WPD INNOVATION
Transforming the electricity network

Spatially Enabled Asset Management (SEAM)

D07 Model Evaluation Report

WESTERN POWER DISTRIBUTION
Serving the Midlands, South West and Wales

Version Control

Issue	Date
1.1	07/06/2021

Publication Control

Name	Role
SEAM project team	Author
Jenny Woodruff	Reviewer
Jenny Woodruff	Approver

Contact Details

Email
wpdinnovation@westernpower.co.uk

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

Disclaimer

Contents

1. Executive Summary

The Spatially Enabled Asset Management (SEAM) project aims to investigate the use of advanced analytical techniques to identify and resolve Geographic Information System (GIS) errors. Inaccuracies in Western Power Distribution's (WPD) 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, one that focussed on creating a connectivity model which was then tested for its capacity to carry the expected loads, the other that recognised that in many cases connectivity information was incomplete, and therefore took a spatial approach to confirming and proposing key asset attributes.

SEAM has developed a Python-based Machine Learning (ML) tool with an Excel front-end which has been installed and tested on a standard WPD laptop with the Anaconda3 software package installed. The models have been 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. Analysis of the results from training and testing the model suggest that both are successful in providing a view of issues that could affect the GIS data.

Model 1: Customer Connectivity Model

The first 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. Errors such as missing normal open points or incorrect cable size could result in sections of the network being shown as overloaded when the network capacity is assessed. Rather than carry out a full power-flow analysis, a max-flow algorithm is used that produces similar results with a simpler, faster algorithm that does not require a separate software package and licence and can be directly created from raw data from disparate systems. Further, making the digital link between customer and asset will potentially enable WPD to make a more informed customer response to queries and complaints.

The algorithm to create the connectivity model identified that for a large proportion of circuits the data was complete and correct enough to build a single graph model. However, the algorithm was also able to detect and correct disconnects in the underlying data that resulted in the feeder initially generating multiple separated graph models. Similarly, where the model connected customer locations to the existing network, this sometimes involved a small extension to the existing modelled service cable, but in other cases, where a simulated service cable needed to be added, this could suggest missing data in the network model (in areas where other service cables were present) or, in areas where service cables were historically included in the network plans, it provided a default value to enable further modelling. The max-flow algorithm itself was seen to be a useful and fast means to identify network anomalies, though many of the anomalies found reflected where missing data had been backfilled in the process to create the connectivity model. Taking an iterative approach where model the model is run, errors investigated and corrected and the model is run again, is expected to give increasingly useful results.

Model 2: Spatial Graph Model

The second model works by creating a spatial graph that represents all EO assets as nodes with feature vectors that encode the attributes of interest, plus additional nodes and edges to represent the point locations of those assets and the relationships between the assets and locations and between the locations, then using a graph neural network machine learning model to predict the correct values of the attributes of interest for all of the asset nodes as a node classification task. The 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.

The model can correctly predict values for the network type, operational voltage, specification material and specification size attributes for all asset types (where applicable) using only these attributes and the asset geometry as inputs. Using the synthetic errors to measure performance showed that the prediction accuracy was high for all cases and very high for most cases with no underlying errors or missing input data. Using the original data, some

individual errors and patterns of errors were identified for investigation as well as some patterns of false alarms. Some simple model improvements have already been identified that should reduce these false alarms. The model is fast to run, with both the model tuning and prediction run in a reasonable time on a standard WPD laptop. The spatial graph model is designed to be a flexible foundation for adding more attributes and relationships of interest and additional data sources and adding more data to the model will improve the quality of all of the predictions, as well as providing additional functionality.

Summary of recommendations and conclusions

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). 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.

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 PoC geography. There were no existing errors from INM in the Barnstaple area that were directly comparable to the errors identified by SEAM. However, SEAM could complement INM by addressing a different set of error types (i.e. the analytical approaches used by SEAM can be used to address errors and improve the underlying GIS data that won’t be picked up by the user-defined validation rules and matching process in INM).

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 recommended set of steps are carried out as part of future development based on gathering feedback from users, implementing “quick win” updates and combining the models. These would all significantly improve the accuracy and performance of the current models for a relatively low additional amount of effort.

The SEAM models have a flexible design that can be incrementally enhanced and extended. In the next phase of development, it is recommended the scope of the model is “scaled-up” to maximise its value as a BaU tool. This includes extending the geographical coverage, introducing additional data sets, and enhancing functionality of the models. There are many different options which are likely to require more effort to implement due to the increased complexity (e.g. the likely prevalence of varied data issues and structures across the different licensed regions due to the history of each network and its ownership and approach to data management). However, these changes are flexible and can be prioritised based on the potential value they will deliver.

The scope of SEAM has been focussed on cleansing EO attributes. In the next phase of development, the underlying graph models could be exploited to pursue additional use cases. The project has considered potential additional uses which in the longer-term could be implemented to create further value from the models.

2. Context and purpose of this document

The purpose of the Model Evaluation report is to review the performance of the two proof-of-concept (PoC) models and summarise key findings. It follows completion of the model development and creation of the final output reports; and provides a detailed description of the findings that will be summarised in the Project Closedown report. The evaluation report covers the following topics:

- Review of results against the project success criteria
- Summary of the key findings from the two SEAM proof-of-concept models
- Detailed description of the model output reports
- Outline of key conclusions from the results for the proof-of-concept area and investigations into sample errors
- Analysis of model outputs for independent test area (not used to train the models)
- Comparison of SEAM with errors identified by the Integrated Network Model (INM)
- Recommended next steps to improve model performance and transition to business-as-usual

The full model output reports that have been evaluated in this report are part of deliverable D06 Cleansed Datasets.

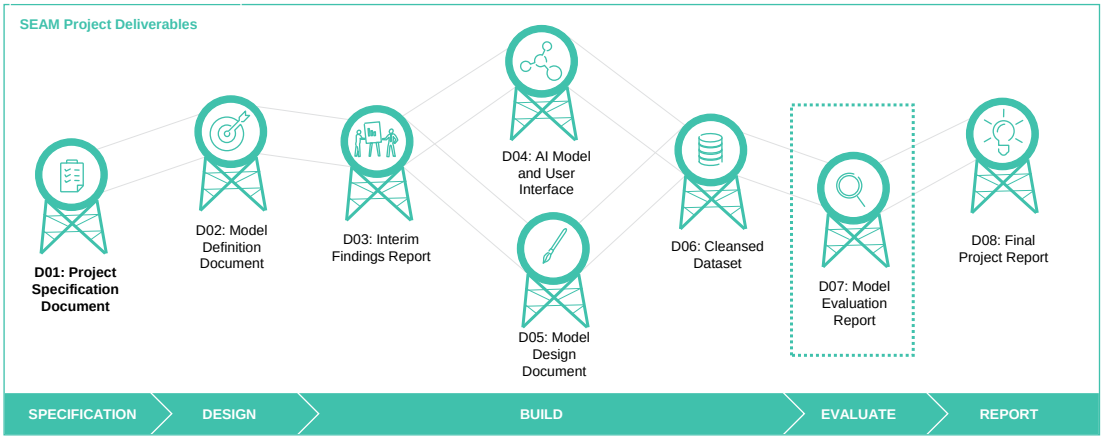


Figure 1: SEAM Project Deliverables

3. Summary of key findings

3.1. High-level

A review of the project success criteria covered by the discussion in this evaluation report are summarised in the following table:

Criteria	Met (Y/N)	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 this report (see Section 7).
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 7). • 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.

Table 1: Project Success Criteria

The remaining success criteria based on sharing of the project learnings with other Distribution Network Operators is not considered directly in this report and will be completed once the final closedown report is published and the project dissemination webinar has taken place.

3.2. Model 1: Customer Connectivity Model

A summary of the key findings for Model 1:

- Max flow is fast (data preparation and post processing phases take the majority of the model running time) for a simple transportation problem which is suitable for studies without the need for considering extreme events and useful within the reconciliation process / data verification using the technical feasibilities of the circuits.
- 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.
- 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.
- 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.
- 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.
- 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.

3.3. Model 2: Spatial Graph Model

A summary of the key findings for Model 2:

- It is possible to train an inductive graph neural network-based machine learning model to identify and correct missing and erroneous data in a power distribution network.
- The model can produce useful suggestions using only a limited number of attributes and the geospatial relationships, which all come directly from the Electric Office (EO) dataset.
- The model has identified some individual cases and some groups of cases where there appears to be errors in the EO extract used for the project.
- The overall performance of the current model supports the aim and learning objectives of the PoC. Further enhancements to the model performance should be considered as part of a transition into Business-as-Usual (BaU).
- Enhancing the model with additional attributes, nodes and edges is expected to incrementally improve the performance. In particular, some candidate enhancements have already been identified that should improve the performance for the main groups of incorrect predictions discussed in this document.
- The spatial graph structure and graph neural network are a flexible basis for adding a range of data and predictions.
- The model is fast: prediction runtime is dominated by reading and writing the GIS files and pre-processing the geospatial data, and model training can be completed in a reasonable time on a standard laptop (for the scope of this PoC).

4. Evaluation of model results

This analysis is based on the “training” area from the PoC region. This given the bounding box with X in [248000, 264000] and Y in [126000, 141000]. Please note that the model results have not been validated by the mapping team within WPD and in some cases data shown as being absent may be present elsewhere as part of a different attribute.

4.1. Model 1: Customer Connectivity Model

Model Processes

Common errors in WPD’s GIS data, such as assets that are connected in real life not being shown as connected in the GIS data. This can prevent the operation of network tracing functions to highlight the extent of a feeder or a particular section of network, and also affects the generation of network models that can be used for power flow analysis. The Customer Connectivity Model can be used to identify incorrectly disconnected sections of network and can also use a capacity validation technique to identify other anomalies. E.g. if an LV feeder has been shown as incorrectly supplying an additional separate feeder due to the omission of a normal open point, this would result in the network being shown as overloaded. Similarly, if a cable type was incorrectly attributed to one with a lower capacity, this would result in a bottleneck being identified by the capacity validation algorithm.

Model 1 is a customer connectivity model which connects Electric Office linear assets (cables and wires), power source point assets (pole mounted and ground mounted substations, link boxes, etc.) and customers (extracted from CROWN). The model takes the technical feasibility of power transportation modelling to verify the composite data sources and find exceptions and technical violations. A max-flow algorithm is used in the model in preference to passing the graph model to a separate tool for power flow analysis as it produces similar results without adding the to the processing time, complexity and licencing costs that would be involved in interfacing with a tool such as LV Connect. There is currently no connectivity data at the low voltage level, apart from that which can be implied by the location of assets extracted from the GIS data. At a high level, missing assets, exceptions and other results are found for these processes:

- Circuit connectivity
- Customer connectivity
- Substation location
- Transportation / maximum flow reporting

Model Output Report

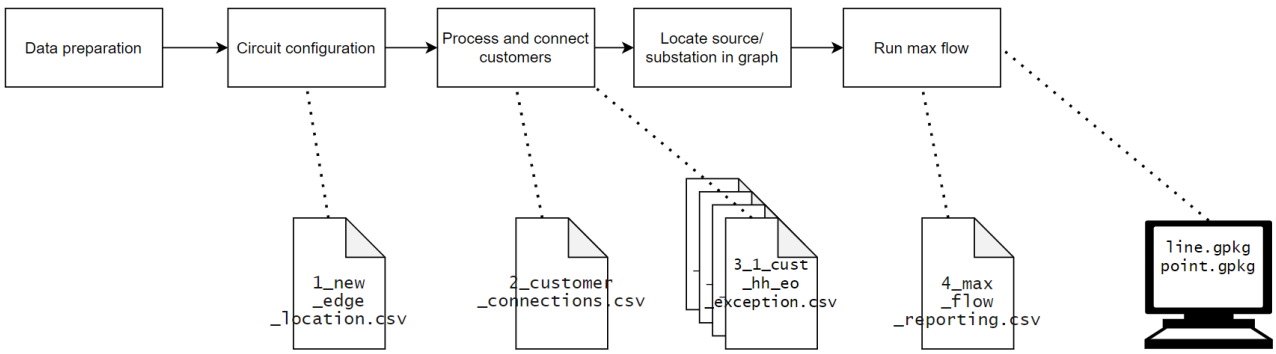


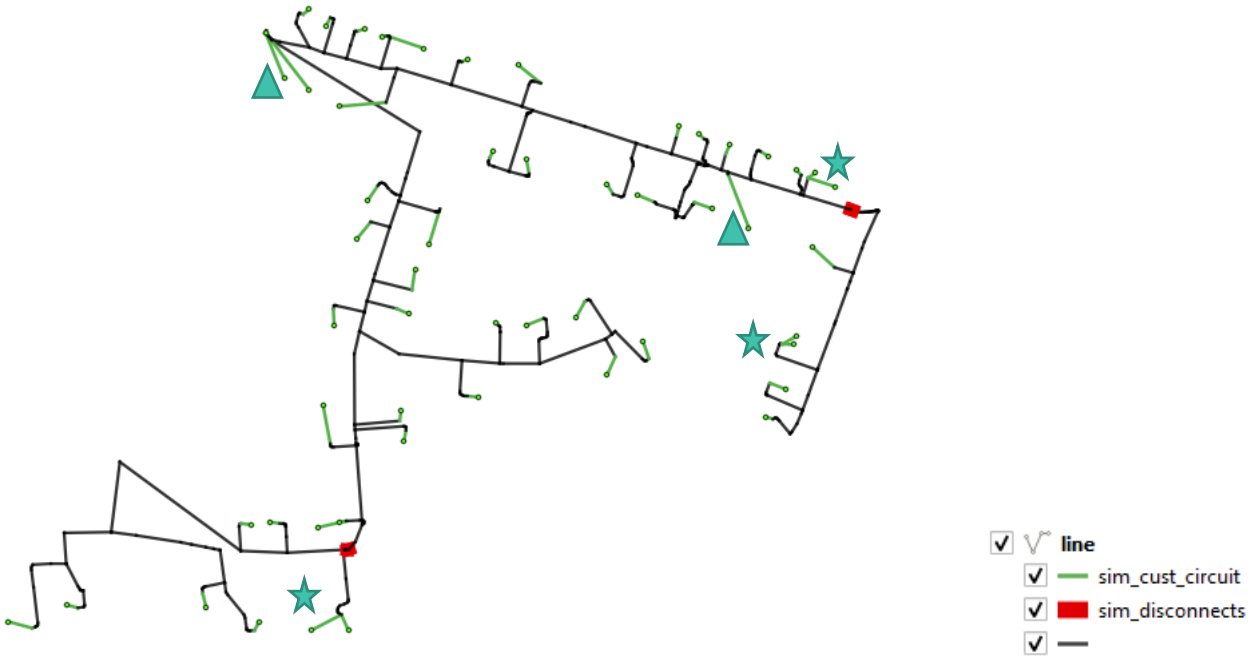
Figure 2: Model 1 high-level process and output reports

The modelling and analysis conducted as part of the customer connectivity model follows the process shown in Figure 2 . Further details of this process are given in the SEAM Model Design report[1].This generates a set of CSV reports to be used alongside the GeoPackage files. Table 2 provides a detailed description of the output reports for Model 1:

Report Name	Purpose	Description
1_new_edge_location	This report contains all edges added to each circuit (**circuit_id**) as part of the circuit config stage. Edges are added if the addition of this set of edges connects all isolated graphs for the circuit with **circuit_id.** These are indicators of where there may be microdisconnects in the underlying data.	**circuit_id** is the Electric Office **circuit_id** **new_edges** is the new edges added to the circuit and is in the format {(node_from_1): [(node from_1_to_node_1), (node from_1_to_node_2),…, (node_from_1_to_node_n)], …,}
2_customer_connections	This report contains all edges added to each circuit (**circuit_id**) as part of the customer connectivity stage. All edges are kept, so that the user can use this report to cross reference against the reports '**line**' and '**point**' and decide on a threshold distance by filtering using standard data processing software such as Excel. This information would help data stewards identify unexpectedly missing service cables in areas where this was generally available.	**circuit_id** is the Electric Office **circuit_id** **circuit_node** is the node on the existing circuit graph which corresponds to the closest node to the customer on the matching **circuit_id** (in the format WKT) **customer_node** is the customer node (in the format WKT) distance is the distance in metres from circuit node to customer node **customer_mpan** is the customer mpan **customer_uprn** is the customer uprn **customer_id** is the internal id generated by this process to differentiate unique customers that may have the same mpan and uprn
3_1_cust_hh_eo_exception **3_4_cust_eac_eo_exception**	These reports contain customers (half hourly and estimated annual consumption) which were not matched to any of the circuits (**circuit_id**) Again this would be indicative of missing service cable data though it may also indicate incorrect network association within CROWN.	**customer_sub** is the substation which is provided for each customer **customer_lv_feeder** is the lv feeder which is provided for each customer **_circuit_id** is the internal **circuit_id** which is generated by: [**customer_sub**]/[**customer_lv_feeder**] from crown and substation and lv feeder elements from **circuit_id** from Electric office
3_2_cust_hh_eo_matching **3_3_cust_eac_eo_matching**	These reports contain customers (half hourly and estimated annual consumption) which were matched to circuits (**circuit_id**).	Same as above

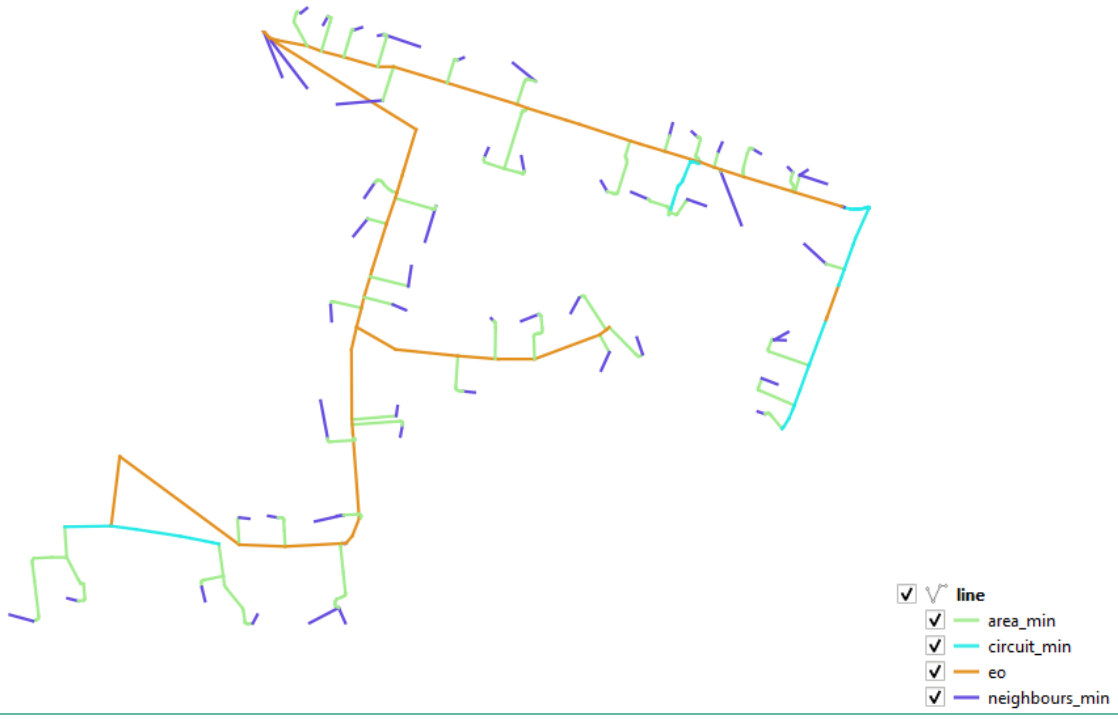
[1] Link to SEAM Model Design Report

Report Name	Purpose	Description
4_max_flow_reporting	This report shows the metrics of each stage of building the circuits per **circuit_id** and the number of customers and wires that are exceptions within each circuit. This report is intended to be used alongside '**line**' and '**point**' to investigate exceptions.	**circuit_id** corresponds to Electric Office **circuit_id** **input_isolated_graph_no** is the number of isolated graphs (when built from wires and cables) for each **circuit_id** **output_isolated_graph_no** is the number of isolated graphs for each **circuit_id** after joining disconnects process **edges_added** is the number of edges added during this process **number_of_customers** is the number of customers found and connected to the circuit **distance_to_sub** is the distance in meters from the nearest node in the circuit to the centroid of the substation geometry **nearest_node** is the nearest node in the circuit to the centroid of the substation geometry type is the type of substation found at the **nearest_node** **n_cust** is the number of customers whose demand is not met **n_headroom** is the number of wires and / or cables which have headroom below the absolute threshold set by the user **n_headroom_pc** is the number of wires and / or cables which have % headroom (i.e. (capacity – flow)/capacity) below the % threshold set by the user
line	This Geopackage comprises of linear assets within Electric Office which were used in the connectivity modelling, detailing a number of attributes which were generated / inferred / backfilled during the modelling process as well as the results of the max-flow analysis. The report is intended to be viewed and analysed alongside point within GIS software and symbology used to visually inspect outputs and results to verify exceptions produced as part of the report **4_max_flow_reporting** The report also can be used to understand where capacity backfilling was applied and the techniques which were used.	**circuit_id** corresponds to Electric Office **circuit_id** **network_type** corresponds to Electric Office **network_type** i.e. LV, MV, HV… usage corresponds to Electric Office usage i.e. Distribution, Service … **specification_description** corresponds to Electric Office cable **specification_description** and Electric Office wire **specification_description_1** size is the extracted and normalised size from the **specification_description** feature in mm^2 **eo_type** is the layer name of the asset capacity in kW corresponds to the Electric Office nominal voltage * WPD Directive rated current matched on the basis of Electric Office columns if this value has been found, otherwise it takes the **capacity_backfill** value **capacity_backfill** is the value backfilled as part of data preparation and circuit configuration stages in the circuit building procedure. **capacity_backfill_type** is the method the unknown capacity for the asset has been backfilled (example below) **gen_type** is, where applicable, the stage at which the asset was generated (example below) **specification_description** and **specification_description_1** for the layers cables and wires, respectively **head_room** is the capacity – optimal flow through the asset **head_room_pc** is the (capacity – optimal flow)/capacity)



Above: **`circuit_id`** = 260654/0/0010 Legend: feature **`gen_type`**, which is labelled according to when the asset was generated.
sim_cust_circuit: simulated customer to circuit;
sim_disconnects: simulated connecting disconnects;
blank: electric office origin

In the example above the data steward would check the disconnects marked with the red boxes. Stars have been added to the image above to show locations where there is more than one customer associated with a service point. This would allow for looped services to be identified and confirmed. (Looped services are likely to be of increasing concern as customers install heat pumps and electric vehicle charging points.) The data steward would also be interested in the services that had been added as simulated customer circuits (marked with triangles added to the image above). Whereas most of the simulated customer circuits are short, extending a marked LV service cable to the location associated with the customer's UPRN, these connections appear to connect the customer to the LV main, suggesting that an LV service cable has been omitted from an area where these have historically been recorded.



