Case Studies in using Consumer Analytics with PHM Strategy

Sameer Vittal[1], Mark Sporer[2]

[1]*GE Power, Atlanta, GA, 30068, USA*
Sameer.Vittal@ge.com

[2]*GE Renewables, Greenville, SC, USA*
Mark.Sporer@ge.com

ABSTRACT

As part of the "Digital-Industrial Revolution", the world is seeing the rapid transformation and digitization of the world's energy value network – from generation, through transmission & distribution, to end user consumption. This new paradigm comprises of new business products and services built on data flows that accompany energy flows; where the insight gained from sensors and analytics drives better decision making and customer outcomes. This is what drives the digital strategies of Original Equipment Manufacturers of large industrial assets like power plants, oil & gas equipment, aviation fleets, etc.

In this paper, we look at how analytical methods originally developed in the consumer industry can be applied to industrial data. This helps guide the development of Prognostics & Health Management strategies that are tuned to customer preferences and value models, in addition to engineering inputs. These methods complement, rather than replace, FMEA-driven strategies that are traditionally used in PHM systems design.

1. INTRODUCTION

Traditional PHM systems are designed "bottom up", starting with a FMEA, and then progressing through a series of trade studies where sensors, anomaly detection and remaining life algorithms are selected and integrated to reduce unplanned failures from specific failure modes. The methods do not typically consider marketing or survey data, qualitative customer information, or other exogenous economic variables that are needed to "sell" the PHM system and realize it's true value. The availability of retail and e-commerce generated data on the other hand, has led to the maturity of many consumer analytics techniques like Latent Class Analysis, Text Mining, Multiple Correspondence Analysis, Choice and Conjoint models, etc., that work with traditional data mining classification and clustering methods to parse out preferences and differentiate products from a customer's perspective. These methods and workflows can also be applied to industrial data, and can help drive PHM systems architectures that can be customized to consumer segments, increasing their adoption, usage and ultimately, business value.

In this paper, we provide an overview of consumer analytics techniques that are relevant to industrial data, and show how they can be applied via two cases studies. The first is based on risk-based segmentation of bearing failures observed in wind turbine fleets while the second case study deals with usage-based segmentation analysis of coal-fired power plants. In both cases, we hope to offer insights that would not be available using traditional PHM design methods. Finally, it is hoped that these case studies would motivate the use of consumer analytics methodologies within the broader toolkit of PHM system design methods.

2. BACKGROUND

2.1. Traditional FMEA based System Design Methods and Actuarial Engineering Approaches

The traditional approach to assessing engineering risk is through the use for FMEA's. They formed the cornerstone of the "RCM" approach to asset management, as originally explained by Moubray (1997). Jardine & Tsang (2006) provide a comprehensive overview of how Weibull-based part lifing methodologies can be used for intelligent asset management. This formed the first generation of statistically-derived asset risk management methods. These approaches were subsequently improved by adding sensors for risk mitigation, as explained by Pecht, (2008). A broad overview of contemporary PHM methods is provided by Vachtsevanos, Lewis, Roemer, Hess, & Wu, (2006), including many state-of-the art algorithms for anomaly detection, diagnosis and prognosis.

In a completely different area, actuarial methods were developed over several decades in the 20th century by the insurance industry to measure, price and manage financial and operational risk inherent in industrial systems. They encompass a broad range of techniques, from simple spreadsheet-driven approaches to more complex stochastic simulations. Typically, they involve the use of discrete distributions to model the frequency of claim occurrences (frequency distributions) as well as continuous distributions of associated costs (severity distributions) to model their financial impact. This resulting distribution of expected losses (the "loss model") forms the baseline by which PHM systems are designed to reduce financial risk by shifting the average of expected losses, and more importantly by sharply reducing the "tail risk" which is often the rare but severe risk that can impact the profitability of an asset. Klugman, Panjer & Wilmott (2008) provide a comprehensive overview of methods used in developing actuarial loss models.

Inspired by both the traditional FMEA-based approaches as well as actuarial methods, Vittal, S. and Phillips, R., (2007) developed engineering approaches to model and mitigate lifecycle costs using actuarially-derived risk calculations, rather than relying on engineering-defined specific failure modes. The advantage of this "Actuarial Engineering" approach was the mitigation of lifecycle risk from unplanned events to cheaper, planned events. This approach was practical, profitable and elements were adopted in pricing long term service agreements and in optimizing M&D systems to help manage portfolio risk with a financial target.

