

# Exploring Migration Patterns Using Digital Trace Data

Carolina Coimbra Vieira  
Ebru Şanlıtürk

*Department of Digital and Computational Demography  
Max Planck Institute for Demographic Research*





## About us



- PhD in Public Policy and Administration from Bocconi University, Italy
- Research scientist at Max Planck Institute for Demographic Research  
Laboratory of Migration and Mobility - MPIDR



- BSc and MSc in Computer Science from Universidade Federal de Minas Gerais (UFMG), Brazil
- PhD student at Max Planck Institute for Demographic Research, Max Planck Institute for Software Systems and Saarland University, Germany  
Laboratory of Migration and Mobility - MPIDR  
Group of Social and Information Systems - MPI-SWS



# Workshop Exploring Migration Patterns using Digital Trace Data

**First edition:**

9th African Population Conference  
(APC) 2024

Lilongwe, Malawi  
21 May 2024

Organized by the IUSSP Panel in Digital and Computational Demography.





# Introduction

- Name
- Affiliation
- Research area
- Why this workshop?



*Exploring Migration Patterns Using Digital Trace Data*



# Introduction

Three pillars in population studies: mortality, fertility, and migration

- **Death** is easy to count and model: it's clear what it is and it happens only once
- **Birth** is easy count, a bit less to model: it's clear what it is, mothers may give birth several times
- **Migration** is a “mess”: there is no single definition, it may happen several times in different directions



# Migration

**What is your definition of migration or migrant?**



# Migration

**Migration:** Moving to a country other than the usual residence for a period of at least a year (12 months), so that the country of destination effectively becomes the new country of usual residence.

(adapted from the definition of “immigrant” by UN DESA)

<https://www.iom.int/key-migration-terms>



# Introduction to migration studies

## Data sources in migration studies

- Administrative data sources
  - Census data
- Statistical data sources
  - Survey data
- **Innovative data sources**
  - Digital trace data

(i.e., data generated — and collected —  
as a by-product of our online activities)



# Introduction to migration studies

## Data sources in migration studies

- Administrative data sources
  - Census data
- Statistical data sources
  - Survey data
- **Innovative data sources**
  - Digital trace data

(i.e., data generated — and collected — as a by-product of our online activities)



The screenshot shows the homepage of the Max Planck Institute for Demographic Research. The header contains the institute's logo and name. A green arrow points from the 'Innovative data sources' section in the text above to this page. The page has a navigation bar with links for Home, About Us, Publications & Databases, Research (highlighted in green), Career, and News & Events. The main content area is titled 'DEPARTMENT' and 'Digital and Computational Demography (Zagheni)'. A sidebar on the right is titled 'Research' and lists categories like Social Demography (Myrskylä), Digital and Computational Demography (Zagheni), and Migration and Mobility.

[https://www.demogr.mpg.de/en/research\\_6120/digital\\_and\\_computational\\_demography\\_zagheni\\_11666/](https://www.demogr.mpg.de/en/research_6120/digital_and_computational_demography_zagheni_11666/)



# Introduction to digital trace data in migration studies

Why do we use digital trace data?

- Traditional data are often **difficult, time-consuming, and costly** to collect.
- Traditional data sources are limited in hard-to-reach contexts and societies.
- Some populations (e.g., immigrants) are often underrepresented in traditional data sources.



# Introduction to digital trace data in migration studies

## Advantages

- Detailed and rich information
- Can provide (almost) real-time data
- Less costly than traditional methods of data collection

## Disadvantages

- Privacy and ethical concerns
- Biases



# Examples of digital trace data that have been used in migration studies



Scopus®



WIKIPEDIA  
The Free Encyclopedia





## Examples of digital trace data that have been used in migration studies



The impact of **Hurricane Maria** on **out-migration from Puerto Rico**: evidence from **Facebook** data. Alexander et al. (2019).

Which definition of migration better fits **Facebook ‘expats’**? A response using **Mexican census data**. Varona et al. (2024).

The interplay of **migration** and **cultural similarity** between countries: Evidence from **Facebook** data on food and drink interests. Coimbra Vieira et al. (2022).



The Value of Cultural Similarity for Predicting Migration: Evidence from Food and Drink Interests in Digital Trace Data. Coimbra Vieira et al. (2024).



# Examples of digital trace data that have been used in migration studies



**Openness to migrate internationally for a job:** evidence from LinkedIn data in Europe. Perrotta et al. (2022).

**Global gender gaps in the international migration of professionals on LinkedIn.** Jacobs et al. (2024).

**Using Facebook and LinkedIn data to study international mobility.** Coimbra Vieira et al. (2022).



## Examples of digital trace data that have been used in migration studies

Desaparecidxs: characterizing the population of **missing children** using **Twitter**. Coimbra Vieira et al. (2022).



Online social **integration of migrants**: evidence from **Twitter**. Kim et al. (2023).

**Twitter** data for **migration studies**. Kim et al. (2022).



## Examples of digital trace data that have been used in migration studies

**Scholarly migration within Mexico:** analyzing internal migration among researchers using Scopus longitudinal bibliometric data.  
Miranda-González et al. (2020).

**Global patterns of migration of scholars** with economic development. Şanlıtürk et al. (2023).

**A gender perspective on the global migration of scholars.** Zhao et al. (2023).



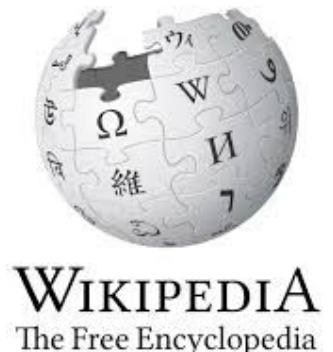


## Examples of digital trace data that have been used in migration studies

Search for a new home: **refugee stock** and **Google search**.  
Şanlıtürk and Billari (2024).

Harnessing the potential of **Google searches** for understanding dynamics of **intimate partner violence** before and after the COVID-19 outbreak. Köksal et al. (2022).

**Wikipedia** as a tool for tracking **mass migration flows**: Insights from the **Russian invasion of Ukraine**. Coimbra Vieira et al. Working paper.





# Examples of digital trace data that have been used in migration studies



Scopus®



# GOOGLE TRENDS DATA IN MIGRATION STUDIES





## WHAT IS GOOGLE TRENDS?

Google Trends is a tool by Google, that shows the **relative** interest over time and/or by subregion for any selected query, time period and location.

(Trends Help, 2021)

(see: [https://support.google.com/trends/answer/4365533?hl=en&ref\\_topic=6248052](https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052))



## WHAT DO GOOGLE TRENDS DATA TELL US?

- Interest for a selected query over time

Search interest for a topic as a proportion of all searches on all topics on Google at the specified time and location

- Interest for a selected query by subregions

Search interest for a topic by subregions as a proportion of all searches on all topics on Google in that same place and time.



## WHAT DO GOOGLE TRENDS DATA TELL US?

- Google Trends **does not** report the overall search volume for a selected query.  
Google Ads – Keyword Planner is meant for insights into monthly and average search volumes, specifically for advertisers to assess the size of the audience  
(<https://support.google.com/google-ads/answer/6325025> )
- It gives us a measure of interest for a query normalized for the selected time and location.

(Trends Help, 2021)



## GOOGLE TRENDS INDEX

- Google Trends normalizes search data to make comparisons between terms easier. Search results are normalized to the selected time and location of a query as follows;
  - “Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity”
  - This process is necessary to avoid the places with the most search volume to always rank the highest.
  - “The resulting numbers are then scaled on a range of 0 to 100 based on a topic’s proportion to all searches on all topics”



## GOOGLE TRENDS INDEX

- Different regions that show the same search interest for a term don't always have the same total search volumes.
  
- The parameters we enter matter. 100 indicates the maximum search interest for a query, only for the time and location selected. Shortening and extending the selected time period may change the minimum and maximum interest points.
  
- Time adjustment for non-real time data

# GOOGLE TRENDS IN THE LITERATURE





## GOOGLE TRENDS DATA IN LITERATURE

### □ Epidemiology:

- Online search data to *nowcast* outbreaks (**Flu Trends!**)

(Ginsberg, et al., 2009) (Pelat, Turbelin, Bar-Hen, Flahault, & Valleron, 2009) (Brownstein, Freifeld, & Madoff, 2009)

### □ Economics:

- Online search data to forecast unemployment rate, economic activity, inflation rate

(Ettredge, Gerdes, & Karuga, 2005) (Askitas & Zimmermann, 2009) (Choi & Varian, 2009)  
(Guzman, 2011)



## GOOGLE TRENDS DATA IN DEMOGRAPHY LITERATURE

### □ Demography

- Online search data to *forecast* abortions, fertility behaviour, suicides and causes of mortality

(Reis & Brownstein, 2010)

(Billari, D'Amuri & Marcucci, 2016)

(Wilde, Chen & Lohmann, 2020)

(McCarthy, 2010)

(Song, et al., 2014)

(Chang, Kwok, Cheng, Yip, & Chen, 2015)

(Solano, et al., 2016)

(Ricketts & Silva, 2017)



## GOOGLE TRENDS DATA IN MIGRATION CONTEXT

□ How do the Google Trends data help us in the migration context?

Main assumptions

### Migration

- Online searches show interest in potential destinations
- Interest in a destination may be a proxy for migration intention

### Forced migration

- People on the move need information
- We look for significant divergences from regular search patterns



## GOOGLE TRENDS DATA IN MIGRATION STUDIES

### □ Use in migration research

- Estimating migration flows
- Estimating migration stocks
- Now-casting and forecasting



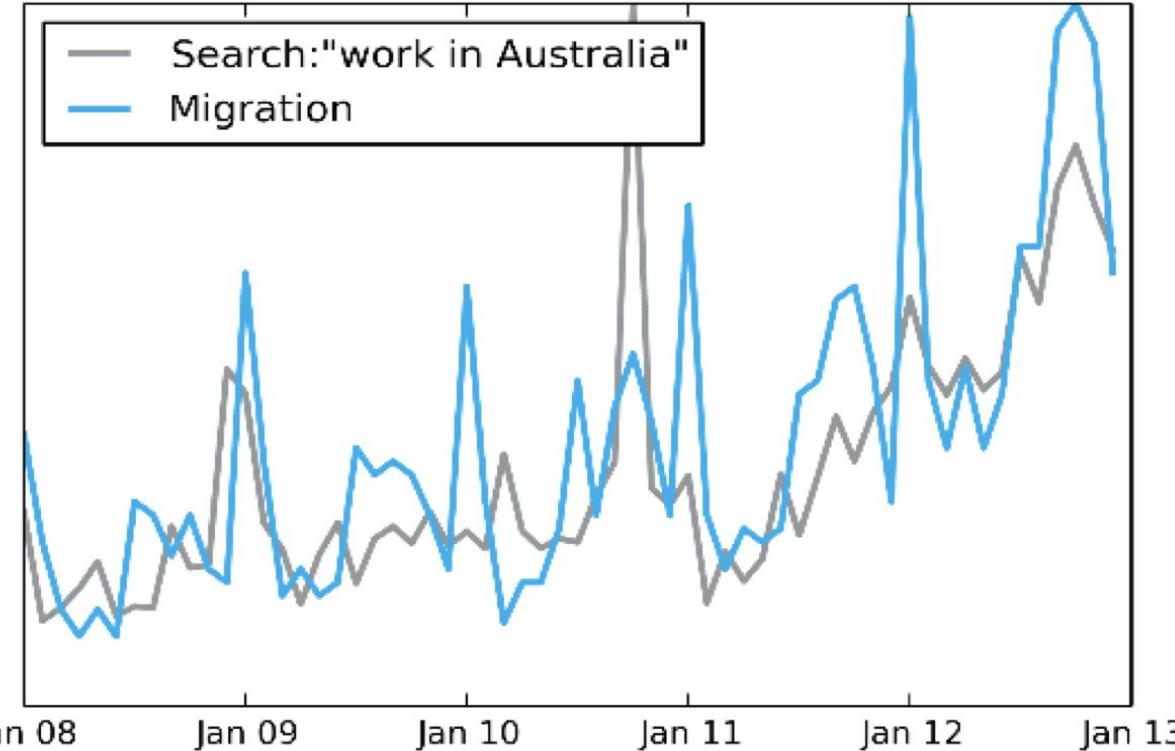
## GOOGLE TRENDS DATA IN MIGRATION STUDIES LITERATURE

- Migration from Latin America to Spain & Google search (Wladyka, 2013)
- UN Global Pulse 2014 – Estimating migration flows using online search data
- Internal migration & Bing search (Lin, Cranshaw & Counts, 2019)
- Syrian refugees & Google search (Connor, 2017)
- Predicting international migration with online search keywords (Böhme, Gröger & Stöhr, 2020)
- Now-casting Romanian migration into the United Kingdom by using Google Search engine data (Avramescu & Wiśniowski, 2021).
- Refugees from Ukraine in Lithuania and the Lithuanian labour market: A preliminary assessment (Deimantas & Sanliturk, 2023)



UN Global Pulse  
(2014), p.13

- **Online search as proxy for migration statistics:** The results of this study demonstrate the potential for online search volumes to be used as proxy for migration statistics. This implies that people interested in migrating conduct online searches to explore employment just prior to migrating, and thus search data could be used as proxy for intent to migrate.

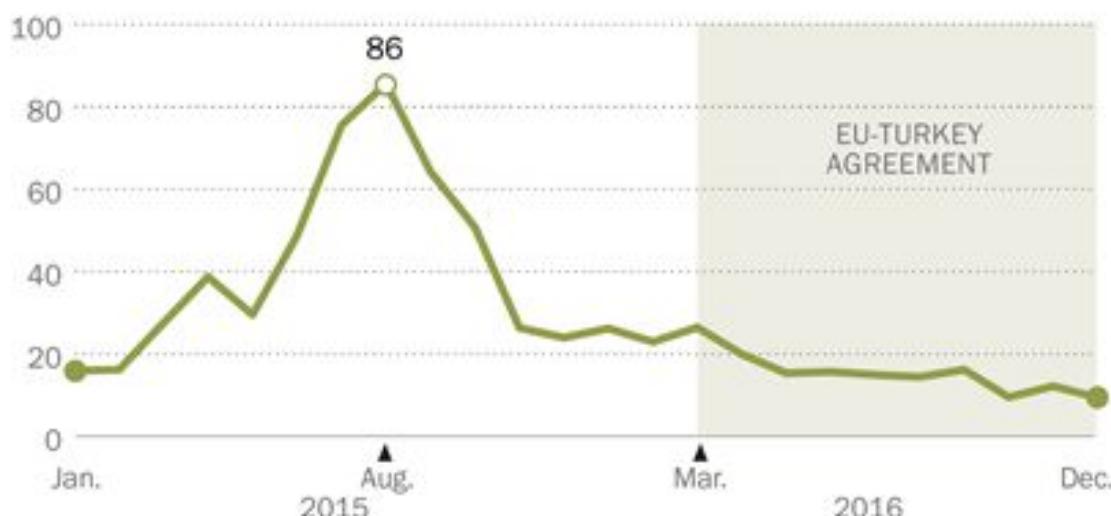


The graph above shows the trend in actual migration from Italy to Australia from January 2008 to December 2013 (blue line) and Google search activity from Italy for the query 'work in Australia' (grey line). The correlation value for migration from Italy to Australia with search query 'work in Australia' is  $r=0.74$ ,  $p<0.001$ .



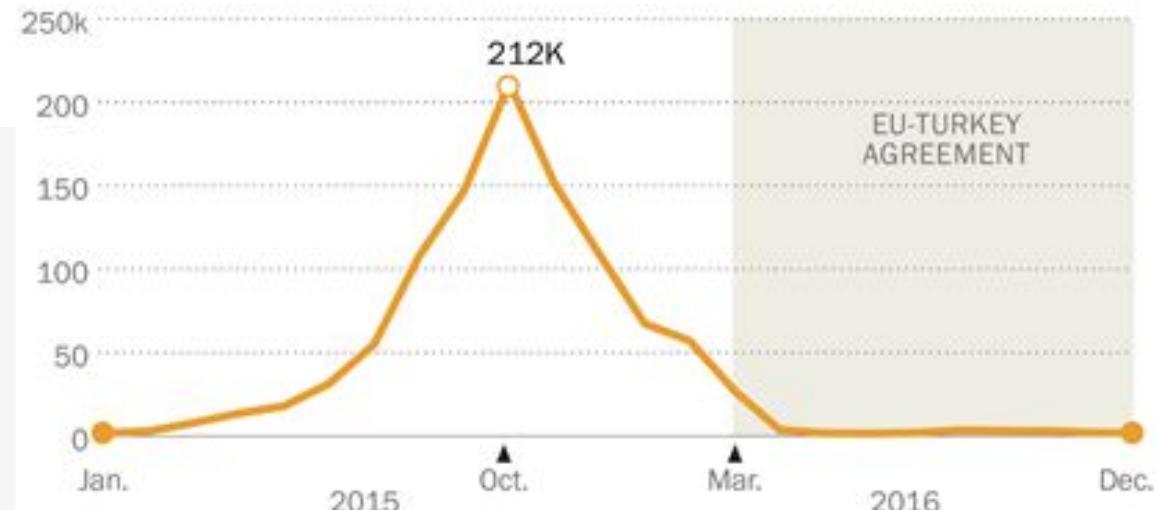
## Surge of Arabic searches for ‘Greece’ in Turkey preceded surge in refugees arriving in Greece

*Google Trends: Relative volume of Arabic-language Google searches for “Greece” by users in Turkey*



Connor. (2017), Pew Research Center

*Monthly arrivals (in thousands) of migrants into Greece*



Note: Google trends data do not indicate the number of searches but instead are standardized data, displaying the relative change in searches over the time period on a 0 to 100 scale. Google trends are monthly averages based on weekly volume. Search data are for the term “Greece” in Arabic (اليونان). Arrivals into Greece are for all nationalities, not only Arabic speakers. See methodology for more details.

Sources: Pew Research Center analysis of Google Trends (accessed on March 3, 2016, at 1:17 p.m.) and United Nations High Commissioner for Refugees (UNHCR) data, accessed March 13, 2017.

“The Digital Footprint of Europe’s Refugees”

PEW RESEARCH CENTER

**Table 1**  
List of main keywords.



English	French	Spanish
applicant	candidat	solicitante
arrival	arrivee	llegada
asylum	asile	asilo
benefit	allocation sociale	beneficio
border control	controle frontiere	control frontera
business	entreprise	negocio
citizenship	citoyennete	ciudadania
compensation	compensation	compensacion
consulate	consulat	consulado
contract	contrat	contrato
customs	douane	aduana
deportation	expulsion	deportacion
diaspora	diaspora	diaspora
discriminate	discriminer	discriminar
earning	revenu	ganancia
economy	economie	economia
embassy	ambassade	embajada
emigrant	emigre	emigrante
emigrate	emigrer	emigrar
emigration	emigration	emigracion
employer	employer	empleador
employment	emploi	empleo
foreigner	etrangeur	extranjero
GDP	PIB	PIB
hiring	embauche	contratacion
illegal	illegal	illegal
immigrant	immigre	inmigrante
immigrate	immigrer	inmigrar
immigration	immigration	inmigracion
income	revenu	ingreso
inflation	inflation	inflacion
internship	stage	pasantia
job	emploi	trabajo
labor	travail	mano de obra
layoff	licenciemment	despido

# **UNDERSTANDING GOOGLE TRENDS DATA IN MIGRATION CONTEXT**





## GOOGLE TRENDS – UNDERSTANDING THE DATA

- Google Trends, while a big data source in itself, limits our access to aggregated and normalized data
- Google Trends gives us a proxy for the intended behavior, i.e. in the case of migration studies intention to move
- Google Trends allows us to form variable for intention to move measured at any given location and any given time
  - as known as Search Popularity Index or Google Trends Index



## OVERLOOK AT GOOGLE TRENDS DATA

- Data does not show the volume of Google searches but its popularity.
- Calculated and normalized by Google
- Data are anonymized, categorized, and aggregated.
- Sample data



## OVERLOOK AT GOOGLE TRENDS DATA

- There are two types of Google Trends data that can be accessed:
- Real-time data covering the last seven days.
  - Time unit: hour
- Non-real time data (a separate sample from real-time data)
  - Between 2004 and up to 36 hours prior



## GOOGLE TRENDS – DATA PROCESSING

- Google processes data prior to reporting Google Trends output.
- The data pre-processing includes;
  - filtering irregular activities (some may still remain),
  - sampling,
  - placing thresholds



## GOOGLE TRENDS – DATA PROCESSING

- Google Trends data excludes;
- Search terms with low volume that cannot pass the threshold (appear as "0")
- Repeated searches from the same person over a short period of time as irregular activity.
- Queries with apostrophes and other special characters.

