Paper SAS2321-2018

Analytics of Things: New Analytical Models for Creating Business Value from IoT Data

Ryan Gillespie, SAS Institute Inc., Cary, NC

ABSTRACT

The number of devices and equipment generating sensor data is rapidly increasing. To intelligently handle this data and create tangible business value requires new analytic techniques and new ways to apply them. This paper covers new SAS® models related to the Analytics of Things. The paper focuses on actions and procedures tailored to Internet of Things (IoT) and includes use cases for predictive maintenance, asset degradation, anomaly detection, and signal processing. We highlight what new models are available in SAS, their related use cases, and how they can be deployed in standard or real-time scenarios for IoT solutions from manufacturing to health care.

INTRODUCTION

The data generated by sensors, equipment, and devices is growing rapidly and requires new methods that will provide business value. These methods must be able to handle new data challenges and offer flexibility to be deployed into streaming environments. The paper focuses on some of the challenges of dealing with this data as well as the use cases and the methods we've used to help analyze data and provide insights. The paper examines use cases around building sensor data, solar farm data, and aircraft engine data. Because this paper is covering a variety of use cases and algorithms, the goal of the paper is to highlight what is available for the user as well as the strengths of each algorithm and how it might be used. You will get a good overview for some of the new methods available, how these methods can be applied, and where you can find additional information for further research.

CHALLENGES WITH INTERNET OF THINGS DATA

Data generated from IoT devices and sensors presents new challenges that must be addressed by the tools and models being used to analyze the data. The following challenges are seen in dealing with IoT data:

- Data contains lots of variables and is generated at high frequency. Contrary to many historical data sets, the data generated by sensors is often wide and occurs at a high frequency. It is not uncommon for sensors or combinations of sensors associated with a process to produce data containing several thousand variables and occurring in millions of events per second.
- Data is generally unexplored. A lot of the devices generating data are producing new types of data that have not been previously analyzed. As such, questions remain on what is possible with the data being generated as well as questions surrounding what use cases are valid for that type of data. Since it is a relatively new field of analytics, the range of what is possible is still in the early stages.
- Sensor data can be noisy and contain missing values. Due to anomalous readings, potential errors can occur with a sensor or loss of data during transmission. As a result, the generated data often contains noise and missing values. Choosing how to handle, flag, or impute these values within a data stream requires decisions on what is valuable and how to interpret what is happening when these types of values occur.
- Not all data might need to be sent back from the edge. Because of the large quantities of data generated by sensors or by a process containing many sensors, it might be desirable to filter the data being streamed at the edge so that transmission and storage fees are reduced. In certain scenarios, the rate at which the data is being generated might be substantially more than what is

required to take action and drive business decisions. Determining how to intelligently filter this data at the edge of the network could prevent the movement of unnecessary data while still retaining all the desired events.

- Not all data might be valuable to store. Similar to determining how much data to transmit, the amount of data that you choose to store is also a concern surrounding many IoT applications. If the amount of data generated represents a storage cost concern and the rate at which it is generated is greater than what is required to drive value, then filtering the data before storing is also a consideration for the use case.
- Data sets can be heavily imbalanced. The nature of the data being generated by sensors might also represent a challenge for building classification models as it can be heavily imbalanced. For asset degradation, predictive maintenance, and fraud applications, the amount of normal events might be substantially more than the amount of failure events. Trying to build classification models with this type of data can pose challenges, so new methods might be more appropriate for modeling IoT data for these use cases.

USE CASES

The following use cases are discussed in this paper:

- Determining Building Power Issues with Moving Windows Principal Component Analysis, which uses the sensor readings from similar parking lot lights to determine when one parking lot light is behaving abnormally.
- Analyzing Solar Farm Data with Robust Principal Component Analysis, which uses Robust Principal Component Analysis for data reduction and anomaly detection while maintaining regression model accuracy.
- Aircraft Engine Degradation using Stability Monitoring and Support Vector Data Description. This use case uses multivariate sensor data to monitor asset degradation and optimize maintenance schedules using two separate techniques.

DETERMINING BUILDING POWER ISSUES WITH MOVING WINDOWS PRINCIPAL COMPONENT ANALYSIS

USE CASE

Buildings on the SAS campus contain sensors that provide a variety of information about the systems and processes of the building. The parking lot contains sensors detailing the power consumption of the lights within the lot. For this use case, we attempt to identify abnormal behavior in the parking lot lights. To do this, we’re using a method called Moving Windows Principal Component Analysis. This method will help us identify an abnormal operating condition for one light based on the usage levels of the surrounding parking lot lights.

METHOD

Moving Windows Principal Component Analysis (MWPCA) performs a series of principal component analyses over a sliding window consisting of a specified number of observations (SAS Institute Inc. 2017). This analysis is ideally suited to measure degradation in systems that contain many correlated measures where the correlation is expected to hold over time. The method returns the normalized first principal component of the data being analyzed. As such, if each variable represents a sensor reading for a different asset (such as a wind turbine), the results of the analysis can be helpful in indicating when one asset is beginning to behave differently than the other assets (when all assets are expected to behave in a correlated or similar manner).

