

```
PROC SQL;
      Create table work.breaks as
      Select break_date, break_year, count(*) as Vol
      From work.import
      Where break_date<>.
      Group by break_date, break_year
      Order by break_date, break_year;
QUIT;
```

Now here is the resulting output, as expected:

Figure 4. Sample Output.

	BREAK_D...	BREAK_YEAR	Vol
1	01/01/1990	1990	29
2	01/02/1990	1990	16
3	01/03/1990	1990	19
4	01/04/1990	1990	12
5	01/05/1990	1990	12

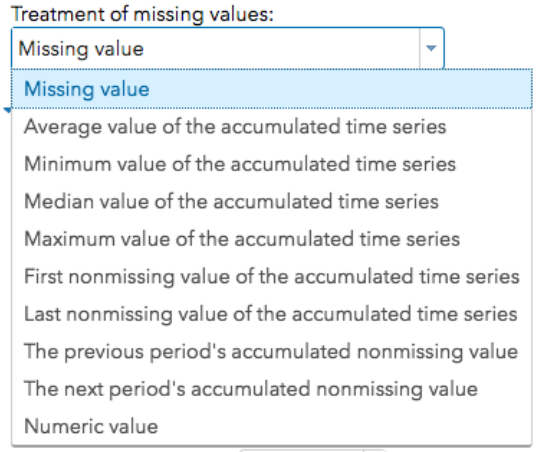
Scrolling through the data, it is no surprise that there are days without any breaks. These may have been days were the temperatures were not cold enough, weekends where no construction was happening, etc.

Figure 5. Sample Output.

42	02/11/1990	1990	6
43	02/12/1990	1990	7
44	02/14/1990	1990	5
45	02/15/1990	1990	3
46	02/16/1990	1990	5

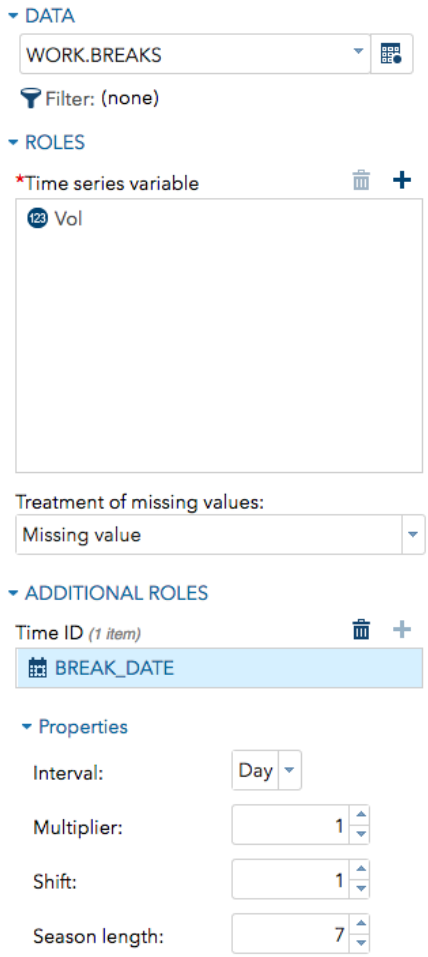
However, in order to build an accurate report, these days need to be taken into account so they can be built into the final model. Although there are a number of different ways to do this (PROC SQL, a DATA Step, etc.) SAS University Edition offers a very nice way to handle these, through the same task that was discussed above. Below are the options available for handling missing data.

Figure 6. List of Missing Value options.



As is obvious, there are a variety of different ways that SAS can handle these missing data. For more information, the books *Flexible Imputation of Missing Data* by Stef van Buuren and *Multiple Imputation of Missing Data Using SAS* by Berglund and Heeringa are highly recommended. For the purposes of the current example, the missing values will be left to missing, as it is important to know that no watermain breaks occurred on those days. Here is the task set up in preparation for this:

Figure 7. Task Preparation.



If you recall, February 13 in the original dataset was missing. Now when we scroll down, we see that there is a . for that date.

Figure 8. Sample Output.