Report Name	Purpose	Description
Above: **circuit_id** = 260654/0/0010 legend: feature **capacity_backfill_type** categorised as shown in the key. Each line asset has the feature **capacity_backfill_type** which is labelled according to how the capacity backfilling was applied. area_min: minimum per usage_type the whole analysis area; circuit_min: minimum per usage_type for the circuit; eo: capacity found using specification_description from Electric Office; neighbours_min: minimum of neighbours		
point	This Geopackage comprises of point assets within Electric Office which were used in the connectivity modelling; specifically customers, **substation_pm** and **substation_gm.** The report is intended to be viewed and analysed alongside line within GIS software and symbology used to visually inspect outputs and results to verify exceptions produced as part of the report **4_max_flow_reporting**	**eo_circuit_id** corresponds to Electric Office circuit_id **MPAN** corresponds to customer MPAN **UPRN** corresponds to customer UPRN **_id** is an internally generated ID to differentiate customers who have the same MPAN, UPRN but different EAC and / or different profile class and **Half_hour_demand** corresponds to peak demand in kW **Cust_type** corresponds to the type of demand available for the customer i.e. EAC from Falcon and HH for half hourly customers Flow corresponds to the amount of power routed to the customer or outgoing from substation **Demand_not_met** corresponds to demand – flow Type corresponds to type of point object, customer, substation_pm or substation_gm

Table 2: Overview of Model 1 output reports

Results

Connecting disconnects

Circuit IDs are labelled in Electric Office by components of a circuit which are downstream from a unique substation and LV feeder. Each record within Electric Office is a separate segment of cable or wire for line asset and are labelled with a circuit id. These segments are connected and processed as connected graphs where the points at the end of the lines are connected if they have exact coordinates. Not all connected assets have exact points which overlap to infer connectivity and small disconnects exist in the data which is a result of lines not extending to connected segments.

Before the connectivity process, 83.3% of all circuits in the Barnstaple area were single isolated graphs, i.e had no micro disconnects for each circuit_id. The percentage of connected circuits is increased to 91.5% (108 additional circuits) after the connectivity process, Table 3 details the count of circuit_id for the input graphs and output graphs.

Number of isolated graphs	Count circuit_id (input graph)	% Count	Count circuit_id (output graph)	% Count
1	1090	83.3%	1198	91.5%
2	69	5.3%	65	5.0%
3	40	3.1%	25	1.9%
4+	110	8.4%	21	1.6%

Table 3: Number of isolated graphs per Electric Office circuit_id, input (before connectivity process) and output (after connectivity process)

When asset data was first digitalised from paper records to GIS data, the process of digitalisation may have introduced a number of types of disconnections as this process was created not necessarily to facilitated automated / digital electrical connectivity. Therefore, not all electrically connected assets will have exact coordinates required to imply connectivity directly from the GIS data. As such, there are many ‘electrical disconnects’ in the GIS data.

Within the scope of this project, vertex to vertex disconnects were explored; these have the characteristic that cables / wires end nodes are a small distance from other cable / wire end nodes. These disconnects could be because of disconnects in the GIS data arising from error in the way the records were digitised, from assets (such as cables being

connected by link boxes or cables and wires stopping at the boundary edge of conduits) not seen by the model or due to actual existing micro-disconnects in the physical asset.

The method depends on the micro-disconnects to be vertex to vertex on ends of line segments, and is successful at connecting isolated graphs which are disconnected in this way. See Figure 3 and Figure 4 for examples where, by inspection, the disconnects appear to be bridged by the addition of new edges correctly. Inspection and analysis of the additional edges by an engineer / subject matter expert is required, on an edge by edge basis. There may be disconnects that are connected which seem reasonable, however the disconnect exists in reality and is due to some configuration of the circuit.

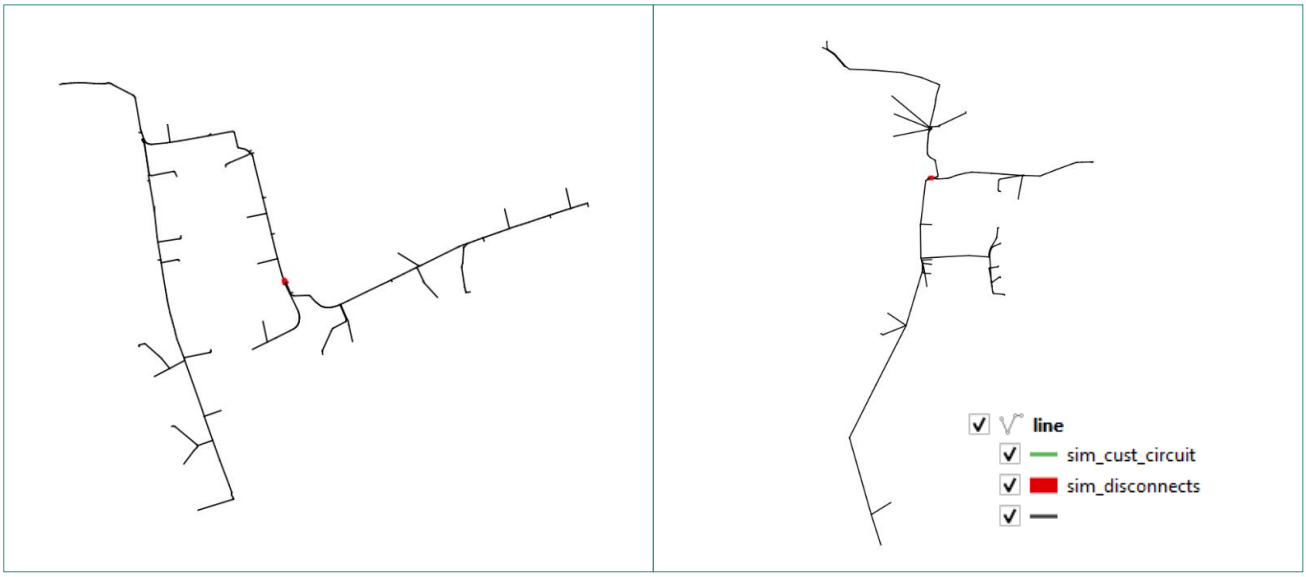


Figure 3: Left: Circuit_id = 260717/0/0030, Right: circuit_id = 260219/0/0020; 4 edges added input_isolated_graph_no = 4 output_isolated_graph_no = 1

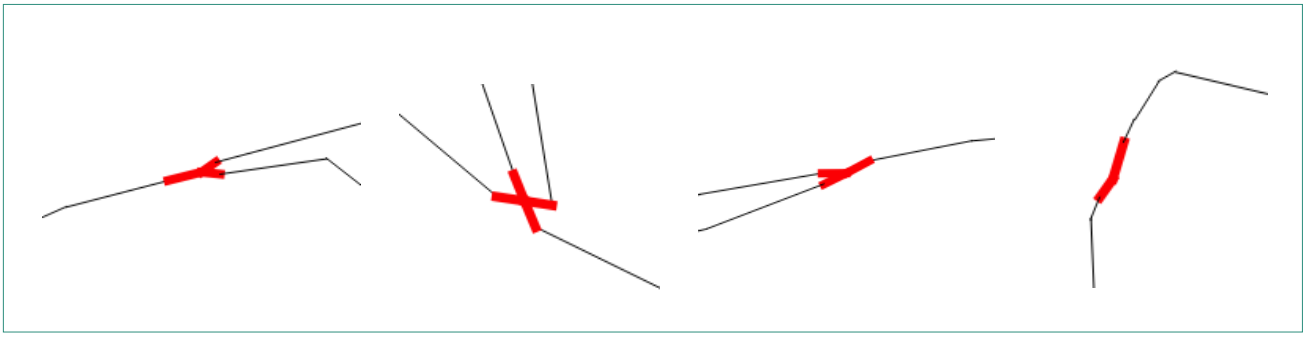


Figure 4: Disconnects which appear upon inspection to be correctly connected, circuit_id from L to R: 263168/0/0010, 263219/0/0040, 263144/0/0040, 260305/0/0030

Some circuits were not fully connected through this method due to the disconnect between isolated graphs being a line to vertex problem, which within the timescale and scope of this study were not studied in more depth. This sort of disconnect is due to a gap in two line geometries where a line does not extend to another line; see figure below for the illustrative diagram of the two types of disconnects which may occur in the GIS data.

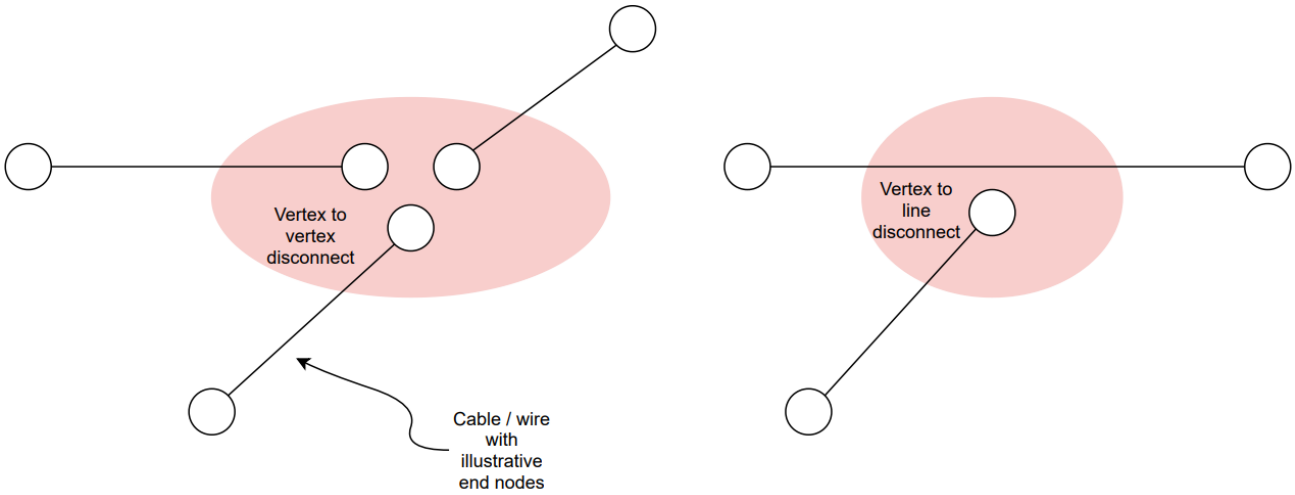


Figure 5: illustration showing the difference between a vertex to vertex disconnect and a vertex to line disconnect

For circuit_id = 260590/0154, the number of disconnected isolated graphs before the connectivity process was 22, the procedure then connected all disconnects apart from a vertex to line problem, leaving the circuit with two isolated connected graphs. See Figure 6 for reference.

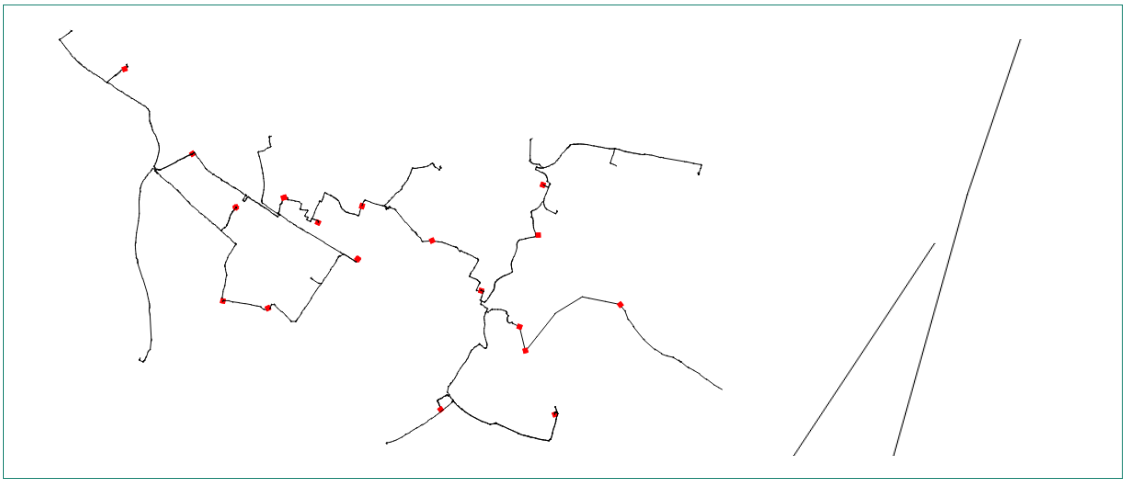


Figure 6: circuit_id = 260590/0154, left showing the successful 39 simulated edges added to make a connected graph, right showing the single disconnect in this circuit which was not connected due to the line to vertex disconnect

The below table shows the detailed summary of output isolated graphs and number of edges added to the circuit in the connectivity process (minimum and maximum for each group). The location of the new edges can be found in the report `1_new_edge_location`.

Number of output isolated graph	circuit_id count	Barnstaple (%)	Number of edges added (min)	Number of edges added (max)
1	1198	91.50%	0	21
2	65	5.00%	2	39
3	25	1.90%	3	44
4	4	0.30%	6	47
5	2	0.20%	14	40
6	6	0.50%	12	71
8	1	0.10%	20	20
11	1	0.10%	59	59
16	1	0.10%	139	139
17	2	0.20%	107	115
18	1	0.10%	115	115
24	1	0.10%	113	113
25	1	0.10%	169	169
79	1	0.10%	194	194

Table *4*: Detailed summary of output isolated graphs and number of edges added to the circuit in the connectivity process (minimum and maximum for each group)

Customer connectivity

Connected customers

Customers in CROWN have an associated UPRN, MPAN, substation and LV Feeder. The substation and LV feeder numbers were used to match against the Electric Office circuit_id to find relevant connections of customers to circuits. The customer's geolocation, provided by matching customer UPRN against the UPRN database, was then used to connect customers to the closest end of line segment within each circuit. All customer connections are kept within the model so that the data steward can use the 2_customer_connections report alongside the geopackages as a reference to examine the connections on a connection basis. The Table below shows an extract from the report with customers that are over 400m away from the nearest node in the circuit. These connections are suspicious and should be flagged up to data stewards as customers are normally connected within 400m of the distribution transformer due to the voltage drop on LV cables. However, the scenario represented by this report is even more extreme as customers are over 400m away from the nearest node in the circuit, which may already be some distance from the distribution transformer. This suggests that the association with the distribution substation as held in CROWN may be incorrect.

circuit_id	circuit_node	customer_node	distance	customer_mpan	customer_uprn	customer_type
261518/0/0010	POINT (255853.91 133159.97)	POINT (262137.445 130031.865)	7019.1	2200014007048	100040245094	EAC
260612/0/0010	POINT (256880.39 129855.24)	POINT (256297.459 132610.065)	2815.8	2200014309790	100040251818	EAC
260612/0/0010	POINT (256873.5 129861.06)	POINT (256297.459 132610.065)	2808.7	2200014309806	100040251817	EAC
262455/0/0010	POINT (255793 141548)	POINT (254950.226 140525.127)	1325.3	2200014155916	10012101838	EAC
260231/0/0010	POINT (257993 135419)	POINT (257169.916 135437.613)	823.3	2200014145341	10000489074	EAC
262260/0/0020	POINT (246516.8 130374.39)	POINT (247252.3 130224.597)	750.6	2200014122089	10023352882	EAC
264563/0/0010	POINT (254610.27 133572.56)	POINT (254031.644 133849.214)	641.4	2200040645138	10012109567	EAC
260179/0/0010	POINT (253795.37 130586.06)	POINT (254107.642 131037.257)	548.7	2200014015642	10012109736	EAC
265109/0/0010	POINT (256234.98 132264.38)	POINT (256181.279 131772.846)	494.5	2200040024886	10012101659	EAC
261708/0/0020	POINT (248823 136087)	POINT (248956.023 136519.616)	452.6	2200014248845	100040258977	EAC
262530/0/0010	POINT (260382 135755)	POINT (260814.164 135655.714)	443.4	2200014143308	10012096204	EAC
263316/0/0020	POINT (264467 137732)	POINT (264100.336 137500)	433.9	2200014152534	10012094583	EAC
264462/0/0010	POINT (263092 136626)	POINT (263076.856 136202.972)	423.3	2200014151521	10000489228	EAC
353221/0/0010	POINT (245825.49 130016.9)	POINT (245958.921 130403.981)	409.4	2200011553814	100040373263	EAC
260041/0/0020	POINT (255626.74 134158.49)	POINT (255223.615 134196.058)	404.9	2200014240855	10012099381	EAC
263316/0/0020	POINT (264444 137713)	POINT (264100.336 137500)	404.3	2200014152701	10012093793	EAC
262286/0/0010	POINT (255603.9 126488.22)	POINT (255224.724 126622.958)	402.4	2200013980053	10012090469	EAC

Table 5: Customers that are > 400m away from the nearest node in the circuit with its associated substation and LV feeder

Circuits and customer exceptions

There were customers which were not identified to match any of the existing circuit_id and vice versa, these exceptions are presented in the exception reports 3_1_cust_hh_eo_exception and 3_4_cust_eac_eo_exception for half hourly customers (Data provided by Durabill) and estimated annual consumption customers (Data provided via the P222 text file data exchange) respectively.

Transportation Modelling / Maximum Flow results

For the 758 circuits with substations located within thresholds and customers connected, 30 circuits were found to have power flow violations with customers not being supplied their full demand and 61 circuits were found to have cables/ wires with head room percentage below threshold set at 20%, using minimum aggregation for capacity backfilling. See Appendix 1 for full results; the next section will explore some of these in more detail.

Observations

Generally, violations where customer demand is not satisfied occurs to a small number of customers at the end of a lines, where the cable upstream / closer to the substation is the bottleneck for distribution of power supply and is the aggregate effect of all customers downstream. Hence, customer violation on an individual basis is not often interesting as this is usually the effect of a cluster of customers or the configuration of the circuit. Where a network has been incorrectly extended, e.g. due to the omission of a normal open point, then we would expect the network sections to be overloaded near the end of the “real” circuit as well as close to the substation.

Many of the violations are happening due to the way capacity is being backfilled for unknown current ratings for wires and cables at LV, where there is data on the wire and cable specifications, this is not conforming to directives. The cable capacities are then backfilled according to the circuit aggregation method chosen by the user, if this is not available then area wide is used; this is potentially a low estimation compared to capacity in reality.

There are some circuits where configuration looks reasonable upon examination and close to threshold at the feeder where headroom is low due to demand supply requirements. This could be potentially alluding to sections of circuitry

where further analysis may be necessary to understand whether this is caused by the GIS data, connectivity, capacity backfilling or where modelled demand may be the root cause.

There are also circuits where there are clearly missing cables / wires connecting to customers and as such the load is concentrated on a line asset where in reality this may be distributed. This should be investigated by a data steward and / or engineer to understand where the additional cables / wires are located, once this is added to the Electric Office dataset, the model can be initiated with the new data to confirm if this violation has been solved.

All violations require further root cause analysis by an engineer / subject matter expert in order to understand the nature of the error since it could be due to a number of reasons:

- Error in the configuration of the circuit
- Under-estimation in capacity backfilling
- Over-estimation of demand (due to demand profiling)
- Customer wrongly assigned to the circuit / feeder / cable
- Actual design of circuit is close to design threshold / asset management required

Investigation of sample errors

This section sets out a sample of detailed examples of violations:

circuit_id = 262417/0/0020

In this circuit, customer with UPRN = 100041037652 has 29% unmet demand, with the cable connecting the customer at 0 headroom. The connection between the substation to this customer appears to be reasonable and due to the non-conforming specification_description field, the capacity has been backfilled by using the area minimum (i.e. Barnstaple minimum, due to the same type of cable not being available in this circuit). This particular exception might be interesting due to the unusually high demand from this half hourly metered customer: 40.3 kW on 17th December.

