

SEMINARIO REGIONAL SOBRE METODOLOGÍAS DE ESTIMACIÓN EN ÁREAS PEQUEÑAS Y DESAGREGACIÓN DE DATOS

DENISE BRITZ DO NASCIMENTO SILVA

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS - ENCE - IBGE

DATOS Y DESARROLLO

- 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)

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

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.

64th ISI World Statistics Congress – Ottawa, Canada

Small Area Estimation For Sustainable Development

Organiser: Ms Clara Aida Khalil

Participants:

Pietro Gennari
Rolando Ocampo Alcántar
Stefano Falorsi
Mr Yakob Seid

Monitoring Progress Towards Sustainable Development Goals With Small Area Estimation Over Space And Time

Category: International Association of Survey Statisticians (IASS)

Organiser: Alice Richardson

Participants:

James Hogg
Justice Moses Kwaku Aheto
Sumonkanti Das
Susanna Cramb
Bernard Baffour-Awuah

Developments In Small Area Statistics Leveraging Non-Random Sampling

Category: International Association of Survey Statisticians (IASS)

Organiser: Dr Snigdhanu Chatterjee

Participants:

Yogita Gharde
Sanjay Chaudhuri
Ms Haoyi Chen
Mahmoud Torabi

Dissagregating Estimates: Small Area Estimation Advances In Latin America

Category: International Association for Official Statistics (IAOS)

Organiser: Mr ANDRES GUTIERREZ

Participants:

Rolando Ocampo Alcántar
ANGELO COZZUBO
LUNA Hidalgo
Caio César Soares Gonçalves
Andrea Diniz da Silva

From Experiment To Production: How National Statistical Offices Adopts Innovative Methods For Producing Disaggregated Data

Category: International Association for Official Statistics (IAOS)

Organiser: Mr ANDRES GUTIERREZ

Participants:

Rolando Ocampo Alcántar
Denise Lopez
Ms Haoyi Chen
Ouedraogo Boureima
Isabel Molina Peralta
Monica Pratesi

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

- **adecuados para la estimación en dominios pequeños en encuestas por muestreo repetidas en el tiempo :**
 - Encuestas Continuas de Hogares/Encuestas de Ocupación y Empleo
- incorpora las observaciones de la serie como información relevante en el proceso de estimación

INICIATIVAS RECIENTES EN BRASIL

- Estimación de los indicadores de las TIC en los estados brasileños (Brazilian Network Information Center - NIC.br)
- Estimación de pequeños dominios en la encuesta anual de servicios del IBGE
- Modelos de Series temporales para encuestas repetidas
 - Encuesta: Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC)
 - Elaboración de estimaciones del desempleo basadas en modelos de espacio de estados (state-space models)

Estimación de los indicadores de las TIC en los estados brasileños - 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.

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

- Estimación los ingresos brutos totales de los servicios por dominios 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.

MODELOS DE SERIES TEMPORALES PARA ENCUESTAS DE HOGARES 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 (FJP), Luna Hidalgo (IBGE),
Denise Silva (ENCE/IBGE), Jan van den Brakel (Statistics Netherlands e Maastrich University)

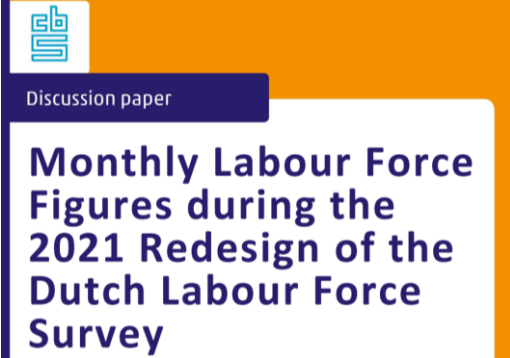
PESQUISA NACIONAL POR AMOSTRA DE DOMICÍLIOS CONTÍNUA (PNADC)

- Mayor encuesta de hogares por muestreo realizada por el IBGE
- Fuente oficial de estadísticas de desempleo en el país
- Muestreo estratificado, por conglomerado bietápico: secciones censales y hogares
- Paneles rotativos 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)

