International Transport Forum

Big Data for Travel Demand Modelling

Summary and Conclusions

186 Roundtable

Big Data for Travel Demand Modelling

Summary and Conclusions

186 Roundtable

The International Transport Forum

The International Transport Forum is an intergovernmental organisation with 63 member countries. It acts as a think tank for transport policy and organises the Annual Summit of transport ministers. ITF is the only global body that covers all transport modes. The ITF is politically autonomous and administratively integrated with the OECD.

The ITF works for transport policies that improve peoples' lives. Our mission is to foster a deeper understanding of the role of transport in economic growth, environmental sustainability and social inclusion and to raise the public profile of transport policy.

The ITF organises global dialogue for better transport. We act as a platform for discussion and pre-negotiation of policy issues across all transport modes. We analyse trends, share knowledge and promote exchange among transport decision makers and civil society. The ITF's Annual Summit is the world's largest gathering of transport ministers and the leading global platform for dialogue on transport policy.

The Members of the Forum are: Albania, Armenia, Argentina, Australia, Austria, Azerbaijan, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Canada, Chile, China (People's Republic of), Colombia, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, India, Ireland, Israel, Italy, Japan, Kazakhstan, Korea, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Mexico, Republic of Moldova, Mongolia, Montenegro, Morocco, the Netherlands, New Zealand, North Macedonia, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Tunisia, Turkey, Ukraine, the United Arab Emirates, the United Kingdom, the United States and Uzbekistan.

International Transport Forum
2 rue André Pascal
F-75775 Paris Cedex 16
contact@itf-oecd.org
www.itf-oecd.org

ITF Roundtables

ITF Roundtables bring together international experts to discuss specific topics notably on economic and regulatory aspects of transport policies in ITF member countries. Findings of ITF Roundtables are published in a Summary and Conclusions paper. Any findings, interpretations and conclusions expressed herein are those of the authors and do not necessarily reflect the views of the International Transport Forum or the OECD. Neither the OECD, ITF nor the authors guarantee the accuracy of any data or other information contained in this publication and accept no responsibility whatsoever for any consequence of their use. This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

Cite this work as: ITF (2021), *Big Data for Travel Demand Modelling: Summary and Conclusions*, ITF Roundtable Reports, No. 186, OECD Publishing, Paris.

Acknowledgements

Eric Jeannière and Alexandre Santacreu of the International Transport Forum (ITF) authored this report. The report's content is based on discussions that took place during the ITF Roundtable entitled Big Data and Transport Models, held by videoconference on 14-16 December 2020. The ITF Secretariat thanks the 37 participants to the Roundtable, who represented 28 different organisations in 14 ITF member countries. Annex B provides a list of all Roundtable participants.

This report also builds on earlier research conducted by the ITF Working Group on Big Data. The authors of the current report would like to thank those who participated in that Working Group for their insights. They include ITF member countries Austria, Canada, Finland, France, Greece, Italy, the Netherlands, Norway, Serbia, the United Kingdom, the United States, and the United Nations Economic Commission for Europe (UNECE). Annex A contains case studies produced by this Working Group.

The ITF Secretariat would like to thank Patricia Hu, Director of the Bureau of Transportation Statistics at the US Department for Transport, for chairing both the Big Data Working Group and the Roundtable on Big Data and Transport Models.

Credits also go to the authors of four discussion papers that provided core material to the present summary report: Patrick Bonnel of the University of Lyon, Norbert Brändle of the Austrian Institute of Technology, Imane Essadeq and Thibault Janik of Systra, and Luis Willumsen of Nommon Solutions and Technologies. Their discussion papers and all other content from the Big Data and Transport Models Roundtable are available online at https://www.itf-oecd.org/big-data-transport-models-roundtable.

The authors are grateful to the following people who provided guidance or material when preparing the Roundtable: Jean Coldefy (Atec ITS France), Damien Verry (Cerema), Antonio Masegosa and Enrique Onieva (Deusto University), Ian Knowles (United Kingdom Departement for Transport), Caroline Almeras (European Conference of Transport Research Institutes), Sergio Fernández Balaguer (Municipal Transport Company of Madrid), Lewis Dijkstra (European Commission), Patrick Gendre (independent), Dominic Paulo and Wolfgang Mühlbauer (INRIX), Dietmar Offenhuber (Northeastern University, Boston), Cristina Pronello (Politecnico, Torino), Alessandro Attanasi (PTV), Markus Friedrich (Institute for Roads and Transportation, Stuttgart University), Joan Roca (Telefónica S.A.), Marianne Stølan Rostoft (Norwegian Centre for Transport Research), Dmitry Pavlyuk (Transport and Telecommunication Institute), Maarten Vanhoof (University College London), Teresa Brell (Umlaut), Ronald Jansen (UN Big Data Global Working Group) and Thomas Deloison (World Business Council for Sustainable Development).

Further acknowledgements go to Lucie Kirstein (ITF) who summarised the Big Data Working Group proceedings, and Luis Willumsen and Aman Chitkara for their manifold and constructive comments during the Roundtable and their thorough review of this report. Final thanks go to Hilary Gaboriau (ITF) who copy-edited the report.

Table of contents

List of abbreviations

AI	Artificial intelligence
API	Application programming interface
CDR	Call detail record
C-ITS	Co-operative Intelligent Transport System
GDPR	General Data Protection Regulation
GSM	Global System for Mobile Communications
GTFS	General Transit Feed Specification
IoT	Internet of Things
MaaS	Mobility-as-a-Service
MNO	Mobile network operator
OBD	On-board diagnostic
OD	Origin-destination
OEM	Original equipment manufacturer
OS	Operating system
RFID	Radio-frequency identification
SDG	Sustainable development goal
SRTI	Safety-related traffic information
TMC	Traffic management centre

Executive summary

What we did

This report examines how big data from mobile phones and other sources can help forecast travel demand. It identifies the strengths and potential use cases for big data in transport modelling and mobility analysis. The report presents ways to address potential biases, commercial sensitivities and privacy threats and offers recommendations for governance arrangements that make data sharing easier.

The report summarises the findings of an ITF Roundtable in December 2020 that brought together experts from 28 different organisations in 14 countries. It draws from four discussion papers that provide practical illustrations of the use of big data from mobile signals, smartphones and smart cards in transport planning.

The report also builds on earlier research conducted by an ITF Working Group on Big Data. Insights for that work were provided by ITF member countries Austria, Canada, Finland, France, Greece, Italy, the Netherlands, Norway, Serbia, the United Kingdom, the United States, and the United Nations Economic Commission for Europe (UNECE). Members of the Working Group submitted case studies on the uses of big data and the benefit of partnerships between the public and the private sectors. The research emphasised how to establish such partnerships and the benefits of big data to national statistics offices. Case studies included the use of data from ticketing, insurance telematics, logistics telematics and other sources and revealed the value of data fusion across different sources.

What we found

Transport planners use big data from mobile network operators, smartphone apps and smart cards to complement traditional travel surveys. The new data sources help transport planners understand and forecast travel demand by estimating how many trips people make, to which destinations they travel and which mode or combination of modes they use. How public authorities use such data varies significantly between countries, but experience suggests that big data could substantially improve transport planning.

Mobile network data are particularly well suited for the analysis of medium- to long-distance trips. Smartphone app data offer higher spatial accuracy than mobile network data and are more likely to infer transport mode, capture short-distance trips and detect short activities. Both data sources offer nearly door-to-door tracking regardless of the transport mode, unlike vehicle telematics and smart-card data.

Compared to traditional data sources, the promptness of big data narrows the gap between data collection and the forecast horizon. It also facilitates the analysis of emerging trends and makes it possible to learn from quasi-experiments and exceptional situations such as the one created by the Covid-19 pandemic. Another positive aspect of big data is its high sample size and continuous sampling that captures seasonal and weekly variability.

The private sector collects much of the big data relevant to transport planners. Access to this data would allow governments to improve important planning tasks and reduce spending on travel surveys.

Establishing the right partnerships between the private and the public sectors is thus critical, as is data protection, often cited as the main barrier to data sharing.

Introducing technical standards could reduce the complexity and cost of sharing big data. Standards would make it easier to merge data from several mobile network operators: for instance, where larger data sets could address biases found in the data coming from a single network operator. However, standards could also hinder innovation in a rapidly changing context where data collection depends on heterogeneous technologies across the industry.

Big data can complement traditional travel surveys but it cannot replace them yet. The analysis of vast amounts of data can provide answers to new questions in transport planning but often does not capture valuable socio-demographics information collected via household travel surveys. Typically, big data must be fused with other sources to compensate for missing information and correct biases. A major concern with big data is whether it is representative: it is poorly documented, so it is difficult to assess.

What we recommend

Collect data only for defined purposes and only the minimum required

The public sector should follow two principles in its transport planning mission to protect private and commercial data: purpose specificity and data minimisation. Purpose specificity ensures that data is collected solely for a precise regulatory task. Data minimisation gives preference to the lightest possible data collection mechanism.

Develop guidelines for the use of big data in transport models

Authorities should periodically revise transport modelling guidelines together with the transport planning community. Creating such guidelines provides an opportunity to reflect on data needs in the rapidly changing landscape of mobility. They should encourage more use of big data where relevant, make users aware of potential biases and prevent misuse. The guidelines should help transport planners define the technical specifications for processing big data. They should be country-specific to account for local survey data formats and legal requirements.

Enable the collection of location data through smartphone apps

Smartphone operating systems should continue letting individuals control which location data and identifier data are shared with each app. Operating systems should continue providing an accurate and steady stream of position data – of tremendous value to transport planners – to the apps that secured user consent for them. Examples of such apps include on-demand transport and ticketing apps, dedicated travel survey apps, journey planners, and apps trading mobility data against a free service.

Protect privacy through multiple solutions

Data protection should not serve as an excuse for not sharing relevant data or for destroying it. Stakeholders should acquire and implement techniques to make data sharing compatible with privacy concerns, such as data aggregation, pseudonymisation, encryption and privacy impact assessments. Partners in a data exchange can involve a trusted third party to facilitate data sharing between them as an “honest broker”. The “safe answers” approach, whereby partners exchange only query results instead of raw data, provides another option.

Define a roadmap for household travel surveys

Authorities and transport planners should continue collecting socio-demographic, behavioural and attitudinal information from traditional household travel surveys. Such disaggregated survey data are essential to validate and recalibrate the protocols that turn big data into accurate mobility data. However, they should develop a roadmap for maintaining and further improving the quality of household travel surveys. Such a roadmap should acknowledge declining response rates and reflect the potential for data fusion with sources such as ticketing, mobile phone signal and smartphone app data.

Design and test smartphone-assisted household travel surveys

Future household travel surveys should also leverage the power of mobile devices. Survey respondents could use dedicated smartphone apps to track and annotate their trips. Apps using motion sensors and precise geolocation services could capture trip details and reduce the burden of maintaining trip diaries for travel survey respondents.

Leverage artificial intelligence for data mining

Research funding should support the development of algorithms, including artificial intelligence (AI) solutions, that infer trip details from mobile phone signals and smartphone sensors. Creating annotated reference data sets would be an important element of this policy. It would support the training of computer models for trip detection and transport mode detection. Universities could organise competitions to make AI applications more transparent and create buy-in. The University of Sussex-Huawei Locomotion Challenge, for instance, invites researchers to identify eight modes of transport based on smartphones' GPS, Wi-Fi and mobile network data.

Create and promote a recognised data steward function in the public and private sectors

Both public and private organisations should designate individuals or teams who proactively initiate, facilitate and co-ordinate data-sharing partnerships. These "data stewards" will initiate pilots, scale them up, promote the exchange and reuse of data in the public interest, and protect potentially sensitive information.

Invest in the data-related training of the public-sector workforce

Governments should recruit big data specialists for transport planning and train public-sector workers in relevant skills. Data governance, the privacy dimension and technical aspects will require more expertise. Also important are skills to assess data quality and determine whether it is adequate for a specific purpose.

The challenging nature of big data

Today's digital revolution generates an unprecedented amount of data that, if used effectively, can inform policy making and deliver better public services (EC, 2020). The public sector has been collecting data on the movement of people and goods for some time already and uses that information to help plan for changes in the transport system. Conventional data collection methods come with a number of limitations. They include high cost, limited sample size, respondent fatigue and non-response. Should big data complement or replace traditional data collection methods? To answer this question, one must take stock of big data's strengths and weaknesses.

Big data offers new opportunities in comparison with traditional survey methods. Its timeliness makes it well suited to a fast-changing mobility landscape. Big data offers a passive data collection method, eliminating the need to survey the population and rely on their sometimes unreliable answers about the trips they make. It tracks travel behaviour at times when unplanned events – a union strike, a natural disaster, or, as the transport world saw recently, a pandemic – unfold and affect the transport system. There are very recent and pertinent examples of big data's use in assessing travel. Public authorities have monitored the effect of Covid-19 on travel thanks to mobile phone data collected by the UK DfT (2021), the US BTS (2021) and Google (2021), and from journey planner requests collected by Apple (2020).

However, big data provides much less control over key aspects of data collection compared to traditional survey methods, resulting in a number of negative consequences:

- Big data from location-based services suffers from inherent representativeness biases due to the lack of a controlled sampling framework. Some age or social groups might have a greater tendency to use a specific mobile network operator, smartphone app or connected vehicle. Some trip purposes and transport modes may attract a more intense use of navigation apps.
- Most big data on mobility is missing the socio-demographic attributes of individuals, preventing researchers from drawing insights that are meaningful and statistically sound. The ITF Working Group on Big Data determined that big data alone would likely be insufficient in developing a robust prediction model, such as a transport demand model.
- Analysts must intervene to detect and remove outliers picked up by big data. For example, devices like shared e-scooters that are part of the Internet of Things (IoT) send and receive data over the mobile network, but they are not carried by people throughout the day like mobile phones are. However, their signal can be caught by nearby antennas creating "noise" in the data that analysts must identify and remove.
- The collection of big data is not always repeatable in a way that offers comparability. This could be due to changes in technology, in the market penetration of mobile devices, and in the market shares of various service providers. The problem could also occur when private sector companies change priorities and stop sharing data.