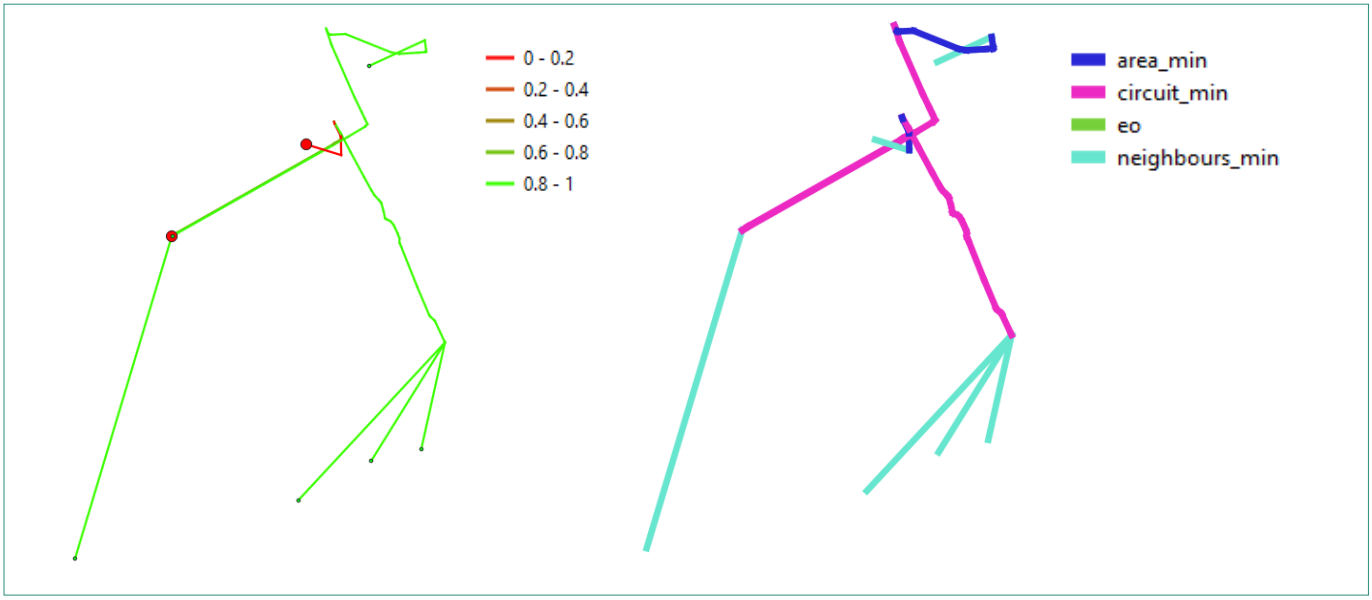


Figure 7: circuit_id = 262417/0/0020: left results from the maximum flow with legend as % percentage head room. Right: legend 'capacity_backfill_type'

circuit_id = 263293/0/0020

The configuration for this circuit looks fairly reasonable; with no obvious cases of significant missing cables as well as customers being distributed along cables and feeders. The customer unmet demand is occurring at the end of a long distribution cable and the cable bottleneck is at a cable directly connected to the substation. The main constraint in this violation is the capacity of the cable directly feeding from the substation, which is backfilled as its specification_description is non-conforming. The cable with specification_description = '185 3c CON' has been backfilled by the minimum capacity of the circuit at 3 phase, distribution, LV: the capacity is 108.4 which corresponds to a cable downstream which is labelled and conforming for the specification_description = '95 3c WCON'; which could be an underestimation in the actual capacity if the specification_description field entry is a data error.

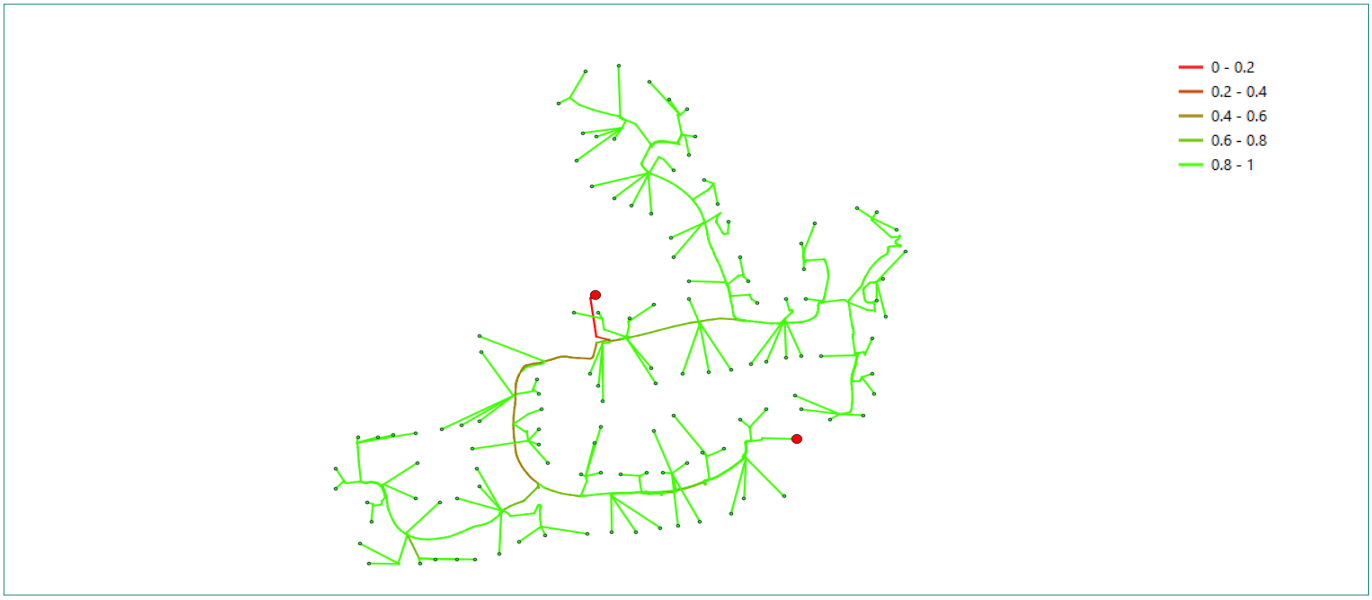


Figure 8: circuit_id = 263293/0/0020 maximum flow output with legend as the headroom = capacity - flow

circuit_id = 262417/0/0020

This exception was interesting as the configuration of the circuit is mesh, in comparison to the majority of circuits in the Barnstaple region. In addition, the violation / threshold boundary (20%) was met on a cable segment with conforming specification_description "95 3c WCON"; with capacity calculated from known maximum current rating found within WPD company directives.

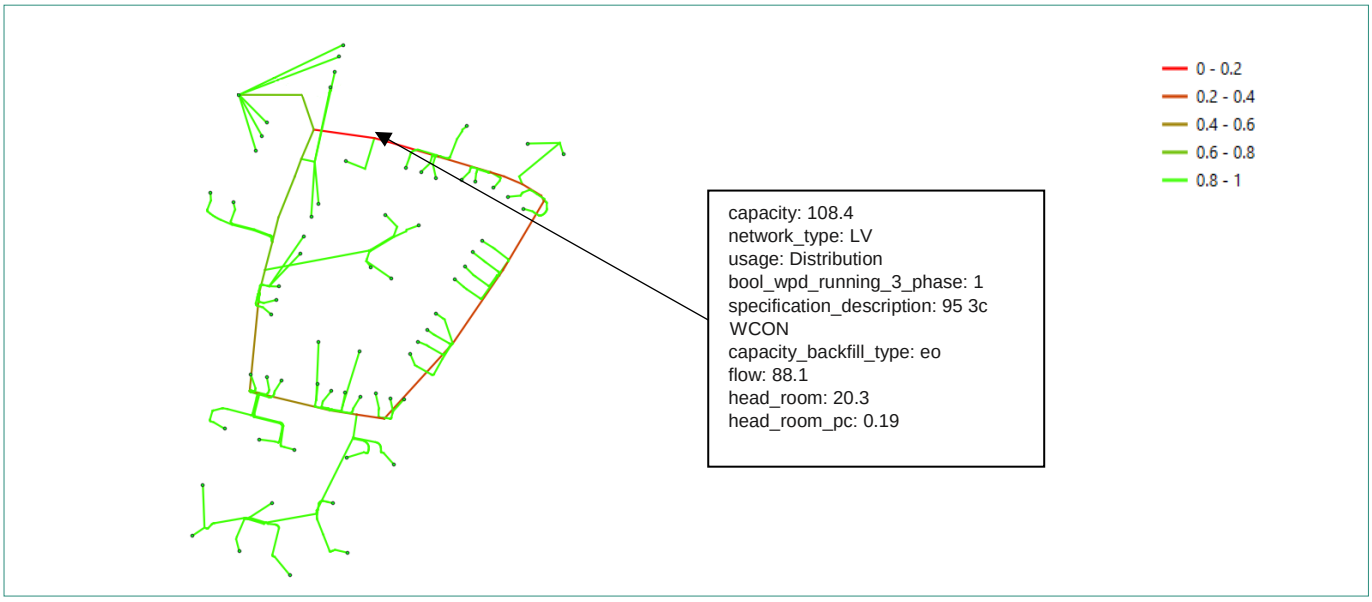


Figure 9: circuit_id = 262417/0/0020 maximum flow output with legend as the headroom = capacity - flow

In this exception example, the circuit appears to have some cables missing to service the cluster of customers assigned to the circuit. This could be the cause of the exception as the bottleneck in capacity appears in the wire leading to the cluster of customers; in reality there may be some missing service cables / wires such that the configuration of the circuit would distribute the flow of power across a number of cables / wires to remove the bottleneck. The root cause of this violation could also be due to the data quality with the specification_description = '4w Unknown' and the capacity_backfill_type being Barnstaple area minimum for LV service wires running 3 phase.

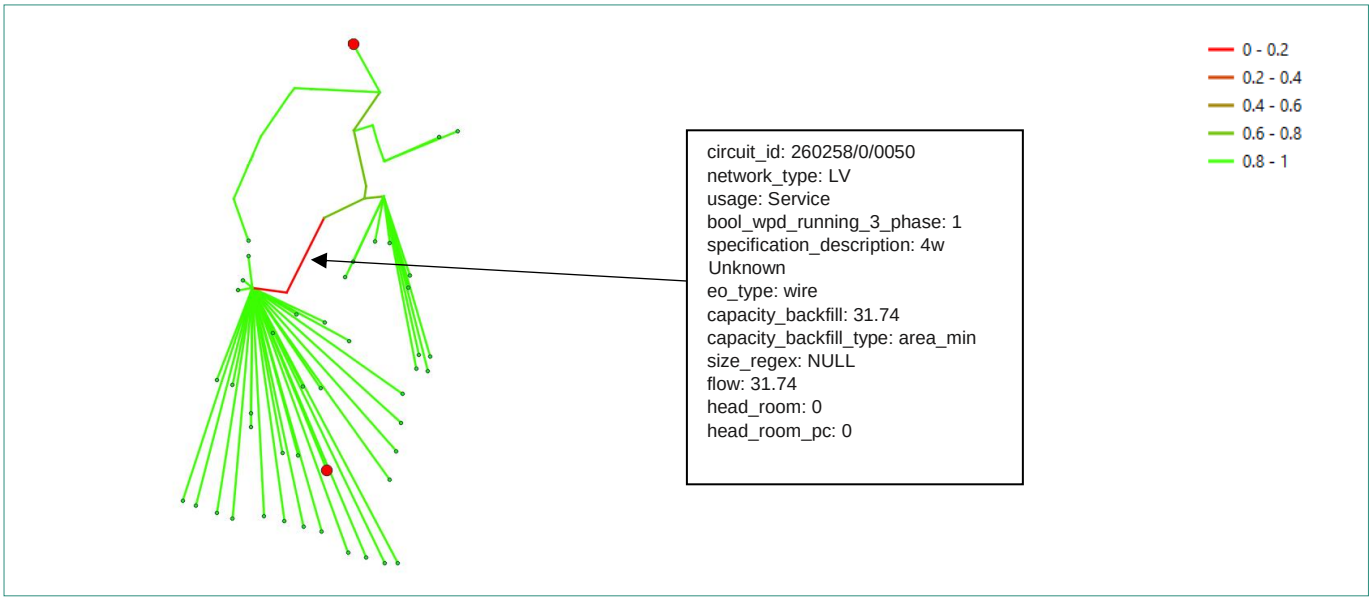


Figure 10: circuit_id = 260258/0/0050 maximum flow output with legend as the headroom = capacity - flow

circuit_id = 260584/0/0030

This exception example is similar to circuit_id = '260258/0/0050', where missing cables / wires to a cluster of customers may be the underlying root cause to bottlenecking at a cable / wire. Again, the data quality regarding the specification_description is poor and the backfill has been applied to be lower bound, i.e. minimum for the Barnstaple area for LV service cable running single phase.

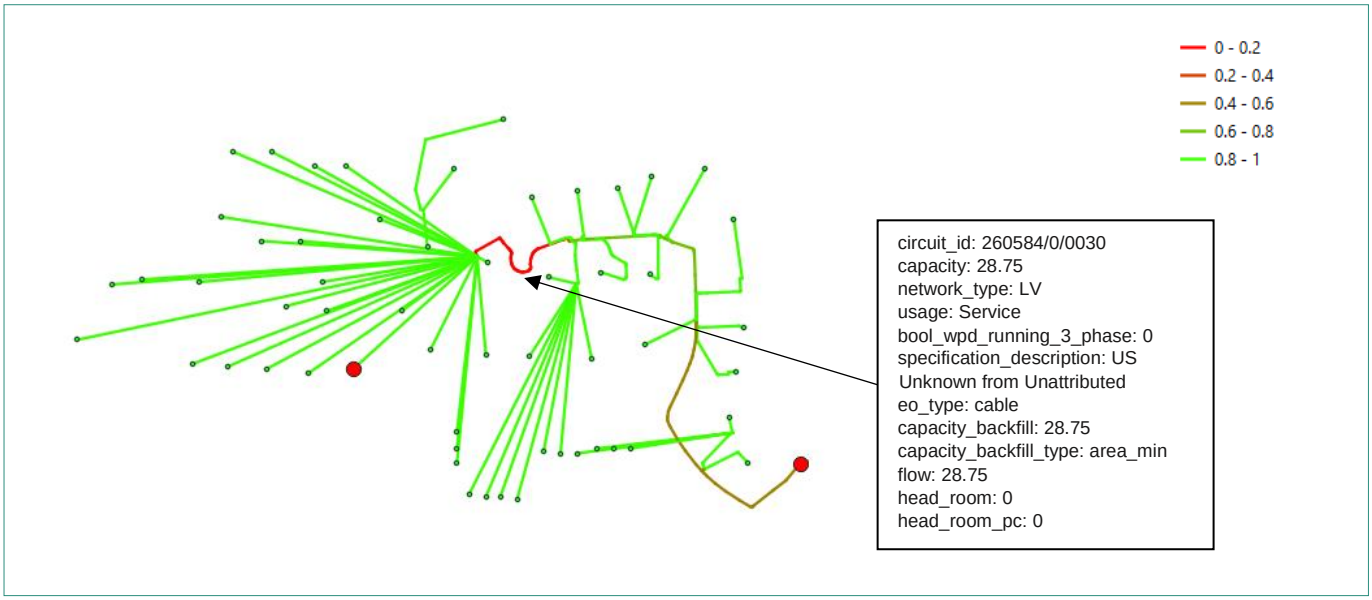


Figure 11: circuit_id = 260584/0/0030 maximum flow output with legend as the headroom = capacity - flow

4.2. Model 2: Spatial Graph Model

Model Processes

The spatial graph model (model 2) is an inductive graph neural network (GNN) based model. As such, there are three separate, but related, processes involved; each with its own purpose and outputs. These are as follows.

- **Training**: This process is used to obtain a new trained model that can be used to generate predictions. The model is trained against the data for a selected region taking the original data as ground truth and adding synthetic errors to form the input data. The GNN model is iteratively optimized to increase the number of output values from the model that match the true (original) values given the corrupted input data.
- **Evaluation**: This process is used to measure the ability of a pre-trained model to make correct predictions, given data with synthetic errors. As for the training process, it uses the original data for the selected region as ground truth and adds synthetic errors to form the input data. It can either be applied in a transductive (i.e. same region as training, but different synthetic errors) or inductive (i.e. different region from training) manner.
- **Prediction**: This process is used to identify errors in the original data and obtain suggested corrections. It uses the pre-trained model to generate predictions using the original data for a selected region as the input data. This process outputs the list of suggested corrections that meet the necessary scoring thresholds. It can either be applied in a transductive (i.e. same region as training) or inductive (i.e. different region from training) manner.

Model Output Reports

Reports

The spatial model produces some reports with detailed results at the level of the attributes of each asset and some summary reports that summarise the results across the entire area covered by the task. These are summarised in the table below.

The detailed results are saved as both CSV files (for use with analytical software[2]) and GeoPackage files (for use with GIS software). The CSV files have one row per attribute per asset, while the GeoPackage files have one layer per attribute, then one feature per asset. The summary results are saved as CSV only.

For the training and evaluation tasks, the evaluation report, evaluation summary and classification report are most relevant. For the prediction task, the exceptions report and exceptions summary are more relevant.

Report Name	Purpose	Description
	Detailed	
`model_outputs`	Understanding all the outputs from the model	All relevant inputs and outputs from model. All the other reports are a subset or summary of the data in this report.
`exceptions_report`	End user investigating suggested changes to the GIS data	All rows from model_outputs where output value is different from input value and score meets criteria. Excludes columns that are not interesting for the end user.
`evaluation_report`	Understanding the behaviour of the model in response to simulated errors	All rows with non-missing values in the original data. Only produced when synthetic errors are added to the input data.
	Summary	
`exceptions_summary`	End user understanding the distribution of identified errors across asset types and attributes	Number of rows with each error_code value (see below) for each asset type for each attribute.
`evaluation_summary`	Understanding the overall performance of the model on simulated errors for different asset types and attributes	Accuracy (proportion exactly correct) for each asset type and attribute for each error_code. Only produced when synthetic errors are added to the input data.

[2] The CSV reports can be opened in Excel but be aware that Excel may incorrectly interpret some of the values. For example, it replaces “1/1” with “01-Jan” and “10-20” with “Oct-20”.

Report Name	Purpose	Description
classification_report	Understanding the overall performance of the model on simulated errors for different attributes and attribute values	Classification report (see below) for each attribute. Only produced when synthetic errors are added to the input data.

Table 6: Model 2 Output Reports

Columns

These are the columns of the model outputs file. All the other reports are a subset or summary of these data.

- **Exceptions report**: all rows with “changed” = True and “score_ok” = True
- **Exceptions summary**: number of rows for each combination of “asset_type”, “attribute_name” and “error_code”
- **Evaluation report**: all rows with “true_ok” = True
- **Evaluation summary**: average of “correct” column for each combination of “asset_type”, “attribute_name” and “error_code” out of all rows with “true_ok” = True
- **Classification report**: classification report (see below) for each “attribute_name” using all rows with “true_ok” = True

The “Excluded from” column indicates when each column is not included in the outputs. For example, it may be constant or not interesting for some reports and it may not be produced for all tasks.

Column	Description	Excluded from
asset_id	Index of associated asset node in spatial graph	Exceptions report: internal
rwo_id	Real world object ID from EO	
asset_type	Asset type from EO	
circuit_id	Circuit ID from EO (if available)	
geometry	Geometry from EO, merged across tiles and parts (WKT)	
attribute_name	Name of attribute for this line	
input_value	Input value of attribute to model (category)	
output_value	Output value of attribute from model (category)	
changed	Whether output value is different from input (Boolean)	Exceptions report: always True
score_abs	Absolute score from model (maximum of scores per category)	
score_rel	Relative score from model (difference between scores for top 2 categories)	
score_ok	Whether score meets criteria (Boolean)	When thresholds is ”none”: treated as True Exceptions report: always True
error_code	Error code for output (see below)	
error_code_simple	Simplified error code for output (see below)	Exceptions report: internal
true_value	Original value of attribute in EO: treated as ground truth by the model (category)	Prediction task
true_ok	Whether original value is not missing (Boolean)	Evaluation report: always True Prediction task
input_modified	Whether input value is different from original value	Prediction task
error_code_true	Simplified error code if output were the original value (i.e. ideal error code value)	Prediction task
correct	Whether output value matches the original value (treated as ground truth)	Prediction task
error_type	What kind of error was added in the input data (see below)	Prediction task
fold	Which training fold the asset is part of (see below)	Prediction or Evaluation task

Table 7: Model 2 Output Columns

Enumerations

The enumerated columns listed above have the following possible values:

Code	Description	Meaning
error_code: Kind of error detected by the model.		
`no_error`	No error	Output value matches the input value.
`missing_value`	Missing value	Input value was missing.
`wrong_value`	Wrong value	Output value different from input value, which was also not missing.
`missing_value`	Missing value - low score	Missing value, but score for predicted value does not meet criteria. (i.e. low confidence score)
`wrong_value`	Wrong value - low score	Wrong value, but score for predicted value does not meet criteria. (i.e. low confidence score)
error_code_simple: This is like error_code but ignoring the score criteria.		
`no_error`	No error	Output value matches the input value.
`missing_value`	Missing value	Input value was missing.
`wrong_value`	Wrong value	Output value different from input value, which was also not missing.
error_type: Kind of error that was simulated in the data.		
`no_error`	No error	Input data matches original data exactly.
`missing_all`	All values missing	All output attributes were replaced with missing values in the input data.
`missing_one`	One value missing	One output attribute (at random) was replaced with a missing value in the input data. If the selected attribute is already missing, then it is not changed.
`corrupt_one`	One value corrupt	One output attribute (at random) was replaced with a random different value in the input data. If the selected attribute is already missing, then it is not changed.
fold: Which training fold the asset is part of.		
`train`	Training set	Asset labels are used for tuning the model.
`validation`	Validation set	Asset labels are used to supervise the model tuning.
`test`	Test set	Hold-out data for final model testing.

Table 8: Enumerations in Model 2 Outputs

Classification report

The classification report file is created by concatenating the multi-label classification reports for each classification output (i.e. one per attribute) from the model using all of the rows that are part of the evaluation. Remember that this assumes that the original values are the ground truth.

Each of the multi-label classification reports are constructed as follows:

- There is one row for each associated attribute value, which is referred to as the “row value” below
- True positives (TP) is the number of rows where the output and true values are both equal to the row value
- False negatives (FN) is the number of rows where the true value equals the row value but the output value does not
- False positives (FP) is the number of rows where the output value equals the row value but the true value does not.
- True negatives (TN) is the number of rows where neither the output or true values are equal to the row value
- Precision is the accuracy out of the rows with output values equal to the row value = TP / (TP + FP)
- Recall is the accuracy out of the rows with true values equal to the row value = TP / (TP + FN)
- F1-score is a balanced accuracy metric, which is the harmonic mean of the precision and recall = 2 TP / (2 TP + FP + FN)
- Support is the number of rows where the true value equals the row value = TP + FN

Results

Training Summary Results

These results were obtained from evaluating the trained model on the input data used for the training. The input data include synthetic errors and were clipped to the range xmin=248000, ymin=126000, xmax=264000, ymax=141000. The “medium” thresholds were used for evaluation.

The input data were split randomly into training, validation and test partitions: the training partition was used to optimize the model parameters, the validation partition was used to monitor the training progress and the test partition was used to confirm the overall model performance at the end. However, the results shown in this section combine the outputs across all 3 folds, so some examples were seen by the model during training, while some were not.

