Bad Data Detection and Data Filtering in Power System

Dhaval Bhatti
Dept of Elect Engg
FTE, M S U of Baroda
Vadodara 390001

Anuradha Deshpande
Deptt of Elect Engg
FTE, M S U of Baroda
Vadodara 390001

ABSTRACT
With increase in advanced metering infrastructure and sensor systems there is increase in data collection. It is hard to handle a large amount of data and assure the quality of data. Good quality of data is essential in power system before taking decision. So data must be cleaned and filtered before operator takes any decision from the data. Otherwise it will cause hazardous condition if poor quality of data affects decision making without knowledge of operator. Bad Data detection and data cleaning is helpful to get over this risk. With use of MATLAB Bad Data can be easily detected. Bad Data can be also removed and Data filtering as well as Data smoothing is also possible. Data smoothing is necessary for some application ex. Load forecasting in power system. Here it is obtained by using Statistical techniques such as OWA (Optimally Weighted Average) and MA (Moving Average).

Keywords
Big Data Analytic Advanced Metering Infrastructure, Load Forecasting, Smart Meter.

1. INTRODUCTION
In recent years large amount of smart meters are installed which collect information of power consumption at an interval of every 15 minutes to 1 hour. In earlier cost of data storage and data collection was significant in power system. Due to enhancement in digital storage system costs are falling rapidly. AMI (Automatic Metering Infrastructure) is introduced to power system with development in communication network technology and it has pushed down the costs of measurement and storage considerably. Smart meters together with communication system and data management system constitute to AMI. With help of Smart Meters there is improvement of data quality and more granular and high quality data is available for many applications. But, with this there is tremendous increment in data to be stored. According to survey 1.1 billion smart meters will be installed worldwide so it will produce 2 petabytes of data per year only from smart meters [1]. With smart meters installed at consumer end detailed information can be achieved. This detailed information unveils many applications by big data analytics. Data analytics has mainly three stages: Descriptive analysis, Predictive analysis and Prescriptive analysis. With this three stages one get to know what data look like, what going to happen with data and what decision can be made from data. Data analytics can be used for many applications in power system with information of domain and off domain data. Domain data includes data coming from telemetry & SCADA, consumption, synchrophasor, waveform, events, metadata, electricity market and pricing. While off domain data includes data coming from social media, videos & images, weather, GIS, traffic, gas usage, water usage, other markets [2]. There are some barriers to adopt big data in power system like addressing discarded & siloed data, to maintain balance between integrated and disintegrated system, Insufficient research on big data analytics system architecture design and advanced mathematics for large amount of data. Application of data analytics in power system are Enhance demand response, Fine granularity forecast, Equipment monitoring and fault detection, Predictive maintenance planning and load profiling.

This large amount of data needs to be stored and it is necessary to process data (filter) before getting worthy information from it. Why there is need of data filtering and cleaning? It is needed because this data may contain bad data or anomalies. What is bad data? Bad data is missing data or failure in data collection due to communication errors. In power system big data analytics is still in developing stage. So for large amount of data, effective data processing and data cleaning techniques need to be developed. Techniques for bad data detection are Probabilistic, Statistical and Machine learning [3].

Type techniques used depend on type and nature of data. In paper [4], have discussed load power data imputation method and it can be helpful for advanced DMS function. Paper has used data of Georgia Tech AMI measurement. Paper has discussed various approaches like historical average and conventional linear interpolation. Paper [5], has discussed real time anomaly detection for short term load forecasting. In VSTLF (Very Short Term Load Forecasting) most recent load information is used. Paper has proposed a method for bad data detection which has two components, a dynamic regression model and an adaptive anomaly threshold. While paper [6], has considered a security related problem with data. Paper has considered false data injection which is threat to power grid with intention of tamper important data. To tackle this problem paper has modeled the false data problem as rank bounded L1 norm optimization and shown online and offline algorithms to remove injected false data and recover original measurement data. Paper [7], has considered big data for power equipment, which are data from equipment itself, monitored data, power grid data and external data like GIS and meteorological data. This data needs to be cleaned so it ensures the validity, consistency and integrity of data. For data cleaning they have considered mainly three process which includes: 1) monitoring and filling of missing data. 2) Bad data detection and correction. 3) Ensuring data quality. They have proposed methods for Missing data imputation are mean method, Regressive method, multiple imputation, K-Nearest distance method. For Anomaly detection and governance neural network/SVM, Multiple Linear Regression, Extreme learning machine, is influencing factors. Paper [8] for load estimation has considered domestic smart meter data and for estimating missing data they have used clustering approach and distance functions.

