Conference Paper

Multi-Level Weighting of Travel Survey Results

Author(s):
Fourie, Pieter Jacobus; Müller, Kirill

Publication Date:
2011

Permanent Link:
https://doi.org/10.3929/ethz-a-006742077

Rights / License:
In Copyright - Non-Commercial Use Permitted
MULTI-LEVEL WEIGHTING OF TRAVEL SURVEY RESULTS

Pieter Jacobus Fourie\textsuperscript{a} and Kirill Müller\textsuperscript{b}

\textsuperscript{a} Future Cities Laboratory, Singapore-ETH Centre for Global Environmental Sustainability (SEC), Singapore
E-Mail: fourie@ivt.baug.ethz.ch

\textsuperscript{b} Institute for Transport Planning, ETH Zurich, Zurich

ABSTRACT

We apply multi-level iterative proportional fitting (IPF) procedures to the problem of inflating travel survey data to simultaneously match control totals in two domains, using the Singaporean Household Interview Travel Survey (HITS) of 2008 as a case in point. In this survey, household trip records need to be inflated to match household control totals that come from census-based sources, and trip control totals coming from observed traffic counts and transit smart card data. Currently HITS uses a two-step procedure to address this problem, producing separate household and trip inflation factors to match externally generated controls. We adapt two recently developed multi-level techniques, namely Iterative Proportional Updating and Hierarchical IPF, to the particular requirements of our case. We compare results with respect to accuracy, bias and algorithm performance, followed by a set of descriptive statistics characterising personal trip-making that illustrate the benefit of using a multi-level approach.

Keywords: Disaggregation, IPF, Hierarchical IPF, IPU, Travel Survey

1. INTRODUCTION

The statistical analysis of travel survey results generally involves an adjustment or “grossing up” of individual trip records to match externally generated control totals, such as those from a cordon or screen-line survey. The contribution of each trip record is indicated by a trip weight, such that the travel demand description obtained by aggregating trip weights by mode and time of day into a set of origin-destination flows will approximate the travel demand of the full population.

On the other hand, the distribution of household demographics in the sample might not match those of the population from which it was drawn, thus requiring a similar adjustment of the contribution of each household to fit the distribution of demographic attributes observed in the region of interest, typically recorded during a full population census. For this stage, iterative proportional fitting is typically employed to associate a household inflation factor with each household, such that weighting an aggregation of demographic attributes by these factors will produce the marginal control totals of the census.

In general, the travel demand produced by inflating trip records by their associated household’s inflation factor, does not produce the observed volumes from the screen-line or cordon survey used as external control. Therefore, a separate inflation procedure is frequently employed to produce the trip weights, which may vary in complexity between different applications. The resulting trip inflation factors for trips belonging to a particular household can therefore differ significantly from the household inflation factor.

This variation in trip inflation factors within a single household introduces difficulty in characterising the individual trip making behaviour of a person belonging to a household with a particular structure. For instance, in the case of activity-based travel analysis, we are interested in the frequency of entire activity chains, such as home—work—home or home—school—shop—home for such a household. If the weight of each of the trips in a chain varies significantly from the household weight, our results may be biased.
Ideally, therefore, one would like to have a single inflation factor for each household, such that aggregation will produce both the observed trip volumes and demographic control totals. In this paper, we explore the application of recently developed, multiple-level iterative fitting techniques to produce such a set of household inflation factors for the Singaporean Household Interview Travel Survey (HITS) of 2008 (Media Research and MVA, 2010).

We take the Hierarchical Iterative Proportional Fitting (HIPF) technique of Müller and Axhausen (2011a) and the Iterative Proportional Updating (IPU) technique of Ye et al. (2009) and adapt it to the peculiarities of the Singaporean case. The paper is organised as follows: we shortly explain the challenges faced with the survey data set, followed by a summary of both techniques and the customisations implemented to address those challenges. We discuss the results produced with respect to accuracy, bias and algorithm performance, followed by a set of descriptive statistics characterising personal trip-making that illustrate the benefit of the multi-level techniques. The paper concludes with a discussion of results and suggestions for further work.

