Searching for a Better Life: Predicting International Migration with Online Search Keywords

Marcus Böhme[a], André Gröger[b], Tobias Stöhr[c]

[a]Organisation for Economic Co-operation and Development (OECD)
[b]Universitat Autonoma de Barcelona (UAB) and Barcelona Graduate School
[c]Kiel Institute for the World Economy (IfW) and IZA

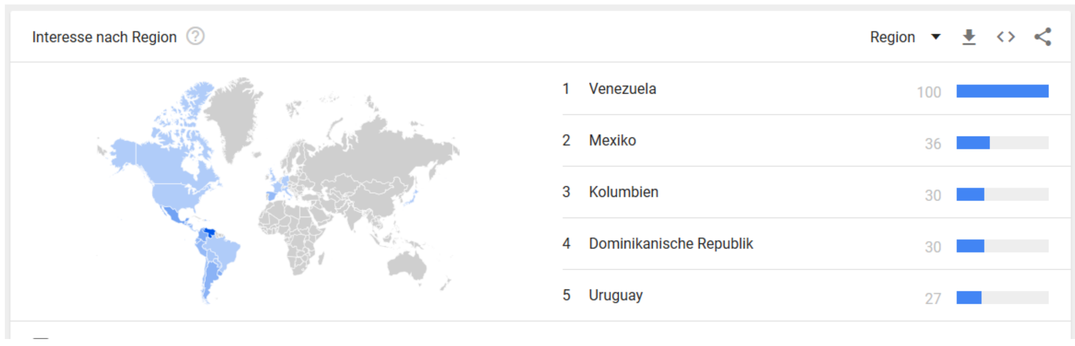
Motivation

- Migration data: scarce, largely inconsistent across countries, often outdated - particularly in developing countries
- Internet usage around the world provides geo-referenced traces of online search queries
- Search query logs may proxy latent demand for information

Research Question:

- *Can online search data be used to measure migration intentions in origin countries in order to predict subsequent outflows?*

Search intensity distribution for "pasaporte" in mid-April 2018



Methodology

1. Can the standard panel model of international migration flows be improved?

- Method 1: Panel regression

$$Y_{ot+1} = \alpha + \beta T_{ot} + \gamma O_{ot} + \eta D_t + \delta_o + \tau_t + \epsilon_{ot}$$

Y_{ot+1}: international migration flows
T_{ot}: aggregated search volumes
O_{ot}: origin country determinants
D_t : destination country determinants
δ_o: origin-FE τ_t: time-FE

2. Can we ensure improvements are not due to overfit?

- Method 2.1: Shrinkage methods
  - Least absolute selection and shrinkage operator (LASSO)
  - Least-angle regression (LARS) algorithm
- Method 2.2: Out-of-sample exercise
  - 10-fold cross validation assessing OOS-R^2 and OOS-RMSE

3. Why is there additional predictive power?

- Method 3: Comparison with survey-data on intention to migrate from the Gallup World Poll
  - Do search volumes and survey instruments correlate?
  - Horse-race specification: Which explains more?

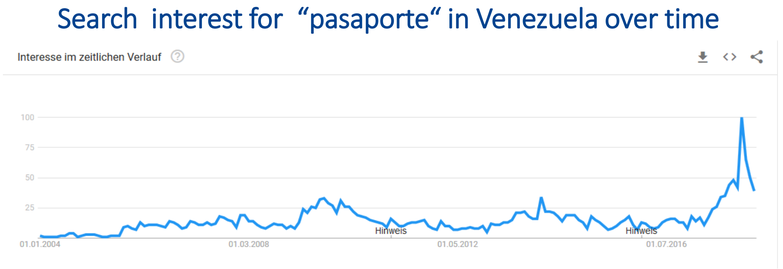
Data

- Google Trends Data
  - Access through API
  - Download average daily search intensity per country
  - Aggregate to yearly average
- Keyword selection
  - "Semantic scholar"

    Top third of 100 most correlated terms for "immigration" and "economic"

    ENG, FRA, ESP
- Additional data
  - OECD yearly migration flows 2004-2015
  - World Development Indicators
    - Minimum setup: GDP, Population
    - Extended setup: unemployment rate, share of young population, share of internet users, mobile phone subscriptions
  - Melitz-Toubal (2015) language matrix
  - EM-DAT weather and non-weather related disasters
  - Polity IV Autocracy Score, State Fragility Index

