# Constrained Optimisation Approach for Adjusting Misclassified Proxy Variables

Fabrizio Solari, Valeria Quondamstefano,
Romina Filippini, Gaia Rocchetti

Istat

Fourth Workshop on Methodologies for Official Statistics



Introduction

Calibration

Optimisation Problem

Case Study

Generalised Calibration

Conclusions

# Motivation

- Target variable $y$ is a partially observed categorical variable
- Proxy $x$ is observed for entire population
- Proxy is affected by misclassification and structural bias but with good micro-record prediction properties
- Goal: produce micro-record adjusted proxy $\tilde{x}$ consistent with target population distribution and preserving $x$ prediction properties

# Notation

- Population $P$, $|P| = N$
- $y$, $x$ categorical variable taking values in $\{1, \ldots, H\}$
- Population distribution ($i = 1, \ldots, H$):

$$N_{y,i} = \sum_{\ell \in P} I(y_\ell = i), \qquad N_{x,i} = \sum_{\ell \in P} I(x_\ell = i)$$

- Sample $S$, sampling weights $w_\ell$
- Sampling distribution ($i = 1, \ldots, H$):

$$\widehat{N}_{y,i} = \sum_{\ell \in S} I(y_\ell = i) w_\ell, \qquad \widehat{N}_{x,i} = \sum_{\ell \in S} I(x_\ell = i) w_\ell$$

# Initial Confusion Matrix

Empirical confusion matrix:

| $y$ | $x$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | $\cdots$ | j | $\cdots$ | H | $\widehat{N}_y$ |
| 1 | $\hat{a}_{11}$ | $\cdots$ | $\hat{a}_{1j}$ | $\cdots$ | $\hat{a}_{1H}$ | $\widehat{N}_{y,1}$ |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| i | $\hat{a}_{i1}$ | $\cdots$ | $\hat{a}_{ij}$ | $\cdots$ | $\hat{a}_{iH}$ | $\widehat{N}_{y,i}$ |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| H | $\hat{a}_{H1}$ | $\cdots$ | $\hat{a}_{Hj}$ | $\cdots$ | $\hat{a}_{HH}$ | $\widehat{N}_{y,H}$ |
| $\widehat{N}_x$ | $\widehat{N}_{x,1}$ | $\cdots$ | $\widehat{N}_{x,j}$ | $\cdots$ | $\widehat{N}_{x,H}$ | $N$ |

$$\hat{a}_{ij} = \sum_{\ell \in S} I(y_\ell = i) I(x_\ell = j) w_\ell$$

$$\widehat{A} = (\hat{a}_{ij}) \in \mathbb{R}_{\geq 0}^{H \times H}, \quad \widehat{A} 1_H = \widehat{N}_y, \quad 1_H^T \widehat{A} = \widehat{N}_x^T$$

$\widehat{A}$ can show systematic bias in $x$ (even when predictive power is strong)

# Calibration Step 1: Population Margins

Enforce proxy margins to match population totals $N_x$:

$$\widetilde{A} = \arg \min_{Z \in \mathcal{Z}(\widehat{N}_y, N_x)} D(Z \| \widehat{A})$$

where $\mathcal{Z}(r, c) = \{Z \in \mathbb{R}_{\geq 0}^{H \times H} : Z\, 1_H = r,\ 1_H^T Z = c^T\}$
(divergence $D$ may be KL (IPF), Hellinger, $L_1$, $L_2$, etc.)

| $y$ | $x$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | $\cdots$ | j | $\cdots$ | H | $\widehat{N}_y$ |
| 1 | $\tilde{a}_{11}$ | $\cdots$ | $\tilde{a}_{1j}$ | $\cdots$ | $\tilde{a}_{1H}$ | $\widehat{N}_{y,1}$ |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| i | $\tilde{a}_{i1}$ | $\cdots$ | $\tilde{a}_{ij}$ | $\cdots$ | $\tilde{a}_{iH}$ | $\widehat{N}_{y,i}$ |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| H | $\tilde{a}_{H1}$ | $\cdots$ | $\tilde{a}_{Hj}$ | $\cdots$ | $\tilde{a}_{HH}$ | $\widehat{N}_{y,H}$ |
| $N_x$ | $N_{x,1}$ | $\cdots$ | $N_{x,j}$ | $\cdots$ | $N_{x,H}$ | $N$ |

# Calibration Step 2: Target Margins

Next, impose consistency with target marginal $\widehat{N}_y$:

$$B = \arg \min_{Z \in \mathcal{Z}(\widehat{N}_y, \widehat{N}_y)} D(Z \| \widetilde{A})$$

