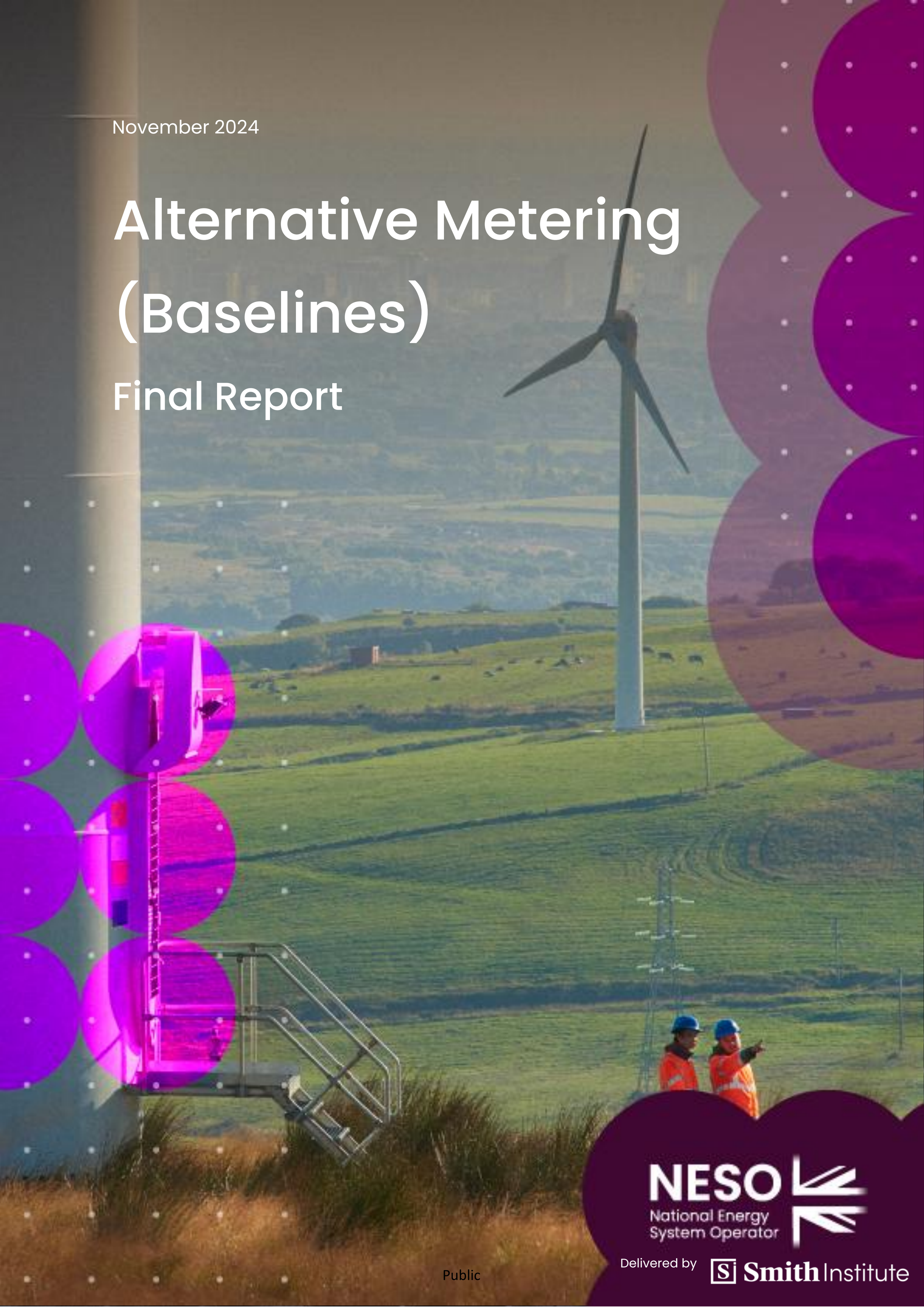
November 2024

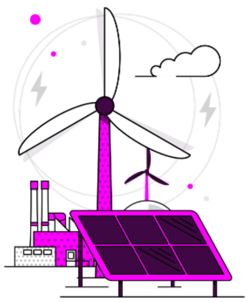
Alternative Metering (Baselines)

Final Report



Delivered by Smith Institute

Contents



1. Executive Summary

Context

What problem does this innovation project address?