(Trends Help, 2021)



## REPRESENTATIVENESS

- Google Trends output is calculated based on a representative sample instead of the entire volume of Google searches. This is due to the too big volume of Google searches, exceeding billions of searches per day.
- We don't know the exact sampling methodology used by Google.
  
- Even if you search for trends using the same parameters, you may get very slightly different results, due to the sample. These are statistically not significant – but can do a robustness check.



## REPRESENTATIVENESS

- Beware of the representation bias while using digital trace data
- Google usage is mostly more widespread than use of a certain social media outlet, but is still bound by the same limitations
- In statistical analyses using an adjustment factor is encouraged
  - such as the Google search engine market share or internet penetration rate



## NON-REAL TIME DATA – REPORTING

□ Time unit of non-real time data reporting depends on the selected time period

- Up to 7 days      hourly data
- **Up to 9 months** daily data
- Up to 5 years      weekly data
- + 5 years            monthly data

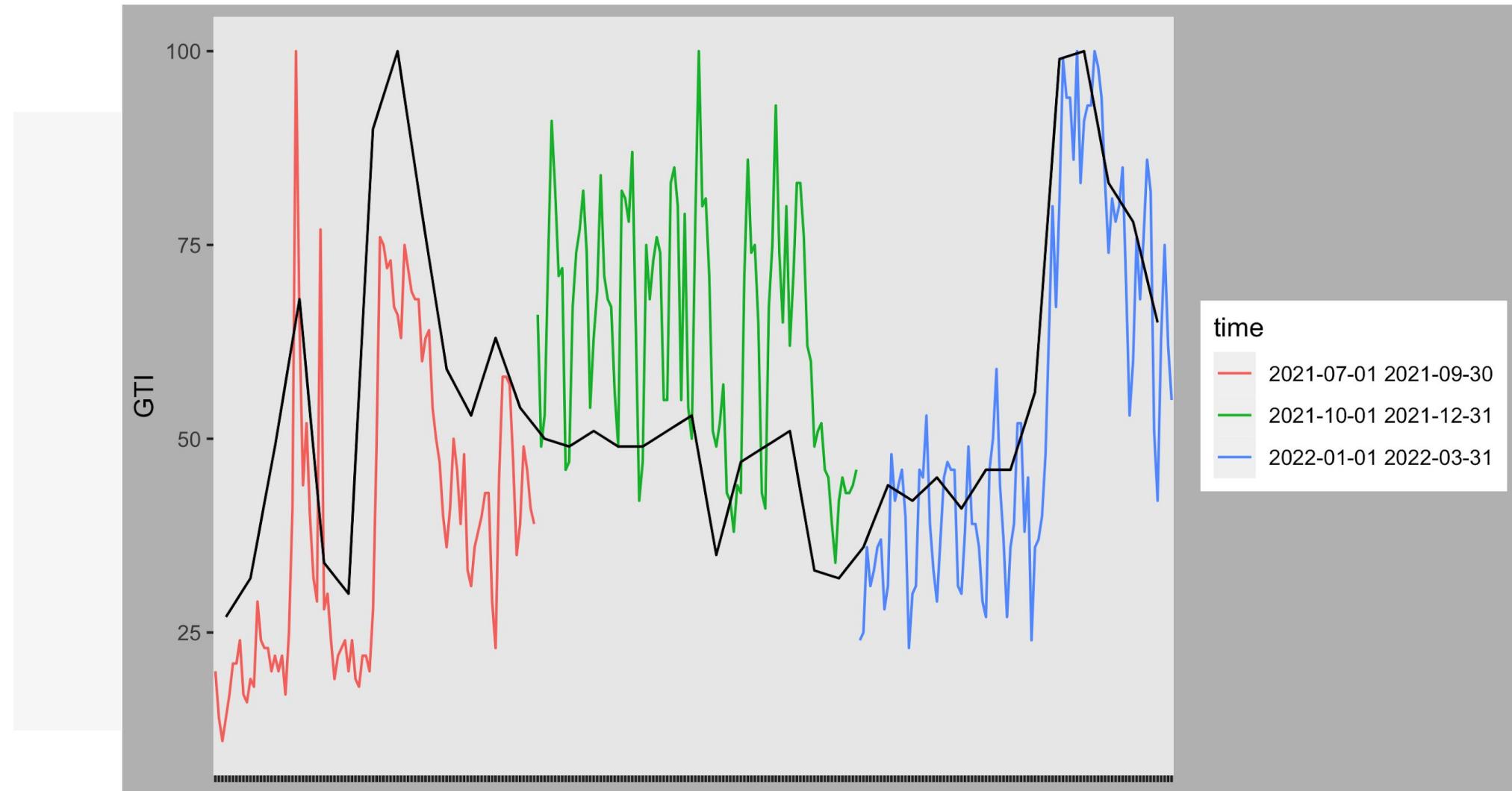


## EXTENDING THE TIME PERIOD

- If we need daily data for longer than 9 months or weekly data for longer than 5 years, we need to download them separately.
- Normalization problem
- Google Trends normalizes the data for the given time period. Merging different time periods requires additional adjustments or normalization.



## MERGED DAILY DATA



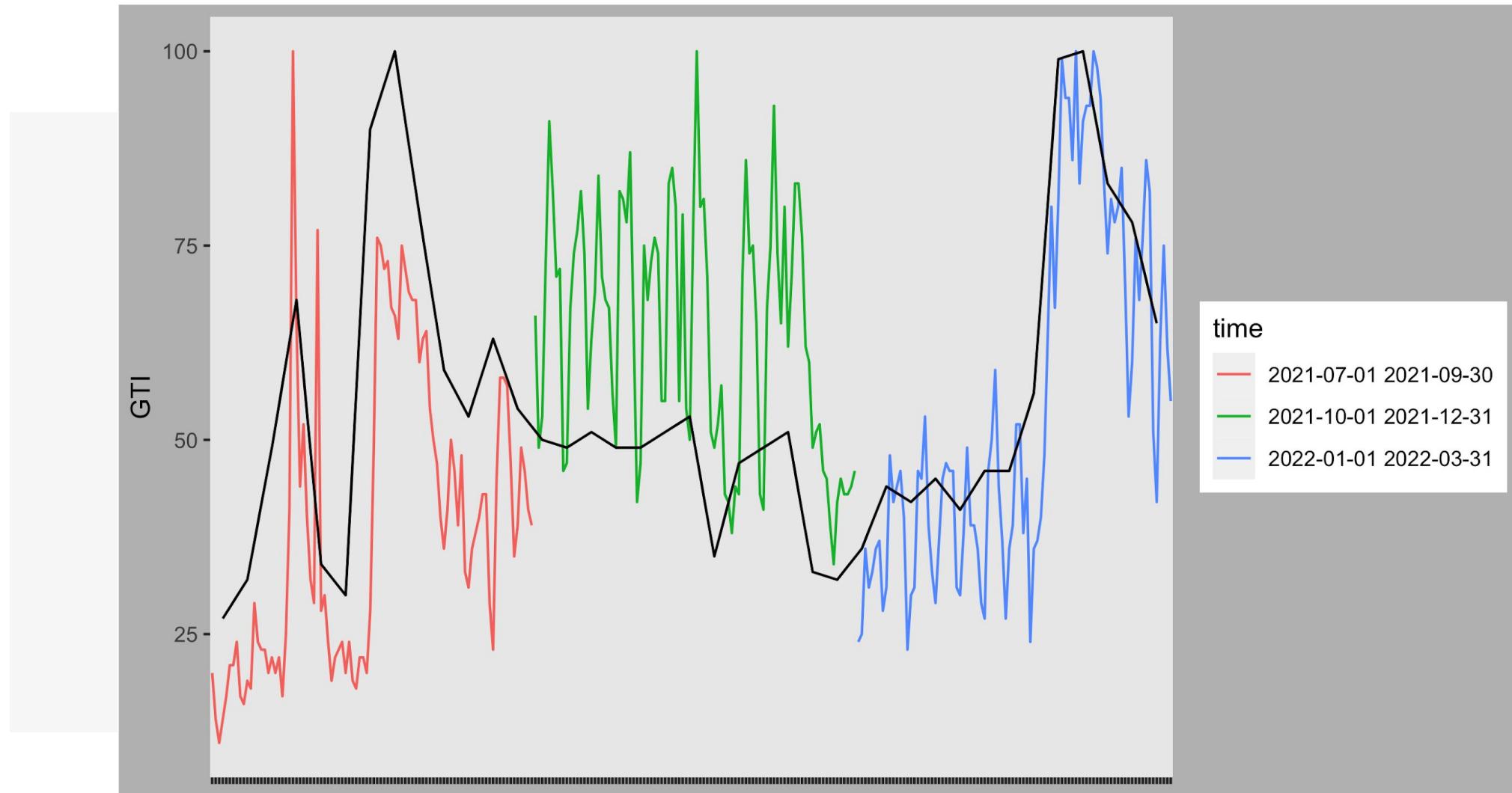


## EXTENDING THE TIME PERIOD

- Create an adjustment factor
  - Combine daily (or weekly) data sets
  - Download weekly (or monthly) data set for the same time period
  - Calculate the adjustment factor by the overlapping dates and apply the adjustment to the daily data of the same week (weekly data of the same month)
    - (Johansson, 2014; Risteski & Davcev, 2014)
  - Rescale to 0-100 range

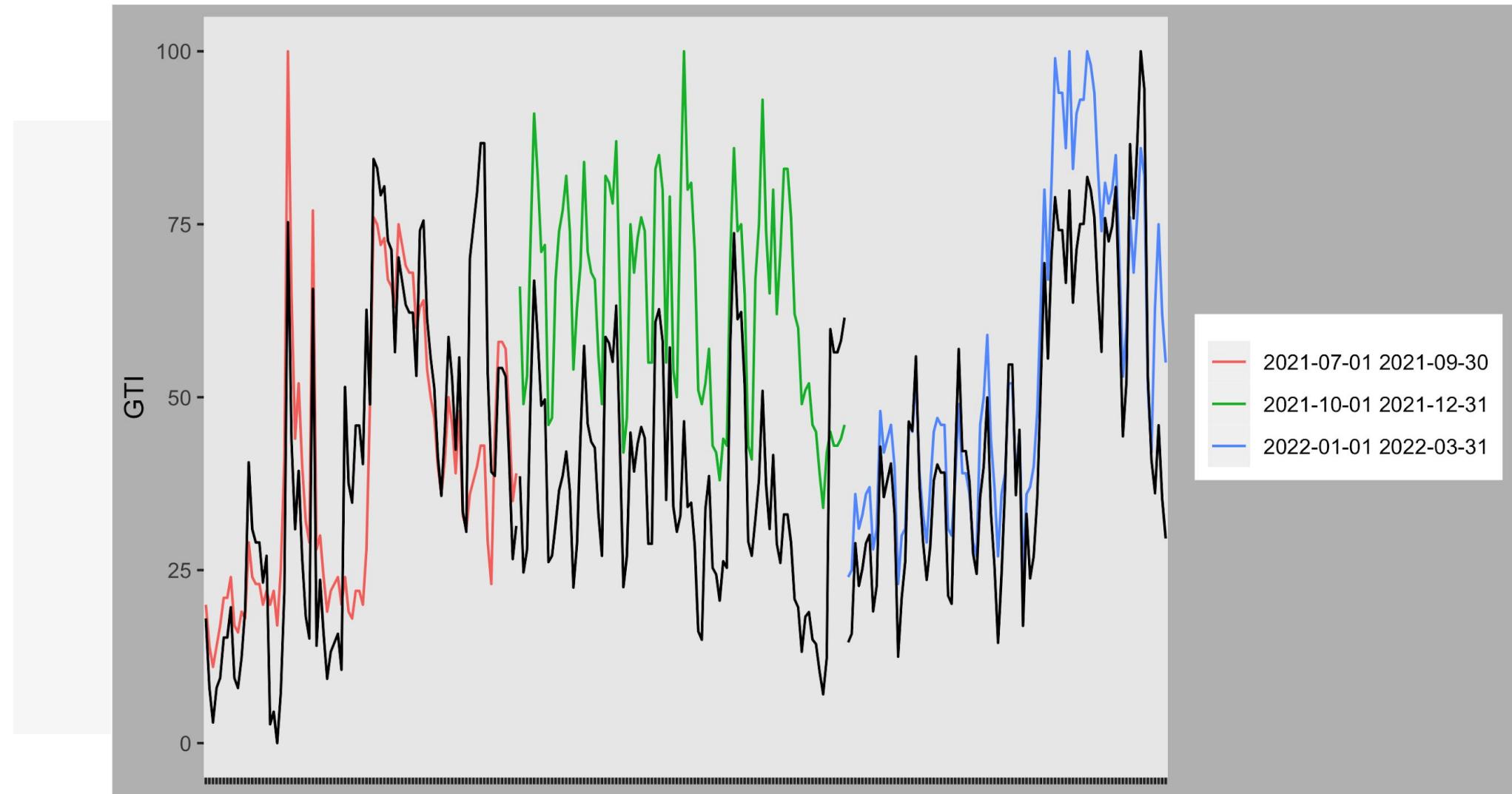


## EXTENDING THE TIME PERIOD





## EXTENDING THE TIME PERIOD



## USING ON GOOGLE TRENDS DATA





## HOW TO MAKE A QUERY

Search term	Results
tennis shoes	<p>Results can include searches containing both tennis and shoes in any order. Results can also include searches like "red tennis shoes," "funny shoes for tennis," or "tennis without shoes."</p> <p>No misspellings, spelling variations, synonyms, plural, or singular versions of your terms are included.</p>
"tennis shoes"	<p>Results include the exact phrase inside double quotation marks, possibly with words before or after, like "red tennis shoes."</p>
tennis + squash	<p>Results can include searches containing the words "tennis" OR "squash."</p>
tennis -shoes	<p>Results include searches containing the word "tennis," but exclude searches with the word "shoes."</p>
center + centre + centere	<p>Results include alternative spellings like "centre" or "centere," and common misspellings like "centere." Trends considers each version of a word a different search, including misspellings.</p>

Google. (2022), Trends Help: Search Tips for Trends, [https://support.google.com/trends/answer/4359582?hl=en&ref\\_topic=4365530](https://support.google.com/trends/answer/4359582?hl=en&ref_topic=4365530)

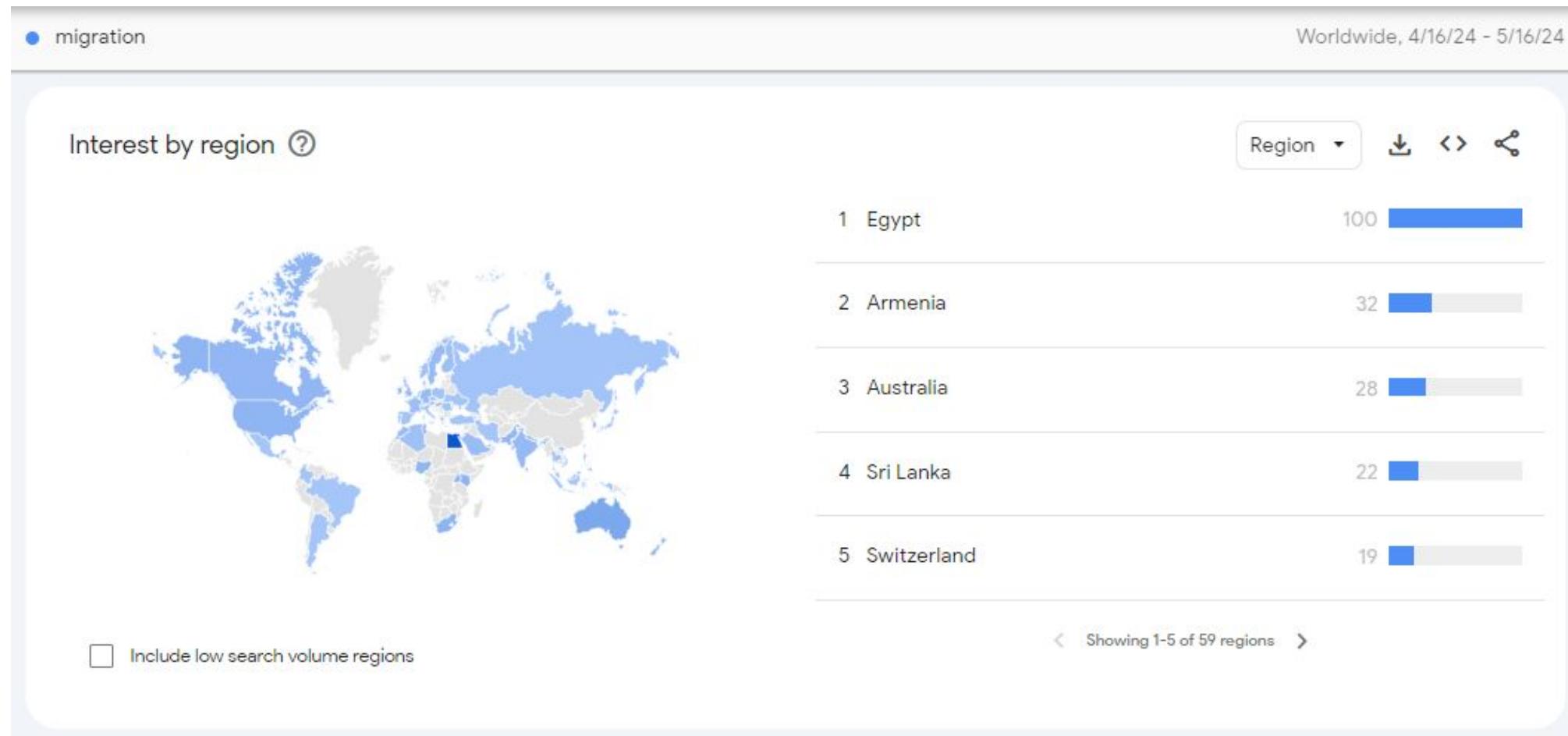


## HOW TO MAKE A QUERY

- Being more specific with selected keywords helps narrow down the focus to the matter of interest
- When you determine the parameters, download the data using *gtrendsR* package by Massicotte & Eddelbuettel (or *pytrends* on Python)
  - For further information, see <https://github.com/PMassicotte/gtrendsR>

