NIA Project Close Down Report Document

Notes on Completion: Please refer to the appropriate NIA Governance Document to assist in the completion of this form. The full completed submission should not exceed 6 pages in total.
Network Licensees must publish the required Project Progress information on the Smarter Networks Portal by 31st July 2014 and each year thereafter. The Network Licensee(s) must publish Project Progress information for each NIA Project that has developed new learning in the preceding relevant year.

Date of Submission

Aug 2022

Project Reference Number

NIA_WPD_054

Project Progress

Project Title

Spatially Enabled Asset Management (SEAM)

Project Reference Number

NIA_WPD_054

Funding Licensee(s)

WPD - Western Power Distribution (East Midlands) Plc

Project Start Date

November 2020

Project Duration

1 year and 0 months

Nominated Project Contact(s)

Jenny Woodruff

Scope

The project will include LV and 11kV networks which represent the bulk of GIS data. 33kV networks are included to provide a comparison between the issues identified by the ML algorithm and those flagged by the Integrated Network Model. To enable the use of INM data, the South West region will be used though the approach could be applied to any area.

The project will investigate model inputs, outputs and different machine learning algorithms with the final model incorporated into a user interface. The design and algorithms used will be documented to enable knowledge transfer.

Objectives

In line with the overall objective of creating and testing a machine learning algorithm to identify and propose fixes for GIS data issues, the project objective are to;

- Generate of potential hypotheses to test and use cases for the tool to be applied to
- Understand the data available to support the machine learning proof of concept
- Outline of the model design including selection of machine learning algorithms.
- Create a final cleaned and prepared dataset that will be used to train and develop the model.
- Provide an interim report that sets out early findings from the modelling and direction for the remainder of the project.
- Develop the final version of the PoC model and front end.
- Carry out statistical evaluation of the model and accuracy through comparison of the model outputs with baseline and training datasets.
- Carry out data cleaning and loading of selected network area, including schematics if available in the format of a connectivity and impedance electrical model of EHV, HV and LV networks.
- Provide a summary of key findings, assessment of outcomes against success criteria, recommendations and learnings to be shared.

Success Criteria

The project will be judged successful if the following criteria are met.

- A standalone AI Model has been developed tested and applied to a dataset in the agreed regional area.
- The model performance has been evaluated and the application to the wider GIS data landscape assessed.
- The approach to roll into business-as-usual has been assessed with recommendations.
- Key learnings have been identified and shared with other DNOs.

Performance Compared to the Original Project Aims, Objectives and Success Criteria

The Project has met its objectives as follows:

- Generation of potential hypotheses to test and use cases for the tool to be applied to.
o Complete - Use cases and hypotheses have been documented in the system specification document.
- Understand the data available to support the Machine Learning Proof of Concept.
o Complete - Several sets of data have been provided including extracts from the Electric Office (EO) GIS system, the Integrated Network Model, extracts from our asset management system CROWN, data from policy documents, aggregated customer numbers and consumption data etc. Exploratory data analysis has been carried out on these datasets and a data dictionary has been created to support the project. The analysis of the datasets was included in the Interim Learning Report
- Outline the model design including the selection of Machine Learning algorithms.
o Complete - The selection of the models is given in the Model Definition Document
- Create a final cleaned and prepared dataset that will be used to train and develop the model.
o Complete - This was delivered to support the User Acceptance Testing and subsequent application of the model during the Trial.
- Provide an interim report that sets out early findings from the modelling and direction for the remainder of the project.
o Complete - An interim learning report has been produced and shared with external parties from Scottish and Southern Electricity Networks and Scottish Power Energy Networks via a webinar
- Develop the final version of the Proof of Concept model and front end.
o Complete - This was delivered and used to carry out the User Acceptance Testing and the SEAM Trial model use.
- Carry out statistical evaluation of the model and accuracy through comparison of the model outputs with baseline and training datasets.
o Complete – The model produces an evaluation report which introduces errors into the data and compares the accuracy of the model in correctly identifying incorrect data and proposing values for missing data. This is documented in the Model Evaluation Report 1(see Tables 12, 13 & 16) The Evaluation Report was output as part of the SEAM trial.
- Carry out data cleaning and loading of selected network area, including schematics if available in the format of a connectivity and impedance electrical model of Extra High Voltage (EHV), High Voltage (HV) and LV networks
o Majority Complete - The connectivity model was used to process and cleanse LV networks. The schematic models created from the data cleansing were viewable via the geopackage results. It had been anticipated that the same model could be used to cleanse HV and EHV networks and enable comparison to issues reported by the Integrated Network Model however this was not the case partly due to the INM issues not being directly comparable and partly as connectivity model could not be used interchangeably with different datasets for different voltage levels.
Given that the INM is better placed than the SEAM model to report GIS connectivity issues for the HV and EHV networks, the impact of not being able to implement this element is minimal.
EHV, HV and LV networks were included in the spatial graph model which allowed for missing cable types to be identified so that impedance electrical models can be populated. Therefore the vast majority of the potential benefits from the models has been achieved.
- Provide a summary of key findings, assessment of outcomes against success criteria, recommendations and learnings to be shared.
o Complete - Key findings, learning, outcomes and recommendations have been captured and shared via the Interim Learning Report, Model Evaluation Report and project webinar. This document also assesses the performance against the success criteria and includes a summary of the key learning.

