Jacobs

Freight Modelling Think Pieces

FAME WP4 - Unlocking Data with Innovation

Think piece report

30th March 2026

FAME WP4 - Unlocking Data with Innovation work package, think piece report

Client Name:	Department for Transport
Document Title:	Freight Modelling Think Pieces, Unlocking Data with Innovation
Document No.:	2
Revision:	1
Revision Date:	30th March 2026
Project Number:	B5017300
Project Manager:	Dan Cornelius
Prepared By:	Pawel Kucharski / Tim Beech / Hayley Pawlowski
Approved By:	Dan Cornelius
File Name:	FAME WP4 - Unlocking Data with Innovation - Thinkpiece Report v2.0 - FINAL

Jacobs UK Limited.

Cottons Centre, Cottons Lane
London SE1 2QG
United Kingdom

T +44 (0)20 3980 2000

www.jacobs.com

Contents

Glossary of abbreviations

CAA	Civil Aviation Authority
CSRGT	Continuous Survey of Road Goods Vehicles
DBT	Department for Business and Trade
DC	Distribution Centre
DfT	Department for Transport
FAME	Freight Analysis and Modelling Environment
FOC	Freight Operating Company
GBFM	Great Britain Freight Model
GPS	Global Positioning System
HGV	Heavy Goods Vehicle
HMRC	His Majesty's Revenue and Customs
ID	Identification
INRIX	A trademark of Inrix, Inc.
LGV	Light Goods Vehicle
LTDS	London Travel Diary Survey
ML	Machine Learning
MND	Mobile Network Data
MSOA	Middle Super Output Area
NTEM	National Trip End Model
NTM	National Transport Model
NRTP	National Road Traffic Projections
O-D or OD	Origin-Destination
ONS	Office of National Statistics
OS	Ordnance Survey
P-C or PC	Production Consumption
RFI	Rail Freight Interchange
TABS	Track Access Billing System
VDM	Variable Demand Modelling
VOA	Valuation Office Agency
UK	United Kingdom

1. Introduction

1.1 Background

Department for Transport (DfT) currently relies on available statistics and modelling tools that inform analyses such as the National Road Traffic Projection 2022. Although this approach has been successfully applied over the years, DfT's freight model (Great Britain Freight Model – GBFMv5) has gradually shown limitations related to cross-modal analysis and freight transport chains following Brexit and the pandemic.

To continue supporting DfT's work with robust evidence base on the current and future freight patterns, DfT decided to pursue the development of the Freight Analysis and Modelling Environment (FAME). The FAME project aims to enhance the Department's analytical capability to support investment decisions, policy development, emergency responses and strategic directions, by developing a new cross-modal freight analysis environment, centred on a new National Freight Transport Model and robust scenario development that will be central to FAME.

To support this aim, a Feasibility Report[1] and the subsequent Peer Review[2] were commissioned to set out the current state of the art in freight modelling and propose the most appropriate approach to FAME. Both these reports provide insights and recommendations, but to maximise the opportunity to consider additional ideas, DfT commissioned four think pieces related to modelling of freight:

- Work Package 1: Base Year Matrices
- Work Package 2: Forecasting Methodology
- Work Package 3: LGV Methodology
- Work Package 4: Unlocking Data with Innovation.

Jacobs was appointed to deliver Work Package 4 and this report forms the deliverable for this task.

1.2 Our approach and structure of this report

Whilst the focus of WP4 is on the innovative use of data in freight modelling, we note that some considerations set out here are relevant to the areas covered by the other work packages and where such links exist, we highlight this in our report. In particular, there are synergies between Work Package 4 and Work Package 1 in relation to the development of the base year matrices, but also with Work Package 2 on forecasting methodology. Our aim was not to duplicate the scope of these work packages, but to provide context for the innovative use of data set out in this think piece.

The remainder of this document is organised as follows:

- Chapter 2 – Summarises our assessment of available data and the gaps in this data relevant to the development of FAME.
- Chapter 3 – Sets out the potential ways to address these gaps.
- Chapter 4 – Sets out conclusions and recommendations.

[1] https://assets.publishing.service.gov.uk/media/6718b881d94d2c219a5405d0/dft-freight-analysis-modelling-environment-feasability-report.pdf
[2] https://assets.publishing.service.gov.uk/media/67977f0ae0edc3fbb0606316/dft-peer-review-freight-analysis-modelling-environment.pdf

2. Current Data Gaps

2.1 Summary of existing data

DfT has access to a wide range of data related to freight. Some of the data are collected by DfT, others are third-party data and these sources are summarised in Table 2-1 based on the information provided by DfT. In the table we also summarise our understanding of the level of spatial aggregation of these datasets, as this is a critical issue for FAME.

Table 2-1: Summary of available data sources

Data source	Creator	Aggregation level
Air Freight by Type of Flight and Nationality of Operator	CAA	Spatially aggregate freight by airport
Air Freight by Airport and domestic/international	CAA	Spatially aggregate freight by airport
Port Freight Quarterly Statistics	DfT	Spatially aggregate freight by port
Port Freight Annual Statistics	DfT	Spatially aggregate freight by port
Domestic Waterborne and Port Freight Annual Statistics	DfT	Spatially aggregate freight by port
UK Port Freight Traffic Forecasts	DfT	Spatially aggregate freight by port
Rail Freight Periodic Statistics	ORR	Spatially aggregate rail freight
Rail Freight Quarterly Statistics	ORR	Spatially aggregate rail freight
Rail Freight Annual Statistics	ORR	Spatially aggregate rail freight
HMRC Trade Data	HMRC	Aggregate international trade flows by value
Road goods vehicles travelling to Europe	DfT	Spatially aggregate ro-ro data
Vacancies by industry	ONS	Aggregate economic data
Transport Statistics Great Britain	DfT	Aggregate statistics, but some underlying data will be available in a spatially disaggregate form
Non-financial business economy, UK	ONS	Aggregate economic data
Domestic Road Freight Activity Annual statistics	DfT	Spatially aggregate freight data by commodity,
International Road Freight Activity Annual Statistics	DfT	Spatially aggregate freight data by commodity.
Daily domestic transport use by mode	DfT	Spatially aggregate transport usage index
Road traffic statistics	DfT	Aggregate traffic flow based on link counts.
Road traffic estimates	DfT	Aggregate traffic flow based on link counts.
Network Rail Freight Origin-Destination (TABS)	NR	Spatially disaggregate rail flow data capable of providing partial OD (for the rail leg)
Ordnance Survey (OS) Warehouse locations	OS	Disaggregate point data
Freight vehicles over 3.5tn GPS Data	INRIX	GPS-based OD/routing data (land use attached)
Shipping Instructions - Bills of Laden	DBT	Container shipping data
Non-domestic rateable properties	VOA	Disaggregate point data
Network Rail routing information	DfT	Information on routes taken by freight trains
Global trade outlook	DBT	Global GDP, productivity and trade forecasts
UK input-output tables: product by product	ONS	Spatially aggregate economic data
UK input-output tables: industry by industry	ONS	Spatially aggregate economic data
CSRGT	DfT	OD data by commodity type carried by road.

