

Uso de R na Estimación en Áreas Pequenas

Un caso real con datos bidimensionais

María Bugallo¹, Esteban Cabello¹, María Dolores Esteban¹,
Domingo Morales¹, Agustín Pérez², Sofía Rodríguez-Ballesteros¹

¹Instituto Universitario “Centro de Investigación Operativa”

²Departamento de Estudos Económicos e Financeiros

Universidade Miguel Hernández de Elche

XII Xornada de Usuarios de R en Galicia
Santiago de Compostela, 16 de outubro do 2025

O mundo dos datos

* O mundo está a xerar tantos datos cada dous días como os que se crearon desde o inicio da civilización ata o ano 2003.

(Eric Schmitt, exCEO de Google).

* A cantidade de datos no mundo duplícase cada 18 meses. *(IBM).*

Pero ao mesmo tempo...

O imperativo de «**Non deixar a ninguén atrás**» implica dispoñer de estatísticas sobre grupos e minorías que hoxe son “*invisibles*”.

* O reto da Estatística Oficial:

Máis detallada, rápida, precisa e máis barata.

(Holt, D.T., *The American Statistician*, 2007, **61(1)**, 1–8)

Pero ao mesmo tempo...

O imperativo de «**Non deixar a ninguén atrás**» implica dispoñer de estatísticas sobre grupos e minorías que hoxe son “*invisibles*”.

* O reto da Estatística Oficial:
Máis detallada, rápida, precisa e máis barata.

(Holt, D.T., *The American Statistician*, 2007, **61**(1), 1–8)



OBXECTIVOS DE DESENVOLVEMENTO SUSTENTABLE



Que é a Estatística Oficial? Por que é importante?

Publicación dunha **Oficina Nacional de Estatística** baseada nunha enquisa, censo, datos administrativos, etc.

- A Estatística Oficial está presente na vida cotiá e inclúe resultados sobre taxas de desemprego, medidas de pobreza, estatísticas de saúde e medio ambiente, etc.
- Para moita xente, a Estatística Oficial é a esencia da estatística!!
- Os responsables políticos deberían usar na toma de decisións.

Na Estatística Oficial, os orzamentos axustados e a demanda de resultados detallados dan lugar a problemas de **Estimación en Pequenas Áreas** (*Small Area Estimation*, SAE).

Cal é o principal problema?

Os **tamaños mostrais** son pequenos!
Ou simplemente... non hai datos!

→ Dominios non planificados → Falta de información

- ★ Variabilidade inaceptablemente alta dos métodos tradicionais
- ★ Coeficientes de variación superiores ao 20–30 %
- ★ Resultados pouco fiables ou... *inexistentes*!



Cal é a proposta estatística?

A **Estimación en Pequenas Áreas** mellora a fiabilidade das estimacións a nivel granular ao “*tomar forza prestada*” de variables auxiliares, información doutros subgrupos e estruturas de dependencia.

O máis habitual é empregar modelos lineais (xeralizados) mixtos con intercepto aleatorio.

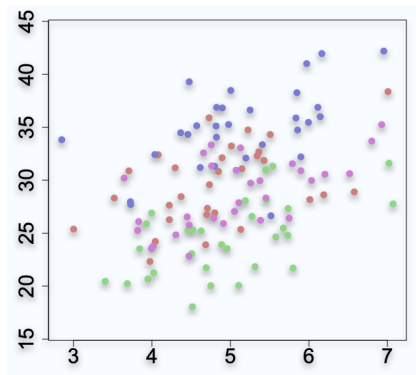
Modelos de área

- Modelo de **Fay-Herriot** (1979)

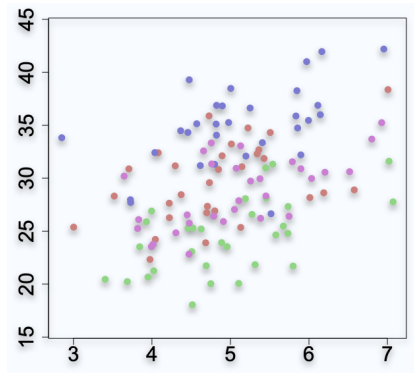
Modelos de unidade

- **Modelo de erro aniñado** (Battese, Harter e Fuller, 1988)

Un exemplo de xoguete no caso univariante...



