Nowcasting with Big Data: is Google helpful in the presence of other information?

Xinyuan Li

London Business School

Modelling with Big Data and Machine Learning
organized by BOE/FRB/KCL
London, Nov 26-27, 2018

Motivation

- Economic decisions in real time should be based on economic data and predictions
- Economic data are often released with significant delays
- The recent surge in big data
- cheap/fast to collect
- may help measure/predict economic conditions better
  - web-based search data: Ginsberg et al (2009 Nature), Choi and Varian (2012, 2014), Da et al (2011 JOF, 2015 RFS)
  - text: Tetlock (2007 JOF), Gentzkow and Shapiro (2011, 2014 AER, econometrica 2010), Hansen et al (2018 AER)
  - picture: Premise x World Bank
  - satellite: Henderson et al (2012 AER)
  - web-scraped prices: Billion prices project

Agenda / Contribution

Main contributions of this paper:

- issues of Google search data and suggestions for its use
- literature review: forecasting with search data
- a framework for real-time forecasting to incorporate these timely data source, as in Bańbura et al (2013, Handbook of Econ. Forecasting)
  - able to deal with a large panel
  - able to deal with mixed frequency
  - able to deal with unsynchronized releases
- two applications: initial claims and unemployment

Literature

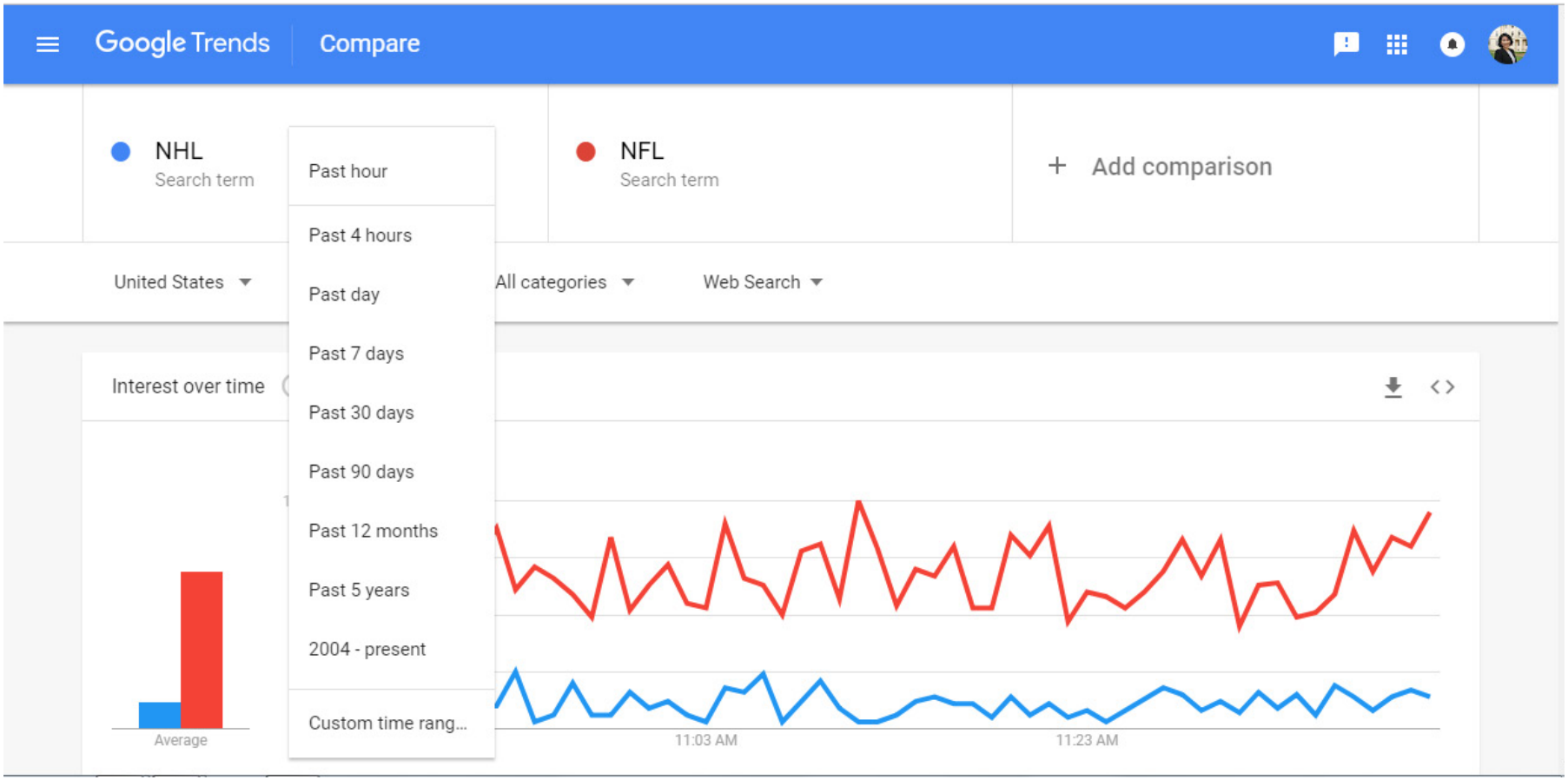
- Choi and Varian (2012): Google Trends can help predict the present
- Choi and Varian (2014): use Google Trends and Google Correlate to forecast initial claims and retail sales
- working papers: Bank of England (McLaren and Schabhogue, 2011), Bank of Italy (D'Amuri and Marcucci, 2015), Bank of Spain (Artola and Galan, 2012), Central Bank of Chile (Carrière-Swallow and Labbé, 2010), Central Bank of the Republic of Turkey (Chadwick and Sengül, 2012), Bank of Isreal (Suhoy, 2009)
- surveys on Nowcasting and Big Data: Bok et al (2018), De Mol et al (2016)

Google Trends (as of end-2016)

- Google Trends gives how often a particular term is searched relative to the total search volume over a certain period of time in a certain geographical region. The daily search share
$$S_{d,r} = V_{d,r}/T_{d,r}$$
- it is an index, not an absolute number
- the weekly share $S_{w,r}$ is the average of seven days
- the index scales the maximum value of share to 100
- available from Jan 4, 2004, weekly, released on Monday

Google Trends (as of JUL2018)

- becoming more granular and more real-time
- data frequency depends on how you request it ⇒ the shorter the requested period, the higher frequency
- e.g.: a 'past hour' request at 2018-07-12 11.44.48 BST ⇒ minute-by-minute data ⇒ identify shocks?



Google Correlate (as of JUL2018)

- it is a tool that allows users to upload their data of interest
- it outputs the queries whose search intensity over time/across states is most correlated with the data of interest
- Google: 'find searches that correlate with real-world data'
- the target can be monthly or weekly
- the output is monthly or weekly; daily data is achievable if reverse-engineered, combining Google Trends

Google Correlate (as of JUL2018)

lj4800617@gmail.com | M

Google Correlate | Search correlations | Enter your own data

Compare US states
Compare weekly time series
Compare monthly time series

Shift series 0 weeks
Country:
United States

Documentation
Comic Book
FAQ
Tutorial
Whitepaper
Correlate Algorithm

Correlate Labs
Search by Drawing

Find searches that correlate with real-world data

Google Correlate finds search patterns which correspond with real-world trends.

Compare time series

Many search terms vary in popularity over time. To find terms that vary in a similar way to your own time series, enter your data using the link above. Or take a look at these examples to see which search terms:

- ...are more popular in winter
- ...were most likely to be issued in 2005
- ...match the pattern of actual flu activity
  (this is how we built Google Flu Trends!)

You can also enter a query into the search box above to find search terms that have a similar pattern of activity, or try one of these:

- mittens
- losing weight
- ribosome

Compare US states

Search terms are often popular in some states and less popular in others. To find terms whose pattern of activity across the United States reflects your own US states dataset, enter your data using the link above. Or, you ca find terms correlated with:

- ...the state's latitude
- ...being in New England
- ...annual rainfall in the state

You can also use the search box above to see which searches correlate state-by-state to any query, or try on of these:

- mittens
- hunting season
- southern cooking

© 2011 Google - Send feedback - Terms of Service - Privacy Policy

Google Correlate/Trends: methodology (Mohebbi et al 2011, Vanderkam et al 2013)

- what is the unit of text?
- n-gram: a contiguous sequence of n items from a given sample of text or speech
- example: `'Bank of England'`
  - `Bank`
  - `of`
  - `England`
  - `Bank of`
  - `of England`
  - `Bank of England`
- each n-gram receives a count when the text is searched

Google Correlate: methodology (Mohebbi et al 2011, Vanderkam et al 2013)

- approximate nearest neighbour: find the first 100 terms that are most correlated
- in theory: loop over every n-gram in the database, and compute the Pearson correlation with the target
- Big Data: the n-gram databse $\sim O(10^7)$, cannot read into RAM $\Rightarrow$ Big Data algorithm
- in practice: approximate distance is computed for each n-gram and the target

Google Correlate: methodology (Mohebbi et al 2011, Vanderkam et al 2013)

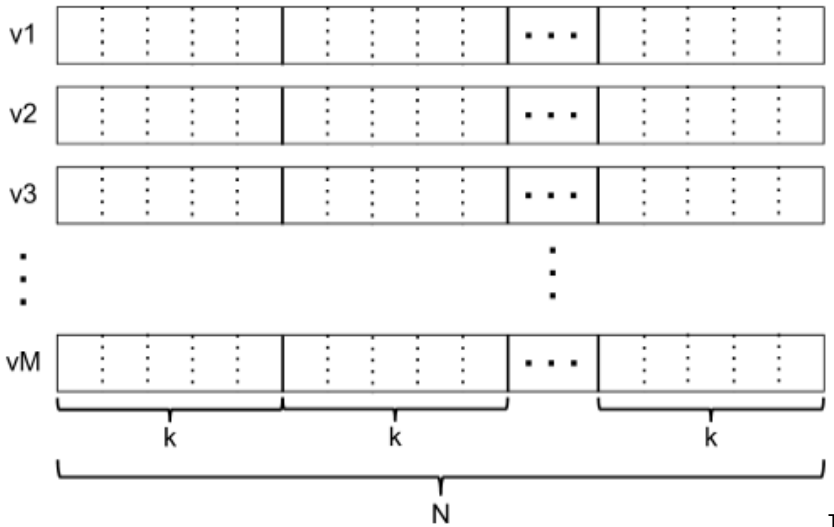


Figure 2: Illustration of how vectors are split into *chunks*

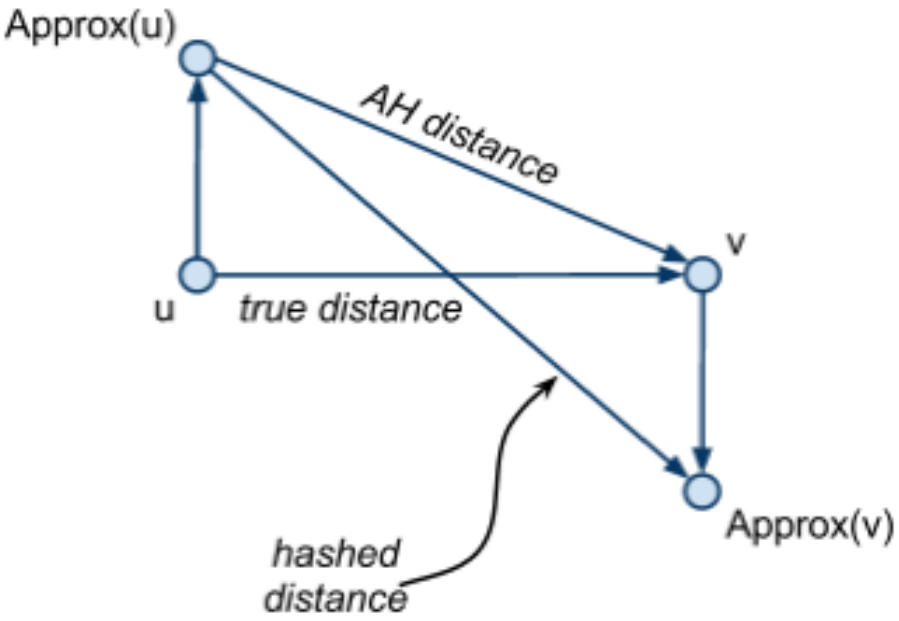


Figure 3: 2 Dimensional visualization of the distance to approximations of vectors.

