M-quantile Area-level Models for Robust Small Area Estimation without Reliance on Design-based Variances

María Bugallo Porto

In collaboration with:
María José Lombardía and Alexandro Aneiros-Batista

MODES research groups & CITIC, University of A Coruña, Spain

❀ Problem of interest

What is our motivation for this research?

Main goal

Answer the requests of the research project: “Small Area Estimation for the use of auxiliary information in survey data disaggregation”.

Funding institution

Spanish National Institute of Statistics (INE)

Work plan (2025–2026)

- Estimate labour and wage indicators in unplanned domains with small sample sizes
- Perform analysis for specific regions and subpopulations

Methodology

Small Area Estimation (SAE) techniques

Spanish Quarterly Labour Cost Survey

Within a project funded by the Spanish National Institute of Statistics

Estimating the effective hours worked in small areas in Spain, benchmarking results against the following aggregates:

- ✔ Autonomous Communities and sectors of activity.
- ✔ Divisions and company size groups.

- Unplanned domains with no sample.

- In the **2º quarter of 2024**, domains with data are reduced to $D = 2500$.

Small Area Estimation Problem

Sample information is scarce in many estimation domains.

	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
n_d	**1**	**3**	**5**	8	11	74	5	25	43	62	79	96
N_d	2	10	29	85	369	37,263	7	54	136	355	1,100	37,604
f_d	0.12	3.03	8.33	16.67	30.77	85.71	0.15	1.90	4.15	7.95	14.51	66.67

(a) NUTS 2 x NACE x Size. (b) NUTS 2 x NACE.

Table 1: Summary of some descriptive results.

Small Area Estimation Problem

Sample information is scarce in many estimation domains.

	0%	20%	40%	60%	80%	100%	0%	20%	40%	60%	80%	100%
n_d	**1**	**3**	**5**	8	11	74	5	25	43	62	79	96
N_d	2	10	29	85	369	37,263	7	54	136	355	1,100	37,604
f_d	0.12	3.03	8.33	16.67	30.77	85.71	0.15	1.90	4.15	7.95	14.51	66.67

(a) NUTS 2 x NACE x Size. (b) NUTS 2 x NACE.

Table 1: Summary of some descriptive results.

The **Hájek estimator** of the population mean is given by:

$$\widehat{\overline{Y}}_d^{dir} = \frac{1}{\widehat{N}_d} \sum_{j \in s_d} \omega_{dj} y_{dj}, \qquad \widehat{N}_d = \sum_{j \in s_d} \omega_{dj}.$$

❀ Background and motivation

Review of existing area-level approaches

Classic: **Fay-Herriot model** (LMM, Fay and Herriot, 1979):

$$\widehat{\overline{Y}}_d^{dir} = \overline{\boldsymbol{X}}_d'\boldsymbol{\beta} + u_d + e_d, \quad u_d \sim N(0, \sigma_u^2), \quad d = 1, \ldots, D.$$

Review of existing area-level approaches

Classic: **Fay-Herriot model** (LMM, Fay and Herriot, 1979):

$$\widehat{\overline{Y}}_d^{dir} = \overline{\boldsymbol{X}}_d'\boldsymbol{\beta} + u_d + e_d, \quad u_d \sim N(0, \sigma_u^2), \quad d = 1, \ldots, D.$$

Recent: **Fay-Herriot M-quantile model** (MQ, Marchetti et al., 2025):

$$\widehat{\overline{Y}}_d^{dir} = \overline{\boldsymbol{X}}_d'\boldsymbol{\beta}_\psi(q_{d,\sigma_d}) + e_{\psi,d}, \quad q_{d,\sigma_d} \in (0,1), \quad d = 1, \ldots, D.$$

Review of existing area-level approaches

Classic: **Fay-Herriot model** (LMM, Fay and Herriot, 1979):

$$\widehat{\overline{Y}}_d^{dir} = \overline{\boldsymbol{X}}_d'\boldsymbol{\beta} + u_d + e_d, \quad u_d \sim N(0, \sigma_u^2), \quad d = 1, \ldots, D.$$

Recent: **Fay-Herriot M-quantile model** (MQ, Marchetti et al., 2025):

