Combining survey data and mobile network operator data for commuting statistics

Tiziana Tuoto

Third Workshop on Methodologies for Official Statistics, Istat

Joint work with E. Cerasti, L. Di Consiglio, D. Filipponi, T. Pichiorri, A. Piovani, L.-C. Zhang

December 4 2024

Outline

Credits

Project 101132744 — 2022-IT-TSS-METH-TOO

Research Project co-funded by the European Commission:

- ESSnet MNO-MINDS Mobile Network Operator data - Methods for Integrating New Data Sources https://cros.ec.europa.eu/mno-minds
- Eurostat initiatives on the reuse of MNO data for official statistics production https://cros.ec.europa.eu/MNOdata4OS

Inference and MNO data

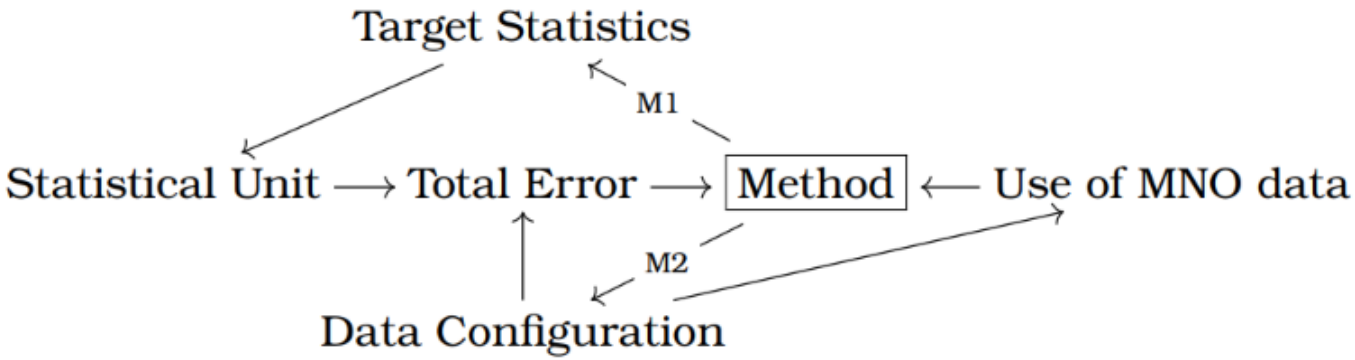

Figure 1: Reference frame for methods to combine MNO and non-MNO data. M1, M-executor methods; M2, M-enabler methods.

Figure: from Deliverable D3.1 *Preliminary report on methodologies* of the MNO-MINDS project

Methods are classified according to the main source of uncertainty:

- Randomisation
- Quasi-randomisation
- Super-population modelling

In this work

- Commuters' Origin Destination matrix by municipality
- Definition of commuter: at least 3 days a week leaving and going back home on the same day. for work or study
- First Step: commuters by municipality without destination

Available information

- Survey Census data: 4 years round, 2018-2021 plan, in reality 2018, 2019, 2021
  - Self representative municipalities: in the sample every year
  - Not self representative: 1 municipality for each stratum of 4
- MNO: Call Detail Records data from one of the three main operators in Italy
  - Generated by calls and SMSs
  - 6 weeks in 2017 in one region (Tuscany, 273 municipalities) provide info on the time and position of the events

Available information

- Previous census (2011) flows
- Administrative data on place of work/study, ad hoc procedure for the purpose of Census production
- Municipality features

Notation

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{bmatrix} \quad (1)$$

- $N \times 1$ stacked vector $y = vec(Y)$ of these flows by origin-centric ordering
- $N = n^2 - n$
- the first n elements reflect flows from origin zone 1 ($i = 1$) to all $n - 1$ destinations and so on

In the first step, only flows originated by each municipality

$$y = (y_{1+}, y_{2+}, \ldots, y_{n+})$$

Interaction Models

Interaction model (LeSage and Pace, 2008) is:

$$\mathbf{y} = \alpha + \mathbf{H}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\phi} + \mathbf{D}\theta + \boldsymbol{\epsilon} \tag{2}$$

where **H** and **G** are features of the origin and destination respectively, and **D** is a friction, such as a distance.
Extension with spatial autoregressive component:

$$(\mathbf{I}_N - \rho\mathbf{W}_d)(\mathbf{I}_N - \lambda\mathbf{W}_o)\mathbf{y} = \alpha + \mathbf{H}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\phi} + \mathbf{X}\xi + \boldsymbol{e} \tag{3}$$

where in X we consider the distance plus other auxiliary variables related to origin and destination, such as MNO flows.

