



Environmental Pollution Prediction of NO_x by Predictive Modelling and Process Analysis in Natural Gas Turbine Power Plants

Alan Rezazadeh*

Applied Research and Innovation Services, Southern Alberta Institute of Technology, 1301 – 16 Avenue NW, Calgary, AB, Canada T2M 0L4

Received: 31 December 2020, Revised: 27 March 2021, Accepted: 30 March 2021
© University of Tehran

ABSTRACT

The main objective of this paper is to propose K-Nearest-Neighbor (KNN) algorithm for predicting NO_x emissions from natural gas electrical generation turbines. The process of producing electricity is dynamic and rapidly changing due to many factors such as weather and electrical grid requirements. Gas turbine equipment are also a dynamic part of the electricity generation since the equipment characteristics and thermodynamics behavior change as turbines age and equipment degrade gradually. Regular maintenance of turbines are also another dynamic part of the electrical generation process, affecting performance of equipment as parts and components may be upgraded over time. This analysis discovered using KNN, trained on a relatively small dataset produces the most accurate prediction rates in comparison with larger historical datasets. This observation can be explained as KNN finds the historical K nearest neighbor to the current input parameters and approximates a rated average of similar observations as prediction. This paper incorporates ambient weather conditions, electrical output as well as turbine performance factors to build a machine learning model predicting NO_x emissions. The model can be used to optimize the operational processes for harmful emissions reduction and increasing overall operational efficiency. Latent algorithms such as Principle Component Algorithms (PCA) have been used for monitoring the equipment performance behavior change which deeply influences process parameters and consequently determines NO_x emissions. Typical statistical methods of performance evaluations such as multivariate analysis, clustering and residual analysis have been used throughout the paper.

KEYWORDS: KNN, ML, Process Degradation, Emissions, PCA, Clustering

INTRODUCTION

The main objective of this paper is introducing K-Nearest-Neighbor algorithm as a candidate to be used in Predictive Emission Monitoring Systems (PEMS), predicting Nitrogen Oxides (NO_x) emissions produced in the process of electricity production of gas turbines (Environment and Climate Change Canada, 2017). This paper uses the gas turbine process dataset from University of California at Irvine (UCI) open data repository (Kaya et al., 2019), which was collected over five year period in north western Turkey. The power generation utility donated the dataset would like to remain anonymous and author would like to extend gratitude for allowing this valuable dataset to be used in industrial analytics research. The power plant location is close to sea level, prone to humidity fluctuations, comprised of mild temperatures occasionally dropping below freezing point (Kaya et al., 2012).

* Corresponding Author, Email: Alan.Rezazadeh@sait.ca

The power generation system is a Combined Cycle Power Plant (CCPP), comprised of gas and steam turbines. Figure 1, depicts schematics of the power plant, comprised of two gas turbines of 160MWh each with a Heat Recovery Steam Generator (HRSG) powering a 160MW steam turbine (Kaya et al., 2012). The exhaust from gas turbines, usually maintain high temperatures are used for driving a steam turbine, which result in a highly efficient power generation system (Tüfekci, 2014), approximately about 60% efficiency in comparison to a simple cycle gas turbine of approximately 35% to 40% efficiency (Poullikkas, 2005).

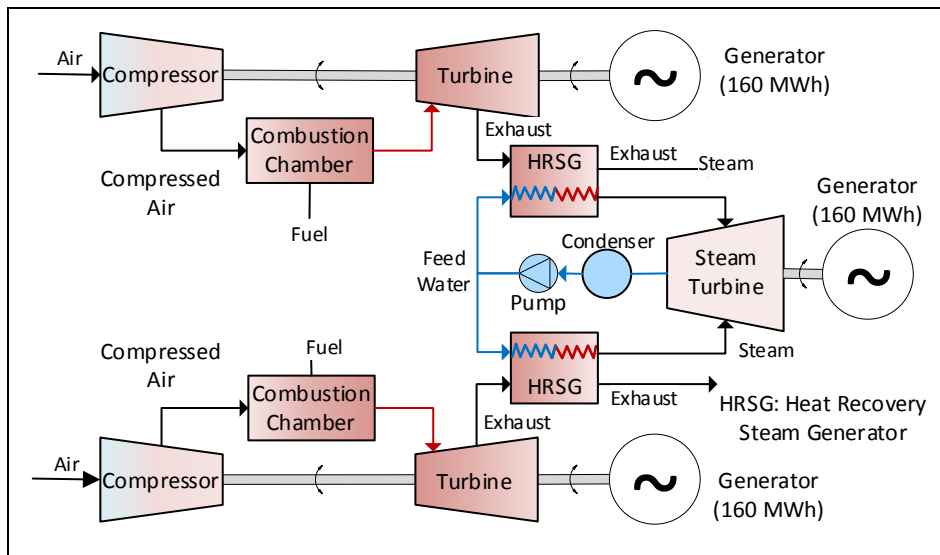


Figure 1. Combined cycle power plant schematic diagram

PEMS have been discussed widely within literature for the last 25 years as a backup to Continuous Emission Monitoring Systems (CEMS) using allocated sensors, directly measuring emissions and pollutants produced by the combustion process (Chien et al., 2010). CEMS based on dedicated hardware and necessary software have been a part of gas turbine design with well-defined legal operational requirements. Many researchers consider PEMS as a backup, or alternate monitoring system to CEMS (Chien et al., 2005). This paper presents another application of PEMS, which is monitoring the electrical generation process efficiency, in addition to predicting emissions such as NO_x under fast changing process conditions.

As the field of industrial data science is evolving, more applications to PEMS are being identified (Si et al., 2019). A new application to use PEMS can be identified as monitoring degradation and process efficiency of gas turbines (Ge et al., 2017). PEMS, naturally using many process parameters such as turbine pressures and temperatures for predicting emissions, which offer great opportunity to monitor the process performance and degradation (Yan et al., 2017).

