MOTT MACDONALD

Transport data for artificial intelligence innovation

Discovering, prioritising and analysing high-potential data sets

March 2026

Contents

1 Introduction

Data is foundational to today's transport system. It is also crucial to the Department for Transport's (DfT) Transport Artificial Intelligence (AI) Action Plan vision for a "resilient transport system delivering cheaper, cleaner and safer journeys for all" and delivery of its Transport Data Strategy.

This report presents a structured assessment of existing and emerging data sets in the transport sector. Through a three-stage methodology, it identifies where improved access, quality and coordination of data could unlock high-value AI use cases and drive innovation aligned with strategic policy goals.

1.1 What is the ambition of the Transport AI Action Plan?

Following on from the UK Government's National AI Strategy, published in 2021, the DfT's Transport Data Strategy focused on "laying the foundations" around five themes:

- Sharing, discovery and access
- Data standards and quality
- Skills, culture and leadership
- User needs and communication
- Governance, protection and ethics

These provide strategic direction but do not go into detail about the applications of data, such as in AI. DfT subsequently published the Transport AI Action Plan in June 2025 which sets out how DfT will work with the transport sector to exploit the opportunities while managing potential risks. It highlights that large data sets are key to enabling AI opportunities that will unlock transport benefits.

This project was commissioned by the DfT to support the implementation of the Transport AI Action Plan, the development of the Transport Data Action Plan and related workstreams, such as digital twins and advanced analytics. It aims to identify and prioritise data sets with potential to support development of novel transport applications using AI and undertake a review of key data sets that were surfaced.

1.2 What this report sets out to achieve

The project set out to understand:

What data is out there? | Explore the breadth of data that allows the exploration of potential AI innovations in transport.

How can the data be used? | Identify potential transport use cases.

Is the data accessible? | Identify the barriers to access and engage with data holders to identify priority data sources and routes to increasing access for benefits realisation.

Focus areas

AI use cases in transport

Unfamiliar, closed data sets

This was achieved through three stages**, the** aims, methodology and findings of which are presented below. The methodology includes references to sections in the Appendix where further details can be found. Sections 2-4 provide a more detailed overview of the key findings.

Stage 1: Discovery and compilation

Identify whether there are data sets that could support valuable AI use cases in the transport sector.

Desktop review of online sources to identify data sets.

Framework Development (A.1.2) to record metadata for discovered data sets. This includes type of access, updated frequency, temporal resolution, etc.

Expert workshop and interviews to identify potential data sets, explore relevant use cases and gather feedback on the proposed framework.

Output: longlist of **156 data sets** across: Connected Vehicle and Sensors; Demographic Movement and Behaviours; Earth Observation and Environmental; Energy Generation, Transmission and Emissions; Freight; Transport Network; Transport Operations; and Other.

Stage 2: Filtering and prioritising

Prioritise a shortlist identified by a set of criteria and recommend data sets suitable for further analysis.

Desktop prioritisation of longlist (A.1.1) using criteria around strategic alignment, AI suitability and access (A.2.1).

Expert workshop and interviews to test the shortlisting methodology, review findings, explore current data use and partnerships, identify data gaps and discuss how DfT could support improved access.

Output: shortlist of 34 consolidated into **18 data sets** across: Connected Vehicle and Sensors; Demographic Movement and Behaviours; Earth Observation and Environmental; Freight; Transport Network; and Transport Operations.

Stage 3: Analysis and insight

Further prioritise “high-potential” data sets, ie data that may not yet be available but expected to have high potential to enable AI innovation in transport.

Developed challenge statements (A.3.1) to further narrow down the list.

Data sets were matched to challenge statements and ranked by the number of associations (depth) and dissimilarity (depth) to identify five data sets offering the best balance between both.

Output: prioritised list of **5 data sets**. This included: last-mile courier and haulier operational data; mobile phone operating system and services providers; CCTV highway-based video analytics; Urban Traffic Control (UTC) and connected infrastructure sensor and configuration data; and connected vehicle data.

2 Discovery and compilation

2.1 Available but not accessible or not available but potentially transformative?

In identifying where AI innovation could be accelerated, it was important to consider data that is on both ends of the spectrum of accessibility and maturity, from available and open data to data that does not yet exist but could have transformational impact.

- **Available data**: Data that exists and can be accessed but may need additional structuring, promotion and support, for example, data available in the Rail Data Marketplace.
- **Data that exists but is not open**: This could be challenging to extract, in private ownership or facing procurement barriers, for example, signal timing data that has technical and procurement barriers to access.
- **Data that is hypothesised to exist**: Data that is likely to exist but not confirmed, for example, data that may be collected and stored by hauliers.
- **Data that does not exist yet but could have transformational impact**: Data that may not be collected currently, for example, detailed data on pedestrian behaviour or supply chain logistics.