2. THE SINGAPOREAN HOUSEHOLD INTERVIEW TRAVEL SURVEY (HITS)

The HITS of 2008 records the trip-making behaviour of approximately 11,000 qualifying households sampled across the island. Qualifying households are Singaporean citizens and permanent residents (the PR population), as well as the legal immigrant population residing in Singapore, be it for work, education or as a legal dependent of a qualifying member of the household (the non-PR population). The PR population of the island was estimated in 2008 to amount to 1.144 million households (Media Research and MVA, 2010). These households contain approximately 3.6 million people. However, the total legal resident population of Singapore in 2008 amounts to 4.839 million people — the difference of 1.2 million people represents the legal immigrant population.

As a matter of policy, the Singaporean statistical authority, SingStat, only releases detailed reports on the demographics of the PR population, in the form of various summary tables compiled from census and household surveys. No information on the non-PR population is released, except for the total number of persons legally residing on the island. On the other hand, the source of validation for trip-making is traffic counts by vehicle category across 18 screen-lines across Singapore, as well as boarding and alighting totals derived from travel smart card data for public transport. These data reflect the total (PR + non-PR) resident-produced traffic for the island on a typical weekday.

The original survey analysts arrived at trip-level inflation factors by first performing a household-level inflation to match the number of households by dwelling type and aggregated Development Guide Plan (DGP) zones (55 planning areas used in urban land-use planning, aggregated to 19 zones for the purpose of the survey). This process produced a grossed-up population of 4.058 million people, 16% less than the total island population of 4.839 million for 2008. Household inflation factors were calculated for all households interviewed, including those that only used non-motorised transport during the entire day, or those that turned out, upon investigation, not to be eligible for the survey. It was expected, therefore, that a dramatic shortfall of trips would be encountered, an issue that was addressed in a separate trip inflation step.

For non-public transport modes, the trip inflation process involved the generation of trip matrices, starting from the aggregation of trip inflation factors initialised to the value of their respective household factors, followed by equilibrium network assignment. A matrix adjustment step followed, to correlate assigned and observed screen-line counts by vehicle class, for each of three time periods; the AM-peak (07h30-09h30), PM-peak (17h30-19h30) and all other times during the rest of the day, or off-peak (OP). The network assignment and matrix adjustment steps were repeated to convergence. For public transport trips, trip inflation factors initialised to their respective household factors were aggregated to OD matrices and iteratively adjusted until they matched the totals from smart card travel data taken from a typical weekday in 2008.
3. ALGORITHMS

We start with a number of definitions, derived from (Müller and Axhausen, 2011a) and adapted to suit our case. We then present our adaptation of the IPU procedure (Ye et al., 2009) as represented in the review by Müller and Axhausen (2011b), as well as the HIPF (Müller and Axhausen, 2011a).

It is the goal of both algorithms shown in this paper to produce household level inflation factors $f_h$, $h \in H$, where $H$ represents a disaggregate sample of records of motorised trip making households, given as a multi-set of category tuples $H = \{(a_h, b_h, \ldots), (a_h, b_h, \ldots), \ldots, (a_h, b_h, \ldots)\}$.

We define the set of households $H_{ab}$ for a given combination of categories $(a, b)$ as $H_{ab} := \{h \in H : (a_h, b_h, \ldots) = (a, b, \ldots)\}$. In addition, we define the sum $F_{ab}$ of all inflation factors for given categories $a, b$ as $F_{ab} := \sum_{h \in H_{ab}} f_h$. For any category $x$, $H_x$ and $F_x$ are defined analogously.

The control total for category $x$ is denoted by $C_x$, $x \in a, b, \ldots$; matching grand totals for all attributes are a trivial precondition for the convergence of any fitting algorithm: $\sum_a C_a = \sum_b C_b$.

Furthermore, the reference sample also features all trips made by the constituent households, denoted by another set of category tuples: $T = \{(\alpha_t, \beta_t, \ldots, h_t), \ldots, (\alpha_t, \beta_t, \ldots, h_t), \ldots, (\alpha_t, \beta_t, \ldots, h_t)\}$. The attribute $h_t$ specifies the household that each trip $t$ belongs to. Accordingly, this introduces an implicit attribute $t_h$ for each household $h$, referring to the total number of trips made by that household. We denote by $T_\alpha$ the set of trips that fall within category $\alpha$ : $T_\alpha := \{t \in T : (\alpha_t, \ldots) = (\alpha, \ldots)\}$. We also define the control totals $C_\alpha$ and the grand total of trips $T_\alpha = \sum_a C_\alpha$.