PEMS as a method of predicting emissions uses operational data for training and building machine learning models (Chien et al., 2010). As the training data becomes longer in time, the electrical generation process may change due to new grid requirements, extreme weather conditions or equipment degradation, hence decreasing the prediction success (Miletic et al., 2004). As a result, contrary to popular believe, shorter training time may actually offer better prediction rates, utilizing more of the recent data points, rather than including all historical data (Qin et al., 2019).

Benchmarking process behavior based on physics of gas turbines and laws of thermodynamics also present another method of monitoring process drift or degradation (Ge

et al., 2017). However, due to lack of sufficient internal turbine data this paper refrains from exploring the degradation in more depth and invites the power generation industry to share more detailed internal process datasets for further analysis and research.

The paper discusses application of factor clustering, latent variables such as principle components for monitoring and early detection of process change (Kourti, 2005), and KNN predictive modelling for electrical production of gas turbines. The main objective is to better understand and predict NO_x resulted from combustion process in gas turbines.

MATERIAL AND METHODS

NO_x is a generic term for emission family of Nitrogen Oxide (NO₂) and Nitric Oxide (NO), which are usually created as a result of combustion process (Smrekar et al., 2013). Although, both transportation and power generation sectors use combustion process, the main objective of this paper is analysis of NO_x resulted from gas power plants, which contribute to smog, acid rain and tropospheric ozone (European Environment Agency, 2019) pollutions.

Even though, natural gas is relatively a cleaner burning fossil fuel, combustion process produces small amount of sulfur, mercury and particulates depending on the quality of fuel (Environment and Climate Change Canada, 2017). These pollutants are considered fuel dependent and can be eliminated by using higher quality and cleaner natural gas (European Environment Agency, 2013). In contrary NO_x considered pollutant which is resulted from higher temperature combustion such as gas turbines and diesel engines (European Environment Agency, 2019). NO_x are considered process dependent, meaning by optimizing combustion process the pollutant can be minimized (Liukkonen & Hiltunen, 2016). Table 1 illustrates typical pollutant emissions from gas turbines.

Table 1. Typical pollution emissions from gas turbines and their source

Pollutant	Gas Turbine Pollution	
	Fuel Dependent	Process Dependent
NO _x		✓
CO		✓
SO _x	✓	

PEMS are software solutions for predicting emissions based on operating process parameters such as internal turbine pressures and temperatures (Lee et al., 2005). Emissions resulted from combustion process are usually monitored and measured using CEMS; hardware sensors in two different methods of periodical or continuous intervals. In either case, specialized hardware is used, and usually maintenance of the sensors are within the operational budgets and schedules (Fichet et al., 2010).

Figure 2 illustrates the schematics of an open cycle gas turbine model (Potter & Somerton, 2019) and the data elements available for this study (Kaya et al., 2019). The utilized dataset for this research misses a few critical elements for a more thorough analysis such as consumed fuel amount, compressor discharge temperature and released Carbon Dioxide amount. Nevertheless, existing dataset offers valuable insights into operations of gas turbines and NO_x predictions. The available data elements can be found in Figure 2 as well as Table 2.

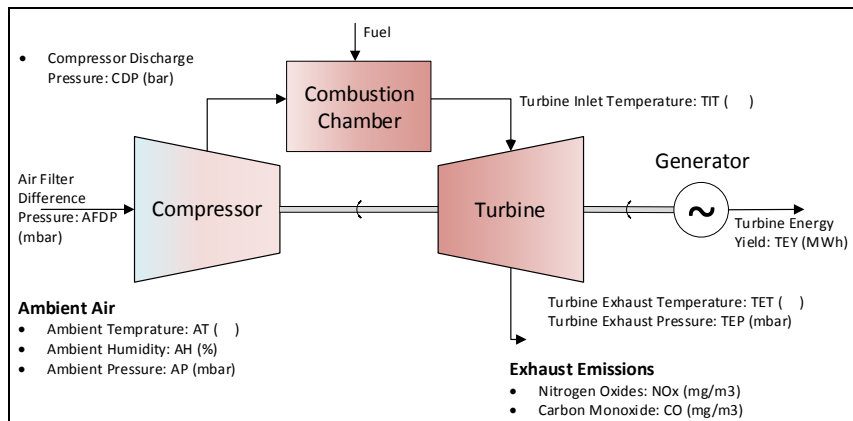
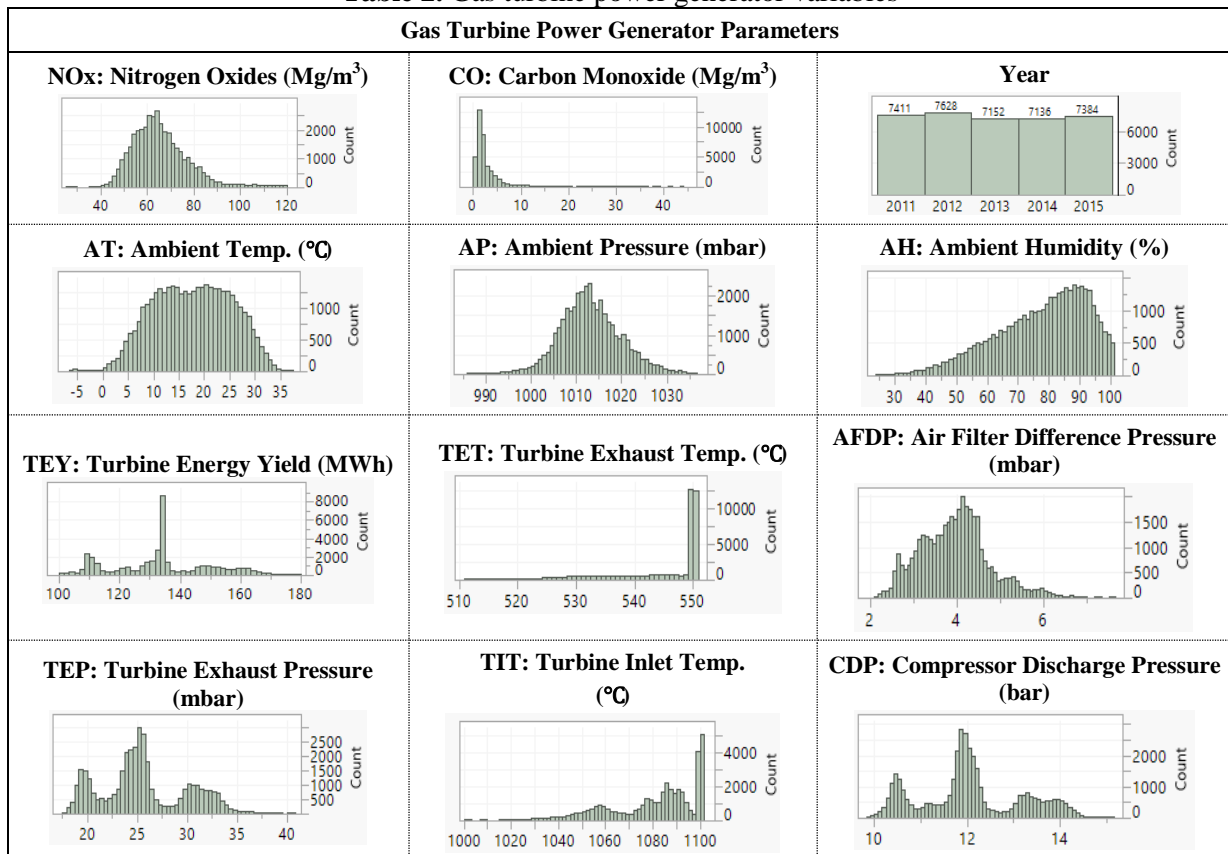