2.2 Compiling the longlist

The initial longlist was composed of 156 data sets. Eight themes were defined to group the material, with each theme illustrated with examples of the associated data. The full list is in A.1.1.

Connected vehicle and sensors	Data from vehicles, including inbuilt and external sensors, providing journey time, speed, location and acceleration, as well as in-vehicle CCTV.	– **In-vehicle CCTV:** Continuous feeds from public transport vehicles. – **Smart cycle sensors:** Bike rider experience, for example, swerving and braking.
Demographic movement and behaviours	Person-level population data, including behaviour and travel.	– **Mobile network data:** Crowd-based insights to understand mobility and footfall. – **Transaction data: D**ebit, credit and prepaid payment data.
Earth observation and environmental	Macroscopic data sets, including different types of satellite and geographical data.	– **High-resolution nightlight remotely sensed imagery with 1m resolution.** – **Met Office Climate Data Portal:** Climate statistics focused on UK.
Energy generation, transmission and emissions	Information about energy supply and infrastructure with implications for how transport is powered.	– **EV charge points data.** – **Energy data:** Data from large electricity network operators.

Freight	Freight movement by land, sea and air, including movement of vehicles and goods.	– **Courier last-mile data,** for example, data including route, stops, package levels. – **Freight data:** Container shipping, ship data, world cargo database.
Transport network	Network-scale data about all parts of the network and its components.	– **Transport API:** Aggregator of transport open data. – **Rail Data Marketplace:** Central platform for finding and sharing rail data sets.
Transport operations	Data about network performance, disruptions and vehicle movements.	– **Bus location data:** Temporal spatial information about bus locations, via Service Interface for Real-time Information Vehicle Monitoring (SIRI-VM). – **TfWM ADEPT Live Labs**: Capture of existing data for new uses.
Other	Data falling outside the first seven themes.	– **GDELT:** Global monitor of broadcast, print and web news. – **Business data**: Track high growth companies and emerging sectors.

2.3 Key findings

The insights below summarise key findings that emerged from the desktop study, workshop and interviews:

- **Not all data of interest is big data:** It is assumed that big data is most appropriate for AI applications. However, smaller data sets may still have potentially innovative applications when combined or analysed using AI/machine learning (ML) technology.
- **Openness of data is a critical issue:** Many valuable data sets, such as Signal, Phase and Timing Extended Message (SPATEM), Map Extended Message (MAPEM), Controller Area Network Bus (CANBus) or mobile phone and freight (especially rail) data, are not publicly accessible due to technical, legal or procurement barriers. This limits their potential for innovation.
- **Ownership and access ambiguity:** For certain data sets, such as train loading data, there is uncertainty around ownership, sensitivity and technical complexity, which complicates their use.
- **Standardisation challenges hinder data use:** Public sector data often lacks standardisation, making interoperability difficult. This issue is compounded by procurement processes that do not prioritise data accessibility.
- **Quality of data is vital:** Poor or uncertain data quality can hinder rather than help AI applications. Knowing whether data is accurate is as important as the data itself.
- **Private sector data presents the greatest opportunity:** While public data access is improving, the most significant untapped potential lies in private sector data, which is often closed but rich in value. Examples include data captured by technology companies, such as Google, Apple and other major mobile mapping and journey planning suppliers, or data on last-mile delivery routing for couriers.

- **Relationships are essential for unlocking data:** Building strong connections with private companies and within the public sector is key to accessing and utilising data without relying solely on regulation.
- **Local authority data is often not exploited: T**his includes CCTV processing, variable message signs (VMS) and other data sources, with barriers to access including data being stored in a way that is not readily accessible and commercial agreements with suppliers.
- **Areas for further exploration have been identified:** Earth Observation, Urban Traffic Control, Freight, Connected Vehicles and Rail are found to be themes for future data-driven innovation, both in terms of existing data with untapped potential and/or new or growing data sets.

3 Filtering and prioritising

3.1 From longlist to shortlist

Desktop prioritisation used the following criteria, listed in full in A.2.1, to reduce a longlist of 156 data sets to a shortlist of 34. These were consolidated where there was significant overlap. After consolidation, 18 data sets remained, which are presented below.