STRENGTHS OF THE METHOD

The MWPCA method has several advantages, and MWPCA can be particularly helpful in monitoring system processes and identifying degradation in environments with similar operating conditions.

- Situations where the data exhibits seasonal or cyclical behavior. Returning to the example of wind turbines, trying to predict the output of a particular turbine at a certain point of the day can be difficult as the weather conditions at that time might not be the same as they were in the previous hour, day, month, or year. However, it is highly likely that the wind turbines in the surrounding area are experiencing the same weather conditions as the turbine being evaluated. As such, we can use the expected correlated nature of the system to help indicate if one asset's operating condition is starting to deviate from the others. We do not have to rely on rules or limits that can be difficult to set for systems with seasonality or unpredictability.
- Situations where training data is limited or unhelpful. In other methods of monitoring asset degradation, historical data is used to determine limits or to build a model that helps assess whether the asset is performing as expected. However, with MWPCA, no historical data is needed while assessing system performance. The method is evaluating the performance of only one asset (or variable) versus the rest over each sliding window. In addition to the seasonality issues discussed above, this can also be helpful in situations where there might be limited data on previous degradations or situations where certain types of degradations have not been recorded or have yet to occur.

RESULTS

Returning to our use case on identifying abnormal behavior with parking lot light sensors, we will examine the sensor data and evaluate how MWPCA identifies certain types of behavior based on the correlated nature of the data.

In Figure 1: Building U Parking Lot Lights Energy Consumption, we see the energy consumption metrics for six different parking lot light sensors. As can be seen by the values for each sensor, there exists a high degree of correlation between lights, and each light is relatively cyclical in its behavior. However, it should be noted that the energy consumption for each light is not consistent in terms of the maximum use during the high points of the cycle. So trying to use rules or other statistical process control techniques to evaluate when a sensor might be behaving abnormally can be a challenge as it requires the individual assessment of each sensor.

For our evaluation, we will focus on the red sensor in the panel that is second to the top in the figure. This sensor shows generally consistent behavior with the exception of a downward spike prior to the 2000th measurement sequence and an upward spike just beyond the 4000th measurement sequence. With MWPCA, we look to identify the first spike as it differs from its counterparts. While the second spike might also represent an issue, it appears to be a system-wide issue. In this case, since the assets are still behaving in a correlated manner, we will not expect it to be seen within the output of our approach.

Figure 2: Building U Parking Lot Lights – Moving Windows PCA shows the normalized first principal component output from our MWPCA approach. As can be seen in the figure, we see the deviation of the red sensor data prior to the 2000th sequence, which indicates that the sensor in question has deviated in behavior from the other sensors it is being evaluated against. Similarly, after the 4000th sequence, we do not see any deviation of the red sensor in the figure, as expected. In this case, MWPCA has managed to correctly identify the areas in the data where the sensor has deviated in behavior from the others. In a business use case, this information could then be used to identify potential problems with the sensor or asset in question. It should be noted that while the discussion is focused on the red sensor, the method is evaluating all sensors simultaneously. So, for example, if the green or orange sensor had also deviated, these issues would have also been identified within the same analysis.

One final point to note is that while the red sensor did not show any deviation after the 4000th sequence in Figure 2, there were slight deviations for both the blue and green sensors. If we return to the data for these sensors in Figure 1, we can see that each of them appears to have a step change in their energy consumption at this time period.

