Text Mining for Procedure-Level Primitives in Human Reliability Analysis

4 authors:

Sarah Ewing
Atos S.A.
17 PUBLICATIONS **302** CITATIONS
SEE PROFILE

Ronald Laurids Boring
Idaho National Laboratory
306 PUBLICATIONS **2,819** CITATIONS
SEE PROFILE

Martin Rasmussen Skogstad
NTNU Samfunnsforskning
60 PUBLICATIONS **838** CITATIONS
SEE PROFILE

Thomas A Ulrich
Idaho National Laboratory
126 PUBLICATIONS **666** CITATIONS
SEE PROFILE

Text Mining for Procedure-Level Primitives in Human Reliability Analysis

Sarah M. Ewing[1], Ronald L. Boring[1], Martin Rasmussen[2], Thomas Ulrich[1]

[1] Idaho National Laboratory, PO Box 1625, Idaho Falls, ID 83415
{sarah.ewing, ronald.boring, thomas.ulrich}@inl.gov
[2] NTNU Social Research, Studio Apertura, Dragvoll Allé 38 B, 7491 Trondheim, Norway
martin.rasmussen@svt.ntnu.no

Abstract. The classification of nuclear power plant procedures at the sub-task level can be accomplished via text mining. This method can inform dynamic human reliability calculations without manual coding. Several approaches to text classification are considered with results provided. When a discrete discriminant analysis is applied to the text, this results in clear identification procedure primitive greater than 88% of the time. Other analysis methods considered are Euclidian difference, principal component analysis, and single value decomposition. The text mining approach automatically decomposes procedure steps as Procedure Level Primitives, which are mapped to task level primitives in the Goals, Operation, Methods, and Section Rules (GOMS) human reliability analysis (HRA) method. The GOMS-HRA method is used as the basis for estimating operator timing and error probability. This approach also provides a tool that may be incorporated in dynamic HRA methods such as the Human Unimodel for Nuclear Technology to Enhance Reliability (HUNTER) framework.

Keywords: Human Reliability Analysis · Computation-Based Human Reliability Analysis · Human Error · GOMS-HRA · Text Mining

1 Introduction

The quantification of nuclear power plant (NPP) anomalous events as a probability over time is called dynamic probability risk assessment (PRA). The use of PRA in NPPs has become commonplace in NPPs, with quantification methods implemented throughout the entire U.S. NPP fleet. Examples of PRA methodology implemented by regulators include the Systems Analysis Programs for Hands-on Integrated Reliability Evaluations (SAPHIRE) and the Standardized Plant Analysis Risk (SPAR) models. However, the human component in each NPP is difficult to quantify due to commission and omission errors. Closer inspection of the NPP operation manuals that are implemented can give real-time quantitative information on human behavior, with insights into the specific human actions that need to take place in order for an NPP to operate.

The classification of NPP procedures can be accomplished via the use of text mining capabilities. This method can then be used to inform dynamic human reliability calculations without the need for tedious manual coding or analysis. This approach includes an objective assessment of the procedures based on previous expert assessments conducted by analysts. Providing an initial objective assessment allows experts to identify anomalies in the algorithms and contribute to an objective result that will provide consistent outcomes.

The application of a Goals, Operation, Methods, and Section Rules (GOMS) model as applied to NPP operator actions is detailed in [1]. And the subsequent application to NPP operation manuals and association to timing data to complete steps are detailed in [2]. In this exploration, the procedures were taken from NPP control room manuals, and as such only GOMS that can be associated with main control room actions were considered. A list of the GOMS procedures as detailed in [1] and [2] is provided in Table 1. The association of GOMS, automatically, can be created through a text mining framework.

Table 1. A list of GOMS primitives as defined by [1] and [2]. GOMS primatives considered are indicated with **.

