WPD INNOVATION

Transforming the electricity network

Spatially Enabled Asset Management (SEAM)

NIA Project Closedown Report
WPD_NIA_054

WESTERN POWER DISTRIBUTION

Serving the Midlands, South West and Wales o

Version Control

Issue	Date
0.1 First Draft	25/06/2021
0.2 Review by Ryan Huxtable	28/07/2021
0.3 Review by Yiango Mavrocostanti	10/09/2021
1 Final version	20/09/2021

Publication Control

Name	Role
Jenny Woodruff/ Ben Neate	Author
Ryan Huxtable	Reviewer
Yiango Mavrocostanti	Approver

Contact Details

Email

wpdinnovation@westernpower.co.uk

Postal

Innovation Team
Western Power Distribution
Pegasus Business Park
Herald Way
Castle Donington
Derbyshire DE74 2TU

Disclaimer

Neither WPD, nor any person acting on its behalf, makes any warranty, express or implied, with respect to the use of any information, method or process disclosed in this document or that such use may not infringe the rights of any third party or assumes any liabilities with respect to the use of, or for damage resulting in any way from the use of, any information, apparatus, method or process disclosed in the document.
Western Power Distribution 2021

Contents

1. Executive Summary

The Spatially Enabled Asset Management (SEAM) project investigated the use of advanced analytical techniques to identify and resolve Geographic Information System (GIS) errors. Inaccuracies in our data could prevent their digitalisation strategy, constrain the future build of network topologies that support smart networks and the transition to a DSO and reduce the value and wider use of their data by third parties. An analysis of the common types of GIS error resulted in several use cases being proposed with SEAM focussing on the “harder to fix” error types that were not simple to fix using algorithms in Electric Office (EO). These use cases were addressed with two modelling solutions which were incorporated into a prototype Proof of Concept (PoC) tool that used an Excel front end, linked to Python-based models, accessed using the Anaconda3 software package.

Model 1 – the Customer Connectivity Model, created network graphs representing point assets such as substation and customers as nodes and linear assets such as underground cables or overhead lines as edges. This process identified disconnects in the underlying data and suggested where these should be connected in order to produce a single graph model for each LV feeder. The resulting network model was then tested for its capacity to carry the expected loads using a max-flow algorithm. Overloaded sections could indicate an error in the network model, for example where a missing normal open point had resulted in the network extending beyond its real end points.

Model 2 – the Spatial Graph Model recognised that in many cases connectivity information was incomplete, and therefore took a spatial approach to confirming and proposing key asset attributes. A graph neural network machine learning model was used to predict the correct values of the network type, operational voltage, specification material and specification size for network assets. The Machine Learning (ML) model is trained by simulating errors in the original data, and optimizing the model to predict the original values, as a semi-supervised learning problem due to the presence of missing values in the original data. This enables the model to both suggest values for missing data and to identify attributes with incorrect values.

Both models were trained using an area of network around Barnstaple and then tested on a reserved area of network that was not used as part of the model training. The models were fast to run, with both the model tuning and prediction run in a reasonable time on a standard WPD laptop.

The results from the proof-of-concept demonstrate that advanced analytical techniques and machine learning can be used to identify and correct potential inaccuracies in GIS data with reasonable confidence (with potential to further improve performance) For example the high confidence results from the Graph model approach showed 98.7% accuracy when predicting the network type when this was missing and 98% accuracy in identifying that a value was incorrect. It also showed 99.9% accuracy when determining that the data present was correct which is important to avoid large numbers of false positives. Accuracies for predictions of the asset voltage and specification material ranged from 88.7% to 99.2% for missing values and identification of correct data though predictions of incorrect data were lower in the 71%to 83% range. In terms of the connectivity model, this was seen to reduce the disconnects in LV feeders improving the number of feeders that were represented by a single connected graph by 109 while reducing the number of feeders with more than four disconnects by 89. Overall 216 feeders had potential disconnects identified out of a sample of 1309 with some feeders having multiple issues identified. The use of the models would complement existing network GIS data cleanse initiatives by reporting predictions on the “harder to fix” inaccuracies that aren’t straight forward to identify and fix using algorithms in EO and offer an additional data point to be evaluated by data stewards when resolving issues.

