

SEMINARIO SOBRE DESAGREGACIÓN DE DATOS MEDIANTE TÉCNICAS DE ESTIMACIÓN EN ÁREAS PEQUEÑAS EN URUGUAY

DENISE BRITZ DO NASCIMENTO SILVA

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS ENCE - IBGE

- Holt, D.T. (2007). The official statistics Olympics challenge: Wider, deeper, quicker, better, cheaper. *The American Statistician*, 61(1, February), 1-8.
- Estadísticas oficiales: herramienta para el desarrollo
- Creciente demanda de información:
 - Actualizada en el tiempo
 - Temáticamente más completa
 - **Espacialmente detallada**
 - **En grupos específicos (ancianos, jóvenes)**



Home

Welcome to the Sustainable Development Goal indicators website

*A robust follow-up and review mechanism for the implementation of the new 2030 Agenda for Sustainable Development will require a solid framework of indicators and **statistical data** to monitor progress, inform policy and ensure accountability of all stakeholders.*

<http://unstats.un.org/sdgs/>

RECOGIDA DE DATOS E PRODUCCIÓN DE INFORMACIÓN

FUENTES DE DADOS

Encuestas por muestreo

Censos

Registros administrativos

Big data - datos orgánicos

ALGUNOS MÉTODOS ESTADÍSTICOS

Muestreo

Métodos de estimación en
áreas pequeñas

Combinación de muestras
probabilísticas y no
probabilísticas

Conexión de registros

RECOGIDA DE DATOS E PRODUCCIÓN DE INFORMACIÓN

FUENTES DE DADOS

Encuestas por muestreo

Censos

Registros administrativos

Big data - datos orgánicos

ALGUNOS MÉTODOS ESTADÍSTICOS

Muestreo

Estimación en áreas pequeñas

Combinación de muestras
probabilísticas y no
probabilísticas

Conexión de registros

TÉCNICAS DE ESTIMACIÓN EN ÁREAS PEQUEÑAS

- Muchos años de desarrollo (modelo Fay-Herriot de 1979)
- Técnicas conocidas y reconocidas en muchos países
- Foco de varios proyectos de colaboración entre el mundo académico, los institutos nacionales de estadística y las organizaciones internacionales
- Esfuerzos para popularizar el uso de las técnicas: guías, cursos, conferencias (desde 1992) , biblioteca de funciones R, UNSD Wiki SAE4SDG, estadísticas experimentales, etc.



SAE4SDG

SPACE SHORTCUTS

SAE4SDG

PAGE TREE

- Why is SAE important for SDG data
- › Producing SAE
- Communicating SAE methods and results
- › From SAE experiment to production
- › SAE practices
- Software packages
- SAE key readings
- Training materials
- SAE projects
- FAQ

Pages

SAE4SDG

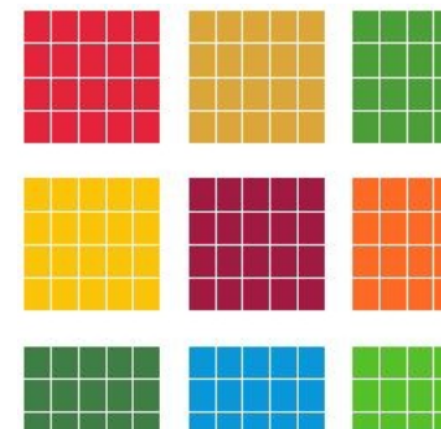
Created by UNSD Clarence Lio, last modified by Haoyi Chen on Apr 08, 2022



Welcome to the Toolkit on Using Small Area Estimation for SDGs!

In committing to the realization of the 2030 Agenda for Sustainable Development, Member States recognized that the dignity of the individuals is fundamental and that the Agenda's Goals and targets should be met for all nations and people and for all segments of society. Ensuring that these commitments are translated into effective action requires a precise understanding of the target populations and progress made in addressing their particular priorities.

To properly measure this, statistics need to be presented for different population groups and geographical areas. The Sustainable Development Goal (SDG) indicator framework has included an overarching principle of data disaggregation: SDG indicators should be disaggregated where relevant by income, sex, age, race, ethnicity, migratory status,



<https://unstats.un.org/wiki/display/SAE4SDG/>



- Why is SAE important for SDG data disaggregation
 - › Producing SAE
- Communicating SAE methods and results
 - › From SAE experiment to production: challenges and the enabling environment
- ✓ SAE practices
 - SAE by SDGs
 - › SAE for countries and partners
- SAE key readings
- Training materials
- SAE projects
- FAQ
- References
- Acknowledgement



<https://repositorio.cepal.org/handle/11362/44214>



Better Data for More Effective Policies and Programmes

SMALL AREA ESTIMATION

https://www.unfpa.org/sites/default/files/pub-pdf/19-310_SAE_Brochure_A4-SINGLE-PROOF6.pdf

SAE Conference 2023
7 al 9 de Junio
PUC Perú – INEI – UN CEPAL

Guidelines on small area estimation for city statistics and other functional geographies

2019 edition



<https://ec.europa.eu/eurostat/documents/3859598/10167610/KS-GQ-19-011-EN-N.pdf/3b56be5d-8266-0ee7-7579-7c3e8e63289d>

2022 International Methodology Symposium



Statistics
Canada

Data Disaggregation: Building a more-representative data portrait of society

Statistics Canada's 2022 International Methodology Symposium "Data Disaggregation: Building a more-representative data portrait of society" will take place virtually from November 2 to November 4, 2022, inclusively.