Data: Issues with Google Trends / Correlate

- Data compression/Variable selection
- Seasonality
- Sampling error

Data Compression/Variable Selection

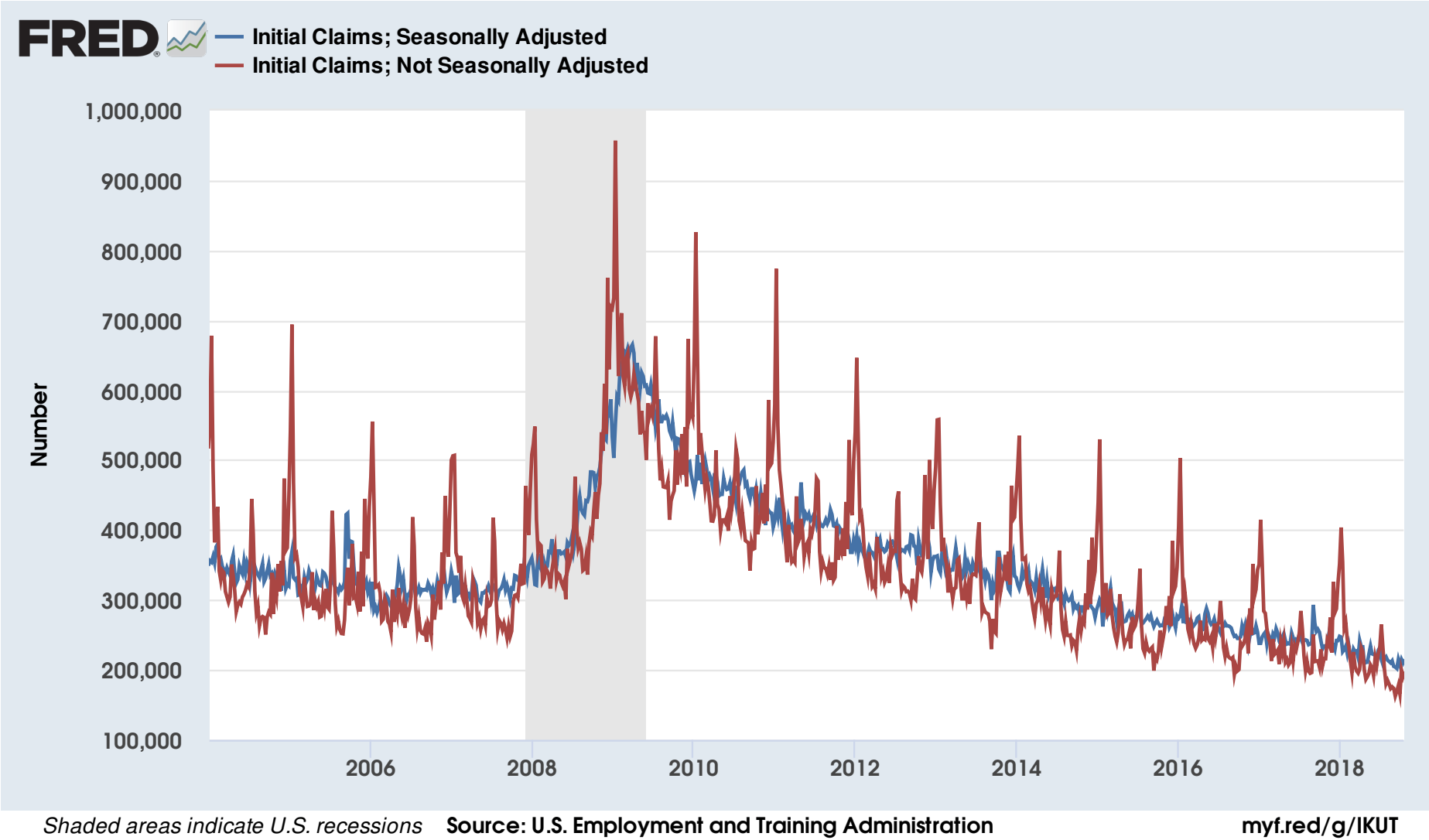
$$\boxed{\text{tens of millions of queries}}$$

$$\overset{Google}{\Rightarrow} \quad \boxed{\text{Google Trends / Correlate}}$$

$$\overset{Economists}{\Rightarrow} \quad \boxed{\text{predictors}}$$

- first arrow: how Google compresses raw data into indices is black-boxy; dynamic algorithm (Lazer et al, 2014 Science) $\Rightarrow$ becoming more stable recently, Google Trends api available
- less so for individual words $\Rightarrow$ select words $\Rightarrow$ robustness?
- second arrow: economists as end-users select variables for prediction/estimation purposes

Initial claims: ICSA (target) vs. ICNSA



Google Correlate (ICSA, vintage=30JUN2015)

...terms can be economically meaningless

Google Correlate

ICSA | Search correlations | Sign in to enter your own data

Compare US states
Compare weekly time series
Compare monthly time series

Shift series 0 weeks
Country:
United States

Documentation
Comic Book
FAQ
Tutorial
Whitepaper
Correlate Algorithm

Correlate Labs
Search by Drawing

Correlated with **ICSA**

0.9468	uncoached
0.9447	loan modifications
0.9417	dare4distance
0.9403	loan modification
0.9391	laguna beach jeans
0.9382	craigslist helper
0.9381	mgmt kids lyrics
0.9365	coomclips
0.9361	att webmail
0.9354	dare4distance lyrics

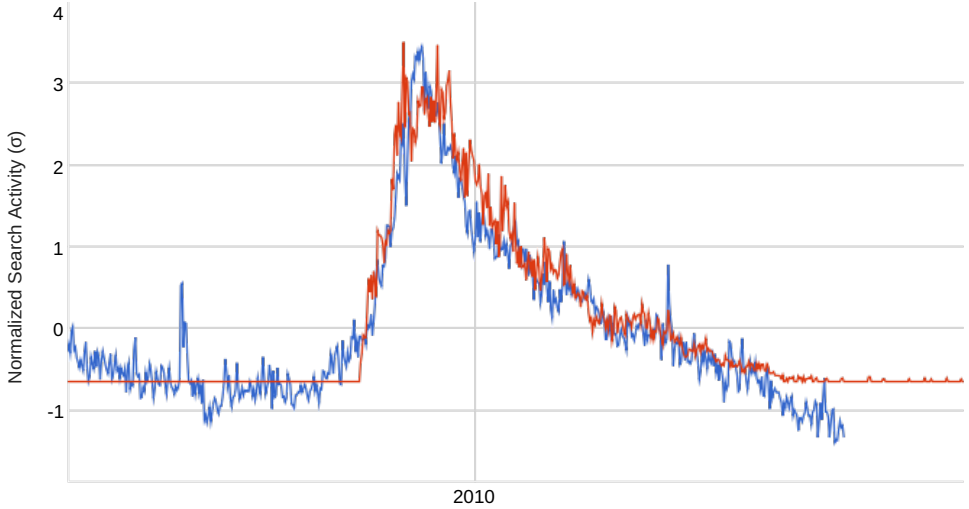
Show more | Sign in to export data as CSV | Share:

User uploaded activity for **ICSA** and United States Web Search activity for **uncoached** (r=0.9468)

Line chart Scatter plot

— **ICSA** — **uncoached**

Hint: Drag to Zoom, and then correlate over that time only.



Google Correlate (ICNSA, vintage=01JUL2015)

...but sometimes can be meaningful

Google Correlate

ICNSA | Search correlations | Sign in to enter your own data

Compare US states
Compare weekly time series
Compare monthly time series

Shift series 0 weeks
Country:
United States

Documentation
Comic Book
FAQ
Tutorial
Whitepaper
Correlate Algorithm

Correlate Labs
Search by Drawing

Correlated with **ICNSA**

0.8697	michigan unemployment
0.8315	idaho unemployment
0.8088	pennsylvania unemployment
0.8043	rhode island unemployment
0.7989	illinois unemployment
0.7966	unemployment filing
0.7914	filing unemployment
0.7913	jersey unemployment
0.7912	state unemployment
0.7908	new jersey unemployment

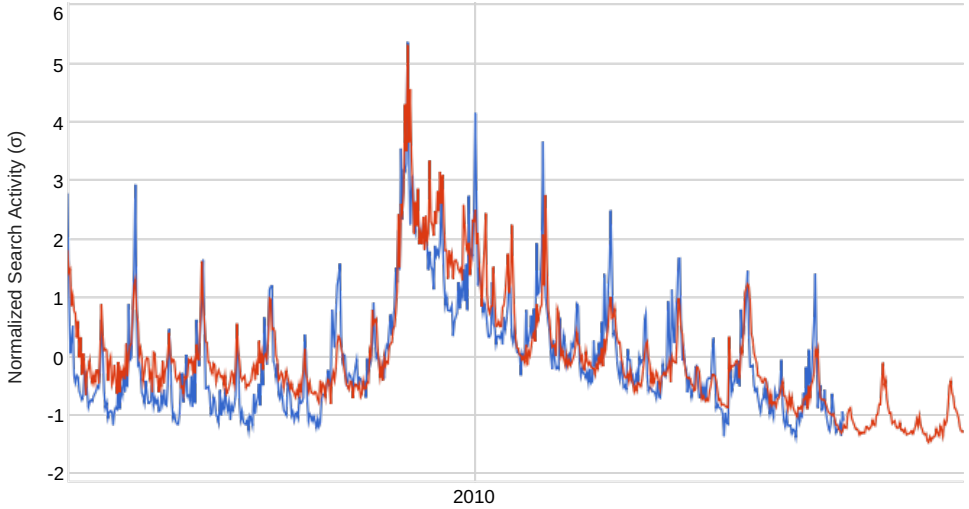
Show more | Sign in to export data as CSV | Share:

User uploaded activity for **ICNSA** and United States Web Search activity for **michigan unemployment** (r=0.8697)

Line chart Scatter plot

— ICNSA — michigan unemployment

Hint: Drag to Zoom, and then correlate over that time only.



Google Correlate (ICNSA, vintage=31MAR2013)

...when the target sample is small

Google Correlate

ICNSA_2013-03-31 | Search correlations | Sign in to enter your own data

Compare US states
Compare weekly time series
Compare monthly time series

Shift series 0 weeks
Country:
United States

Documentation
Comic Book
FAQ
Tutorial
Whitepaper
Correlate Algorithm

Correlate Labs
Search by Drawing

Correlated with **ICNSA_2013-03-31** between 2010-11-05 and 2013-02-07

0.8594	free winter desktop
0.8478	winter scene wallpaper
0.8471	discount calendars
0.8448	winter lesson plans
0.8430	free winter
0.8425	winter desktop wallpaper
0.8412	free winter desktop wallpaper
0.8400	calendar free
0.8386	browntrout.com
0.8385	desktop wallpaper winter

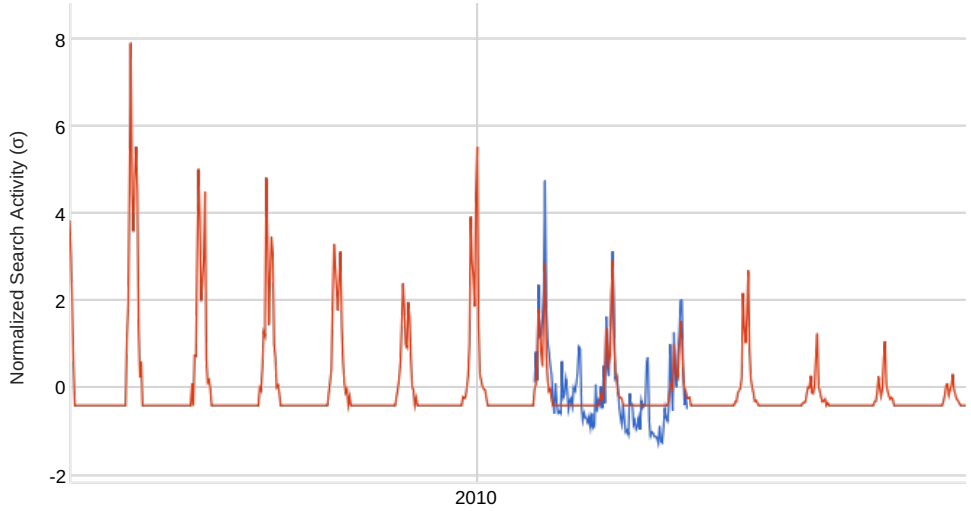
Show more | Sign in to export data as CSV | Share:

User uploaded activity for **ICNSA_2013-03-31** and United States Web Search activity for **free winter desktop** (r=0.8594)

Line chart Scatter plot

— ICNSA_2013-03-31 — free winter desktop

Click to search over all time.