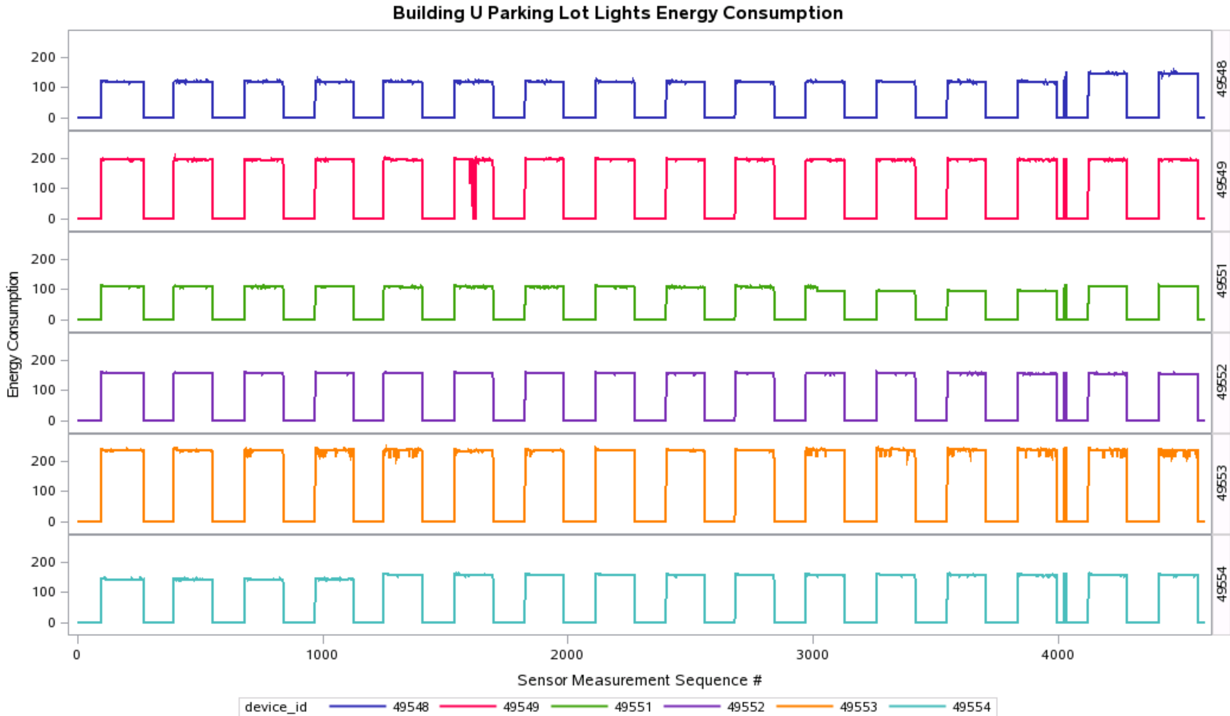

Figure 1. Building U Parking Lot Lights Energy Consumption

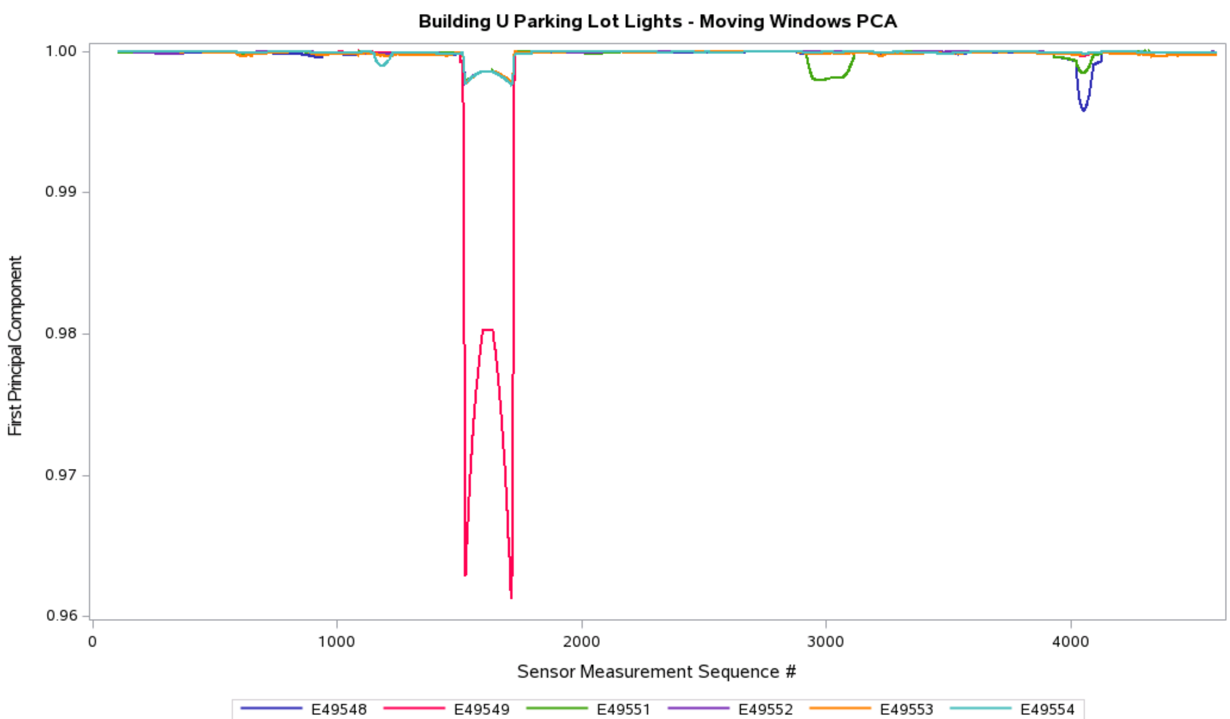

Figure 2. Building U Parking Lot Lights – Moving Windows PCA

ANALYZING SOLAR FARM DATA WITH ROBUST PRINCIPAL COMPONENT ANALYSIS

USE CASE

The next use case examines the benefits of analyzing data from solar farm panels using a technique called Robust Principal Component Analysis (RPCA). There are three areas of the analysis. The first area discusses the data reduction possibilities associated with RPCA relative to a standard principal component analysis (PCA). The second area evaluates the model validity of a linear regression model built on the data with RPCA. The final area discusses how RPCA can also be used to identify anomalies within a data set.

METHOD

Robust Principal Component Analysis decomposes an input matrix into a summation of a low-rank matrix and a sparse matrix, with the low rank matrix then being used with PCA or singular value decomposition (SVD) (SAS Institute Inc. 2017). The robustness of the method comes from its ability to handle anomalies and remove them prior to applying the PCA or SVD. As a side benefit of removing these anomalies, they are stored within the sparse matrix and can be used to examine potentially abnormal behavior within the data set. While we are using the method in this case to analyze solar farm data, it can also be used for other use cases such as image processing, ranking, and matrix completion.