2. DATA STRUCTURE
Data used for analysis is smart meter data of one year power consumption data of Newyork of 365 days from 1st February 2005 to 1st January 2006. This data is time series data with measurement interval of 5 min to 1 hour.

3. METHODOLOGY

With using big data there are mainly three approaches for bad data detection and data cleaning 1) Probabilistic 2) Machine Learning 3) Statistical. Selection of approach depends on type of data and data structure.

Advantage of Statistical method is that it is suitable for real valued datasets or at very least quantitative ordinal data distribution where the original data can be transformed to suitable numerical values for statistical processing. But if complex data transformation is necessary before processing then it limits their applicability and increase processing time.

Data used here is Smart meter data which is essentially time series data, so statistical techniques is used for bad data detection and data cleaning. Two approaches are considered: 1) Weighted Average (WA) [8], 2) Moving Average (MA).

3.1 Weighted Average

Weighted Average method is efficient for power system data for online and offline application. This method is performed on historical load data measurement from smart meter.

A. Linear Interpolation:

Linear Interpolation is technique for adding new data points within a range of set of known data points. It is effective with time series data of smart meters. This technique can be used to fill in missing data, smoothing existing data and also for making predictions. To estimate missing data this method uses nearest non missing values (neighbor values) for linear interpolation.

Linear Interpolation estimates missing value from previous and next available value, Yi and Yj with

$$\hat{y}_i^{LI} = y_h + \frac{y_j - y_h}{x_j - x_h}(x - x_h),\ x_h < x_i < x_j \ldots \text{ [1]}$$

This method requires only two samples to estimate missing data. If length of data period increases then accuracy of Linear Interpolation decreases.

B. Weighted Average:

Weighted average is similar to mean, the difference in that is instead of all data samples contribute equally to final average, some sample contribute more than other. Problem of missing data for longer data period can be tackled from this method. If weights of all data samples are equal then weighted average will be same as mean of that.

For Yi data samples and Wi as their weights

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} \text{ [2]}$$

$$\bar{y} = \frac{w_1 y_1 + w_2 y_2 + w_3 y_3 + \ldots + w_n y_n}{w_1 + w_2 + w_3 + \ldots + w_n} \ldots \text{ [3]}$$

Therefore sample with high weight contribute more to the weighted mean then other samples with lower weight.

3.2 Moving Average

Moving average is obtained by performing series of average of different subsets of the whole data set. It is also called as moving mean. There are three approaches according to applications: Simple Moving Average (SMA), Weighted Moving Average (WMA) and Exponential Moving Average (EMA).

Moving average is useful with time series data to smoothing sudden variation and spikes in values. Moving average is obtained with given series of data samples and fixed subset size. First element is obtained by taking average of the initial fixed subset of sample series. Then after subset is modified with going forward by removing first sample of series and adding next sample in the subset. Whenever new data sample arrives previous one is removed from subset. Accordingly variation in data is aligned with variation in average.

If we consider data samples $a_1, a_2 \ldots a_n$ for n number of samples then moving average can be obtained from

$$M.A. = \frac{1}{n}\sum_{i=i}^{i+n*1} a_j \text{ [4]}$$

4. RESULTS AND DISCUSSION

One year data of Newyork city electricity consumption from February 2005 to January 2006 is used here. Data has total 365 data sheets of whole year with per day and sampling interval is of 5 minute to 1 hour. MATLAB is used to analyze data. Data is of measurements of total 11 regional meters and various information like station id, Load, Name, Time & date and Time zone are included. So dimensionality is high with all these information. Some information is not needed like time zone and station id in our application. It is tedious process to go from all that information and process all of them even if we do not need all of that. To overcome this problem firstly dimensionality reduction is performed on data to refine only worthy information. So, only required data is taken into account for further analysis. It also helps to reduce computation and simulation time.

From all the regional data one region is selected and its power consumption is plotted against time. Here power consumed in DUNWOD region is plotted against time as shown in figure 1. As here can be seen in plot it contains so many zeros and sudden spikes in it which is indeed a bad data. In particular region there are many households and at any particular instant of time load consumed do not become zero in all house. So, this zero indicates missing data or data measurement error.

This bad data needs to be removed before this data is taken into account for any application. So, data filtering is necessary process.

