AI Decision Support Prognostics for IoT Asset Health Monitoring, Failure Prediction, Time to Failure

Kenny Gross
Oracle Physical Sciences Research Center
Oracle Corp.
San Diego, CA
guang.wang@oracle.com

Guang Chao Wang
Oracle Physical Sciences Research Center
Oracle Corp.
San Diego, CA
kenny.gross@oracle.com

Abstract*—This paper presents a novel tandem human-machine cognition approach for human-in-the-loop control of complex business-critical and mission-critical systems and processes that are monitored by Internet-of-Things (IoT) sensor networks and where it is of utmost importance to mitigate and avoid cognitive overload situations for the human operators. We present an advanced pattern recognition system, called the Multivariate State Estimation Technique-2, which possesses functional requirements designed to minimize the possibility of cognitive overload for human operators. These functional requirements include: (1) ultralow false alarm probabilities for all monitored transducers, components, machines, subsystems, and processes; (2) fastest mathematically possible decisions regarding the incipience or onset of anomalies in noisy process metrics; and (3) the ability to unambiguously differentiate between sensor degradation events and degradation in the systems/processes under surveillance. The prognostic machine learning innovation presented herein does not replace the role of the human in operation of complex engineering systems, but augments that role in a manner that minimizes cognitive overload by very rapidly processing, interpreting, and displaying final diagnostic and prognostic information to the human operator in a prioritized format that is readily perceived and comprehended.

***Keywords*—Data Preprocessing; Machine Learning Algorithms; Anomaly Detection.**

I. INTRODUCTION

In this paper we present a novel approach to human-machine cognition for human-in-the-loop supervisory control applications [1], to assist the operators of complex engineered systems by (1) helping them to deduce the state of the system from the (possibly faulty) sensor data, and (2) providing expert advice on possible actions, even in the face of incomplete knowledge. For human-in-the-loop supervisory control applications, it is the job of the operator to interpret the state of the subject system from the monitored parameters and take appropriate actions. This is hard enough to do with a complex system, and almost impossible under the pressure of an emergency when there may be multiple faults, numerous alarms, conflicting data and missing or incomplete information. Of utmost importance in such scenarios is avoiding the onset of cognitive overload for the human operators.

The new AI-based system proposed in this paper implements an advanced pattern recognition framework, called the Multivariate State Estimation Technique (MSET) [2-5], to very rapidly process, interpret, and display final diagnostic and prognostic information to the human operator in a prioritized format that is readily perceived and comprehended.

Model-based reasoning is reasoning about the behavior of a system using a model based on the *empirical* structure and function of the system. Ideally, well-constructed models will also aid in providing explanations of the state and behavior of the system. Expert systems are sophisticated computer programs that manipulate knowledge to solve problems efficiently and effectively in a narrow problem area. An expert system provides high-level expertise to aid in problem solving. The expertise (knowledge) is explicit and accessible. Two capabilities of expert systems that are particularly important in this work are predictive modeling and "root cause" explanation. A vital element of the root cause explanation is disambiguation between false alarms (also called Type-I errors in statistical process control), from real anomalous behavior in the monitored systems or processes. A predominant cause of false alarms in conventional machine-learning (ML) prognostics is the fact that conventional ML surveillance methodology works on threshold based actuation. We discuss later that how threshold-based ML prognostics result in either lower sensitivity for annunciation of anomalies, or higher false alarm rates, or both.

It is in these kinds of real-time problem-solving situations that many of the limitations of human subject matter experts (SMEs) are at their most apparent. Their tendency to overlook relevant information, to respond too slowly and to panic when the rate of information flow is too great, all contribute to lower than desired levels of performance. It is the goal of the research presented in this paper to provide effective decision support in order to transform the environment from an inefficient, data-intense, high-cognitive-demand situation to an efficient, information-rich, high-performance human-machine system.

