INL/CON-24-76335-Revision-0

Methods and Tools To Assess Robustness of Nuclear Plant Outages

September 2025

Changing the World's Energy Future

Diego Mandelli

Idaho National Laboratory

INL/CON-24-76335-Revision-0

Methods and Tools To Assess Robustness of Nuclear Plant Outages

Diego Mandelli

September 2025

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Methods and Tools To Assess Robustness of Nuclear Plant Outages

D. Mandelli*, **A. Al Rashdan, C. Wang, S. St. Germain,**
S. Lawrence, N. Mapes, J. Cogliati

Idaho National Laboratory, Idaho Falls, ID

[leave space for DOI, which will be inserted by ANS – do not highlight this line]

ABSTRACT

Refueling outages are one of the most challenging phases in a nuclear power plant (NPP) operating cycle. They are extremely costly for an NPP due to the large amount of required resources and lost revenue from the plant being off the grid. Outage durations have steadily decreased across the industry over the last few decades primarily due to improved planning and coordination, but there are still many plants that struggle to meet the performance metrics accomplished by other utilities. Schedule resilience is one of the issues. NPP outages require scheduling thousands of activities within an average of 30 days. Despite detailed planning, once the outage starts, numerous emergent issues typically appear along with schedule delays, requiring continuous replanning and adjustment. When a schedule disruption occurs during an outage, plant staff make urgent efforts to recover but are often not able to maintain the planned outage duration. These outage delays can cost a utility several millions of dollars per day. Tools that could help outage schedulers create a more resilient schedule and allow them to optimally reschedule emergent work could significantly reduce outage delays. One key aspect of creating a resilient schedule is to have accurate estimates of activity duration. Another important outage scheduling capability is the ability to schedule emergent work with minimal disruption. This paper focuses on developing tools and methods to support NPPs with outage schedule optimization and describes the initial development of tools to support outage management that leverage computational and machine learning methods.

Keywords: outage, NLP, robustness

1. INTRODUCTION

Nuclear power plant (NPP) refueling outages represent one of the more challenging aspects of managing a light-water reactor facility. Refueling outages typically require completing over 10,000 scheduled activities representing several thousand work orders in around 30 days. Most utilities use numerous contract workers to support outage activities, adding to the complexity and cost. Since the reactor is taken offline to perform this work, the utility is not generating revenue during the outage. The cost of lost revenue can be up to $2 million per day.

To complete outages as efficiently as possible, planning begins more than a year before the start of the outage. Significant effort is placed on creating a schedule that minimizes outage duration with all the required work to be completed. In addition to the known work being scheduled for the outage, there may be several hundred emergent activities added to the schedule during the outage. While the outage managers have nearly a year to create the initial schedule, they may only have a few hours to reschedule each day. This difficulty in rescheduling is one of the reasons outage durations almost always exceed the original

* Corresponding author: diego.mandelli@inl.gov

plan. When a refueling outage has a significant delay, the cost is substantially increased due to the necessity for the contract workers to stay onsite longer.

The objective of this research project is to develop methods and tools to support plant staff in creating an outage schedule that has a high probability of completion in the desired timeframe. To improve on the current critical path (CP) methodology, we are investigating a method to calculate schedule resilience by creating a model of the schedule. A resilient schedule is one that has been analyzed and adapted to account for the duration uncertainty capable of reorganizing activities to better absorb duration variability. A resilient schedule also presents margins for highly uncertain, non-CP activities. Last, a resilient schedule should have the capability to absorb the expected amount of scope growth without a significant disruption to the planned duration.

In order to effectively model schedule resilience, we need activity duration uncertainty information along with the usual planned activity duration currently used in the CP methodology. In this case, each activity is assigned a duration distribution rather than a simple duration estimate. Various machine learning and artificial intelligence (ML/AI) and natural language processing (NLP) methods are investigated to automatically assign activity duration distributions based on the analysis of historical outage performance data. In cases where data are not sufficient to assign a duration distribution, a schema will be developed of a standard distribution of the assigned duration based on generic average completion time distributions for common types of work activities, such as valve refurbishment, erecting scaffolding, circuit breaker refurbishment, etc.

