Institute of Transport Economics
Norwegian Centre for Transport Research

ENGLISH Summary

Valuation based on Big Data and revealed preference data

An assessment for Norwegian transport appraisal

TØI Report 1882/2022 • Authors: Stefan Flügel, Christian Weber, Askill H. Halse, Ingunn O. Ellis •
Oslo 2022 • 90 sider

For several decades, stated preference (SP) studies have been the dominant method for transport valuation. However, there are many indications that revealed preference (RP) data is making a strong comeback due to access to Big Data and new analysis possibilities such as machine learning.

In this report, we assess the capability of different RP data sources. We find that app-panel with GPS-tracking give the broadest and most precise bases for valuation. In order to accommodate current segmentation of unit values in Norwegian transport appraisal, one does, however, need to collect additional background surveys. The use of traditional travel surveys is also ranked high, in particular when synergies with the estimation of transport models can be realized.

Background and motivation

While SP studies build on an analysis of hypothetical choices in experimental settings without real-world consequences to the respondents, RP-choices are observed in real-world settings and therefore the preferred method to derive preferences. However, with RP data the researcher has little control over the data and little variation and/or high correlation in is a persistent challenge in RP-based estimation of unit values. This challenge can partly be overcome with larger data volume, which is more and more available due to the raise of Big data. Figure S1 summarises main advantages of RP data in general and Big Data in particular and how this may contribute to more valid and more up-to-date unit values for Norwegian appraisal.

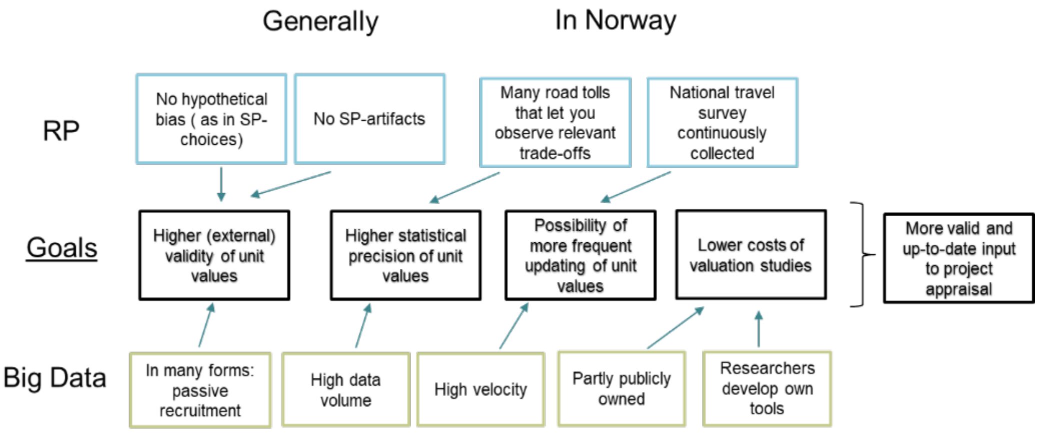
Generally
In Norway
RP
No hypothetical bias (as in SP-choices)
No SP-artifacts
Many road tolls that let you observe relevant trade-offs
National travel survey continuously collected
Goals
Higher (external) validity of unit values
Higher statistical precision of unit values
Possibility of more frequent updating of unit values
Lower costs of valuation studies
More valid and up-to-date input to project appraisal
Big Data
In many forms: passive recruitment
High data volume
High velocity
Partly publicly owned
Researchers develop own tools

Figure S1: Overview over motivation of use of revealed preference (RP) data and Big Data for transport valuation.