II. METHODOLOGY

A. *MSET2 Overview*

MSET2 for prognostic health monitoring of business-critical systems comprises a comprehensive methodology for proactively detecting and isolating failures, recommending condition-based maintenance (CBM) [6], and estimating in real time the remaining useful life (RUL) [7] of critical components. Over the last 18 years, Oracle has developed and patented a suite of advanced pattern recognition innovations that leverage MSET2 prognostics for components, subsystems, and for integrated hardware-software systems in enterprise data centers [8-10]. The key enabler for achieving MSET2-based Electronic Prognostics capabilities is a continuous system telemetry harness (CSTH), which collects and preprocesses any/all types of time series signals relating to the health of dynamically executing components and subsystems. These time series provide quantitative metrics associated with physical variables (a typical data center now contains up to one million physical sensors inside the IT assets measuring distributed temperatures, voltages, currents, power metrics, fan speeds, vibration sensors), and performance variables (CPU & memory loads, throughputs, queue lengths, process metrics, etc.). The CSTH signals are continuously archived to an offline circular file (i.e. the "Black Box Flight Recorder"), and are also processed in real time using the advanced pattern recognition technique MSET2 for proactive anomaly detection and for RUL estimation with associated quantitative confidence factors.

The prognostic research initiative presented in this paper shows how MSET2 based prognostics developed for enterprise data center applications are now being spun off for human-in-the-loop control applications involving dense-sensor IoT business critical applications in the fields of Oil&Gas, smart-manufacturing, utilities, and transportation (including aviation). CSTH (real-time) plus BBR (offline) telemetry coupled with MSET2 pattern recognition help to increase asset reliability margins and system availability goals while reducing (through improved root cause analysis) costly sources of "no trouble found" events from spurious false alarms that can cause costly down time in customer's critical assets.

Oracle's suite of MSET2-related innovations bring significant advantages over conventional and competitive machine monitoring and ML approaches for real time surveillance of business-critical assets, the most significant of which are:

1) The ability to proactively catch very subtle incipient disturbances, even when the disturbance signature is a tiny fraction of the inherent variance in the monitored metrics
2) Ultra-low False-alarm and Missed-alarm probabilities (FAPs and MAPs)
3) Separately Specifiable FAPs and MAPs [note: conventional equipment surveillance approaches have a "sea saw" relationship between false- and missed-alarms.]
4) Real Time signal validation and sensor operability validation [note: most FAPs and MAPs in prognostic health management of business-critical and even safety-critical systems are due to sensor degradation events.]
5) Low compute cost for large-scale prognostic monitoring applications, i.e. lots of sensors and/or high sampling rates. (In many past "bake off" comparisons between MSET and neural networks, MSET achieves an order of magnitude higher sensitivity for catching subtle disturbances in noisy process variables, with an order of magnitude lower compute cost)
6) Remaining Useful Life estimation with quantitative confidence factors [note: RUL capability is a key enabler for "Condition Based Maintenance" of customer IoT assets.]
7) Highly accurate "inferential variable" capability. (i.e. one doesn't have to shut down a million dollar asset because a $2 internal sensor failed. MSET can swap in a highly-accurate inferential variable, so the sensor fix/replacement can be postponed to a scheduled maintenance window).

By extending the prognostic surveillance envelope to include an IoT customer's production assets, programmable logic controllers, power supplies, motor-operated valves, and interconnecting networks, all of the benefits in the bullets above helps IoT PHM applications achieve higher availability with lower Operation and Maintenance costs.

B. *The Sequential Probability Ratio Test (SPRT): Avoidance of False Alarms*

MSET2 provides a superior surveillance tool because it is sensitive not only to disturbances in signal mean, but also to very subtle changes in the statistical moments of the monitored signals and the patterns of correlation between/among multiple types of signals. MSET employs a statistical pattern recognition technique called the Sequential Probability Ratio Test (SPRT) [11-13], which provides the basis for detecting very subtle statistical anomalies in noisy process signals at the earliest mathematically possible time, thereby providing actionable warning-alert information on the type and the exact time of onset of the disturbance. Instead of simple threshold limits that trigger faults when a signal increases beyond some threshold value, the SPRT technique is based on user-specified FAPs and MAPs, allowing the end user to control the likelihood of missed detection or false alarm. For sudden, gross failures of sensors or system components the SPRT annunciates the disturbance as fast as a conventional threshold limit check. However, for slow degradation that evolves over a long time period (e.g., gradual decalibration bias in a sensor; very subtle voltage drift from the variety of aging mechanisms that cause resistances to change very slowly with age; bearing degradation, lubrication dryout, or buildup of a radial rub in all types of rotating machinery; the gradual appearance of new vibration spectral components in the presence of noisy background signals), the SPRT raises a warning of the incipience or onset of the disturbance long before it would be apparent to any conventional threshold based rules.

