Estimación de indicadores muestrales de FOP a nivel micro-territorial

Christian Haedo
Director técnico-científico
Fundación Observatorio PyME (FOP)

Instituto de Economía Aplicada de la ANCE - FIEL
1 de Octubre de 2025

Proyecto Small Area Estimation (SAE) FOP-UNIBO

Por FOP:

Christian Haedo

Juan Pablo Rosiello

Por UNIBO (Department of Statistical Sciences):

Silvia De Nicolò

Maria Rosaria Ferrante

Lorenzo Mori

SAE Objetivo

Comprender la distribución de un indicador genérico en subgrupos específicos de una población: mapear un indicador.

Los subgrupos se definen técnicamente como áreas o dominios, tales como:

- ➔ áreas geográficas específicas (regiones, municipios, ditritos, ...)
- ➔ grupos socio-demográficos (niños en edad escolar, familias monoparentales, ...)
- ➔ cualquier interacción entre lo geográfico y otras características (diferentes especies de árboles en un parque nacional, ...)

Áreas o dominios pequeños

La fuente de datos es una muestra que implica:

- muy pocas observaciones para algunos subgrupos/áreas/dominios,
- quizás CERO observaciones para algunos dominios (out-of-sample).

Por lo tanto, las estimaciones estándar (directas) se miden con un error de muestreo grande $\Rightarrow$ queremos reducir ese error y hacer predicciones confiables.

Regla práctica: se recomienda utilizar SAE si el **CV** de las estimaciones directas superan el 16,6% (Statistics Canada, `https://www.statcan.gc.ca/en/start`).

Debemos apoyarnos en una técnica de reducción de la varianza

Modelos de estimación de áreas pequeñas

Integrar datos de encuestas con información secundaria/auxiliar (censos, datos de infraestructura, ambiente, telefonía móvil, ...) para mejorar la estimación.

Modelos a nivel de área

Vinculan estimaciones muestrales a nivel de área con covariables de las área $\rightarrow$ supuestos restrictivos.

Modelos a nivel de unidad

Vinculan observaciones individuales con covariables individuales $\rightarrow$ más exigentes en términos de datos.

$\Rightarrow$ Los modelos SAE no son espaciales (dominio no necesariamente geográfico), a menos que se introduzca explícitamente una estructura de dependencia espacial.

DATOS PRIMARIOS: Encuestas FOP

Encuestas a empresas (5.000 por año) realizadas por equipos propios:

1. **ENCUESTA ESTRUCTURAL Anual - EE (desde 1996) a 1.400 empresas.**
2. **ENCUESTA COYUNTURAL Trimestral - EC (desde 2004) a 500 empresas.**

- Muestras probabilísticas estratificadas mediante afijación proporcional.
- Diseño de panel con rotación parcial anual (20%) y completa cada 5 años.
- Cuestionarios estructurados en secciones fijas (serie histórica) y móviles.
- Métodos de recolección de datos: CAWI, CATI y CAPI.
- Cobertura objetivo de cuotas por estrato (228 estratos EE y 27 EC).

3. **ENCUESTAS ESPECIALES (financiamiento externo): BID, BM, UE, OIT, MINCyT, Gobiernos y organismos públicos.**

DATOS SECUNDARIOS: Bases de datos públicas

Más de 400 indicadores secundarios sobre **Demografía**, **Dinámica empresarial**, **Economía**, **Educación**, **Ambiente**, **Infraestructura**, **Político-electoral**, generados mediante técnicas de procesamiento geoestadístico a escala micro-territorial:

1. Análisis de necesidades, fuentes y documentación (data lakes);
2. Armonización y normalización para asegurar comparabilidad;
3. Homologación y validación de integridad y calidad;
4. Almacenamiento en data warehouse, con distribución en datamarts.