Google Correlate (ICNSA, vintage=31MAR2013)

...in the bust

Google Correlate

ICNSA_2013-03-31 | Search correlations | Sign in to enter your own data

Compare US states
Compare weekly time series
Compare monthly time series

Shift series 0 weeks
Country:
United States

Documentation
Comic Book
FAQ
Tutorial
Whitepaper
Correlate Algorithm

Correlate Labs
Search by Drawing

Correlated with **ICNSA_2013-03-31** between 2008-05-17 and 2009-06-25

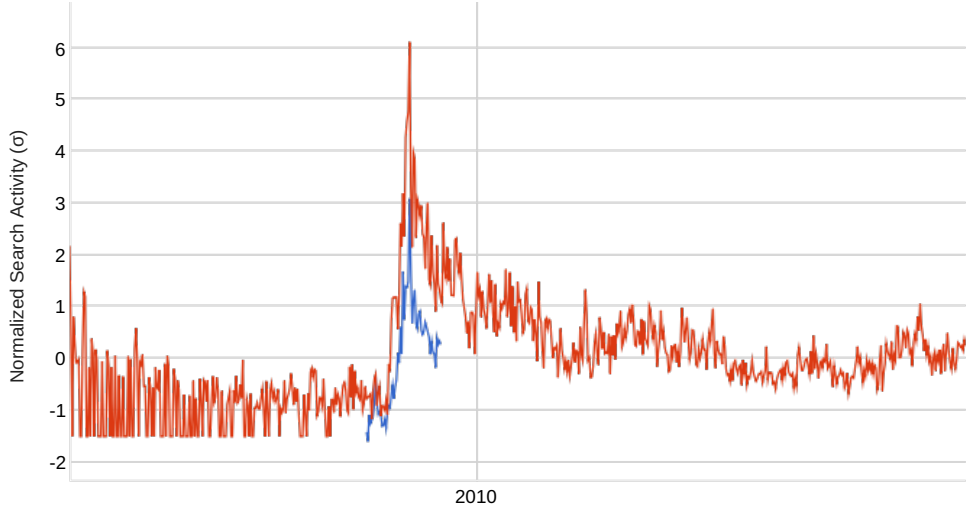
0.9525 corporate bond fund
0.9427 unemployment application
0.9423 new jersey unemployment
0.9419 jersey unemployment
0.9397 unemployment number
0.9395 snap fitness locations
0.9394 eapply4ui
0.9382 2009 calendar
0.9377 pennsylvania unemployment
0.9358 bni foreclosures

Show more | Sign in to export data as CSV | Share:

User uploaded activity for **ICNSA_2013-03-31** and United States Web Search activity for **corporate bond fund** (r=0.9525)

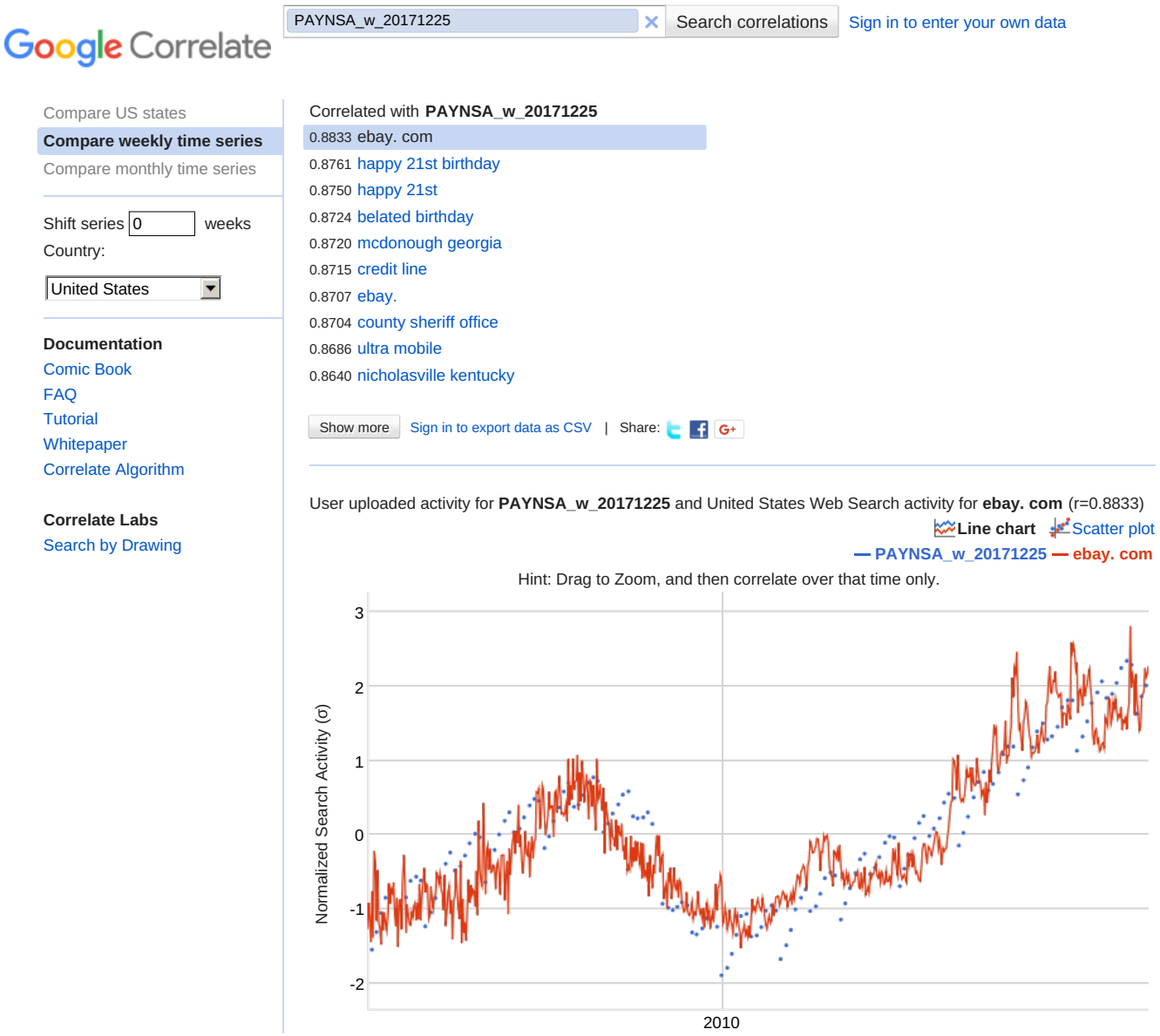
Line chart Scatter plot
— ICNSA_2013-03-31 — corporate bond fund

Click to search over all time.



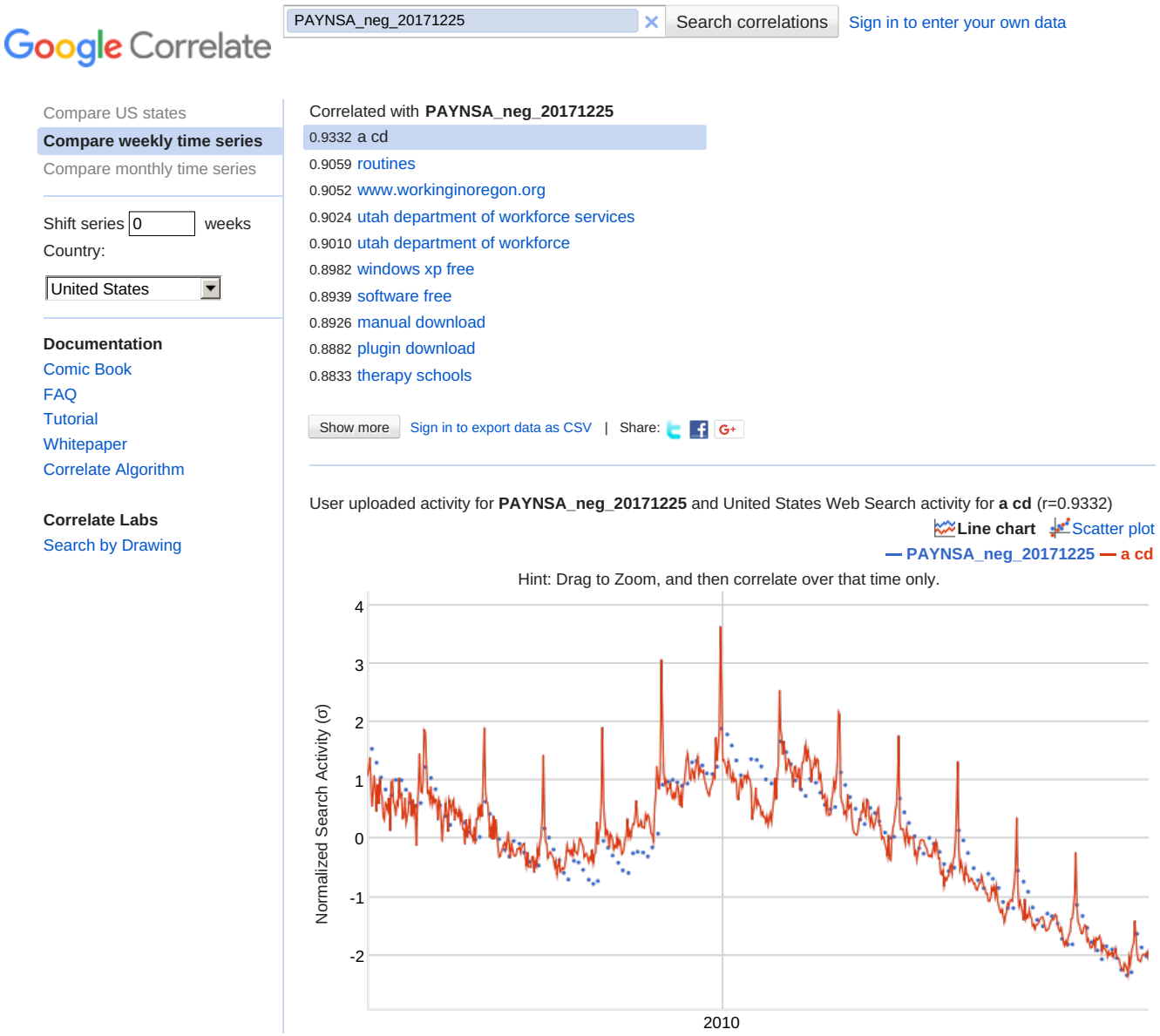
Google Correlate (PAYNSA, vintage=25DEC2017)

...when the target is monthly



Google Correlate (PAYNSA, vintage=25DEC2017)

...when the target is the inverse of employment



Google Correlate: suggestions for in-sample searches

More likely to get meaningful terms (as a result, stickier/tighter correlation) if

- enough sample size: at least one business cycle (JAN2004 - present) $\Rightarrow$ do we need to roll windows?
- sub-sample: spurious in boom times; (more) meaningful in bust times
- the target is more frequent: weekly instead of monthly
- the target is not transformed: not diff, not %chg, not SA
- where there is a 'problem': job destruction instead of job creation, 'mens shirt' (retail sales in Scott and Varian (2014))
- variation is large compared to seasonal variation

Seasonality

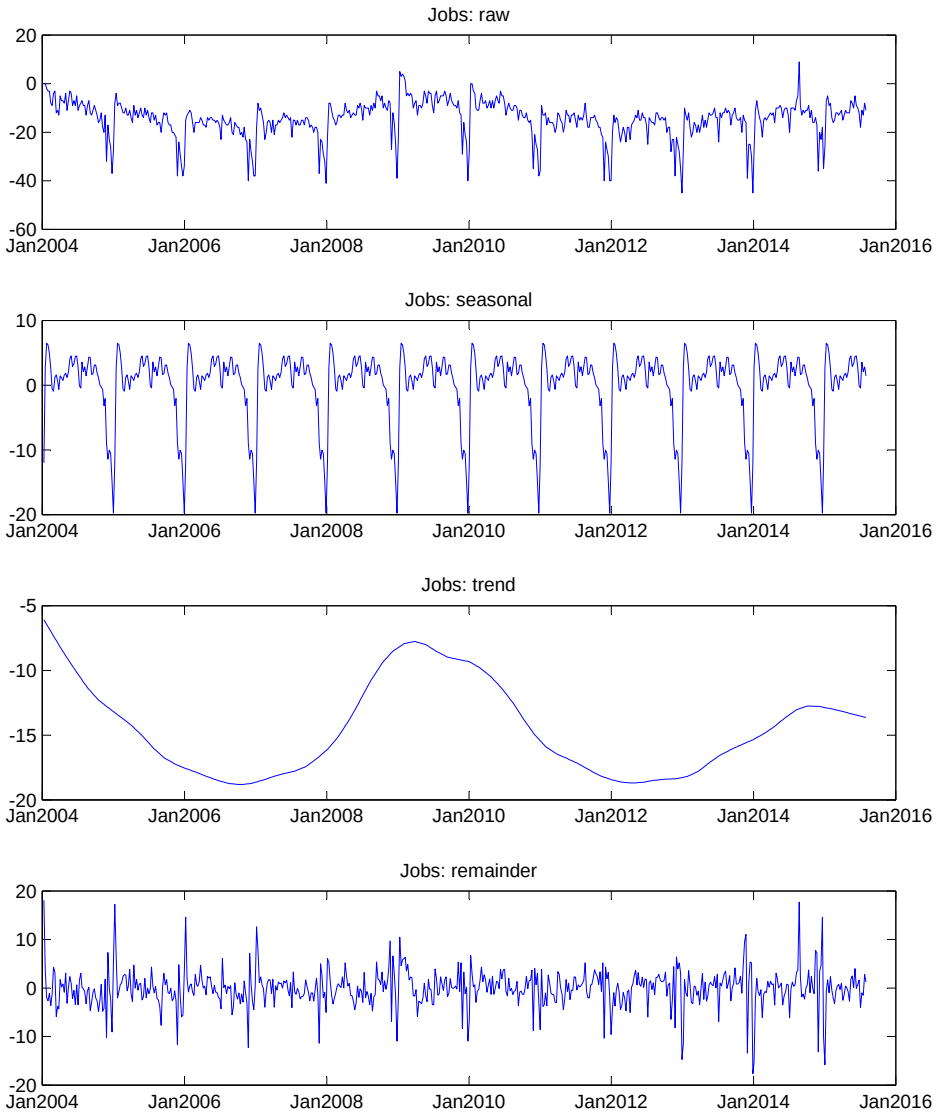
- Seasonality
  - Choi & Varian (2009, 2012) deseasonalized $\texttt{Jobs}_t$ and $\texttt{Welfare}_t$ using `stl` command in `R`
  - `stl` is a local regression method (Cleveland et al, 1990 JOS)
  - discretion in choosing smoothing parameters, s.window

  $\Rightarrow$ A more general question: how to model data with seasonality? How will it affect forecast?