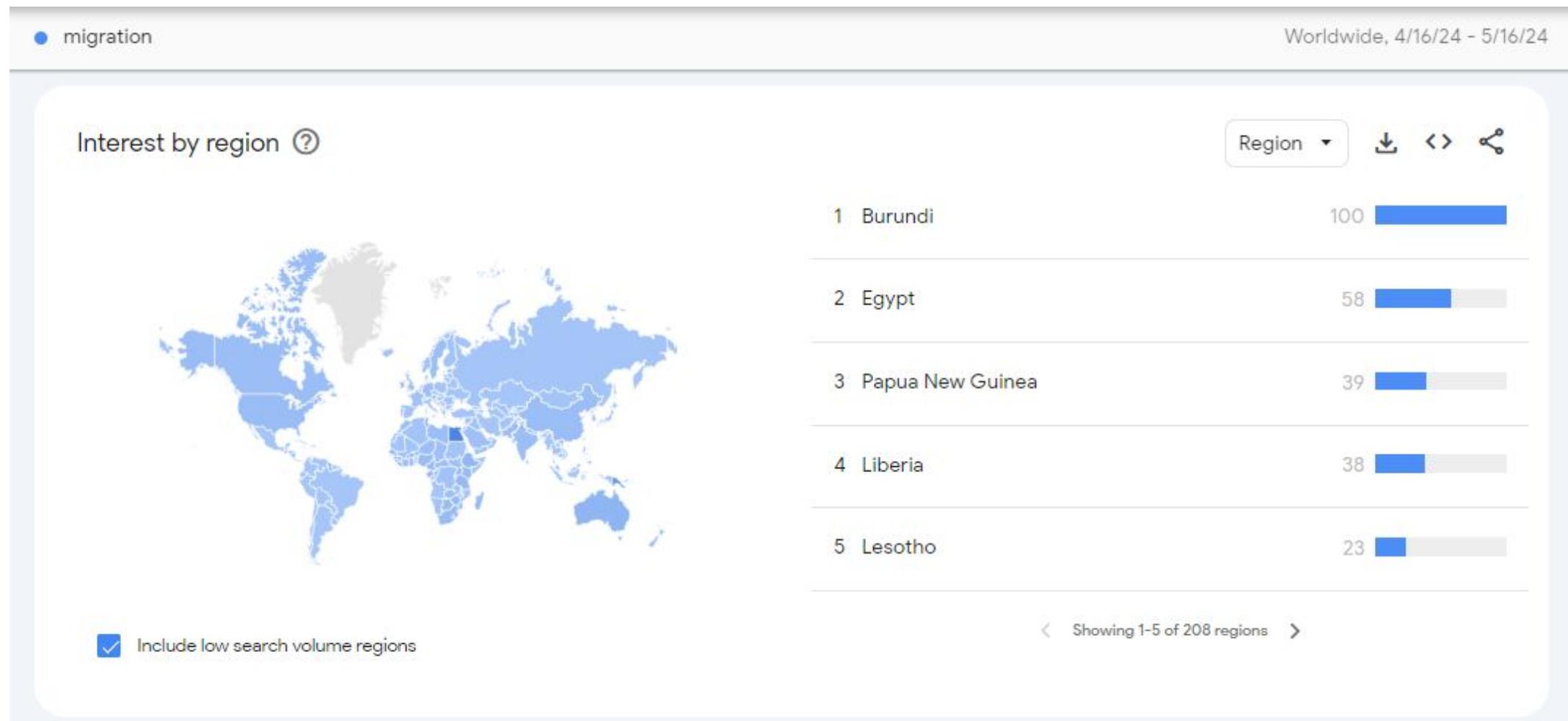


# GOOGLE TRENDS VISUALIZATION





# GOOGLE TRENDS VISUALIZATION



# ALTERNATIVES?





## YANDEX KEYWORD STATISTICS

- Useful for research on Russian-speaking communities
- Used in several countries apart from Russia, but Google is clearly the market leader
- <https://wordstat.yandex.com>



## YANDEX KEYWORD STATISTICS

- Monthly data for 2 years, weekly data for 1 year
- Custom date selection is not possible
- Provides absolute numbers of searches as well as relative figures
- Provides a distinction between searches made on all and mobile devices



# YANDEX KEYWORD STATISTICS

[Direct](#) [Directory](#) [Metrica](#) [Advertising Network](#) [Market](#) [more](#)

[Logout](#)



migration

By keyword

By region

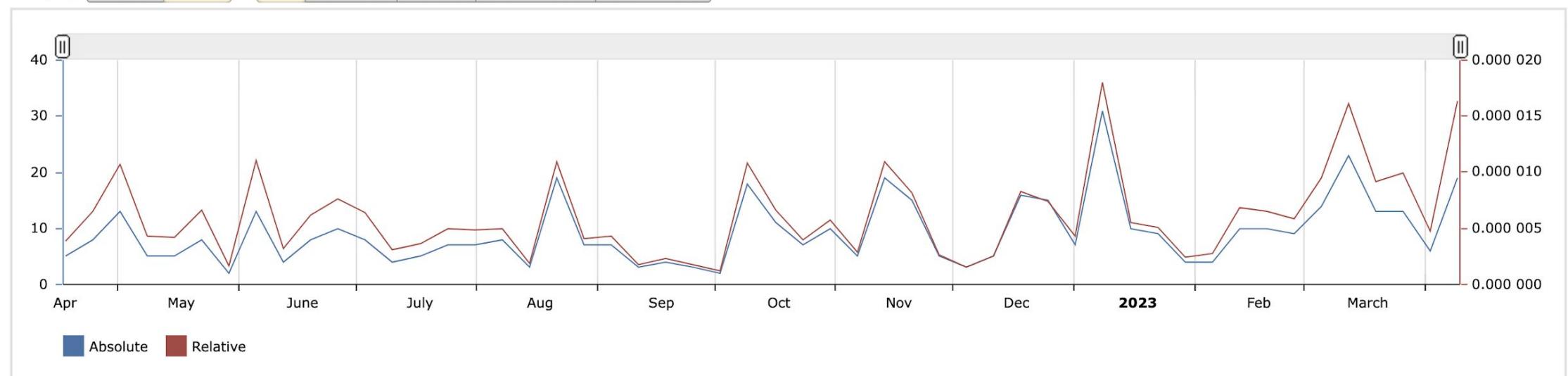
Query history

United Kingdom

[Submit](#)

Impressions history for keyword "migration"

Group by: [month](#) [week](#) [All](#) [Desktop](#) [Mobile](#) [Phones only](#) [Tablets only](#) [?](#)

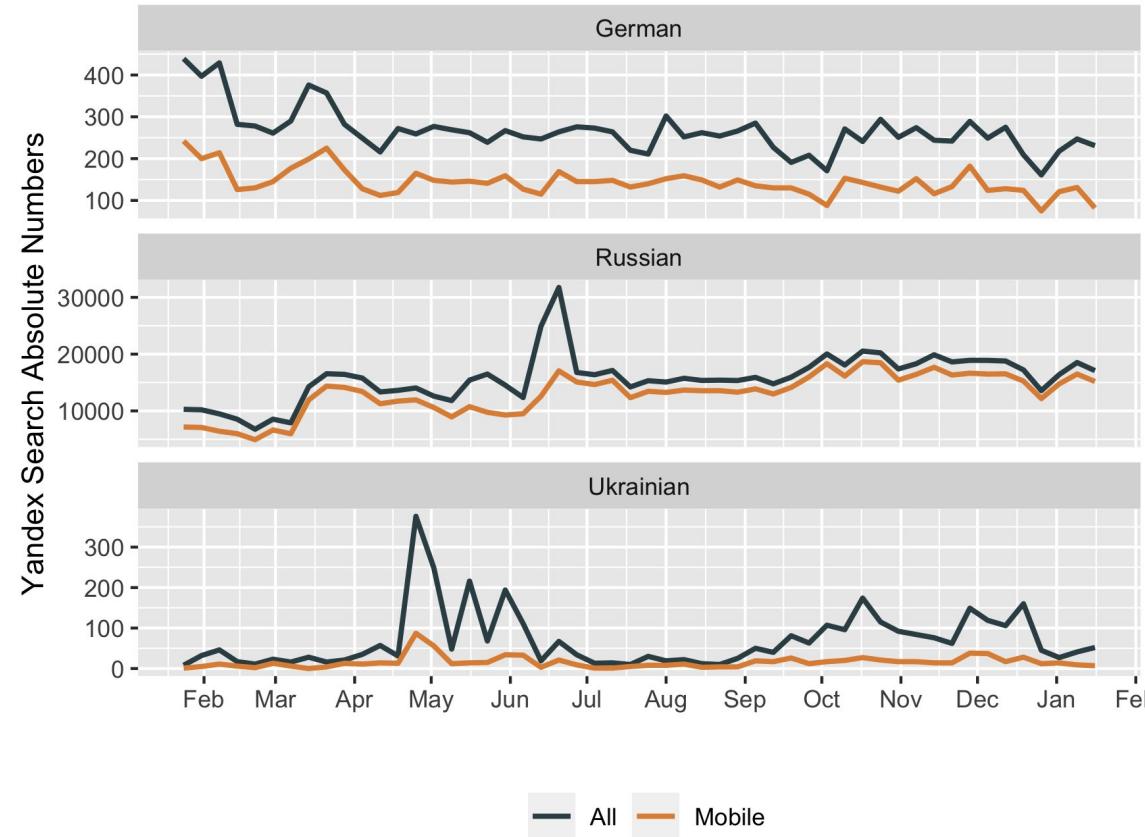


Period	Absolute	Relative	Period	Absolute	Relative
11.04.2022 - 17.04.2022	5	0.000 003 859 177	10.10.2022 - 16.10.2022	11	0.000 006 683 209
18.04.2022 - 24.04.2022	8	0.000 006 478 142	17.10.2022 - 23.10.2022	7	0.000 003 924 607
25.04.2022 - 01.05.2022	13	0.000 010 695 592	24.10.2022 - 30.10.2022	10	0.000 005 731 116
02.05.2022 - 08.05.2022	5	0.000 004 285 441	31.10.2022 - 06.11.2022	5	0.000 002 855 059

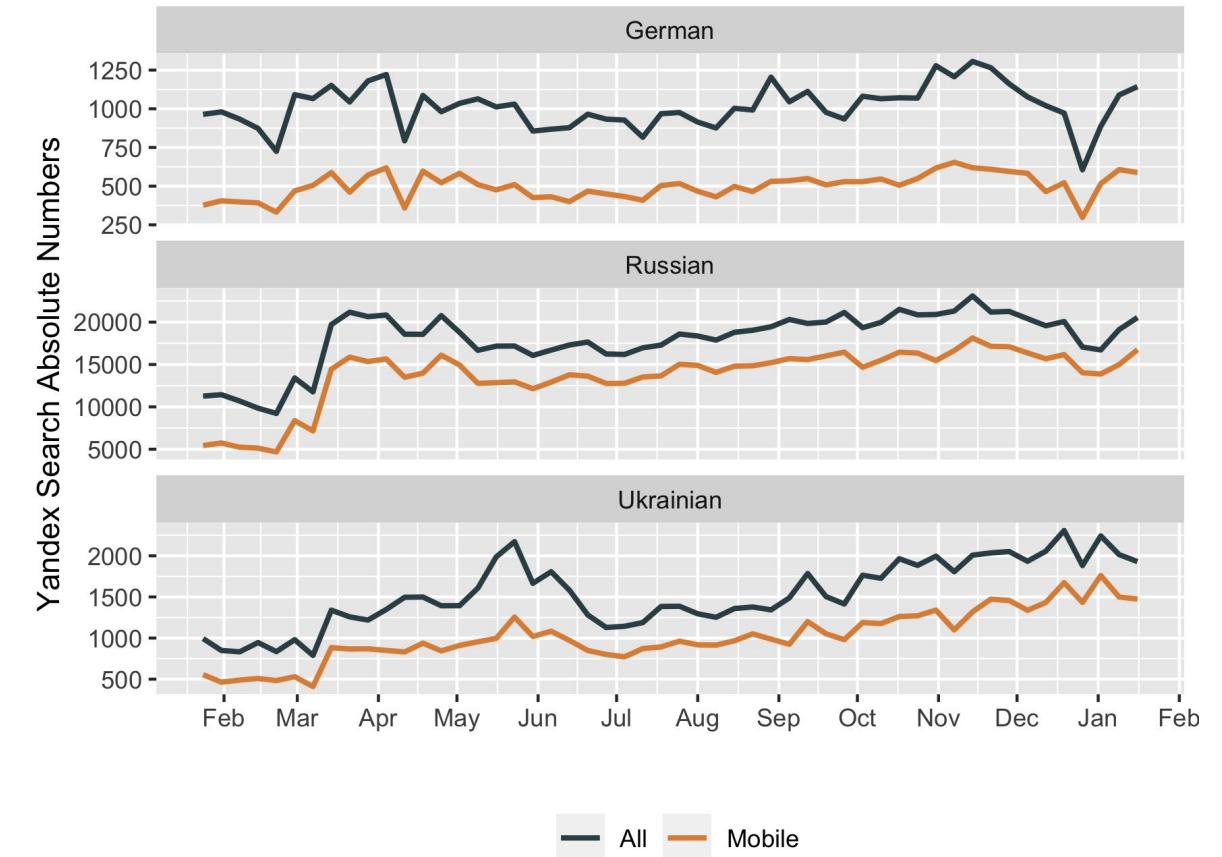


# YANDEX KEYWORD STATISTICS

## Health-related keywords



## Job-related keywords

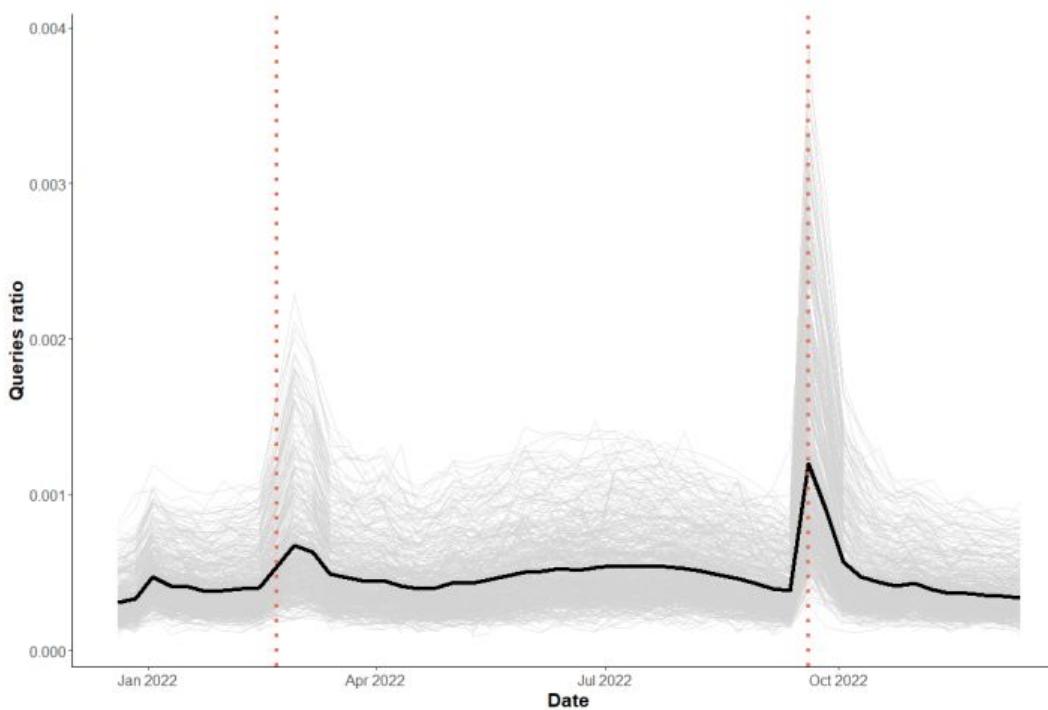




# YANDEX KEYWORD STATISTICS

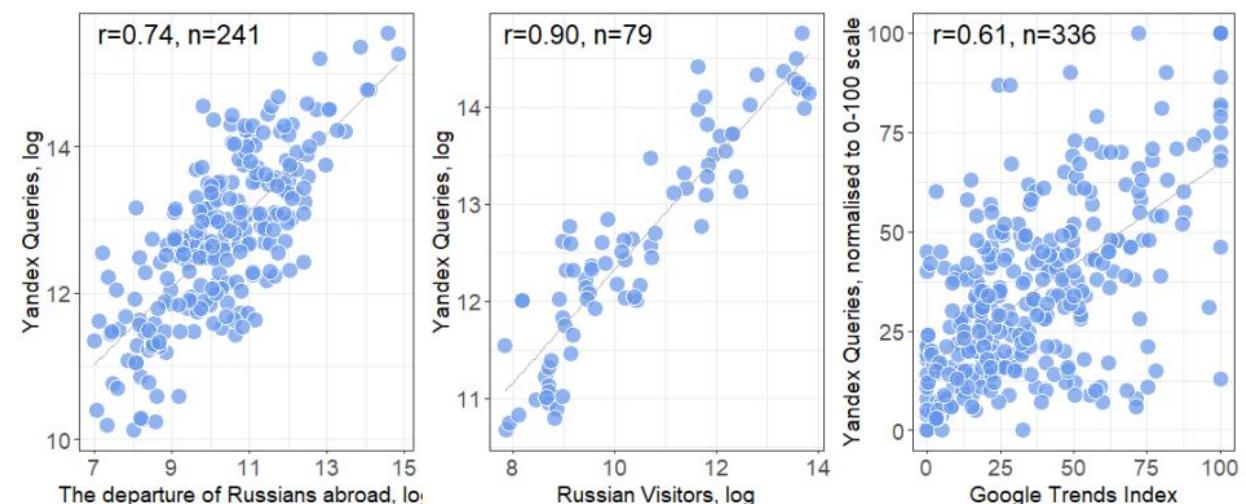
*Demographic Research: Volume 50, Article 8*

**Figure 1:** Mobility-related internet searches across Russia



Note: Mobility-related searches (share of total weekly searches) performed at the city level (grey lines) in Russia and their average (black line). Red dots indicate February 24, 2022 (beginning of the full-scale invasion) and September 21 (beginning of mobilization).

**Figure 3:** Correlations between Yandex web searches and other mobility indicators



Note: Correlations between total mobility-related searches on Yandex and departures of Russians abroad, border crossings into selected destination countries, and respective Google Trends Indices indicate that Yandex web searches are a reasonable proxy for measuring migration intentions, similar to Google Trends. The left plot shows the number of departures of Russians abroad in each quarter from 2020 Q4 to 2022 Q3 for each potential destination country in our sample and the corresponding quarterly number of Yandex searches for those countries.

The middle plot shows the monthly number of Russian visitors to Turkey, Georgia, and Serbia from January 2021 to November 2022 and the corresponding monthly number of Yandex searches for those countries.

The right plot depicts the monthly value of the Google Trends Index for mobility-related queries in each country in the sample from December 2020 to November 2022 and the corresponding monthly number of Yandex searches for those countries normalized within the country.