1. Small area estimation in the context of equity-deserving groups
2. Integrating multiple data sources (including those from probability and non-probability sources)
3. Innovation in methods for collection and analysis of data for hard-to-reach populations (questionnaire design, non-traditional sampling techniques, issues for rare populations, ...)
4. Adaptations of methods for traditional survey steps (such as sampling, imputation, estimation, coding, classification, measurement errors, record linkage, disclosure control, ...)
5. Methods for Data equity: Issues of ethics and privacy in the application of disaggregated data methods in official statistics
6. Methods for Data equity: Advances in analysis of populations that have been made vulnerable (such as Fair AI in data science, missing data analysis, ...)

ENCUESTAS POR MUESTREO ¿ÁREA O DOMINIO PEQUEÑOS?

- Problema con el tamaño de la muestra en el área o dominio de interés
- Tamaño de la muestra no es lo suficientemente grande como para producir estimaciones directas con la precisión deseada

Solución

- "Toma prestada" información auxiliar de otros conjuntos de datos:
 - áreas o dominios similares (información transversal)
 - la misma área en otras ocasiones (modelos de series de tiempo)

TIPOS DE MODELOS

Modelos transversales

- relaciona estimaciones de la muestra con la información auxiliar en una ocasión

Modelos de series temporales

- adecuado para la estimación en dominios pequeños en encuestas por muestreo repetidas en el tiempo (como el la Encuesta Continua de Hogares)
- incorpora las observaciones de la serie como información relevante en el proceso de estimación

Estimación de los indicadores de las TIC en los estados brasileños (Brazilian Network Information Center - NIC.br)

- Estimador directo: promedio de estimaciones de años consecutivos y la agregación de muestras de años consecutivos
- Estimador sintético de un solo año
- Estimador compuesto: combinación de muestras de dos años consecutivos
- Enfoques más sencillos debido a la amplia gama de indicadores que deben producirse rápidamente tras la recogida de datos

Bertolini Coelho, I., Trindade Pitta, M. and do Nascimento Silva, P. S. (2020). Estimating state level indicators from ICT household surveys in Brazil, *Statistical Journal of the IAOS* 36, 495–508.

INICIATIVAS RECIENTES EN BRASIL

Proyecto GradePop: IBGE - UNFPA – Worldpop

Elaboración de estimaciones de población para secciones censales, utilizando modelos de aprendizaje automático y modelos jerárquicos bayesianos con integración de:

- Datos geoespaciales (imágenes de satélite)
- Datos de encuestas
- Registros administrativos

Coordinadores (IBGE): Maria Bueno, Luís Felipe Barros y Renata Curi

Estimación de pequeños dominios en la encuesta anual de servicios del IBGE

- Estimación basada en modelos de los ingresos brutos totales de los servicios por estado y por clasificación de la actividad económica que no se publican actualmente debido al plan de muestreo de la encuesta
- Pequeños dominios: clasificación económica de 4 dígitos para los estados de la región Nordeste
- Variables auxiliares del Directorio de Empresas (personas ocupadas, número de establecimientos, etc.)
- Datos de la encuesta y del registro de 2007 a 2016

Estimación de pequeños dominios en la encuesta anual de servicios del IBGE

- Modelos lineales a nivel de área (Fay-Herriot)

Neves, A. F. A. ; Silva, D. B. N. ; Corrêa, S. T., Small domain estimation for the Brazilian Service Sector Survey. *Estadística*, 65, 185, pp. 13–37, Instituto Interamericano de Estadística, 2013.

- Distribución normal asimétrica:

- modelos con efectos aleatorios de dominio y tiempo

Moura, F. A. S.; Neves, A. F. A.; Silva, D. B. N. Small area models for skewed Brazilian business survey data. *Journal of Royal Statistical Society*, 180, Part 4, pp.1039-1055, serie A, 2017.

Neves, A. F. A. ; Silva, D.B.N. ; Moura, F.A.S. Skew normal small area time models for the Brazilian annual service sector survey. *Statistics in Transition new series*, vol. 21, pp. 84-102 , 2020.

Estimación de pequeños dominios en la encuesta anual de servicios del IBGE

- Distribución normal asimétrica:
 - modelos con efectos aleatorios de dominio y tiempo
 - modelos espacio temporales (efectos estructurados sectorialmente - actividades económicas similares declaradas como áreas contiguas/vecinas)

Neves, A. F. A. Modelos assimétricos para estimação em pequenos domínios na Pesquisa Anual de Serviços. Tesis doctoral, ENCE , IBGE, 2021.

