Unemployment Estimates for the Brazilian Labour Force Survey Using State-Space Models

Caio Gonçalves
João Pinheiro Foundation and National School of Statistical Sciences (ENCE/IBGE)

Luna Hidalgo
Brazilian Institute of Geography and Statistics (IBGE)

Denise Silva
National School of Statistical Sciences (ENCE/IBGE)

Jan van den Brakel
Statistics Netherlands and Maastrich University

IASS Webinar Series: Webinar #25
25 January 2023

Motivation

- Holt, D.T. (2007).The official statistics Olympics challenge: Wider, deeper, quicker, better, cheaper. *The American Statistician*, 61(1, February), 1-8
- Challenges to the NSOs: demand for more timely and more detailed information, budget limitations, and increase of auxiliary data sources
- Research project to develop model-based estimates for the Brazilian Labour Force Survey (BLFS)
- BLFS is the largest survey household survey conducted by IBGE and is the source of official unemployment figures since 2016

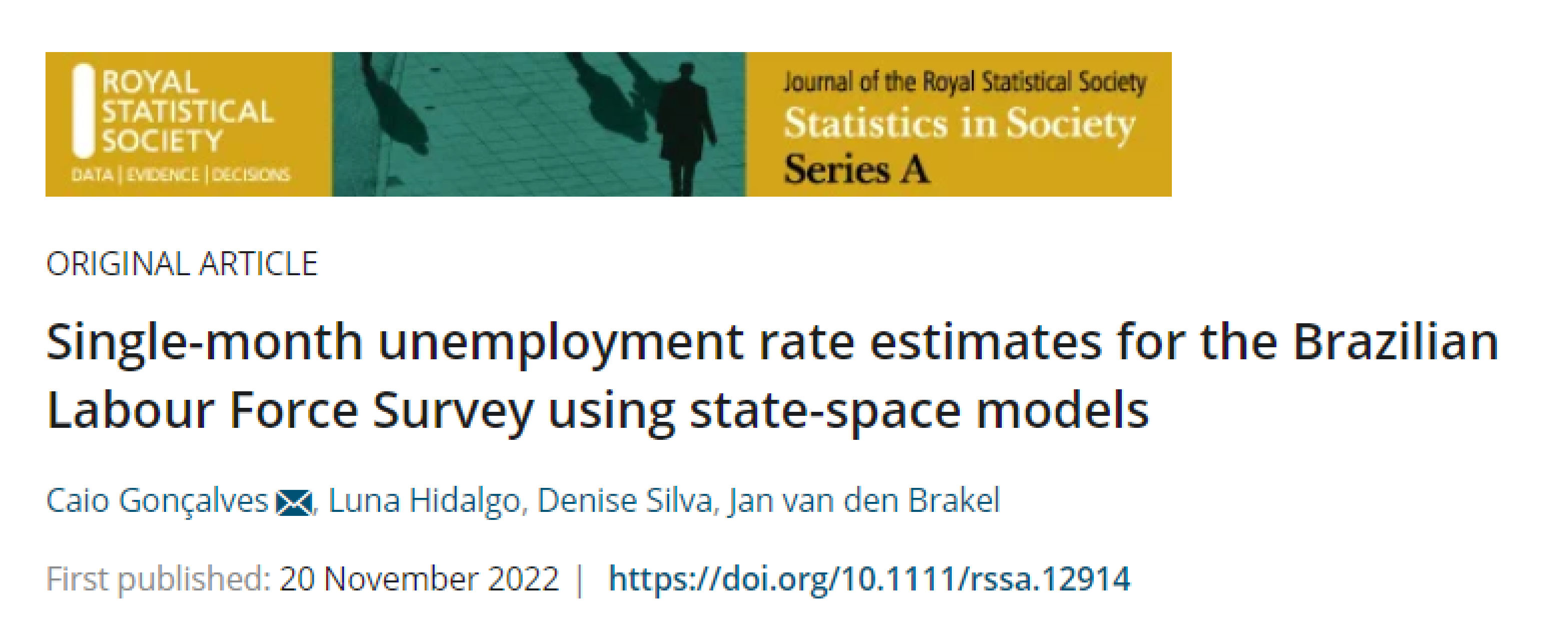
The Brazilian Labour Force Survey (BLFS)

- BLFS has a complex sample: two-stage cluster design – census enumeration areas are PSUs and households are SSUs
- Rotating panel survey with a partially overlapping sample of households – scheme 1-2(5) – (five rotation groups)
- Planned sample overlap between quarters: 80% of households
- BLFS national estimates are released monthly (based on rolling quarterly data), and subnational estimates are published quarterly
- Each household is interviewed once every three months

Context

- BLFS national estimates are released monthly (based on three-month rolling data), and subnational estimates are published quarterly
- The need to produce single-month estimates to monitor the labour market emerged markedly after the COVID-19 pandemic for both national and state levels
- Users have been calling for estimates based solely on a single-month sample, and for a greater frequency of subnational releases
- The amount and variety of alternative data sources, such as big data, are rapidly growing and consolidating as a potential data source for producing official statistics

First project output



ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

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

Users have been calling for estimates based solely on a single-month sample, and for a greater frequency of subnational releases

What followed next?