## BING SEARCH?

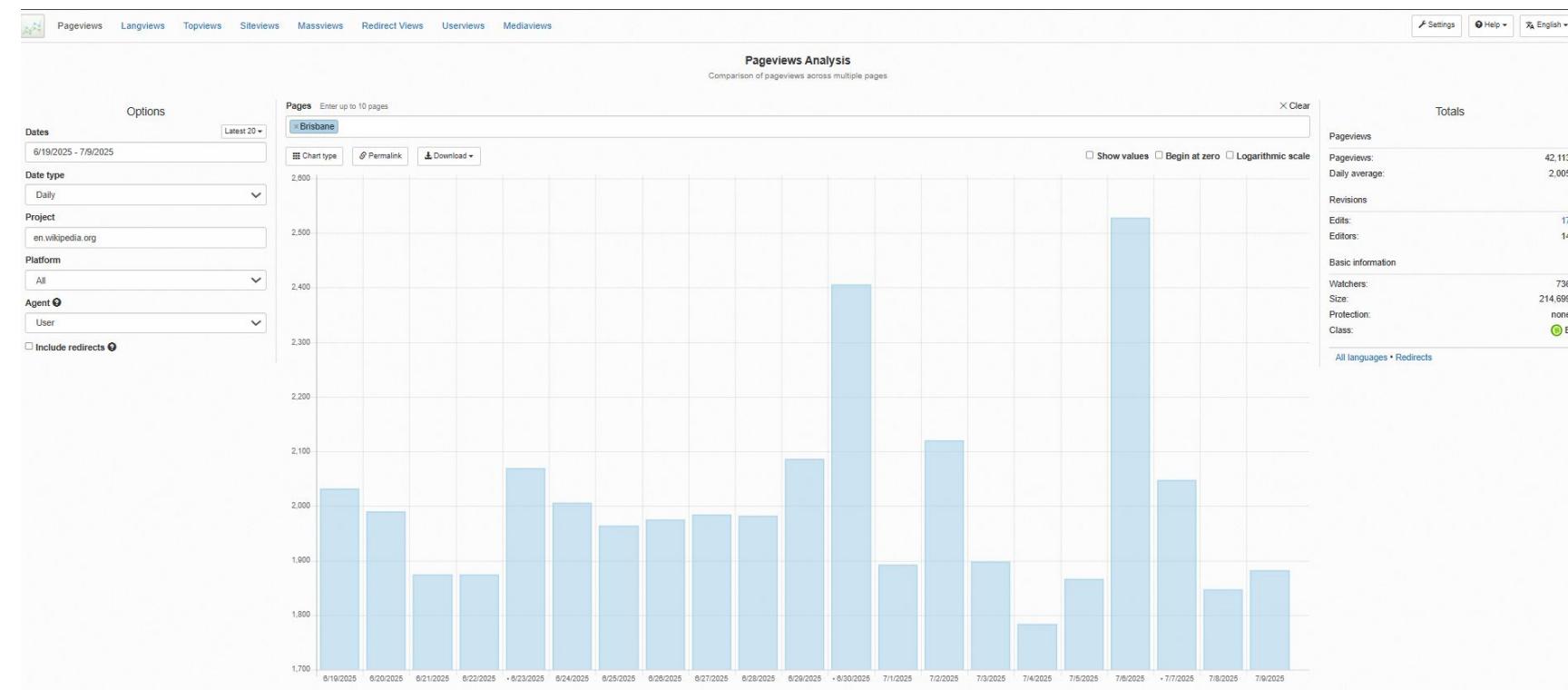
- See:
  - Lin, A. Y., Cranshaw, J., & Counts, S. (2019). Forecasting US Domestic Migration Using Internet Search Queries. Proceedings of the 2019 World Wide Web Conference (WWW'19), (pp. 13-17).



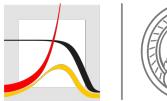
# WIKIPEDIA DATA

## PAGEVIEW STATISTICS

- Allows for language distinction
- Provides daily data on total number of views (and total number of edits)
- Does not have location information



# WIKIPEDIA DATA FOR MIGRATION STUDIES

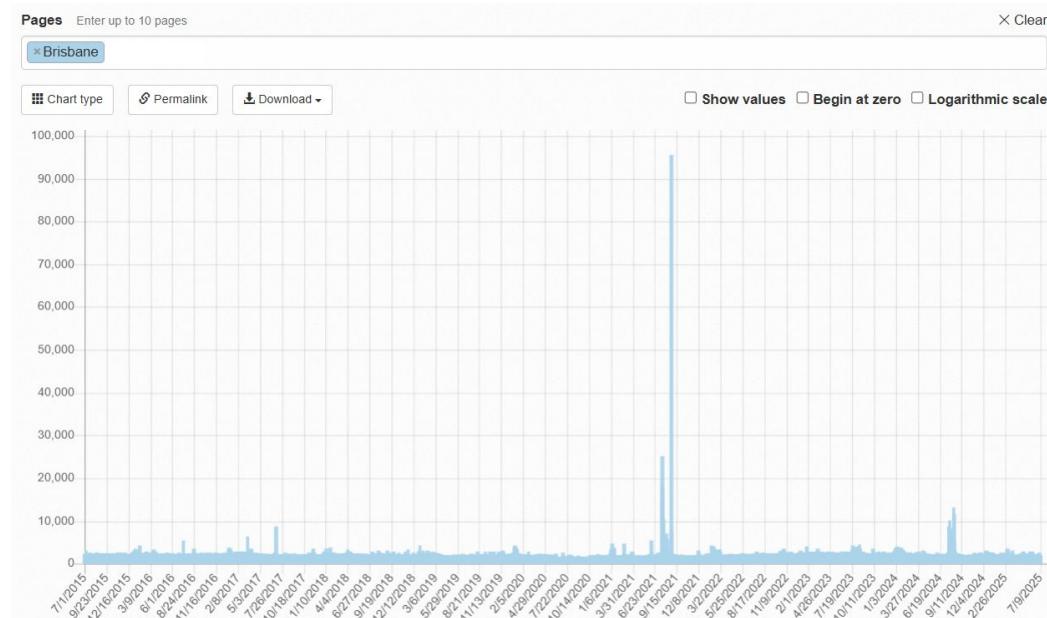




# Wikipedia data

Wikipedia Pageviews: <https://pageviews.wmcloud.org/>

Provides **daily view counts** for each **Wikipedia page** of interest across all available languages, starting **from July 2015**.



WIKIPEDIA

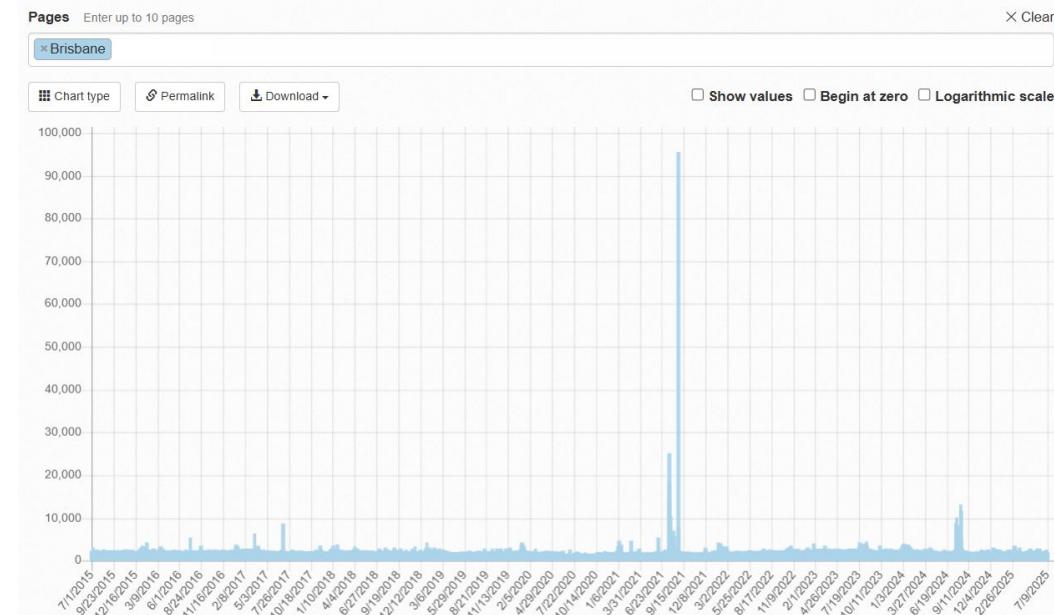


# Wikipedia data

Wikipedia Pageviews: <https://pageviews.wmcloud.org/>

Provides **daily view counts** for each **Wikipedia page** of interest across all available **languages**, starting from **July 2015**.

This data do not have the location where the views come from. Language is used as a proxy for origin.



WIKIPEDIA



## Langviews Analysis

Pageviews of a page across all languages

Dates	Source project	Platform	Agent
7/1/2015 - 4/14/2025	en.wikipedia.org	All	User
Page			
Star Wars	<input type="button" value="Submit"/>		

# Langviews Analysis

Pageviews of a page across all languages

[◀ Do another query](#)

Brisbane 7/1/2015 - 7/9/2025

List

Chart

[🔗 Permalink](#)

[⬇️ Download ▾](#)

#	Language	Page title	Badges	Pageviews ↓ <sup>?</sup>	Daily average
Totals	144 languages	49 unique titles	✖️ × 1, + × 1, × 1	15,177,701	4,145
1	en	Brisbane		8,355,192	2,282
2	de	Brisbane		873,404	239
3	es	Brisbane		720,433	197
4	zh	布里斯班		686,026	187
5	ja	ブリスベン		676,842	185

# Langviews Analysis

Pageviews of a page across all languages

◀ Do another query

Brisbane 7/1/2015 - 7/9/2025

🔗 Permalink

⬇️ Download ▾

#	Language	Page title	Pageviews ↓ <sup>z</sup>	Daily average
Totals	144 languages	49 unique titles	15,177,701	4,145
1	en	Brisbane	8,355,192	2,282
2	de	Brisbane	873,404	239
3	es	Brisbane	720,433	197
4	zh	布里斯班	686,026	187
5	ja	ブリスベン	676,842	185

▼ array [144]

▼ 0 {9}

▼ data [3662]

0 : 2086

1 : 2104

2 : 2032

3 : 2056

4 : 2445

5 : 3015

6 : 2370

7 : 2492

Wikipedia page about Brisbane is available in 144 languages

For each language, each data point corresponds to the number of views for the Wikipedia page about Brisbane in language L on the date D.



# How could we use Wikipedia data to study migration?



## Context

Imagine a situation of **crisis**, such as conflicts or war, resulting in **large-scale migration** of people.



## Context

Imagine a situation of **crisis**, such as conflicts or war, resulting in **large-scale migration** of people.

...who **seek for information** quickly and efficiently to **decide where to go**

and receive all the **support** they need at the **destination**.



## Context

Imagine a situation of **crisis**, such as conflicts or war, resulting in **large-scale migration** of people.

...who **seek for information** quickly and efficiently to **decide where to go**

and receive all the **support** they need at the **destination**.

There is a growing literature that claims that **online sources of information** are used during their journey and after arriving at the destination, as important sources of information.

(Zimmer and Scheibe, 2020; Merisalo and Jauhainen, 2020)



## Hypothesis

While **moving** and right after crossing the border out of their countries, refugees might **search** for places to move in to.



# Hypothesis

While **moving** and right after crossing the border out of their countries, refugees might **search** for places to move in to.

In this study, we aim to shed light on the relationship between **refugee flows** and **online sources of information (Wikipedia\*)**



\* Wikipedia free online encyclopedia and a worldwide well-known source of information.

WIKIPEDIA

## Contents [hide]

(Top)

[Toponymy and names](#)[History](#)[Geography](#)[Cityscape](#)[Demographics](#)[Government and politics](#)[Economy](#)[Education](#)[Transport](#)[Culture](#)[Sports](#)[Famous people](#)[Rankings](#)[International relations](#)[See also](#)[Notes](#)[References](#)[Bibliography](#)[Further reading](#)[External links](#)

# Warsaw

[Article](#) [Talk](#)[Read](#) [Edit](#) [View history](#) [Tools](#) ▾

From Wikipedia, the free encyclopedia

Coordinates: 52°13'48"N 21°00'40"E

*For other uses, see [Warsaw \(disambiguation\)](#).**Several terms redirect here. For other uses, see [Warszawa \(disambiguation\)](#), [Warschau \(disambiguation\)](#), and [City of Warsaw \(disambiguation\)](#).*

**Warsaw**,<sup>[a]</sup> officially the **Capital City of Warsaw**,<sup>[6][b]</sup> is the **capital** and **largest** city of Poland. The metropolis stands on the [River Vistula](#) in east-central Poland. Its population is officially estimated at 1.86 million residents within a greater metropolitan area of 3.1 million residents, which makes Warsaw the 6th most populous city in the European Union.<sup>[2]</sup> The city area measures 517 km<sup>2</sup> (200 sq mi) and comprises 18 districts, while the metropolitan area covers 6,100 km<sup>2</sup> (2,355 sq mi).<sup>[7]</sup> Warsaw is an alpha global city,<sup>[8]</sup> a major cultural, political and economic hub, and the country's seat of government. It is also capital of the [Masovian Voivodeship](#).

Warsaw traces its origins to a small fishing town in [Masovia](#). The city rose to prominence in the late 16th century, when [Sigismund III](#) decided to move the Polish capital and his royal court from [Kraków](#). Warsaw served as the de facto capital of the [Polish–Lithuanian Commonwealth](#) until 1795, and subsequently as the seat of [Napoleon's Duchy of Warsaw](#). The 19th century and its [Industrial Revolution](#) brought a demographic boom which made it one of the largest and most densely populated cities in Europe. Known then for its elegant architecture and boulevards, Warsaw was [bombed](#) and [besieged](#) at the start of [World War II](#) in 1939.<sup>[9][10][11]</sup> Much of the historic city was destroyed and its diverse population decimated by the [Ghetto Uprising](#) in 1943, the general [Warsaw Uprising](#) in 1944 and systematic razing.

Warsaw is served by two international airports, the busiest being [Warsaw Chopin](#) and the smaller [Warsaw Modlin](#) intended for low-cost carriers. Major public transport services operating in the city include the [Warsaw Metro](#), [buses](#), [commuter rail service](#) and an extensive [tram network](#). The city is a significant centre of research and development, [business process outsourcing](#), and [information technology outsourcing](#). The [Warsaw Stock Exchange](#) is the largest and most important in Central and Eastern Europe.<sup>[12][13]</sup> [Frontex](#), the European Union agency for external border security, and [ODIHR](#), one of the principal institutions of the Organization for Security and Co-operation in Europe, have their headquarters in Warsaw. As of 2022, Warsaw has one of the highest number of skyscrapers in Europe while [Varso Place](#) is the tallest building in the European Union.

<b>Warsaw</b>
<a href="#">Warszawa (Polish)</a>
<a href="#">Capital city and county</a>
<a href="#">Capital City of Warsaw</a> Polish: <i>miasto stołeczne Warszawa</i>



Warsaw business district

Royal Castle and  
Sigismund's ColumnNicolaus Copernicus  
Monument

Łazienki Park

Main Market Square



# Варшава [ред. | ред. код]

239 мов ▾

Зміст сховати

## Вступ

Етимологія

> Географія

Історія

Демографія

> Уряд

Економіка

Освіта

> Культура

> Транспорт

> Українці у Варшаві

Відомі варшав'яни

Галерея

Міста-побратими

Стаття

Обговорення

Читати

Редактувати

Редактувати код

Переглянути історію

Інструменти

Зовнішній вигляд

сховати

Текст

Малий

Стандартний

Великий

Ширина

Стандартний

Широкий

Кольорова тема (бета)

Автоматична

Світла

Темна

Матеріал з Вікіпедії — вільної енциклопедії.

**Варшава** (пол. Warszawa), офіційно **столичне місто**

**Варшава** (пол. miasto stołeczne Warszawa) — столиця

Польщі з 1596 року, порт на річці Вісла,

адміністративний центр Мазовецького воєводства.

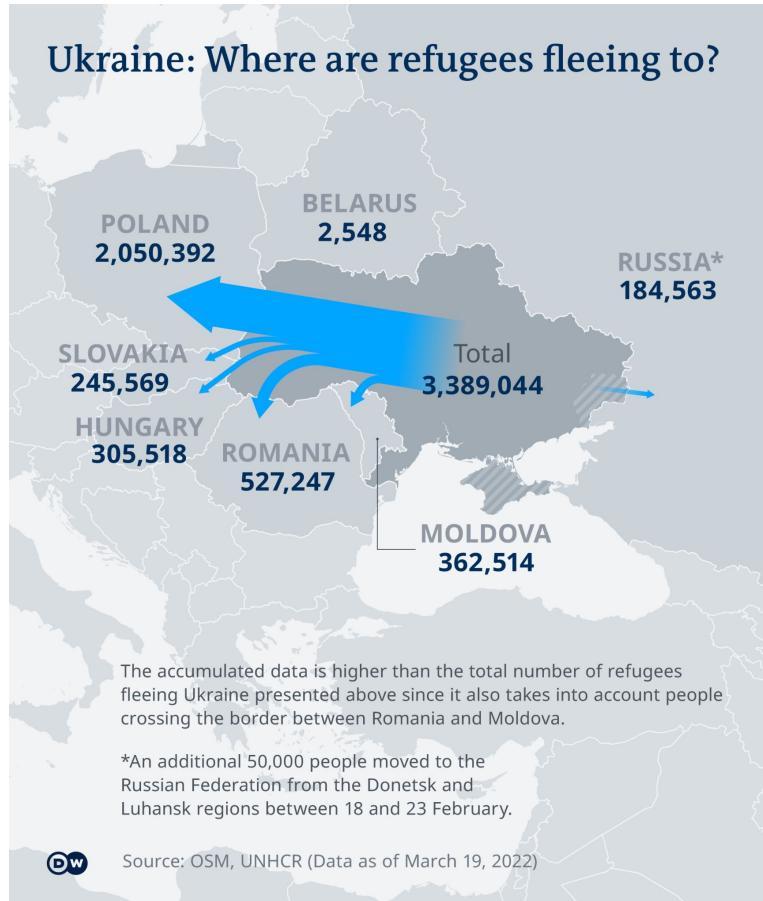
Місто є місцем розташування центральних органів влади Республіки Польща, іноземних місій, штаб-квартир значної кількості підприємств та громадських об'єднань, що працюють в Польщі.

Варшава з населенням 1 860 281<sup>[20]</sup> жителів є найбільшим містом Польщі й утворює другу найбільшу агломерацію в країні (після Катовицької конурбації).

Розвинені **машинобудування** (автобудування, електротехнічне та ін.), металообробка і металургія,



## Case study:



### The Ukrainian refugee crisis in 2022

~ 6.5 million Ukrainian refugees recorded globally.  
Poland has welcomed the greatest number (~ 60%) of Ukrainian refugees.