What technical approach did the project take?

What recommendations can we take from this?

Context

NESO handles the difficult challenge of keeping the lights on for GB – whilst at the same time ensuring security of supply and enabling the transition to a low carbon future. Frequency Response Services enable NESO to keep the grid at the required frequency of 50Hz on a second-by-second basis.

Providers of Frequency Response are required to submit baselines when qualifying to participate in the services. NESO and industry have identified that the current method of handling baselines serves as a barrier to entry for assets with complex delivery requirements (e.g. batteries) or for assets that participate in more than one Ancillary Service.

To remove this obstacle, NESO would like to allow providers to submit delivery profiles with baselines that vary from their declared position at gate closure. To ensure system security, however, NESO must first develop a reliable algorithm and scoring system which flags the use of varying baselines to game performance data.

What problem does this innovation project address?

This project scopes, designs and implements a process, proof-of-concept algorithmic model and scoring system which would enable NESO to allow for the submission of varying baselines securely, through the detection of patterns of submission that may be anomalous and therefore unhelpful for system security. This will ensure that NESO are able to ensure fair market participation and facilitate Clean Power 2030 by utilising the services of more LCTs.

What technical approach did the project take?

Smith Institute engaged with NESO through a series workshops, which ascertained where gaming opportunities may arise in the system and what data sources might be analysed to reveal anomalous behaviour. We then reverse-engineered analysis pipelines from these scenarios, selecting relevant data inputs from the available streams of data. For example, it is suspicious if a baseline is correlated with the ideal response curve. We analyse data from DFR contracts and grid frequency and raise a flag when this correlation is above a threshold.

As well as correlation analysis, we also flag when the active power versus the ideal response is outside performance bounds. We use anomaly detection methods to flag more complex behaviours. These include: comparing reported and expected baselines, comparing active power between performance monitoring and operational metering, detecting high/periodic levels of unavailability, and price incentives. These were selected to provide the appropriate balance between detective power and explainability.

The checks require data from: DFR contracts, performance monitoring, operational metering, PNs and BOAs (BMUs only), market prices and unavailability. This data is read from APIs or local files. The gaming detection model performs each check to determine an aggregated gaming score, between 0 and 1, representing how suspicious a unit's behaviour is for a given settlement period. Each check has a score of 1 if any flags are present, 0 otherwise, and the aggregated score is the mean of these.

There is no historical baseline data for units with varying baselines, so we synthesised this for testing and demonstration purposes. In addition to synthetic data generation, the process relies on anomaly detection methods. These include changepoint detection, to identify abrupt changes in the statistical properties of a timeseries, and periodic anomaly detection, to flag points where autocorrelation exceeds a certain threshold. We also used timeseries forecasting to predict future data points based on historical patterns, flagging significant deviations from these predictions.

What recommendations can we take from this?

Optimising model parameters, e.g. thresholds for flagging suspicious behaviour, relies on high-quality data indicative of gaming. However, this is currently limited for units with varying baselines and delivery profiles. ESO should consider whether checks need to be live before varying baseline units are incorporated into DRS, or how soon after they are needed. There will be a trade-off between accuracy of detections and length of time without any detections available. We recommend gathering as much real varying baseline data as possible, and recording any suspicious incidents, including those that prove to be legitimate behaviour.

2. Frequency Response Baseline Gaming Solution Design

Introduction

Technical Approach

Introduction

In the future, ESO plans to allow the submission of variable baselines. This will lower the barrier to entry for participation, particularly for assets that participate in multiple Ancillary Services, or with complex deliveries (e.g. batteries). There is a possibility that participants in this market may choose to behave in ways which would not be in keeping with market rules.