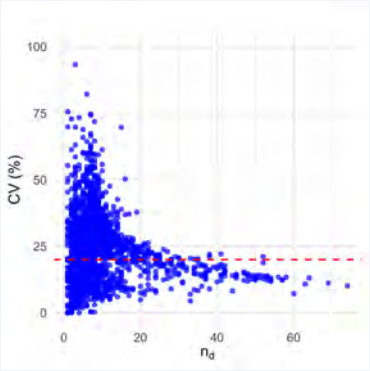
$$\widehat{\overline{Y}}_d^{dir} = \overline{\boldsymbol{X}}_d'\boldsymbol{\beta}_\psi(q_{d,\sigma_d}) + e_{\psi,d}, \quad q_{d,\sigma_d} \in (0,1), \quad d = 1, \ldots, D.$$

Common features ($d = 1, \ldots, D$)

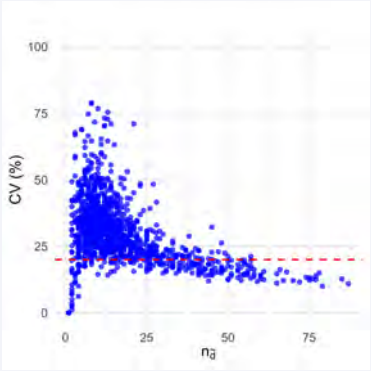
- Model errors: $e_d, e_{\psi,d} \sim N(0, \sigma_d^2)$
- Known sampling variances

$$\widehat{\sigma}_d^2 := \frac{1}{\widehat{N}_d^2} \sum_{j \in s_d} \omega_{dj}(\omega_{dj} - 1)\big(y_{dj} - \widehat{\overline{Y}}_d^{dir}\big)^2.$$

Why not rely on the design-based variances?



(a) NUTS 2 × NACE × Size.



(b) NUTS 2 × NACE.

Figure 1: Design-based CV estimates for the Hájek estimator of effective hours versus small-area sample sizes. Red line: CV = 20%.

❀ Methodological proposal

Less restrictive area-level MQ models

Extending the unit-level approach by Chambers and Tzavidis (2006)

Novel: For each quantile index $q \in (0,1)$,

$$\widehat{\overline{Y}}_d^{dir} = \overline{\boldsymbol{X}}_d' \boldsymbol{\beta}_\psi(q) + e_{\psi,d}(q), \quad d = 1, \ldots, D,$$

where $Q_q(e_{\psi,d}(q); \sigma_q, \psi | \overline{\boldsymbol{X}}_d) = 0$; and common variance σ_q^2.

Less restrictive area-level MQ models

Extending the unit-level approach by Chambers and Tzavidis (2006)

Novel: For each quantile index $q \in (0,1)$,

$$\widehat{\overline{Y}}_d^{dir} = \overline{\boldsymbol{X}}_d' \boldsymbol{\beta}_\psi(q) + e_{\psi,d}(q), \quad d = 1, \ldots, D,$$

where $Q_q(e_{\psi,d}(q); \sigma_q, \psi | \overline{\boldsymbol{X}}_d) = 0$; and common variance σ_q^2.

For SAE: A **distribution-free** strategy captures inter-area variability:

$$q_d = \underset{0<q<1}{\text{solution}} \left\{ \widehat{\overline{Y}}_d^{dir} = \overline{\boldsymbol{X}}_d' \boldsymbol{\beta}_\psi(q) \right\}.$$

After fitting, a grid-search is performed in practice for each domain.

Linear prediction and MSE estimation

Plug-in: Based on the new AMQ models for $q = \widehat{q}_d$,

$$\widehat{\overline{Y}}_d^{amq} = \overline{\boldsymbol{X}}'_d \widehat{\boldsymbol{\beta}}_\psi(\widehat{q}_d), \quad d = 1, \ldots, D.$$

Linear prediction and MSE estimation

Plug-in: Based on the new AMQ models for $q = \widehat{q}_d$,

$$\widehat{\overline{Y}}_d^{amq} = \overline{\boldsymbol{X}}'_d \widehat{\boldsymbol{\beta}}_\psi(\widehat{q}_d), \quad d = 1, \ldots, D.$$

Two options for the MSE estimation of the predictors:

Analytical: Based on a Taylor series expansion:

$$mse_d^{anlt} = (\boldsymbol{a}'_d - \mathbf{1}'_d) \operatorname*{diag}_{1 \le g \le D} (\widehat{\sigma}_g^2)(\boldsymbol{a}'_d - \mathbf{1}'_d)' + (2a_{dd} - 1)\widehat{\sigma}_d^2$$
$$+ \Big(\sum_{g=1}^{D} a_{dg} \overline{\boldsymbol{X}}'_g \widehat{\boldsymbol{\beta}}_\psi(\widehat{q}_g) - \widehat{\overline{Y}}_d^{dir} \Big)^2.$$