Many industrial processes have embedded diagnostic systems and online statistical process control techniques that perform real-time analysis of process variables. Most of these systems employ simple tests (e.g., threshold, mean value + three-sigma, SPC control-chart thresholds) that are sensitive only to gross changes in the process mean, or to high step changes or spikes that exceed some threshold-limit test to determine whether or not a failure has occurred or a process is drifting out of control. These conventional methods suffer from either large false-alarm rates (if thresholds are set too close) or high missed (or delayed) alarm rates (if the thresholds are set too wide). For new dense-sensor IoT monitoring applications in industrial manufacturing facilities, utilities, and transportation assets, false alarms are very costly in terms of plant or physical-asset down time. Missed alarms can be even more costly when incipient problems are not identified and expensive assets fail catastrophically.

The overall MSET2 framework consists of a training phase and a monitoring phase (Fig. 1). The training procedure is used to characterize the monitored equipment using historical, error-free operating data covering the envelope of possible operating regimes for the system variables under surveillance. This training procedure evaluates the available training data and automatically selects a subset of the data observations using a similarity operator that are determined to best characterize the monitored asset's normal operation. It creates a stored model of the equipment that is used in the monitoring procedure to estimate the expected values of the signals under surveillance. In the monitoring step, new observations for all the asset signals are first acquired. These observations are then used in conjunction with the previously trained MSET2 model to estimate the expected values of the signals. MSET2 estimates are extremely accurate, with error rates that are usually only 1 to 2 percent of the standard deviation of the input signal. (BTW, the MSET2 estimate for a signal originating from any physical transducer is more accurate than the transducer itself).

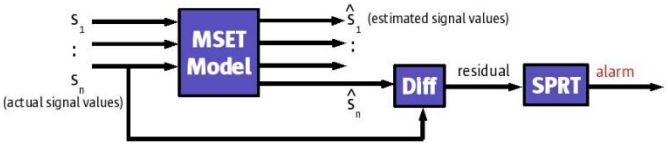

Figure 1: MSET2 surveillance-phase block diagram

The difference between a signal's real-time MSET estimate and its directly sensed value is termed a residual. The residuals for each monitored signal are used as an anomaly indicator for sensor and equipment faults. Instead of using simple thresholds to detect fault indications, SPRT is able to determine whether the residual error value is uncharacteristic of the learned process model and thereby indicative of a sensor or equipment fault. The SPRT algorithm is a significant improvement over conventional threshold detection processes in that it provides more definitive information about signal validity with a quantitative confidence factor with statistical hypothesis testing. This approach allows the user to specify FAPs and MAPs, allowing SME control over the likelihood of false alarms or missed detection.

With MSET2 plus SPRT, the ML surveillance framework achieves:

a) Ultra-low MAPs, which boosts the overall availability for critical production assets by avoiding serious outages.

b) For IoT industries where prognostic alerts lead to automatic shutdowns of revenue-generating assets, we additionally benefit from ultra-low FAPs.

Moreover, just the fact that MSET2 prognostic solutions allow FAPs and MAPs to be separately controlled is a very large win for IOT use cases (because their conventional prognostics are most likely threshold based, meaning they have to pick FAP or MAP to minimize, which for threshold-based prognostics causes the other one to go up). Oracle's solutions avoid the tyranny of the "Quality-Control sea-saw effect" between anomaly-detection sensitivity and false alarm probabilities.

C. *Intelligent Data Pre-processing (IDP) Innovations*

Oracle's IDP innovations serve as front-end data preprocessing in reference to the back-end MSET and SPRT algorithms. Herein we highlight some features of key IDP algorithms for maximizing the value-add for customers of Prognostic ML and Data Mining techniques.

a) Analytical Resampling Process (ARP)

Because the various data streams may originate with differing sampling rates, this step uses interpolation-based upsampling and downsampling methods to generate uniform sampling intervals for all telemetry time series. Moreover, it is very common that the various internal asset clocks, control network clocks, and environmental variable-monitoring clocks are out of synchronization. Clock mismatch issues will cause almost all ML prognostic algorithms to fail. Oracle's ARP prevents this issue with real-time empirical phase synchronization. Refer to [14, 15] for details.