Figure 2. Schematic diagram of a simple cycle gas turbine, and the available parameters

Industrial data analytics begins with understanding the operational dataset including their internal relationships, trends and statistical distributions. Most industrial application datasets may contain tens of thousands of data points (i.e. records, rows) with tens of variables (i.e. predictors, columns) or more. The variables are typically sensor readings within a process and may contain strong collinearity, meaning groups of predictors may move together under specific conditions (Cuccu et al., 2017,).

Table 2, illustrates the histogram of power generation process parameters for the period of five years, 2011 to 2015, beginning from January 1st of each year. As can be seen there are overall 12 variables, including six internal variables to the gas turbine, which are used to describe the status of power generation process.

Table 2. Gas turbine power generator variables

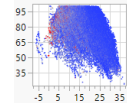
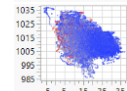
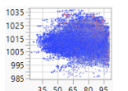
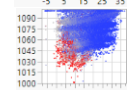
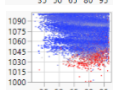
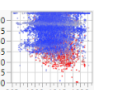
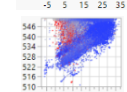
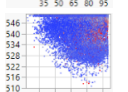
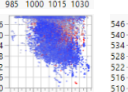
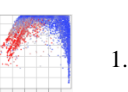
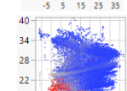
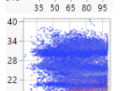
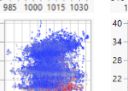
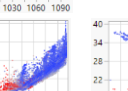
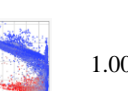
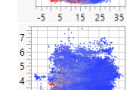
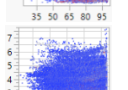
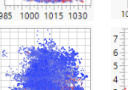
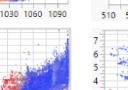
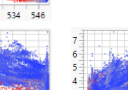
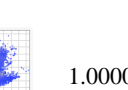
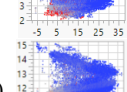
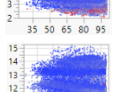
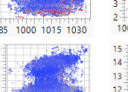
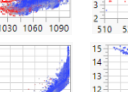
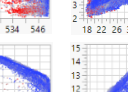
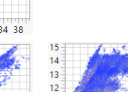

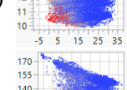
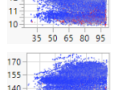
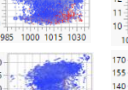
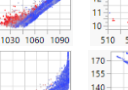
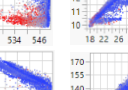
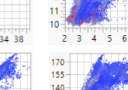
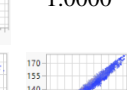
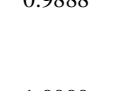


RESULTS AND DISCUSSION

Multivariate analysis is a set of techniques to analyze the multidimensional data, seeking patterns within data elements (Kano et al., 2004). Table 3 illustrates the multivariable correlation between the variables in two formats of numerical correlation and visual scatter plot. The red data points in the scatter plots indicate data readings with higher NO_x values.

As can be seen in Table 3, the higher values of NO_x are usually clustered closer to lower temperatures and lower power production yields, which result lower turbine temperatures and pressures. Visual inspection of scatter plots in Table 3, and correlation values indicate the process variables are more correlated with each other and less with the weather condition parameters. Therefore, clustering parameters in a scientific and quantitative method can clarify the relationships in more details.

Table 3. Multivariate correlations of available predictors. Red data points indicate higher NO_x values.