Un exemplo de xoguete no caso univariante...

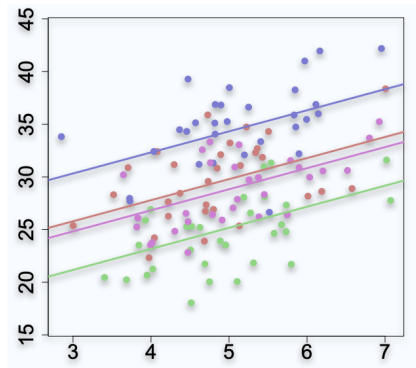


Librerías de R

- Modelos clásicos:
sae (Molina e Marhuenda, 2015)
- Indicadores de pobreza e desigualdade: emdi (Kreutzmann et al., 2019)
- Modelos bayesianos:
hbsae (Boonstra, 2022)

Un exemplo de xoguete no caso univariante...

$$y_{dj} = \mathbf{x}'_{dj}\beta + u_d + e_{dj}, \quad u_d \sim N(0, \sigma_u^2), \quad e_{dj} \sim N(0, \sigma_e^2), \quad \sigma_u^2, \sigma_e^2 > 0.$$



Librerías de R

- **Modelos clásicos:**
`sae` (Molina e Marhuenda, 2015)
- Indicadores de pobreza e desigualdade: `emdi`
(Kreutzmann et al., 2019)
- Modelos bayesianos:
`hbsae` (Boonstra, 2022)

Índice

✿ Motivación

✿ Problema de interese e datos

✿ Problema estatístico e proposta

✿ Aplicación a datos reais

✿ Consideracións finais

Proxecto de transferencia de coñecemento

Estimación en áreas
pequenas de indicadores
multivariáveis en España

OBXECTIVO XERAL

Mellorar a calidade da información estatística oficial a partir de datos de enquisas.

INE

Instituto Nacional de Estadística

Estimación en áreas pequeñas de indicadores multivariados en España

OBJECTIVO XERAL

Mellorar a calidade da información estatística oficial a partir de datos de enquisas.



Instituto Nacional de Estadística

PRIORIDADES 2025

Predicción de indicadores no lineais:

02. Custo laboral (CTB) e salarial (S), por traballador (TR) e hora efectiva (HE), a nivel de CCAA $\times$ división CNAE $\times$ tamaño empresarial.

Enquisa Trimestral de Custo Laboral

Sobre a enquisa...

- Mostra representativa de empresas do territorio nacional
- Periodicidade trimestral
- Resultados INE moi agregados e xerais

Enquisa Trimestral de Custo Laboral

Sobre a enquisa...

- Mostra representativa de empresas do territorio nacional
- Periodicidade trimestral
- Resultados INE moi agregados e xerais

Sobre o problema de interese...

- Variables obxectivo **bivariantes**:
 $(CTB_{dj}, TR_{dj}); (S_{dj}, TR_{dj});$
 $(CTB_{dj}, HE_{dj}); (S_{dj}, HE_{dj});$
 $d = 1 \dots, D$ cruces de 17 CCAA,
 82 divisións CNAE e 3 grupos de
 tamaño empresarial (1-49, 50-199,
 200 e máis traballadores); e
 $j = 1, \dots, N_d$ empresas.