Fay-Herriot Model

To reduce the bias, we can consider a FH model (Fay and Herriot, 1979).
We can assume that the areas observations obey some population model, e.g. the (3).
A simplification is to consider the autoregressive component only for the us, for example depending only on the origins or destinations:

$$\hat{\mathbf{y}} = \mathbf{H}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\phi} + +\mathbf{X}\xi + (I - \rho_j \mathbf{W}_j)^{-1}\boldsymbol{u} + \boldsymbol{e} \qquad (4)$$

FH only for origin

- In the first step, where instead of O/D we consider the number of commuters originated from a municipality, regardless their destination
- the previous models apply considering only the origin
- y_i is the total commuters originated from municipality i
- Similarly as auxiliary variables we consider the outbound MNO flows or administrative counts originated from municipality i.

A Transfer Learning approach for the estimation for unobserved flows

Let k be the domain, $k = ij$ if origin-destination, $k = i$ if only origin.

- $\mathcal{D}_1 = \{k : n_k > 0\}$ $(\hat{\xi}, \hat{\sigma}_u^2)$ are the parameter estimator and $\hat{Y}_k^{FH} = x_k^\top \hat{\xi}_k + \hat{u}_k$ are the resulting EBLUP of Y_k.
- $\mathcal{D}_0 = \{k : n_k = 0\}$ consider estimating $\{u_k : k \in \mathcal{D}_0\}$ by minimising

$$L(\mathcal{D}_0) = \sum_{k \in \mathcal{D}_0} u_k^2 + \alpha \sum_{k \in \mathcal{D}_0} (x_k^\top \hat{\xi} + u_k - Y_k^*)^2 \qquad (5)$$

α is a tuning constant, $\alpha \geq 0$, and Y_k^* a proxy to Y_k, which is available for all the domains

A Transfer Learning approach for the estimation for unobserved flows

- Choice of α: based on $\mathcal{D}_1$ and the EBLUP $\{\hat{u}_k : k \in \mathcal{D}_1\}$
- Given α, let $\tilde{u}_k(\alpha)$ be obtained by minimising $L(\mathcal{D}_1)$ for the domains with $n_k > 0$, just like one would have done for the domains with $n_k = 0$.
- Choose the value of α, such that the corresponding $\{\tilde{u}_k(\alpha) : k \in \mathcal{D}_1\}$ are closest to the EBLUP $\{\hat{u}_k : k \in \mathcal{D}_1\}$ according to some chosen metric.

Quasi Randomization - 1

An estimate of the pseudo probability of being included in the MNO sample is given by:

$$m_{i+|z}/\hat{Y}_{i+|z} \tag{6}$$

where $\hat{Y}_{i+|z}$ are the direct estimates of commuters originating from area i in the socio-economic group z and $m_{i+|z}$ the corresponding MNO count. Then

$$\hat{Y}^M_{ij|z} = \hat{Y}_{i+|z}\frac{m_{ij|z}}{m_{i+|z}} \tag{7}$$

and the target estimate based on MNO is

$$\hat{Y}^M_{ij} = \sum_z \hat{Y}^m_{ij|z} \tag{8}$$

Unfortunately we do not have socio-economic groups in our MNO data, this can be applied only considering geo variables in z.

Quasi Randomization - 2

In addition of breaking down the direct estimates $\hat{Y}_{i+|z}$ by the MNO ratios $m_{ii|z}/m_{i+|z}$, we can use some adjusting factor(s) to derive MNO-based target estimates, e.g.:

$$\hat{Y}^{QR}_{ij|z} = \frac{m_{ij|z}}{\tau * \phi_z * \tilde{\alpha}} \tag{9}$$

- τ is the adjustment relative to the target population - from ADL survey
- ϕ_z is the proportion of subscribers of our MNO in area z - provided by the MNO itself.
- $\tilde{\alpha}$ is the deduplication adjustment - from ADL survey

The sampling estimates from the Aspect of Daily Life 2022, can be broken down by socio-demographic features, they are not planned domains. So far, national estimates have been used.