INICIATIVAS RECIENTES EN BRASIL

MODELOS DE SERIES TEMPORALES PARA ENCUESTAS REPETIDAS

Elaboración de estimaciones del desempleo basadas en modelos de espacio de estados (state-space models)

Encuesta: Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC)

Caio Gonçalves (ENCE/IBGE), Luna Hidalgo (COREN/IBGE), Denise Britz (ENCE/IBGE)
Jan van den Brakel (Statistics Netherlands e Maastrich University)

- Mayor encuesta de hogares por muestreo realizada por el IBGE
- Fuente oficial de estadísticas de desempleo en el país
- Muestreo estratificado e conglomerado en dos etapas de selección: secciones censales e hogares
- Panel rotativo con periodicidad trimestral – un hogar permanece en la PNADC durante cinco trimestres
- Esquema de rotación con solapamiento parcial de hogares:
1-2(5) – cada hogar se entrevista una vez al trimestre
- Estimaciones mensuales se obtienen a partir de medias móviles (trimestres móviles)

Satisfacer algunas demandas históricas y desafíos recientes:

- Estimación de la tendencia y la estacionalidad de las series
- Producción de indicadores basados únicamente en los casos encuestados en el mes de referencia
- Incorporación de fuentes de datos auxiliares y alternativas, como el big data
- Nowcasting
- Estimación del efecto de la pandemia SARS-COV-2 (mayor volatilidad)
- Estimación en áreas pequeñas

Satisfacer algunas demandas históricas y desafíos recientes:

- ✓ Estimación de la tendencia y la estacionalidad de las series
- ✓ Producción de indicadores basados únicamente en los casos encuestados en el mes de referencia (JRSS –A)
- ✓ Incorporación de fuentes de datos auxiliares y alternativas, como el big data (IASS Webinar - 25/01/2023) - Nowcasting
- ✓ Estimación del efecto de la pandemia SARS-COV-2 (mayor volatilidad)
- Estimación en áreas pequeñas (ya se ha iniciado)



Discussion paper

Monthly Labour Force Figures during the 2021 Redesign of the Dutch Labour Force Survey

3. Time series model for official monthly labour force figures

Jan van den Brakel

Januari 2022



Research Paper

1351.1.55.160

Office for National Statistics

English (EN) | Cymraeg (CY)

Release calendar | Methodology | Media | About | Blog

Home | Business, industry and trade | Economy | **Employment and labour market** | People, population and community | Taking part in a survey?

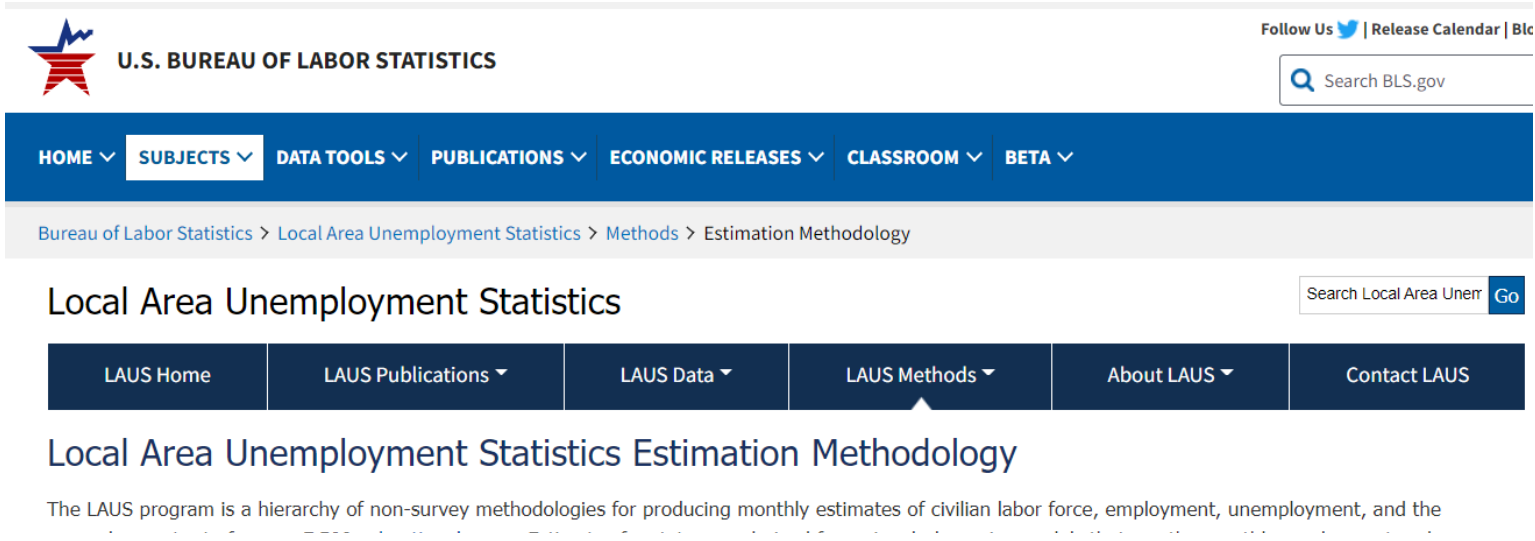
Search for a keyword(s) or time series ID

Home > Employment and labour market > People in work > Employment and employee types >

Experimental model-based single-month estimates for the Labour Force Survey: methods explained

Experimental model-based single-month estimates for the Labour Force Survey: methods explained

Exploration of state space modelling approaches for statistical impact measurement in ABS time series: The Labour Force Survey as a case study



The screenshot shows the U.S. Bureau of Labor Statistics website. At the top left is the BLS logo and the text "U.S. BUREAU OF LABOR STATISTICS". To the right are social media links and a search bar for "Search BLS.gov". Below this is a navigation menu with items: HOME, SUBJECTS, DATA TOOLS, PUBLICATIONS, ECONOMIC RELEASES, CLASSROOM, and BETA. A breadcrumb trail reads: Bureau of Labor Statistics > Local Area Unemployment Statistics > Methods > Estimation Methodology. The main heading is "Local Area Unemployment Statistics" with a search bar for "Search Local Area Unemp". Below this is a secondary navigation menu: LAUS Home, LAUS Publications, LAUS Data, LAUS Methods (highlighted), About LAUS, and Contact LAUS. The main content area is titled "Local Area Unemployment Statistics Estimation Methodology" and begins with the text: "The LAUS program is a hierarchy of non-survey methodologies for producing monthly estimates of civilian labor force, employment, unemployment, and the".

