Analysis of household survey sample size in trip modelling process

Tomasz Kulpa a,*, Andrzej Szarata a

*Politechnika Krakowska, Warszawska 24, 31-155 Kraków, Poland

Abstract

Large cities are developing sophisticated trip models that consist of demand and supply part. Since supply model is quite easy to obtain and develop, demand model needs more input data. In particular Comprehensive Travel Surveys (CTS) are conducted to feed the demand model. The core part of the CTS are household trip questionnaires which are used, among others, to identify residents’ trip patterns. These surveys are very expensive and often determine CTS frequency. Many cities, also large ones, cannot afford to conduct these kinds of survey as often as it should be done.

In this paper the authors analyzed possible ways to reduce CTS cost. They verified whether survey sample size reduction would not cause rapid error increase. Based on CTS for Krakow metropolitan area authors compared results between full sample (6000 households inquired) and reduced samples (2400, 1000 and 750 households inquired). The differences between results obtained from different samples are perceptible but not significant. Therefore, the authors proved that household survey sample size might be reduced even to 1000 inquired households with acceptable loss of results quality.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of Road and Bridge Research Institute (IBDiM)

Keywords: comprehensive travel survey; sample size; trip modelling

* Corresponding author. Tel.: +48 12 628 23 25; fax: +48 12 628 25 35.
E-mail address: tkulpa@pk.edu.pl
1. Introduction

Nowadays, during the development process of transportation systems, it is necessary to conduct sophisticated analysis of interaction between different modes of transportation. When we take into consideration urban sprawling process and changes in travel behaviors of residents we obtain complex trip patterns. These must be modeled using reliable input data and proper simulation tools. Normally in Poland a four stage approach is used for trip modeling, as suggested in JASPER (2014), which could be modified according to data availability. The four stage model structure consist of following parts (Garber, 2001):

- Trip generation. According to obtained results of Comprehensive Travel Survey (CTS), it is possible to define the relationship between generated trips and spatial development for particular trip purposes. These formulas usually have a linear regression character.
- Trip distribution. The result of this stage is O-D (origin – destination) matrix which represent spatial distribution of generated trips in the analyzed area. O-D matrix is calculated for particular trip purposes using gravity model, with parameters estimated on the basis of CTS results.
- Modal split (Mode choice). In this step the share of different modes of transport in total trips is estimated. Mode choice models usually have a logit character and use either generalized travel cost or travel times quotient of different transport modes as a utility function.
- Trip assignment. This step involves searching the shortest path between defined TAZ’s. There are plenty of procedures, from the simplest ones (single route, all or nothing), to more sophisticated (multi route, stochastic equilibrium).

Steps listed above are commonly called demand model. On the other hand, we have a supply model, which covers the development of transportation network. Usually, it is assumed, that the street network is modeled using a graph theory. Intersections are represented by nodes while streets are represented by links, connecting adequate nodes. Each link has its own parameters (e.g. capacity, free flow speed (PTV, 2014)) which allows classification of street network. Public transport lines can be treated as a separated network but also it is possible to share same links as for private transport systems. Bus and tram lines (as well as other modes of PT) require detailed route modeling, locations of stops, travel times between stops and timetables or headway. To support the process of supply model data collection GIS datasets are widely used (Johnston and Barra, 2000; Zhong and Hanson, 2009).

After applying the mentioned approach (demand and supply), the model must be assessed using measurement datasets, both for public and private transport systems. Measurements refer to number of PT passengers on links (as a total volume and volume for each PT line) as well as to traffic volume, traffic composition and turns on junctions for private transport. A common approach is to verify correlation between measured and modeled volumes as well as assessment of GEH factor (DMRB, 1996). The model verification quite often requires a more detailed approach such as comparison of trip length and duration distribution or compliance between measured and respective modeled values for particular screenlines (Schiffer and Rossi, 2009). Measurement datasets are the basis for further model calibration and provides information about its quality (Szarata, 2011).