$B$ describes the confusion matrix between $y$ and the adjusted proxy $\tilde{x}$

| $y$ | $x$ | | | | | |
|---|---|---|---|---|---|---|
| | 1 | $\cdots$ | j | $\cdots$ | H | $\widehat{N}_y$ |
| 1 | $b_{11}$ | $\cdots$ | $b_{1j}$ | $\cdots$ | $b_{1H}$ | $\widehat{N}_{y,1}$ |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| i | $b_{i1}$ | $\cdots$ | $b_{ij}$ | $\cdots$ | $b_{iH}$ | $\widehat{N}_{y,i}$ |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| H | $b_{H1}$ | $\cdots$ | $b_{Hj}$ | $\cdots$ | $b_{HH}$ | $\widehat{N}_{y,H}$ |
| $\widehat{N}_y$ | $\widehat{N}_{y,1}$ | $\cdots$ | $\widehat{N}_{y,j}$ | $\cdots$ | $\widehat{N}_{y,H}$ | $N$ |

# Minimum Transition Problem

- The transition model induces movements from class $j$ of $x$ to class $k$ of $\tilde{x}$
- To ensure integer transitions, the matrices $\widetilde{A}$ and $B$ must be converted into integer-valued matrices
- Let $\widetilde{A}^*$ and $B^*$ denote the integer-valued versions of $\widetilde{A}$ and $B$, obtained via controlled rounding procedures (Cox and Ernst, 1982; Causey et al., 1985)

# Transition Array

Define a 3-dimensional transition array:

$$T = (t_{ijk}) \in \mathbb{Z}_{\geq 0}^{H \times H \times H}$$

where $t_{ijk}$ = number of records with true class $i$ whose proxy transitions from $j$ (original) to $k$ (adjusted)

Constraints:

$$\sum_k t_{ijk} = \tilde{a}^*_{ij}, \quad \forall\, i, j$$

$$\sum_j t_{ijk} = b^*_{ik}, \quad \forall\, i, k$$

# Minimum Transition Problem

Minimise the number of modified proxy values:

$$T^* = \arg \min_{T \in \mathcal{T}(\widetilde{A}^*, B^*)} \sum_{i,j,k:j \neq k} t_{ijk}$$

This is a special case of the 3-index transportation problem

Proposition: $T^*$ exists iff $\sum_j \tilde{a}^*_{ij} = \sum_k b^*_{ik}, \quad \forall\, i$

This condition is satisfied since $\sum_j \tilde{a}^*_{ij} = \sum_k b^*_{ik} = \widehat{N}_{y,i}\,, \forall\, i$

Marginal transitions matrix $\widetilde{T}^* = (\tilde{t}^*_{jk}) \in \mathbb{Z}^{H \times H}_{\geq 0}$ ignoring the true class $i$ in $T^*$:

$$\tilde{t}^*_{jk} = \sum_i t^*_{ijk}$$

Used when $y$ unobserved in population. Represents required number of individuals to shift from $j$ to $k$

# Adjustment Mechanism

Two strategies:

1. Marginal-only reassignment: select uniformly from units with $x = j$ to apply $\tilde{t}^*_{jk}$ transitions
2. Model-assisted reassignment: use MAP predictions $\tilde{y}_\ell = \max_i \Pr(y_\ell = i)$ to target transitions, where $\Pr(y_\ell = i)$ derives from a model $\mathcal{M}$

Sampling may follows:

- SRS
- Stratified Sampling
- Balanced Sampling
- PPS with $\Pr(y_\ell = i)$ as size
- Alternative solutions: assignment algorithms, propensity scores

# Case Study: Educational Attainment in Italy

- Target variable: 8-level educational attainment: 1 Illiterate, 2 Literate but no formal educational attainment, 3 Primary education, 4 Lower secondary education, 5 Upper secondary education, 6 Bachelor degree or equivalent level, 7 Master degree or equivalent level, 8 Doctoral degree (PhD)
- Population: Italian citizens in the 2011 Census without administrative education signals in 2012–2023
- Proxy variable $x$: educational attainment from the 2011 Census
  - Undercoverage in administrative archives $\Rightarrow$ possible unrecorded qualifications $\Rightarrow$ proxy $x$ potentially biased
- True values $y$: observed in the 2023 census sample survey
- Model probabilities from Di Zio et al. (2019)

# Classification Performance Metrics

Metrics for confusion matrices:

- Accuracy: proportion of correctly classified cases

$$\text{Accuracy} = \frac{1}{N} \sum_i a_{ii}$$

- Precision: correctness of predictions per category

$$\text{Precision} = \frac{1}{H} \sum_i \frac{a_{ii}}{\sum_j a_{ji}} = \frac{1}{H} \sum_i \text{Precision}_i$$