	AT (C)	AH (%)	AP (mbar)	TIT (C)	TET (C)	TEP (mbar)	AFDP (mbar)	CDP (bar)	TEY (MWh)
AT (C)	1.0000	-0.4763	-0.4067	0.1840	0.2821	0.0458	0.2519	0.0152	-0.0914
AH (%)		1.0000	-0.0153	-0.2218	0.0235	-0.2350	-0.1478	-0.1961	-0.1371
AP (mbar)			1.0000	-0.0043	-0.2252	0.0580	-0.0403	0.1031	0.1190
TIT (C)				1.0000	-0.3867	0.8750	0.6928	0.9093	0.9106
TET (C)					1.0000	-0.7026	-0.4680	-0.7094	-0.6861
TEP (mbar)						1.0000	0.6786	0.9785	0.9641
AFDP (mbar)							1.0000	0.7027	0.6658
CDP (bar)								1.0000	0.9888
TEY (MWh)									1.0000

Clustering of variables (Table 4), is performed by using principle components, based on application of eigenvalues and eigenvectors (Miletic et al., 2004). Clustering process begins by assigning all variables to one cluster, if the second eigenvalue of the cluster is larger than a predefined threshold, the variables split into two clusters, since second large eigenvalue indicates existence of significant variance among the second group of variables. The process continues until the second eigenvalues of all clusters fall below a predefined threshold (SAS Institute Inc., 2014).

Table 4 shows the power generation variables can be split into three clusters. Interestingly as illustrated all members of cluster 1, are the internal parameters of gas turbine, meaning the variables are highly correlated. The second cluster consists of Ambient Humidity (AH) and

Ambient Temperature (AT), meaning these two variables are correlated and move together. The third cluster consists only of one variable which is Ambient Pressure (AP), meaning this variable is independent of other factors.

Table 4. Clustering of process parameters

Cluster	Members	RSquared with Own Cluster	RSquared with Next Cluster	1 – RSquare Ratio	Comments
1	CDP (bar)	0.983	0.015	0.017	Process Dependent
1	TEY (MWH)	0.959	0.014	0.041	Process Dependent
1	TEP (mbar)	0.951	0.027	0.050	Process Dependent
1	TIT (C)	0.816	0.056	0.195	Process Dependent
1	AFDP (mbar)	0.602	0.054	0.421	Process Dependent
1	TAT (C)	0.523	0.051	0.503	Process Dependent
2	AH (%)	0.738	0.034	0.271	Weather Dependent
2	AT (C)	0.738	0.165	0.314	Weather Dependent
3	AP (mbar)	1.000	0.052	0.000	Weather Dependent

For the first cluster, Compressor Discharge Pressure (CDP) contains 98.3% of the variation within the group. Meaning using CDP is the best variable of this cluster (gas turbine parameters) to explain the cluster variance. Interestingly CDP is one of the most important factors in predicting efficiency of gas turbine Brayton cycle thermodynamics (Potter & Somerton, 2019) and also electrical yield has very strong linear relationship with CDP.

R-Square for the cluster variables are defined as the ratio of explained variance on a variable to its own cluster component (SAS Institute Inc., 2014). R-Square with next cluster is the proportion of explained variance within a variable with the next cluster. The value of 1-RSquare ratio is defined by Equation 1 as the ratio of 1 minus its own cluster R-Square to 1 minus next closest cluster's R-Square.

$$1 - RSquared = \frac{1 - RSquared \text{ with own Cluster}}{1 - RSquared \text{ with Next Closest}} \quad (1)$$

Equation 1. Definition of 1 – RSquared Ratio

Based on Table 4, there are three clusters identified, which CDP, AH and AP are the most significant parameters with the most variance for each group.

Predictor screening is used to find the contribution of each predictor to the response variable, NOx values. This technique is specifically advantageous for variables with potentially weak direct correlation with the response (NOx); however, with stronger interaction through other variables. Predictor screening is based on Bootstrap Forest (Proust, 2019) (fitting a model by averaging many trees similar to Random Forest), finding contribution and percentage portion of contribution to NOx values (Hundi & Shahsavari, 2020). Table 5 shows the most contribution was made by AT (Ambient Temperature) to NOx production, with overall 31.8% of all effects.

Excluding the weather parameters from the predictor screening and running the analysis with only turbine internal process predictors, results Table 6. As can be seen the order of parameters are still similar to Table 5 which included weather parameters as well; however, only contributions to NOx production are different.

Table 5. Identification of most related predictor to NO_x production, including weather parameters







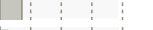

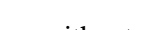






Predictor Screening - NO_x (mg/m³) (Process and Weather Parameters)				
Predictor	Contribution	Portion	Portion Ratio	Rank
AT (C)	631626	0.3185		1
TIT (C)	305719	0.1541		2
TEP (mbar)	221869	0.1119		3
TET (C)	205984	0.1039		4
AFDP (mbar)	181085	0.0913		5
TEY (MWH)	173212	0.0873		6
CDP (bar)	136176	0.0687		7
AP (mbar)	93584	0.0472		8
AH (%)	34009	0.0171		9

Table 6. Identification of most related predictor to NO_x production, without weather parameters

Predictor Screening - NO_x (mg/m³) (Only Process Parameters)				
Predictor	Contribution	Portion	Portion Ratio	Rank
TIT (C)	234449	0.2205		1
TEP (mbar)	202488	0.1904		2
AFDP (mbar)	174308	0.1639		3
TET (C)	172821	0.1625		4
TEY (MWh)	156293	0.1470		5
CDP (bar)	122904	0.1156		6

The predictor screening, Table 5 indicates the strongest factor including the weather data is ambient temperature, which may influence consumer demand for using more power during colder hours. Increased power demand, may force power generation to operate on higher yield modes, increasing TIT and CDP which result in reduction of NO_x, and increased NO_x production during lower demand hours in combination to colder air intake (Lee et al., 2005).

