Wikipedia as a Tool for Tracking Mass Migration Flows: Insights from the Russian Invasion of Ukraine

Carolina Coimbra Vieira*, Ebru Sanlitürk, Emilio Zagheni

[1]Max Planck Institute for Demographic Research
coimbravieira@demogr.mpg.de

Abstract

Tracking and predicting migration flows, especially those arising from unforeseen factors like conflicts and wars, can be challenging. Digital trace data offer an innovative approach to track, predict, and assess real-time changes in migration flows faster than traditional data sources. In this study, we propose a methodology to use Wikipedia data to determine how the number of views on Wikipedia pages about cities change over time due to rising interest, in response to big migration events. As a case study, we focus on recent mass migration events caused by the Russian invasion of Ukraine in 2022. Our results show a correlation between the number of views on Ukrainian Wikipedia pages dedicated to European capitals and the number of Ukrainian refugees in European countries. Similarly, we observed a high correlation between the number of views on Ukrainian Wikipedia pages dedicated to Polish cities and the number of Ukrainian refugees in Poland. Our findings reveal opportunities in the use of Wikipedia as a proxy to study and predict mass migration flows.

Introduction

"At the end of 2022, 108.4 million people worldwide were forcibly displaced as a result of persecution, conflict, violence, human rights violations and events seriously disturbing public order".[1] Despite these large numbers, migrants in need of international protection (MNP)[2] - refugees, asylum-seekers, displaced persons, other persons in need of international protection, and stateless persons are part of a distinct migrant subgroup often not captured by traditional data (Robinson 1998).

Tracking migrants in need of international protection is crucial to ensure their safety and well-being by providing necessary assistance and protection throughout their journey and resettlement process. In addition to that, governments and aid agencies can better plan on how to allocate resources efficiently based on population movements and needs.

However, migration flows are especially challenging to track and predict when these migration events happen due to unforeseen factors like conflicts and wars, resulting in forced migration (Cesare et al. 2018; Leurs and Smets 2018; Tjaden 2021). Given the difficulty of measuring migrants in need of international protection in traditional data recording mechanisms, there is a continuous effort to expand data recording capabilities.[3] Digital trace data propose an alternative data source to track migrants faster than traditional data sources. It is also often combined with official statistics to improve the estimates of migration flows, opening up new avenues for research and policy development (Leasure et al. 2023; González-Leonardo et al. 2024; Hsiao et al. 2023).

Digital trace data from Google Trends (Böhme, Gröger, and Stöhr 2020), Facebook Ads (Spyratos et al. 2018; Alexander, Polimis, and Zagheni 2019) and LinkedIn Ads (Zhu, Fritzler, and Orlowski 2018; Perrotta et al. 2022; Vieira et al. 2022) have been shown as good data sources to track and predict migration. However, most of the social media platforms often operate as black boxes, complicating interpretation and reproducibility (Lazer et al. 2014). Moreover, the trend of restricting or closing APIs by these platforms has made accessing data increasingly challenging (Davidson et al. 2023; Lazer et al. 2014). In contrast, data access to Wikipedia is free of cost and easy to obtain, which motivates our work to check whether Wikipedia can be used in the study of migration.

Wikipedia is the largest and most popular free online encyclopedia, aiming to provide equal access to information about current events and media coverage of a topic worldwide (Singer et al. 2017). Wikipedia pages are created and edited by volunteers around the world (Graham and Dittus 2022), and volunteer contributions to Wikipedia can be affected by events in the real world. For instance, the death of Queen Elizabeth II, on 8 September 2022, and the rapid reaction of Wikipedia editors to update Wikipedia pages related to the queen and the Royal family[4] is a good example of how zealous Wikipedia editors are making quick changes in response to real-world changes.