The public sector must accept that big data does not replace traditional data sources; it will not be able to answer the same research questions. However, big data could provide answers to new questions in a

timely manner. Conversely, more conventional sources are often essential to complement and validate the information derived from big data.

The seven characteristics of big data

The term *big data* was coined in the late 1990s to refer to the growing volumes of stored data. Laney (2001) defined the commonly known "three Vs" of big data as volume, velocity and variety. The ITF Big Data Working Group revealed a further four characteristics of big data. This report, therefore, proposes a list of seven characteristics or "seven Vs":

- Volume: Increased storage capacity and decreased storage costs (e.g. through cloud storage) have facilitated a trend towards the collection and handling of higher volumes of data.
- Velocity: Technology such as radio-frequency identification (RFID) tags, sensors, smart metering and ubiquitous connectivity are driving the need to handle torrents of data in a timely manner.
- Variety: Data come in all formats, either structured, with defined metadata fitting easily into relational databases, or unstructured, such as IoT data, Twitter feeds, media and text files, communications, search items, etc.
- Variability: Periodic peaks, such as trending topics on social media or event-triggered data, can make data flows irregular or inconsistent.
- Veracity: This aspect refers to the trustworthiness of the data. Large discrepancies might exist in data collected from multiple sources.
- Value: A key issue is the value proposition for the use of big data analytics, i.e. what can be learned from it, what traditional data collection or estimation can be replaced, and how. One needs to factor in the time needed for data cleaning and analysis.
- Visualisation: Big data analytics may be of little direct value to decision makers without effective visualisation tools to capture the richness of the information available in a form that is succinct but accessible and resilient to misinterpretation.

The characterisation of big data raises several points that affect data quality and usability by government in a regulatory and planning role. The next section assesses data quality across seven dimensions.

The seven dimensions of data quality assessment

The widely accepted benchmarks for data quality presented in the "Quality Framework and Guidelines for OECD Statistical Activities" are relevance, accuracy, credibility, timeliness, accessibility, interpretability and coherence (OECD, 2012).

Relevance

Relevance refers to a qualitative evaluation of how well the data respond to the objectives of the study. In other words, can data answer the questions posed (OECD, 2012)? Some researchers believe this quality benchmark is the biggest obstacle in the field of big data because the users and producers of the data are, in most cases, no longer a part of the same organisation (Merino et al., 2016). Since the primary producer of big data is the private sector, those data are usually not collected with the intent of providing

comprehensive evidence to support policy decisions. Indeed, most big data is collected for purposes different from providing information about mobility.

Accuracy

The data's ability to estimate values from a traditional statistical perspective, often quantified by terms of error, determines its accuracy. Big data eliminates a number of sources of error, such as human errors, recall bias and respondent fatigue. However, it introduces new sources of errors due to outliers and the lack of representative sampling. Confirming statistical validity in traditional surveys is usually a standardised procedure with minimal cost. With big data, there is an issue with the precision and level of error of the raw data and the accuracy of the processed mobility indicators: for example, aggregate origin-destination (OD) trip matrices. The processing of the *raw* big data is critical to the elimination of or compensation for errors and the production of accurate results. It is the validity of these results that can be tested and validated.

Credibility

The reputation and scientific rigour of the data's producer largely determines the data's objectivity and quality. This criterion can be difficult to satisfy when working with big data from the private sector; most of these data producers tend not to have the same scientific reputation as governments and research institutions.

Timeliness

For data to be pertinent, their information must not be obsolete or irrelevant by the time the study is published. Big data has led to the production of massive amounts of information in a short time. Whereas data collection is often the time-consuming phase of traditional research methods, the cleaning and analysis of big data – especially when it lacks structure and consistency – is where most time is lost.

Some statistical authorities see a challenge in the use of big data, where the near-daily refresh rate is much quicker than the statistical publication process; they believe big data becomes redundant more quickly. In fact, it may well be that the tremendous strength of big data lies therein and will likely stimulate new ways of producing and disseminating transport statistics.

Accessibility

Accessibility is how quickly the researcher can access the data and metadata. It is also the cost of that data. This criterion should be adapted for big data to include the cleaning and processing costs, which will often consume most of the budget for a big data study. The key challenge for researchers working with big data that they did not produce will be accessing thorough and consistent metadata; data producers have varying methods of data documentation and many require that collection methods and metadata remain confidential (Liu et al. 2016). However, "transparency regarding the nature of data and the conditions under which it was collected is crucial for data-driven transport policy making" (ITF, 2015: 18).

Interpretability

Has the study adequately defined the dimensions of the data it requires? This includes definitions of variables, concepts used in the data collection process and descriptions of the target population. If the researcher is unable to understand these critical factors, the study will be unable to produce meaningful

results. The challenge with satisfying the interpretability benchmark when using private-sector big data is that researchers generally have less control over the conditions of the study and the variables than with traditional statistical surveys.

Even the producers of big data do not always have the freedom to define and choose the variables. In particular, big data often lacks socio-demographic information on individuals, which diminishes the meaningfulness of the results it produces. This also makes findings very difficult to generalise or interpret since the sample population is unknown (Liu et al., 2016).

Coherence

Terms and methodology must remain logically consistent within the data, across multiple data sets, over time, and across countries (if applicable). Unfortunately, some studies have shown that inconsistency is an inherent trait of big data due to a lack of scientific rigour in big data collection methods and the complexities of establishing such methods given the extreme fluctuations in the inputs and selection of the sample population (Liu et al., 2016).

Consistency is another essential standard for quality data, particularly if the data are to be used by official statistics organisations. Surveys compare changes in indices over the years and give insight into how countries develop over time, a practice that requires extremely reliable time-series data. The fact that big data collection methods tend to change often can be problematic. Data providers are aware of this and fine-tune their algorithms to process the *raw* big data in the most sensible way. It is the role of transport planners to conduct the validation, as frequently as needed, of the resulting big data products against alternative sources.

Table 1. Applying OECD data quality standards to big data

Dimension	Big data benefits	Big data challenges
Relevance	Big data allows researchers to study subjects that previously could only be theorised through extrapolation.	Big data tends to lack information needed to effectively respond to some research objectives (variables of interest, socio-demographic variables, metadata, etc.).
Accuracy	Data are more precise and detailed than ever before and human errors in data inputs no longer exist.	Evaluations of the statistical validity of the results have to be adapted to the particular case of big data processing.
Credibility	---	The processing of most big data to reduce noise and bias and thus produce useful mobility indicators often lacks transparency and results in outputs of variable credibility. Rigorous validation should be applied to outputs.
Timeliness	Data are collected at faster rates than ever before.	Big data appears to become out-of-date faster than ever before, a challenge that is inspiring great innovation in analytic methods.
Accessibility	Big data can be quickly or even instantaneously collected and accessed in certain cases.	Documentation on data collection and processing methods can be confidential and costly to obtain, assuming proper documentation has been recorded.

Interpretability	In cases when the necessary variables are accessible, the level of detail of big data can facilitate interpretations and allow a more precise understanding of the population of interest.	Big data from devices or social media is often missing information on the sample population and can lack the socio-demographic variables needed to generate sound statistical extrapolations from the data. In addition, it can be difficult to find data with the variables needed to guide policy decisions, as there is often less control over what information is collected.
Coherence	A number of data providers could build on their global presence to produce internationally comparable transport statistics. Such providers include Apple, Google, traffic information companies, social media companies, and the data aggregators using the same software development kits (SDKs) in smartphone apps.	There is often a lack of procedures in place to ensure the consistent collection and documentation of big data. In many cases, the complication of setting up such procedures is due to fluctuations particular to big data.

At first glance, it might seem dangerous to accept lower levels of data quality. However, traditional survey data are not free from certain biases and limits to their validity, and they is nevertheless used as evidence to advise policy decisions. Whether the information comes from surveys or big data, inputs are rarely exact and unbiased. Analysts must simply be realistic and seek to minimise these problems.

As highlighted in ITF (2016), the particular skill sets to understand, format, clean, deconstruct and analyse large, unstructured and real-time data are not typically present in the public sector. In addition, public authorities compete with the private sector for data scientists, including statisticians. These highly remunerated activities can put further strain on public budgets, which might raise the question of value for money of using big data in a government context. Governments should be aware of this and consider investing in the relevant skills or procuring dashboard-type solutions rather than raw data.

New data sources for transport modellers

Current trends are shaping a future with ubiquitous connectivity. Smartphone penetration is near 80% in almost every developed country, and more than three billion people use mobile internet globally. The general public is widely adopting location services for navigation, identifying nearby services, the weather forecast, social networking and online dating. Vehicles and wearable technologies such as health monitors also come with location technologies.

Location services become more accurate as the world becomes more connected, as every fixed Wi-Fi access point serves as a location beacon. A growing number of novel location-sensing technologies and global satellite positioning systems contributes to the increased accuracy of location services.

Sensors are increasingly embedded in mobile devices, vehicles and public spaces. As the connected network of objects expands, the amount of data it generates grows exponentially. Opportunities exist for big data to contribute to a better understanding of transport demand in particular and transport systems in general.

Transport data can support real-time applications through incident detection, congestion management and route guidance. However, the Big Data and Transport Models Roundtable focussed on transport data that support planning applications, primarily related to travel demand forecast. This report prioritises solutions to procure trip matrices, also called origin-destination or OD matrices. It examines whether big data can provide OD matrices by mode and by socio-demographic groups.

This section highlights existing solutions that derive OD matrices from big data. It examines questions of data quality, bias and privacy protection, leading to some technical recommendations.

Sources of big data for transport planning

Roundtable participants examined the most relevant use cases for big data in transport planning. In particular, they explored the possibility of big data to complement or replace traditional travel surveys, reflecting how many trips people make, to which destinations, and using which mode or combination of modes.

Participants looked at those sources of big data that seem most relevant to estimate passenger volumes between origins and destinations: mobile phone network data, smartphone data, vehicle data, ticketing and smart-card data.

Data from mobile network operators

Mobile network operators (MNOs) collect call detail records (CDR) for customer billing purposes. CDRs log the connections between the telecommunication network and any mobile device. Mobile phones are the most common of these devices, but CDRs also collect information from other connected devices equipped with a SIM card, such as shared bikes, smartwatches or tablets.

Originally limited to phone calls and text messages, CDRs now include the transmission and reception of any data to or from the device, including emails, internet traffic and app data. Some MNOs also log operational events, such as the handover of one mobile device from one cell to another, that facilitate the location of a device[1]. To estimate trip volumes from MNO data, the core concept consists of detecting periods of time when a device is stationary (with a dwell time above a predefined threshold). A trip is inferred whenever a device leaves a location where it was stationary and is later found stationary in another location. Depending on the algorithm, the raw data can be analysed to identify a path, a transport mode, a break or a transport change in particular locations (a train station or other transport hubs) allowing to define different legs for a trip.

Roundtable participants reported that there are at least two sources of trip matrices from mobile network data. Some MNOs are able to process their own raw data and produce trip matrices. On the other hand, there are independent data analytics companies that have developed transport planning expertise and partnered with MNOs to process their raw data to produce trip matrices. In general, providers offer little transparency or documentation on their algorithms. For greater control and transparency, transport planners can develop partnerships with MNOs and run queries directly on the raw data they securely host. Third parties such as transport consultancies can facilitate such partnerships.

Spatial accuracy

MNOs define a cell as the area covered by a given antenna.[2] Usually, the size of a cell depends on population density, with its radius ranging from 100 m to 10 km (ITF (2015), Box 8). The dynamic and overlapping nature of cell boundaries further reduces the accuracy of cell-based positioning. Mobile devices may not necessarily communicate with the nearest antenna. Obstacles may have weakened that antenna's signal or peaks in demand may have saturated it. As a result, the mobile device will connect with a more distant antenna that is less encumbered. These more distant connections are known as "cell jumps" or "phantom trips".

Triangulation or other such techniques could refine device positioning from MNO data. However, MNOs rarely undertake such computationally intensive tasks. The most realistic ways for position accuracy to improve over the next ten years are through a denser network of antennas and the adoption of shorter-range 5G antennas. For now, common 3G/4G networks seem best suited to the needs of transport planners developing national models for inter-city travel based on a coarse zone system.

Data from smartphone apps

For service-related or pure monetisation purposes, smartphone apps often request access to the location data of a device, as provided by the smartphone operating system (OS). When a user grants an app limited access to location data, e.g. only when the app is used, the app collects very intermittent traces of this particular user. Yet even patchy data become valuable when accumulated across several apps and linked through a marketing identifier that is unique to a device. Privacy considerations have called into question the future of this marketing identifier.

In exchange for free apps and services, a number of smartphone users consent to their position being tracked and monetised. Data aggregators primarily monetise this information for marketing purposes, but can also partner with transport consultancies and contribute to the estimation of trip data. Umlaut (2020) collects data from thousands of apps worldwide to provide transport planning services.

