



Small Area Estimation for Poverty at *comuna* level in Chile



Gobierno
de Chile

UNSC 53rd Session Side Event: Leaving no one behind:
Adopting a systematic approach of using small area
estimation for SDG monitoring

Social Observatory Division
Ministry of Social Development
Santiago, 1 February, 2022

National Socioeconomic Characterization Survey

- Chile's official data source for poverty statistics is the National Socioeconomic Characterization Survey (Casen).
- Casen has been sponsored by the Social Development Ministry every 2 or 3 years since 1987.
- Approximately 70,000 surveyed households and 324 selected comunas.
- Comuna poverty estimates are used for the allocation of public funds to the local administrations.
 - Municipal Common Fund: provides unconditional grants to all municipalities with the aim of fiscal equalization and is self-financing.
- Since 2011 the Ministry has implemented a small area estimation method for the poverty rates at the comuna level.
 - Area-level model: Fay-Herriot (1979)
 - Advisory of Partha Lahiri and UNDP






Ministry-ECLAC collaborative work

- The Ministry elaborated a diagnostic based on the procedures applied on the poverty estimations during 2015 and 2017. From this diagnostic, the Ministry decided to update some aspects of the SAE methodology implemented for the poverty estimation.
- In 2019, the Ministry signed a cooperation agreement with ECLAC, in which the Statistics Division would give technical advisory based on the Ministry diagnostic.
- The main improvements in the SAE procedures were:
 1. Selection of **quality indicators to evaluate the direct estimates** used in the Fay-Herriot model.
 2. Estimation of sampling variance through a **Generalized Variance Function**.
 3. Estimation of **MSE** using a parametric Bootstrap.

The document that summarises this joint effort has already been published (in Spanish). It is possible to access to the document following this [link](#).



Area level model for poverty estimates

- SAE: through modelling, information from auxiliary variables and from other domains is incorporated to “borrow strength”.
- The area level model used is based on Fay and Herriot proposal (1979), which used a two-level model.
- The linking model relates the parameter of interest (θ_d) to known auxiliary variables (\mathbf{x}_d) for each of the domains (d) that constitute the partition of the whole population (D).

$$1) \quad \theta_d = \mathbf{x}_d' \beta + u_d, \quad u_d \sim^{iid} (0, \sigma_u^2)$$

- Since the parameter of interest (θ_d) is unobservable, a direct estimator is used ($\hat{\theta}_d^{Dir}$), which carries a sampling error (e_d).

$$2) \quad \hat{\theta}_d^{Dir} = \theta_d + e_d, \quad e_d \sim^{ind} (0, \psi_d^2)$$

$$\hat{\theta}_d^{Dir} = \mathbf{x}_d' \beta + u_d + e_d, \quad u_d \sim^{iid} (0, \sigma_u^2), \quad e_d \sim^{ind} (0, \psi_d^2)$$

Area level model for poverty estimates

- The model assumes that σ_u^2 and ψ_d^2 are known parameters. However, in practice they have to be estimated based on sample data.
- The model can be expressed as a weighted average of the direct estimator ($\hat{\theta}_d^{Dir}$) and a regression synthetic estimator ($\hat{\theta}_d^{syn}$):
 - $\hat{\theta}_d^{FH} = \hat{\gamma}_d \hat{\theta}_d^{Dir} + (1 - \hat{\gamma}_d) \hat{\theta}_d^{syn}$, $d = 1, \dots, D$
 - $\hat{\theta}_d^{syn} = \mathbf{x}_d' \hat{\beta}$
 - $\hat{\gamma}_d = \widehat{\sigma}_u^2 / (\widehat{\sigma}_u^2 + \widehat{\psi}_d^2)$
- Smaller sampling variances imply more weight is placed on the direct estimator of the area d . In other words, the bigger the sample size is (small $\widehat{\psi}_d^2$) the closer $\hat{\theta}_d^{FH}$ is to the direct estimator.