Primitive	Description	
Ac	Performing required physical actions on the control boards	**
Af	Performing required physical actions in the field	
Cc	Looking for required information on the control boards	**
Cf	Looking for required information in the field	
Rc	Obtaining required information on the control boards	**
Rf	Obtaining required information in the field	
Ip	Producing verbal or written instructions	**
Ir	Receiving verbal or written instructions	**
Sc	Selecting or setting a value on the control boards	**
Sf	Selecting or setting a value in the field	
Dp	Making a decision based on procedures	**
Dw	Making a decision without available procedures	
W	Waiting	

2 Methods

Data mining is the extraction of meaningful patterns and information from large amounts of data. In the same respect, text mining refers to the process of defining intriguing and relevant conclusions from text [3]. The application of text mining was applied to NPP control room procedures so that a better understanding of the ‘procedure’ performance shaping factor can be achieved. Seven procedural manuals were acquired from a U.S. NPP [4 - 10]. The text was captured out of portable document format (PDF) files using the suite of Microsoft products, R 3.2.2 and SAS 9.3 [11] [12].

After the text was pulled from the PDF files, it was formatted into four different levels. These levels are defined by expert HRA analysts and will be referred to as a Levels 1 through 4; an example is provided in Table 2. For the purpose of analysis, the procedure manual was analyzed at the fourth level, because this is where most of the control room instructions are clearly defined. Additionally, the fourth level is the level at which GOMS-HRA most naturally translates. The seven different operation manuals contained more than 2,100 fourth-level procedures. Table 2 is an example of the differing levels as defined in the NPP procedural manual regarding the main turbine.

Table 2. An example of the levels of actions defined in the procedural manual for the NPP main turbine.

	Levels			
Procedural Manual Text	**Level 1**	**Level 2**	**Level 3**	**Level 4**
Instructions	6			
Main turbine startup	6	6.1		
Prerequisites	6	6.1	6.1.1	
The feedwater system is in service per feedwater and condensate.	6	6.1	6.1.1	6.1.1.1
The main turbine lube oil system is in service per main turbine lube oil system.	6	6.1	6.1.1	6.1.1.2
The generator seal oil system is in service per generator seal oil system	6	6.1	6.1.1	6.1.1.3
The main generator is filled with hydrogen per generator hydrogen.	6	6.1	6.1.1	6.1.1.4
The stator cooling system is in service per stator cooling system.	6	6.1	6.1.1	6.1.1.5
The stator cooling water system trips have been reset per stator cooling system.	6	6.1	6.1.1	6.1.1.6

These procedures were then altered into a format that is easier to text mine. This is completed via the removal of stop words; these are typically conjunctive words that have little meaning when content is analyzed (e.g., “and,” “but,” “or,” and “with”). Then each non-conjunctive word in the manual had the suffix removed; this is called stemming. This is completed so that similar words would be counted as the exact same word. For example, “charge,” “charging,” “charged,” and “charges” would all be defined as differed words before stemming is completed; after stemming, they are all “charg-”. In addition to implementing stemming and stop word removal, all punctuation and numbers are removed as software identifies RCP’s, “RCPs”, and RCPs as different from one another when no content difference exists. An example of stop word removal, stemming, and punctuation removal on a procedural step can be seen in Table 3.

Once the text has been prepared, a text matrix is generated that identifies the number of times a word stem is in a subsection. The seven procedural manuals produced more than 2,000 unique word stems. A bag-of-words approach was taken such that the context of each word was ignored, except in special cases. One such case was due to the frequency use of the term charging pump; it was analyzed as "chargingpump." The context of the word stems is integral, because two different words can mean the same thing (synonymy) and the same word can have two or more meanings in different contexts (polysemy). While this realization exits, it is difficult to quantitatively capture this information. An example of a text matrix with five word stems can be seen in Table 4.

Table 3. An example of stemming, stop word removal, and deletion of numbers and punctuation performed on a procedural manual step.

Before	IF BOTH of the following occur at any time during this procedure: • Pressurizer level lowers to 33% • Restoration of charging is NOT Impending THEN trip the reactor. **NOTE:** Multiple indications and SM/CRS discretion should be applied to diagnosing Charging Pump gas binding.
After	follow occur any tim procedur pressur level lower restor charg impend trip reactor multipl indic smcrs discret appli diagnos charging pump gas bind

Table 4. A text matrix with the original procedure and formatted procedure, along with a selection of five stem words and their respective counts.