A comparison to the Integrated Network Model (INM) was considered by the project. SEAM has primarily focused on the LV and 11kV networks because these are where most of the GIS data exists. In contrast INM does not cover the LV network as it aligns to existing network coverage in the Distribution Management System. SEAM can be used to target the improvement of LV network data quality by using only a limited number of attributes and the geospatial relationships, which all come directly from the EO dataset. It was seen that for 11kV,,33kV and 132kV systems, the Graph model approach suggests assets without issues made up 84%, 91% and 97% of the populations however the INM identified far fewer issues suggesting that the incorrect and missing values may be more numerous than the inconsistencies between systems

The SEAM models are designed to be scalable across all network types and operational voltages – the spatial graph model covers all operational voltages for the target attributes within the selected trial area geography. The SEAM models have a flexible design that can be incrementally enhanced and extended.

A series of prioritised recommendations have been made by the project team to implement the models into business-as-usual and improve model performance. A key step to transitioning the models to BaU is to form a process for reviewing the reported errors/fixes and establish how these can be represented within EO to reflect they are modelled/predicted values with an associated level of confidence. Alternative methods such as the work carried out by SPEN using smart meter data to validate LV network connectivity and cable types could be used as an additional sense-check of the information without a need for site visits. It is also believed that as missing data items are populated the model performance would also improve as more accurate data would be used to construct and train the models. Therefore one of the “quick wins” identified is to use the results of one model to improve the performance of the other.

2. Project Background

Geographic Information Systems for utilities have been created by the digitisation of paper records. These have then undergone transformations as data has been moved between one system and another and manipulations such as the corrections to the Master Map background. It is expected that there will be inaccuracies in the current records and these can persist for many years partly due to the lack of visibility of buried assets and partly due to the length of time between sites requiring site visits that would be expected to update the GIS.

Inaccuracies in the GIS system have the potential to impact:

- Accuracy of network modelling
- Accuracy of regulatory reporting
- Field safety
- Network operational efficiency
- Network upgrade/maintenance efficiency
- Accuracy of New Connections information
- Accuracy of information provided to third parties

Similarly, when performing power systems analysis on the networks, data gaps (such as missing cables and asset types) can be a key issue and can greatly increase the analysis time due to time spent fitting data and cleansing the set prior to carrying out the analysis. With the data having potentially many users within and outside of WPD, every time data gaps are filled this not only duplicates effort but is likely to result in different assumptions being made and potentially different conclusions being drawn.

While there is a clear need to improve data quality, process to correct identified GIS problems, for example a micro-disconnect issue may be highly manual and therefore costly and time consuming. Therefore, the use of an Artificial Intelligence (AI) engine to identify potential data issues and propose corrections would make the process far more practical and affordable.

To date there have been very few instances of using AI with GIS data within UK utilities, so determining whether AI can contribute to improving data quality was an area of research likely to be of value beyond WPD. During registration it became apparent that some previous work had been carried out by Scottish Power Energy Networks (SPEN) and that future work was being planned by Scottish and Southern Electricity Networks (SSEN). Workshops were held to ensure that work was not duplicated and an interim findings phase added to the project to ensure that useful information was passed to SSEN as soon as possible.

The project was structured into the following work packages with the associated deliverables as given below and ran from November 2020 to (October 2021).

Figure 2-1: Project Phases and Deliverables

Work Package	Deliverable
WP1 Specification	D01: Project Specification Document
WP2 Design	D02: Tool Design Document
WP3 Build (Phase 1)	D03: Interim Findings Report
WP4 Build (Phase 2)	D04: AI Model and User Interface
	D05: Model Design Document
	D06: Cleansed dataset
WP5 Evaluate	D07: Model Evaluation Report
WP6 Report	D08: Final Project Report

3. Scope and Objectives

The project has met the objectives set out when the project was registered as given below.

Figure 3-1: Status of project objectives

Objective	Status
Generate 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 the model design including the 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 Proof of Concept 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	Majority Delivered
Provide a summary of key findings, assessment of outcomes against success criteria, recommendations and learnings to be shared.	✓

4. Success Criteria

The project has met all of the Success Criteria set out when the project was registered as given below.

Figure 4-1: Status of Project Success Criteria

Success Criteria	Status
A standalone Artificial Intelligence (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 out into business-as-usual has been assessed with recommendations given.	✓
Key learnings have been identified and shared with other DNOs	✓

5. Details of the Work Carried Out

One of the key risks that was identified when registering the project was that the data may not be suitable to support the development of an AI model. Therefore one of the early project activities, was to provide a variety of data samples even before a trial area had been finalised.

