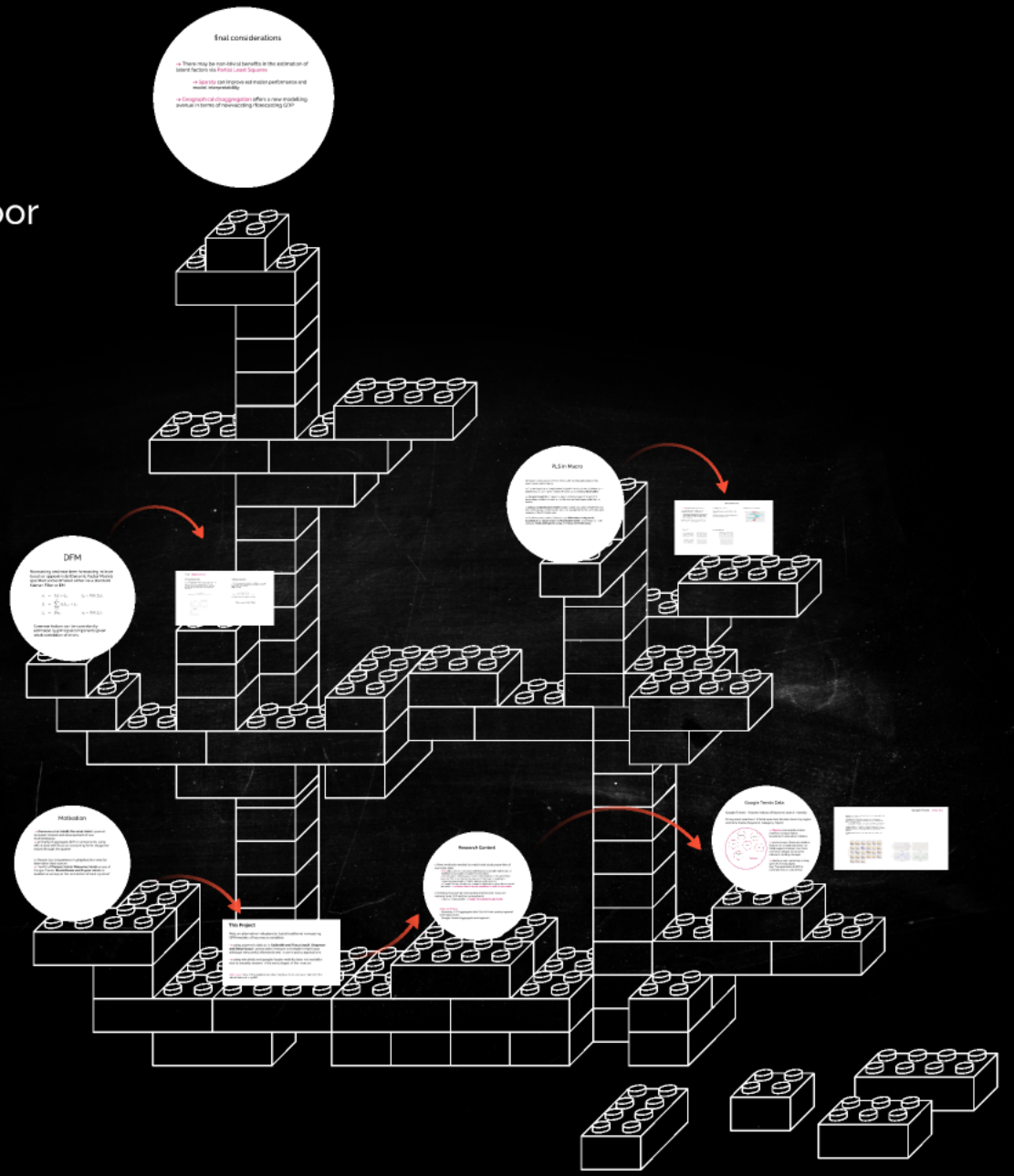
Sparse Warcasting

Forecasting in a data-rich but statistics-poor environment



Mihnea Constantinescu
National Bank of Ukraine and
University of Amsterdam

June 2023
NBU-NBP ARC

Motivation

-> **Giannone et al. (2008); Doz et al. (2011)** spurred renewed interest and development of new methodologies
-> primarily of aggregate GDP or components, using official data with focus on accounting for its staggered nature through the quarter

>> Recent Covid experience highlighted the need for alternative data sources
-> VoxEU of **Diebold (2020); Woloszko (2020)** on use of Google Trends; **Blanchflower and Bryson (2021)** on qualitative surveys or the "economics of walking about"

Motivation

-> **Giannone et al. (2008); Doz et al. (2011)** spurred renewed interest and development of new methodologies
-> primarily of aggregate GDP or components, using official data with focus on accounting for its staggered nature through the quarter

>> Recent Covid experience highlighted the need for alternative data sources
-> VoxEU of **Diebold (2020); Woloszko (2020)** on use of Google Trends; **Blanchflower and Bryson (2021)** on qualitative surveys or the "economics of walking about"

>> The Feb 24th russian invasion led to a freeze of all official data gathering by local and national statistical agencies
--> only left with alternative data sources

This Project

Rely on alternative indicators to build traditional nowcasting DFM models of key macro variables

-> using payments data as in **Galbraith and Tkacz (2018)**, **Chapman and Desai (2021)** : unavailable timespan and depth in April 2022 although very useful afterwards and in use in policy applications

-> using electricity and google/apple mobility data: not available due to security reasons in the early stages of the invasion;

Main issue: many of the available variables may have no, or very weak, relation to the latent factors driving GDP

Research Context

1. New methods needed to match statistcal properties of available data
   -> No official statistics released until mid 2022; use alternative data as identified in development economics literature
   -> Lit considers primarily peace-time GDP estimation: NL and Twitter possibly "flip" signs during the course of the invasion leading to inconsistent parameters; Twitter data recently for a fee
   -> Google Trends not entirely immune to these issue given the nature of the shock **->> different factor model specification and/or estimation**

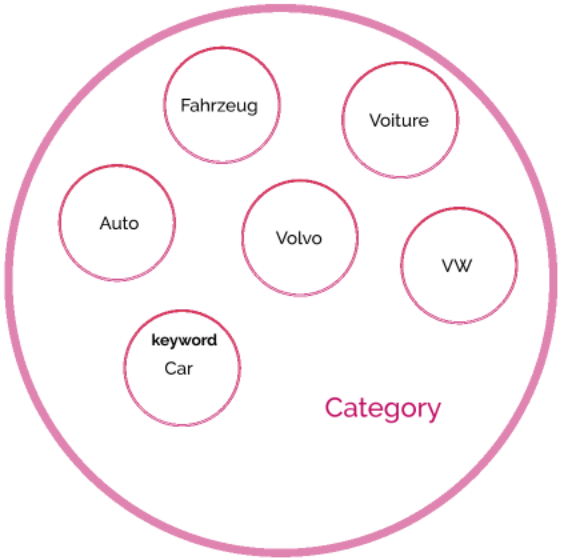
2. Existing forecasting/nowcasting frameworks focus on national level GDP and/or components
   -> Spatial heterogeneity **->> scope for regional factor model**

data-summary:
- Quarterly GDP aggregate data (Q4 2021) and yearly regional GDP data (2020)
- Google Trends (aggregate and regional)

Google Trends Data

Google Trends = Volume indices of keyword search intensity

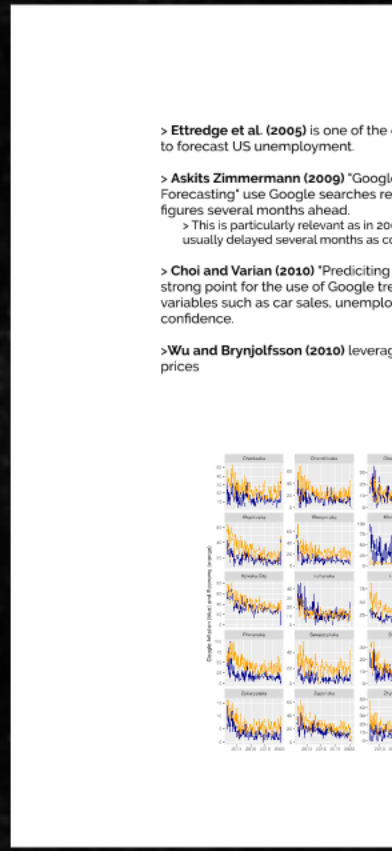
(# keyword searches/ # total searches) broken down by region and time frame {Keyword, Category, Topic}



-> Topics: consumption, labor markets, transportation, investment, education, inflation

-> some issues: these are relative indices (w.r.t. total searches); as total usage increases over time and new categories become relevant, ranking changes.

-> during a war, some topics may give the wrong signal (eg. Transportation & GDP in normal times vs. war times)



Google Trends - empirics

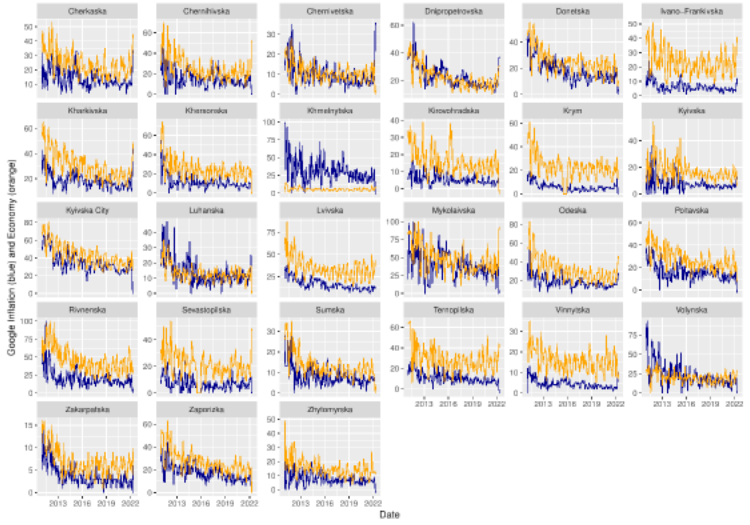
> **Ettredge et al. (2005)** is one of the earlier references using Google Search activity to forecast US unemployment.

> **Askits Zimmermann (2009)** "Google Econometrics and Unemployment Forecasting" use Google searches related to unemployment to forecast offical figures several months ahead.

> This is particularly relevant as in 2008-2009, data releases on key macrovariables are usually delayed several months as compared to observed macro and financial shocks

> **Choi and Varian (2010)** "Prediciting the Present with Google Trends" makes a strong point for the use of Google trends data to nowcast a multitude of economic variables such as car sales, unemployment claims, travel, and consumer confidence.