41	10FEB1990	11
42	11FEB1990	6
43	12FEB1990	7
44	13FEB1990	.
45	14FEB1990	5

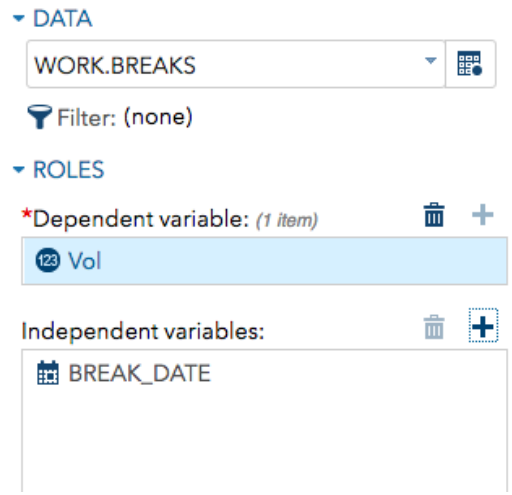
Now that our data is prepared, we'll move onto the next step of exploring our data.

Time Series Exploration

SAS University Edition provides an easy way to get a sense of what your data looks like using the Exploration task. This is different from a preliminary data exploration as there are statistical tests that can be performed specifically for time series data (Decomposition Analysis, Cross-correlation analysis, etc). These analyses are beyond the scope of this paper, and the books *SAS for Forecasting Time Series* by Brocklebank and Dickey, *Forecasting Examples for Business and Economics Using SAS* by SAS Institute and Practical Time Series Analysis Using SAS by Milhøj are all very good at explaining these concepts. The SAS/ETS User's Guide (available here) is also extremely useful.

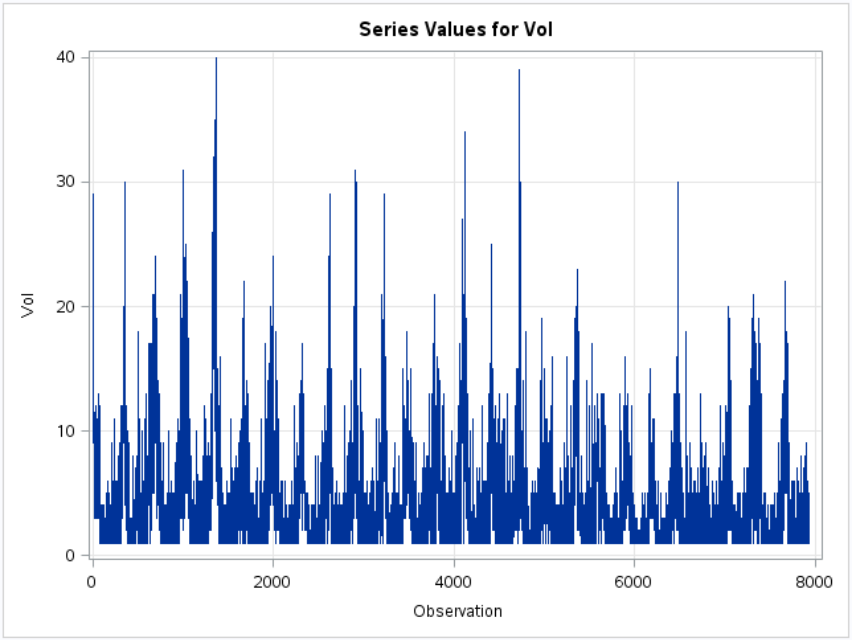
Using the task is, as with other tasks, relatively simple – the hard part is knowing where to put your variables, and how to interpret the output. Here is the task set up for our aggregated data.

Figure 9. Task Preparation.



When the task is run, I get a number of different graphs but the most interesting one is shown below, where it is very apparent there are cycles to the number of watermain breaks.

Figure 10. Output of the line graph.

