New Data Sources for Official Statistics

with applications at Statistics Netherlands

Jan A. van den Brakel [1] [2]

Third Workshop on Methodologies for Official Statistics
Statistics Italy
Rome, 4-5 December 2024

[1] Statistics Netherlands, Department of Statistical Methods
[2] Maastricht University, Department of Quantitative Economics

Outline

- Introduction
- Non-probability data as a primary data source
- Non-probability data as covariates in model-based inference
  - Time series models for the Dutch LFS
  - Predicting poverty from aerial images
- Machine learning to improve sampling strategy
- Discussion

The views expressed in this paper are those of the author and do not necessarily reflect the policy of Statistics Netherlands!

Introduction

Introduction

Official statistics:

- Traditionally probability samples with design-based or model-assisted inference
- Advantage:
  - Robust for model-misspecification
  - Designed data, which implies control over:
    - Precision (through sample design)
    - Availabilty and stability over time of the data
    - Operationalization of concepts (through questionnaire design)
  - Low risk level of impairment under this approach

Introduction

- Problems with survey sample data:
  - expensive
  - not very timely
  - high variances for small domain estimates (under design-based inference methods)
  - non response
  - response burden
- Increasing interest in alternative data sources (big data):
  - time and location of network activity available from mobile phone companies,
  - social media messages from X (Twitter) and Facebook
  - internet search behaviour from Google Trends
  - sensor data and satellite or aerial images
  - administrative data like tax registers

Introduction

Strong points non-probability data

- Large amount of records
- Cost effective
- High frequency (in real time)
- Detailed level
- Direct measurement of behaviour instead of asking

Introduction

Weak points non-probability data

- Selection bias / DGP unknown
- Unstructured
- No/poor auxiliary variables
- Often suboptimal construct for the intended target variables
- High risk level
  - No design-phase to control accuracy
  - Model-based inference procedures to combine non-probability data with survey data or to correct for selection bias
  - No control over availability, stability, and consistency of the data source

Introduction

Use of non-probability data in official statistics:

- Primary data source
- Covariates in model-based estimation methods
- Auxiliary information to improve sampling strategy in a design-based approach

Non-probability data as primary data source

Non-probability data as primary data source

- Selection bias
- X. Meng (2018), Statistical paradises and paradoxes in big data :
  - Estimation error product of
    - Data-defect correlation: $\rho = cor(Y_i, R_i)$
    - Data-quantity measure: $\sqrt{(N-n)/n}$
    - Problem difficulty measure: $\sqrt{S_Y^2}$
  - Small deviation of $\rho = 0$ result in large bias
  - Effective sample size of a NP sample with n = 2,300,000 reduces to SRS with n = 400 (if ρ = -0.005)

Non-probability data as primary data source

Literature for bias correction in non-probability samples:

- Weighting and calibration (Särndal et al., 1992)
- Informative sampling (Pfeffermann and Sverchkov, 2003, 2009)
- Quasi-randomization: explicit model for estimating inclusion probabilities (Elliot and Valliant, 2017, Valliant and Dever, 2011, Chen et al. 2020, Rafei et al., 2020)
- Super population model to predict target variables not included in the sample (Valliant et al. 2000)
- Reference sample to assess the selectivity of a non-probability sample (Kim and Wang, 2018, Beaumont et al., 2024)

Non-probability data as primary data source

Literature for bias correction in non-probability samples (cnt.):

- Literature inference methods for NP data
  - Rao (2020), On Making Valid Inferences by Integrating Data from Surveys and Other Sources, *Sankya, B.*
  - Beaumont (2020), Are probability samples bound to dissapear for the production of official statistics? *Survey Methodology*
  - Wu (2022), Statistical inference with non-probabilty survey samples, *Survey Methodology*

Non-probability data as primary data source

- Issues with big data as primary data source:
  - Methods assume structured data (identify units of the target population in the non-probabilty data source)
  - Methods assume availability of auxiliary information in the non-probabilty data source
- Aforementioned literature: opt-in panels

Non-probability data as primary data source

- Appropriate for official statistics?
  - Strong assumptions
  - No control over availability and comparability over time of the data
  - Results in a higher risk level
- Applications:
  - Scanner data for measuring price indices
  - Income statistics from tax and social benefit registers
  - Business statistics from VAT registers
  - Aerial images for measuring solar power panels and land use
  - ...

Non-probability data as covariates in a model-based inference approach

Non-probability data as covariates

- Relevance of statistical information increases with the level of detail, its timeliness and its frequency
- Design-based methods: large variances under small sample sizes
- Model-based inference procedures
  - Small area estimation models
  - Nowcasting models
- Potential way to use new data sources:
  - Potential covariates in SAE and nowcast models
  - High frequency to produce more timely nowcasts
  - Manageable risk level: primary data are survey data

Non-probability data as covariates

- Relevant literature
  - Marchetti et al. (2015): mobility of cars tracked with GPS as a covariate for poverty in an Area Level Model
  - Blumenstock et al. (2015): mobile phone data to predict poverty
  - Steel at al. (2017): mobile phone data and satellite images to predict poverty
  - Schmid et al. (2017): mobile phone data for estimating literacy with an Area Level Model
  - ...

Non-probability data as covariates

- Official statistics
  - Repeated surveys
  - Time series models
    - borrow strength over time and space as a form of SAE
    - combine with auxiliary series derived from big data sources
      - further improves precision
      - high frequency: estimation in real time (nowcasting)
- Example: Dutch Labour Force Survey
  - Structural time series model for monthly labour force figures
  - Extensions
    - Claimant counts
    - Google trends
  - Time varying correlations

Application: Time Series Models for the Dutch LFS

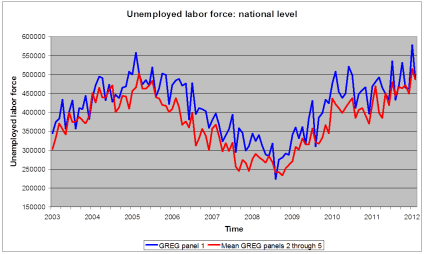
Joint work with Caterina Schiavoni, Stephan Smeekes and Siem Jan Koopman

Dutch Labour Force Survey (LFS)

- Rotating panel design
- Each month: stratified two-stage sample of addresses
- Sample observed 5 times at quarterly intervals
- Data collection
  - First wave: CAWI, CATI CAPI
  - Follow-up waves CATI
- Inference: general regression (GREG) estimator

LFS: Problems

- Sample size:
  - Too small for monthly figures with GREG estimator
  - Rolling quarterly figures as an alternative
- Rotation group bias (RGB)
- Discontinuities due to survey redesigns



Figuur: RGB unemployed labour force

LFS: Solution

- Multivariate state space model (Pfeffermann, 1991)
  - Small area estimation to borrow strength over time and space
  - Account for RGB
  - Account for serial correlation due to panel overlap
  - Modelling shocks due to survey redesigns
  - Extend the model with auxiliary series