STRENGTHS OF THE METHOD

RPCA offers several benefits to users depending on the use case that is being analyzed.

- The ability to decrease data size, transmission, and storage requirements. In situations where data volumes might represent an issue and methods are being evaluated to decrease the amount of data being transmitted and stored, RPCA is an option for dimensionality reduction. This can be particularly helpful with IoT sensor data that contains thousands of variables with many of the variables being highly correlated.
- The sparse matrix generated from the process can be used to identify anomalies within the data. This can be used in a variety of situations from identifying a system process event to separating the foreground from the background in a video stream or series of images. The user also has the ability to adjust the hyper-parameters of the method to provide a variety of levels of sparsity within the matrix. This can help to reduce false positives or to maximize coverage of anomaly detection depending on the use case.

RESULTS

The first area of focus is on the dimensionality reduction of the data with RPCA relative to the original data set and relative to standard PCA. Figure 3: Data Reduction with Robust Principal Component Analysis shows the relative amount of data required for both PCA and RPCA to explain 95% of the variance of the original solar farm data. As seen in the figure, standard PCA requires 26 components to explain 95% of the variance in the data set while RPCA requires only 13 components to explain the same amount of variance. RPCA can be useful as an option in situations where data reduction is necessary for transportation or storage requirements. RPCA also provides an extra lift over traditional PCA in terms of the number of components required to capture nearly the entirety of the variance in the data.

The components from each of the traditional PCA and RPCA transformations were then used as inputs in a linear regression model to predict the energy output from the panels. The purpose of the model construction was to evaluate whether the decreased amount of data generated from the RPCA transformation would have any substantial negative impact on the model's predictive ability.

Figure 4: Residuals from Linear Regression Model shows the output of the residuals from the linear regression models built with PCA in red and with RPCA in blue. Both represent the residuals from testing the model on holdout data.