This can be achieved by method discussed earlier. But data is not regularly sampled. Sampling time differ from 5 minute to 1 hour interval. So before processing data it needs to be in regular manner. So retiming of data is done on an interval of 5 minute. But at some points data will be missing due to there is no measurement of data at every 5 minute interval. So this data points needs to be filled. This problem can be solved with linearly interpolating missing data points. With short missing data period Linear Interpolation method is effective. With help of this method missing data and zero load condition is removed and at place of missing data interpolant values are filled.

With Linear Interpolation bad data can be removed and fill in missing data points using linear interpolant values. This can be seen in figure 2. As one can see in this figure zero load is removed and data at regular interval is obtained. But still there are sudden spikes of due to some bad measurements. This is not acceptable for application like load forecasting. So this spikes need to be smoothed out.

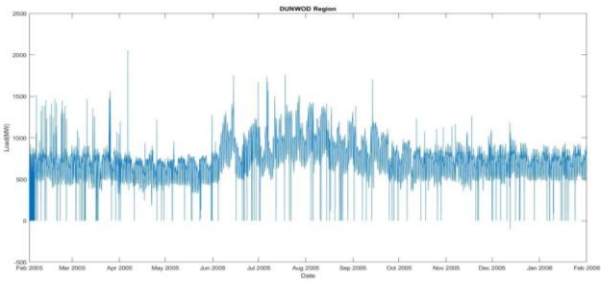

Fig 1: Plot of DUNWOD region (With Bad Data)

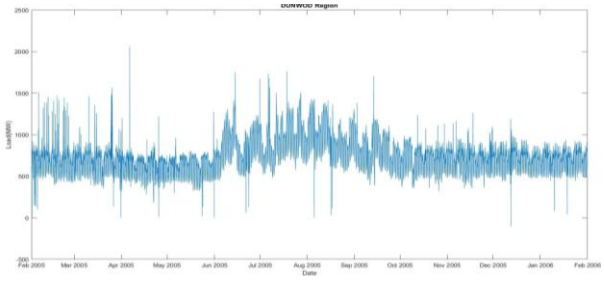

Fig 2: Plot of DUNWOD region (With removal of Zero Load)

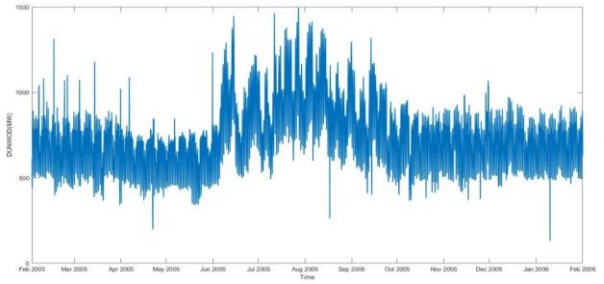

Fig 3: Plot of DUNWOD region (without bad data)

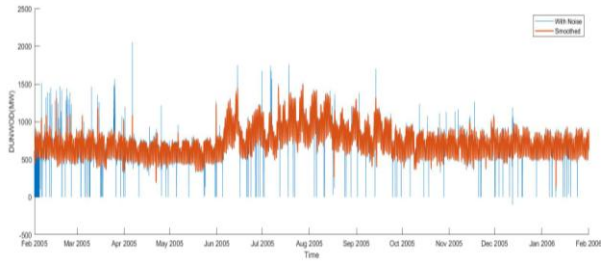

Fig 4: Plot of DUNWOD region (Comparison with noise and smoothed)

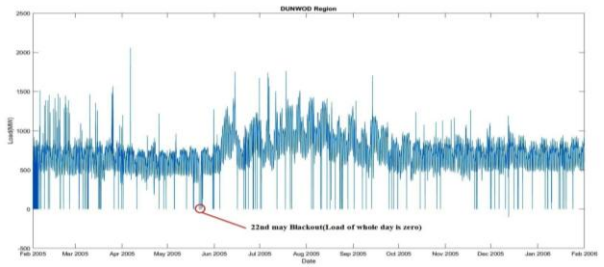

Figure 5: Plot of DUNWOD Region (Black out at 22thmay)

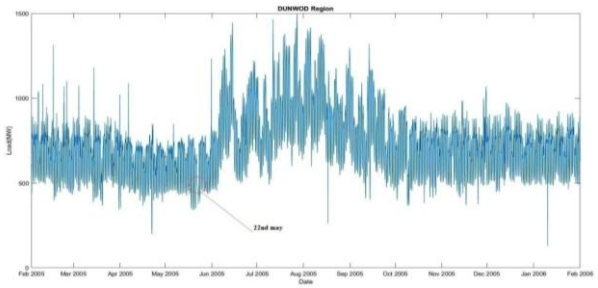

Figure 6: Plot of DUNWOD Region (data at black out filled M.A. Method)

