I Z A Institute
of Labor Economics

IDSC
INTERNATIONAL
DATA SERVICE CENTER

Doing and Enabling Research with Data

IDSC Research Data Center of IZA

Nikos Askitas, IZA – Institute of Labor Economics

March 7, 2019

The future of labour market research in Central-Eastern European countries,
INGRID/CELSI, Falkensteiner hotel Bratislava

Table of contents

1. Introduction
2. IDSC - The RDC of IZA
3. The internet as a data source for Social Science
4. Unemployment and Health
5. Shares: referenda, Case/Shiller Index
6. Hourly Data: Traffic Jams, [Askitas, 2016a]
7. Argentinian Inflation
8. IDSC focus on Internet data

Introduction

About myself

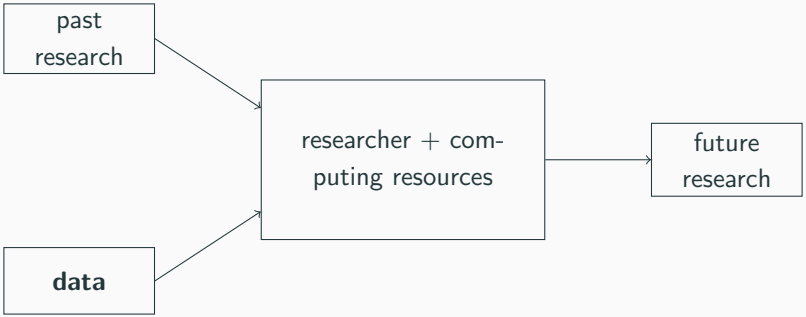
I am a mathematician by training but also by conviction currently working as an Economist/Data Scientist. I am in charge of the ICT and the RDC at IZA. In my life as a mathematician I was a *low dimensional topologist* with research on *four dimensional topology and classical knot theory*. As strange as that might sound my transition from mathematics, through technology to economics has been quite natural although at times serendipitous and overall the product of economic disruption. Every hat I ever wore comes into play to one extend or the other in my current work.

About this talk

Everything I say is the current snapshot of close to 20 years of work at the Institute. I hope that this will be useful to one extend or the other for a good portion of those present. There may be some ideas here and there that apply to filling the gaps in the representation of minorities and vulnerable groups in the data.
Making good on the many hats I have put on over the years let me start with a research model. It might help us put everything I will say into perspective.

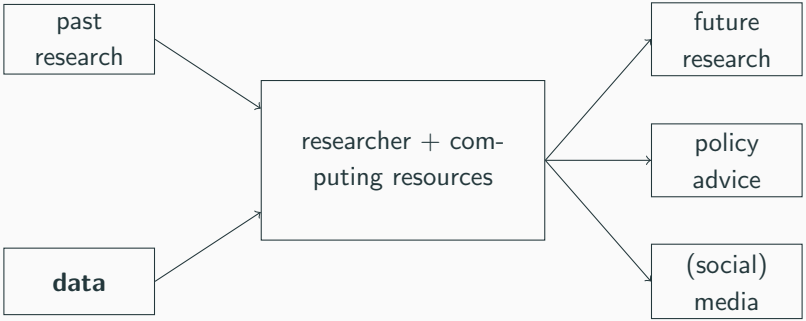
scientific research as an I/O process

Data: *sine qua non*



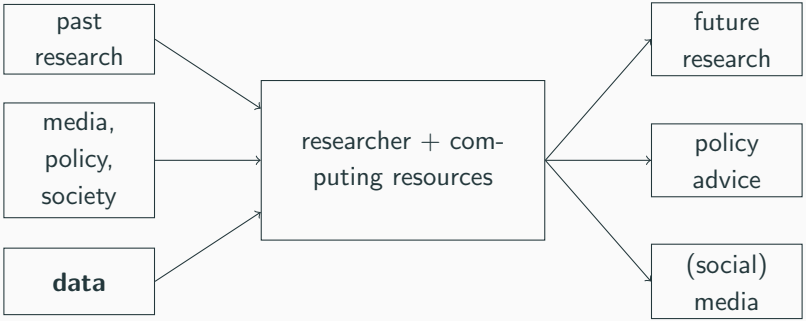
empirical scientific research as an I/O process

Output must have many dimensions



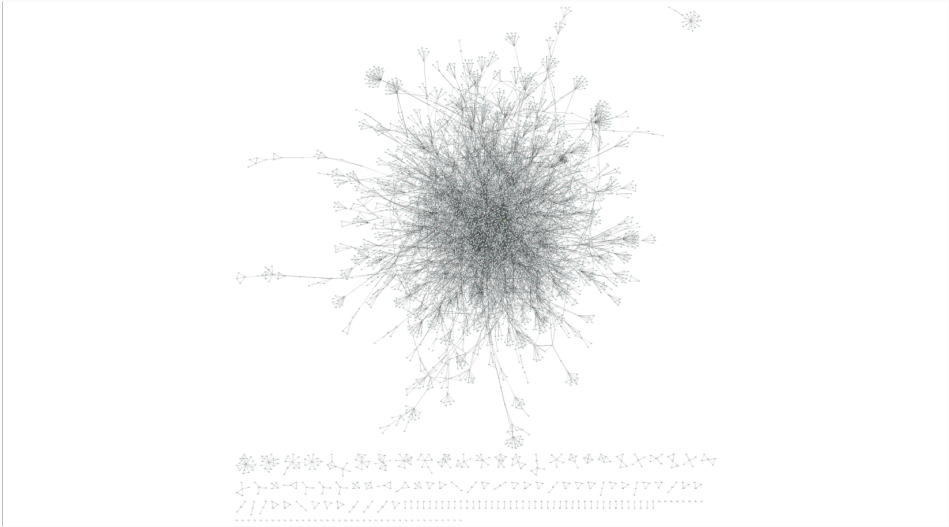
multidimensional output

Input must have many dimensions



multidimensional input

IZA DP co-authorship network



Follow the data

One can arguably play around with some of the ingredients modulating how much of each one takes into the mix. The one ingredient you cannot do without in empirical research (and consequently in evidence based policy advice) is **data**. This is particularly important and evident at a time such as ours where we observe irrationalities such as **post-truth politics** and the resurgence of **identity politics** but also in the presence of the so called **replication crisis** in (social) science.

IDSC - The RDC of IZA

IDSC – The Research Data Center of IZA

Empirical research is the foundation of evidence based policy advice, and data is its key ingredient. However, availability of empirical data can be limited for various reasons. The IDSC contributes two major pillars to the Institute's own mission of bridging the divide between academic research and policy advice on the one hand and between data silos and researchers on the other: research data and digital technology

- development of tools for secure remote analysis of sensitive datasets,
- dataset storage/provisioning for replication/secondary analysis, and
- active participation in communities and efforts in all things data

Our special focus is on developing tools for extracting internet data as an existing and innovative source of data for empirical research on labor market issues. Our own focused research underscores the relevance of data research for social science disciplines.

