SESUG Paper 264-2018

Understanding Crime Pattern in United States by Time Series Analysis using SAS Tools

Soumya Ranjan Kar Choudhury, Oklahoma State University, Stillwater, OK; Agastya Komarraju, Sam's Club, Bentonville, AR

ABSTRACT

Crime, be it personal or public, property or violent has always been a social evil and a drawback to the inclusive development of the society. The United States, fortunately, has had a significant reduction in both property and violent crime over the past quarter century. Uniform Crime Reporting (UCR), an organization within the Federal Bureau of Investigation (FBI), has a data repository constituting different kinds of crimes for 20 years spanning 1995 through 2014 at a state level, collected by several city agencies across the country. Using this dataset, it is possible to forecast crime levels employing time series analysis in order to understand the severity of crimes and provide an estimate of upcoming crimes over the next 3 years (from 2015 to 2017). This forecast can then be compared with the actual crime numbers to realize the accuracy of the forecast. SAS® Enterprise guide was used for running the basic forecasting model by state and crime type. SAS® Enterprise Miner and SAS® 9.4 were then used to build exponential smoothing models by state and crime type. Visualization of the trends was done using SAS® Viya. Several SAS® tools were leveraged all through the project to showcase how a combination of SAS® products can help build better systems and analytics. This paper illustrates the simplicity of explaining time series forecasting in SAS® tools and since UCR doesn't forecast crimes on its website, this paper's idea can be used to provide an estimation for future crime rates in all states. The FBI can use this setup to understand the trend of various types of crimes in each state and thus necessitate preventive measures.

INTRODUCTION

The Uniform Crime Report (UCR) division is responsible for crime data compilation at a yearly level, which is then published by the Federal Bureau of Investigation (FBI). Over 18000 law enforcement agencies across the country come together to donate data to this cooperative effort of collating crime data. The data is then published annually in a "Hate Crime Statistics" publication by the FBI. The UCR website is pretty scalable and flexible with customized downloadable tables based on the dimensionality of the required data. The fields consist of the name of the state, population and the individual numbers for different types of crimes as defined by FBI.

This paper relies on the UCR repository data and acknowledges its efforts in providing customizable data for research purposes. This paper makes use of the statistical crime data from 1995 to 2014, cleaned and grouped for a simpler look, in order to make time series forecasted models of seven different types of crimes across all the 51 states. The data is available at a yearly level, state-wise and divided into the different categories of crime; Violent crimes (Murder, Rape, Assault and Robbery) and Property crimes (Larceny, Burglary and Motor Vehicle Theft). The data set is brought to a format where it has been grouped by state, year and by crime type.

First, SAS® 9.4 was used to merge all the individual years' datasets and then transpose them in order to bring them to a more SAS® readable format, just like a flat file. Secondly, this dataset was imported into SAS® Enterprise Miner™ where the time series models were built at a state and crime level using the Exponential Smoothing node. Exponential smoothing is an extrapolation procedure based on a weighted moving average in which the weights decrease exponentially as

data becomes older. Thirdly, the forecasted values and estimates of all the models were exported by writing a SAS® code and the exported excel files were imported into SAS® Viya in order to create visualizations of the forecasted crime numbers versus the actual numbers.

METHODOLOGY AND RESULTS

SAS® 9.4 was used to prepare the data set into the desired time series format. The original data was present at a state level in different sheets of an excel file; individual files having year in the rows and type of crime in the columns. The data was collated into one file using the APPEND procedure and then certain other procedures like PROC SORT and PROC TRANSPOSE were used in order to transpose the data so that it is in a time series format where the data looks like as if grouped by year, state and crime type and the dependent being number of crimes. The year data was converted into a readable date format for SAS® Enterprise Miner™ to spontaneously recognize the information. The code is as follows:

```
proc sort data = trans out = sorted;
        by State Year Population;

run;

proc transpose data = sorted
        out = TransposeOut(rename = (COL1 = CrimeNumber );
        var Tot_Violent_Crime Murder Rape Robbery Assault Tot_Property_Crime Burglary
        Larceny Motor_Theft;

        by State Year Population;

run;
```

After the data was harmonized to the prerequisite format, it was stored in a SAS® dataset, which was then imported into SAS® Enterprise Miner™ where a new project is created for forecasting the crime numbers up to 3 years. Since the current dataset had numbers available until 2014, the forecasting was done for the next three years ending in 2017. The layout of nodes in the SAS® Enterprise Miner is shown in Figure 1.