- Research to investigate the potential of producing:
    - ✓ precise estimates of monthly unemployment figures based on multivariate time series models that integrate survey data with big data
    - ✓ nowcast estimates since Google Trends series are available one month before the single-month unemployment estimates
- Common trend models with survey data and Google Trends time series might have some potential, especially for small samples at the state level or for specific population groups such as young people

Google Trends

- Google Trends provides a series of word queries for several countries, and in some cases, queries can be specific for states or provinces

- The data is also grouped into categories using a natural language classification engine, such as health, employment, sports, travel, etc.

- Google Trends does not inform the exact search volume. The data is normalised on a scale from 0-100, where the maximum (100) represents the query’s highest point considering a specific beginning and end (GOOGLE, 2022)

Literature using Google Trends

- Choi and Varian (2009a), Choi and Varian (2009b) and Choi and Varian (2012) revealed the potential of Google Trends time series for producing nowcast estimates of economic indicators and for identifying turning points in the series

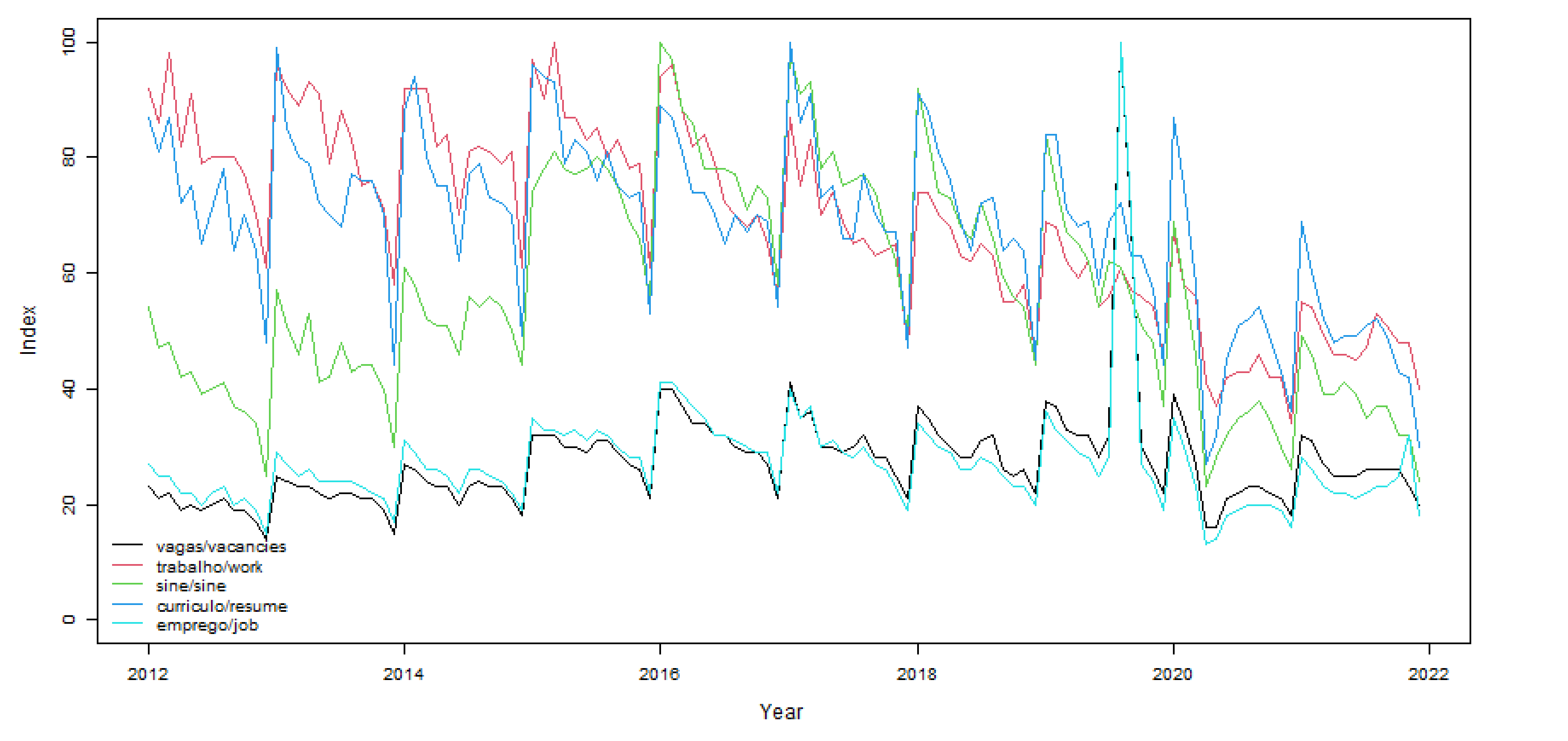
- Jun, Yoo and Choi (2017) listed several areas of study that used Google Trends query data - ten years summary (2006-2016)

- Schiavoni et al. (2020) proposed a dynamic factor approach to incorporate auxiliary information from high-dimensional data sources in a state space model also including survey and administrative data

Google Trends word selection

A different set of word series were considered for each state as Brazil is a large country and has a diverse workforce behaviour

Figure 1: Selected Google Trends series - Brasil - Jan. 2012 - Dec. 2021



Source: Google Trends.