>**Wu and Brynjolfsson (2010)** leverage Google Search Data to forecast house prices



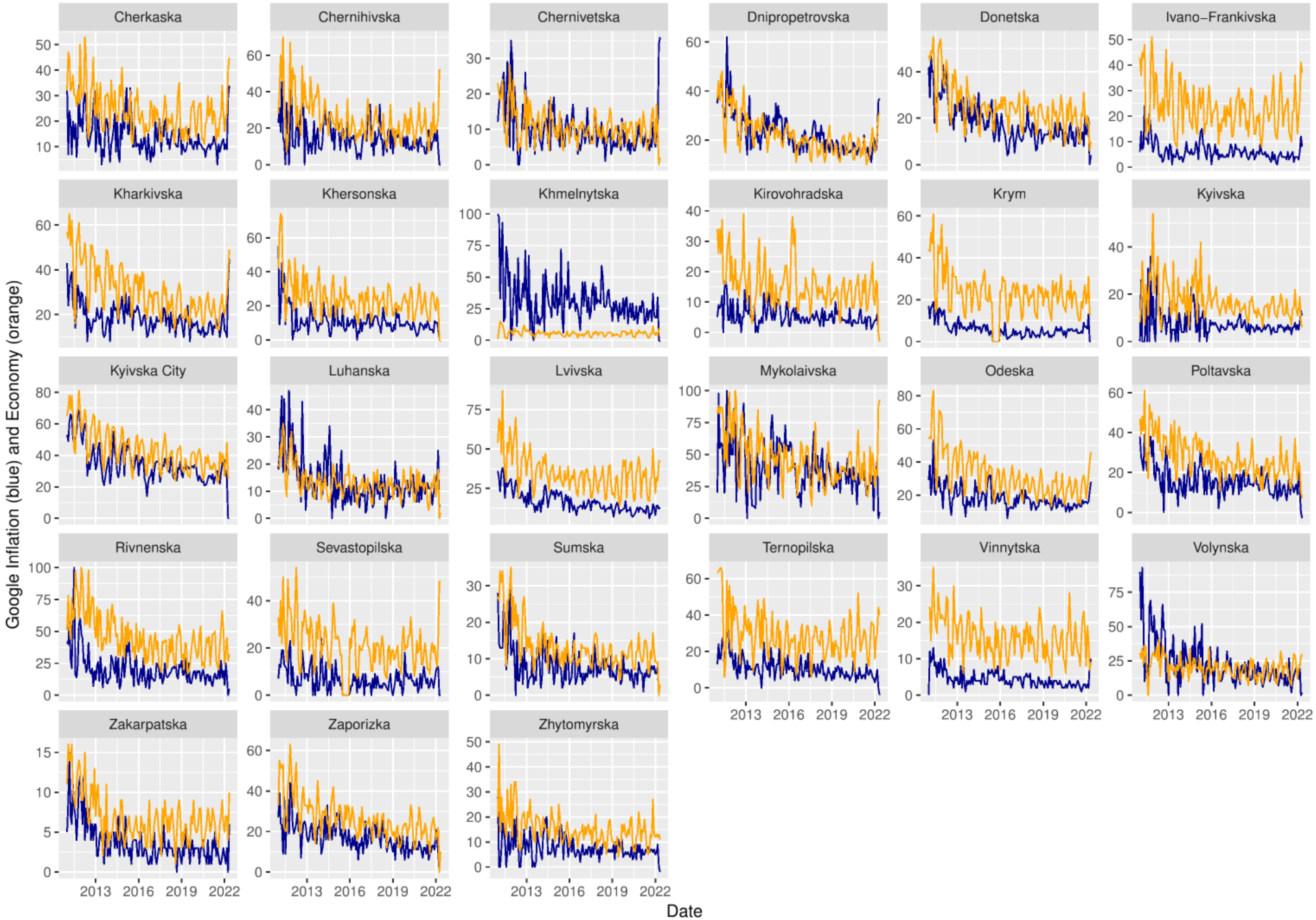
> **Ettredge et al. (2005)** is one of the earlier references using Google Search activity to forecast US unemployment.

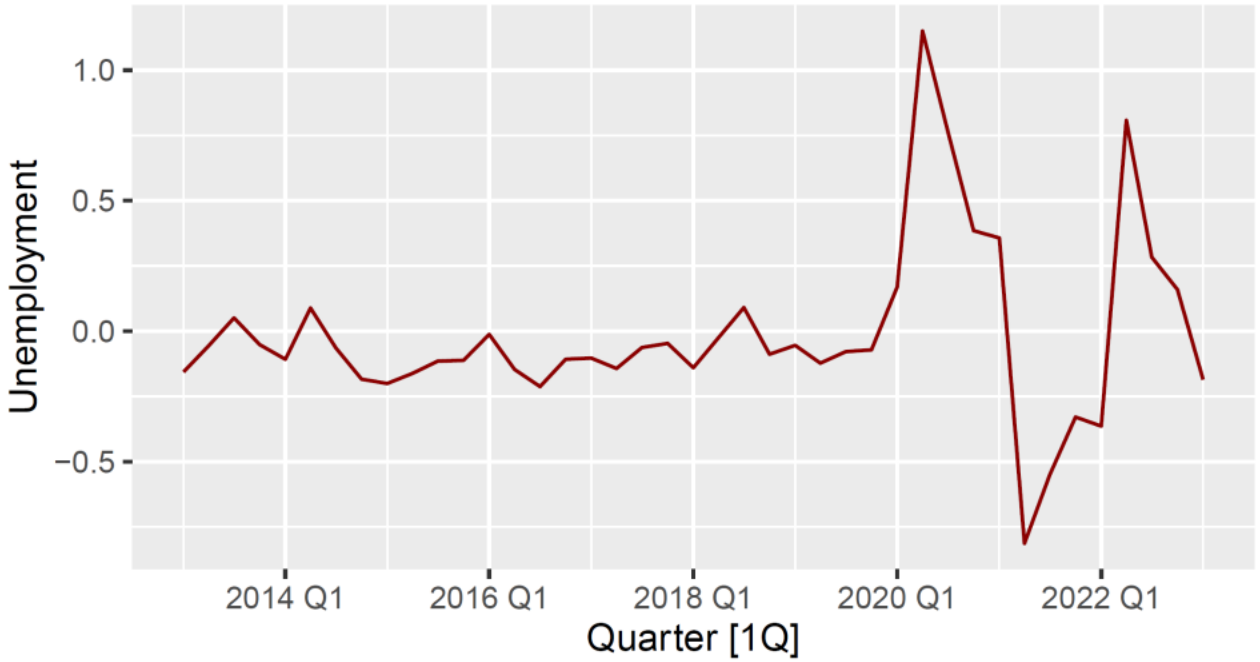
> **Askits Zimmermann (2009)** "Google Econometrics and Unemployment Forecasting" use Google searches related to unemployment to forecast offical figures several months ahead.

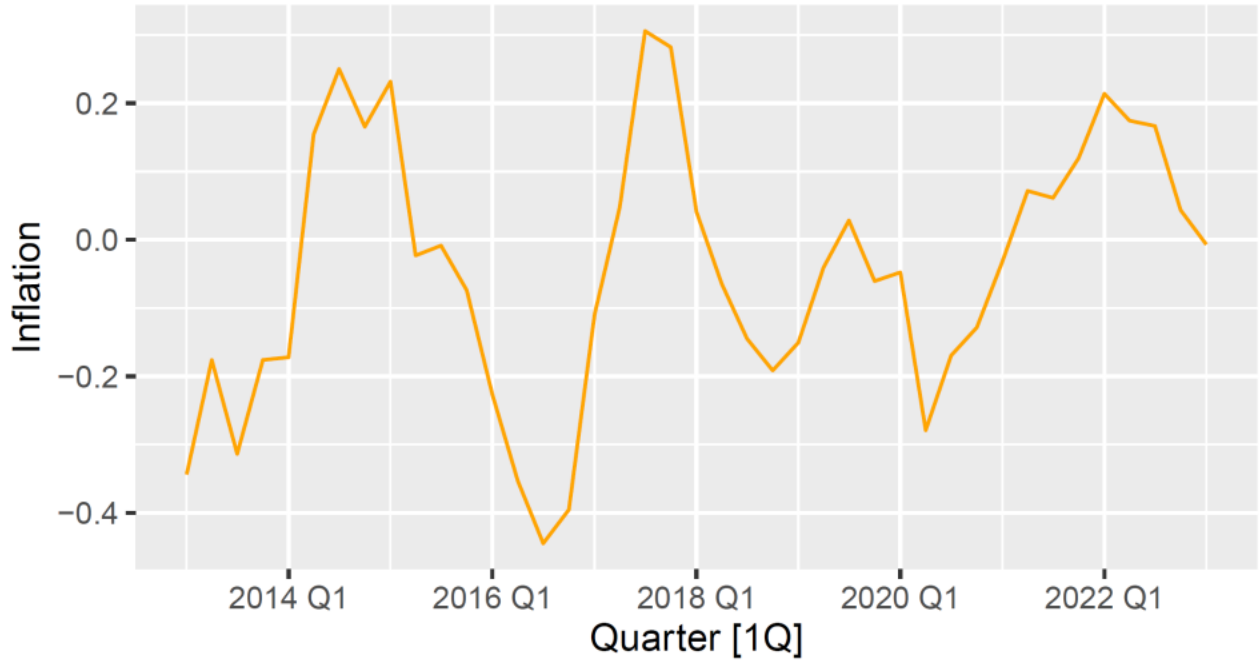
> This is particularly relevant as in 2008-2009, data releases on key macrovariables are usually delayed several months as compared to observed macro and financial shocks

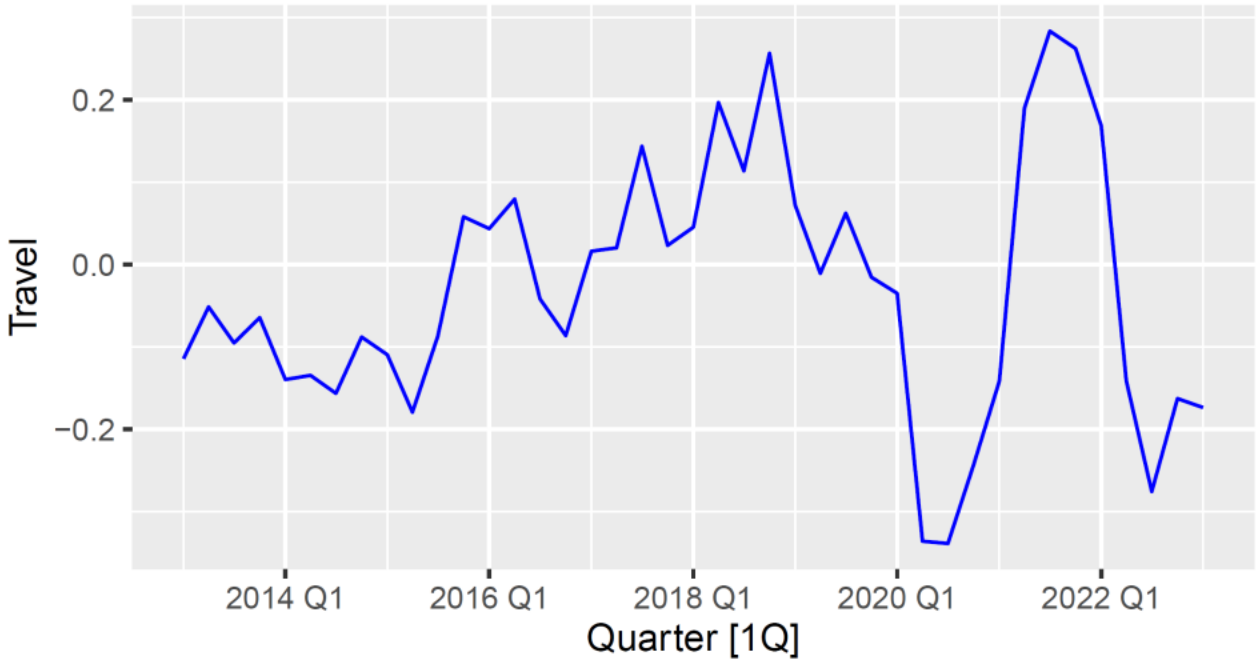
> **Choi and Varian (2010)** "Prediciting the Present with Google Trends" makes a strong point for the use of Google trends data to nowcast a multitude of economic variables such as car sales, unemployment claims, travel, and consumer confidence.