ESTIMADORES BASADOS EN MODELOS DE ESPACIO DE ESTADOS

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



3. Time series model for official monthly labour force figures

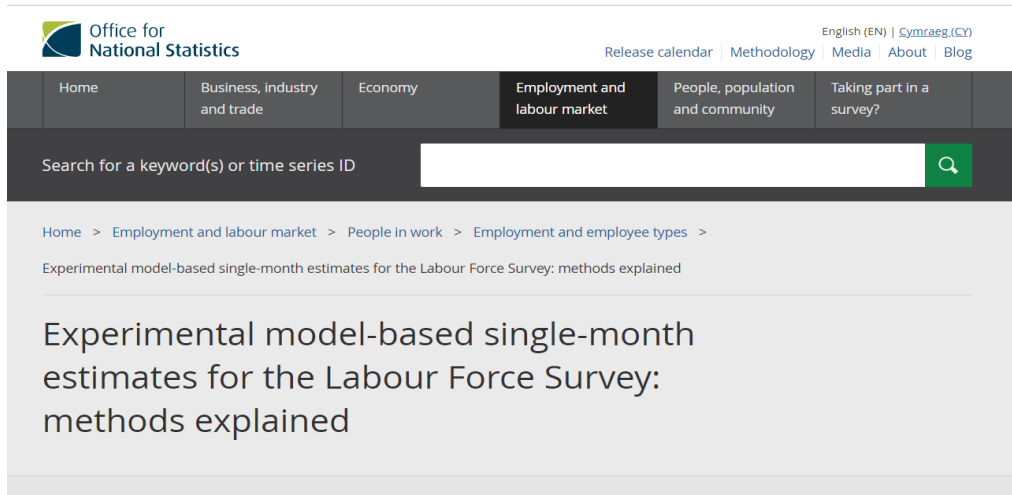
Jan van den Brakel

Januari 2022

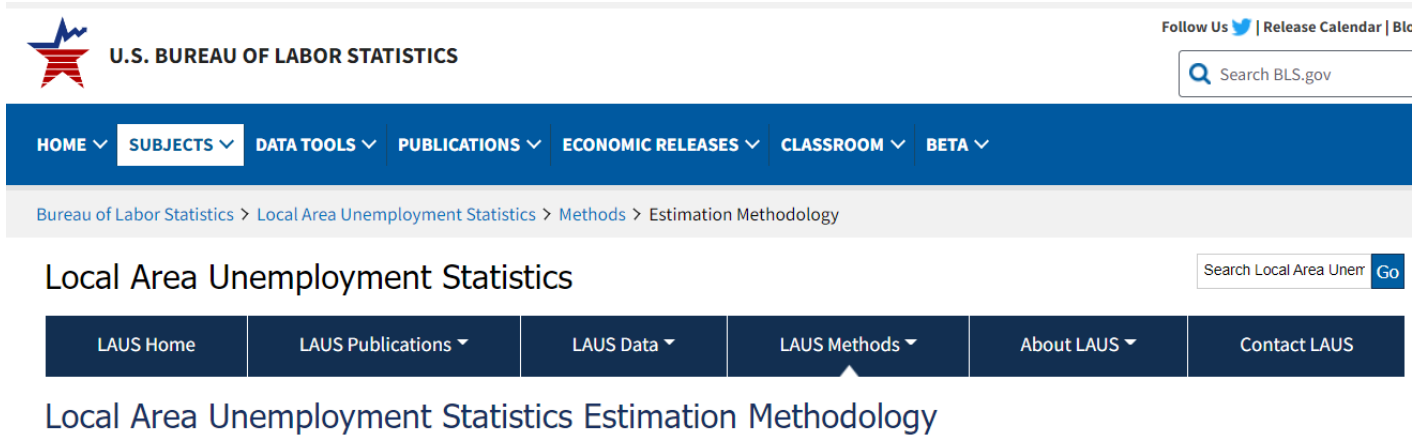


Research Paper

1351.1.55.160



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 links for "Follow Us" (with a Twitter icon), "Release Calendar", and "BLS.gov". Below this is a search bar with the text "Search BLS.gov". A dark blue navigation bar contains several menu items: "HOME", "SUBJECTS", "DATA TOOLS", "PUBLICATIONS", "ECONOMIC RELEASES", "CLASSROOM", and "BETA". Below the navigation bar is a breadcrumb trail: "Bureau of Labor Statistics > Local Area Unemployment Statistics > Methods > Estimation Methodology". The main heading is "Local Area Unemployment Statistics" with a search bar "Search Local Area Unem" and a "Go" button. Below the heading is a dark blue navigation bar with six items: "LAUS Home", "LAUS Publications", "LAUS Data", "LAUS Methods", "About LAUS", and "Contact LAUS". The page title is "Local Area Unemployment Statistics Estimation Methodology".

Estimates for states

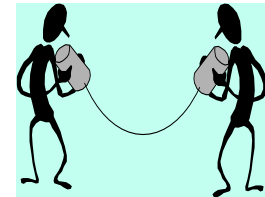
Monthly labor force data for all states and the District of Columbia are based on the time-series approach to sample survey data.¹

The Local Area Unemployment Statistics (LAUS) program is a federal-state cooperative effort in which monthly estimates of total employment and unemployment are prepared for over 7,600 areas.

ENCUESTAS POR MUESTREO REPETIDAS EN EL TIEMPO

PNADC - encuesta con un diseño de panel rotativo:

- Parte de la muestra se mantiene fija entre dos o más ocasiones de encuesta
- Se sustituyen hogares en cada ocasión por un nuevo conjunto de unidades muestrales seleccionadas :
 - solapamiento parcial de la muestra
- Mismas unidades muestrales:
 - respuestas correlacionadas en el tiempo



ENCUESTAS POR MUESTREO REPETIDAS EN EL TIEMPO

- El análisis de las series temporales debe tener en cuenta:
 - ✓ el diseño de la muestra
 - ✓ solapamiento de la muestra
 - ✓ correlación de las respuestas
- Series observadas sujetas a errores de muestreo