Local Area Unemployment Statistics: *Estimation*

The LAUS program uses a hierarchy of nonsurvey methodologies to produce monthly estimates of the civilian labor force, employed persons, unemployed persons, and unemployment rates for approximately 7,500 subnational areas. Estimates for states

Estimates for states

Monthly labor force data for all states and the District of Columbia are based on the time-series approach to sample survey data.¹ This

PRODUCCIÓN DE INDICADORES BASADOS ÚNICAMENTE EN LOS CASOS ENCUESTADOS EN EL MES DE REFERENCIA

Estimaciones más desagregadas, pero a lo largo del tiempo



ORIGINAL ARTICLE

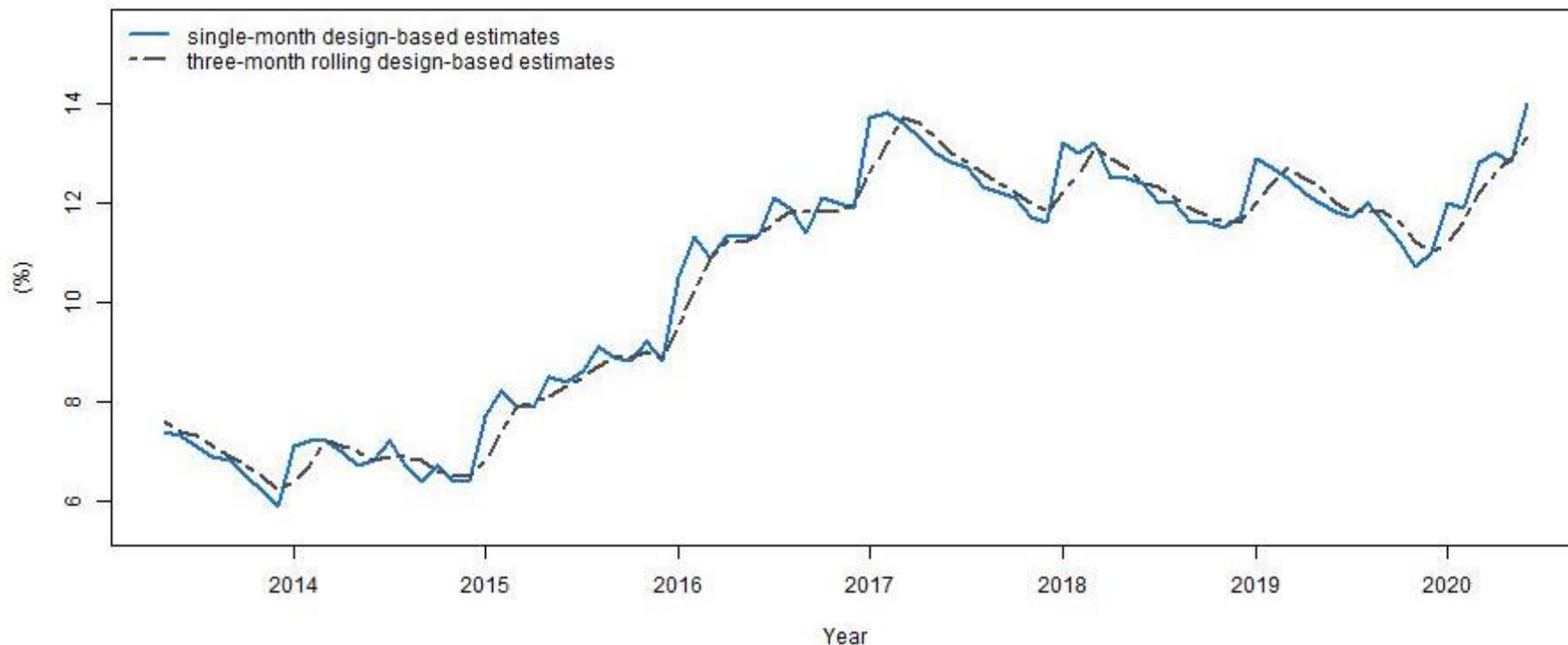
Single-month unemployment rate estimates for the Brazilian Labour Force Survey using state-space models