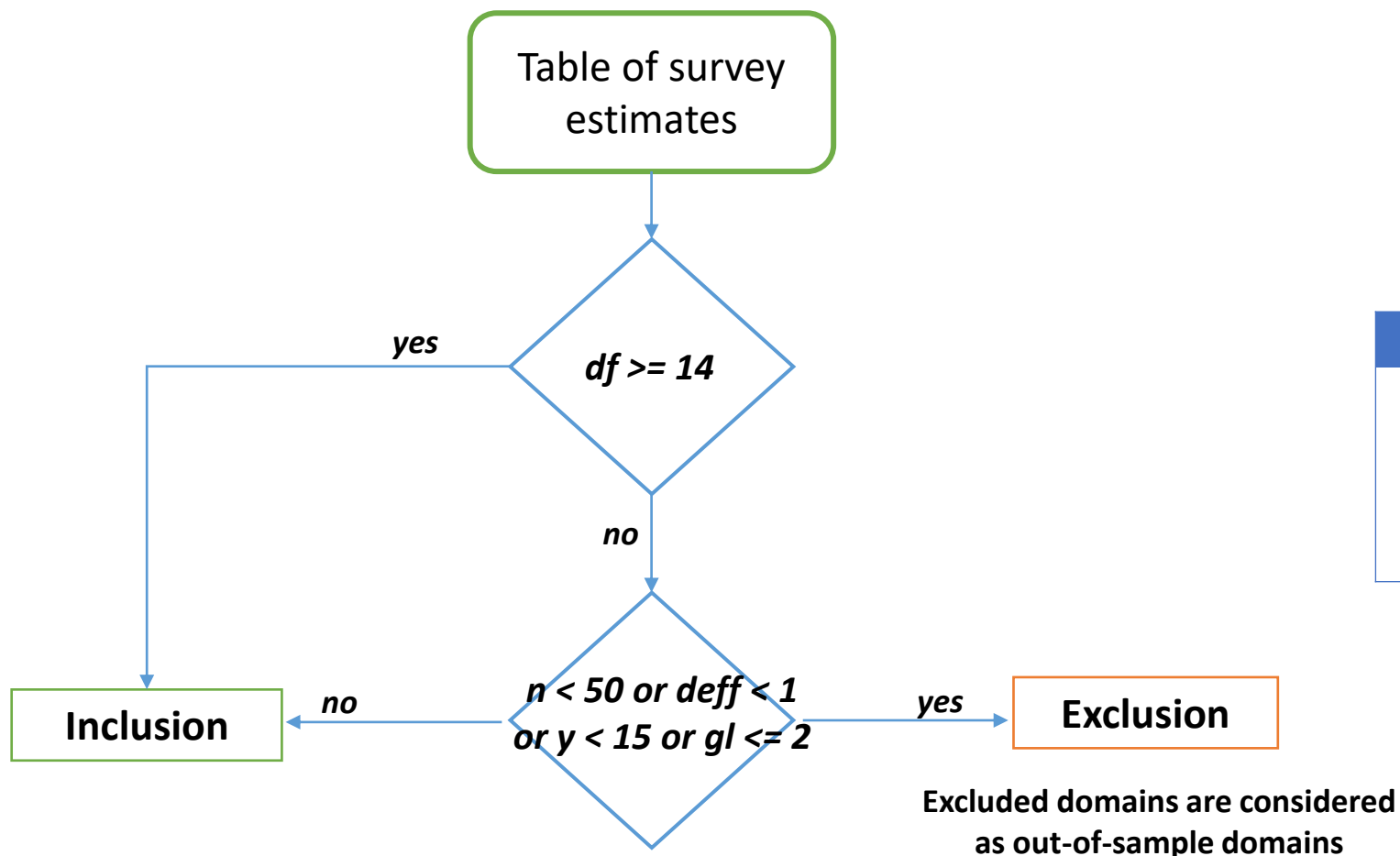
Data requirements for an area level model

- To produce reliable small area estimates it is necessary to have:
 - **Auxiliary information** correlated with the variable of interest, in the form of further survey data, administrative registers or census data.
 - Auxiliary variables used in the poverty estimation come from:
 - 2017 Census.
 - Socio-economic indicators based on administrative records.
 - All these auxiliary variables are aggregated at the comuna level.
 - **Reliable direct estimates.** It is assumed that the sampling variance estimates of the target domains are unbiased and consistent.