Enabling Research

- Tools for research, access, collaboration etc
- Data services (provisioning, documentation, citation etc)
- Teaching, data advocacy, networking

Tools (see econ.tools)

- josua.iza.org
- statsdirect.org
- GPU lab (nvidia GPUs) for high end computing (ML etc).
- cloud.iza.org/cloud.briq-institute.org and git.econ.tools
- jobs.iza.org, Conference Management System (under development)

Services

- www.iza.org/apps/citation/
- datasets.iza.org/
- ed.iza.org

Teaching, data advocacy

- Text Mining with Python for Economists
- Data Seminar at IZA
- www.eddi-conferences.eu
- Matching workers and Jobs Online
- wol.iza.org/opinions/a-data-tax-for-a-digital-economy

Networking and Cooperations

- RatSWD – German Data Forum (founding and actively contributing member): network of 32 RDCs in Germany covering a large portion of Social Science.
- GESIS – Leibniz Institute for the Social Sciences (EDDI conference)
- IAB – Institute for Employment Research at the German Federal Employment Agency (JoSuA, IZA Evaluation Dataset, etc)
- IQB – Institute for Educational Quality Improvement, Humboldt University (JoSuA)
- WIF – Wage Indicator Foundation (Data provisioning)
- BiBB – Federal Institute for Vocational Education and Training (JoSuA etc)

Policy Projects

- Social Media in SME
  - Partners: T-Systems, IHK Bonn/Rhein-Sieg, Synergie GmbH, Fraunhofer FIT
  - Funding: Federal Ministry for Economic Affairs and Energy (BMWi)
  - Duration: 2012-2013
- Online talent platforms, labour market intermediaries and the changing world of work
  - Partners: IZA, CEPS
  - Funding: EU Sectoral Social Dialogue Committee (SSDC), World Employment Confederation-Europe, UNI-Europa
  - Duration: 2017
- IRSDACE: Industrial Relations and Social Dialogue in the Age of Collaborative Economy
  - Partners: IZA, CELSI, CEPS, COP, FA
  - Funding: European Commission DG Employment
  - Duration: 2017-2018
- Emerging Labour Market Data Sources towards Digital TVET
  - Partners: PARIS21
  - Funding: German Federal Ministry of Economic Cooperation and Development
  - Duration: 2018

The internet as a data source for Social Science

"own data"

For the last 10 years we ask people to tell us which if any datasets they used in their IZA DP when they upload it. Most people do. The most popular of the well known dataset, SOEP, comes in a distant third behind "own data" and "experiment"!

Some examples from (mostly) my own research

As an economist I have been thematically all over the place because:

- (internet) data has been the driving force and
- by running ICT and the RDC at IZA I have a day job granting me more degrees of freedom than usual research wise.

On the other hand of course Economics is a derivative discipline founded on Math, Psychology, Sociology, etc., emerging data contains so much more than just economics and it all feeds into the economy so multidisciplinarity is essential. This is particularly acute at a time when the public debate is shifting from being about economics (Ricardian globalisation) to being about trade tariffs and identitiy politics.

A data taxonomy for the working economist

Many ways to talk about Internet as a data source. Lets start descending into it:

- Official Data: recording transactions of the state (e.g. with its citizens, companies and other economic entities)
- Non-Official data: e.g. proprietary transaction data of non state entities (firms), internet data, new data.

Historically the collection of data in both categories was meant to serve a rather narrow, operations-specific purpose but we now know that all of these data are vital for understanding and fine tuning our economies and for making sure we keep them in desirable equilibria. State of data still sub-optimal.

Figure 1: The diary of Merer (4,500 years old) describes the daily life of workers who took part in the building of the Great Pyramid of Giza

Official Data: occasionally rigged

- Argentinian Inflation: "bogus accounting 2007-2017" **(economist.com, May 2017)**
- *China's 2015 GDP Was Exaggerated By Fake Data, Analysis Shows*, **(bloomberg.com, February 2018)**
- *"... the fact that basic data on domestic debt are so opaque and difficult to obtain is proof that* **governments will go to great lengths to hide their books when things are going wrong**, *just as financial institutions have done in the contemporary financial crisis"* **Reinhart, Carmen M.. This Time Is Different: Eight Centuries of Financial Folly . Princeton University Press.**
- Lies, damned lies and Greek statistics, **January 2013 ft.com** *"Greece has brought criminal charges against, Andreas Georgiou, the official responsible for measuring the country's debt"*
- German official data not all not always immune...

Need for multiple sources of data

- Measure differently, in various contexts and check whether the results hold.
- New sources of data (internet) can play an important role

Examples

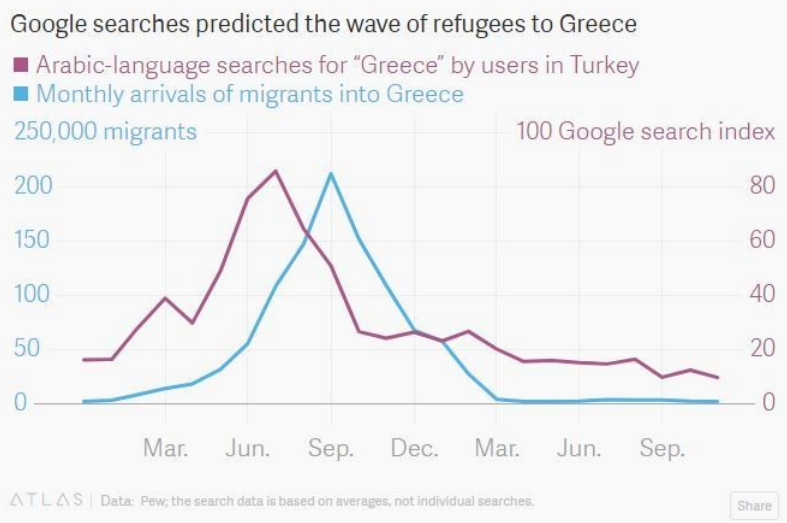
- Billion Prices Project
- Satellite images reveal which countries cheat on their economic statistics, qz.com, May 2018.
- Finding Swimming Pools with Google Earth – Greek Government Hauls in Billions in Back Taxes, **spiegel.de, August 2010**
- Toll-Index: Real Time Economic Telemetry

The Internet

- Starting in the 1990s, digitisation and the internet in particular is impacting our world in a completely new and different way and will continue to be doing so in the foreseeable future.
- It places networked computing into an increasing number of objects that are neatly integrated into our daily lives thus creating a data driven society and a data driven economy.
- Perspective of a complete recording of all aspects of our life. Since demand and supply, all activities of companies and individuals, as well as the matching process are documented in the internet, not only the complete market economy but also all social aspects are reflected in a huge data cloud (big data).
- When made accessible to social scientists, it provides a universe of research potential. Not only do we remember the past but we can rewind and replay recordings of the socioeconomic process. [Askitas and Zimmermann, 2015b]

Markets go online

Internet data can be applied to a very wide range of research areas including forecasting (e.g. of unemployment, consumption goods, tourism, festival winners and the like), nowcasting (obtaining relevant information much earlier than through traditional data collection techniques), detecting health issues and well-being (e.g. flu, malaise and ill-being during economic crises), documenting the matching process in various parts of individual life (e.g. jobs, partnership, shopping), and measuring complex processes where traditional data have known deficits (e.g. international migration, collective bargaining agreements in developing countries).


Google searches predicted the wave of refugees to Greece
Arabic-language searches for "Greece" by users in Turkey
Monthly arrivals of migrants into Greece
250,000 migrants
100 Google search index
200
150
100
50
0
80
60
40
20
0
Mar.
Jun.
Sep.
Dec.
Mar.
Jun.
Sep.
ATLAS | Data: Pew; the search data is based on averages, not individual searches.
Share

Digitisation is pervasive (*Traveling in Europe's River of Migrants*, The New York Times, August 2015.)

