# Suggestions for estimation and prediction of PWC matrices

Gunnar Flötteröd, VTI

March 22, 2019

## Preliminaries

This note is written in continuation of the questions raised in my presentation during the March 15 seminar. It is a combination of (i) an attempt to summarize the PWC matrix modeling process and (ii) suggestions that may improve this process without having to systematically change the existing model structures.

### Singular flows

Singular flows are completely excluded from this discussion. This means that they are taken out of all data sets and subtracted from all marginal information prior to estimation. In prediction, they are assumed to be separately extrapolated and added to the model-based prediction. This is based on the assumption that the notion of "singularity" implies incompatibility with all other structural modeling assumptions made here; it obviously would be more desirable to have a structurally meaningful model even for singular flows.

### Import/export/transit and domestic flows

The following presentation only considers domestic flows to keep the presentation simple and to bypass the need to navigate around ambiguities in the available report. The described models and methods can, however, be expected to carry over to the estimation and prediction of import/export/domestic flows.

## Estimation

### Estimation of a gravity model from VFU (varuflödesundersökning)

The gravity model may be estimated exclusively from data considered in and available from the VFU. One gravity model per commodity (goods type) $t$ is estimated. The VFU provides flow and cost samples for (a subset of) all origin/destination pairs $rs$. Given a parameter vector $\boldsymbol{\beta}_t$, an attribute vector $\mathbf{x}_{t,rs}$, and a domestic commodity value $\nu_t$, a gravity model of the following form can estimated from this data:

$$f_{t,rs} = e^{\boldsymbol{\beta}_t^T \mathbf{x}_{t,rs}} \quad (1)$$

$$c_{t,rs} = e^{\boldsymbol{\beta}_t^T \mathbf{x}_{t,rs} + \ln \nu_t} \quad (2)$$

with $f_{t,rs}$ and $c_{t,rs}$ being expected flows respectively costs of commodity $t$ in origin/destination relation $rs$ as predicted by the gravity model. Superscript $T$ denotes the transpose.

Estimating this model jointly from flow and cost data yields a point estimator of $\ln \nu_t$, including an estimate of its confidence interval. From this, a point estimate and a confidence interval of $\nu_t$ can be derived. Having the value $\nu_t$ within the model even provides a starting point for further developing the value estimator, e.g. by adding explanatory variables that interact $\nu_t$ with shipment size (possibly capturing discounts) or coupling $\nu_t$ to zonal parameters (possibly capturing regional variations).

The result are commodity-specific gravity-based flow matrices $F_t = F(\mathbf{x}_t \mid \boldsymbol{\beta}_t)$ as well as commodity values $\nu_t$.

**Supplementary marginal data**

Marginal information may refer to more than one commodity type and may refer to flows or costs. This information is expressed as a set of constraints, with the $j$th constraint being specified as follows:

- $u_j$ is an upper bound on the $j$th marginal (flow or cost). If there is no upper bound, it may be set to infinity. $u_j$ is set to a finite value if one believes that the underlying data source *overestimates* the true marginal.
- $l_j$ is a lower bound on the $j$th marginal (flow or cost). If there is no lower bound, it may be set to zero. $l_j$ may be used if the underlying data source *underestimates* the true marginal.
- If one believes to have reliable data, one may set $u_j = l_j$.
- The parameters $b_{j,t,rs}$ are set to
  - one if the $j$th marginal refers to a flow, includes commodity $t$, and contains matrix entry $rs$;
  - $\nu_t$ if the $j$th marginal refers to a cost, includes commodity $t$, and contains matrix entry $rs$;
  - zero otherwise.
- These quantities are combined in the following interval constraint:

$$\sum_{t,rs} b_{j,t,rs} g_{t,rs} \quad \in \quad [l_j, u_j]. \tag{3}$$

Denote by $\Omega$ the set of PWC matrices that comply with all constraints. If this set is empty, meaning that there are contradictory constraints, an automatic correction of $\Omega$ is necessary and possible. This correction minimizes some deviation between corrected and original interval constraints while ensuring a nonempty $\Omega$.

The currently used marginal data is a special case of the setup described here.

**Balancing of the flow matrix**

Let $G_t(\Omega, F) = (g_{t,rs}(\Omega, F))$ be the result of jointly balancing all gravity-based flow matrices $F = \{F_t\}$ subject to all interval constraints (the feasible set $\Omega$).

The usage of the more general bound constraints defining $\Omega$ bypasses the need for additional matrix calibration steps. All relevant information is already included in $\Omega$. This is a more general setting than what vanilla iterative proportional fitting (IPF) can handle because (i) multiple matrices (for different commodities) may be balanced at the same time, (ii) the constraints are not limited to row and column sums and (iii) inequality constraints are allowed for.

In the special case of the currently used marginal data, however, these particular features disappear and vanilla IPF is applicable.

That the more general case considered here has a unique solution is guaranteed from (i) the constraint set being convex and (ii) the sum of the entropy functions per commodity type $t$ being strictly concave. Generalizations of IPF in this direction exist, for instance Range-RAS,[1] which is not much more complicated than vanilla IPF but can handle bound constraints.

The result of this balancing are the complete (apart from singular flows) commodity-specific status-quo matrices $G_t = G_t(\Omega, F)$.

# Prediction

### Gravity model

Given predicted gravity model attributes $\mathbf{x}'_t$, one can predict new gravity model outputs

$$F'_t \quad = \quad F(\mathbf{x}'_t \mid \boldsymbol{\beta}_t). \tag{4}$$

*Without additional marginal information*, one now can make different assumptions about how the deviation between predicted complete matrix $G'$ and gravity model prediction $F$ changes, leading to different predictors of the complete matrix $G'$.

- The deviation between complete matrix and gravity model matrix is constant:

$$G' \quad = \quad F' + (G - F). \tag{5}$$

- The complete matrix is element-wise proportional to the gravity model matrix:

$$G' \quad = \quad \left( \frac{f'_{rs}}{f_{rs}} g_{rs} \right). \tag{6}$$

- Combinations of the two aforementioned models.
- ...

*With updated marginal information*, one may balance the predicted gravity model $F'$ against the predicted marginal information $\Omega'$ to obtain complete (apart from singular flows) predicted matrices

$$G'_t \quad = \quad G_t(\Omega', F') \tag{7}$$

per commodity $t$.

[1] Censor, Y. and S.A. Zenios (1991). Interval-Constrained Matrix Balancing. *Linear Algebra and its Applications* 150:393-421.

# Summary

An added value of the formulation suggested here is that all information supplementing the gravity model is encoded in one large set of constraints, which then is included in one single matrix balancing step. Given methods for automatically (i) ensuring the existence of a solution by suitably widening the constraints and (ii) solving a multi-matrix balancing problem subject to general linear inequality constraints, the matrix modeling work would be reduced to *describing* available data sources instead of *manipulating* them. The necessary computer implementations appear not particularly complicated, given the rather clear structure of the underlying mathematical problems. An extension of this approach to import/export/transit matrices appears feasible.