SAE - FOP: Objetivo

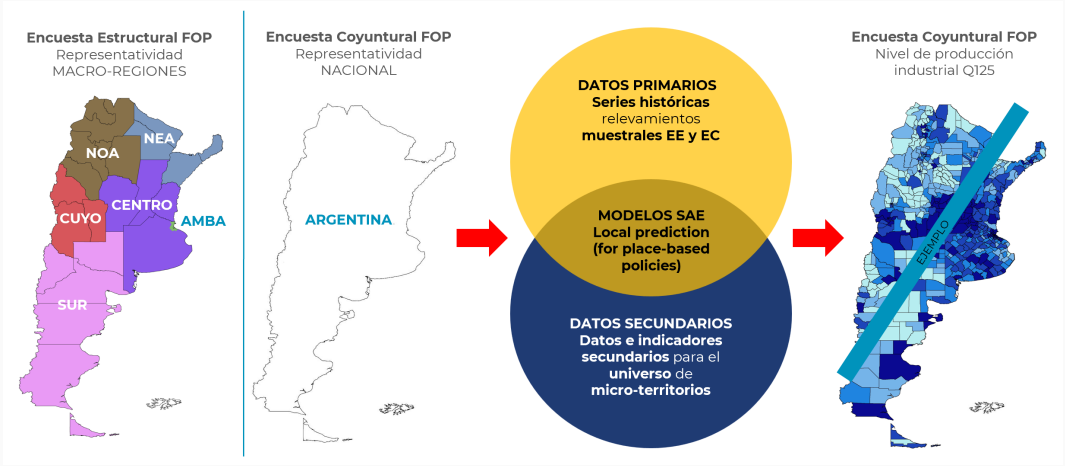


Figure 1: SAE para estimación de indicadores FOP a nivel micro-territorial (departamentos).

Índice de Confianza Empresaria - ICEPyME

Se compone de las siguientes cinco variables con ponderación uniforme:

1. Situación económica de la empresa respecto a un año atrás (**Emphace**) y dentro de un año (**Empdentro**),
2. Situación del sector respecto a un año atrás (**Sechace**) y dentro de un año (**Secdentro**),
3. Situación del país hace un año (**Paishace**) y dentro de un año (**Paisdentro**),
4. Rentabilidad de la empresa dentro de un año (**Rentdentro**),
5. Momento para invertir en maquinaria y equipo (**Momento**).

Las cuatro primeras variables se evalúan con 5 opciones: Sustancialmente mejor (2), Levemente mejor (1), Igual (0), Levemente peor (-1) y Sustancialmente peor (-2); mientras que la última con 2: Buen momento (1) y Mal momento (-1).

ICEPyME: Cálculo e indicadores derivados

1. Expectativas futuras:
   ICEdentro = **Empdentro** + **Rentdentro** + **Secdentro** + **Paisdentro**
2. Situación de la empresa:
   ICEempresa = **Emphace** + **Empdentro** + **Rentdentro**
3. Situación actual con respecto a un año atrás:
   ICEhace = **Emphace** + **Sechace** + **Paishace**
4. Situación del país:
   ICEpais = **Paishace** + **Paisdentro**
5. Situación del sector:
   ICEsector = **Sechace** + **Secdentro**
6. ICEPyME total:
   ICEtotal = **ICEhace** + **ICEdentro** + **Momento**

ICEPyME: Normalización

En todos los casos, se calcula el promedio ponderado (inversa de la probabilidad de selección) y se normaliza para que varíe entre 0 y 100.

$$NICE\ldots_t = \frac{ICE\ldots_t - Min(ICE\ldots_{it})}{Max(ICE\ldots_{it}) - Min(ICE\ldots_{it})} \times 100$$

El ICEPyME puede también estimarse directamente para cada variable de estratificación muestral (actividad manufacturera y tamaño de la empresa) y para cada estrato k (cross-analysis).

Índice de gestores de compras - PMIPyME

El PMIPyME es un índice promedio ponderado compuesto formado por las siguientes cinco variables (ponderaciones entre paréntesis)::

1. Volumen físico de producción (0,25),
2. Ocupados (0,2),
3. Cartera de pedidos (0,3),
4. Stocks de materias primas e insumos (0,1),
5. Tiempo de entrega de los proveedores (0,15).

Cada variable se evalúa con 3 opciones: Aumentó (1), No varió (0) y Disminuyó (-1).