Matching

The reason offline markets rapidly migrate online and new markets are directly born online is that the internet and digitisation are much more efficient with matching of supply and demand than analogue implementations. Moreover they provide market participants with non invasive access to the market at hand and they have the ability to record transactions seamlessly. This latter feature makes online market invaluable for socioeconomic research.

The offline version of the so-called **gig economy** is day labor hotspots.

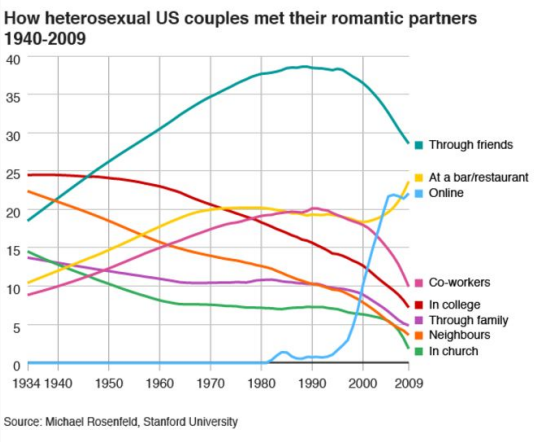


Figure 2: *The graphs that show the search for love has changed*, bbc.com, February 2016.

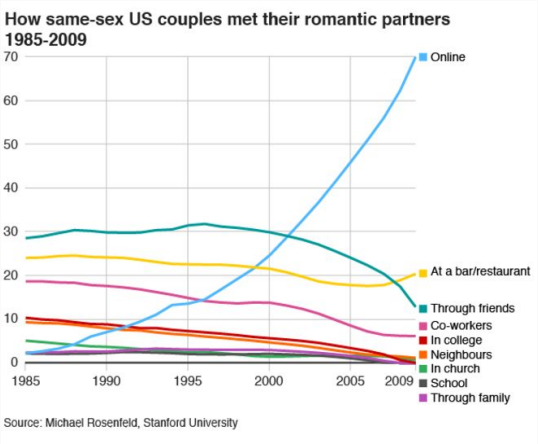


Figure 3: *The graphs that show the search for love has changed*, bbc.com, February 2016.

Matching in the Information Market

The best known example of course is Google itself which matches the demand for with the supply of documents. In the information market attention is the scarce resource as noticed first by Herbert A. Simon:

"...in an information-rich world, the **wealth of information** *means a* **dearth** *of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the* **attention of its recipients**. *Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it"*

This is the reason for the early success of Google's business model.

Google Search and the state of the Google user

The demand for information on a certain topic by an individual reveals information about the state of that individual. If you search for "symptoms of depression" you might be self diagnosing or acting in diagnosing a person close to you. If you search for "side effect of prosac" you or a loved one is probably taking the antidepressant. In aggregate form we might then be able to develop indicators on the proliferation of certain phenomena. In the remainder of this talk I will show you some examples. I summarised the properties of the Google Trends data in [Askitas, 2015b] (https://wol.iza.org/).

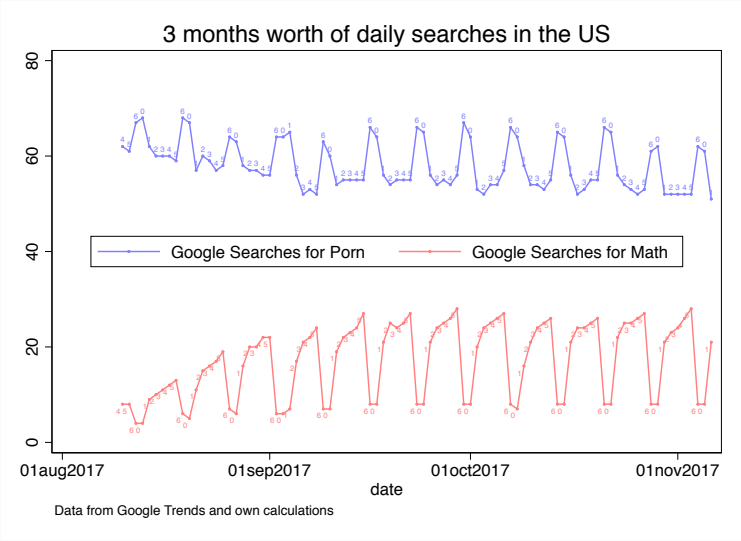
Google searches are utterances

Google searches may be viewed as continuous, irregular, unsolicited surveys with a Global cross section and a longitudinal resolution which is timestamped to the second.

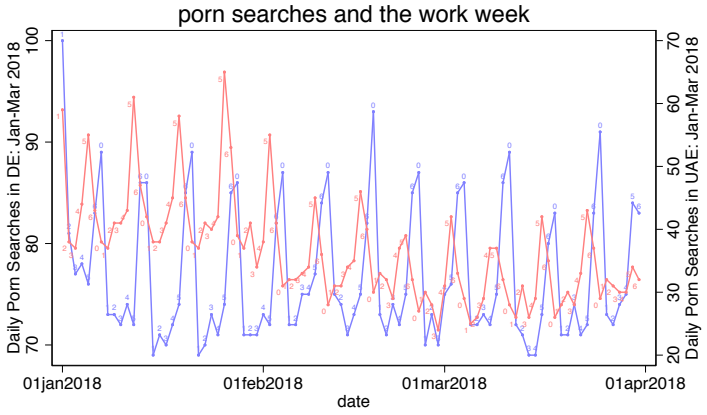
In traditional surveys we devise a question to which we seek an answer and get a (possibly) strategic one whereas in Google search we have unsolicited and hence more revealing answers to which we seek to reverse engineer the fitting question.

Example of strategic answering: we know that people search for "child porn" but nobody would admit doing so in a traditional survey.