Some limitations in this approach became apparent as energy markets saw the rise of a widely distributed and unpredictable energy mix (E.g. Renewables – Wind and Solar). In this market, power generation asset missions were a function of market forces, weather effects, customer preferences and "portfolio choices" in addition to traditional maintenance and operational risk. The attendant services revenue stream was *not* driven by a combination of traditional planned and unplanned maintenance risk only, and the "probability of an asset being deemed economical to dispatch" became a critical variable. In simple terms, if customers chose not to run their assets, there would be an impact to the services revenue stream, and global customers are widely different in their preferences of choosing an asset mix in each market condition.

From the preceding discussion, the PHM community has started to realize that additional approaches to PHM systems design were needed to understand market & consumer preferences in a statistically rigorous way, using power plant operational data as well as "soft" measurements like configuration, economic forecasts, preferences, risk tolerance etc. For example, Vogt L.J, (2009) provides an overview of electricity pricing using traditional engineering principles, and Weron, R. (2006) provides an overview of stochastic-models and data-driven approaches to forecasting electricity loads and prices. Most of these measurements come from surveys, unstructured text, third party datasets, market indicators, etc., and do not fit easily into the tools PHM engineers typically use to design systems. The authors of this paper have been working on this problem, and believe that this is where techniques from consumer and marketing research can have an important role to play.

2.2. Consumer and Marketing Research Analytics

The field of Digital Marketing is rapidly growing, driven by the availability of massive datasets regarding customer transaction histories, their social media preferences and internet browsing user experiences. This has led to the development of many extremely sophisticated statistical methods designed to model the activity of existing and future customers, and Chaffey, D., and Chadwick, F.E., (2015) provide a comprehensive overview of the field. At the heart of this is *Predictive* Customer Lifetime Value, (CLV) which is the total amount a customer is expected to pay over the course of his relationship with a supplier. In an engineering context, this would be the expected revenue generated from asset sales, part sales, maintenance services and other value-added services (software subscriptions, upgrades) that go along with the asset.

To model CLV, a variety of methods are used. One popular method is the Pareto/NBD approach, popularized by Fader and Hardie (2016), where the probability of a customer churning is modeled using a Pareto distribution and the expected number of items ordered are modeled using a Negative Binomial Distribution (NBD). In order to improve the accuracy of CLV calculations, a variety of methods are used to "understand" customers from their digital traces. This includes segmenting the modeling space and input distribution parameter modeling for CLV analysis.

Analytics to "understand customers" include,

a) Methods to model transaction histories (E.g. Association Rule Mining, Cohort Analysis)

b) Methods to segment your customers (Persona Analysis, Clustering methods like Normal Mixtures, Hierarchic, Latent Class Analysis, Kohenen Maps, Multinomial Logit Models, etc.)

c) Models to understand preferences (Conjoint Analysis, A/B testing), and

d) Models to predict customer attrition (Churn Analysis, Markov Processes, Survival Analysis) and their drivers.

The goal of segmentation analysis is to move beyond mass marketing and gain a data-mining based insight into distinct peer-groups who operate assets in a similar manner, or who purchase services in a similar way. Using lifecycle segmentation, one can identify customers likely to stay active and those likely to switch. Using persona-based segmentation, we can define customer clusters based on what

they purchase, or how they operate. A lot of sensor and configuration data lends itself to persona-based segmentation, usually achieved using clustering algorithms that tend to operate on quantitative input parameters. However, given the preponderance of *qualitative* data on customer behavior (configuration data, purchase histories), it also becomes necessary to use methods based on qualitative inputs, and Latent Class Analysis (LCA) has become one of the most popular methods for this.

2.3. Latent Class Analysis

As described by Magidson, J., and Vermunt, J.K., (2002), "…Traditional models used in regression, discriminant and log-linear analysis contain parameters that describe only relationships between the *observed* variables. LCA models (also known as finite mixture models) differ from these by including one or more discrete *unobserved* variables. In the context of marketing research, one will typically interpret the categories of these latent variables, the latent classes, as clusters or segments". In simplistic terms, LCA allows one to identify clusters when the data is categorical (E.g. qualitative), where each level of the unobserved variable is called a "latent class". It is assumed that these "unobserved" (hence latent) clusters are responsible for generating the observed measurements, and these latent classes can then provide insights into customer behavior that would not be apparent otherwise.