The dataset used in this study contains five years of hourly process parameters as described in Table 2. The multivariate analysis of predictors (Table 3) illustrates the aggregated correlation between predictors over five years of study, without including the effects of time. During five years of operations many correlations may change due to variety of factors such as weather (i.e. abnormally low or high temperatures), consumer demand change, grid requirements or equipment degradation, which would require operators to readjust process parameters for most efficient operations. For a complete analysis the effects of time and process change, detailed operational datasets are required (Kuhn & Johnson, 2013). The exact prognosis of process change will be considered out of scope for this paper, which requires more detailed turbine parameter data and power grid requirements.

As illustrated in Table 5, the most contributing factor to NO_x production is ambient temperature. The ambient temperature also affects the consumer demand on power which ultimately defines other process parameters such as turbine energy yield, operational pressure and temperature parameters (Biagioli & Güthe, 2007).

Illustrates the total energy yield versus ambient temperature. As can be seen the red data point indicating higher levels of NO_x, are consistently at lower temperatures and lower power production range. Reviewing Table 3, TEY versus TIT chart indicates NO_x are mostly produced at lower temperatures and lower electric power yields that can be due to lower consumer demand (off peak hours) or operational start-up and shut-down periods (Biagioli & Güthe, 2007).

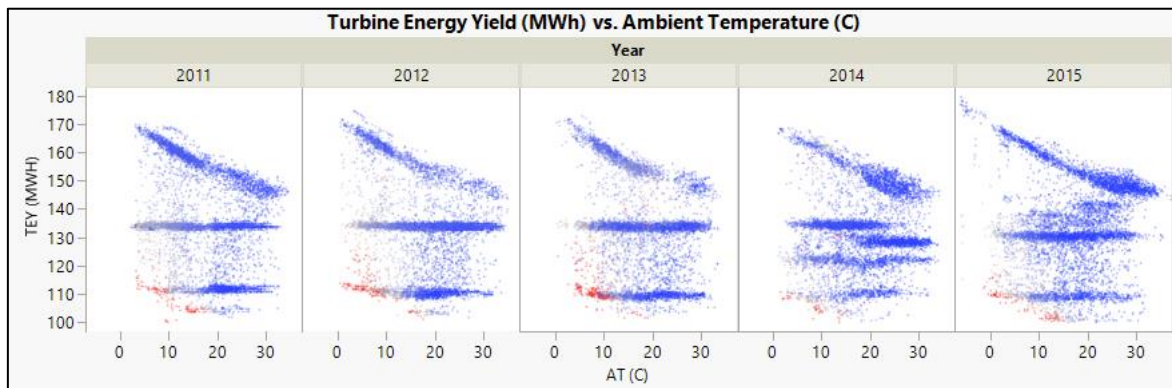


Figure 3. Bivariate analysis of turbine power output versus ambient temperature by year. Red data points indicate higher values of NOx

Analysis of equipment parameters such as CDP versus TEP (Turbine Exhaust Pressure) by year depicts the physical turbine operations characteristics change over time. As can be seen in Table 7, the relationship of CDP versus TEP, which is a hardware parameter and defined by laws of thermodynamics have changed over the life of dataset. This change highlights the reason static prediction models may lose accuracy over time since the process is changing; instead adaptive models will be required to be trained only on smaller, more recent data points predicting smaller range into future (Kourti, 2005).

As the power generation process is changing over time (Table 7), a quantitative benchmark should be used for accurate and unbiased process monitoring (Tüfekci, 2014). The existing dataset was originally intended for predictive modeling of NOx production, which lacks required data elements for efficiency analysis of the gas turbine. Identifying the exact reasons and root cause analysis of process drift requires more detailed data points, for instance more turbine pressure and temperature readings to cross reference with thermodynamics and gas laws of turbines (Potter & Somerton, 2019).

Table 7. Process change and degradation over time

2011	2012	2013	2014	2015
$CDP = 5.44 + 263.60 * TEP$	$CDP = 5.44 + 260.40 * TEP$	$CDP = 5.46 + 259.22 * TEP$	$CDP = 5.46 + 254.93 * TEP$	$CDP = 5.86 + 238.34 * TEP$
R-Square: 0.9891	R-Square: 0.9885	R-Square: 0.9923	R-Square: 0.9808	R-Square: 0.8801

Principle Components Analysis (PCA) is a dimension reduction method to reduce redundancy in a larger set of variables, generating smaller number of orthogonal vectors, preserving the information as much as possible (Kourti, 2005), while decreasing the number of variables. The analysis of CDP versus TEP; although, very valuable to show the process drift over time, only illustrates the change among two variables. PCA in contrary compresses the information into smaller number of variables, which can be used for early change detections among all variables (Kano et al., 2004).

Plotting principle component 1 versus principle component 2 by years, visualizes the

maximum information among the variables over the age of dataset (Figure 4) in a two dimensional chart. As can be seen, the principle component plots indicate change in the process over years. Application of latent variables (e.g. PCA) over time is a method for discovering process change over time which can be used for early detection of degradation for preventative maintenance (Kourti, 2005).

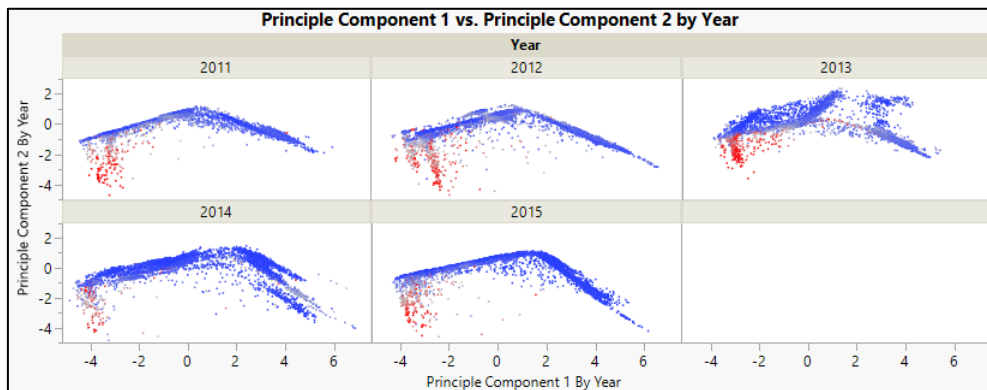


Figure 4. Principle component 1 versus 2 by year. Red data points indicate higher values for NOx.