Work tasks and method

The conclusions and recommendations in this report are based on:

1) A literature review on valuation based on RP/Big Data
2) A list up of possible data sources and a discussion of their relevance for valuation.
3) An assessment of relevant combination of data sources and unit values based on 19 different criteria. Scores are given on a scale of 1 to 5. The scoring was partly based on an internal Delphi survey.
4) A synthesis of the assessment in three groups of criteria: “Access and general quality”, “Analysing opportunity for valuation” and “Flexibility, synergies and future perspective”
5) A practical description of three of the most promising approaches
6) A case study to illustrate some challenges of aggregated data sources

Data sources

For a data source to be relevant for valuation, the following need to apply:

1) The data need to be available in Norway or there needs to be clear path to how it can be made available.
2) The data set must give direct or indirect information on the behaviour of travellers, either in the form of individual choices or in the form of aggregated market shares.
3) The data set needs to enable the attachment of relevant and sufficiently precise attributes to the different alternatives in the choice set.
4) Some of the choices that are observed need to imply an actual trade-off between at least two attributes that are relevant for the underlying unit value. Attributes, like time and cost, can be positively correlated (and they often are in practice), but there needs to be some mechanisms (at least for a subset of choices) where variation in the data is invoked (e.g. through road tolls).Table S1 provides an overview of the included data sources and their main characteristics.

Table S1: Overview of data sources.

Data source	Technology	(Assumed) data owner / access for researcher	(Assumed) level of aggregation of data	Most applicable choice context / unit value
National RVU	Traditional travel survey	Transport authority / free	Disaggregated (trips of single persons)	Mode choice / various
Mobile data	Call Detail Record via cell towers	Commercial providers as Telia / costly	Aggregated (BSU or routes)	Route choice (mainly long distance) / VTTS car
App panel with GPS-tracking	GPS/A-GPS , GNSS	Researchers / free access to own panels	Disaggregated (trips of single persons)	Mode- and Route choice / various
Automatic traffic counters (ATC)	Sensors (typically electrical induction)	NPRA / free	Aggregated (points)	Route choice / VTTS car
Toll transaction data	ANPR cameras and RFID devices	NPRA / free (limited as of today)	Disaggregated (cars over different points)	Route choice / VTTS car
Tracking data from commercial providers	Various (GPS, Navigation devices,..)	Commercial provider as TomTom or fitbit / costly	Aggregated (BSU or routes)	Route choice / VTTS car
Dedicated cameras and sensors	Various (ANPR, RFID, bluetooth tracking and magnetic sensors	Researchers / free access to self-installed hardware and data	Disaggregated (cars over different points)	Route choice / VTTS car
Mobility-as-a-Services ordering data	Stored data from apps	MaaS providers as Bolt or Ruter / unclear of today	Disaggregated (trips of single persons)	Various / VTTS (waiting time)
Automatic passenger counts (APC)	Various (camera technology, mobile phone tracking and/or light barriers)	PT providers / free (some restrictions)	Aggregated (station-pair/departure)	Submode-departure choice / crowding multiples
Camera-based crowd counts at stations	Cameras (supported by machine learning)	Researchers / free access to self-installed hardware and data	Aggregated (station/departure)	Wait for next departure at station / crowding multiples

Summary of assessment

Data access and general quality was assessed based on the following criteria:

- Access to relevant and updated RP data
- Resources required for data access and maintenance (high score for low resources needed by the executing body of the valuation study; original costs by others not included)
- Resources required for data processing (high score for low resources needed by the executing body of the valuation study; original costs by others not included)
- Data volume
- Coverage (high score if all of Norway is covered)
- Representativity

While the latter 3 criteria may depend on the unit value of interest, the total scores for this group of criteria is rather stable across different relevant unit values.

The criteria for Opportunities for analysis for valuation were:

- Observation of actual choices
- Quantification of attributes and costs of chosen alternative
- Identification/modelling of non-chosen alternatives (choice set)
- Quantification of attributes and costs of non-chosen alternative
- Variation and correlation in central attributes
- Possibility to control for other effects
- Possibility to segment (current segmentation)
- Possibility for combined RP-SP models and other advanced estimation methods

The last group of criteria encompasses flexibility, synergies and future perspective of the data sources. This group is assessed from a general perspective and not from the perspective of the researchers (as the two previous groups). The following criteria where included:

- Possibility to frequent and continuous data collections in future
- Possibility to segment results beyond current segmentation
- Synergies with transport models
- Other synergies
- Relevance for new trends/technologies

Figure S2 gives an overall ranking of the evaluated data types. The scores for opportunity for analysis for valuation apply to the unit value with the best score within each data type.