Time series models for repeated surveys

- The model proposed by Schiavoni *et al.* (2020) combines a univariate signal extraction model for unemployment estimates from a repeated survey with the Google Trends series modelled by a dynamic factor model

- It is a type of Seemingly Unrelated Time Series Equations (SUTSE) model

Modelling the survey data ($\hat{y}_t$)

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

Signal extraction:

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

Unobserved components of unknown population quantity θ_t:

$$\theta_t = T_t + S_t + I_t \qquad I_t \sim N(0, \sigma_I^2)$$

Trend:

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

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

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

Modelling the survey data ($\hat{y}_t$)

Seasonal component S_t : $S_t = \sum_{l=1}^{\frac{s}{2}=6} S_{l,t}$

Sampling error e_t: $e_t = \hat{c}_t \tilde{e}_t$

$\hat{c}_t$: standard error of design-based estimates

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

↓

each household is interviewed once every quarter

Modelling the Google Trends data (x_t)

$\boldsymbol{x_t}$: Google Trends series
$\boldsymbol{f_t}$: factors

$$\boldsymbol{x_t} = \widehat{\boldsymbol{\Lambda}}\,\boldsymbol{f_t} + \boldsymbol{\xi_t} \qquad \boldsymbol{\xi_t} \sim \mathbf{N}\left(\mathbf{0}, \widehat{\boldsymbol{\Psi}}\right)$$

$$\boldsymbol{f_t} = \boldsymbol{f_{t-1}} + \boldsymbol{u_t} \qquad \boldsymbol{u_t} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{I_r})$$

- $\boldsymbol{f_t}$ is obtained via principal components analysis of $\boldsymbol{x_t}$
- $\widehat{\boldsymbol{\Psi}}$ and $\widehat{\boldsymbol{\Lambda}}$ are estimated by OLS regression
- $\boldsymbol{f_t}$ is reestimated as state variables of the dynamic factor model that incorporates survey data using Kalman filter

Combining survey data and Google Trends

$$\begin{pmatrix} \hat{y}_t \\ \boldsymbol{x_t} \end{pmatrix} = \begin{pmatrix} \theta_t \\ \boldsymbol{\hat{\Lambda}}\, \boldsymbol{f}_{\boldsymbol{t}} \end{pmatrix} + \begin{pmatrix} e_t \\ \boldsymbol{u}_t \end{pmatrix}$$

The joint model includes the correlation/covariance between the slope disturbance term $\eta_{R,t}$ and the factor's disturbance term $\boldsymbol{u_t}$.

The covariance is:

$$cov\left(\eta_{R,t}, u_{i,t}\right) = \rho^R_{y,fi}\ \sigma_{R,t} \qquad\qquad i = 1, \dots, r$$

since $\sigma_{fi,t} = 1$.

Combining survey data and Google Trends

$$\begin{pmatrix} \hat{y}_t \\ \boldsymbol{x_t} \end{pmatrix} = \begin{pmatrix} \theta_t \\ \widehat{\boldsymbol{\Lambda}}\, \boldsymbol{f}_{\boldsymbol{t}} \end{pmatrix} + \begin{pmatrix} e_t \\ \boldsymbol{u_t} \end{pmatrix}$$

Selecting the Google Trends (x_t)

- Tests to detect outliers were performed and outliers were replaced by linear interpolation
- The seasonal component was removed from Google Trends series
- The series must be I(1). Augmented DickeyFuller (ADF) stationarity tests were performed
- Cases of highly correlated series were identified due to slight differences in query terms such as word order

Strategies for targeting the predictors

- Selection of series before conducting the PCA is required
- Adding a series with no impact on factors deteriorates the PCA estimation process

Bai e Ng (2008) found that targeting the predictors procedure is always better

$$\boldsymbol{x_t} = \ \widehat{\boldsymbol{\Lambda}}\,\boldsymbol{f_t} + \boldsymbol{\xi_t} \qquad\qquad \boldsymbol{\xi_t} \sim \mathbf{N}\left(\mathbf{0}, \widehat{\boldsymbol{\Psi}}\right)$$

- Elastic Net
- Time Series Clustering
- Bivariate structural model

Elastic Net

- Bai e Ng (2008) indicated the use of the elastic net proposed by Zou e Hastie (2005)

- The method consists in a penalised regression that sets the coefficients of the irrelevant series to zero or to a close value

- Series with non-zero coefficients are selected for inclusion in the PCA estimation procedure

Time Series Clustering

- The cluster analysis handles the series of total unemployment (first difference of slope) together with the time series from Google Trends (also in first difference)
- The Dynamic Time Warping (DTW) distance is a type of shape-based time-series clustering
- One of the initial studies of DTW is from Sakoe and Chiba (1978) for speech recognition
- Google Trends series classified in the same group of the unemployment series are selected

Bivariate Structural Model

- Bivariate structural models are fitted to determine whether there is a significant correlation between the slope disturbances of the unemployment series and each Google Trends series

- Google Trends series for which there is evidence to reject the null hypothesis of $\rho_{y,x}^{R} = 0$ are selected