It is clear from the summary in Table 2-1 that a significant amount of data is available for ports and airports and for the international trade flows. This type of data lends itself to:

- Developing travel demand matrices in a Production-Consumption (P-C) format due to rich information about commodities. Also this data can be assigned to point zones (and so represent trip ends in matrix building and in forecasting).

- Developing forecasting models, as it is collected regularly and lends itself to the development of econometric models where consistently collected temporal commodity data is needed.

However, such data is spatially aggregate and rarely offers information about the ultimate destinations of freight within the UK (exceptions are discussed below) and also temporally aggregate (it typically comes as annual or quarterly estimates). This limits the options for the development of reliable spatially and temporally disaggregate matrices, which in turn affects the level of detail at which FAME will be able to operate. However, three of the sources contain useful information and require special attention. These are:

- **CSRGT data**, which contains the information about end points of road freight trips and is the only source that can directly support the development of road freight matrices in the P-C format. However, its primary weakness is the relatively low sample size (DfT suggest that it covers approximately 1% of all HGV journeys). It allows expanding the sample at a relatively aggregate spatial level (such as region-to-region), but in many respects this is not detailed enough to support the development of spatially disaggregate demand matrices. The expansion process could no doubt be refined to infer additional information and attach it to CSRGT (and some suggestions were made in the peer review of the FAME specification), but many weaknesses will remain: lack of information about routes taken, low sample and the lack of information about the distribution centres and inter-modal freight flows.
- The **Network Rail TABS data**, which provides information about the end points of rail freight (origins and destinations of the rail legs) by commodity type. This is already useful in that it provides partial freight trip information and can help support the calibration of mode choice models as well as the development of matrices at a relatively aggregate spatial level (such as region-to-region or sector-to-sector flows). However, it does not provide insights into how rail freight trips interact with road freight (last mile deliveries) and it is difficult to link these trips to the consumption (C) end for demand modelling purposes. It is worth noting that this dataset could additionally be supported by Network Rail train routing information to help refine the matrices spatially and validate assignment models.
- The **GPS-based data**. The dataset currently available to DfT is provided by Inrix and the data is collected through data feeds from GPS devices. This is a commercial product where the supplier processes the data to a defined specification and typically contains:
    - Trip origin and destination (subject to appropriate treatment of privacy at the trip ends)
    - Link flows
    - Journey times
    - Type of vehicle (HGV/LGV, fuel type)
    - Additionally, land use information attached to the ends of recorded tips.

  This type of data is capable of providing the level of granularity needed for the development of a spatially detailed model. However, two major limitations associated with this specific dataset need to be addressed. These are:

    - The data is supplied to users in line with a defined specification. This means that certain types of information are pre-aggregated and the underlying data that is behind this product is not available and the users would not be able to re-analyse it to search for trends or reprocess it into a different format or level of aggregation. In other words, the data would be supplied as is, and the flexibility of exploring new methods such as Machine Learning could face some constraints.
    - The data does not offer information about the industry that the vehicles serve, which could help improve its usefulness to forecasting and matrix development. The land use information can help, but with MSOA-level aggregation, there may be some ambiguity around specific locations. This links to the previous point that after the data is pre-processed, its usefulness may be reduced.

    - Following on the above, whilst readily processed data available to DfT currently will provide link-level information, It is unclear if it can provide reliable routing information for given vehicle in a way where a vehicle ID along the full length of the journey is associated with every link along the route (without breaks such as tunnels), which is important for reliable detection of logistic chains. There may still be uncertainty over the detection of intermediate stops (distribution centres, etc) as the analyst will not have insights into what rules were used to classify stops (as either intermediate or end points). Retention of unique vehicle identifiers across all stops, and ability to re-analyse this data could be key to implementing multi-modal/multi-leg supply chain modelling.

It is worth noting that GPS-based datasets that retain most of the original information are also available on the market and could be used to remove some of these limitations.

There are also potential other sources not listed in Table 3-1. A key source to bear in mind here is Mobile Network Data (MND) where the providers of the data increasingly develop methods to identify HGV and LGV vehicle trips and offer this as a commercial product. There are also attempts to deduce the routing of the vehicles using probabilistic methods. The properties of such data will be similar to MND data used in the development of car matrices, but in terms of other characteristics of trips it is likely to capture less detail and rely on inference more heavily than some of the GPS-based sources currently available. However, it may offer additional insights in terms of detecting regular start and end points of the freight trips and help increase the LGV sample. Jacobs is currently engaged in work for Transport for London where we will undertake comparative analysis of MND and GPS sources, and which may lead to the development of methodologies that will draw on strengths of both datasets.

2.2 Identification of gaps

Some of the limitations associated with the available data sources have already been raised in Section 2.1. In this section, for clarity, we summarise them under selected themes:

- Spatial detail – most of the datasets are aggregate, and only a handful contain origin and destination data. However, even then, they have specific limitations around spatial detail.
    - CSRGT – It has a relatively low sample and so the spatial distribution of trips has to be synthesized or obtained from other sources.
    - NR TABS data – contains only partial information about the freight trips (rail leg).
    - GPS-based data – provides granular information about ODs of road freight and has the potential to remove the need to synthesize the last leg of trips, but needs to be linked with information about commodities and mapped to the underlying network (such as OS maps) in a way that provides continuous routing for each record in the sample.
- Supporting P-C format for forecasting - here it is not only necessary to have information in the right format (linking the Production and Consumption ends), but also a sufficiently detailed commodity split to support the forecasting and strategic-level mode choice modelling:
    - The various spatially aggregate sources such as the data from ports and airports can provide some information in the suitable format for at least one of the trip ends, but they typically do not contain information about the other trip end.
    - CSRGT and TABS could potentially support the P-C format but have limited sample and do not contain all required information.
    - GPS-based data can only support OD format of trips.
- Multi-modality – here it is necessary to be able to link the trips for a particular P-C commodity flows to individual legs by mode. It may not be critical for strategic P-C level mode share modelling, where it would still be appropriate to focus on the main modes. However, capturing first/last mile mode will be important (also for assignment modelling). The gaps here are:

  - Some of the spatially aggregate sources (for instance data for ports and airports) will provide information about the main mode, but are unlikely to offer insights about the last leg at the other end of the journey.
  - CSRGT is a roads-only dataset.
  - TABS is a rail-only dataset with no information about the first/last legs.
  - GPS-based data is roads only, and it is unclear if the source currently available is capable of supporting the modelling of the last/mile legs.

- Multi-leg trips – where it is necessary to link freight trips that consist of multiple segments. This is closely related to multi-modal freight flows, but specifically it relates to the distribution centres (DCs), where long haul freight (typically HGV) delivers the goods and then they are distributed over shorter distances to local destinations. The local trips can be made by HGVs or LGVs and the short distance and granularity of the local deliveries makes them particularly difficult to represent in freight modelling due to the limitations of data. Looking at the most promising datasets, the remaining limitations are:
  - CSRGT records origins and destinations, but it does not currently include information about the land use at the origin or the destination. This can be overcome with some geo-coding and fusion with other data sources (so land use could be inferred), but then the limitations associated with the sample size are likely to come into play and lead to relatively sparsely populated matrices ('lumpy' matrix problem).
  - With GPS-based data, it is unclear if the currently available source (Inrix) is capable of supporting the development of last/mile delivery legs in the modelling process (how reliable it is in detecting deliveries to/from distribution centres).
- Access to the full detail behind the data and flexibility to support the application of novel techniques. This relates only to the big data source and not the survey-based data, which under the acceptance of data licence terms, can usually be made available to the developers in the form of individual records. With big data sources they usually exist in two forms:
  - The original data before it is processed and aggregated into a product.
  - The product supplied to customers, which aggregates the data to an agreed specification. Provision of data at different levels of aggregation requires defining new specification and preparing a new data extract.