The topic of PEMS predictive modelling have been discussed in literature by many authors for the last 25 years (Shakil et al., 2009). Most successful algorithms have been non-parametric, which do not model the process based on statistical distributions or mathematical models, simulating behavior of process under study. Instead relying on application of previous history for finding an approximation to current parameters (Ge et al., 2017).

K-Nearest-Neighbor (KNN) is a machine learning algorithm, which could be used for both classification and regression, approximating a value based on K closest training data points. KNN is considered a non-parametric algorithm, meaning the approximation is not based on any specific distribution (e.g. Poisson, Gamma or Normal), instead utilizing training dataset and finding weighted average value of K nearest neighbors based on distance to the closest neighbors, identified in training dataset (Li et al., 2020).

As discussed earlier, gas turbine electrical generation process goes through subtle changes among relationship between process parameters. As a result success rate of models predicting NOx production over time decreases, due to effects of process change, also known as process drift. This research compares two different modeling approach based on KNN, first approach for all years together and then for each year analyzed separately.

Creating KNN predictive model for all years together produced the best accuracy using three neighbors, which gives the lowest root average squared error (RASE). Figure 5, illustrates relationship between RASE and number of K for predicting NOx using KNN algorithm. This chart indicates by using only 3 neighbors for calculating distance averaging NOx, the lowest error is achieved.

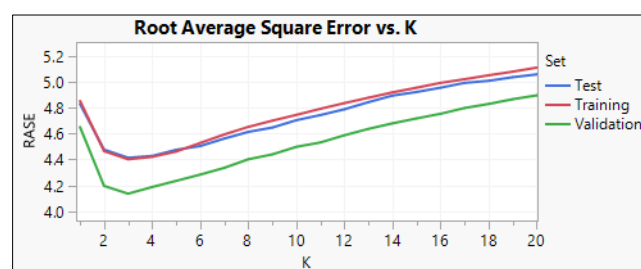


Figure 5. Root average square error versus K

KNN is a powerful algorithm for datasets with localized data point concentrations, due to reduction of average distance to each neighbor (Li et al., 2020). As observed in Figure 3 and Table 7, gas turbine operations are mostly performed within a small number of operational modes which causes data points to be relatively concentrated and forming high density locales. This characteristic, increases performance of KNN for predicting process outcomes; i.e. NOx production (Chen et al., 2018). Meanwhile, increasing number of variables (i.e. dimensions), decreases KNN performance known as “Curse of Dimensionality”, which has been discussed among academia in depth (Pestov, 2013). In case of high number of variables (dimensionality), KNN prediction performance usually drops (Bagheri et al., 2010); therefore, dimension reduction techniques such as principle components or similar methods are highly recommended for large number of data variables (Skiena, 2017).

Table 8, illustrates performance comparison of KNN for all years together versus yearly generated models. As Table 8 indicates KNN performance is slightly higher using yearly models and indicators such as R-Square, RASE and AAE (Average Absolute Error) exhibit better performance when the prediction models are based on annual parameters. This is an important observation by this paper indicating the process is changing over time and although annual models have less training data; however, produce better results.

Table 9 illustrates NOx prediction performance of KNN for each given year, indicating total annual R-Square were above 90 percent, with the lowest for 2013 (90%) and highest for 2015 (94%). This observation is an indication of process characteristics change over time which using smaller more localized datasets produce higher prediction rates than using larger datasets.

Table 8. Model performance comparisons of overall data together versus year by year

	Predicted KNN NOX (mg/m3) All Years			Predicted KNN NOX (mg/m3) By Year			Freq
	R ²	RASE	AAE	R ²	RASE	AAE	
Training	0.9349	2.9770	1.7471	0.9469	2.6874	1.5945	28554
Validation	0.8693	4.1359	2.5693	0.8919	3.7603	2.3441	4078
Test	0.8634	4.4142	2.6098	0.8934	3.8999	2.3345	4079
Total	0.9196	3.3104	1.9343	0.9348	2.9796	1.7600	36711

Table 9. Comparison of predictive model performance by year

Year	KNN By Year Details												Freq
	Training			Validation			Test			Total			
	R ²	RASE	AAE	R ²	RASE	AAE	R ²	RASE	AAE	R ²	RASE	AAE	
2011	0.9516	2.3292	1.4023	0.8960	3.3790	2.1079	0.8940	3.7366	2.1268	0.9383	2.6531	1.5611	7411
2012	0.9392	2.5264	1.5945	0.8756	3.6526	2.4075	0.8881	3.3106	2.1944	0.9267	2.7686	1.7516	7628
2013	0.9215	3.3745	2.0861	0.8574	4.5536	2.9462	0.8373	4.8497	3.0821	0.9051	3.7115	2.2924	7152
2014	0.9254	2.6972	1.4452	0.8321	3.6502	2.1397	0.8393	4.3880	2.2232	0.9054	3.0456	1.6086	7136
2015	0.9539	2.4099	1.4557	0.8955	3.4772	2.1291	0.9244	2.9827	2.0707	0.9447	2.6170	1.5988	7384

Analysis of residuals, Figure 6 illustrate consistently scattered residuals across the plot without any specific pattern, which is a positive sign of strong KNN prediction model, independent of any significant bias. The data points were broken down for each year to Training (70%), Validation (15%) and Test (15%). The exact number of the break downs can be seen under column Frequency in Table 8 and Table 9.