Both algorithms require the operation of a basic fitting step to operate on household inflation factors, which also forms the basis for two important adaptations explained in paragraphs to follow. We refer to this procedure as FIT, as shown in Figure 1. For each household record, we denote the number of persons belonging to that household as $p_h$, and we denote the grand total of persons for the study area by $T_p$. Consequently, we define a procedure FITPAX shown in Figure 2 to adjust household inflation factors such that the weighted sum of persons in each household equals $T_p$.

---

**Figure 1:** Procedure FIT($f_h, C_a, C_h, \ldots$) — the Fitting Step in our List-Based IPF procedures

**Figure 2:** Procedure FITPAX($f_h, p_h, T_p$) — fitting step to match total population.
3.1 Adapting multi-level IPF algorithms to the Singaporean case

We adapt the IPU (Ye et al., 2009) and HIPF (Müller and Axhausen, 2011a) to the Singaporean case in order to produce fractional household-level inflation factors \( f_h \) for all motorised trip-making households \( h \) that aim to:

- satisfy trip-level controls of trip volumes by origin-destination pair, time of day and main mode of travel;
- satisfy the total population control, \( T_p := 4,839,000 \), such that \( \sum_h f_h \cdot p_h = T_p \), where \( p_h \) represents the number of persons in each household \( h \);
- satisfy household-level controls exactly where said control categories pertain to the number of households residing in state-built housing;
- for any other household control category \( x \), produce factors that at least sum to the control total for that category, e.g. \( \sum_{h \text{total}} f_h \geq C_x \).

The rationale behind the last two constraints is that, while the procedure should generate at least the number of households per category specified in the census- and survey-based summary tables released by SingStat, it should satisfy any categories pertaining to government-built housing exactly because such housing is only available for resale to permanent residents and citizens. Therefore, we need to alter the household fitting step to only operate under these conditions, as detailed by procedure FITTH shown in Figure 3.

Due to space constraints, we cannot detail a full exposition of the IPU and HIPF procedures, and instead refer the reader to the aforementioned authors for details. We summarize both procedures in Figures 4 and 5 and highlight the alterations made to either procedure to accommodate our requirements for this case.

---

**Figure 3: Procedure FITTH\((f_h, C_x, C_h, \ldots)\) — adjusted fitting procedure for households**

```
Require: Reference sample, expansion factors \( f_h \), control totals \( C_x, C_h \) and possibly others
Ensure: Expansion factors \( f_h \) for the next iteration
for \( C_x \in C_a, C_b, \ldots \) do
  if \( \sum_{h \in H} f_h^{(k+1)} < C_x \) for household control \( x \), or control relates to government housing then
    \( f_h^{(k+1)} \leftarrow \text{FIT}(f_h^{(k)}, C_x) \) for all \( h \in H \),
  end if
end for
return \( f_h \)
```

**Figure 4: Procedure IPU\((f_h, T_p, C_a, C_b, C_{\alpha}, C_{\beta}, \ldots)\) — IPU adapted to the Singaporean case**

```
Require: Reference sample, control totals \( T_p, C_a, C_b, C_{\alpha}, C_{\beta} \) and possibly others
Ensure: Expansion factors \( f_h \) obeying a selection of control totals
\( k = 0 \)
\( f_h^{(0)} \leftarrow 1 \) for all \( h \in H \)
repeat
  \( f_h^{(k+1)} \leftarrow \text{FIPAX}(f_h^{(k)}, p_h, T_p) \) for all \( h \in H \)
  \( f_h^{(k+2)} \leftarrow \text{FITTH}(f_h^{(k+1)}, C_a, C_b, \ldots) \) for all \( h \in H_{ob} \)
  \( f_h^{(k+3)} \leftarrow \text{FIT}(f_h^{(k+2)}, C_{\alpha}, C_{\beta}, \ldots) \) for all \( h : \frac{t_h}{t_e} \in T_{\alpha\beta} \)
  \( k \leftarrow k + 3 \)
until convergence
Return \( f_h^{(k)} \)
```
3.2 Adjusting for trips- and persons per household in the HIPF-based procedure

Analogous to Müller and Axhausen (2011a), we introduce a step to estimate new expansion factors \( f_h' \) from those produced after averaging trip-level inflation factors \( f_h \), subject to the restriction that \( \sum_h p_h f_h' = T' \).