  The best example is the Mobile Network Data (MND) familiar to many model developers today. The data behind the MND matrices is vast and very granular and allows the analyst to interrogate many dimensions, analyse hidden patterns or perform advanced fusion with other data sources. But when an MND matrix is commissioned, it is pre-aggregated into origins and destinations in line with agreed dimensions (temporal, spatial and categorical). The GPS-data can be supplied as a ready product, similar to the MND matrix, but an ideal position is to have the full flexibility of adjusting these dimensions to exploit the full menu of analytical options.
- Modelling of LGVs. This point has not been covered so far in this report, but it is an important consideration for FAME given the growth in LGV traffic in recent years. The CSRGT data focuses on HGVs and so a separate LGV survey is needed. However, some insight into LGV movements at an OD level are available from the GPS-based sources and we are aware of providers who offer approximately 5% sample of LGV vehicles in the country. This is significantly above the sample sizes offered by CSRGT. The fact that this type of data does not provide any information about commodities, may be a limitation, but given that LGVs are less relevant to long-haul and more relevant to OD level mode choice modelling to/from distribution centres, this limitation may not be critical. Additionally, OD-level GPS data can offer extremely accurate information about the routing of LGV delivery vans and be an invaluable source of information about the milage driven by road type, which is one of the levels of segmentation of the National Road Traffic Projections (NRTP).

3. Addressing identified gaps

3.1 Summary of the requirements

Whilst this document is not intended to form a specification of FAME or its data flows, and many details that would be required in such a specification are likely to be missed here, we thought it would be useful to set out at the high level the desired characteristics of the freight modelling framework that are most relevant to data innovations. This will provide a clearer context for the discussion about the innovative uses of data, which we cover later in this section. The main aspects to consider are:

1) **Demand forecasting** - freight growth will need to be forecast at the "trip end" level so that differences in growth by region or entry point to the UK can be captured. Demand forecasting needs to be undertaken at the level of commodity groups, to reflect commodity-specific demand drivers and elasticities. **The available data sources at commodity level are well suited to support this step of modelling.** However, it is important to note that forecasting is likely to be undertaken at an annual level, and additional data sources are needed to convert the annual demand to daily and where necessary time period flows (see subsequent points). This means that much more temporally disaggregate data will also be needed.

2) **Mode choice at P-C level.** This should be modelled at the commodity level so that commodity characteristics are considered (value, perishability, physical characteristics) as well as transport costs at the high level. It is most likely that at this level only the main mode (or family of modes) is considered to filter out infeasible modes, and the level of spatial detail is still relatively high (sectors/regions). **The available data sources at commodity level are best suited to support this step of modelling.** However, to support the full range of analytical needs, the network-level mode choice is also needed to allocate trips to O-D legs by mode and there is a need for a two-level mode choice modelling approach. Also, similarly to the demand forecasting, mode choice at P-C level is likely to be undertaken annually, and conversion into O-D legs will need to consider daily and time period level flows.

3) **Conversion from P-C to O-D level.** Here, to go from the P-C to O-D level, some form of logistics chains sub-model is needed beneath the P-C level macro mode choice to allocate trips to ODs taking into account ports, rail terminals, distribution centres, parcel hubs, frequency of service and empty running. The distribution of trips around logistic centres is sometimes covered by gravity models, but these may lack accuracy and reliable data to calibrate them. They are also unlikely to build realistic chains and trips at the OD level may be disjointed from the P-C level. **We think that this area offers the most significant opportunity to leverage new data sources and we will focus on this later in this section.**

4) **Mode choice at O-D level.** Following the logistic chains sub-model, some adjustments to the allocation to modes on individual legs may still need to be modelled, after the detailed networks costs can be assigned to individual trips. This should not influence the macro choice (such as modes in scope for the commodity in question) and may be applicable only to certain commodities. But regardless of the ultimate format of the mode choice model, at this level of modelling, a much more spatially detailed data is needed than that offered by CSRGT, but this data still needs to be linked to commodity flows. **Again, this is an area for innovative approaches to be considered.**

5) **Assignment modelling.** This could be approached in two ways. Theoretically one could imagine a multi-modal assignment which also performs O-D level mode choice (discussed in the bullet above). However, limitations around suitable software availability are likely and different assignment requirements apply to rail and road. In case of rail, detailed network assignment beyond OD mode choice may not be necessary as feedback from congestion on the railway is not required (does not apply) and route choice options are very limited to the freight paths secured by the Freight Operating Companies (FOCs). However, in case of highway, we need to represent the impact of highway congestion and a classic highway assignment model is required. To avoid the duplication of effort, the use of National Transport Model version 5 (NTM v5) traffic assignment is likely, but reliable assignment matrices are

needed to produce reliable traffic estimates. Information about the routing will help calibrate route choice within the assignment model. It is also possible to use GPS-based routing data as the first cut source of travel costs for input into the choice modelling procedure in case FAME is developed in phases, and the early phases do not facilitate fully functioning links to NTMv5 assignment model. **This is why GPS-based data is highly beneficial.**

6) **Analysis of transport metrics**, which will be a function of all of the above modelling steps but will rely on carefully designed flow of data through the modelling system. For instance, it may not be necessary to carry our full assignments to produce output metrics from the forecasting and mode choice models. Metrics such as vehicle-kilometres, tonnes lifted, freight speeds, freight emissions, spatial distribution (by sector and road type) and many other, can be derived directly from the demand and cost data, **provided that appropriate spatial detail and metrics are available and linked to trip and in turn to the commodity data**.

We are conscious that there are also some discussions in the industry around the modelling of changes in the spatial distribution of freight (destination choice response in VDM). Considerations around the destination choice response are similar to those associated with mode choice and are not discussed here further, given the speculative nature of this element in the context of FAME.

Below we focus on the potential avenues to close some of the identified data gaps using the key processes summarised above to couch this discussion. Whilst the summary above is set out in the order of the typical modelling steps, below we focus with the most prominent opportunity to leverage the innovative use of data - the links between P-C and O-D trips within the mode choice model.

3.2 Approaches to addressing the data gaps

As discussed in the previous sections, the key gap is the lack of the link between the spatially aggregate data in P-C format and the spatially and temporally disaggregate data in O-D format. CSRGT can provide that link, but is limited in sample, and lacks some of the information included in other OD sources as discussed earlier. This is compounded by the fact that it is collected for road freight only. There is therefore a need for innovative methods to links multiple data sources.