2. ASSESS ACTIVITY DURATION VARIANCE

2.1. NLP Methods for Duration Prediction

This effort used two datasets from nuclear facilities. Both datasets contained a brief activity description, usually of less than ten words, and timestamps of when the activity was predicted to start and end and when it actually started and ended. Each activity had an alphanumeric code that seemed to follow a certain undefined structure. Given the sparsity of the activity description text, it is desired to connect each activity to the actual work scope performed. However, there was no connection established between an activity and the corresponding work order, and the outage work management data were not available. This was the case for both datasets. This section describes the performed evaluation determining whether it is possible to predict activity durations from the short activity description. This evaluation was conducted using two approaches: a natural language processing approach and a semantic text-mining approach.

Using the provided dataset, this section describes an exploration of a direct correlation between word occurrence in the activity description and the activity's predicted duration. This is followed by establishing the correlation between word occurrence in the activity description and activity actual duration. To examine how well the scheduler is predicting the activity durations, a plot is generated to correlate actual times needed to complete the job versus the predicted times in the schedule. Three different correlation methods were used as shown in Table 1, and each method was compared to a random data correlation. Pearson's r [1] is a normalized covariance between the planned hours and actual hours taken to complete the activity. If the predicted and actual hours are centered (mean of zero), the prediction multiplied by the actual hours for each datapoint averaged is covariance. To normalize covariance, it is divided by both the standard deviation of the predicted and actual hours, which gives Pearson's r. The correlation coefficient found using Pearson's r coefficient is 72.8%. Details on the other two methods can be found in [2],[3]. Spearman's Rho is based on the monotonic correlation of the two variables (i.e., if one increases, the other increases as well). Kendall's Tau is similar but differs in the way the monotonic behavior is mathematically described.

Table 1. Summary of the correlation results using three different correlation methods.

Correlation Type	Original Forecast	Machine Learning Forecast
Log10 Transformed Data Pearson's r	72.8%	77.1%
Spearman's Rho (transformed or untransformed)	67.7%	73.5%
Kendall's Tau (transformed or untransformed)	55.5%	57.5%
Random Data (r, Rho or Tau)	0%	0%

The data are also shown in Figure 1 (left). The number of datapoints used for this evaluation was approximately 1,000. The red error margins represent the margin for where 95% of the data are located and is calculated as: red upper/lower margin = mean actual value ± 1.96 of the error standard deviation. The 95% uncertainty is on the order of 10 hours to over 100 hours. This is especially visible for the activities around the typical mean time of an activity. From this analysis, we can conclude that either the scheduler is unable to accurately predict the activity duration or that the plant staff are not logging the actual completion times, causing this very large discrepancy.

The hypothesis tested here is whether a machine can perform better in predicting activity durations. A machine learning model was developed using a fine-tuned Sentence-BERT (SBERT) transformer [4] on the activity description using a masked language model. A masked language model has the task of deleting words from a task description in the input to the neural network and then trying to predict the correct words that were deleted. This is a type of autoencoder and is therefore unsupervised.