Original Procedure with Punctuation	**Procedure Formatted**	**Text Matrix**				
		action	**charg**	**chargingpump**	**chbhs523**	**Close**
IF BOTH of the following occur at any time during this procedure • Pressurizer level lowers to 33% • Restoration of charging is NOT Impending THEN trip the reactor. **NOTE**: Multiple indications and SM/CRS discretion should be applied to diagnosing Charging Pump gas binding.	follow occur any tim procedur pressur level lower restor charg impend trip reactor multipl indic smcrs discret appli diagnos chargingpump gas bind	0	1	1	0	0
IF Charging Pump gas binding is indicated by ANY of the following: • Charging header flow fluctuations • Charging header pressure fluctuations • Charging header flow less than expected for running charging pumps • Charging suction source (VCT, RWT) level lost THEN perform Appendix G, Responding to Gas Binding of Charging Pumps.	chargingpump gas bind indic follow charg header flow fluctuat charg header pressur fluctuat charg header flow less expect run chargingpump charg suction sourc vct rwt level lost perform appendixg	1	4	2	0	0

Additionally, while the context of word stems was not able to be captured, parts of speech were captured (i.e., noun, verb, and adjective). This was conducted with a hybrid of natural language processing algorithms and excerpts of tables from the Professional Procedure Association's manual [13]. The context of the word was briefly considered as an analytical approach but was not retained due to inaccuracies and time constraints. All of the above techniques were applied to the analysis of the seven procedural manuals, which had more than 2,000 unique word stems for more than 2,100 Level 4 procedures. Thus, more than 4,200,000 observations were considered in matrix form. The most frequent word stems in the 2,100 fourth-level procedures are provided in Fig. 1.

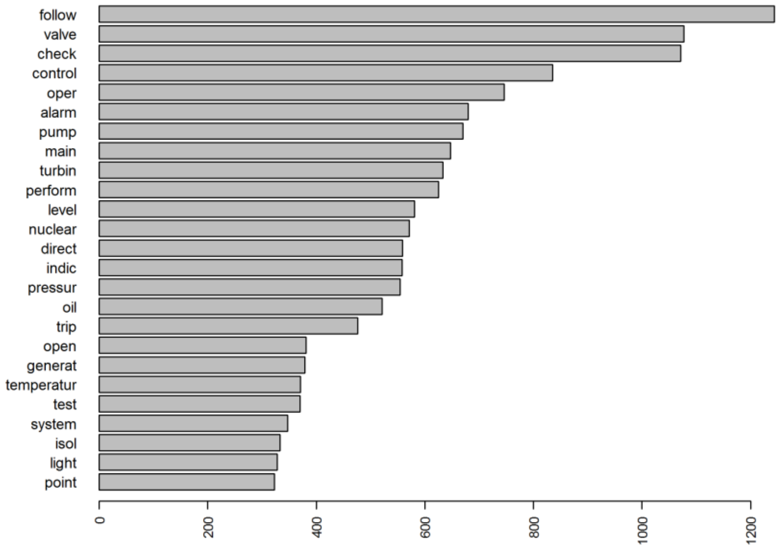

Fig. 1. Bar chart of the top 25 occurring word stems in the seven manuals for fourth-level procedures.

3 Analysis and Results

There are many analysis methods that can be implemented on a text matrix. Due to the large data nature of this analysis, reduction of dimensions or noise is desired. Some methods considered include principal component analysis (PCA), ridge regression, single value decomposition (SVD), and expert judgment [14] [15]. Then analytical methods were further implemented to the text matrix codex to define the GOMS primitives. While all these methods were explored, only the details of PCA, SVD, and expert judgment are detailed herein. To provide meaningful results, a randomly selected subset of 148 of the 2,100 procedures was mapped to GOMS; this created a codex upon which meaningful conclusions can be mapped. The top-occurring word stems are provided in Fig. 2. As such, the analytical methods are applied to the subset of 148 procedures. For the methods to be confirmed as more generalizable, a larger codex needs to be considered.