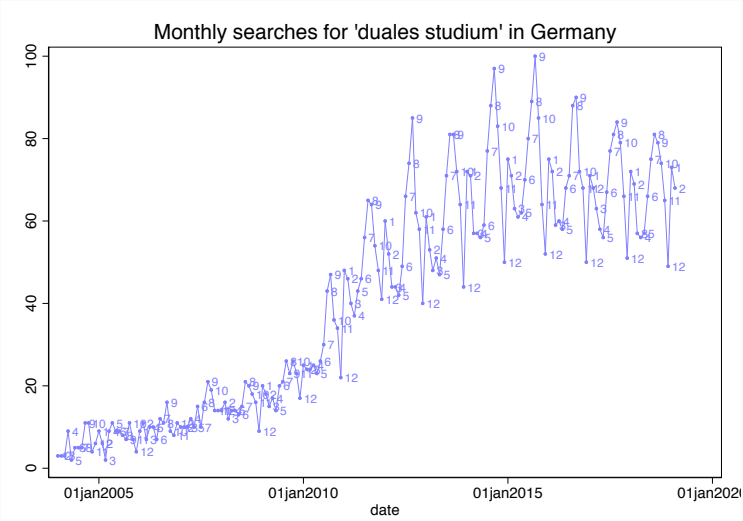
Capturing behavior: porn vs math



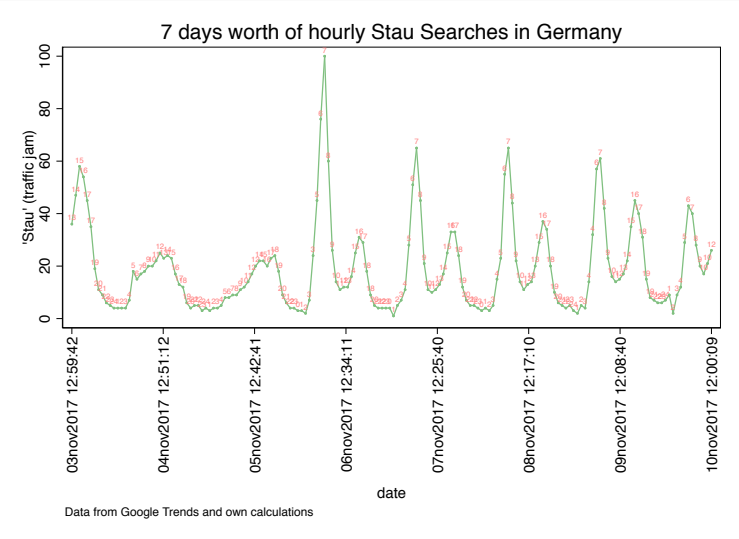
Capturing behavior: porn and the work week



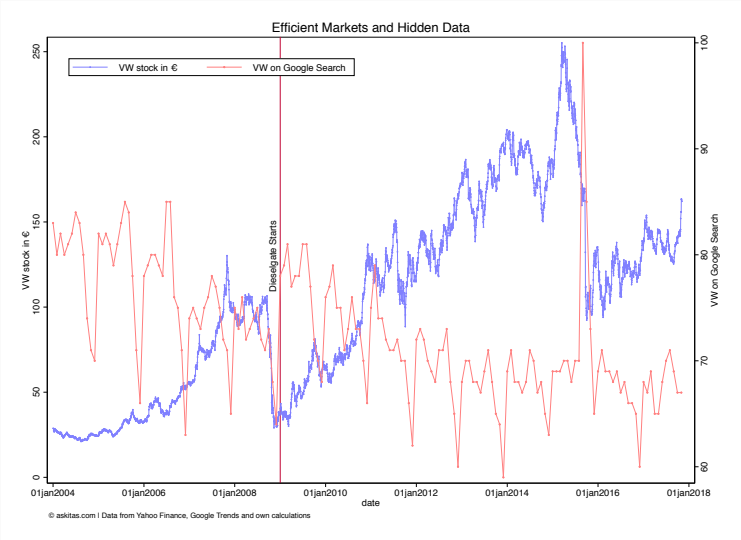
Capturing behavior: “duales studium”



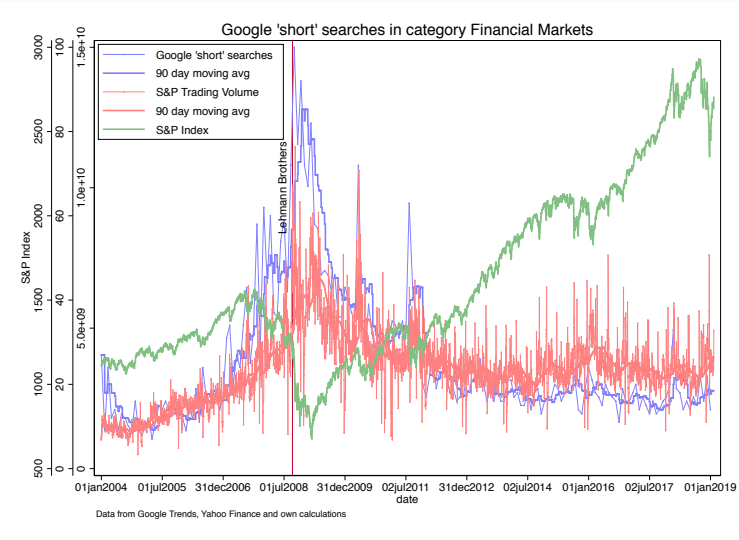
Capturing behavior: traffic jams



Capturing behavior: VW defeat device, dieselgate

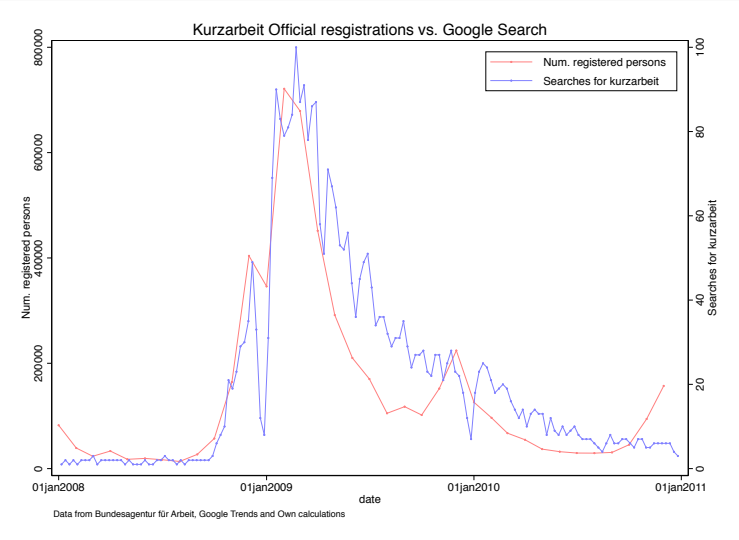


Capturing behavior: S&P 500 (Is the market going to crash any time soon?)

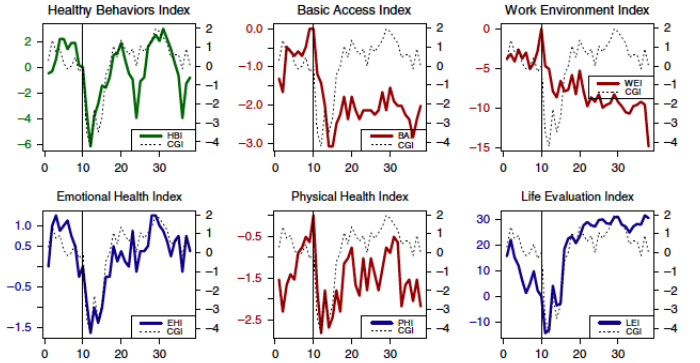


Unemployment and Health

Kurzarbeit (short-work)



Gallup [Askitas and Zimmermann, 2015a]



Notes: Data were obtained from www.well-beingindex.com where information on the definitions may also be found. All series are expressed as percentage of change relative to their value at time t = TARP (so they are all zero on the TARP date $x = 10$). Each of the six plots contains the composite index in addition to a component of the Gallup Well-Being Index

Symptoms and Side Effects

Use Google searches for "symptoms" as a proxy to self diagnosis and searches for "side effects" as a proxy for medication. Test what happens around the TARP date (Troubled Asset Relief Program – October 3, 2008), [Askitas and Zimmermann, 2015a].