Linear prediction and MSE estimation

Plug-in: Based on the new AMQ models for $q = \widehat{q}_d$,

$$\widehat{\overline{Y}}_d^{amq} = \overline{\boldsymbol{X}}'_d \widehat{\boldsymbol{\beta}}_\psi(\widehat{q}_d), \quad d = 1, \ldots, D.$$

Two options for the MSE estimation of the predictors:

Analytical: Based on a Taylor series expansion:

$$mse_d^{anlt} = \left(\boldsymbol{a}'_d - \mathbf{1}'_d\right) \operatorname*{diag}_{1 \le g \le D} (\widehat{\sigma}_g^2) \left(\boldsymbol{a}'_d - \mathbf{1}'_d\right)' + (2a_{dd} - 1)\widehat{\sigma}_d^2$$

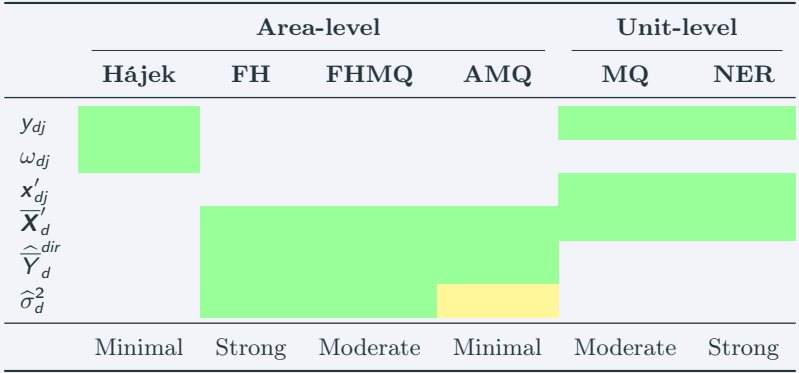
$$+ \Big(\sum_{g=1}^{D} a_{dg} \overline{\boldsymbol{X}}'_g \widehat{\boldsymbol{\beta}}_\psi(\widehat{q}_g) - \widehat{\overline{Y}}_d^{dir} \Big)^2.$$

Bootstrap: More flexible semiparametric approach (...):

$$mse_d^{boot} = \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{\overline{Y}}_d^{amq*(b)} - \overline{Y}_d^{*(b)} \right)^2.$$

Data Requirements

Linear Prediction and MSE estimation



Color coding: green = required, yellow = optional, white = not required.

❀ Empirical results

Summary of the simulation results (I)

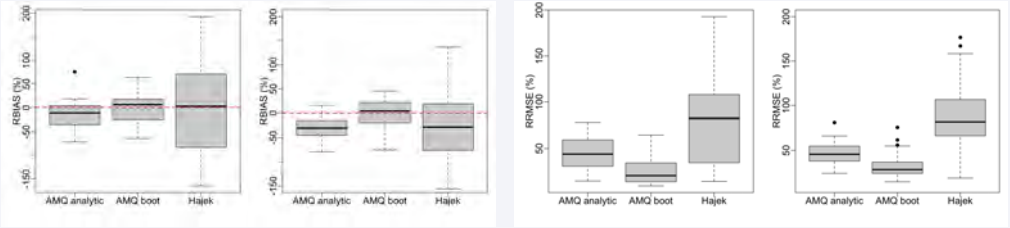
Simulation 1: Performance of population mean predictors

	No outliers	**Outliers**	
	[0, 0]	[3% e, 0]	[3% e, 10% u]
FH model	✔✔		
FHMQ model		✔✔	✔
AMQ model		✔	✔✔

Table 2: Prediction performance in model-based simulations; $n_d = 5$.

Summary of the simulation results (II)

Simulation 2: Reliability of the variance and MSE estimators



(a) Relative bias: [0,0] & [e,u].

(b) Relative root-MSE: [0,0] & [e,u].

Figure 2: Performance of the variance and MSE estimators; $n_d = 5$.