Relationships

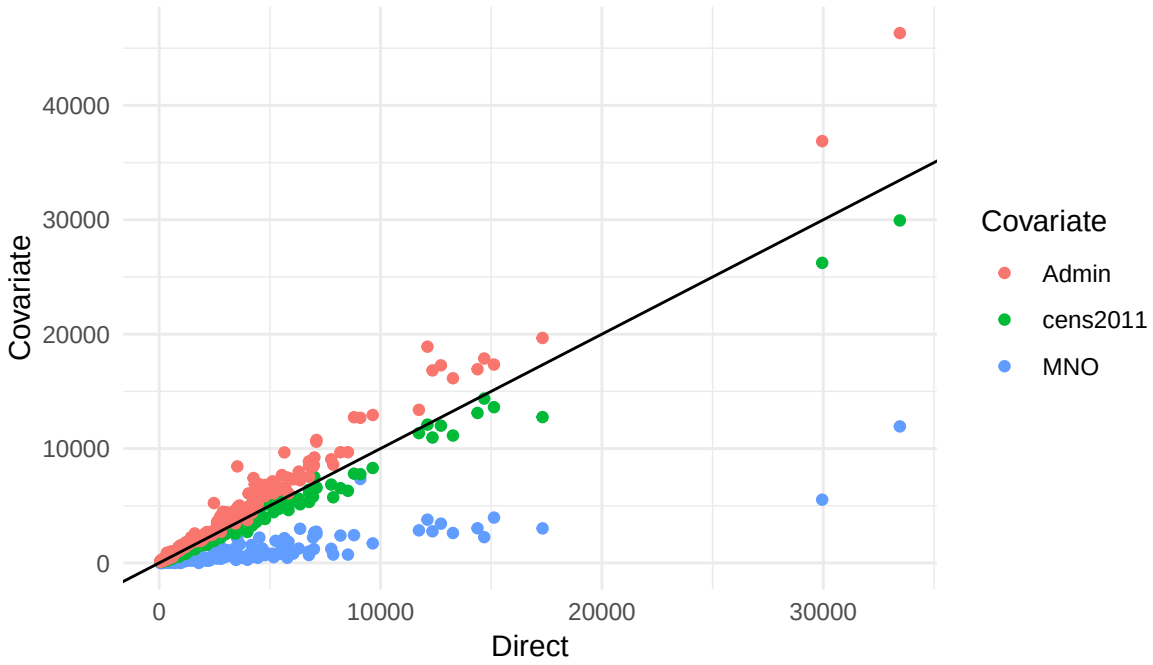

Figure: Relationship between direct estimates and covariates for outbound commuters, survey 2021

Similar relationship for origin/destination flows, and in survey 2018.

Direct Estimates

Year	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2021	2.180	5.092	6.828	7.972	9.532	20.650
2018	2.222	5.132	6.791	8.739	10.023	47.113

Table: CV direct estimates

The municipality direct estimates of outbound commuters are good both in 2021 and 2018, but about 100 (150) municipalities are not in the sample in 2021 (2018).
The CVs of the direct estimates of the OD flows are higher.

Model choice

Covariates	BIC
census2011	-63.33
Admin	-54.48
MNO	456.94
census2011+Admin	-79.12
cens2011+MNO	-64.79
Admin+MNO	-51.03
All	-77.36

Table: BIC of linear models for the outbound commuters - 2021

MSEs - Direct and FH estimates - 2021

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	93	3930	15824	72712	79646	1715768
FH full	91	2559	13441	63157	68016	1803387
FH MNO+Admin	123	3042	14980	68748	74820	2057839

