Informal document WP.6 No. 1 (2023)

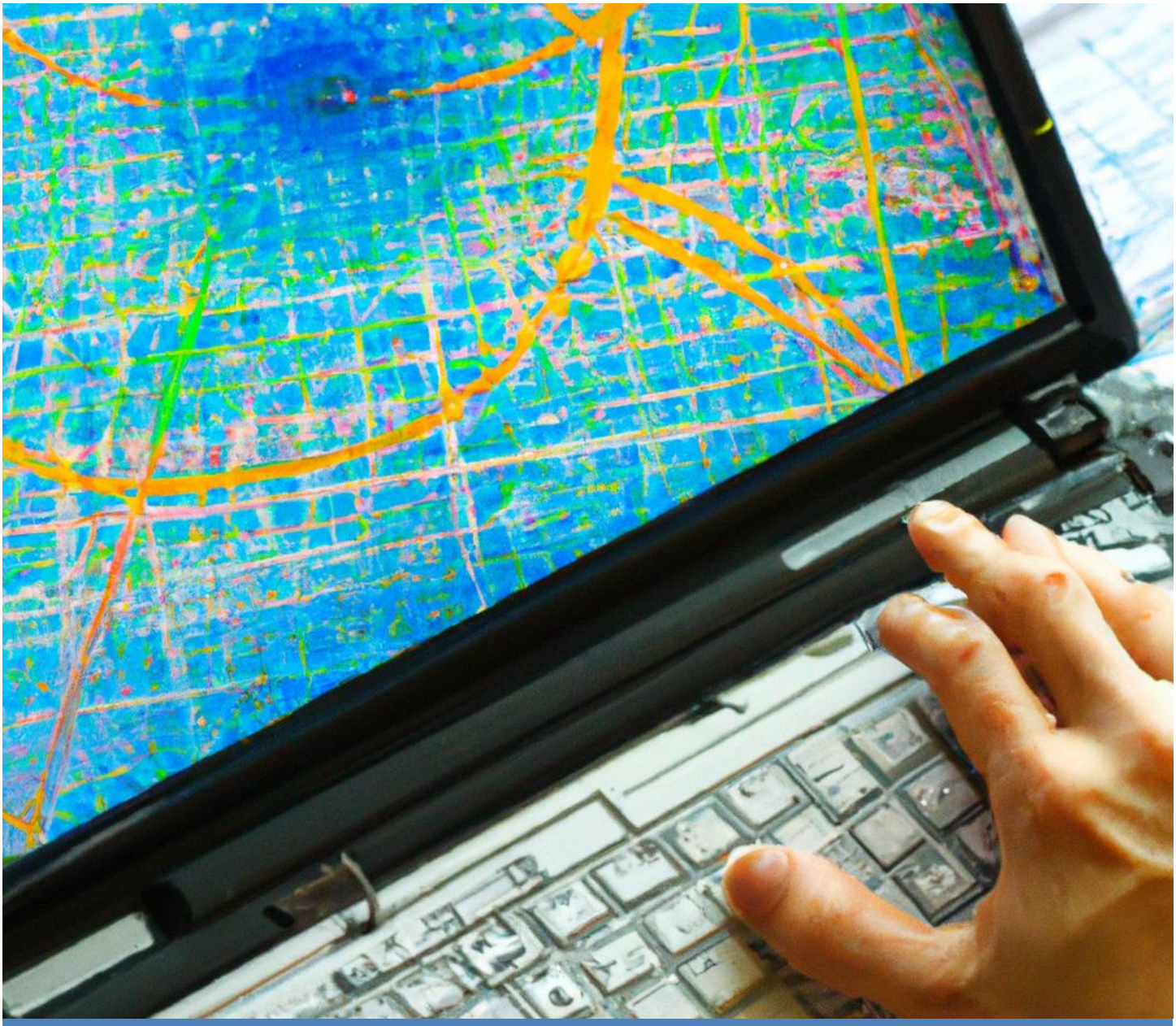
Distr.: Unrestricted
17 May 2023
English only

**Working Party on Transport Statistics**

**Seventy-fourth session (in-person only)**
Geneva, 15-17 May 2023
Item 3 of the provisional agenda
**Big Data and innovation in transport statistics production**

# Draft handbook on use of mobile phone data for transport and commuting statistics

1. This document contains a draft of the handbook on use of mobile phone data for transport and commuting statistics, as developed by the Task team on Mobile Phone Data of the UN Committee of Experts on Big Data. Delegates are encouraged to provide comments and additional examples for inclusion in the following weeks.



# Use of Mobile Phone Data in Transportation

UN Council of Experts in Big Data

# Contents


Executive Summary ..... 3
Abbreviations ..... 4
Introduction ..... 5
Sources of Transport Data ..... 7
  Legacy Sources & Types ..... 7
  Emerging Sources & Types ..... 8
Mobile Phone Data Landscape ..... 10
  Types of Mobile Phone Data ..... 10
  Typical Data Flow and Processing ..... 12
    Data Cleaning Techniques ..... 16
  Understanding the Limitations ..... 19
Applications of Mobile Phone Data in Transportation ..... 20
  Transport Statistics ..... 20
    Case Study: Replacing Indonesia's Household Travel Surveys ..... 21
  Transport and Urban Planning ..... 22
    Case Study: UK's Use of MPD to Understand Modal Usage ..... 23
  Transit & Public Transport Planning ..... 24
    Case Study: Tracking Rail Passengers in Austrian Railways ..... 24
    Case Study: Seoul's Use of MPD in Planning Night Bus Service ..... 25
  Traffic Analysis ..... 26
    Case Study: Poland's Use of MPD in Traffic Analysis ..... 26
  Operational Decision Support ..... 26
    Case Study: To be Added ..... 27
  Events and Crises Management ..... 27
    Case Study: Dubai's Use of MPD in Events Management ..... 27
  Traffic & Transit Demand Prediction ..... 28
Conclusions ..... 29
References ..... 30

# Executive Summary

Mobile phones have become ubiquitous in today's world, with nearly 8 billion mobile devices in use worldwide. With this widespread adoption comes a wealth of data that can be used for transport planning, operations, and regulation. In this handbook, we will explore the different techniques for collecting, processing, and analyzing mobile phone data and discuss the role of different stakeholders in leveraging this data for transportation, in particular for production of official statistics.

The handbook is primarily aimed at transport statisticians, planners, operators and regulators to understand the nuances of using MPD in decision-making by shedding light into the types of mobile phone data, their pros and cons, type of processing they go through before they are useful as well as the applications in transportation. We have also identified case studies from around the world to help understand how some of the agencies and cities are currently using MPD in their planning and operations.

# Abbreviations

<Add list of abbreviations here>

# Introduction

In 2019, the Task Team on Mobile Phone Data (MPD) drafted the first handbook on the use of MPD for official statistics (UNSD, 2019) which provided a broad and general background and guidance for countries in starting out with such projects.

In its continuing effort to provided clear guidance to countries and promote the use of MPD for official statistics, the UN Committee of Experts on Big Data and Data Science for Official Statistics (UN-CEBD) Task Team on MPD embarked on a project that aims to develop handbooks that are designed to guide countries who plan to use MPD for statistics and indicators. Handbooks in displacement and disaster, information society, dynamic population, migration, and tourism have already been produced, and this handbook covers transport statistics.

