DNV·GL

# Theft Analysis 2019/20 – Proposed Approach V1.0

## Background

The purpose of this analysis is to use detected theft records to create a set of factors that can be used to split the Balancing Factor between the 36 combinations of EUC/Product Class. This approach is based on the assumption that is made throughout the Unidentified Gas analysis: that the Balancing Factor is composed mostly of undetected theft and should therefore be split based on our best estimate of the relative incidence of such theft.

The issue with using detected theft records to split undetected theft is that detected theft patterns are not necessarily consistent with wider (undetected) theft. Theft will only be detected where it is looked for, and so detected theft rates are heavily influenced by the detection activity that each Supplier chooses to carry out. There is a lot of detected theft information available and it is a potentially useful resource for deriving information about undetected theft, but the challenge of this process is to remove the bias caused by the targeting of the Suppliers' detection activities to produce unbiased factors that will be indicative of overall theft.

The process of detecting theft has three stages, as follows:

- Lead
  Suspicious meter read pattern or tip-off. Meter reads can be identified as suspect via the Supplier's own analysis, TRAS outlier analysis, or a notification from Meter Asset Manager (MAM), Meter Reading Agent (MRA) or a Gas Transporter (GT).

- Investigation
  The Supplier decides which leads, if any, to investigate further. Each Supplier uses their own criteria and investigation rates vary widely – from investigating nothing to investigating everything. On average, around 35% of leads are investigated (2016 SPAA Theft of Gas report).

- Detection
  The proportion of investigations that lead to a detected theft again varies between Suppliers, from 0% to 40% (2016 SPAA Theft of Gas report). On average, approximately 20% of investigations result in a detection of theft.

## Principles

In order to produce objective factors that can be applied to undetected theft (and hence the Balancing Factor), the bias needs to be removed from two of the above stages:

1. The bias caused by what detection activity (or lack of it) the Supplier chooses to carry out.
2. The bias caused by the Suppliers' selection of which leads to investigate.

At this point in the development of the methodology, the relationship between investigation and detection (as recorded in the raw data) can be assumed to be constant and representative for any given EUC/Product Class category, although Investigation → Detection rates will vary from category to category. This assumption of single Investigation → Detection rates for each category may not be strictly true, because if the Suppliers' selection criteria are effective there is likely to be a

"diminishing returns" effect – i.e. the highest ranked (and therefore first) investigations are more likely to relate to a real theft, whilst the more investigations are carried out the lower the quality of the lead being followed and the lower the probability of it relating to an actual theft. The data currently available does not allow this effect to the quantified, however, and very detailed data from Suppliers covering the ranking of leads and the sequence of investigations would be required in order to do this.

Therefore the Investigation → Detection relationships are assumed to be constant. This assumption will remain valid whatever "diminishing returns" effect really does exist as long as Supplier detection behaviour remains consistent (i.e. they maintain a similar strategy over time).

As with all elements of the Unidentified Gas analysis, the undetected theft factors must be split by EUC and Product Class as described above, and in this case by Meter Type (traditional/Smart/AMR) as well. The need to split by Meter Type only creates 3 more categories because the remaining 33 all have mandatory requirements for the meter: Smart Meter or AMR for anything except Product Class 4, and AMR for EUC 04B and above. This leaves just Product Class 4 EUCs 01B-03B that can take either a Smart or traditional meter.

The general principle is therefore to base our method on existing leads and investigations (split by EUC/Product Class/Meter Type), but to adjust the numbers of these to what they would have been if the investigations had been carried out equitably based only on:

1. Population
2. Propensity to suspicious meter reads

Once these figures have been derived, category-by-category Investigation → Detection rates can be used to convert to unbiased detection numbers. The relative size of these adjusted detection figures provides the basis for undetected theft factors which can then be used to split the Balancing Factor.

**Data**

The data specified below represents a subset of the contents of the TRAS Outcome files that are provided by Suppliers. This data has been formally requested by the AUGE, and this request was approved by the SPAA Change Board on 29 November 2018.

The removal of bias from the "leads" data and the production of category-by-category Investigation → Detection rates both require record by record theft data from the TRAS Outcome files. This must include **all** leads (not just those that were investigated or led to a detection): for 2016 this would therefore consist of a total of 57,099 individual theft records (i.e. one for each Supplier "suspected incident"). The minimum information each record must contain is as follows. Items in italics are to be provided by the CDSP based on the supplied data.

- (Dummy) MPRN
  Real MPRN to be supplied to the CDSP, who will convert to dummy MPRNs consistent with other datasets.

- Meter Serial Number
  Provided to CDSP only, not to the AUGE.

- *EUC*
  *To be provided by the CDSP, queried using MPRN.*

- *Product Class*
  *To be provided by the CDSP, queried using MPRN.*

- *Meter Type (traditional/Smart/AMR)*
  *To be provided by the CDSP. This will be queried from the asset data using MPRN, and also calculated using the Meter Serial Number based rule set provided by the AUGE. If either source returns AMR or Smart, this is the assigned value.*

- *Meter installation date*
  *To be provided by the CDSP, queried using MPRN.*

- Source of lead (MAM, MRA, GT, TRAS, own analysis, tip-off)
  From TRAS Outcome files.