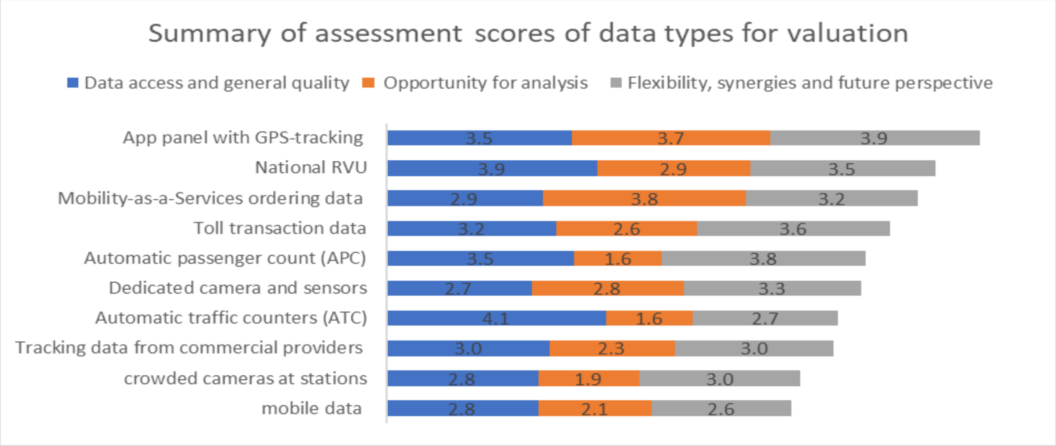

Figure S2: Overall ranking of RP-data types for valuation.

App panel with GPS-tracking is ranked highest overall.

The scores for Opportunities for analysis vary with the underlying unit values.

Besides the total scores, an important information is also how many unit values the data source in applicable for. Table S2 summaries our findings.

Table S2: Number of applicable unit values and range of total scores for Opportunity of analysis for estimation of unit values

Data source	Number of unit value data is applicable*	Total score	Main advantage	Main disadvantage
National RVU	6	2.2- 2.9	covers current requirement for segmentation	imprecise spatial information
mobile data	2	1.7-2.1	somewhat better control over routes compared to ATC, at least for long distance	little control and possibility for segmentation; works poorly for short distance routes
App panel with GPS-tracking	10	3.3-3.7	detailed information on routes	trip purpose unreliable observed
Automatic traffic counters (ATC)	1	1.6		routes not directly observed
Toll transaction data	2	2.6	can distinguish car types	works only in networks that contain road tolls
Tracking data from commercial providers	2	2.1-2.3	better control over route than mobile data and ATC	little background information
Dedicated cameras and sensors	4	2.7-2.9	good control over routes given good sufficient coverage of cameras	trip purpose not observed
Mobility-as-a-Services ordering data	2	3.5-3.8	direct and precise information on attribute values	trip purpose not observed, open the app likely endogenous
Automatic passenger counts (APC)	1	1.6		OD not directly observed
crowded cameras at stations	1	1.9		Works only under specific conditions

Illustrations and case study

The report also contains a more practical description of three of the most promising approaches (National RVU, Fotefar, which is a upcoming GPS-app tracking software, and toll transaction data) as well as a case study using aggregated data sources (traffic counts, mobile data and data from TomTom). The latter illustrates some of the practical difficulties in using aggregated data to derive unit values.

Conclusion and recommendation

Below we summarise our main conclusions:

1) As of today, travel surveys such as **national RVU** are the most relevant data source with regard to the current segmentation of unit values

which require information about travel purposes. There are large potential synergies with transport models and one should consider aligning the next RTM estimation with the next valuation study. In this connection, it may be appropriate to move away from the current RVU, and rather design a more tailored survey that is better suited for both demand modelling and valuation.