Caio Gonçalves ✉, Luna Hidalgo, Denise Silva, Jan van den Brakel

First published: 20 November 2022

<https://doi.org/10.1111/rssa.12914>

TASA DE DESEMPLEO NACIONAL BASADA EN EL DISEÑO ESTIMACIÓN DIRECTA



FOLHA DE S.PAULO

23.set.2020 às 9h30
Atualizado: 23.set.2020 às 12h45

Diego Garcia

Desemprego aumentou 27,6% em quatro meses de pandemia, diz IBGE

População desocupada foi de 10,1 milhões em maio a 12,9 milhões em agosto

REUTERS®

June 22, 2021
6:07 AM -03
Last Updated 14 days ago

Eduardo Simões

Brazil passes half a million COVID-19 deaths, experts warn of worse ahead



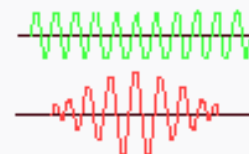
MODELOS DE SERIES TEMPORALES PARA ENCUESTAS REPETIDAS

Evolución de la señal en el tiempo (componente no observable)

Relación entre los errores de muestreo (ruido)

$$y_t = \theta_t + e_t$$

$$\theta_t = T_t + S_t + I_t$$



El modelo de series temporales para la estimación de la muestra es una combinación de dos modelos diferentes:

Modelo que describe la evolución del valor real del indicador $\{\theta_t\}$

Modelo que representa la autocorrelación de los errores de muestreo $\{e_t\}$

Es necesario identificar el modelo adecuado para el proceso $\{e_t\}$ a partir de la estimación de la estructura de correlación de los errores de muestreo

Modelos de Series Temporales para PNADC

State-space model for the Brazilian unemployment rate



\hat{y}_t : design-based estimate for unemployment rate at month t

Accounting for sampling error:

$$\hat{y}_t = \theta_t + e_t$$

Scott and Smith (1974),
Scott et al. (1977)

Unobserved components of unknown population quantity θ_t

$$\theta_t = T_t + S_t + I_t$$

Trend:

$$T_t = T_{t-1} + R_{t-1}$$

$$R_t = R_{t-1} + \eta_{R,t}$$

Durbin and Koopman (2012)

$$\eta_{R,t} \sim N(0, k_t \sigma_R^2)$$

k_t fator increases the variance to allow more flexibility in trend

van den Brakel et al. (2021)

Proposed model for PNADC



$$\begin{aligned}
 \{y_t\} & \left\{ \begin{aligned} y_t &= \theta_t + e_t \\ \theta_t &= T_t + S_t + I_t, & I_t &\sim N(0, \sigma_I^2) \end{aligned} \right. \\
 \{\theta_t\} & \left\{ \begin{aligned} T_t &= T_{t-1} + R_{t-1}, \\ R_t &= R_{t-1} + \eta_{R,t}, & \eta_{R,t} &\sim N(0, k_t \sigma_R^2) \\ S_t &= -\sum_{j=1}^{11} S_{t-j} + \eta_{S,t}, & \eta_{S,t} &\sim N(0, \sigma_S^2) \end{aligned} \right. \\
 \{e_t\} & \left\{ \begin{aligned} e_t &= c_t \tilde{e}_t \\ \tilde{e}_t &= \phi_3 \tilde{e}_{t-3} + \eta_{e,t}, & \eta_e &\sim N(0, \sigma_e^2 \cong 1) \end{aligned} \right.
 \end{aligned}$$

$$y_t = \boxed{T_t + S_t + I_t} + \boxed{e_t}$$

Structural components
noise

seasonally adjusted series

signal: $\theta_t^* = T_t + S_t$

trend: T_t

seasonally adjusted series: $\theta_t^s = T_t + I_t$

Time-series models for BLFS

State-space regional multivariate model for unemployment rate

$\hat{y}_{j,t}$: design-based estimate for unemployment rate at month t in the state j

$$\begin{pmatrix} \hat{y}_{1,t} \\ \vdots \\ \hat{y}_{J,t} \end{pmatrix} = \begin{pmatrix} \theta_{1,t} \\ \vdots \\ \theta_{J,t} \end{pmatrix} + \begin{pmatrix} e_{1,t} \\ \vdots \\ e_{J,t} \end{pmatrix} \quad j = 1, \dots, J$$

Borrowing strength from the time and space:

$$\text{COV}(\eta_{R,y_j,t}, \eta_{R,y_{j'},t}) = \rho_{y_j,y_{j'}}^R \sigma_{R,y_j,t} \sigma_{R,y_{j'},t} \quad j \neq j'$$