Quality indicators for direct estimates

- 4 quality indicators were selected:
 - Sample size (n)
 - Degrees of freedom (df)
 - Number of people that have the characteristic of interest (y)
 - Design effect ($deff$)
- Indicators used at the NSOs for deciding when to suppress inaccurate survey estimates were used as a starting point (Gutierrez, et. al, 2020). The thresholds were adapted to be used for SAE purposes.
- These indicators are a measure of:
 - How well the collection process followed the planned sample design.
 - How well the direct estimates measure the poverty situation at the target domain (in terms of their point estimates and their variances).

Quality indicators for direct estimates



Number of comunas according to the quality classification criteria

Survey's year	Included	Excluded
2011	232	92
2013	279	45
2015	234	90
2017	242	82
2020	256	68

Excluded domains are considered as out-of-sample domains

Quality indicators for direct estimates

- Example of 2 comunas: Camiña and Vitacura. In 2017, both comunas had a 0% poverty rate according to the survey estimate. The estimated sampling variance for both years was also 0%.

Income Poverty rate direct estimates

Survey year	Camiña	Vitacura
2017	0.0%	0.0%

Camiña - Quality indicators (2017)

n	df	y
82	1	0

Vitacura - Quality indicators (2017)

n	df	y
867	87	0

Direct variance estimation

- Target variable: **proportion** of individuals under the poverty line.
- When working with proportion estimates it is necessary to resolve two main issues to apply the Fay-Herriot model:

1. **High variability** of the sampling variance estimates: sampling variances are often noisy due to the small sample size.

2. **Endogeneity problem**: sampling variances are related to their corresponding small area proportions.