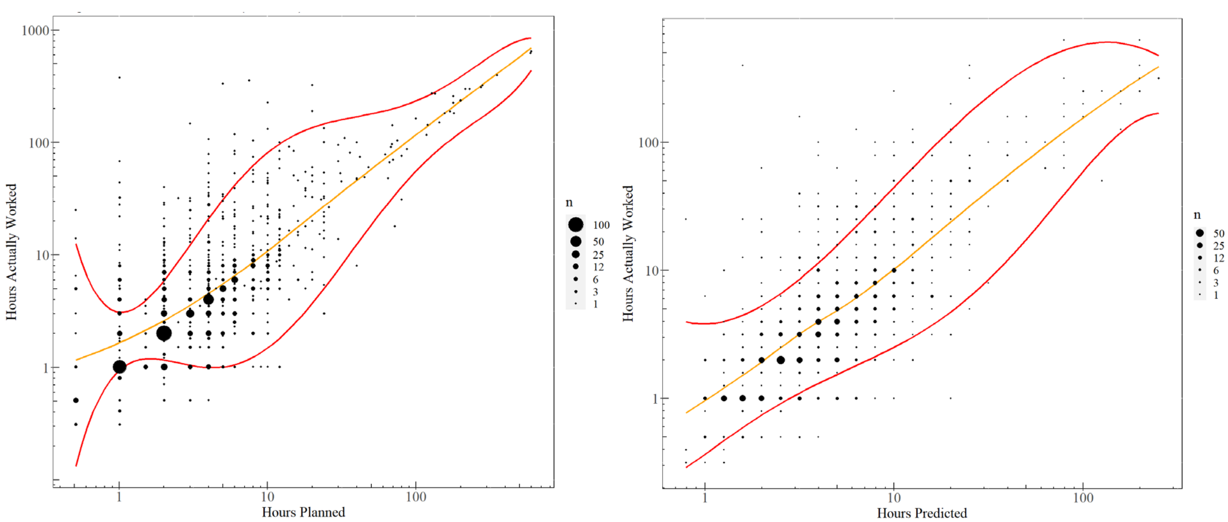

Figure 1. Plot of work hours forecasted by the scheduler (left) and by machine learning (right) versus the actual hours worked.

Next, the supervised task of predicting activity durations using a regressor is performed. The previously fine-tuned transformer neural network would output 7,168 dimensional embeddings. These embeddings were used to train a CatBoost regressor [5] on 80% of embeddings, then the CatBoost model forecasted 20% of the embeddings. The number of datapoints used for this evaluation was also approximately 1,000. All activities with fewer than 0.25 hours worked were discarded. The prediction results are shown in Table 1. The data are shown in Figure 1 (right). The correlation coefficient is 0.77, indicating more consistent results. However, given the log scale use, this also represents 10 hours to over 100 hours of uncertainty. Therefore, it is apparent that a machine cannot predict the activity duration given the short activity description.

To further compare the performance of the scheduler versus the machine, the differences between scheduler-predicted planned and actual activity durations are plotted against the difference of the machine-predicted durations. This is only performed for 20% of the data since the remaining 80% of the data were used for training the CatBoost + SBERT transformer algorithm. The error statistics of both the planned and scheduled and predicted durations show that a machine tends to predict higher values, while the scheduler tends to predict smaller values. Machine predictions resulted in a smoother prediction since it was able to predict fractions of hours, unlike the scheduler, often in 0.25–1 hour increments.

2.2. Text Similarity Search for Activity Duration Assessment

A parallel research direction to the one shown in Section 2.1 was to evaluate the completion time of an activity using text semantic similarity. The basic idea is to identify the subset of activities performed in previous outages that are similar to the activity being queried. Then, the temporal distribution of the queried activity can be determined by collecting the historical completion time from the subset of past activities.

An example of textual similarity is shown in Figure 2 where two activities with similar semantic meanings are compared that brings up the importance of data cleaning and curation. The example provided in Figure 2 suggests that, if we were to perform a simple word-to-word similarity between those two activities, they would be very dissimilar. On the other hand, if the historical activity were to be cleaned (e.g., through spell checking, and abbreviation identification and expansion), it would be transformed into "[ACC01-B] PRESSURE TRANSMITTER CALIBRATION." Consequently, the two activities would be very similar. The elements required for the semantic similarity analysis are:

- The set of past outage activities. This set might be partitioned on several datasets, one for each outage. The outage of different plant units, different plants, or different utilities can be gathered to improve analysis results.
- A computational method designed to compute the semantic similarity between two activities (i.e., the queried and historic activity) would generate a point value that measures "how similar" the two activities are. An important note here is that the computational time for such a method needs to be very small since the similarity search for a queried activity in a database of tens of thousands of past activities needs to be performed within minutes.