PMIPyME: Cálculo e indicadores derivados

Considera los porcentajes de cada opción de respuesta y solo las diferencias en el período t de referencia con respecto al período anterior $(t-1)$, es decir, solo el índice de difusión de Situación (SDI_t):

$$SDI_{jt} = \left(p_{1j_{t-1}} \times 1\right) + \left(p_{2j_{t-1}} \times 0,5\right) + \left(p_{3j_{t-1}} \times 0\right) \qquad j = 1, \ldots, 5$$

Por lo tanto,

$$PMI_t = SDI_{1t} \times 0,25 + SDI_{2t} \times 0,2 + SDI_{3t} \times 0,3 + SDI_{4t} \times 0,1 + SDI_{5t} \times 0,15$$

Si el 100% de las empresas declara un mejoramiento, $PMI_t = 100$, disminución, $PMI_t = 0$, y sin cambios, $PMI_t = 50 \rightarrow PMI_t > 50$ indica una mejora, $PMI_t \approx 50$ no hubo cambios, y $PMI_t < 50$ indica disminución ($PMI_t = \leq 42$ recesión, `https://www.ismworld.org/`).

EC - FOP Industria manufacturera: Serie ICEPyME vs. PMIPyME Q215-Q125

Confianza vs. Actividad real

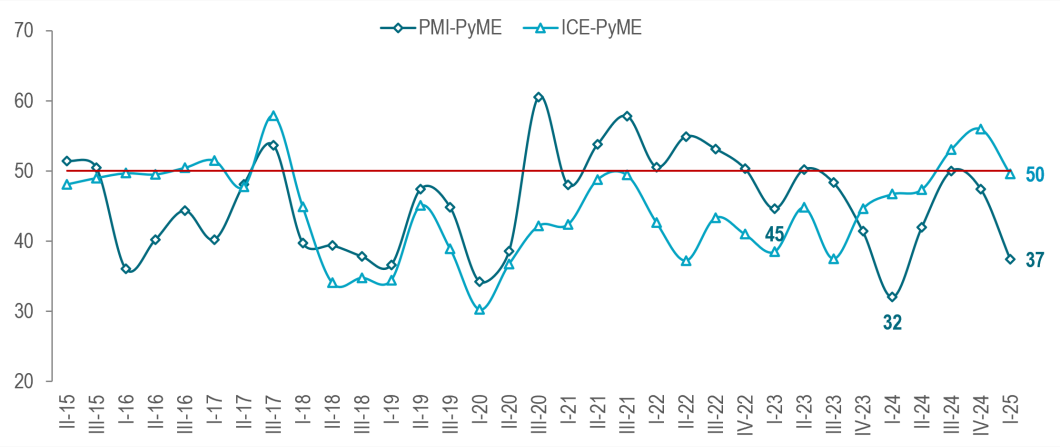


Figure 2: Serie ICEPyME vs. PMIPyME Q215-Q125.

El modelo a nivel de área de Fay-Herriot (FH) con transformación logit

Modelo muestral

$$\text{logit}(\hat{Y}_m)|\theta_m \overset{ind}{\sim} \mathcal{N}(\theta_m, \mathbb{V}_m^\star)$$

Variable target

$$\theta_m = z_m^T\beta + u_m$$

Linking model

$$\theta_m|\boldsymbol{z}_m \overset{ind}{\sim} \mathcal{N}(\boldsymbol{z}_m^T\beta, \sigma^2)$$

$$\text{logit}(\hat{Y}_m) = z_m^T\beta + u_m + \epsilon_m, \quad m = 1, \dots, M$$

$$u_m \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\epsilon_m \overset{ind}{\sim} \mathcal{N}(0, \mathbb{V}_m^\star)$$

Modelo lineal mixto, donde los efectos fijos provienen de las covariables y los aleatorios representan la variabilidad específica de cada área m.

Notación

- Supongamos que queremos estimar los verdaderos indicadores de área ICEPyME o PMIPyME para M áreas $1, \dots, M$,
- de modo que los estimadores genéricos (directos) de la encuesta $\hat{Y}_m$ tengan una gran varianza muestral $\mathbb{V}_m^{design}$ (dada por el modelo muestral y conocida $\forall m$),
- y un vector de $\boldsymbol{z}_m$ características conocidas del área, por ejemplo, demográficas (variables secundarias/auxiliares), que han sido registradas sin error.