## **Research questions:**

**RQ1: How has the Ukrainian refugee crisis affected the information-seeking behavior on Wikipedia?**

**RQ2: What is the temporal relationship between the information-seeking behavior on Wikipedia and Ukrainian refugee flows?**

## Datasets

**Wikipedia Pageviews data:** daily number of views on Wikipedia pages about European capitals and the most populous cities in Poland

**United Nations High Commissioner for Refugees (UNHCR) data:**  
Europe: total number of Ukrainian refugees in European countries as of October 2023

Poland:

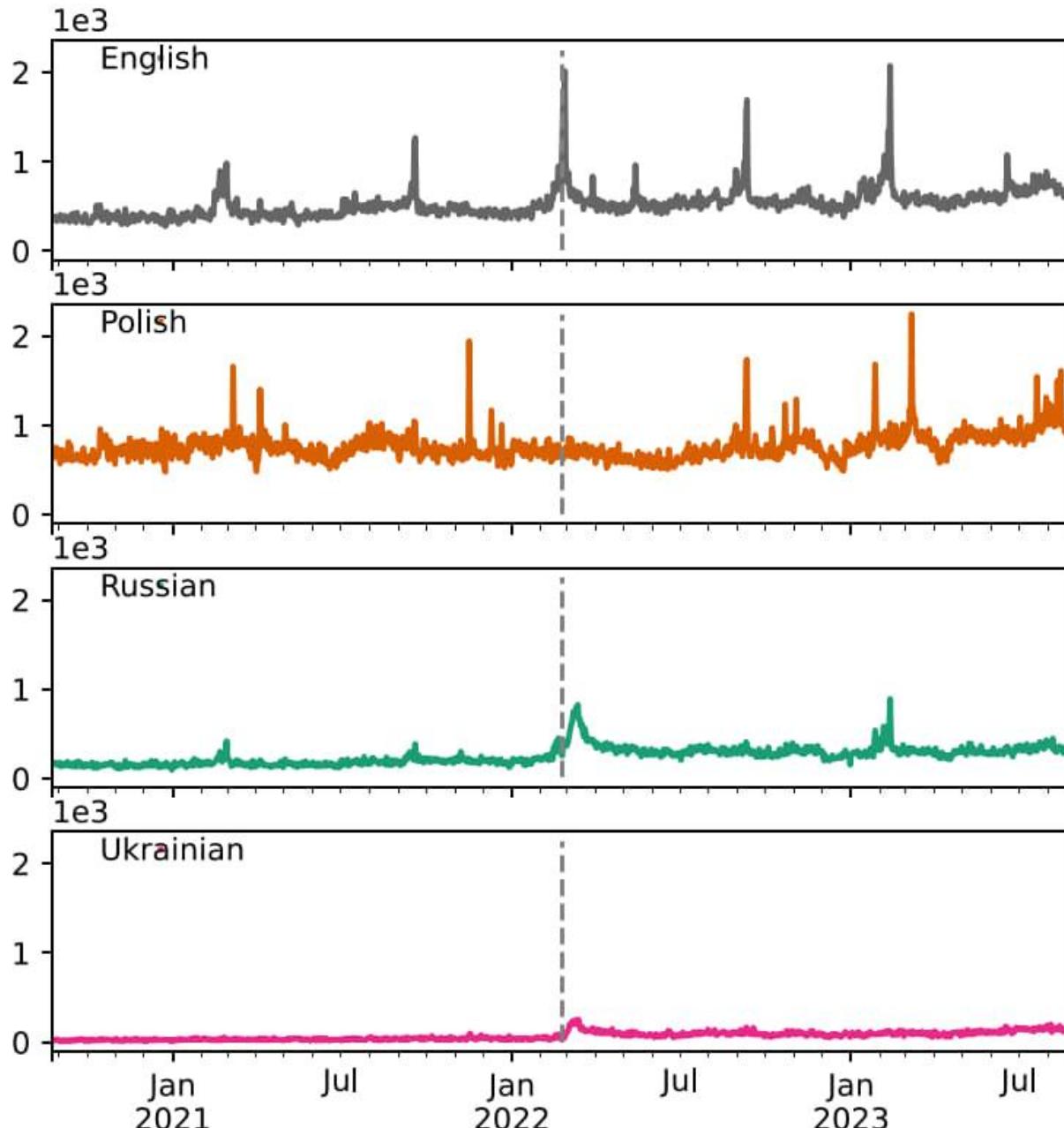
daily number of Ukrainian refugees crossing the border to Poland  
since February 24th, 2022

(weekly) number of Ukrainian refugees who have been assigned a PESEL residence registration number in Polish cities since April 2022

# Data: Wikipedia Pageviews

Daily number of views on  
Wikipedia pages about:

- European capitals and
- the 19 most populous cities in Poland



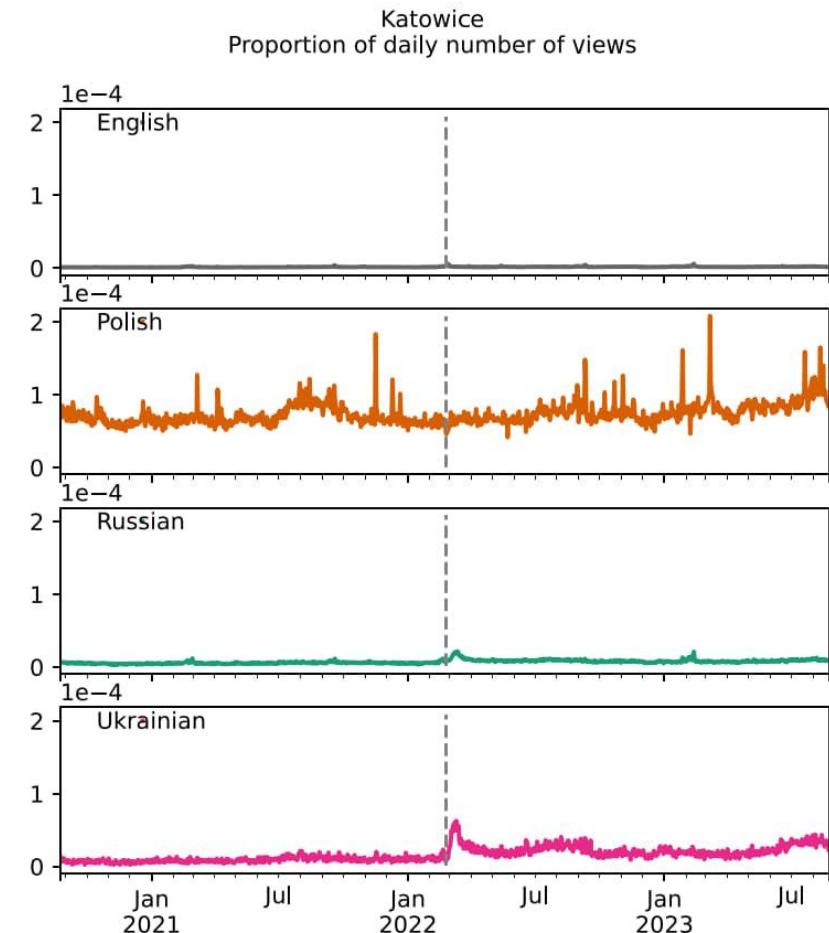
# Data: Wikipedia Pageviews

Daily number of **views** on Wikipedia pages about European capitals and the 19 most populous cities in Poland.

## Normalization process:

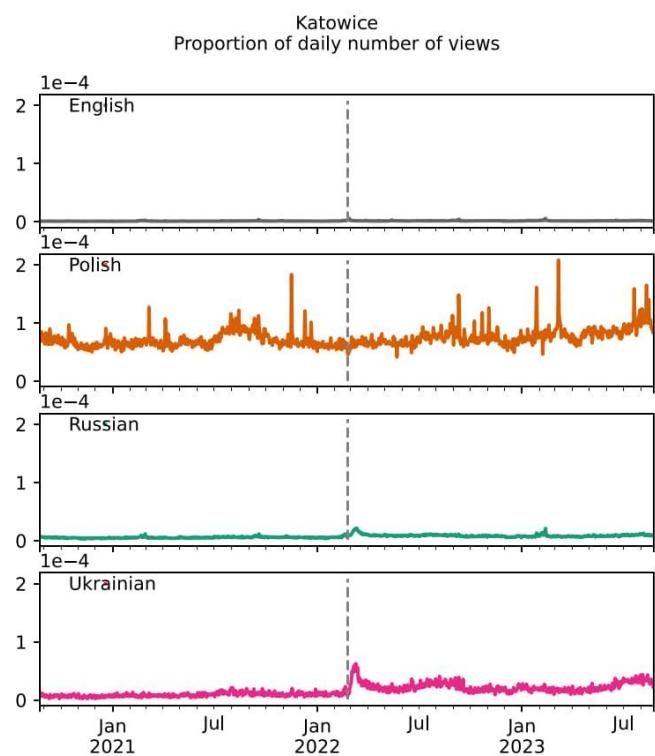
$$\frac{\text{\# daily views (page P, language L)}}{\text{\# daily views (language L)}}$$

**Proportion** of the daily number of **views** on Wikipedia pages across different **languages**.



# Relative change Wikipedia Pageviews

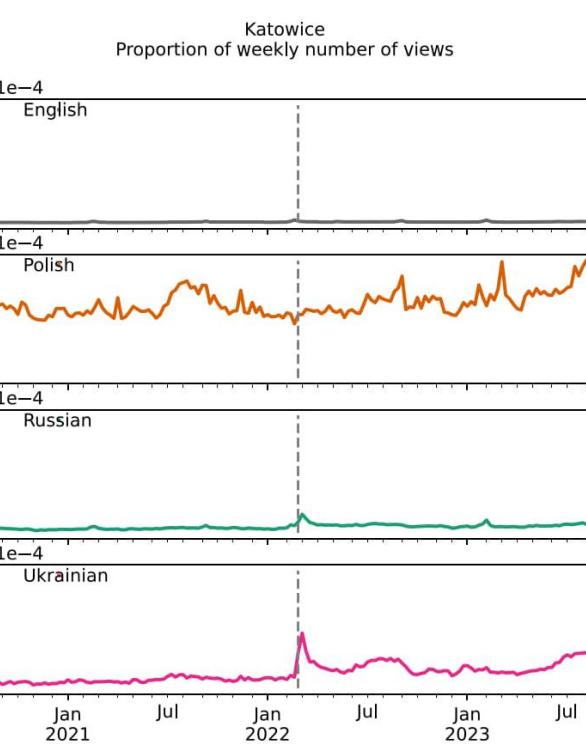
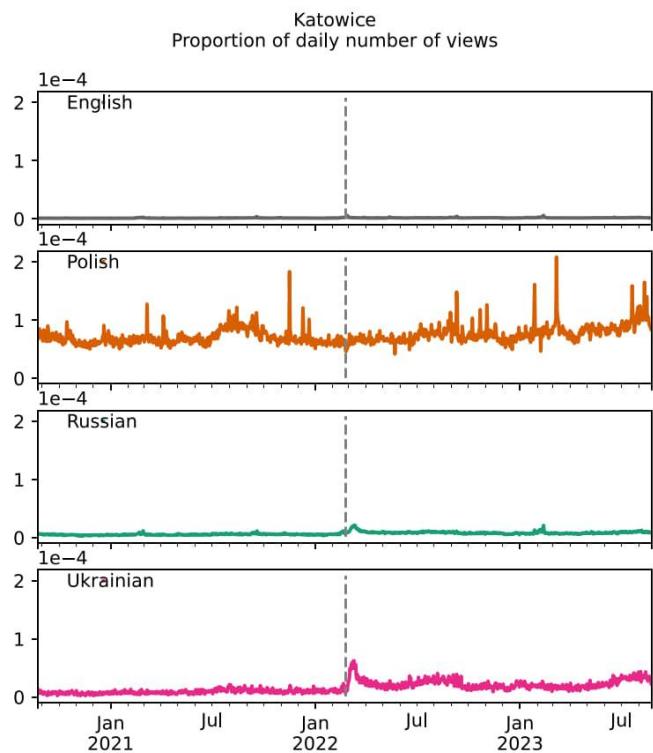
Daily data



# Relative change Wikipedia Pageviews

Daily data

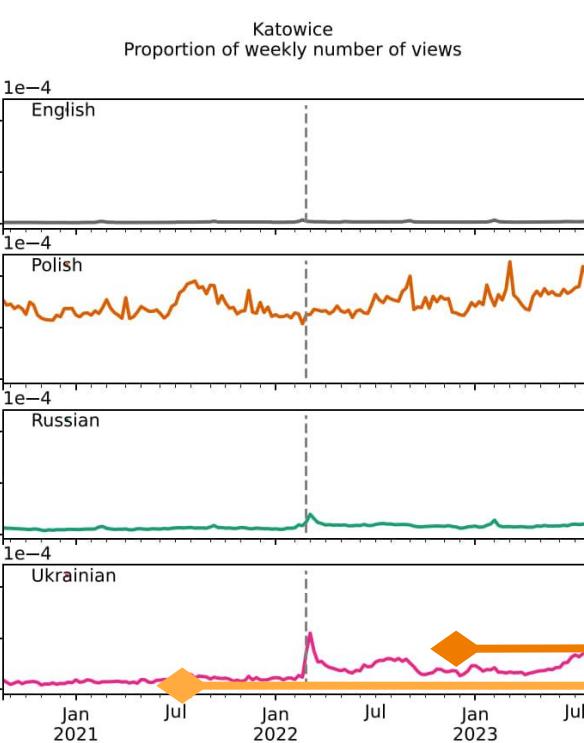
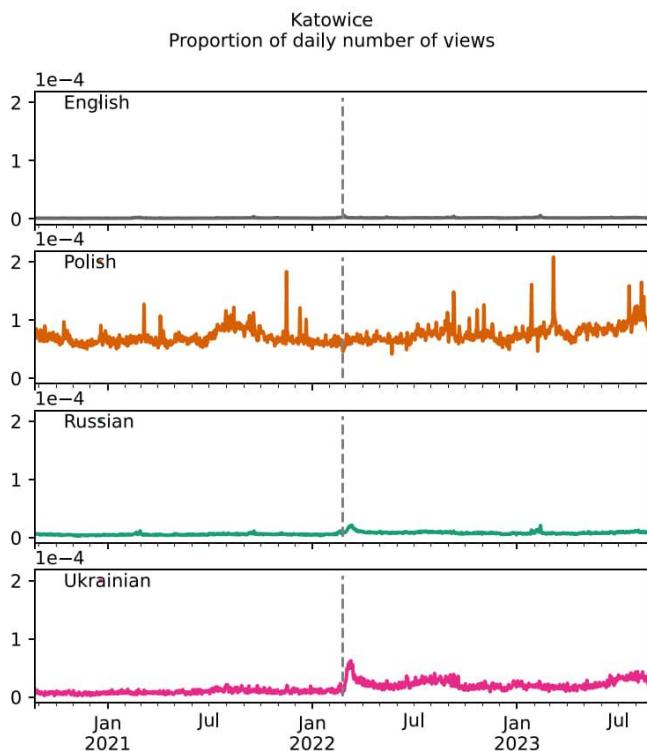
Weekly data



# Relative change Wikipedia Pageviews

Daily data

Weekly data



Increase in the number of views:  
***median post-war - median pre-war***



# Ranking European capitals according to the increase in Wikipedia Pageviews in Ukrainian

The increase in the number of views on  
**Ukrainian Wikipedia pages** dedicated  
to **European capitals** after February  
24th, 2022

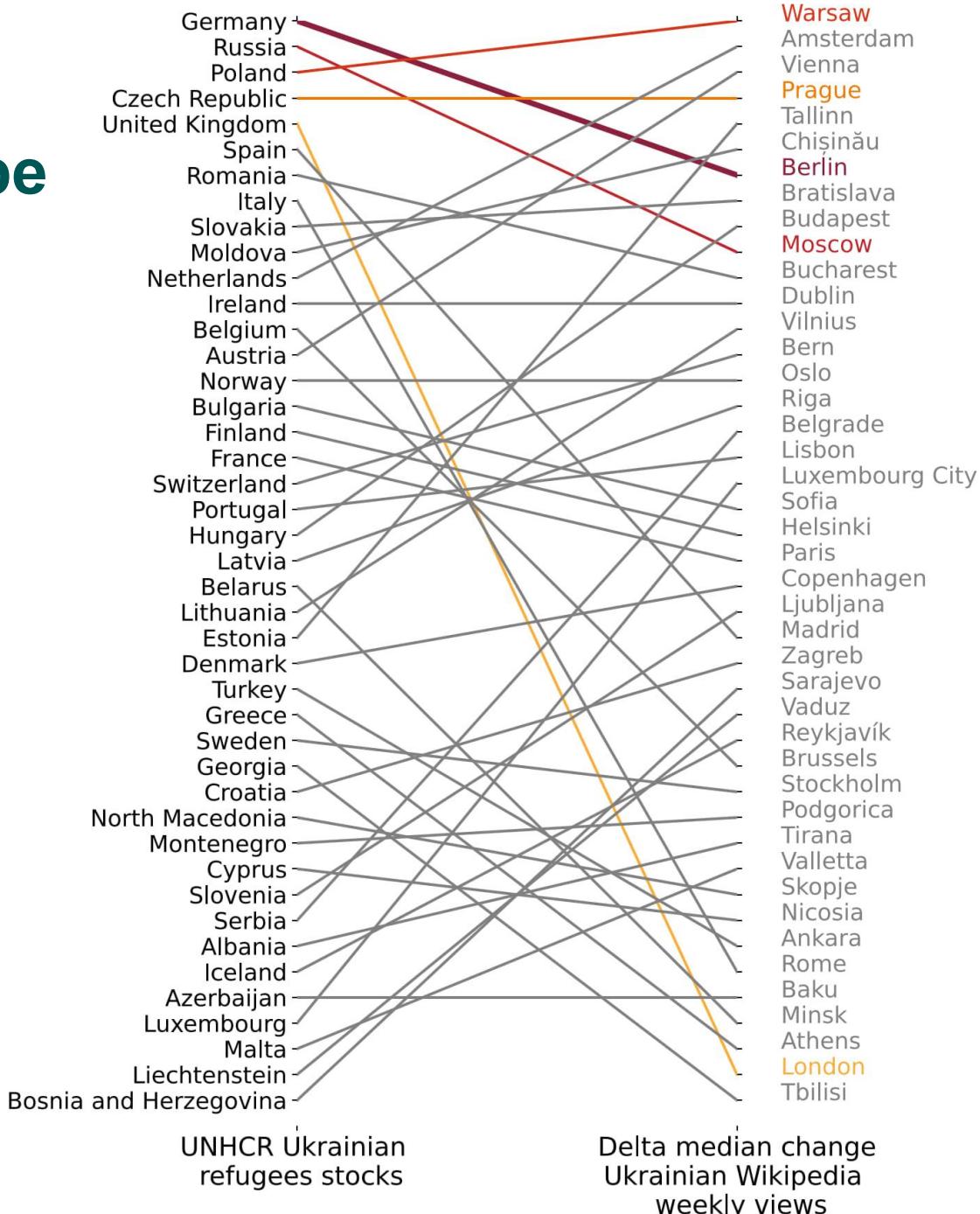


Warsaw  
Amsterdam  
Vienna  
Prague  
Tallinn  
Chisinau  
Berlin  
Bratislava  
Budapest  
Moscow  
Bucharest  
Dublin  
Vilnius  
Bern  
Oslo  
Riga  
Belgrade  
Lisbon  
Luxembourg City  
Sofia  
Helsinki  
Paris  
Copenhagen  
Ljubljana  
Madrid  
Zagreb  
Sarajevo  
Vaduz  
Reykjavík  
Brussels  
Stockholm  
Podgorica  
Tirana  
Valletta  
Skopje  
Nicosia  
Ankara  
Baku  
Minsk  
Rome  
Athens  
London  
Tbilisi

# Findings I: Rank correlation Europe

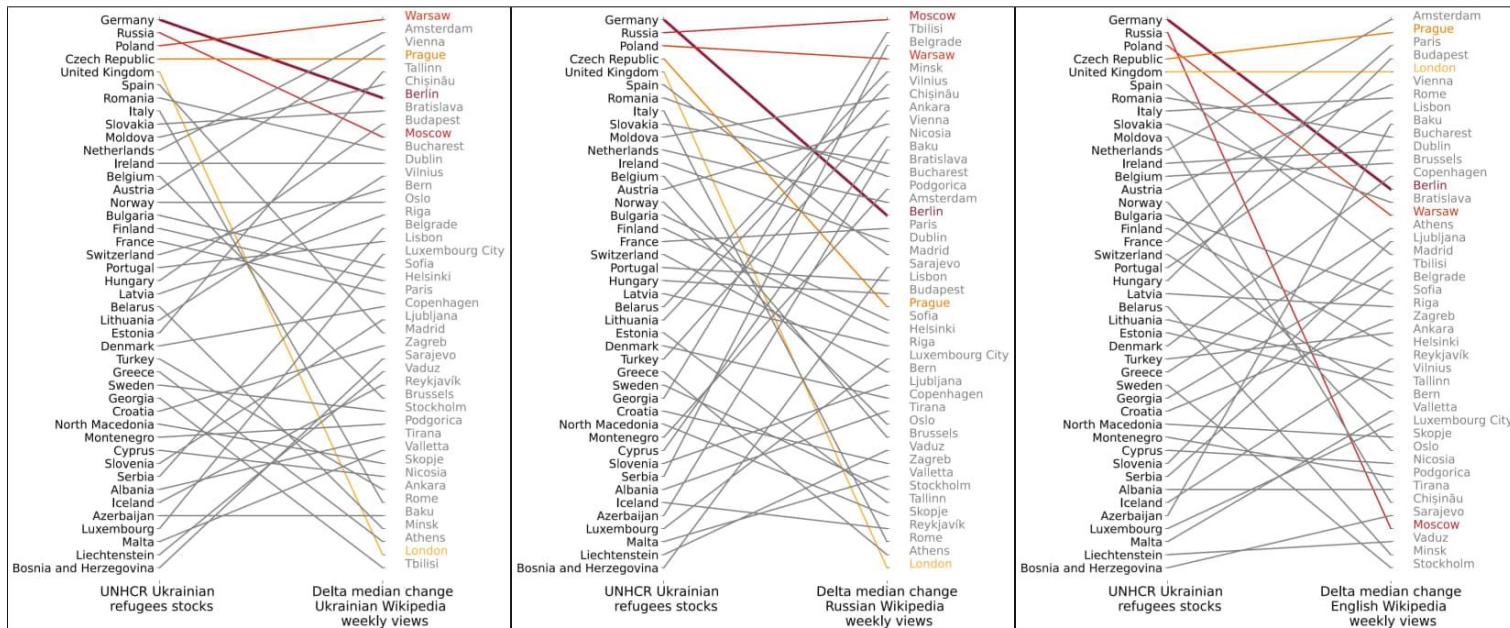
The increase in the number of views on **Ukrainian Wikipedia** pages dedicated to **European capitals** after February 24th, 2022 is highly correlated with the current **number of Ukrainian refugees in European countries**.

