Searching for a Better Life: Nowcasting International Migration with Online Search Queries

Tobias Stöhr (Kiel Institute for the World Economy)

joint work with
André Gröger (Universitat Autònoma de Barcelona)
Marcus Böhme (OECD)

UNU WIDER conference - Accra - 5.10.2017

Motivation and Research Question

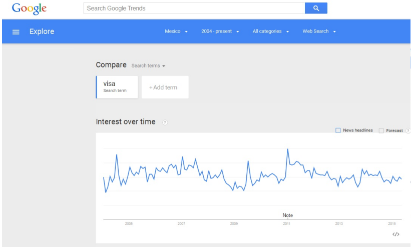
Lack of migration data

- *inconsistent* across countries
- typically *outdated*
- often *inexistent*, especially problematic: time dimension
- Geo-located online search data provides new opportunities for predicting current human behavior (**now-casting**)
- Potential migrants search the internet for information about migration prior to departure (e.g. Maitland & Xu 2015)

Is online search behavior in origin countries predictive of international migration flows?

Might it be a proxy of interest in emigration?

Google Trends Index (GTI)



Search Google Trends

Explore

Mexico ▾ 2004 - present ▾ All categories ▾ Web Search ▾

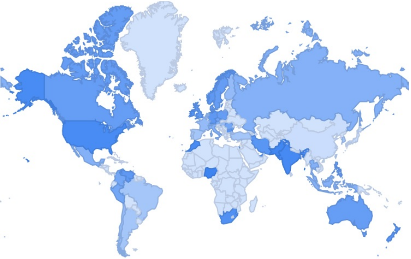
Compare Search terms ▾

visa
Search term

+ Add term

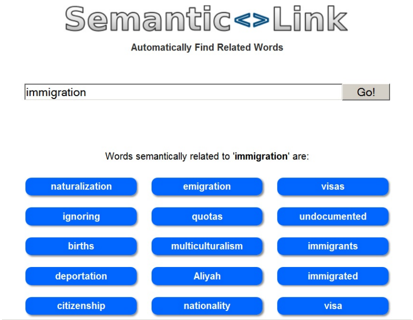
Interest over time

□ News headlines □ Forecast



- Google is the most common search engine (market share: 73%)
- GTI reflects revealed demand for information

To decrease very large p to $p < n \cdot T$



Semantic<>Link

Automatically Find Related Words

immigration

Words semantically related to '**immigration**' are:

Translated into all three UN working languages that use the Latin alphabet (i.e. ENG, FRA, and ESP)

Data: Keywords

Migration

applicant
arrival
asylum
border control
citizenship
consulate
customs
deportation
diaspora
embassy
emigrate
emigration
foreigner
illegal
immigrant
legalization
migrant
nationality
naturalization
passport
quota
refugee
requirement
Schengen
smuggler
smuggling
tourist
unauthorized
undocumented
unskilled
visa
waiver

Economics

benefit
business
compensation
contract
discriminate
earning
economic
economy
employer
employment
GDP
hiring
income
inflation
internship
job
labor
layoff
minimum
payroll
pension
recession
recruitment
remuneration
salary
tax
unemployment
union
vacancy
wage
welfare

Note: Translated into all three UN working languages that use the Latin alphabet (i.e. ENG, FRA, and ESP). Always A.E. and B.E. spelling, singular and plural. Analogous for FRA and ESP.

Additional Data

OECD International Migration Database

- Yearly panel (2004-2013) with inflows of foreign nationals (regular and asylum) to OECD
- 198 origin to 33 OECD destination countries (excl. Mexico and Turkey)
- Some gaps and missing values for certain countries

WDI: GDP, internet users, literacy, population, unemployment, human capital
Melitz and Toubal (2012): Spoken language
Gravity variables, Polity IV, and more

Estimation Strategy

Specification 1: Unilateral flows to OECD (Panel FE)

$$Y_{o,t+1} = \alpha + \beta T_{ot} + \gamma O_{ot} + \eta D_t + \delta_o + \tau_t + \varepsilon_{ot}$$

with:

- Y_{ot}: Log inflow to OECD by foreign nationality.
- T_{ot}: Trends search terms at origin.
- O_{ot}: Vector of origin-specific control variables.
- D_t: Vector of destination-specific control variables.
- δ_o: Origin country FE.
- τ_t: Time FE.
- ε_{ot}: Robust error term, clustered at the origin country level.

Estimation Strategy

Specification 2: Nowcasting equation

$$Y_{o,t+1} = \alpha + \delta_1 Y_{ot} + \delta_2 \Delta Y_{ot} + \beta T_{ot} + \gamma O_{ot} + \eta D_t + \varepsilon_{ot},$$

with:

- Y_{ot}: Log inflow to OECD by foreign nationality.
- $\Delta Y_{ot} = Y_{ot} - Y_{ot-1}$
- T_{ot}: Trends search terms at origin.
- O_{ot}: Vector of origin-specific control variables.
- D_t: Vector of destination-specific control variables.
- ε_{ot}: Robust error term, clustered at the origin country level.