In our paper, we have used the LCA implementation in JMP/Pro V14 software. The various steps involved in fitting LCA models are,

1. Assemble the dataset, "N" observations, each having a vector of variables $X_{1,n}$, …, $X_{k,n}$.
2. Convert all the numerical X's into "categories" using either standard univariate binning, or a mixture-based binning algorithm.
3. Run the LCA analysis across a range of clusters. Pick the model with optimal number of clusters as the one which has the lowest BIC score. In some cases, the software may report an AIC score, though that is not typically used in industry practice for this type of analysis
4. Assign cluster ID's to the N observations
5. Provide a narrative to the meaning of these clusters, by studying them in the context of the inputs variables in them (E.g. using share charts). This narrative provides the "persona" inherent in the cluster
6. Uses these clusters as inputs variables for life-regression analysis (to model segment based risk), or as a segmentation mechanism where actuarial methods, or CLV models can be run within each cluster.
7. We also compare the results from LCA clustering with those from traditional survival (E.g. Weibull) analysis and highlight advantages of the proposed approach.

This is explained with the help of the following two case studies.

3. WIND TURBINE BEARING CASE STUDY

In this case study, the dataset consists of N = 879 unique wind turbines where 44 bearing failures have been recorded. **To protect proprietary information, all variables have been anonymized and age has been scaled**. This does not impact any resulting analysis or conclusion. The turbine age at failure has been recorded, along with other variables like the bearing type, customer name, wind farm name, the type of service relationship (full maintenance, partial, no agreement or unknown), and the capacity of the turbine (two levels, dependent on upgrades). This is summarized below.

Table 1: Wind Turbine Input Parameters

Input Variable	Description	Categorical Variable Levels
SerialNumber	Not used	879
PartType	Part that's failing	2
CustID	Customer ID	19
ParkID	Wind Farm ID	19
Censor	Failed or OK	55
Age	Age in Years	Continuous
ServiceType	Service Agreement	4
Capacity	Wind Turbine Type	2
AgeBin	Binned from Age	3

Age was a continuous variable, and needed to be binned intelligently. A distribution analysis of age, as shown in Figure 1 below, indicated that a Three-Mixture Normal distribution had the lowest AICc score, when compared to 13 distribution types, and was selected as the best fit.

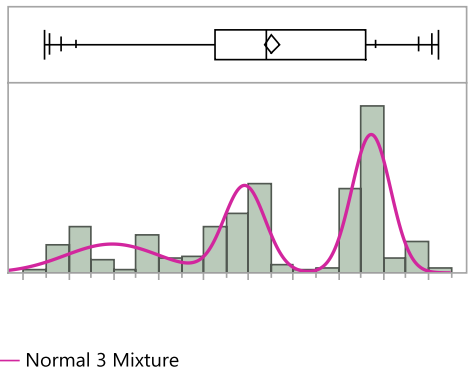

Figure1: Histogram and the fitted 3-Mixture Normal Density Function of Age (X axis, Years)

The next step was to convert this into three "bins" using a Normal Mixtures algorithm, which calculates the probability that the given observation will fall into (in this case) one of the three underlying normal distributions. This probability-model based cluster assignment approach is superior to other distance-based classifiers (E.g. K-Means, Hierarchic) particularly when the clusters overlap and results need to be robust. In addition, the previous distribution analysis did confirm the existence of three normal mixtures, which adds more confidence in the resulting clusters. The assignment of age to these clusters is shown in the figure below.

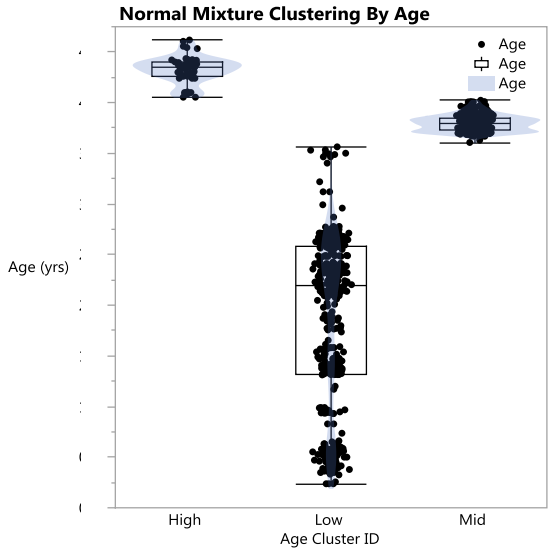

Figure 2: Box Plot & Contour Plot of Age binned into three bins using Normal Mixtures Clustering

The next step was to run the LCA model using PartType, CustID, ParkID, ServiceType, Capacity and AgeCluster as inputs. Note that "Censor" which had the bearing failure indicator was omitted, as we were interested in seeing the natural segments that emerged before associating them with failure risk. The number of clusters were selected iteratively from 2 to 12, and the optimal number of latent classes was found to be 5, which had the lowest BIC scores.