Roundtable participants reported that population sample sizes of smartphone app data seem smaller than those of MNO data. Kisio (2020) reports a sample of four million smartphone users in France, for instance. StreetLight and Cuebiq report over 30 million devices in their MNO data set (Cuebiq, 2016) and roughly

12% of the adult population in the United States (Streetlight Data, 2017). For the vast majority of transport planning applications, however, the representativeness of the sample is the point that deserves the most attention. One risk is that apps are often specific to certain types of users, resulting in data that misrepresent the population as a whole. The "Data quality and potential biases" section below discusses whether MNO and smartphone app samples are representative of the real population.

Smartphones include sensors such as accelerometers, compasses and gyroscopes that can help identify a mode of transport when the device is in motion. This is particularly useful in mobility survey apps, assuming the user has consented to position tracking and the device allows for higher battery use, in order to prefill any trip diary (Brändle, 2021).

Spatial accuracy

Smartphone location data collected through apps are particularly accurate. They are fed mainly through a GPS signal with a range of precision of 10 to 50 metres (ITF, 2015). Where buildings obscure GPS signals, nearby Wi-Fi access points serve as beacons and deliver similar location accuracy.

The spatial accuracy of smartphone app data is so high that it generates particularly high commercial interest: it can reveal which billboard someone could see on the road towards which retail unit. A vast business ecosystem has developed using smartphone location data for advertisement purposes. However, the spatial accuracy of smartphone app data makes it particularly sensitive in terms of privacy protection (Thompson and Warzel, 2019).

Data from vehicles

The majority of drivers have access to GPS positioning and navigation services, supplied by either smartphone apps (e.g. Waze), in-car displays (e.g. TomTom) or fleet telematics (e.g. Geotab). Navigation service providers collect trip details, including origin and destination, start and end times, waypoints, and travel time along each road segment. Some monetise OD matrix estimates and other transport planning data, offering a disaggregation between passenger cars, local commercial fleets and long-haul trucks.

Estimating a real OD matrix from a sample of users requires a careful approach since several biases may exist: the user base may not be representative (e.g. bias towards recent or high-end vehicles equipped with navigation systems) and the trip selection may not be representative (e.g. lower use of navigation services for short or routine trips).

Spatial accuracy

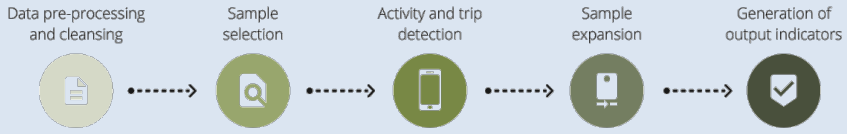
The precision of GPS positioning has made vehicle data particularly sensitive in terms of privacy protection. INRIX, one of the largest providers of traffic-related data, has taken steps towards mitigating the risk of re-identification in personal data and preventing their misuse. It destroys the exact origin and destination[3] of each trip and allocates a random vehicle identifier to each new trip.

Data from ticketing and smart-cards

Public transport systems generate a stream of transaction data that reveals valuable insights on public transport trips. However, those data are not without their limits: ticketing data are not integrated across transport modes; the use of heterogeneous media, including paper tickets, smart-cards, debit cards and smartphones, make it difficult to compile data; the data miss entirely users who travel illicitly and do not pay for the journey; and little data are available on where the passenger gets off the transit network. Bonnel (2021) illustrates how transport planners can fuse survey data with incomplete ticketing data for a more complete picture of travel by public transport.

Box 1. Five steps to estimate an origin-destination matrix from mobile phone data

Estimating trip matrices and other mobility indicators from mobile phone data typically involves five important steps. This box describes the purpose and the challenges involved in each of the five steps. What follows applies to both mobile network data and smartphone app data.



1. Pre-process and cleansing. Mobile network and app data are not error free and must be cleaned before further analysis. Analytics companies providing these indicators have developed their own filters and algorithms to detect errors and eliminate data points considered unreliable or faulty.

2. Sample selection. A useful sample excludes devices that belong to the Internet of Things but are most likely not mobile phones. It also excludes devices that are turned off for an excessive amount of time. The selection of devices is based on a set of criteria related to their activity, which is enough to determine the mobility patterns of their users with an adequate level of accuracy and reliability.

3. Activity and trip detection. One must identify stationary periods, also called stays or activities, since trips are defined as what connects two consecutive stays in different locations. In the case of MNO data, one can identify a stay when two or more connection events (such as phone calls and data exchange) occur in the same location during a particular period. Many algorithms use a threshold to the order of 30 minutes to this end, but the exact value depends on the frequency of events. In the case of smartphone app data, one can identify a stay from a cluster of GPS "pings" in a set time threshold of, say, 15 minutes.

To infer a mode of transport for each trip, one seeks to match the mobile traces to transport network characteristics: speeds, timetables, routes, stations, platforms, etc. In urban areas, this has shown little success so far because transport mode characteristics are very close to each other. However, mode identification works much better on the interurban side, as the longer distances provide better results. Sensors in smartphones (mainly the accelerometer) can help identify the mode in smartphone data, though it is not without its own challenges. The inference of transport mode from smartphone data is a fruitful research domain, as illustrated by Brändle (2021).

4. Sample expansion. The fourth step is the expansion of the sample to the total population while at the same time compensating for any bias in the data. The approach used for the upscaling of the sample depends on the characteristics of the study. In the case of the residents in the country for which the mobile network data are available, the expansion of the sample involves factors based on the home location, typically at the level of census tracts.

One can infer the place of residence based on the user's longitudinal behavioural patterns during several days or weeks. However, due to privacy constraints in some countries, the anonymisation process resets the device identifier every 24 hours and makes it harder to infer home location.

5. Generation of output indicators. Finally, having expanded the sample to the total population, data analysts process the activity-travel diaries to produce the information requested by the project (e.g. OD matrices) with the required level of spatial and temporal aggregation. Note that at this stage, the data provided to end-users are always aggregate.

Source: Willumsen (2021), Brändle (2021).

Ticketing data can produce station-to-station trip matrices in the best case and does not capture door-to-door movements. This level of spatial precision is nevertheless sufficient in most transport models.

These data can be useful in conjunction with other data sets to build transport models and analysis, as well shown by Vinayak et al. (2019) for public transport mode share analysis, or for shared mobility analysis (Aifadopoulou et al., 2020).

Data quality and potential biases

Big data comes potentially with big biases, especially when sourced from uncontrolled samples of the population, as it may underrepresent certain types of households. Examples include low-income households that may not have or use smartphones, the elderly, or ethnic minorities. Such samples may carry a specific socio-economic or behavioural bias, which could result in a distorted vision of reality, and which the most complex correction mechanisms may only partly correct.

Unpublished research in France and Norway suggest that estimated trip numbers can vary dramatically from one MNO to another. This is a reminder for transport planners to invest time and resources in the validation of new data sources against the most robust data available, such as passenger traffic counts. MNO data use is reportedly more common in the United Kingdom, the United States and Spain, resulting in better-quality data processing in those countries.

Independent sources for calibration and validation

Roundtable participants considered more traditional data sources as essential to the elaboration and validation of newer big data sources. Traditional sources are not only essential in the short term; they will remain so. This is because the technology and the population samples that generate big data keeps changing. In such conditions, the accuracy and representativeness of big data rely on the frequent recalibration of the process and its validation against ground truth.

For Willumsen (2021), however, no data source is error-free. In the same way as Household Travel Surveys suffer from underreporting, mobile network data and smartphone app data could also miss some trips. Considering traditional data sources as ground truth could be misleading.

All of the roundtable participants agreed that traditional survey data are needed to validate new sources of data. Traditional survey data are often used in the calibration process of OD matrix estimation from these new data sets. Friedrich et al. (2010) and Bonnel (2021) provide good examples where results from mobile network data and smartphone app data were compared with ground truth from traditional sources.

Main limitations and sources of error of mobile phone data

This section identifies some of the limitations of mobile phone data for trip matrix estimation. It comments on ways analytics companies have overcome at least some of these limitations.

Sample size

MNOs generate large data samples that are mainly limited by the penetration rate of mobile phones and more importantly by the market share of a given telecommunication company. In most countries, the sample can represent between 15% and 40% of the population, not just for one day but for 365 days a year. Even if this is not uniform and represents a maximum because the process works on a subset of

devices that have an adequate level of accuracy and reliability, the final sample is generally much higher than the one achieved with household travel surveys (Willumsen, 2021).

Based on a review of about 80 articles on mobile phone data, Chrétien et al. (2018) found it possible to secure a larger sample size over a longer duration using MNO data rather than smartphone app data, due to user acceptance constraints.

Socio-demographics and expansion factors

Having access to some socio-demographic characteristics of mobile phone users in a sample is important for two main reasons. First, it helps diagnose and correct for biases in the sample. Second, it enables the analysis of travel statistics by population group, which transport planners often require.

Mobile phone data's major weakness is that it does not capture socio-demographic details at the source but indirectly through linkages with various other sources. For instance, mobile traces may reveal each users' home location. From there, linkages can be made with other sources of data so to attach socio-economic characteristics to each traveller. It probably depends on local circumstances whether someone's address, which can be common to tens or hundreds of dwellings, is an accurate predictor of an individual's socio-demographic details. In the case of MNO data, one may retrieve further socio-demographic information from the customer database, such as gender and type of contract.

Roundtable participants remained divided over the severity of biases in mobile phone data. The debate will likely continue due to the lack of transparency from MNOs, from big data analytics companies and from aggregators of smartphone app data. Any gap in market penetration is indeed commercially sensitive.

Spatio-temporal precision

Location accuracy of data from MNO is dependent on the size of the cells; this can be quite small in dense urban areas and much bigger in rural, underpopulated regions. The number of events or communications also give the spatial (and temporal) precision. MNO data provide more and more traces as people intensify the use of mobile data. There was a huge increase in accuracy with 4G (two to three times more events than previous generations). There will be again an increased spatial resolution with the coming 5G, where cells cover a much smaller area.

App data usually provide more frequent data points, depending on the number and duration of the interruptions in the use of the apps. Smartphone apps offer better geolocation than mobile network data.

Algorithms' accuracy

The algorithms used in the treatment process of raw data play a key role in identifying activities, trips and trip chains obtained from MNO or app data. Their parameters influence how well they capture a trip's origin and destination and the time of travel. Transport modellers would have more confidence in estimated trip matrices if data providers were more transparent with their algorithms and validated their results against conventional data.

Short trips and phantom trips in MNO data

Short-distance trips are particularly difficult to detect from MNO data, as such trips tend to remain within the same network cell. In many transport models, however, the zoning is coarser than the cells of a mobile network: the lack of data on short-distance trips is, therefore, accepted. These short trips are mostly walking trips and intra-zonal trips in a transport model.

More problematic could be the issue of "phantom trips" mentioned earlier. Some roundtable participants claimed they could identify and filter out the cell jumps that create these fictitious trips. However, this process remains one of the least documented aspects of MNO data processing.

Continuity

The lack of continuity over time degrades the usefulness of the data. This may happen in different ways. Some MNOs reanonymise the data with some frequency, once a day or even every couple of hours, for example. This makes it impossible to track sufficient movements and stays to identify the place of residence with confidence, which, in turn, makes correcting for bias and expansion of the sample much more difficult. Smartphone app data seem to suffer less from this issue as the marketing IDs used are more persistent.

Quality indicators

Transport modellers should be mindful of the limitations of big data. Quality assurance could involve the computation of quality indicators for each OD pair. With the specific aim of estimating trip matrices from MNO data, Essadeq and Janik (2021) define quality indicators that reveal how the inference of transport mode becomes more reliable on longer-distance trips.

Especially when ground truth is not well known, it seems to be a good practice to define some quality indicator related directly to the estimation process of the trip matrix. Future research may be needed to define the most useful indicators for practitioners.

Privacy protection and its implications on transport modelling

Protecting privacy is often cited as the main obstacle to the use of big data in the analysis of travel patterns. Big data creates privacy threats, especially with the growing risk of re-identification of individuals in anonymised data sets.

De Montjoye et al. (2013) observed that one can characterise a unique individual from their trip traces. An attacker having observed the timed whereabouts of a target individual at four random moments during a whole year has a 95% probability of recognising his target in a data set of 1.5 million individuals.

Data collection and analysis should be fully in line with existing privacy frameworks and should evolve as those frameworks themselves evolve. Roundtable participants discussed the balance between the need to protect privacy and that of measuring, managing and planning for transport demand.

When considering data sharing, parties need to ensure compliance with privacy protection regulations. In Europe, countries can adopt national laws authorising extensive reprocessing of personal data without violating the EU General Data Protection Regulation (GDPR) in cases where the data collection serves the public interest. Compliance of an application with the GDPR requires, for example, data minimisation (an application must collect only necessary personal data), purpose limitation (the purpose of personal data collection must be clear and limited) and transparency to the end-user.

SuM4All (2021) proposes to develop a “personal data sharing tool” that helps individuals give and revoke consent for personal data sharing, such as the reuse of mobile network data. With such a tool, individuals would be in constant control of their personal data. They could access and contribute to a global layer of personal data for cross-border and cross-sectoral sharing. This tool would encourage and enable citizens to donate information about their trips to transport planning authorities without the fear of being tracked.

Even when anonymous, geo-spatial data collected by devices such as smartphones can reveal a lot about individuals, including medical conditions or visits to political or religious gatherings (Thompson and Warzel, 2019). Hence, the most robust data protection methods should be applied to location, trajectory and other personal data (ITF, 2015).

Protecting privacy is a central concern of both mobile network operators, app aggregators and data analytics companies. The key idea is to provide this protection by design and never to disclose data that could be traced back to an individual.