Suitability for use in AI applications

Alignment with the new government's strategic priorities

Private/ restricted access

Connected vehicle and sensors	– UTC and connected infrastructure sensor and configuration data – CCTV – in-vehicle/at stop – Connected vehicle data – CCTV highway-based video analytics
Demographic movement and behaviours	– Mobile phone operating system and service providers – Public transit analytics data – Card payment organisations – Wi-Fi connection data
Earth observation and environmental	– Earth observation and environmental data portals
Freight	– Last-mile courier and haulier operations data
Transport network	– Rail sector data, including data in Rail Data Marketplace – Public transport mobility data – High-definition mapping
Transport operations	– Rail passenger loading data – Ride-sharing and micromobility trip and ops data – CCTV track and environs video feeds – Highways analytics data

3.2 Key findings

The following section summarises insights from expert interviews that focused on three main areas: current data use and enabling partnerships, gaps in desirable but unavailable data and potential ways DfT could support improved access in the future.

3.2.1 How data is being used

There is huge diversity in the way that data is being used across the sector. One emerging area is the use of floating vehicle data (FVD) in network management, particularly when integrated with Urban Traffic Management and Control (UTMC) systems or used to validate other data sources, such as sensors. Data from sources such as Waze, TomTom and INRIX is considered to be reliable, providing real-time insights into congestion, collisions and road conditions. However, data-sharing agreements can take time to establish. ML applications currently under development, or in use, focus on tasks such as congestion forecasting, stopped-vehicle detection and inferring signal timings. Meanwhile, CCTV data from stations, stops and on-board public transport vehicles is increasingly used to monitor crowding and identify areas requiring interventions, such as bus priority measures. More information on use cases can be found in A.1.1.

3.2.2 Desirable but currently unavailable data

While real-time signal phasing data is highly valued for its potential to improve network management and traffic efficiency, it remains largely inaccessible despite existing procurement arrangements. Data on active travel modes, particularly pedestrian movement, also lacks the necessary granularity to support innovation in services designed for walking and wheeling. Although the demand for data is increasing, it is not always fully exploited. For example, EU legislation mandating data sharing from connected vehicles has not been matched by consistent data use, leaving some manufacturers hesitant to increase engagement. Linking movement data to human activity remains crucial for understanding travel motivations and influencing behaviour to reduce transport-related harms.

3.2.3 Barriers to access

3.2.3.1 Technical

Access to real-time signal phasing and UTC sensor data is widely seen as challenging to access due to legacy system designs that did not anticipate external data use, vendor-controlled “black box” systems, unclear data availability, lack of open by default procurement practices and limited market competition, all of which constrain innovation. Additionally, there seems to be a lack of a compelling business case for vendors to invest in working with these systems to extend them beyond “business as usual”.

3.2.3.2 Structural

Access to data is often constrained by a mix of structural challenges. Business case processes frequently undervalue the role and benefits of data, with procurement for sensor-based infrastructure often failing to ensure service continuity or prevent vendor lock-in. Siloed departmental and organisational working was perceived as a significant barrier to access and exploitation of data. Examples include transport and health service planning. There is a need for common metadata, terminology and an assumption of cross-department data use. Additionally, inconsistent powers among public authorities restrict scalability and the wider application of innovations developed locally. Sustainable data use ultimately depends on strong partnerships, with collaborations, such as that between Transport for West Midlands (TfWM) and West Midlands Police, a model example.[1]

[1] Safer Travel Partnership | Transport for West Midlands

3.2.3.3 Capacity and skills

Control rooms face coordination challenges due to limited capacity, particularly when working with national agencies, such as National Highways and Network Rail. Consolidating these functions within public authorities could help improve efficiency and foster innovation by reducing duplication across multiple highway authorities. However, ongoing skills and recruitment challenges continue to limit innovation and resilience, where difficulties in attracting young professionals have led to staff shortages over time.

3.2.4 How DfT could help

3.2.4.1 Governance

Improving data access will require stronger governance and clearer frameworks for data management. Establishing an information governance team could help streamline privacy and data sharing processes, while encouraging open data by default across all procurements, including UTC systems and sensors. Legal and funding levers should also be explored to unlock data from existing systems, such as real-time signal phasing, and to support legacy contracts rather than waiting for expiry. Where central funding is unavailable, establishing common licensing terms and a shared public sector data portal would enable more consistent and efficient access across organisations.

3.2.4.2 Procurement

Procurement processes play a critical role in improving the quality and accessibility of transport data. Clear procurement guidance should be developed to ensure that data suitability requirements are established. In addition, procurement practices should support the onward sharing of network performance data, including FVD.