As seen in figure 3 sudden spikes are removed and curve is smoothed out. This smoothed curve is obtained by Moving Average (MA) method. Now this data can be used for load forecasting other applications as bad data is removed from it and data is in more granular form. In figure 4 comparisons between plot of data with containing bad data and plot after removing bad data and smoothing is shown. Here one can clearly see the difference between raw data and processed data. So it is advantageous to filter raw data before taking it into account for making decision or getting information from it. Here the one more case which has to be considered is what if there is black out in whole region, if black out occurs then load of that region becomes zero and smart meter reading of that is indeed not a bad reading. So here case of one day black out at 22th may have been considered. Due to black out condition there is no electricity consumption in particular region and load reading at that whole day is zero.

As one can see in figure 5 loads at 22th may is zero due to black out. This black out needs to be considered, because it is actual condition not a bad or missing data, but for load forecasting fine granular data needed so in further process this black out zero load data is replaced with moving average value and we get result as shown in figure 6.

Big Data technology is now developing now a days and it is now possible to harness large amount of data coming from millions of consumer meters and also information of dwelling, information from social media in the various forms like excel sheets, pictures and videos etc. Due to emerging trends in BDA, it is now possible to fetch this information and it can be used to take important decision. This information after undergoing some process like bad data detection and data cleaning is more dependable and homogeneous. This cleaned and smoothed data further can be used for transmission network planning and expansion for future planning in power system.

5. CONCLUSION

In this paper methodology to detect bad data according to type of data has been discussed. Here it can be seen that for time series data statistical methods are easy to use and less complex. Linear Interpolation, weighted average and moving average methods are used for removing bad data and data smoothing and results are obtained as shown in figures. So as seen in figure bad data has been filtered out and data smoothing is done. The reason behind smoothing the data is in many application of data analytics fine and granular data is requirement. This can be done with using MATLAB software easily and time requirement for this is very less.

6. ACKNOWLEDGMENTS
Thanks to the electrical engineering department, of Faculty of Technology and Engineering M.S.U.

7. REFERENCES
[1] Hossein Akhavan-Hejazi, Hamed Mohsenian-Rad "Power systems big data analytics: An assessment of paradigm shift barriers and prospects", Energy Reports, Volume 4, Pages 91-100, November 2018.

[2] N. Yu and S. Shah and R. Johnson and R. Sherick and M. Hong and K. Loparo "Big data analytics in power distribution systems" in IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT), Vol.0, No.0, pp.1-5, Feb2015.

[3] V. Hodge and J. Austin, "A survey of outlier detection methodologies", Artificial intelligence review, vol. 22, no. 2, pp. 85–126, 2004.

[4] J. Peppanen, X. Zhang, S. Grijalva, and M. J. Reno, "Handling bad or missing smart meter data through advanced data imputation," in IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), 2016, pp. 1–5.

[5] L. J., T. H., and Y. M., "Real-time anomaly detection for very short term load forecasting," Journal of Modern Power Systems and Clean Energy, vol. 6, no. 2, pp. 235–243, 2018.

[6] H. Huang, Q. Yan, Y. Zhao, W. Lu, Z. Liu, and Z. Li, "False data separation for data security in smart grids," Knowledge and Information Systems, pp. 1–20, 2017.

[7] Chongqing Kang, Yi Wang, Yusheng Xue, Gang Mu, and Ruijin Liao "Big Data Analytics in China's Electric Power Industry" in IEEE Power and Energy Magazine, Volume 16, Pages 54 - 65, April 2018.

[8] A.Al- Wakeel, J Wu, and N Jenkins," K means based load estimation of domestic smart meter measurements." Applied Energy, Vol 194,pp 333-342, 2017

[9] Yi Wang, Qixin Chen, Tao Hong, Chongqing Kang"Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges" in IEEE Transactions on Smart Grid(Early Access) Pages 1-1, March 2018.

[10] Mingyang Sun "Big Data Analytics in Power System", Imperial Collage of London, November 2016.

[11] Fintan McLoughlin, Aidan Duffy, Michael Conlon "A clustering approach to domestic electricity load profile characterization using smart metering data", Energy Reports, Volume 141, Pages 190-199, 1 March 2015.