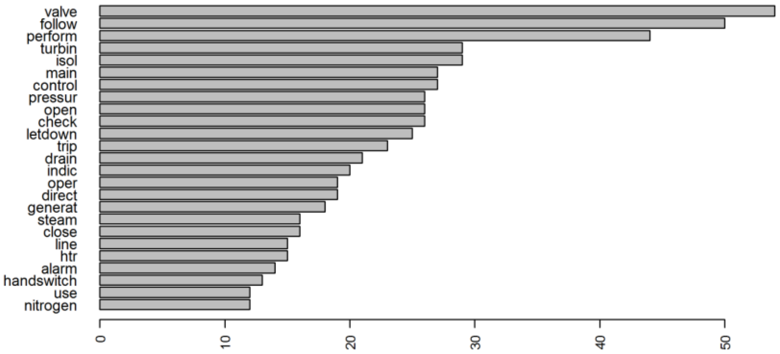

Fig. 2. Bar chart of the top 25 occurring word stems in the GOMS codex of procedures.

3.1 Dimension and Noise Reduction

Principal Component Analysis. PCA uses a text matrix of the words to create linear combinations of word stems. These new variables are called Eigen vectors and are orthogonal to one another. The number of Eigen vectors created is equal to the number of variables, or word stems, that are in the initial text matrix. With 33 Eigen vectors, 90% of the variance is explained. A way to visualize the first two Eigen vectors that explain the most variation is provided in a bi-plot in Fig. **3.**

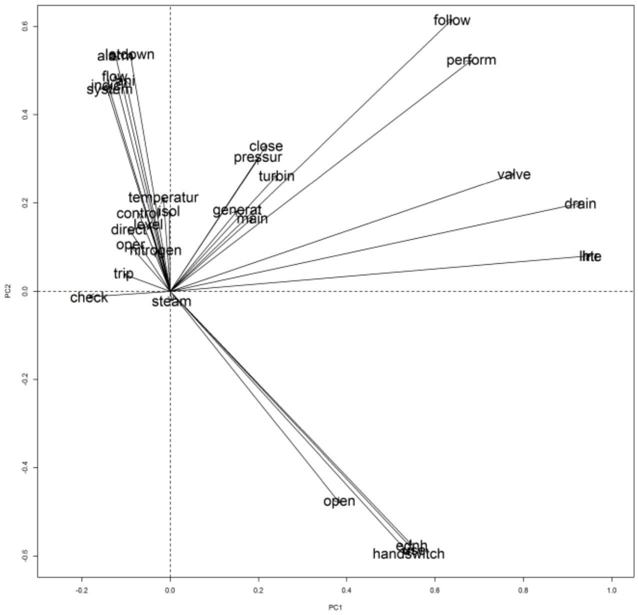

Fig. 3. A PCA bi-plot of the first two Eigen vectors with only the top 30 word stems considered.

The word stems have meaning based on their angle to one another (Fig. 3). The arrows in the figure are at different angles to one another, indicating the level of correlation. When the arrows are at 90 degrees, this indicates orthogonality, or a lack of correlation between word stems. And parallel arrows are considered to be highly correlated. Arrows at 180 degrees from one another are inversely related. Based on this, words like “follow” and “perform” are considered essentially parallel. “Check” and “drain” are 180 degrees from each other, indicating an inverse relationship.

While this method provides informative descriptive statistics and dimension reduction, identifying the stems that are strongly correlated with the GOMS primitives is not straightforward in this form. Thus, other methods were considered for auto calculating GOMS primitives to NPP procedures.

Single Value Decomposition. SVD is a statistical method to reduce the noise of irrelevant variables. SVD describes data by reducing the sum of the difference between the text matrix vectors, the details of which are described in [15]. The positive aspect is that SVD does not overrate the similarity between two words, in contrast to the PCA approach. Unlike other methods, SVD does not automatically toss out highly correlated word stems. However, the output for SVD is similar to that of PCA and as such was not utilized as a method to reduce dimensions in GOMS.