Supply and demand model are developed in parallel. On the one hand data to develop supply model is widely available, also from on-line and free of charge resource. On the other, to obtain a reliable demand model the most important part is data collection. Exceptionally important are interviews with the residents (household travel surveys). These are the basis for mathematical description of the trip patterns in the analyzed area. The quality of trip patterns model is highly depended on the sample size of household travel survey. Additionally, sample size depends on analyzed area size (HTM 2005). Within the proposed study the provision of a different sampling procedure for the travel survey is assumed which leads to a significant decrease in the total costs of the survey.

2. Travel surveys

The most expensive and time consuming part of the transportation model development process is travel data collection. There are many ways of data collection (Richardson et al., 2005), such as worksite interviews, on-board surveys (Garber and Hoel, 2001), GPS household survey diaries, application of mobile phone trajectories (Friedrich
et al., 2010), internet survey etc. Jeon et al. (2010) proposed an approach in which TAZ’s are aggregated into bigger ones (for city of Seoul the aggregation level was from 522 TAZs to 86). The results showed that it is possible to create such aggregate models and the survey was not aimed at a data collection process. Crevo (1991) presented similar results, providing evidence of a comparable level of accuracy for forecasting of travel demand in network with reduced TAZ’s number. Similar problems were mentioned in Bovy and Janses (1983) and Khatib et al. (2001). Reducing of the sample size was also investigated in Stopher et al. (2008), where application of GPS was proposed to validate results and weighting adjustment. In fact, GPS is quite often applied to the household survey (Swamm and Stopher, 2008; Giaimo et al., 2010) but in this case it is used as a tool for data collection and the authors are focused on different ways of improving the obtained results e.g. focus group meetings or direct interviews supporting GPS data. In Polish conditions usage of GPS or mobile phone data is still hardly possible due to personal data protection. Conducting one of the surveys (CTSMV, 2015) authors made an attempt to obtain data from mobile devices via Google applications to estimate through, inbound and outbound trips in analyzed area and unluckily defeated. Within other study, which ought to be carried out till 2016, mobile phone data from base transceiver stations is planned to use for trip modeling purposes. Nevertheless, still the most popular and reliable surveys in Polish conditions are traditional PAPI (Paper And Pen Interview) household travel surveys. It is a quite expensive and relatively complicated method, but the results used for further demand model development gives satisfactory results. Comprehensive Travel Study (CTS) has revealed a preferential character and is conducted as a household survey in the modeled area divided into smaller units (TAZs). In Polish conditions such surveys usually have retrospective character and the respondent is asked about all his trips from the previous day. The CTS questionnaire consists of questions concerning travelling (e.g. origin and destination address, mode of transport, travel time, trip purpose etc.) as well as questions concerning travel preferences and opinion about the transport system. Full interview lasts approximately 30 minutes. The second important issue is the sample size defined as a number of households which are randomly selected to be inquired (HTM, 2005; Schnabel and Lohse, 1997). According to HTM (2005), the recommended sample size depends on number of residents in the analyzed area (see Table 1).

<table>
<thead>
<tr>
<th>Population in analyzed area</th>
<th>Suggested sample size (%)</th>
<th>Minimum sample size (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>below 50 000</td>
<td>20.0</td>
<td>10.0</td>
</tr>
<tr>
<td>50 000–150 000</td>
<td>12.5</td>
<td>5.0</td>
</tr>
<tr>
<td>150 000–300 000</td>
<td>10.0</td>
<td>3.0</td>
</tr>
<tr>
<td>300 000–500 000</td>
<td>6.7</td>
<td>2.0</td>
</tr>
<tr>
<td>500 000–1 000 000</td>
<td>5.0</td>
<td>1.5</td>
</tr>
<tr>
<td>above 1 000 000</td>
<td>4.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

In practice, the suggested sample size (according to Table 1) should be used for the first CTS. If CTS was conducted previously, it is acceptable to apply minimum values. Nevertheless, sample size along with long lasting questionnaires implies a high level of costs of CTS. Depending on the area size and survey scope the cost of CTS in Polish cities can be about $ 0.4 million (Szczećin, population of 400 thousand, CTSS (2013)) or even exceed $ 0.5 million (Gdansk, population of 460 thousand, CTSG (2009)). Of course the mentioned costs also consists traffic volume measurements, but still household shares comprise the most significant amount in the total budget (even one third).