Table 2: Performance metrics for the Optimal LCA Model with 5 clusters

Measure	
-LogLikelihood	4861.505
Number of Parameters	404
BIC	12461.64
AIC	10531.01

Table 3 shows the impact that the various inputs had on the clustering model. As expected, customer ID's and wind farm ID's had a significant effect as maintenance practices and local weather effects can be significant. It's worth mentioning that these are qualitative variables that would have been completely ignored in conventional risk or segmentation models that use numerical inputs only.

Table 3: Relative effect of input variables (LCA model)

Column	Effect Size	LR Logworth
PartType	0.7228	105.09
CustID	1.6395	408.55
ParkID	1.9948	437.08
ServiceType	1.1044	235.5
Capacity	1.0323	241.51
AgeCluster	0.9458	194.87

Table 4: Variation in Performance scores of the LCA algorithm Vs. Number of clusters

NCluster	-LogLikelihood	BIC	AIC	Best
2	6167.92	13427.2	12657.8	
3	5506.92	12654.3	11497.8	
4	5249.44	12688.4	11144.9	
5	4861.51	12461.6	10531	Smallest BIC
6	4591.6	12470.9	10153.2	
7	4554.51	12945.8	10241	
8	4292.99	12971.9	9879.98	
9	4158.27	13251.5	9772.53	
10	3974.62	13433.3	9567.25	Smallest AIC
11	3980.79	13994.7	9741.58	
12	3928.76	14439.7	9799.53	

A useful output is the share chart below. This, along with numerical parameter estimates, helps provide the narrative underlying the various segments. For example, Cluster1 is mainly dominated by Part Type 2 (94%), Customer ID 13 (78%), has a variety of parks in it (nothing dominates), and mostly has customers with service type 3 (99%), capacity type 2 (85%) and units in the low age cluster (99%). Using domain and market knowledge, these "shares" can be mapped to appropriate customer personas.

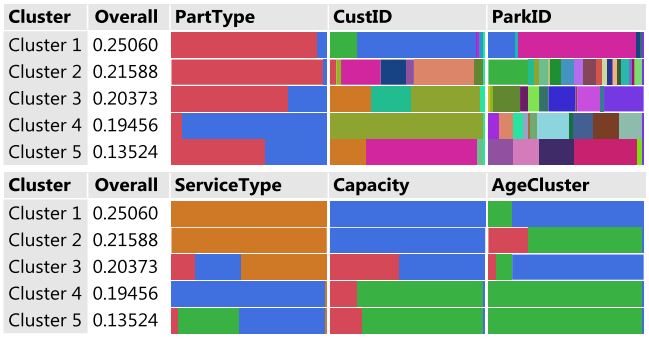

Figure 3: Cluster Share Chart, showing the relative proportion of input variables in each cluster.

Other useful diagnostics include Multidimensional Scaling charts as shown in Figure 4. As a simplistic interpretation, a well-segmented model should have cluster spread across this

space, not bunched up (which indicates similarities). MDS is often referred to as the qualitative analog of numerical PCA.

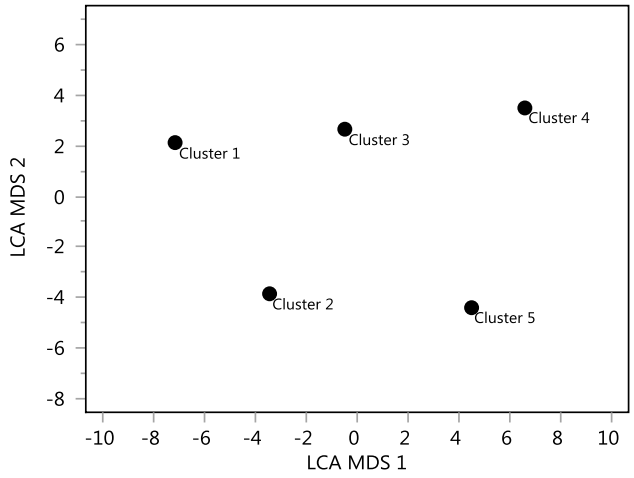

Figure 4: Multi-Dimensional Scaling Plot for Clusters from the Latent Class Model

The final stage of the analysis was to see how these clusters can explain the observed pattern of bearing failures, and/or provide information that can improve failure risk prediction for the fleet.