LFS: Multivariate state space model

- Rotation scheme: data collected in 5 independent samples
- $\hat{y}_t^{(j)}$ GREG estimate month t, based on the panel that is observed for the j-th time
  - $\hat{y}_t^{(1)}$: sample entered the panel in t, observed for first time
  - $\hat{y}_t^{(2)}$: sample entered the panel in $t-3$, observed for second time
  - $\hat{y}_t^{(3)}$: sample entered the panel in $t-6$, observed for third time
  - $\hat{y}_t^{(4)}$: sample entered the panel in $t-9$, observed for fourth time
  - $\hat{y}_t^{(5)}$: sample entered the panel in $t-12$, observed for fifth (and last) time

LFS: Multivariate state space model

Pfeffermann (1991)

$$\begin{pmatrix} \hat{y}_t^{(1)} \\ \hat{y}_t^{(2)} \\ \hat{y}_t^{(3)} \\ \hat{y}_t^{(4)} \\ \hat{y}_t^{(5)} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \theta_t + \begin{pmatrix} 0 \\ \lambda_t^{(2)} \\ \lambda_t^{(3)} \\ \lambda_t^{(4)} \\ \lambda_t^{(5)} \end{pmatrix} + \begin{pmatrix} e_t^{(1)} \\ e_t^{(2)} \\ e_t^{(3)} \\ e_t^{(4)} \\ e_t^{(5)} \end{pmatrix} + ... \Leftrightarrow \hat{\boldsymbol{y}}_t = \mathbf{1}_{[5]}\theta_t + \boldsymbol{\lambda}_t + \mathbf{e}_t$$

Population parameter: $\theta_t = L_t + S_t + I_t$

- L_t: trend-cycle modelled with smooth trend model

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1} \\ R_t &= R_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_{\eta_y}^2) \end{aligned}$$

- S_t: trigonometric seasonal component
- I_t: population white noise $I_t \sim N(0, \sigma_{I_y}^2)$

LFS: Multivariate state space model

Pfeffermann (1991)

$$\begin{pmatrix} \hat{y}_t^{(1)} \\ \hat{y}_t^{(2)} \\ \hat{y}_t^{(3)} \\ \hat{y}_t^{(4)} \\ \hat{y}_t^{(5)} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \theta_t + \begin{pmatrix} 0 \\ \lambda_t^{(2)} \\ \lambda_t^{(3)} \\ \lambda_t^{(4)} \\ \lambda_t^{(5)} \end{pmatrix} + \begin{pmatrix} e_t^{(1)} \\ e_t^{(2)} \\ e_t^{(3)} \\ e_t^{(4)} \\ e_t^{(5)} \end{pmatrix} + \ldots \Leftrightarrow \hat{\boldsymbol{y}}_t = \mathbf{1}_{[5]}\theta_t + \boldsymbol{\lambda}_t + \mathbf{e}_t$$

RGB: $\boldsymbol{\lambda}_t$

- First wave unbiased
- Difference follow-up waves and first wave: $\lambda_t^{(j)} = \lambda_{t-1}^{(j)} + \mu_t^{(j)}, \;\; \mu_t^{(j)} \sim N(0, \sigma_\mu^2)$

LFS: Multivariate state space model

Pfeffermann (1991)

$$\begin{pmatrix} \hat{y}_t^{(1)} \\ \hat{y}_t^{(2)} \\ \hat{y}_t^{(3)} \\ \hat{y}_t^{(4)} \\ \hat{y}_t^{(5)} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \theta_t + \begin{pmatrix} 0 \\ \lambda_t^{(2)} \\ \lambda_t^{(3)} \\ \lambda_t^{(4)} \\ \lambda_t^{(5)} \end{pmatrix} + \begin{pmatrix} e_t^{(1)} \\ e_t^{(2)} \\ e_t^{(3)} \\ e_t^{(4)} \\ e_t^{(5)} \end{pmatrix} + ... \Leftrightarrow \hat{\boldsymbol{y}}_t = \mathbf{1}_{[5]}\theta_t + \boldsymbol{\lambda}_t + \mathbf{e}_t$$

Sampling error: $\mathbf{e}_t$

- Heteroscedasticity: $e_t^{(j)} = \sqrt{var(\hat{y}_t^{(j)})}\tilde{e}_t^{(j)}$
- First wave $\tilde{e}_t^{(j)} \sim N(0, \sigma^2_{e_1})$
- Serial correlation follow-up waves: $\tilde{e}_t^{(j)} = \rho\tilde{e}_{t-3}^{(j-1)} + \varepsilon_t^{(j)}, \quad \varepsilon_t^{(j)} \sim N(0, \sigma^2_{e_j})$

LFS: Multivariate state space model

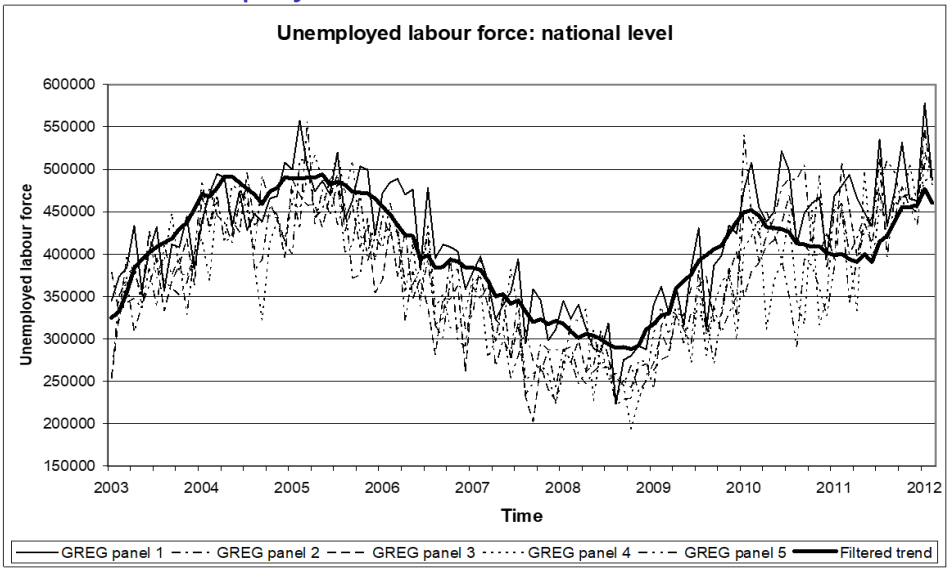
- Models in state space form:
  - $\mathbf{y}_t = \mathbf{Z}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{H})$
  - $\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\xi}_t, \qquad \boldsymbol{\xi}_t \sim N(\mathbf{0}, \boldsymbol{\Omega})$
    - $\boldsymbol{\alpha}_t = (L_t, R_t, S_t^1, ..., \lambda_t^{(2)}, ...)'$
- Kalman filter: optimal estimates for $\boldsymbol{\alpha}_t$
- Software: OxMetrics and Ssfpack