Sample size of the CTS derives from further necessity of sample weighting to provide reliable results for whole population. A very important issue is sampling error. Sampling error is an error that results from choosing to measure a sample as representative of a much larger population (Garber and Hoel, 2001). It is assumed, that the error is random in occurrence, usually has a normal distribution and decreases accordingly to number of measurements made (Kish, 1965).
In Fig. 1 sampling error for CTS in Krakow is presented, depending on sample size (number of households inquired) and confidence level. In Fig. 1, it may be seen that the sampling error increases with sample size reduction.

![Sampling error for analyzed area.](image)

The data obtained from household surveys forms the basis for demand model development. Usually it uses single or multiple regression (linear or nonlinear) formulas to estimate parameters of particular trip modeling procedures, e.g. number of trips generated by TAZ, gravity and mode choice models parameters etc. The calibration process of mentioned values requires data from household survey sample, which later, using statistical tools, is weighted. The background of such an approach is that a proper sampling process (adequate sample size, random choice of household and their representative character) gives results which could describe relationship in the whole survey area.

3. Analysis of sample size reduction

As shown before (see Fig. 1) one of the significant factors influencing survey error is sample size. On the other hand sample size generates additional costs which, as was already highlighted, is the main part of CTS cost.

Based on the results of CTS in Krakow metropolitan area (CTSK 2014) comparison of results for four sample sizes was done. In particular, following samples were considered: full (6000 households) and reduced (2400, 1000 and 750 households). Reduced samples were representative and drawn from full sample. For the comparison of different sample sizes the demand model for each was calculated according to 4-stage approach. In particular, following issues were compared: trip generation, trip distribution and modal split.

3.1. Trip generation

Analysis was conducted for seven trip purposes (see Table 2). For particular trip purposes trip generation formulas were estimated, both for production and attraction. Single and multiple linear regression models were used. Next, for each sample size and trip purpose trip generation rates were calculated. These were compared to trip generation rates calculated using the full sample. In majority of cases R-squared values exceeded 0.7, which is satisfactory.
Table 2. Relative errors in trip generation [%].

<table>
<thead>
<tr>
<th>Trip purpose</th>
<th>Sample 2400</th>
<th>Sample 1000</th>
<th>Sample 750</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home-Work</td>
<td>2.0</td>
<td>5.9</td>
<td>2.7</td>
</tr>
<tr>
<td>Work-Home</td>
<td>-9.1</td>
<td>-4.6</td>
<td>-8.0</td>
</tr>
<tr>
<td>Home-Education</td>
<td>-2.8</td>
<td>-6.6</td>
<td>-1.8</td>
</tr>
<tr>
<td>Education-Home</td>
<td>-30.0</td>
<td>-24.7</td>
<td>-19.0</td>
</tr>
<tr>
<td>Home-Other</td>
<td>3.1</td>
<td>4.5</td>
<td>14.5</td>
</tr>
<tr>
<td>Other-Home</td>
<td>-5.1</td>
<td>1.8</td>
<td>8.3</td>
</tr>
<tr>
<td>Non-Home Based</td>
<td>-15.9</td>
<td>-12.2</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Depending on trip purposes the difference between regression coefficients for full and reduced samples was between 2 and 30%. Relative Errors in Table 2 were calculated as a comparison between trip generation calculated for whole sample with trip generation calculated for reduced sample.

Negative values in Table 2 mean that trip generations for reduced sample are lower than for the full sample. The highest relative errors where achieved for two trip purposes: Education-Home and Non-Home Based. For other cases errors are lower than 10% which is satisfactory. Similar results were obtained for attractions, thus it will not be presented in this paper. Nevertheless, it can be stated that, in general, trip generation rates obtained from reduced sample are close to those from the full sample.