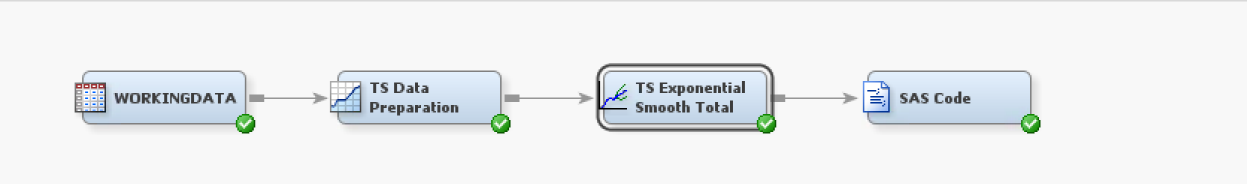

Figure 1: Process Flow for Time Series Forecasting

The manipulated and cleaned crime dataset used in this paper has one time ID, one target variable called CrimeNumber, two cross-sectional variables (Cross IDs: 51 states and 7 crime types) and one input variable called Population. The individual roles of the variables are specified in the initial stages of creating a data source in SAS® Enterprise Miner™. The TS Data Preparation node in this case is simply used to select the variables to be used for exponential smoothing, since our data has already been prepared using SAS® 9.4. The roles of each variable as specified can be seen in Figure 2.

Name	Role	Level	Report
CrimeNumber	**Target**	**Interval**	**No**
CrimeType	**Cross ID**	**Nominal**	**No**
Population	**Input**	**Interval**	**No**
State	**Cross ID**	**Nominal**	**No**
Year	**Time ID**	**Interval**	**No**

Figure 2: Variable Role Setting for Crime Time Series Data

The exponential smoothing node was used for the transformed dataset with the following settings as shown in Figure 3.

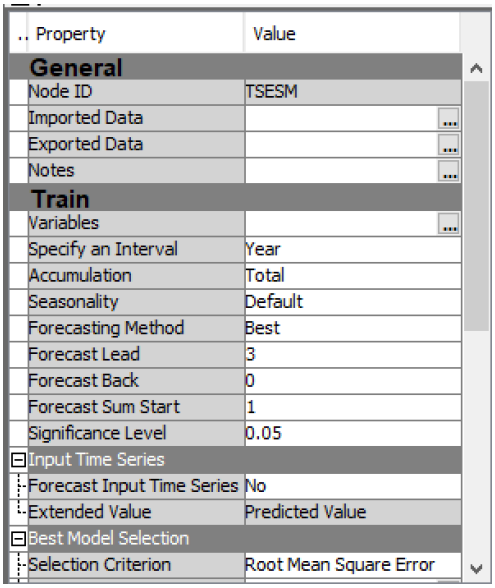

Figure 3: Settings for Running the Exponential Smoothing Node

The results after running the above node are divided by TSIDs. TSIDs are created when we add cross-sectional variables in the model. Since there are 51 states and 7 crime types in the dataset, 357 TSIDs will be created and the model will have forecasts for each and every TSID. Instead of using the results shown in SAS® Enterprise Miner™, the forecast values and estimates are exported through a SAS® code node onto an Excel file. The code used for export is as below:

```
libname TS 'M:\skarcho\TimeSeries\Workspaces\EMWS1';

proc export data = TS.tsesm_outforcst
        outfile = 'M:\skarcho\TimeSeries\Forecast.xlsx'
        dbms = xlsx replace ;
run;
```