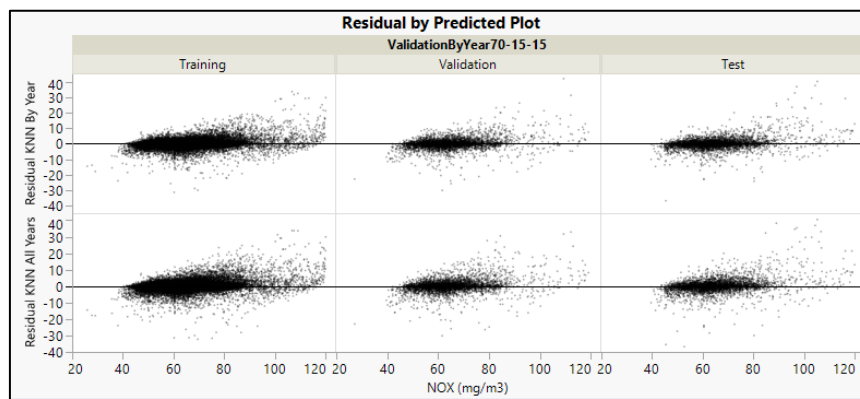


Figure 6. Analysis of residuals

Figure 7 displays the plot of actual NO_x emissions versus predicted over time. As can be seen the predictions are very close to actual, which is also supported by high R-Square values of above 90%. KNN, a non-parametric algorithm provides high success rates for NO_x prediction, by finding the previous K similar observations at the training data and then approximating the new value based on the distance from each observation.

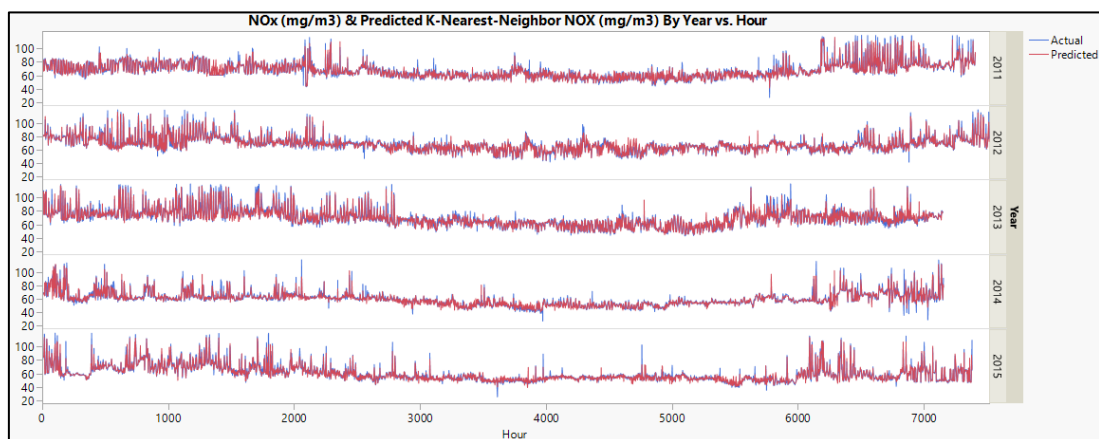


Figure 7. NO_x K-Nearest-Neighbor prediction vs. actual

CONCLUSION

The process of power generation is dependent on many dynamic factors including power demand, weather, equipment efficiencies and operational conditions. Therefore, predicting NO_x which is a process dependent pollutant can be more effectively accomplished using shorter training datasets which are more similar to current operating parameters. Hence, adaptive algorithms may offer more advantage since they assign heavier weight to more recent training data.

Industrial data analytics can only be accomplished upon availability of accurate datasets describing a dynamic process. Applications of machine learning and predictive modelling will further advance by availability of data for finding the most effective methodologies, openly discussing the results in dissertations, conferences and scientific publications. Sharing non-confidential industrial data creates the opportunity for a larger community of research and training enthusiasts, creating next generation of digital savvy workers, benefitting the very same industries by accelerating adoption of new technologies.

Throughout the paper there were discussions of process degradation. However, degradation analysis needs to be quantitatively defined, monitored and if possible minimized. Formulation of degrading process requires more detailed parameters than were available in this dataset. For instance Brayton cycle thermodynamics of gas turbine parameters could be used for monitoring the gas turbine performance, predicting degradations and efficiencies.

Adaptive KNN algorithm in pollutant release prediction as well as other applications that have a changing behavior can be explored. Modelling a dynamic process requires a resilient adaptive algorithm to incorporate gradual change as it might be due to social and economics evolution or an industrial equipment degradation.

ACKNOWLEDGEMENT

Wayne Hovdestad, M.Eng., P.Eng., has kindly collaborated with the research on the engineering analysis of the gas turbine parameters. His insight added extreme values to better understanding the gas turbine thermodynamics and detailed operations.

The data used in this analysis was produced by a power generation plant in north western Turkey which is a well populated and industrialized section of the country. This research could not be possible without the shared data with the scientific community. Author would like to express gratitude to the power generator operator for publishing and sharing the operational data. My best gratitude to Dr. Heysem Kaya, for facilitating this dataset to be publicly available and also collaborating with the author for this research.

Author would like to acknowledge CFREF (First Canadian Research Excellence Fund) for providing the opportunity to research new topics helping environmental protection as well as academia.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The author declare that there is not any conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy has been completely observed by the author.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

- Bagheri, B., Ahmadi, H. and Labbafi, R. (2010). Application of data mining and feature extraction on intelligent fault diagnosis by Artificial Neural Network and k-nearest neighbor. *The XIX International Conference on Electrical Machines - ICEM 2010*, 1–7.
- Biagioli, F. and Güthe, F. (2007). Effect of pressure and fuel–air unmixedness on NO_x emissions from industrial gas turbine burners. *Combustion and Flame*, 151(1–2), 274–288.
- Chen, X., Wang, P., Hao, Y. and Zhao, M. (2018). Evidential KNN-based condition monitoring and early warning method with applications in power plant. *Neurocomputing*, 315, 18–32.