The evaluation summary (which is described in the previous subsection) is shown below. This can be used to help a user to understand what the performance of the model was during training for each asset type and attribute and each error_code output. The results should be considered qualitatively, rather than quantitatively, since they are strongly dependent on the simulation of the errors in the input data for training, and they assume that the original data are always the ground truth.

asset_type	attribute_name	no_error	missing_value	wrong_value	missing_value_low	wrong_value_low
cable	**network_type**	99.89%	99.54%	98.42%	93.24%	84.25%
cable	**nominal_voltage_pp**	98.41%	95.00%	84.59%	79.72%	57.63%
cable	**spec_material**	98.85%	79.68%	62.59%	49.60%	37.45%
cable	**spec_size**	97.74%	62.43%	37.03%	40.75%	20.41%
connector_point	**network_type**	99.95%	100.00%	99.82%	97.49%	95.28%
connector_point	**nominal_voltage_pp**	98.13%	95.72%	92.17%	80.32%	65.79%
connector_segment	**network_type**	99.86%	97.25%	98.57%	83.71%	62.50%
connector_segment	**nominal_voltage_pp**	99.35%	96.76%	96.28%	79.23%	80.56%
energy_consumer	**network_type**	100.00%	100.00%	100.00%	92.00%	100.00%
energy_consumer	**nominal_voltage_pp**	98.39%	90.00%	100.00%	96.55%	66.67%
energy_source	**network_type**	100.00%	100.00%			
energy_source	**nominal_voltage_pp**	100.00%				
isolating_eqpt	**network_type**	99.87%	97.28%	90.48%	88.61%	39.71%
isolating_eqpt	**nominal_voltage_pp**	99.29%	96.26%	88.89%	78.77%	50.75%
keypole	**network_type**	100.00%	100.00%	100.00%	87.10%	80.00%
keypole	**nominal_voltage_pp**	99.97%	99.33%	97.97%	70.73%	50.00%
pole	**network_type**	99.83%	99.49%	100.00%	85.88%	86.05%
pole	**nominal_voltage_pp**	99.65%	98.16%	94.17%	69.09%	64.29%
protective_eqpt	**network_type**	99.01%	69.46%	65.28%	66.40%	28.80%

asset_type	attribute_name	no_error	missing_value	wrong_value	missing_value_low	wrong_value_low
protective_eqpt	**nominal_voltage_pp**	98.26%	42.17%	28.77%	33.22%	22.03%
service_point	**network_type**	99.84%	100.00%	100.00%	96.34%	97.56%
service_point	**nominal_voltage_pp**	98.28%	98.81%	86.36%	87.90%	90.16%
tower	**network_type**	100.00%	100.00%	100.00%	100.00%	
tower	**nominal_voltage_pp**	96.72%	100.00%	100.00%		100.00%
wire	**network_type**	99.87%	99.64%	99.42%	89.37%	58.67%
wire	**nominal_voltage_pp**	99.23%	93.52%	64.87%	62.00%	15.81%
wire	**spec_material**	99.64%	93.38%	78.48%	55.96%	23.94%
wire	**spec_size**	98.10%	80.53%	51.59%	47.27%	17.95%

Table 9: Model 2 Evaluation Summary from Training Process

Observations from the evaluation summary:

- Accuracy for rows where the error_code is “no_error” is generally very high (>98%).
- Accuracy amongst the high scoring error codes (i.e. “missing_value” and “wrong_value”) is almost always greater than or equal to the accuracy amongst the low scoring error codes (i.e. “missing_value_low” and “wrong_value_low”)—sometimes considerably—indicating that the threshold on the confidence score is doing its job.
- Accuracy for cases with missing values is much higher than the majority class percentages, especially for the high scoring error code. These are network type = 72% (LV), nominal_voltage_pp = 47% (230V), spec_material = 23% (HDC), spec_size = 24% (20-30).
- Accuracy for cases identified as “wrong_value” are also generally higher than the majority class percentages, especially for the high scoring error code, but with some exceptions.
- Accuracy for the protective_eqpt asset type is lower than expected. Since the model does not explicitly discriminate between asset types, this most likely means that there are unusual patterns involving protective_eqpt assets specifically (i.e. ones that don’t apply to other asset types) and this will be investigated later using examples from the model predictions.

The classification report (which is described in the previous subsection) is shown below. This can be used to help a user to understand the ability of the performance of the model with respect to false positives and false negatives for each output value. As before, the results should be considered qualitatively, rather than quantitatively, since they are strongly dependent on the simulation of the errors in the input data for training, and they assume that the original data are always the ground truth.

attribute_name	attribute_value	TP	FP	FN	TN	precision	recall	f1_score	support
network_type	**LV**	65973	206	800	25213	99.69%	98.80%	99.24%	66773
network_type	**MV**	20999	774	252	70167	96.45%	98.81%	97.62%	21251
network_type	**HV**	4159	81	9	87943	98.09%	99.78%	98.93%	4168
nominal_voltage_pp	**110**	29	374	5	91784	7.20%	85.29%	13.27%	34

attribute_name	attribute_value	TP	FP	FN	TN	precision	recall	f1_score	support
nominal_voltage_pp	**230**	40412	1256	2639	47885	96.99%	93.87%	95.40%	43051
nominal_voltage_pp	**400**	22062	2342	1619	66169	90.40%	93.16%	91.76%	23681
nominal_voltage_pp	**11000**	20918	569	340	70365	97.35%	98.40%	97.87%	21258
nominal_voltage_pp	**33000**	3521	57	70	88544	98.41%	98.05%	98.23%	3591
nominal_voltage_pp	**132000**	576	76	1	91539	88.34%	99.83%	93.73%	577
spec_material	**aaac**	116	12	0	16965	90.63%	100.00%	95.08%	116
spec_material	**abc**	1301	154	82	15556	89.42%	94.07%	91.68%	1383
spec_material	**acsr**	232	38	6	16817	85.93%	97.48%	91.34%	238
spec_material	**al**	2268	191	389	14245	92.23%	85.36%	88.66%	2657
spec_material	**c/c**	1442	114	234	15303	92.67%	86.04%	89.23%	1676
spec_material	**cad cu**	104	50	3	16936	67.53%	97.20%	79.69%	107
spec_material	**cu**	561	63	96	16373	89.90%	85.39%	87.59%	657
spec_material	**hdc**	3884	53	101	13055	98.65%	97.47%	98.06%	3985
spec_material	**hyb**	2241	176	126	14550	92.72%	94.68%	93.69%	2367
spec_material	**s/c**	587	157	91	16258	78.90%	86.58%	82.56%	678
spec_material	**sac**	120	328	18	16627	26.79%	86.96%	40.96%	138
spec_material	**solidal**	60	22	3	17008	73.17%	95.24%	82.76%	63
spec_material	**wcon**	2651	168	377	13897	94.04%	87.55%	90.68%	3028
spec_size	**0-10**	3	15	1	21349	16.67%	75.00%	27.27%	4
spec_size	**10-20**	1791	584	192	18801	75.41%	90.32%	82.19%	1983
spec_size	**20-30**	4610	262	541	15955	94.62%	89.50%	91.99%	5151
spec_size	**30-60**	2447	684	364	17873	78.15%	87.05%	82.36%	2811
spec_size	**60-90**	1484	97	406	19381	93.86%	78.52%	85.51%	1890
spec_size	**90-140**	3247	267	927	16927	92.40%	77.79%	84.47%	4174
spec_size	**140-280**	3844	689	446	16389	84.80%	89.60%	87.14%	4290
spec_size	**280-550**	860	339	109	20060	71.73%	88.75%	79.34%	969
spec_size	**550-900**	88	57	8	21215	60.69%	91.67%	73.03%	96

Table 10: Model 2 Classification Report from Training Process

Observations from the classification report:

- For most outputs, the accuracy is good: f1-scores >97% for network_type, >91% for nominal_voltage_pp, >82% for spec_material and >79% for spec_size.
- The f1-scores only drop below these number for rare classes (those with less than about 200 examples in the training dataset) and some of these have very low scores. This is to be expected since it's harder for the model to learn the patterns around these attribute values.
- The "spec_size/0-10" class is so small that it should probably be merged with the "10-20" range. Note that the breaks for this conversion from numbers to categories are customizable at training time via the model parameters file.
- Especially for these rare classes, there tend to be lots of false positives and few false negatives, which means that the precision tends to be low while the recall is much better. This reflects on the synthetic error generation process, which takes any other value at uniformly random independent of the original value when corrupting an attribute. This artificially increases the number of assets with rare values observed in the corrupted input data, which increases the likelihood that the model will suggest these, leading to more false positives for these categories.

Overall observations:

- These results suggest that the model is performing correctly and that the quality of the suggested corrections is generally good, especially considering that the model has access to very limited information about the assets in the network and there is significant scope to optimize the GNN model.
- Further investigation is required into some of the false predictions to understand whether there are straightforward improvements that could be made to the model to allow the model to learn the underlying patterns better.

Evaluation Summary Results

These results were obtained from evaluating the trained model on the input data including synthetic errors. The input data were clipped to the range xmin=248000, ymin=126000, xmax=264000, ymax=141000. The "medium" thresholds were used for evaluation.

While the input data come from the same area as for the training, the synthetic errors added are different since a different random number seed was used. This provides a useful check on the performance of the model obtained from the training process and demonstrates that the performance of the model can be measured using synthetic errors separately from the model training process.

The evaluation summary and classification report (which are described in the previous subsection) are shown below. These results are very similar to the results from the training, which confirms those previous results and observations. In particular, it demonstrates that those results were not dependent on the specific errors added in the training data.

asset_type	attribute_name	no_error	missing_value	wrong_value	missing_value_low	wrong_value_low
cable	**network_type**	99.89%	99.38%	98.31%	94.21%	86.50%
cable	**nominal_voltage_pp**	98.26%	94.53%	83.37%	79.80%	58.12%
cable	**spec_material**	98.70%	81.25%	58.90%	49.45%	30.23%
cable	**spec_size**	97.41%	63.05%	35.45%	39.43%	15.90%
connector_point	**network_type**	99.96%	99.87%	99.82%	97.80%	95.58%
connector_point	**nominal_voltage_pp**	98.11%	95.40%	91.75%	77.99%	60.35%
connector_segment	**network_type**	99.88%	98.41%	98.58%	82.21%	50.94%
connector_segment	**nominal_voltage_pp**	99.34%	97.56%	97.40%	78.73%	71.43%

asset_type	attribute_name	no_error	missing_value	wrong_value	missing_value_low	wrong_value_low
energy_consumer	**network_type**	99.43%	100.00%	100.00%	88.89%	100.00%
energy_consumer	**nominal_voltage_pp**	97.22%	100.00%	100.00%	75.76%	100.00%
energy_source	**network_type**		0.00%		100.00%	
energy_source	**nominal_voltage_pp**		0.00%		100.00%	
isolating_eqpt	**network_type**	99.87%	97.53%	89.63%	83.19%	43.06%
isolating_eqpt	**nominal_voltage_pp**	99.30%	96.09%	92.31%	70.81%	46.58%
keypole	**network_type**	99.94%	100.00%	100.00%	97.73%	83.33%
keypole	**nominal_voltage_pp**	99.94%	99.88%	98.22%	70.37%	38.10%
pole	**network_type**	99.88%	99.55%	98.28%	94.67%	77.78%
pole	**nominal_voltage_pp**	99.57%	98.73%	93.30%	76.74%	65.00%
protective_eqpt	**network_type**	98.73%	67.10%	78.79%	72.30%	23.08%
protective_eqpt	**nominal_voltage_pp**	98.52%	39.63%	25.33%	32.77%	27.43%
service_point	**network_type**	99.95%	100.00%	100.00%	95.00%	95.83%
service_point	**nominal_voltage_pp**	97.67%	94.85%	74.07%	89.26%	86.57%
tower	**network_type**	100.00%	100.00%	100.00%	0.00%	100.00%
tower	**nominal_voltage_pp**	100.00%	94.12%	100.00%		
wire	**network_type**	99.94%	99.34%	99.81%	85.84%	66.67%
wire	**nominal_voltage_pp**	99.35%	93.01%	64.95%	60.45%	19.77%
wire	**spec_material**	99.30%	92.08%	71.11%	58.22%	28.66%
wire	**spec_size**	98.27%	79.49%	50.56%	48.02%	18.49%

Table 11: Model 2 Evaluation Summary from Evaluation Process

attribute_name	attribute_value	TP	FP	FN	TN	precision	recall	f1_score	support
network_type	**LV**	65987	165	786	25254	99.75%	98.82%	99.28%	66773
network_type	**MV**	21059	766	192	70175	96.49%	99.10%	97.78%	21251
network_type	**HV**	4155	60	13	87964	98.58%	99.69%	99.13%	4168
nominal_voltage_pp	**110**	26	434	8	91724	5.65%	76.47%	10.53%	34
nominal_voltage_pp	**230**	40307	1284	2744	47857	96.91%	93.63%	95.24%	43051
nominal_voltage_pp	**400**	22023	2362	1658	66149	90.31%	93.00%	91.64%	23681
nominal_voltage_pp	**11000**	20970	578	288	70356	97.32%	98.65%	97.98%	21258
nominal_voltage_pp	**33000**	3531	39	60	88562	98.91%	98.33%	98.62%	3591

attribute_name	attribute_value	TP	FP	FN	TN	precision	recall	f1_score	support
nominal_voltage_pp	**132000**	573	65	4	91550	89.81%	99.31%	94.32%	577
spec_material	**aaac**	115	32	1	16945	78.23%	99.14%	87.45%	116
spec_material	**abc**	1277	157	106	15553	89.05%	92.34%	90.66%	1383
spec_material	**acsr**	217	42	21	16813	83.78%	91.18%	87.32%	238
spec_material	**al**	2256	173	401	14263	92.88%	84.91%	88.71%	2657
spec_material	**c/c**	1442	132	234	15285	91.61%	86.04%	88.74%	1676
spec_material	**cad cu**	99	54	8	16932	64.71%	92.52%	76.15%	107
spec_material	**cu**	560	70	97	16366	88.89%	85.24%	87.02%	657
spec_material	**hdc**	3868	76	117	13032	98.07%	97.06%	97.57%	3985
spec_material	**hyb**	2252	197	115	14529	91.96%	95.14%	93.52%	2367
spec_material	**s/c**	577	123	101	16292	82.43%	85.10%	83.74%	678
spec_material	**sac**	116	341	22	16614	25.38%	84.06%	38.99%	138
spec_material	**solidal**	57	28	6	17002	67.06%	90.48%	77.03%	63
spec_material	**wcon**	2663	169	365	13896	94.03%	87.95%	90.89%	3028
spec_size	**0-10**	4	14	0	21350	22.22%	100.00%	36.36%	4
spec_size	**10-20**	1779	632	204	18753	73.79%	89.71%	80.97%	1983
spec_size	**20-30**	4612	255	539	15962	94.76%	89.54%	92.07%	5151
spec_size	**30-60**	2444	684	367	17873	78.13%	86.94%	82.30%	2811
spec_size	**60-90**	1498	86	392	19392	94.57%	79.26%	86.24%	1890
spec_size	**90-140**	3213	258	961	16936	92.57%	76.98%	84.05%	4174
spec_size	**140-280**	3809	739	481	16339	83.75%	88.79%	86.20%	4290
spec_size	**280-550**	845	335	124	20064	71.61%	87.20%	78.64%	969
spec_size	**550-900**	86	75	10	21197	53.42%	89.58%	66.93%	96

Table 12: Model 2 Classification Report from Evaluation Process

Prediction Summary Results

These results were obtained from using the trained model to predict using the input data without synthetic errors. The input data were clipped to the range xmin=248000, ymin=126000, xmax=264000, ymax=141000. The “medium” thresholds were used.

Since this process runs with the original EO data unmodified, the identified errors and suggested values relate to actual errors in the EO database (or at least as it was when the extract was made). The summary is discussed here and sample errors are discussed in the next subsection.

The exceptions summary (which is described in the previous subsection) is shown below.

asset_type	attribute_name	no_error	missing_value	wrong_value	missing_value_low	wrong_value_low
cable	**network_type**	39001	0	31	0	45
cable	**nominal_voltage_pp**	38596	0	171	0	310
cable	**spec_material**	9228	19467	41	10248	93
cable	**spec_size**	12181	12600	197	13850	249
connector_point	**network_type**	15442	0	1	0	0
connector_point	**nominal_voltage_pp**	15392	0	9	0	42
connector_segment	**network_type**	7688	0	2	0	40
connector_segment	**nominal_voltage_pp**	7712	0	7	0	11
energy_consumer	**network_type**	231	0	0	0	0
energy_consumer	**nominal_voltage_pp**	230	0	0	0	1
energy_source	**network_type**	2	0	0	0	0
energy_source	**nominal_voltage_pp**	2	0	0	0	0
isolating_eqpt	**network_type**	3984	0	64	0	24
isolating_eqpt	**nominal_voltage_pp**	3992	0	27	0	53
keypole	**network_type**	4724	0	0	0	0
keypole	**nominal_voltage_pp**	4715	0	0	0	9
pole	**network_type**	4534	3161	1	753	1
pole	**nominal_voltage_pp**	4523	2235	4	1679	9
protective_eqpt	**network_type**	2435	0	16	0	155
protective_eqpt	**nominal_voltage_pp**	2493	0	21	0	92
service_point	**network_type**	2512	0	0	0	0
service_point	**nominal_voltage_pp**	2511	0	0	0	1
tower	**network_type**	82	0	0	0	0
tower	**nominal_voltage_pp**	82	0	0	0	0
wire	**network_type**	11150	0	3	0	24
wire	**nominal_voltage_pp**	10561	0	298	0	318
wire	**spec_material**	7586	1211	66	2235	79
wire	**spec_size**	8389	1214	140	1222	212

Table 13: Model 2 Exceptions Summary from Prediction Process

Observations from the exceptions summary:

- Vast majority of attributes were identified as “no error”.
- Only pole assets are missing any voltage attributes.
- About two thirds of cable assets and one third of wire assets have both specification parts used for this model missing.
- The number of “wrong_value” or “wrong_value_low” errors are small compared to the number of attributes used. This is expected since the model is trained to predict those original values from the dataset. However, some errors are still identified.
- The accuracy of the identified errors can only be assessed by spot-checking each one individually.

Investigation of Sample Errors

These results are sample exceptions identified from the prediction task above.

Missing values: network type and operational voltage

The only assets in EO with missing network type and operational voltage are poles. In most cases, the suggested values appear to be correct. For example:

- RWO 70937259 (no pole number[3]) has no network type or operational voltage attributes and the model suggests 11kV and MV for this, which matches the neighbouring assets.

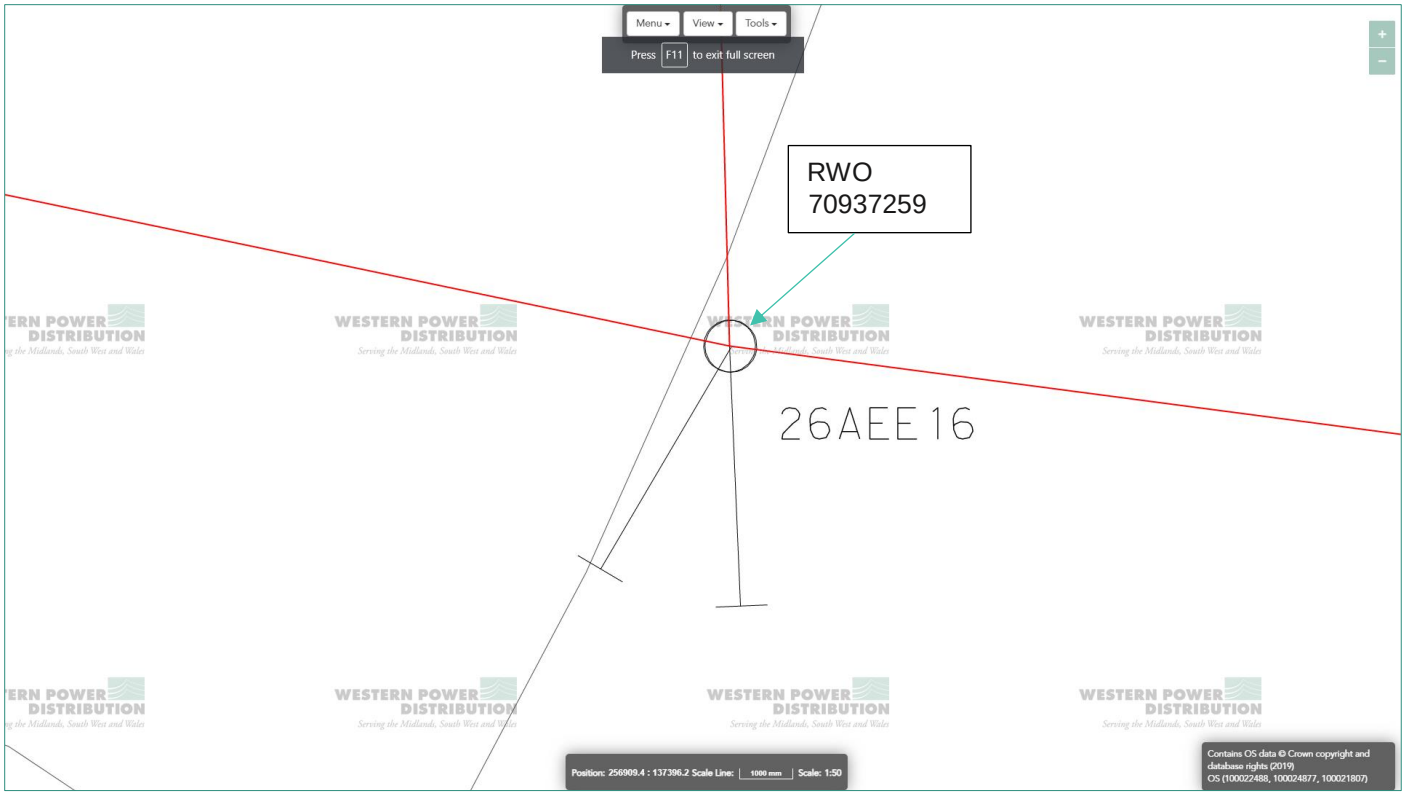


Figure 12: Map of RWO 70937259

[3] Poles are shown as circles in WPD DataPortal2. Note that pole 26AEE16 (RWO 70937257) has the same coordinates as RWO 70937259, which means that the symbols for the two poles overlap completely.