1. Generalized Variance Function (Hidiroglou, 2019; Wolter, 2007)

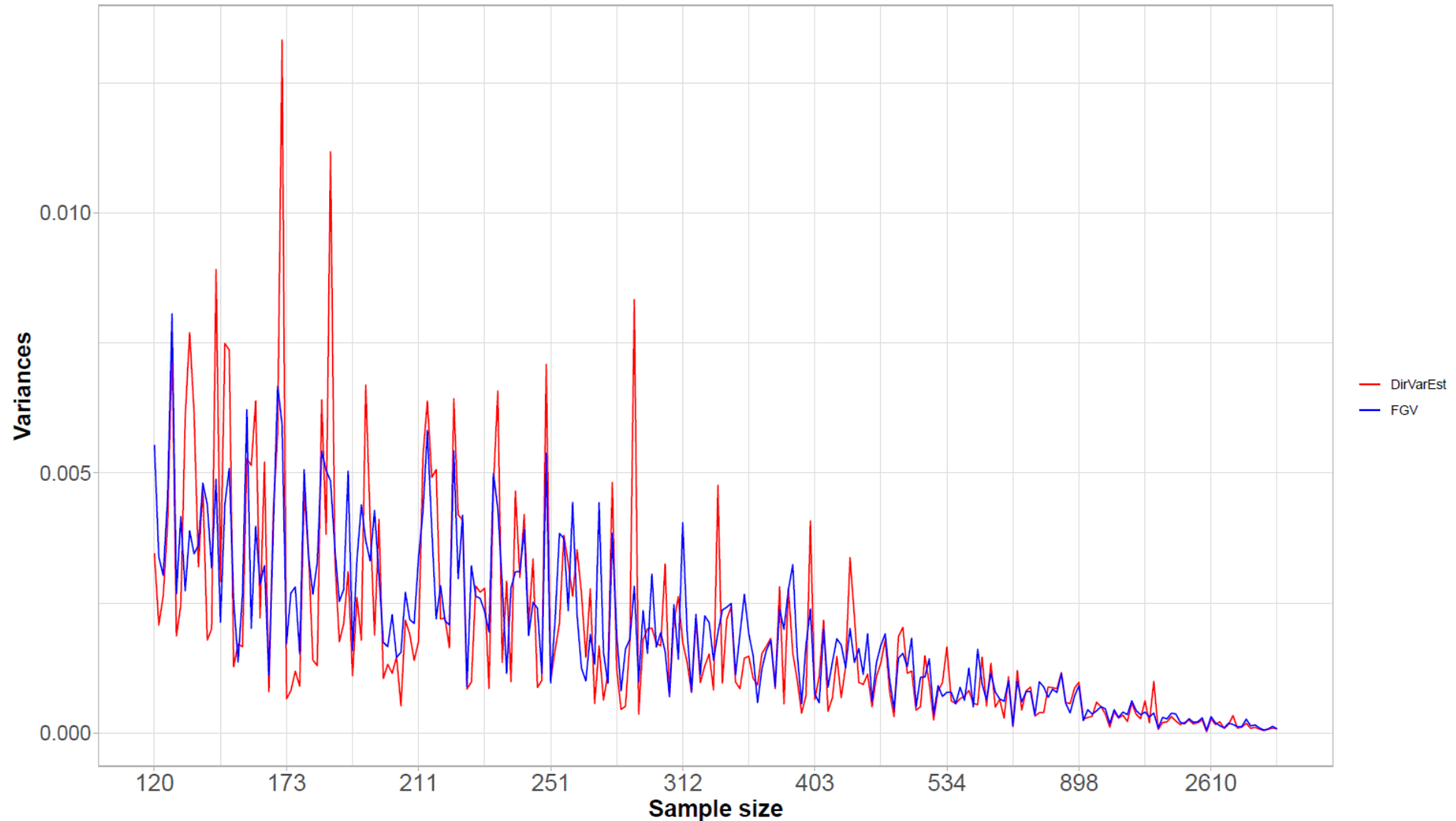
GVFs tend to smooth out the noisiness that may be present in the underlying data.

2. Arcsin transformation (Jiang, Lahiri, Wan & Wu, 2001)

$$\hat{z}_d = \arcsin\left(\sqrt{\hat{\theta}_d^{Dir}}\right)$$

$$Var_p(\hat{z}_d) = \frac{\widehat{DEFF}_d}{4 \times n_d}$$

Smoothed variance versus direct variance



Source: MDFS, Casen en Pandemia 2020

MSE estimation

- The mean squared error (MSE) is the most common measure to assess the uncertainty associated with the area-specific prediction under the model that has been assumed (Tzavidis et al., 2018).
- Bootstrap methods are very promising to estimate a measurement of uncertainty when using a FH model with an arcsine transformation.
 - Parametric bootstrap implemented in Casen 2020: allowed to estimate MSE and the estimation of confidence intervals for all comunas. The procedure considered a bias corrected back-transformed Fay-Herriot estimate. The procedure was developed by Hadam, Würz & Kreutzmann (2020).
- To produce estimations for MSE it was used the R package **emdi** (Kreutzmann et al., 2018).

$$MSE_B(\hat{\theta}_d^{FH,trans}) = \frac{1}{B} \sum_{b=1}^B \left(\hat{\theta}_{d,(b)}^{FH,trans} - \theta_{d,(b)}^{trans} \right)^2$$

MSE estimation

- Using the estimated MSE it is possible to obtain a relative measure, similar to the coefficient of variance, with the aim to evaluate the quality of the Fay-Herriot estimates. This is also known as relative root mean square error (RRMSE).