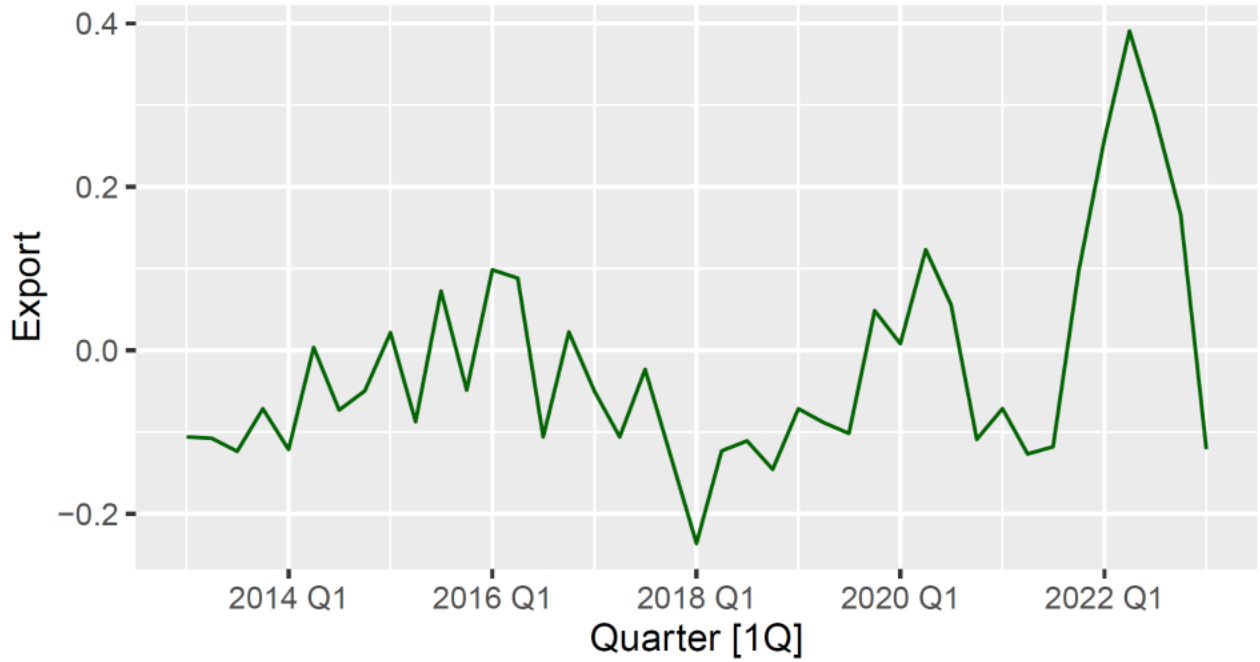
>**Wu and Brynjolfsson (2010)** leverage Google Search Data to forecast house prices

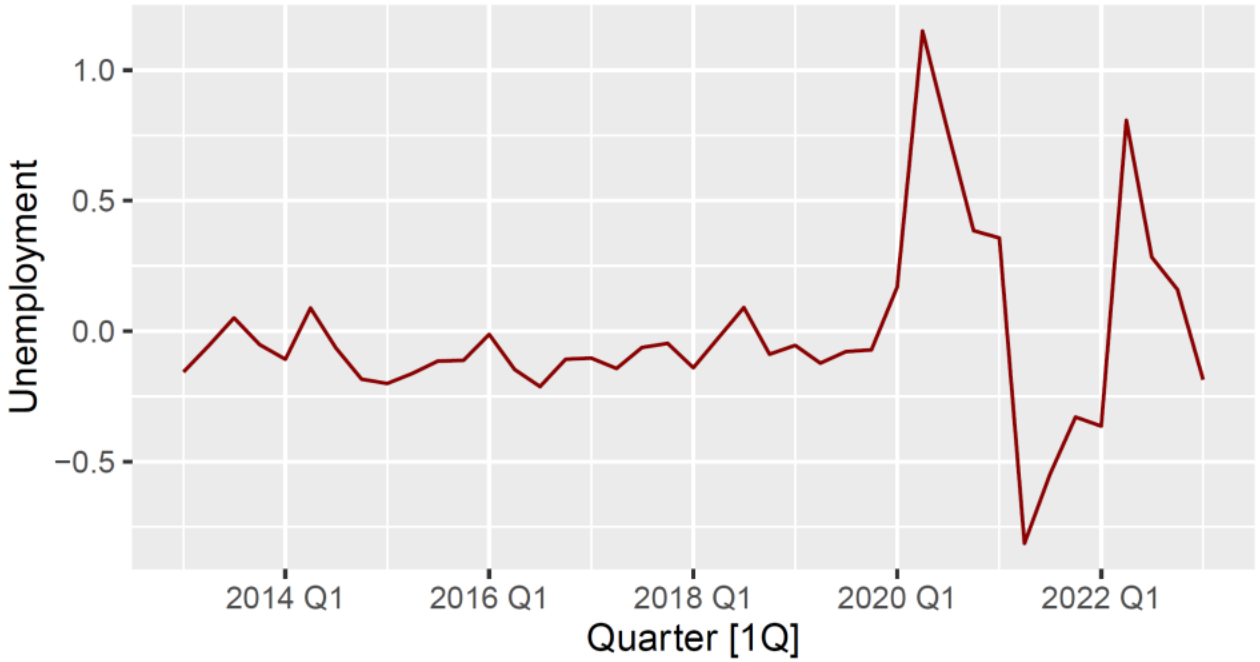

Cherkaska
Chernihivska
Chernivetska
Dnipropetrovska
Donetska
Ivano−Frankivska
Kharkivska
Khersonska
Khmelnytska
Kirovohradska
Krym
Kyivska
Kyivska City
Luhanska
Lvivska
Mykolaivska
Odeska
Poltavska
Rivnenska
Sevastopilska
Sumska
Ternopilska
Vinnytska
Volynska
Zakarpatska
Zaporizka
Zhytomyrska
Google Inflation (blue) and Economy (orange)
Date
2013
2016
2019
2022

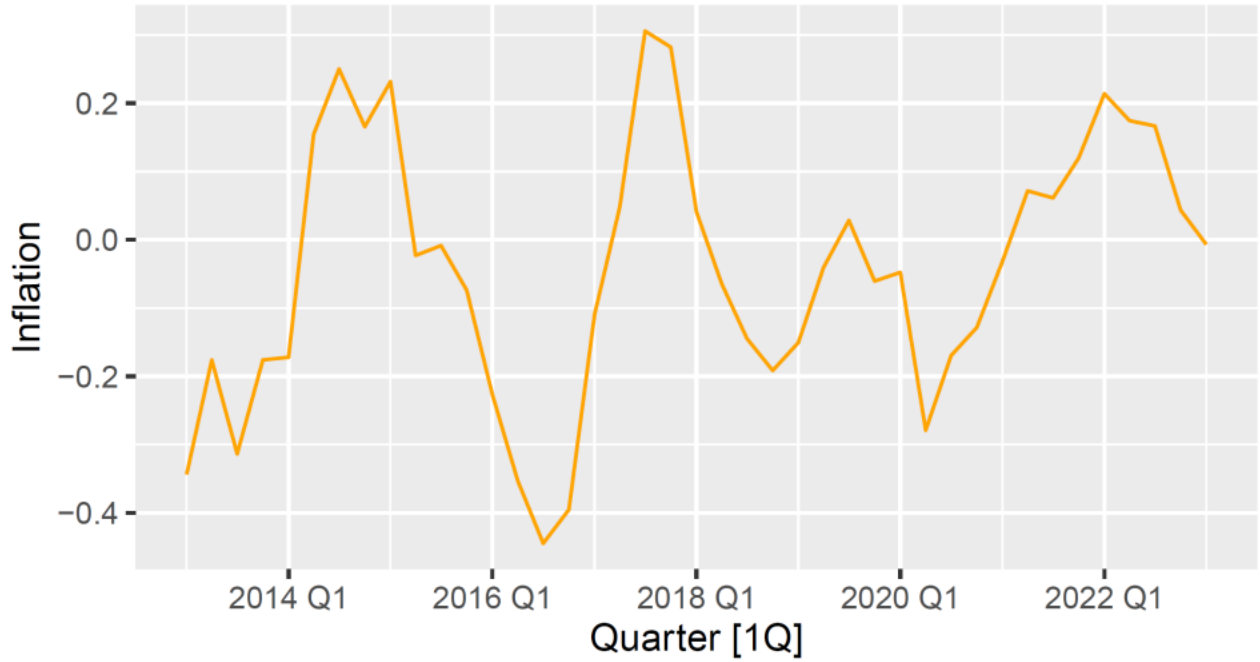

Unemployment
1.0
0.5
0.0
−0.5
2014 Q1
2016 Q1
2018 Q1
2020 Q1
2022 Q1
Quarter [1Q]

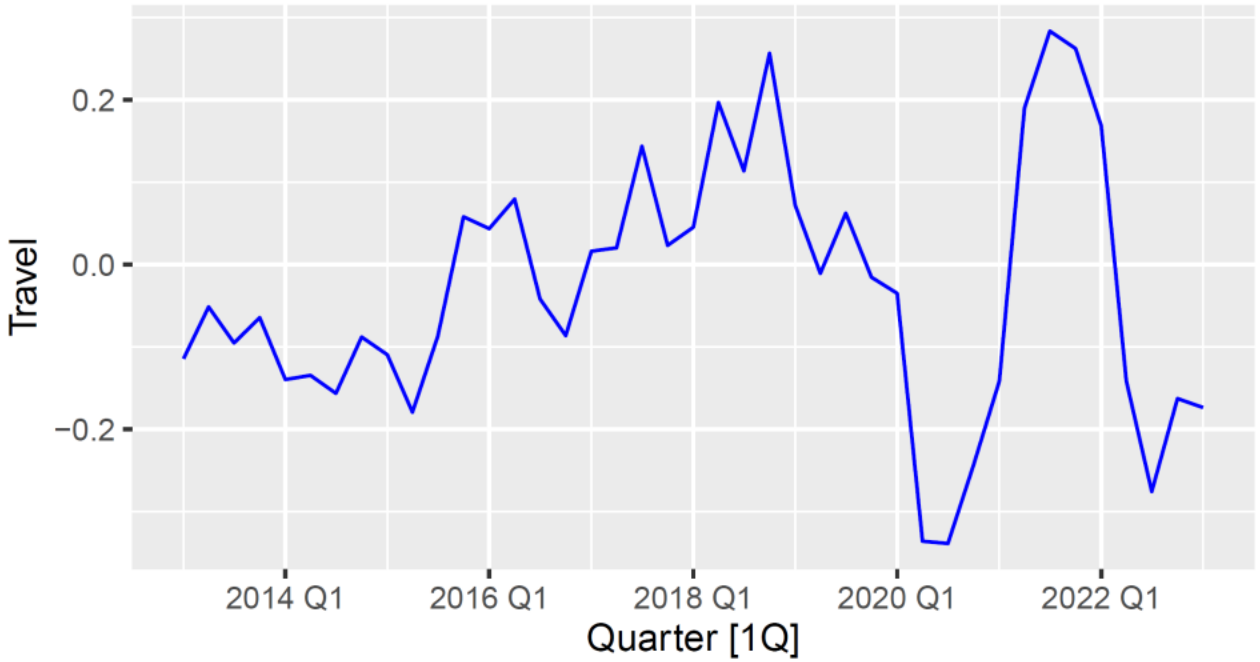

Inflation
0.2
0.0
−0.2
−0.4
2014 Q1
2016 Q1
2018 Q1
2020 Q1
2022 Q1
Quarter [1Q]