Summary of the simulation results (III)

Effect of the sample sizes on the MSE estimation

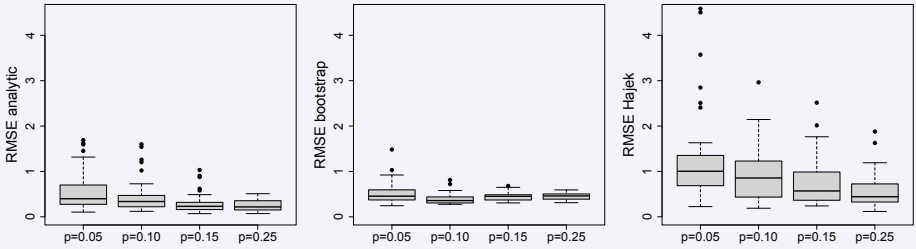


Figure 3: Performance of MSE (analytical & bootstrap) and design-based variance estimates for $D = 40$, $N = 200$ and $n = Np \in \{10, 20, 30, 50\}$.

❀ Application to real data

Spanish Quarterly Labour Cost Survey

Within a project funded by the Spanish National Institute of Statistics

Estimating the effective hours worked in small areas in Spain, benchmarking results against the following aggregates:

- ✔ Autonomous Communities and sectors of activity.
- ✔ Divisions and company size groups.

- Unplanned domains with no sample.

- In the **2º quarter of 2024**, domains with data are reduced to $D = 2500$.

Prediction and root-MSE estimation

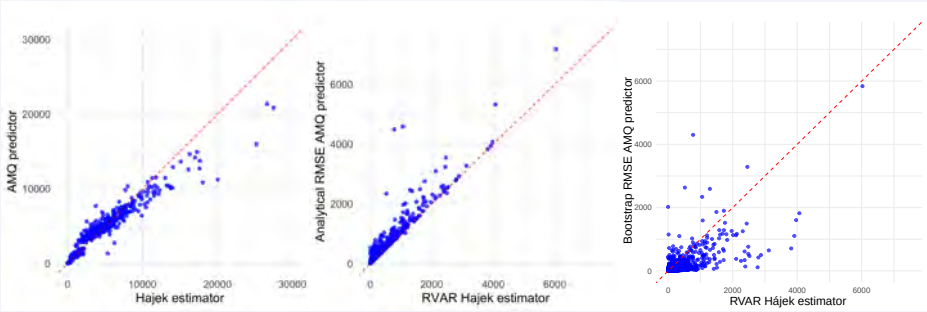


Figure 4: Left: AMQ predictor vs Hájek estimator for population means of effective hours; Center & right: analytical and semiparametric bootstrap RMSEs of AMQ ($B = 100$) vs design-based RVARs.

- Aligned AMQ predictions and Hájek estimates.
- Reduced root-MSE estimates for the AMQ predictor.

Bootstrap & design-based coefficients of variation

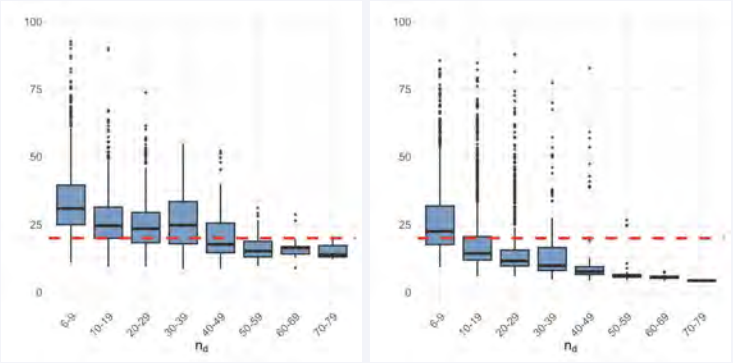


Figure 5: Left: CV bootstrap of the AMQ predictor; Right: Design-based CV of the Hájek estimator. Red line: CV=20%.

Key Findings

Main contribution

This work extends Small Area Estimation methods to business surveys using flexible M-quantile area models.

- ✓ **More reliable inference:** Hájek design-based variances are poorly estimated.
- ✗ **Benchmarking warning:** Not recommended under high heterogeneity across domains.
- ☞ **Next steps:** Temporal extensions and deeper analysis of benchmarking.

Improving robustness and flexibility in business survey estimation

5th ISI REGIONAL STATISTICS CONFERENCE MALTA 2026

THANK YOU

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Cofinanciado por la Unión Europea
AGENCIA ESTATAL DE INVESTIGACIÓN
gmodes DPTO. DE MATEMÁTICAS
citic
UNIVERSIDADE DA CORUÑA