- Lead investigated? (Yes/No)
  From TRAS Outcome files.

- Theft detected? (Yes/No)
  From TRAS Outcome files.

- Assessed Losses
  From TRAS Outcome files.

Data will therefore initially be supplied to the CDSP, and then from the CDSP to the AUGE once it has been anonymised and the additional fields added.

**Calculation**

The first step of the analysis is to produce a set of unbiased leads, split by the 39 population categories (i.e. the original 36-way EUC/Product Class split plus 3 for meter type). Leads from the following sources can be considered to be free from Supplier bias:

- TRAS
- MRA
- Tip-off

Leads from all of these sources either come from the whole population where any given theft is equally likely to be flagged (MRA and tip-off) or are the result of dedicated analysis that is applied to the whole population without Supplier specific targeting (TRAS). Leads from other areas (own analysis, MAM, GT) *may* have inherent targeting that could skew the number of leads coming from these sources and create bias across the population categories. These are therefore discounted in the “unbiased leads” calculation.

The unbiased leads derived in this way (as a subset of the overall leads) and split by EUC/Product Class/Meter Type therefore reflects a combination of both the population of each category and the propensity to suspicious meter reads. At this point the option exists to scale the "unbiased leads" total to the overall leads total, which would result in estimates for each population category of what the number of leads would have been without any targeting. The final output from the overall detected theft analysis is a set of factors that are used to split the Balancing Factor, which therefore operate on the basis of their relative level rather than their absolute level. As such, scaling the leads in this way will not result in any tangible difference in the output, but it would nevertheless ensure that the leads total remained the same - this may aid industry parties in understanding the process. This is therefore strictly speaking an optional step, but one which will be applied for this reason.

These "unbiased leads" figures must now be converted first into "unbiased investigations" and from there into "unbiased detections".

Whilst the method for the Investigation → Detection step has already been defined in the "Principles" section above, and appropriate rates for each population category can be calculated from the record-by-record theft data, the category-by-category Lead → Investigation step has not yet been defined.

Whilst we now have unbiased figures for leads, as described above, we cannot use category-by-category Lead → Investigation rates calculated from the raw data because the decision to investigate certain leads but not others still lies with the Supplier and hence may still contain an element of targeting. If such Lead → Investigation rates were calculated from the raw data (and aggregated across Suppliers), any targeting effect would manifest itself as a deviation from uniform values: if there was no targeting, all the rates would be the same.

At this point an assumption needs to be made that the quality of leads (which can be regarded as the likelihood of any given lead meriting further investigation) does not vary between population categories: so, for example, leads from Population Category A will have a similar probability of meriting further investigation as leads from Population Category B. This is a reasonable assumption because whilst the data granularity (i.e. meter read frequency) will vary between population categories, this will affect only the speed with which a lead can be identified rather than the quality of the lead itself.

Therefore, based on this assumption, any deviation from uniformity in the Lead → Investigation rates across population categories reflects different Supplier behaviour in following up these leads. If, for example, a large domestic Supplier rarely acts on any leads but a large Supplier for small commercials follows up almost all leads, this will knock on into differences between the Lead → Investigation rates for EUCs 01B-03B (all relevant Product Classes).

With no differences in Supplier behaviour there would be a constant Lead → Investigation rate across all population categories, and therefore this needs to be the basis for the step from unbiased leads to unbiased investigations in the theft analysis. A single rate calculated as the ratio of all leads to all investigations (both aggregated across all categories and all Suppliers) should therefore be used as a constant value to convert unbiased leads to unbiased investigations.

As described in the "Principles" section above, category-by-category Investigation → Detection rates derived directly from the raw data can now be applied to the unbiased investigations figures. These rates can be calculated from the raw data without the need for further manipulation because it is

the identification of leads and the decision to investigate that are affected by the different theft regimes of different Suppliers – once an investigation is under way, its likelihood of resulting in a detection of theft is unaffected by the decision process that led to the investigation.

The result of this stage of the calculation is a set of estimated unbiased theft detections for the time period covered by the raw theft dataset, split by the 39 population categories (36 EUC/Product Class plus 3 for meter type). In the final steps of the process these must first be converted to kWh of theft rather than the number of thefts, and then projected forward to the forecast year. At this point, the dual figures for Product Class 4 EUCs 01B-03B are combined into single figures for each, and these final figures are converted to factors. This final output is used to split the Balancing Factor. This process is carried out as follows:

1. For each EUC/Product Class/Meter Type category, calculate the average kWh stolen per theft. These figures will reflect not only the higher quantities of gas consumed by larger sites, but also any effects caused by meter read frequency and data granularity affecting theft duration.
2. Multiply the number of thefts by the average kWh to give the unbiased total stolen energy from detected thefts.
3. Calculate the change in population for each population category from the theft dataset year to the forecast year, as a percentage $P_n$% for each category.
4. Scale each unbiased stolen energy figure by each $P_n$ – this is the best estimate of unbiased total stolen energy from detected thefts for the forecast year.
5. Add the individual component unbiased stolen energy figures for PC4 01B, PC4 02B and PC4 03B to give single estimates for each of these EUC/Product Class categories.
6. Convert these raw figures to proportions for each EUC/Product Class category.
7. Apply these proportions to split the Balancing Factor estimate for the forecast year.