In this project, we focused on developing the semantic similarity method and testing this method on several outage databases. The following sections provide details of the development and present a high-level overview (to mask proprietary data) of the obtained results.

Figure 2. Example of semantic similarity between a queried and historical outage activity.

Word, sentence, and document similarity analyses are an active part of recent NLP method development, and these analyses play a crucial role in text analytics, such as text summarization and representation, text categorization, and knowledge discovery. There is a wide variety of methodologies that have been proposed during the past two decades. Mainly, these techniques can be classified into five groups: lexical knowledge base approach, statistical corpus approach (word co-occurrence), ML and deep learning approach, sentence-structure-based approach, and hybrid approach. However, there are several common major drawbacks for these approaches: computationally inefficient and lack of automation, adaptability, and flexibility. In this research, we are trying to address these drawbacks by developing a tool that can be used generally in any application requiring a similarity analysis.

As shown in Figure 3, we are trying to leverage parts of speech (POS), disambiguation, lexical database, domain corpus, word embedding and vector similarity, sentence word order, and sentence semantic analyses to calculate sentence similarity. POS is used to parse a sentence and tag each word and token with a POS tag and syntactic dependency tag. This information provides syntactic structure information (i.e., negation, conjecture, and syntactic dependency) about the sentence that can be used to guide the similarity measuring process. The disambiguation approach is employed to determine the best sense of the word, especially when coupled with a specific domain corpus. It will ensure the right meaning of the words (e.g., the right word synsets in the lexical database) among the sentence for comparison.

Then, a predefined word hierarchy from lexical database (i.e., WordNet) is used to compute the word similarity. However, some words are not contained in the lexical database since it only connects four types of POS—nouns, verbs, adjectives, and adverbs. Moreover, these words are grouped separately and do not have interconnections. For instance, nouns and verbs are not interlinked (i.e., the similarity score between "calibration" and "calibrate" is 0.091 when using WordNet). In this case, ML-based word embedding is introduced to enhance the similarity calculation. For the "calibration" and "calibrate" example, the similarity score becomes 0.715 instead. The next step is to compute sentence similarity by leveraging both sentence semantic information and syntactic structure. The semantic vectors are constructed using the previously introduced word similarity approach, while the syntactic similarity is measured by word order similarity.