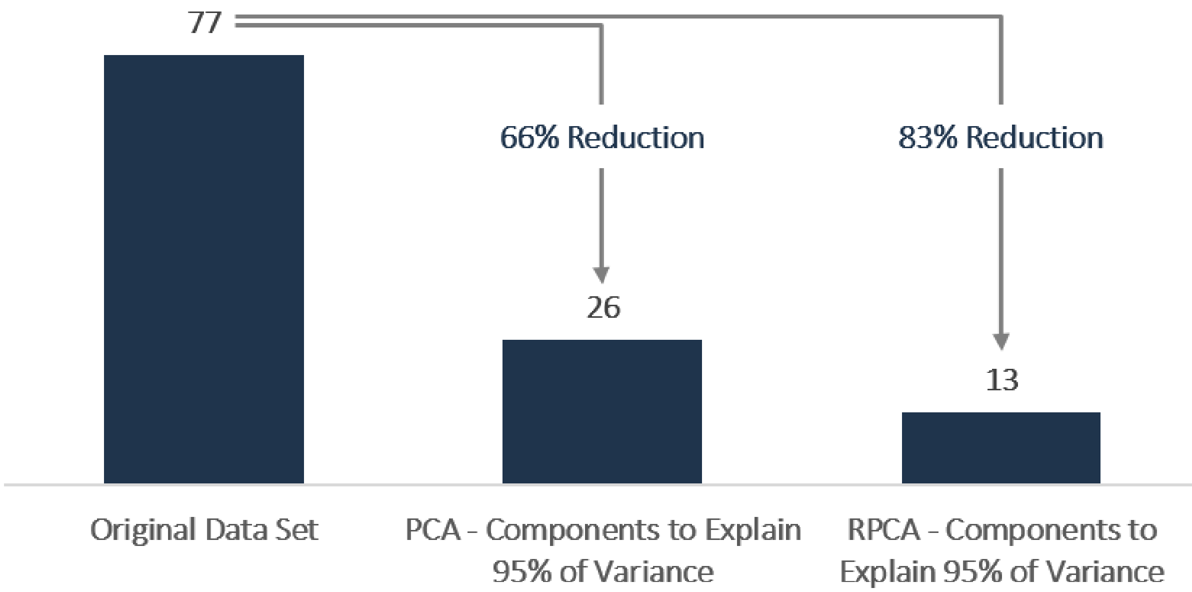

Figure 3. Data Reduction with Robust Principal Component Analysis

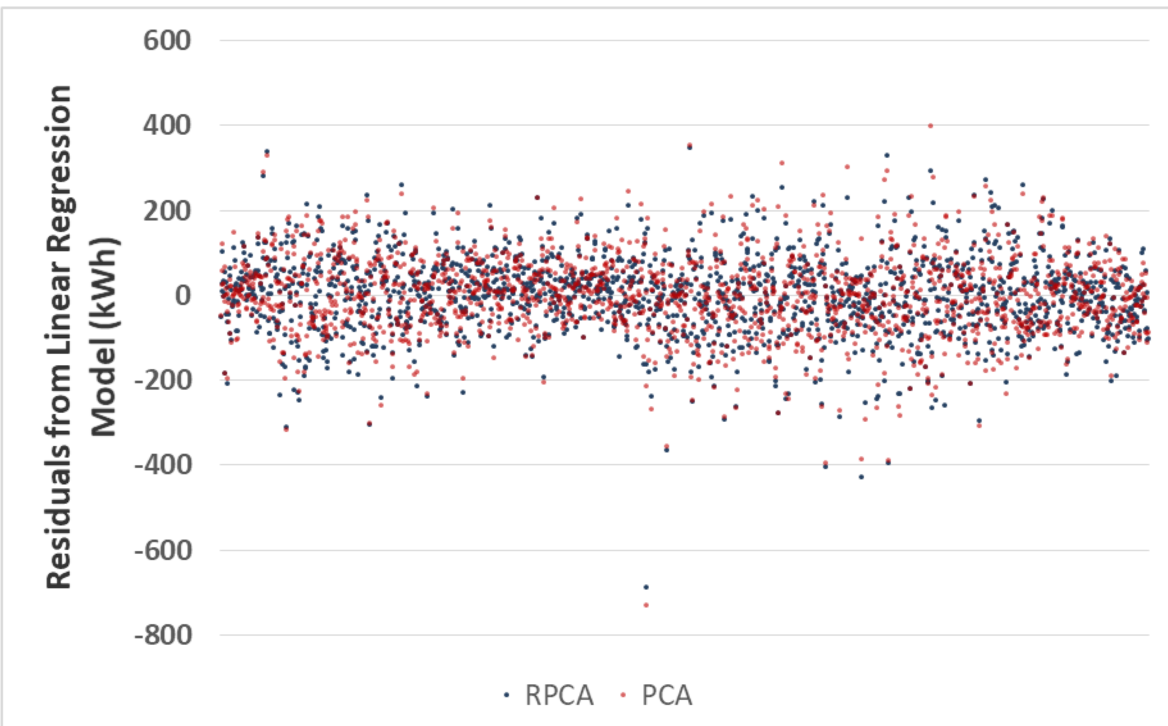

Figure 4. Residuals from Linear Regression Model

Figure 4 shows a similar spread of residuals for each transformation, and this result matches the holdout root mean square error (RMSE) for each model as well. For the RPCA model, the RMSE was 92.17 kWh, and for the PCA model, the RMSE was 93.21 kWh. While the RMSE associated with the RPCA model is slightly less, both models are fairly equivalent in terms of accuracy. This helps illustrate that, for this analysis, while the amount of data generated by the RPCA transformation is approximately 50% less than the data generated by the PCA transformation, there is no reduction in predictive ability for the linear regression model built with each data set.

The sparse matrix from the RPCA transformation was also used to identify observations and time frames that might be good to further evaluate for anomalous behavior in the solar panels. The extreme values within this matrix can be used to identify observations that might represent the anomalies. For example, see the kWh Delivered data in Figure 5: Sparse Matrix Data for kWh Delivered. While all observations in the figure might represent something of interest (since observations with a value of zero have been removed), several observations represent large deviations from what might be expected. These two observations are located just after April 6th, 2015 and just prior to December 12th, 2015.

One further point of note is that while the figure represents a relatively long time scale, the analysis could be run in a shorter batch periodicity to try to capture anomalies as they occur on a more immediate level.