As a next step, we have used a "Weibull Regression" model to estimate the failure probabilities in each cluster. Here the parameters of a Weibull (or other life model) are modelled as linear combination of the LCA Cluster ID's (Variable LCA_5cluster). i.e. the customer segments are used to describe the survival model parameters. The Likelihood Ratio based test indicates the Weibull model with LCA cluster's as covariates is statistically significant

Table 5: Survival-LCA Model Fit Summary

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
LCA_5cluster	4	4	21.7492458	0.0002 *

Figures 5 and 6 show the survival (Weibull) probability plots for cases where both location scale parameters are impacted by the LCA clusters, and one where the location parameter only can vary by cluster (Weibull shape is held constant across clusters)

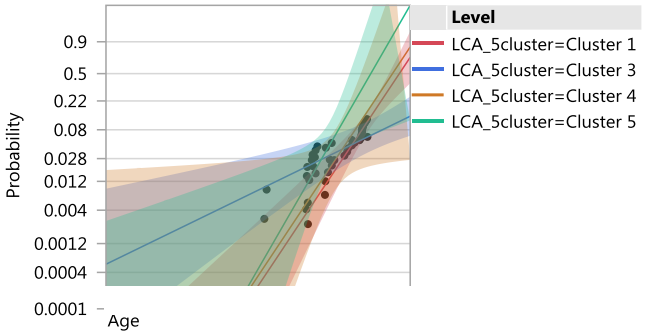

Figure 5: Weibull probability plot with both location and scale parameters varying by cluster type

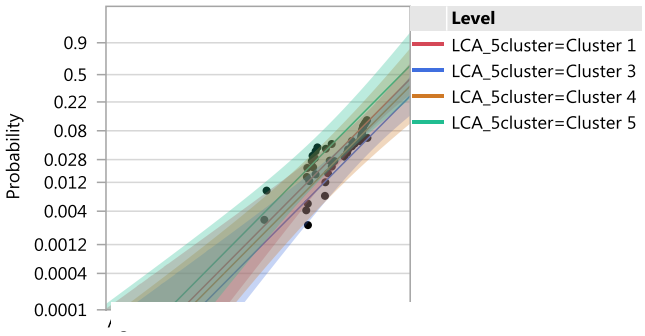

Figure 6: Weibull probability plot with only the location parameter varying by cluster type

It is helpful to compare the results of this approach with standard reliability analysis used in industry. Figure 7 below shows results from a standard Weibull model that does not include any LCA-based clusters as covariates.

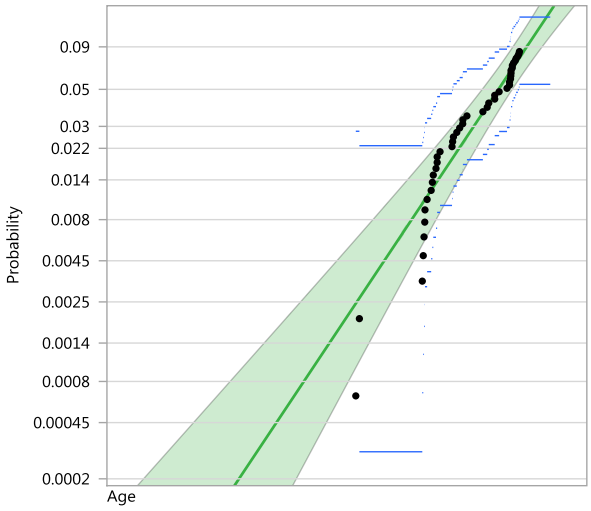

Figure 7: Standard Weibull Analysis

The table below shows the failures predicted by standard Weibull analysis, as well as those from the proposed LCA-Cluster based Weibull model ("LCA-Weibull").

Table 6: Comparison of failure predictions

LCA	Actual Data		Standard		LCA-Weibull	
Cluster	Failed	OK	Failed	Error	Failed	Error
Cluster 1	25	194	16.2	35%	23.5	6.0%
Cluster 2	0	191	7.1	high	0.0	0.0%
Cluster 3	11	168	13.5	-23%	10.6	3.4%
Cluster 4	3	168	3.0	-1%	3.0	1.3%
Cluster 5	5	114	2.4	52%	4.9	2.2%

It is interesting that Cluster 1 has a disproportionate share of failures followed by Cluster 3 with clusters 4 & 5 having a lower failure rate. It would also be helpful to compare the environmental, usage variables as well as maintenance policies between clusters 1 and 2 which provide the highest "contrast" in risk. Finally, we also see a vast improvement in failure prediction accuracy across the fleet segments, which can help provide actionable insights into failure cause drivers as well as personalize maintenance policies for each segment.