- RWO 244987101 (no pole number) has no network type or operational voltage attributes and the model suggests 230V and LV for this, which matches the neighbouring assets. While the other three services are shown as supplying nearby buildings, this service terminates at the edge of the footpath and the “PL” within the cable type suggests public lighting which may explain the difference in completeness of data.

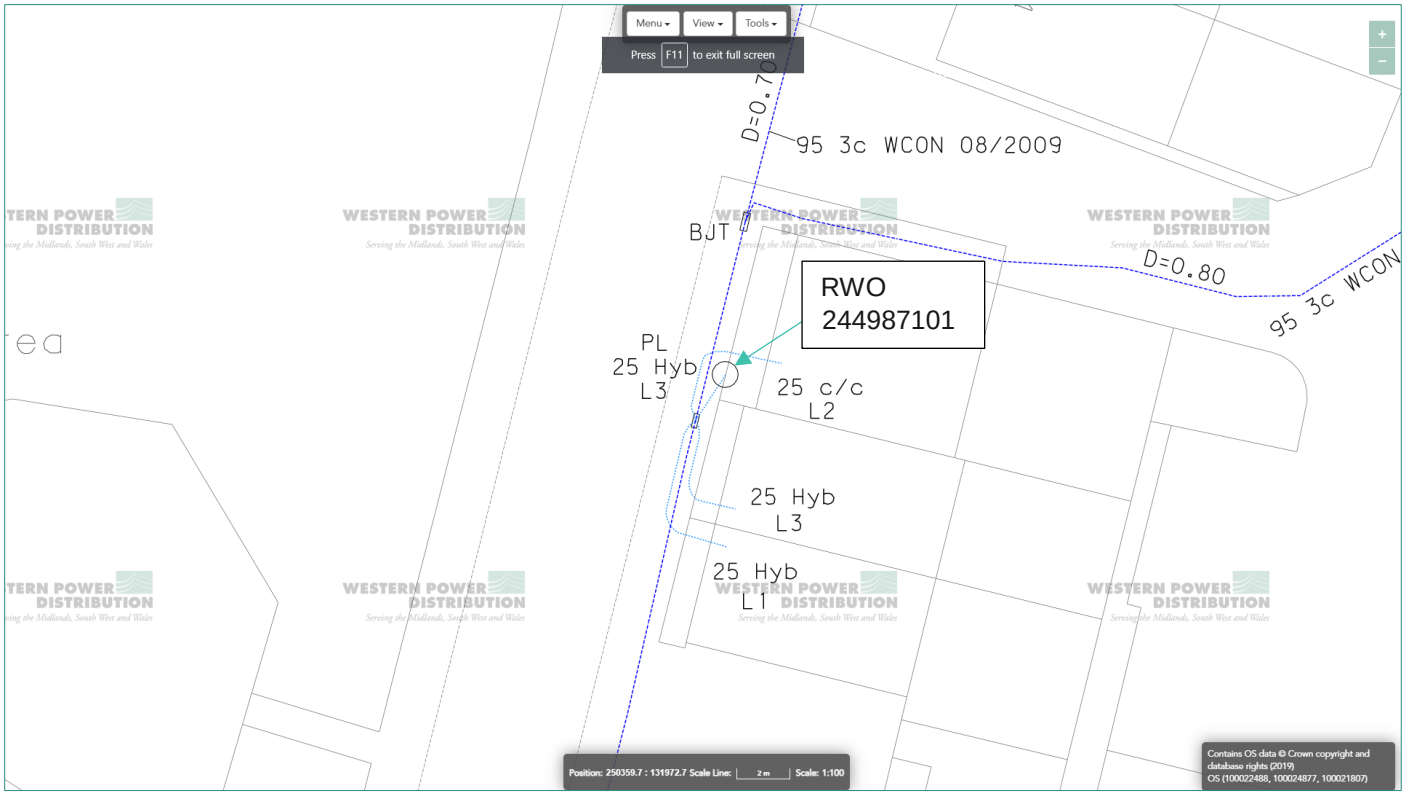


Figure 13: Map of RWO 244987101

However, occasionally the suggested values appear to be incorrect. For example:

- RWO 312462016 (pole 26-4094-3) has no network type or operational voltage attributes and the model suggests the operational voltage should be 11kV, which is not correct. In this case, the pole is at the end of a very short, very straight LV circuit in countryside, which means that a significant number of the neighbouring locations in the spatial mesh are associated with 11kV assets. Note that the suggested value for the network type (MV) did not meet the relevant scoring threshold.

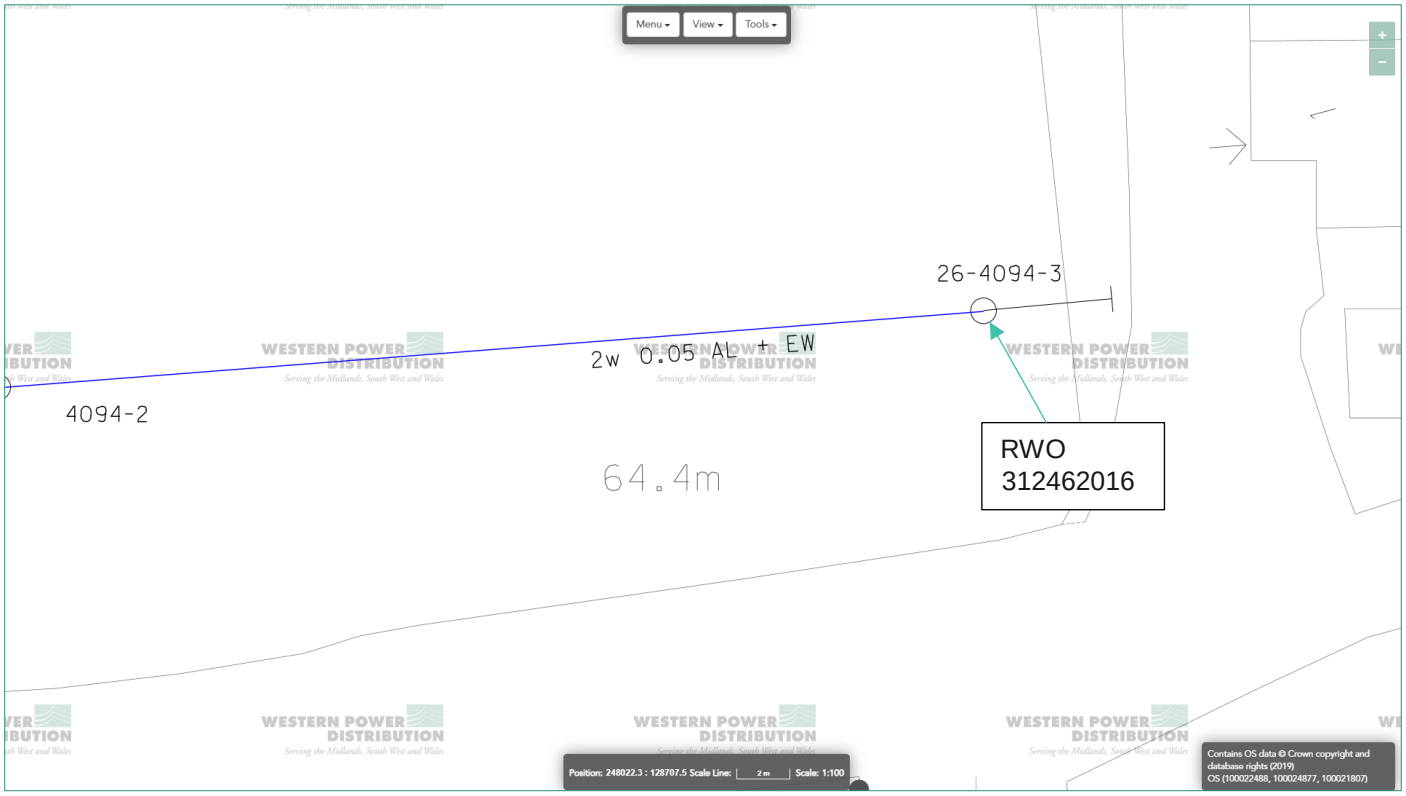


Figure 14: Map of RWO 312462016

Missing values: specification material and size

In EO, the cable specification for many service cables are marked as "unknown"[4]. This is common in the area shown below (e.g. RWO 444935212). Note that the service specification and phase annotations are not part of the EO extract. Where the outputs meet the scoring threshold, the model suggests these have material "Hyb" and sizes in the "30-60" category. The WPD DataPortal2 map below shows that this is often the correct values for the missing data, but sometimes the correct value have size "20-30". The mix of 35 and 25 size cables reflects that the properties are served by looped services. The first property has a 35 service with the subsequent one or two properties having 25 services. In this case, the phase for each service provides confirmation of the connectivity of the services which would suggest that this area would be better suited to validation with model 1. Note that the information on this map associated with the properties is not part of the EO export used for this project. There is work planned to determine how well cable specification information that is given as a text item can be associated with the nearest cable.

[4] More specifically, the text is "SV Unknown from Unattributed".

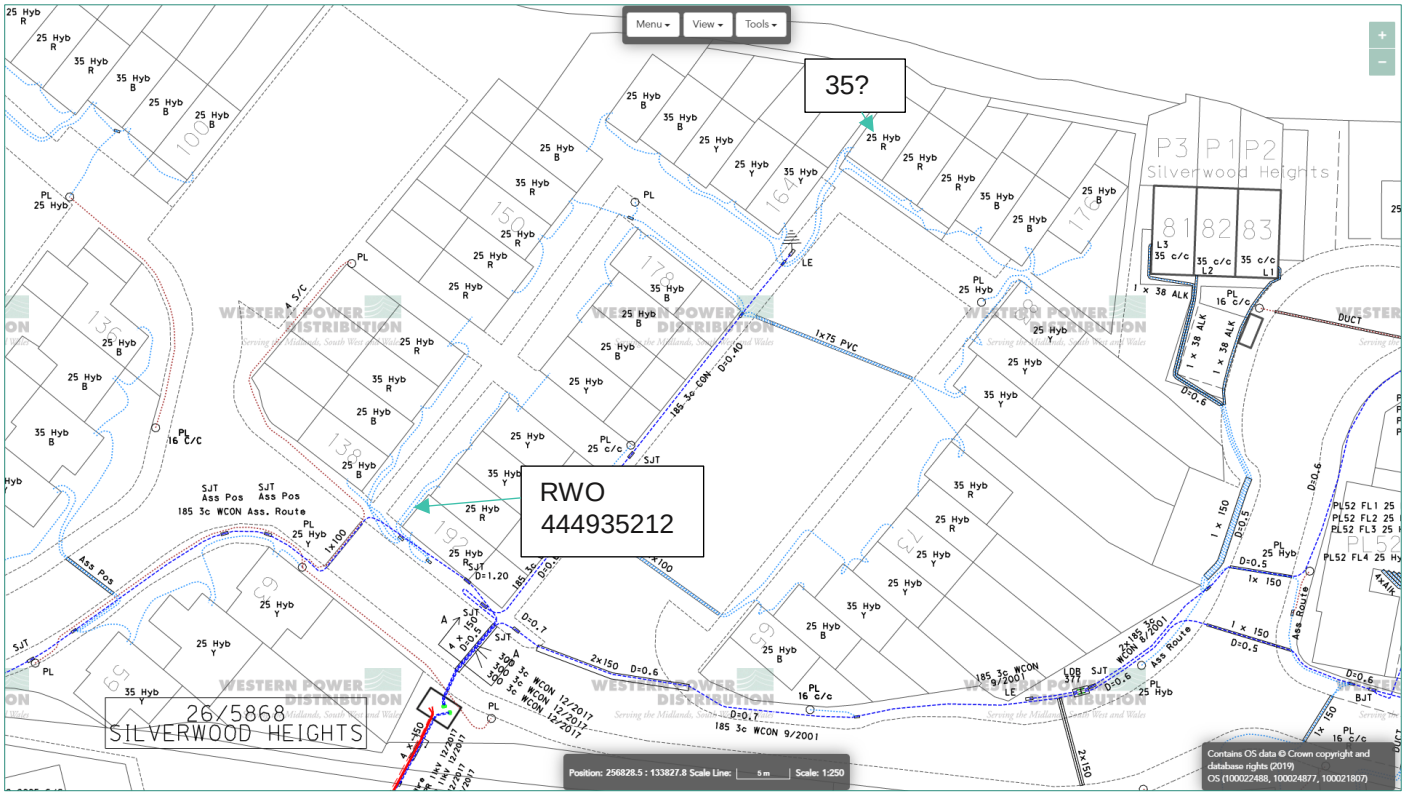


Figure 15: Map of RWO 444935212 and surroundings

For the LV (dark blue) cable close to the centre of the map below (RWO 445053204), the specification is in EO as “4C UNKNOWN SIZE”. The model suggests that this should have material “WCON” and a size in the “140-280” category. This matches the specification of the following cable and other similar cables in the neighbourhood. The cable is one of a set of three from the substation “Silverwood Heights” which start as 300 3c wavecon and pass through a duct under the roadway. The other two feeders also have straight joints soon after the ducted crossing where the cable transitions to 185 wavecon, increasing the likelihood that the transition for this cable from 300 to 185 occurs at the first straight joint (by house number 192) rather than downstream of this unknown cable section.

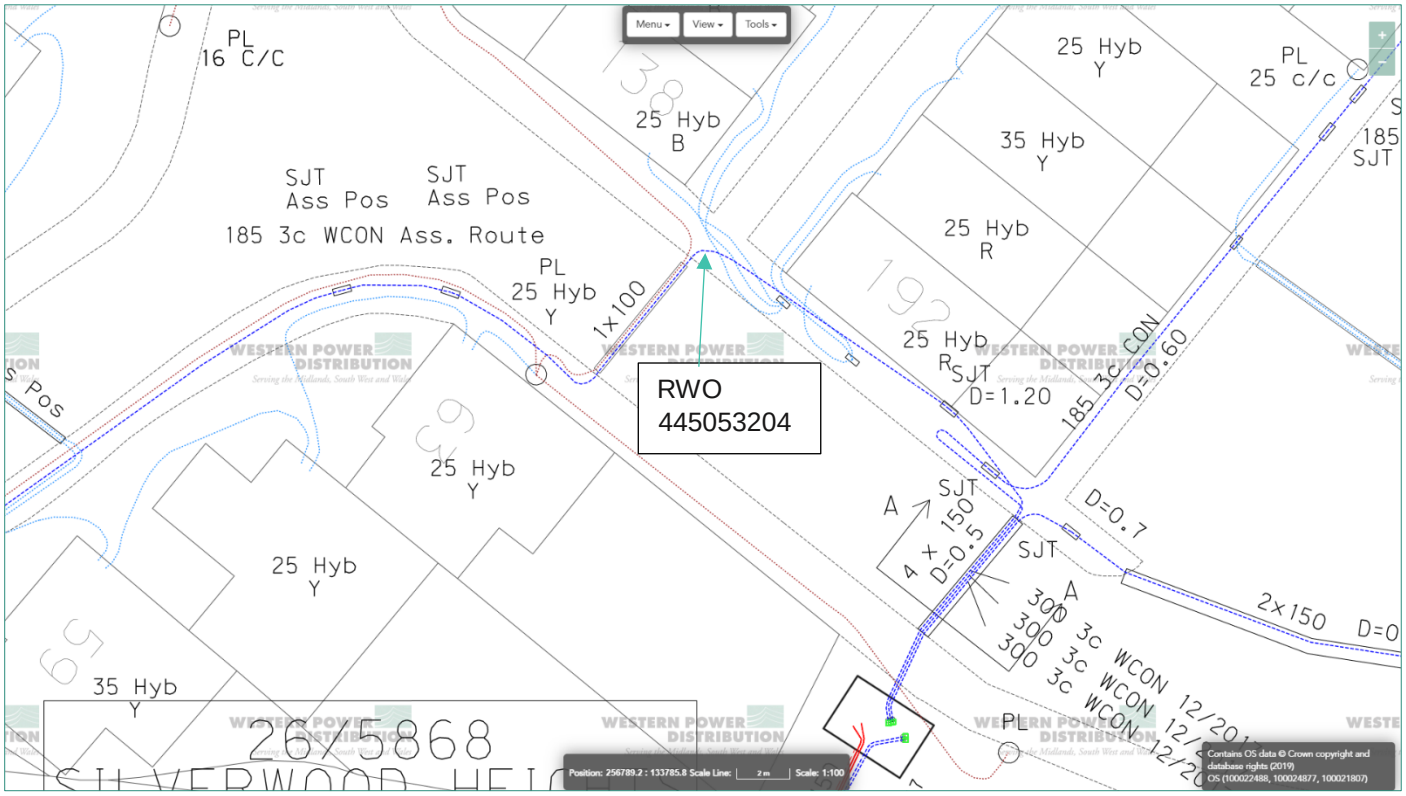


Figure 16: Map of RWO 445053204

Missing values: Earth wires

The specification parsing in PoC model is very simple: searches the specification description attribute for one of a list of known materials (e.g. “wcon”, “solidal”, “hyb”) and for word that is a decimal number or integer with 2 or more digits. Since the specification of Earth wires are stored as “11kV Earthwire” or similar, it will treat this as unknown material and unknown size and will try to suggest specification parts for Earth wires that do not match this convention. In particular, the model has no inputs or outputs that indicate that the relevant asset is an Earth wire.

There are a number of possible enhancements that could be made to improve this, such as suppressing specification changes when “Earth” is part of the input value or separating separate Earth wires into a distinct asset type in the graph with no specification description attribute.

Wrong values: 11kV -> 230V

Some connector points (e.g. RWO 428883346) have attributes indicating that they have an operational voltage of 11kV but are part of the LV network, and the model suggests replacing the operational voltage with 230V. The model has correctly identified that these assets should have the operational voltage fixed rather than the network type, and, for the examples checked, it has correctly identified that the connected assets are also 230V.

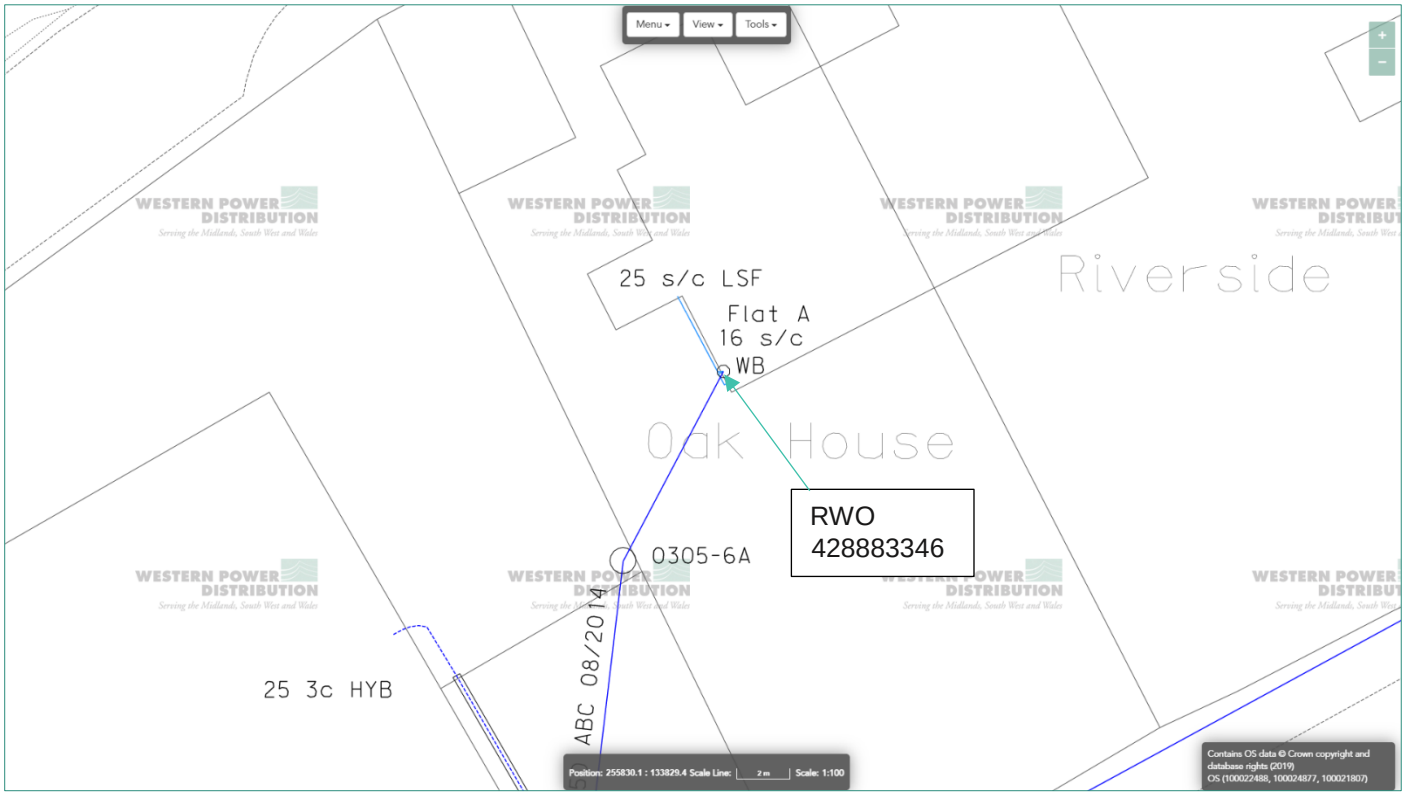


Figure 17: Map of RWO 428883346

Wrong values: 11kV -> 33kV, MV -> HV

Pole 26ZAJ1 (RWO 42537478) has attributes indicating that it has an operational voltage of 11kV (MV), and the model suggests replacing these attributes with 33kV (HV). Since it supports 33kV (HV) wires, this is considered to be a correctly identified error.