Anonymisation techniques

Data producers employ a range of anonymisation techniques to prevent the misuse of personal information. To do so, they often choose to *pseudonymise* all big data traces that record the position of individuals over time. In practice, they replace all information that directly reveals a person's identity, such as name and address, with an artificial identifier.

Fiore et al. (2020) point out that pseudonymisation only provides a mild level of data protection, and describe the risks of re-identification using location tracking data. For Brändle (2021), one must regard as personal data the pseudonymised fine-grained location data representing the daily routines of individuals.

In response to the threat of re-identification, data producers deploy various levels of additional protection. Typically, they reset the artificial identifier at regular intervals: for instance, after each trip, each day or each week. The longer an identifier is retained the more robust and insightful the analysis (UN, 2019).

Once the data leave the hand of their producer and are more widely disseminated, further techniques help protect privacy. Aggregation is the most frequently used solution, whereby only trip totals are provided. Willumsen (2021) describes how data providers strengthen the protection using K-Anonymity, in which no trip count is provided if this count is smaller than K. It is sometimes the case that two providers in the same country apply different criteria for selecting the value of K. In the case reported by Essadeq and Janik (2021), Orange sets the value of K at 20.

An alternative to aggregation consists in generating an entirely synthetic population that perform activities and trips modelled on the observations. This is, in essence, what is done when using Household Travel Surveys to produce disaggregate models except that, in this case, the sample size is at least an order of magnitude bigger and the level of detail includes variability of activities and behaviour on different days and times of the year. (Willumsen, 2021)

There are instances where big data analytics companies run their trip detection and travel demand analysis algorithms directly on the raw data hosted on the data producer's servers. Aggregation only occurs later, when exporting final outputs. However, it can take years to build enough trust between partners to formalise authorisations and to develop the technique.

Among researchers who mine smartphone positioning and sensor data to survey people trips, some report a lack of comprehensive reference data sets (Brändle, 2021). These are raw data sets, prior to any aggregation, which are precisely annotated with the relevant trip details for the training of trip detection algorithms. Institutions that wish to support research should be aware of the shortage of reference data sets and should address it in a way that does not represent a privacy threat.

The role of smartphone operating systems

For transport planners to benefit from smartphone app data, companies (such as Apple and Google) developing operating systems and controlling the apps marketplace must keep granting access by those apps to relevant data such as fine-grained smartphone location. Competing to provide higher levels of privacy protection to their customers, those companies could alter some of the features that enable the analysis of travel demand through app data.

Brändle (2021) reports that Google defines "stalkerware" as "code that transmits personal information off the device without adequate notice or consent and does not display a persistent notification that this is happening." The company defines a set of compliance rules for application developers distributing mobile applications on the app store (Google, 2020). It appears that access to fine-grained location data is getting increasingly difficult, which favours the end-users but restricts options for smartphone application developers from including travel mode identification into their applications.

Apple also communicates and takes action towards limiting access to private information by apps on their marketplace. With iOS 14.5, smartphone owners will be asked to authorise explicitly the access by each app to the device's unique marketing ID (Apple, n.d.). Failing to access the marketing ID, aggregators of app data could find it increasingly difficult to merge data coming from various apps used by the same individual (Koetsier, 2020). At the cost of greater computational power and some loss of transparency, they may, however, identify patterns in traces collected by various apps and keep providing a similar level of service to transport planners.

Roundtable participants, therefore, call on smartphone OS developers to consider the needs of the transport planning profession. In particular, they ask developers to acknowledge that transport authorities now develop smartphone apps to conduct travel surveys (Pronello and Kumawat, 2021) and that those apps require access to high-frequency fine-grained location and sensor data in the background.

Technical recommendations for the use of mobile phone data in transport models

This section builds on existing guidance on the estimation of trip matrices using mobile phone data. Authorities, research institutes and other stakeholders have explored the use of MNO data in this context and offered some guidance on their adoption. Such guidance includes:

- "Utilising Mobile Network Data for Transport Modelling" in the United Kingdom (Transport Systems Catapult, 2016)
- "Synopsis of New Methods and Technologies to Collect Origin-Destination (O-D) Data" in the United States (FHWA, 2016)
- "Cell Phone Location Data for Travel Behavior Analysis" also in the United States (NCHRP, 2018)
- "Handbook on the Use of Mobile Phone Data for Official Statistics" (UN, 2019)
- the technical note on the use of big data in urban transport planning by the Inter-American Development Bank (BID, 2019) and
- ongoing work by Positium (Estonia) and Cerema for France (Cerema, 2019).

Frequent technological changes have an impact on big data sources. Changes include new mobile network technology and new antennas such as 4G and 5G, changing behaviours such as the intense use of social media, or increased computing powers for the processing of big data. The legal framework also changes to keep up with technology. This ever-evolving landscape makes it particularly hard for transport planners to invest in the risky use of big data sources.

Not only do the legal and technological environments keep changing, they also tend to vary locally. Even in Europe, where privacy protection is harmonised under the GDPR, each national government and national privacy protection agency can interpret EU-level guidance in its own way. Adding to the regulatory

fragmentation, individual American states can develop their own privacy protection rules. The California Consumer Privacy Act (CCPA)[4] gives consumers more control over the personal information that businesses collect about them.

Authorities and the transport planning community should periodically revise transport modelling guidelines in light of this heterogeneous and fast-changing context. Doing so will encourage greater use of big data where relevant and prevent misuse. These guidelines should help transport planners define the technical specifications of big data processing.

Such guidelines should be country-specific to account for specific data sets used in data fusion (e.g. national census, national travel survey, etc.) and for the local privacy protection laws or case law.

Preparing and updating national modelling guidelines is also an opportunity to reflect on data needs in a rapidly changing transport planning landscape. Doing so allows policy makers to adapt to:

- new forms of mobility (e.g. personal mobility devices, app-based services)
- new commercial traffic, especially with home-delivery services
- new transport modelling tools, such as activity-based models.

Identifying trips and activities in mobile network operator data

To detect trips in MNO data, Essadeq and Janik (2021) suggest examining how long a device remains in a cell and deciding whether this duration, or dwell time, reflects an activity (when it is above the predefined dwell time threshold) or a trip between two activities. This is the most frequently documented approach to identifying trips in MNO data.

There is no convention among professionals for setting this threshold. A relatively short duration of a few minutes would be necessary to capture a trip to a short-lasting activity, e.g. dropping off children at school. Yet choosing a short duration will improperly classify as an activity the time spent crossing a congested cell or waiting for a public transport connexion. A threshold of several hours will allow air travellers to be coded as making a single trip, despite their precautionary time at the airport, but will ignore a vast number of trips motivated by an activity whose duration is shorter than that. No single value can fit all purposes. Data producers should thus refine their algorithms so to handle all types of trips.

Essadeq and Janik (2021) found that using a three-hour threshold leads to ignoring nearly half of the trips otherwise observed using a one-hour threshold in the short- to medium-distance trip segments.

If the aim is to develop agent-based models, it is best to request data on the sequence of activities and trips. This is a more complex data structure than a trip matrix but provides the full linkage of activities and trips required for such a model. This area only recently explored by modellers is likely to be particularly rich in insights. Willumsen (2021) reports an example where MNO data served as input to an activity-based model for Barcelona, Spain.

Modes of travel

When using mobile phone data, one can seek to detect the mode of transport by comparing three sources of information:

- the trajectory of the user observed from mobile phone records, be it from MNO or from smartphone apps
- the geography of the transport network (road network, transport hubs, etc.) and

- the travel times for different transport modes and route choices, obtained from diverse data sources such as online travel planners' Application Programming Interfaces (API).

Limited by the spatio-temporal resolution of mobile phone data, this type of approach performs best for medium- and long-distance travel. Mode identification is particularly problematic in urban areas, due to the density of the transport network and the coexistence of different transport services. For some OD pairs, the travel times and trajectories for different modes (e.g., car, bus and bicycle) can be very similar, making it practically impossible to reliably identify transport mode from location data.

It remains unclear if some providers perform better mode detection than others and if detection becomes more accurate as the technology matures. The transport planning profession would thus benefit from independent benchmarking exercises. Competitions organised by universities for the benchmarking of various trip and mode recognition techniques are welcome. Such competitions stimulate the research community towards the most accurate interpretation of big data and help transport planners have some transparency on the limitations of big data. Brändle (2021) reports on the Sussex-Huawei Locomotion Challenge, in which research teams compete to infer trip details from an annotated set of smartphone sensor data. To increase the quality of their algorithms, researchers, developers and practitioners would benefit from having more training data sets available, with mode data analysis competitions organised.

In comparison with MNO data, smartphone app data could offer the advantage of a higher spatial accuracy. Brändle (2021) states that one can integrate mode-detection algorithms to different mobile applications, such as implicit autonomous ticketing for public transport journeys, incentive schemes for behaviour change, the capture of actual travel in MaaS applications, and many others.

Artificial intelligence

Artificial intelligence (AI) could play an increasing role in the accurate detection of trips and activities since the application of simple methods offers incomplete results. AI could identify and learn from patterns that human intelligence may not notice. Likewise, AI could outperform other algorithms to detect transport mode from geolocation and sensor data. Researchers should thus embrace the potential for artificial intelligence in the mining of big data.

Artificial intelligence is considered a key investment to unlock innovation. Countries and corporations are engaged in a global race to take a leading role in the unfolding AI revolution. This interest in AI, and the funding that comes with it, should be captured to the benefit of the transport planning profession.

Research sponsors should foster the development of algorithms, including AI solutions that infer trip details from mobile phone signals and smartphone sensors. Creating annotated reference data sets would be an important element of this policy. It would support the training of computer models for trip detection and transport mode detection. The organisation of AI competitions could be part of this strategy.

New generation of household travel surveys

Participants at the ITF Roundtable did not foresee big data replacing household travel surveys. However, they envisaged two main ways big data could reshape travel surveys: first, through smartphone-assisted data collection and later, through data fusion with crowd-sourced data.

Pronello and Kumawat (2021) examined 55 smartphone apps that either support a travel survey with controlled sampling or simply collect travel data donated by the crowd. Travel survey apps detect trip departure and arrival times automatically, sometimes also inferring the mode of transport, creating a prefilled travel diary for the respondent to review and submit. Future Mobility Sensing is one such app that

demonstrated not only the need to calibrate the trip detection algorithms carefully but also to design the most user-friendly interface for the review and validation of the prefilled travel diary (Carrion et al., 2014).

Willumsen (2021) describes the growing reluctance from respondents to complete traditional travel questionnaires in general and travel diaries in particular. Some people choose to simplify their responses and under-report trips. The genuine difficulty to recall all trips and activities further contributes to the under-reporting of trips. Smartphone-assisted travel surveys could thus address several weaknesses of traditional survey methods.

Smartphone-assisted travel survey protocols come with specific challenges, such as intense battery use and heterogeneous hardware and OS configurations. Considering the high development costs involved, local and national governments should join forces when developing such tools and involve experienced academic partners. Such innovation is nevertheless a promising prospect that could address the notoriously high cost of household travel surveys.

The time is right for household travel surveys to include dedicated smartphone apps for the tracking and annotation of trips. It should be among the priorities of transport planning authorities to design and test smartphone apps to capture trip details and reduce the burden of trip diaries on respondents.

Willumsen (2021) also proposes that crowd-sourced mobile phone data complement household travel surveys in such a way that authorities could reduce the sample size and thus the cost of the survey. Travel surveys could power good trip/tour/activity generation, destination and mode choice models, while big data generated by mobile phones provides origin-destination matrices. Bonnel and Munizaga (2018) have already proposed research to specify optimal sampling for household travel surveys when good mobile phone data are available.

Mobile phones could have several transformative impacts on travel surveys, as discussed above. One could also mention the possibility of creating smartphone-assisted or MNO-assisted longitudinal surveys. There is thus an opportunity for transport planners to partner with the research community to define a roadmap for household travel surveys that accounts for the opportunities created by big data. This roadmap should seek to maintain and improve the quality of household travel surveys. It should acknowledge the falling response rates and the potential for data fusion with other sources such as ticketing, mobile phone signals and smartphone app data.

Lifting the barriers to data sharing

Public administrations in the transport sector already collect and analyse large quantities of data from transport service operators, surveys and many other sources such as phones, connected vehicles and smart infrastructure (WBCSD, 2020). However, the unprecedented quantity, complexity and availability of data collected from and about transport, together with advances in analytics, present new opportunities for transport policy making. With the digitalisation of transport services, the ubiquitous use of smartphones, increased mobile internet access, social media, and the vast amounts of data collected by vehicles, transport infrastructure and various mobile devices comes the potential for a paradigm shift in how and for what purpose data are used.

Governments in many countries are trying to improve transparency and encourage innovation. One strategy for doing so in the transport sector is by making data more widely available and linking them with data from other sectors. The ITF Big Data Working Group discussed the use cases based on open data.

The use of big data has the potential to change the way governments undertake planning and regulatory functions in areas including economic efficiency, maintenance and safety, and environmental protection. Data-driven policy making can also facilitate the monitoring and enforcement of transport-related legislation.

The private sector collects much of the big data that is relevant to transport policy makers. Access to this data would allow governments to improve key planning and regulatory tasks while reducing survey costs. Establishing the right partnerships between the private and the public sector is critical. This section gives recommendations on how to do so. It draws from discussions held during the Big Data and Transport Models Roundtable and the 2016-2017 meetings of the ITF Big Data Working Group.

The principles to facilitate data sharing

One of the challenges that authorities face is how to design public policy in an era where an increasing amount of privately sourced and owned data chart the location and movement of individuals (ITF, 2016). Location data such as those produced by smartphones and navigation devices have led to a range of new business models. However, a considerable gap in the collection of mobility data has emerged between the private and public sectors in recent years, with the private sector holding the most timely and accurate data on mobility. Were public authorities to have access to such data, they would gain insight that would allow them to improve the management of transport systems and make better-informed decisions for new investments. This makes the development of new models of public-private data-sharing partnerships indispensable.