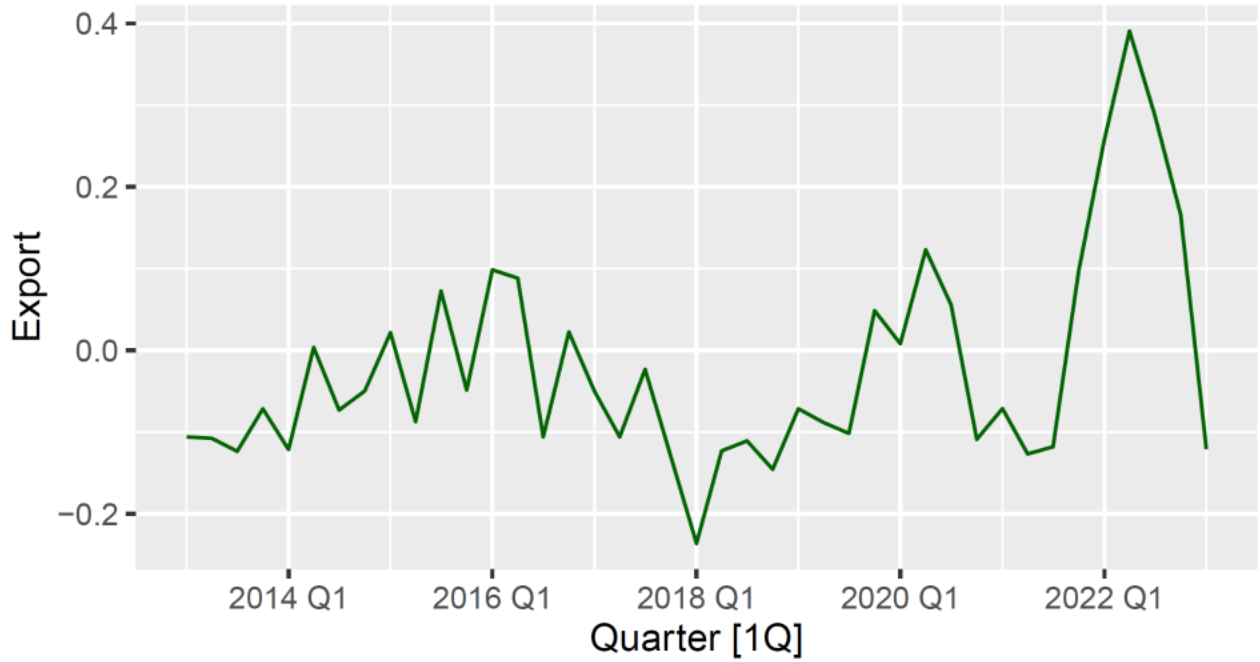

Travel
0.2
0.0
−0.2
2014 Q1
2016 Q1
2018 Q1
2020 Q1
2022 Q1
Quarter [1Q]


Export
0.4
0.2
0.0
−0.2
2014 Q1
2016 Q1
2018 Q1
2020 Q1
2022 Q1
Quarter [1Q]


Unemployment
1.0
0.5
0.0
−0.5
2014 Q1
2016 Q1
2018 Q1
2020 Q1
2022 Q1
Quarter [1Q]


Inflation
0.2
0.0
−0.2
−0.4
2014 Q1
2016 Q1
2018 Q1
2020 Q1
2022 Q1
Quarter [1Q]


Travel
0.2
0.0
−0.2
2014 Q1
2016 Q1
2018 Q1
2020 Q1
2022 Q1
Quarter [1Q]


Export
0.4
0.2
0.0
−0.2
2014 Q1
2016 Q1
2018 Q1
2020 Q1
2022 Q1
Quarter [1Q]

Google Trends - empirics

> **Ettredge et al. (2005)** is one of the earlier references using Google Search activity to forecast US unemployment.

> **Askits Zimmermann (2009)** "Google Econometrics and Unemployment Forecasting" use Google searches related to unemployment to forecast offical figures several months ahead.

> This is particularly relevant as in 2008-2009, data releases on key macrovariables are usually delayed several months as compared to observed macro and financial shocks

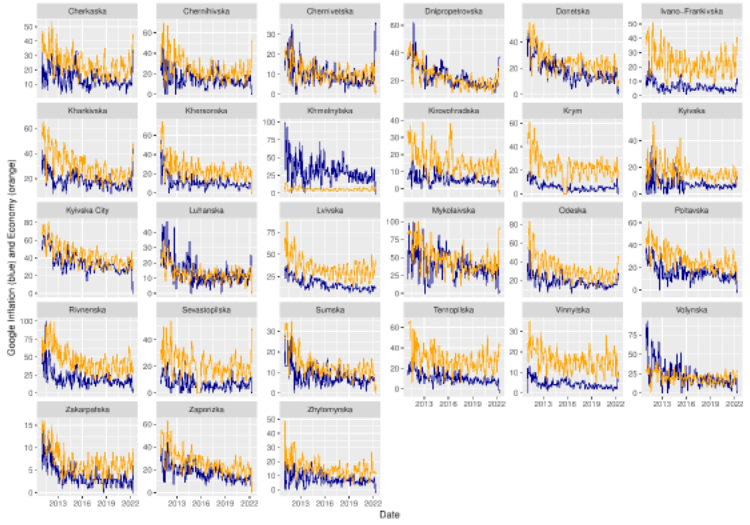
> **Choi and Varian (2010)** "Prediciting the Present with Google Trends" makes a strong point for the use of Google trends data to nowcast a multitude of economic variables such as car sales, unemployment claims, travel, and consumer confidence.

>**Wu and Brynjolfsson (2010)** leverage Google Search Data to forecast house prices

-> About 35 trends are used

-> Monthly time series contain plenty of variation (too much?) >> Q

-> Number implies some shrinkage/ dimensionality reduction is needed

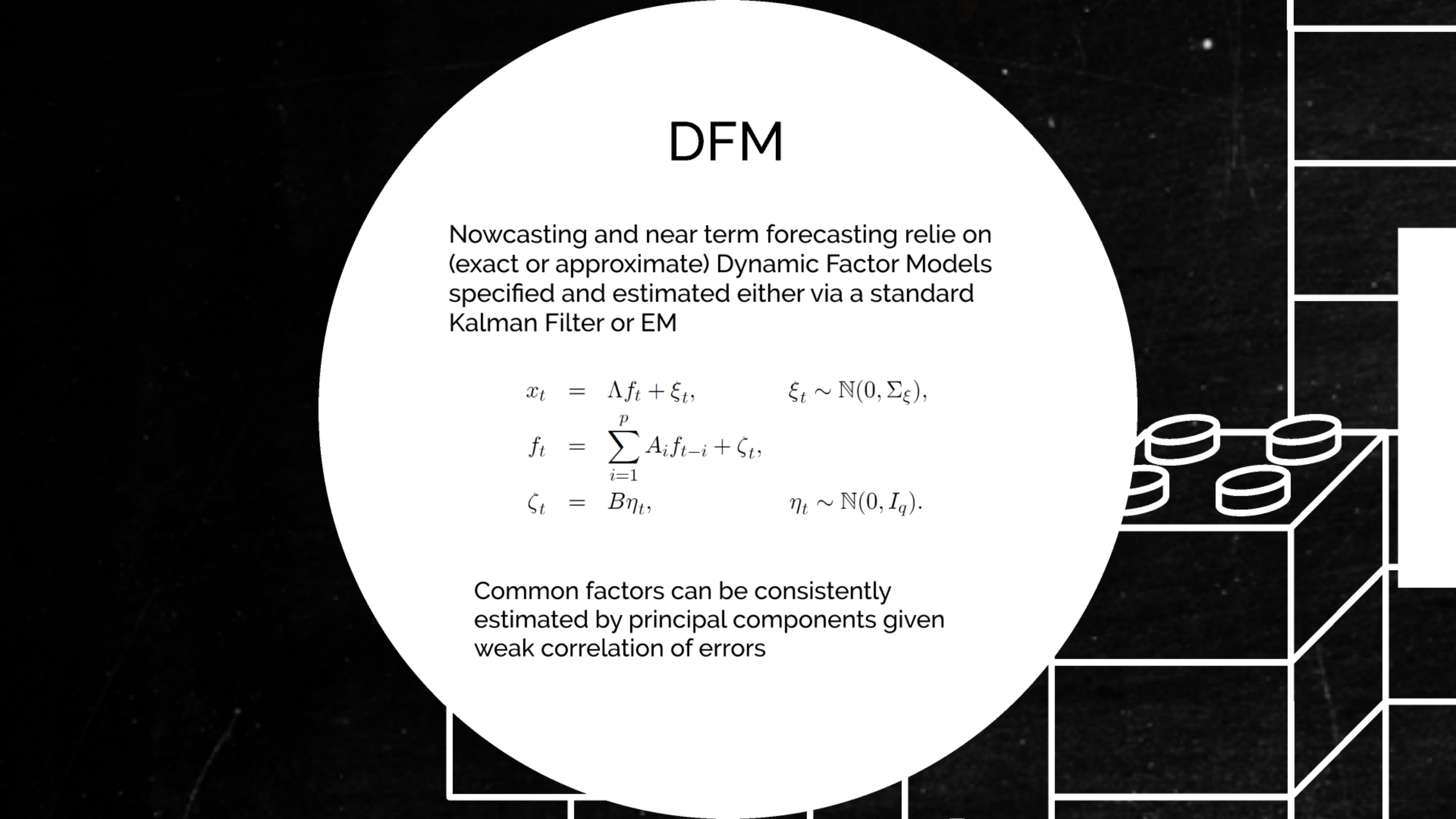


DFM

Nowcasting and near term forecasting relie on (exact or approximate) Dynamic Factor Models specified and estimated either via a standard Kalman Filter or EM

$$
\begin{aligned}
x_t &= \Lambda f_t + \xi_t, & \xi_t &\sim \mathbb{N}(0, \Sigma_\xi), \\
f_t &= \sum_{i=1}^{p} A_i f_{t-i} + \zeta_t, & & \\
\zeta_t &= B\eta_t, & \eta_t &\sim \mathbb{N}(0, I_q).
\end{aligned}
$$

Common factors can be consistently estimated by principal components given weak correlation of errors

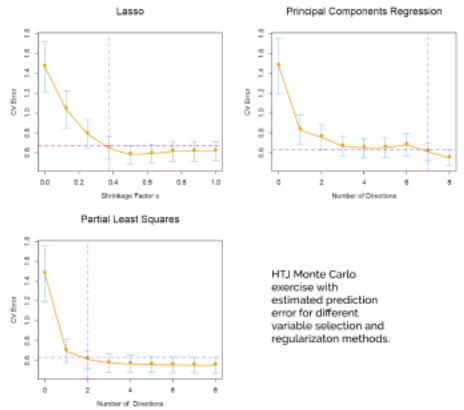


PCR - digging deeper

Principal Components

-> Given a data matrix **X** (N obs x p variables). PCA will perform a SVD of the centered matrix **X*** to find directions in the column space of **X*** that have small variance (with direction vectors *v* independent of each other)

$$\max_{\alpha} \mathrm{Var}(\mathbf{X}\alpha)$$
$$\text{subject to } ||\alpha|| = 1,\ \alpha^T \mathbf{S} v_\ell = 0,\ \ell = 1, \ldots, m-1,$$



HTJ Monte Carlo exercise with estimated prediction error for different variable selection and regularizaton methods.