Expert Opinion. While the previously defined methods for noise reduction in the number of word stems may be more descriptive, their implications are not always straightforward. As such, expert opinion was employed that involved dropping all the word stems that had three or less occurrences. Three was decided upon because it was the median number of occurrences of word stems in the codex. This resulted in 84 word stems remaining in the text matrix. Further dimension and noise reduction was completed using analytical techniques, with unique results being applied to each GOMS primitive type.

3.2 Analysis Methods

The results of the analysis provide word stems that are strongly correlated with GOMS primitives. The methods considered include naive Bayes, random forest, logistic regression, heat map algorithms, Euclidian hierarchical clustering (EHC), correlation networks, and Bayesian discrete discriminant (BDD) analysis. Details from EHC, correlation network, and BDD are provided below.

Euclidian Hierarchical Clustering. The first step to EHC is to calculate the distance matrix by the Euclidian method. The distance matrix provides the distance between two vectors such that it is implemented between the rows of the text matrix [14].

Once the distance between rows is computed, the resulting matrix is considered a matrix of dissimilarities. The distance, or dissimilarity, matrix does not necessarily calculate the literal distance between words and is a way to empirically evaluate the data. The rows of our text matrix are the stem word, so when the dissimilarity matrix is calculated, it is calculating the difference of the procedures based on the frequency of the word stems. This matrix can be represented graphically as a dendrogram, as seen in Fig. 4.

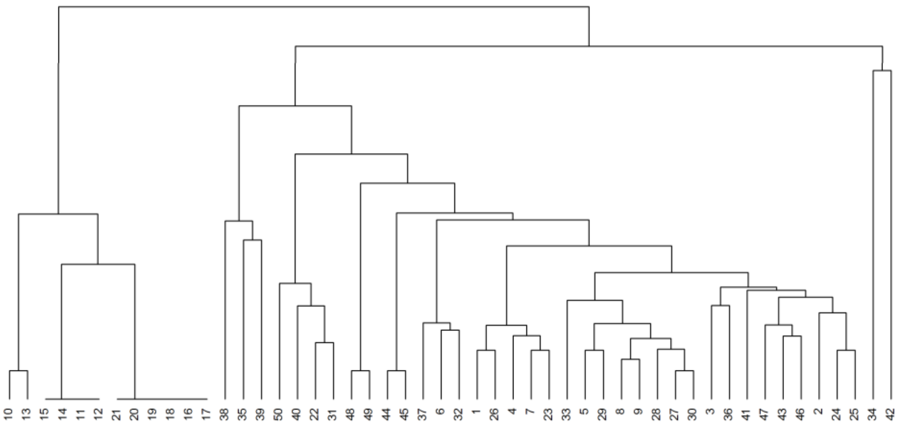

Fig. 4. A Euclidian single cluster dendrogram on the NPP procedures, where the numbers at the bottom are NPP procedures in the codex.

The numbers at the bottom of Fig. 4 are the identification numbers associated with the procedures in the codex. A hierarchical cluster analysis is applied to the dissimilarity matrix for n clusters, where n is defined subjectively by the expert. Based on data configuration, the number of clusters selected is seven, corresponding the number of GOMS that are being investigated. This is then examined against the GOMS groups, which resulted in 11% accuracy. As such, further methods were considered for defining the GOMS type.

Correlation Network. When investigating the dependence between multiple variables, a correlation matrix can be constructed. In this case, the correlation between procedures is being evaluated. The result is a matrix containing the correlation coefficients between each of the procedures. While a matrix contains a lot of information, visualization of that data can be difficult and chaotic. Thus, a network was constructed to better visualize the correlation relationships between the stem words, as in Fig. 5.

The thickness of the lines between the stem word nodes denotes the strength of the correlation. In addition to line thickness, the colors of the lines indicate if the correlation is positive (black) or negative (grey). Oddly enough, there are no strong negative correlations, or thick grey lines, whereas there is a strong positive relationship between clumps of procedures. These clumps may lend themselves to mapping to the GOMS primitives; however, there only appear to be 4 or 5 clumps at most, while seven GOMS primitives are defined in the codex. As such, another method to define the GOMS primitives was explored.