Seasonality

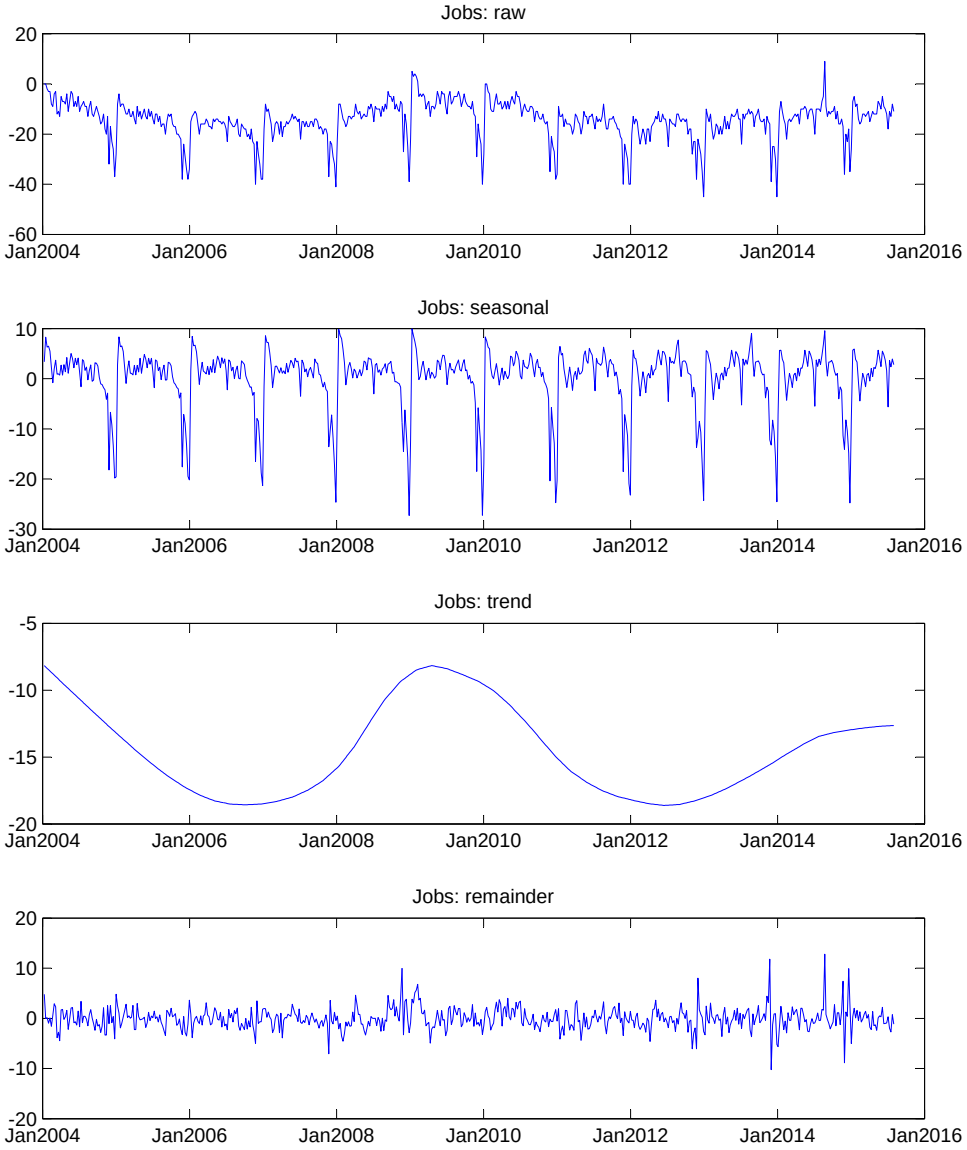
stl: s.window=periodic (Choi and Varian, 2012) (a)

There is still seasonal component in the remainder!!!



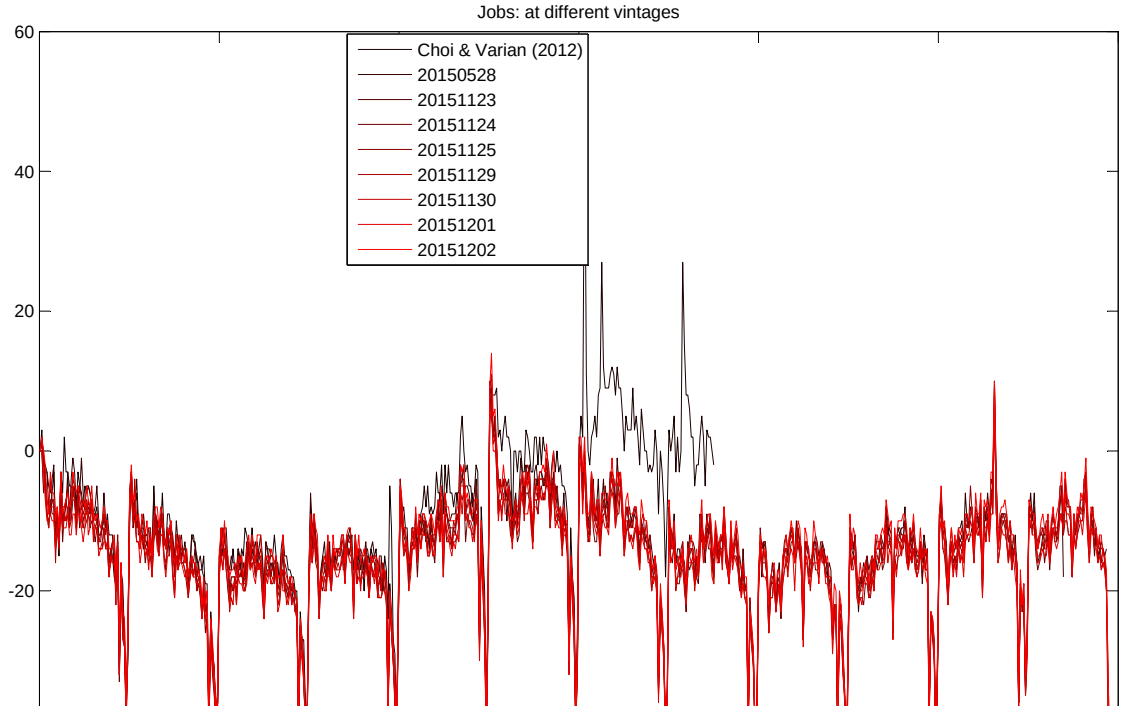
Seasonality

stl: s.window=7 (b) This smoothing parameter is much better!



Sampling error

- 'jobs' category index: data available in 2015 vs. Choi & Varian (2012)
- big differences might be caused by algorithm changes (Lazer et al (2014, Science))
- small differences are sampling errors ⇒ average out if necessary
- a single search term series is less subject to algorithm changes, than a category



Common approach in the literature of search data

- A common approach: compare the forecast performance of an augmented AR(1) with an AR(1) benchmark

$$\phi(L)y_t = \alpha + \beta x_t + \theta(L)\epsilon_t.$$

- y_t: target to forecast
- x_t: Google Trends
- A rule of thumb to pick $\phi(.)$ and $\theta(.)$ or
- A horserace of AR, ARMA models, to get the smallest RMSE
- mixed frequency
  - average weeks over a month or a quarter
  - select one or two specific weeks

$\Rightarrow$ Fondeur and Kameré (2013): the dataset is generally 'impoverished' by doing so

Issues with this approach

- most exercises are not real-time
- data issue: dynamic algorithm
- simple time series models cannot:
  - deal with mixed frequency
  - deal with ragged edge
  - it is not clear whether Google data is helpful, given other data source $\Rightarrow$ main question in this paper!

Example: Choi and Varian (2012)

- Data:
- Initial claims (ICSA): weekly, released on Thursday
- Google Trends data: weekly, released on Monday
  - `Local/Jobs`
  - `Society/Social Sevices/Welfare & Unemployment`
- An augmented AR(1) model:

$$y_t = \alpha + \beta y_{t-1} + \gamma Jobs_t + \delta Welfare_t + \epsilon_t$$

is compared with an AR(1)

Initial Claims ICSA: Baseline forecast

Compare the forecast performance of:

- a random walk (RW)
- a random walk with drift
- an AR(1)
- an AR(1) Augmented with Google search data
  - (a) Choi & Varian data, s.window='periodic' (Choi & Varian, 2012)
  - (b) Choi & Varian data, s.window=7
  - (c) Data available on 28MAY2015, s.window='periodic'

Initial Claims ICSA: Baseline forecast

Evaluation period		RW	RW with drift	AR(1)	AR(1) Aug.(a)	AR(1) Aug.(b)	AR(1) Aug.(c)
from	to	MAE	1-ratio	1-ratio	1-ratio	1-ratio	1-ratio
03 Nov 2007	02 Jul 2011	0.0313	1.0038	1.1117	1.1778	1.1750	1.1591
01 Dec 2007	30 Jun 2009	0.0312	0.9982	1.2754	1.1024	1.0630	0.9937
01 Mar 2009	01 May 2009	0.0301	1.0013	1.0182	0.7956	1.0630	0.9520
01 Dec 2009	01 Feb 2010	0.0354	0.9980	1.0072	0.8827	0.9390	1.1222
15 Jul 2010	01 Oct 2010	0.0255	1.0037	0.9911	0.9591	0.9676	0.9543
01 Jan 2011	01 May 2011	0.0516	1.0007	0.9959	0.9894	0.9691	0.9972

from	to	RMFSE	1-ratio	1-ratio	1-ratio	1-ratio	1-ratio
03 Nov 2007	02 Jul 2011	0.0402	1.0017	1.0996	1.1214	1.1187	1.1194
01 Dec 2007	30 Jun 2009	0.0413	0.9986	1.2139	1.0421	1.0241	0.9757
01 Mar 2009	01 May 2009	0.0376	1.0085	1.0132	0.7637	0.8063	0.9144
01 Dec 2009	01 Feb 2010	0.0423	0.9987	1.0085	0.9750	0.9742	1.0982
15 Jul 2010	01 Oct 2010	0.0296	1.0002	1.0032	1.0351	1.0120	1.0298
01 Jan 2011	01 May 2011	0.0586	0.9998	0.9976	0.9664	0.9519	0.9909

All evaluation windows are the same as Choi & Varian (2012). The first row is the whole evaluation window. The rest are the windows, in which Choi & Varian (2012) claim AR(1) Aug (a) outperforms AR(1) and these windows are the 'turning points' for initial claims.

But their result is not very robust.

Model

The dynamic factor model developed in Bańbura et al (2013)

$$X_t = CF_t + \epsilon_t, \epsilon_t \sim \mathcal{N}(0, Q)$$
$$F_t = AF_{t-1} + u_t, u_t \sim \mathcal{N}(0, R)$$

and it is able to

- to maintain parsimony with large panels
- to deal with missing data at the beginning and the end of the dataset
- to handle data of mixed frequencies

There are the features of a nowcasting problem formally framed in Giannone, Reichlin and Small (2008, JME)