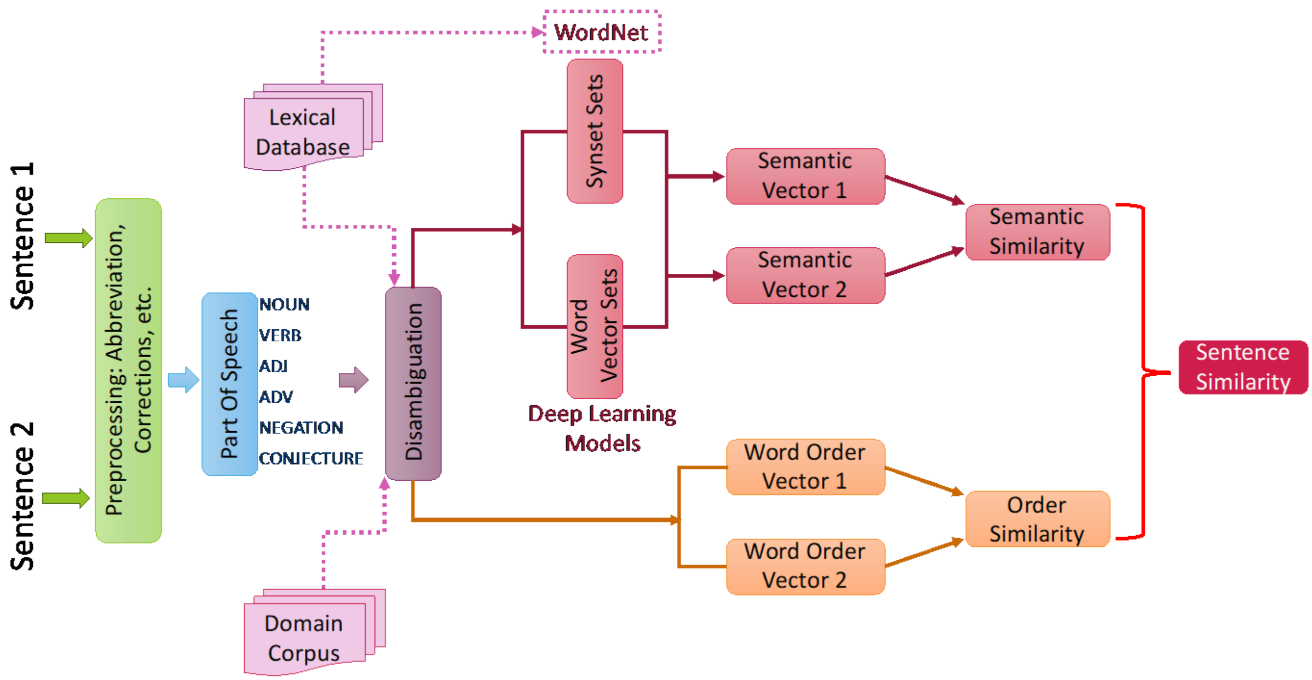

Figure 3. Illustration of sentence similarity calculation.

We want to highlight here the importance of data curation (e.g., cleaning and reconstruction) of the textual elements that describe each activity and how it might impact the search for similar activities. More specifically, the process of data curation performed for all historical outage activities includes the following steps:

1. *Remove component IDs*. The presence of specific asset or system IDs (e.g., Accumulator ACC-01B in Figure 2) does not necessarily provide any type of information from a semantic point of view and, hence, they can be removed from the actual text. This can be accomplished by either parsing the activity text or providing to the algorithm a list containing the full list of plant asset or system IDs. During our

testing, such a list was not available, and we relied on an empirical method designed to remove all words containing a mixture of characters, numbers, and symbols.

2. *Handle abbreviations*. NPP outage activities are usually short sentences that often contain abbreviations. The presence of abbreviations negatively impacts the ability to extract knowledge from such texts. Hence, we have developed an NLP pipeline designed to identify abbreviations and replace them with their corresponding complete words. The starting point is a library of abbreviations that have been collected from documents available online. This library is basically a dictionary that relates an identified abbreviation to the corresponding set of words. A challenge here is that a single abbreviation might have multiple words associated with it. Similarly, a word might have multiple ways to be reduced. Abbreviations are handled in each sentence by first identifying misspelled words. Then, each misspelled word is searched in the developed library. If an abbreviation in the library matches the misspelled word, it is replaced by the corresponding complete word. If no abbreviation in the library matches the word, we proceed by searching for the closest abbreviation. If multiple words match the obtained abbreviation, the word that fits most of the sentence context is chosen.
3. *Spellcheck*. After the abbreviation handler method is completed, the remaining misspelled words are parsed through our spellchecking methods for a last correction.

Once historical plant outage data have been cleaned, the similarity value between the queried activity and each historical activity is determined. This results in an array of similarity values with dimensionality identical to the number of historical activities and the corresponding array (with identical dimensionality) containing the activity durations.

The predicted duration computation of the queried activity is determined by considering both the similarity and duration arrays. More precisely, by setting a similarity threshold (typically in the 0.7–0.9 range[†]), we are collecting elements of the duration array so the corresponding similarity measure is greater than . A relevant note to be highlighted here is that, if the queried activity has never been completed in past outages, no similar past activities with a similarity value above will be found. This approach does not, in fact, perform any type of regression.