4. COAL PLANT OPERATIONAL ANALYSIS CASE STUDY

This case study focusses on cluster analysis of operational data obtained for a year from a fleet of 152 coal-fired power generators, across 93 unique power plants, all operating in the USA. Data was collected in hourly intervals for a year and was available in commercial data feeds and government sources. The goal of the study was to see if outlier detection and cluster analysis would identify segments in the population that would benefit from targeted PHM based monitoring. To protect proprietary information, all customer-identifiable information has been anonymized.

The analysis steps are like the ones detailed earlier with a few exceptions. The raw inputs feeds are hourly time series, and a series of statistically derived features need to be extracted to provide one snapshot per observation. A Normal-Mixture binning algorithm was used to bin continuous variables. The following features were used as inputs to the cluster analysis.

1. Plant Name (Categorical, 93 levels, not used)
2. Unit ID (Categorical, 152 levels, not used)
3. Primary Fuel Code (Categorical, with 4 levels)
4. Heat Rate, in BTU/kW-Hr. both as continuous measurements and binned into two categories
5. Total CO2 emissions (tons), both as continuous measurements and binned into two categories
6. Total SO2 emissions (lbs.), both as continuous measurements and binned into two categories
7. Total NOx emissions (lbs.), both as continuous measurements and binned into two categories
8. Number of major load swings, both as continuous measurements and binned into two categories
9. Number of minor load swings, both as continuous measurements and binned into three categories
10. Unit age, in years. Continuous and binned into three categories

A typical exploratory analysis plot of emissions vs plant age and colored by coal type, is shown in Figure 8. It is hard to identify any patterns in the data, other than the fact that most of the units are over 30 years old and there is a wide variability in their heat rate and annual emissions produced.

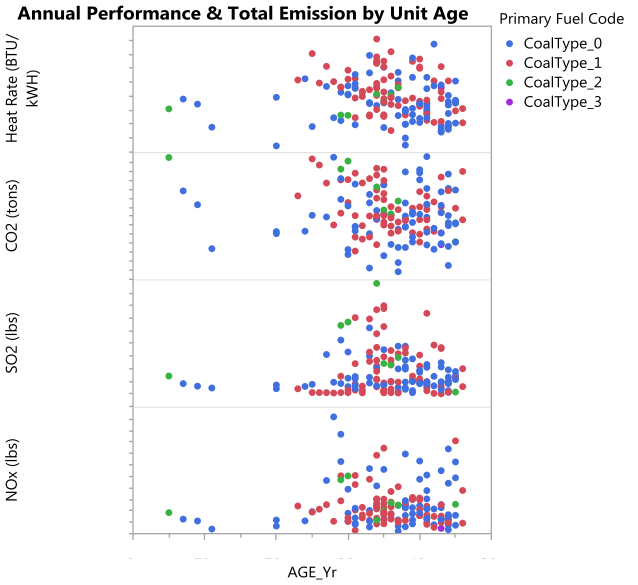

Figure 8: Exploratory performance and emissions analysis

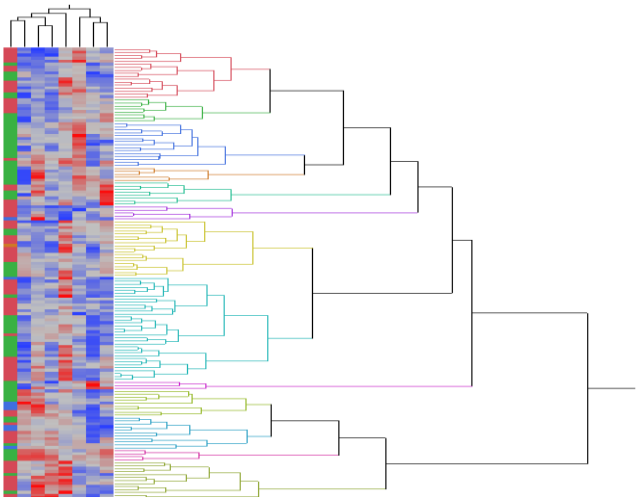

Figure 9: Two-Way Hierarchic Cluster Analysis plot with inputs in columns, and power generation units as rows.

