Benchmarking a HAR(1) Fay-Herriot Model for Estimating Labour Indicators from the Spanish Quarterly Labour Cost Survey

Alexandro Aneiros-Batista, María José Lombardía Cortiña, and Esther López Vizcaíno

Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain.
IGE, Instituto Galego de Estatística.
Correspondence: alexandro.aneiros.batista@udc.es

DOI: *https://doi.org/10.17979/spu.23.c11*

Abstract: We develop a scalable multivariate small-area framework to produce publication-grade labour-cost indicators from Spain's Quarterly Labour Cost Survey (QLCS) over fine domains defined by Autonomous Communities × 2-digit NACE divisions × three enterprise-size classes. A bivariate Fay–Herriot model with a heteroscedastic autoregressive HAR(1) structure jointly links each cost component to its natural "base" (labour cost with workers; wage cost with effective hours), exploiting cross-equation dependence while remaining parsimonious and stable. Estimation proceeds via REML; uncertainty is quantified with closed-form MSE expressions specific to the HAR(1) setting, reported as coefficients of variation (CV). To ensure tractability at national scale, we introduce a *Divide-et-Impera* strategy: models are fitted in parallel by Autonomous Community and optimally recombined using Fisher-information weights, delivering large speed-ups with negligible efficiency loss. For institutional coherence, we implement a fast triple-conformity benchmarking heuristic—with positivity constraints—aligning model-based totals with official aggregates across enterprise size, industry and territory; ratios are reconstructed from benchmarked numerators and denominators. Using 2024Q2 data and auxiliary covariates, the approach substantially stabilises estimates: median CVs are $\leqslant 20\%$ and seldom exceed 40%, meeting international dissemination standards. The pipeline—HAR(1) BFH + Divide-et-Impera + positivity-preserving benchmarking—offers a reproducible, policy-ready solution for releasing coherent labour- and wage-cost indicators at granular levels.

1 Introduction

Labour cost per worker and per effective hour, together with wage cost, are key indicators for understanding the dynamics of the labour market. They summarise both gross remuneration and employer contributions, thus reflecting the real economic effort behind each unit of work. Reliable measurement of these costs is crucial for evaluating competitiveness, guiding employment policies, and supporting collective bargaining.

The Quarterly Labour Cost Survey (QLCS), conducted by the Spanish National Statistics Institute (INE), is the official source for these indicators. However, its sample design ensures precision only at broad levels (e.g. national or sectoral). When disaggregated into smaller domains—such as Autonomous Community, 2-digit from the Statistical Classification of Economic Activities in the European Community (NACE) division, and enterprise size—the direct estimators become highly variable and often unsuitable for publication. This motivates the use of Small Area Estimation (SAE) methods, which combine survey data with auxiliary sources to stabilise estimates and extend their usability in official statistics.

2 Problem Description and Objective of the Study

The main statistical challenge lies in producing coherent and reliable estimates of total labour cost per worker (LC.TR) and wage cost per effective hour (WC.EH) across a detailed cross-classification: *Autonomous Community* $\times$ *2-digit NACE division* $\times$ *enterprise-size group* (*1–49, 50–199, 200+*) This design leads to $D \approx 2500$ domains, many with very small or even zero sample sizes, making direct estimators unstable and their coefficients of variation (CV) too large for dissemination.

To address this, we propose a bivariate Fay–Herriot (BFH) model with a heteroscedastic autoregressive (HAR(1)) structure, which jointly models labour and wage costs with their corresponding "bases" (workers and effective hours). This joint specification borrows information across related outcomes, reducing variance and preventing incoherent results.

However, direct application of BFH to thousands of domains generates computational bottlenecks and numerical instabilities. To overcome this, we design a Divide-et-Impera (DeI) strategy: the domain set is partitioned by Autonomous Community ($K = 17$), models are fitted locally in parallel, and results are aggregated using Fisher-information weights. This approach preserves asymptotic efficiency while delivering substantial reductions in runtime.

Finally, to guarantee coherence with official INE totals, we implement a triple-conformity benchmarking heuristic with positivity constraints, aligning model-based numerators and denominators across enterprise size, industry, and territory before reconstructing ratios. The objective is thus twofold:

1. Produce *publication-grade* small area estimates of labour and wage costs at highly disaggregated levels, meeting international CV standards ($\leqslant 20\%$ for most domains).
2. Ensure institutional coherence and reproducibility of the results, making them suitable for direct use in official statistics.

3 Literature Review

The use of SAE methods in official statistics has been widely studied over the last three decades. The FH model remains the cornerstone for area-level inference (Fay and Herriot (1979); Prasad and Rao (1990)), offering stable estimators through shrinkage toward regression predictions. Extensions to multivariate settings, such as the models of Benavent and Morales (2015), allow information to be shared across correlated outcomes and are especially suited to economic indicators.

Recent developments address computational scalability through divide-and-conquer strategies (Banerjee et al. (2019)) and reinforce institutional coherence with benchmarking procedures (González-Manteiga et al. (2008)). Applications in labour statistics highlight the relevance of these methods for producing reliable disaggregated indicators (Morales et al. (2021)).