An initial application of the developed methods was performed using the dataset provided by an existing U.S. NPP. This dataset contains activities performed during five outages. The data cleaning was performed for the activities contained in each of the five outages. A relevant feature of the provided datasets is that some activities are categorized using plant-specific labels. A label indicates the type of work performed in an activity (e.g., electrical, chemical, instrumentation and control). Note that a good portion of outage activities (about 30%) is not labeled. For those activities, the label NaN was assigned. About a hundred unique labels were identified. An example of similarity search results is shown in Figure 4; the output consists of an histogram representing the duration variance to complete the queried activity provided past outage data. Given these results, the analysis now has the possibility to statistically analyze the actual duration of similar activities in order to identify possible outliers obtained from the similarity search, track the historic trend in activity completion time, and evaluate the impact of employed human resources on completion time.

3. MEASURING OUTAGE SCHEDULE ROBUSTNESS

As indicated earlier, current outage schedule optimization methods rely on an activity duration expressed as a point value. However, the activity duration can be affected by many factors that can be either internal (e.g., number of personnel performing the task, plant crew workload) or external (e.g., discovery of a

[†] Recall that a similarity measure is in the (0,1] range where perfect similarity is indicated with the unitary value, while very low similarity values (near 0) will be assigned for dissimilar textual elements.

component failure during the inspection). Representing the effect of these factors on the activity duration via a single point value might have a major negative impact on schedule management during the outage.

The ability of a planned outage schedule to withstand a delay in an activity completion time (indicated with the term robustness) or to be able to counteract a change in an activity completion time (indicated with the term resilience) significantly increases the probability of the outage completion as predicted during the planning phase. From a plant operational standpoint, this also has an economic impact since, on average, each day of lost production (e.g., caused by outage delay) costs up to $2 million in terms of revenue. While the estimation of activity duration values from past outage data is presented in Section 3, this section focuses on how uncertainties associated with activity completion time can be used to measure the robustness and resilience of an outage schedule.

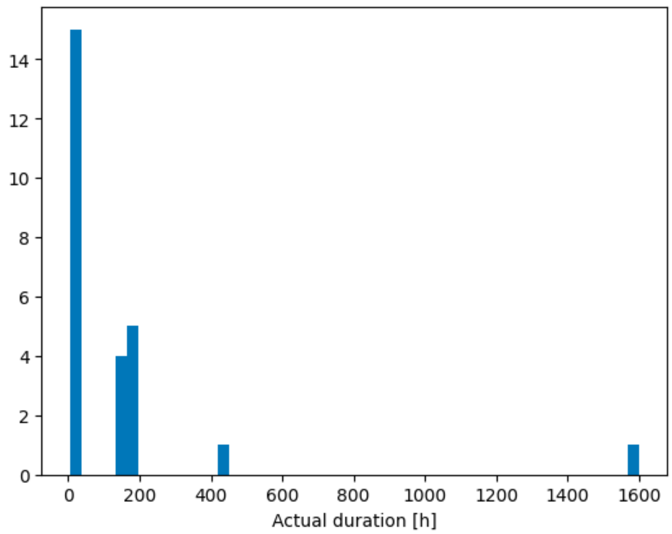

Figure 4. Example of similarity search results: a histogram representing the duration variance to complete the queried activity.

As a starting point, we need the ability to propagate activity duration uncertainties through an outage schedule (here indicated as CP uncertainty) and evaluate how uncertainties associated with an activity duration affect CP completion time, activity drag, and activity total float (TF). The propagation of activity duration uncertainties can be performed by assigning a probability distribution function (pdf) to an activity duration instead of reliance on a point value. This uncertainty quantification can then be easily performed through a classical Monte Carlo sampling method. Note that, when activity duration uncertainties are propagated through an outage schedule, the actual CP structure might change depending on the chosen activity duration values. In other words, depending on the sampled activity duration values, the sequence of activities that are part of the CP can differ.