National statistical systems (NSS) are under increasing pressure to produce timely, high-quality data to monitor how people and goods move around their country. This came to the fore in particular in 2020 during the start of the COVID-19 pandemic, where lockdowns led to drastically changed mobility patterns and traditional official statistics that would be released with a large time lag would not have been useful. It is believed that Mobile Phone Data (MPD) as a big data source offers a great promise for its use in measuring transport and mobility. MPD may be able to fill data gaps that are required for policy making related to the SDGs, in particular related to transport-related Sustainable Development Goal (SDG) targets like 3.6, 9.1 and 11.2. Further, the concept of no-one being left behind makes these data particularly valuable when they can provide information on certain population or geographic sub-groups, that may not be available at the necessary level of granularity from traditional sources. These new data sources could be used as a complement or even substitution to traditional/conventional data sources such as travel surveys, traffic counts, and other administrative data such as public transport ticket information. Traditional or conventional data sources can have a number of drawbacks, including timeliness, granularity and others including cost.

This handbook on the Use of MPD for Transport Statistics provides practical guidance on how to collect and aggregate data from MPD so they can be used for producing transport statistics. It uses non-domain-specific material developed from the other handbooks where relevant, and focusses this application on a limited number of use cases, which have either succeeded already in producing data to the standards of official statistics or are expected to in the future. It reviews these use cases, analysing the approach taken in terms of data access, ensuring privacy maintenance, data quality control, and any relevant comparisons with existing datasets.

MPD will not be replacing traditional transport statistics sources in the near future. In particular data on gender, trip purpose and income level obtained from national travel surveys are not likely to be replicated from MPD or other big data sources soon. But travel surveys are costly, and it may be that MPD can complement travel survey data through greater granularity in geographical

coverage and more timely data on movements. This may mean for example, that lower and middle income countries can justify doing a travel survey every five or ten years, with integrated MPD data filling in the gaps for non-surveyed years.

# Sources of Transport Data

Transport statisticians, planners, and operators rely on several sources of data to understand how people and goods are moved around and between countries using roads, railways, inland waterways, seas and the air. Information on this is crucial to know how people access jobs, services and leisure, have access to transport services, average distances travelled by mode and the number of trips taken. In parallel, data allow understandings of the quantities of goods shipped on different modes in different countries. All this information is useful for its own sake, but equally feeds in to measuring other areas such as access to jobs and services, climate and energy goals, pollution, logistics chains management, international trade, gender aspects and road safety. Traditionally, the primary sources of transport data included household travel surveys to understand the travel demand in terms of trip purpose, departure time etc., traffic data to understand the private vehicle transport statistics, and public transport ticketing data to understand the origin-destination of public transport users. Transport planners & traffic engineers typically use simulation-based models that base trip generation on land-use data from census to estimate travel demand. Two major categories of data sources are described below:

## Existing Sources & Types

Some of the legacy data types and sources used in transport statistics, planning and operations. They are:

1. **Travel Surveys & Census Data**, covering the types of data collected via census or household travel surveys. Given the stated preference nature of this data collection, the sample size is low, but the available data related to traveler preference is of high quality. The dataset could have inherent biases associated with it based on how statistically distributed, the respondent profiles are. Some of the information collected are number of trips, typical trip departure times, origin and destinations, trip chaining, mode and trip purpose.
2. **Transport Network Information**, covering the information (aggregated or geospatial) on transport networks such as roadway, lane-level information; traffic control devices; rail lines, public transport route, stops, schedules etc; Depending on the jurisdictional characteristics, this information could be aggregated to KPIs such as centerline kilometers, public transport coverage etc. Transport Network Information can also be available as kml maps for road networks, GTFS files pertaining to public transport, GBFS information related to shared mobility etc.
3. **Vehicle Traffic Data**, could be statistical or historical information about the usage of the transport network, or real-time information such as travel-time, congestion index etc. The source of this information could be roadway-based sensors or other Intelligent Transport System (ITS) equipment. Bluetooth sensors or Automatic Plate-Number Recognition (APNR) systems can also be used in estimation of traffic volumes and Origin Destination Matrices.

4. **Public Transport & Shared Mobility Ticketing Data** are generally a source of transit usage statistics or provide OD matrices on public transport and shared mobility modes. Depending on the ticketing system, the data could just provide station/stop counts or traceable user-movement patterns. (Some cities use pay-per-use tickets, whereas some other cities use smartcards that link all check-ins and check-outs). In addition, based on jurisdictional restrictions, cities may or may not mandate shared mobility providers to share usage data with them for planning and regulatory purposes and with the public for transparency purposes.
5. **Safety & Accidents Data** can provide insights into transportation safety metrics, as well as accident hotspots, depending on data availability. Typical sources of safety data are from accident reports that are either managed by traffic policing units or a jurisdictional safety database. For example, USDOT/NHTSA manages a Fatality Analysis Reporting System at country-level that manages all fatalities associated with traffic accidents/crashes.
6. **Logistics Data** are collected and aggregated at different jurisdictional levels. This can include smart-tags that shows the type and weight of commodities carried to weigh-station that weigh vehicles and log them over different fixed locations in the country/city.
7. **Vehicle & Driver Registration Data** are typically used to understand transport statistics with respect to vehicles, vehicle types, mobility modes and modal split under different jurisdictions. Such data is generally managed through Department of Motor Vehicles (or similar) database when drivers and vehicles are licensed to operate. Insurance database also provides reliable source of vehicle and driver registration data.
8. **Transport Regulatory Data** typically include fines and points associated with driving violations. Such data are typically used by regulatory agencies to ensure safe transport operations, but also to study potential hotspots for unsafe transport operations.

Please note that the listing above is not exhaustive and is only meant to provide an overarching classification of transport-related data sources and types.

## Emerging Sources & Types

In addition to the sources of data that are used traditionally in transport statistics, planning and operations, there are certain emerging sources in the last decade that has proven to be of importance in the context of transport. Some of the them are:

1. **Probe Vehicle Data** representing the data from GPS devices of vehicles or occupants of the vehicles that are collected at high frequency to represent the state of the roadway or public transport route that is being used. For example, Google Maps use user-location, speed, and user-provided inputs to enhance its mapping data. (This is also called **Floating Car Data**)
2. **Connected and Autonomous Vehicles Data**
3. **High Definition Mapping** represents the dynamic maps used by Connected and Autonomous Vehicles in their dynamic driving decisions. This can include lane-level traffic information, curb-space information, dynamic work-zone information etc.

4. **Mobile Phone Data** represents the use of location (and other attributory) information from mobile phones in transport and mobility system design. Mobile Phone Data could either be collected from telecom providers as tower association data, or from smartphone applications as user-location data, or even as digital surveys as user provided inputs to mobility-related questions. The purpose of this document is to enable readers to better understand Mobile Phone Data in the context of transport and commuting, its types, uses, applications and limitations.