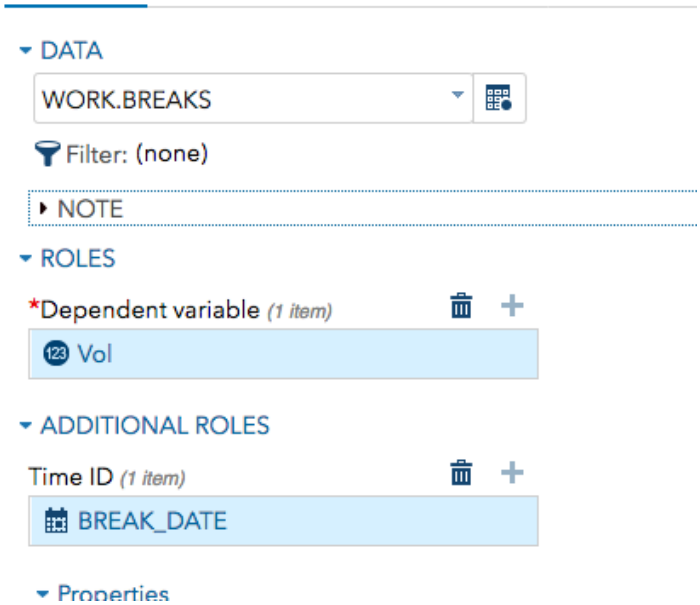


Modelling and Forecasting

The last step, and the most complicated, is obviously building the forecast. But having said that, SAS University Edition makes it very easy to build as it's simply a matter of selecting options from a series of drop-downs. New users must be forewarned that a "good" looking forecast may be wrong, as there may be an underlying confounding factor that was overlooked. As was stated above, having someone that is experienced that can be used as a resource is essential.

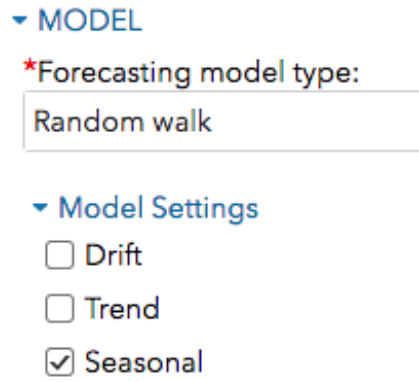
Using the BREAKS dataset, the first tab is filled in with the variables:

Figure 11. Step One of the Forecasting Task.



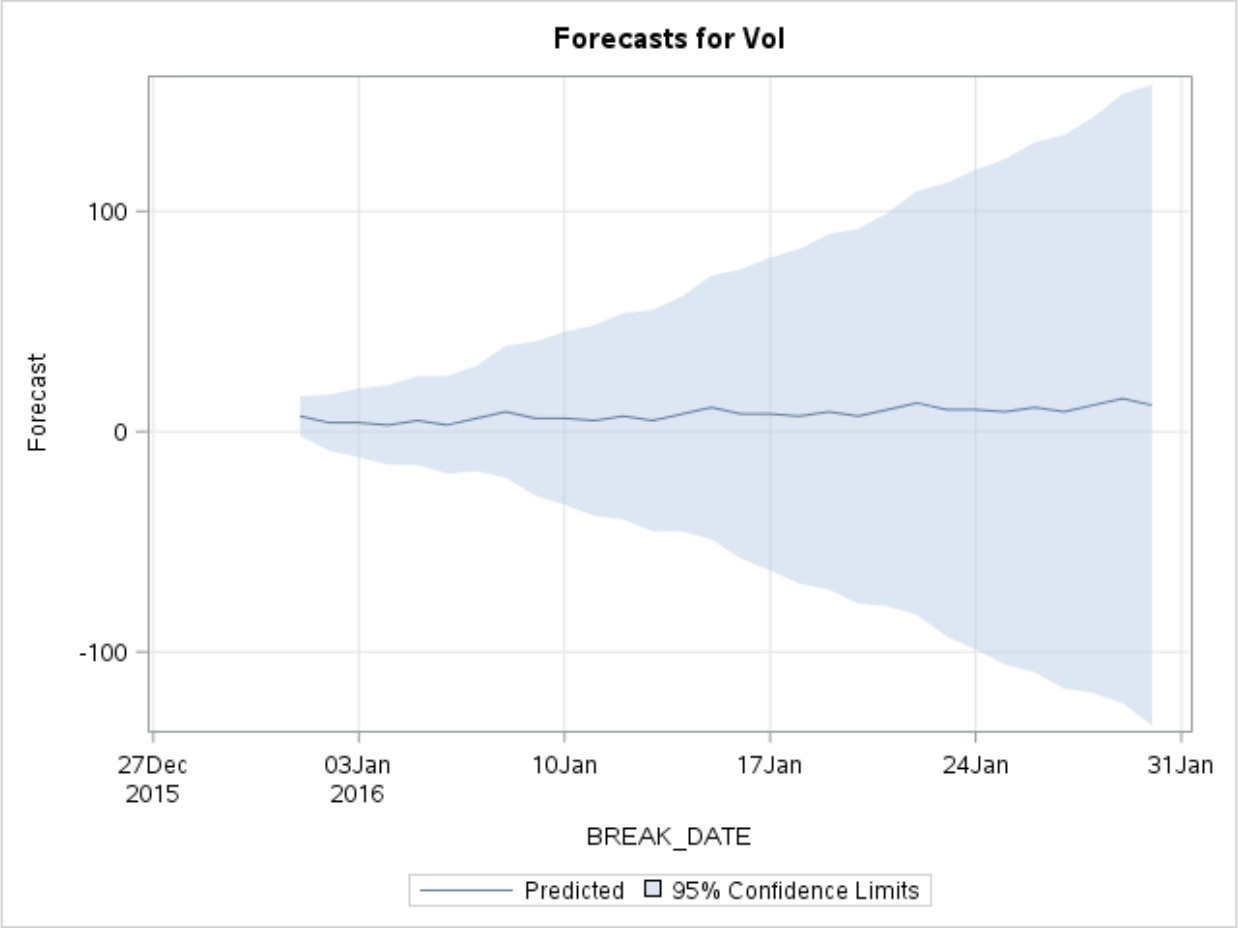
The next piece of the puzzle is to select the model; after much trial and error, the Random Walk with Seasonal effect was found to be the most reliable for the Watermain data.