A distance based clustering method (Hierarchic) and a LCA were applied to features 3 through 10 from the preceding list. Except for coal type, the Hierarchic Clustering method used

the continuous form of the inputs. Results from the hierarchic model are shown in Figures 9 and 10 below. The algorithm identified 13 clusters in the population.

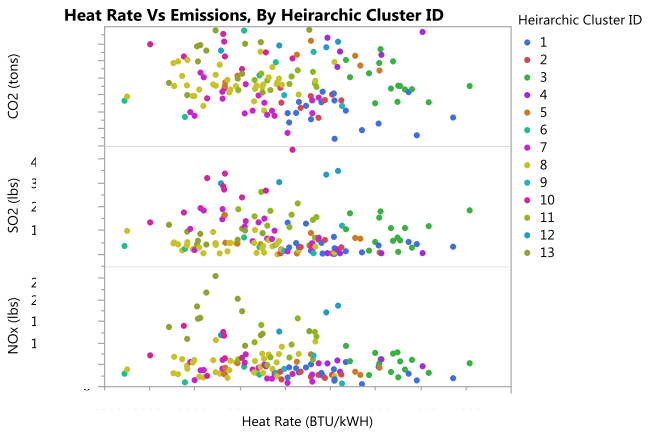

Figure 10: Performance-Emission segments for coal-fired units identified using Hierarchic Clustering

In the next step, we completed a Latent Class clustering analysis on the same dataset, where the continuous variables we binned based on optimal normal mixture density clustering. The algorithm picked 3 clusters, which is more tractable from a marketing segmentation and maintenance planning perspective. Details on the composition of each fleet segment is shown in the share chart below.

Visually, a pattern starts to emerge with a few segments tracking high-efficiency (low heat rate), low emissions units while we see clusters of units that would benefit from targeted upgrades to improve performance and reduce emissions. A combination of these two methods can then be used to identify segments where PHM requirements and deployment strategies can be used, in conjunction with FMEA and Actuarial Engineering methodologies.

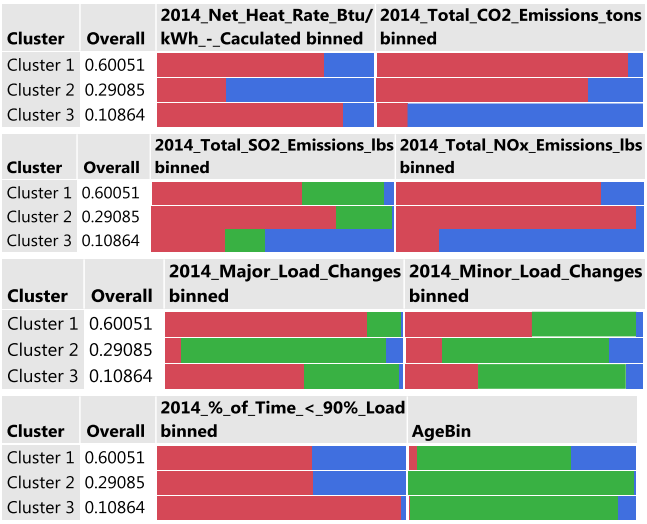

Figure 11: LCA cluster share chart for coal plant data

Table 7: Relative effect of input variables on the LCA model for coal plant data

Column	LR Logworth	Effect Size
2014_Major_Load_Changes binned	17.236	0.718
2014_Total_CO2_Emissions_tons binned	11.386	0.6504
2014_Total_NOx_Emissions_lbs binned	9.3515	0.5733
2014_Net_Heat_Rate_Btu/kWh_-_Caculated binned	6.5044	0.4464
2014_Total_SO2_Emissions_lbs binned	6.0448	0.5908
AgeBin	4.5797	0.3634
2014_Minor_Load_Changes binned	3.8116	0.377
2014_%_of_Time_<_90%_Load binned	2.6479	0.2359

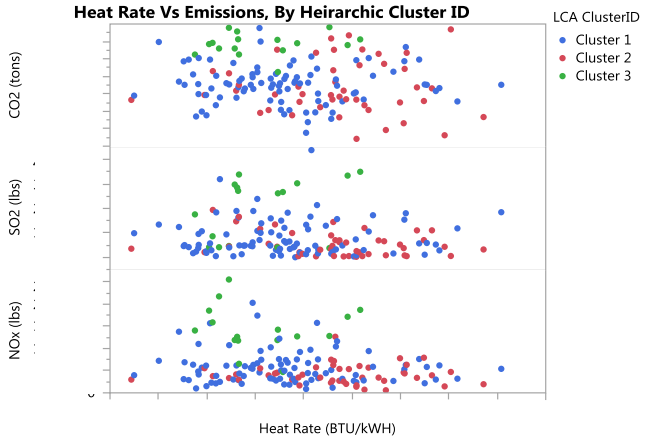

Figure 12: Performance-Emission segments for coal-fired units identified using LCA Clustering

In summary, we would like to propose the inclusion of customer analytics techniques as part of the fleet asset manager's toolkit, as it provides better risk and failure predictions as seen in the wind turbine case study. In addition, methods like LCA include both qualitative and quantitative variables to segment the fleet in a practical and actionable manner, where targeted PHM systems can be developed in a cost-effective way.