Spearman's rank correlation: 0.52



# Findings I: Rank correlation Europe

**Spearman's rank correlation** on the increase in the number of views on **Wikipedia** pages dedicated to **European capitals** after February 24th, 2022 and the **number of Ukrainian refugees** in European countries.



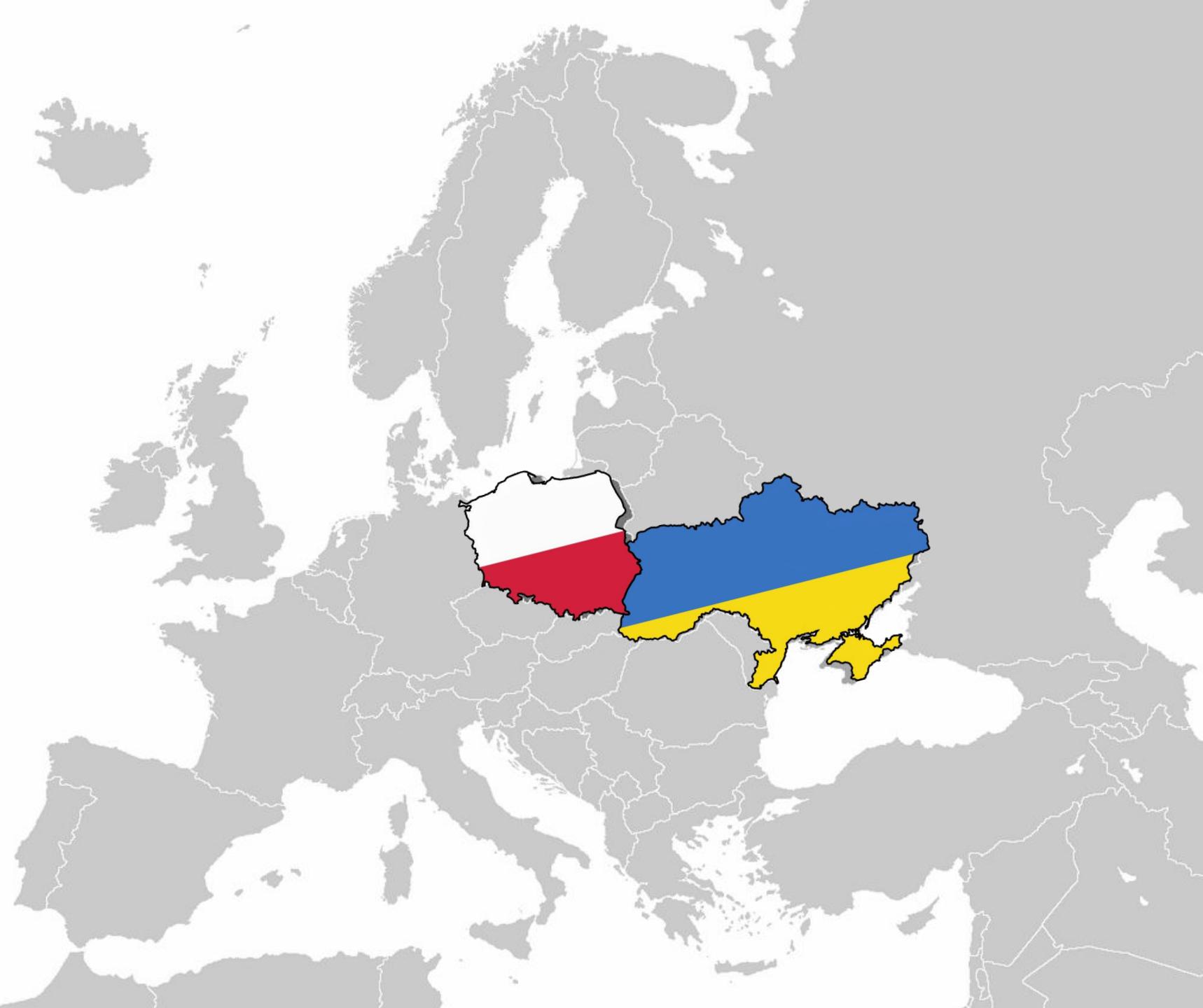
Ukrainian Wikipedia: 0.52

Russian Wikipedia: 0.22

English Wikipedia: 0.47

# From Europe to Poland

Warsaw  
Kraków  
Wrocław  
Gdańsk  
Rzeszów  
Katowice  
Lublin  
Poznań  
Łódź  
Szczecin  
Bydgoszcz  
Częstochowa  
Gdynia  
Toruń  
Białystok  
Kielce  
Gliwice  
Radom  
Sosnowiec



# Ranking Polish cities according to the increase in Wikipedia Pageviews in Ukrainian

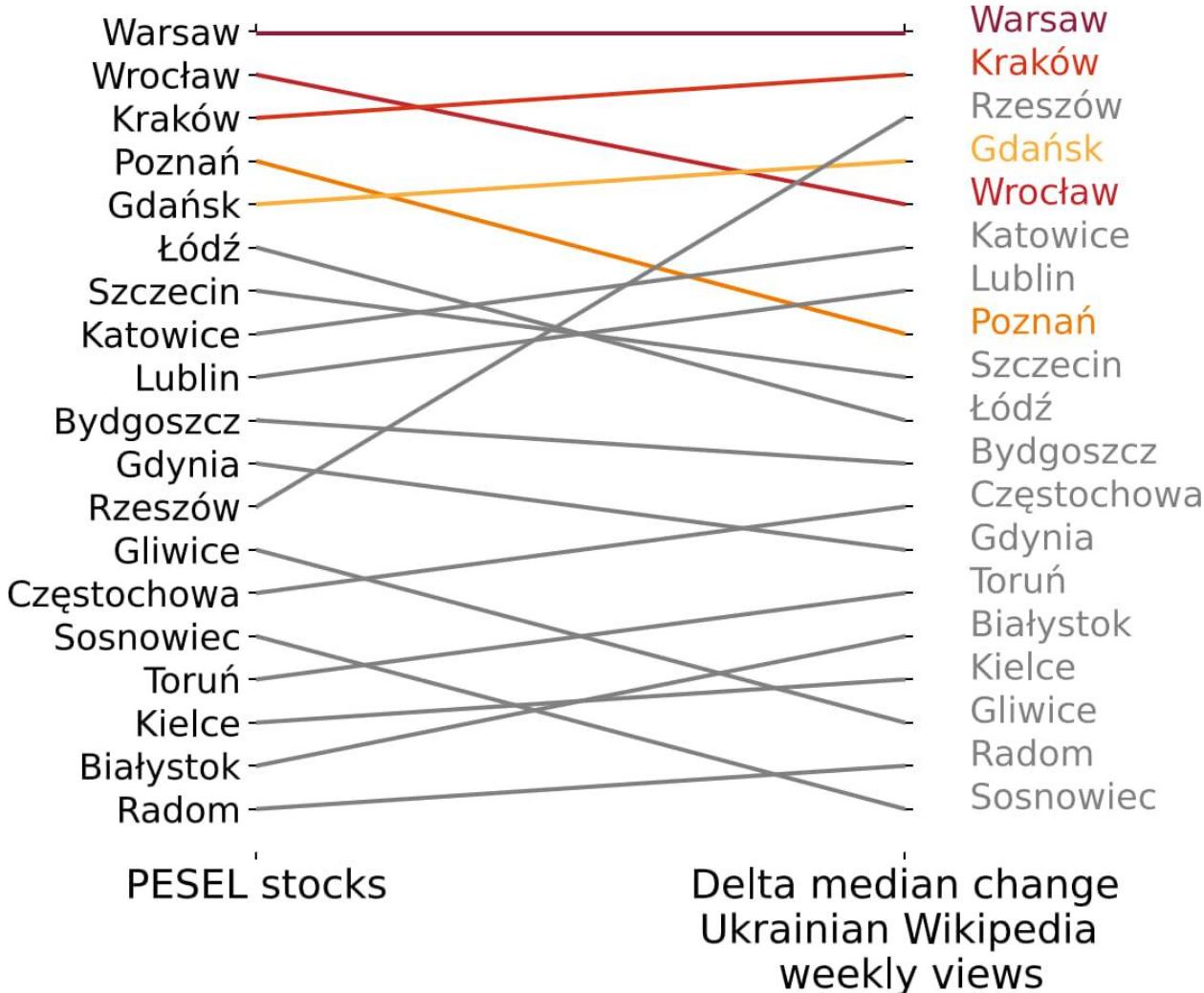
The increase in the number of views on  
**Ukrainian Wikipedia pages** dedicated  
to the 19 most populous **cities in**  
**Poland** after February 24th, 2022

Warsaw  
Kraków  
Wrocław  
Gdańsk  
Rzeszów  
Katowice  
Lublin  
Poznań  
Łódź  
Szczecin  
Bydgoszcz  
Częstochowa  
Gdynia  
Toruń  
Białystok  
Kielce  
Gliwice  
Radom  
Sosnowiec

# Findings I: Rank correlation Poland PESEL

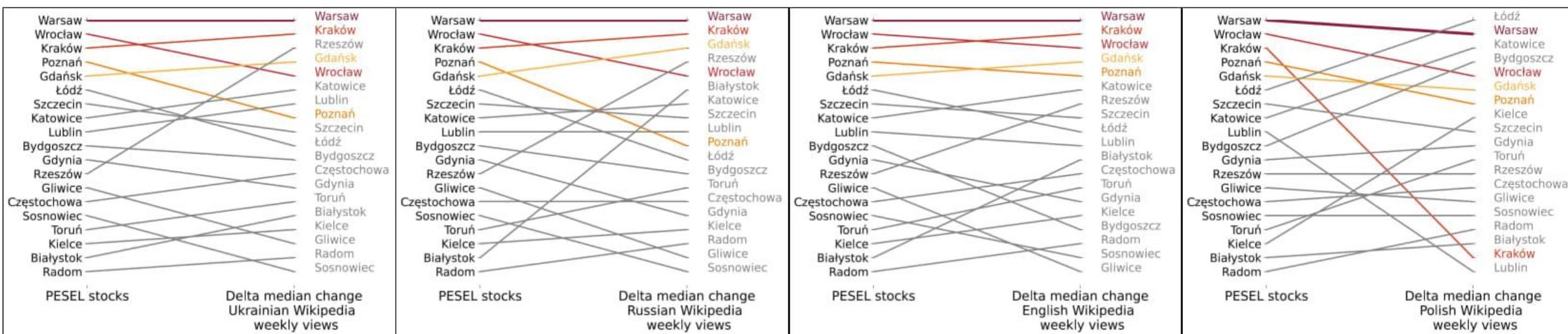
The increase in the number of views on **Ukrainian Wikipedia pages** dedicated to the 19 most populous **cities in Poland** after February 24th, 2022 is **correlated** with the total number of **Ukrainian refugees** who have been assigned a **PESEL residence registration number** in **Polish cities**

Spearman's rank correlation: 0.83



# Findings I: Rank correlation Poland PESEL

**Spearman's rank correlation** on the increase in the number of views on **Wikipedia** pages dedicated to **the 19 most populous cities in Poland** after February 24th, 2022 and the **number of Ukrainian refugees who have been assigned a PESEL number** in Polish cities



Ukrainian Wikipedia: 0.83

Russian Wikipedia: 0.68

English Wikipedia: 0.82

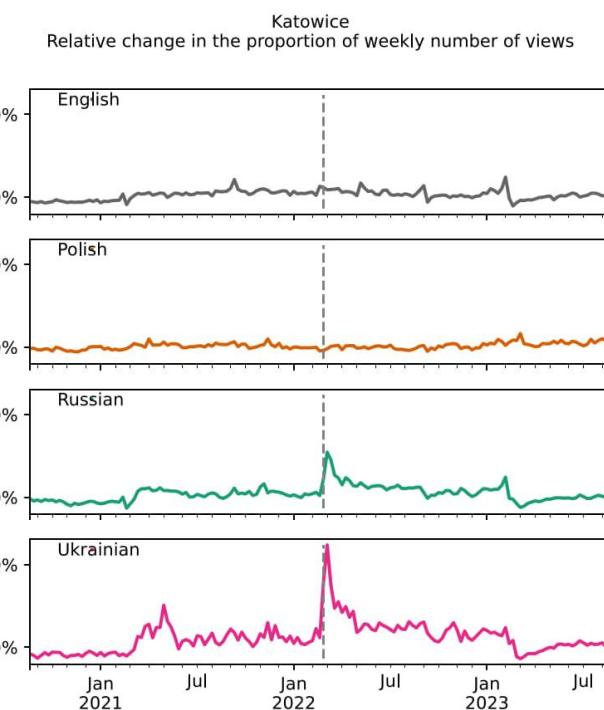
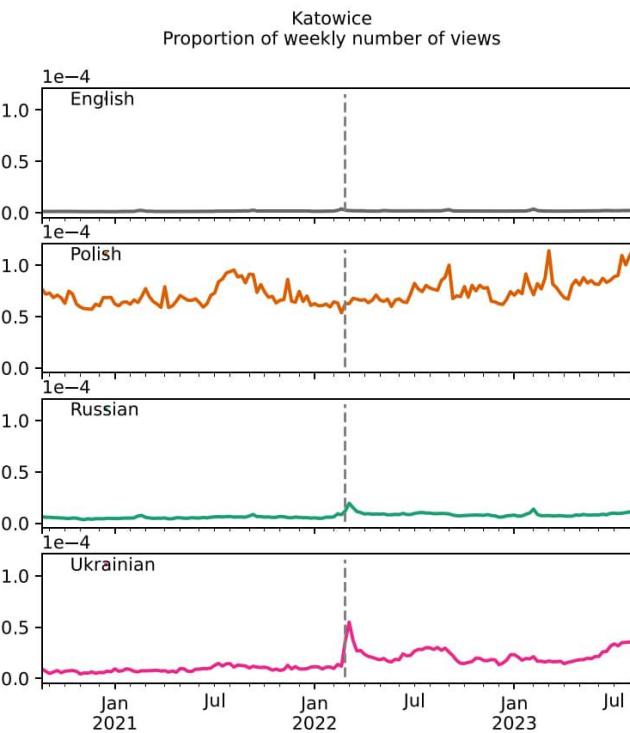
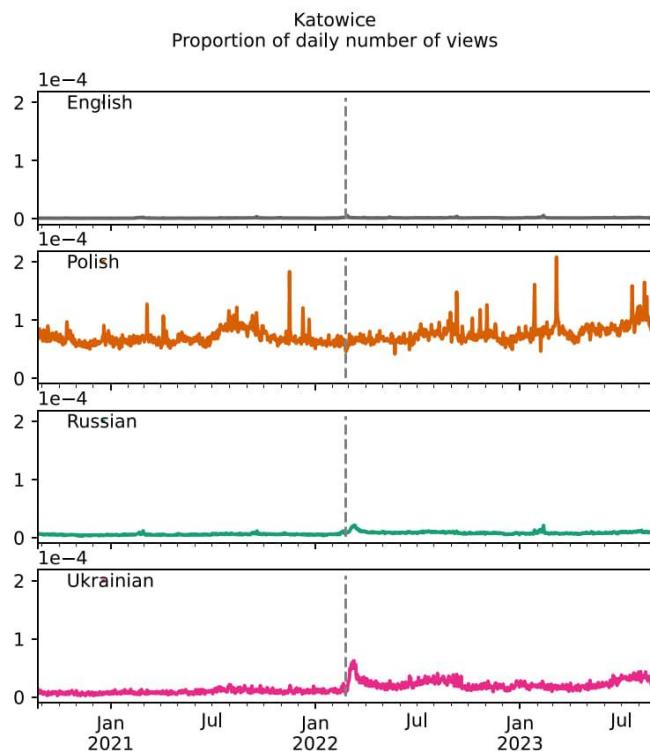
Polish Wikipedia: 0.51

# Relative change Wikipedia Pageviews

Daily data

Weekly data

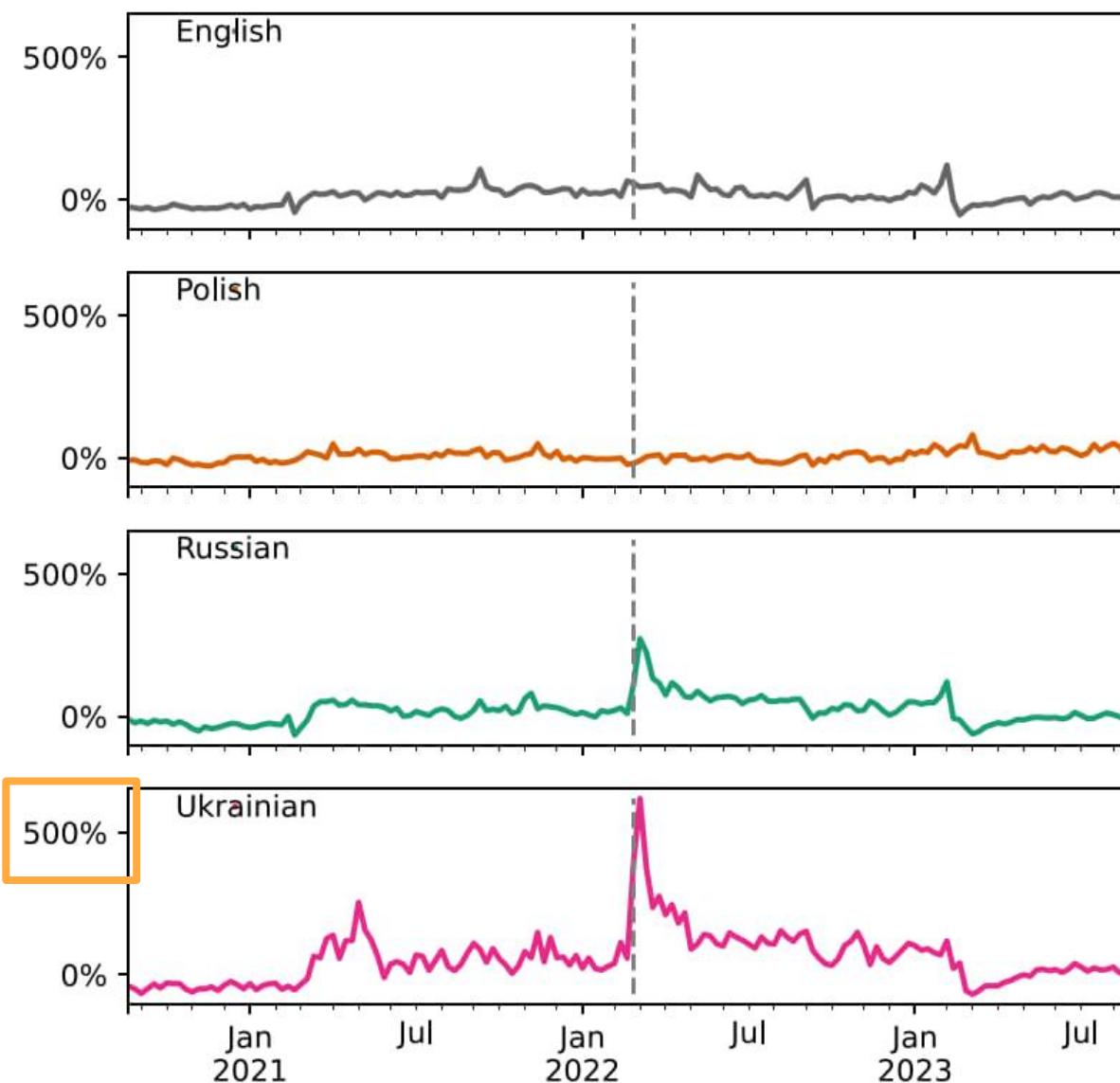
Relative change  
\*to the same week one year before