The project has met all of its success criteria.

- A standalone Artificial Intelligence (AI) Model has been developed tested and applied to a dataset in the agreed regional area.
o Criteria met – the AI model was developed and tested in the agreed regional area.
- The model performance has been evaluated and the application to the wider GIS data landscape assessed.
o Criteria met – the model performance has been evaluated and the way in which it could be applied to the GIS data has been considered
- The approach to roll out into business-as-usual has been assessed with recommendations given.
o Criteria met – a full set of recommendations has been given in the closedown report
- Key learnings have been identified and shared with other DNOs.

o Criteria met – the learning has been disseminated via a webinar and sharing the closedown report.

Required Modifications to the Planned Approach During the Course of the Project

The original objectives assumed that having created a model to validate the connectivity of LV networks the same tool could be applied to validate the connectivity of HV and EHV networks and enable a comparison of any errors found to the errors identified by the Integrated Network Model. Unfortunately the different data structures for
network information at different voltages meant that this was not the case. However, during the project it was also found that the type of issues that were being reported in the Integrated Network Model for the trial area e.g. “No matching CROWN asset for PowerOn Transformer” or “ Cable termination is not between OHL and underground
cable” would not reflect the issues reported by the connectivity model which was designed to support use cases appropriate for LV networks and was therefore focussed on identifying LV network micro-disconnects, customers with no network attachment, overloaded sections of network indicating a missing normal open point etc.
Given that the Integrated Network Model already makes use of the dataset considered to be the master source for connectivity, i.e. PowerOn, it was unlikely that SEAM, which did not use that dataset, would be able to spot errors in the GIS data that had not already be identified by INM with the exception of HV connected sites that do not have customers associated with them. This has been identified as a data issue as part of the work on the EPIC innovation project but a report highlighting these sites could be generated within the asset management system, CROWN, which contains records of sites as well as customer to site mapping data. Therefore, while it would have been interesting to have made the comparison, this element was not core to the success of the SEAM project.
Creating an impedance model is therefore reliant on having complete data set for asset types and sizes. As the spatial graph model validates these items across LV, HV and EHV networks this element of the objective was able to be delivered.
All significant changes were subjected to the normal change request assessment as part of the usual project governance arrangements.

Lessons Learnt for Future Projects

Lessons learned related to two main areas, modelling methodology and data.

Modelling Methodology

· In order to effectively test data-driven and machine learning methods to identify data errors, the process had to introduce data quality exceptions in a way that simulates real life. Therefore an understanding of the prevalence of data errors is required to inform the synthetic errors added to train the Spatial Graph model.

· The data contains multiple distinct asset types with attributes that are not comparable. This means there are a limited number of features to apply traditional Machine Learning imputation techniques so an alternative more-complex approach is required, e.g. merging information about the other asset types into one main table per asset type, or some form of multi-model graph model.

· There are a number of key attributes that are high-cardinality categorical features, such as specification description, structure number, site number, circuit ID, etc. Basic approaches for encoding categorical features, such as one-hot encoding, can only be use effectively with categorical features with low cardinality. It may be possible to simplify some of these features by splitting them into separate columns, i.e. separating composite features like specification description.

· Where possible, rules-based algorithms should be preferred to data-driven ones, since these are easy to verify and understand, and the resulting suggestions have a very high probability of being correct.

· The methods used in SEAM are robust to different topologies and configurations of the networks, accommodating radial and mesh and can be used in a number of different scenarios where data on network topology may not be of high quality or complete.

· The use of this model could be more iterative in nature, with a data steward checking violations, updating Electric Office where violations may be caused by configuration, specifications and re-running the model to see the improvements made and reduction in violations.

· Traditional Machine Learning approaches that work with table-based observations (e.g. regression techniques such as k-nearest neighbours) will have limited usefulness. This was an initial hypothesis and a proposed approach for Use Cases 2 and 3. In context of geospatial data the absolute location of each asset is of limited utility on its own: what matters more is the local neighbourhood of each asset, i.e. what are the attributes of the other assets in the surrounding area? Therefore our approach will utilise a graph model (i.e. based on a connected graph of nodes and relationships with properties and labels).