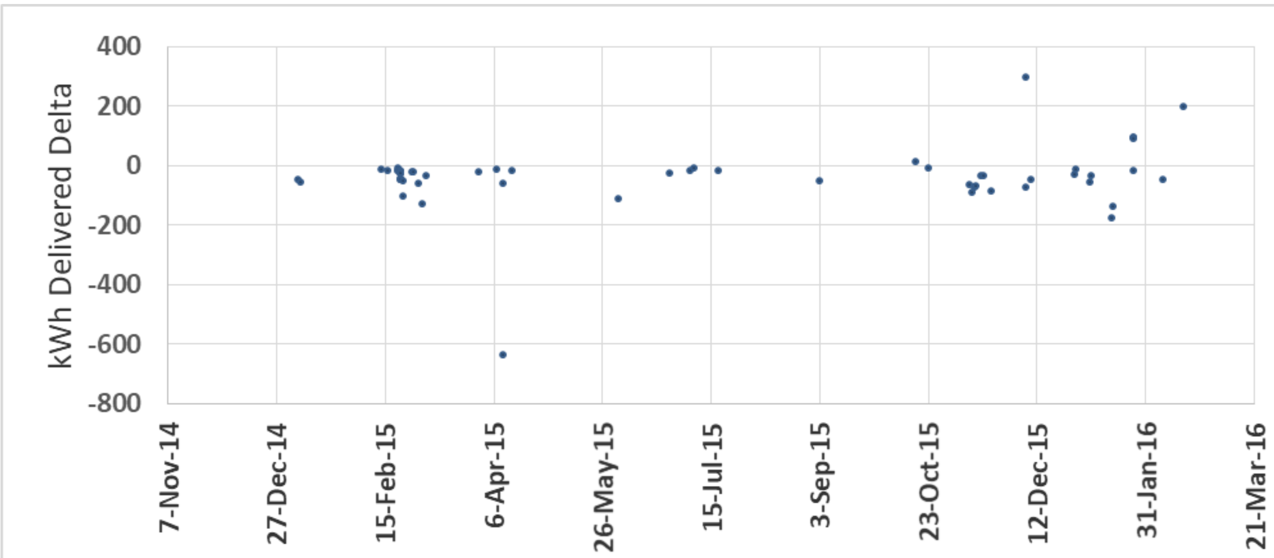

Figure 5. Sparse Matrix Data for kWh Delivered

AIRCRAFT ENGINE DEGREDATION USING STABILITY MONITORING AND SUPPORT VECTOR DATA DESCRIPTION

USE CASE

This next use case focuses on how sensor data generated from aircraft engines can be used to monitor asset degradation and aid with predictive maintenance efforts. If the multivariate sensor data can be used to determine when an asset is starting to break down, it can help in optimizing the time it should be scheduled for maintenance and also help to reduce the number of breakdown events that lead to costly repairs, lost productivity, and potential injury. We will use the sensor data and two separate methods, Stability Monitoring and Support Vector Data Description, to show how engine degradation can be identified for a turbofan engine data set produced by NASA for the 2008 conference on Prognostics and Health Management (Saxena and Goebel 2008).

METHOD – STABILITY MONITORING

Stability monitoring uses statistical monitoring methods to help identify anomalies (SAS Institute Inc. 2017). Here is how this approach works:

- Choose a target variable (sensor) whose behavior should be explained by other sensor variables within the data set.

- Specify and calibrate a variety of statistical models.
- Select the best performing model based on holdout analysis.
- Forecast target sensor values for new input values based on the calibrated model.
- Compare the forecasted value versus the actual value with a rule definition to indicate an anomaly dependent on the business context. (For example, the actual value is outside of the 95% confidence band.)

STRENGTHS OF THE METHOD – STABILITY MONITORING

Here are some advantages of using stability monitoring:

- Multiple models can be evaluated and specified for calibration.
- The method can handle multivariate data and automatically performs variable selection for each model.
- The method uses holdout analysis to evaluate the selected model types and perform the calibration.

RESULTS – STABILITY MONITORING

The aircraft engine used in this example had 240 cycles associated with it. The first 80 cycles of the engine were used for model training with the next 10 cycles used to perform the holdout analysis. The model was then further evaluated on the remaining cycles.

The target variable sensor chosen was the ratio of fuel flow to static pressure at the high-pressure compressor outlet. This variable was chosen as the target with the assumption that as equipment performance begins to deteriorate, it will be reflected in changes in the fuel consumption.

Figure 6: Stability Monitoring Results shows the output from the method on the turbofan data. Each graph represents a different time frame with blue circles indicating actual values, red circles indicating anomalies, and the shaded zone indicating the 95% prediction limits.

The top left quadrant of the figure represents the holdout zone that was used to calibrate the model. As can be seen in the figure, all of the actual values are within the predicted limits for this area, so no anomalies are present. Looking at the second time frame in the top right of the figure, we begin to see some anomalies represented by the indication of the red circles. These anomalies become more prevalent in the following time frame within the lower left quadrant of the figure and continue to become more prevalent within the final time frame located in the lower right quadrant. By the time the engine is in this final quadrant, most of the observations are occurring as anomalies. The progression of the anomalies to this stage indicates that the engine is deteriorating and allows the business to make decisions to optimize when the piece of equipment should be repaired and also when might be a good time to remove it from service.