These methods can be broadly divided into two areas:

a) Inference or imputation of information from one dataset on the basis of another dataset.

b) Fusion of datasets to create a more complete and more useful synthesized dataset.

Such methods are typically used when one dataset offers spatially and temporally disaggregate sample, but lacks detailed characteristics recorded in datasets that cover the same or similar population, but with a much lower sample size. This primarily relates to big data sources, which offer very large samples, but are unable to accurately provide the same information as survey-based data. The most common example in transport modelling is Mobile Network Data (MND) where missing information about trips is inferred from travel diary surveys (e.g. journey purpose) or imputed from ancillary sources such as gravity models (e.g. missing short-distance trips).

Alternatively, parallel datasets that cover the same domain, but are collected differently or cover different populations can be fused to provide a better estimate. For instance, using MND alongside ticket sales data to construct multi-modal personal trip matrices for transport models.

In our view, these types of methods will be applicable to freight data discussed in earlier chapters and the obvious candidates to form the core of this process are CSRGT, TABS and the GPS-based data, potentially with the additional use of counts (this list is not exclusive, and linking these sources as well as the trip end data by commodity should be expected).

Technical methodologies that can be used in these two broad areas vary and depend on the individual characteristics of the datasets in question. Examples of technical methods include:

- **Rules based methods** where the analyst identifies patterns in the datasets and uses them to infer missing information. Typically, the inference flows from survey data into big data, which

offers much larger samples, spatial and temporal disaggregation[3]. In case of freight this would involve inferring information from CSRGT and TABS into the GPS-based data so that a larger sample that carries information recorded in CSRGT and TABS can be obtained.

- **Parallel datasets**, developed independently to assess the quality and reliability of one dataset over another. This approach can be used to identify biases not revealed through other verification tests, but also provides an opportunity to draw strengths from each of the respective datasets, for instance through fusion. An example here could be deriving a synthetic dataset (matrix) from CSRGT (potentially, drawing on methodologies used to date) and comparing it with GPS-based sources or MND sources (in places where the providers offer estimates of HGV and LGV traffic). Some such parallel datasets can also be fused, and fusion techniques range from simple imputation of missing records, statistical methods, through to Machine Learning methods, and the choice of the method will depend on the properties of the data and the nature of the gaps.
- **Mathematical modelling** where the inference of the missing characteristics into the larger dataset is undertaken with the use of models calibrated from the survey data. A good candidate for this type of work are ML classifiers which allocate discrete characteristics to the big data source using a model trained on survey data. For instance, the commodity information associated with the records in CSRGT can be assigned to GPS-records, provided that sufficiently reliable explanatory variables can be found in both datasets.

These technical methos will rely on:

- Verification of the quality and reliability of the big-data source to make sure that significant biases are not imported into the process. This will involve comparative analysis of the dataset against survey data as well as other sources.
- Sufficient number of 'common fields' across the datasets to allow inference. An obvious one will be the spatial patterns (usually a geographical sector), but ideally also vehicle type, size/weight, fuel type and ideally industry sector to which the vehicle belongs.
- Existence of information to expand the big data source to the total population, which may involve some conversions and transformations of the data into a common metric (such as tonnes of freight lifted).

These technical approaches are not exclusive and a combination of them may be required. In case of freight, it is likely that the process will involve several distinct steps, and in each of these steps a different technical approach may be more applicable or preferred.

Rules-based methods, simple imputation or statistical fusion methos are more straightforward and are commonly applied in car matrix building using MND. ML methods are more complex and require careful consideration before embarking on them. This is discussed below alongside examples of how this approach could be applied to FAME.

3.3 Considerations around the use of Machine Learning

3.3.1 Typical uses of Machine Learning

Machine learning is most likely to be useful, in situations where the volume of data to analyse, infer or fuse is too large to handle using rules-based or statistical techniques. There are three areas where ML models could come into play. These are:

- **Trend detection.** This is relevant to large datasets with an array of variables, where trends may not be easy to detect using standard descriptive statistics. In this instance unsupervised learning models are used to perform **clustering analysis** – they assign records to groups that

[3] In some cases it may relate to expansion of the smaller datasets to a larger sample, but his relies on the other dataset recording total for the complete population in question. This almost never true with big data sources – they capture much larger samples, but still only a subset of the total population and also rely on their expansion to control totals.

share similar characteristics. These clusters can help segment the data and provide insights into which variables may be important for building predictive models.

For instance, we used these type of models to undertake clustering analysis of the distribution of HV traffic arriving at the Port of Dover, where trends were being detected from disparate and large data sources that included counts, port records, HGV car park records, weather forecasts, ferry timetables across the whole year and for all times of day, which then helped inform predictive models developed for the Port of Dover.

- **Classifiers.** Where ML models are used to assign data records to discrete categories. This type of model is likely to be most relevant when inferring characteristics from one dataset onto another. This is most commonly undertaken with the use of logistic regression models. In the case of FAME it would seek to infer information from surveys into GPS data.

  A case study which used classifiers to derive London-wide travel matrices was Project EDMOND for Transport for London. One of the key elements of the project was to reliably assign MND records to transport mode and journey purpose so that they can be applied in the development of the travel demand matrices for use in the multi-modal demand model for London (MoTiON). The classifier model (logit) was trained using LTDS.

- **Predictive models**. Here predictive ML models are used to forecast the dependent variable that takes continuous form, for instance traffic flow or traffic speed. The functional form of these models varies from relatively simple linear regressions to decision tree models such as Random Forest or Gradient Boosting Machines (e.g. XCBoost or LightGBM)[4].

  In the context of developing synthetic datasets, predictive models are useful for fusing different data that describe the same characteristic, but are collected differently, cover overlapping, but slightly different populations, or have different strengths, weaknesses and biases. For instance, estimation of passenger flows from Bluetooth/WiFi signals, passenger counts in certain locations and ticket sales, where none of these sources offers a reliable estimate, but synthesizing patronage from all of them could return more accurate estimates.

3.3.2 Areas in FAME development where ML could be considered

In the earlier section we summarised three types of analysis where ML models can be deployed. Here we consider which areas of FAME these could relevant to. The discussion is then followed by our assessment of what would be most appropriate for FAME and what steps would be required to deploy these ML models (Table 3-1).

Potential areas of FAME where ML could be considered include:

- **Trend analysis.** Here one of the key activities would be to establish the characteristics of vehicle trips and identify logistic chains associated with a particular commodity type. Taking a purely hypothetical example - do heavy trucks in the food industry have a long road chain before shorter lighter chains? If the analysis identifies clear characteristics of vehicle flows associated with these commodities, it can be used to inform the training of the classifier model to assign these characteristics to vehicles in the larger GPS-based dataset. To assist with this, the ML method that could be deployed here is clustering analysis which can detect such relationships. There are two considerations here:

  - The primary dataset that would be analysed in this step is CSRGT, and the size of it may not warrant the application of clustering algorithms. It may be sufficient to

[4] There are also Deep Learning Models using neural networks, which are the most powerful models often used to handle the most complex data such as image recognition, text and audio analysis. The scale of data that justifies the use of these models is unlikely to be relevant in the development of FAME. Also given that these types of models are the least transparent to users, we do not believe they are in scope for consideration for FAME.

undertake standard analysis of the dataset informed by industry knowledge of logistics chains by commodity to identify and summarise key relationships.