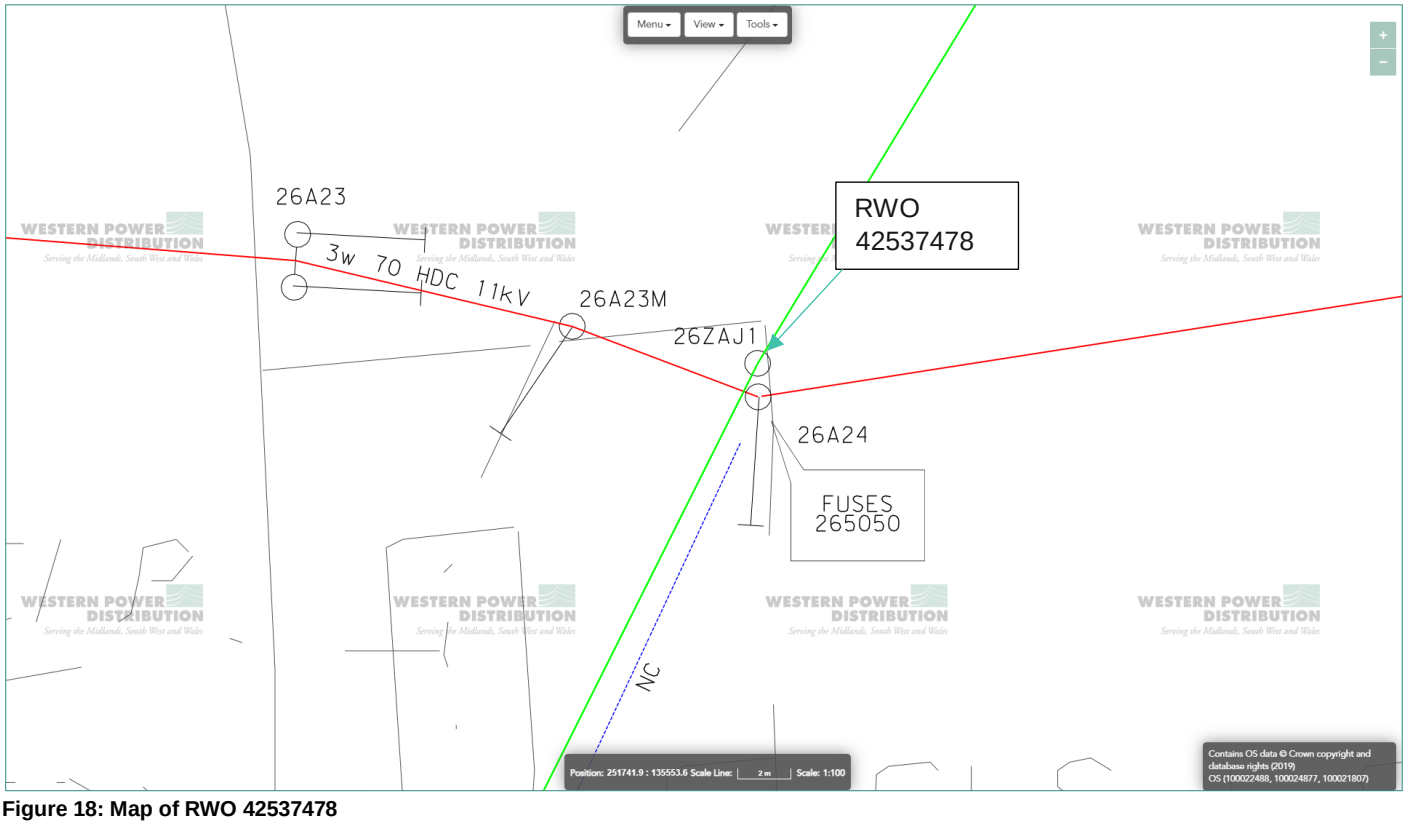


Figure 18: Map of RWO 42537478

Wrong values: 230V -> 11kV, LV -> MV

There are several situations where the model suggests that an asset marked as 230V (or 400V) and LV should be 11kV and MV instead, such as the service cable shown below (RWO 172507162), but where this appears to be incorrectly identified as an error. Most of these are associated with short, straight, rural LV circuits with very few LV assets, surrounded with MV assets. Such situations are hard for the model, because it does not have enough information to correctly classify this pattern, and because the assets at the pole-mounted substation all share a small number of distinct coordinates.
Note that earth wires shown here do not appear in the EO extract for the PoC.

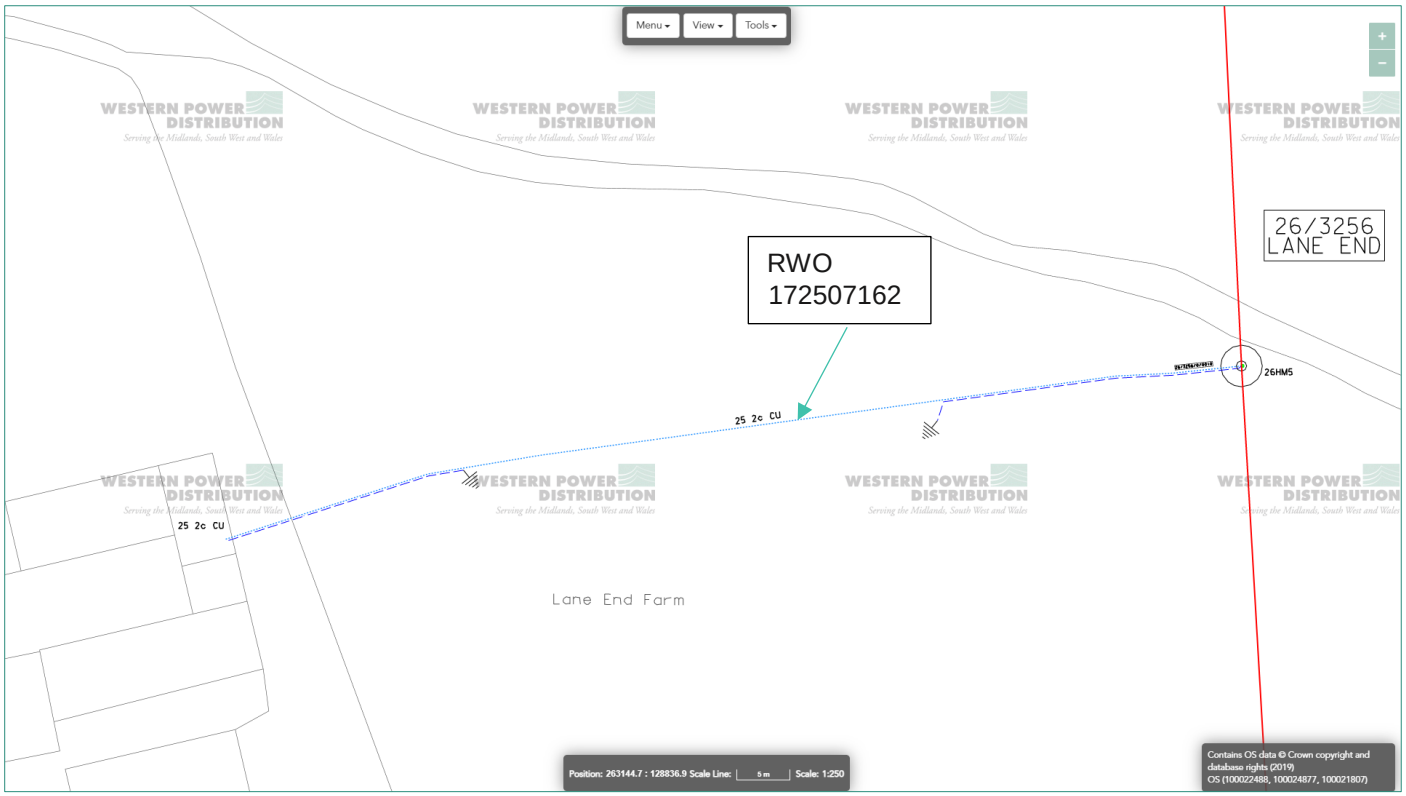


Figure 19: Map of RWO 172507162

There are also examples of the same issue affecting the predictions for point assets at the pole-mounted substation location, such as isolating equipment.

There are several possible enhancements to the spatial graph network that might improve this. Examples include:

- Including power transformer attributes in the graph data. These are currently excluded because they have “load_network_type” and “source_network_type” assets rather than just “network_type”. These could be included by splitting the power transformer assets into a “power_transformer_source” node and a “power_transformer_load” node.
- Adding asset type and usage attributes to nodes.
- Adding circuit ID nodes and edges.

Wrong values: 230V -> 400V

The single largest group of suggested changes in the data concern assets with a nominal_voltage_pp stored as 230V which the model suggests should be 400V. These cases are hard for the model to predict, given the information available to it, since 230V and 400V assets may be directly connected together without a substation, power transformer and other associated assets, since 230V is 1 phase from a 400V 3 phase circuit. Hence, while some of these suggestions seem correct (e.g. asset has running_3_phase attribute set to True) or at least plausible, many seem incorrect, given the context.

There are several possible enhancements to the spatial graph network that might improve this. Examples include:

- Adding usage attributes to nodes
- Adding phasing (e.g. running 3 phase) attributes to nodes

Wrong values: protective equipment

There are a number of situations where protective equipment is flagged as having the wrong voltage attributes. Upon investigation, there appear to be inconsistencies in the way that the attributes of protective equipment in the database. These inconsistencies may be the cause of the relatively low accuracy scores for protective equipment observed during training.
For example:

- 400V PME earth spike connected to 230V earth wire (like RWO 82721188, associated with pole 26-3085-1)
- LV PME earth spike connected to 11kV earth wires (like RWO 204059937, 204059934 and 210298738, near substation 262634)
- LV PME earth spike connected to 11kV pole (like RWO 127967227, associated with pole 26FAA1[5])

Further inspection of the EO data has shown cases of 11kV earth spike connected to LV circuit and LV earth spikes connected to 33kV circuits that were not detected by the model.

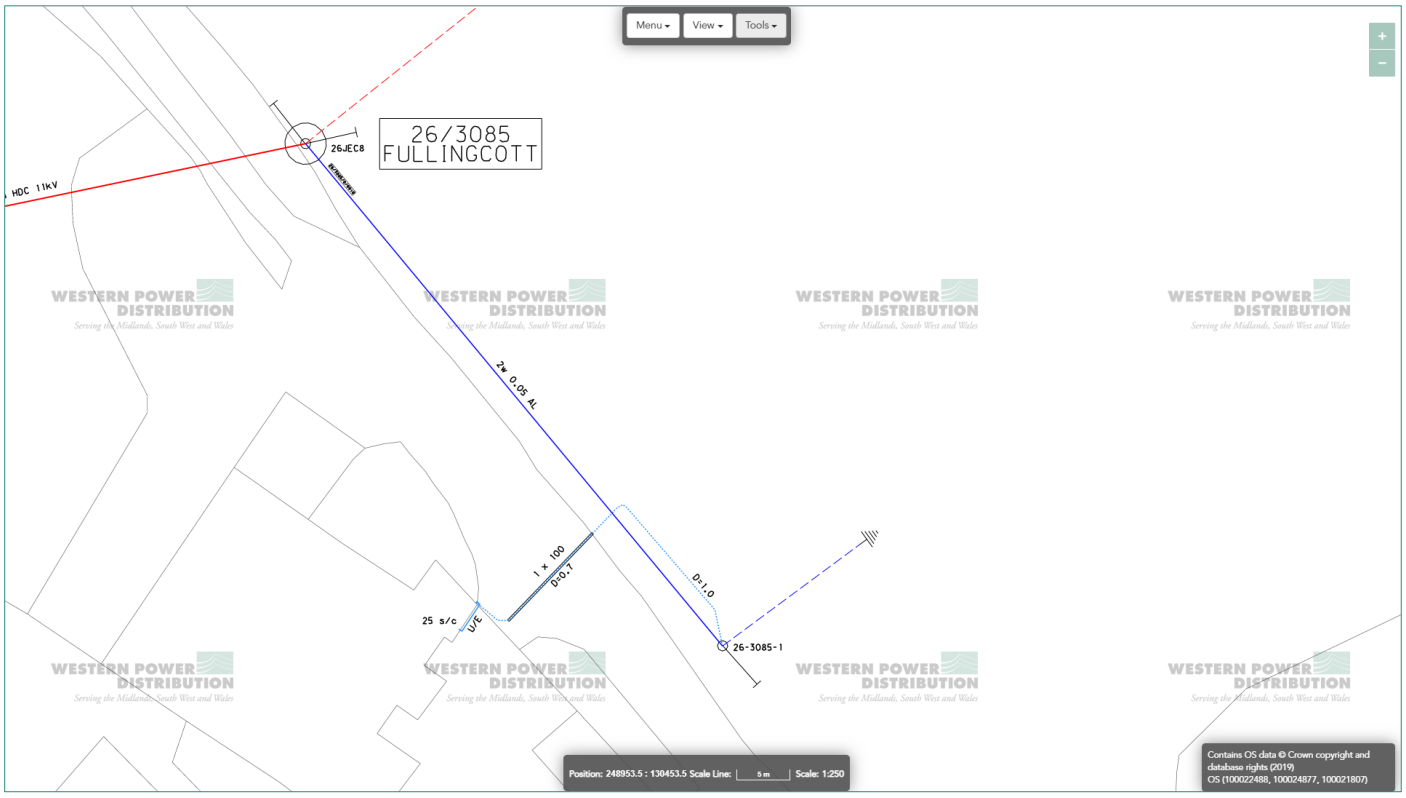


Figure 20: Map of RWO 82721188

[5] This is a slightly complicated scenario because pole 26FAA1 has both MV and LV keypoles associated with it (RWO 21711499 and 21711504 respectively), although only the MV keypole shares the same coordinates as the earth spike.

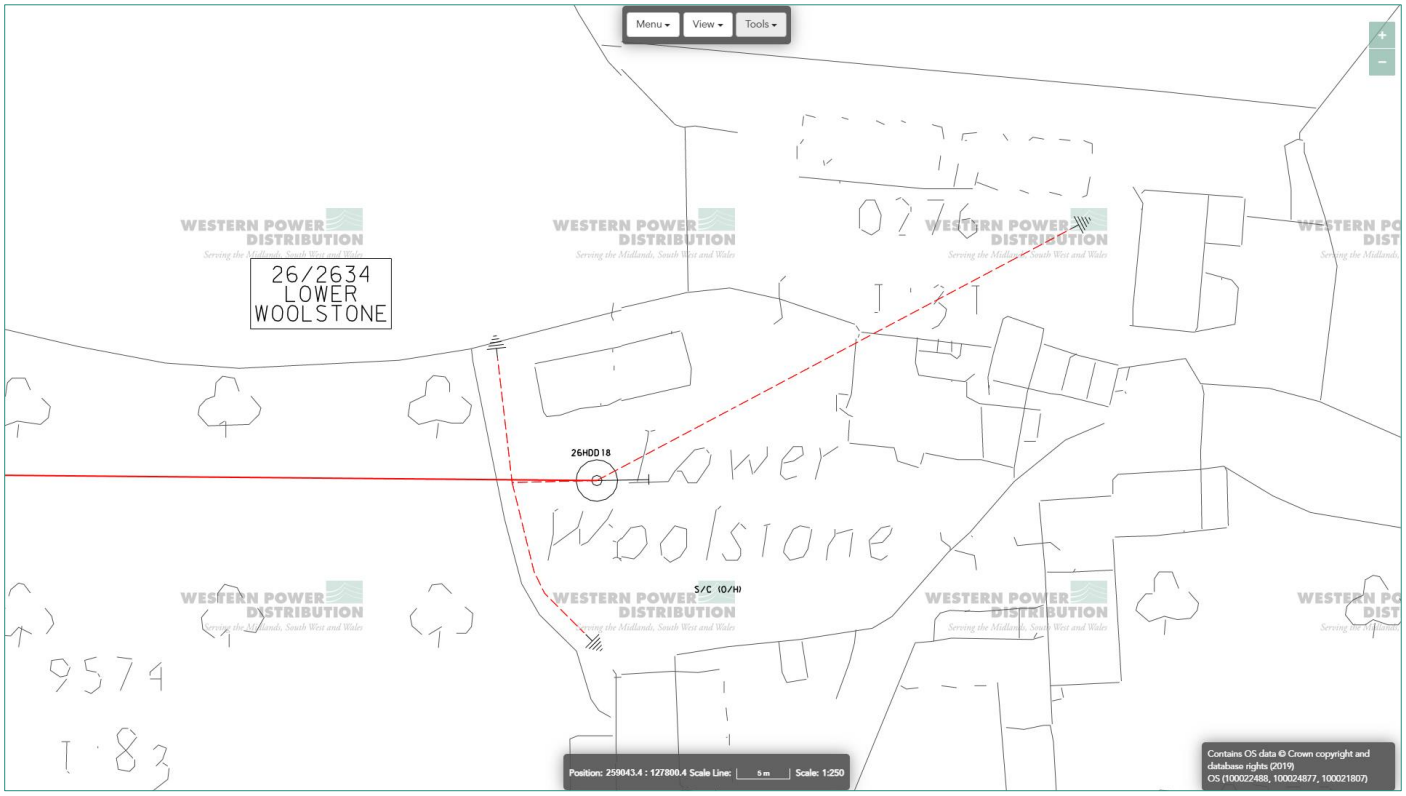


Figure 21: Map of RWO 204059937, 204059934 and 210298738

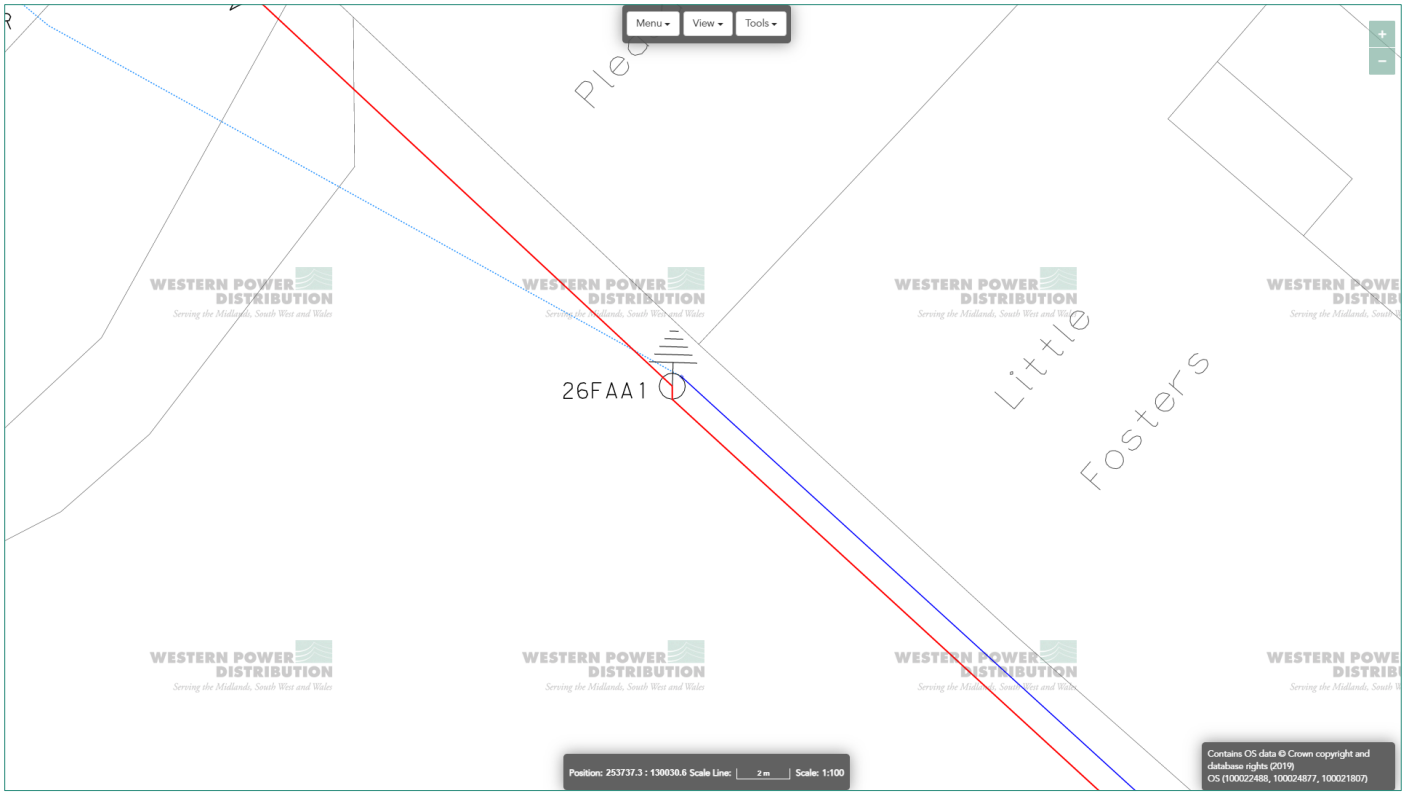


Figure 22: Map of RWO 127967227

If it is determined that some of these patterns are desirable (such as 400V earth spike with 230V earth wire), then adding asset type attributes to the spatial graph structure will enable the model to learn patterns that are specific to each asset type, rather than treating all asset types interchangeably. Improved modelling of earth wires, as previously discussed, would also be beneficial here.

Wrong values: specification material and size

It is hard to assess these outputs without specific input from an SME, except that some identified errors seem plausible, e.g. replacement with a category that matches some neighbouring assets, and some identified errors seem unlikely, like replacement with material that is not used for wires/cables as appropriate or replacement with a very large or very small size.

The first area for improvement that affects the over-prediction of unlikely values is the generation of synthetic errors. As discussed before, the current method of corrupting attribute values replaces the current value with any other value at random, without any other consideration. In particular, it means that there are observed values in the input data for training that match the errors observed above. Hence, improving the synthetic error generation to generate more realistic errors should lead to better suggested corrections for the original dataset.