This study builds directly on these contributions, combining multivariate FH modelling, scalable Divide-et-Impera estimation, and benchmarking heuristics to deliver coherent, publication-ready labour cost estimates for Spain.

4 Methodology

The methodological strategy combines BFH modelling with a HAR(1), supported by auxiliary administrative and census information. The direct estimators are Hájek ratios (Hájek (1971)), which constitute the official design-based baseline in the QLCS.

The key methodological components are:

- Integration of auxiliary information from large-coverage sources: census registers of workers, the Central Business Register (CBR/DIRCE) turnover, and survey-based indicators. These covariates stabilise small-domain estimates and capture structural variation.

- Application of a HAR(1) bivariate Fay–Herriot model to jointly estimate pairs of outcomes (labour cost with number of workers, wage cost with effective hours). This multivariate setting exploits cross-equation dependence and reduces incoherent results across related variables.
- Estimation of variance components and regression parameters by restricted maximum likelihood (REML). Domain-level uncertainty is quantified through closed-form mean squared error (MSE) expressions specific to the HAR(1) model, reported as CVs.
- Implementation of a DeI strategy: the $D \approx 2500$ domains are partitioned into $K = 17$ Autonomous Communities. Models are fitted independently in parallel, and local results are combined through Fisher-information weights, preserving global efficiency while achieving major computational gains.
- Benchmarking heuristic with positivity constraints, enforcing triple conformity with INE aggregates across enterprise size, NACE division, and Autonomous Community. Since direct benchmarking of ratios is not feasible, adjustments are applied separately to numerators and denominators, after which ratios are recomputed.

This combined framework improves precision and guarantees coherence with official aggregates. Following international standards (Office for National Statistics (2006)), a CV threshold of 20% is taken as the benchmark for publication-quality estimates, with most domains meeting this criterion and rarely exceeding 40%.

5 Application to Official Statistics — Our Contribution

5.1 Model specifications

Using the HAR(1) bivariate Fay–Herriot model, two specifications were estimated:

- **Model A** considers the response vector $\mathbf{y}_d = \{LC.Dh, TR\}$, i.e. labour cost relative to census workers and the direct estimation of the number of workers. The derived indicator, that is, the quotient of interest, is the total labour cost per worker ($LC.TR$).
- **Model B** considers $\mathbf{y}_d = \{WC.Dh, EH.Dh\}$, i.e. wage cost per census worker and effective hours per census worker. The derived indicator is the wage cost per effective hour ($WC.EH$).

Both models share the same HAR(1) covariance structure but differ in their response pairs. The auxiliary covariates are chosen to capture the size and intensity of labour input and enterprise activity:

$$x_1 = D_h \text{ (census workers)}, \quad x_2 = \sqrt{TR/D_h}, \quad x_3 = \frac{(CBRT/D_h)}{\text{med}(\cdot)} \text{ (scaled turnover)}.$$

These regressors provide structural anchors to stabilise domain-level predictions, reflecting both the labour supply base (D_h) and economic turnover (CBRT).

5.2 Estimated coefficients

We use data from the second quarter of 2024. The tables below report the estimated regression coefficients (β) and their standard errors for Models A and B.

The results confirm the role of turnover (x_3) as a positive driver of both labour and wage costs, while census workers (x_1) and x_2 capture domain size and workforce composition. In

	LC.Dh	TR
Intercept	-45.7 (1.27)	47.8 (1.91)
x_1	–	0.0028 (0.0002)
x_2	1088.3 (10.29)	–
x_3	0.31 (0.05)	0.0083 (0.0091)

Table 1: Model A: Labour cost per worker ($LC.TR$). Standard errors in parentheses. Variable x_3 was retained in the model despite not being statistically significant, for reasons of economic interpretation and coherence.

	WC.Dh	EH.Dh
Intercept	-23.73 (0.77)	6.70 (0.22)
x_1	–	-7.89×10^{-5} (1.05×10^{-5})
x_2	633.94 (6.96)	–
x_3	0.375 (0.032)	0.0317 (0.0028)

Table 2: Model B: Wage cost per effective hour ($WC.EH$). Standard errors in parentheses.

Model A, the strong coefficient for x_2 (1088.3) reflects the close link between labour cost per worker and the intensity of employment relative to the census base. In Model B, the significant effect of x_3 and the negative sign of x_1 on effective hours highlight the inverse relationship between workforce scale and individual hourly costs.

5.3 Goodness of fit and uncertainty

Both specifications were assessed under the full HAR(1) + DeI + benchmarking framework. Diagnostics confirmed the normality of random effects and the homoscedasticity assumption. Precision was quantified by CV. Following Office for National Statistics (2006) standards, a 20% threshold is considered the upper limit for publication. Our results show that CVs do not exceed 20% at the median and are below 40% in most domains, even under strong disaggregation, thus meeting international criteria for dissemination of official statistics.