- That said, this step is also likely to cover the TABS data. This could also be done manually and does not in itself require clustering analysis. But the need to discover patterns across both CSRGT and TABS adds additional dimensions and with each dataset added to the mix, the use of ML to identifies patterns and group data into clusters is more and more justified.

The decision about the applicability of clustering analysis would need to be made after the detailed inspection of the relevant datasets and understanding of their format, data structure and the size of the datasets. This is discussed further at the end of this section.

- **Classification of the GPS-data records.** After the identification of trends, the inference of information onto larger sample GPS data will be needed. Let's take the example of commodity information, which will need to be assigned to the vehicle trip records in the GPS data to link the PC-level forecasting by commodity group and the OD-level modelling. This could be done using an ML classifier that uses explanatory variables associated with this commodity group identified in the survey data. The same variables need to be identifiable in both datasets, and we expect that additional information will need to be appended to both datasets – for instance land use data. The specific steps required here are:

  - The classification would commence with hypotheses derived from the trend analysis. Before hypothesis testing starts, a training and validation datasets would need to be created. It will require additional information to be appended to the survey records – in particular about the land use. Creating this dataset will require rules-based methods to allocate records from the survey to land use in a way that matches ODs, commodity type and industry sector. Given that CSRGT is a road vehicle survey, certain trips would have to be linked to rail interchanges (ODs recorded in the TABS data).

  - The classification of GPS records to different commodity types, best performed using ML classifier models, with the best candidate being logistic regression.

  The output of this process would be a larger dataset suitable for the creation of P-C and O-D matrices, and the development of mode choice models. To create the matrices, the dataset would still need to be expanded to totals implied by freight statistics at ports and airports, and this would be enabled by linking the data to global freight statistics.

- **Fusion and forecasting.** This is another area where ML models could be considered and the approach would differ from the steps described above in that instead of detecting trends, and classifying big data into categories, a single dataset could be built with the use of ML fusion. An attempt could be made to use all available data (trip records from GPS, survey data, TABS, land use, ports and airports) to directly predict trips by mode for each PC and/or OD pair. This is akin to creating a digital twin of the multi-modal freight flows across the country. Creation of such digital twin would make sense if it could also directly support forecasting activities. Two separate areas could be considered here:

  - Mode choice prediction – instead of a traditional mode choice model based on logit, ML from the decision tree family could be considered. It would ingest a range of different data sources to return prediction of freight flows by mode. Such models have a tendency to overfit the prediction to the base data and their predictive power over long term horizons is limited. Another key limitation is that they are less transparent about the variables that drive the prediction and may be challenging to explain to stakeholders and policy customers.

- Demand forecasting – predictive ML could be used to forecast total demand (based on time series data) or predict freight trips rates (using cross-sectional data if sufficient amount of freight data by land use type can be identified). Such ML model would most likely collapse to more traditional regression models. Without knowing the considerations around demand forecasting covered by Work Package 2, we believe that this application of ML is one of the most likely in FAME.

Admittedly, whilst we have experience with undertaking clustering analysis and applying classifier models, we did not undertake any feasibility testing of using fusion to build the single dataset. But we envisage some challenges associated with it. The primary challenge is that such models focus on predicting a single dependent variable (say traffic flows) at a given point in time, based on explanatory variables directly linked to it. Freight datasets describe many dimensions (spatial and temporal): some are annual flows; some represent time stamps by second, some represent aggregated total tonnage, while others could provide detailed vehicle movements. In other words, the datasets do not lend themselves to developing digital twins in the same way as real-time traffic data does. The model would be complex and the causal relationships may not be transparent enough to users of FAME. The effort spent on developing freight-specific digital twin may not provide a good foundation for a long-term forecasting model.

Based on the considerations set out above, below we focus on how we think these ML models could be applied to FAME, as well as where we do not think they are applicable. We also summarise the key risks, opportunities and limitations associated with them (Table 3-1).

Table 3-1: Areas where ML should be applied in FAME

FAME Area / ML type	**Recommendations**	**Risks, Opportunities & Limitations**
Trend analysis	The development of the single FAME dataset should commence with: 1. Preparation of data sources for trend analysis. It will involve processing, cleaning and reformatting of data in such a way that the clustering algorithms can ingest it and use it to detect trends. 2. Application of the clustering algorithms that will assign records into groups with similar characteristics. This will provide information about the most appropriate segmentation of the models, insights into logistic chains and initial hypotheses for testing in the classifier model.	This is area is promising . However, it is possible that: • Data may be too “messy” to allow trend analysis. • Detected trends are implausible. This should not be the reason for not pursuing trend analysis, but means that using rules-based approach supported with the knowledge of the logistic chains to derive hypotheses will need to me kept as a fall-back option.
Classification	After the trends analysis FAME should assign characteristics of freight flows identified in trends analysis to the vehicle trip records from the GPS data. This would formally test the hypotheses established earlier and develop the classifier model that links GPS records with these hypothesis. This would effectively lead to the creation of a single, multi-modal dataset for use in FAME.	We expect that the use of ML here will be strictly needed as it is the most likely route to constructing a high sample, single dataset for use in FAME. Such approaches have been implemented successfully before, but like with any model estimation, there is still a risk that a statistically significant classifier cannot be developed. Again, the mitigation would be to fall back on the rules-based approach.
Fusion & forecasting	In our view, FAME should not pursue the application of ML in this area at this stage. Whilst ML lends itself to the development of	This is the least likely area of success for FAME as it has a number of major limitations.

	digital twins that describe the current state of the system, application of ML in forecasting has not yet been proven and there are major limitations associated with it. Pattern change analyses will enable testing of what-if scenarios. The single dataset, if long enough GPS data series is available can support it and it does not require the development of the forecasting or digital twin models.	• Fusion would work best if the different data sources describe the same metric. • Applicability of ML would be constrained to forecasting short-term effects, rather than long term strategic policy modelling. • ML-based forecasting models run a risk of being overfitted to observed data and may not be able to respond to future demand drivers.

Table 3-1 provides a succinct summary of where in our view ML should be pursued in the development of FAME. It is worth clarifying at this point:

- **Data preparation requirements** for the application of trend analysis - It is not possible to just "throw all data at ML" and expect that it will return the information about the trends. Whilst trend analysis is performed with clustering algorithms that use unstructured learning about the trends (no trends are specified upfront by the model operator), it will require data to be structured appropriately for input into the model for the trend analysis to be successful. This still requires all data to be reviewed and assessed by a skilled ML analyst who understand the workings of the clustering algorithm and therefore knows if data is of suitable quality for input. This would cover review of gaps and completeness of the data, units used and metrics described, spatial and temporal dimensions and availability of "common" fields that the clustering algorithm can refer to (such as dates, zones, coordinates, land use definitions etc). This activity is necessary to determine if clustering analysis is likely to be successful, but also to enable it.