# Mobile Phone Data Landscape

The use of Mobile Phone Data (MPD) in advancing the mobility systems has seen a rise in the recent years with increasing cellphone and smartphone access. Examples from around the world include Indonesia's use of MPD to replace aspects of National Household Travel Surveys, India's use of MPD by Ola Mobility Institute in deriving ease of moving index across the metropolitan cities, and case-studies from US where cities and administrations are using data from Air Sage (an MPD provider) for advancing their mobility goals.

<Add an illustration>

## Types of Mobile Phone Data

Typically, mobile phone data represents the data collected via mobile phones, and can be of three types (or a combination of these):

1. **Data from Telecom Operators or Mobile Network Operator (MNO)**, constituting the cell-tower associations of cellphones collected from telco providers. The advantage of telco data is that the cell-tower switching happens every time a user moves, however the accuracy of telco data depends on the density of cell-towers and the additional processing

required to improve this resolution is still under development. Jurisdictions may also implement additional privacy protection rules on telco providers that may reduce the location accuracy even further.

There are three type of data probing done by telecom operators:

| Call Detail Records | Passive Signaling Data | Active Signaling Data |
|---|---|---|
| This refers to data produced whenever the subscriber makes or receives a call, sends or receives a SMS, or accesses data using their phone. The data typically include at least the subscriber/phone ID, the time of the event, and a location (the location of a cell tower, not the device itself). Data are collected and stored for billing purposes and so no additional collection would be required. | This data includes logs of every connection of a mobile device to the network, which is typically more frequent than CDR events. The greater volume of data requires more processing and storage capabilities, and as such the data may only be stored for a limited time (weeks/months). | These are generated when MNOs decide to send (or "ping") signals to mobile devices. The MNO normally needs a specific justification for doing this (e.g. a mandate from law enforcement) therefore there is little scope for using these data for statistical purposes. |

2. **Data from Smartphone Applications**, consisting of data collected by data suppliers by tapping into smartphone applications that consume user locations. The advantage of this is the higher accuracy of location data, since it is collected from the GPS sensors of the phones. However, due to increasing concerns on user privacy and few data abusers, an increasing number of users are opting out from sharing accurate locations with their applications.

Similar to the telecom data, there are three types of location data collected by the smartphone applications. Depending on the privacy settings for each application, the user may choose one over the other, either voluntarily or involuntarily.

| User Check-in Data | Foreground Location Data | Background Location Data |
|---|---|---|
| Applications may trigger users to check-in at locations either as part of the social media postings, or as part of incentivizing the users to contribute to the community. | Applications may collect precise or approximate user-location based on user setting when applications are running (for example to support mobility services, navigation services etc.). | Applications may also collect data from users in the background (if settings allow), to support third-party data sharing. |

3. Digital Survey Data, consisting of user responses to generalized or targeted surveys pushed to applications by app developers. Typically, digital surveys are used as a value-add on top of the first two types of MPD to understand personalized experience.

Regardless of the type of data collection, the useability of the data is attributed to the following factors:

1. **Latency**: This refers to the time delay between collection of mobile phone data and its availability for the end-user. If the data is available in near real-time, it could be used for operational decision-making. Otherwise, the data use is limited to statistics and planning.
2. **Precision**: The accuracy of the mobile phone data is also of importance. When it comes to location data, the data from tower-associations will have relatively low precision based on the coverage area of each tower. Microcell towers will be able to provide improved location accuracy. The app-based data could have higher accuracy based on user setting.
3. **Ping Rate**: The rate of location pings is also an important factor, and depends on a variety of factors, such as type of data collection, user-chosen factors, and privacy laws of the country on whether passive probing is allowed or not. Lower ping rates will limit the use of data, while higher ping rates could increase complexity and legality of data usage.
4. **User Penetration**: This represents the percent of commuters who actively contribute to the MPD data provisioning pool. Typically, MNO-provided data will have a higher user penetration rate, especially if only few number of operators are allowed in the market. App-based data collection is hence done through brokers that implement data collection SDKs in multiple apps that allow increasing the user penetration.
5. **Completeness**: This represents the number of attributes present in the mobile phone data. Typically, MNO data includes at least a time-stamp, tower location, and an anonymized subscriber ID. Auxiliary information such as demographic details of the user, billing address etc. can support in increasing the usability of this data. App-based data can also collect auxiliary information from other sensors to detect mode of commute.

The mapping below shows the typical range of data quality for the three subsets of mobile network operator data and data collected from smart phones:

| | **MNO Data** | | | **Application Data** | | |
|---|---|---|---|---|---|---|
| | **Call Detail Records** | **Passive Signaling Data** | **Active Signaling Data** | **User Check-in Data** | **Foreground Location Data** | **Background Location Data** |
| **Latency** | Low-Med | Low-Med | Low-Med | Med | Med | Med-High |
| **Precision** | Low-Med | Low-Med | Low-Med | High | High | Med-High |
| **Ping Rate** | Low | Med | High | Low | Low-Med | High |
| **User Penetration** | Med-High | Med-High | Low | Low | Low-Med | Low-Med |
| **Completeness** | Low-Med | Low-Med | Low | Med-High | Med-High | Low-Med |

## Typical Data Flow and Processing

The use of MPD in transport is a complex process, and involves a variety of stakeholders and processing types. In this section, we discuss the typical data flow between the different stakeholders, and the processes involved.

<Add an illustration>

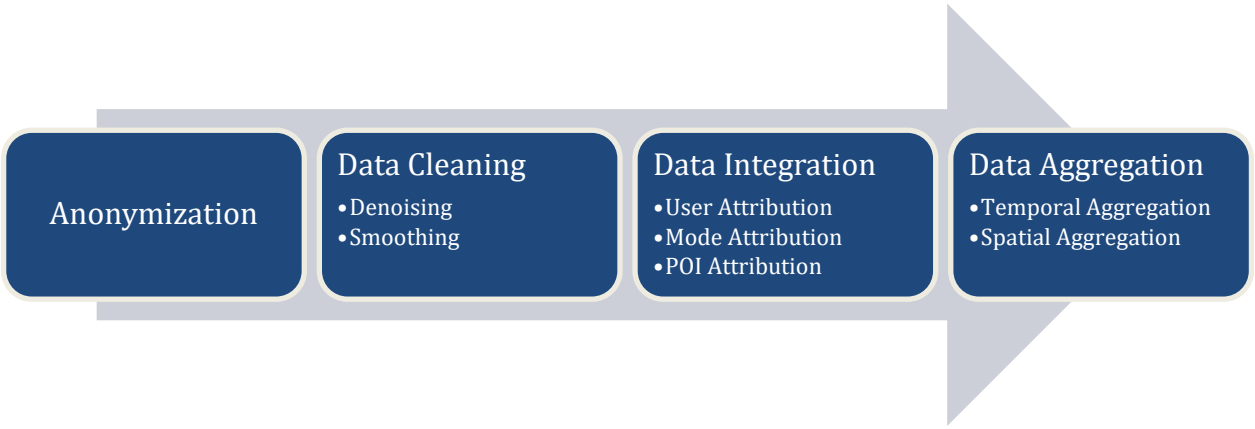
Typically, the dataflow of MPD occurs between different stakeholders listed below (non-exhaustive):