5.4 Graphical diagnostics

Figures 1 and 2 illustrate key diagnostics for both models:

- CV distributions across domains (a), showing substantial variance reduction compared to direct estimators.
- Scatter plots of direct vs. model-based estimates (b), illustrating how the EBLUPs shrink noisy direct values toward stable regression predictions. For clarity of presentation, values below zero are truncated in the plots.
- Benchmarking plots (c), demonstrating how positivity constraints and triple conformity align model-based totals with INE aggregates.

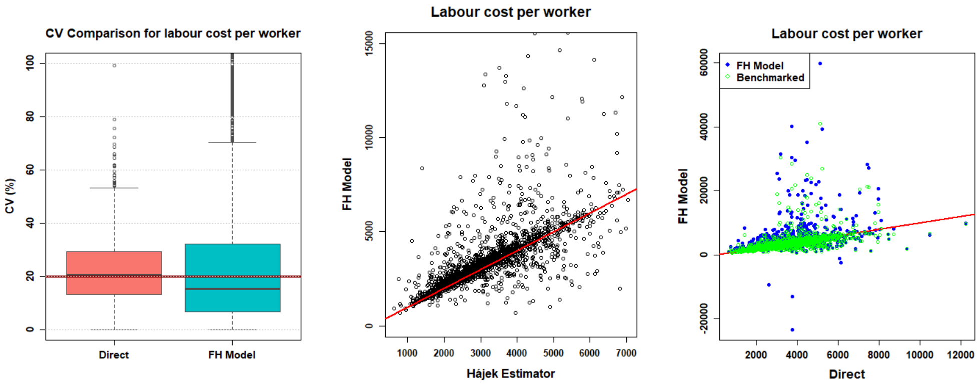

Figure 1: Diagnostics for Model A (LC.TR): (a) CV plots, (b) model vs. direct (or Hájek estimator), (c) benchmarking.

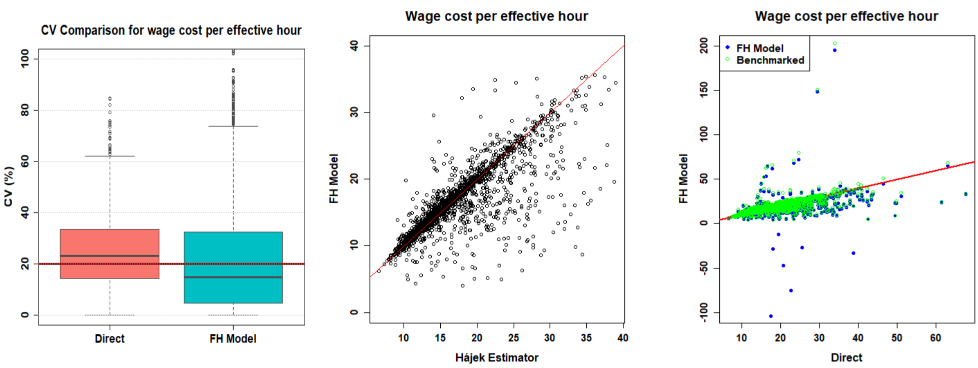

Figure 2: Diagnostics for Model B (WC.EH): (a) CV plots, (b) model vs. direct (or Hájek estimator), (c) benchmarking.

6 Conclusions

The evidence shows that the HAR(1) FH model, combined with Divide-et-Impera partitioning and a fast benchmarking heuristic, delivers:

1. Publication-ready small area estimates of labour and wage costs at highly disaggregated levels.
2. Precision consistent with official statistics standards (CVs generally under 20% at the median).
3. Coherent totals aligned with INE aggregates, guaranteed by triple-conformity and positivity constraints.
4. A scalable and reproducible methodology that makes national deployment feasible in practice.

In summary, our contribution is the development of a multivariate SAE framework that is at once statistically rigorous, computationally efficient, and directly usable in official statistics production.

Bibliography

M. Banerjee, C. Durot, and B. Sen. Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *Annals of Statistics*, 47(2):720–757, 2019.

R. Benavent and D. Morales. Multivariate fay–herriot models for small area estimation. *Computational Statistics & Data Analysis*, 94:372–390, 2015.

R. E. Fay and R. A. Herriot. Estimates of income for small places: An application of james–stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277, 1979.

W. González-Manteiga, M. J. Lombardía, I. Molina, D. Morales, and L. Santamaría. Estimation of the mean squared error of predictors based on the fay–herriot model. *Journal of Statistical Planning and Inference*, 138(2):308–321, 2008.

J. Hájek. Comment on "an essay on the logical foundations of survey sampling, part one". In V. P. Godambe and D. A. Sprott, editors, *The Foundations of Survey Sampling*, page 236. Holt, Rinehart and Winston, New York, 1971.

D. Morales, M. D. Esteban, A. Pérez, and T. Hobza. *A Course on Small Area Estimation and Mixed Models: Methods, Theory and Applications in R*. Springer, 2021.

Office for National Statistics. Model-based estimates of ilo unemployment for lad/uas in great britain: Guide for users. Technical report, Office for National Statistics, 2006.

N. G. N. Prasad and J. N. K. Rao. The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85(409):163–171, 1990.