- Errores muestrales correlacionados en el tiempo
- Autocorrelación de la serie observada generada por el diseño muestral (superposición de unidades muestrales/hogares)

SERIES TEMPORALES DE ENCUESTAS POR MUESTREO

“When a times series of population values is estimated from a survey, the sampling errors complicate the analysis. Complex rotation patterns give rises to complex covariance structures which are superimposed on the covariance structure of the times series and should be taken account of in any analysis” (T.M. Smith, 1999)

El esquema de solapamiento y rotación de las muestras de la encuesta genera una estructura de correlación para los errores de muestreo que se confunde con la estructura de correlación de las series temporales y puede distorsionar el análisis.

SERIES TEMPORALES DE ENCUESTAS POR MUESTREO

Para obtener información sobre el verdadero valor poblacional es necesario descomponer la serie observada en dos procesos:

Señal + Error muestral

Sea \hat{y}_t una estimación del parámetro poblacional θ_t obtenida a partir de la muestra observada en el momento t

Extracción de señal en presencia de ruido

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

- \hat{y}_t la estimación directa (basada en el diseño muestral)
- θ_t **señal** - valor poblacional (parámetro desconocido)
- e_t error muestral – ruido introducido en una serie temporal debido a las características del diseño muestral

SERIES TEMPORALES DE ENCUESTAS POR MUESTREO

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

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

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

- Modelo que describe la evolución del valor real del indicador $\{\theta_t\}$
Evolución de la **señal** en el tiempo (componente no observable)
- Modelo que representa la autocorrelación de los errores generada por el diseño de paneles rotativos $\{e_t\}$

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

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

Estimaciones más desagregadas, pero a lo largo del tiempo

Los usuarios solicitan estimaciones basadas únicamente en una muestra mensual y una mayor frecuencia de publicaciones subnacionales.