Partial Least Squares

-> PLS is a supervised method which identifies the components or factors (phi) to be independent of each other but also have high correlation with a target **y**
{Wold et al. 1984}

$$\max_{\alpha} \mathrm{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\mathrm{Var}(\mathbf{X}\alpha)$$
$$\text{subject to } ||\alpha|| = 1,\ \alpha^T \mathbf{S} \hat{\varphi}_\ell = 0,\ \ell = 1, \ldots, m-1.$$

Hastie, Tibshirani, Friedman 2nd ed. "The Elements of Statistical Learning"

PLS as a Latent Factor Model

Principal Components

-> Given a data matrix **X** (N obs x p variables). PCA will perform a SVD of the centered matrix **X*** to find directions in the column space of **X*** that have small variance (with direction vectors *v* independent of each other)

$$\max_{\alpha} \mathrm{Var}(\mathbf{X}\alpha)$$
$$\text{subject to } ||\alpha|| = 1,\ \alpha^T \mathbf{S} v_\ell = 0,\ \ell = 1, \ldots, m-1,$$

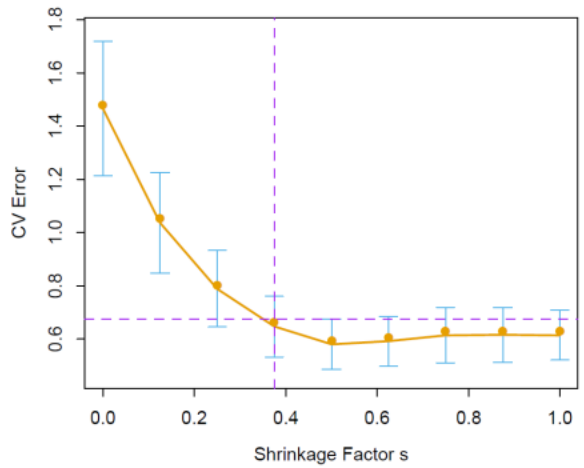
Partial Least Squares

-> PLS is a supervised method which identifies the components or factors (phi) to be independent of each other but also have high correlation with a target **y**
{Wold et al. 1984}

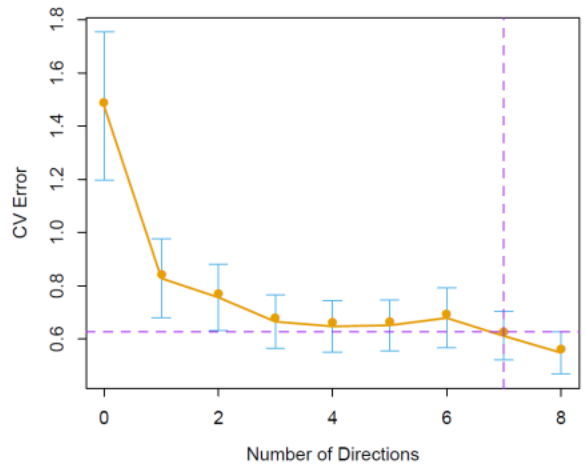
$$\max_{\alpha} \mathrm{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\mathrm{Var}(\mathbf{X}\alpha)$$
$$\text{subject to } ||\alpha|| = 1,\ \alpha^T \mathbf{S}\hat{\varphi}_\ell = 0,\ \ell = 1, \ldots, m-1.$$

Hastie, Tibshirani, Friedman 2nd ed. "The Elements of Statistical Learning"

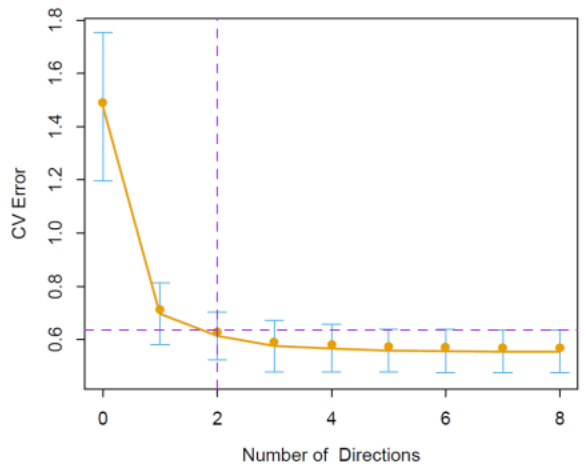
Lasso



Principal Components Regression



Partial Least Squares



HTJ Monte Carlo exercise with estimated prediction error for different variable selection and regularizaton methods.

PLS as a Latent Factor Model

PLS as a Latent Factor Model

PLS as a Latent Factor Model

$$
\begin{aligned}
T &= X \times W^*, \quad X \in \mathbb{R}^{n \times p}, T \in \mathbb{R}^{n \times K}, K < p \\
X &= T \times P' + \epsilon, \quad P \in \mathbb{R}^{p \times K} \\
y &= T \times C' + \xi, \quad C \in \mathbb{R}^{1 \times K}
\end{aligned}
$$

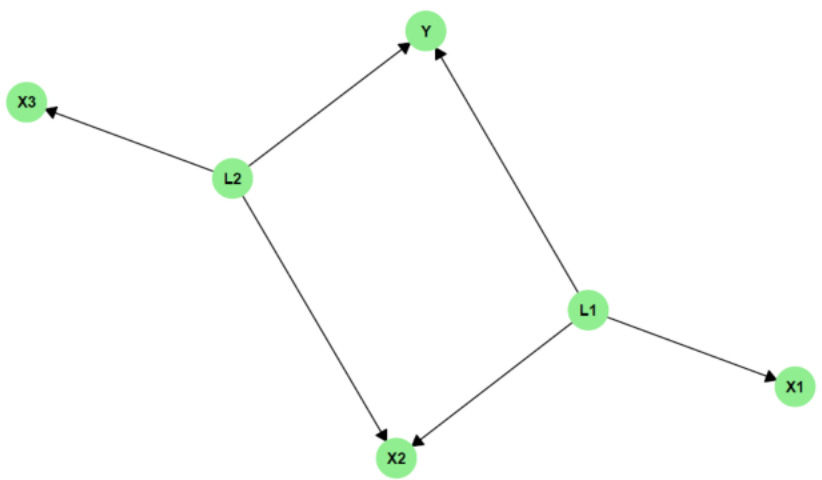


Figure 2 – Network Graph of a Simple Latent Factor Model

PLS in Macro

Although a close cousin of PCR, PLS is a bit the Harry&Meghan of the macro econometric family

>> first emerged as a *computational algorithm* to tackle multicolinearity in bioinformatics lacking the needed theoretical scafolding **Wold (1984)**

>> **Helland (1990)** illuminates the relationship between PLS and PCR presenting conditions under which the two methodologies yield similar results.

>> **Stoica and Söderström (1998)** explore conditions under which PCR and PLS produce equivalent results, deriving asymptotic formulas for bias and variance of the PLS estimator.

>> PLS in macroeconomic forecasting in **Eickmeier and Ng (2011), Cubadda et al. (2013), Groen and Kapetanios (2016)**, and in finance in the study by **Preda and Saporta (2005)** and **Kelly and Pruitt (2015)**

Although a close cousin of PCR, PLS is a bit the Harry&Meghan of the macro econometric family

>> first emerged as a *computational algorithm* to tackle multicolinearity in bioinformatics lacking the needed theoretical scafolding **Wold (1984)**

>> **Helland (1990)** illuminates the relationship between PLS and PCR presenting conditions under which the two methodologies yield similar results.

>> **Stoica and Söderström (1998)** explore conditions under which PCR and PLS produce equivalent results, deriving asymptotic formulas for bias and variance of the PLS estimator.

>> PLS in macroeconomic forecasting in **Eickmeier and Ng (2011), Cubadda et al. (2013), Groen and Kapetanios (2016)**, and in finance in the study by **Preda and Saporta (2005)** and **Kelly and Pruitt (2015)**

ML model overview

Sparsity to tackle asymptotic inconsistency risk

-> Chun and Keles (2010) indicate challenges to asymptotic consistency of the PLS estimator in a "large p small n" context, with *fixed* **p1** *relevant* and *increasing* **p – p1** *irrelevant* variables.

-> The intuition for the lack of asymptotic consistency comes from the ridge-like nature of the PLS algorithm. Given that PLS latent factors load on all variables available in X, a larger fraction of irrelevant variables weaken the ability of the algorithm to identify the true factor directions.

-> Sparsity is achieved via variable selection in a multitude of ways, depending on the joint specificities of data sample and machine learning model (Lasso like, FWD or BWD Variable Selection, GA)

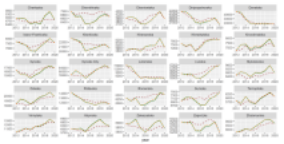
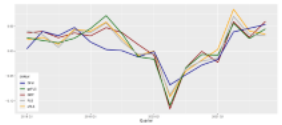
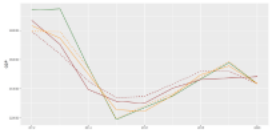
Variable Selection - overview

-> A wrapper (GA) and an embbeded method (sPLS) are used to induce sparsity: as a results, variable selection leads to improved interpretability

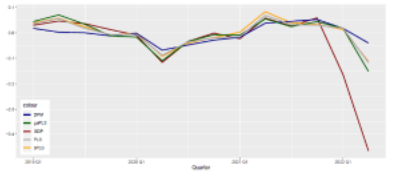
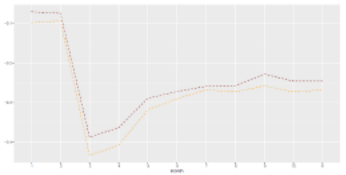
-> sPLS of Chun and Keles (2010) introduces a LASSO penalty in the optimization problem and jointly selects the optimal number of latent factors and the amount of penaty

Variable Selection - the GA algorithm

Insample Fit

Out of sample Forecast

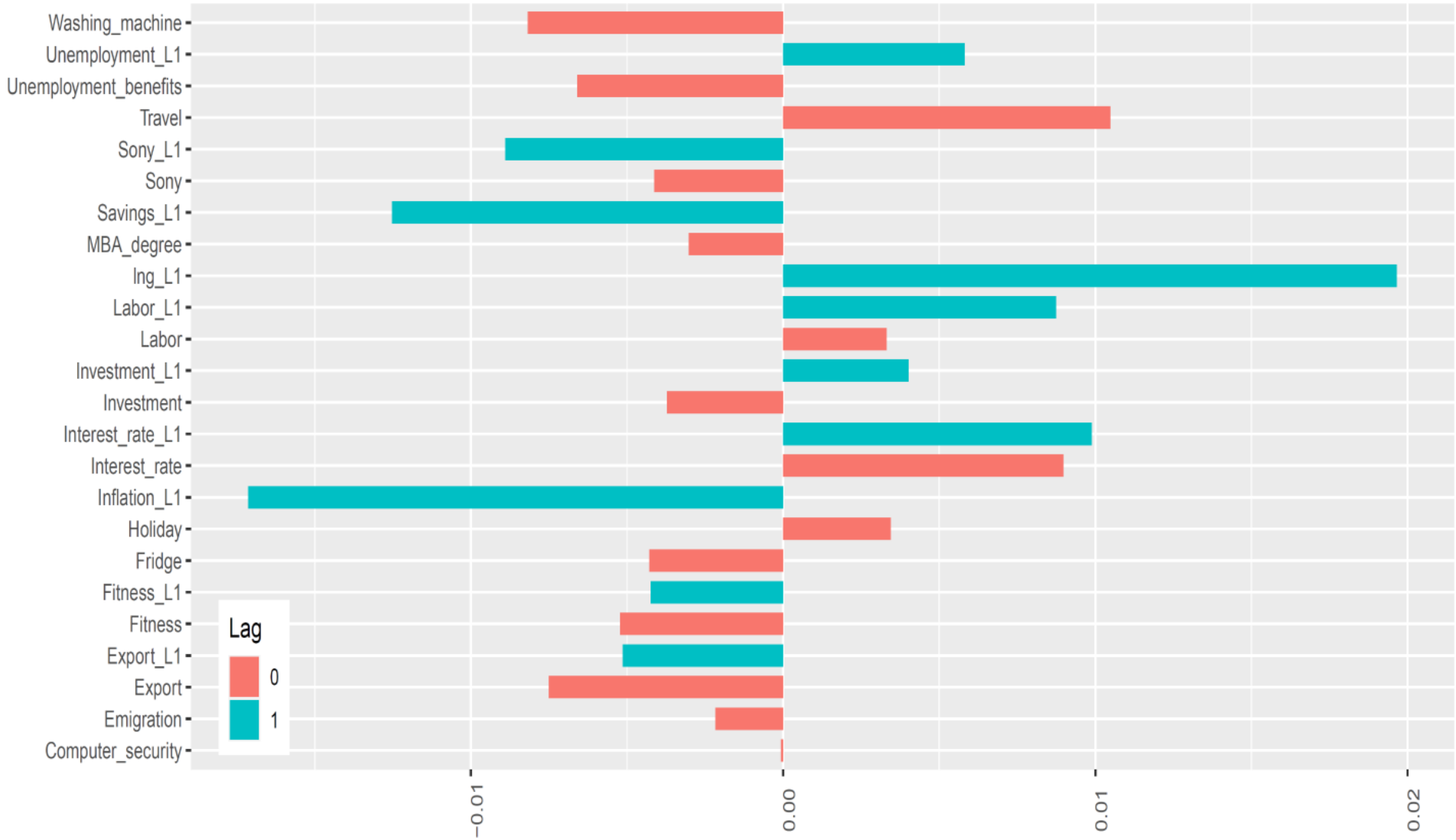
-> Chun and Keles (2010) indicate challenges to asymptotic consistency of the PLS estimator in a "large p small n" context, with *fixed* **p1** *relevant* and *increasing* **p – p1** *irrelevant* variables.