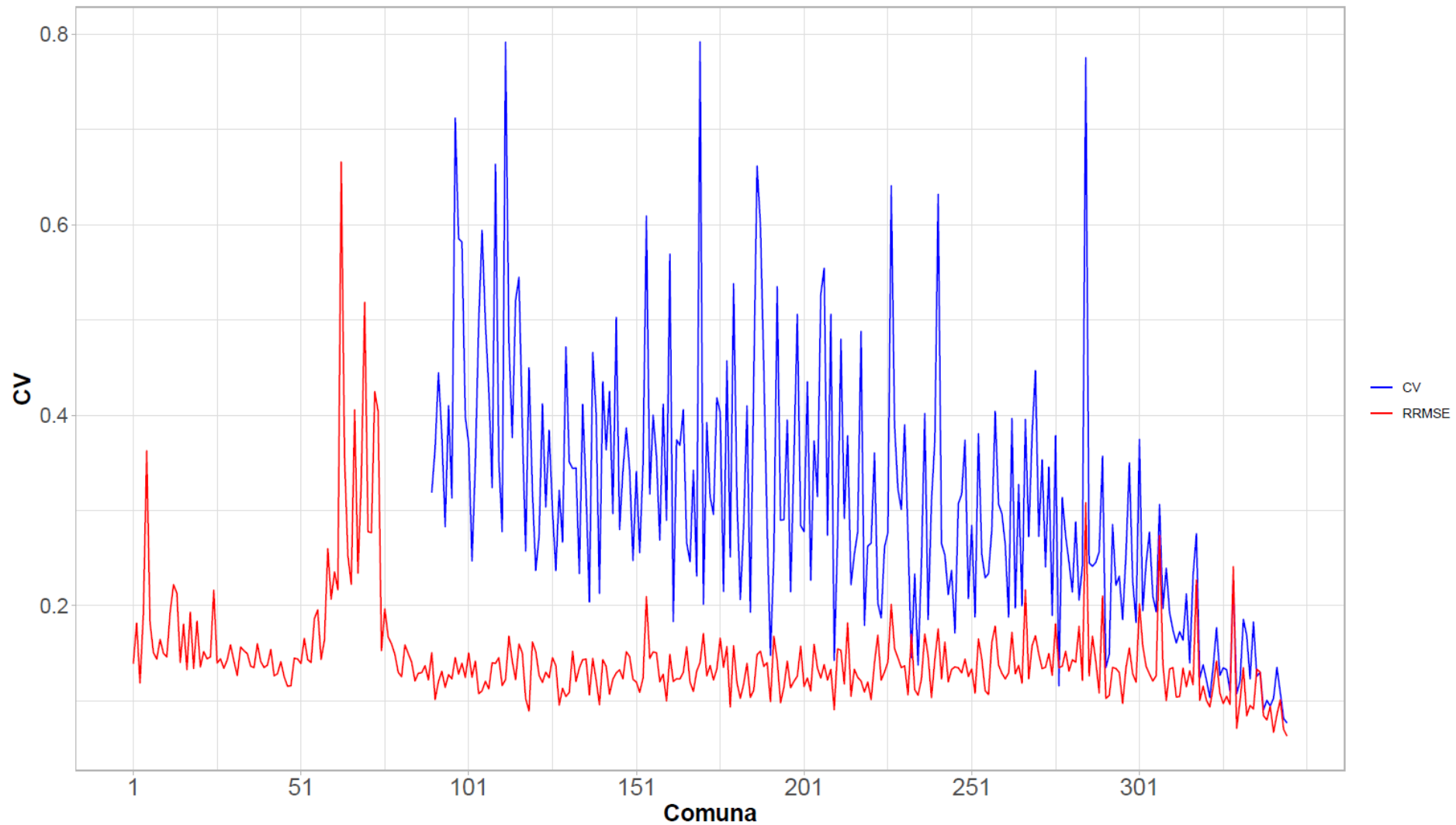
$$\widehat{CV} = \frac{\sqrt{MSE_B(\hat{\theta}_d^{FH,trans})}}{\hat{\theta}_d^{FH,trans}} * 100$$