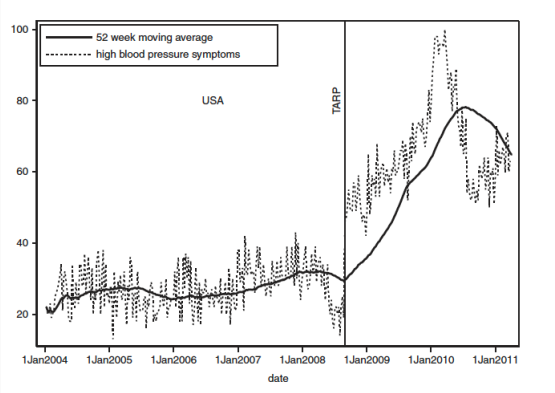
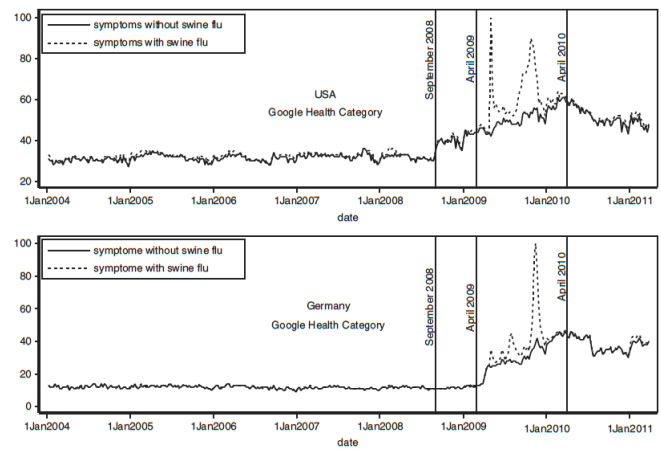


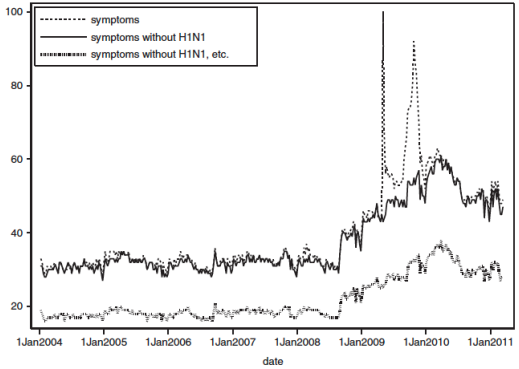
Figure 4: TARP – Troubled Asset Relief Program – (October 3, 2008)

“symptoms” searches in the Health Category, US and DE.



Notes: While in the USA they surge as early as September 2008 (first vertical line), for Germany they do not do so until April 2009 (second vertical line). By April 2010 (third line) the searches more than doubled compared to the level before September 2008. We use “symptoms” and “symptoms flu – swine – H1N1” in the USA and “symptome” and “symptome – grippe – H1N1 – schweinegrippe” in Germany

“symptoms” breakdown

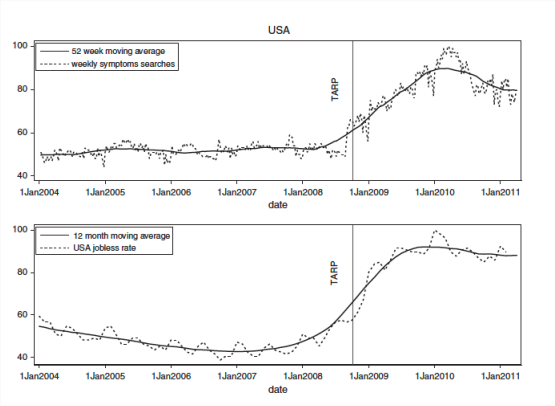


Notes: This holds true even after we remove the following terms: H1N1, swine, flu, old, kidney, menopause, medical, herpes, withdrawal, Salmonella, and, allergy, infection, lyme, diabetes, AIDS, HIV, stroke, yeast, herpes, thyroid, autism, cancer, heart, attack, depression, anxiety, early, pregnancy. We stopped this exercise at the 30 word per query limit of Google Trends. The marvels of the long tail remain at this point a mystery

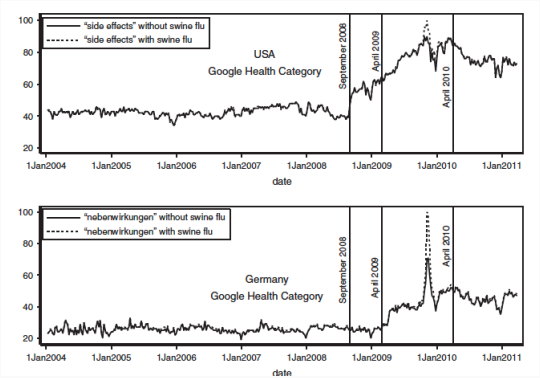
Remarks on breakdown

- The break down is useful to disentangle various causes for deviations from the mean (H1N1).
- It can be used as we will see in an other example later to know whether your identification strategy is good enough.
- In this case we see that after taking things out which might be less prone to respond to the crisis shock the spike remains as pronounced.

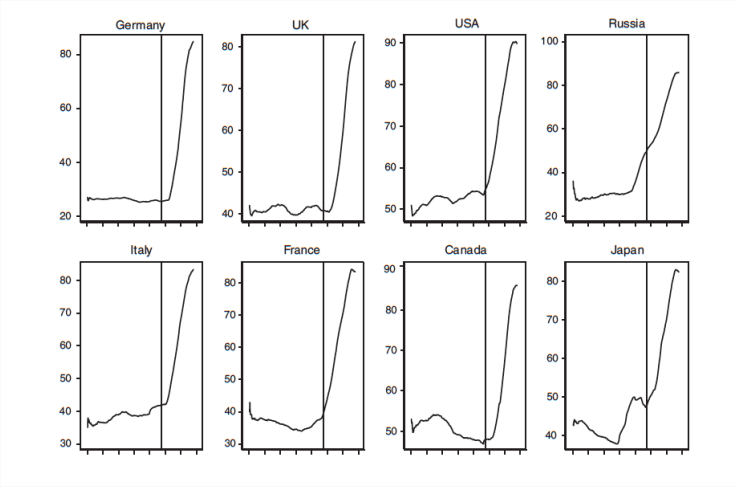
Symptoms searches vs Labor Market, US.



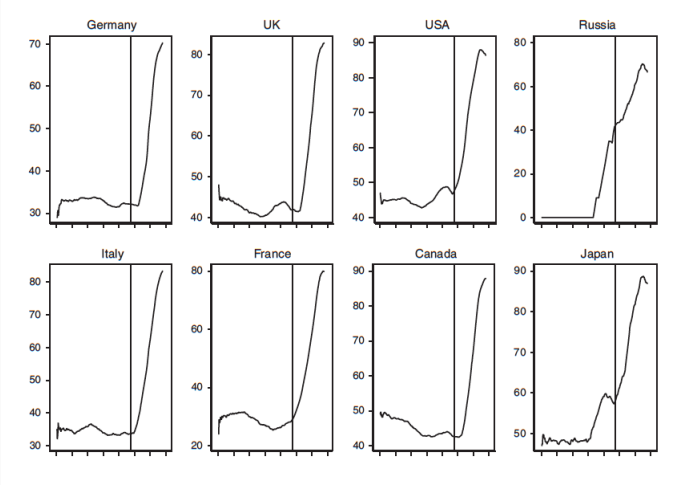
Side effects searches , US vs Germany: phase difference