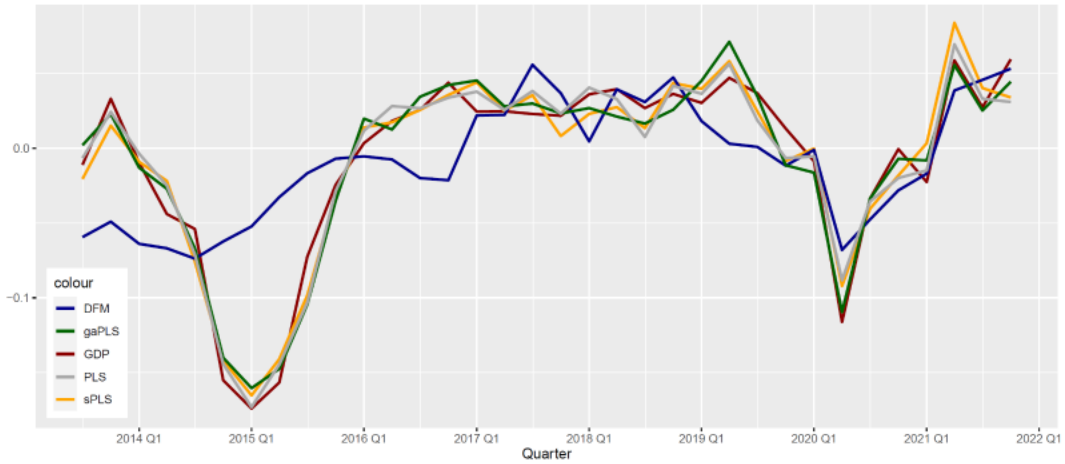
-> The intuition for the lack of asymptotic consistency comes from the ridge-like nature of the PLS algorithm. Given that PLS latent factors load on all variables available in X, a larger fraction of irrelevant variables weaken the ability of the algorithm to identify the true factor directions.

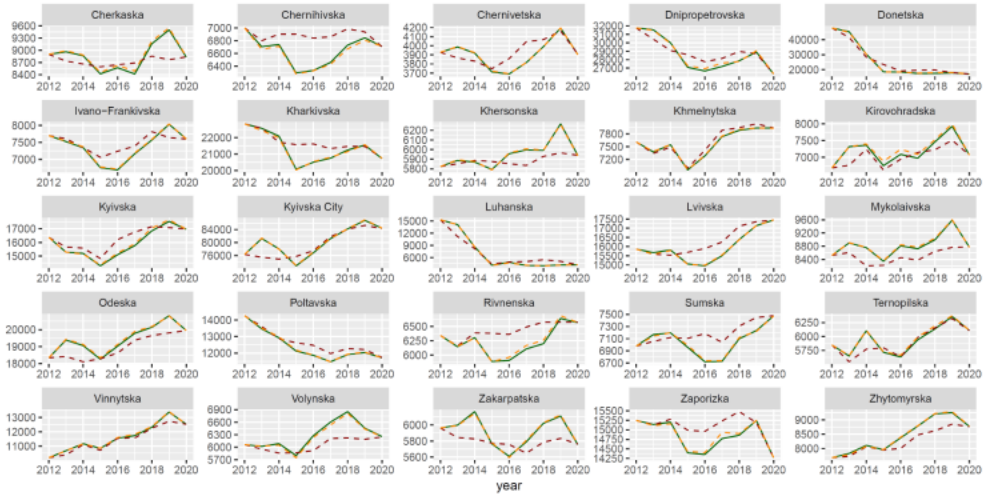
-> Sparsity is achieved via variable selection in a multitude of ways, depending on the joint specificities of data sample and machine learning model (Lasso like, FWD or BWD Variable Selection, GA)

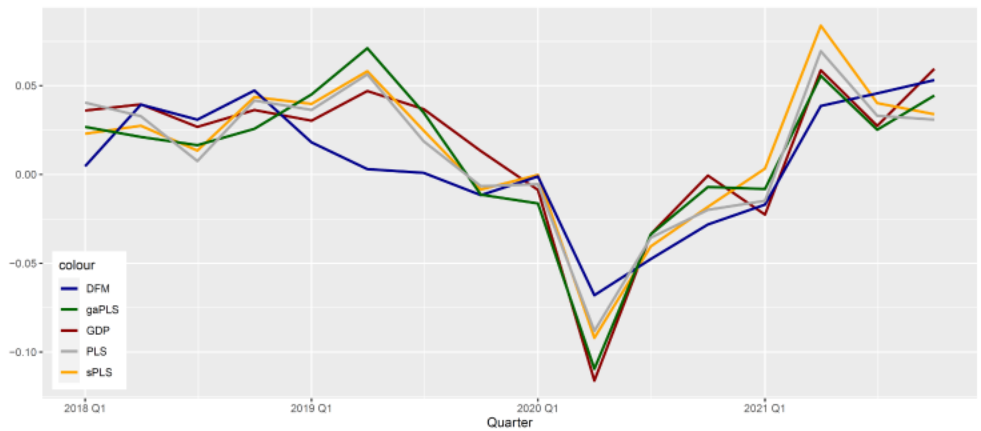
-> A wrapper (GA) and an embbeded method (sPLS) are used to induce sparsity: as a results, variable selection leads to improved interpretability

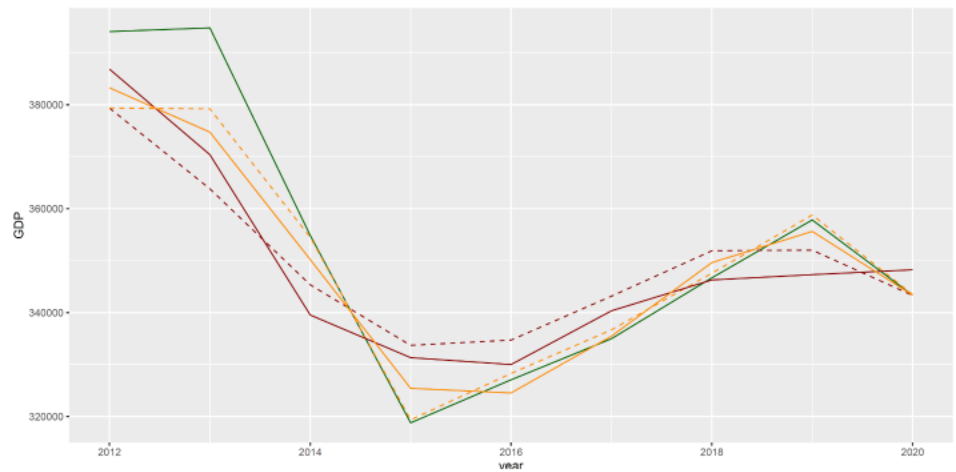
-> sPLS of Chun and Keles (2010) introduces a LASSO penalty in the optimization problem and jointly selects the optimal number of latent factors and the amount of penaty

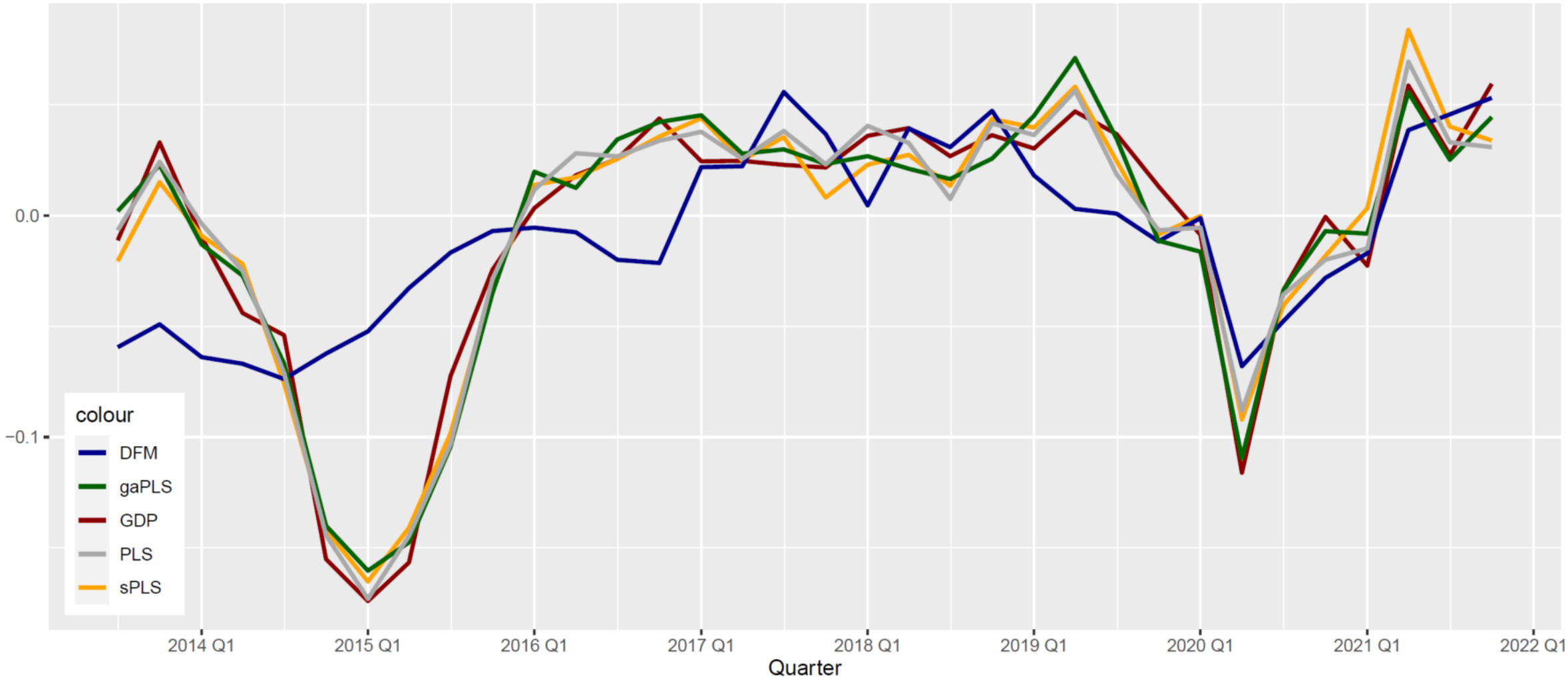

Washing_machine
Unemployment_L1
Unemployment_benefits
Travel
Sony_L1
Sony
Savings_L1
MBA_degree
Ing_L1
Labor_L1
Labor
Investment_L1
Investment
Interest_rate_L1
Interest_rate
Inflation_L1
Holiday
Fridge
Fitness_L1
Fitness
Export_L1
Export
Emigration
Computer_security
Lag
0
1
−0.01
0.00
0.01
0.02