- Given the theoretical distribution of the Fay-Herriot estimator is normal and considering an unbiased back-transformation, it is possible to estimate the confidence intervals using:

$$CI(\hat{\theta}_d^{FH,trans}) = \left(\hat{\theta}_d^{FH,trans} \pm 1.96 RMSE(\hat{\theta}_d^{FH,trans}) \right)$$

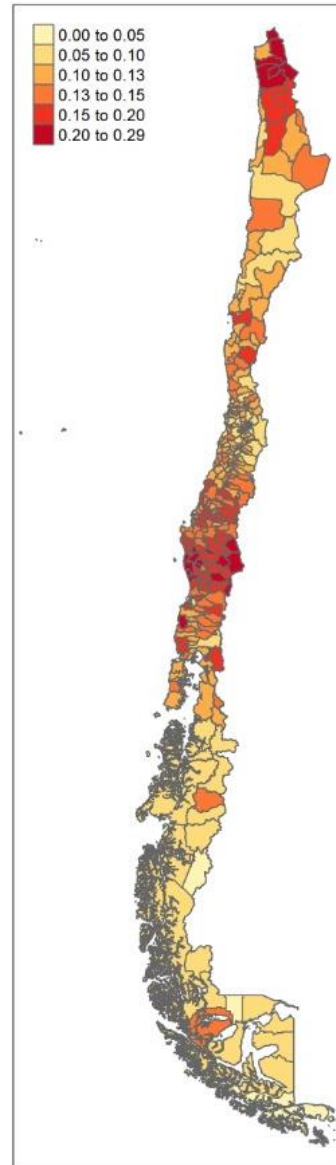
- Where RMSE: $\sqrt{MSE_B(\hat{\theta}_d^{FH,trans})}$