Custo laboral por traballador e salarial por hora efectiva

Os parámetros de interese son aditivos, como as **medias poboacionais**:

$$\begin{aligned}CTB_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} CTB_{dj}, & TR_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} TR_{dj}, \\ S_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} S_{dj}, & HE_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} HE_{dj}.\end{aligned}$$

Custo laboral por traballador e salarial por hora efectiva

Os parámetros de interese son aditivos, como as **medias poboacionais**:

$$\begin{aligned} \text{CTB}_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} \text{CTB}_{dj}, & \text{TR}_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} \text{TR}_{dj}, \\ S_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} S_{dj}, & \text{HE}_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} \text{HE}_{dj}. \end{aligned}$$

e non aditivos, como os **ratios de medias poboacionais**:

$$\text{CTB}.\text{TR}_d = \frac{\text{CTB}_d}{\text{TR}_d}, \quad \text{S}.\text{HE}_d = \frac{S_d}{\text{HE}_d},$$

nos cruces de 17 CCAA, 82 divisións CNAE e 3 grupos de tamaño empresarial (1-49, 50-199, 200 e máis traballadores).

Custo laboral por traballador e salarial por hora efectiva

Os parámetros de interese son aditivos, como as **medias poboacionais**:

$$\begin{aligned} \text{CTB}_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} \text{CTB}_{dj}, & \text{TR}_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} \text{TR}_{dj}, \\ S_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} S_{dj}, & \text{HE}_d &= \frac{1}{N_d} \sum_{j=1}^{N_d} \text{HE}_{dj}. \end{aligned}$$

e non aditivos, como os **ratios de medias poboacionais**:

$$\text{CTB}.\text{TR}_d = \frac{\text{CTB}_d}{\text{TR}_d}, \quad \text{S}.\text{HE}_d = \frac{S_d}{\text{HE}_d},$$

nos cruces de 17 CCAA, 82 divisións CNAE e 3 grupos de tamaño empresarial (1-49, 50-199, 200 e máis traballadores).

Pódense estimar con métodos directos, pero espérase alta variabilidade.

Índice

✿ Motivación

✿ Problema de interese e datos

✿ Problema estatístico e proposta

✿ Aplicación a datos reais

✿ Consideracións finais

Estimación en áreas pequenas de indicadores bivariantes

- Estimacións directas (INE)
- Modelo de Fay-Herriot
- Modelo de erro aniñado

PERO... as propostas máis sinxelas non se adaptan á natureza do problema!

Estimación en áreas pequenas de indicadores bivariantes

- Estimacións directas (INE)
- Modelo de Fay-Herriot
- Modelo de erro aniñado

PERO... as propostas máis sinxelas non se adaptan á natureza do problema!

POR ISO... xeralízase o modelo de erro aniñado a datos bivariantes (modelo BNER) ([Esteban et al., 2022a](#), [2022b](#))

$$\begin{pmatrix} y_{dj1} \\ y_{dj2} \end{pmatrix} = \begin{pmatrix} \mathbf{x}'_{dj1} & 0 \\ 0 & \mathbf{x}'_{dj2} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_{d1} \\ u_{d2} \end{pmatrix} + \begin{pmatrix} e_{dj1} \\ e_{dj2} \end{pmatrix},$$

$$\begin{pmatrix} u_{d1} \\ u_{d2} \end{pmatrix} \sim N_2(\mathbf{0}_2, V_{ud}), \quad \begin{pmatrix} e_{dj1} \\ e_{dj2} \end{pmatrix} \sim N_2(\mathbf{0}_2, V_{edj}), \quad V_{ud}, V_{edj} \in \mathcal{M}_{2 \times 2}.$$

O axuste do novo modelo, o cálculo dos predictores e as medidas de erro programáronse en R.

Índice

✿ Motivación

✿ Problema de interese e datos

✿ Problema estatístico e proposta

✿ Aplicación a datos reais

✿ Consideracións finais

Comparativa das prediccións co enfoque directo

- Conformidade: Método Nelder-Mead (Nelder and Mead, 1965): `constrOptim` + `optim`