Smith Institute is designing a proof-of-concept for a method to detect manipulation or gaming in Frequency Response Services performance data, based on submissions of variable baselines and response profiles post-delivery. To qualify for participation, units must submit baselines showing that they can deliver the required response. The methods designed by Smith Institute aim to detect gaming behaviours in the submission of these variable baselines.

In this document, we outline the approach to designing the method which will be demonstrated in a proof-of-concept. The input data collated for the solution is described, and we provide details on the algorithmic process for gaming detection and synthetic data generation.

For ease of reading, a list of abbreviations used in this document are listed below.

Abbreviation	In full
API	Application programming interface
ASDP	Ancillary service dispatch platform
BM	Balancing mechanism
BOA	Bid/offer acceptance
DC	Dynamic containment
DM	Dynamic moderation
DR	Dynamic response
EFA	Electricity forward agreement
FPN	Final physical notification
NIV	Net imbalance volume
OM	Operational metering
PM	Performance monitoring
PN	Physical notification
SP	System price

Technical Approach

Smith Institute proposed to complete this project by moving through a three-stage process with NESO, which is outlined below:

1) A process map and agreed scope	Smith Institute: • Held joint SI-ESO workshops for requirements capture and scoping. • Created a modelling process map. • Delivered a scoping conclusions document for agreement of ESO stakeholders.
2) A summary of the design of the agreed solution, including a scoring system	Smith Institute: • Evaluated potential approaches, from low to high data needs, balancing power with computational and IT complexity. • Constructed a protocol for categorisation of manipulation risk in data series. • Delivered a document describing the solution to be implemented, agreed with ESO.
3) End to end development scripts	Smith Institute: • Implemented the agreed algorithm. • Built a synthetic data generation engine, to augment the evaluation and appraisal of model performance. • Ensured all code is clear and fully commented, ready to be picked up by ESO IT SMEs in productionisation and Delivered to ESO the final model scripts.

The result of the first workshop was an agreed list of candidate approaches, which were then evaluated to assess their data needs. In order to perform these evaluations, Smith Institute collated input data from a number of input sources, which is described in the section below.

1. Collated Input Data

The data that was collated for analysis is:

1.1 DC, DR and DM Results by unit

This data is the Dynamic Containment, Dynamic Regulation & Dynamic Moderation contracts of units participating in Dynamic Frequency Response services. This service manages second-by-second changes in grid frequency, for maintaining 50Hz +/- 1%.
This data is fetched from ESO's data portal via an API call. For a given unit, key fields include the service, delivery start and end time, and the cleared volume.
For each unit providing the FR service, this data contains the time from which delivery commences and ends, and the volume of response that was accepted as per the cleared price. The time between service commencement and end is always 4 hours, as per the Electricity Forward Agreement, and that the week is split into 42 blocks with 6 four-hour blocks each day.

1.2 Notification of Availability/Outage

This data is submitted by units which communicates their availability to provide frequency response services and notifies ESO of any service specific outages. Non-BM units communicate availability information via an API with the ASDP platform, which is different from the API used by BM units.
The Smith Institute has access to examples of this data through four anonymised csv files, shared on the 27th of November. The fields are: start time, end time, available capacity and service. Availability is reported per-settlement period.
In the future, the Notification of Availability/Outage data will replace the 'Availability' field in historic Operational Metering data.

1.3 Physical Notifications (PNs)

This data describes a unit's expected generation or demand for a given settlement period. It is a series of MW figures and associated times, forming a profile of intended input or output of Active Power. Final PNs

(FPNs), which are submitted at gate closure 1 hour before delivery, give ESO a forward plan of projected imbalance levels.
PNs for BM units are accessed via an API on bmrs.elexon.co.uk. Non-BM units submit PN data via the ASDP, which the Smith Institute does not currently have access to. The data contains fields for: time from, time to, level from and level to.