- Recall: ability to identify cases in each category

$$\text{Recall} = \frac{1}{H} \sum_i \frac{a_{ii}}{\sum_j a_{ij}} = \frac{1}{H} \sum_i \text{Recall}_i$$

# Classification Performance Metrics

Metrics for confusion matrices:

- F1-Score: harmonic mean of Precision and Recall
- MAE / MSE: incorporate distances $L_1$ and $L_2$ between categories, capture overall error

$$\text{MAE} = \frac{1}{N} \sum_{i,j} a_{ij} \, |i - j|,$$

$$\text{MSE} = \frac{1}{N} \sum_{i,j} a_{ij} (i - j)^2.$$

# Classification Performance: Summary

| Measure | Model | Census | | |
|---|---|---|---|---|
| | | Pre-Adjustment | Post-Adjustment ($T^*$) | Post-Adjustment ($\widetilde{T}^*$) |
| Accuracy | 0.7352 | 0.8357 | 0.8453 | 0.8435 |
| Precision | 0.5664 | 0.7071 | 0.7353 | 0.7323 |
| Recall | 0.5659 | 0.7086 | 0.7321 | 0.7300 |
| F1-Score | 0.5662 | 0.7072 | 0.7330 | 0.7305 |
| MAE | 0.3405 | 0.2044 | 0.1867 | 0.1899 |
| MSE | 0.5377 | 0.3101 | 0.2703 | 0.2773 |

- Census proxy variable performs better than model based predicted probabilities
- Adjustments improve all metrics compared to both raw census data and model probabilities
- $T^*$ performs slightly better than $\widetilde{T}^*$

# Generalised Calibration

- Minimise the number of transitions could not be the optimal solution $\Rightarrow$ force a larger number of transitions
- Target function with diagonal control:

$$Z^* = \arg \min_{Z \in \mathcal{Z}(r,c)} \alpha D_1(Z \| M) + (1 - \alpha) D_2(\operatorname{diag} Z \| \nu).$$

  - $\alpha$ tuning parameter balancing global fidelity and diagonal constraints
  - Choosing $\alpha$: Monte Carlo selection over a set of candidates
  - Efficient search strategies: ternary search (when the adopted accuracy measure is unimodal), Bayesian optimisation, ...

# Conclusions

- Method ensures marginal consistency
- Analogies with optimal transport but implementing individual level transitions
- Minimally alters individual proxy values
- Framework interpretability and flexibility: can incorporate domain calibration, diagonal targets, model-assisted assignment
- Domain-size sensitivity may affect the validity of the estimates
  - Regularising domain-specific confusion matrices toward larger-domain matrices could help stabilise the estimates and borrowing strength from larger populations (in the spirit of SAE)

# Short References


- Causey, Beverley D., Lawrence H. Cox, and Lawrence R. Ernst (1985). Applications of transportation theory to statistical problems. In: *Journal of the American Statistical Association* 80.392, pp. 903–909. DOI: 10.1080/01621459.1985.10478201.
- Cox, Lawrence H. and Lawrence R. Ernst (1982). Controlled rounding. In: *INFOR: Information Systems and Operational Research* 20.4, pp. 423–432. DOI: 10.1080/03155986.1982.11731877.
- Cuturi, Marco (2013). Sinkhorn distances: lightspeed computation of optimal transport. In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges et al. Vol. 26. Red Hook, NY, USA: Curran Associates, pp. 2292–2300. URL: https://papers.nips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

# Short References


- Di Zio, Marco, Romina Filippini, and Gaia Rocchetti (2019). An imputation procedure for the Italian attained level of education in the register of individuals based on administrative and survey data. In: *Rivista di Statistica Ufficiale* 2.3/2019, pp. 143–174. DOI: 10.1481/istat.rivista.2019.3.143.
- Haley, Kenneth B. (1963). The multi-index problem. In: *Operations Research* 11.3, pp. 368–379. DOI: 10.1287/opre.11.3.368.
- Hofman, René (2000). Multidimensional assignment problems: a survey. In: *European Journal of Operational Research* 128.3, pp. 467–486. DOI: 10.1016/S0377-2217(99)00409-1.
- Kuno, Yusuke (2005). Concave cost multidimensional transportation problems. In: *Mathematical Programming* 102.1, pp. 135–150. DOI: 10.1007/s10107-004-0563-1.

# Short References


- Peyré, Gabriel and Marco Cuturi (2019). Computational optimal transport: with applications to data science. In: *Foundations and Trends in Machine Learning* 11.5-6, pp. 355–607. DOI: 10.1561/2200000073.
- Villani, Cédric (2009). *Optimal Transport: Old and New*. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-540-71050-9.

# Thank you for your attention!



Introduction

Calibration

Optimisation Problem

Case Study

Generalised Calibration

Conclusions