1. **Commuters** (Mobile Phones), represents the commuters whose mobile phones are used as the raw data source. The data could be MNO-data or app-based location data. Their role is to understand what data is collected, by whom and what privacy controls are available at their disposal, so as to protect personally identifiable information (PII).
2. **Data Collectors & Providers**, represents the stakeholder(s) who collect data from the commuters are provide it to the end-users in a clean, usable way that not only protects their business, but also their commuter privacy. In the case of MNO-data, this could be the MNO operator, or a consortium of operators with other technology companies with the know-how of how to collect and process this information. In the case of app-based data, it could be the app developer or a consortium of developers with the potential to reliably monetize this data.
3. **Data Regulators**, represents the governmental or non-governmental organizations that actively or passively regulate such activity and ensure protection of user rights, data protection laws and citizen privacy. Active regulation is done in certain markets where telecom regulators have algorithms which actively detect privacy violations and data collection within the ecosystem of telecom operators.
4. **Data Brokers**, represents the companies that work with multiple data collectors to collect data, process them according to end-user needs, potentially synthesize value-added attributes and provide them to the end-users. Data brokers typically are the ones who deduplicate data received from multiple collectors. Data brokers also may operate data monetization platforms that the end-users use to get the data and pay for their usage.
5. **Data End-Users**, represents the users of MPD data, and could be statistics departments, transport operators, metropolitan planning organizations (MPOs) or private companies who use MPD to target and improve their business.
6. **Benefactors**, represents the stakeholder groups that are benefitted from the end-users' use of MPD data. Typically, when MPD is used for transport planning and operations, the

benefactors are the commuters itself, were the data originated, completing the full cycle of data flow.

It is important to know that the same organization could have multiple roles in this data flow chain.

When it comes to the type of processing the raw data goes through different types of processing before it can be used for transport statistics, planning and operations. The typical data processing pipeline is showing below:




1. **Anonymization** – Typically, the first step in using Mobile Phone Data is to anonymize the data, and it refers to the process of removal of personally identifiable information (such as name, age, address, ID numbers etc.), addition of proxy identifiers (demographic categories such as age group, internal IDs etc.), removal of surrogate identifiers (precise home location) etc. Techniques such as k-anonymity, l-diversity, and differential privacy can be used to protect individual privacy while still allowing for meaningful analysis

2. **Data Cleaning** – This is the process of detecting and correcting (or removing) corrupt or inaccurate records from the data set. This can include inconsistencies, inaccuracies, duplicates, or irrelevant data. For instance, in mobile phone data, GPS inaccuracies may occur due to signal loss when a user is in a building or tunnel. These inaccuracies should be detected and handled appropriately, either by correction or removal. Some of the techniques used in data cleaning are denoising and data smoothing. These data cleaning techniques are explained later in this section.

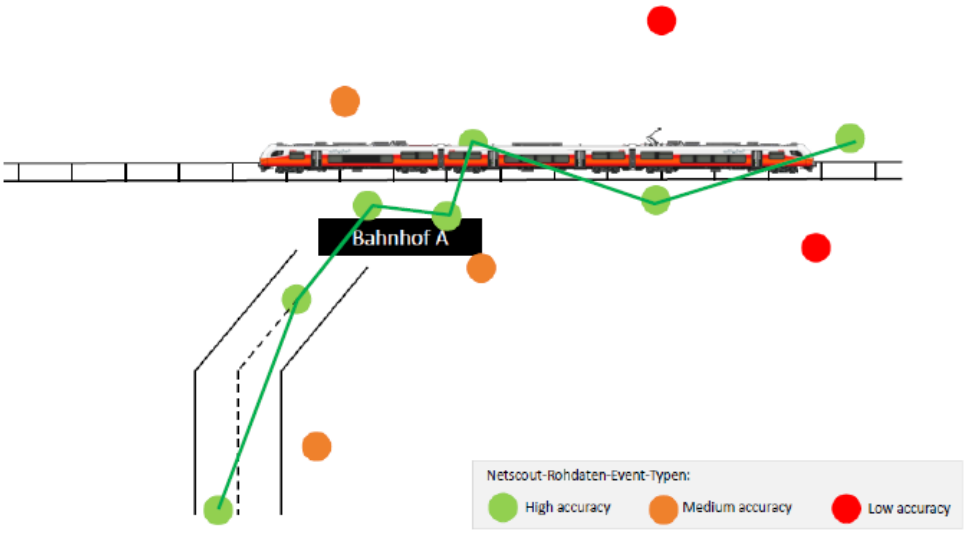



Figure showing smoothing and denoising MPD data to attribute them to the railway line in Austria (Source: Invenium)

3. **Data Integration -** Data integration involves combining data residing in different sources and providing users with a unified view of the data. Given the diverse sources of mobile phone data (e.g., call detail records, GPS data, app-based data), these sources must be integrated into a coherent dataset. This can be achieved using a common identifier, such as a user ID or a device ID. Some of the examples are:

    a. **User Attribution** – refers to the process of attributing data to a specific user or internal ID. This is important, especially if commuters use multiple mobile phones, or other datasets would be merged to create a value-added dataset, such as mode-detection sensor data.
    b. **POI Attribution** – refers to the process of attributing map-based points of interests to people movement data. This could be used for mode detection, home/work detection, trip purpose detection etc.
    c. **Mode Detection** – refers to the process of using auxiliary data to detect mode of travel for each leg of the journey.

4. Data Aggregation: This refers to the process of reducing the datasets for use by transport planners, regulators and operators for their specific use. High-resolution datasets are typically difficult to work with. Hence, based on the use-case, data aggregation is performed. For example:
    a. **Temporal Aggregation:** Mobile phone data is often collected at high temporal resolutions, such as every second or minute. Temporal aggregation involves combining data over time intervals to produce a summary statistic for that interval (may be hour or day or minute). For instance, mobile phone data collected every second might be aggregated into hourly or daily intervals to analyze trends over time. The appropriate aggregation level will depend on the specific research question

b. **Spatial Aggregation:** Like temporal aggregation, spatial aggregation involves aggregating data over spatial units. This can help reduce the size of the dataset and protect individual privacy. For example, mobile phone data might be aggregated to the level of census tracts or neighborhoods

## Data Cleaning Techniques

Five data cleaning techniques are discussed in this section.