3.2. Trip distribution

Trip distribution was conducted using a gravity model. It was analyzed for four trip purposes groups, namely: Home-Work-Home, Home-Education-Home, Home-Other-Home, Non-Home Based. Since all trips were considered (motorized and non-motorized) exponential distance-decay function was used:

\[ f(l_{ij}) = a \cdot e^{-cl_{ij}} \]  

(1)

where: \( l_{ij} \) – distance between traffic zones, \( a, c \) – model parameters.

Parameters of distance-decay function for full and reduced samples for all trip purposes groups are shown in Table 3. In all cases R-squared are higher than 0.80. Regression parameters do not differ significantly from particular samples. However, the biggest difference is observed for trips related with education and other.

Table 3. Distance-decay function parameters.

<table>
<thead>
<tr>
<th>Trip purpose</th>
<th>Full sample</th>
<th>Sample 2400</th>
<th>Sample 1000</th>
<th>Sample 750</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>c</td>
<td>a</td>
<td>c</td>
</tr>
<tr>
<td>Home-Work-Home</td>
<td>0.084</td>
<td>-0.133</td>
<td>0.079</td>
<td>-0.124</td>
</tr>
<tr>
<td>Home-Education-Home</td>
<td>0.158</td>
<td>-0.308</td>
<td>0.122</td>
<td>-0.254</td>
</tr>
<tr>
<td>Home-Other-Home</td>
<td>0.143</td>
<td>-0.300</td>
<td>0.136</td>
<td>-0.253</td>
</tr>
<tr>
<td>Non-Home Based</td>
<td>0.096</td>
<td>-0.179</td>
<td>0.103</td>
<td>-0.184</td>
</tr>
</tbody>
</table>

In Fig. 2 and Fig. 3 comparison of distance-decay function graphs are presented. Obtained shapes of distance-decay functions for particular groups of trip purposes are very similar. Considering Home-Education-Home and Non-Home Based trip purposes difference of 10% between full and reduced samples can be seen in the first part of function graph. The difference disappears at the length of 4 km. The most alike function graphs were obtained for Home-Work-Home trips. For Home-Other-Home trips function graphs differ between full and reduced samples at trips length range from 5 to 10 km.
3.3. Modal split

First step in modal split is separation of pedestrian trips. Most often exponential function based on trip length is used, as shown below:

$$ u_p = a \cdot e^{b \cdot l_{ij}} $$

(2)

where: $l_{ij}$ – distance between traffic zones, $a$, $b$ – model parameters.

Pedestrian trips share function obtained using full sample was very similar for all trip purposes (CTSK 2014). Thus one function was considered. For particular samples model parameters and R-squared values were calculated (see Table 4). For all samples, except the smaller one, estimated parameters does not differ significantly.

Table 4. Pedestrian trips share function parameters.

<table>
<thead>
<tr>
<th>Trip purpose</th>
<th>a</th>
<th>c</th>
<th>R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full sample</td>
<td>1.365</td>
<td>-0.563</td>
<td>0.97</td>
</tr>
<tr>
<td>Sample 2400</td>
<td>1.394</td>
<td>-0.558</td>
<td>0.95</td>
</tr>
<tr>
<td>Sample 1000</td>
<td>1.317</td>
<td>-0.545</td>
<td>0.95</td>
</tr>
<tr>
<td>Sample 750</td>
<td>1.521</td>
<td>-0.631</td>
<td>0.96</td>
</tr>
</tbody>
</table>

For the confirmation of the results given in Table 4 pedestrian trips share function graphs were compared. As it can be seen in Fig. 4 for all samples function graphs are very similar in whole range.
Next step of modal split is division of trips between public transport (PuT) and private transport (PrT). A variation of logit model was used, where utility function is based on travel time. Following function was used:

\[ u_{ki} = \frac{e^{-I_{PrT/PuT} c}}{1 + e^{-I_{PrT/PuT} c}} \]

where \( I_{PrT/PuT} \) – PrT to PuT travel times quotient, \( c \) – model parameter.

Analyzing modal split for the full sample no significant difference was obtained between Home-Work-Home, Home-Other-Home and Non-Home Based trip purposes. Therefore, these has the same model parameters (see Table 5).

Table 5. Modal split function parameters.