  However, there is still a risk that it will not detect statistically reliable trends. Here, the initial review of the data by a skilled analyst does help, and provides an opportunity to have a fall-back option of using the rules-based approach to formulating the initial hypothesis. This is because a skilled analyst with the knowledge of the logistics industry will be able to identify key trends on the basis of descriptive statistics. Whilst some more nuanced trends may go unnoticed and the use of ML is highly desirable, the in-depth understanding of the data will provide options for alternative approaches and de-risks the development of FAME.

- **Pattern-detection analysis** as an enabler of "what-if" analysis. In Table 3-1 we say that ML is not suitable for long-term forecasting, but may have some applications in short-term forecasting. However, significant resource and time are still required to develop reliable digital twins that could support such limited use cases. "What-if" analysis could already be supported with the creation of a single dataset that is highly disaggregate spatially and temporally. For instance, the effects of roads closures will be visible from GPS records (before and after analysis). Impacts of port disruption can also be identified from past data (in terms of impact on congestion and routing) and the single dataset will provide insights which industries and commodities affected. The effects of COVID would be visible not only from the GPS data, but also from the aggregate statistics and the changes in freight patterns that followed the event can be established.

  Based on this, it is possible to postulate that a similar event in the future would have a similar effect. The cause and effect would not be formally established and so this could not constitute formal forecasting and should be treated as "what-if" scenario analysis but this could already provide a powerful tool and be enabled by the single dataset. Also, it is worth noting that digital twins would also carry similar limitations. It is not possible to forecast the effect of events that themselves cannot be forecast and longer data series rather than one-off events are needed to establish causal relationships between these events and the outcomes.

Based on these considerations, in Section 3.4 we present a potential approach to the development of FAME datasets.

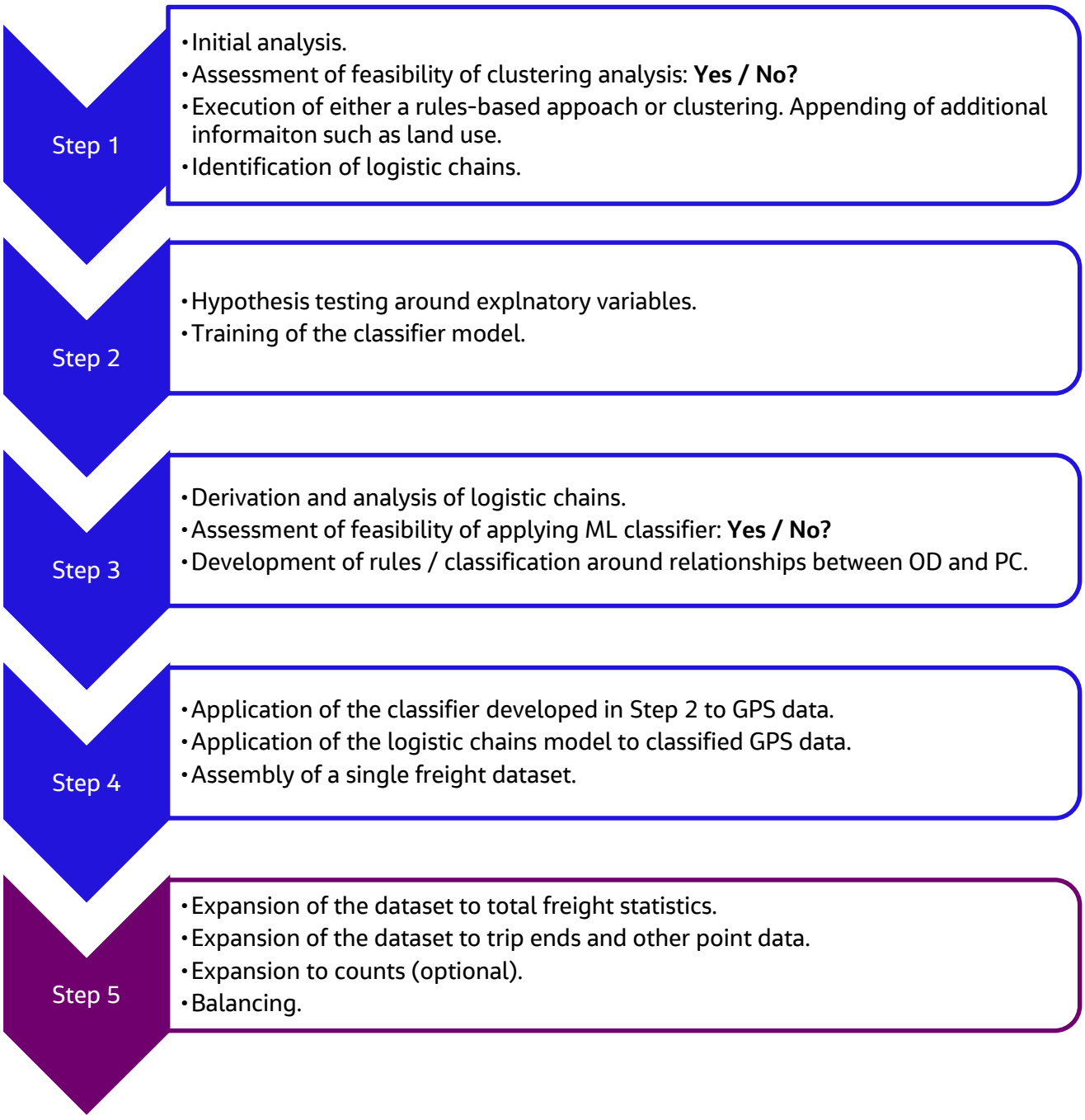
3.4 Possible approach to FAME development

In this section we postulate that the most likely approach would be the use of classifier models to assign survey characteristics to 'big data' and significantly expand the sample of data available to FAME. Also, given the complexity of the interactions between modes and land-uses in FAME, we believe that a step-wise approach that uses a mix of traditional data analyses, rules-based approaches and targeted use of ML is most likely to improve on the current freight modelling methods, offer transparency and provide evidence base for FAME that can be assured. The approach to data in FAME would consist of the following steps:

- Step 1 - Analysis and processing of the datasets.
- Step 2 - Hypothesis testing and classifier model development.
- Step 3 – Development of the logistic chains model.
- Step 4 – Application of the models to data and validation.
- Step 5 – Expansion of the dataset to a complete multi-modal matrix of freight flows.

Steps 1-4 would lead to the development of a single database, from which specific components of the model could be developed further, for instance, a multi-modal demand matrix or the dataset for the development of the P-C and O-D level choice models. In case of base matrices, this could be integrated with the process of creating a single dataset (Step 5), subject to recommendations from Work Package 1. The potential process that takes into account decisions about the use of ML is illustrated in Figure 3-1.

Figure 3-1: Potential process to develop the FAME dataset



Each of these steps is described in more detail below.

Step 1 - Analysis and processing of the datasets

This step will involve the analysis of multiple data sources discussed in Chapter 2 that have different formats and characteristics and which need to be understood by the analyst. An initial step will therefore be required to structure the comparisons. The datasets are disparate and have different structures and an obvious question here will be if unsupervised ML can handle these data differences successfully. The setup of data for input into the clustering analysis will still require a significant amount of initial 'manual' review of data, cleaning as well as reformatting of the datasets. It will also require establishing if the data has enough in common (common fields such as temporal or spatial dimensions) for the clustering algorithm to be able to successfully 'link' the datasets to identify trends.