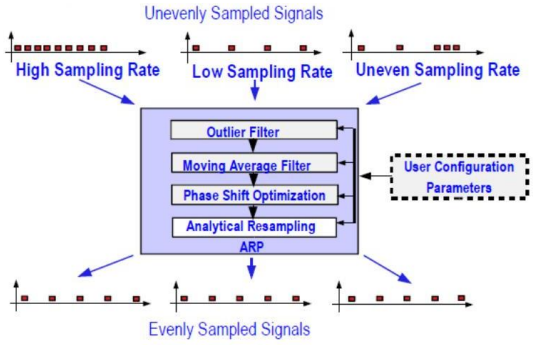

Figure 2: The schematic of ARP technique.

b) Unquantization of Quantized Sensor Signals

Another big challenge with using telemetry signatures in computational machine learning algorithms is quantization, which can severely affect the resolution of the telemetry signals (and hence accuracy of the computed results) [16-17]. Quantization occurs from low-bit A/D chips typically used in industrial and high-tech equipment transducers. Oracle's prognostic solution UnQuantize has built-in techniques to "un-quantize" signals in real time, in effect producing high-accuracy output signatures from low-resolution input signals. Figure 3 presents a typical use case where the unquantization technique is applied to a quantized data.

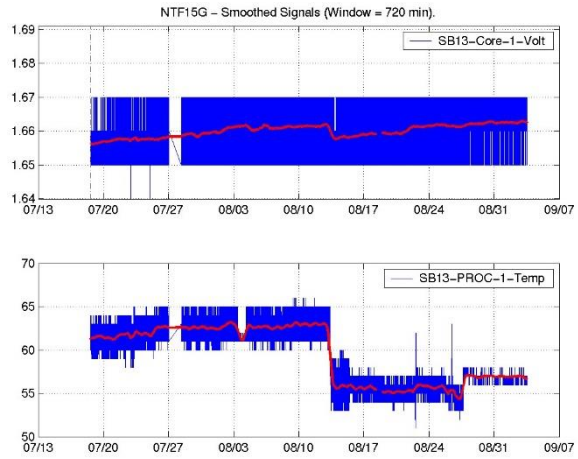

Figure 3: Blue signals show the raw signals are reported from 8-bit A/D chips frequently used in industrial machinery (and in most enterprise computing servers as well). Upper plot is a typical voltage; lower plot is a typical temperature. The red signal shows the high-accuracy value of the variable being monitored after ARP with Oracle's "unquantization" innovation to attain high-accuracy prognostics from low resolution sensors.

*c) **Missing Value Imputation (MVI)***

The last primary challenge for dense-sensor IoT application is the missing values in sensor time-series signals. Conventional approach for "filling in" the missing values is doing interpolation. When the ultimate end-goal for the customer is prognostic anomaly discovery (or certifying the absence of anomalies), the reality is that no matter how cleverly one fills in a "blind spot" in a sequence of measured observations through interpolation, it is still a "blind spot" in terms of whether some anomalous event occurred in the asset under surveillance at the times coinciding with missing values in individual sensor measurements.

MVI is a special case of inferential sensing, where individual observations that are missing during surveillance are computed using MSET in the inferential mode, exactly as is done when sensors fail. Note that whatever mechanisms cause there to be missing observations in the surveillance data will likely also be present when signals are being collected for training. This is a challenge for all ML prognostics, not only for applications involving MSET.

Specifically, during the signal preprocessing phase, the training dataset, which may contain missing observations, is divided into two halves, A and B. The missing observations in A are first replaced with conventional interpolation. A is then used to train MSET, and MSET is applied to B to "fill in" the missing values in B with MVI. B is then used to train MSET, which is now used on A to replace the prior interpolated values with MVI values.

Figure 4 illustrates an example where the blue signals are high-accuracy measured values used for the Training dataset, and the red signals are high-accuracy measured values used for the Estimation dataset in Phase I of the MVI procedure. The black observations are the randomly selected values to be removed from the data streams to create missing values. The actual values are "held back" as "ground truth" values for assessing the accuracy of the new MVI procedure.

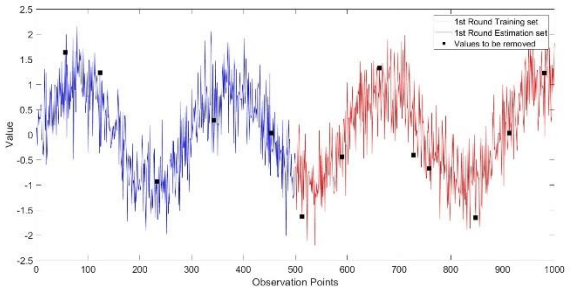

Figure 4: Training and estimation set of ground truth data for the first round MVI analysis with artificial missing values indicated.