This was followed by workshops with the Mapping Technology Team to identify the issues that had been identified as affecting the GIS data which were unlikely to be easily addressed by the teams existing programme of work, which was largely focussed on relating drawing objects for the LV networks to LV circuits. The issues were captured as use cases which were then grouped and incorporated in the Project Specification Document, which clarified the expectations for the Proof of Concept system. The groups of use cases are given in Figure 5-1

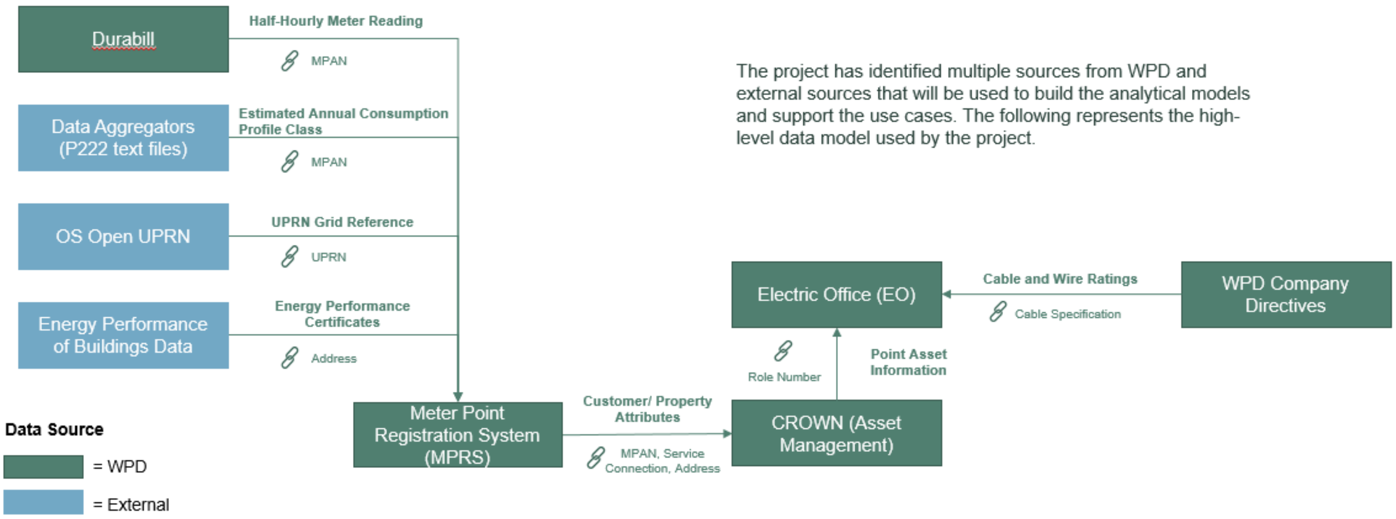
Figure 5-1: Use Case Groups

Use Case Group	Investigation Methods
Customer Connectivity	• Identify potential errors and link customers to LV circuitry through connection points, where existent in Crown or Electric Office and also using spatial proximity if there are un-mapped customers. • Estimate peak demand (i.e. on a winter's evening) up to substation level to identify potential areas in the LV network that the demand surpasses the capacity. • Infer inaccurate or missing customer attributes with enriched customer datasets.
Incorrect or missing asset attributes and missing assets	• Identify missing or incorrect LV asset attributes using EO data and suggest corrections. • Develop a graph model and apply Machine Learning techniques to identify relationships and patterns in the way that point and linear assets as well as their attributes are connected to highlight potential errors and suggest corrections.
Inconsistency across systems	• Where there are disparate systems that store data on the same asset (e.g. EO and CROWN) and there is associated locational data, it can be tested whether this can be used to verify any attributes that are stored in each system. • Build on the graph model developed in Use Case group 2 by linking with CROWN dataset.
Technical inconsistencies	• Building a technical ruleset as well as 'learning' rules from the data to highlight where circuitry built from the GIS data is infeasible. • Use point/line topological graph modelling to assess possible paths from substation to customer to find exceptions where this is technically infeasible, illogical or unlikely.

Exploratory Data Analysis (EDA) continued with additional potential datasets being identified once the use cases were defined. The datasets are shown in Figure 5-2: High-level logical data model. The potential methods and models which could be used to satisfy the use cases were considered with the optimal approach being formulated to use two different types of model. The way in which the models would satisfy the functional requirements was recorded in the Model Definition Document[1].