*I thank Now-casting Economics Ltd. for making the code available

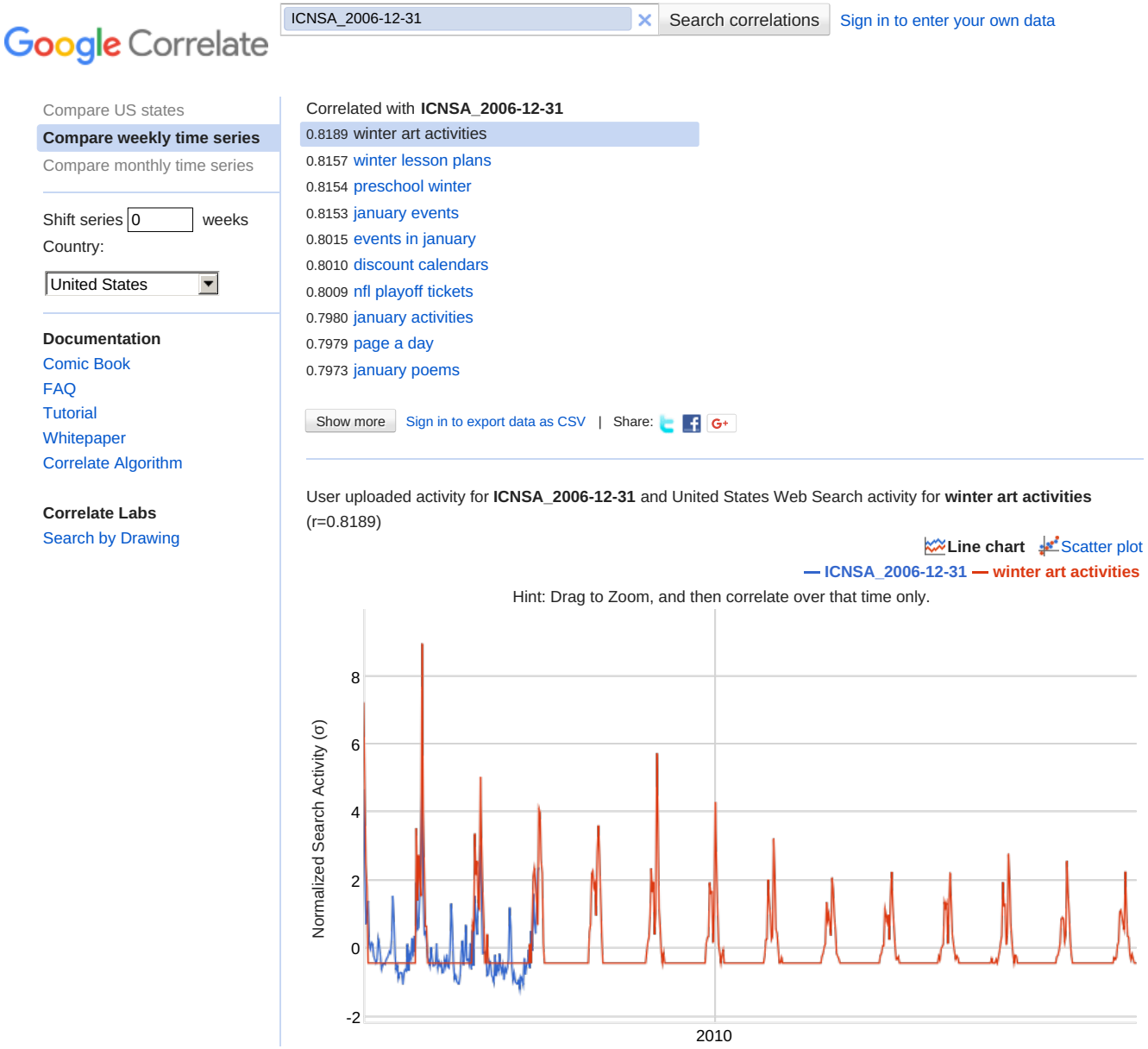
Forecast ICSA: models and variables

Variable	DFM daily	DFM daily w/Trends (c)	DFM daily w/Correlate	DFM weekly w/ Correlate	RW	RW w/drift	AR(1)	AR(1) w/Trends (c)
Industrial Production Index	x	x	x					
Capacity Utilization: Total Industry	x	x	x					
ISM Manufacturing: PMI Composite Index	x	x	x					
Real Disposable Personal Income	x	x	x					
Civilian Unemployment Rate	x	x	x					
All Employees: Total Nonfarm Payrolls	x	x	x					
Real Personal Consumption Expenditures	x	x	x					
Housing Starts: Total: New Privately Owned Housing Units Started	x	x	x					
New One Family Houses Sold: United States	x	x	x					
Manufacturers' New Orders: Durable Goods	x	x	x					
Producer Price Index by Commodity for Final Demand:Finished Goods	x	x	x					
Consumer Price Index for All Urban Consumers: All Items	x	x	x					
Exports of Goods and Services, Balance of Payments Basis	x	x	x					
Imports of Goods and Services: Balance of Payments Basis	x	x	x					
PA Fed Manufacturing Business Outlook Survey: general activity	x	x	x					
Conference Board consumer confidence index	x	x	x					
Advance Retail Sales: Retail and Food Services, Total	x	x	x					
Value of Manufacturers' Total Inventories for Durable Goods Industries	x	x	x					
Value of Manufacturers' Unfilled Orders for Durable Goods Industries	x	x	x					
Bloomberg consumer comfort index	x	x	x					
Chicago Fed National Financial Conditions Index	x	x	x					
Initial Claims, Seasonally Adjusted	x	x	x	x	x	x	x	x
Continued Claims (Insured Unemployment)	x	x	x					
Covered Employment	x	x	x					
Google Trends search category: jobs		x						x
Google Trends search category: welfare & unemployment		x						x
Google Correlate (ICNSA, 1-50)			x	x				
Crude Oil: West Texas Intermediate (WTI) - Cushing, Oklahoma	x	x	x					
10-Year Treasury Constant Maturity Rate	x	x	x					
3-Month Treasury Bill: Secondary Market Rate	x	x	x					
Trade Weighted Exchange Index: Major Currencies	x	x	x					
S&P 500	x	x	x					

Notes. The Google Correlate (ICNSA, 1-50) series are the top 50 most correlated series with initial claims seasonally unadjusted (ICNSA), downloaded on Dec 25, 2017, deseasonalized with s.window='periodic', vintage by vintage.

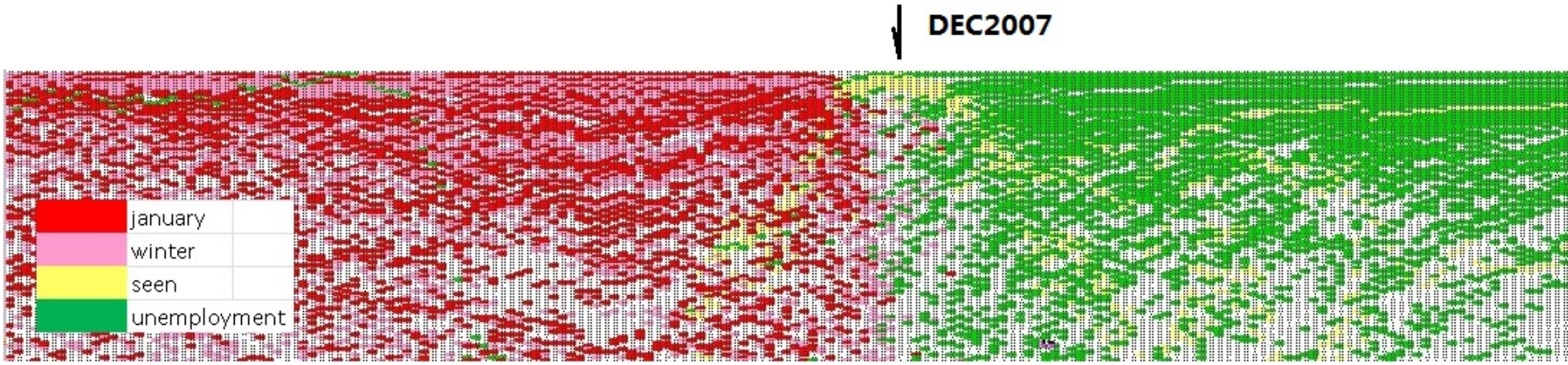
Google Correlate: ICNSA, vintage=31DEC2006

...upload ICNSA available on 31DEC2006, get 100 terms



Google Correlate: ICNSA, 31DEC2006-25DEC2011

...repeat this until 25DEC2011...



Google Trends: alabama unemployment

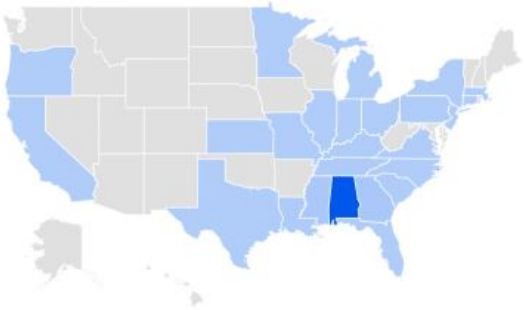
alabama unemployment
Search term

+ Compare

United States | 2004 - present | All categories | Web Search

Interest by subregion

Subregion



1	Alabama	100
2	Georgia	5
3	Mississippi	4
4	Louisiana	2
5	Tennessee	2

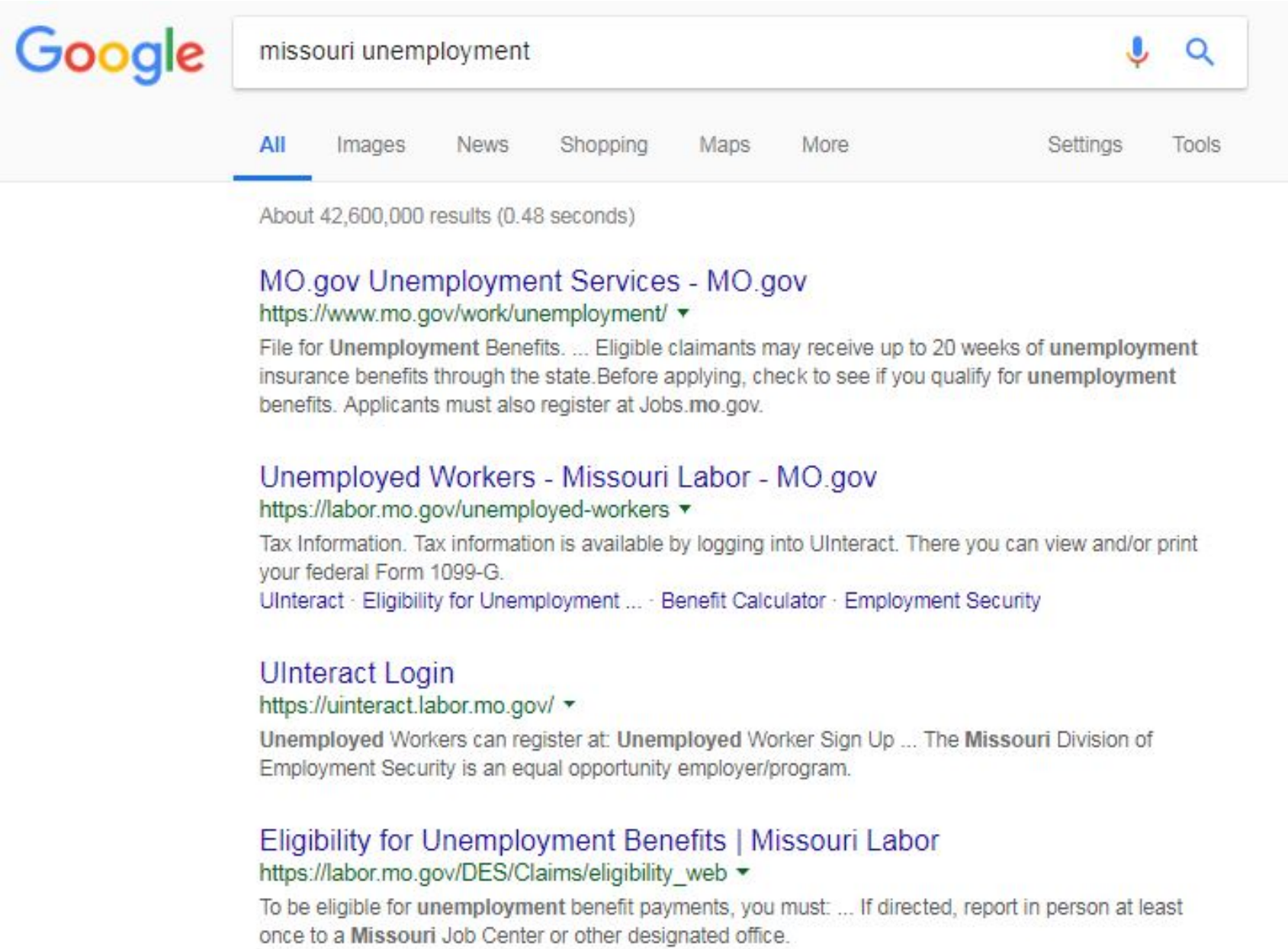
Related topics — Top

1	Alabama - US State	100
2	Unemployment - Topic	97
3	Unemployment benefits - Topic	8
4	Filing - Topic	5
5	Office - Building use	5

Related queries — Top

1	unemployment in alabama	100
2	unemployment office alabama	54
3	unemployment office	53
4	unemployment claim alabama	48
5	unemployment claim	46