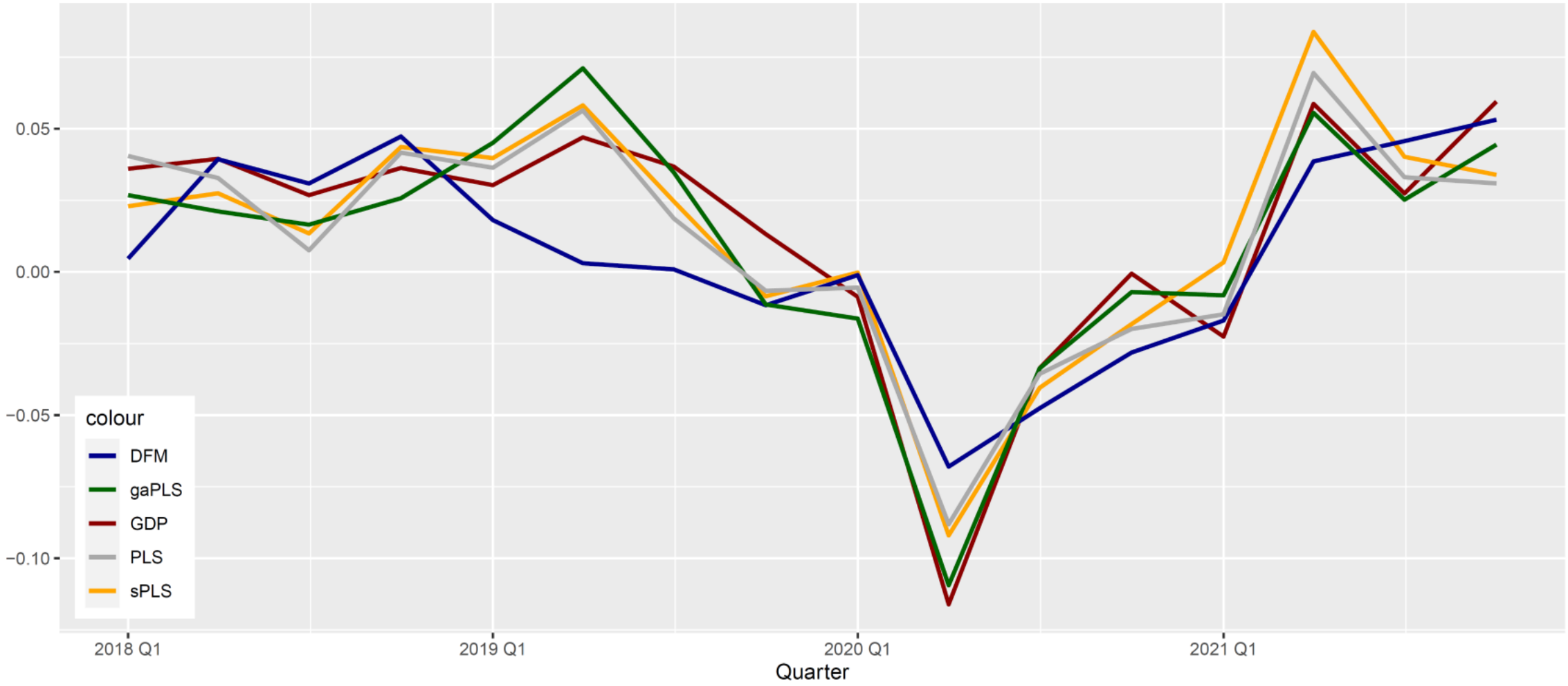

0.0
−0.1
colour
DFM
gaPLS
GDP
PLS
sPLS
2014 Q1
2015 Q1
2016 Q1
2017 Q1
2018 Q1
2019 Q1
2020 Q1
2021 Q1
2022 Q1
Quarter

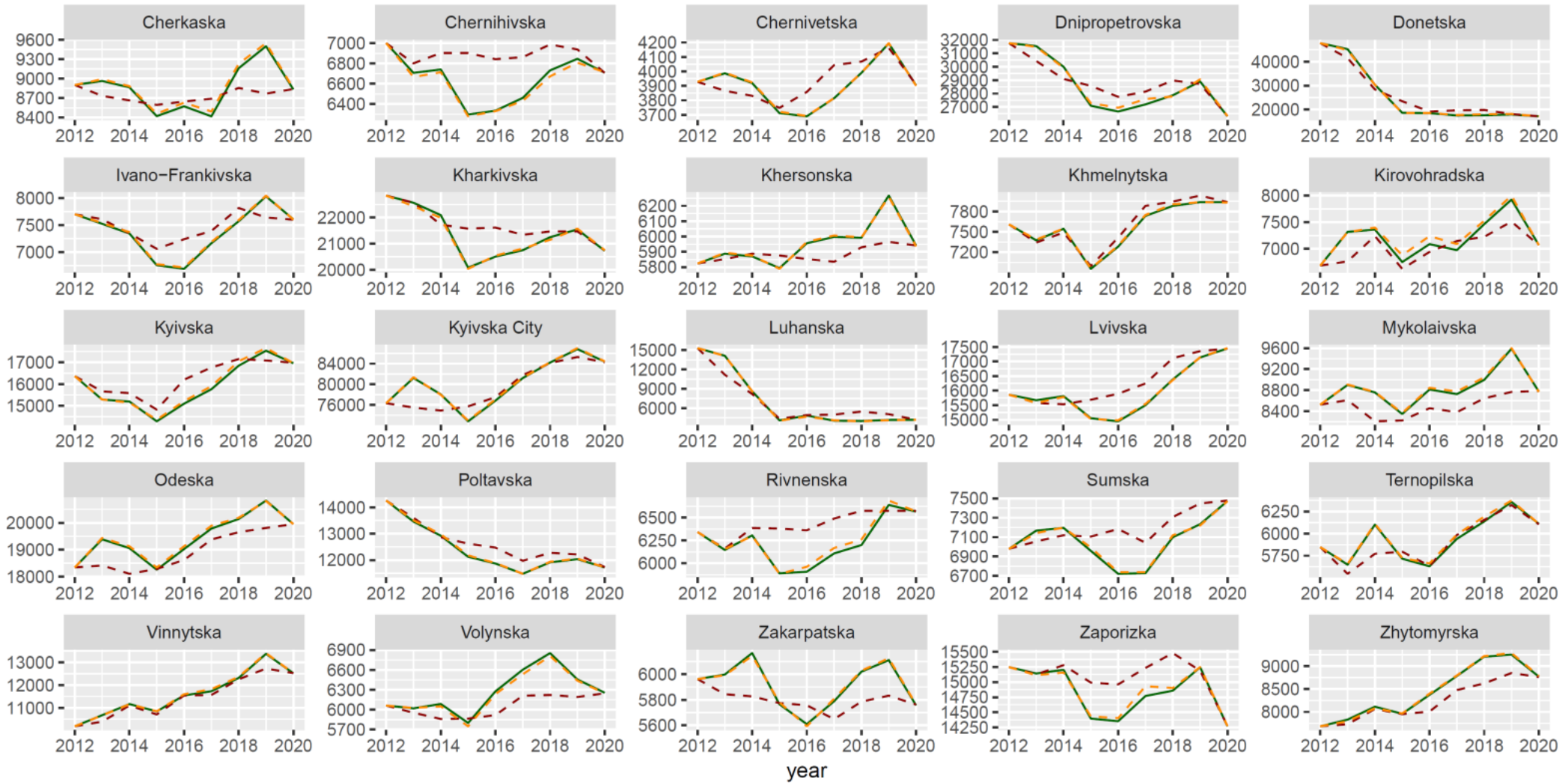

Cherkaska
Chernihivska
Chernivetska
Dnipropetrovska
Donetska
Ivano−Frankivska
Kharkivska
Khersonska
Khmelnytska
Kirovohradska
Kyivska
Kyivska City
Luhanska
Lvivska
Mykolaivska
Odeska
Poltavska
Rivnenska
Sumska
Ternopilska
Vinnytska
Volynska
Zakarpatska
Zaporizka
Zhytomyrska
2012
2014
2016
2018
2020
year

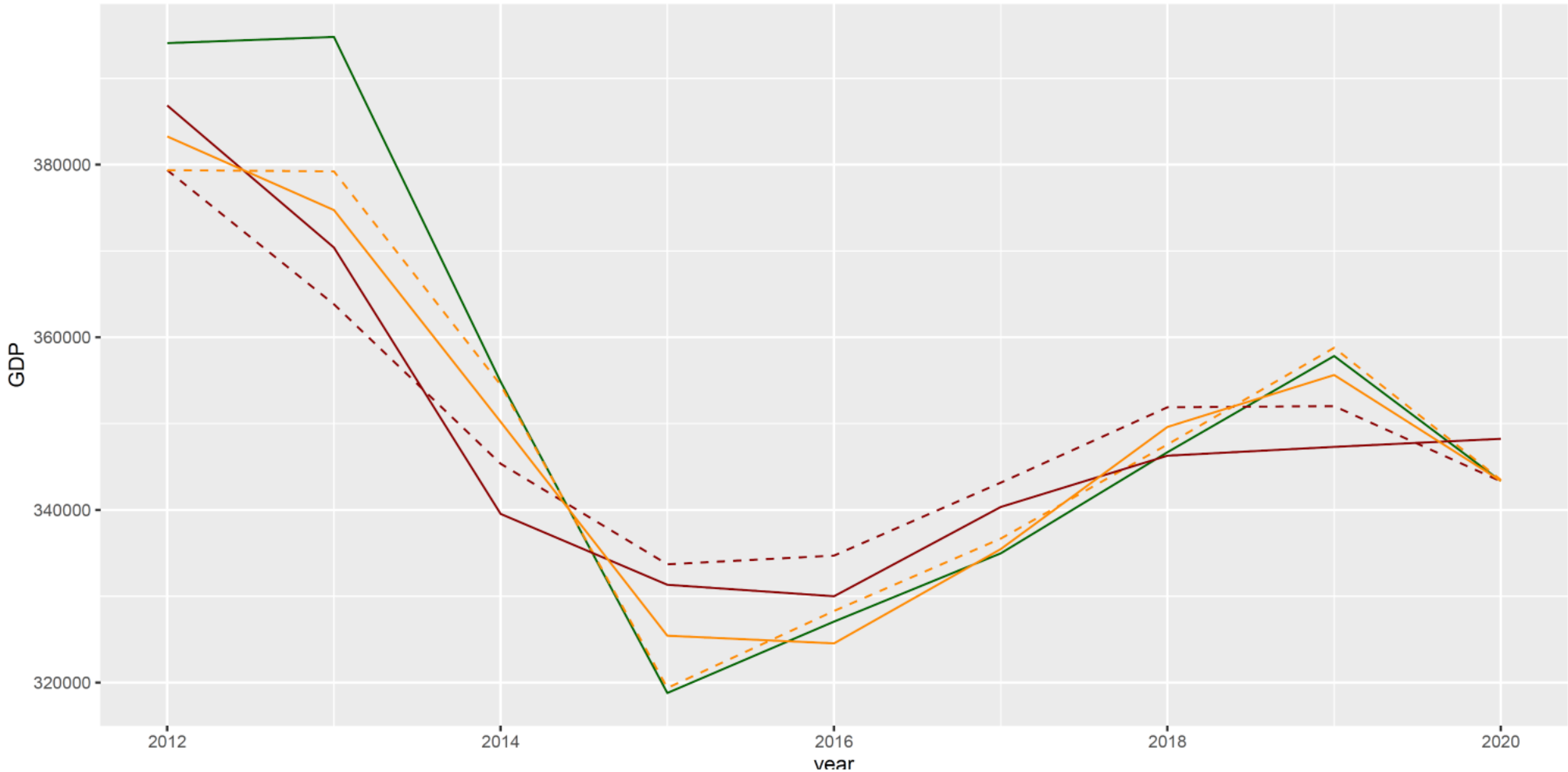

0.05
0.00
−0.05
−0.10
colour
DFM
gaPLS
GDP
PLS
sPLS
2018 Q1
2019 Q1
2020 Q1
2021 Q1
Quarter