From the infinite set of feasible solutions, we choose the one that minimizes the relative entropy from \( f_h \) to \( f_h' \), defined as \( D(f_h' \| f_h) = \sum_h f_h' \ln \frac{f_h'}{f_h} \).

The solution of the underlying optimisation problem was derived in a similar fashion as detailed in Müller and Axhausen (2011a).

Figure 5: Procedure HIPF(\( T, \alpha, \beta, \ldots \)) — HIPF adapted to the Singaporean case

4. RESULTS

We generated a number of controls for our application to the HITS reference sample using the original survey analysts’ categories, to make direct comparison to the original survey inflation factors possible. For households, we control for dwelling type, aggregated DGP zone, ethnicity and household size; for trips, we control for trip destination DGP, mode and time of day. From the outset, the strong correlation between household DGP and trip origin appeared to be the cause of divergence. Therefore, the trip origin was excluded from the controls but included in the validation of results against the full set of trip volumes by origin, destination, mode and time of day.

We ran each procedure twice, once using cross-classified control variables where available, and a second time using only marginal controls for all categories. In the case of households, cross-classified controls for dwelling type by DGP were available, while we used a single control of destination by mode by time of day for trip-level adjustment. Convergence was defined as a target compliance with the total population control.

4.1 Bias of inflation factors

We compare the distribution of inflation factors for the two integrated approaches against the separate household and trip inflation factors produced by the original survey analysts. In the first two plots shown in Figures 6 and 7, we arrange the trip records in decreasing order of \( f_h' \) (or \( f_i \) in the case of
the original trip inflation factors produced by the original survey analysts) and plot their cumulative sum. Increasing deviation from linearity indicates the bias introduced by stricter enforcement of the person control total for each method. The original trip-wise inflation factors therefore represent the minimum amount of bias possible for this sample.

![Figure 6: IPU-based procedure vs. original trip and household weights](image1)

![Figure 7: HIPF-based procedure vs. original trip and household weights](image2)

![Figure 8: Number of households produced by both procedures for a 10% error against the person control total.](image3)

![Figure 9: SRMSE vs. slack on the person control total for both cross-classified (joint) and marginal (marg) runs of both procedures.](image4)

The results shown for the IPU-based procedure was produced using cross-classified controls. In general, this algorithm produced less bias and converged faster when using cross-classified controls. The results shown for the HIPF-based procedure, on the other hand, were produced using marginal controls only, which seems to be its requirement for less bias and quicker convergence. However, even when using cross-classified controls, the HIPF-based procedure introduces less bias than the IPU-based one, as can be seen in Figure 8. This figure shows the bias introduced at the household level, and for purposes of comparison, the dark line shows the projected number of households produced when averaging the original trip inflation factors for each household.
4.2 Goodness-of-fit

We use the distance-based SRMSE measure of goodness of fit as suggested by Knudsen and Fotheringham (1986), and compare predicted trip volumes for the different origin-destination, modal and time period combinations against the target matrices used to produce trip inflation factors in the original survey analysis. Figure 9 shows a plot of SRMSE vs. increasing error with respect to \( T_p \), the total population control. Dark lines are for runs using the cross-classified controls, while grey lines are for marginal controls only. In both cases, the HIPF-based procedure outperforms the IPU-based one.

4.3 Number of households by dwelling type and DGP

Stricter enforcement of the total population constraint increases bias in the distribution of inflation factors, while reducing goodness-of-fit against the trip matrices produced in the original survey analysis. Therefore, one has to decide on a suitable trade-off between these factors for the application at hand. In our case, we selected the inflation factors produced by the HIPF-based procedure using marginal controls for a 10% error against the total population control.