In the next step, the Excel file is imported into SAS® Viya where SAS® Visual Analytics is used to draw various graphs and charts showing the forecasted values of crime numbers, grouped by state and crime type. SAS® Viya was preferred over SAS® Enterprise Miner™ because the latter shows visualizations only for the first 100 TSIDs and does not carry an option to view all of the TSIDs (In this case, 357 of them). Another advantage of SAS® Viya is that it has a plethora of customizable visualizations that can be built depending on our own requirements, including animations. Integrating SAS® Viya and SAS® Enterprise Miner™ can be a very efficient way to show accurate modeling results along with beautiful visualizations for our data.

After importing the forecast results, a visualization for the actual and predicted crime numbers was built at a year level keeping State as a filter, grouped by Crime type. For example, Figure 4 shows a forecast built on SAS® Viya for Alaska:

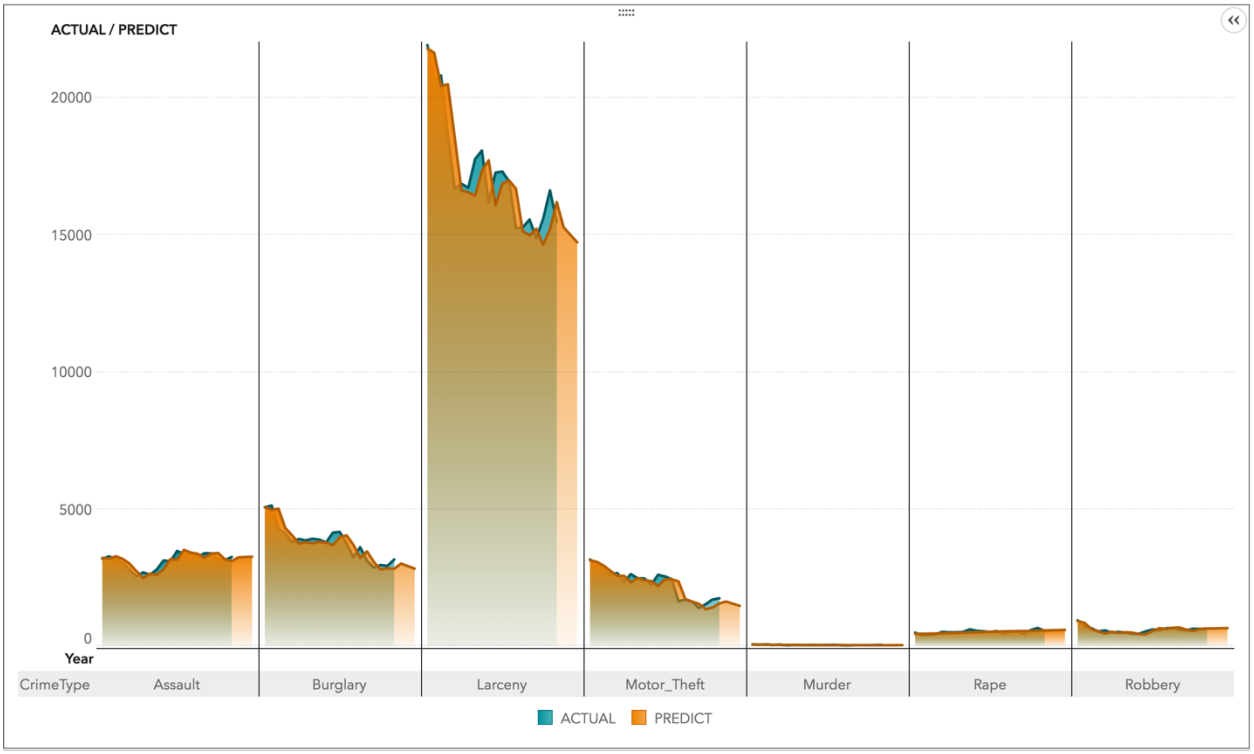

Figure 4: Forecast for Alaska in SAS® Viya

Another visualization that can be created is an animation of forecasted data for all states, filtered by crime type and year as the animation variable. An example of this is as shown in Figure 5 for Murder. Pressing the play button in the application plays a mix of graphs for each year starting 1995 until 2017.