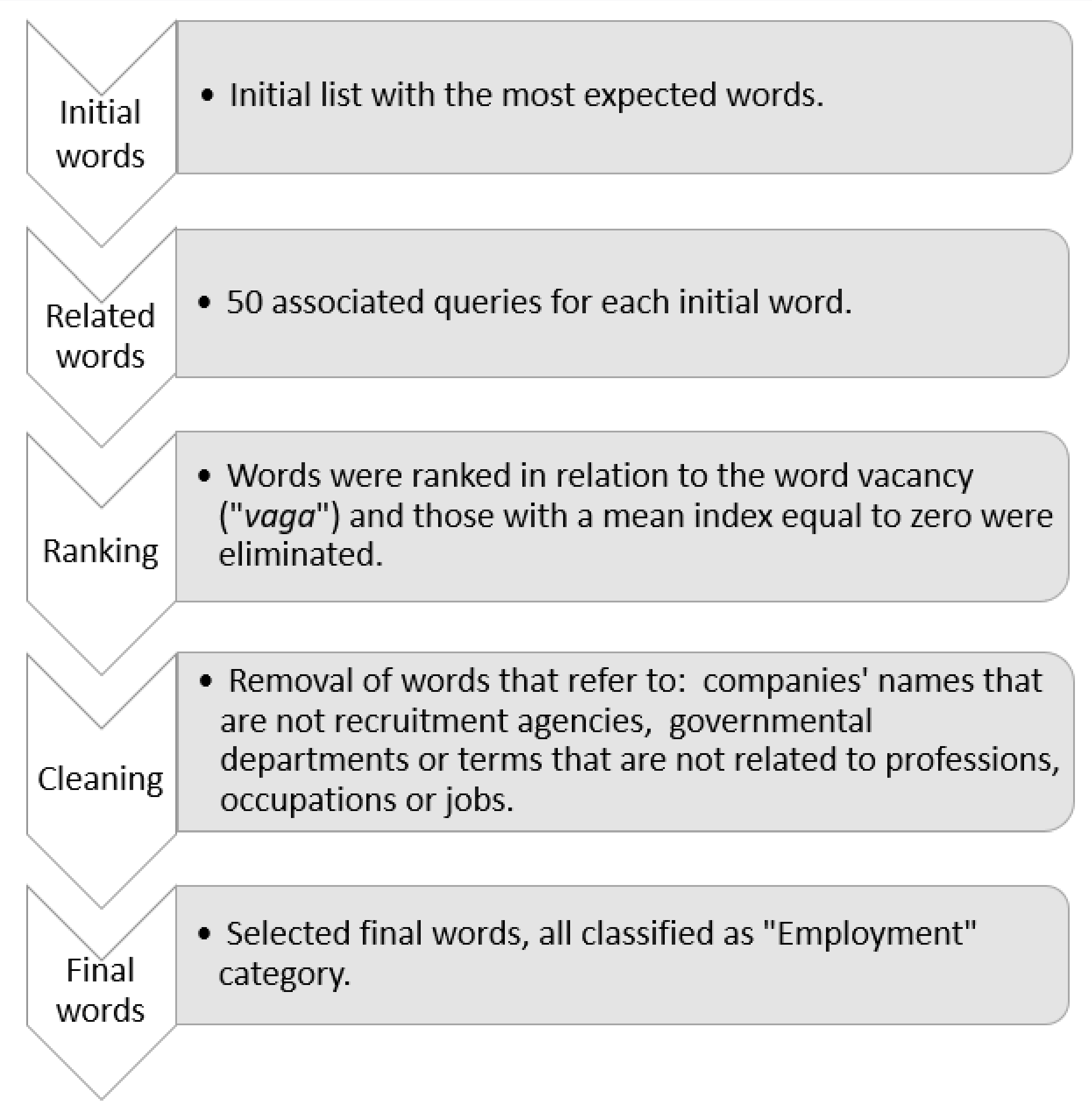
Models

Dynamic factor models with different strategies for targeting the predictors (elastic net, time series clustering and bivariate model) are examined:

- UN: univariate model (baseline)
- DFM EL: dynamic factor model with the elastic net
- DFM CL: dynamic factor model with the time series clustering
- DFM BI: dynamic factor model with the bivariate structural model
- DFM EL+BI: dynamic factor model with the elastic net combined with the bivariate structural model
- DFM CL+BI: dynamic factor model with the time series clustering combined with the bivariate structural model

Google Trends word selection prior to targeting procedure

Figure 2: Steps for word selection



Initial words

- Initial list with the most expected words.

Related words

- 50 associated queries for each initial word.

Ranking

- Words were ranked in relation to the word vacancy ("*vaga*") and those with a mean index equal to zero were eliminated.

Cleaning

- Removal of words that refer to: companies' names that are not recruitment agencies, governmental departments or terms that are not related to professions, occupations or jobs.

Final words

- Selected final words, all classified as "Employment" category.

Results: selection of words

The analysis period spans from Jan2012 until Dec2021 for Brazil, Minas Gerais and Roraima

Table 1: Number of words considered in the series selection and treatment stages prior to targeting procedure

Stages	Brazil	Minas Gerais	Roraima
Initial words	82	82	82
Related words	1964	1246	176
Most searched words	527	560	112
After removing inadequate words	448	430	85
I(1)	181	65	5
Final words	150	49	2

Results: targeting strategies

Table 2: Selected words by targeting strategies - working age population - Minas Gerais