ORIGINAL ARTICLE

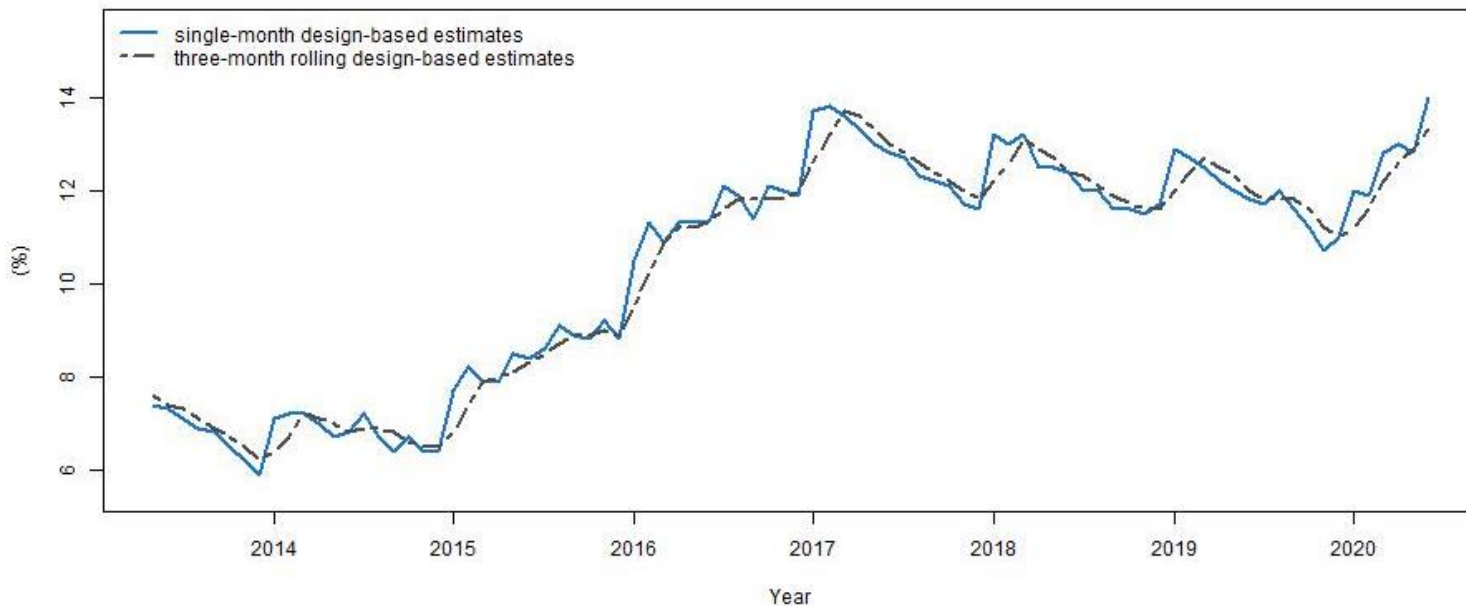
Single-month unemployment rate estimates for the Brazilian Labour Force Survey using state-space models

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



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

FOLHA DE S.PAULO
 23.set.2020 às 9h30
 Atualizado: 23.set.2020 às 12h45
 ★★★
 Diego Garcia

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 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)

Model for Sampling Error

$$e_t = \hat{c}_t \tilde{e}_t$$

Binder e Dick (1989)

\hat{c}_t : standard error of direct estimate

\tilde{e}_t : scaled sampling error

$$\tilde{e}_t = \phi \tilde{e}_{t-3} + \eta_{\tilde{e},t}, \quad \eta_{\tilde{e}} \sim N(0, \sigma_{\tilde{e}}^2)$$