Another opportunity for improvement is the inclusion of features from the connectivity graph model (model 1), such as consumer nodes and data, electrical connection edges, assumed or calculated capacity attributes, calculated headroom, etc. These sorts of features should aid the spatial graph model to make better predictions of the specification parts.

Overall observations:

- Many of the suggestions from the model for missing values seem correct, or at least plausible, and warrant further investigation.
- Many of the suggestions from the model for wrong values seem correct, or at least plausible, and warrant further investigation. While there are plenty of isolated examples, two patterns have been identified which need to be reviewed: one concerns the voltage attributes of protective equipment, and another concerns a group of connector points with incorrect nominal_voltage_pp attributes.
- Where there are patterns of suggested values seem wrong, some potential improvements have been identified and will be summarised in the “Next Steps” section. In particular this concerns earth wires, short rural circuits and distinction between 230V and 400V.

5. Comparison to the Integrated Network Model (INM)

Overview of the Integrated Network Model (INM)

The INM is a Master Data Management solution which: creates a single, canonical, reconciled version of electricity network asset master data that is mastered across a number of discrete systems and makes this available to other applications via data services, including CIM extracts; and tracks data anomalies, mismatches and other discrepancies while doing so and reports these to data stewards so the offending source data can be corrected.

The process for INM includes taking data from three source systems: Distribution Management System (DMS), Geographic Information System (GIS) and Enterprise Asset Management (EAM) and loading these extracts into relational database staging tables before transferring to the INM graph database to perform transformation and data reconciliation tasks. The transformed DMS data is used as the 'anchor' dataset from which the GIS and EAM datasets are matched to and reconciled.

Comparing the scope of INM and SEAM

A common objective of INM and SEAM is to improve network data quality by identifying inaccuracies and fixing the errors. The methodologies and scope of data for cleansing differ between the projects and are compared in this section of the report.

Component	INM	SEAM
Network type	11 kV, 33kV, 66kV and 132 kV networks	All network types (currently limited to assets in the Barnstaple area covered by the PoC)
Target data source cleanse	PowerOn (DMS) Electric Office (GIS) CROWN (EAM)	Electric Office (GIS)
Core system data sources	PowerOn (DMS) Electric Office (GIS) CROWN (EAM)	Electric Office (GIS) CROWN (EAM)
Additional data sources		UPRN (OS Open UPRN) Half Hourly Meter Readings (Durabill) Estimated Annual Consumption & Profile Class (Data Aggregators) Cable and Wire Specifications (WPD Company Directives)
Methodology summary	Master Data Management solution that creates a single, canonical, reconciled version of electricity network asset master data across DMS, GIS and EAM systems.	Model 1: Power system network topology (built using asset GIS data) and asset attributes used to detect erroneous customer connections where demand is above asset capacity. Model 2: Spatial graph model (built using asset GIS data) focussed on predicting asset attributes and relationships with an emphasis on the spatial relationship between assets.
Data error identification	Validation rules applied during the data transformation process. Automated matching rules to correlate data across systems (e.g. fuzzy matching and proximity).	Model 1: Assesses the technical feasibility of power transportation to verify the composite data sources and find exceptions and technical violations. Model 2: An inductive graph neural network (GNN) based model that performs node classification to predict asset attributes and relationships based on spatial relationships.

Component	INM	SEAM
Data error fix	Validation errors are reported and resolved by IT and Data Steward staff by investigating the issue and resolving it in the source data.	Model 1: Connecting isolated graphs where micro-disconnects are vertex to vertex on ends of line segments; proximity match customers to circuits; exceptions and technical violations. Model 2: Predicted missing or incorrect values reported with an associated confidence score. An agreed method of reviewing and implementing potential corrections suggested by the models has not been defined in the PoC.

Table 14: INM and SEAM comparison

The INM project uses a rules-based approach to validate, reconcile, and master data across the three core WPD systems. The solution identifies potential errors and issues through two means:

- A set of validation rules (e.g. rating value invalid or out of range) that are applied during the transformation of the data into the canonical model. These issues are reported to IT and WPD Data Stewards to investigate and attempt to resolve.
- A matching process is carried out to correlate data from different source systems that describe different aspects of the same core network components. This process uses a range of automated rules (e.g. direct, inferred, fuzzy matching, etc.) to match the data components and then with the DMS as the 'anchor' the data is reconciled across GIS and EAM. Corrections are applied directly where a confidence threshold is met or added to a review list for IT and WPD Data Stewards where there is a lower level of confidence.

The primary focus of SEAM is on the LV and 11kV networks which represent the majority of the GIS data. The INM requires use of DMS data which currently doesn't include a representation of the LV network (except assets on the LV side of a transformer). A direct comparison cannot therefore be made with the results from SEAM on the LV network– but comparisons on the 11kV and higher voltage networks can potentially be made with the Spatial Graph Model.

An extract of unresolved data issues reported in INM for Barnstaple (identified by substations located within the geographic area covered by the PoC) were supplied for comparison. The most recent extract (May 2021) reveals that all issues in the area relate to POF14 ('Invalid value or units') or POF15 ('No circuit name found in POF'). This limits the comparison of results from SEAM so a historical extract was used which includes a broader set of issues that were previously identified by INM and subsequently resolved.

Table 15 presents a summary of the unresolved data issues from the historical INM report for the Barnstaple area. This shows no issues that directly relate to GIS data (issue code EO). The issues reported relate to the matching and validation of data between PowerOn and CROWN (issue code INM) and violation of the PowerOn validation criteria (issue code POF). These issue types are not errors that would potentially be identified by the Spatial Graph Model and cannot therefore be used for direct comparison.

Issue code	Description	Issue count
INM10	No matching CROWN asset for PowerOn Transformer	85
INM11	No matching PowerOn Transformer for CROWN asset	21
INM23	Only one side of transformer is connected	65
POF6	Cable termination not between OHL and underground cable	133
POF8	Cable termination does not have the expected number of connected nodes	117
POF19	Switch component has more than two connections and forms a parallel with a secondary transformer	48

Table 15: Historical INM unresolved data issues reported for the Barnstaple area

Table 16 compares the prevalence of data issues identified by INM and SEAM (high confidence). These are measured by the percentage of assets (as a proportion of the asset count within EO) with a reported error. The summary for SEAM includes independently the percentage of assets that have at least one attribute identified as

missing and at least one attribute predicted to be a wrong value at high confidence. The 'All correct' figures represent the percentage of assets that have no reported errors at any confidence level.

Network voltage[6]	Number of assets	SEAM Spatial Graph Model			INM
		Any missing attributes	Any wrong attributes	All correct	Data issues
400V[7]	66,766	30.8%	1.3%	53.2%	0.3%
11kV	21,258	12.9%	0.6%	84.1%	1.2%
33kV	3,591	6.5%	0.9%	91.2%	0.4%
132kV	577	0.0%	0.5%	96.5%	1.7%

Table 16: Summary of data issue prevalence for INM and SEAM

Given there is no overlap in the error types identified by SEAM and INM for the Barnstaple area, the comparison demonstrates the extent to which additional potential errors have been identified by SEAM during the PoC on 11kV and above networks – and potential errors identified on LV network assets that aren't covered by INM.

The Spatial Graph Model could be used to target the improvement of LV network data quality – which isn't covered by INM – by using only a limited number of attributes and the geospatial relationships, which all come directly from the EO dataset.

SEAM also has the potential to complement the INM approach on the 11kV and higher networks by identifying errors in the underlying data would not be picked up by the INM rules. In general, several of the data issue types identified by INM are dependent on the underlying values being correct in PowerOn or pre-defined validation rules. Further, there is potential for common error types identified by SEAM being used to form rules that are then added to INM.

[6] For SEAM this is based on original network voltage value in EO
[7] Note that 110V, 220V and 400V are all grouped into the '440V' category

6. Model outputs on independent test area

An independent test area was partitioned from the Barnstaple proof-of-concept data and was not used as part of the model development and evaluation. This has proven that the models can be applied to a different area of the network not used as part of their development and that they deliver comparable results. A summary of the results is presented in this section.

The training area has a bounding box with X in [248000, 264000] and Y in [126000, 141000]. The independent hold-out test area has a bounding box with X in [246000, 248000] and Y in [126000, 141000].

6.1. Model 1

As an independent test, the input data was clipped to the bounding box as described above and circuits within this area where analysed.

Results

For the 199 circuits with substation located within threshold and customers connected, 5 circuits were found to have power flow violations with customers not being supplied their full demand and 10 circuits were found to have cables/ wires with head room percentage below threshold set at 20%, using minimum aggregation for capacity backfilling.

circuit_id	n_cust	n_headroom_pc
352844/0/0010	2	25
353643/0/0030	2	13
352844/0/0020	2	6
352994/0/0010	2	4
350600/0/0040	2	2
355411/0/0020	0	39
355411/0/0040	0	21
355566/0/0010	0	7
353350/0/0010	0	5
350501/0/0030	0	4

Table 17: Independent test area Maximum Flow results

The majority of the violations were due to a combination of unknown capacity and missing service cables to customers, which appears similar to the area studied in the training set.

Many in the select area had ‘UNKNOWN SIZE’ attribute within the specification_description column; this means that the backfill applied is minimum of this type of cable in the circuit. Almost all of the violations are caused by bottlenecks at these cables / wires, leading to customers whose demand was not satisfied. This combined with low availability of service cables in the Electric Office data set in the exception cases highlight above create violations in the circuits.

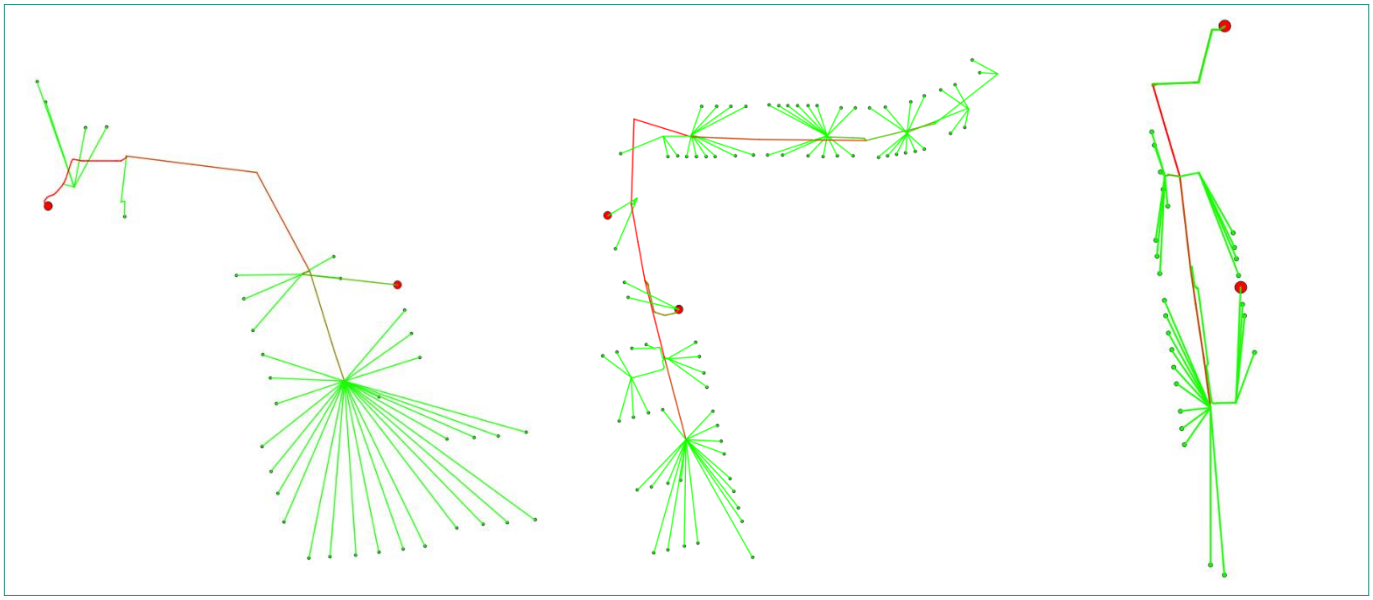


Figure 23 - Three circuits in the test area with customer demand not met; circuits cover a similar geographical area and have similar patterns of unknown cable properties. From left to right circuit_ids: 352844/0/0010, 352844/0/0020, 350600/0/0040.

For the circuit with circuit_id = 352844/0/0010, the bottleneck occurs in the cable directly connected to the substation which has specification_description = '3ph Unknown'. The capacity is then backfilled to the minimum of this type of distribution cable for this area. This bottleneck leads to a customer not being served further downstream. Also apparently in this circuit are a number of missing service cables, leading to bunching of customers at the end of a distribution cable. This pattern is similar to other circuits in this geographical area due to the pattern of missing cable specifications.

In the circuit with circuit_id = 352844/0/0020, the bottleneck in this circuit occurs in the distribution cable downstream from the first cable segment from the substation. Similar to the example above, the specification_description = '4w 3ph UNKNOWN SIZE' has no distinctive features and as such the capacity is backfilled to the minimum of this type of distribution cable. The example, circuit_id = 350600/0/0040, shares similarity to the previous two examples, with the bottleneck cable having specification_description = '4w 3ph UNKNOWN SIZE'.

6.2. Model 2

As an independent test, the model was retrained and evaluated on the training area and then predictions were created using the new model for the hold-out test area.

Training and Evaluation Results

The results of the training and evaluation processes are very similar[8] to the original model results shown in section 4. There are exceptions for the rare classes (<200 examples) and some low confidence outputs, which is to be expected since these are more sensitive to the exact errors that are added. In addition, the new trained model seems to do worse for service_point/nominal_voltage_pp and slightly worse for spec_material and spec_size.

asset_type	attribute_name	no_error	missing_value	wrong_value	missing_value_low	wrong_value_low
cable	**network_type**	99.91%	99.50%	99.58%	94.32%	92.25%
cable	**nominal_voltage_pp**	98.34%	92.05%	74.37%	78.41%	46.06%
cable	**spec_material**	98.76%	72.54%	52.15%	50.66%	24.88%
cable	**spec_size**	97.14%	49.81%	25.00%	33.28%	15.34%
connector_point	**network_type**	99.98%	100.00%	100.00%	98.76%	94.37%
connector_point	**nominal_voltage_pp**	97.98%	96.07%	92.06%	75.94%	58.63%
connector_segment	**network_type**	99.81%	98.93%	99.69%	84.68%	87.50%
connector_segment	**nominal_voltage_pp**	99.16%	96.62%	97.97%	83.83%	70.59%
energy_consumer	**network_type**	100.00%	100.00%	100.00%	95.00%	100.00%
energy_consumer	**nominal_voltage_pp**	97.11%	92.86%	100.00%	77.78%	60.00%
energy_source	**network_type**	100.00%			100.00%	
energy_source	**nominal_voltage_pp**	100.00%				
isolating_eqpt	**network_type**	99.87%	97.47%	96.67%	84.05%	69.05%
isolating_eqpt	**nominal_voltage_pp**	98.95%	91.15%	82.79%	83.38%	70.15%
keypole	**network_type**	99.97%	100.00%	100.00%	96.64%	95.00%
keypole	**nominal_voltage_pp**	99.94%	99.52%	99.00%	78.95%	81.82%
pole	**network_type**	99.74%	99.22%	98.66%	89.87%	87.50%
pole	**nominal_voltage_pp**	99.54%	98.61%	97.97%	74.42%	54.55%
protective_eqpt	**network_type**	98.72%	75.00%	62.96%	64.71%	54.55%
protective_eqpt	**nominal_voltage_pp**	98.01%	27.60%	38.81%	35.92%	54.55%
service_point	**network_type**	99.95%	100.00%	100.00%	98.02%	100.00%
service_point	**nominal_voltage_pp**	97.65%	65.38%	58.33%	64.86%	59.21%
tower	**network_type**	100.00%	100.00%	100.00%	66.67%	
tower	**nominal_voltage_pp**	100.00%	100.00%	50.00%	60.00%	100.00%
wire	**network_type**	99.94%	100.00%	99.80%	87.18%	83.93%

[8] Most accuracy values within a few percent

asset_type	attribute_name	no_error	missing_value	wrong_value	missing_value_low	wrong_value_low
wire	**nominal_voltage_pp**	99.34%	94.89%	61.51%	66.00%	15.56%
wire	**spec_material**	99.43%	89.54%	63.95%	49.84%	20.47%
wire	**spec_size**	98.23%	80.66%	43.39%	48.73%	21.40%

Table 18: Model 2 Evaluation Summary from Independent Training Process

Prediction Results

The results from the prediction process are hard to compare with the previous results since they apply to different areas of different sizes and different characteristics. However, there are a few observations that can be made by comparing ratios.

- Missing value:
  - Lower proportion of high confidence results for cable/spec_size compared with low confidence ones
  - Lower proportion of high confidence results for pole/nominal_voltage_pp and wire/spec_size compared with low confidence ones
  - Higher proportion of high confidence results for wire/spec_material compared with low confidence ones
  - Lower proportion of high confidence results for wire/spec_size compared with low confidence ones
- Wrong value:
  - Higher proportion of high confidence results for cable/nominal_voltage_pp compared with low confidence ones
  - Higher proportion of high confidence results for cable/spec_size compared with low confidence ones
  - Lower proportion of high confidence results for isolating_eqpt/network_type compared with low confidence ones
  - Lower proportion of wrong values (especially low confidence) for protective_eqpt/network_type and protective_eqpt/nominal_voltage_pp compared with no error

Given the shape of the hold-out test set area, which is only 2 km wide, it is perhaps to be expected that the distribution of confidence scores is different from the training set. Since the data are clipped at the boundary of the test area, assets that are close to this boundary may have different neighbourhoods compared with those well within the interior of the region. This may either mask relevant context for assets, reducing confidence score or mask confounding information, increasing confidence scores.

It is slightly unexpected that so few wrong values were detected for the protective_eqpt assets, since there are certainly examples of LV earth spike connected to 11kV earth wire appears in the hold-out test set too. However, since there are so many examples in the training data, it is not surprising that the model treats many of these as correct.

asset_type	attribute_name	no_error	missing_value	wrong_value	missing_value_low	wrong_value_low
cable	**network_type**	5985	0	0	0	2
cable	**nominal_voltage_pp**	5929	0	36	0	22
cable	**spec_material**	1292	2879	6	1793	17
cable	**spec_size**	1515	822	49	3580	21
connector_point	**network_type**	1987	0	0	0	0

asset_type	attribute_name	no_error	missing_value	wrong_value	missing_value_low	wrong_value_low
connector_point	**nominal_voltage_pp**	1982	0	2	0	3
connector_segment	**network_type**	877	0	0	0	0
connector_segment	**nominal_voltage_pp**	877	0	0	0	0
energy_consumer	**network_type**	45	0	0	0	0
energy_consumer	**nominal_voltage_pp**	45	0	0	0	0
energy_source	**network_type**	0	0	0	0	0
energy_source	**nominal_voltage_pp**	0	0	0	0	0
isolating_eqpt	**network_type**	472	0	0	0	9
isolating_eqpt	**nominal_voltage_pp**	471	0	7	0	3
keypole	**network_type**	400	0	0	0	1
keypole	**nominal_voltage_pp**	400	0	0	0	1
pole	**network_type**	388	481	0	72	1
pole	**nominal_voltage_pp**	388	120	1	433	0
protective_eqpt	**network_type**	281	0	0	0	2
protective_eqpt	**nominal_voltage_pp**	279	0	1	0	3
service_point	**network_type**	422	0	0	0	0
service_point	**nominal_voltage_pp**	422	0	0	0	0
tower	**network_type**	2	0	0	0	0
tower	**nominal_voltage_pp**	2	0	0	0	0
wire	**network_type**	1284	0	0	0	0
wire	**nominal_voltage_pp**	1222	0	31	0	31
wire	**spec_material**	795	260	16	195	18
wire	**spec_size**	839	76	9	326	34

Table 19: Model 2 Exceptions Summary from Independent Prediction Process

Investigation of Sample Errors

Overall, the sample of exceptions investigated follow the same patterns as for the training area, some of which are described here.

Missing values: network type and operational voltage

Appear correct:

- RWO 453230922 (pole number 26-0067-10) is classified as 400V and LV, which matches the surrounding assets.
- RWO 453230862 (pole number 26-0449-3) is classified as LV, which matches the surrounding assets. Operational voltage is predicted to be 400V with low confidence and the pole is between a 400V wire and a 230V cable.
- RWO 174290830 (no pole number) is classified as 11kV and MV, which matches the surrounding assets.