Transformación logit

Si $\hat{Y}_m$ es una proporción, aplicamos la transformación logit:

$$y_m^\star = g(\hat{Y}_m) = \log \frac{\hat{Y}_m}{1 - \hat{Y}_m}$$

En Fay–Herriot, el supuesto es que después de la transformación los errores son aprox. normales $\Rightarrow$ la varianza se transporta mediante el método delta (primer orden):

$$V_m^\star \approx \frac{V_m^{\text{design}}}{\hat{Y}_m^2 (1 - \hat{Y}_m)^2}$$

Estimación de σ^2 en el modelo Fay–Herriot

Método de Máxima Verosimilitud Restringida (REML)

REML ajusta la verosimilitud para tener en cuenta la pérdida de grados de libertad en la estimación de β.

$$\ell_{\text{REML}}(\sigma^2) \;=\; -\tfrac{1}{2}\Big\{ \log|V(\sigma^2)| + \log|Z^\top V(\sigma^2)^{-1} Z| + (y^\star)^\top P(\sigma^2)\, y^\star \Big\}$$

donde

$$V(\sigma^2) = \operatorname{diag}(V_m^\star + \sigma^2), \qquad P(\sigma^2) = V(\sigma^2)^{-1} - V(\sigma^2)^{-1} Z (Z^\top V(\sigma^2)^{-1} Z)^{-1} Z^\top V(\sigma^2)^{-1}$$

- Al maximizar $\ell_{\text{REML}}(\sigma^2)$ se obtiene un estimador insesgado de σ^2.

Estimación (II)

REML vs. ML en modelos SAE (Maestrini et al., 2024):

- **Número de áreas**: clave para estimar la varianza de efectos aleatorios.
- **Problema de ML**: con muchas covariables y pocas áreas, se consumen muchos grados de libertad para estimar los efectos fijos, y al no descontar esa pérdida, subestima la varianza de los efectos aleatorios.
- **Ventaja de REML**: corrige el sesgo en muestras finitas, ajustando la verosimilitud y eliminando la influencia de los efectos fijos, de modo que la estimación de la varianza del área no depende de la cantidad de covariables que se hayan incluido.

Estimación (III)

Estimación GLS de β dado σ^2

En Fay–Herriot, además del componente aleatorio $u_m \overset{ind}{\sim} \mathcal{N}(0, \sigma^2)$, cada área tiene un error de muestreo con varianza conocida $V_m^\star$ distinta $\Rightarrow$ errores heterocedásticos.

Definiciones

- Vector de respuestas en escala logit:

$$y^\star = (y_1^\star, \dots, y_M^\star)^\top$$

- Matriz de diseño:

$$Z = \begin{bmatrix} z_1^\top \\ \vdots \\ z_M^\top \end{bmatrix} \quad \in \mathbb{R}^{M \times p}$$

Estimación (IV)

Definiciones (cont.)

- Varianzas conocidas de muestreo transformadas:

$$V_m^\star, \qquad m = 1, \dots, M$$

- Con σ^2 dado, la varianza total por área es:

$$V_m = V_m^\star + \sigma^2, \qquad V = \operatorname{diag}(V_1, \dots, V_M)$$

- Matriz de pesos:

$$W = V^{-1} = \operatorname{diag}\left(\frac{1}{V_1}, \dots, \frac{1}{V_M}\right)$$

Estimación (V)

Intuición

Cada observación de área se pondera por

$$w_m = \frac{1}{V_m} = \frac{1}{V_m^\star + \sigma^2}$$

- Si $V_m^\star$ es grande (estimador directo ruidoso), w_m es pequeño $\Rightarrow$ menor influencia,
- Si $V_m^\star$ es chico (estimador directo preciso), w_m es grande $\Rightarrow$ mayor influencia,
- El término σ^2 actúa como un colchón que evita diferencias extremas en los pesos entre áreas.

Estimación (VI)

Estimador GLS

El estimador de mínimos cuadrados generalizados (GLS) para β, dado σ^2, es

$$\boxed{\hat{\beta}(\sigma^2) \;=\; (Z^\top W Z)^{-1} Z^\top W\, y^\star}$$

Varianza del estimador

La matriz de varianza-covarianza de $\hat{\beta}(\sigma^2)$ está dada por

$$\mathrm{Var}\Big(\hat{\beta}(\sigma^2)\Big) = (Z^\top W Z)^{-1}$$

Predicción EBLUP y Back-transformation

Predicción EBLUP (escala logit)