Symptoms searches in the G8



Side Effects in the G8



Recovering "Theorems" of Health Economics

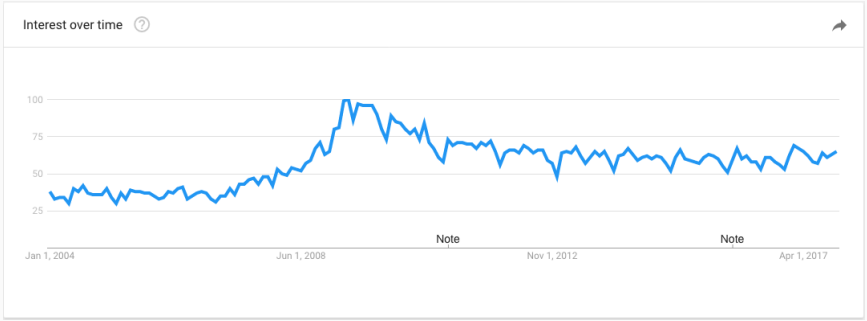
This work is relevant for Business cycle forecasting in the tradition of Joseph Schumpeter who in his 1939 volume "Konjunkturzyklen" includes church going in his list of 41 variables to monitor the business cycle:

"number of church goers is in inverse proportion to the degree of economic development "

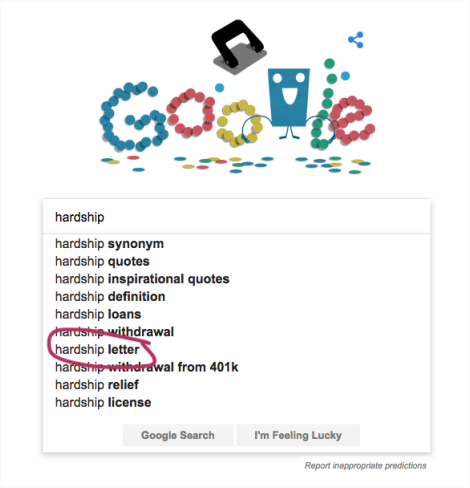
Schumpeter would have a ball with internet data!

Hardship searches

In the course of this work we noticed, as we had expected, and is known in the literature that searches for "chocolate cake", "inspirational quotes" or "obituaries" also respond to the crisis shock. We also noticed that searches for **hardship** also spike.



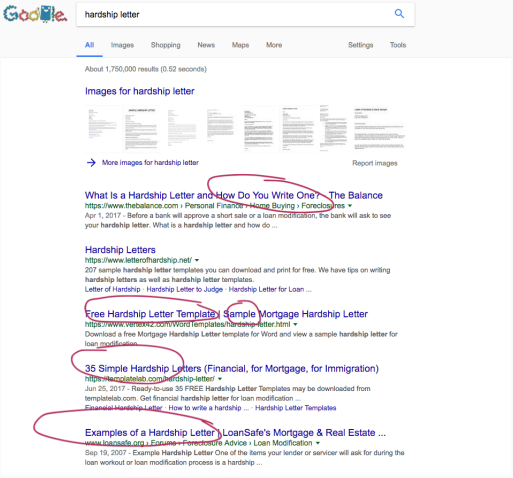
Hardship Letter



What is Hardship Letter?

A hardship letter is a loan modification request you write to your mortgage bank to explain that you are in temporary financial distress in order to avoid foreclosure.

Google results for "hardship letter"



"Obviously they are being counseled by the Internet. Disgusting"

In May of 2008, Angelo Mozilo, then Chairman of the board and CEO of a mortgage company called Countrywide Financial, made the news by accidentally hitting "reply" instead of "forward" in response to an e-mail from a distressed homeowner in North Carolina. The homeowner, who apparently was facing or anticipating an inability to make mortgage payments on time, had written a letter to request a loan modification and emailed it directly to the Office of the President of Countrywide. Angelo Mozilo who apparently thought he was forwarding the email to one of his people commented it as follows:

"This is unbelievable. Most of these letters now have the same wording. Obviously they are being counseled by some other person or by the internet. Disgusting."

Angelo Mozilo

Apparently, the homeowner who received the inadvertent reply with the comment made the email exchange public. Countrywide Financial, which in 2006 was the "largest mortgage lender" in the US, was by then failing and was purchased by Bank of America. Mozilo who at that moment must have been confronted with a prolonged rise in loan delinquencies saw it right. As more and more homeowners were entering financial hardship they sought help by writing loan modification requests to the banks holding their mortgages and the internet was their prime resource: a number of sites offered advice on the subject matter, loansafe.org being among the most prominent.

What if we were able to count Mozilo's letters?

Counting hardship letters

We can do one better than counting the letters. We can count the number of people who are in the process of fearing they might eventually need to write one!

Nowcasting Mortgage Delinquencies

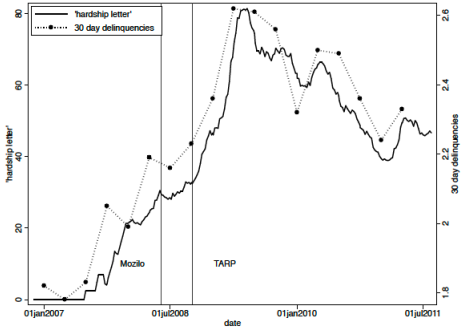


Figure 2: Searches for hardship letter predict 30-day mortgage delinquencies.
Sources: Searches for "hardship letter" (Google (2008)) are projected on the Google Category "Finance & Insurance / Credit & Lending / Debt Management" and are compared to 30 day prime loan delinquencies (Mortgage Bankers Association (2011)). We take a 12 week moving average smoothing of the hardship letter searches without forward terms.

Shares: referenda, Case/Shiller Index

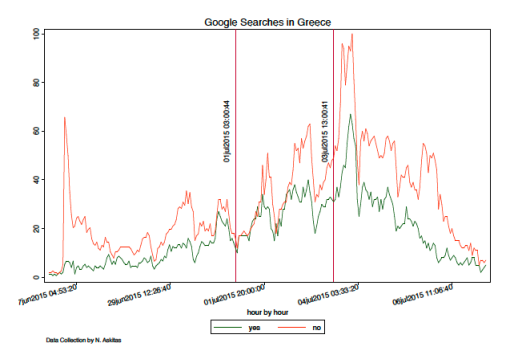
General Strategy

- Google trends data are relative data i.e. for a keyword K in time unit t, Google Trends delivers $g(K) = K_t/T_t$ where K_t are the number of searches for K at time t and T_t are the total number of searches at time t.
- So if your target variables are shares e.g. yes or no, sell or buy, put or call and you have a good identification strategy for representing them in terms of searches you might hope to get exact numbers.
- E.g. in a yes no referendum you need to know Y/N where Y is the yes vote and N is the no vote.
- So if you can proxy the yes voting intent and the no voting intent with $g(\tilde{Y})$ and $g(\tilde{N})$ for some keywords $\tilde{Y}$ and $\tilde{N}$ respectively then

$$\frac{g(\tilde{Y}_t)}{g(\tilde{N}_t)} = \frac{\tilde{Y}_t/T_T}{\tilde{N}_t/T_T} = \frac{\tilde{Y}_t}{\tilde{N}_t} \approx \frac{Y}{N}$$

  for t close to voting date.