· Traditional graph models for power networks are focused on power systems analysis and network management, rather than on

asset management. They typically rely on electrical properties and require complete electrical connectivity – ignoring spatial relationships. This approach is well suited where the physical connectivity of the model is central to conducting the modelling or forms a part of the pattern identification

· The performance of the model was evaluated in line with set of synthetic errors. The model was trained/optimised to correct those synthetic errors which were reflected in any "confidence scores" assigned. The confidence scores were found to be useful in separating results into high and low confidence groups and the high confidence groups were seen to have greater accuracy.

· There is a trade off in the depth of the neural network (which requires greater processing and potential overfitting) and performance of the model.

The full spatial graph model is a heterogenous graph (or heterograph), which necessitates the usage of Relational Graph Convolutional Layers (RGCN), and types of layers derived from them, in the neural

· network model. Furthermore, each asset attribute to be predicted by the neural network requires a separate "head", resulting in a multi-headed architecture. This meant that the neural network had a backbone of several RGCN (or similar) graph convolutional layers followed by several Machine Learning Process heads taking the asset node embeddings as inputs.

· The application of Neural Network to predict asset attributes and relationships based on spatial relationships continues to be a viable method, as additional complexity is added. Proven to produce level of performance for predicting network type and operational voltage of each asset.

· As a side-effect of the process of creating the spatial mesh, it is possible to create a report of coordinates from all geometries in the region of interest that are very close each other (about 10% of edges in the spatial mesh are under 10cm). This could be used to make minor modifications to the geometry of some assets to ensure that they "snap together" exactly.

· For the purposes of the Proof of Concept, it is simplest if all of the node attributes to be predicted/corrected are categorical ones. While it is relatively straightforward to support both classification and regression heads in the neural network, it is not a high-enough priority, especially given the time remaining. Also, creation of "confidence scores" is easier for classification tasks. This means that Numerical attributes to be predicted, e.g. conductor rating, must be binned into fixed ranges.

· Evaluation of the spatial model has shown that the performance is good for the Proof of Concept. Furthermore, this model has the ability to be trained on one subset of the network and then used to identify and correct errors in another subset of the network. It is also able to be extended with more data, functionality and optimization. This proves that Generalised Neural Networks are a good approach for data cleaning for asset management.

· Evaluation of the spatial model has also identified some patterns of false alarms that are consequences of the highly limited data available to the current version of the model. Some "quick wins" have been identified that should bring significant performance benefits.

· The spatial and connectivity model are complementary. For example, the features calculated for the connectivity model (e.g. electrical connectivity, electrical properties) are valuable inputs for the prediction of the cable/wire specifications. Similarly, the electrical properties can be back-filled using the spatial model in order to get better estimates of the network capacity for the connectivity model.
Hence, combining the models will bring significant performance benefits to both.

· Network flow provides an exact characterization of network capacity in the single commodity case (i.e. real power flow). The linear formulation of network flow is functional and efficient as well as fast to solve and is sufficient for analysis that does not consider: analysis of systems away from their operating points (blackouts, instabilities), losses and coupling between real and reactive power.

· Where there are multiple demand and supply points in a network; commonly known as the circulation with demands problem, can be reduced to a network flow problem (which has many fast and efficient algorithms) by adding a synthetic 'super' source and 'super' sink nodes with edges that lead to actual source and sink nodes with capacity equal to demand / supply. This was used in the SEAM solution.

· Max flow is fast (data preparation and post processing phases take the majority of the model running time), robust and efficient. If the processing is done on a circuit by circuit basis, these procedures can be done in parallel.

Data

· There is no direction / parent-child relationships within the GIS data and so for connected point assets / line segment elements there is no indication within the data of the direction of flow of power expected within the asset. This meant that the connectivity graph needed to be undirected graph and any method chosen needed to be robust to the lack of direction / parent-child relationships within the data.

· The ability to eliminate reasons for violations (customer wrongly assigned, profile class wrongly assigned, Estimated Annual Consumption or half hourly consumption error, for example) is diminished due to the level of missing assets (cables and wires to create connectivity and connections to customers) and missing labels for cable and wire specifications. Again, this suggests that an iterative approach may be useful where this data is progressively added. Some improvements to the specification descriptions for cables and wires was required (potentially as an input from model 2); as well as assessment / review of the missing / synthetic service cables.

· There are few 'true' violations of network capacity indicated in the data as mostly the components of the network flagged as bottlenecks are where capacity values have been or reflect simulated cables / wires or the simplifying assumptions used to model ways in which customers are connected. This meant there was a need to combine outputs from the spatial graph model / manual interventions in the quality of technical data for circuits to reduce the number of false positives

· A sufficient completeness of physical circuits is required to understand the relationship between assets in different locations and how this can be pooled and used to improve the data quality in all of those areas. While work was ongoing during the project to build LV network connectivity in Electric Office (the circuits in our dataset include the outcome of Phase 1 of this project),a significant number of LV cables, wires and point assets with no circuit ID remained.