ACKNOWLEDGEMENT

The authors would like to thank Brian Marriner (GE) for helping us with data for the coal study, and our colleagues in GE Power and Renewables for their support. We would like to acknowledge the JMP Division of SAS Institute, Inc, for providing us with early adopter versions of JMP/Pro 14 which was used in the analysis.

NOMENCLATURE

AICc	Akaike Information Criteria, Corrected
BIC	Bayes Information Criteria
CO2	Carbon Dioxide
kW-Hr	Kilo Watt Hour
CBM	Condition Based Maintenance
CLV	Customer Lifetime Value

FMEA	Failure Modes and Effects Analysis
FMM	Finite Mixture Model
LCA	Latent Class Analysis
LR	Likelihood Ratio
MCA	Multiple Correspondence Analysis
MDS	Multi-Dimensional Scaling
NOx	Nitrous Oxides
NBD	Negative Binomial Distribution
OEM	Original Equipment Manufacturer
PCA	Principal Component Analysis
PHM	Prognostics & Health Management
RCM	Reliability Centered Maintenance
SO2	Sulphur Dioxide
WLR	Weibull with Life Regression

REFERENCES

Moubray, J., (1997), *Reliability Centered Maintenance II*, Oxford, UK: Butterworth -Heinemann, Inc

Jardine, A.K. S., & Tsang, A.H.C., (2006). *Maintenance, Replacement and Reliability: Theory and Applications*, Boca Raton, FL: Taylor & Francis. LLC

Pecht, M.G., (2008), *Prognostics and Health Management of Electronics*, Hoboken, NJ: John Wiley & Sons, Inc

Vachtsevanos, G., Lewis, F. L., Roemer, M., Hess, A., & Wu, B. (2006). *Intelligent fault diagnosis & prognosis for engineering systems*. Hoboken, NJ: John Wiley & Sons

Klugman, S., Panjer, H.H., Willmot, G.E., (2008), *Loss Models: From Data to Decisions, 3rd Ed*., Hoboken, NJ: John Wiley & Sons, Inc

Vittal S. & Phillips, R., *Modeling and Optimization of Extended Warranties Using Probabilistic Design, RAMS2007*, Orlando, FL (2007)

Vogt, L.L., (2009), *Electricity Pricing: Engineering Principles and Methodologies*. Boca Raton, FL: Taylor & Francis. LLC

Weron, R., (2006), *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*., Hoboken, NJ: John Wiley & Sons, Inc

Chaffey, D., and Ellis-Chadwick, F., (2012), *Digital Marketing: Strategy, Implementation and Practice, 5th Ed*., Pearson Education, USA.

Fader, P.S., and Hardie G.S., (2016), An Introduction to Probability Models for Marketing Research, 27th Annual Advanced Research Techniques Forum, London Business School, London, UK.

Magidson, J., and Vermunt, J.K. (2002). Nontechnical introduction to latent class models. Statistical Innovations White Paper #1

BIOGRAPHIES

Sameer Vittal is the Director of Data & Analytics for GE Power – Global Fleet Services, having been in a variety of reliability, CBM and data analytics roles with GE since 2000. He manages global teams responsible for the development, implementation and lifecycle management of analytics related to Monitoring & Diagnostics, Asset Performance Management, Product Services and related Software Solutions for the Industrial Internet. He has a PhD in Mechanical Engineering with a strong research and industrial background in data mining, reliability engineering, predictive modeling and optimization. He is based in Atlanta, GA.

Mark Sporer is the Technical Leader for Customer Application Engineering at GE Renewables, having been in a variety of reliability, risk management and related engineering roles in GE. Mark is responsible for risk and actuarial analytics used in Wind Maintenance Services Agreements, in addition to reliability analysis of emerging and top issues. Mark has a MS in Electrical Engineering with a background in data mining and predictive modeling, and is based in Greenville, SC