LFS: Multivariate state space model

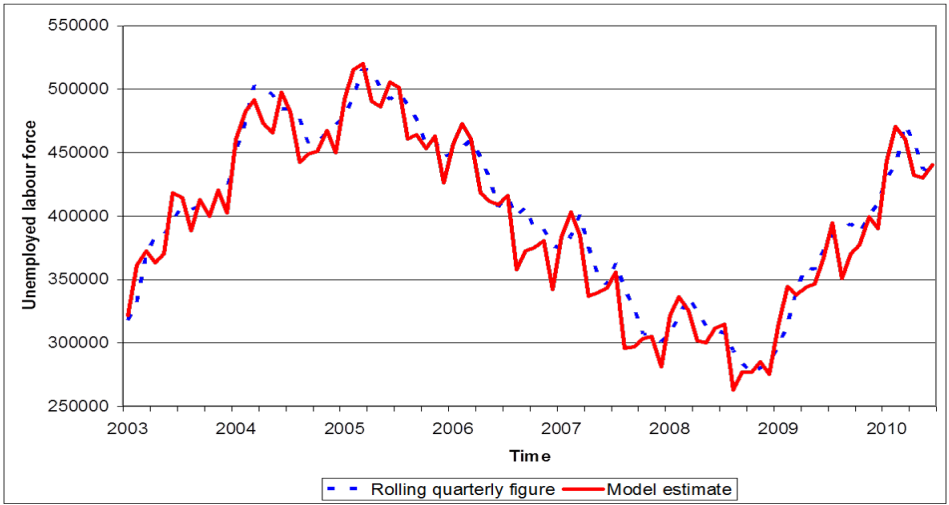
Official monthly labour force figures:

- Unemployed labour force
- Employed labour force
- Total labour force
- Publication levels:
  - National level
  - Breakdown in 6 domains $gender[2] \times age[3]$
- Unemployment rate: ratio of model predictions unemployed and total labour force

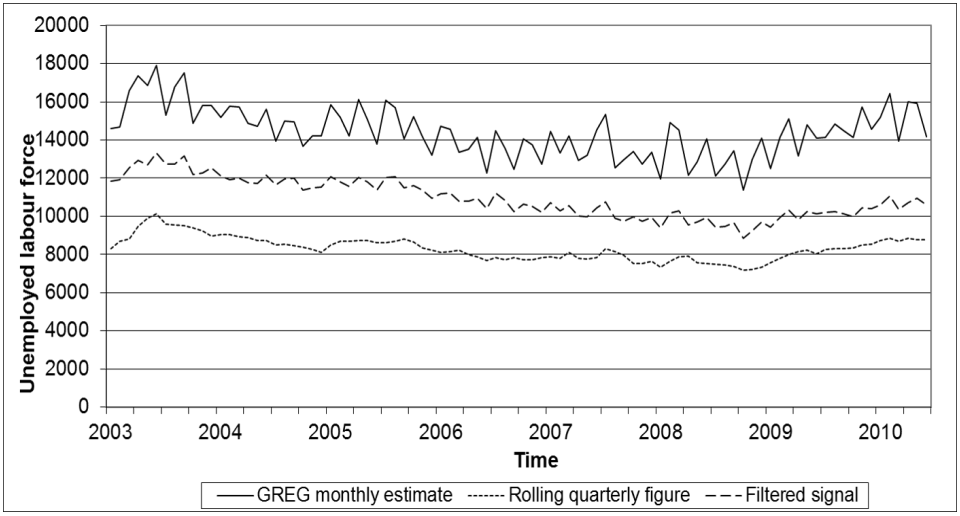
Unemployed labour force national level: Trend



Unemployed labour force national level: Trend+Seasonal



St. Error Unemployed labour force national level



LFS: Multivariate state space model

Possible extensions:

- Combine the 6 domains in a 30 dimensional state space model
  - Model cross-sectional and temporal relations
- Combine the LFS series with auxiliary series
  - Claimant counts
  - Google trends

LFS: Multivariate state space model for 6 domains

$$\begin{pmatrix} \hat{\boldsymbol{y}}_{t,1} \\ \hat{\boldsymbol{y}}_{t,2} \\ \hat{\boldsymbol{y}}_{t,3} \\ \hat{\boldsymbol{y}}_{t,4} \\ \hat{\boldsymbol{y}}_{t,5} \\ \hat{\boldsymbol{y}}_{t,6} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{[5]}\theta_{t,1} \\ \mathbf{1}_{[5]}\theta_{t,2} \\ \mathbf{1}_{[5]}\theta_{t,3} \\ \mathbf{1}_{[5]}\theta_{t,4} \\ \mathbf{1}_{[5]}\theta_{t,5} \\ \mathbf{1}_{[5]}\theta_{t,6} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_{t,1} \\ \boldsymbol{\lambda}_{t,2} \\ \boldsymbol{\lambda}_{t,3} \\ \boldsymbol{\lambda}_{t,4} \\ \boldsymbol{\lambda}_{t,5} \\ \boldsymbol{\lambda}_{t,6} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_{t,1} \\ \mathbf{e}_{t,2} \\ \mathbf{e}_{t,3} \\ \mathbf{e}_{t,4} \\ \mathbf{e}_{t,5} \\ \mathbf{e}_{t,6} \end{pmatrix}$$

$$\theta_{t,d} = L_{t,d} + S_{t,d} + I_{t,d} \quad d = 1, \ldots, 6.$$

Correlated trends:

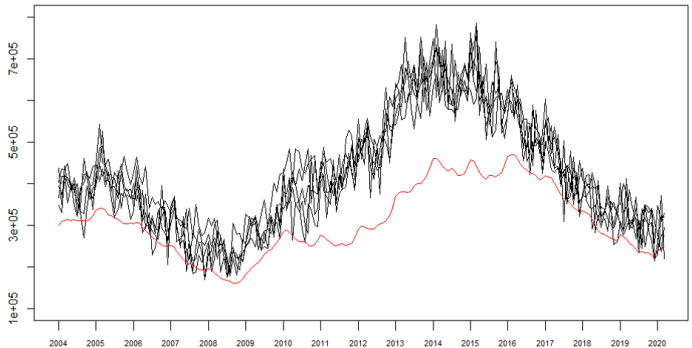
- $L_{t,d} = L_{t-1,d} + R_{t-1,d}$ and $R_{t,d} = R_{t-1,d} + \eta_{t,d}$
- $\boldsymbol{\eta}_t \sim N(\mathbf{0}_{[6]}, \boldsymbol{\Sigma})$
  - $\boldsymbol{\eta}_t = (\eta_{t,1}, \eta_{t,2}, \ldots, \eta_{t,6})'$
  - $\boldsymbol{\Sigma}$: 6×6 full covariance matrix