In Table 1, we see the increase in households by dwelling type produced by the selected procedure. In the first four column headings, ‘hdb’ refers to government-built or Housing Development Board (HDB) housing, while the following number refers to the number of rooms in the unit. ‘landProp’ and ‘privFlat’ refers to privately owned landed property and condominiums. In spite of always controlling for government housing, we still see a significant increase in the weighted sum of inflation factors for these households. However, as expected, the largest increase is for the privately-owned and ‘other’ categories, the household inflation factors of which are only adjusted if they fall below their control total. Consequently, the most significant increase in number of households by DGP is for the centre of the island, where the number of HDB dwelling units is lowest, and private housing dominates.

<table>
<thead>
<tr>
<th>Table 1: Households by dwelling type</th>
</tr>
</thead>
<tbody>
<tr>
<td>hdb1.2</td>
</tr>
<tr>
<td>PR only control</td>
</tr>
<tr>
<td>HIPF weighted sum</td>
</tr>
<tr>
<td>Difference</td>
</tr>
<tr>
<td>% increase</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Trips per person by mode of transport.</th>
</tr>
</thead>
<tbody>
<tr>
<td>bus</td>
</tr>
<tr>
<td>Orig. tot. trips</td>
</tr>
<tr>
<td>Orig. tot. pax</td>
</tr>
<tr>
<td>HIPF tot. trips</td>
</tr>
<tr>
<td>HIPF tot. pax</td>
</tr>
<tr>
<td>Orig. trips/pax</td>
</tr>
<tr>
<td>HIPF trips/pax</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Trips per person by trip purpose.</th>
</tr>
</thead>
<tbody>
<tr>
<td>biz</td>
</tr>
<tr>
<td>Orig. tot. trips</td>
</tr>
<tr>
<td>Orig. tot. pax</td>
</tr>
<tr>
<td>HIPF tot. trips</td>
</tr>
<tr>
<td>HIPF tot. pax</td>
</tr>
<tr>
<td>Orig. trips/pax</td>
</tr>
<tr>
<td>HIPF trips/pax</td>
</tr>
</tbody>
</table>
4.4 Trips per person

The mismatch between household- and trip-level inflation factors in the original survey analysis meant that, essentially, we had a large number of trips produced by a relatively small number of households. Say we have a person \( p_h \) belonging to a household \( h \) with household inflation factor \( f_h \), making several trips \( t_i \), each having a trip inflation factor \( f_{it} \). Then, for the population as a whole, the number of trips made by a person of this type, belonging to a household of the given structure, amounts to \( \sum f_{it} + f_h \). This distortion disappears with our simultaneous adjustment procedure, as can be seen in the results shown in Table 2, where we compare the number of trips and persons making trips using a particular main mode of travel. ‘Orig.’ refers to the number of persons and trips produced by the original household and trip inflation factors. Assuming that people, in general, make tours (from home to a major activity and back) and sub-tours (from home or a major activity to a secondary activity and back), we generally expect the number of trips per person to be closer to an even number, as can be seen for the HIPF-based result.

Table 3 shows a similar analysis of the trip-making behaviour for the 7 major reported trip purposes. Once again, we have less bias in this case, especially for pick-up and drop-off type trips (‘pikUpDrop’), which show that qualifying persons generally make two such trips per day.

5. DISCUSSION

Even though our procedure manages to remove the bias of separate household and trip inflation factors when analysing personal trip-making behaviour, there is still a significant bias in the distribution of inflation factors. This bias can be seen for the original trip inflation factors as well, and is ascribed to the under-reporting of trips in the survey (especially for secondary activities), as well as a lack of interest in survey participation observed for the non-PR population. Even though this group contributes 25% towards the total island population, we estimate that they make up only 3% of trip records, based on Singapore’s ethnicity profile.

Further work will include a simple model of permanent residency, in order to explicitly control for this quantity. We also plan to produce an in-depth investigation into personal trip-making behaviour, specifically with respect to the distribution of personal travel distances and times for various trip modes and purposes. These results will ultimately be used to estimate the choice behavior of the commuting population in a disaggregate, activity-based, household travel demand generation procedure.

REFERENCES