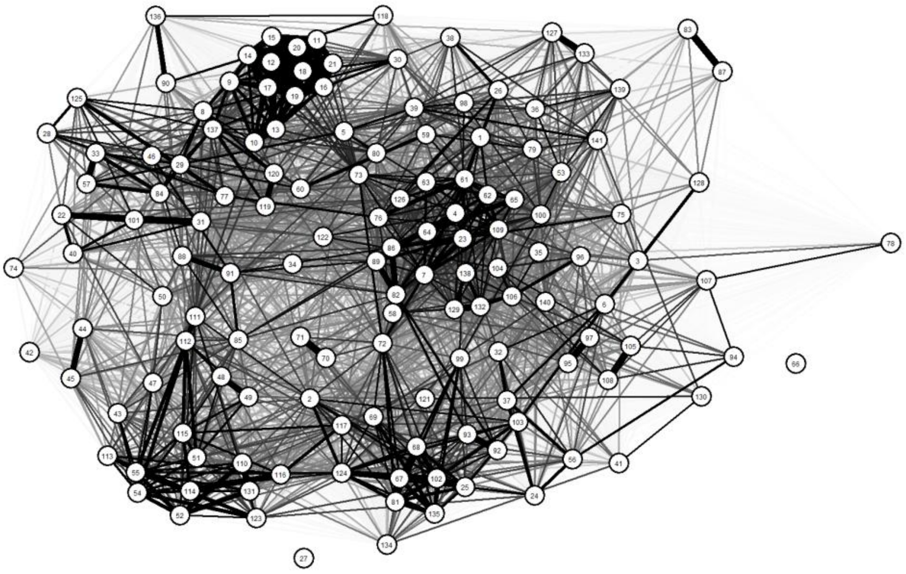

Fig. 5. Network of the word stems based on the correlation matrix. Black indicates a positive correlation, and grey a negative. The nodes, or circles, are the procedures in the codex.

Discrete Discriminant Analysis. BDD is implemented with the assumption that the frequency of GOMS primitives in the codex is representative of all NPP manuals. Initially, a discrete multinomial distribution of GOMS primitives was assumed; however, this produced low accuracy. Thus, each GOMS primitive was dummy coded and assessed individually, which is in line with expert opinion; the details of discrete discriminant analysis are provided in [16]. Each procedure in an NPP manual may be composed of multiple primitives. A binominal BDD for each GOMS primitive lends itself to identification of multiple GOMS per a procedure.

To further reduce the word stems utilized, stepwise selection based on an Akaike information criterion was applied [17] [18]. Then an algorithm to fit all possible discrete discriminant analysis combinations was executed, with the best performing model defined based on the lowest Akaike information criterion value. The resulting word stems were retained; the accuracy for each GOMS is provided in Table 5. Due to Ir and Sc having such a low frequency in the codex, any results for Ir and Sc are not considered accurate and are not presented.

Table 5. Results from the discrete discriminant analysis for each GOMS primative. The frequency of example procedures are provided along with the analysis accurecy and the word stems that were included in the model.

GOMS Primitive	Frequency	Prediction Accuracy	Discriminant Analysis Results
Ac	30	95%	cool exist manual trbl leak regen ensur high output refer bottl place air test ani level handswitch alarm close trip letdown check control turbin perform valve
Cc	45	88%	instal low speed gate initi leak run output bottl place action flow system level handswitch close direct trip letdown pressur isol turbin follow valve
Rc	26	94%	cool cooldown greater instal low gate suppli breaker reactor section flow ani steam generat direct drain trip letdown check pressur
Ip	18	95%	enter smcrs mainten regen auxiliary direct pressur turbin
Ir	5	100%	NOT ACCURATE
Sc	2	94%	NOT ACCURATE
Dp	15	98%	speed leak lpturbin mainten end loss output rcs refer breaker place section servic ani perform follow

4 Results and Conclusions

Text mining, as applied to NPP control room operation manuals, provides a lot of descriptive statistics that can better inform the future development of manuals and error quantification methods. The number of unique word stems, more than 2,000, in NPP control room operations manuals is relatively low compared to other invocations of the English language in everyday life. Experts have suggested that this is because NPP manuals need to be easily understood, even in situations of extreme stress and when English is a second language. Many other interesting findings may still come to light from these documents that will give unique insights to NPP control room interworkings.