Greek Referendum 2015: yes no searches



Greek Referendum 2015: identification

TABLE I

TOP SEARCHES FOR NO I.E. ¨οχι -ναι'

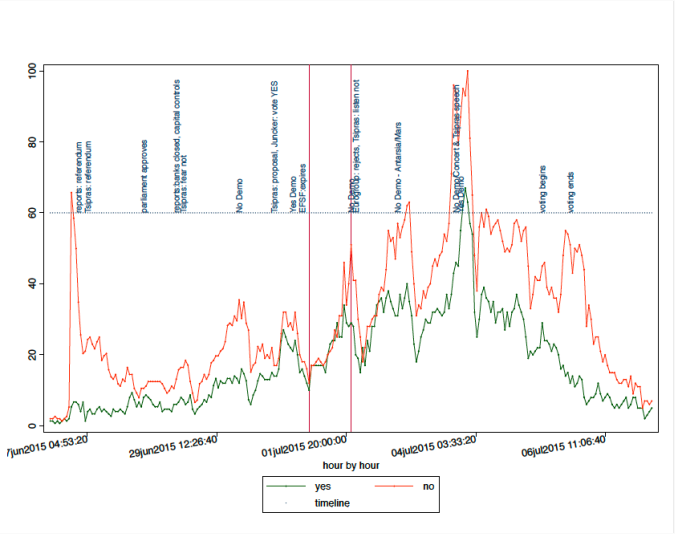
βεγγος οχι	Veggos No
δημοψηφισμα οχι	referendum no
οχι στο δημοψηφισμα	no in the referendum
συγκεντρωση υπερ του οχι	demonstration in favour of no
υπερ του οχι	in favour of no
συγκεντρωση οχι	demonstration no
λεμε οχι	we say no
δημοψηφισμα οχι	referendum no
οχι	no
ψηφιζω οχι	I vote no

TABLE II

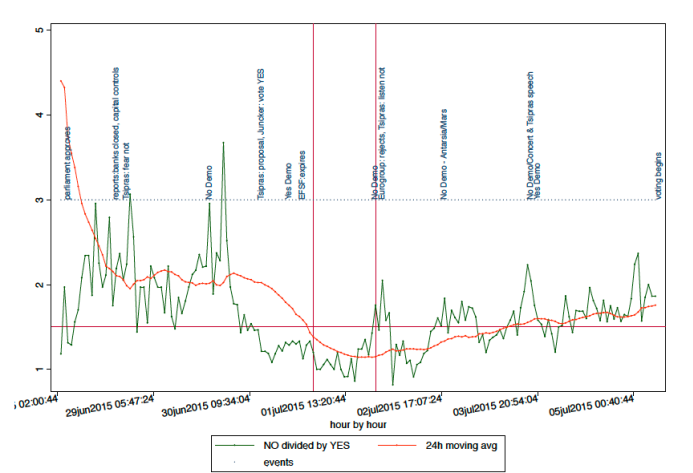
TOP SEARCHES FOR YES I.E. ¨ναι -οχι'

ναι στην Ευρωπη	yes to Europe
δημοψηφισμα ναι	referendum yes
ναι στο δημοψηφισμα	yes in the referendum
συγκεντρωση υπερ του ναι	demonstration in favour of yes
συγκεντρωση για το ναι	demonstration for yes
ναι στο ευρω	yes to euro
επιτροπη υποστηριξης του ναι	committee in support of yes

Greek Referendum 2015: yes no timeline



Greek Referendum 2015: ratio timeline

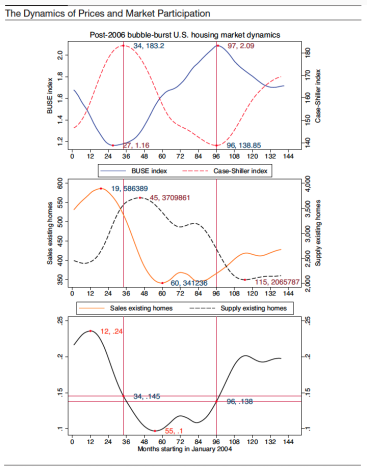


Irish Gay Marriage Referendum 2015: ratio timeline

- I applied the method of [Askitas, 2015a] to the Irish Gay Marriage referendum in [Askitas, 2015c] using an undocumented access to the data shared with me by the Google Trends team and it was successful as well.
- I tried the method to a number of national elections and it did NOT work although it captured surprising outcomes some times.
- Reason: not alway possible to proxy reliably. E.g. in Greek, German French elections searches for the Populist Right were topping the results not because it meant support for them but because everybody searched for the dystopian option. In the Greek and Irish referendum there was a high degree of polarisation, it was simpler due to dichotomy and could identify the two camps very well.
- I did find another dichotomy to apply the method to and returned interesting results

In [Askitas, 2016b] I formed what I called the BUSE index (i.e. BUy SEll index) by taking the point wise ratio of buy to sell Google searches in the Real Estate category.

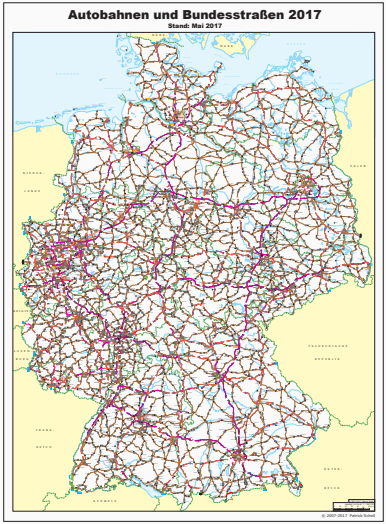
Nowcasting the Case/Shiller Housing Index



BUSE index = buyers sellers index. Case-Shiller index = Standard & Poor's/Case-Shiller® U.S. National Home Price Index.
Notes: The series have been smoothened to better recognize the cyclical pattern by a method described in appendix B.
Series used: Case-Shiller index and Google BUSE index (top), the inventory of existing homes for sale and the sales of existing homes (middle), and the probability of a house for sale being sold (bottom). The latter probability is simply the ratio of sales to inventory.
Data sources: Google Trends (http://www.google.com/trends); FRED (http://www.research.stlouisfed.org); S&P Dow Jones Indices (http://www.us.spindices.com); author's calculations

Hourly Data: Traffic Jams, [Askitas, 2016a]

The German Highway Network



Traffic Jam (stau) searches

- These searches take the form "stau Highway Number" or
- "stau Radio Station Name" or
- "stau TV Station Name"
- This means they are drivers preparing to inject themselves into the traffic
- Combined with IP geolocation we have their complete itinerary.
- Google searches are an open ledger of upcoming itineraries.

Traffic Jams

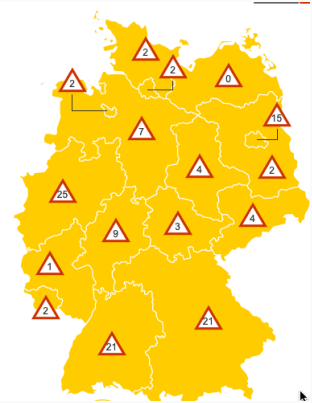
Adverse road conditions such as **traffic congestion** are known to cause a host of undesired side effects. These vary from **increased pollution (e.g. gas emissions), energy waste, additional transportation and production costs to waste of labor and delays in product deliveries**. Traffic congestion is also known to impact **public health negatively, to cause stress and road rage and to even affect the unborn**. If we thought of a city, a region, a country or any other social unit as a large living organism road traffic would be one of its circadian rhythms and traffic jams would be an obstruction to its entrainment. It is hence not surprising that obstructing the smooth flow of traffic sends ripples of negative effects deep into many aspects of socioeconomic life. Understanding, forecasting and preventing adverse road conditions is therefore important for the benefit of the drivers but also obviously for economic reasons.