1. **Outlier Detection and Removal**

   Outliers are data points that are significantly different from others. Outliers can occur due to measurement errors, data processing errors, or simply due to natural variability. They can distort the analysis and lead to incorrect conclusions. Various statistical methods and machine learning algorithms can be used to detect and remove outliers. For instance, the Z-score method, the IQR method, or DBSCAN can be used for outlier detection in mobile phone data. Examples of using this technique:

| Use Case | Details | Reference |
|---|---|---|
| **Urban Mobility Study** | In a study by Zheng et al. (2008), GPS data collected from mobile phones was used to analyze urban mobility patterns. The researchers used a distance-based outlier detection method to identify and remove erroneous GPS records that were significantly different from the user's typical location patterns | Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W. Y. (2008). Understanding mobility based on GPS data. In Proceedings of the 10th international conference on Ubiquitous computing (pp. 312-321) |
| **Call Detail Record Analysis** | A study by Becker et al. (2011) used call detail records (CDRs) to analyze social networks. They identified outliers by comparing each user's call frequency and duration against the overall distribution, and removed users with extremely high or low values that were likely due to recording errors or special circumstances (e.g., business accounts) | Becker, R. A., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., ... & Volinsky, C. (2013). Human mobility characterization from cellular network data. Communications of the ACM, 56(1), 74-82 |
| **Network Traffic Analysis** | In a study conducted by Erman et al. (2006), network traffic data collected from a mobile service provider was analyzed to detect anomalies and faults. The researchers used a clustering-based outlier detection method to identify unusual traffic patterns that were significantly different from typical usage patterns | Erman, J., Mahanti, A., Arlitt, M., & Williamson, C. (2007). Identifying and discriminating between web and peer-to-peer traffic in the network core. In Proceedings of the 16th international conference on World Wide Web (pp. 883-892) |

2. **Smoothing**

   Smoothing techniques can be used to reduce high-frequency noise in the data. For example, moving averages or exponential smoothing can be used to smooth time series

data from mobile phones. Kernel smoothing or Gaussian smoothing can be used to smooth spatial data. Examples of using this technique:

| Use Case | Details | Reference |
|---|---|---|
| **Mobility Pattern Analysis** | In a study by Asakura and Hato (2004), GPS data from mobile phones was used to analyze mobility patterns. The raw GPS data, which can be quite noisy due to factors like signal loss or multi-path distortion, was smoothed using a technique known as Kalman filtering. This helped to produce more accurate and stable estimates of the user's location and velocity | Asakura, Y., & Hato, E. (2004). Tracking survey for individual travel behaviour using mobile communication instruments. Transportation Research Part C: Emerging Technologies, 12(3-4), 273-291 |
| **Traffic Flow Estimation** | In a study by Herrera et al. (2010), mobile phone data was used to estimate traffic flow on highways. The data, which included measurements of the speed of individual vehicles, was smoothed using a technique called Local Regression (LOESS). This helped to reduce the noise in the speed measurements and produce more accurate and reliable estimates of traffic flow | Herrera, J. C., Work, D. B., Herring, R., Ban, X. J., Jacobson, Q., & Bayen, A. M. (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. Transportation Research Part C: Emerging Technologies, 18(4), 568-583 |

### 3. Dimensionality Reduction

Dimensionality reduction techniques can be used to reduce the number of variables in the dataset, which can help remove noise and improve computational efficiency. For instance, Principal Component Analysis (PCA) can be used to transform the original variables into a smaller number of uncorrelated variables that capture most of the variance in the data. Examples of using this technique:

| Use Case | Details | Reference |
|---|---|---|
| **Human Mobility Characterization** | A study conducted by Gonzalez, Hidalgo, and Barabasi (2008) used mobile phone data to study human mobility patterns. The researchers used Principal Component Analysis (PCA), a popular dimensionality reduction technique, to simplify the representation of the data and reveal the primary modes of variability in human travel patterns | Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. Nature, 453(7196), 779-782 |
| **Behavioral Pattern Analysis** | In a study by Dong et al. (2015), mobile phone data was used to analyze behavioral patterns. They utilized Non-negative Matrix Factorization (NMF), a dimensionality reduction technique that provides a parts-based, sparse, | Dong, Y., Yang, Y., Tang, J., Yang, Y., & Chawla, N. V. (2014). Inferring user demographics and social strategies in mobile social networks. In Proceedings of |

| | | |
|---|---|---|
| | and non-negative representation of the data. This helped to identify key behavior patterns and simplify the representation of complex, high-dimensional mobile phone data | the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 15-24) |
| **Social Network Analysis** | In a study by Blondel et al. (2015), mobile phone data was used to analyze social networks. They used a dimensionality reduction technique called t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize high-dimensional data in two dimensions, revealing key patterns and clusters in the data | Blondel, V. D., Decuyper, A., & Krings, G. (2015). A survey of results on mobile phone datasets analysis. EPJ Data Science, 4(1), 10 |

**4. Data Imputation**

Missing data can introduce noise into the analysis. Various data imputation methods can be used to estimate missing values, such as mean imputation, regression imputation, or more sophisticated methods like multiple imputation or K-Nearest Neighbors (KNN) imputation. Examples of using this technique:

| Use Case | Details | Reference |
|---|---|---|
| **Mobility Analysis** | In a study by Pappalardo et al. (2013), mobile phone data was used to analyze human mobility patterns. The researchers had to contend with missing data due to non-uniform sampling intervals, which they addressed using data imputation. They used an imputation method based on trajectory similarity, where missing data points were filled in using the trajectories of similar users | Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., & Barabási, A. L. (2015). Returners and explorer's dichotomy in human mobility. Nature communications, 6, 8166 |
| **Transportation Planning** | In a study by Toole et al. (2015), mobile phone data was used to study commuting patterns and inform transportation planning. The researchers faced missing data due to users turning off their phones or losing signal. They used a multiple imputation technique to estimate the missing data, based on the known characteristics of each user's commuting pattern | Toole, J. L., Herrera-Yaqüe, C., Schneider, C. M., & González, M. C. (2015). Coupling human mobility and social ties. Journal of The Royal Society Interface, 12(105), 20141128 |

**5. Signal Processing Techniques**

Techniques from signal processing, such as wavelet transforms or Fourier transforms, can be used to remove noise from the data. For example, a Fourier transform could be used to identify and remove periodic noise in the data. Examples of using this technique:

| Use Case | Details | Reference |
|---|---|---|
| **GPS Trajectory Noise Reduction** | In a study by Zheng et al. (2008), they used a method called the Kalman filter to remove noise from GPS trajectories obtained from mobile phones. This process greatly improved the accuracy of the GPS data, which was crucial for their goal of understanding human mobility patterns | Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008). Learning transportation mode from raw GPS data for geographic applications on the web. In Proceedings of the 17th international conference on World Wide Web (pp. 247-256) |
| **Indoor Positioning Systems** | In a study by Liu et al. (2012), they used a signal processing technique called the Wiener filter to denoise the signal strength of Wi-Fi signals received by mobile phones. This helped improve the accuracy of an indoor positioning system that used the strength of Wi-Fi signals to determine a user's location within a building | Liu, H., Darabi, H., Banerjee, P., & Liu, J. (2007). Survey of wireless indoor positioning techniques and systems. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 37(6), 1067-1080 |

## Understanding the Limitations

While considering using MPD for transport analytics functions, agencies and jurisdictions must be cognizant about the limitations it places on the data and outcome from the data. Some of them are listed below:

1. Access – Due to access restrictions from telecom regulatory authorities, not all MPD data may be available for use. Sometimes, regulatory authorities may impose restrictions on the currency of data, resolution of data etc.
2. Privacy – Due to privacy restrictions of different jurisdictions, MPD data may not be attributed to specific demographics. In addition, MNO operators may mask the real origin and destination of trips.
3. Biases - For transport statistics to be relevant, they need to be representative of movements of the complete population. Whether the MPD come from a single MNO, or from multiple MNOs, it is likely that mobile phone users will have different mobility patterns than those without mobile phones. Although the share of the population without a mobile phone is shrinking, these are likely to over represent the elderly and children, recent immigrants, people without fixed accommodation etc. Adjusting post-facto any insights from the MPD to reflect any sub-populations excluded is necessary in order to meet quality assurance.
4. Precision – Based on the source of transport data from MPD, the precision may vary. In general, app-based data would have higher precision than MPD from MNOs. Some dense urban areas utilize microcells which may improve precision.
5. Processing Requirements – Unlike traditional data, MPD data are high-noise, high-resolution datasets. Hence IT infrastructure and big data expertise required to process them into meaningful outcome are higher.