Within-dimension only (Panel FE)

Main results

- Depending on the specification the coefficient of determination increases between 120% to 280%, from a very low 0.05-0.06.
- In-sample performance better if ENG, FRA, ESP more widely spoken in country of origin

Risk: Overfit

With "large p, small N, small T" risk of mechanical overfit

Possible steps towards solution

- Variable selection methods
- Out-of-sample estimation
- Reduce dimensions

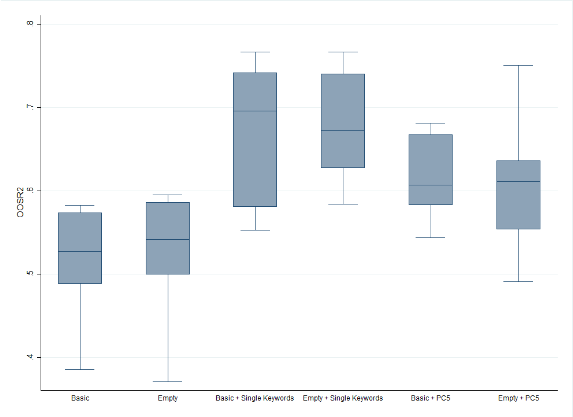
Variable selection models

- LASSO: Least absolute shrinkage operator (Tibshirani, 1996)
- LARS: Least angle regression (Efron, Hastia, Johnstone and Tibshirani, 2004)
- Information criterion: Mallows' Cp
- Suggests: Keep over half of the single keywords in the model

Out-of-sample (OOS) estimation

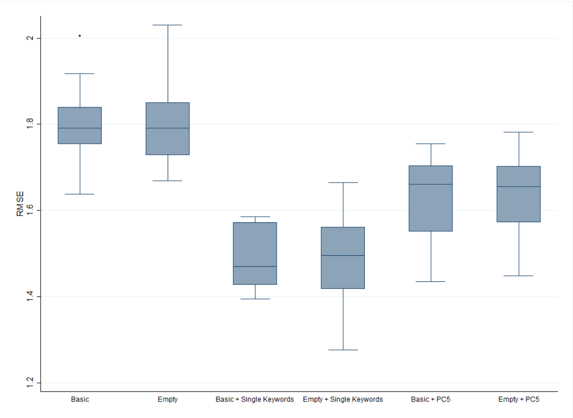
- Idea: if mechanical overfit, should not hold up out-of-sample
- Approach: k-fold cross validation
    - Draw k=10 random samples without replacement
    - Use 9/10 to estimate model
    - Apply model with estimated parameters in remaining fold
    - Estimate statistics such as R^2 and RMSE

Explaining Levels: Crossfold Validation R^2



Note: Out-of-sample Pseudo R2 based on 10-fold cross validation without variable selection procedure

Levels: Crossfold Validation RMSE



Note: Out-of-sample RMSE based on 10-fold cross validation without variable selection procedure

Dimension reduction using PCA

- Principle component 5 has very good in-sample and out-of-sample performance
- Disadvantage of method: very abstract
- Proposed solution: Correlates of principal components, i.e. understanding the variation we are using for prediction

Beyond Predictive Power

Test correlations with Gallup World Poll

- "Ideally, if you had the opportunity, would you like to move permanently to another country, or would you prefer to continue living in this country? And if yes: To which country would you like to move?"
- Add log country-level migration intention to our model
- n=330, GWP has estimated coefficient of 0.18-0.26
- Adding GTI reduces GWP coefficient considerably, suggesting imperfect overlap
- Specification 2: GWP insignificant, GTI as before

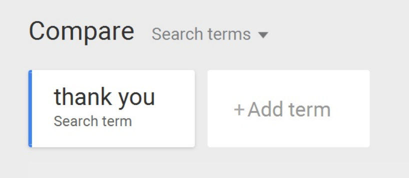
Findings and Contributions

Findings

- Provide evidence that the GTI has **substantial predictive power** for estimating international migration
- Relating our GTI to available survey data provides preliminary evidence that it reflects migration intentions

Contributions

- Providing *consistent* data on **migration intentions** worldwide
- Potential for short-term **now-casting analyses** (e.g. humanitarian crises)



Compare Search terms

thank you
Search term

Data Access: Google Trends API

- Short proposal to Google to get non-profit status
- ID with free download contingent per day
- Python code to scrape data from Trends API
- Output as delimited text files

Summary and outlook

- Providing **consistent and worldwide! indicators** for **prediction of migration** (and many other things).
- Many possible **micro-level applications** for geospatial **analysis of disasters**:

Examples

1. Man-made disasters: **Syrian Refugee Crisis** - GT for "Migration + Turkey" at origin in Syria are positively correlated with refugee arrivals in Turkey
2. Natural disasters: **2015 Earthquake in Nepal** - Indicating demand for information on survival strategies (labor, credit, migration, etc)