[1] Model Definition Document https://www.westernpower.co.uk/downloads-view-reciteme/360802

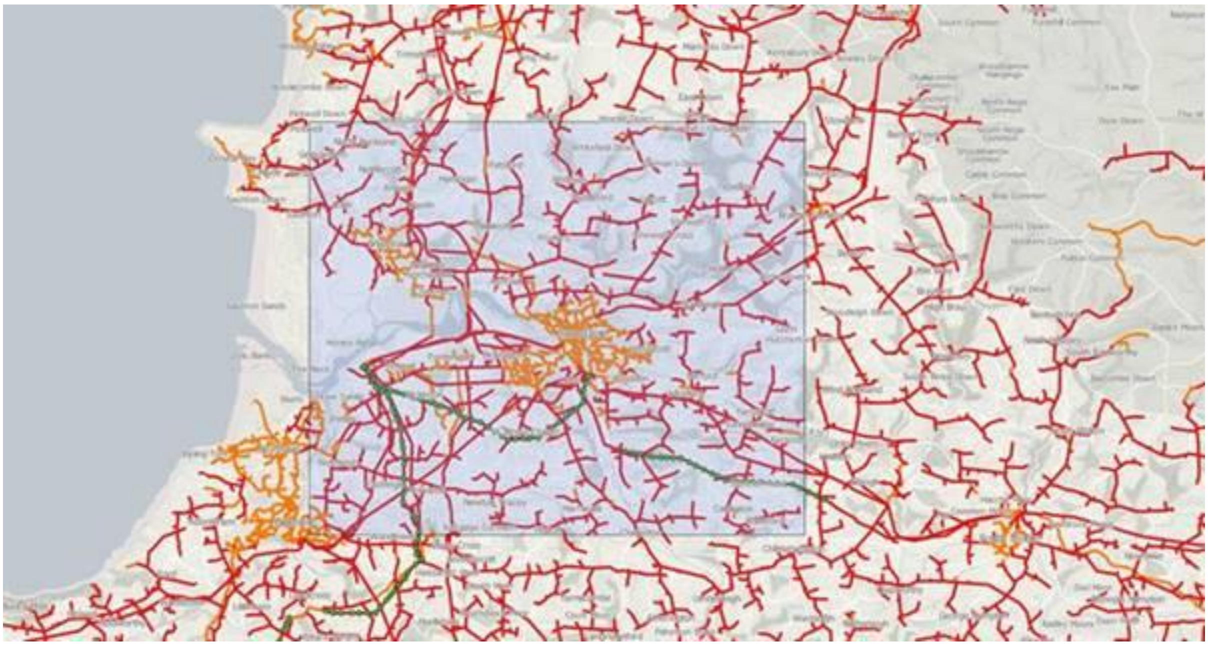
Figure 5-2: High-level logical data model


Durabill
Half-Hourly Meter Reading
MPAN
Data Aggregators (P222 text files)
Estimated Annual Consumption Profile Class
MPAN
OS Open UPRN
UPRN Grid Reference
UPRN
Energy Performance of Buildings Data
Energy Performance Certificates
Address
Meter Point Registration System (MPRS)
Customer/ Property Attributes
MPAN, Service Connection, Address
CROWN (Asset Management)
Point Asset Information
Role Number
Electric Office (EO)
Cable and Wire Ratings
Cable Specification
WPD Company Directives
Data Source
= WPD
= External

The project has identified multiple sources from WPD and external sources that will be used to build the analytical models and support the use cases. The following represents the high-level data model used by the project.

After the EDA was complete the trial area was finalised which included a rectangular area centred on the Barnstaple area of our South West licence area as shown in Figure 5-3 : SEAM Trial Area

Figure 5-3 : SEAM Trial Area



SEAM Trial area - The area of interest is a 18km by 15km rectangle centred on Barnstaple and aligned with the 1km OS grid; the SE corner has OS grid ref SS 46000 26000.

In order to inform the upcoming work by SSEN, an additional output had been added to the project to share learning at an early stage via an Interim Learning Report [2]and associated workshops.

This contained:

- an assessment of the observed data quality issues from the sample data provided;
- an assessment showing how the requirements for using Machine Learning with GIS data were met;
- a data map showing the different data extracts that had been used, their relationship to each other and their expected use within the proposed models;
- a summary of the use cases;
- the benefits of using both a spatial model and traditional graph model, and
- the future direction of the project.