Figure 5 "reverses" the Training and Estimation data sets, where now the temporary interpolates that were used for the training operation in Phase-I are now replaced with optimal MVI values in Phase-II.

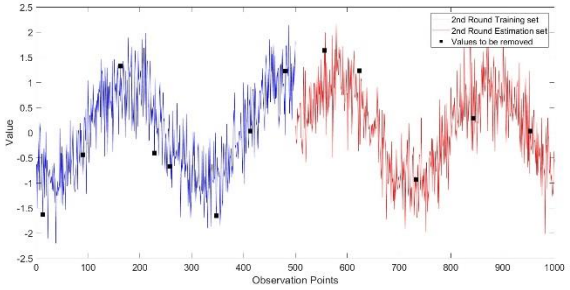

Figure 5: Training and estimation set of ground truth data for the second round of MVI analysis with artificial missing values indicated.

Figure 6 plots the original Ground Truth values in black, the Interpolation values in red, and the MSET optimal MVI values from the new 2-phase MVI data-flow framework in green.

For this case study, we noticed the average uncertainty of the new MVI approach is 0.41 while the average uncertainty of conventional interpolation is 0.73, showing that for this set of signals, the reduction in uncertainty by the new MVI approach is 44%. We have also tested this technique against very many datasets with varying degrees of cross correlation and varying

signal-to-noise ratios and from all experiments we obtained a reduction in uncertainty of from 39% to 51%.

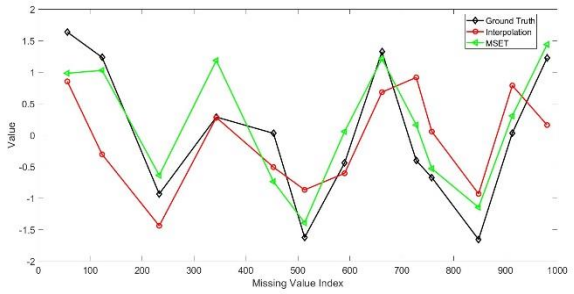

Figure 6: Validation of MVI using simple averaging interpolation and MSET against ground truth signal

The value for MVI is not just that the imputed missing values are significantly more accurate than conventional interpolation can achieve, but lies in the fact that if any degradation events occur at the precise narrow time window during which a missing value occurs, the MVI estimate will reflect the degradation condition, whereas the conventional interpolated values cannot.

III. CONCLUSIONS

In summary, the MSET system comprises a synergistic integration of the SPRT technique with a data-driven modeling method to produce a system with unique surveillance capabilities which is expected to outperform the conventional approaches, including neural networks, autoassociative kernel regression, and regularized kernel regression, in sensitivity, reliability, robustness to unreliable and possibly degrading sensors, simplicity of training, adaptability when sensor configurations change, and computational efficiency. In addition, Oracle's intelligent data preprocessing (IDP) innovations assure optimal ML performance for prognostics, streaming analytics, prognostic cyber security, and real time signal validation and sensor-operability validation across a variety of industries for which human-in-the-loop supervisory control of complex engineering assets is standard practice. MSET2 and SPRT, together with the suite of IDP algorithms that mitigate and avoid common sensor and signal anomalies that cause excessive false-alarm and missed-alarm rates in conventional Machine Learning prognostics play a vital role as an integrated and autonomous operator decision aide because as shown in this paper, the integrated system substantially reduces the probabilities of false and missed alarms that increase "cognitive overload" events for expert human operators.

REFERENCES

[1] Gross, K. C., Baclawski, K., Chan, E. S., Gawlick, D., Ghoneimy, A., & Liu, Z. H. (2017, March). "A supervisory control loop with Prognostics for human-in-the-loop decision support and control applications." ***2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*** (pp. 1-7). IEEE.

[2] Gross, K. C., Singer, R. M., Wegerich, S. W., Herzog, J. P., VanAlstine, R., & Bockhorst, F. (1997). "Application of a model-based fault detection system to nuclear plant signals." ***Proc. 9th Intnl. Conf. On Intelligent Systems Applications to Power Systems***, pp. 66-70, Seoul, Korea.