Direct CV versus Relative RMSE

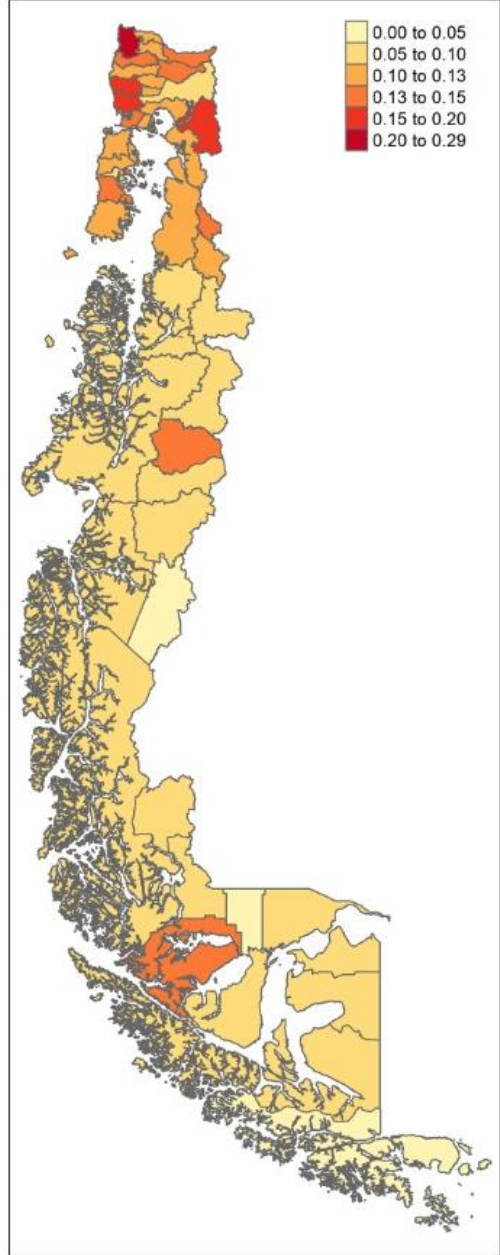
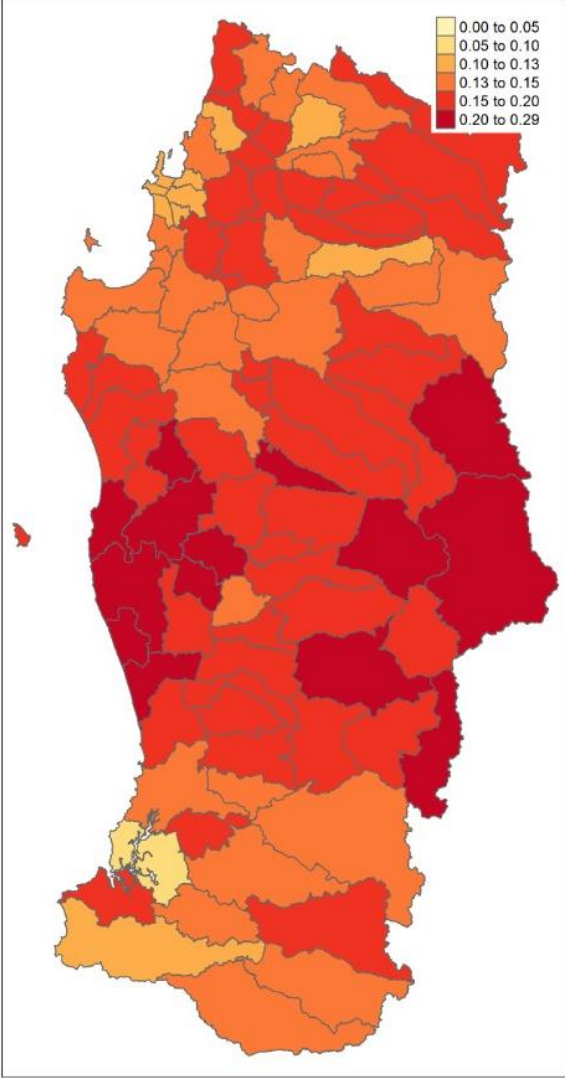
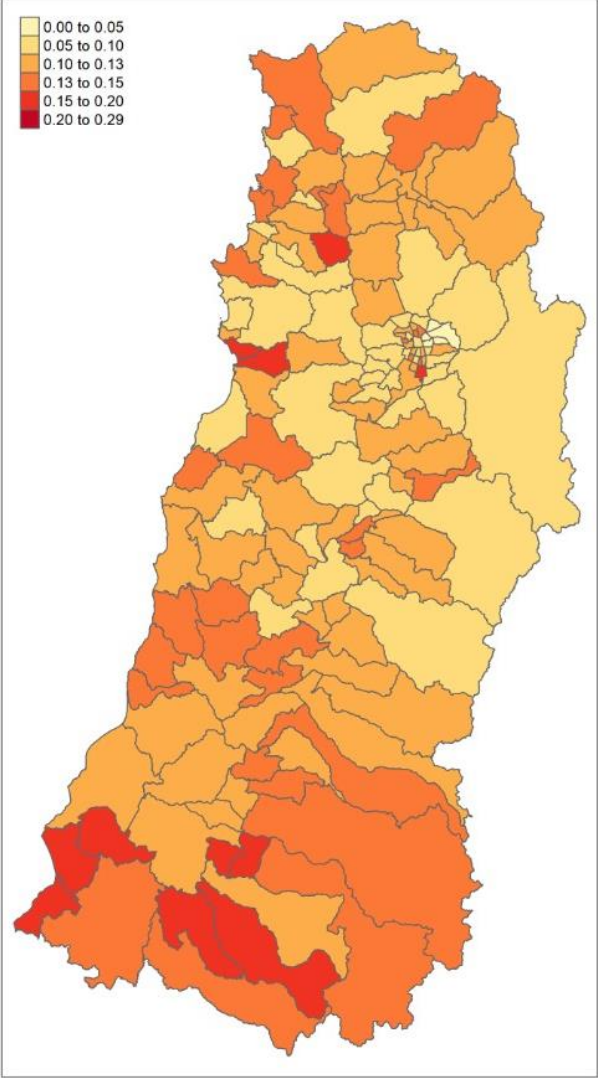
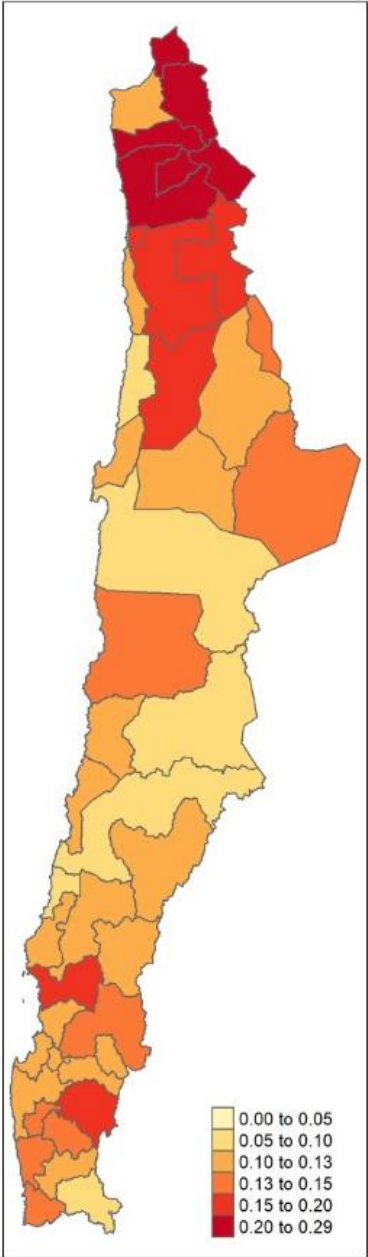