Words	EN	CL	BI	Words	EN	CL	BI
agencia de empregos		x	x	sine bh			x
cadastrar curriculo			x	sine contagem	x		x
catho			x	sine de bh		x	x
como fazer um curriculo	x			sine empregos		x	
currículo	x			sine vagas bh			x
curriculum	x			trabalho			x
emprego contagem		x		vagas bh	x		
emprego em bh		x		vagas de emprego bh		x	
empregos	x	x		vagas de emprego em bh			x
empregos bh		x		vagas de emprego sine		x	x
empregos uberlandia	x		x	vagas no sine	x		x
infojobs	x		x	vagas sine			x
jf emprego			x	varginha online	x		
sine		x	x				

Results

Table 3: Estimated hyperparameters, by targeting strategy and number of factors - working age population unemployment - Brazil and Minas Gerais

Model	DFM CL	DFM CL	DFM BI	DFM BI	DFM CL+BI	DFM CL+BI
Number of factors	1	2	1	2	1	2
Brazil						
Number of Google Trends series	22	22	18	18	5	5
$\rho^{R}_{y,f1}$	-0.59	-0.13	0.64	0.23	0.86	0.85
$\rho^{R}_{y,f2}$		0.84		-0.83		-0.24
$\boldsymbol{\rho}^{\boldsymbol{R}}_{\boldsymbol{y},\boldsymbol{f}} = 0$ (p-value)	0.05	0.01	0.01	0.00	0.00	0.00
Minas Gerais						
Number of Google Trends series	11	11	17	17	5	5
$\rho^{R}_{y,f1}$	0.43	0.17	0.44	0.40	0.54	0.42
$\rho^{R}_{y,f2}$		0.71		-0.13		-0.43
$\boldsymbol{\rho}^{\boldsymbol{R}}_{\boldsymbol{y},\boldsymbol{f}} = 0$ (p-value)	0.35	0.36	0.19	0.48	0.27	0.45

Results

Relative mean squared error (RMSE)
Relative mean squared forecast error (RMSFE)

Table 4: Accuracy measures for selected unobservable components - working age population - Brazil

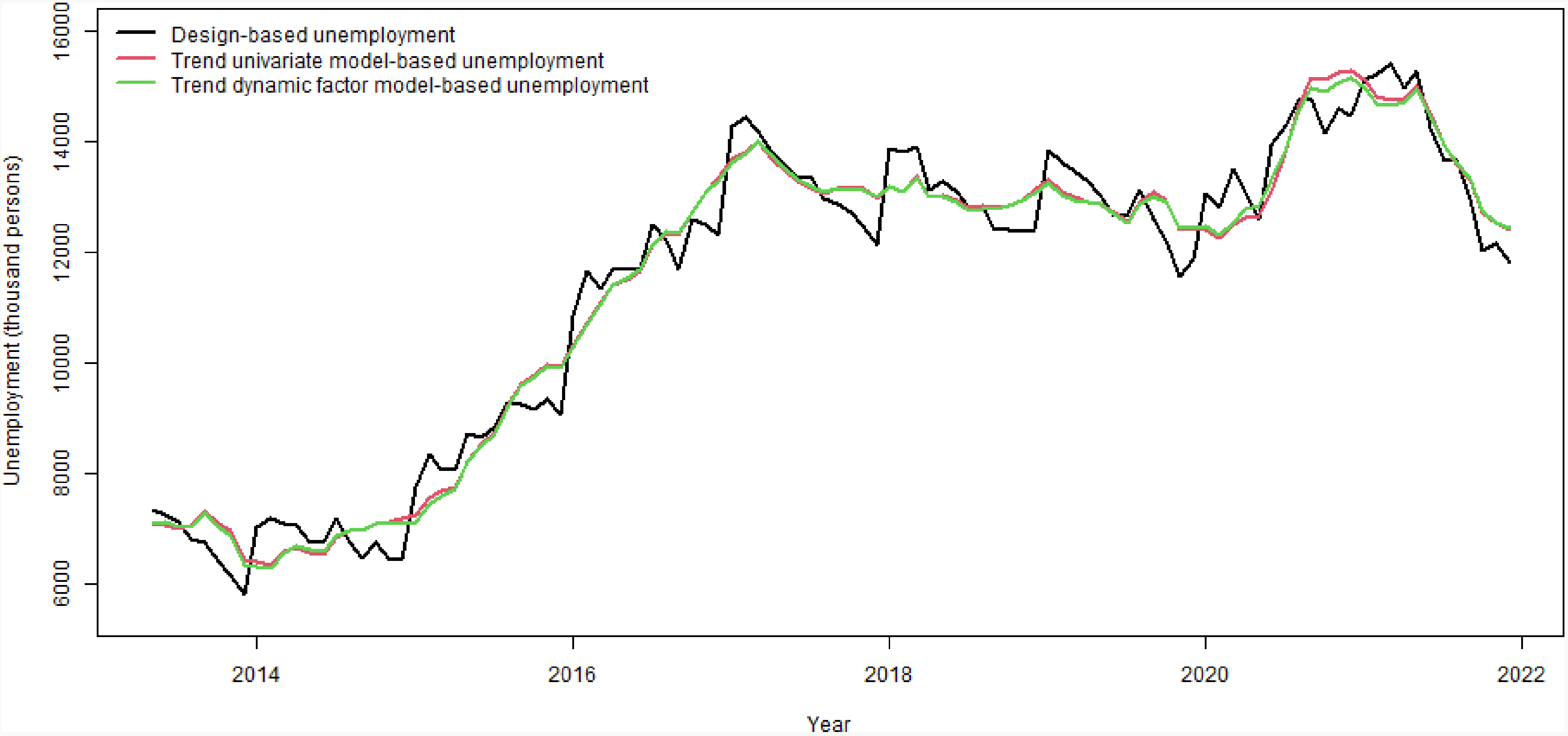
Models	Number of factors	RMSE			RMSFE		
		T	R	θ	T	R	θ
UN	-	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
DFM CL	1	0.9545	1.0091	0.9533	0.9524	0.9950	0.9516
DFM CL	2	**0.8977**	**0.8648**	0.9337	**0.8691**	**0.8251**	**0.8776**
DFM BI	1	0.9181	0.9100	0.9324	0.9013	0.8890	0.9049
DFM BI	2	**0.8822**	**0.8653**	**0.9171**	**0.8682**	**0.8230**	**0.8750**
DFM + CL+BI	1	0.9240	0.9808	0.9370	0.9141	0.9349	0.9170
DFM + CL+BI	2	0.9025	0.9408	**0.9195**	0.8978	0.9129	0.9008