1.4 Bid-Offer-Acceptances (BOAs)

This data contains the formal instructions of the acceptance of offers and bids by ESO in operating the Balancing Mechanism (BM). The BM spans from gate closure to the end of the Settlement Period. During this, ESO issues BOAs which instruct units to deviate from their PN position. The issuing of BOAs aims to balance generation and demand within acceptable frequency limits without the need for response.
This data is accessed via an API on bmrs.elexon.co.uk.

1.5 Operational Metering (OM)

Operational metering is used in the Balancing Mechanism and Ancillary Services markets to communicate asset output in real-time to the ESO control room. It is used not only to inform real time grid balancing, but also downstream for system modelling and to look at short term demand.
The Smith Institute relies on ESO to send us OM data, in line with the example received on the 13th of October, of an anonymised csv file. The key field in this dataset is the unit's Active Power in MW at a given time.

1.6 Performance Monitoring (PM)

This is data submitted by units participating in frequency response services. It can be used to assess how well a unit is meeting its obligations under its agreement with ESO. It describes a unit's availability, its actual response to frequency deviations, and any periods of declared unavailability. It also contains information on whether a unit is armed for a given service (DC/DM/DR, low or high frequency).
The Smith Institute relies on ESO to send us PM data, in line with the example received on the 13th of October, of two anonymised parquet files. Key fields include: the frequency, the unit's reported baseline, its reported

active power, availability, and armed status.

1.7 System Price and Net Imbalance Volume (SP & NIV)

This is market-wide data that describes the net imbalance volume (in MWh) of the total system for a given Settlement Period. The System Sell Price (SSP) and System Buy Price (SBP) are used to settle the difference between contracted generation, or consumption, and the amount that was generated or consumed in each period.

When units cannot or do not deliver according to their FPNs and BOAs, additional volumes are found to make up for shortfalls. These "out of position" units are charged at the system price, charged per MWh. If a BM participant generates less or uses more power than dictated by FPNs and BOAs, it buys the shortfall at the system price. Vice-versa, if a participant generates more or uses less power than dictated, it sells the excess at the system price.

This data is accessed via an API on bmreports.com (section 5.2.43, "Derived System Wide Data").

This is summarised in Table 1. Data sources summary.

Data source	Location
DC, DR, DM Results by Unit	Publicly available via API
Notification of availability/outage	ESO to provide
Physical Notifications	Publicly available via API
Bid-Offer-Acceptances	Publicly available via API
Operation metering	ESO to provide
Performance monitoring	ESO to provide
System Price and Net Imbalance Volume	Publicly available via API

Table 1. Data sources summary

2. Algorithm

Once candidate evaluation was complete, a number of checks to include in the algorithm were agreed. These checks are described in the section following. They are divided into two categories: basic checks and anomaly detection.

2.1 Basic checks

2.1.1 PM versus OM active power outside threshold

The active power of a unit as reported in PM and OM data should generally be the same. Differences may be due to either gaming, or some other legitimate circumstances which required altering of the active power submitted in PM data.

Input(s)	Method	Output(s)
OM data PN data	Active power from PM data will have to be down-sampled to the sample rate of the OM data, for example by taking the average over a given time window. Calculate the difference between OM active power and (down-sampled) PM active power.	All timestamps when the difference is above some threshold value. If this flags up too much data, then alternatively the output could be a percentage of time when the difference was above some threshold, with perhaps a flag raised when the percentage is above some other threshold.

2.1.2 Reported active power versus ideal response outside performance bounds

The reported active power in PM data should be equal to the baseline plus the ideal response for the measured frequency, within performance bounds.

Input(s)	Method	Output(s)
PM data	The ideal response curve can be used to calculate what response a provider should have given for	Flag any times, or what proportion of the time, a unit's response

Ideal response curve Performance bounds definition	the observed frequency, taking account of which services they are signed up to. Deviations are expected, but they should be within performance bounds defined by ESO.	is outside the response performance bounds. Also flag if response is continually set very close to the upper or lower bound.