Forecasting Road Conditions two hour in advance

Since I did not have a target variable to predict, in order to demonstrate the informational value of these data, I created one by programmatically harvesting the number of traffic jams every 5 minutes as reported by the German automobile club ADAC on their website. I did this for several months and I also collected the hourly stau searches for the same time period. I then wrote down a model which included day of the week and hour of day fixed effects plus the Google data.

The main result of the paper is that after controlling for time-of-day and day-of-week effects we can still explain a significant additional portion of the variation of the number of traffic jam reports with Google Trends and we can thus explain well over 80% of the variation of road conditions using Google search activity. A one percent increase in Google stau searches implies a .4 percent increase of traffic jams.

Main Ingredients: ADAC



Main Ingredients: python code for parsing the html page and harvest the data

```
SOURCE: <a title="Hessen" href="/reise_freizeit/verkehr/aktuelle_verkehrslage/suchergebnis.aspx?
search=CB6YECNuJoIo4nAgDfLe4nzkxz1SzUydFBy74fPLDnRX8uqovfTgCB6YEA__">9</a>

CODE:  hessen = re.match(r"<a title=\"Hessen\".+? (\d+)</a>", "SOURCE")

RESULT: hessen.group(0) = 9
```

Argentinian Inflation

The story in a nutshell

- Under president Cristina Fernandez de Kirchner, INDEC, the Argentinian Statistical Agency, published fake inflation data for many years.
- In 2012 the Economist stopped including official Argentinian inflation numbers in its weekly list of the 42 largest economies. It used data from a company called PriceStats instead.
- IMF followed in 2013 censuring Argentina for "inaccuracy" in its data
- The fraud was uncovered by the Billion Prices Project which collected online prices and formed its own inflation index and showed that the government index was severely misrepresenting reality.
- INDEC started producing a new real inflation index since a little over a year ago (The Economist resumed publication in May 2017).

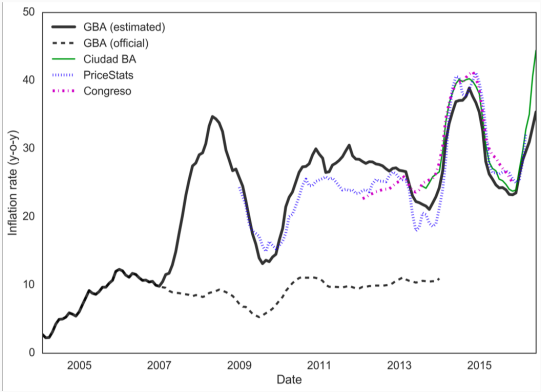


Figure 5: R. Campos, A Reconstruction of Argentina's Consumer Price Index using a Dynamic Factor Model, 2016

Billion Price Project

The Billion Price Project is the crown jewel of research with internet data by any metric conceivable. In terms of:

- Its real world impact (brought down a government and restored credibility of official data).
- Its academic impact: publications in top journals.
- Entrepreneurial spin off: pricestats.com (business model: multi country online inflation rates)

All by running a few lines of Python code to harvest price data from online shops.

IDSC focus on Internet data

- First workshop in 2018 at IZA. We had academic researchers but also private firm researchers and reps. (Linkedin, Burning Glass Tech, Indeed Hiring Lab). We saw papers with online RCTs as well as online work in official data.
- CfP for 2019 workshop is out. In cooperation with the Center of Advanced Internet Studies, in Bochum in September 20-21. Keynotes by: Julia Lane (NY Univ.), Ioana Marinescu (UPenn), Jessica Yu (Stanford) and CEDEFOP.
- In 2020 we will take it back to Bonn and it will be jointly organised with the Univ. of Luxembourg. We will broaden it to include Computation Social Science and more multi-disciplinarity to cross-fertilise with Labor Economics.

It's all about data!

- **Regulatory arbitrage**: In 2014 the CEO of the DT (built the MAUT system in 2007) pointed out that by not retaining MAUT data the government wasted a chance: many applications e.g. road conditions. While German data protection law imposes severe restrictions on data retention in an effort to protect individual privacy (history...), Germans offer their personal data to the FANGs voluntarily which then export the data to US based data refineries which in turn sell the resulting data-based services back to German companies (e.g. Google Traffic).
- **Official vs Internet Data**: Typically we seek to validate the latter using the former. This might some times be overturned. Think of official data on hospital visits vs "symptoms" Google searches in an economic downturn.
- **Data Tax (**see: Askitas, A data tax for a digital economy, wol.iza.org opinion article **)**: Companies whose business models involve recording personal data of large numbers of people (Platform Economy, Credit Cards etc) could be taxed in the form of having to make aggregate forms of their data available, in a form suitable for policy making, very much in the spirit of trends.google.com or movement.uber.com.
- **Data Lobbying**: We ought to successfully communicate to policy makers that successful forecasting and evidence based policy does not start with the researcher but with the policy maker: every project financed by taxpayer money needs to have a data retention plan, data reuse plan as well as RCTs built in.

Important Class of Questions

In a data driven economy/society where we are able to see data (of our activities) compiled and visualised in real time how do our activities change? What is the downside? What is the upside? Might we see "micro bubbles" form and burst?

Big Data (Irrational Exuberance ([Shiller, 2015]))

- The first bubbles found coincide with the advent of the newspaper
- The stock market crash in the 1930s coincides with improvements in long distance telephone technology (boiler rooms)
- The crash of the new economy in 2000 coincides with the proliferation of the Internet (think Netscape 1995)
- The crash of 2008 coincides with the peak in Social Media?
- Internet removes communication barriers.
- Accelerates contagion of ideas.
- World Economic Forum 2013 report: *"digital wildfires in a hyperconnected world"* one of the major global economic risks ahead.

Thank your for your attention

Askitas, N. (2015a).
Calling the greek referendum on the nose with google trends.
IZA Discussion Paper Series No 9569.

Askitas, N. (2015b).
Google search activity data and breaking trends.
IZA World of Labor.

Askitas, N. (2015c).
Predicting the irish "gay marriage" referendum.
IZA Discussion Paper Series No 9570.

Askitas, N. (2016a).
Predicting road conditions with internet search.
PLoS ONE11(8): e0162080. https://doi.org/10.1371/journal.pone.0162080.

Askitas, N. (2016b).
Trend-spotting in the housing market.
Cityscape: A Journal of Policy Development and Research, 18(2):165–178.

Askitas, N. and Zimmermann, K. F. (2013).
Nowcasting business cycles using toll data.
Journal of Forecasting, 32(4):299–306.

Askitas, N. and Zimmermann, K. F. (2015a).
Health and well-being in the great recession.
International Journal of Manpower, 36(1):26–47.

Askitas, N. and Zimmermann, K. F. (2015b).
The internet as a data source for advancement in social sciences.
International Journal of Manpower, 36(1):2–12.

Shiller, R. J. (2015).
Irrational exuberance.
Princeton university press.