The main build phase of the project involved developing, testing and refining the AI models used and the User Interface.

Another challenge was the need to prove the Proof of Concept model could operate on a WPD standard computer. The standard laptop met the basic requirements in terms of processor and RAM but our security requirements prevented the system being able to access the usual Python Libraries that would normally be available via an internet connection. This meant that the packaged solution provided needed to include all the associated libraries which would be stored and accessed locally on the WPD laptop.

Installation of the system on the WPD laptop was also complicated by the security features preventing the SEAM team from connect to the WPD machine remotely. This resulted in installation taking place by the transfer of a number of very large installation files with very clear instructions to ensure that the various components were installed in the correct folders.

User Acceptance Testing (UAT) took place over a series of days with updated files being provided to correct the small number of issue that were identified. After the UAT was completed the WPD project manager was able to configure the tool to run the training, evaluation and testing phases against the given training area and reserved testing area. Having run the tools to create the various output reports, sample errors from these reports were then investigated using the information in the Excel reports to identify areas to examine using our GIS tool, EMU. The geopackage output files were inspected using QGIS.

The results from using the tools on the test and trial areas were evaluated in terms of the measured accuracy of the model outputs. An internal workshop was held to feedback the results to the GIS team before the results were captured in the Model Evaluation Report. [3]

As an additional item, SEAM project partners provided a knowledge sharing workshop open to all our staff with an interest in understanding more about AI which covered the various different types of AI, their suitability for different applications and examples of their application within utilities. This was very helpful at providing a layman-friendly introduction to the area which is likely to feature in future innovation projects.

Finally the project held a dissemination event[4] to share the learning from the project at the end of June 2021. This was ahead of the planned schedule to ensure the availability of key team members.

[2] Interim Learning Report https://www.westernpower.co.uk/downloads-view-reciteme/360820

[3] Model Evaluation Report https://www.westernpower.co.uk/downloads-view-reciteme/360805

[4] Webinar recording https://www.westernpower.co.uk/projects/spatially-enabled-asset-management-seam

6. Performance Compared to Original Aims, Objectives and Success Criteria

6.1. Objectives

The Project has satisfied the original aims and objectives as detailed in Figure 6-1.

Figure 6-1 : Performance Compared to Original Aims and Objectives

Objective	Met	Summary of evaluation and evidence
Generation of potential hypotheses to test and use cases for the tool to be applied to.	Yes	Use cases and hypotheses have been documented in the System Specification Document and Model Definition Document.
Understand the data available to support the Machine Learning Proof of Concept.	Yes	Several sets of data have been provided including extracts from the Electric Office 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.	Yes	The selection of the models is given in the Model Definition Document The details of the model design is specified in the Model Build Document.
Create a final cleaned and prepared dataset that will be used to train and develop the model.	Yes	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.	Yes	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.	Yes	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.	Yes	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 (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	Majority Met	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 has been anticipated that the same model could be used to cleanse HV and EHV networks and enable

Objective	Met	Summary of evaluation and evidence
connectivity and impedance electrical model of EHV, HV and LV networks.		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.	Yes	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.

6.2. Success Criteria

The Project has satisfied the original aims and objectives as detailed in Figure 6-1.

Figure 6-2 Performance Compared to Original Success Criteria

Success Criteria	Met	Summary of evaluation and evidence
A standalone AI Model has been developed tested and applied to a dataset in the agreed regional area.	Yes	• A standalone solution has been developed that can be run through a single interface that allows a user to run both models and customise parameters, inputs, and outputs. • The models use Machine Learning and advanced analytics techniques but does not require coding or data science expertise to run and adjust them. • Relative to the complexity of the tasks being performed by the models, the run time is fast and can be processed on a standard laptop. • The models have been trained and evaluated on the agreed Barnstaple area in the South West region. • An independent hold-out area was removed from training the models and used to demonstrate the models can be applied to a different geographic area (with identical data structures) with comparable predictive results.
The model performance has been evaluated and the application to the wider Geospatial Information System (GIS) data landscape assessed.	Yes	• The results from the models on the proof-of-concept area demonstrate that Machine Learning can be used to identify and correct potential missing or erroneous GIS data. • Graph models are central to the project's modelling approach. This comprises a traditional graph model that relies on electrical connectivity (Model 1) and a spatial graph model focussed on predicting asset attributes and relationships which emphasises the spatial relationship between assets (Model 2). • The performance of the models is discussed in detail within this report with evidence from the output reports. This also includes discussion on the confidence levels and potential thresholds for reporting the predictions from Model 2.