Results

DFM BI: dynamic factor model with bivariate structural model as targeting strategy

Figure 3:

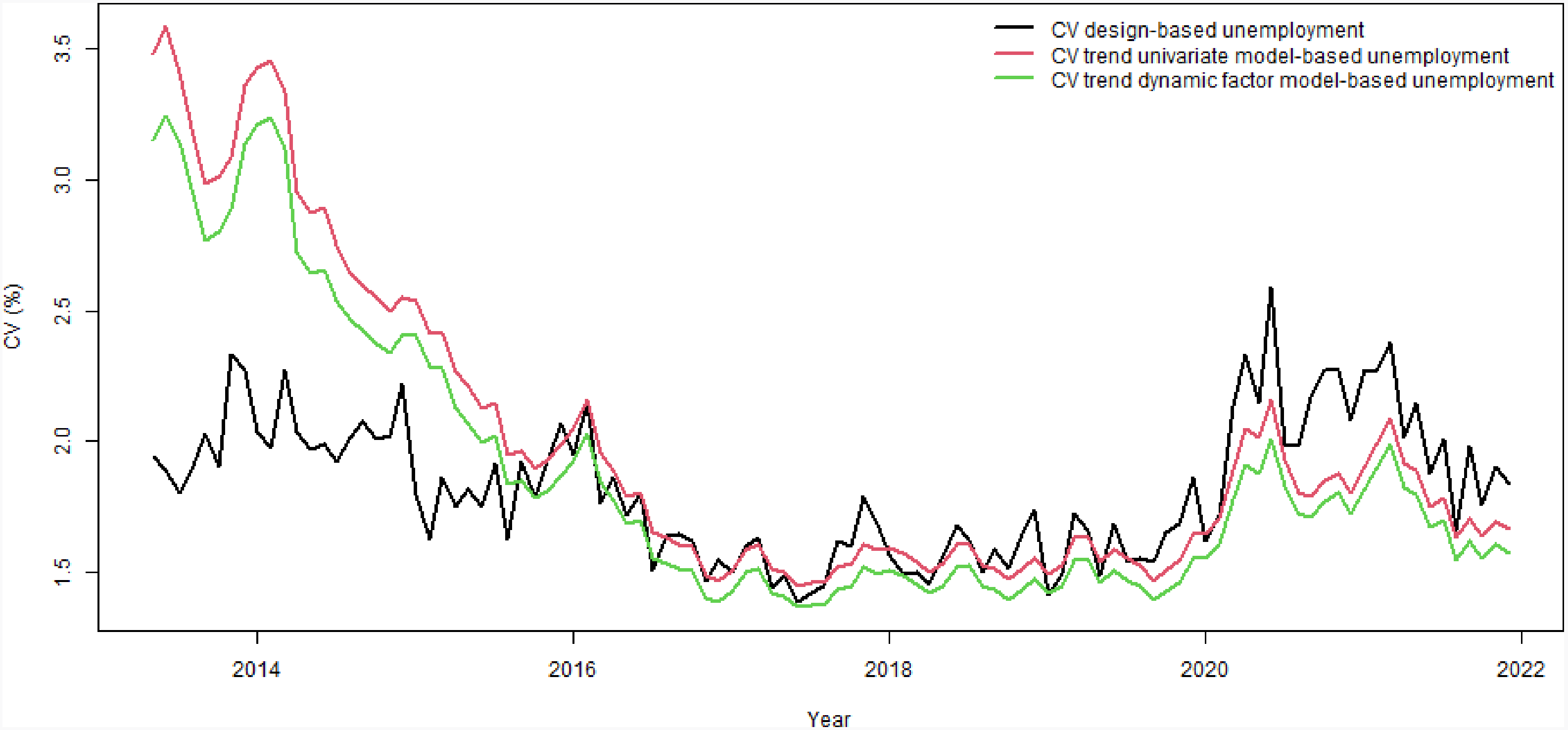
Design-based and model-based estimates for unemployment - Brazil