· Complete and detailed data dictionaries/catalogue do not currently exist for all our data sources. This slowed the process of forming a detailed understanding of the data to determine the best suited modelling approaches.

· There are different naming conventions for attributes across the different systems/sources which introduced an element of confusion. A data catalogue is being created to support the project data model to ensure there is a clear understanding of the relationship between datasets. This includes mapping to the Common Information Model outputs from the Integrated Network Model project.

· The project would have liked to use the Energy Performance Certificate (EPC) dataset to enrich the customer features (issued for domestic and non-domestic buildings constructed, sold or let since 2008). There is a challenge linking this to Meter Point Registration System (MPRS) data because EPC does not contain Unique Property Reference Number (UPRN) and the address data available is not well structured. Either including a UPRN on EPCs or improving the quality of the address data in CROWN would be beneficial.

· Lack of standard formats for data extracts from CROWN and Electric Office would need to be resolved to make the data analysis repeatable.

· The cable and wire specification attributes in EO are a concatenation of three associated components (size, type/material, number of conductors). A significant number of these contain at least one component that is 'unknown'

Note: The following sections are only required for those projects which have been completed since 1st April 2013, or since the previous Project Progress information was reported.

The Outcomes of the Project

The main outcomes of the project are;

1. An assessment of our initial data evaluation in the trial area has been made as part of the Interim Learning Report. This has provided an overview of the completeness of the different datasets and proportions of different asset types between the voltage layers rather than commenting on the accuracy of the data presented.

2. The types of error that can affect the GIS data have been documented as use cases and grouped together to allow for mapping

between use case groups and potential evaluation methods. This is likely to be transferrable knowledge to other DNOs.

3. Interim learning has been shared with other DNOs that have already carried out work in this area or are planning to in order to avoid duplication of effort.

4. The applicability of AI approaches to identifying and suggesting corrections to GIS errors has been confirmed.

5. Two complementary modelling approaches have been selected and the rationale for their selection has been documented and shared.

6. The Proof of Concept model has been developed and tested both by Capgemini staff and on our hardware with a configuration that does not require access to the internet by our staff.

7. The accuracy of the Proof of Concept models has been evaluated and shown to be above that achieved by assuming the most frequently occurring result.

8. The results of the models have been evaluated and confidence metrics have been used to separate values with high and low confidence. The separate groups are seen to differ in accuracy with the high confidence group achieving better results than the low confidence group, confirming the usefulness of the confidence metrics.

9. Reports from using the model have been passed back to the business to allow identified errors and proposed corrections to be examined further with a view to correcting the errors identified.

10. Comparison with INM errors has shown that the different approaches are complementary.

11. Suggested priorities for Business As Usual implementation and further analysis likely to improve data accuracy have been proposed

12. Learning has been disseminated via published reports and a webinar enabling other DNOs to build on the learning generated by the project without duplicating the work

Data Access

No new data about the network or consumption has been gathered in the course of this Project, but use has been made of existing data within WPD's systems. Network model data, such as that contained within our Integrated Network Model can be accessed via our Energy Data Hub. https://www.westernpower.co.uk/our-network/energydata-hub

Detailed network plans are available via our Data Portal. https://www.westernpower.co.uk/ournetwork/ network-plans-and-information

Access to the data generated during the project about specific identified GIS errors can be obtained via our normal data sharing policy using our on-line form at https://www.westernpower.co.uk/Innovation/Contact-us-andmore/ Project-Data.aspx.

https://www.westernpower.co.uk/Innovation/Contact-us-andmore/ Project-Data.aspx.

Foreground IPR

Default IPR arrangements apply to the project.

The two items of Foreground IPR developed by the project team are;
1) The Proof of Concept Model comprising the Excel front end, supporting Python code and data templates.
2) The documentation supporting the Proof of Concept Model i.e. the Specification Document, Model Design document and Model Build Document and model installation instructions.
These are jointly owned in equal shares by the project partners WSP and Cap Gemini

Planned Implementation

We have recently created a new Data and Digitalisation team. Master data improvement is one of the elements of our data and digitalisation action plan2. The team are currently determining the scale of missing data items and prioritising data items to be backfilled. The way in which modelled data will be incorporated within systems along with provenance information identifying the model and version that was used to generate the estimated values and any confidence metrics has not yet been determined. It has been recognised that this data must be clearly distinguishable from that which was captured on-site or inherited from legacy systems.

Other Comments

Not Applicable

Standards Documents

Not Applicable