- Chien, T. W., Chu, H., Hsu, W. C., Tu, Y. Y., Tsai, H. S. and Chen, K. Y. (2005). A performance study of PEMS applied to the Hsinta power station of Taipower. *Atmospheric Environment*, 39(2), 223–230.
- Chien, T. W., Hsueh, H. T., Chu, H., Hsu, W. C., Tu, Y. Y., Tsai, H. S. and Chen, K. Y. (2010). A Feasibility Study of a Predictive Emissions Monitoring System Applied to Taipower's Nanpu and Hsinta Power Plants. *Journal of the Air & Waste Management Association*, 60(8), 907–913.
- Cuccu, G., Danafar, S., Cudre-Mauroux, P., Gassner, M., Bernero, S. and Kryszczuk, K. (2017). A data-driven approach to predict NO_x-emissions of gas turbines. *2017 IEEE International Conference on Big Data (Big Data)*, 1283–1288.
- Environment and Climate Change Canada. (2017, November). *Guidelines for the Reduction of Nitrogen Oxide Emissions from Natural Gas-fuelled Stationary Combustion Turbines*.
- European Environment Agency. (2013). Reducing air pollution from electricity-generating large combustion plants in the European Union: An assessment of potential emission reductions of NO_x, SO₂ and dust
- European Environment Agency. (2019). Assessing the effectiveness of EU policy on large combustion plants in reducing air pollutant emissions
- Fichet, V., Kanniche, M., Plion, P. and Gicquel, O. (2010). A reactor network model for predicting NO_x emissions in gas turbines. *Fuel*, 89(9), 2202–2210.
- Ge, Z., Song, Z., Ding, S. X. and Huang, B. (2017). Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access*, 5, 20590–20616.
- Hundi, P. and Shahsavari, R. (2020). Comparative studies among machine learning models for performance estimation and health monitoring of thermal power plants. *Applied Energy*, 265, 114775.
- Kano, M., Hasebe, S., Hashimoto, I. and Ohno, H. (2004). Evolution of multivariate statistical process control: application of independent component analysis and external analysis. *Computers & Chemical Engineering*, 28(6–7), 1157–1166.
- Kaya, H., Tufekci, P. and Gürgen, F. S. (2012, March). Local and Global Learning Methods for Predicting Power of a Combined Gas & Steam Turbine. *International Conference on Emerging Trends in Computer and Electronics Engineering*, 13–18.
- Kaya, H., Tufekci, P. & Uzun, E. (2019). Predicting CO and NO_x emissions from gas turbines: novel data and a benchmark PEMS. *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, 27(6), 4783–4796.
- Kourti, T. (2005). Application of latent variable methods to process control and multivariate statistical process control in industry. *International Journal of Adaptive Control and Signal Processing*, 19(4), 213–246.
- Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling* (1st ed. 2013, Corr. 2nd printing 2018 ed.). Springer.
- Lee, Y.-H., Kim, M. and Han, C. (2005). Application of Multivariate Statistical Models to Prediction of NO_x Emissions from Complex Industrial Heater Systems. *Journal of Environmental Engineering*, 131(6), 961–970.
- Li, W., Zhang, C., Tsung, F. and Mei, Y. (2020). Nonparametric monitoring of multivariate data via KNN learning. *International Journal of Production Research*, 1–16.
- Liukkonen, M. and Hiltunen, T. (2016). Monitoring and analysis of air emissions based on condition models derived from process history. *Cogent Engineering*, 3(1), 1174182.
- Miletic, I., Quinn, S., Dudzic, M., Vaculik, V. and Champagne, M. (2004). An industrial perspective on implementing on-line applications of multivariate statistics. *Journal of Process Control*, 14(8), 821–836.
- Pestov, V. (2013). Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 65(10), 1427–1437.
- Potter, M. and Somerton, C. (2019). *Schaums Outline of Thermodynamics for Engineers, Fourth Edition (Schaum's Outlines)* (4th ed.). McGraw-Hill Education.
- Poullikkas, A. (2005). An overview of current and future sustainable gas turbine technologies. *Renewable and Sustainable Energy Reviews*, 9(5), 409–443.

- Proust, M. (2019). JMP® 15 Predictive and Specialized Modeling. SAS Institute Inc.
- Qin, Z., Cen, C. and Guo, X. (2019). Prediction of Air Quality Based on KNN-LSTM. *Journal of Physics: Conference Series*, 1237, 042030.
- SAS Institute Inc. (2014). *SAS/STAT® 13.2 User's Guide The VARCLUS Procedure*.
- Shakil, M., Elshafei, M., Habib, M. A. and Maleki, F. A. (2009). Soft sensor for and using dynamic neural networks. *Computers & Electrical Engineering*, 35(4), 578–586.
- Si, M., Tarnoczi, T. J., Wiens, B. M. and Du, K. (2019). Development of Predictive Emissions Monitoring System Using Open Source Machine Learning Library – Keras: A Case Study on a Cogeneration Unit. *IEEE Access*, 7, 113463–113475.
- Skiena, S. S. (2017). *The Data Science Design Manual (Texts in Computer Science)* (1st ed. 2017 ed.). Springer.
- Smrekar, J., Potočnik, P. and Senegačnik, A. (2013). Multi-step-ahead prediction of NO_x emissions for a coal-based boiler. *Applied Energy*, 106, 89–99.
- Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60, 126–140.
- Yan, J., Meng, Y., Lu, L. and Li, L. (2017). Industrial Big Data in an Industry 4.0 Environment: Challenges, Schemes, and Applications for Predictive Maintenance. *IEEE Access*, 5, 23484–23491.