We envisage that to enable this step, additional information will need to be imputed into the datasets (i.e. land use at origins and destinations of trips). This will likely require some level of rules-based inference or imputation of data first. For instance, as a minimum it will be necessary to infer the land use information at the trip ends associated with the CSRGT, GPS data, and where possible TABS. Considerable thought will be needed around the inclusion of other data. The purpose of this will be to allow a more meaningful detection of trends within the freight flows. The output of this step will be:

- Descriptive statistics that explain the spatial and temporal characteristics of the data as well as any other segmentation available.

- Training dataset that will be used to develop classifier models later and enable hypothesis testing around the explanatory variables that drive the characteristics of freight trips. A test dataset will also be required (likely to be created by dividing the data in the training and test proportions) to validate the model on the data that the model did not see in training. This will require identifying 'common fields' between the training dataset, test dataset and the dataset to which the classifier model will be applied.

- Understanding of the 'behaviours' of freight flows associated with specific commodities, commodity groups or sub-groups as well as their inter-modal interactions, i.e. heavy trucks in the food industry have a long road chain that then link to shorter chains performed by lighter vehicles. The formulation of these links will be key to the development of the logistic chains model necessary for developing links between P-C and O-D level components of FAME.

In summary, this will be one of the most important steps in the data-related activities and facilitate all the subsequent steps. It will also de-risk the project, if the application of clustering analysis is not successful – this is because the manual analysis of data will already provide some understanding of patterns and provide a fall-back option that mitigates risks around the performance of the clustering analysis.

This step should also be combined with verification tests of the GPS data. The key insight from this will be the degree to which the data sources agree, identification of biases and the preferred way of correcting them. The verification process will bear a lot of similarities to the initial steps of the MND matrix building process for car trips and is therefore a necessary first step before any modelling starts, and it will require a hands-on engagement of the analyst with the data.

Step 2 - Hypothesis testing and classifier model development

This process will be familiar from regression modelling where the functional form of models and explanatory variables need to be established first. The process will start with plausible hypotheses that emerge from Step 1 – such as the hypothesis around the variables linked with the characteristics of the freight trips (following the example of commodity splits, a set of hypothesis tests would focus on the variables that can predict which commodity of a particular freight trip recorded in big data). The tests would focus on statistical significance of the variable and consider if the variable was directly observed, imputed or inferred. Two possible outputs from this exercise are possible:

- A set of candidate models for implementation and deployment on the actual dataset to test their performance using the final verification tests described below, or

- A confirmation that no reliable model can be estimated (no statistically significant relationships can be found). In this case the method would revert to a fall-back approach that could include:

    - A rules-based approach, which makes plausible assumptions about the links between the datasets and assess the overall accuracy of the approach through more traditional validation steps in model development (accuracy of freight traffic assignment, mode shares by region, trip length distributions etc). Much of this would be informed by analysis of trends in Step 1.

- Consider if the reverse approach of expanding the survey with big data is possible technically and can offer a way forward.

- Consider a more classic approach (similar to MND car matrices) where big data is categorised using information derived from survey at sector level (such as trip length distributions and commodity splits) and controlling to trip ends derived from land use data, and point data such as ports and airports. Similarly to the MND-based car matrix approach where irrelevant modes are removed from the MND dataset, this process could isolate the trips associated with RFIs and then subsequently link them to the rail freight trips recorded in TABS. Also, this approach should still allow distinguishing between end destinations and DCs to create logistic chains as the vehicle trips in GPS data are linked by unique IDs.

The output from this step would be a larger dataset, based primarily on GPS trip records, but including other information about the freight trips inferred from other datasets, such as the commodity information from CSRGT or multi-modal information derived with the help of the TABS data.

Step 4 – Development of the logistic chains sub-model

Whilst there are case studies that used the methods described earlier, the development of the logistics chains model with the use of big data will be a step unique to FAME. We envisage that the single unified dataset would readily provide information about commodity type, industry sector, length of haul vehicle type including type, size, fuel and axles (some GPS datasets provide information about the number of vehicle axles), origins and destinations, day of travel, time of day, as well as the land use and therefore a probability that the origins and destinations used DCs or RFIs.

Furthermore, the GPS-based data includes unique identifiers of vehicles across the full length of the route and time stamps and allows linking the individual trips in logistic chains. All this information will allow the detection of a long-haul trip visiting multiple sites or a direct trip (for instance between ports and DCs), as well as short-haul trips where round trips by LGVs are more likely. Again similarly to the earlier stages the logistic chains could be developed using:

- Trend analysis in Step 1. Or if this fails, or

- A rules-based approach which will use insights from Step 1 as well as practical knowledge of the logistics industry operations. The way such model could operate in forecasting mode that it would divide freight flows forecast at the P-C level into logistic chains categories and modify proportions within each logistic chain category to reflect changes in freight flows forecast at the level above, or

- Depending on the structure and the variables determined in the classifier, form of logit model could be used to further split records that already had commodity allocated to them into OD-level mode/leg combinations. This way there is a possibility that a nested logit structure could be developed within the choice models. However, the feasibility of this approach much depends on the variables selected through hypothesis testing, the number of logistic chain categories that need to be forecast and the feasibility of using ML classifiers.

The output of this stage would provide logistic chains information that would allow calculating more accurate travel costs based on the detailed understanding of individual legs and modes used as well as accurate estimates of the distance travelled. This would then be capable of supporting the OD based mode choice modelling, but will link to the P-C level forecasting and mode choice.

Step 4 – Application of the model to data and validation.

After the classification of the GPS and TABS data into the desired commodity groups and logistic chains, a second round of verification tests will need to be undertaken on the final dataset in a similar

way to the final verification tests typically performed on prior car matrices based on MND data. However, it is important to note that in case of freight there will be more limitations associated with independent secondary data sources to undertake certain verification tests. For instance:

- In case of MND matrices, trip ends can be verified independently against sources such as NTEM or local estimates of trip rates and land use data. In case of freight, the land use information will likely be directly appended to both the GPS data and CSRGT and will not be truly independent. Similarly, attractions to distribution centres will be second order derivatives from land use data and the trip record data.

- Truly observed trip length distributions and average trip lengths can only be obtained with CSRGT and this data will be used to train the models. No other independent data will therefore remain to additionally verify the final dataset.

The only true validation of the expanded freight dataset would be the collection of road and rail based routed markers which are independent of either of the datasets used in the process to compare against the expected routes through these points from the expanded data. For instance, this could be undertaken at the screenlines or cordons at strategic sites etc. But this type of validation is only relevant to the verification of data which contains the routing information and is only possible if GPS-based data is used to create the single dataset.

Step 5 – Expansion of the single dataset to a complete multi-modal matrix of freight flows

The creation of the database described above would directly facilitate a number of FAME processes. It would enable:

- The development of the demand model matrices. Assuming that the demand model would be applied incrementally.

- The development of the assignment model matrices. This process will require dividing the data into time periods (including the allocation of the freight trips, some of which are long-distance) into appropriate time periods in the assignment model).