3.2.4.3 Standards

Agree standards for integrating FVD with UTMC. Furthermore, define standards and advocate for consistent transport powers for all authorities to improve scalability and transferability of data and innovative data-driven solutions. Where data standards are needed, always look for suitable global or pan-national standards before creating new ones.

3.2.4.4 Technical solutions

Developing a national FVD data set covering journey times and speeds would provide the necessary granularity for multiple applications, including effective monitoring and evaluation of schemes. Establishing a national FVD portal could further improve accessibility for public sector users and represents a key opportunity for DfT support, particularly in standardising integration with existing UTMC systems. In parallel, national methods for filling data gaps, such as incomplete ANPR coverage, should be implemented using validated statistical approaches to ensure reliable data sets across the transport network.

3.2.4.5 Strategic

Seek opportunities for better data linking observed movements to other human activity to understand motivations and purposes and provide an evidence base to support policy formation.

4 Analysis and insight

4.1 From shortlist to high-potential data sets

The selection process involved the development of a series of challenge statements. These statements posed questions aimed at tackling specific transport problems in part or in whole. For example: ***How might we enhance transport conditions for all road users and meet policy objectives through network management that prioritises movement of people and goods?*** See A.3.1 – A.3.3 for more information. The five data sets that have been further prioritised are presented below:

Last-mile courier and haulier operational data	Real-time information on vehicle locations (parked, in use or in transit), vehicle load/capacity information, goods priority information (time-sensitive deliveries), origins and destinations.
Mobile phone operating system and service providers	Apple (iOS) and Google (Android) system services, such as Global Positioning System (GPS), accelerometer, other sensor data, mode inference, Apple/Google Pay, Apple/Google Maps.
CCTV highway-based video analytics	Continuous feeds of video from CCTV cameras on the highway network. Applications of Computer Vision (CV) data extraction across single to multiple cameras for object detection and ANPR reading. Suitable for vehicle/person identification and classification, speed and trajectory detection.
UTC and connected infrastructure sensor and configuration data	UTC feeds from sensors on instrumented highway to include configuration information and, where possible, dynamic, real-time (traffic signal) phasing information. UTC asset status information.
Connected vehicle data	Data from connected vehicle sensors, including general vehicle telematics (for example, acceleration/deceleration, speed, location), safety-related information (such as unusual braking/stopped vehicle information), road surface conditions, object detection and proximity/trajectory information and occupancy data.

Further analysis has also been carried out on these data sets to identify the range of challenge statements they address, their specific use cases, key stakeholders and barriers to access. See A.3.3.

4.2 Interpreting the results

The stage 3 selection criteria were tailored according to the project priorities and the nature of the shortlisting process. Some familiar and well-used portals have not made the final five. This does not indicate a lack of value, only that they did not score highly enough against the specific criteria agreed during this commission. Below are some considerations to help interpret the results:

- Professional judgement was applied at different stages of the process, for example in identifying the strategic priorities to which each data set aligns. Although this was done with group discussions, there remains a risk of bias.

- Data portals were considered as single entries which is a limitation of the methodology. This was an intentional decision to make the process of longlisting manageable. However, it means the full richness of portals and details of individual component data cannot be fully accounted for.
- We have tried to represent the breadth and the depth of the shortlist when selecting five entries from a final shortlist of 18.
- The focus has been on finding data sets that are not as well-known, well-exploited or as readily accessible. Consequently, well-established open data sets and sources, such as the Rail Data Marketplace, score lower than more closed examples, and so have not made the final shortlist. This is not an indication that they are less valuable or not worth further exploration and investment. An added value of the approach is the flexibility in the methodology that would allow re-weighting to, for example, test whether open (rather than closed) data sets are being fully exploited.

4.3 Key findings

- **Many worthwhile potential data sets:** There was a great deal of richness in the range of data identified. The priorities of this commission and the challenge statements helped to reveal the innovation potential for data that is not yet fully or actively exploited. While five data sets have been identified as high-potential, other data should not be neglected.
- **Initial understanding of the barriers highlights a complex system:** Deep understanding is needed to unpick the complex barriers of procurement, issues of standardisation, high data volumes and commercial sensitivity.
- **Few surprises in results:** While this was undoubtedly a useful exercise producing new findings and insights, there were few true surprises in the longlisting and shortlisting exercises, showing a good existing understanding of the data landscape.

5 Summary

In summary, our analysis identified a wide range of data sets, explored their potential use cases and assessed the barriers affecting access and utilisation.

What data is out there?

Across the transport sector, there is no shortage of potentially valuable data. The study confirmed that much of this data is already known to DfT, rather than newly discovered, suggesting that the exercise largely validated, rather than transformed, existing understanding of the data landscape.