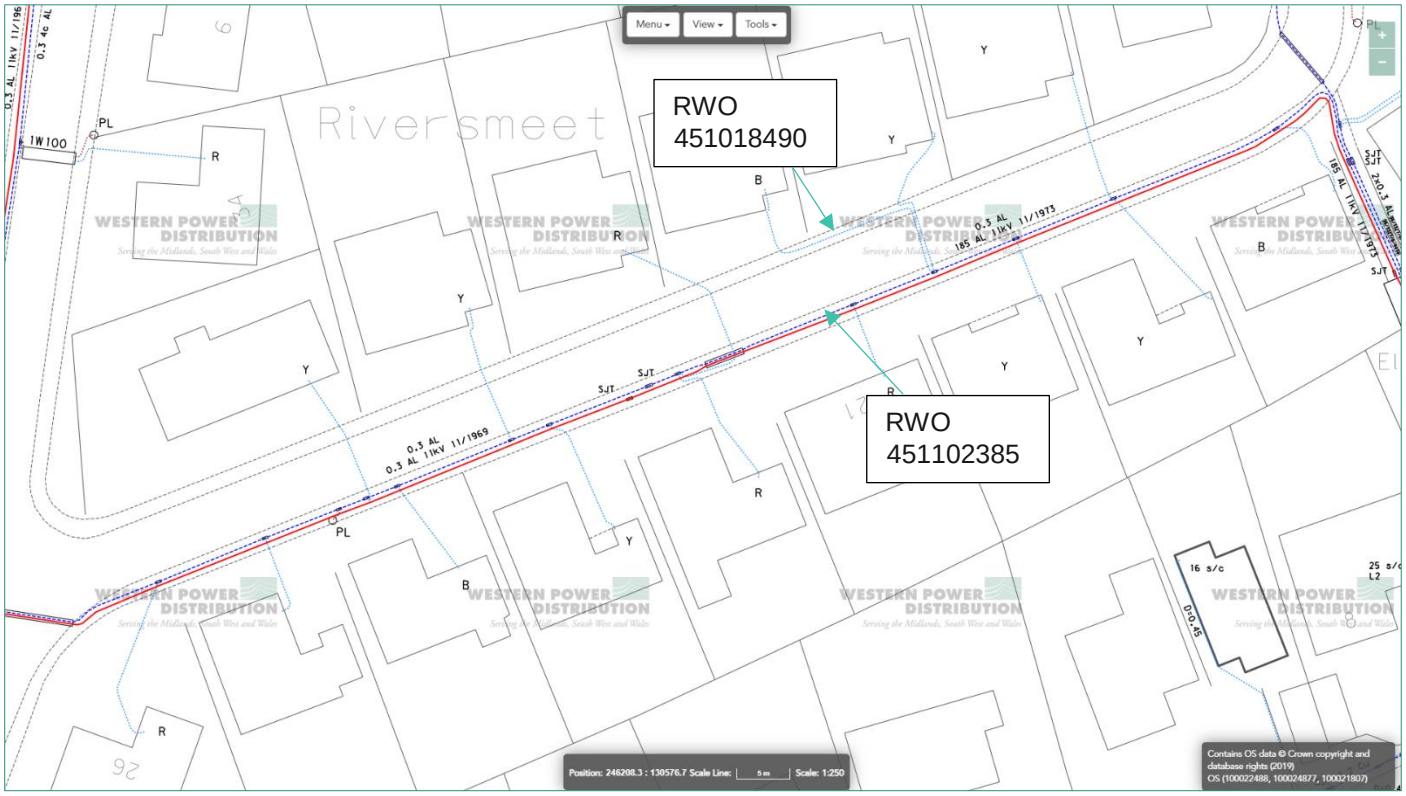
Appear incorrect:

- RWO 457217403 (pole number 26-0849-4) is classified as 11kV, but the associated wire is 230V. Network type was predicted to be MV with low confidence.

Missing values: specification material and size

Appear plausible:

- RWO 451102385 is a 400V distribution cable (dark blue) with the original specification description of “3ph unknown”. This is classified by the model as “WCON”, which is a common material for LV distribution cables. The size is predicted to be “140-280” with low confidence, which is also plausible. The map annotation seems to suggest that the correct specification is “0.3 AL”, which matches the size but not the material.
- The service cables surrounding RWO 451102385 all have original specification values of “SV Unknown from Unattributed” and are classified as “c/c”, which is a common material for service cables. The map annotations do not contain any hints.
- For most of the service cables surrounding RWO 451102385, the predicted size has low confidence. However, for RWO 451018490, which is a 230V service cable (light blue), the predicted value is “10-20” with high confidence. This is also consistent with its usage since there are plenty of examples of “16 c/c”.



Wrong values: 11kV -> 230V or 400V

All of the exceptions involving assets currently attributed as 11kV are listed below.

- RWO 450733713 is a zero-length cable. This is in the dataset because it starts at the edge of the hold-out test area and extends away from it. Hence, when the assets are clipped to the selected region, it ends up as a point. Assets like these should be removed from the dataset. Nevertheless, this particular location is within a conduit carrying 400V and 11kV cables and an 11kV Earthwire that are all aligned with the centre of the conduct. Hence, given the information available to the model, 400V is a plausible value even if 11kV is the correct one.
- RWO 451106018 is a connector point (wall box) like the ones discussed previously. The connected assets are all 230V, which matches the prediction from the model. Hence, this is regarded as correct.
- RWO 89475513 (pole number 26JF5) is a pole that is associated with a 11kV to 400V pole-mounted substation (number 260020). The co-ordinates of the pole match the 400V wires and cables, while the co-ordinates of the 11kV wires and key pole are very close by, but not exactly the same. By pure proximity, the pole operational voltage should be 400V, as identified by the model, but the original value of 11kV is also justifiable. Hence, this is regarded as plausible.
- RWO 343890016 is an Earth spike surrounded by 230V assets. The model identifies that the operational voltage should be 230V. Hence, this is regarded as correct.

Wrong values: 400V → 132kV

RWO 311749079 is an LV distribution cable, but the model suggests this should be 132kV, which is wrong. It is not close to any 132kV assets and most of the surrounding assets are 11kV or 33kV, since it associated wit an isolated small pole-mounted substation. It is unexpected that this prediction should be made, but it is likely to be a consequence of the distribution of simulated errors being uniform across all possible values.

There is also a question about whether edges in the spatial mesh above a certain distance should be removed. The closest distance between RWO 311749079 and the 33kV circuit is about 630m and edges like this are probably not useful for message passing in the graph neural network.

Wrong values: specification material and size

As before, it is hard to assess these outputs without specific input from an SME. As before, there appears to be a mixture of outputs that seem plausible and outputs that seem wrong.

7. Next steps: Business-as-usual implementation and model performance improvement

The recommended next steps for the project are grouped as follows with an assessment by the project team of the relative priority and effort (low/medium/high) to implement each step.

Recommendation		Priority	Effort
1.	Transition to BaU	High — puts the tool into the hands of users	Medium — depends on business requirements and objectives
2.	Feedback from users	High — sets direction for next phase	Low — based on existing deliverables from this phase
3.	"Quick wins"	High — directly addresses some findings from this report	Low — changes have already been identified and are relatively straightforward
4.	Combine models	High — increases the performance of the models	Low — mostly makes use of existing functionality
5.	Scale-up	Medium — increases the scope of the model	High but flexible — many different options within this category
6.	Blue skies	Low — exploits the model for new use cases	N/A

Table 20: Summary of recommendations

Improvements that have already been identified above as part of the evaluation are highlighted in **bold**.

7.1. Transition to BaU

(Priority = High, Effort = Medium)

Process for resolving data issues

The proof of concept model can identify potential errors and predict missing values in the data on the Barnstaple area within a matter of minutes. However, the processes to resolve the identified data issues are currently evolving within WPD. For some issues there may be sufficient contextual information in the data for a person to confirm or reject proposed values without requiring a site visit. E.g. an unknown cable type is proposed and a text label on the GIS (not used in the predictive modelling) confirms the cable size. For overhead assets it may be possible to gather information from the field if there is work in that area, but for underground assets it may be a very long time before the assets in question are exposed and therefore direct validation in the field will not be an option.

There is a need to keep source data in the GIS system (assumed to be captured at installation with an assumed level of accuracy) separate from predicted values from the models. It is therefore not appropriate to take the model outputs at face value and overwrite the original data. However, providing this data, so that it can be optionally substituted by the user would provide the intended benefits of the project, ensuring that incomplete or incorrect values do not prevent the data being useful and providing a standardised way of overcoming the issues so that third party use of the data is consistent. It is likely that there will be a range of possible methods to backfill data so it may be useful to extend the ability to record predicted values for asset data to also include the methodology used and a confidence metric so that the best alternative can be selected. As models are re-run and confidence metrics change the "best" backfill value may change so recording the date at which the backfill value was calculated may also be useful.

The work currently being carried out by Scottish Power to use smart meter data to validate LV network connectivity and cable types is of particular interest as this provides an alternative means to sense-check the LV network information without the need for site visits. In their modelling, the Thevenin's equivalent line impendence for the total circuit is used as an input variable, along with distance from the smart meter to the substation, power consumption and voltage levels at the smart meter (which does not require anonymisation by being grouped with data from other customers) to create a predictive model for the voltage distribution in time and distance for the circuit. This is used to

monitor for potential voltage violations to compare the actual voltages seen on the network / from the smart meters. This methodology, alongside using PSSE to assess the full power flow, is used to verify network topology and monitor potential violations. Their distance based neural network model enables the monitoring of parts of the network where the penetration of smart meters may be low; utilising the spatial relationships between the smart meter and the substation.

Technical integration of the tool

The SEAM model was developed as a standalone, proof-of-concept tool. In the next phase, this should be transitioned to an integrated, business-as-usual tool. These activities have a similar priority to those in the "scale-up" group. Due to the modular, incremental design of the SEAM model, activities in this group can be implemented in parallel with those in the other groups.

There are four key areas of activities.

- **Integration** with other WPD systems, e.g. read asset data from PostGIS directly, export cleaned data to new database[9].
- **Productionisation**, i.e. increasing software quality to level required for BaU, including software testing, monitoring, model and data versioning, scheduling and batch processing, historical records of anomalies detected, etc.
- **User interface** improvements to support BaU, including presentation and tracking of output reports.
- **Deployment** to WPD systems, including whether to move to server-hosted service rather than standalone desktop application.

7.2. Feedback from users

(Priority = High, Effort = Low)

The models have been evaluated by the SEAM project team. In the next phase of development, getting feedback from end users and SMEs is the first priority. This should cover both retrospective topics, such as the errors that are being identified and the contents of the output reports, and prospective topics, such as types of error to focus on for the next phase. The deliverables from this phase of the project form the basis of this step.

Some specific topics for discussion with stakeholders are listed below; some of these may translate into "quick wins".

Summary of user feedback topics	
Model 1	• Review of connected graphs and method • Review of connected customers and method • Review of max flow outputs to align outputs with engineering and subject matter expert's experience • Review parsing of specification description into parts and derivation of electrical properties
Model 2	• Investigate identified exceptions. • **Review handling of protective equipment and Earth wires.** • **Review breakpoints for conversion of conductor sizes to categories.** • **Review simulation of data errors**, especially correlation between original and corrupted values.

Table 21: User feedback topics

7.3. "Quick wins"

(Priority = High, Effort = Low)

The SEAM model delivers good performance for the purposes of the PoC. In the next phase of development, there some straightforward improvements, already identified here, that can be made to the model that target the patterns of

[9] Having separate "original" and "cleansed" databases is one approach for ensuring that assumed values are only used for applications that don't require positive confirmation of any proposed changes.

false alarms identified in this report. These improvements should deliver significant benefit by increasing accuracy and confidence of the model predictions. It is recommended that these changes would be a high priority in future development of the model.

Model 1	
Enhancing connectivity	Further work can be done to develop the connectivity process in more detail. Acknowledging the ongoing work within WPD's digital team which uses line-extension based methods to create connectivity, the work in this project focused on methods which specifically addresses vertex-to-vertex disconnects; which were the most common type of disconnect in the Barnstaple area. Connectivity at LV level is a priority and investigating more methods of connection would improve modelling performance for this model as well as for wider WPD use.
Specification description analysis and rating matching	Create a more robust method for prediction / backfilling unknown or non-conforming values for specification descriptions for cables and wires to match against the WPD directives. Alongside this, increase the knowledge base of potential wire and cable specifications to ratings by using online resources / engineering models. This will enable more accurate predictions for capacity and to reduce the number of false positives in the model as currently the model produces many violations based on low approximations for unknown cable specifications.
Capacity bottleneck suggested rating	For bottleneck cables and wires, compute the minimum capacity with the threshold headroom to enable demand satisfaction for all consumers.
Model 2	
Improve modelling of assets	For example: • **Split power transformers into nodes for source and load sides** • **Separate Earth wires into distinct asset type**
Additional attributes as asset node features	Especially: • **Asset type** • **Usage** • **Running 3 phase**
Review breakpoints for conversion of conductor sizes to categories.	
Review simulation of data errors, especially correlation between original and corrupted values.	
Add circuit ID nodes and edges	
Exclude invalid assets	For example: • **Zero length linear assets** — these are artifacts of the way that the assets are selected then clipped to the area of interest • "special" circuit IDs — some circuit IDs (e.g. 979997/9997 and 989998/9998) are for assets that are not part of the network proper

Table 22: Summary of recommended "quick wins"

7.4. Combine models

(Priority = High, Effort = Low)

The SEAM model is currently implemented as two separate parts. In the next phase of development, these should be combined, since they have complementary functionality and combining them will enhance the quality of information that each has available to work with, thereby increasing accuracy and confidence of the model predictions. For example:

- The electrical information (e.g. electrical connectivity and conductor capacity) and consumer data that are inputs to model 1 can be added as features in the spatial graph for model 2. This additional information will enable model 2 to make better predictions for the attributes in the model.
- Outputs of model 2 can be used to backfill the missing conductor attributes required for model 1. This will lead to better assumptions about the capacity of each conductor, and hence to higher quality predictions of the headroom available for each conductor.

Some specific tasks that could be included in this step are listed below.

Model 1	
Improve backfilling capability	Improving the backfilling capability by using the graph structure, i.e. by converting the line assets into connected nodes, graph neural networks / model 2 can be used to verify existing / predict missing attributes.
Projections for customer power consumption	Using Model 1's spatial mesh to create projections for customer power consumption without smart meters using existing smart meter data.
Model 2	
Basic features from model 1	For example: • Edges for (assumed) electrical connectivity between electrical assets • Add basic attributes like capacity • Add nodes for (metered) consumers and attributes for simplified profiles, e.g. EAC
Additional outputs	For example: • Predict remaining specification parts, such as number of conductors • Regression output for capacity
Advanced feature from model 1	For example: • Add complex attributes like headroom • Add detailed consumer profiles, e.g. half-hourly demand

Table 23: Tasks for combining SEAM models

7.5. Scale-up

(Priority = Medium, Effort = High)

The SEAM model has a flexible design that can be incrementally enhanced and extended. In the next phase of development, the scope of the model should be "scaled-up" to maximise its value as a BaU tool. The model was developed as a proof-of-concept for a limited test area using a subset of the data available targeting a subset of the attributes in the GIS dataset, and there are lots of opportunities to extend this. This group has a very flexible scope and has similar priority to the "transition to BaU" group.

The main directions that the model can be "scaled-up" are as follows.

Enhancements to scale-up the model	
Additional area	The end goal is to be able to use the SEAM tool throughout the WPD-licensed areas. However, this involves roughly 3 orders of magnitude more data, judging by area, so the current approach cannot scale directly. Various techniques must be explored to control the memory usage and the time required to train the model and to analyse any given area, and the design of the model will need to be adapted to handle the increased range of patterns observed in the larger area. This also provides the opportunity to train and test the model on separate areas, e.g. train on Devon and test on Cornwall or target a more complex section of the network (e.g. a city centre such as Exeter).
Additional data (Model 2)	The current data sources include attributes that are currently unused but could be added to the model (e.g. pole–keypole links, wire: specification_description_2). It would also be valuable to include assets from a "buffer zone" around the area of interest, in order to ensure that all assets in that area have full contextual information. There are additional external data sources that could be integrated into the graph model to provide additional data. For example: • Map annotations in WPD DataPortal2 (e.g. service data from properties, public lighting locations[10]) • Property data (e.g. from OS AddressBase)
Additional functionality (Model 2)	The model can be extended to identify and correct errors in other EO attributes, such as usage, phasing, circuit membership, etc, to provide regression outputs (as appropriate) and to output the top few suggestions (rather than just one) when the scores are similar. The max-flow analysis can be extended to more circuits, such as 11kV and higher.
Additional performance (Model 2)	Since the focus of this project was to develop a proof-of-concept, there is scope to optimize the design and hyperparameters of the model to obtain higher overall performance for the same input data. These aspects of the model will need to be investigated to some extent as part of any other change to the model, in order to handle more data and more complex patterns. For example:

[10] Public lighting is useful to understand it represents the unmetered consumers.

	• Simulation of data errors (e.g. distribution of errors, correlations, multiple errors) — to simulate more realistic errors and hence to better detect and correct real errors. • Neural network architecture (e.g. number, type and size of layers; learning rate and number of epochs) — to improve the capacity of the model to learn different patterns in the data. • Spatial mesh construction (e.g. choice of edges to include, edge weight calculation, handling of locations within tolerance) and additional message-passing layers (e.g. higher-level grouping of locations) — to improve the ability of the model to gather information about the local neighbourhood of each asset.
Extend model to substation, MV, HV (Model 1)	The modelling is currently only conducted at the LV level per unique feeder. This modelling could easily be extended to the substation, this the potential to be extended through MV and HV. Acknowledging that the connectivity data is better at HV levels, this part of the network can be used to verify the circuitry and assets downstream.
Whole area integration (Model 1)	Test the whole area of data without iterating through circuit_id to test connectivity and how customers are connected, particularly if a circuit has mislabelled circuit_id. A number of stranded assets and segments of cables and wires are not currently assigned an operational circuit_id, and instead are bucketed within an error code which captures segments from across areas. These segments, for example, could be connected / create connectivity for other circuit_ids. Similarly, if a segment of wire is incorrectly labelled, the current method would not be able to create connectivity correctly. Therefore, an investigation into whole area connectivity could potentially improve the overall connectivity as there would be better visibility over the entire area dataset.
Increasing the asset base	• Add additional assets into model to verify connectivity / distinguish how things are connected including junctions, open points, isolating equipment. Additional demand sinks, streetlights, DERs, new connections to better reflect real life and identify where there are currently false negatives in the model. • This will also enable more accurate customer connections as it eliminates more connections on the circuit; currently customers are connected to the nearest cable / wire which is not connected to any other asset; where there are existing assets that can be added to the model, eliminating nodes where no assets are connected and improving accuracy of customer connections.

Table 24: Potential enhancements to "scale-up" SEAM models

7.6. Blue Skies

(Priority = Low, Effort = N/A)

The SEAM model is focussed on cleansing EO attributes. In the next phase of development, the underlying graph models could be exploited to pursue additional use cases. Potential examples are included in the table below:

Potential additional uses of the model	
Link to INM master data (e.g. via CIM export)	To provide more data (especially links) at 11kV+ and bridge functionality with INM
Identify and correct attributes in linked datasets e.g. CROWN attributes, consumer profile, etc.	To increase data quality in linked datasets too
Group circuits based on characteristics via graph clustering	To identify groups of similar circuits
Identify non-compliant circuits via graph classification	To provide an alternative method for circuit analysis
Clustering of consumers	To analyse and predict consumer profiles
Network simulation	To see how predictions change if selected attributes are modified or if assets are added or removed
Graph queries	To provide an approach for querying the network based on relationships as well as attributes
Link LCT data	To analyse LCT installation rates and impact
Link work orders and/or wayleaves	To analyse changes being made to the network
Link smart meter data	To analyse consumption patterns and validate LV connectivity
Increase modelling complexity	Consider modelling voltage and other circuit attributes, extending the simplistic transportation model; other optimisation models exist to apply to

	more complex formulations of the problem, this could include moving from a transportation model to the use of a full power flow analysis using a tool such as WinDEBUT or LV Connect which provide a consistent assessment of the network capacity as is used by the network planners.
Improve accuracy of customer demand data	As the smart meter roll out continues, more accurate half hourly data would better reflect actual diversity of peak demand – currently this is approximated by using a formulation and is not exact. Having accurate customer demand data would resolve issues around assumed levels of diversity that are built into the customer profiles and would also reflect the difference between individual customers more accurately. It would be expected to reduce the number of cases where the network capacity was incorrectly flagged as being insufficient and / or detect unusual high demand areas.

Table 25: Potential additiona*l use cases*

Appendix 1: Transportation Modelling / Maximum Flow results

Transportation Modelling / Maximum Flow results

circuit_id	n_cust	n_headroom_pc
352844/0/0020	3	8
260584/0/0030	2	60
263168/0/0010	2	54
263219/0/0010	2	30
352844/0/0010	2	25
353643/0/0030	2	13
263297/0/0020	2	9
260650/0/0010	2	8
263293/0/0020	2	7
262417/0/0020	2	6
260201/0/0010	2	4
261584/0/0010	2	3
260304/0/0020	2	3
352994/0/0010	2	2
260258/0/0050	2	2
265201/0/0010	2	2
260304/0/0010	2	2
350600/0/0040	2	2
263130/0/0010	2	2
352994/0/0010	2	2
260974/0/0040	2	1
350501/0/0030	1	4
263278/0/0020	1	3
260452/0/0010	1	2
262687/0/0010	1	2
262919/0/0030	0	43
264452/0/0030	0	39
355411/0/0020	0	39
264862/0/0030	0	28
262670/0/0020	0	25
263015/0/0020	0	25
355411/0/0040	0	21

circuit_id	n_cust	n_headroom_pc
263258/0/0030	0	19
261039/0/0060	0	19
262982/0/0010	0	16
263049/0/0020	0	15
260452/0/0020	0	13
261722/0/0040	0	13
260033/0/0010	0	12
262806/0/0030	0	12
263373/0/0010	0	12
262968/0/0030	0	12
263278/0/0010	0	11
265557/0/0020	0	9
264945/0/0010	0	8
262745/0/0010	0	8
262246/0/0020	0	8
262968/0/0040	0	7
262246/0/0030	0	7
355566/0/0010	0	7
261297/0/0040	0	6
353350/0/0010	0	5
263045/0/0020	0	4
265557/0/0010	0	3
262806/0/0010	0	3
262592/0/0010	0	2
261708/0/0010	0	1
260594/0/0010	0	1
260041/0/0020	0	1

Appendix 2: QGIS tips

A2.1 Opening GeoPackage files

In QGIS terminology, “opening a file” is described as “creating a layer”. The current QGIS manual has a section on creating a new GeoPackage layer, which is also valid for QGIS 3.18. It is also possible to open a GeoPackage file by dragging-and-dropping it into an open QGIS window.

A2.2 Finding assets from well-known text (WKT)

The WKT contains pairs of XY co-ordinates in OSGB 1936 / British National Grid. If the QGIS project is set to use this CRS, then the coordinates of any point from the asset geometry can be pasted into the “Coordinate” text box on the bottom status bar to quickly navigate to that asset.

A2.3 Finding assets by attributes

The “Search Layers” plugin for QGIS provides the functionality to search for assets by attribute names (https://plugins.qgis.org/plugins/searchlayers/). This can be installed and accessed via the “Plugins” drop-down menu.

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	WPD 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)	WPD's geospatial system which displays the network layout at all voltages
Geospatial Information System (GIS)	A data system capable of capturing, storing, analysing, and displaying geographically referenced information.
Integrated Network Model (INM)	WPD's 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	WPD's 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
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 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.

WPD INNOVATION

Transforming the electricity network

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