Google search: alabama unemployment

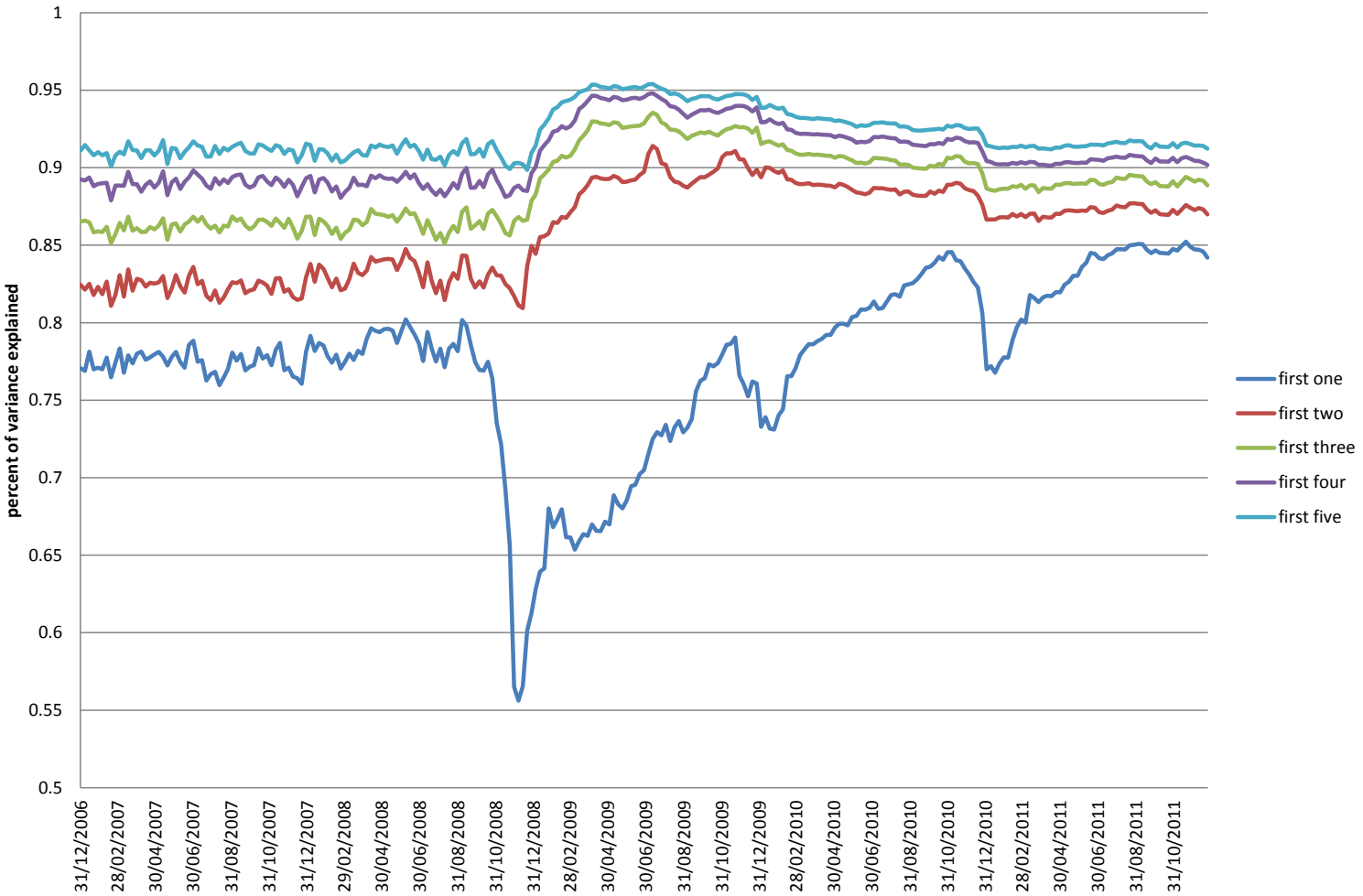


Google Correlate: ICNSA, 31DEC2006-25DEC2011

...the first five principal components of the Correlate data, vintage by vintage

⇒ a structural break around the crisis

⇒ at least two factors are needed

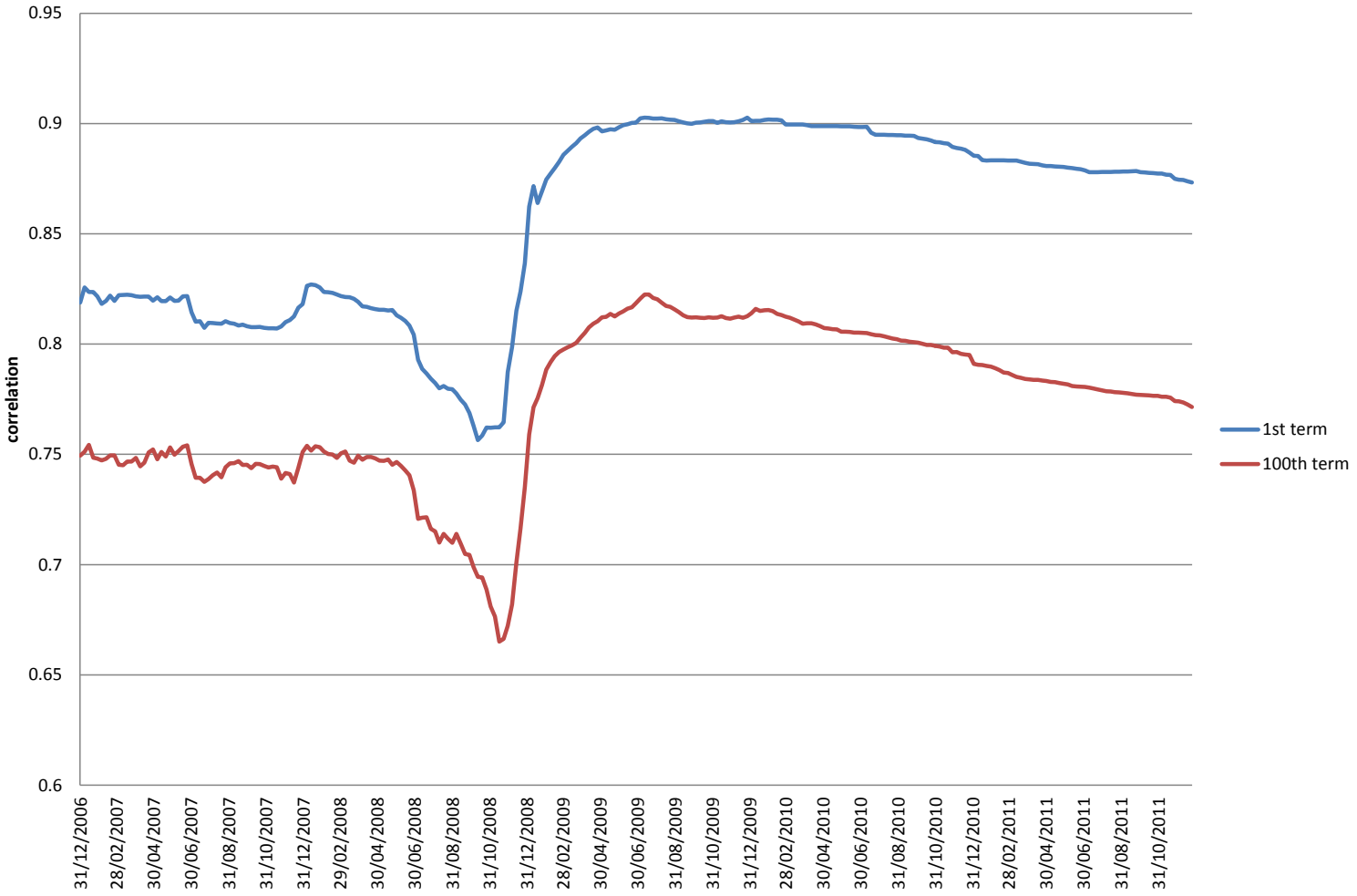


Google Correlate: ICNSA, 31DEC2006-25DEC2011

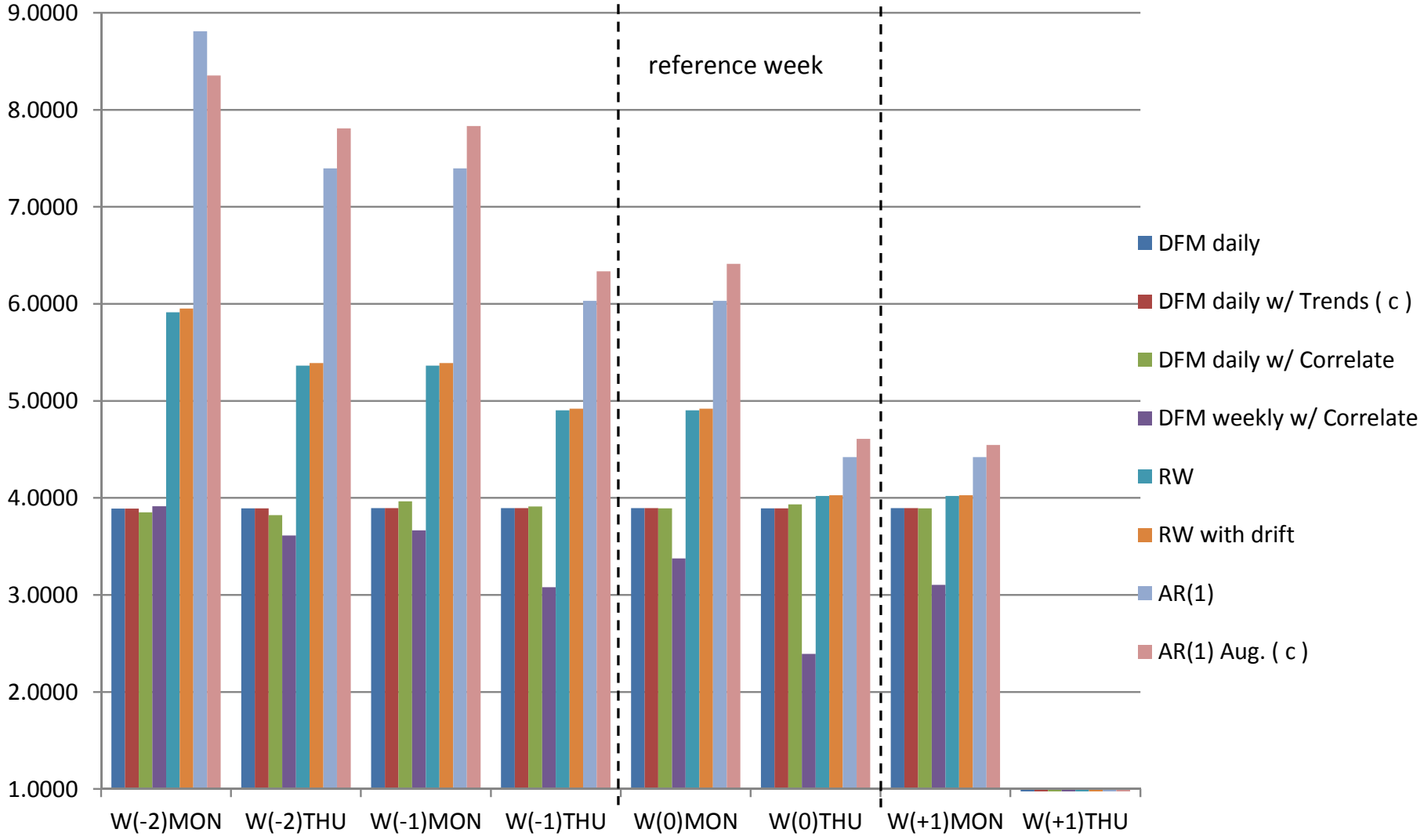
...the correlation between the target and the 1st and the 100th term

⇒ the correlation is tight, c.f. GDP x consumption

⇒ rolling window?



Forecast ICSA: RMSE



Prediction accuracy test: ICSA

Panel B: DM test, p-value, DFM weekly w/ Correlate against alternative models

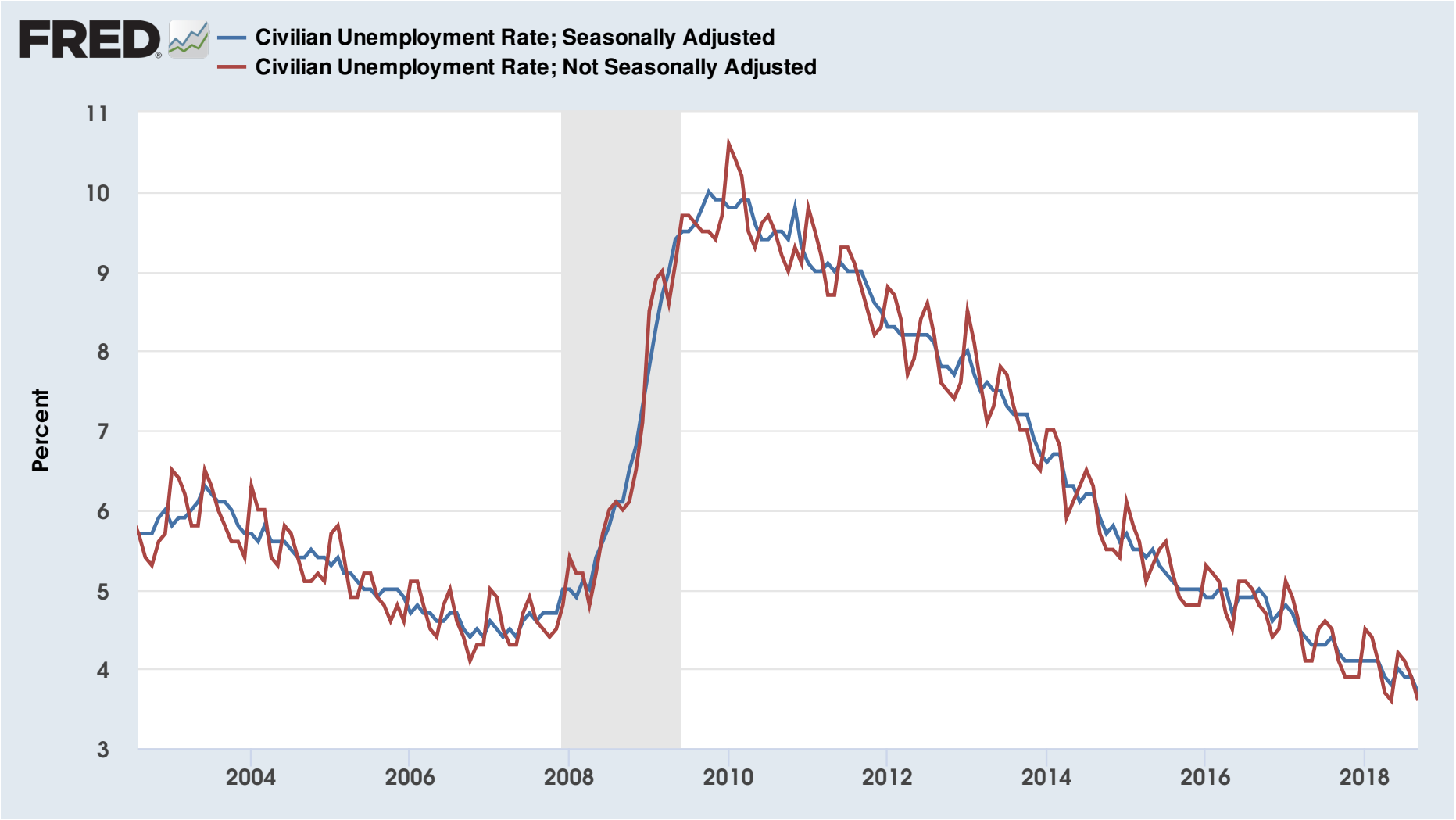
	W(-2)MON	W(-2)THU	W(-1)MON	W(-1)THU	W(0)MON	W(0)THU	W(+1)MON	W(+1)THU
DFM daily	0.81	0.69	0.76	0.25	0.54	0.00	0.02	release
DFM daily w/ Trends (c)	0.81	0.69	0.76	0.25	0.54	0.00	0.02	release
DFM daily w/ Correlate	0.81	0.69	0.76	0.25	0.53	0.00	0.02	release
DFM weekly w/ Correlate	NaN	NaN	NaN	NaN	NaN	NaN	NaN	release
RW	0.12	0.00	0.08	0.01	0.04	0.00	0.03	release
RW with drift	0.13	0.01	0.09	0.01	0.04	0.00	0.03	release
AR(1)	0.03	0.02	0.01	0.00	0.00	0.00	0.00	release
AR(1) w/ Trends (c)	0.04	0.01	0.03	0.00	0.01	0.00	0.00	release