2) **Data from apps that can track individuals** with GPS or other high resolution/frequency sensors score overall best in our assessment. The ability to add background information is important. This may require additional data collection, for instance in form of surveys.

3) A combination of surveys (and/or register data) and GPS tracking is considered the best option and something that is recommended to work towards.

4) **Aggregated data** (e.g. counting data on roads and public transport) place great constraints on analysis opportunities and will hardly be sufficient for national unit values given requirements coverage and in the current segmentation. That said, it can – based on appropriate case studies – help to validate the absolute level of the value of time (VTTS).

5) Aggregated **mobile data** provides better analysis options compared to counting data, at least for intercity travel, but is quite expensive to get access to. As other aggregated data sources it has clear limitations compared to more disaggregated data sources.

6) **Toll transaction data that tracks individual cars** will be able to provide information of route choice of individuals or groups in areas with a good coverage of road tolls and there are different possibilities to add individual background variables. Such data would in most cases not be completely anonymous, but access to non-anonymous data for research purposes would most likely be feasible under the current data protection legislation. However, facilitating access to data would require some goodwill and effort of the owners of the data. A more flexible (but more expensive) alternative to this data is to set up d**edicated cameras for automatic number plate recognition** (ANPR).

7) **Aggregated App-data from commercial enterprises** can also be a promising alternative. NPRA has access to aggregated tracking results from e.g. TomTom, a data source which could be utilized more for studying route choice behaviour, e.g. at toll roads across the country. In order to use TomTom data for research, access to more information about data collection and data processing, and the possibility the share this information with the public, are crucial. There are currently also major limitation in sharing data and publishing results from data analysis.

8) Most data sources mentioned under 4) – 7) have a fundamental advantage in their passive recruitment. The data sources are therefore interesting for the quality assurance of survey and app-based studies where unobservable factors can affect the level of the VTTS due to sample selection bias. That said, there can also be some biases in the sample of mobile companies and app-data providers.

9) A disaggregated data source with great potential are **MaaS ordering data** (e.g. from raid-hailing services). It is currently limited in access and application. In Norway studying choices/preference for micro-mobility seems most applicable. This type of data might also be made available via future versions of more traditional PT apps (e.g. via a future version of the Ruter-app that may let travellers pick, order and pay for all available transport solutions).

We see three approaches for the next valuation study. They are given below in ranked order.

1) **GPS-tracking data plus background surveys.** The recruitment should come from a combination of large (existing) samples or – preferably – the population register. Economic incentives should be given for donating tracking data to the project as this is likely to attract a broader sample and can therefore reduce the danger of sample selection biases. From a modelling perspective, combined mode and route choice models are likely to give the best and broadest basis for unit value estimation. The background survey should include questions on mode, car type and ticket type availability and include information about the location of home, work and other points of interest of respondents such that trip purpose can be derived from the spatial information in the GPS data. In addition, small SP experiments could be included in the background survey for cross-validation and for estimation of unit values that may be difficult to estimate from RP data.
2) **National RVU or – preferable – a tailored travel survey in a joint estimation with the RTM model.** Compared to suggestion (1.), this approach puts less weight on precise data and emphasizes consistency and synergies with transport models. The unit values would be derived from the mode choice utility function of the mode/destination choice models that are part of the RTM model system. Fitting route choice models in the network assignment tool (e.g. CUBE) against aggregated data sources can in addition support the estimation/recommendation of unit values. It is highly recommended that spatial information from the travel survey data is provide with 8 digit BSU ("grunnkrets") codes throughout (i.e. annul the current practice of providing BSUs with less than 100 inhabitants with 6 digit codes). With that, the level of precision will still be far below GPS-tracking, but should be acceptable within this approach.
3) A third approach would be to **keep the stated preference** approach. In this case, we would recommend **to use several well-crafted RP case studies to validate/adjust the overall level of VTTS**. Combined RP-SP models would be recommended in order to utilized the advantages of both data types. In this connection it would be preferable to recruit part of the SP sample from the areas where the RP case studies are conducted.