The case studies in Annex A show that partnerships for data sharing between the public sector and the private sector vary in type, size and scope. They include large-scale data dumps as part of surveys, fixed-term research projects and commercial schemes. Partnerships have included different types of value propositions: logistics data analytics, improved route planning algorithms, innovative data usage for roads

safety, information on tourism traffic and usage-based car insurance. Partners include various government agencies, universities and research institutes, and private businesses (car manufacturers and telecoms companies among others).

The data-sharing partnerships in Annex A mainly revealed a mix of research projects, business development activities from the private sector and cases where the data sharing with the public sector by the private sector was mandatory. In the case of research projects, the co-operation among partners tended to be straightforward but often limited to the project's terms of reference and the pre-set limited duration. In a range of cases, the partnerships were used to try and test novel ideas and concepts, e.g. in the form of pilot projects. The cases that included a business development element were of a limited duration, as free data access was only granted for a certain period of time. However, parties could further negotiate a continuing partnership with a clear cost structure following the initial period.

Box 2. Drivers and principles for data sharing

Working in partnership with a range of mobility stakeholders, including auto manufacturers, operators and industry experts, the World Business Council for Sustainable Development (WBCSD) has identified principles and drivers that can shape a model and standards for data sharing. There are five principles to which mobility system stakeholders should strive to adhere to create data sharing architectures that benefit all parties:

- data sharing should enable all stakeholders to create and capture value
- data sharing must be ethical, inclusive and unbiased
- data sharing should incorporate privacy by design
- data sharing should embrace cyber-security by design
- data sharing should be adaptive and iterative.

To make mobility data-sharing principles as relevant and widely applicable as possible, they should embody four drivers that stand as prerequisites:

- vision: the end-goal for a mobility use case and the shared data that enable it
- value: thinking expansively and holistically about the different ways data create benefits and for whom
- future-proofing: crafting principles that are sufficiently flexible to cope with a rapidly changing environment
- trust: essential to the future of mobility, especially in the realm of shared data.

Source: WBCSD (2020).

Data were shared only between the parties involved in almost all case studies. This was an effort to protect private and commercial data. Data or insights gained from the data also remained between the parties. The Working Group nevertheless recommended that governments turn any data they receive into open data unless specific reasons make it impossible to do so (e.g. protection of personal or commercial data). Parties need to ensure compliance with privacy protection regulations when sharing data. In Europe, the

GDPR allows countries to adopt national laws authorising extensive reprocessing of personal data if the data collection serves the public interest. Before sharing, pseudonymisation and encryption of data should be considered.

Trust is a prerequisite to data sharing. To build trust, partners in a data exchange can develop and endorse non-disclosure agreements. Partners can also involve a trusted third party, sometimes called an "honest broker". Another solution consists of developing a "safe answers" approach, whereby only query results are exchanged instead of raw data. A privacy-proof, ethical, unbiased and inclusive sharing of data facilitates trust.

For the sake of protecting private and commercial data, the public sector should seek to follow two principles: 1) purpose specificity, so that no data are collected without a precise regulatory or planning task and 2) data minimisation, giving preference to the lightest possible data collection mechanism.

Governments can form partnerships with the private sector for the provision of data. This requires incentives for the private sector. Governments could offer to produce timely open data in exchange for receiving data from the private sector. This public data can make businesses more efficient and more profitable.

Another solution is the supplier-client relationship, which could be an incentive for the collection of higher-quality data at the source. Financial compensation might allow companies to recover the cost of collecting or adapting the data. This would address the company's concern that other operators may freely benefit from the data collection they funded.

Governments have the option of mandating the sharing of data to fulfil a number of regulatory tasks. In this case, the principles of purpose specificity and data minimisation are particularly important to limit the burden imposed on the private sector and the loss of commercially sensitive information. Mandatory data sharing implies that the data are of public interest. The Working Group recommends that only cases of clear public interest require a mandatory sharing of data.

However strong the incentives, however simple the partnership, data sharing involves substantial transaction costs. Partners often lack the highly qualified personnel for this. Klein and Verhulst (2017) recommend that corporations elect "data stewards" to act as focal points for data access:

> A key challenge to engage with companies to access corporate data sources is the current lack of clarity as to which individuals are tasked or have the authority to consider data requests from third parties. In order to streamline data collaboration, corporations should consider creating the role of "data stewards" to act as focal points for handling requests to access corporate data. These data stewards would be responsible for responding in a more effective and consistent manner to external demand for data, as well as co-ordinating with the various data actors within a company.

Following Roundtable and Working Group discussions, it becomes apparent that both public and private organisations should designate individuals or teams who proactively initiate, facilitate and co-ordinate data-sharing partnerships. Those "data stewards" will initiate and scale up from pilots, promote the exchange and reuse of data in the public interest, and protect potentially sensitive information. This recommendation is also made by the High-Level Expert Group on Business-to-Government (B2G) data sharing – an independent expert group set up by the European Commission to make more data available and increase their reuse for the common good (EC, 2020). Data stewards should also work to break the silos that represent a barrier to internal data sharing. Silos are frequent in all kinds of organisations in both the private and public sectors.

Governments should thus invest in the training of the public-sector workforce and in the recruitment of big data specialists. Skills should address the governance, privacy and technical dimensions of big data. Public sector staff must have the ability to assess if data quality is adequate for each operational purpose. Governments should prioritise skills development in areas such as artificial intelligence, machine learning and cloud computing, which are necessary for data processing and data sharing (SuM4All, 2021).

Examples of data sharing

In the context of the ITF Working Group on Big Data, several ITF member countries submitted case studies on public-private transport data-sharing experiences. The cases illustrate various levels of ambition, from short-term collaborations to large research projects (see Annex A). The legal frameworks used to specify each partner's roles and responsibilities ranged from licensing to different types of partnership agreements, e.g. confidentiality agreements, memorandums of understanding (MoU) and contracts.

Countries provided information on a specific partnership, including its value proposition, the costs involved, data sources, data ownership and the methods used to analyse the data. Countries also described the lessons learned, the relevance of the results and the innovations born out of the data-sharing projects.

Participants in the ITF Roundtable discussed other, more recent examples of data sharing. Some examples could inspire new data-sharing agreements; others could serve as templates for elaborating the technical or legal conditions of data sharing.

SuM4All (2021) details ten other case studies from four different continents. They illustrate how concerted policy-making approaches can guide the development of an economically viable, secure, and ethical data-sharing ecosystem for both public and private stakeholders.

Several member cities of the ITF Safer City Streets network[5] have engaged in a partnership with the navigation app Waze for a two-way data exchange. In this protocol, the city administration provides timely information on planned closures or events and receives relevant data on traffic conditions and travel times over entire cities.[6]

In Madrid, the mobility agency EMT collaborates with Moovit, a transit navigation app and analytics company. The agency's goal is to retrieve big data on Madrid's public transport use to ultimately model the effect of new mobility services and improve the mobility system as a whole.[7]

Today, transport planning agencies have access to vehicle trip origin-destination data from various providers.[8] Such commercial products are standardised in a way that limits transaction costs. They offer agencies seeking to calibrate transport models an alternative to costly intercept surveys.

Tools and libraries

The GovLab offers a library of data collaboratives, including several cases of mobile phone data use in transport planning. It reports that Seoul's metropolitan government analysed late-night call records provided by the Korean telecom KT Company to plan bus routes in an effort to create more affordable transport options. It also reports that the non-profit organisation OpenTraffic collects real-time traffic data from individuals and organisations across the transport sector to provide a complete data set of traffic and routing data to the public free of charge.

Templates for data sharing agreements are compiled in an online library known as Contracts for Data Collaboration (C4DC). These offer the opportunity to learn from fields other than transport and to find examples of legal terms used in a wide variety of countries.[9]

Templates for the *technical* dimension of data sharing also help reduce transaction costs. This is particularly important to shared mobility start-ups, operating in tens or hundreds of different cities in the world. Los Angeles released a Mobility Data Specification (MDS)[10] that ensures that data and regulations can be shared automatically, bi-directionally and in machine-readable format between mobility operators and public authorities. More precisely, it is a data standard and API specification for providers of mobility services, such as shared e-scooter companies. One goal is for local governments to collect the data to enforce, evaluate and manage the provision of mobility services. Another goal is that local governments may publish and refresh their regulations in a geo-coded machine-readable format, thus facilitating their seamless adoption by service providers. (ITF, 2019)

Shared Streets, a non-profit consortium that includes public- and private-sector representation, aims to facilitate data sharing and common standards by creating a set of open tools and data sets to aid cities and businesses. Its pilot projects include the analysis of origin and destination data from taxi and micromobility companies.[11]

Data standards development and adoption is one of the major challenges in assimilating data from various sources. Recognition of innovation in the industry should not be a barrier to standards development. (SuM4All, 2021)

The European cities of Madrid, Spain; Leuven, Belgium; Regensburg, Germany; and Thessaloniki, Greece are part of a project Modelling Emerging Transport Solutions for Urban Mobility (Momentum). The project investigates the use of big data sources in transport planning and modelling. An inventory of transport data sources reveals existing data-sharing partnerships for the provision of vehicle data (e.g. taxi and e-scooter position) and mobile network data (Momentum, 2019).

Examples of data sharing for sustainable development

Transport planning is a key aspect of delivering sustainable development goals (SDGs); there is a strong interest in the sharing of mobility data among the development community. The Development Data Partnership[12] involves 25 data partners, together with the OECD and other development organisations. The Partnership unlocks proprietary data in a secure, responsible manner towards the public good.

Mapbox Movement is one of the data sets partnering with the Development Data Partnership and used by the World Bank. It draws from over 600 million monthly active users of Mapbox-powered apps worldwide. Weather, social media and fitness apps all contribute to the data collection, resulting in global coverage and fine spatial accuracy.[13]

The Global Partnership for Sustainable Development Data connects hundreds of partners to foster collaboration, spur innovation and build trust in data sharing. Among other things, the initiative resulted in a workshop where participants discussed the use of MNO data in delivering SDGs in Africa. A partnership between Vodafone and the Ghana Statistical Service involved capacity building and access to call details records.[14]

Unlocking business-to-government data sharing

Having looked at the specific nature of big data, the barriers to data sharing, the principles for data sharing and several examples of partnerships, this report outlined several keys to unlock business-to-government data sharing. First, one should acknowledge the sheer diversity of possible data sharing frameworks, from the mandated release of data for the public interest to the commercial trade of big data. Various other forms of partnerships exist; these will thrive with the creation of a data stewardship role in both public

and private organisations. In addition, a workforce of knowledgeable data technicians will reassure stakeholders in their partnerships and create greater trust in big data. Governments should recruit and train skilled data scientists for the handling, validation and purposeful use of big data. Data analytics consultancies could offer assistance in setting up a framework that complies with privacy regulations without excessively degrading data quality. This report cited “honest broker” and “safe answers” as two examples, among many, where such third parties can help elaborate complex frameworks. Last, one should bear in mind the threats (real or perceived) that data sharing presents to privacy and commercial interests. A limited scope is thus a good place to start. Concepts of purpose specificity and data minimisation should guide an emerging data-sharing framework.

Notes

1 Operational events most often used by mobile network operators include “probes”, “handovers” and “location area updates”.

2 A common configuration is a tower (technically called Base Transceiver Station) with three antennas, each covering 120 degrees and generating three cells.

3 In this case, the trace ends with a non-zero speed, meaning the trip has been cut close to its destination, not revealing the true OD.

4 See https://oag.ca.gov/privacy/ccpa

5 See https://www.itf-oecd.org/safer-city-streets

6 See https://www.waze.com/en-GB/ccp

7 See https://moovit.com/wp-content/uploads/2019/07/Moovit-and-EMT-Announcement-News-release.pdf

8 Providers include INRIX (see: https://inrix.com/resources/inrix-insights-trips/) and TomTom (see: https://www.tomtom.com/products/origin-destination-matrix-analysis/)

9 See https://contractsfordatacollaboration.org/

10 See https://github.com/openmobilityfoundation/mobility-data-specification

11 See https://sharedstreets.io/pilots/

12 See https://datapartnership.org/

13 See https://datapartnership.org/updates/introducing-mapbox-movement-data/ and https://blog.mapbox.com/global-movement-data-for-mobility-insights-680955ee42d1

14 See https://www.data4sdgs.org/index.php/resources/mobile-data-social-impact-summary-report

References

Aifadopoulou et al. (2020), "Management of resource allocation on vehicle-sharing schemes: The case of Thessaloniki's bike-sharing system" *Operational Research*, p. 1-16. http://dx.doi.org/10.1007/s12351-020-00569-3.

Apple (2020), "Apple makes mobility data available to aid COVID-19 efforts", www.apple.com/newsroom/2020/04/apple-makes-mobility-data-available-to-aid-covid-19-efforts/ (accessed 06 May 2021).

Apple (n.d.), User Privacy and Data Use, App Store webpage, https://developer.apple.com/app-store/user-privacy-and-data-use/ (accessed 06 May 2021).

Brändle, N. (2021), "Inferring Modal Split from Mobile Phones: Principles, Issues and Policy Recommendations", *International Transport Forum Discussion Papers*, No. 2021/07, OECD Publishing, Paris, https://doi.org/10.1787/c752ca76-en.