		• The models have been developed to a level of performance that supports the aim and the key learning objectives of the proof-of-concept. Further enhancements and extending the scope of GIS data included in the model have been considered in the Model Evaluation Report.
The approach to roll into business-as-usual has been assessed with recommendations	Yes	• The project has developed a set of recommendations that cover the key steps to transition the current SEAM models into BaU (see Section **Error! Reference source not found.** of the Model Evaluation Report). • Key components of the BaU implementation are the review and validation process, and how to reflect the predicted values and the associated confidence levels in EO alongside the original values. • Additional recommendations cover scaling-up the models and potential additional use cases.
Key learnings have been identified and shared with other DNOs.	Yes	• The Model Evaluation Report has been published and the project learning has been shared via a webinar.

7. Required Modifications to the Planned Approach during the Course of the Project

There have been minor changes in the scheduling of some activities within the project reflecting the length of time required for data provision, which has taken longer than expected.

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.

8. Project Costs

The project has been completed within the original budget as given in **Error! Reference source not found.**, below.

Figure 8-1: Project Spend

Activity	Budget	Actual	Variance
WPD Project & Programme Management costs	£37,544	£30,976	£6,568
Contractor Costs	£350,000	£350,000	0
Contingency	£38,754	£0	£38,754
Total	**£426,298**	**£380,976**	**£45,322**

The most significant variation to budget is the approximately £39k reserved as project contingency. This was not used by the project which was relatively short and focussed.

There is also a minor variance in project management costs is due to marginally less time being required to support the project than planned resulting in an underspend of approximately £6.5k.

9. Lessons Learnt for Future Projects

The learning from the project has been captured and presented to stakeholders via the;

- Interim Learning Report
- Model Evaluation Report, and
- SEAM Project Webinar

The key points have been summarised in the Figure 9-1: Project Learning Summary below.

Figure 9-1: Project Learning Summary

Area	Learning Detail	Internal Outcomes
Modelling methodology	In order to effectively test data-driven and machine learning methods to identify data errors, the process will need to be able to introduce data quality exceptions in a way that simulates real life to most effectively measure the success of testing any algorithm. The sensitivity of each modelling algorithm will be tested against pervasiveness of errors as to test the limitations of each methodology.	Missing values and erroneous values are the two types of errors introduced to the 'cleansed data set' before testing each algorithm. Missing values will be informed by a data profiling activity.
Modelling methodology	The data contains multiple distinct asset types and the attributes of each are not necessarily comparable. This means that it is not possible to just create a table with all of the assets and there are a limited number of features to apply traditional Machine Learning imputation techniques.	An alternative more-complex approach is required, such as merging information about the other asset types into one main table per asset type, or some form of multi-model graph model.
Modelling methodology	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.
Modelling methodology	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 project is looking to leverage domain knowledge first as it maximises the available information in the training data for the data-driven algorithms to extract more subtle or more complex patterns that can be used to address harder-to-fix data errors, by narrowing the scope of the remaining errors present.

Data	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. Our Technology Mapping Team are currently undertaking a project to build LV network connectivity in EO (the circuits in our dataset include the outcome of Phase 1 of this project), but there remain a significant number of LV cables, wires and point assets with no circuit ID	Our assessment of the data suggests there is sufficient completeness of circuits to evaluate Machine Learning based methods for the purposes of the proof-of-concept. Increased completeness of the physical connectivity will likely enhance the performance of the models.
Data	Complete and detailed data dictionaries/catalogue do not currently exist for all our data sources. This would provide the foundation for forming a detailed understanding of the data to determine the best suited modelling approaches.	Further investigation and ongoing engagement with data owners has been undertaken to establish an understanding of the data in sufficient detail to design the analytical approaches.
Data	There are different naming conventions for attributes across the different systems/sources.	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.
Data	The project is exploring the use of 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 MPRS data because EPC does not contain UPRN (it includes a unique building reference number that has no relationship with UPRN). Further, the address data in CROWN is not well structured (the address is inconsistently ordered across 9 address lines).	The project is carrying out an address matching analysis. This is unlikely to produce 100% accuracy due to differences in formats or structure across datasets. The use of UPRN and address (from OS AddressBase premium) to match would provide a more structured dataset but this is not available to support the proof-of-concept.
Data	The EO and CROWN data extracts are not created as standard reports. This will require a system user to select the data attributes and geographical coverage each time an extract for a model run is required.	To make the analysis repeatable in the PoC tool the data inputs will need to be in the same format to read into the model and evaluate the relevant features. This will be defined as part of the Model Design document.
Data	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’.	Extracts from our Directive for cables and wires is used to map current ratings to the assets. The cable and wire ratings depend on size, material, number of conductors, location (i.e. in ground, duct or in air) and season/loading type.