Panel C: DM test, p-value, DFM daily w/ Correlate against alternative models

	W(-2)MON	W(-2)THU	W(-1)MON	W(-1)THU	W(0)MON	W(0)THU	W(+1)MON	W(+1)THU
DFM daily	0.64	0.55	0.45	0.51	0.53	0.66	0.53	release
DFM daily w/ Trends (c)	0.67	0.61	0.47	0.53	0.53	0.67	0.54	release
DFM daily w/ Correlate	NaN	NaN	NaN	NaN	NaN	NaN	NaN	release
DFM weekly w/ Correlate	0.19	0.31	0.24	0.75	0.47	1.00	0.98	release
RW	0.00	0.00	0.00	0.00	0.00	0.16	0.18	release
RW with drift	0.00	0.00	0.00	0.00	0.00	0.15	0.17	release
AR(1)	0.05	0.04	0.03	0.01	0.01	0.00	0.00	release
AR(1) w/ Trends (c)	0.00	0.00	0.01	0.00	0.00	0.00	0.01	release

Notes. Sample = 01JAN1983-02JUL2011 for daily DFMs and 10JAN2004-02JUL2011 for the rest, evaluation = 10NOV2007-02JUL2011. A week is defined to end on Saturday, so 10NOV2007(SAT) indicates the week 28OCT2007-10NOV2007. Rolling window = NO. Figure 9 plots the results of Panel A.

Forecast UNRATE: UNRATE vs. UNRATENSA



Forecast UNRATE: models and variables

Variable	DFM daily	DFM daily w/ Trends(c)	DFM daily w/ Correlate	DFM weekly w/ Correlate	RW	RW w/ drift	AR(1)
Real Gross Domestic Product	x	x	x				
Industrial Production Index	x	x	x				
Capacity Utilization: Total Industry	x	x	x				
ISM Manufacturing: PMI Composite Index	x	x	x				
Real Disposable Personal Income	x	x	x				
Civilian Unemployment Rate	x	x	x	x	x	x	x
All Employees: Total Nonfarm Payrolls	x	x	x				
Real Personal Consumption Expenditures	x	x	x				
Housing Starts: Total: New Privately Owned Housing Units Started	x	x	x				
New One Family Houses Sold: United States	x	x	x				
Manufacturers' New Orders: Durable Goods	x	x	x				
Producer Price Index by Commodity for Final Demand: Finished Goods	x	x	x				
Consumer Price Index for All Urban Consumers: All Items	x	x	x				
Exports of Goods and Services, Balance of Payments Basis	x	x	x				
Imports of Goods and Services: Balance of Payments Basis	x	x	x				
PA Fed Manufacturing Business Outlook Survey: general activity	x	x	x				
Conference Board consumer confidence index	x	x	x				
Advance Retail Sales: Retail and Food Services, Total	x	x	x				
Value of Manufacturers' Total Inventories for Durable Goods Industries	x	x	x				
Value of Manufacturers' Unfilled Orders for Durable Goods Industries	x	x	x				
Bloomberg consumer comfort index	x	x	x				
Chicago Fed National Financial Conditions Index	x	x	x				
Initial Claims, Seasonally Adjusted	x	x	x				
Continued Claims (Insured Unemployment)	x	x	x				
Covered Employment	x	x	x				
Google Trends search category: jobs		x					
Google Trends search category: welfare & unemployment		x					
Google Correlate (UNRATENSA, 1-50)			x	x			
Crude Oil: West Texas Intermediate (WTI) - Cushing, Oklahoma	x	x	x				
10-Year Treasury Constant Maturity Rate	x	x	x				
3-Month Treasury Bill: Secondary Market Rate	x	x	x				
Trade Weighted Exchange Index: Major Currencies	x	x	x				
S &P 500	x	x	x				

Notes. Same variable as in Table 2, except for Google Correlate (UNRATENSA, 1-50), but the deseasonalization is the same as before.

Google Correlate: UNRATENSA, vintage=07FEB2007

Google Correlate

UNRATENSA_2007-01-01 | Search correlations | Sign in to enter your own data

Compare US states
Compare weekly time series
Compare monthly time series

Shift series 0 weeks
Country:
United States

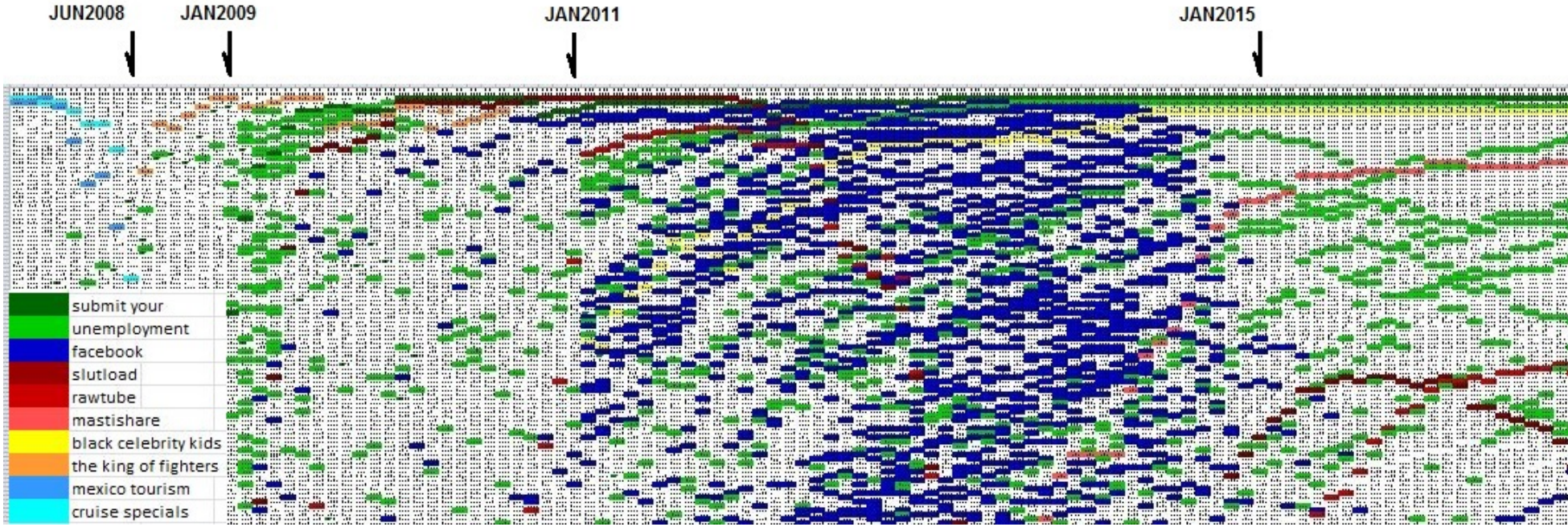
Documentation
Comic Book
FAQ
Tutorial
Whitepaper
Correlate Algorithm

Correlate Labs
Search by Drawing

Correlated with **UNRATENSA_2007-01-01**

0.9275	cruise specials
0.9214	mexico tourism
0.9181	charter flights
0.9161	travel packages
0.9160	buddy icons for aim
0.9131	travel agencies
0.9110	bass tablature
0.9098	discount vacations
0.9091	osex
0.9087	sonic mydvd
0.9061	mydvd
0.9050	discount travel
0.9050	vst
0.9047	discount cruises
0.9045	eas
0.9043	las vegas vacation packages
0.9038	travel package
0.9024	interior decorating
0.9023	package
0.9021	discount vacation
0.9021	allegro
0.9018	grand flamenco
0.9010	guides
0.9000	flamingo hilton
0.8998	free chinese
0.8996	nht speakers
0.8983	web page software
0.8976	speed tweaks
0.8971	las vegas travel
0.8968	torisan

Google Correlate: UNRATENSA, 07FEB2007-07JAN2018

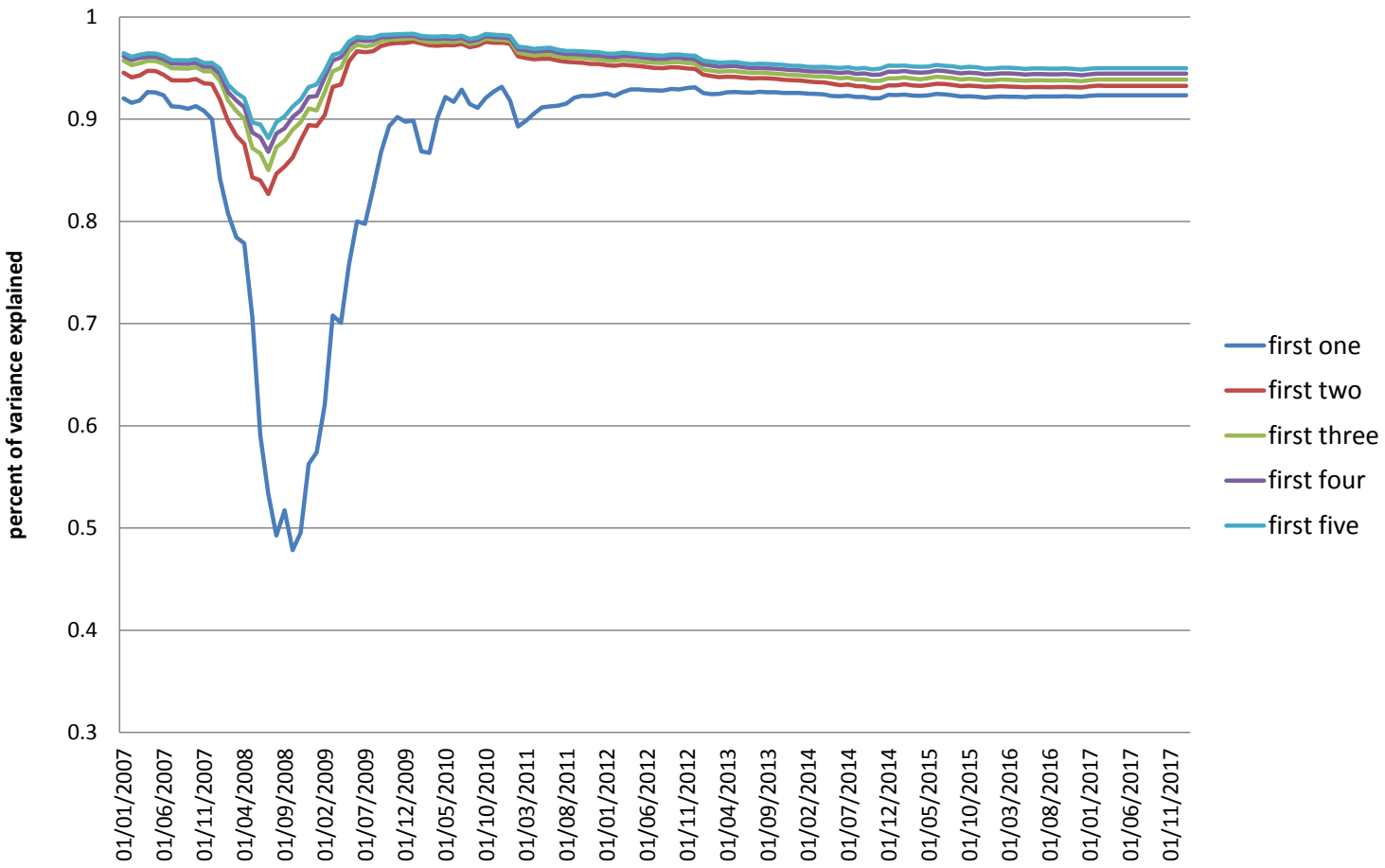


other grids: other social networks (#DeleteFacebook?), other porn sites, other hoilday destinations, online games, online tv streaming

during crisis: 'where can i download'