Many dimension-reduction methods were employed with the final technique executed, including expert opinion, stepwise selection, and creation of all possible models. Analysis methods for identification of the GOMS primitives to the procedures are accomplished by associating multiple GOMS to a procedure. While the examination only considered the mapping of the one GOMS to procedures, applying a BDD analysis is highly effective with all models, indicating 88% or greater accuracy. To have more accurate results, more examples of GOMS primitive mappings need to be provided so that more generalizable results can be obtained that apply to more than just seven NPP operation manuals.

The highly accurate automation of typing NPP procedures into multiple GOMS primitives is a step toward creating a dynamic framework that can calculate a realistic human error probability in real time. This real-time assessment will be based on the procedures that control-room and field operators implements. Further quantitative research needs to be completed describing the event trees and the other associated performance shaping factors that will have an impact on a control room operator's ability to complete tasks.

Acknowledgments. Every effort has been made to ensure the accuracy of the findings and conclusions in this paper, and any errors reside solely with the authors. This work of authorship was prepared as an account of work sponsored by Idaho National Laboratory, an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Idaho National Laboratory is a multi-program laboratory operated by Battelle Energy Alliance LLC for the United States Department of Energy under Contract DE-AC07-05ID14517.

5 Works Cited

1. R. Boring, M. Rasmussen, T. Ulrich, S. Ewing and D. Mandelli. "Task and Procedure Level Primitives for Modeling Human Error," Proceedings of the 8th Applied Human Factors and Ergonomics, Los Angeles, In Press 2017.
2. T. Ulrich, R. Boring, S. Ewing, M. Rasmussen and D. Mandelli, "Operator Timing of Task Level Primitives for Use in Computation-Based Human Reliability Analysis," Proceedings of the 8th Applied Human Factors and Ergonomics, Los Angeles, In Press 2017.
3. V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence,* pp. 60-76, 2009.
4. U.S. NPP Generating Station, "EXCESSIVE RCS LEAKRATE."
5. U.S. NPP Nuclear Generating Station, "STANDARD POST TRIP ACTIONS."
6. U.S. NPP Nuclear Generating Station, "The Steam Generator Tube Rupture."
7. U.S. NPP Nuclear Generating Station, "Panel 6 Alarm Responses."
8. U.S. NPP Nuclear Generating Station, "Panel 7 Alarm Responses."
9. U.S. NPP Nuclear Generating Station, "Loss of Charging or Letdown."
10. U.S. NPP Nuclear Generating Station, "Main Turbine."
11. R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2016.
12. SAS Institute Inc, "Base SAS® 9.3 Procedures Guide," SAS Institute Inc, Cary, NC, 2011.
13. Procedure Professionals Association, "Procedure Writer's Manual PPA AP-907-005 R2," 2016.
14. F. Murtagh, "Multivariate Data Analysis with Fortran, C and Java Code," Queen's University Belfast, and Astronomical Observatory Strasbourg, Belfast.
15. R. Albright, "Taming Text with the SVD," in *SAS Institute Inc.*, Cary, NC, 2004.
16. J. D. Knoke, "Discriminant Analysis with Discrete and Continuous Variables," *Biometrics,* vol. 38, no. 1, pp. 191-200, 1982.
17. R. R. Hocking, "A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression," *Biometrics,* vol. 32, no. 1, pp. 1-49, 1976.
18. D. J. Beal, "Information criteria methods in SAS for multiple linear regression models," in *15th Annual South East SAS Users Group (SESUG) Proceedings*, South Carolina, 2007.