BID (2019), "Como aplicar big data en la planificación del transporte urbano" (How to apply big data in urban transport planning), Technical note No. IDB-TN-1773, Banco Interamericano de Desarrollo (Inter-American Development Bank), http://dx.doi.org/10.18235/0002009.

Bonnel, P. and M. Munizaga (2018), "Transport survey methods – in the era of big data facing new and old challenges", *Transportation Research Procedia*, Vol. 32, pp. 1-15, https://doi.org/10.1016/j.trpro.2018.10.001.

Bonnel, P., M. Fekih and Z. Smoreda (2018), "Origin-Destination estimation using mobile network probe data", *Transportation Research Procedia*, Vol. 32, pp. 69-81, https://doi.org/10.1016/j.trpro.2018.10.013.

Bonnel, P. (2021), "Benefits of Cellular Telecommunication and Smart Card Data for Travel Behaviour Analysis", *International Transport Forum Discussion Papers*, No. 2021/06, OECD Publishing, Paris, https://doi.org/10.1787/3884255b-en.

Carrion et al. (2014), "Evaluating Future Mobility Survey: Preliminary Comparison with traditional travel survey", in *TRB 93rd Annual Meeting Compendium of Papers*, Transport Research Board, Washington, D.C.

Cerema (2019), "Quel apport des données issues des nouvelles technologies à la modélisation des transports ?", Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement, https://www.cerema.fr/system/files/documents/2019/07/cerema_panorama_donnees_passives_modeles.pdf (accessed 06 May 2021).

Chrétien et al. (2018), "Using Cell Phone Data to Understand Travel Behavior and Transportation Systems", in Aguilera A. and V. Boutueil, *Urban mobility and the smartphone: Transportation, travel behavior and public policy*, Elsevier.

Cuebiq (2016), "StreetLight Data Partners with Cuebiq to Help Transportation, Retail and Logistics Companies Use Location-based Data", https://www.cuebiq.com/press/streetlight-data-partners-cuebiq-help-transportation-retail-logistics-companies-use-location-based-data/ (accessed 06 May 2021).

de Montjoye, Y.-A., et al. (2013), "Unique in the crowd: The privacy bounds of human mobility", *Scientific Reports*, Vol. 3, https://www.nature.com/articles/srep01376 (accessed 06 May 2021).

EC (2020), *Towards a European Strategy on Business-to-Government Data Sharing for the Public Interest*, Final report prepared by the High-Level Expert Group on Business-to-Government Data Sharing, http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=64954.

Essadeq, I. and T. Janik (2021) "Use of Mobile Telecommunication Data in Transport Modelling: A French Case Study", *International Transport Forum Discussion Papers*, No. 2020/34, OECD Publishing, Paris, https://doi.org/10.1787/90483afc-en.

FHWA (2016), *Synopsis of New Methods and Technologies to Collect Origin-Destination (O-D) Data*, Report FHWA-HEP—16-083, Federal Highway Administration, United States Department of Transportation, https://www.fhwa.dot.gov/planning/tmip/publications/other_reports/origin-destination/fhwahep16083.pdf.

Fiore, M. et al. (2020), "Privacy of trajectory micro-data publishing: A survey", arXiv:1903.12211, https://arxiv.org/abs/1903.12211 (accessed 05 November 2020).

Friedrich, M. et al. (2010), "Generating Origin-Destination Matrices from Mobile Phone Trajectories", *Transportation Research Record: Journal of the Transportation Research Board*, p. 93-101, https://doi.org/10.3141/2196-10.

Google (2020), "Stalkerware – effective 1 October 2020", *Developer Programme Policy: 16 September 2020 announcement*, https://support.google.com/googleplay/android-developer/answer/10065487 (accessed 06 May 2021).

Google (2021), COVID-19 mobility reports, https://www.google.com/covid19/mobility/ (accessed 06 May 2021).

ITF (2019), "Governing Transport in the Algorithmic Age", *International Transport Forum Policy Papers*, No. 82, OECD Publishing, Paris, https://doi.org/10.1787/eec0b9aa-en.

ITF (2016), "Data-Driven Transport Policy", *International Transport Forum Policy Papers*, No. 20, OECD Publishing, Paris, ~~https://doi.org/10.1787/5jlwvz8g4vbs-en~~.

ITF (2015), "Big Data and Transport: Understanding and Assessing Options", *International Transport Forum Policy Papers*, No. 8, OECD Publishing, Paris, https://doi.org/10.1787/5jlwvzdb6r47-en.

Kisio (2020), "GPS et WiFi, des traces qui en disent long sur notre mobilité", https://kisio.com/uploads/2020/09/Kisio_dossierspecial_gps_wifi_VF.pdf (accessed 06 May 2021).

Klein, T. and S. Verhulst (2017), "Access to new data sources for statistics: Business models and incentives for the corporate sector", *OECD Statistics Working Papers*, No. 2017/06, OECD Publishing, Paris, https://doi.org/10.1787/9a1fa77f-en.

Koetsier, J. (2020) "IDFA Stay of Execution: Apple Delays New iOS 14 Privacy Measures Until 2021", Forbes, https://www.forbes.com/sites/johnkoetsier/2020/09/03/idfa-stay-of-execution-apple-delays-new-ios-14-privacy-measures-until-2021/ (accessed 15 July 2021).

Laney, D. (2001), *3D Data Management: Controlling Data Volume, Velocity and Variety*, META Group, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 05 November 2020).

Liu, J. et al. (2016), "Rethinking big data: A review on the data quality and usage issues", *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 115, p. 134-142, https://doi.org/10.1016/j.isprsjprs.2015.11.006.

Merino, J. et al. (2016), "A Data Quality in Use Model for Big Data", *Future Generation Computer Systems*, Vol. 63, p. 123-130, https://doi.org/10.1016/j.future.2015.11.024.

Momentum (2019), *Deliverable D3.1: Data Inventory and Data Quality Assessment*, https://h2020-momentum.eu/wp-content/uploads/2020/03/MOMENTUM-D3.1-Data-Inventory-And-Quality-Assessment-Issue-1-Draft-5.pdf.

NCHRP (2018), "Cell Phone Location Data for Travel Behavior Analysis", National Cooperative Highway Research Program, https://www.camsys.com/sites/default/files/publications/nchrp_report_868_Cell%20Phone%20Location%20Data_Cambridge%20Systematics.pdf

OECD (2012) "Quality Framework and Guidelines for OECD Statistical Activities", https://www.oecd.org/sdd/qualityframeworkforoecdstatisticalactivities.htm (accessed on 06 May 2021).

PIARC (2019), "Big data for road network operations", available free of charge for registered visitors at: https://www.piarc.org/ressources/publications/11/84bcf66-31346-2019R18EN-Big-Data-for-Road-Network-Operations.pdf.

Pronello, C. and P. Kumawat (2021), "Smartphone Applications Developed to Collect Mobility Data: A Review and SWOT Analysis", in Arai K., S. Kapoor and R. Bhatia (eds.), *Intelligent Systems and Applications, IntelliSys 2020, Advances in Intelligent Systems and Computing*, Vol. 1251, Springer, https://doi.org/10.1007/978-3-030-55187-2_35.

Streetlight Data (2017), Location-Based Services Data Beats Cellular on Spatial Precision, www.streetlightdata.com/cellular-data-vs.-location-based-services-data-spatial-precision/ (accessed 06 May 2021).

SuM4All (2021), *Sustainable Mobility: Policy Making for Data Sharing*, Sustainable Mobility for All, Washington DC, License: Creative Commons Attribution CC BY 3.0, https://www.wbcsd.org/contentwbc/download/11663/176259/1.

Thompson, S. and C. Warzel (2019), "One Nation, Tracked: Twelve Million Phones, One Dataset, Zero Privacy", New York Times, https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html (accessed 6 May 2021).

Transport Systems Catapult (2016), "Utilising mobile network data for transport modelling: Recommendations paper", document reference: MV7/RPT001/D14, Transport Systems Catapult, https://www.gov.uk/government/publications/mobile-phone-data-in-transport-modelling (accessed 10 November 2020).

UK DfT (2021), "COVID-19 transport data: Methodology", Department for Transport, United Kingdom, https://www.gov.uk/government/statistics/transport-use-during-the-coronavirus-covid-19-pandemic/covid-19-transport-data-methodology-note (accessed 06 May 2021).

Umlaut (2020), "Mobility behavior and app trends in times of Covid-19", https://www.umlaut.com/en/stories/mobility-behavior-and-app-trends-in-times-of-covid19 (accessed 06 May 2021).

UN (2019) "Handbook on the Use of Mobile Phone Data for Official Statistics", UN Global Working Group on Big Data for Official Statistics, United Nations, https://unstats.un.org/bigdata/task-teams/mobile-phone/MPD%20Handbook%2020191004.pdf.

US BTS (2021), "Daily Travel during the COVID-19 Public Health Emergency", Bureau of Transportation Statistics, United States Department of Transportation, https://www.bts.gov/daily-travel (accessed 06 May 2021).

Vinayak, P. et al. (2019), "Using Smart Farecard Data to Support Transit Network Restructuring: Findings from Los Angeles", *Transportation Research Record: Journal of the Transportation Research Board*, https://doi.org/10.1177/0361198119845661

WBCSD (2020), "Enabling data-sharing: Emerging principles for transforming urban mobility", World Business Council for Sustainable Development, https://www.wbcsd.org/contentwbc/download/8198/127647/1 (accessed 06 May 2021).

Willumsen, L. (2021), "Use of Big Data in Transport Modelling", *International Transport Forum Discussion Papers*, No. 2021/05, OECD Publishing, Paris, https://doi.org/10.1787/86a128c7-en.

Annex A. Case studies

Several countries that participated in the ITF Working Group on Big Data submitted case studies illustrating the use of big data in transport planning and transport statistics. Many of these case studies also illustrated the benefit of partnerships between the public and the private sectors.

There were eight complete case studies, covering examples from five countries and a variety of transport modes and sectors:

1. Finland: NordicWay Project
2. France: E-Logistics database
3. France: Using passive data sources to estimate traffic on regional trains
4. France: SNCF Route planner
5. France: Using GSM data to estimate tourism traffic
6. Greece: Delivering Driving Analytics to Insurance Companies
7. Italy: PLUG-IN Project
8. The Netherlands: KiM Mobility Panel

This material was written in 2016 and 2017 and reflects the state-of-play at the time.

Finland: NordicWay Project

Country		Finland
Case study		NordicWay Project
Partnership	**Problem description**	Hazardous locations such as a slippery road surface or a crash scene pose a threat to road users. Yet Safety Related Traffic Information (SRTI) is currently provided only in separate silos by different organisations. If SRTI messages could be shared among all organisations, stakeholders and entities, it could scale SRTI distribution. Drivers could be warned in advance and adjust their behaviour to enhance traffic safety. NordicWay is a pilot project that researches solutions for these problems in the cellular network (3G and 4G/LTE) and also in hybrid situations including ITS-G5. Cellular infrastructure is already available and is a cost-effective and scalable solution.
	Value proposition for data sharing	The NordicWay project goal is to pilot and facilitate specific Co-operative Intelligent Transport System (C-ITS) functionalities through a common architecture. A key asset in the architecture is the NordicWay interchange server which distributes SRTI between the stakeholders of National Road Administrations TMC SRTI data (partly open data) and OEM data. The goal of the project is to lay the foundation for interoperable automated cloud communication via the cellular network with data generated by vehicle on-board sensors, road users and the surrounding infrastructure. Communication will be established between vehicles, smart devices on the road, service providers, road administrators and other public administrations. A business model and a detailed scenario for the roll-out of cellular-based C-ITS services will also be developed. The first phase is expected to provide information about the functionality of the service through feedback from the users. Additionally, the functionality of the system, its capability for wider use, its commercial potential and its ecosystem model will be evaluated. The authorities are also able to expand their knowledge about channels for critical information delivery.
	Partners involved	The NordicWay project is a collaboration between National Road Authorities in Finland, Norway, Sweden and Denmark, and private partners including Ericsson, HERE, Kapsch, Scania and Volvo. It is co-financed by the European Union within the Connecting Europe Facility programme 2015-2017. In June 2015, a three-year pilot project between the Finnish Transport Agency, Transport Safety Agency Trafi and a HERE-led consortium was launched as part of the NordicWay. During the project, cars will utilise the mobile network to share specific and low-latency traffic safety information. Voluntary drivers will download an Android mobile application in their smartphones to connect and share information with other vehicles on the road.
Data and analytics	**Types of data/ data sources/ access to data**	During the NordicWay project, cars will utilise cellular networks to share hazardous locations, weather and slippery road DATEX2 SRTI messages. The pilot project shares obstacles on the road, weather conditions, slippery surfaces, road works warning, reduced visibility and accidents SRTI messages. NordicWay's goal is to have as many as 2 000 vehicles that will connect and share SRTI with other vehicles on the road and the surrounding infrastructure.
	Open data (legislative framework required?)	During the pilot, data are not shared outside the pilot partners. NordicWay's first phase builds on the SRTI open data platform between public and private stakeholders. The long-term vision is to share other C-ITS service data. The current pilot is intended to research this architecture and potential business models.
	Methods, algorithms, tools for analysis	Vision is a platform that shares C-ITS data with stakeholders so they may analyse it and create services to improve traffic safety, data value chain or the C-ITS cloud-to-cloud data market.