Each household is interviewed once every quarter

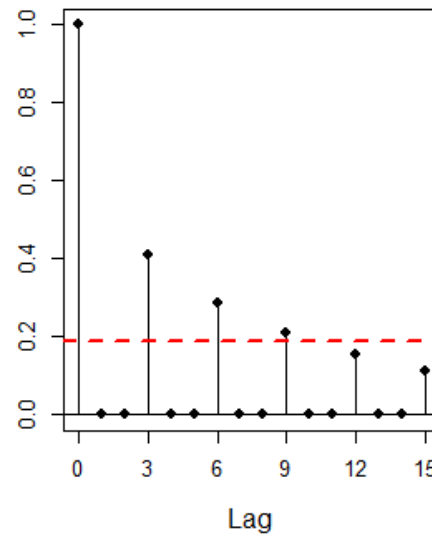
$\{e_t\} \sim \text{AR}(3)$

$$e_t = \phi_3 e_{t-3} + \eta_t^{(e)},$$

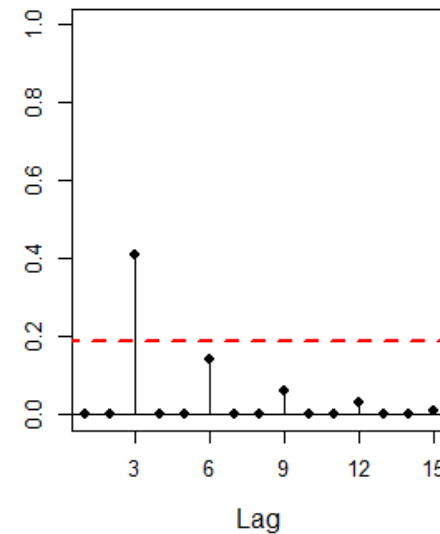
$$\phi_1 = \phi_2 = 0$$

$$\hat{\rho}_3 = \hat{\phi}_3 = 0.4106$$

ACF



PACF



Proposed model for PNADC



$$\begin{aligned}
 \{\hat{y}_t\} & \left\{ \begin{aligned} \hat{y}_t &= \theta_t + e_t \end{aligned} \right. \\
 \{\theta_t\} & \left\{ \begin{aligned} \theta_t &= T_t + S_t + I_t, & I_t &\sim N(0, \sigma_I^2) \\ T_t &= T_{t-1} + R_{t-1}, \\ R_t &= R_{t-1} + \eta_{R,t}, & \eta_{R,t} &\sim N(0, k_t \sigma_R^2) \\ S_t &= \sum_{l=1}^{\frac{s}{2}=6} S_{l,t} + \eta_{S,t}, & \eta_{S,t} &\sim N(0, \sigma_S^2) \end{aligned} \right. \\
 \{e_t\} & \left\{ \begin{aligned} e_t &= c_t \tilde{e}_t \\ \tilde{e}_t &= \phi_3 \tilde{e}_{t-3} + \eta_{e,t}, & \eta_e &\sim N(0, \sigma_e^2 \cong 1) \end{aligned} \right.
 \end{aligned}$$