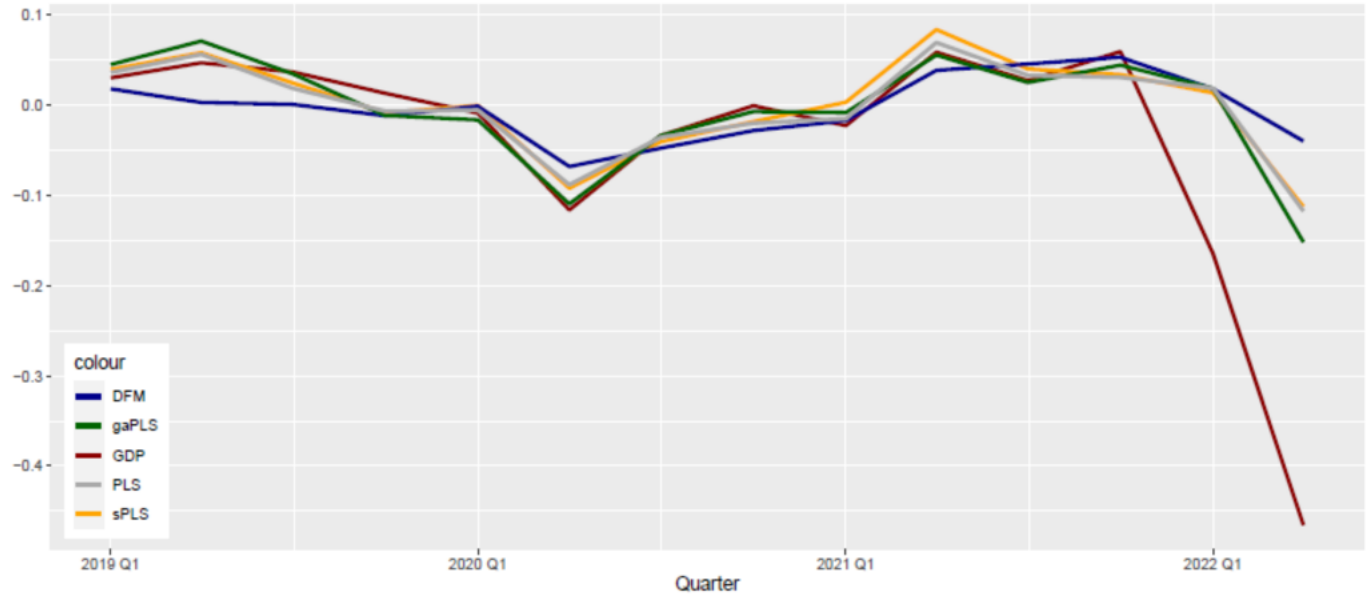

380000
360000
340000
320000
GDP
2012
2014
2016
2018
2020
year

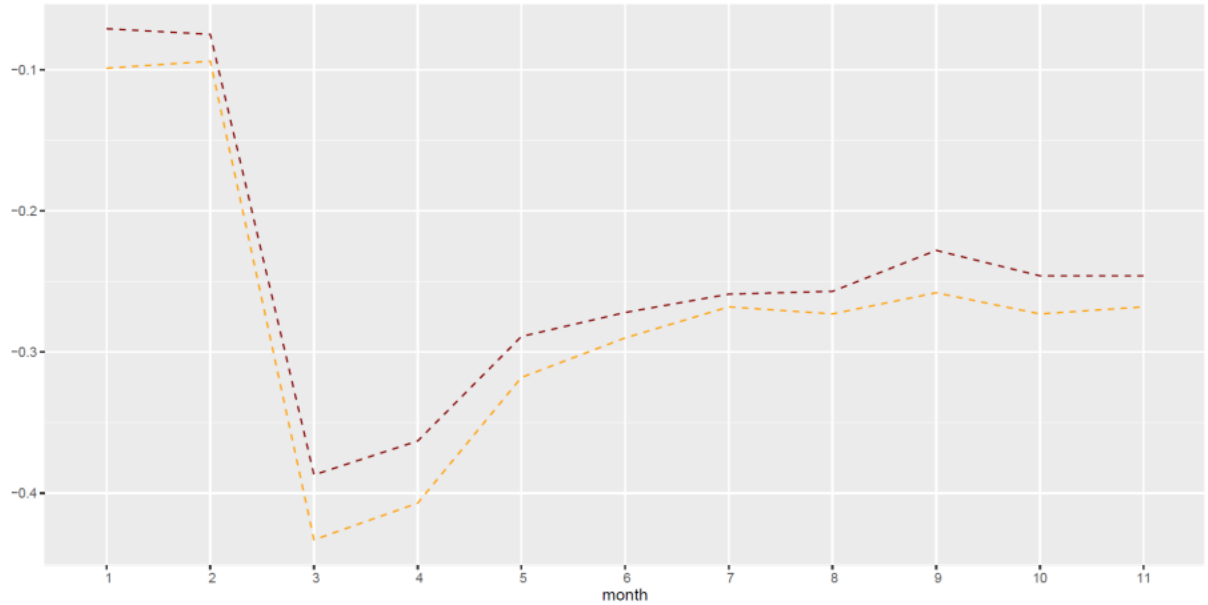

0.0
−0.1
colour
DFM
gaPLS
GDP
PLS
sPLS
2014 Q1
2015 Q1
2016 Q1
2017 Q1
2018 Q1
2019 Q1
2020 Q1
2021 Q1
2022 Q1
Quarter


0.05
0.00
−0.05
−0.10
colour
DFM
gaPLS
GDP
PLS
sPLS
2018 Q1
2019 Q1
2020 Q1
2021 Q1
Quarter


Cherkaska
Chernihivska
Chernivetska
Dnipropetrovska
Donetska
Ivano−Frankivska
Kharkivska
Khersonska
Khmelnytska
Kirovohradska
Kyivska
Kyivska City
Luhanska
Lvivska
Mykolaivska
Odeska
Poltavska
Rivnenska
Sumska
Ternopilska
Vinnytska
Volynska
Zakarpatska
Zaporizka
Zhytomyrska
year


GDP
380000
360000
340000
320000
2012
2014
2016
2018
2020
year


0.1
0.0
−0.1
−0.2
−0.3
−0.4
colour
DFM
gaPLS
GDP
PLS
sPLS
2019 Q1
2020 Q1
2021 Q1
2022 Q1
Quarter


−0.1
−0.2
−0.3
−0.4
1
2
3
4
5
6
7
8
9
10
11
month

final considerations

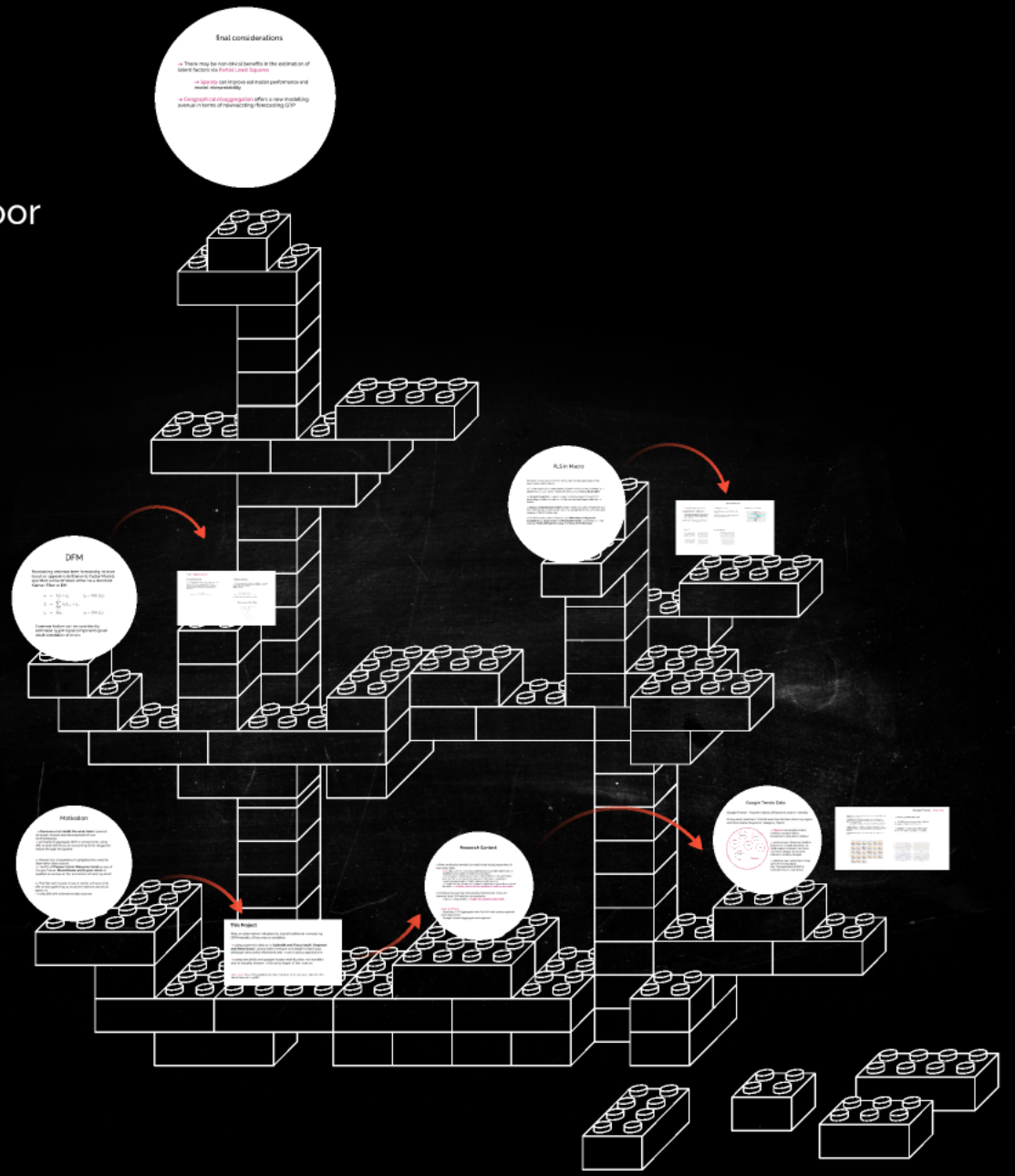
-> There may be non-trivial benefits in the estimation of latent factors via Partial Least Squares

-> Sparsity can improve estimation performance and model interpretability

-> Geographical disaggregation offers a new modelling avenue in terms of nowvacsting/forecasting GDP

Sparse Warcasting

Forecasting in a data-rich but statistics-poor environment



Mihnea Constantinescu
National Bank of Ukraine and
University of Amsterdam

June 2023
NBU-NBP ARC