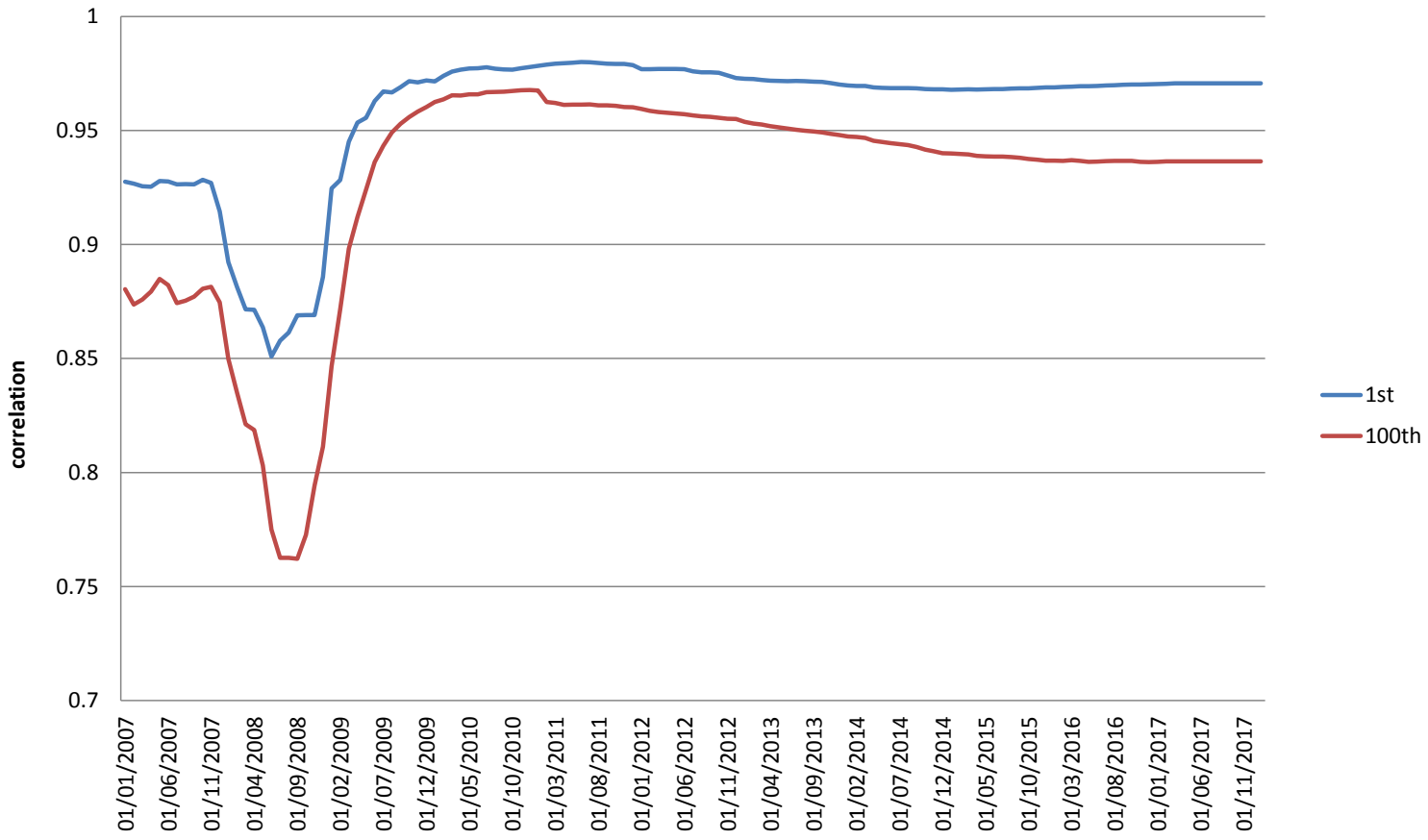
Google Correlate: UNRATENSA, 07FEB2007-07JAN2018

...first five principal components of the Correlate data, vintage by vintage

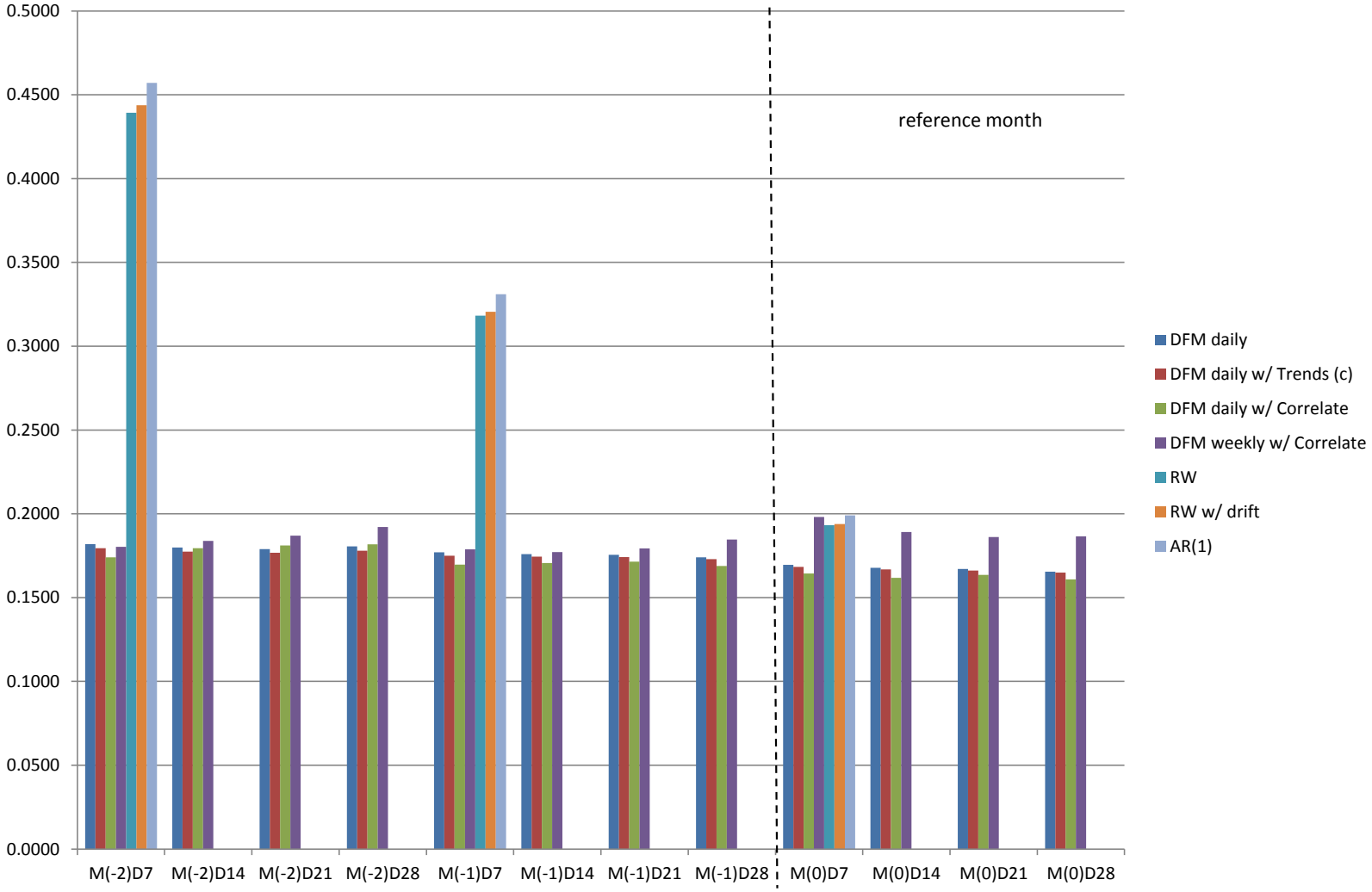


Google Correlate: UNRATENSA, 07FEB2007-07JAN2018

...the correlation between the target and the 1st and the 100th term



Forecast unemployment rate UNRATE: RMSE



Prediction accuracy test: UNRATE

Panel A: RMSE

	M(-2)D7	M(-2)D14	M(-2)D21	M(-2)D28	M(-1)D7	M(-1)D14	M(-1)D21	M(-1)D28	M(0)D7	M(0)D14	M(0)D21	M(0)D28	M(+1)D7
DFM daily	0.1819	0.1799	0.1789	0.1806	0.1770	0.1760	0.1755	0.1740	0.1696	0.1678	0.1671	0.1655	release
DFM daily w/ Trends (c)	0.1795	0.1775	0.1767	0.1779	0.1750	0.1744	0.1742	0.1729	0.1683	0.1669	0.1662	0.1649	release
DFM daily w/ Correlate	0.1740	0.1794	0.1811	0.1818	0.1697	0.1706	0.1715	0.1688	0.1644	0.1618	0.1635	0.1608	release
DFM weekly w/ Correlate	0.1803	0.1839	0.1870	0.1922	0.1787	0.1772	0.1793	0.1846	0.1982	0.1891	0.1862	0.1865	release
RW	0.4393				0.3183				0.1932				release
RW w/ drift	0.4438				0.3205				0.1939				release
AR(1)	0.4571				0.3310				0.1991				release

Panel B: DM test, p-value, DFM daily w/ Correlate against alternative models

	M(-2)D7	M(-2)D14	M(-2)D21	M(-2)D28	M(-1)D7	M(-1)D14	M(-1)D21	M(-1)D28	M(0)D7	M(0)D14	M(0)D21	M(0)D28	M(+1)D7
DFM daily	0.00	0.36	†NaN	†NaN	0.05	0.00	0.07	0.06	†NaN	†NaN	0.28	0.31	release
DFM daily w/ Trends (c)	0.00	0.49	†NaN	†NaN	0.04	0.00	0.03	0.03	†NaN	†NaN	0.27	0.30	release
DFM daily w/ Correlate	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	release
DFM weekly w/ Correlate	0.00	0.00	0.00	0.00	0.01	0.00	0.05	0.01	0.08	0.10	0.08	0.04	release
RW	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.03	0.08	0.11	release
RW w/ drift	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.04	0.06	0.10	0.12	release
AR(1)	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.02	0.06	0.08	0.12	0.15	release

Panel C: DM test, p-value, DFM daily w/ Trends (c) against alternative models

	M(-2)D7	M(-2)D14	M(-2)D21	M(-2)D28	M(-1)D7	M(-1)D14	M(-1)D21	M(-1)D28	M(0)D7	M(0)D14	M(0)D21	M(0)D28	M(+1)D7
DFM daily	0.00	0.01	0.02	0.00	0.37	0.67	0.90	0.85	0.65	0.95	0.97	0.96	release
DFM daily w/ Trends (c)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	release
DFM daily w/ Correlate	1.00	0.51	†NaN	†NaN	0.96	1.00	0.97	0.97	†NaN	†NaN	0.73	0.70	release
DFM weekly w/ Correlate	†NaN	0.00	†NaN	†NaN	†NaN	†NaN	0.15	0.00	0.09	0.11	0.11	0.04	release
RW	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.04	0.06	0.06	release
RW w/ drift	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.05	0.07	0.08	0.08	release
AR(1)	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.02	0.08	0.09	0.11	0.11	release

Notes. Sample = JAN1983-DEC2016 for daily DFMs and JAN2004-DEC2016 for the rest, evaluation = MAY2007-DEC2016. Rolling window = NO. † p-value is not available because the spectral density at zero is not estimatable with the finite sample. Figure 16 plots the results of Panel A.

Google Correlate: in-sample vs. out-of-sample

- for in-sample,
  - already discussed: how to find meaningful terms in-sample
  - in-sample findings (not rolling window):
  - vintages before 2008, spurious correlation
  - vintages after 2008, more meaningful terms, some perhaps never expected
  - do we want rolling window?
- for out-of-sample,
  - do we actually mind the correlation being spurious???
  - in the short run, no, as long as the correlation is sticky/has inertia for a while
  - in the long run, yes
  - in the past, when a correlation collapses, we did't know what else is out there $\Rightarrow$ unknown unknowns (BOE: JSA/Universal Credit)
  - but now, we can re-search $\Rightarrow$ value-added of Google Correlate

Next...

- state-level
- measure time use/job search of the unemployed people

Litterman (1986): the problem of economic forecasting

- The sense in which there is only a limited amount of data is perhaps not so obvious. After all, the total quantity of economic data processed and available on computer data bases today is enormous.
- When we talk of forecasting the economy, we usually are referring the problem of predicting ... economic aggregates.
- Most forecasts are short to medium term, and much of the variation in these aggregate variables at these horizons seems to be generated by an underlying phenomenon, the business cycle.
- The sense in which data are scarce is that the entities that we are really trying to measure and forecast are business cycles, and the number of observations of business cycles relevant for use in forecasting today's economy is relatively small.

Stephens-Davidowitz (2017)[1]: Four power of Big Data

- the re-imagining of what qualifies data: ... it can offer you new kinds of information to study - information that had never previously been collected
- certain online sources get people to admit things they would not admit anywhere else
- Big Data allows us to meaningfully zoom in on small segments of a dataset to gain new insight on who we are
- it makes randomized experiments, which can find truly causal effects, much, much easier to collect

1

Economist Best Book of the Year
PBS NewsHour Book of the Year
Amazon Best Book of the Year in Business and Leadership
New York Times Bestseller

Conclusion

This paper

- gives a comprehensive literature using search data for forecastings/nowcasting
- discusses various issues of Google search data and provides some suggestions for its use
- provides a framework of using new type of web-based search data into a nowcasting model
- Google Correlate help improve the predictions of initial claims and unemployment