	Data ownership after processing	Data shared in the NordicWay Interchange server between project partners would be unprocessed SRTI messages using DATEX2 standard. The OEMs own the data during the pilot. National road authorities have their own already existing national open data that they are sharing.
Outcome	Lessons learned: why did it work or fail?	The trial phase began in May 2016 on E18, the main road between Helsinki and Turku (it also includes Ring I and Ring III roads), and expanded on 15 November 2016 to main roads from Helsinki to Espoo, Tampere (including the ring road) and Lahti. The pilot will last one year.
	Does it meet national obligations?	n/a
	Efforts to educate general public/ staff	The pilot is open for all drivers using Android smartphones along South Finland's main roads.
	Innovations it provides	Interoperable SRTI message distribution via the NordicWay Interchange server.
	New innovations	Interoperable SRTI message distribution via Interchange server.
	Costs/ cost structure involved	NordicWay is EU CEF funded, with a total budget of EUR 5.2 million.
	Relevance of the results	n/a

France: E-Logistics database

Country		France
Case study		E-logistics database
Partnership	Problem description	The project consists of collecting data from logistics service providers. Data from more than 1 000 e-commerce shippers (e-shippers) is gathered and innovative indicators like "click to possession" time or rate of successful deliveries at first attempted delivery are calculated. Logistics service providers supply the data which is then analysed. The aggregated results are shared with the participating e-commerce shippers through their logistics service provider. The database might be shared with government authorities and researchers if the remaining challenges with their statistical framework and with anonymisation for data sharing are overcome.
	Value proposition for data sharing	n/a
	Partners involved	Logistics data providers WelcomeTrack, Neopost Shipping and DDS Logistics collect data from more than 1 000 e-shippers working with 25 carriers to deliver more than 40 million shipments (10% of the market). The logistics service providers collect, track and trace data from their clients, standardise and send it to Deliver, an e-commerce logistics services platform. Deliver sent aggregated indicators back to its three partners who, in turn, provide that information to their clients with their own indicator, which allows them to compare themselves with their competitors. Access to aggregated and anonymised data from the database is restrained to logistics service providers who feed the database. Aggregated indicators are published at the national level.
Data and analytics	Types of data/ data sources/ access to data	Data source is raw track-and-trace data at each reloading, and aggregated data contains activity of the e-commerce shipper, the size of the shipment, its origin and destination, sector, B2C or B2B, stock or event. Since data is track-and-trace, anonymisation is challenging: researchers wish for a very precise location of shipments while Deliver only provides aggregated indicators. Indeed, aggregated indicators from Deliver, by region (NUTS3) and type of shipment, would be enough to trace the shipper, which is not acceptable for e-shippers.
	Open data (legislative framework required?)	Data is only shared between the project's partners. However, there are discussions about whether those data sharing should include researchers and ministries. The database could be enriched with existing open data on transport.
	Methods, algorithms, tools for analysis	A scientific committee with researchers, administrations and logistic specialists discusses statistical framework of an e-logistics database. Access to the database is proposed through the cloud with data visualisation tools.
	Data ownership after processing	Deliver owns treated data and calculated indicators but raw data is owned by the e-shippers.
Outcome	Lessons learned: why did it work or fail?	Getting data from logistics service providers makes data-gathering and trust-building for data sharing more efficient than collecting data directly from e-shippers. Different levels of aggregation can be applied with regional or sectoral indicators for national statistics and more disaggregated indicators for e-shippers, which allow them to rank themselves, give consistent shipment times to their client and optimise shipments. Anonymisation processes and statistical framework still needs to be improved through R&D developments in order to further build confidence in indicators and data anonymisation. Track-and-trace data might not be given by some carriers since it is not regulatory for them to share them with the e-shippers.

	Does it meet national obligations?	New indicators and data are gathered through this project, which will be helpful to get a better knowledge of e-shippers logistics.
	Efforts to educate general public/ staff	Data visualisation tools are used to make data easy to read and analyse for e-shippers.
	Innovations it provides	New, useful indicators are provided by this database (click-to-possession times, rate of successful deliveries at first attempted delivery, etc.) that will improve the transparency of shippers' efficiency and help optimise logistics. E-shippers can compare themselves with the rest of the market and optimise their rate of successful deliveries at first attempted delivery by better choosing its carriers. This project also provides transparency to the final client as to the quality of service; currently, the reliability of the time that passes between the order and the reception of the package is usually not known by users. This data can also be useful for decarbonisation analyses since new data, like the number of passages by carrier for each shipment or its itinerary, is included in the database.
	New innovations	n/a
	Costs/ cost structure involved	The majority of the costs are for research and development. Maintenance costs run about EUR 20 000 per year.
	Relevance of the results	n/a

France: Using passive data sources to estimate traffic on regional trains

Country		France
Case study		Using passive data sources to estimate traffic on regional trains
Partnership	**Problem description**	The aim of this study is to gain knowledge of rail traffic flows in order to enrich market analyses. Furthermore, several stakeholders share the transport market in the Paris region and several types of pricing exist, making it difficult to evaluate the traffic of SNCF trains. Therefore, SNCF needs a new analytical tool to analyse traffic on various time frames (hourly, daily, monthly, etc). Various data collecting devices exist (automated ticket control, ticket sales, automatic riders counts, etc.) but are not yet analysed because of the large volumes (about nine million validations per day), and because each type of device holds only a part of the information and is not sufficient for a comprehensive understanding of traffic (for instance, the Transilien train stations are not all equipped with automated ticket control devices).
	Value proposition for data sharing	n/a
	Partners involved	Research programme
Data and analytics	**Types of data/ data sources/ access to data**	Three types of data have been used in this project: data from automated ticket control, sales data, and data from automated laser counting of users from equipped trains (not all trains).
	Open data (legislative framework required?)	No open data is used in this project.
	Methods, algorithms, tools for analysis	The aim is to develop estimations of traffic by analysing data from various types of devices. The definition of the estimators faces two challenges. First, historic data can be seen as samples from a sampling design to be determined. An *ex-post* definition of the sampling design as a stratified multistage design was proposed. Secondly, methods for integrating multiple data types in the database are used to benefit from the richness of available data and consolidate results.
	Data ownership after processing	The data and analytics tool remain the ownership of SNCF.
Outcome	**Lessons learned: why did it work or fail?**	The future of mobility analysis includes various types of data, providing a higher volume of data but also rather superficial and sparse data. In order to enrich and fully use the potential of this data, the question arises of who or what structure can gather together the stakeholders who each hold some specific data. Even inside a structure, some barriers exist to prevent data sharing. Data is often not shared because of competition concerns related to the knowledge of rail traffic per station or per OD. The project helped acknowledge the potential value of data sharing and of combining passive data which is currently under-used by SNCF.
	Does it meet national obligations?	No, but it could enrich existing databases.
	Efforts to educate general public/ staff	Data visualisation tools are used to help understand and explore the database. Traffic indicators are computed coherently from disaggregated levels to macro levels, providing both detailed analyses and synthetic views.

	Innovations it provides	The statistical framework includes an innovative methodology for the model calibration which consists in considering that passive data can be treated like data from a traditional survey based on a stratified sampling design. The use of these combined passive data sources is innovative for SNCF.
	New innovations	n/a
	Costs/ cost structure involved	n/a
	Relevance of the results	n/a

France: SNCF route planner

Country		France
Case study		SNCF route planner
Partnership	Problem description	The aim of the project is to analyse if potential demand can be inferred from route planner queries. The project analyses data from route planners (route requests): how the requests are structured, whether they can reflect the mobility, whether the request volume is constant over time, what are the most frequent OD and how early are travels planned.
	Value proposition for data sharing	n/a
	Partners involved	This project involves various internal SNCF partners: SNCF Innovation and Research unit, Kisio Digital (in charge of apps and route request servers), Kisio Analysis (consultant in mobility analysis). The data-sharing framework is a research programme.
Data and analytics	Types of data/ data sources/ access to data	Data includes all digital devices from SNCF in the Paris region (mobile app, mobile phone website, and website). Three months of data were analysed (about 100 million requests). Requests were associated with data from ticket sales and ticket controls.
	Open data (legislative framework required?)	The data was enriched with public and open data: socio-demographic data, transport network using the GFTS format, points of interest (POI) from OpenStreetMap, data on cultural and sporting events from start-up Mapado.
	Methods, algorithms, tools for analysis	Data visualisation tools were used to analyse big data and proved necessary to understand whether and how this type of data could be used.
	Data ownership after processing	The data and analytics tool remain the ownership of SNCF.
Outcome	Lessons learned: why did it work or fail?	n/a
	Does it meet national obligations?	n/a
	Efforts to educate general public/ staff	n/a
	Innovations it provides	Route requests reveal most characteristics of transport demand (peak periods). There is a strong relationship between requests and traffic (estimated by ticket validations). This relationship allows data analysts to use route requests to anticipate mobility, especially for special events. Although it is rarely used, the route requests data allow to analysts to study how users anticipate their trips: 2% of requests are made a week prior to departure, 8% are made three days prior to departure and 40% are made the day before departure (up to 73% for a Monday morning between 7am and 9am). For a special event, the route request is made earlier: 15 to 20% of requests are made three days prior to the event. This anticipation for special events can be analysed more precisely by event type and date, making it conceivable to anticipate punctual demand for special

		events a few hours or days before the event with “real-time” analysis of route requests and adapt the transport offer. A deeper analysis of these data could be combined with machine-learning frameworks to model types of users in order to provide mode-focused information that is more relevant for the user.
	New innovations	The project provides surprisingly good outcomes in terms of traffic estimations. It shows potential for dynamic traffic operations for special events, although the research on this source of data is in its early stages.
	Costs/ cost structure involved	n/a
	Relevance of the results	n/a

France: Using GSM data to estimate tourism traffic

Country		France
Case study		Using GSM data to estimate tourism traffic
Partnership	Problem description	The aim of this project was to assess the feasibility of using mobile positioning data for generating statistics on domestic outbound and inbound tourism flows, and to address the strengths and weaknesses related to access, trust, cost, and the technological and methodological challenges inherent in the use of such a new data source. Tourism statistics is one of the domains in which the opportunities are rather clear as the properties of the data correspond to the nature of the tourism activities. Inbound, outbound, roaming and domestic data stored by mobile network operators (MNO) clearly correspond to the respective inbound, outbound and domestic domains of tourism; however, not without some methodological reservations.
	Value proposition for data sharing	n/a
	Partners involved	Research project.
Data and analytics	Types of data/ data sources/ access to data	Access to mobile positioning data is currently very limited mainly because of the regulatory limitations. There are big differences between the EU countries. The regulatory framework in the European Union is currently evolving with the upcoming General Data Protection Regulation. The study concludes that there is a need for a central framework for national statistical institutes and other stakeholders in order to obtain the data legally and according to an approved methodology in order to be able to produce comparable and reliable tourism statistics.
	Open data (legislative framework required?)	n/a
	Methods, algorithms, tools for analysis	Call Detail Records (CDR) are the basic type of data that are the most easily accessible by MNOs that represents phone activity, i.e. calls and messaging. Alternative data types include Data Detail Records, location updates and others. This study concentrated on the possibilities of longitudinal data that present the best methodological options for compiling tourism statistics. In the sphere of tourism statistics the resulting data can provide the following indicators and breakdowns: number of trips/visits; number of nights spent; number of days spent; number of unique visitors; broken down by: country of residence/place of residence; aggregation of time; aggregation of space; duration of trip/stay (same-day/overnight trip); main destination, secondary destination, transit pass-through; collective movement patterns; repeat visits.
	Data ownership after processing	n/a
Outcome	Lessons learned: why did it work or fail?	Based on the analysis of the methodology, it can be concluded that mobile positioning data, at present, cannot replace, but rather supplement the official tourism indicators required in the current Regulation 692/2011 concerning European statistics on tourism.
	Does it meet national obligations?	Longitudinal data is a must for reliable tourism statistics in order to assess the whereabouts of the subscribers over a long period of time (e.g. usual environment, differentiation of the trips by length, identification of overnight stays, etc.). Based on the outcomes of this study, it can be concluded that, at present, mobile positioning

		data can be used as a supplement rather than as a replacement source of data for the current official tourism indicators required.
	Efforts to educate general public/ staff	n/a
	Innovations it provides	The use of mobile positioning data has the potential to improve several aspects of tourism and other statistics, such as timeliness, access to previously unavailable statistical information (new indicators), calibration opportunities for existing data, space and timely resolution and accuracy.
	New innovations	n/a
	Costs/ cost structure involved	Initial implementation and automation is possibly expensive for both MNOs and national statistical institutes, offset over the years by lower costs of maintaining the system. After the automated processes are in place, the annual work to process mobile positioning data can be drastically less compared to current methods of gathering tourism statistics. A combination of mobile and traditional methods can prove to be cost-efficient, e.g. combining mobile data with (small) demand surveys may radically reduce survey sample sizes and provide cost-savings
	Relevance of the results	n/a

Greece: Delivering driving analytics to insurance companies

Country		Greece
Case study		Delivering driving analytics to insurance companies
Partnership	**Problem description**	Usage-based insurance (UBI) schemes, such as Pay-as-you-drive (PAYD) and Pay-how-you-drive (PHYD), have recently started to be commercialised around the world. The main idea is that instead of a fixed price, drivers must pay a premium based on their travel and driving behaviour. For these schemes to be effective, one must develop a system that continuously monitors the driving behaviour of users. This system entails a high degree of complexity related to the difficulty of developing technological solutions to crowdsource, collect, store and retrieve driving data, converting the massive data into meaningful driving analytics and developing a methodology to rate drivers according to their behaviour.
	Value proposition for data sharing	The project's goal is to develop a platform, methodologies and an algorithmic toolbox to deliver driving analytics and a driver rating system that will be further used by the insurance companies to form Usage-Based Insurance (UBI) schemes. The project has two major phases that complement each other. The first phase relates to the development of the proper indicators to monitor driving behaviour. Advanced computational intelligent techniques are implemented to detect events and mobile usage, and to identify the driver or passenger and the mode of transport. A significant part of the modelling relates to fraud analytics. The second phase develops a comprehensive system to rate drivers. This delivers a consistent driving rating system which quantifies the accident risk, so as the insurance companies can manage their premiums, but also allows the drivers to evaluate and improve their driving behaviour.
	Partners involved	This project is a collaboration between the National Technical University of Athens (NTUA) – a public institution –, and the private company Oseven Telematics, which provides insurance companies with telematics to apply UBI based on the methods and algorithms provided by the NTUA's research team in the Traffic Engineering Laboratory.
Data and analytics	**Types of data/ data sources/ access to data**	Two different advanced technologies are used to gather massive data on driving behaviour: 1) OBD-II devices installed in vehicles and 2) smartphones from users located freely inside the vehicles. Apart from those sources, other complementary data are constantly gathered and analysed based on questionnaire surveys, driving simulator experiments, real driving experiments to extract the perceived and revealed driving behaviour and correlate them to the massive data collected by smartphones and OBD-II devices.
	Open data (legislative framework required?)	During the project, the collected data are not shared outside the partners. Nevertheless, there is a memorandum of understanding between the company and the university which allows the use of the collected data for research purposes and in publications and presentations to conferences with the consent of Oseven.
	Methods, algorithms, tools for analysis	In-house apps are developed that automatically detect the beginning of the trip, record it and automatically send it to the server. The app is user friendly and provides useful information about the driving analytics per trip, per day, week month and overall. It also includes several gamification features: • Platform: A cloud service platform is developed to store and retrieve data for computations. All computations conducted are cloud-based. • Data cleaning and preparation: Filtering algorithms are implemented for noise reduction. Advanced detection algorithms are developed to detect erroneous data.