2.1.3 Correlation between baseline and ideal response

In normal circumstances, the baseline reported in PM data should be independent of the ideal response. A correlation in these two variables indicates a provider may be adjusting their reported baseline to give the appearance of having delivered more response than was actually provided.

Input(s)	**Method**	**Output(s)**
PM data Ideal response curve	Compute the correlation coefficient between the reported baseline and the ideal response, given the observed frequency measurements. Some more sophisticated checks may be worthwhile, such as the mutual information score between the two variables.	Flag if the correlation coefficient (or mutual information score) is above some threshold.

2.2 Anomaly detection

A series of more sophisticated checks involving anomaly detection algorithms can be carried out to identify more complex gaming behaviour. The basic assumption here is that the majority of units are not engaging in gaming behaviour; providers who are gaming will therefore exhibit behaviour which appears anomalous. In this way it is not necessary to have access to labelled gaming data, and unsupervised techniques can be used to detect outlying behaviour.

There are two ways of approaching this task: direct anomaly detection on time series data, for example identifying step changes in a quantity or a departure from periodic behaviour, and clustering/density-based algorithms implemented on windowed data. In the latter case, some transformation of the available data will be required to extract features which describe the windowed data.

The anomaly detection checks described in this section will use many of the same methods. To avoid repetition, these are outlined here and will be referred to in the "Method" section for each check. The direct anomaly detection checks which will be carried out are listed in Table 2, and the clustering/density-based checks are listed in Table 3. In both cases, additional methods may be added depending on the performance of those listed. For the checks in Table 3, the time series data will need to be split into windows and features generated on this windowed data. A natural starting point in carrying out this procedure is to use EFA blocks as window lengths, with a step size of half an hour. Summary statistics (mean, variance, interquartile ranges, etc.) can be used as a simple way of extracting features. This process may be refined depending on performance.

Method
Violation of periodic behaviour
Significant jumps
Level shifts
Step changes in variance
Values outside interquartile ranges
Values outside thresholds
Changes of autoregressive behaviour

Table 2. Direct anomaly detection checks on time series data.

Method	Description
Isolation forest	Data points are plotted in feature space. For each data point, it is isolated from all others by recursively generating partitions. Outlying windows require fewer partitions to isolate. This algorithm scales well to large datasets.
BIRCH (balanced iterative reducing and clustering using hierarchies)	A tree data structure is constructed from the input data points and used to define clusters. Can be used for anomaly detection by setting threshold and branching factor parameters. This algorithm is optimized for performance on large data sets.
KMeans	K-Means is a relatively simple general purpose clustering algorithm. Observations are partitioned into *k* clusters, with each data point residing in the cluster with the closest mean value. If the distance between a data point and its nearest cluster centre is above some threshold, then it may be classed as an anomaly.
LOF (local outlier factor)	The distance of each data point to *k* nearest neighbours is used to estimate the local density of feature space. An outlier will have lower density than its nearest neighbours.

Table 3. Clustering/density-based anomaly detection methods.

2.2.1 Reported baseline different from expected baseline

In PM data, the baseline should be equal to the PN baseline plus any BOAs. For units with fixed baselines, differences should be flagged by ESOs performance monitoring activities and are out of scope of this project. For units with varying baselines, there will be a characteristic behaviour observed in the differences in baselines. Changes in this behaviour may be indicative of gaming.

Input(s)	Method	Output(s)
PM data PN data BOAs	For each unit, generate a time series for the difference in reported and expected baselines, and run the checks in Table 2.	All timestamps when the checks signify anomalous behaviour.

	Split the time series into windows and generate features from the windowed data. Run the checks in Table 3.	

2.2.2 High level of unavailability

Units can report going unavailable for legitimate reasons, but unavailability may also be an indicator of gaming since it can be used to hide activities from ESO monitoring. ESO will be checking for high levels of unavailability, so this is out of scope for this project. We will check for periodic unavailability.

Input(s)	**Method**	**Output(s)**
Notice of availability and outage data	Generate time series of availability for each unit and identify any periodic behaviour.	Flag if a unit is periodically reporting unavailability, and potentially the time range in which this behaviour was observed.