Table: MSEs - 2021

Estimates 2021

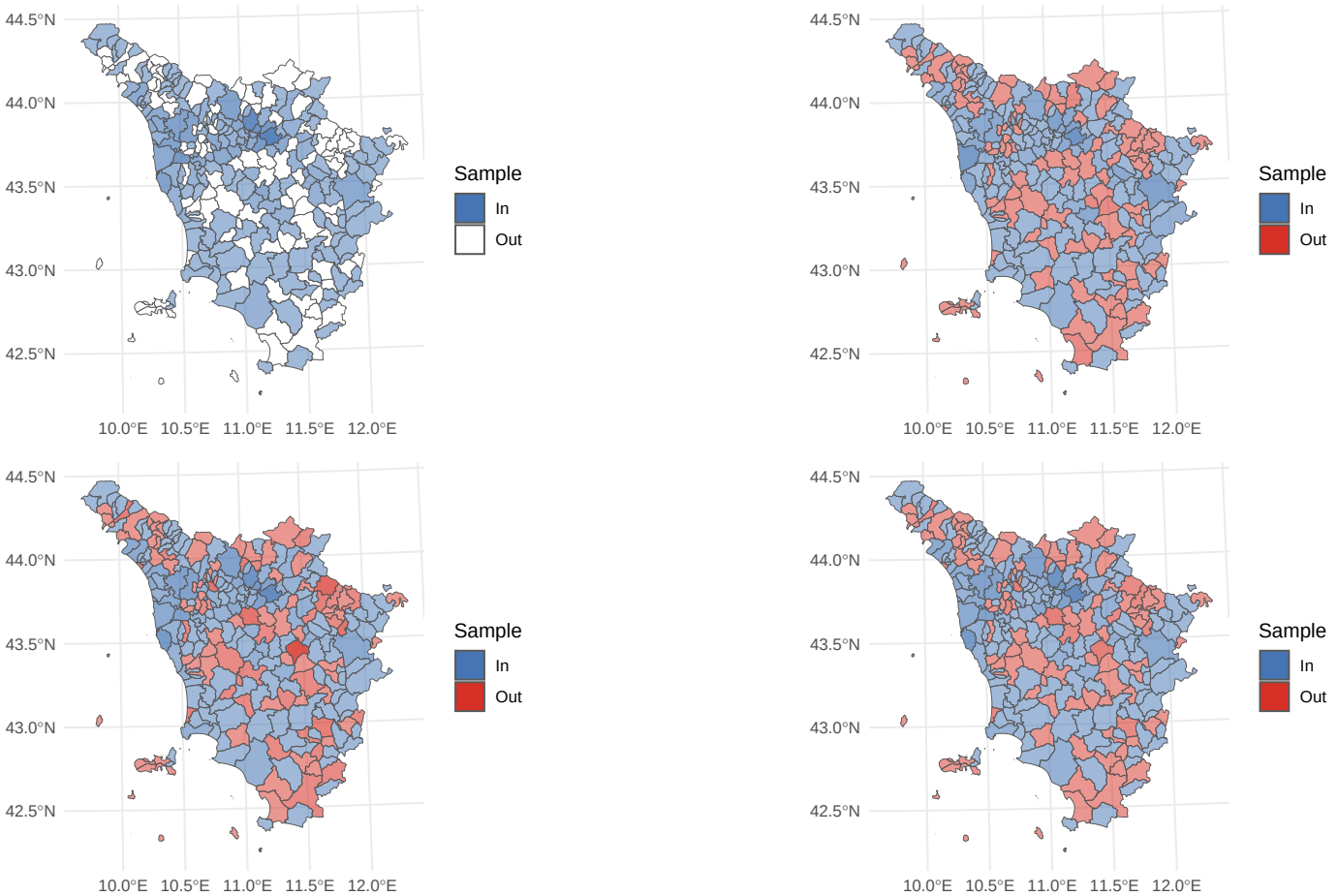

Figure: Top: Direct and QR-2 estimates. Bottom: EBLUP and TL estimates.