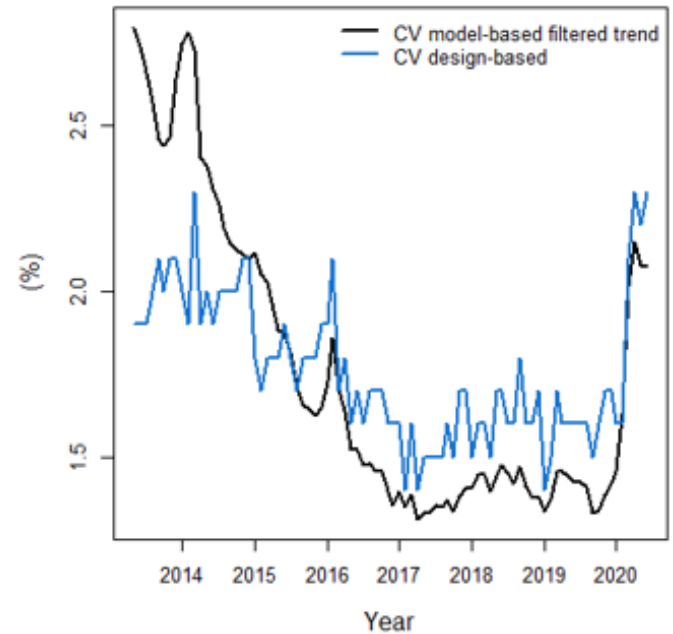
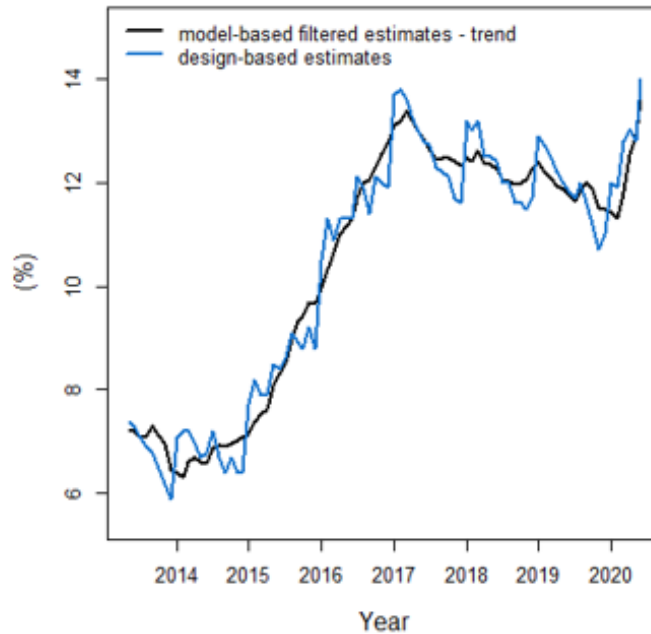
$$\hat{y}_t = \underbrace{T_t + S_t + I_t}_{\text{Structural components}} + \underbrace{e_t}_{\text{noise}}$$

signal: $\theta_t^ = T_t + S_t$*
trend: T_t
seasonally adjusted series: $\theta_t^s = T_t + I_t$

Results



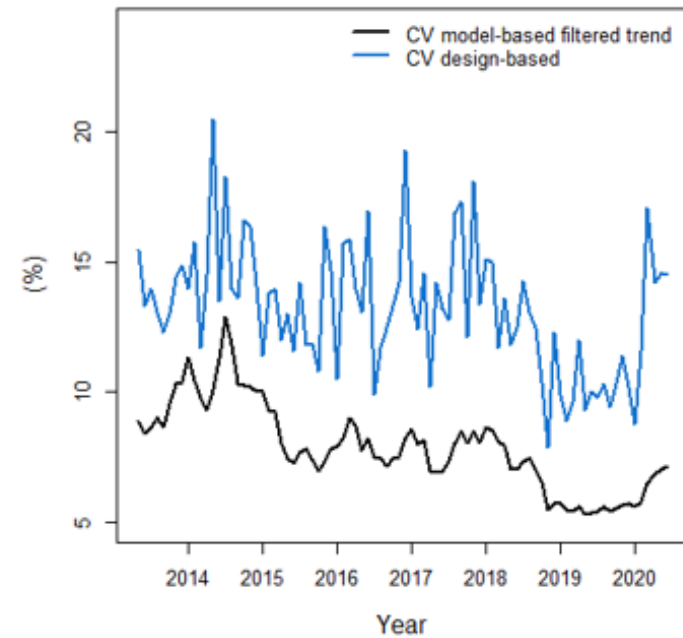
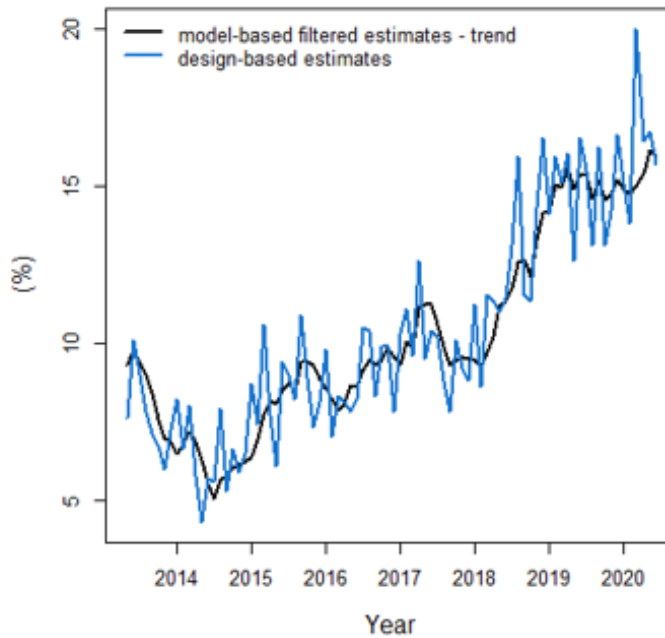
Unemployment rate design-based and model-based (trend) estimates, and respective coefficients of variation – Brazil