Definimos el factor de reducción (shrinkage) para asignar el peso de la estimación directa en relación a la prediccón basada en el modelo:

$$B_m = \frac{V_m^\star}{V_m^\star + \hat{\sigma}^2}$$

$\Rightarrow$ EBLUP en escala logit es un promedio ponderado entre la estimación directa y la predicción basada en el modelo de regresión:

$$\boxed{\widehat{\theta}_m^{\text{EBLUP}} = B_m \, z_m^\top \hat{\beta} + (1 - B_m) \, y_m^\star}$$

Back-transformation

En la escala original (proporción):

$$\widehat{p}_m = g^{-1}\left(\widehat{\theta}_m^{\text{EBLUP}}\right), \qquad g^{-1}(x) = \tfrac{1}{1+e^{-x}}$$

Proceso de selección de covariables

Más de 200 posibles covariables a nivel de área, agrupadas por Temática: **Demografía**, **Dinámica empresarial**, **Economía y finanzas**, **Educación y talentos**, **Ambiente**, **Infraestructura** y **Político-electoral**.

- El primer paso del proceso de selección consiste en una depuración inicial destinada a evitar problemas de multicolinealidad y de error de medición.
- En el segundo paso, se aplicó para cada **Índice** un procedimiento de selección de variables paso a paso (stepwise) basado en el **BIC**.

Estimaciones preliminares ICEPyME y PMIPyME Q125: Errores estándar

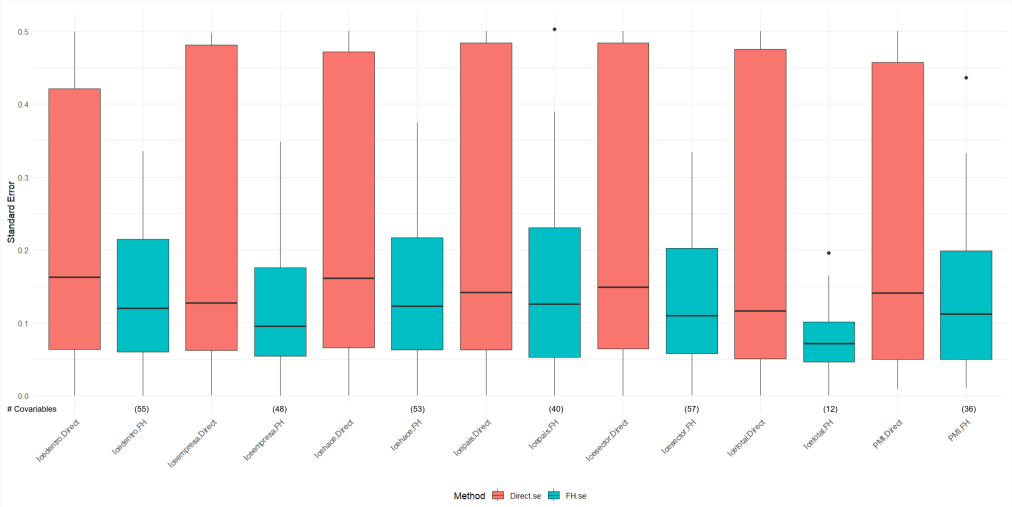


Figure 3: Errores estándar de las estimaciones directas y FH con $|Z|$ entre paréntesis.

Estimaciones preliminares ICEPyME y PMIPyME Q125: Mapas

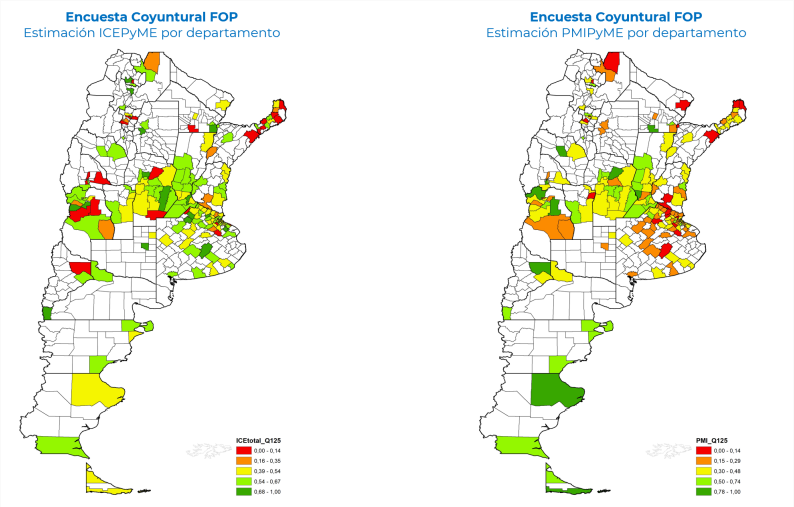


Figure 4: Estimación ICEPyME y PMIPyME Q125 a nivel micro-territorial.