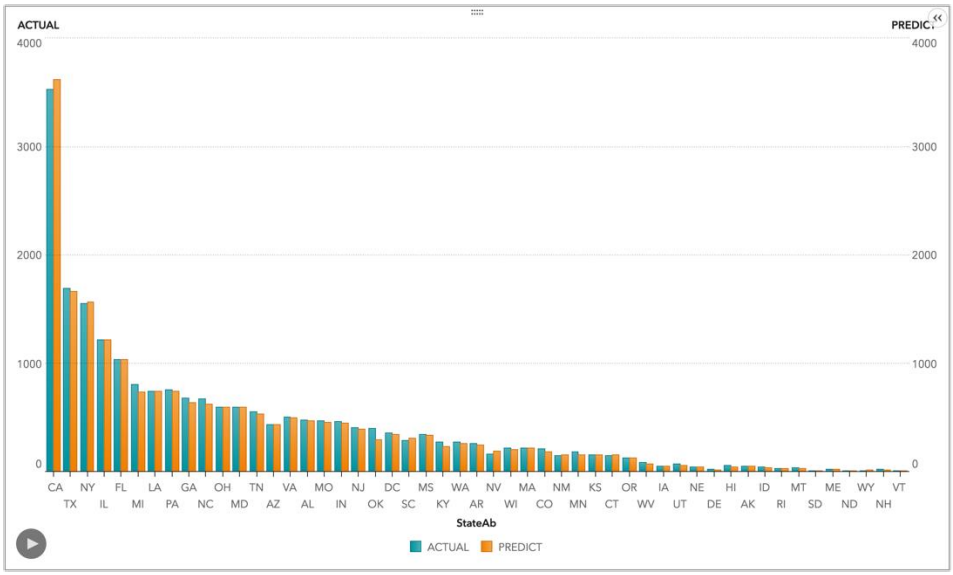

Figure 5: Animation of Murder for 51 States

CONCLUSION

This paper illustrates the usage of a combination of SAS® products in order to come up with cool visualizations that not only show the trend of crimes in the whole of United States but also forecast the crime numbers for the next 3 years. The visualizations show us many insights like:

1. New York has almost same population as Florida but in the recent times has had drastic decrease in the number of crimes. Florida outranks New York in terms of crimes and has had more or less a consistent number of crimes in the past 20 years
2. North Dakota has the least number of total crimes even if Wyoming has the lowest population along with District of Columbia.
3. District of Columbia has the highest proportion of murders.
4. The most committed crime is larceny throughout all the states.
5. Crimes have an overall decreasing trend. The trend and forecast can be seen in Figure 6.