# Applications of Mobile Phone Data in Transportation

Transportation agencies and statistics commissions across the world are moving towards either solely using high-resolution people movement data from mobile devices in their reporting, planning and operational needs, or at least use it in supplementing other available data (such as travel surveys, ticketing data etc.). Some of the benefits from the MPD data in transportation, as seen from cases around the world are:

a. Lower data acquisition costs compared to travel surveys.
b. Higher accuracy in transportation statistics and performance indicators.
c. Higher resolution and more accurate data means better planning decisions.
d. Predictive and operational capabilities improve passenger experience for residents and citizens.
e. Better situational awareness to make more informed operational decisions.

## Transport Statistics

One of the most common use of MPD in transportation is the use of MPD in statistics or census-related use-cases. The type of statistics utilized are:

1. Population Density within communities and districts (based on cell-tower attribution of users).
2. People Movement Data between communities and districts (based on changes in cell-tower attribution of users).
3. Total PMT (Personal Miles Traveled), estimated based on the location data of users, with some allowable correction factors to account for lower ping rates.

MPD can be used in both Public Transport Statistics as well as Traffic Statistics. While public transport statistics are easier to gather through use of ticketing machines, turnstiles, passenger counting systems etc., use of MPD can add a lot of value to public transport statistics data. In addition, as cities move to ticket-free or virtual-ticket travel, MPD will form the primary source of public transport data. Traffic statistics on the other hand rely on expensive sensors that are placed across roadways to measure traffic, congestion, etc. Detailed O-D patterns are hard to arrive at using such traffic data. Metrics such as AADT are easier to calculate, where as metrics such as VMT are estimated based on several other inputs such as fuel use, toll use, etc. Use of MPD can help gain more accurate traffic statistics including breakdown by communities, user personas etc. Cities are already moving from collecting such information using travel surveys to relying completely on MPD due to its reliability and ease of collection.

Additionally, data and statistics related to active mobility (walking, biking, and other soft-mobility modes) were historically difficult to estimate due to the skew of population who are active. Use of MPD has been found very beneficial in calculating these statistics.

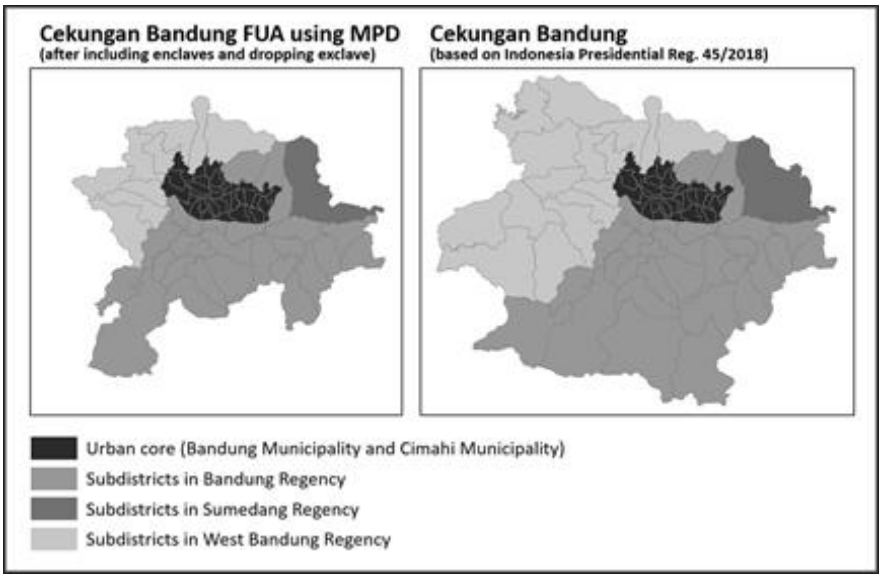
## Case Study: Replacing Indonesia's Household Travel Surveys

BPS-Statistics Indonesia has already produced commuting statistics using a conventional survey since 2014. However, due to the budget constraints on conducting surveys, commuting statistics can only be produced every 2 years in 5 metropolitan areas consecutively. The statistics are also limited to regency/municipality level (the second level of local area unit in Indonesia). In addition, it takes a long time to conduct conventional surveys, from designing it to disseminating the results. These limitations have driven BPS-Statistics Indonesia to begin exploring another source of data to produce official statistics. From experience of working with MPD, it has been shown that it is possible to produce statistics in a shorter time lag, at a lower level of administration area, and to increase the coverage without increasing the budget.

In 2019, BPS-Statistics Indonesia worked with the Indonesian Ministry of National Development Planning (BAPPENAS) to conduct a pilot project. The project explored the use of MPD to delineate metropolitan areas, with a case study in Cekungan Bandung, one of the metropolitan areas that have been delineated. Cekungan Bandung was used because the MPD results were initially planned to be compared with the last commuting survey when the project was conducted. The data itself was obtained from Telkomsel, one of the largest MNO in Indonesia, which is a state-owned private enterprise. There were 50,907 Telkomsel subscribers used as the sample within November 2019. To analyze commuter patterns using MPD, BPS-Statistics Indonesia developed algorithms to identify commuters and estimate commuting flows at sub-district level (the third level of local area unit in Indonesia). The commuting flows were used to measure the integration between the urban core and surrounding hinterlands in Cekungan Bandung.

BPS-Statistics Indonesia developed the algorithms and logic of processing by exploring raw data samples of MPD. After it is ensured that the algorithm work correctly in producing the expected output, Telkomsel then run the algorithm against all collected data. To verify the result of MPD algorithms, the validation process was carried out using some volunteers who were a subset of Telkomsel subscribers. The volunteers verified or annotated their movements through a travel diary survey. The algorithms results from the MPD were then compared with the results of the volunteers' annotations that used an application called MEILI. Therefore, there is a repetition of process from the Design phase until the Process phase in the statistical business process for the MPD to ensure data validity.

Furthermore, from the final aggregate data obtained from Telkomsel, BPS-Statistics Indonesia then analyzed it to determine which subdistricts included in Cekungan Bandung should be considered as part of its metropolitan area, by calculating their commuting rate. The delineation from MPD results were then compared to the delineation already determined by the Government of Indonesia, to identify which sub-districts were included in the metropolitan area by law but have a low rate of commuting flow. The results will be a recommendation for the Government of Indonesia to help determine a more appropriate delineation by using MPD, especially in Cekungan Bandung. In the future, similar projects will be carried out in other metropolitan area to improve the MPD algorithms and delineation thresholds. Figure below shows the visualization of MPD delineation and Indonesian Government delineation comparison on Cekungan Bandung metropolitan area in this study.