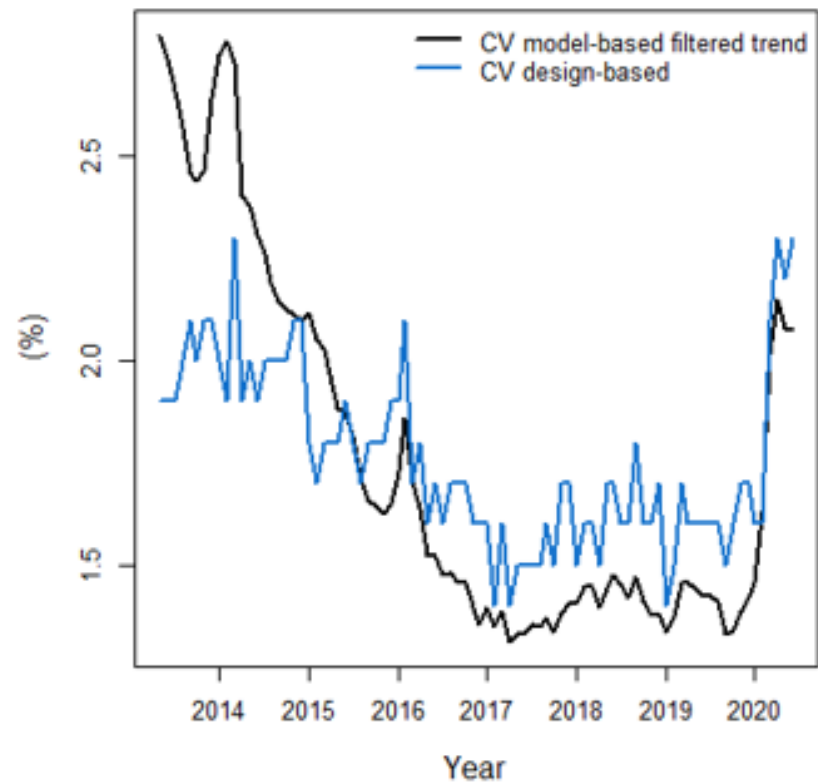
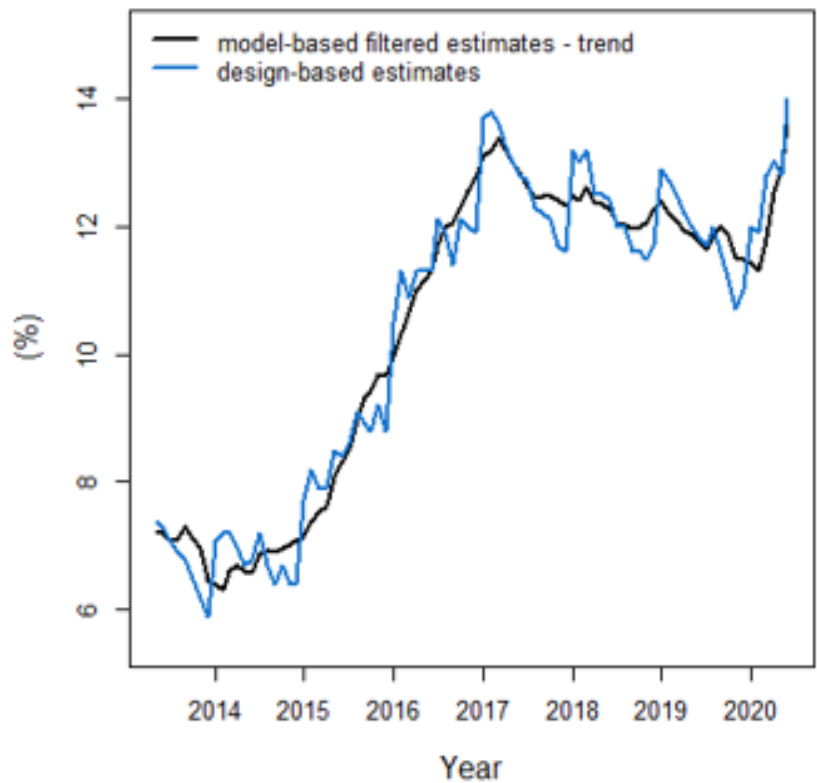
$\rho_{y_j,y_{j'}}^R$ is the correlation between the slope disturbance terms of the unemployment rate of states j and j'

Multivariate state-space model incorporating auxiliary information

$$\begin{pmatrix} \hat{y}_t \\ x_t \end{pmatrix} = \begin{pmatrix} \theta_{y,t} \\ \theta_{x,t} \end{pmatrix} + \begin{pmatrix} e_t \\ 0 \end{pmatrix}$$