**Smart Meter Theft Adjustment**

The above sections describe the proposed theft calculation method in full. In addition to this, the following extension to the methodology will be considered, and will be implemented if the data supports it.

The Smart Meter population is young, and existing theft work shows that there is an approximate lead time of 8 years until all thefts that are going to be detected have been detected. This timescale may be reduced for Smart Meters due to the more detailed information that comes from them but this is yet to be proven.

This phenomenon will not affect the Investigation → Detection rates for Smart Meter population categories, but it will affect the number of leads, i.e. where "young" thefts haven't yet produced enough suspicious meter readings to generate a lead, and so they will not yet be investigated and detected. Therefore an adjustment for this effect will be considered, which would use the meter installation date record for each detected theft in the training data year with the following logic applied:

- For Smart Meters up to 1 year old, only $P_1$% that will generate a lead have yet done so.
- For Smart Meters 1-2 years old, $P_2$% that will generate a lead have done so.
- ⋮
- For Smart Meters 7-8 years old, $P_8$% that will generate a lead have done so.
- For Smart Meters over 8 years old, all that will generate a lead will have done so.

where $P_8 > P_7 > \ldots > P_1$

The installation date field in the detected theft data allows the total number of Smart Meter leads from the raw data (for any given population category) to be further stratified by meter age. The factors $P_1$, $P_2$, …, $P_n$ as defined above will then be applied to these stratified figures to scale the number of (raw, untargeted) leads to what they would have been if the population was mature, i.e. it estimates the actual number of Smart Meter sites with suspicious meter reads rather than just those it has been possible to identify at this early stage. Raw Smart Meter leads from all sources should first be scaled in this manner to give a set of revised targeted leads before the processes detailed above are applied to remove bias and output the sets of unbiased leads, investigations and detections.

This calculation will only be possible if sufficient data exists to support it and allow theft detection rate curves (similar to those used in the existing Detected Theft calculation for the whole population) to be generated specifically for Smart Meters. This will be assessed on receipt of the data.

If the reasonable assumption is made that propensity to steal is steady over time, the Smart Meter population is the only one that requires this pre-adjustment. The AMR and traditional meter populations are mature and so in these cases the issue will not occur.

**Process Summary**

Figure 1 below shows a simplified graphical representation of the theft analysis process steps. The start point for the process is the dataset containing all leads, with the final output being the split of the Balancing Factor in line with proportions of unbiased theft (from each EUC/Product Class category) for the forecast year. The additional potential steps of the Smart Meter population adjustment, which will be carried out when sufficient data is available, are shown in grey.

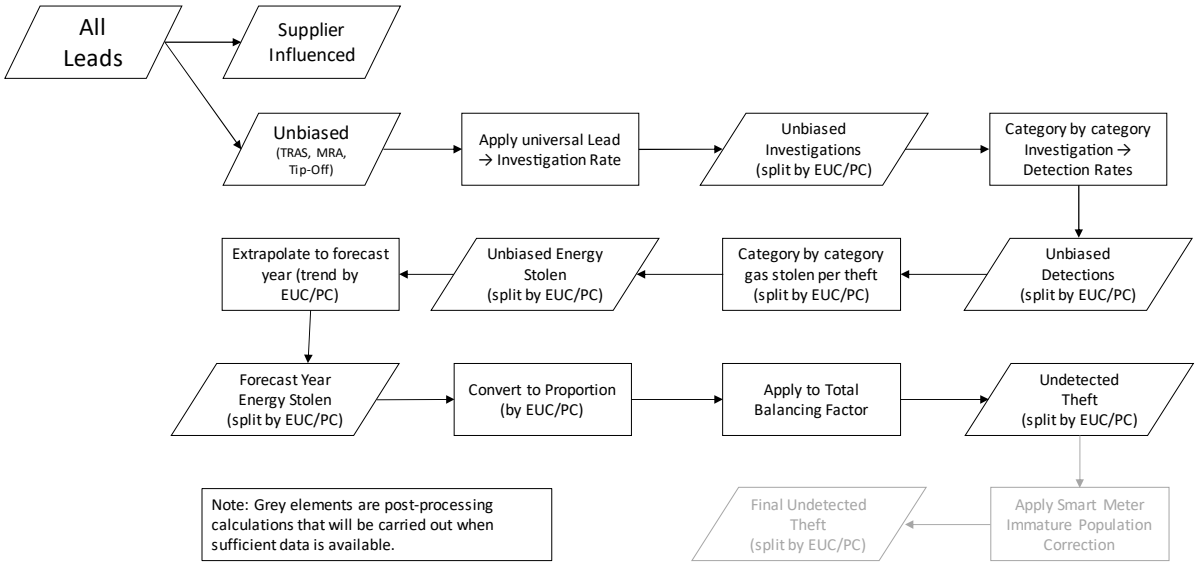


FIGURE 1: THEFT ANALYSIS PROCESS STEPS