The difference between these two delineations depend on the commuting rate threshold used in MPD delineation. The higher the threshold, the fewer subdistricts are included. Based on defined threshold in this study, there were 40 subdistricts included in MPD delineation compared to 52 subdistricts from Indonesian Government delineation. The result suggests that the metropolitan area delineation determined by Indonesian Government on Cekungan Bandung has a lower commuting rate than the threshold. Therefore, it is important to identify the best commuting rate threshold to delineate metropolitan areas in Indonesia.

## Transport and Urban Planning

For transport and urban planning use-cases, in addition to transport statistics data, the following are essential:

1. Land-use, including POI types, zoning, affluence level, etc.
2. Origin-Destination Matrix including Modal Split, Trip Purpose and Departure Time
3. Dwell Time, Travel Time Index and other temporal factors.
4. Commuter attribution and personas, including permenant residents, temporarly residents, visitors, tourists etc.
5. Spatial activity analysis as well as seasonal and annual dynamics of people movement.

While MPD also has limitations and caveats, it does provide:

- **Passive data collection.** MPD does not require user interaction, and a larger sample size, so is likely to be less biased to some users.
- **Travel by time of day.** Ticket sales do not provide information on when people are travelling, and surveys are not granular enough to look at hourly trends. MPD allows analysis of peak travel times, how these change by mode and how they were affected by the pandemic.

- **Demographics.** Detail included in the MPD divided users by age, gender and socioeconomic status, and means impacts of policies on specific user groups could be considered.
- **Trip chains.** MPD allows considerations of total mobility across modes, and the behaviour of transport users who may change mode on longer journeys.
- **Additional modes and user groups.** MPD can, to a limited extent, begin to help understand the behaviour of electric vehicle users and international HGV drivers. This would be difficult with a roadside survey, as these users cannot be identified before the vehicles are stopped.
- **Timeliness and location.** Surveys must be planned in advance and traffic counters may not be in appropriate locations to measure the immediate impacts of national crises or short-term events. Mobile phone data can be available in near real-time to support rapid decision making for any location.

## Case Study: UK's Use of MPD to Understand Modal Usage

Since the pandemic, the UK Department for Transport (DfT) have continued to explore how MPD can support monitoring, evaluation and planning in areas where the ability to collect or survey evidence is limited. These have included:

- Understanding the demographics of users of the rail network, and how this varies by time of day and throughout the week.
- Understanding the mobility of international freight drivers within the UK and the impact they have on the movement of goods.
- Supporting national crises with real-time MPD to monitor whether the transport network was overloaded.
- In future, they hope to use MPD to understand users of Electric Vehicles, to better evaluate policies that facilitate their use.

The DfT routinely collect data on transport users through surveys, traffic counters and ticket sales, depending on mode. Long term and national trends are also collected through the National Travel Survey, which covers all modes, travel patterns and behaviours, but is much less frequent as a result. These provide some information on travel patterns and changes by mode, but are limited in some ways.

The DfT purchased data directly from a MNO, and each subsequent dataset currently requires an additional purchase. Different providers have been explored but they have largely maintained a relationship with one provider to ensure consistency of data and coverage. The MPD procured is based on mobile event data. Mobile phones generate "events" as they communicate with the operator's national cell network. These are collected on an anonymised basis and processed by the operator before we receive the data. The operators so far used each have market shares of 20-30%, but processing involves scaling this up to be representative of the whole population.

The final datasets received by DfT vary in granularity but are typically provided for zones within local authority boundaries and at a daily or hourly level. The time periods covered have been

specific to the questions they are using the data to answer, usually covering 2-3 months, though near real-time data providing information for 5-minute intervals with around 5 minutes latency over the period of a week have also been accessed. The MPD is processed, anonymised and aggregated by the provider before it reaches the department, hence no sensitive data is held by the DfT. This imposes restrictions on the structure and granularity of the datasets available, as the information must comply with data protection laws and small numbers of trips in very granular data cannot be provided. However, this means no additional privacy controls are needed within DfT beyond standard data handling and storage policies. It also means that datasets are ready to be used in modelling, analysis and visualisation to support decision makers.

Comparability over time was ensured, where needed, by procuring consistently formatted and processed datasets from the same mobile network operator each time. Where changes were implemented or made, an evaluation of the difference was conducted to understand the impact on reporting.

To ensure quality, the DfT compared the MPD to existing published and internal statistics. Further, they did not use the MPD in isolation, but typically combined the datasets with other statistics such as census data. The exercise has shown that MPD largely show similar trends to other sources, and where differences do exist these have been explainable by the collection methods used and caveats associated with one or more of the sources. But it was noted that the MPD is less well suited to measuring smaller trips, for example local walks.

The data are not being used to produce official statistics, but instead are being used as part of a suite of evidence and analysis to support policy decisions and internal reporting. While it is unlikely to become part of official statistics production, largely due to the costs involved and processes that would need to change, it is used increasingly to support analysis on the evaluation of policies across multiple modes. Feedback so far has been positive, and any negative comments are generally focussed on reliability and coverage where the mobile phone data cannot be compared to other sources (e.g., for Electric Vehicles, as there are very few other sources to compare to).

## Transit & Public Transport Planning

One of the main challenges in planning transit or public transport is to understand the travel behavior of non-public transport user. Cities typically have enough data to optimize their services for public transport users through data from ticketing and passenger counting services. However, adapting the service to a wider audience, or having tailored services for other personas have traditionally been challenging. Multiple cities have shown potential to close this gap by using MPD data.

### Case Study: Tracking Rail Passengers in Austrian Railways

Austrian Railways (OBB) in cooperation with Invenium focused on passenger flow analysis on to improve their internal planning capabilities relating to passenger demand. The project has been built around six different use cases, namely:

- The number of passengers entering and exiting at each station
- Railway station-based origin-destination matrices, excluding transfer passengers
- The loads of passengers on specific sections of the network
- Analysis for management of delays
- Origin-Destination matrices as input to the Austrian National Transport Demand Model
- To ascertain the real catchment areas of stations.

In addition to these use cases, the project has considered some special analyses such as looking at demand peaks.While incomplete passenger count data already existed alongside ticketing information, the MPD was considered as adding value in this task due to its comprehensiveness (covering every station) and timeliness (data were available on a daily basis).

The project has involved a partnership between the A1 Telkom MNO (with around 38% of the mobile phone market in Austria), Invenium and OBB Infrastruktur. A1 provide the already-anonymized MPD (so there are no additional privacy preserving techniques to action), Invenium operates and enhances the algorithm platform (which splits signals between stationary and moving devices, and then assigns an estimated transport mode), and OBB provides the actual train timetable (trains run, not trains scheduled) on a daily basis using their Advanced Railway Automation Management Information System (ARAMIS).

## Case Study: Seoul's Use of MPD in Planning Night Bus Service

Because the metro system in Seoul closed from midnight to 5:00 am, the only option for many commuters was to take a taxi, which is expensive. Night taxis were also notoriously difficult to catch, putting mostly many at risk on the road. Needless to say, this problem disproportionately affected the socially and economically disadvantaged population. The solution was to introduce a night time bus service. However, designing optimal routes for the night bus service proved difficult since daytime traffic data could be misleading as commuter behavior at night wouldn't necessarily be that at night.