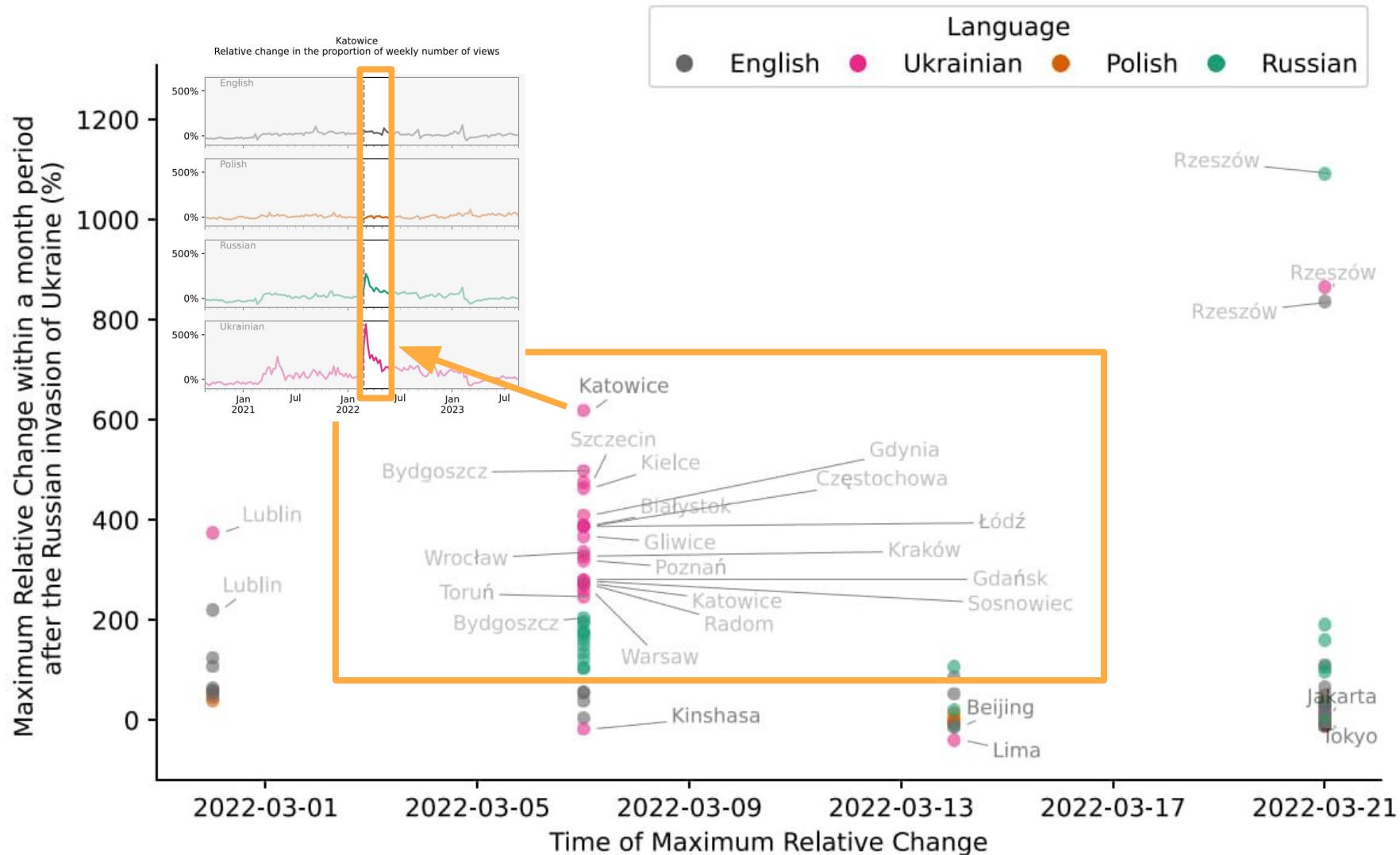
## Findings II: Relative change

The Wikipedia page about 18 of the 19 most populous cities in Poland experienced an **increase of at least 200%** in weekly number of views after the Russian invasion of Ukraine, compared to the baseline period from February 24th to August 24th, 2020.

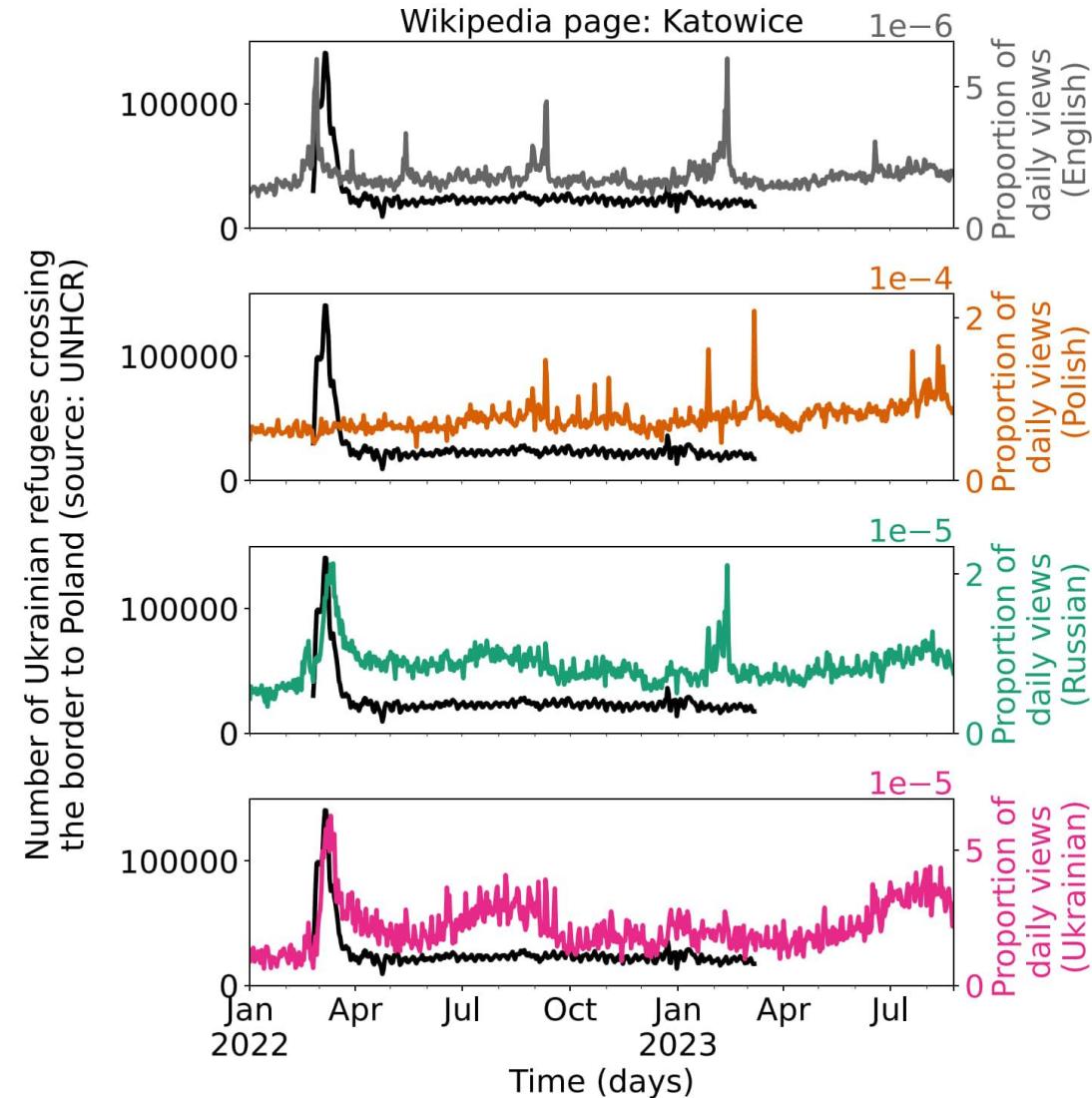
Katowice  
Relative change in the proportion of weekly number of views



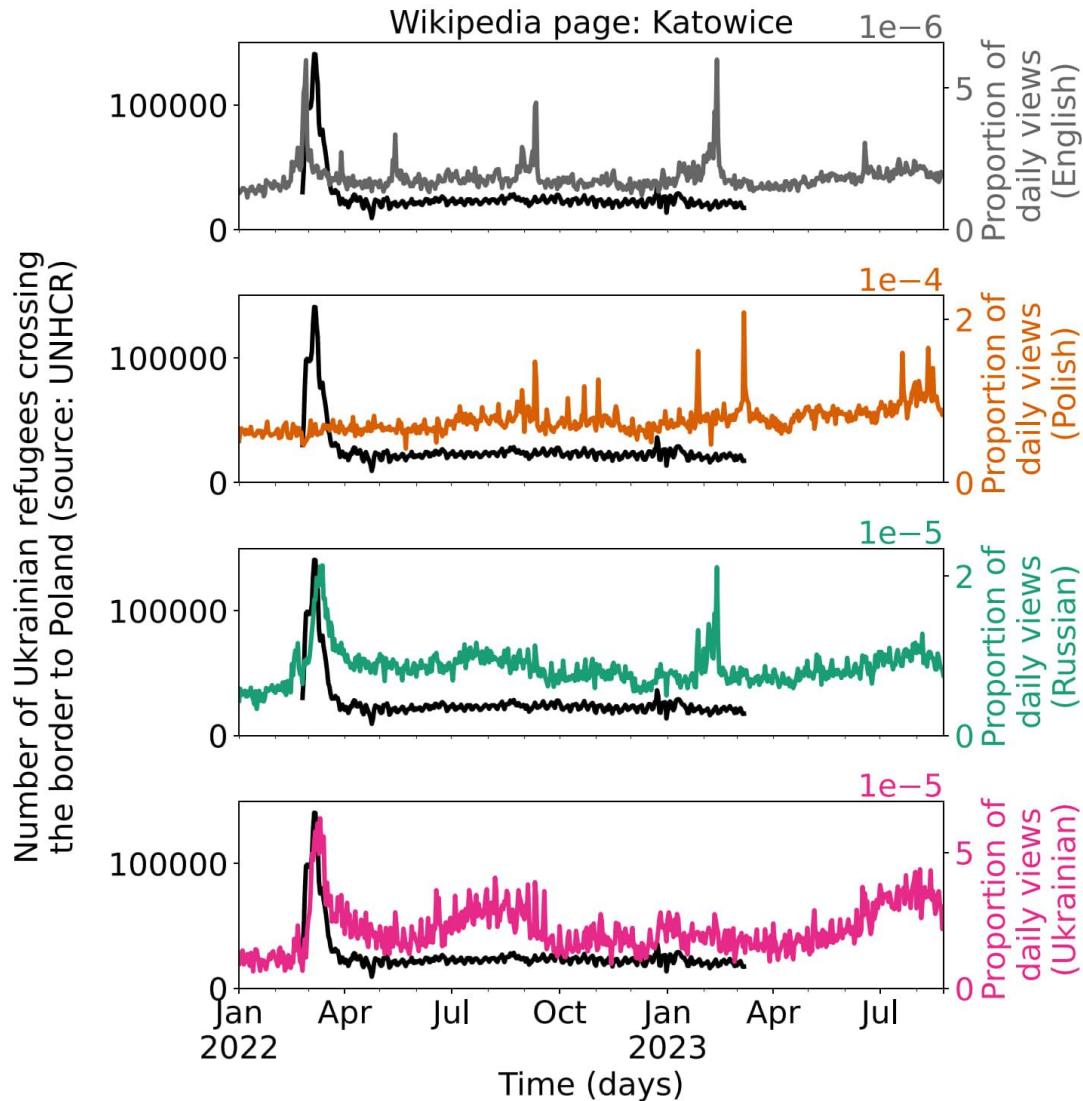
## Findings II: Relative change



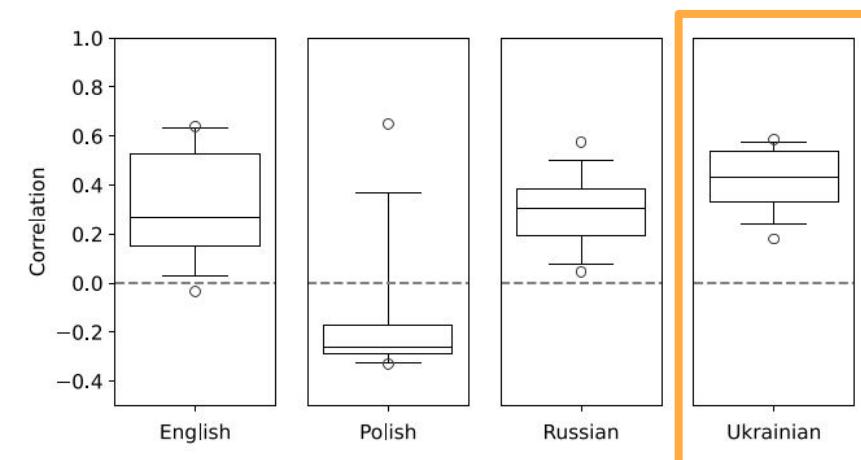
# Wikipedia Pageviews and UNHCR border crossing



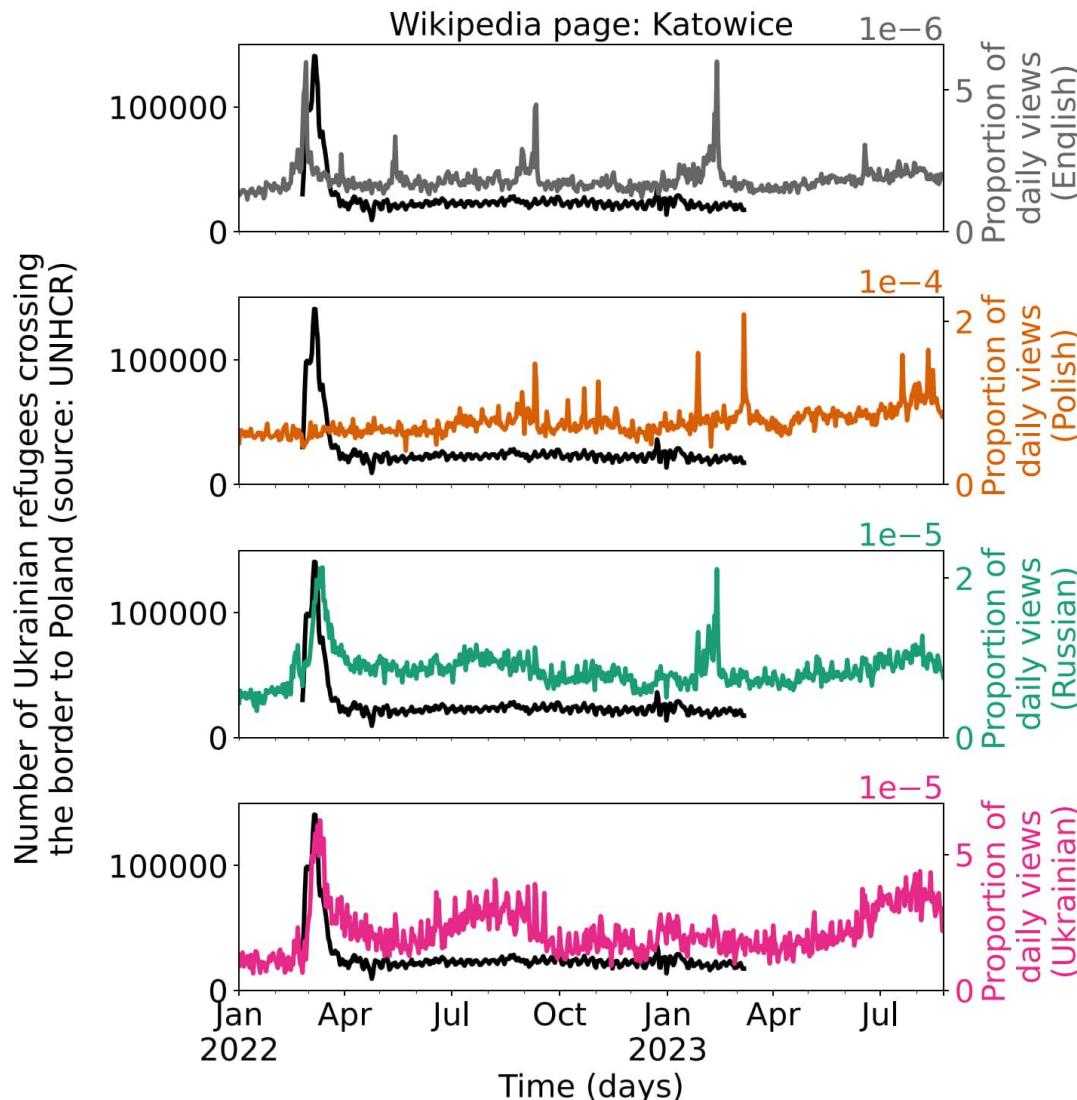
## Findings III: Correlation



The **number of views on Ukrainian Wikipedia pages** dedicated to the 19 most populous cities in Poland is **always positively correlated** with the **number of Ukrainian refugees who crossed the border to Poland**



## Findings IV: Granger causality



Statistical hypothesis test for determining whether one time series is useful in forecasting another.