Dinámica empresarial-FOP

Monitoreo permanente de la demografía y el dinamismo empresarial a partir de una metodología armonizada internacionalmente (BM, Eurostat, OCDE).

Directorio de empresas de FOP

- Actualización mensual de **6 millones de CUITs activos** (acceso a datos abiertos públicos y privados a nivel municipal, provincial y nacional);
- Descarga de datos masiva remota (sistemas SOAP y APIs REST): **500 mil constancias** de inscripción de empresas empleadoras;
- Informes trimestrales de **indicadores sintéticos por sector y micro-territorios** y **modelos de expectativas de vida** de las empresas empleadoras de Argentina (`https://comunidad.observatoriopyme.org.ar/demografia-empresarial`).

Dinámica empresarial-FOP: Indicadores básicos

Algunos indicadores básicos

- **Tasa de permanencia**: porcentaje de **registros permanentes** en t en relación al stock de registros empleadores en $t-1$, siendo los **permanentes** aquellos registros empleadores tanto en el período de referencia t como en el anterior $t-1$.
- **Nacimientos** (altas totales): registros empleadores en t, NO empleadores en $t-1$.
- **Cierres** (bajas totales): registros empleadores en $t-1$, NO empleadores en t.
- **Tasa de natilidad agregada**: Nacimientos - Cierres entre t y $t-1$ en relacición al stock de registros empleadores en $t-1$.

Dinámica empresarial-FOP: Industria manufacturera Q125

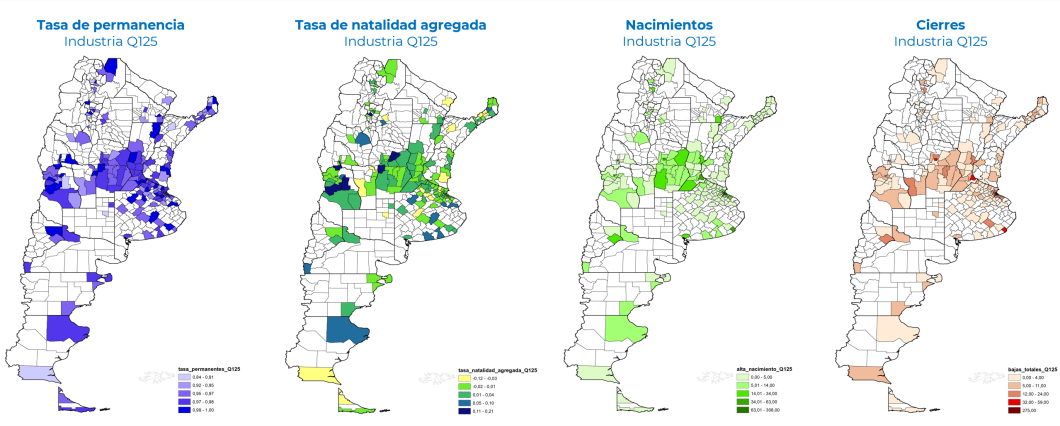
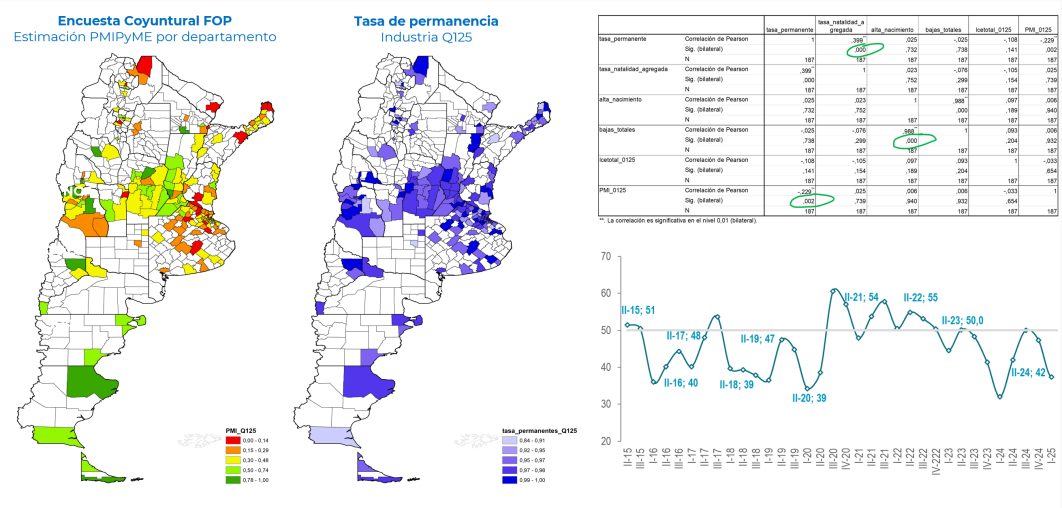