Figure 12. Step Two of the Forecasting Task.



Finally, when the task is run, there are a number of different graphs and tables that are output; although important to understanding Time Series analysis and validation of the model, the piece of key interest is the forecast itself.

Figure 13. Output of the Forecasting Task.



Based on the data in the dataset, forecasting out to the end of January 2015 shows a couple of small spikes, but nothing too abnormal. The Confidence Interval, however, which was set at 95%, does go well beyond 100; this is to be expected given the wide variance in the data.

Although there is more detailed analyses that can be done, and a lot of information that needs to be studied and learnt before being comfortable with this powerful analytical tool, Time Series forecasting can give students a huge advantage in the industry they are interested in working in.

CONCLUSION

SAS University Edition offers students and others wanting to learn SAS a great tool, free of charge. Being able to use the platform, combined with the datasets available in the SASHELP library, allows users to move forward a great deal in their knowledge. However, exploring open datasets, as in this paper, is where students can really explore their newfound skills. Time Series Analysis and Forecasting is a highly complex skillset that takes years to develop; getting started as early as possible will give SAS Analysts a head start compared to their peers.

REFERENCES/RECOMMENDED READING

Allison, P. (2012). Survival Analysis Using SAS: A practical guide 2nd Edition. Published by SAS Publishing, Cary, NC.

Brocklebank, J. and Dickey, D. (2009). SAS for Forecasting Time Series, 2nd Edition. Published by SAS Publishing, Cary, NC.

Cody, R. (2001). *Longitudinal Data and SAS: A Programmer's Guide***. Published by SAS Publishing, Cary, NC.**

Jablonski, K. and Guagliardo, M. (2016). Data analysis plans: A Blueprint for Success using SAS. Published by SAS Publishing, Cary, NC.

Milhøj, A. (2013). Practical Time Series Analysis Using SAS. Published by SAS Publishing, Cary, NC.

SAS Institute. Forecasting Examples for Business and Economics Using SAS. (2007). Published by SAS Publishing, Cary, NC.

Stokes, M. and Staff, S.R. (2015) SAS/STAT ® 14.1: Methods for massive, missing, or multifaceted data. Available at: https://support.sas.com/resources/papers/proceedings15/SAS1940-2015.pdf (Accessed: 3 April 2016).

van Buuren, S. (2012). Flexible Imputation of Missing Data. Published by Chapman * Hall/CRC, Boca Raton, FL.

ACKNOWLEDGMENTS

As with any project, a SAS Paper takes a great deal of team work. From the SAS Global Forum executive, to my newfound partner in Analysis Lucy D'Agostino McGowan, this paper would not have been possible without their support and willingness to have Lucy present the paper in my stead. I also owe a great deal of gratitude to my friends who proofed this paper a number of times (I even managed to get one friend to switch from R ☺). Thank you all for your dedication.

Chris

CONTACT INFORMATION

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

Christopher Battiston
Darth.pathos@gmail.com