Results



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



$\hat{y}_{j,t}$: unemployment rate direct estimates for state j and month t

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

Model borrows strength across both time and space
(useful formulation for small area estimation)

$$\theta_{j,t} = T_{j,t} + S_{j,t} + I_{j,t}$$

$$T_{j,t} = T_{j,t-1} + R_{j,t-1}$$

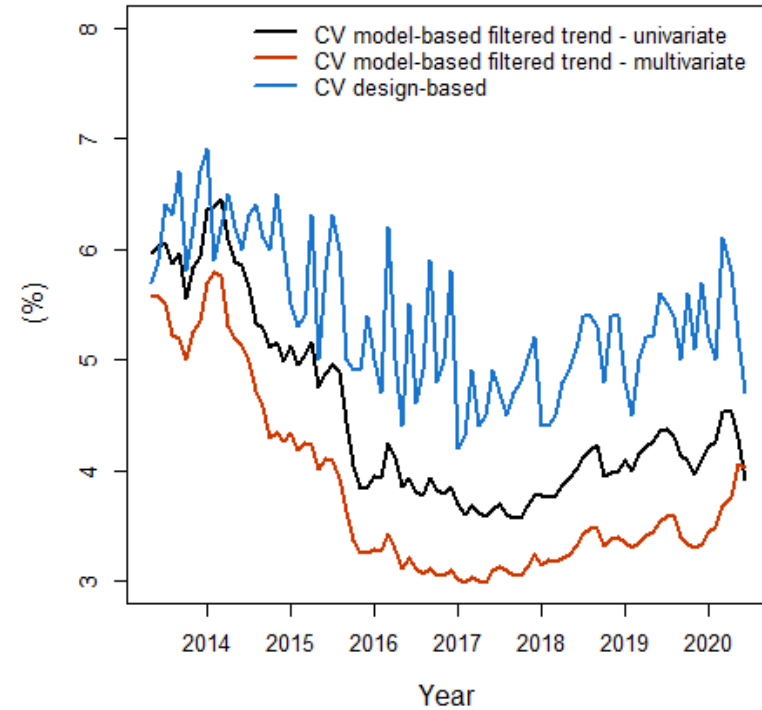
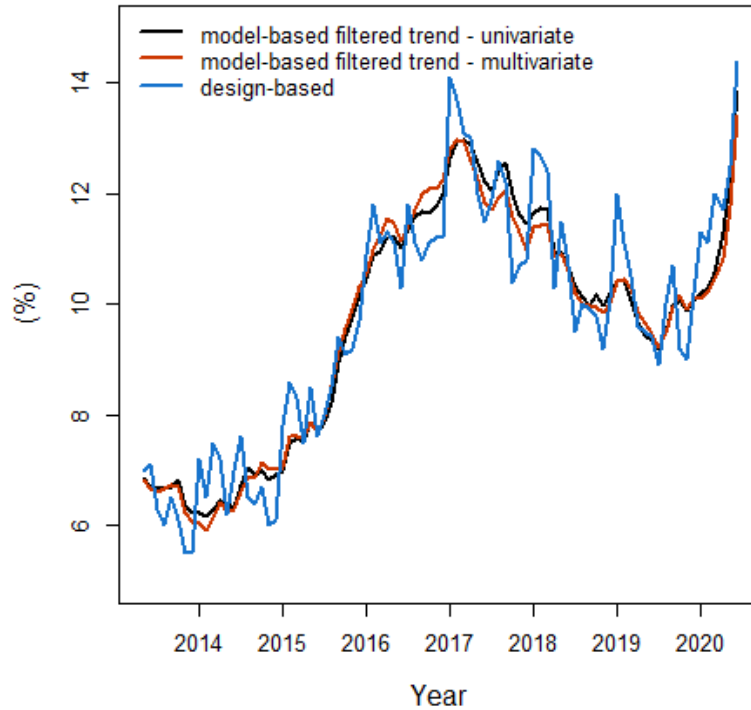
$$R_{j,t} = R_{j,t-1} + \eta_{Rj,t}$$