Search interest for "pasaporte" in Venezuela over time



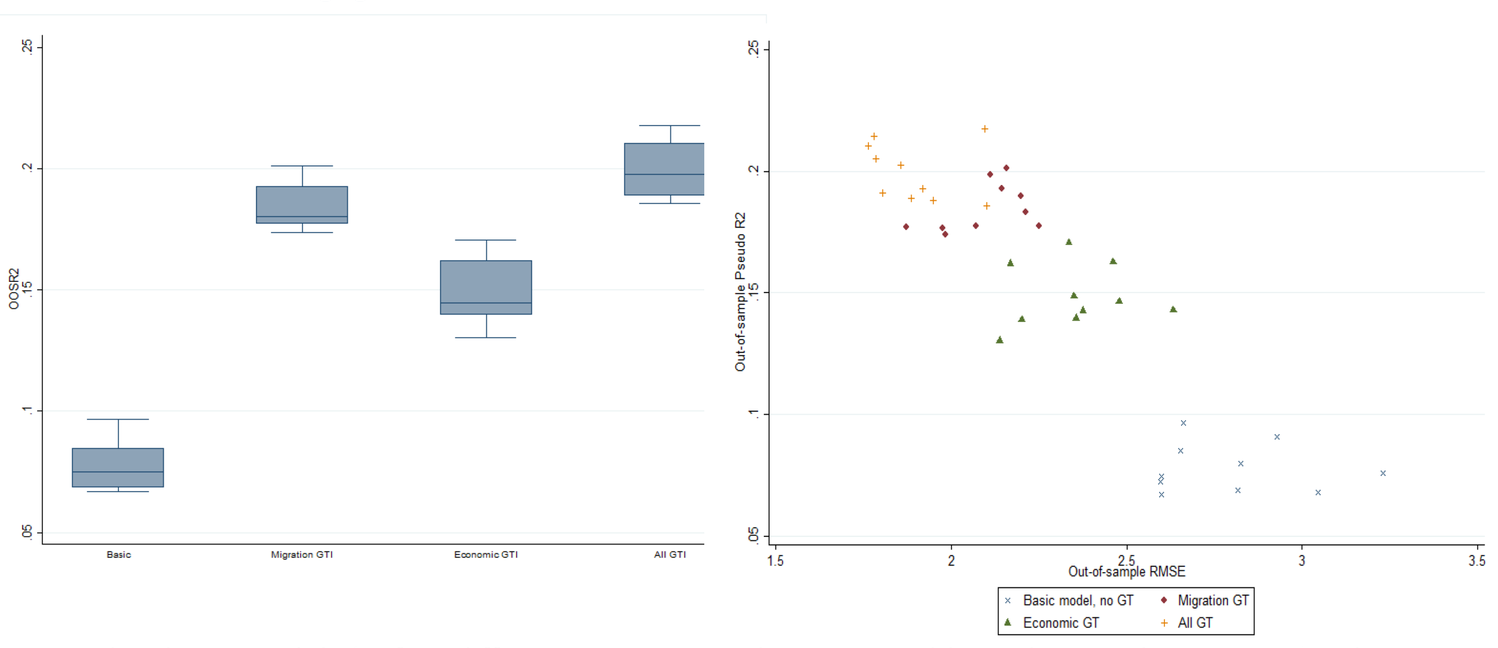
Results #1

Google Trends	(1) None	(2) Migration	(3) Economic	(4) Mig+Econ
GTI Migration keywords (37)		√		√
GTI Economic keywords (37)			√	√
Origin FE	√	√	√	√
Year FE	√	√	√	√
Observations	1,068	1,068	1,068	1,068
Joint significance GTI keywords (p-value)	–	0.000	0.0002	0.000
within-R^2	0.077	0.2080	0.167	0.258
Number of Origins	98	98	98	98
Panel B: Spoken Language Share > 50% at Origin				
Google Trends	(1) None	(2) Migration	(3) Economic	(4) Mig+Econ
GTI Migration keywords (37)		√		√
GTI Economic keywords (37)			√	√
Origin FE	√	√	√	√
Year FE	√	√	√	√
Observations	732	732	732	732
Joint significance GTI keywords (p-value)	–	0.000	0.000	0.000
within-R^2	0.089	0.255	0.213	0.312
Number of Origins	67	67	67	67

Results #2

Large gains in OOS-R^2

No trade-off between OOS-R^2 and OOS-RMSE



- Shrinkage-models (in first differences to get as close as possible to the panel regression) indicate explanatory value of additional keywords, leaving about 15 of 37 migration keywords in optimal model
- OOS-results indicate increased share of variation explained with <u>improved</u> forecasting error

Results #3

- Average of „the" three GWP migration intention items explain variation in panel regression
- Adding our search volumes more than halves their coefficient, yet GWP item remains statistically significant
- Interpretation: The latent variation we capture overlaps considerably with the latent concept the GWP seeks to measure. Further research needed

Conclusion

- Search volumes can help improve models of international migration
- Gains in explained variance will be higher if tested languages are widely spoken and if more people have access to the internet
- Could be developed into approach that might be useful in forecasting/nowcasting emigration