LFS: Multivariate state space model for 6 domains

Correlations trend unemployed labour force

	M 15-24	W 15-24	M 25-44	W 25-44	M 45-64	W 45-64
M 15-24	1					
W 15-24	0.76	1				
M 25-44	0.93	0.94	1			
W 25-44	0.65	0.99	0.88	1		
M 45-64	0.47	0.93	0.75	0.98	1	
W 45-64	0.10	0.70	0.41	0.80	0.81	1

LFS: Multivariate state space model with claimant counts



Black: input series LFS

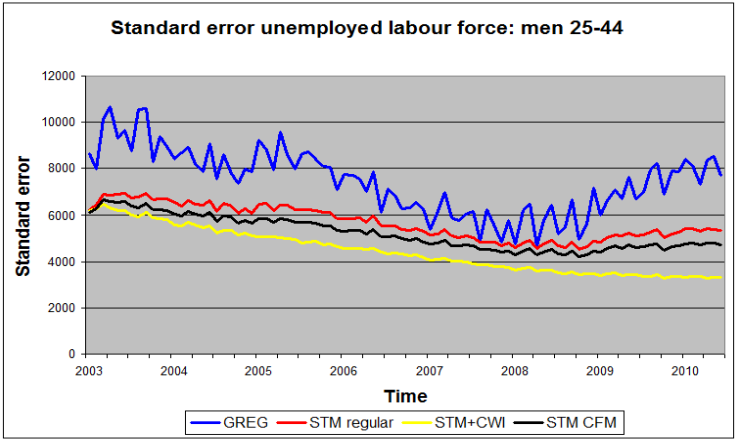
Red: claimant counts

LFS: Multivariate state space model with claimant counts

$$
\begin{aligned}
\begin{pmatrix} \hat{\boldsymbol{y}}_t \\ x_t \end{pmatrix} &= \begin{pmatrix} \mathbf{1}_{[5]}\theta_t^y \\ \theta_{t,x} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ 0 \end{pmatrix} \\
\theta_t^y &= L_t^y + S_t^y + I_t^y \\
\theta_t^x &= L_t^x + S_t^x + I_t^x
\end{aligned}
$$

- Trend target series: $L_t^y = L_{t-1}^y + R_{t-1}^y$ and $R_t^y = R_{t-1}^y + \eta_t^y$
- Trend auxiliary series: $L_t^x = L_{t-1}^x + R_{t-1}^x$ and $R_t^x = R_{t-1}^x + \eta_t^x$
- Correlation between slope disturbance terms:
  $\begin{pmatrix} \eta_t^y \\ \eta_t^x \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\eta_y}^2 & \rho\sigma_{\eta_y}\sigma_{\eta_x} \\ \rho\sigma_{\eta_y}\sigma_{\eta_x} & \sigma_{\eta_x}^2 \end{pmatrix}\right)$, $\hat{\rho} > 0.9$

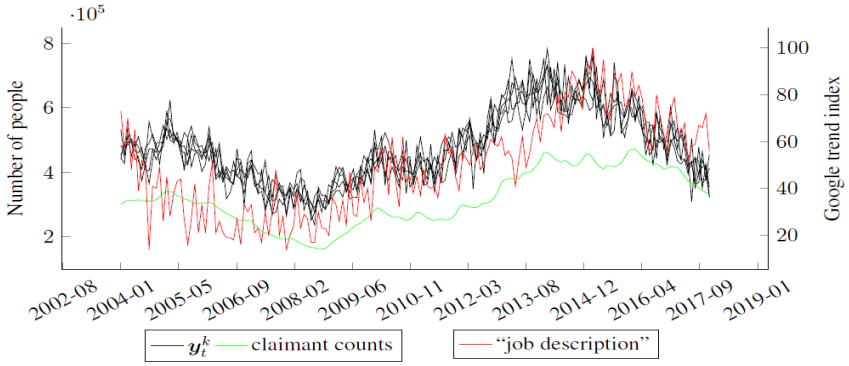
LFS: Comparison multivariate state space models



LFS: Dynamic factor models

Further extension model with claimant counts:

- Google trends search terms related to LFS figures



LFS: Dynamic factor models

- Google trends:
  - High frequency
  - Potentially improve accuracy and timeliness
- Issue:
  - Large amount of potential series: 80 series in this application
  - High dimensionality problems
- Solution:
  - Dynamic factor model fitted with a two-step estimator
  - Giannone et al. (2008), Doz et al. (2011)

Dynamic factor model: Estimation step 1

- Estimate common factors in Google trends (GT)

$$\mathbf{x}_t^{gt} = \mathbf{\Lambda}\mathbf{f}_t + \boldsymbol{\epsilon}_t \quad Var(\boldsymbol{\epsilon}_t) = \mathbf{\Psi}$$
$$\mathbf{f}_t = \mathbf{f}_{t-1} + \boldsymbol{\mu}_t$$

  - $\mathbf{x}_t^{gt}$: n vector with auxiliary series
  - $\mathbf{f}_t$: r vector with common factors $r << n$ assumed to be I(1)
  - $\mathbf{\Lambda}$: $n \times r$ matrix with factor loadings
  - ϵ_t: n vector with idiosyncratic components of $\mathbf{x}_t^{gt}$
  - for identifiability reasons: $E(\boldsymbol{\mu}_t \boldsymbol{\mu}_t^{'}) = \mathbf{I}_{[r]}$
- $\mathbf{f}_t$, $\mathbf{\Lambda}$, $\mathbf{\Psi}$ are estimated with Principal Component Analysis applied to the weekly data of GT

Dynamic factor model: Estimation step 2

- State space model entire data set

$$\begin{pmatrix} \hat{\mathbf{y}}_t \\ x_t^{cc} \\ \mathbf{x}_t^{gt} \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{[5]}(L_t^y + S_t^y) \\ L_t^{cc} + S_t^{cc} \\ \hat{\mathbf{\Lambda}}\mathbf{f}_t \end{pmatrix} + \begin{pmatrix} \mathbf{1}_{[5]} I_t^y \\ I_t^x \\ \epsilon_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ 0 \\ 0 \end{pmatrix}$$

$$L_t^z = L_{t-1}^z + R_{t-1}^z \quad R_t^z = R_{t-1}^z + \eta_t^z \quad z = (y, cc)$$

$$\mathbf{f}_t = \mathbf{f}_{t-1} + \boldsymbol{\mu}_t$$

$$Cov \begin{pmatrix} \eta_t^y \\ \eta_t^{cc} \\ \boldsymbol{\mu}_t \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \sigma_{y,cc} & \sigma_{y,f_1} & \dots \\ \sigma_{y,cc} & \sigma_{cc}^2 & 0 & \dots \\ \sigma_{y,f_1} & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & 1 \end{pmatrix}$$

- $\hat{\mathbf{\Lambda}}$, $\hat{\mathbf{\Psi}}$ obtained in step 1 are kept fixed
- $\mathbf{f}_t$ are re-estimated with the Kalman filter

Dynamic factor models: results

Models:

- Production model: no auxiliary series
- Production model + claimant counts
- Production model + google trends
- Production model + claimant counts + google trends