The estimation of the robustness and resilience of a CP is performed by comparing the pdf associated with activity duration and the corresponding drag or TF values. For the activities that are part of the CP (see Figure 5), the activity duration pdf is compared with the duration point value used for the outage schedule (indicated in green in Figure 5). The portion of the pdf greater than the employed duration point value automatically adds a delay to the CP. The portion of the pdf lower than the employed duration point value indicates the possibility that plant resources (i.e., crew personnel) can become available to reduce the completion time of parallel or subsequent activities (see CP resilience).

A similar discussion applies for activities not on the CP (see Figure 5); the portion of the pdf lower than the employed duration point value would allow plant resources to be available to reduce completion time of other activities. The portion of the pdf greater than the TF implies that such activity would become part of the CP. In this situation, the CP would change, which would negatively impact outage completion time. The remaining portion of the pdf (i.e., located within the TF) is characterized by the fact that such an activity duration delay would not affect the CP (see CP robustness).

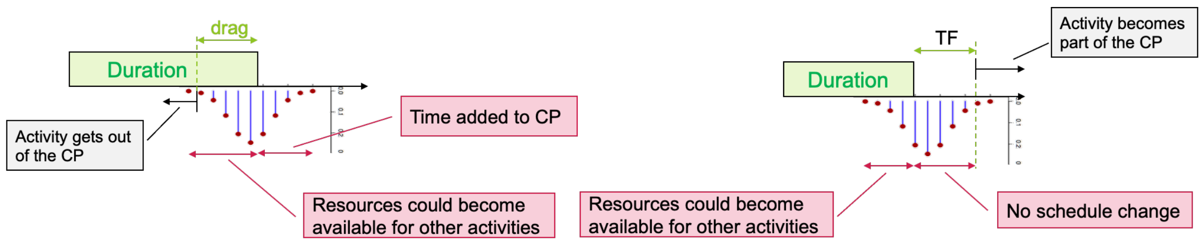

Figure 5. Analysis of activity duration uncertainty (represented through a histogram shown in blue) for an activity that is part of the CP (left) and for an activity that is ouside the CP (right).

The development of analytical tools designed to assess CP uncertainty, robustness, and resilience started during Fiscal Year 2023. Such tools are based on the Idaho National Laboratory–developed open-source code Risk Analysis and Virtual ENviroment (RAVEN) [9]. Here, the user can import the outage schedule into RAVEN through an input file. Using RAVEN, it is possible to propagate uncertainties associated with activity duration through the outage schedule. This can be performed by defining a pdf for each activity duration and choosing the desired sampling strategy (e.g., Monte Carlo or Latin Hypercube Sampling). Once the sampling has been completed, RAVEN provides a database that contains a pdf of the CP completion time and an alternative set of CPs.

The next step is the assessment of CP robustness, which can be computed by analyzing the fraction of the samples characterized by a CP completion time under the base CP completion time. More precisely, provided the pdf of CP completion time, CP robustness (indicated as Rob) can be solved analytically by integrating it up to base CP completion time (indicated as CP_time_base):

$$Rob = \int_{-\infty}^{CP\_time\_base} pdf^{CP\_time}(t)\, dt \quad (1)$$

4. CONCLUSIONS

In summary, the objective of this research project is to develop methods and tools to help plant staff create an outage schedule with a high probability of completion in the desired timeframe. Many plants continue to struggle with completing outages within the planned duration. Most utilities use the CP methodology to analyze and optimize schedules. To improve the current CP approach used by most utilities, we are investigating a method to calculate schedule resilience by creating a model of the schedule. A resilient schedule is one that is analyzed and adapted to find task duration uncertainties that can be reorganized to absorb completion time variability. This tool would provide the staff with margins for highly uncertain, non-CP activities. A resilient schedule is also one that has the capability to absorb the expected amount of scope growth without significant disruption to the planned duration.