Five data sets emerged as having particularly high potential for AI innovation:

- Last-mile courier and haulier operational data
- Mobile phone operating system and service providers
- CCTV highway-based video analytics
- UTC and connected infrastructure sensor and configuration data
- Connected vehicle data

It is worth noting that, while AI applications are often associated with big data, not all data of interest qualifies as big data; much is small-scale, yet capable of significant impact when integrated with other data sets.

The key takeaway is there is already a rich but, in many cases, underutilised data ecosystem. The challenge is less about discovery and more about unlocking, standardising and integrating what is already known and available, or known but not currently accessible.

What can the data be used for?

These data sets support a wide variety of AI transport use cases across the eight identified themes. Examples drawn from the 18 shortlisted data sets include:

Connected vehicles and sensors *Connected vehicle data, including CANBus data*	Understand network-wide data to enhance safety, including near miss data. This could also provide granular information on the autonomous vehicle fleet, including usage of typical vehicles, which could inform policy.
Demographic movement and behaviours *GPS data*	Understand which routes passengers choose during disruptions. This can aid with crowd control and planning.
Earth observation and environmental *Weather models*	Assessment/forecasting of the impact of weather on transport infrastructure.
Energy generation, transmission and emissions *EV charging point data*	Historical usage of charge points could be used to inform policy and development by improving understanding of in-demand locations, helping to prioritise the

	installation of future charge points. This could be combined with data from a relevant energy portal to cross reference the capacity of the electricity grid against need for charge points.
Freight *Courier/haulier - planning and operation data*	With access to a live version of haulier operations, data could be used to feed into UTMC and/or a digital twin of the highway network to enable real-time, dynamic network management that prioritises the movement of goods (and people) rather than just vehicles.
Transport network *Rail Data Marketplace*	Predictive analytics for disruptions, live train time communications, operations support, planning and other passenger experience improvements. Possibility to combine data sets to add value.
Transport operations *UTMC systems*	Support dynamic transport (for example, highway network) management and optimisation (for example, through UTMC or digital twins) and more robust transport modelling (consistency of time and resolution of capture of counts across a fully modelled area).

Is there access?

Access to transport data remains inconsistent, with significant barriers across technical and structural aspects.

Openness of data remains a critical issue; many high-value data sets are locked behind procurement or commercial agreements that were never designed for data sharing. Ownership is often ambiguous, particularly for data sets involving shared systems (UTCs) or third-party suppliers. This creates confusion around who can publish or use them.

Legacy system designs, vendor-controlled “black box” setups, pose another barrier. For example, real-time signal phasing and UTC sensor data is difficult to access because the systems were never built to share information externally. Procurement practices reinforce these silos, locking authorities into vendor-specific systems and discouraging interoperability.

On the structural side, siloed working and inconsistent data sharing powers between authorities limit the scalability of local innovations. Business cases often undervalue the role of data, treating it as an afterthought. Even where sharing is possible, data quality issues, such as inconsistency and lack of standardisation, can undermine AI applications that rely on accuracy and granularity. There are also practical challenges. Local authorities may have the data, but not the time, skills, tools or funding to make it open and usable.

Despite this, there is progress. Public access to some data sets is improving, and there is growing recognition that relationships can unlock data. Strong partnerships between public and private companies, combined with clearer governance, licensing and national portals (such as for FVD data), could standardise access and enable cross-sector innovation.

6 Glossary

Terms

Big data	Extremely large and complex data sets that are too vast, fast-changing or varied to be efficiently processed and analysed using traditional data management tools or methods.
Data set	A structured collection of related data, typically organised in a specific format, such as tables, files or databases, and used for analysis, modelling or decision-making.
Data source	The origin or provider of data or a data set. This could be an organisation, system or platform that generates, collects or maintains the data.
High-potential data sets	Data that may not yet be available but has high potential to enable AI innovation in transport.
Use case	A specific application or scenario where a data set can be used to solve a problem, generate insights or support decision-making.

Abbreviations

ADEPT	Association of Directors of Environment, Economy, Planning and Transport
AI	Artificial Intelligence
ANPR	Automatic Number Plate Recognition
API	Application Programming Interface
CA	Combined Authority
CANBus	Controller Area Network Bus
CCTV	Closed Circuit Television
EV	Electric Vehicle
FVD	Floating Vehicle Data
GDELT	Global Database of Events, Language and Tone
ML	Machine Learning
T&C	Terms and Conditions
UTC	Urban Traffic Control
UTMC	Urban Traffic Management and Control