[3] Singer, R. M., Gross, K. C., Herzog, J. P., King, R. W., & Wegerich, S. (1997). "Model-based nuclear power plant monitoring and fault detection: Theoretical foundations." ***Proc. 9th Intnl. Conf. On Intelligent Systems Applications to Power Systems***, pp. 60-65, Seoul, Korea.

[4] Gross, K. C., & Li, M. (July 2017). "Method for Improved IoT Prognostics and Improved Prognostic Cyber Security for Enterprise Computing Systems." ***International Conference on Artificial Intelligence (ICAI)*** (pp. 328-334).

[5] Wang, G. C., Gross, K. C., & Subramaniam A. (December 2019). "ContainerStress: Autonomous Cloud-Node Scoping Framework for Big-Data ML Use Cases." ***IEEE 2019 Intn'l Symposium on Big Data and Data Science (CSCI-ISBD)***, Las Vegas, NV.

[6] Garvey, D. J., Hines, J. W. and Gross, K. C. (April 2007). "Real-Time Remaining Useful Life (RUL) Estimation of Computer Server Power Supplies." ***Proc. 61st Meeting of the Machinery Failure Prevention Technology Soc. (MFPT61)***, Virginia Beach, VA.

[7] Gross, K. C., & Li, D. (2018). "Machine Learning Innovation for High Accuracy Remaining Useful Life (RUL) Estimation for Critical Assets in IoT Infrastructures." ***International Conference on Internet Computing (ICOMP)***, Las Vegas, NV.

[8] Gross, K. C., Whisnant K. W., and Urmanov A. M. (Feb 2006). "Electronic Prognostics Techniques for Mission Critical Electronic Components and Subsystems." ***Proc. 2006 Components for Military and Space Electronics Symposium***, Los Angeles, CA.

[9] Gross, K. C., Whisnant K. W., and Urmanov A. M. (Feb 2006). "Prognostics of Electronic Components: Health Monitoring, Failure Prediction, Time To Failure." ***Proc. New Challenges in Aerospace Technology and Maintenance Conf***. Suntec City, Singapore.

[10] Vaidyanathan, K. and Gross, K. C. (Sept 17-19, 2004). "Proactive Detection of Software Anomalies through MSET." ***Proc. IEEE Workshop on Predictive Software Models (PSM-2004)***, Chicago.

[11] Gross, K. C., & Lu, W. (2002, June). "Early Detection of Signal and Process Anomalies in Enterprise Computing Systems." ***Proc. 2002 IEEE Int'l Conf. on Machine Learning and Applications (ICMLA)***.

[12] Gross, K. C., & Dhanekula, R. (2012, April). Multivariate SPRT for Improved Electronic Prognostics of Enterprise Computing Systems. ***Proc. 65th Meeting of the Machinery Failure Prevention Technology Society (MFPT2012)***.

[13] Masoumi, T., & Gross, K. C. (2016, December). "SimSPRT-II: Monte Carlo simulation of sequential probability ratio test algorithms for optimal prognostic performance." ***International Conference on Computational Science and Computational Intelligence (CSCI)*** (pp. 496-501). IEEE.

[14] Wang, G. C., & Gross, K. C. "Real Time Empirical Synchronization of IoT Signals for Improved AI Prognostics." ***IEEE 2018 Intn'l Symposium on Internet of Things & Internet of Everything (CSCI-ISOT)***, Las Vegas, NV (Dec 13-15, 2018).

[15] Gross, K. C., & Vaidyanathan, K. (2010, August). "Improved Energy Monitoring and Prognostics of Servers via High-Accuracy Real-Time Synchronization of Internal Telemetry Signals." ***Proc. IEEE World Congress in Computer Science, Computer Engineering, and Applied Computing (WorldComp2010)***, Las Vegas, NV.

[16] Gross, K. C., Dhanekula, R., Schuster, E., and Cumberford G. "Spectral Synthesis of Telemetry Signals to Remove Signal Quantization Effects." Case ID SUN060179, **U.S. Patent 7,248,980** (Jul 24, 2007).

[17] Zhang, F., Boring, S., Hines, J. W., Coble, J., and Gross, K. C. (Nov 2017). "Combination of Unquantization Technique and Empirical Modelling for Industrial Applications." ***2017 American Nuclear Society Intn'l Conf.***, Washington D.C.