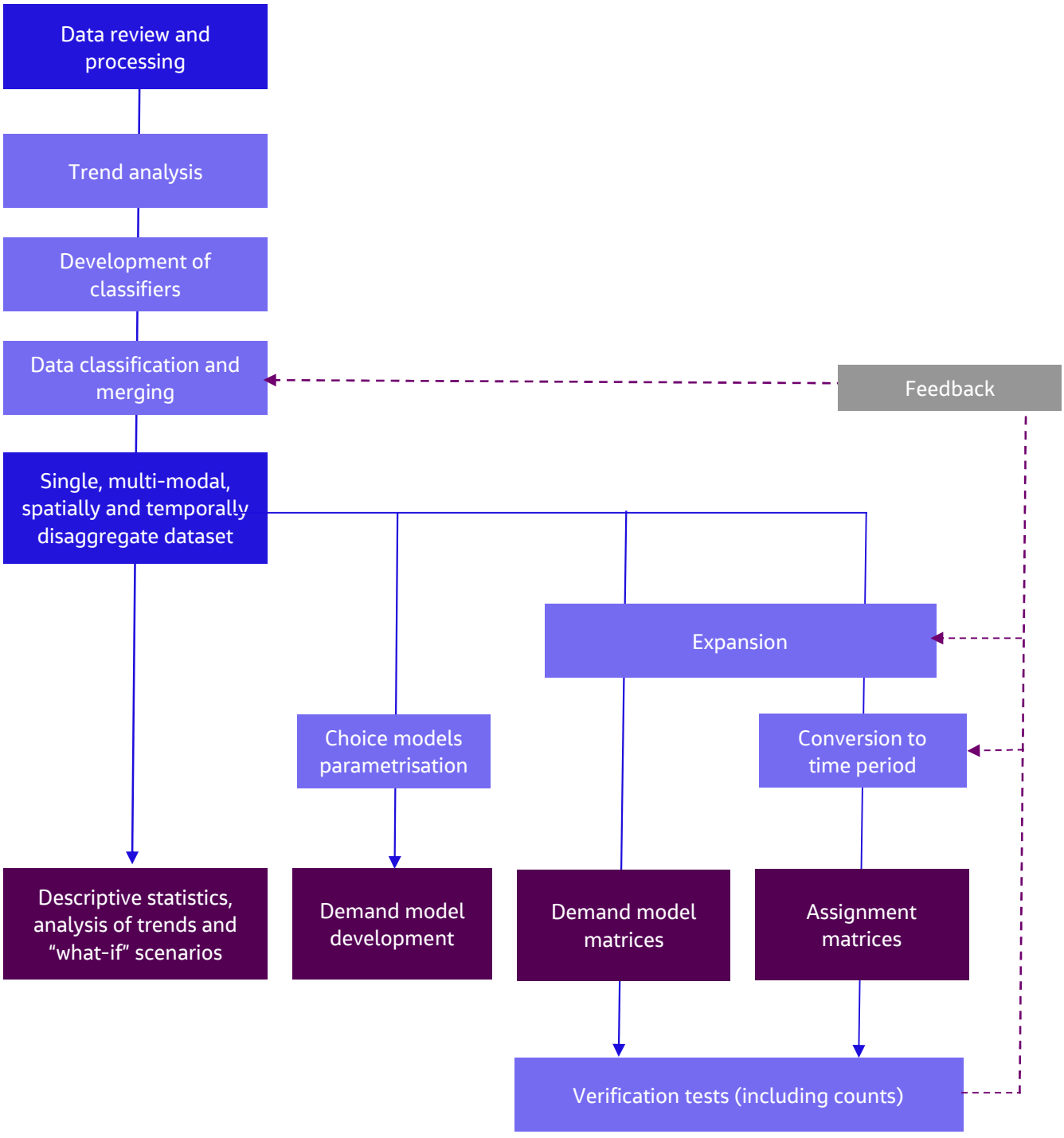
- The training of the mode choice models to estimated their parameters.

- Ad hoc analyses that the not require full model application.

Without prejudging the findings of Work Package 1 on the base year matrices, there is clearly a link between the data flows described in this think piece and the derivation of the matrices. These flows could be automated by appending data expansion to the steps described in this think piece. This could use total freight statistics, as well as freight vehicle counts on the road.

Count data could be used as an additional source in the process of expanding the dataset, but there would likely be enough counts available from DfT data as well as National Highways data to allow some independent verification at the end of the process. The final set of verifications would then need to follow in the assignment model (again similarly to MND matrix development) using NTMv2 assignment. And example of the data flows through FAME that a single dataset could facilitate is presented in Figure 3-2 below.

Figure 3-2: Potential FAME data flow



4. Conclusions

4.1 Summary

In this think piece we set out innovative uses of data that can enable the development of a spatially detailed multi-modal modelling freight framework such as FAME. We reviewed summaries of available data provided to the study team by DfT to identify the key gaps that currently limit the development of FAME.

We then proceeded with the discussion on the potential ways of closing these gaps, focussing on the key modelling steps that a comprehensive freight modelling framework would likely need to follow, as well as the different types of ML models relevant to these steps. We then set out our recommendation for which of these ML uses should be taken forward in the development of FAME and what the associated risks, opportunities and limitations are.

Following this discussion we set out an example of the potential step-wise approach to the development of a single, consistent, spatially detailed multi-modal dataset that could be used in FAME development. The recommendations for the next steps are provided below.

4.2 Recommendations

Based on the potential ways forward set out in this think piece report, we believe that the development of FAME will need:

- **A revised or supplementary scoping document focussed on data** that aims at a more comprehensive freight modelling framework that addresses the concerns of the peer review, but also sets out in more detail the flow of data through FAME to inform individual steps. The aim of this document would be to make sure that all areas covered by the FAME Think Pieces (WP1 – WP4) are aligned and provide a feasible pathway for the development. Preparation of such a scoping document would provide an opportunity for an internal challenge within the scoping team around the linkages between data, matrix, demand model and forecasting and provide an opportunity to iron out inconsistencies.
- **Detailed review of all data sources** to confirm if the approach envisaged in this think piece is feasible, and allow the development of the detailed specification of the data work plan. Such specification should include a series of fall-back options – alternative plans that will allow FAME development to move forward if some of the ML methods discussed in this report turn out to be infeasible. We already offer a few ideas for a fall-back approach in this document, but the full set of options could only be formulated upon the review of all available data sources.
- **Identification of additional data sources** beyond those currently available to DfT. For instance, at the time of writing this report, both TfL and National Highways were commissioning studies to assess the viability of using MND data to model freight trips. Jacobs has been commissioned to undertake the TfL study and the scope will include leveraging both MND and new sources of GPS data[5], which offers high sample (10% of HGVs, 5% of LGVs nationally) and more information about the vehicles and the industry sectors.

The findings of this review should then be used to update the full specification of FAME. We assume that DfT would be looking to undertake the development of FAME in phases and so the approach recommended above lends itself to supporting phased delivery.

[5] Jacobs Freight Solutions data uploaded onto OMNIFlow – Jacobs' agentic AI data analysis environment that enables querying large datasets with the use of natural language, based on Large Language Models (LLM).