<table>
<thead>
<tr>
<th>Trip purpose</th>
<th>Full sample</th>
<th>Sample 2400</th>
<th>Sample 1000</th>
<th>Sample 750</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home-Work-Home</td>
<td>-0.537</td>
<td>-0.349</td>
<td>-0.601</td>
<td>-0.301</td>
</tr>
<tr>
<td>Home-Other-Home</td>
<td>-0.537</td>
<td>-0.349</td>
<td>-0.601</td>
<td>-0.301</td>
</tr>
<tr>
<td>Non-Home Based</td>
<td>-0.537</td>
<td>-0.349</td>
<td>-0.601</td>
<td>-0.301</td>
</tr>
</tbody>
</table>

For Home-Education-Home trips modal split function graphs are very similar (see Fig. 5b). There are only slight differences in range of quotient from 0.5 to 1.0. In other trip purposes there is a significant difference between particular sample pairs: full & 1000 and 2400 & 750. For the full and 2400 sample modal split curve is more steep than for other sample sizes (see Fig. 5a). The difference may results from chosen frequency intervals or sample structure.

Fig. 4. Comparison of pedestrian trips share function graphs.

Fig. 5. Comparison of PrT/PuT modal split function graphs: (a) Home-Work-Home, Home-Other-Home and Non-Home Based; (b) Home-Education-Home.
4. Summary

Summarizing the results given in this paper it can be stated that there is a possibility of sample size reduction with acceptable error increase. Four samples were analyzed: 6000 (full), 2400, 1000 and 750 households. Except modal split function for Home-Work-Home, Home-Other-Home and Non-Home-Based trips and trip generation for Education-Home trip the results are satisfactory. Errors occurred are noticeable, but still acceptable.

In general, at every step of demand model development similar results were obtained for reduced samples as for full sample. There is an odd difference in modal split functions which might be the result of sampling or considering all purposes together (except Home-Education-Home) instead of its separation. For trip generation the biggest errors might be caused by underestimation of Home-Education-Home and Non-Home Based types of trips in reduced samples.

Nevertheless, authors proved the possibility of sample size reduction in Comprehensive Travel Survey. As a result CTS household survey cost may be reduced by almost 50% (fixed cost are not depended on sample size). Consequently, municipalities may repeat reduced CTS for instance every two years and conduct full CTS every 10 years. The suggested size of reduced sample for two-year annual survey is 1000 households. This gives acceptable results with reasonable cost effort. Reduction to 750 is not much cheaper, but causes bigger error increase. On the other hand, reduction from 2400 to 1000 gives still good results but significant reduction of cost. The other issue are commercial vehicle trips. Since CTS are focused mainly on passenger trips, goods deliveries are not investigated in acceptable range. Just two freight studies were conducted in last ten years in Poland. Additionally goods deliveries are more random than passenger trips. Thus analysis of sample size reduction in commercial vehicle trips is troublesome. In this case more likely would be to use other data collection methods, for instance as listed below.

There’s a big potential in usage of internet, mobile phone data, GPS devices, ITS (e.g. tolling systems) and crowdsourcing. In case of mobile phone and GPS devices the main barrier is still data availability due to personal data protection. However, step-by-step, the attitude of data managing bodies is changing and each year more datasets are available (mainly measurements results). On the other hand internet and crowdsourcing as well as open data seems to be inexpensive way of data collection. However some concerns might be raised on sample randomness and representativeness.

References


Comprehensive Travel Study in Kraków, CTSK 2014, Kraków (in Polish).

Comprehensive Travel Study in mazowieckie vovoideship, CTSMV 2012, Kraków (in Polish).


Design Manual for Roads and Bridges, DMRB 1996, Volume 12, Section 2, Part1,


Schiffer, R., Rossi, T., 2009, New calibration and validation Standards for Travel Demand Modeling, Transportation Research Board Annual Meeting, Washington D.C.
Zhong, M., Hanson, B., 2009, GIS-based travel demand modeling for estimating traffic on low-class roads, Transportation Planning and Technology, Vol. 32, No. 5.