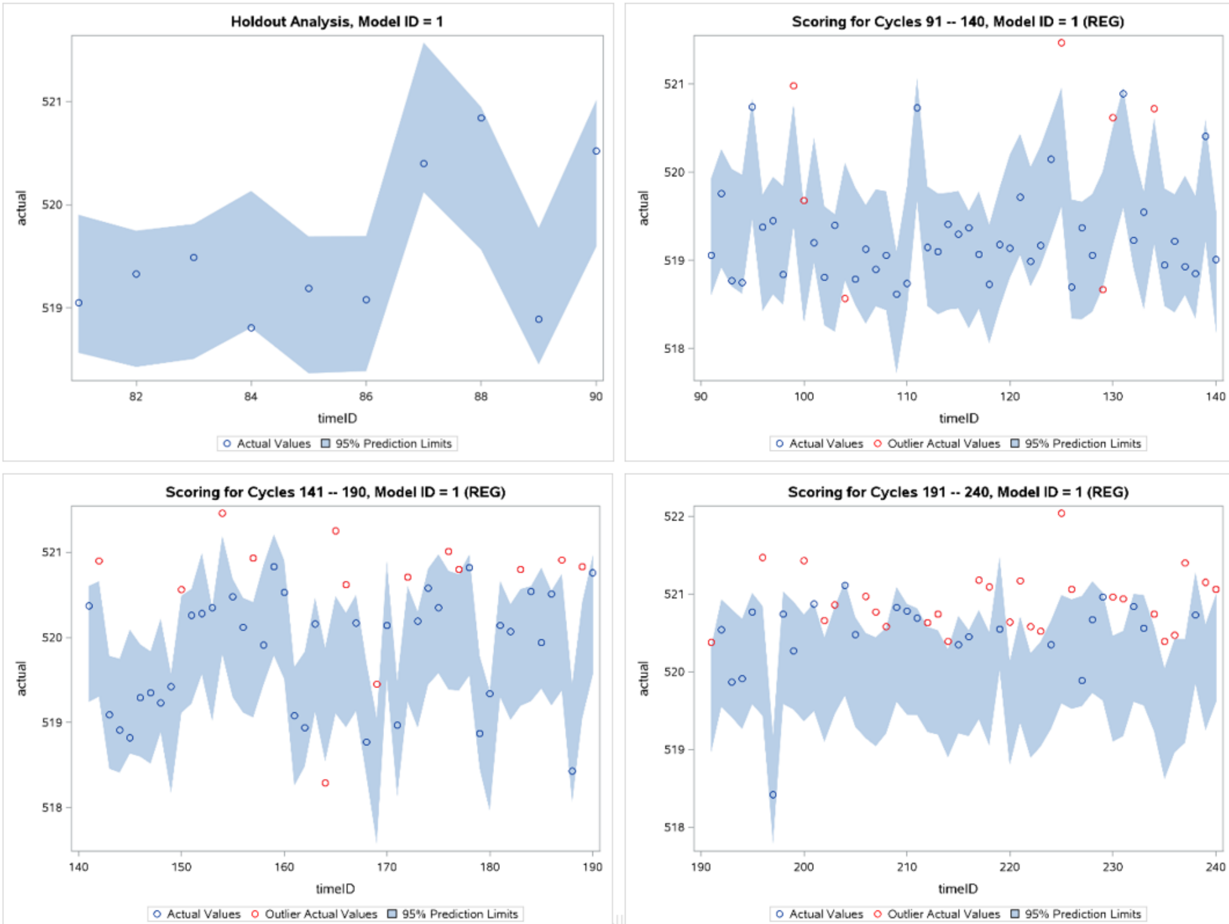

Figure 6. Stability Monitoring Results

METHOD – SUPPORT VECTOR DATA DESCRIPTION

Support Vector Data Description (SVDD) is a single class classification technique that can be used for anomaly detection. The model creates a minimum radius hypersphere around the training data used to build the model with kernel functions being used to add flexibility to the data description (Chaudhuri et al. 2016). Since the model is a single class classification technique, data from normal operating conditions are used to create the model and then observations lying outside of the boundary created by the model can be considered to be potential anomalies.

STRENGTHS OF THE METHOD – SUPPORT VECTOR DATA DESCRIPTION

Here are the advantages of using Support Vector Data Description:

- The model works with multivariate data.
- The model does not require an assumption of normality for the data.
- The model has an option to automatically select the value for the kernel bandwidth parameter.
- Since the model does not require labeled data and identifies anomalies that occur outside of the normal operating conditions that were used to train the model, Support Vector Data Description can potentially identify anomalies that are rare to occur or that might not have occurred to this point.
- Similarly, due to being trained on only one class of conditions, this analysis can be helpful in

applications where the data set is severely imbalanced between examples of normal operating conditions and examples of deteriorating conditions.

RESULTS – SUPPORT VECTOR DATA DESCRIPTION

The SVDD model was trained on the first 25% of the measurements for 30 randomly selected engines in the data set. It was then tested on the remaining 188 engines (Gillespie et al. 2017). Four of the 188 engines were randomly selected, and their results are shown in Figure 7: Sample SVDD Scoring Results. The X axis indicates the cycle of the flight, and the Y axis indicates the distance metric output by the SVDD model.