$$\eta_{R,j,t} \sim N(0, \sigma_{R,j}^2)$$

$$\text{cov}(\eta_{R,c,t}, \eta_{R,j',t}) = \rho_{j,j'}^R \cdot \sigma_{R,j,t}^2 \cdot \sigma_{R,j',t}^2 \quad j \neq j'$$

correlation between disturbance terms
of the slope of $\hat{y}_{j,t}$ for states j e j'

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

Modelo con datos auxiliares

x_t : dato auxiliar del mes t

Series auxiliares: seguro por desempleo, Google Trends

Modelo multivariante con una variable auxiliar :

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

Modelo de ecuaciones de series temporales aparentemente no relacionadas (SUTSE)

$$\text{COV}(\eta_{R,\hat{y},t}, \eta_{R,x,t}) = \rho_{x,\hat{y}}^R \sigma_{R,\hat{y}} \sigma_{R,x}$$

correlación entre los errores de las inclinaciones de \hat{y}_t e x_t

Gonçalves, C. C. S. Produção de indicadores do mercado de trabalho com modelos de séries temporais de pesquisas repetidas. Tesis doctoral, ENCE, IBGE, 2023.

Modelo para series Google Trends (x_t)

x_t : series Google Trends

f_t : factores

$$x_t = \hat{\Lambda} f_t + \xi_t$$

$$\xi_t \sim \mathbf{N}(\mathbf{0}, \hat{\Psi})$$

$$f_t = f_{t-1} + u_t$$

$$u_t \sim \mathbf{N}(\mathbf{0}, I_r)$$

f_t se obtienen mediante el análisis de componentes principales de x_t

Doz, Giannone and Reichlin (2011)

Bai (2003)

ESTIMADORES BASADOS EN MODELOS DE ESPACIO DE ESTADOS

Formulación versátil

Permite:

- elaborar modelos de estimación en áreas pequeñas que incorporan componentes de error de muestreo y variables auxiliares
- estimar los componentes de tendencia y estacionalidad de las series y sus intervalos de confianza



MODELOS DE ESTIMACIÓN EN ÁREAS PEQUEÑAS

$\hat{y}_{j,t}$: estimación directa para el mes t en el área j

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

El modelo adquiere fuerza tanto en el tiempo como en el espacio

$$\text{COV} \left(\eta_{R,y_j,t}, \eta_{R,y_{j'},t} \right) = \rho_{y_j,y_{j'}}^R \cdot \sigma_{R,y_j,t}^2 \cdot \sigma_{R,y_{j'},t}^2, \quad j \neq j'$$

correlación entre los errores de las inclinaciones de $\hat{y}_{j,t}$ de las áreas j e j'

MODELOS DE ESTIMACIÓN EN ÁREAS PEQUEÑAS

Mañana...Escenas del próximo capítulo en este mismo canal.. presentación de Caio Gonçalves

Estimaciones trimestrales de desocupación para la Encuesta de Fuerza Laboral de Brasil utilizando modelos de espacio-estado en áreas pequeñas