Dynamic factor models: results

- Period January 2004 until December 2017 (168 months)
- Estimation accuracy: MSE filtered estimates over last 56 months
- Nowcast accuracy
  - MSFE (one-step-ahead prediction error) over last 56 months
  - nowcast for t: LFS and CC missing, only GT available
  - Hyperparameter estimates based available information in t

Dynamic factor models: results

- Number of common factors for Google trends: 2
- Correlations trend disturbance terms:

Model	$\hat{\rho}_{1,gt}$ (p-value)	$\hat{\rho}_{2,gt}$ (p-value)	$\hat{\rho}_{cc}$ (p-value)
CC			0.92 (0.002)
GT	-0.785 (0.000)	-0.591 (0.014)	
GT+CC	-0.381 (0.082)	-0.456 (0.015)	0.80 (0.001)

p-value: LR test $H_0 : \rho_x = 0$

Dynamic factor models: results

Results trend L_t^y relative to production model

	model		
	CC	GT	CC+GT
$\widehat{MSE}(L_t^y)$	0.869	0.861	0.849
$\widehat{MSFE}(L_t^y)$	0.815	0.935	0.889

Results signal $\theta_t^y = L_t^y + S_t^y$ relative to production model

	model		
	CC	GT	CC+GT
$\widehat{MSE}(\theta_t^{[y]})$	0.889	0.899	0.881
$\widehat{MSFE}(\theta_t^{[y]})$	0.827	0.942	0.899

Dynamic factor models: results

- Claimant counts
  - Strong correlation with LFS
  - Improves estimates but also nowcasts
  - Relative simple model
- Google trends
  - 2 common factors have strong correlation with LFS
  - Improves estimates but also nowcasts
  - Complex model: worthwhile the effort?
  - Useful for countries without register of claimant counts

State space model with claimant counts

$$
\begin{aligned}
\begin{pmatrix} \hat{\boldsymbol{y}}_t \\ x_t \end{pmatrix} &= \begin{pmatrix} \mathbf{1}_{[5]}\theta_t^y \\ \theta_{t,x} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ 0 \end{pmatrix} \\
\theta_t^y &= L_t^y + S_t^y + I_t^y \\
\theta_t^x &= L_t^x + S_t^x + I_t^x
\end{aligned}
$$

- Trend: $L_t^z = L_{t-1}^z + R_{t-1}^z$ and $R_t^z = R_{t-1}^z + \eta_t^z$ for $z = (y, x)$
- Correlation between slope disturbance terms:

$$
\begin{pmatrix} \eta_t^y \\ \eta_t^x \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\eta y}^2 & \rho\sigma_{\eta y}\sigma_{\eta x} \\ \rho\sigma_{\eta y}\sigma_{\eta x} & \sigma_{\eta x}^2 \end{pmatrix} \right)
$$

- Strong assumption: ρ is time invariant

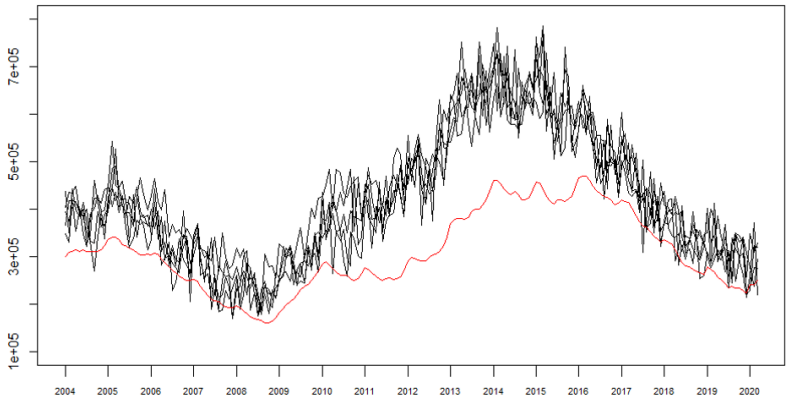
State space model with claimant counts

Relationship between LFS and claimant counts change over time:

- Legislative changes:
  - February 2015: people that find a job keep unemployment benefits for 2 additional months
- Global Financial Crisis 2008:
  - Unemployment benefits for a maximum of 2 years
  - Long-term unemployment as a result of a crisis is not picked up by claimant count series

State space model with claimant counts

Relationship between LFS and claimant counts change over time:



Time varying state correlations

State space model with time varying state correlations:

$$\begin{pmatrix} \hat{\boldsymbol{y}}_t \\ x_t \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{[5]}\theta_t^y \\ \theta_{t,x} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ 0 \end{pmatrix}$$
$$\theta_t^y = L_t^y + S_t^y + I_t^y$$
$$\theta_t^x = L_t^x + S_t^x + I_t^x$$

Trend

- $L_t^z = L_{t-1}^z + R_{t-1}^z$ and $R_t^z = R_{t-1}^z + \eta_t^z$ for $z = (y, x)$
- $\begin{pmatrix} \eta_t^y \\ \eta_t^x \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\eta_y}^2 & \rho_t \sigma_{\eta_y}\sigma_{\eta_x} \\ \rho_t \sigma_{\eta_y}\sigma_{\eta_x} & \sigma_{\eta_x}^2 \end{pmatrix}\right)$

Time varying state correlations

Two specifications for ρ_t:

- Deterministic specification based on cubic splines:
  - State space model remain linear
  - Estimation proceeds with Kalman Filter
- Stochastic specification
  - State space model becomes non-linear
  - Estimation proceeds with indirect inference using cubic splines as an auxiliary model followed by a particle filter that is based on the Rao-Blackwellised bootstrap filter (RBBF)

Time varying state correlations

Deterministic specification based on cubic splines:

- Transformation: $\gamma_t = \tanh(\rho_t)$
- Cubic splines: $\gamma_t = \mathbf{w}_t'\phi$
  - $\mathbf{w}_t$: $k \times 1$ vector with weights
  - ϕ: $k \times 1$ vector with coefficients estimated with maximum likelihood (like the other hyperparameters)
- Advantage
  - γ_t known/not random
  - Computationally fast
- Disadvantage
  - Prior selection of the knots
  - Large uncertainty of the splines at the beginning and the end of the series

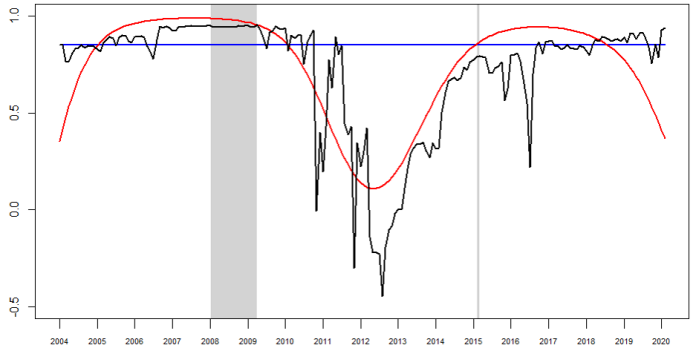
Time varying state correlations

Stochastic specification:

- Model: $\gamma_t = \gamma_{t-1} + \zeta_t, \quad \zeta_t \sim N(0, \sigma_\zeta^2)$
- γ_t: treated as an additional state variable
- makes the state space model non-linear
  - ME: $\mathbf{y}_t = \mathbf{Z}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t$, with $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{H})$
  - TE: $\boldsymbol{\alpha}_t = \mathbf{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\xi}_t$, with $\boldsymbol{\xi}_t \sim N(\mathbf{0}, \boldsymbol{\Omega}_t)$
  - Kalman filter not applicable
- Proposed solution:
  - Estimate σ_ζ^2 by indirect inference with cubic spline model as an auxiliary model (Gourieroux et al. 1993)
  - Estimate the state variables with a particle filter (Rao-Blackwellised Bootstrap filter)
  - Details: Schiavoni, Koopman, Palm, Smeekes and van den Brakel (2021)

Application to unemployed labour force

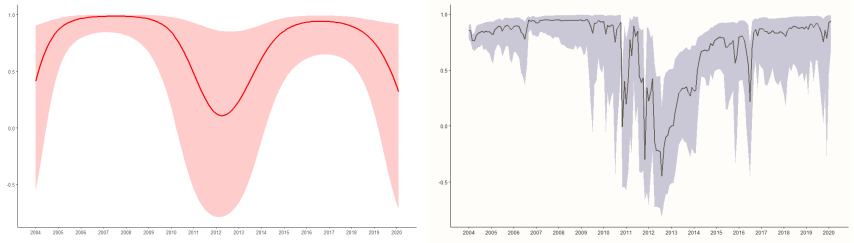
Estimated correlation



constant in blue, with splines in red, with RBBF in black

Application to unemployed labour force

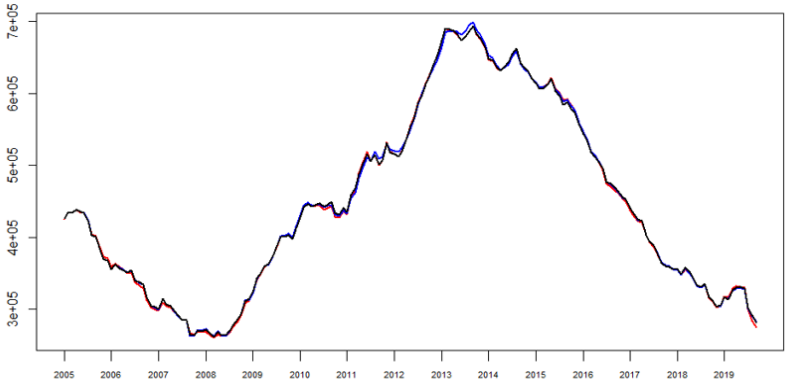
Correlation estimates based on T with 95% confidence intervals



Left: cubic splines, right: RBBF

Application to unemployed labour force

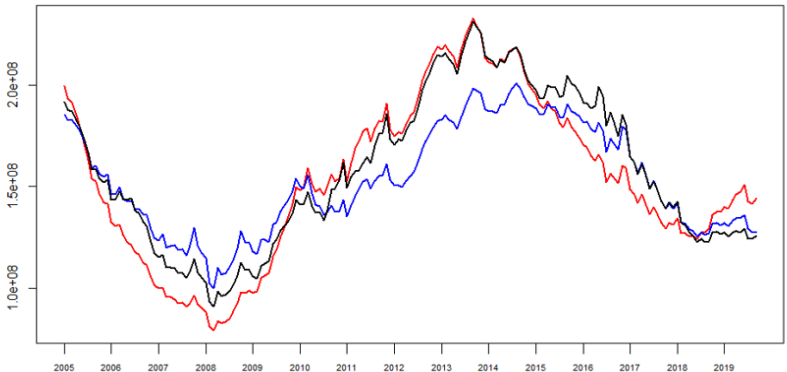
Filtered trend (L_t^y)



constant in blue, with splines in red, with RBBF in black

Application to unemployed labour force

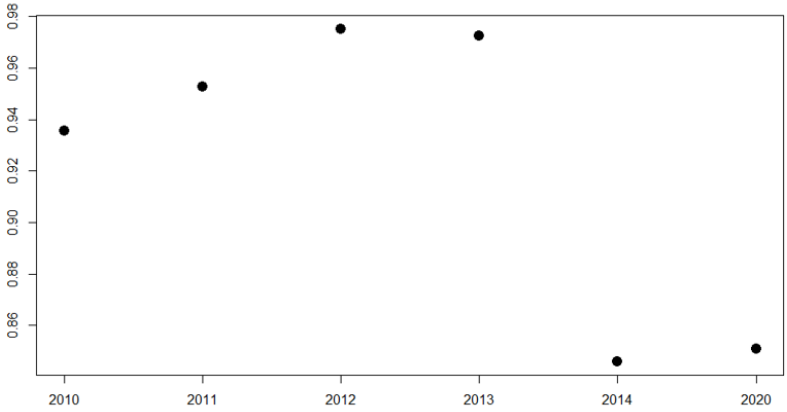
Variance filtered trend



constant in blue, with splines in red, with RBBF in black

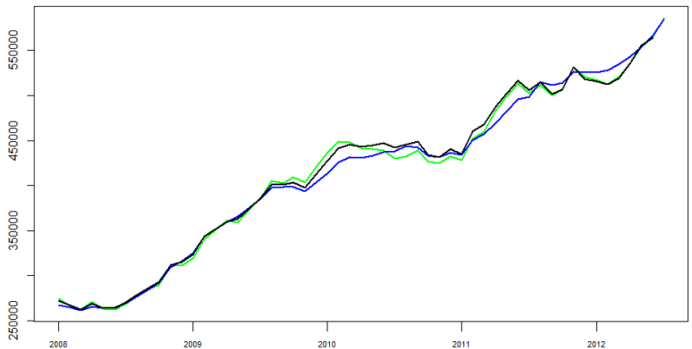
Application to unemployed labour force

Constant correlation estimates in real time



Application to unemployed labour force

Trend in real time (until December 2012)



constant in blue, with RBBF in black, model without claimant counts

Application: Predicting poverty from aerial images

Joint work with Joep Burger and Harm Jan Boonstra

Predicting poverty from aerial images

Refined regional estimates for poverty

- Literature to predict poverty from satellite or aerial images
- Machine learning (ML) algorithms trained using sample data
- Predictions are based on ML only
- Popular for developing countries and combat areas
- Overview: Anderson et al. 2017
- Suboptimal to base predictions on the ML algorithm only