The tools supporting outage activities may be used during outage planning and during outage progression. The outage planning can be improved by identifying activities where a planned duration is significantly different compared to the historical actual duration for the same activity. This would help to make the outage schedule more realistic. The planning tool would also identify non-CP strings of activities that have high uncertainties in their durations and therefore have a high probability of becoming a CP. Being informed about the high-risk activity strings allows alternative schedule planning options, which improves schedule resilience. The tool could highlight highly uncertain strings of activities that must remain in the schedule so that outage staff can maintain proper focus and oversight on those specific tasks to improve the chances of on-time completion. During the outage execution phase, the proposed tool would make recommendations for the best schedule for emergent work activities to minimize the chances or the magnitude of outage extensions.

In the initial stages of this project, we investigated various ML/AI methods to assess the possibility of automatically assigning activity duration distributions based on a historical outage performance data analysis. Facilities provided schedule data and several methods were used to evaluate the schedule duration and variability. In general, we determined that many ML/AI techniques fall short in interpreting the activity descriptions assigned by the utilities. In many cases, the use of abbreviations for activity descriptions limited the ability of the ML/AI tools to match an activity to other similar activities in the data set. The investigation of ML/AI capabilities will continue in the next project phase.

We also investigated using ranges of values described by a probability distribution function to represent an activity duration instead of a single value duration. This work will also be expanded in the next project phase. In cases where data are not sufficient to determine a duration distribution, a schema will be developed to assign a standard duration distribution based on a generic average completion time for common types of work activities, such as valve refurbishment, erecting scaffolding, circuit breaker refurbishment, etc.

While the concept of schedule resilience is understandable, metrics need to be developed based on schedule modeling for automated optimization and recommendations. These metrics for schedule resilience will assist outage schedulers in visualizing potential issues and provide useful information for schedulers to evaluate alternative scheduling options. Initial concepts for resilience metrics are presented in this paper and will be refined and expanded as the research continues. Additional future work will be done to integrate the information available in the work management databases and improve the assignment of distributions to the activity durations. The project team will also create example use-cases using representative example schedules to develop user interfaces and demonstrate concepts for schedule resilience presentation and recommendations for optimization.

ACKNOWLEDGMENTS

This manuscript has been authored by Battelle Energy Alliance, LLC under Contract No. DE-AC07-05ID14517 with the U.S. Department of Energy. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

REFERENCES

[1] K. A. Bollen and K. H. Barb, "Pearson's r and coarsely categorized measures," *Am. Sociol. Rev.*, vol. 46, no. 2, pp. 232-239, 1981, doi: 10.2307/2094981.

[2] P. K. Sen, "Estimates of the Regression Coefficient Based on Kendall's Tau," *J. Am. Stat. Assoc.*, vol. 63, no. 324, pp. 1379-1389, 1968, doi: 10.1080/01621459.1968.10480934.

[3] W. W. Daniel, *Applied Nonparametric Statistics*, 2nd ed. Cengage Learning, 2000.

[4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing*, 2019, doi: 10.18653/v1%2FD19-1410.

[5] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J. Big Data*, vol. 7, no. 1, pp. 94, 2020, doi: 10.1186/s40537-020-00369-8.

[6] R. Navigli and F. Martelli, “An overview of word and sense similarity,” *Nat. Lang. Eng.*, vol. 25, no. 6, pp. 693–714, 2019, doi: 10.1017/S1351324919000305.

[7] Y. Li et al., “Sentence similarity based on semantic nets and corpus statistics,” *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1138–1150, 2006, doi: 10.1109/TKDE.2006.130.

[8] Y. Li, Z. A. Bandar, and D. McLean, “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources,” *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 871–882, 2003, doi: 10.1109/TKDE.2003.1209005.

[9] C. Rabiti et al., “RAVEN User Manual,” Idaho National Laboratory, Idaho Falls, ID, USA, Rep. INL/EXT-15-34123, 2017.