Results

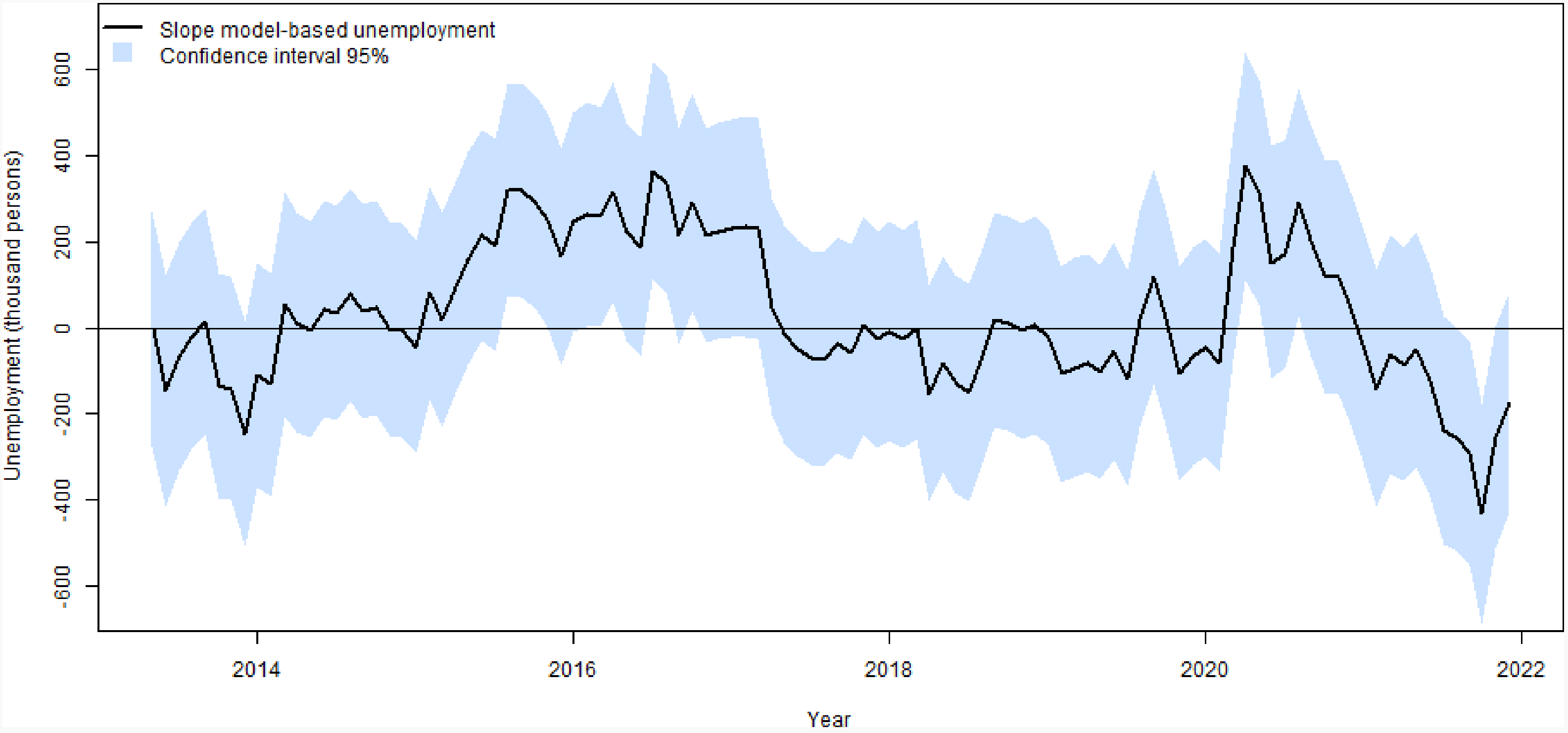
Figure 4:

Design-based and model-based precision estimates for unemployment - Brazil



Results

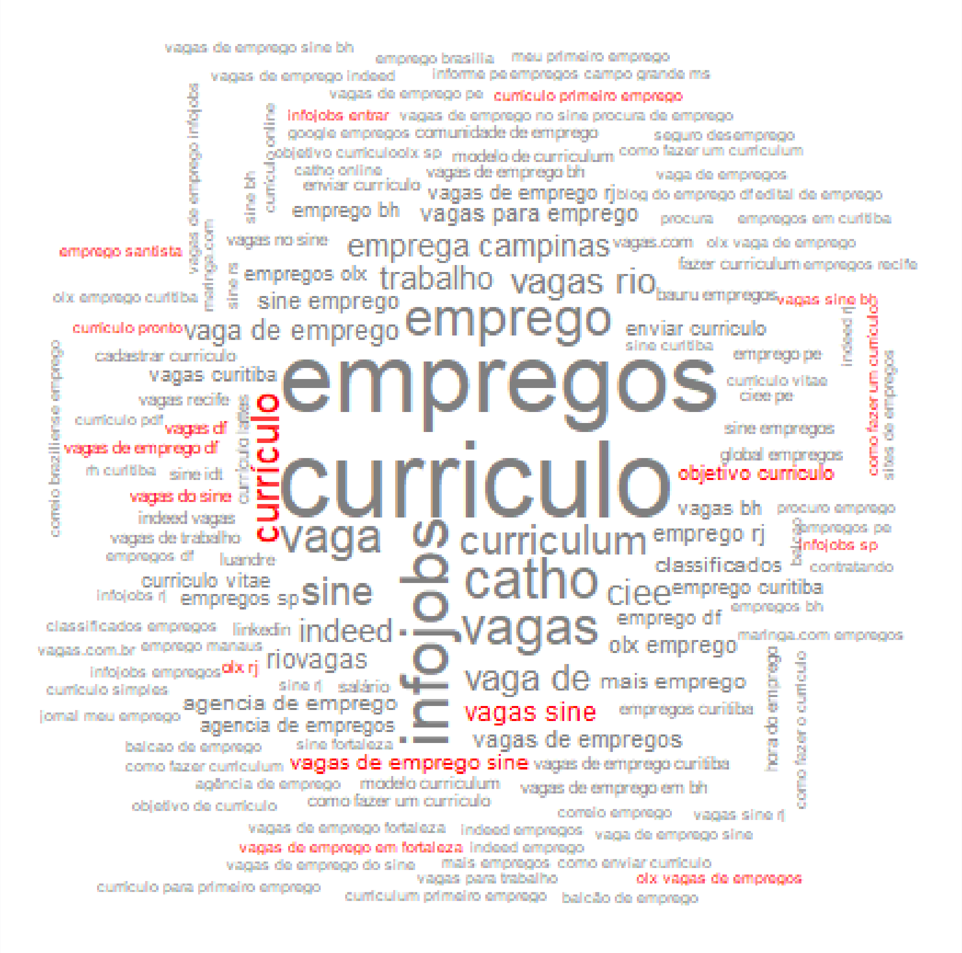
Figure 5: Model-based estimates for month-to-month changes in unemployment figures at national level - Brazil