Predicting poverty from aerial images

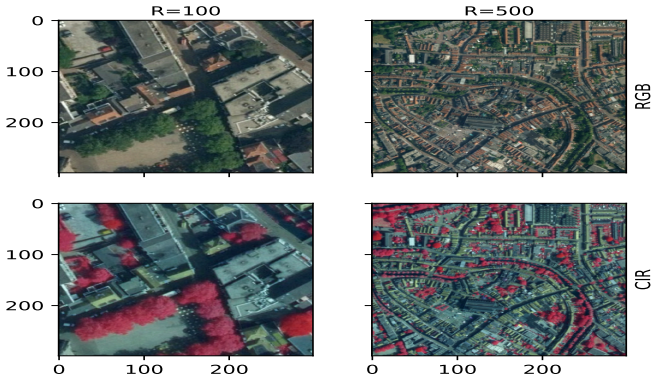
Netherlands

- Income and poverty statistics: registers
- Unique data set to test this approach
- Purpose this project:
  - Deep learning algorithm to predict poverty
  - Simulations with sample designs
  - Illustrate how to use in SAE models
  - Compare outcomes with the truth

Predicting poverty from aerial images

Images

- 2 spectral bands: RGB and near infrared (CIR)
- 2 spatial scales: 100m*100m and 500m*500m



Predicting poverty from aerial images

Data

- Disposable household income all households
- Households are geocoded
- Annotate all images:
  - Inhabitation
  - Indicator for poverty
  - Poverty rate

Predicting poverty from aerial images

Methods

- Training and prediction: convolutional neural network
- Architecture: Interception V3 (48 layers deep)
- Minimize negative log lik. of a binomial distribution
- Performance: relative loss and RMSE with respect to random guessing (latter is based on the fraction of the label in the training set)
- Training and prediction on a 32 GB Nvidea Tesla V100 GPU

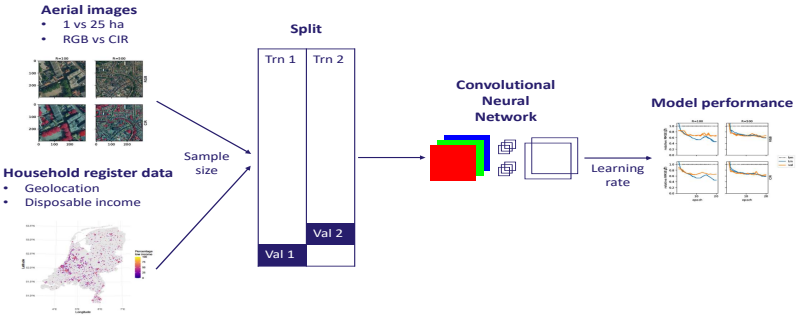
Predicting poverty from aerial images

Methods

- Difficulty: poverty is not an object at an image
- Training and prediction of the following tasks:
  - inhabitation
  - existence of poverty given inhabitation
  - poverty rate given there is poverty

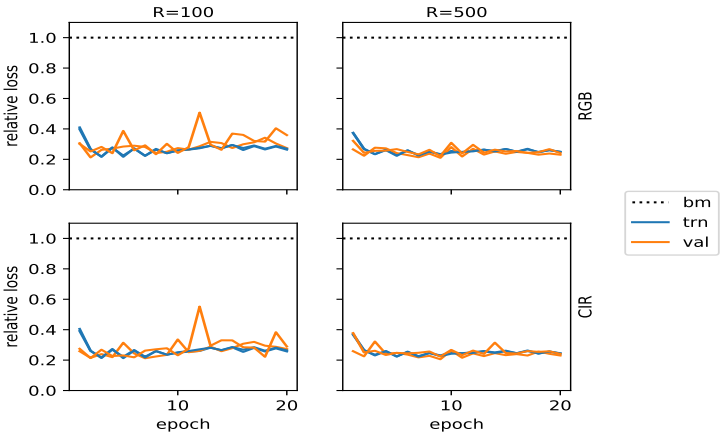
Predicting poverty from aerial images

Process flow:



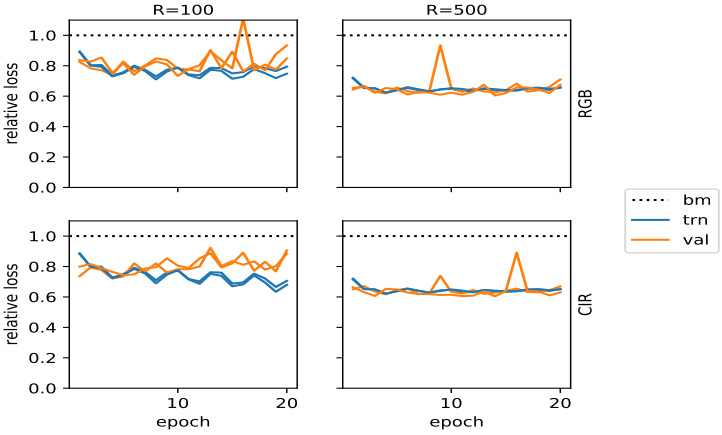
Predicting poverty from aerial images

Results for inhabitation



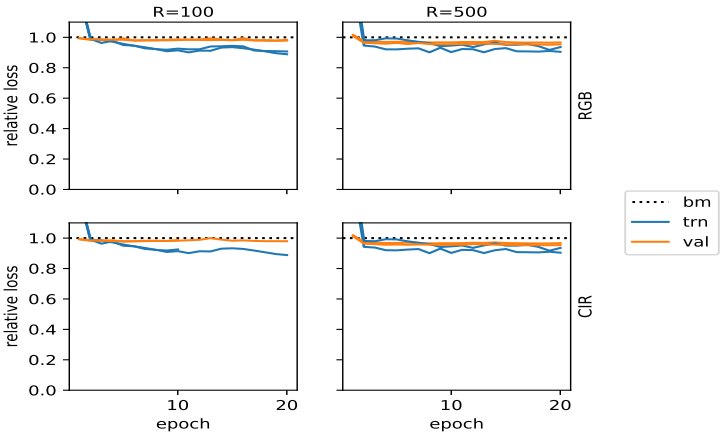
Predicting poverty from aerial images

Results for poverty given inhabitation



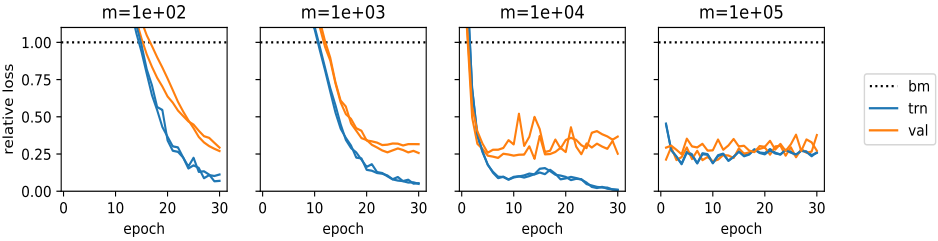
Predicting poverty from aerial images

Results for poverty rate given poverty



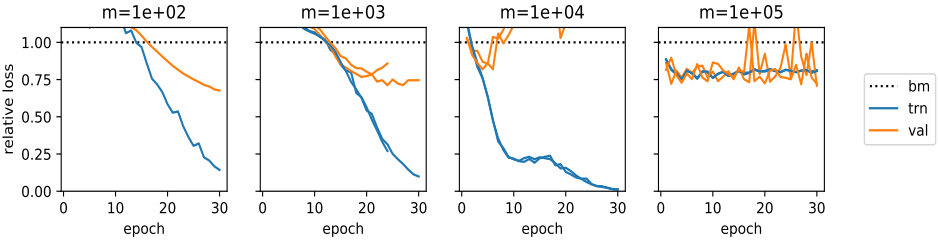
Predicting poverty from aerial images

Results different sample sizes for predicting inhabitation



Predicting poverty from aerial images

Results different sample sizes for predicting poverty given inhabitation



Predicting poverty from aerial images

First results

- Good results for predicting inhabitation
- Reasonable results for predicting poverty given inhabitation
- Predicting poverty rate is difficult
- Best results for 500m*500m images
- No difference between RGB and CIR
- Predicting inhabitation and poverty also possible with small samples
- Details: Burger et al. (2024)

Predicting poverty from aerial images

Ongoing work

- Simulations with different sample designs
- Two stage sampling:
  - m out of M images
  - for sampled image i: n_i out of N_i households
  - $\hat{y}_i$ direct estimate poverty
- Training a CNN
- $\tilde{y}_i$: prediction poverty for M images
- Estimation:
  - Prediction based on images only
  - SAE: $\hat{y}_i = \beta' \tilde{y}_i + \nu_i + e_i$
- Benchmark: true values y_i from tax registers

Predicting poverty from aerial images

Final remarks

- Purpose
  - Evaluate this approach that is frequently found in literature
  - Unique data set in the Netherlands
  - Simulation with different sample designs
  - Illustrate the benefits of combining sample data with ML predictions
- It is understood
  - Not to predict poverty in the Netherlands
  - CNN not directly applicable in other countries due to differences in social and urban structures

Machine learning (ML) to improve the sampling strategy

Joint work with Jonas Klingwort, Kees van Berkel and Piet Daas

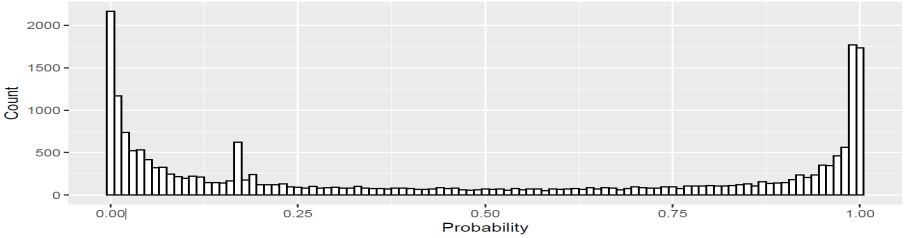
ML to improve the sampling strategy

Community and innovation survey (CIS)

- Publish information about innovative companies
- Bi-annual survey
- Stratified sample among 10,000 companies
- Daas and van der Doef (2020):
  - ML algorithm to predict innovative companies
  - Input text scraped websites
  - CIS: used to annotate web scraped texts
  - ML applied to classify companies based on web texts

ML to improve the sampling strategy

Predicted probabilities that companies have innovation activities
Source: Daas and van der Doef (2020)



ML to improve the sampling strategy

How to use this information:

- Primary data source to predict number of innovative companies
  - Cost-effective way of data collection
  - High risk appetite to extrapolate to future periods
  - Future editions: survey data are required to update the ML algorithm
  - Sub optimal not to use survey data

ML to improve the sampling strategy

How to use this information (cont.):

- Auxiliary information in a model based inference approach
  - SAE models to predict at refined regional level
  - $\hat{y}_i^{CIS} = \beta' \tilde{y}_i^{ML} + \nu_i + e_i$
    - $\hat{y}_i^{CIS}$ sample estimate based on the CIS for region *i*
    - $\tilde{y}_i^{ML}$ prediction based on ML for region *i*
  - Similar to Marchetti et al. (2015)
  - Lower risk appetite since primary data are survey data
  - Inference still model-based

ML to improve the sampling strategy

How to use this information (cont.):

- Auxiliary information to optimize sampling strategy of the CIS in a design based inference approach
  - Improve the sample design
    - PPS using predicted probabilities that companies have innovative activities
    - Stratification based on predicted innovation probabilities
  - Improve weighting scheme GREG estimator
  - Lowest risk appetite since survey data are the primary data with design-based inference approach

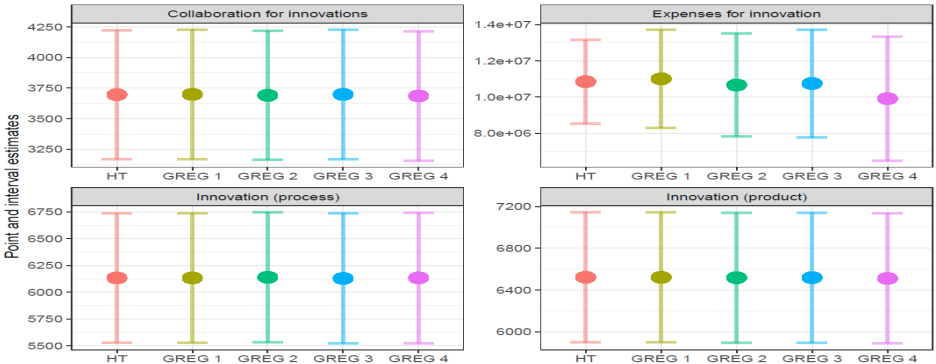
ML to improve the sampling strategy

Example: GREG estimator with probabilities of innovative activities

- HT estimator
- GREG1: probability
- GREG2: probability + number of web sites
- GREG3: probability + language website
- GREG4: probability + number of websites + language website

ML to improve the sampling strategy

Estimates four different innovation variables



Discussion

Discussion

Use of new data sources in official statistics

- As a primary source
- Generally a higher risk level
  - No control over availability, stability and consistency of a data source
  - Selection bias
  - Increased risk of relying on model assumptions
  - Requires structured data with strong auxiliary information: opt-in panels

Discussion

Use of new data sources in official statistics

- Covariates in model-based inference procedures
  - Primary data are still designed data
  - Useful to produce more detailed and timely data
  - Time series models versus cross-sectional models
  - Example: Dutch LFS
    - Multivariate state space model for official monthly LFS figures
    - Extensions to use claimant counts and Google trends as auxiliary series
    - Strong assumption: time invariant state correlations
    - Modelling time varying state correlations
  - SAE for poverty using information from aerial images
  - Dynamic factor model for economic growth

Discussion

Use of new data sources in official statistics

- Auxiliary information to improve sampling strategy
  - Traditional design-based approach
  - Control over availability, stability and consistency of a data source

Thank you for your interest!