Source: MDFs, Casen en Pandemia 2020

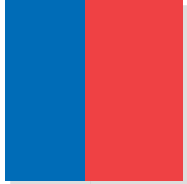
Poverty mapping in Chile



Poverty mapping in Chile



Note: the different 4 maps are not at the same scale



References

- Casas-Cordero, C., Encina, J. & Lahiri, P. (2016), Poverty Mapping for the Chilean Comunas. In Analysis of Poverty Data by Small Area Estimation (ed. M. Pratesi), 379–403. Hoboken: John Wiley & Sons.
- EUROSTAT (2019). Guidelines on small area estimation for city statistics and other functional geographies. 2019 edition
- Fay, R.E., & Herriot, R.A. (1979), Estimates of income for small places: An application of James-Stein procedure to census data, Journal of the American Statistical Association 74, 269-277.
- Gutierrez, A., Fuentes, A., Mancero, X., Molina, F., & Lopez, F (2020). Criterios de calidad en la estimación de indicadores a partir de encuestas de hogares: una aplicación a la migración internacional. Serie Estudios Estadísticos, N° 101 (LC/TS.2020/52), Santiago, CEPAL
- Hadam, S; Würz, N. & Kreutzmann, A (2020). Estimating regional unemployment with mobile network data for Functional Urban Areas in Germany.
- Hidiroglou, Michel A. (2019). “Development of a Small Area Estimation System at Statistics Canada.” Survey Methodology 45 (1): 101–26.
- Jiang, Jiming, P Lahiri, Shu-Mei Wan, & Chien-Hua Wu. (2001). “Jackknifing in the Fay-Herriot Model with an Example,” 36
- Kreutzmann, A., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M. & Tzavidis, N. (2018), The R package emdi for the estimation and mapping of regional disaggregated indicators. Journal of Statistical Software. 91. 10.18637/jss.v091.i07.
- Rao, J. N. K. & Molina, I. (2015): Small area estimation. John Wiley & Sons
- Wolter, K M. 2007. Introduction to Variance Estimation. 2nd ed. Statistics for Social and Behavioral Sciences. Springer.