Modelling methodology	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?	Our approach to all use cases will utilise a graph model (i.e. based on a connected graph of nodes and relationships with properties and labels). Other applications of graph models in this field include the Integrated Network Model data process and Scottish Power Energy Networks for representing an LV Network as a Network Tree Graph to verification of network cables and topologies.
Modelling methodology	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 project is evaluating the use of a traditional graph model for Use Cases 1 and 4.
Modelling methodology	A new graph model is required that is focussed on predicting asset attributes and relationships and emphasises the spatial relationship between assets.	The project is developing a novel approach based on a spatial graph model that contains a layer of point location nodes, with distance relationships between them to create a spatial mesh, and edges between each asset or feature and the location nodes that are part of it. This will support Use Cases 2 and 3.
Modelling methodology	The application of Neural Network to predict asset attributes and relationships based on spatial relationships is a viable method. Proven to produce level of performance for predicting main network type at each location.	Take forward the Spatial Graph methodology for further development in the PoC.
Modelling methodology	An understanding of the prevalence of data errors is required to inform the synthetic errors added to train the Spatial Graph model.	The performance of the model will be evaluated in line with set of synthetic errors. Model will be trained/optimised to correct those synthetic errors. Those synthetic errors will be reflected in any "confidence scores" assigned.
Modelling methodology	There is a trade off in the depth of the neural network (which requires greater processing and potential overfitting) and performance of the model	
Modelling methodology	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.	The neural network will have a backbone of several RGCN (or similar) graph convolutional layers followed by several MLP heads taking the asset node embeddings as inputs.

Modelling methodology	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.	Take forward the Spatial Graph methodology for further development in the PoC.
Modelling methodology	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.
Modelling methodology	For the purposes of the PoC, 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.	Numerical attributes to be predicted, e.g. conductor rating, must be binned into fixed ranges.
Modelling methodology	The GNN approach developed for the voltage attributes mentioned above also gives good performance on the selected specification parts. No changes were made to the model architecture except for a small increase in the width of the hidden layers.	Move on to formal model evaluation.
Modelling methodology	Evaluation of the spatial model has shown that the performance is good for the PoC. 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 GNNs are a good approach for data cleaning for asset management.	Define Next Steps to further investigate, develop and extend the GNN-based model.
Modelling methodology	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.	"Quick wins" described in Next Steps for the project.
Modelling methodology	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.	"Combine models" described in Next Steps for the project.

Modelling methodology	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.	An electrical connectivity model will be utilised to approximate network topology where real power demand and supply data can be used to find feasible flows within our network GIS and customer data. The model will need to be configured graph: - Peak power demand at customer nodes - Aggregate supply at substation node - Wire and cable segments with capacity
Modelling methodology	In the case of 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.	When the graph has been built for each circuit with customers connected and substation located, a new super source and sink will need to be added to the graph with the demand and supply attribute on the nodes transferred to capacity attribute on synthetic edges from the actual to super nodes.
Data	There is no connectivity data at LV level; whilst some connectivity can be implied for end / start exact coordinate overlap between GIS cable and wire segments, there are disconnects in the circuits. These disconnects are either due to existence of other assets not currently in the model or as an artefact of the digitalisation of paper records. Most of the disconnects are happening where the ends of wires are not exactly meeting; some are due to wires not extending to within other wires. Our digital team are doing work to fix some extension based disconnects.	To explore some methods of connecting disconnects to increase the number of complete, connected circuits that can be analysed as part of this workstream.
Data	Vertex to vertex disconnects in the datasets tend to have shorter distances compared to distanced between other nodes in the circuit. The number of clusters should be calculated such that the clusters join up the disconnected nodes of separate connected graphs to create a single connected graph for each circuit. This works well for cases where there are micro-disconnects and the distances between the disconnected nodes are shorter than actual disconnected unconnected nodes.	To use K-means to connect isolated graphs and take completed connected graphs. Noting that there may be false positives within the sample if the disconnects which should be connected are where nodes have a greater distance than those which should remain disconnected.
Modelling methodology	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.	For a simple transportation problem which is suitable for studies without the need for considering extreme events maximum flow can be use in the reconciliation process / data verification process of the technical feasibilities of the circuit data.