		• Driving Analytics: In house machine learning and other advanced computational algorithms are developed based on Deep Learning Neural networks, Decision Trees, Bayesian Networks to solve time series, classification and detection problems.
	Data ownership after processing	The ownership of the data (raw, processed) belongs to the private company, but the public can reuse the data for research purposes.
Outcome	Lessons learned: why did it work or fail?	The project started in 2015 and is ongoing. The first results were delivered in 2016. Regarding the data sharing, experience shows that there is a thin line between what can be published and what is confidential material, as well as what are the exact terms of data reuse.
	Does it meet national obligations?	n/a
	Efforts to educate general public/ staff	There is a constant research and educational collaboration with the University.
	Innovations it provides	The extended toolbox on methodologies, modelling approaches and code for driving analytics using smartphone data and driver's rating based on accident risk.
	New innovations	The delivery of extra features on data reliability involving the detection of driver/passenger, fraud analytics and the custom-made mode detection.
	Costs/ cost structure involved	The project costs amount to EUR 1.2 million and are paid by private venture capital funds.
	Relevance of the results	n/a

Italy: PLUG-IN Project

Country		Italy
Case study		PLUG-IN Project
Partnership	**Problem description**	For the real-time management of road traffic, a major challenge is to integrate data coming from independent heterogeneous sources.
	Value proposition for data sharing	The project's goal was to design and put in place an urban mobility platform to manage information from disparate sources and determine the status of current traffic. It was hoped it would reliably predict traffic conditions in the short to medium terms, define possible strategic intervention in the event of congestion and provide real-time information – collective or individual – to users. More specifically, the project aimed at: collecting and integrating data from heterogeneous sensors and media; elaborating on those data to aid decision making; and diffusing information.
	Partners involved	Unige, the National Research Council (CNR), Leonardo Finmeccanica, Softeco, Aitek, Ansaldo STS, and others, gathered in a consortium.
Data and analytics	**Types of data/ data sources/ access to data**	Examples of data collected were customised for the city of Genoa, which boasts the biggest Italian seaport and two large train stations. Those data included: traffic flows and travel times on segments (via distributed sensors in the network); forecasted peaks of mobility demand due to the arrival of ferries and trains (via the relevant schedule, also considering possible delays); significant sport or social event events (via the analysis of social media); energy consumption of trains; import/export freight flows to and from the seaport gates and unforeseen events, such as accidents.
	Open data (legislative framework required?)	Raw data and preprocessed data after data fusion and filtering were only shared between the projects partners. Filtered data have been used by partners to develop and test traffic-related applications (route guidance, road and railway state prediction, etc.). As the project has ended, the data may be unreachable but not confidential.
	Methods, algorithms, tools for analysis	Data was collected continuously, but a quality-based algorithm selected which data deserved storing and updating at any time.
	Data ownership after processing	n/a
Outcome	**Lessons learned: why did it work or fail?**	The most significant obstacles that occurred during the project were: harmonising the different time scales of data; understanding the reliability of data, especially for those collected in social media; the "translation" of raw data into traffic-related data; estimating the present and near-future state of the whole network (considering private traffic, train and ferry positions, etc.); determining easy-to-apply control policies for traffic optimisation, modes synchronisation, energy consumption reduction, etc.
	Does it meet national obligations?	No, it was an industrial research project.
	Efforts to educate general public/ staff	n/a

	Innovations it provides	Quality-based algorithm data selection and processing. Partners share a common database and communicate with it by means of drivers, still using their proprietary architectures.
	New innovations	n/a
	Costs/cost structure involved	The total cost of the project was EUR 4.5 million.
	Relevance of the results	The project demonstrated the capability of sharing a common platform decoupling different data providers (road, rail, and logistic data) and data users (application developers).

The Netherlands: KiM Mobility Panel

Country		The Netherlands
Case study		KiM Mobility Panel
Partnership	Problem description	This particular case is an example of open data: the Mobility Panel data are available for use by third parties. This contribution focuses on difficulties related to data dissemination and privacy issues.
	Value proposition for data sharing	Data sharing enables third parties to do research, e.g. identifying relationships between changes in travel behaviour, personal and household characteristics, and other mobility-influencing factors. This is of interest to governments, universities, research institutes and consultancy firms.
	Partners involved	The Mobility Panel is an initiative of the Netherlands Institute for Transport Policy Analysis (KiM), the consulting firm Goudappel Coffeng and the University of Twente. At the moment, TNS NIPO conducts the field work and the research institute CentERdata disseminates the data via the platform Survey Data Netherlands.
Data and analytics	Types of data/ data sources/ access to data	Each year the Mobility Panel produces a set of data from household questionnaires, individual questionnaires and travel diaries. A fieldwork report is delivered that includes: the data; questionnaires, travel diaries, survey instructions and reminders; a description of the sample design and sampling method; a description of the set-up and maintenance of the panel file; a description of the non-response analysis and the results obtained; a description of the data collection methods; an account of the data processing steps and; sets of cross-sectional weighting factors. The data sets are available for use by third parties free of charge. Retrieval, consultation and use of data subject to certain restrictions related to use, confidentiality and publications.
	Open data (legislative framework required?)	The data are available for use by third parties. There are and have been several difficulties in sharing the data. To protect the privacy of respondents, the data are stripped of all confidential or privacy-sensitive information before they are shared. This was more difficult than expected. The common privacy-sensitive data like date of birth and address were easily removed. But on top of that, combinations of privacy-insensitive data turned out to be privacy-sensitive as well. This was caused by the large amount of data, the data of households instead of individuals and the longitudinal character of the data. There have been technical difficulties as well. The data are disseminated via the Questacy system on a platform called Survey Data Netherlands. This system and the platform are particularly meant for disseminating longitudinal data. However, the system was designed based on questionnaires and corresponding answer categories.
	Methods, algorithms, tools for analysis	Not elaborated here, so to focus on data sharing challenges.
	Data ownership after processing	The data are owned by KiM. However, the panel that is used to collect the data is owned by the fieldwork company. This turned out to be a difficulty in continuing the Mobility Panel when the contract ends since the work has to be put out to tender under EU rules. This will probably be dealt with by continuing a partnership with the fieldwork company specifically for the panel.
Outcome	Lessons learned: why did it work or fail?	At the moment, the Mobility Panel data collected in 2013 and 2014 are available for use by third parties. As such, it has succeeded despite all the difficulties in sharing the data.

	Does it meet national obligations?	The resulting data do comply with the Dutch privacy legislation.
	Efforts to educate general public/ staff	There were no particular efforts to educate public or staff. However, it required a considerable effort to get the data ready for use by third parties. Preparing data for personal use is much easier than the process of getting the data ready for use by third parties. It requests much more extensive documentation. For example, the data collection and processing have to be explained to users that are not aware of these processes for the Mobility Panel. Furthermore, tracking changes and registering them systematically requires a big effort.
	Innovations it provides	Before introducing the Mobility Panel, cross-sectional travel data were already collected in the Netherlands. However, to identify the relationship between changes in travel behaviour, personal and household characteristics, and other mobility-influencing factors, longitudinal data are needed.
	New innovations	n/a
	Costs/ cost structure involved	KiM bears the costs of the fieldwork and the data dissemination of the Mobility Panel. KiM, Goudappel Coffeng, and the University of Twente all contribute to the Mobility Panel by delivering labour.
	Relevance of the results	With the Mobility Panel, there will be insights available into the factors that play roles in changing people's travel behaviour. This will lead to knowledge of the mobility of various groups of Dutch people, for example, adolescents, families with small children, and the elderly. When formulating policy, the Ministry can be more responsive to changes in mobility. Furthermore, these new insights can be included in modifications made to existing traffic and transport models. The government can use these models when taking decisions about traffic and transport investments.

Annex B. List of Roundtable participants

Patricia HU (Chair), Director, Bureau of Transportation Statistics at US Department of Transportation

Marie ARBOUET, Systra, France

Mario BARRETO, Lead Statistician, International Transport Forum (ITF)

Florian BERKES, Data Scientist, Umlaut, Germany

Patrick BONNEL, Chief of Transport Department, –Graduate School of Civil, Environmental and Urban Engineering (ENTPE), Lyon University, France

Aurelie BOUSQUET, Project Officer Travel Survey and Modelling, Cerema, France

Norbert BRAENDLE, Senior Scientist, Austrian Institute of Technology (AIT), Austria

Robin CAMBERY, Chief Transport Modeller and Head of Modelling and Appraisal Methods, Department for Transport, United Kingdom

Aman CHITKARA, Manager, Mobility, World Business Council for Sustainable Development (WBCSD), Switzerland

Julie CHRÉTIEN, Project Manager, 6-t, France

Rodrigo CONTRERAS, Head of Methodological Development Unit, Ministry of Transport and Telecommunications, Chile

Philippe CRIST, Advisor, Innovation and Foresight, ITF

Markus FRIEDRICH, Chair of Transport Planning and Traffic Engineering, Professor, Stuttgart University, Institute for Roads and Transportation, Germany

Fredrik GREGERSEN, Researcher, Norwegian Centre for Transport Research (TOI), Norway

Paulo HUMANES, Vice President Business Development, PTV, Germany

Thibault JANIK, Systra, France

Eric JEANNIERE, Modeller/Analyst, ITF

Jari KAUPPILA, Head of SG's Office and Head of Quantitative Policy Analysis and Foresight, ITF

Anurag KOMANDURI, Principal Consultant, Cambridge Systematics, United States

Fabien LEURENT, Research Director, Ecole des ponts ParisTech, Laboratoire Ville-Mobilité-Transport, France

Tomas LEVIN, Senior Principal Engineer, Norwegian Public Roads Administration (Vegvesen), Norway

Patrick MALLEJACQ, Secretary-General, PIARC, France

Vasco MORA, Advisor to the Deputy Mayor, City of Lisbon, Portugal

Viviana MUÑOZ, Head of Big Data for Transport Unit, Ministry of Transport and Telecommunications, Chile

David O'NEILL, CEO, Kisio Consulting, France

Christophe PAUWELS, Attaché, SPF Mobilité, Belgium

Stephen PERKINS, Head of Research and Policy Analysis, ITF

Florence PRYBYLA, Head of Modelling Department, SNCF-Transilien, France

Marek RANNALA, Mobility Statistics Product Manager, Positium, Estonia

Josep Maria SALANOVA GRAU, Associate Researcher, Centre for Research and Technology Hellas, Hellenic Institute for Transport (CERTH HIT), Greece

Alexandre SANTACREU, Policy Analyst, Road Safety, ITF

Danyang SUN, PhD Candidate, École des ponts ParisTech, Laboratoire Ville-Mobilité-Transport, France

Per-Olof SVENSK, Project Manager, Swedish Transport Administration, Sweden

Cristina VALDES, Consultant, Municipal Transport Company of Madrid, Spain

Stefaan VERHULST, Co-Founder, The GovLab, United States

Vladimir VOROTOVIC, Senior Manager - Innovation and Deployment, ERTICO - ITS Europe, Belgium

Luis WILLUMSEN, Board Member, Nommon, United Kingdom

International Transport Forum

Big Data for Travel Demand Modelling

This report examines how big data from mobile phones and other sources can help to forecast travel demand. It identifies the strengths and potential use-cases for big data in transport modelling and mobility analysis. It also examines potential biases, commercial sensitivities and threats to privacy. The report presents approaches to resolve such issues and offers recommendations for governance arrangements that make data sharing easier.

All resources from the Roundtable on Big Data for Travel Demand Modelling are available at:
https://www.itf-oecd.org/big-data-transport-models-roundtable

International Transport Forum
2 rue André Pascal
F-75775 Paris Cedex 16
+33 (0)1 73 31 25 00
contact@itf-oecd.org
www.itf-oecd.org

OECD

Published: 08/2021 | Photo credit: Maria Savenko/Shutterstock