Results

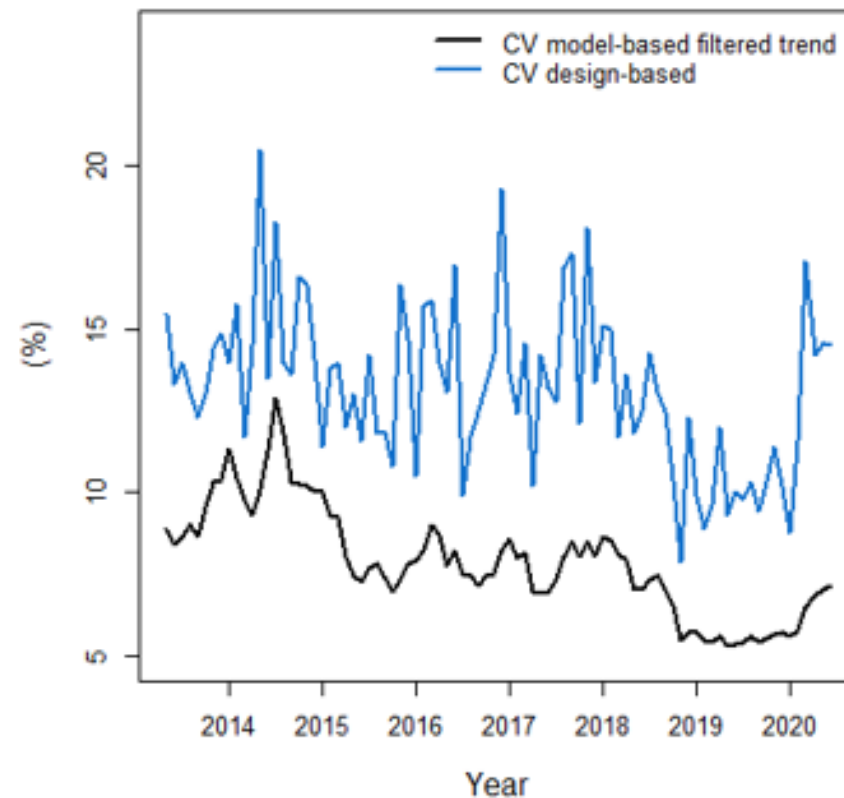
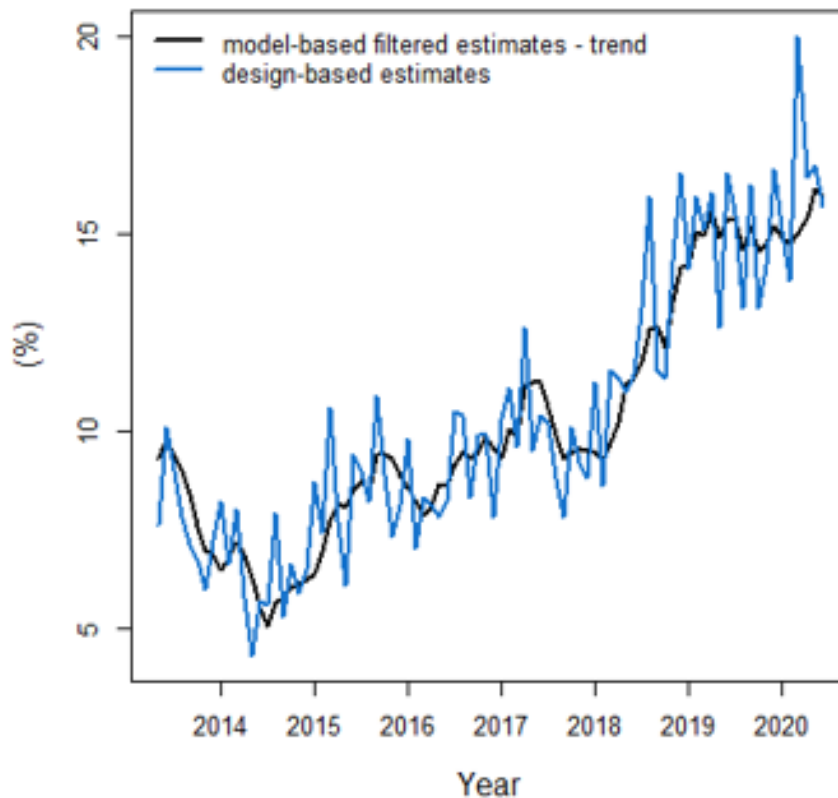
Unemployment rate design-based and model-based (trend) estimates, and respective coefficients of variation – Brazil



Results

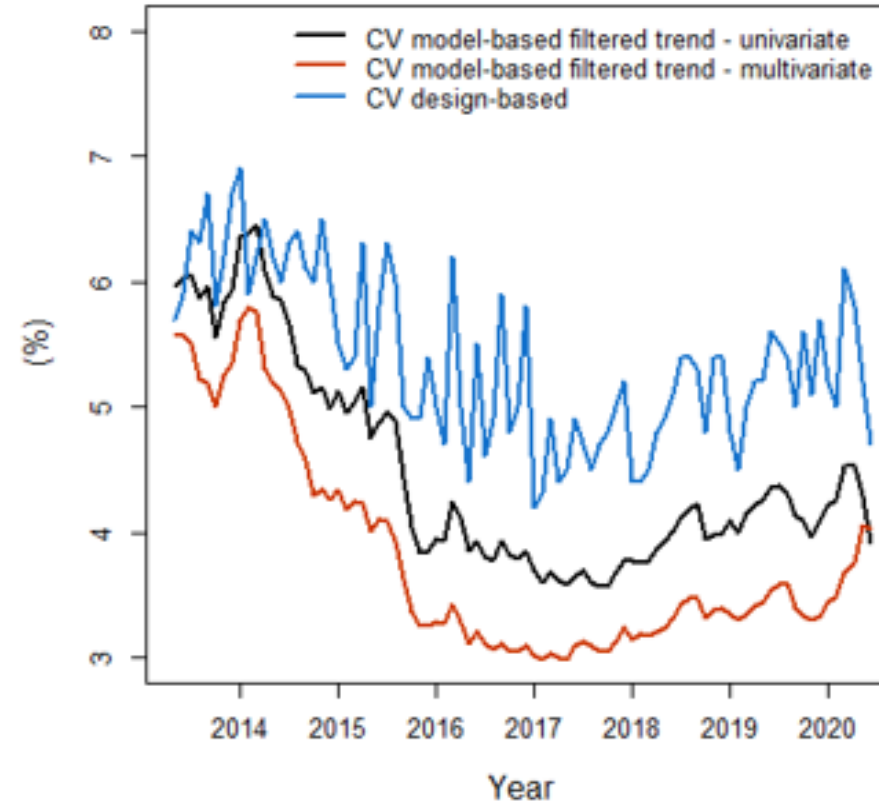
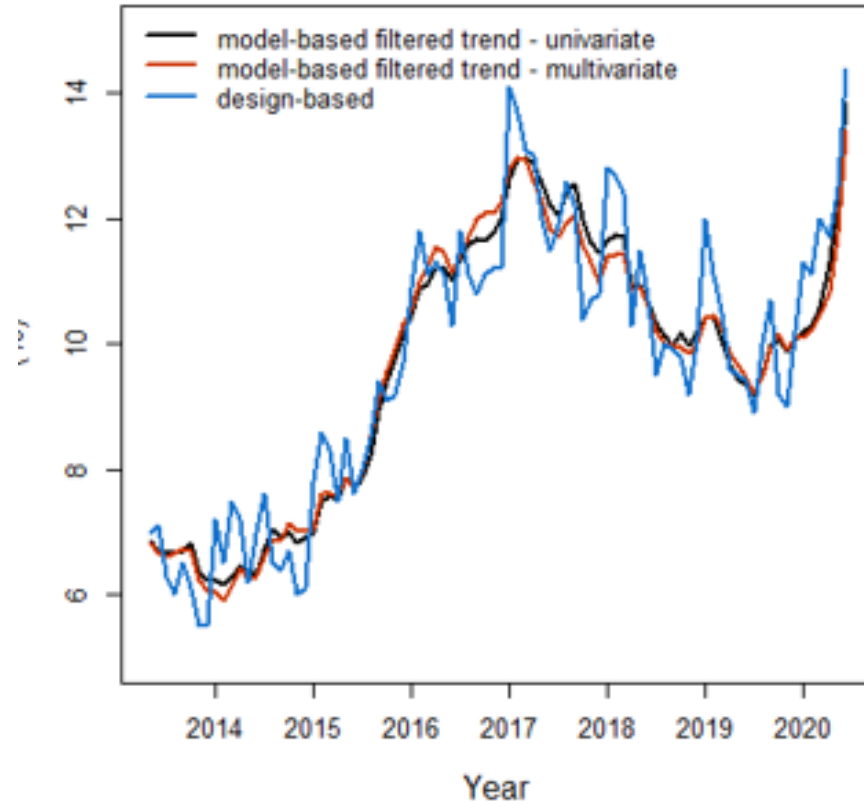