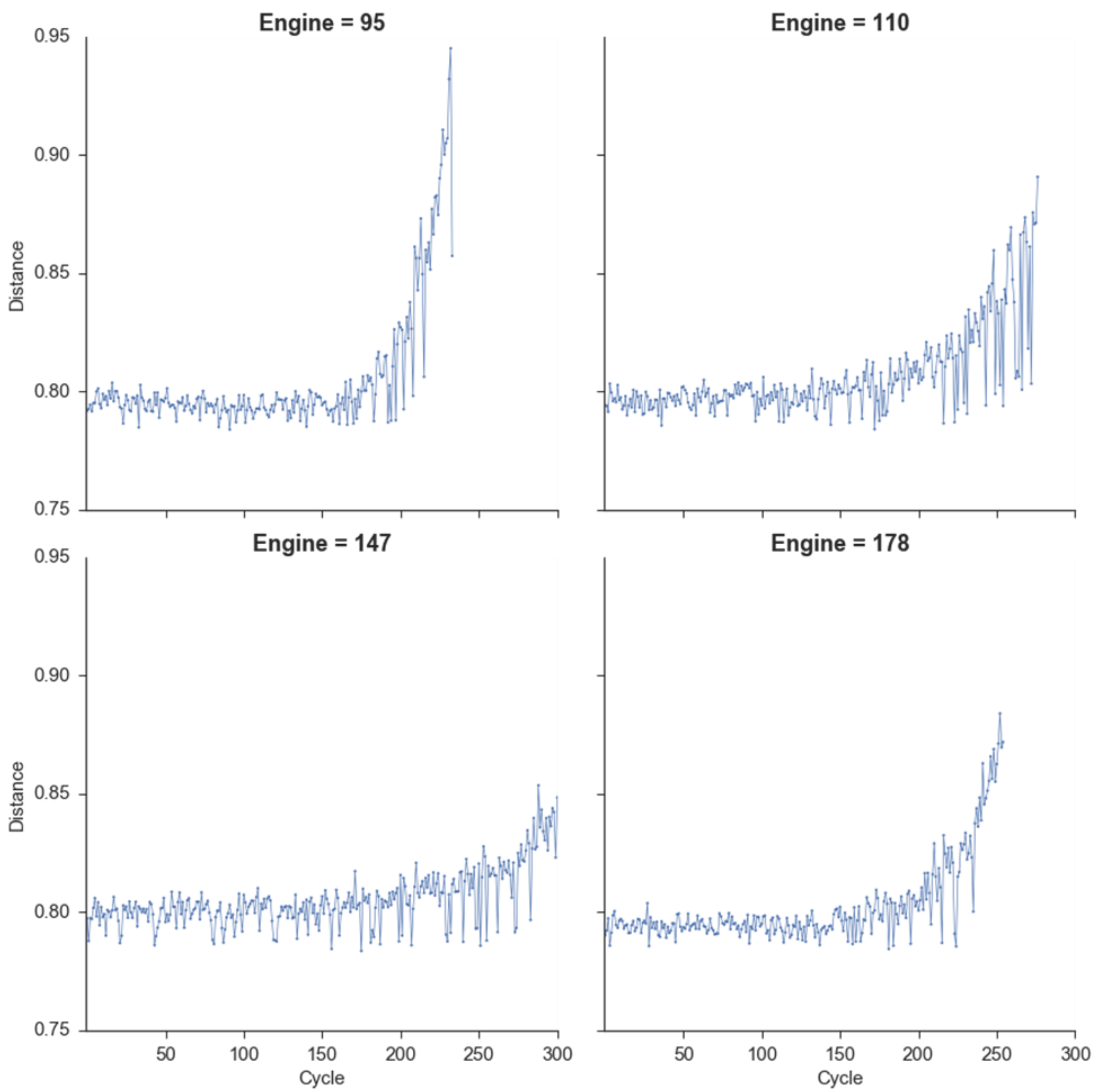

Figure 7. Sample SVDD Scoring Results

As shown in Figure 7, each of the randomly selected engines exhibits a pattern of increasing distance as it gets closer to the end of its life. This pattern begins with some small volatility around a consistent level and then begins to become increasingly volatile and rises in value as the engine begins to deteriorate. With this knowledge, decisions can be put into place to trigger maintenance or shutdown activities once the distance metric is above a certain level or remains above a certain level for a time period applicable to the use case.

CONCLUSION

This paper showcased different models and procedures available from SAS® by focusing on use cases related to the Internet of Things. It examined several techniques and methods to address anomaly detection and asset or process degradation when faced with challenges related to IoT data, such as lack of labeled failure data and uncertainty regarding the nature of historical data. The methods used included Moving Windows Principal Component Analysis, Robust Principal Component Analysis, Stability Monitoring, and Support Vector Data Description. Robust Principal Component Analysis was also used to illustrate possibilities regarding data reduction efforts for applications where transmission and storage are of concern.

REFERENCES

Saxena A. and K. Goebel. 2008. "PHM08 Challenge Data Set." NASA Ames Prognostics Data Repository. NASA Ames Research Center. Moffett Field, CA. Available https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/ Accessed February 24, 2018.

Chaudhuri, A., D. Kakde, M. Jahja, W. Xiao, H. Jiang, S. Kong, and S. Peredriy. 2018. “Sampling Method for Fast Training of Support Vector Data Description.” *Proceedings of the 2018 Annual Reliability and Maintainability Symposium (RAMS)*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Gillespie, Ryan and Saurabh Gupta. 2017. “Real-time Analytics near the Edge: Identifying Abnormal Equipment Behavior and Filtering Data near the Edge for Internet of Things Applications.” *Proceedings of the SAS Global Forum 2017 Conference*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings17/SAS0645-2017.pdf

SAS Institute Inc. 2017. *SAS® Visual Data Mining and Machine Learning 8.2: Procedures*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2017. *SAS® Visual Forecasting 8.2: Programming Guide*. Cary, NC: SAS Institute Inc

ACKNOWLEDGMENTS

A special thank you to Anya McGuirk, Zohreh Asgharzadeh, Sergiy Peredriy, Arin Chaudhuri, Deovrat Kakde and Gül Ege, whose help and work on these use cases made them possible.

CONTACT INFORMATION

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

Ryan Gillespie
100 SAS Campus Drive
Cary, NC 27513
SAS Institute Inc.
Ryan.Gillespie@sas.com
http://www.sas.com