Figure 5: Indicadores FOP de Dinámica empresarial. Industria manufacturera Q125.

Estimación PMIPyME Q125: Validación “naive”



		tasa_permanente	tasa_natalidad_agregada	alta_nacimiento	bajas_totales	ioetotal_0125	PMI_0125
tasa_permanente	Correlación de Pearson	1	,399**	,025	-,025	-,108	-,229
	Sig. (bilateral)		,000	,732	,738	,141	,002
	N	187	187	187	187	187	187
tasa_natalidad_agregada	Correlación de Pearson	,399**	1	,023	-,076	-,105	,025
	Sig. (bilateral)	,000		,752	,299	,154	,736
	N	187	187	187	187	187	187
alta_nacimiento	Correlación de Pearson	,025	,023	1	,988**	,097	,006
	Sig. (bilateral)	,732	,752		,000	,189	,940
	N	187	187	187	187	187	187
bajas_totales	Correlación de Pearson	-,025	-,076	,988**	1	,093	,006
	Sig. (bilateral)	,738	,299	,000		,204	,932
	N	187	187	187	187	187	187
ioetotal_0125	Correlación de Pearson	-,108	-,105	,097	,093	1	-,033
	Sig. (bilateral)	,141	,154	,189	,204		,654
	N	187	187	187	187	187	187
PMI_0125	Correlación de Pearson	-,229**	,025	,006	,006	-,033	1
	Sig. (bilateral)	,002	,739	,940	,932	,654	
	N	187	187	187	187	187	187

**. La correlación es significativa en el nivel 0,01 (bilateral).

Figure 6: Estimación PMIPyME Q125 vs. Tasa de Permanencia Q125.

Conclusiones preliminares y próximos pasos

Conclusiones preliminares

- Validación "naive" del modelo **PMIPyME** micro-territorial con variables EXTERNAS $\rightarrow$ a MENOR nivel de actividad industrial PyME (Q125 recesivo), MAYOR tasa de permanencia de empresas industriales (núcleo duro).
- Los modelos **ICEempresa** y **PMIPyME** incluyeron covariables de TODAS las Temáticas $\rightarrow$ indicadores complejos que requieren más información.
- **ICEdentro**, **ICEhace**, **ICEpais**, **ICEsector** e **ICEtotal** solo de Dinámica empresarial, Demografía e Infraestructura $\rightarrow$ dependen más de factores esctructurales.

Próximos pasos

- Implemetación de SAE para todos los indicadores derivados de los relevamientos EC y EE (cross-section y forecasting) de FOP, incorporando inferencia causal.
- Desarrollo de modelos SAE con dependencia espacial (ICAR).

Referencias bibliográficas

De Nicolò, S., Ferrante, M. R., and Pacei, S. (2024). Small area estimation of inequality measures using mixtures of Beta. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **187** 85–109.

Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N. (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, **91** 1–33.

Maestrini, L., Bhaskaran, A., and Wand, M. P. (2024). Second term improvement to generalized linear mixed model asymptotics. *Biometrika*, **111** 1077–1084.

Rao, J. N. and Molina, I. (2015). *Small area estimation*. John Wiley & Sons, New Jersey.

Runge, M. (2023). Estimating intra-regional inequality with an application to German spatial planning regions. *Journal of Official Statistics*, **39** 203–228.

Más trabajo en curso...

Muchas gracias por la atención

...y por los comentarios!