2.2.3 Price incentives

Unusually high prices on the wholesale electricity market may provide incentives for providers to engage in gaming behaviour. If this occurs at the same time as other potentially suspicious behaviour, the probability of gaming is higher.

Input(s)	**Method**	**Output(s)**
Wholesale market buy/sell prices	Fit an auto-regressive integrated moving average (ARIMA) model to the wholesale market price time series, and use this to identify times when the price has deviated significantly from the expected value.	Flag timestamps when the wholesale electricity price is particularly high.

2.2.4 PM versus OM active power

Gaming may show up as anomalous behaviour in the difference between metered versus reported active power. This will not necessarily show up in basic checks of the difference, and so some anomaly detection will be carried out.

Input(s)	Method	Output(s)
PM data OM data	For each unit, generate a time series for the difference in PM and OM active power. The OM active power will need to be up-sampled to match the time resolution of the PM data, or vice versa. Run the checks in Table 2. Split the time series into windows and generate features from the windowed data. Run the checks in Table 3.	All timestamps when the checks identify anomalous behaviour, along with a description of which specific check triggered the flag to be raised.

2.2.5 Reported active power versus ideal response plus baseline

The expected active power for a unit can be built up from the ideal response (given the measured grid frequency) and PNs plus BOAs. The difference between the expected and reported active power will be non-zero, and may be quite large for units with varying baselines, but gaming behaviour may be evident in patterns observed in this data.

Input(s)	Method	Output(s)
PM data PN data BOAs	For each unit, generate a time series for the difference reported and	All timestamps when the checks identify anomalous behaviour, along with a description of

Ideal response curve	expected active power. Run the checks in Table 2. Split the time series into windows and generate features from the windowed data. Run the checks in Table 3.	which specific check triggered the flag to be raised.

3 Synthetic data generation

The anomaly detection methods discussed above require training on data in order to calibrate them sufficiently to be able to distinguish suspicious behaviour from random events. For the detections to be accurate, the training data needs to include a sufficient range of normal, non-gaming scenarios as well as examples of units exhibiting suspicious behaviour. This is beyond the scope of this project, but we note that significant data will be required to improve accuracy beyond the proof-of-concept stage. For this project, we are demonstrating the tests on a small sample of available data.

Given that varying baselines were not permitted in the past, there is no historical data available for units with varying baselines. With this being the case, we will need to synthesise additional data to supplement the available historical data.

3.1 Synthesizing legitimate behaviour

For varying baseline units, we will need to synthesise varying reported baselines and reported active power. We will assume that their unavailability, expected baselines and reported frequency are the same as for existing units.

For balancing mechanism units with non-varying baselines, several years of data on physical notifications and bid/offer acceptances are available from Elexon, which provides a large dataset of expected baselines. We cannot access physical notifications for non-balancing mechanism units and do not expect to model these within the scope of this project.

To generate reported baselines for varying baseline units, we will modify existing non-varying baseline data. We will modify the reported baselines and reported active power through random events which are uncorrelated to ideal response and market prices where this lack of correlation should represent legitimate variation of reported baselines.

4.2 Synthesizing 'gaming'

In order to generate 'gaming' data, we can overlay gaming behaviour by adjusting normal data, including a unit's baseline, active power and availability. These adjustments will not be based on the historical record since the permitting of varying baselines is not yet in effect and so cannot yet have been gamed or exploited.

Therefore, we shall synthesise the gaming behaviours based upon the set of plausible actions that a malicious actor might take to game the system. This may include unique or regularly occurring events tied to the need to provide response or baseline variations which correlate with market prices or response requirement. We expect the creation of the full set of plausible actions to be a collaborative effort between the Smith Institute and ESO.

National Energy System Operator
Faraday House
Warwick Technology Park
Gallows Hill
Warwick
CV34 6DA

[insert email address]

www.neso.energy

Delivered by Smith Institute