Seoul reverted to use of MPD data in detecting commuter behavior at night. Specifically CDRs were used to understand origin-destination patterns. Without any physical data collection efforts, the city used MPD and taxi trips data to design the late-night bus routes. Nine late-night bus routes (so called owl bus routes) were designed, and they are currently operating between midnight and five o'clock in the morning.

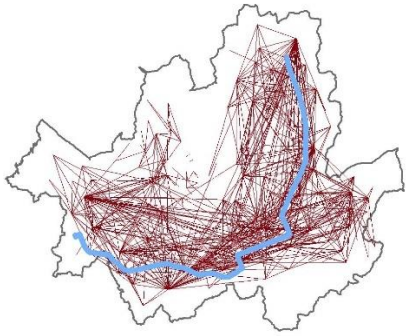


Figure showing travel patterns traced in Seoul for owl bus service.

# Traffic Analysis

Cities and jurisdictions have started using MPD in traffic analysis. Traditionally, traffic data has been collected by traffic counters and other spot sensors. This can give unlinked traffic data at multiple locations within the transportation network. However, to develop linkage, cities have used bluetooth sensors, APNR readers etc. While this provides trajectory traces, supplementary data from travel surveys were relied on, to understand the complete travel matrix, including travel time, origin-destination etc. Use of MPD have been found to supplement this further owing to the fact that it carries people movement information within and across jurisdictions.

## Case Study: Poland's Use of MPD in Traffic Analysis

In 2021 the Poland General Directorate for National Roads and Highways (GDDKIA) conducted a tender to purchase data from mobile carriers, but it did not proceed due to the prohibitively high cost.

The data in question would not have been used to calculate traffic intensity, because they don't cover the entire traffic on a road section. However, they may have offered significant value in terms of origin-destination matrices by day, hour etc. Moreover, trip purpose (work, education etc) could be derived, by analysing patterns in data and combining it with location data.

The tender planned to obtain information on the movement of SIM cards for the entire country, divided into particular poviats (local administrative units) and border crossings, so as to derive averages for each day of the week.

Planned scope of results:

- The size of the survey sample, divided into poviats.
- A sample extension coefficent to extend the obtained data samples to the entire population.
- The number of trips, both internal within each poviat and between poviats.
- Daily and aggregated travel matrices between poviats and country borders, for each day of the survey.
- Matrices with average number of travels between poviats and country borders, for each day of the week (days Tuesday-Thursday calculated together)

According to the MNO, it would have been impossible to divide travels into road, train or airplane traffic using these data. The importance of filtering data to exclude situations when the SIM card roams to another tower, but in fact doesn't change its location, was noted.

# Operational Decision Support

Real-time or near-real-time MPD can be used to understand operational demand on transportation networks to make traffic management, public transport operations and crisis management decisions. Typically, cities are using data from its own sensors, but mobility companies such as Google, TomTom and INRIX monetize on real passenger demand data for operational use. TomTom relies on GPS measurement from its users (similar to Google) to

provide data to traffic management centers along with additional decision-support tools. INRIX relies on commercial vehicle fleet data and smart vehicles to gather realtime GPS measurements (location, speed and heading). Other companies such as Moovit uses MPD to provide public transport delay information to passengers based on data mined from its own users.

### Case Study: To be Added

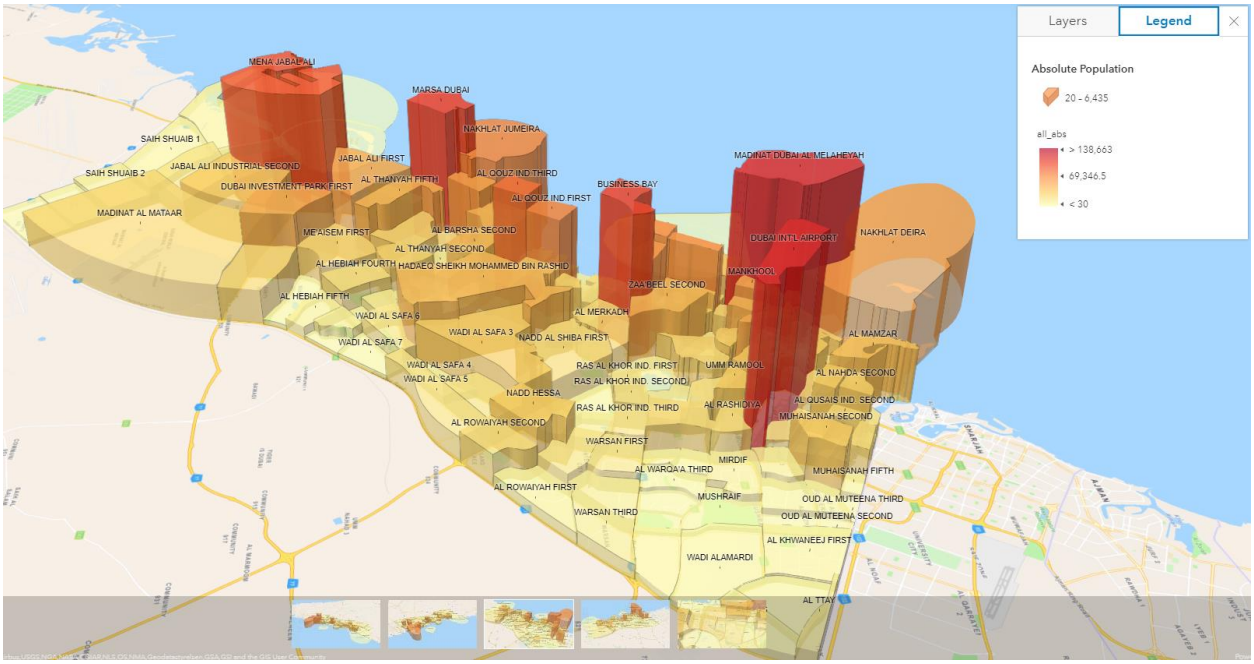
<Add TFL Use-case>

## Events and Crises Management

Realtime and Historic data (both population density and population movement) are also being used for events and crises management functions. Recurring events are being planned based on historic behavior of visitors during their last visit, and real-time data are being used for better management of events. During crises, MPD data is also used to understand the number of captive population between crises areas for better response plans.

### Case Study: Dubai's Use of MPD in Events Management

Dubai's Roads and Transport Authority uses MPD in its planning and operations. With access to population density data at 15-minute resolution, RTA makes real-time decisions to support major events such as EXPO 2020 in deploying additional passenger transport vehicles such as buses and taxis. The data used is anonymized passive signaling data from microcell towers placed across the city.



## Traffic & Transit Demand Prediction

Once real-time MPD is used in operational decision-making, advancements can be made to use probe-mobile data in generating rolling forecasts on operational parameters such as link travel time, public transport demand etc.

# Conclusions

In conclusions, MPD is being used around multiple use-cases across the transport analytics domain to understand multimodal passenger movements. While there are several types and subtypes of mobile phone data, agencies must use the correct data type based on their intended use. It is also important to understand the type and extent of processing required prior to using MPD for the range of applications of interest, ranging from statistical reporting, planning and analysis, operational support use and even predictive use-cases. MPDs are known to reduce resource requirement in data collection, increased data coverage and reduced biases.

# References