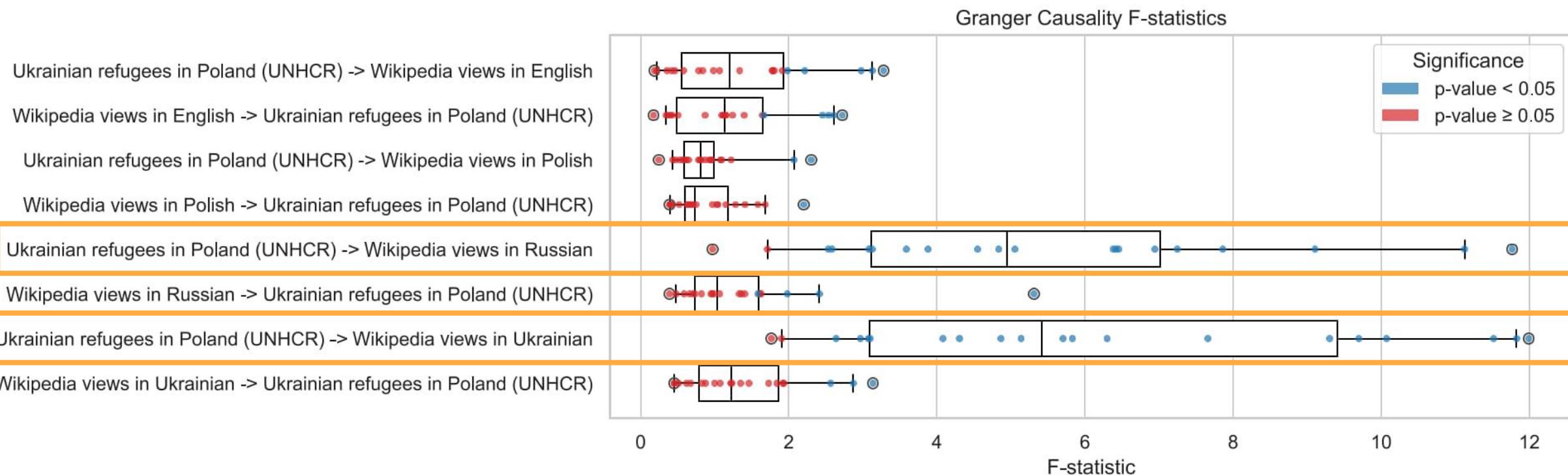
**Number of Ukrainian refugees who crossed the border to Poland**

|  
*(optimal lag of 8 days)*  
↓

**Number of views on Ukrainian Wikipedia pages dedicated to the 19 most populous cities in Poland.**

## Findings IV: Granger causality

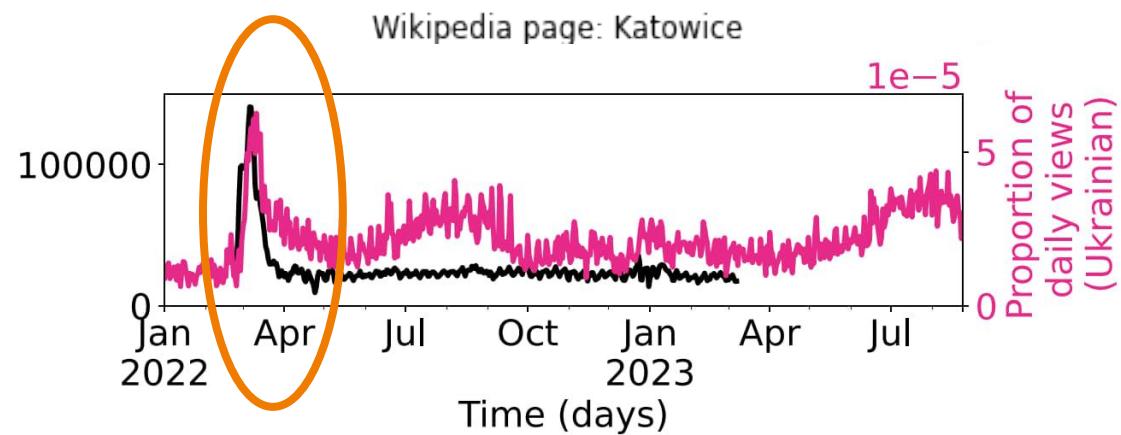
Relationship



## Findings IV: Granger causality

There is a “delay” in the increase of the **number of views on Ukrainian Wikipedia pages** dedicated to the 19 most populous cities in Poland compared to the increase in the **number of Ukrainian refugees who crossed the border to Poland**

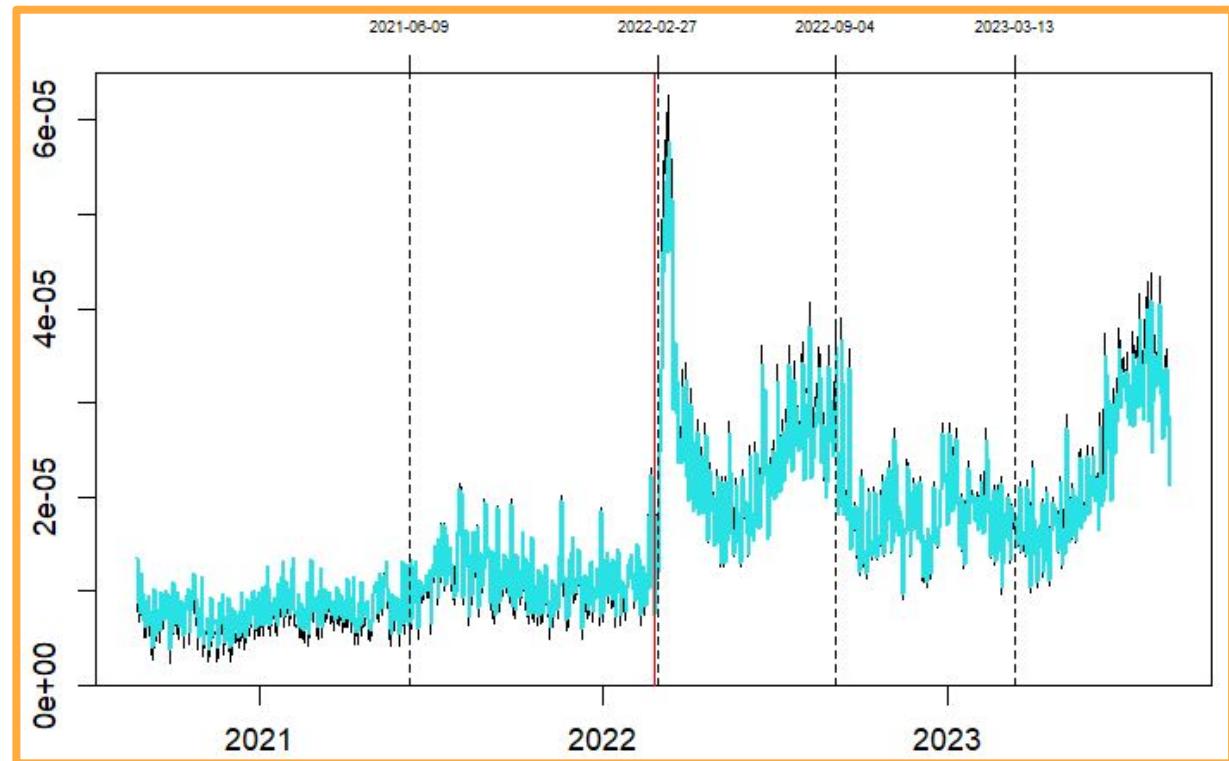
Our hypothesis is that people **crossed the border first, then searched for the city to move to...**



## Findings V: Structural breaks

Of the 19 most populous cities in Poland, 12 show a **break within one week** following the start of the war.

**Sensitivity check:** Tokyo, Seoul, Mexico City and Jakarta. We do not observe break points around the time of the Russian invasion of Ukraine.



### Break points for Katowice:

2021-06-09

**2022-02-27** (3 days after the Russian invasion of Ukraine)

2022-09-04

2023-03-13

## Conclusion

The **Wikipedia** pages views **reflect real-world migration patterns**.

Regarding the temporal dynamics, we identify a lag of approximately 8 days between the onset of mass migration to Poland and the peak in Wikipedia views, confirming that spikes in **information-seeking behavior tend to follow refugee arrivals**.

Our results reveal opportunities to use **Wikipedia views** as a tool for monitoring population movements, understanding the **decision-making processes** and **space distribution** of refugees. Hopefully improving humanitarian responses in a situation of crises.

## Limitations and Considerations

The use of **language** as a proxy for the country of origin.

**Generalizability** of this methodology to other contexts.

Internet penetration rate and **biases** in the use of Wikipedia.

**Privacy** and **ethical** implications of using digital traces to track refugees or vulnerable groups who may be in precarious situations.

# DISCUSSION





## Discussion

- Can you think of any cases for potential research that would benefit from using Google Trends and/or Wikipedia data?
- How can we ensure that digital trace data are used in a responsible and ethical manner?
- In what ways can digital trace data sources help us identify and respond to emerging migration challenges, such as forced displacement due to conflict or environmental disasters?
- How can we improve data on online search patterns for research purposes?

# HANDS-ON





## Hands-on

**Google Trends:**

<https://trends.google.com/trends/>

**Wikipedia Pageviews:**

<https://pageviews.wmcloud.org/>



<https://github.com/carolcoimbra/workshop-exploring-migration>

- ◦
- ◦
- ◦
- ◦
- ◦
- ◦
- ◦
- ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦
- ◦ ◦ ◦

# Exploring Migration Patterns Using Digital Trace Data



**Carolina Coimbra Vieira**  
[@carol\\_coimbra2](https://twitter.com/carlo_coimbra2)   
[coimbravieira@demogr.mpg.de](mailto:coimbravieira@demogr.mpg.de) 

**Ebru Şanlıtürk**  
[@EbruSanliturk](https://twitter.com/EbruSanliturk)   
[sanlituerk@demogr.mpg.de](mailto:sanlituerk@demogr.mpg.de) 



# Exploring Migration Patterns Using Digital Trace Data

**THANK YOU!**



<https://github.com/carolcoimbra/workshop-exploring-migration>



# APPENDIX



# USING GTRENDSR TO RETRIEVE DATA



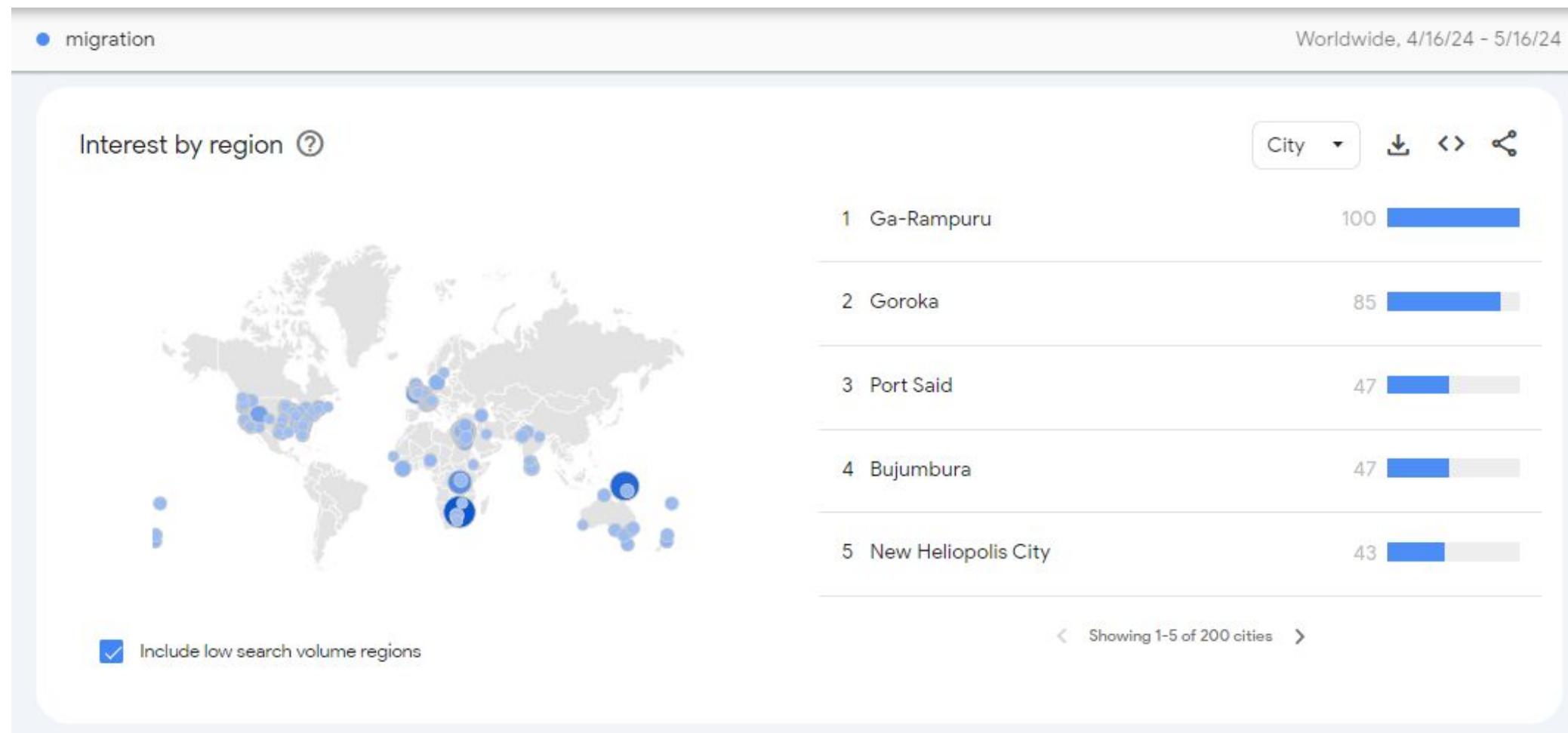


## HOW TO MAKE A QUERY

- Determine the keyword(s)
- <https://trends.google.com/> can be helpful in the brainstorming phase
- See which keywords produce a meaningful result
- Check visualizations



# GOOGLE TRENDS VISUALIZATION





## HINTS

- Consider alternative spellings for search queries
- Consider the use of accented characters
- Google may aggregate results for a query with and without accented characters for local language
  - Same filter / aggregation may not apply in other locations, make a few trials
- Consider regional dialects that may apply
- Be careful when using phrases as search queries
- Consider the justification of query selection



## GTRENSR

- `gtrends(keyword = "", geo = "", time = "", gprop = "", hl = "", low_search_volume = TRUE, compared_breakdown = FALSE)`
- time – default is last 5 years
  - “now 7-d” (last seven days), “today 1-m” (past 30 days), “today 3-m” (past 90 days), “today 12-m” (past 12 months), “Y-m-d Y-m-d”
- `compared_breakdown` can only be used to compare multiple keywords in a single location.



# GTRENSR OUTPUT

## □ Interest over time

- use “onlyInterest = TRUE”

## □ Interest by country (or region)

## □ Interest by dma (designated market area)

## □ Interest by city

## □ Related topics

## □ Related queries



## GTRENSR OUTPUT: KEYWORD VS. TOPIC

- Search for keyword = Columbus
- Search for topic = Columbus (city in Ohio)
- Uses Freebase ID (can be found on Wikidata)
- `gtrends(keyword = "/m/01smm", time = "2024-01-01 2024-03-31", low_search_volume = TRUE)`



## GTRENSDR – COMMON ERRORS

- Error in get\_widget(... : widget\$status\_code == 200 is not TRUE
  
- Make sure you use geo identifiers as given in the *countries* data
- Try downloading the developer version
  - `devtools::install_github("PMassicotte/gtrendsR")`
- It's possible that you have exceeded a limit with Google Trends, try dividing the sets included in your code (keywords, locations etc.)
- It's possible that you have exceeded a limit with Google Trends, try again a some time later.



## GTRENSDR – COMMON ERRORS

- `gtrends(keyword = "asylum", time = "2024-01-01 2024-03-31",  
gprop = "web", hl = "en", low_search_volume = TRUE,  
onlyInterest = TRUE)`
  
- Unless specified, geo is considered worldwide
- hl – language, important for related queries, related topics and location names



## GTRENSDR

```
keys = c("migration", "residence permit", "asylum")
time = " 2024-01-01 2024-03-31"

for (i in keys) {
  trendsoutput = gtrends(keyword=i, gprop ="web", geo="US", time = time,
  onlyInterest = TRUE, low_search_volume = FALSE)
  Sys.sleep(5)
  results [[i]] = trendsoutput$interest_over_time
}
```

## REFERENCES





## REFERENCES – GOOGLE TRENDS & MIGRATION

- Avramescu, A., & Wiśniowski, A. (2021). Now-casting Romanian migration into the United Kingdom by using Google Search engine data. *Demographic Research*, 45, 1219-1254.
- Anastasiadou, A., Volgin, A., & Leisure, D. R. (2024). War and mobility: Using Yandex web searches to characterize intentions to leave Russia after its invasion of Ukraine. *Demographic Research*, 50, 205-220.
- Böhme, M. H., Gröger, A., & Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142, 102347.
- Connor, P. (2017). *The Digital Footprint of Europe's Refugees*. Pew Research Center.
- Lin, A. Y., Cranshaw, J., & Counts, S. (2019). Forecasting US Domestic Migration Using Internet Search Queries. *Proceedings of the 2019 World Wide Web Conference (WWW'19)*, (pp. 13-17).
- Risteski, D., & Davcev, D. (2014, October). Can we use daily Internet search query data to improve predicting power of EGARCH models for financial time series volatility. In *Proceedings of the International Conference on Computer Science and Information Systems (ICSiS'2014), October 17–18, 2014, Dubai (United Arab Emirates)*.
- UN (2014). Estimating migration flows using online search data. *Global Pulse Project Series*, 4, 1-2.
- Wladyka, D. (2013, October). *The Queries to Google Search as Predictors of Migration Flows from Latin America to Spain*. University of Texas at Brownsville.



## REFERENCES – GOOGLE TRENDS & MIGRATION

- Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *IZA Discussion Paper No. 4201*. Institute for the Study of Labor (IZA).
- Billari, F., D'Amuri, F., & Marcucci, J. (2016). Forecasting births using Google. *CARMA 2016: 1st International Conference on Advanced Research Methods in Analytics*. Valencia: Editorial Universitat Politècnica de València.
- Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection – Harnessing the Web for public health surveillance. *The New England Journal of Medicine*, 360(21), 2153–2157.
- Chang, S.-S., Kwok, S. S., Cheng, Q., Yip, P. S., & Chen, Y.-Y. (2015). The association of trends in charcoal-burning suicide with Google search and newspaper reporting in Taiwan: a time series analysis. *Social Psychiatry and Psychiatric Epidemiology*, 50(9), 1451-1461.
- Choi, H., & Varian, H. 8.-9. (2009, April 10). Predicting the present with Google Trends.
- Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, 48(11), 87-92.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012-1014.
- Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of economic and social measurement*, 36(3), 119-167.



## REFERENCES – GOOGLE TRENDS & MIGRATION

- Johansson, E. (2014). *Creating daily search volume data from weekly and daily data*. Retrieved from: <http://erikjohansson.blogspot.com/2014/12/creating-daily-search-volume-data-from.html>
- McCarthy, M. J. (2010). Internet monitoring of suicide risk in the population. *Journal of Affective Disorders*, 122(3), 277-279.
- Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., & Valleron, A. J. (2009). More diseases tracked by using Google Trends. *Emerging infectious diseases*, 15(8), 1327-1328.
- Reis, B. Y., & Brownstein, J. S. (2010). Measuring the impact of health policies using Internet search patterns: the case of abortion. *BMC Public Health*, 10(1), 514.
- Ricketts, C. F., & Silva, C. G. (2017). An analysis of morbidity and mortality using Google Trends. *Journal of Human Behavior in the Social Environment*, 27(6), 559-570.
- Solano, P., Ustulin, M., Pizzorno, E., Vichi, M., Pompili, M., Serafini, G., & Amore, M. (2016). A Google-based approach for monitoring suicide risk. *Psychiatry research*, 246, 581-586.
- Song, T. M., M., Song, J., An, J. Y., Hayman, L. L., & Woo, J. M. (2014). Psychological and social factors affecting Internet searches on suicide in Korea: a big data analysis of Google search trends. *Yonsei Medical Journal*, 55(1), 254-263.
- Wilde, J., Chen, W., & Lohmann, S. (2020). COVID-19 and the future of US fertility: what can we learn from Google? (No. 13776). IZA Discussion Papers.