Model & TL

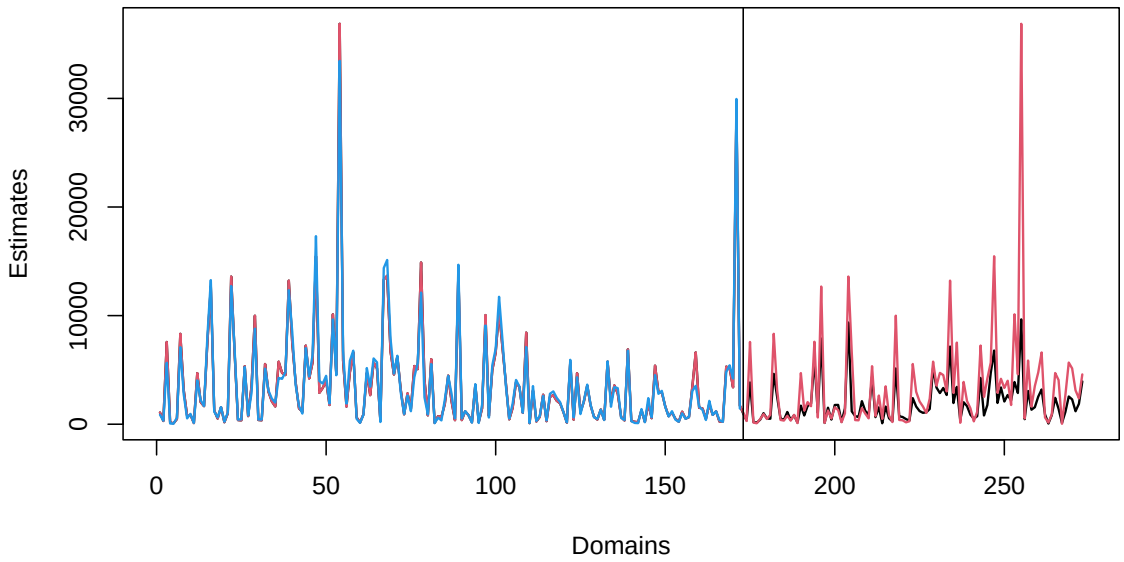

Figure: Direct Estimates, FH covariates MNO+Admin ,TL with census2011 - survey 2021

Model & TL

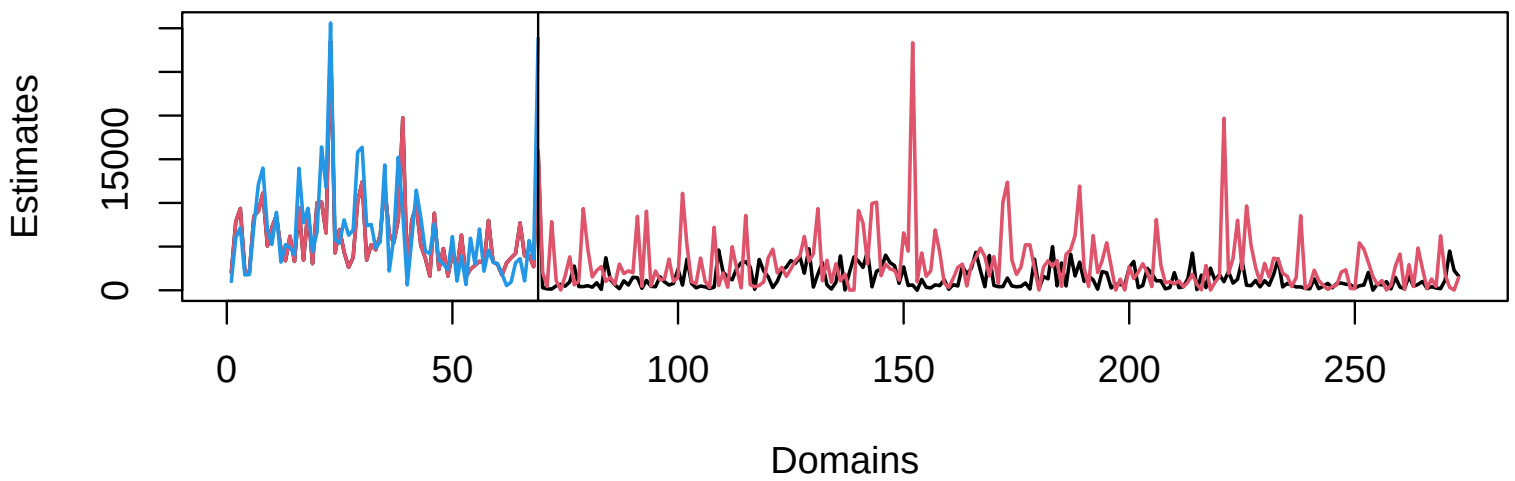

Figure: Direct Estimates, FH covariate MNO ,TL with census2011 - survey 2018

Next steps

- Spatial Models
- More on Transfer Learning
- Analysis on rates rather than on absolute figures
- O/D estimation

Thank you for your attention

https://cros.ec.europa.eu/mno-minds

tuoto@istat.it