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.	The connectivity graph will need to be undirected graph and any method chosen will need to be robust to the lack of direction / parent-child relationships within the data.
Modelling methodology	The method can be used to highlight particularly important assets that have a high impact on the circuit, i.e. where there may be potential bottleneck in a circuit and verification is required that its specifications are correct for the technical operation of the circuit. The method is also useful to ensure that the most critical assets are highlighted.	This model / methodology can be used on partially complete / imperfect data and will be able to assess where the most important attributes are required for operation of the circuits.
Modelling methodology	The method is 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.	This model / methodology can be used on partially complete / imperfect data to highlight where data errors / wrongly connected customers may be causing violations.
Modelling methodology	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.	Some ownership / feedback loop is required in order for the full benefits of the use of this model.
Data	The ability to eliminate reasons for violations (customer wrongly assigned, profile class wrongly assigned, EAC 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 is required (potentially as an input from model 2); as well as assessment / review of the missing / synthetic service cables.
Data	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.	The 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.

10. 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 PoC 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 PoC 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 BAU 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.

11. Data Access Details

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 our 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/energy-data-hub
Detailed network plans are available via our Data Portal. https://www.westernpower.co.uk/our-network/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-and-more/Project-Data.aspx.

12. 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.

13. 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 plan[5]. 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.

[5] Hyperlink to data and digitalisation action plan on website.

14. Contact

Further details on this project can be made available from the following points of contact:

Innovation Team
Western Power Distribution,
Pegasus Business Park,
Herald Way,
Castle Donington,
Derbyshire
DE74 2TU
Email: wpdinnovation@westernpower.co.uk

Glossary

Abbreviation	Term
Artificial Intelligence (AI)	The training of computer systems with human intelligence traits like learning, problem solving, and decision making.
Command-line interface (CLI)	A text-based user interface used to view and manage computer files.
CROWN	Our enterprise asset management system. Holds data about assets which includes data defining the assets, condition data and defect data. It also records inspection and maintenance activities on the assets as ‘events’.
Data Cleanse	The action of identifying and then removing or amending any data within a database that is incorrect or incomplete.
Electric Office (EO)	Our geospatial system which displays the network layout at all voltages
Exploratory Data Analysis (EDA)	Analysing data sets to summarise their main characteristics.
Geospatial Information System (GIS)	A data system capable of capturing, storing, analysing, and displaying geographically referenced information.
Integrated Network Model (INM)	Our combined dataset for 11kV and above that merges data from CROWN, GIS and PowerOn.
Machine Learning (ML)	A subset of AI, the study and application of algorithms that improve automatically through experience.
Meter Point Administration Number (MPAN)	A unique 21-digit reference number used in the UK that identifies each electricity supply point.
PowerOn	Our distribution management system used for system operations.
Proof of Concept (PoC)	An exercise or demonstration to verify that concepts or theories have the potential for real-world application.
Python	An open-source general-purpose programming language.
QGIS	A free and open-source cross-platform desktop geographic information system (GIS) application that supports viewing, editing, and analysis of geospatial data
SPEN	Scottish Power Energy Networks
SSEN	Scottish and Southern Electricity Networks
Unique Property Reference Number (UPRN)	A unique number (1-12 digits in length) created by the Ordnance Survey for every addressable location in the UK.
User Acceptance Testing (UAT)	Testing by the user to confirm that the functionality of the system matches the specified functionality
User Interface (UI)	The means by which the user will interact with the model.
Well-known text (WKT)	A text markup language for representing vector geometry objects.

Western Power Distribution (East Midlands) plc, No2366923
Western Power Distribution (West Midlands) plc, No3600574
Western Power Distribution (South West) plc, No2366894
Western Power Distribution (South Wales) plc, No2366985

Registered in England and Wales
Registered Office: Avonbank, Feeder Road, Bristol BS2 0TB

wpdinnovation@westernpower.co.uk
www.westernpower.co.uk/innovation

WESTERN POWER
DISTRIBUTION
Serving the Midlands, South West and Wales