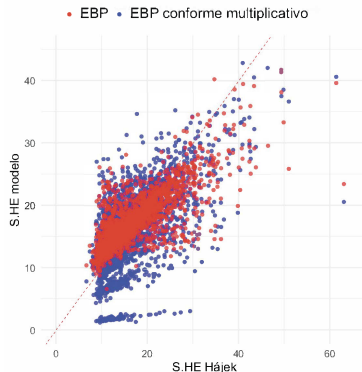
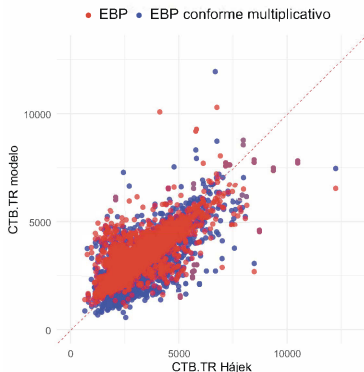


Figure: Comparación entre o EBP e o EBP conforme multiplicativo para CTB.TR (esquerda) e S.HE (dereita), fronte ás estimacións de Hájek.

Índice

✱ Motivación

✱ Problema de interese e datos

✱ Problema estatístico e proposta

✱ Aplicación a datos reais

✱ Consideracións finais

Consideracións finais

Estimación en áreas pequenas de indicadores multivariables en España

METODOLOXÍA. Aplicouse un modelo de erro aninado bivalente para estimar o custo laboral por traballador e o custo salarial por hora efectiva en áreas pequenas.

REDUCCIÓN DA VARIABILIDADE

- Respecto ás estimacións directas
- Temporal (entre trimestres)

Consideracións finais

Estimación en áreas pequenas de indicadores multivariables en España

METODOLOXÍA. Aplicouse un modelo de erro anidado bivalente para estimar o custo laboral por traballador e o custo salarial por hora efectiva en áreas pequenas.

REDUCCIÓN DA VARIABILIDADE

- Respecto ás estimacións directas
- Temporal (entre trimestres)

CONFORMIDADE cos resultados publicados polo INE en niveis superiores de agregación.

RESULTADOS GEOGRÁFICOS

Observáronse diferenzas significativas entre CCAA, divisións CNAE e tamaños empresariais.

E RECORDEMOS...
Programouse todo en R.

Referencias I

Battese, G. E., Harter, R. M., Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J Amer Statist Assoc*, **83**(401), 28–36.

Esteban, M.D., Lombardía, M.J., López-Vizcaíno, E., Morales, D., Pérez, A. (2022a). Small area estimation of expenditure means and ratios under a unit-level bivariate linear mixed model. *J Appl Stat*, **49**(1), 143–168.

Esteban, M.D., Lombardía, M.J., López-Vizcaíno, E., Morales, D., Pérez, A. (2022b). Empirical best prediction of small area bivariate parameters. *Scand J Stat*, **49**(4), 1699–1727.

Fay, R. E., Herriot, R. A. (1979). Estimates of income for small places. *J Amer Statist Assoc*, **74**(366a), 269–277.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., Santamaría, L. (2008). Bootstrap mean squared error of small-area EBLUP. *J Stat Comput Simul*, **78**(4), 443–462.

Hájek, J. (1971). Comment on “An essay on the logical foundations of survey sampling” by Basu. In *Foundations of Statistical Inference*, 236–242.

Referencias II

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Hernangómez, D. (2025). `mapSpain`: Administrative Boundaries of Spain. R.

Kreutzmann, A., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., Tzavidis, N. (2019). R package `emdi` for estimating and mapping regionally disaggregated indicators. *J Stat Softwe*, **91**(7), 1–33.

Molina, I., Marhuenda, Y. (2015). `sae`: An R package for Small Area Estimation. *The R Journal*, **7**(1), 81–98.

Morales, D., Esteban, M. D., Pérez, A., and Hobza, T. (2021). *A Course on Small Area Estimation and Mixed Models*. Springer.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, **7**(4), 308–313.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York, 2nd Edition.

María Bugallo Porto

Centro de Investigación Operativa
Universidade Miguel Hernández de Elche

Repositorio de
código en  GitHub :



INē
Instituto Nacional de Estadística



UNIVERSITAS
Miguel Hernández
RESEARCH INSTITUTE



UNIVERSITAS
Miguel Hernández