Results

Figure 6:

Words selected for modelling represented by relative ranking to “*vaga*”



Against expectations, popular searched words did not compose time series related to trend variation of unemployment in Brazil.

Conclusions

National level

- The results provide favourable evidence, presenting estimates with good precision, borrowing strength from Google Trends series and producing nowcast estimates with higher precision than the univariate models

State and youth unemployment

- At the state level, there was no evidence of benefits of modelling Google Trends series in conjunction with the unemployment series.

- The same can be said for young unemployed people. It was expected that, as a group that tends to use the internet more, youth unemployment series would have greater adherence to the behaviour of word searches on the Google Trends platform.

References I

BAI, J.; NG, S. Forecasting economic time series using targeted predictors. Journal of Econometrics, Elsevier Science, v. 146, p. 304-317, 2008. Disponível em: <http://doi.org/10.1016/j.jeconom.2008.08.010>.

CHOI, H.; VARIAN, H. Predicting initial claims for unemployment benefits. [S.l.], 2009. Disponível em: <https://static.googleusercontent.com/media/research.google.com/pt-BR//archive/papers/initialclaimsUS.pdf>.

CHOI, H.; VARIAN, H. Predicting the present with Google Trends. [S.l.], 2009. Disponível em: <https://static.googleusercontent.com/media/www.google.com/pt-BR//googleblogs/pdfs/google_predicting_the_present.pdf>.

CHOI, H.; VARIAN, H. Predicting the present with google trends. Economic Record, v. 88, 2012. Disponível em: <http://doi.org/10.1111/j.1475-4932.2012.00809.x>.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, v. 33, n. 1, p. 1-22, 2010. Disponível em: <https://www.jstatsoft.org/v33/i01/>.

GOOGLE. Google Trends lessons. 2022. Disponível em: <https://newsinitiative.withgoogle.com/training/lessons?tool=Google%20Trends&image=trends>.

HARVEY, A.; CHUNG, C. Estimating the underlying change in unemployment in the uk. Journal of the Royal Statistical Society: Series A (Statistics in Society), John Wiley and Sons, v. 163, p. 303-309, 2000. Disponível em: <http://doi.org/10.1111/1467-985x.00171>.

HYNDMAN, R. J.; KHANDAKAR, Y. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software, v. 26, n. 3, p. 1-22, 2008.

JUN, S.; YOO, H. S.; CHOI, S. Ten years of research change using google trends: From the perspective of big data utilizations and applications. Technological Forecasting and Social Change, p. S0040162517315536, 2017. Disponível em: <http://doi.org/10.1016/j.techfore.2017.11.009>.

PFAFF, B. Analysis of Integrated and Cointegrated Time Series with R. Second. New York: Springer, 2008. ISBN 0-387-27960-1. Disponível em: <https://www.pfaffikus.de>.

PFEFFERMANN, D. Methodological issues and challenges in the production of official statistics. Journal of Survey Statistics and Methodology, v. 3, p. 425-483, 2015. Disponível em: <http://doi.org/10.1093/jssam%2Fsmv035>.

References II

SAKOE, H.; CHIBA, S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics Speech and Signal Processing, IEEE, v. 26, p. 0-49, 1978. Disponível em: <http://doi.org/10.1109/tassp.1978.1163055>.

SARDA-ESPINOSA, A. dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance. [S.l.], 2022. R package version 5.5.10. Disponível em: <https://CRAN.R-project.org/package=dtwclust>.

SCHIAVONI, C. et al. A dynamic factor model approach to incorporate big data in state space models for official statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society), p. rssa.12626, 2020. Disponível em: <http://doi.org/10.1111/rssa.12626>.

TOEWS, M. W.; WHITFIELD, P. H.; ALLEN, D. M. Seasonal statistics: the 'seas' package for r. Computers & Geosciences, v. 33, n. 7, p. 1895, 2007.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B (Methodological), v. 67, p. 301-230, 2005.

Thank you!