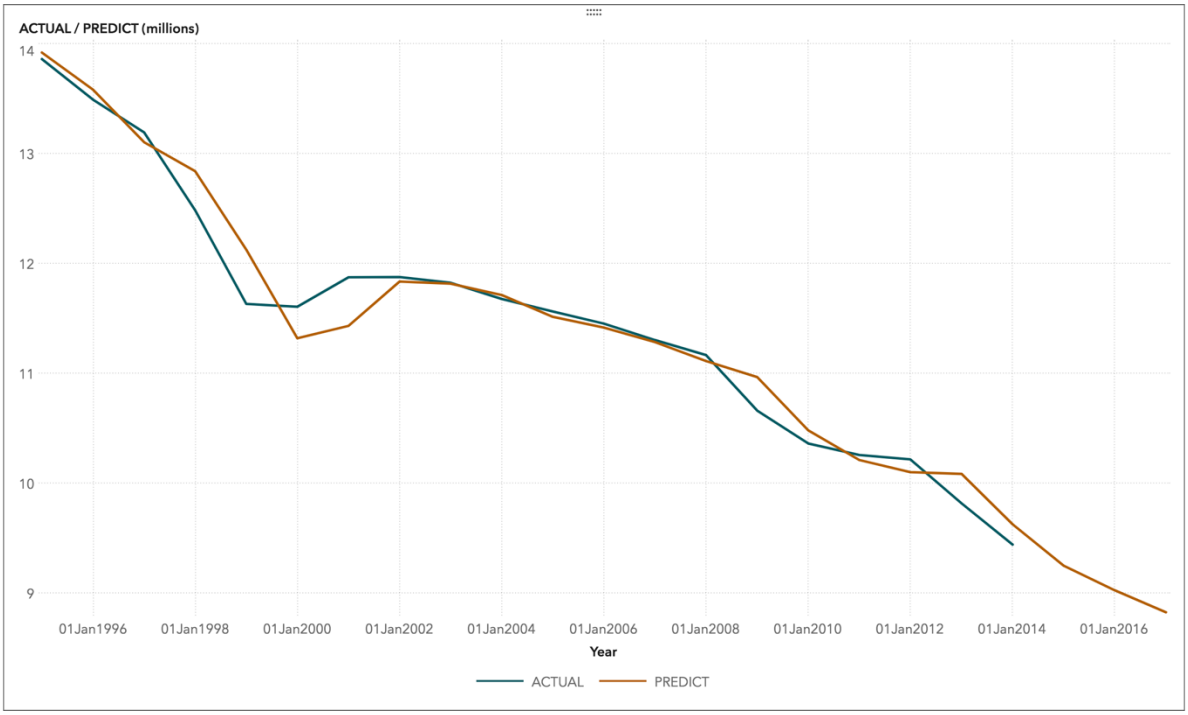

Figure 6: Forecast Chart of all the States and Total Crimes

Future work on this data can encompass auto-regressive time series models with exogenous variables like population, income, age, employment rate etc. where the effect can be quantified into getting a better forecast result for each state. The current dataset can be packaged into one macro which when run, can create these visualizations on SAS® Viya. The capabilities of SAS® Viya are limitless considering its superb user interface and the ability to create numerous statistical models while at the same time visualizing the results in a scalable manner for the end user to draw insights. With the amassed enterprise data available in a time series format, the ability to assimilate this data into scalable analysis workflows with visualizations can help data mongers to easily build valuable models.

REFERENCES

[1] Selukar, Rajesh. 2009. “Structural Analysis of Time Series Using the SAS/ETS® UCM Procedure.” *Proceedings of the SAS Global 2009 Conference*, 306-2009. Cary, North Carolina. Available at http://support.sas.com/resources/papers/proceedings09/306-2009.pdf

[2] Battiston, Christopher and McGowan, Lucy. 2017. “Time Series Analysis and Forecasting in SAS® University Edition” *Proceedings of the SAS Global 2017 Conference*, 1270-2017. Vanderbilt, Tennessee. Available at http://support.sas.com/resources/papers/proceedings17/1270-2017.pdf

[3] Chen, Peng & Yuan, Hongyong & Shu, Xueming. (2008). Forecasting crime using the ARIMA model. Proceedings - 5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008. 5. 627 - 630. 10.1109/FSKD.2008.222.

[4] U.S. Department of Justice, FBI. (2017): https://ucr.fbi.gov/ucr-publications

ACKNOWLEDGMENTS

We thank Dr. Goutam Chakraborty, SAS® Professor of Marketing Analytics and Dr.Miriam McGaugh, Clinical Professor at Oklahoma State University for their guidance and support throughout the research.

CONTACT INFORMATION

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

1. Soumya Ranjan Kar Choudhury, Oklahoma State University, Stillwater, OK
   Email: skarcho@okstate.edu
2. Agastya Komarraju, Associate Director, Sam’s Club, Bentonville, AR
   Email: Agastya.Komarraju@samsclub.com