Unemployment rate design-based and model-based (trend) estimates, and respective coefficients of variation – Roraima



Results

Unemployment rate design-based, trend model-based (univariate and multivariate) estimates, and respective coefficients of variation – Minas Gerais



Correlation matrix of slope disturbance terms - Southeast region

States	Minas Gerais	Espírito Santo	Rio de Janeiro	São Paulo
Minas Gerais	1			
Espírito Santo	0.9520	1		
Rio de Janeiro	0.5436	0.7677	1	
São Paulo	0.8162	0.8581	0.7587	1

APRENDIENDO DE LOS QUE LO HACEN

Xavier Mancero, CEPAL

Limitaciones de las encuestas y posibilidades de SAE

Andrés Gutiérrez, CEPAL

Aplicaciones prácticas de SAE en América Latina

Jenny Encina, División Observatorio Social, MDSF de Chile

Estimación de la pobreza a nivel comunal en Chile

Juan Pablo Ferreira y Gastón Presto, INE del Uruguay

Una aproximación a la pobreza total e infantil para Montevideo a nivel de áreas pequeñas

APRENDIENDO DE LOS QUE LO HACEN

- La importancia de elaborar estimaciones desagregadas
- La viabilidad de la aplicación de los métodos
- La importancia de las estimaciones basadas en modelos para las políticas públicas
- El éxito del trabajo en asociación para el desarrollo y la innovación en la estimación de áreas pequeñas en nuestra región

- Modelos a nivel de área

¿También en el caso de las estimaciones para el IPM (Privación multidimensional) ?

Multiple deprivation index using small area estimation methods: an application for the adult population in Colombia (Stats Can Symposium)

Alejandra Arias-Salazar, Freie Universität Berlin; Andrés Gutiérrez, Stalyn Guerrero-Gómez, Xavier Mancero, (ECLAC); Natalia Rojas-Perilla, United Arab Emirates University and Hanwen Zhang, Universidad Autónoma de Chile

- Cuidado al decidir qué áreas deben incluirse en la estimación del modelo de previsión
- Elección de las variables auxiliares
- Estabilidad de las variables en los modelos que se utilizarán durante varios años
- Diferentes métodos de estimación según el tipo de indicador
- Una buena encuesta por muestreo es un punto de partida esencial para todos los proyectos presentados
- Se destaca el éxito del trabajo de estimación de pobreza en Chile a lo largo de los años

APRENDIENDO: DESAFÍOS

- Las estimaciones basadas en modelos están sujetas a sesgos
- Dificultad para encontrar variables auxiliares
- Posibilidad de integrar la información geoespacial
- Necesidad de evaluar la calidad y adecuación del modelo
- complejidad en el cálculo de la precisión de las estimaciones
- Un modelo para cada variable de interés (solución a medida – tailor made solution)

APRENDIENDO: PREGUNTAS...

....”si los consejos fueran buenos no se daban, se vendían”

Todavía pido



¿ De todas las precauciones al aplicar el método, hay alguna que le gustaría destacar?

¿Hay alguna parte del proyecto que, con sus conocimientos actuales, habría decidido de forma diferente (y habría ahorrado tiempo y esfuerzo)?

¿Cuál es su palabra de recomendación o estímulo para un equipo que aún no ha trabajado con estos métodos?