In this study, we aim to shed light on the relationship between online sources of information and migration flows. Due to its nature, forced migration increases the need to access information quickly and efficiently. The use of smart-

[1]https://www.unhcr.org/global-trends

[2]https://www.refworld.org/policy/legalguidance/unhcr/2017/en/121440

[3]https://www.unhcr.org/blogs/statistics-refugee-numbers-highest-ever/

[4]https://www.npr.org/2022/09/15/1122943829/wikipedia--queen-elizabeth-ii-death-deaditors-editors-article

phones connected to the internet is known to help forced migrants and refugees to access information they need (The GSM Association 2017; Ulutürk, Uysal, and Varol 2019). Among many online sources of information, Wikipedia is a well-known free online source of information. By examining the association between the number of views on Wikipedia by language and recent refugee flows, we aim to determine whether (i) mass migration events affect the number of views on Wikipedia pages and (ii) the number of views on Wikipedia pages can be used to track and as a predictor for mass migration flows.

Our methodology consists of using Wikipedia data to assess how the number of views on Wikipedia pages dedicated to cities change across time in response to migration events. As a case study, we focus on a recent migration event – the Ukrainian refugee crisis in 2022 – caused by the Russian invasion of Ukraine. We collected the number of views in Ukrainian (i.e., the official language in Ukraine), Russian, and English on Wikipedia pages dedicated to European capitals and the most populous cities in Poland, hosting many Ukrainian refugees.

Overall, we found a strong correlation between the number of views on Ukrainian Wikipedia pages dedicated to European capitals and the number of Ukrainian refugees in European countries. We observed the highest increase in the number of views on Ukrainian Wikipedia pages dedicated to Polish cities. There is a high correlation between the number of views on Ukrainian Wikipedia pages dedicated to Polish cities and the number of Ukrainian refugees registered in Polish cities. Additionally, we observed a strong association between the number of views on Ukrainian Wikipedia dedicated to Polish cities and the number of Ukrainian refugees crossing the border from Ukraine to Poland, as reported by the United Nations High Commissioner for Refugees (UNHCR). Finally, our results reveal opportunities for the use of Wikipedia as a proxy to study and track mass migration flows. We also contribute to the literature on the study of forced migration and understanding the decision-making process as well as the space distribution of migrants after large-scale migration events.

The contribution of this study is considered twofold. First, we introduce Wikipedia data as a novel source to analyze mass migration flows, exploiting the increased need for information and condensed information-seeking patterns. Second, we determine the timing of information seeking on Wikipedia and mass migration events, thus showing how it can reflect and be used as a tool for tracking and predicting mass migration flows.

Background

Migration and online sources of information

Information communication technologies, as well as online sources of information and social media platforms, are popular tools on which migrants who seek information quickly and efficiently base their decisions on whether to migrate and the destinations where to settle (Dekker et al. 2018; Felton 2015). Studies about Ukrainian refugees reported that 92% of Ukrainian refugees in Poland have a mobile phone and 86% have consistent internet connection (Imperative 2022). Furthermore, International Migration Organization (IOM) reports show that Ukrainian refugees in Germany (International Organization for Migration 2022) and Romania (International Organization for Migration 2023) consider social media and the internet as their top sources of information. On one side, online sources of information help migrants with their decision-making processes (The GSM Association 2017; Ulutürk, Uysal, and Varol 2019; Dekker et al. 2018; Felton 2015). On the other side, the availability of online search data paved the way for a growing literature to predict and analyze migration processes (Lin, Cranshaw, and Counts 2019; Böhme, Gröger, and Stöhr 2020; Golenvaux et al. 2020; Avramescu and Wiśniowski 2021; Anastasiadou, Volgin, and Leasure 2024).

Online search engines provide a useful tool to measure interest in a topic that can be used as a proxy for intention to move in migration studies. However, there are some limitations involving the use of online search engine data. One limitation is the availability of various search engines and the difficulty of comparing online search data from different sources. While Google is the most popular search engine worldwide, Bing and Yandex also appear as competitors in different regions. Each search engine reports online search interest in its own way, using different parameters and algorithms. These algorithms may create a bias themselves, for instance, Google Trends provides only a normalized index for the given place and time and applies an unobservable threshold that prevents results for very low interest. The lack of access to logs containing absolute numbers of searches by Google Trends constitutes a significant shortcoming for detailed comparative research using the most popular search engine in the world (see Appendix for further comparison between Google Trends and Wikipedia data).

In contrast, Wikipedia data regarding the number of views on pages are easily accessible through the API or via download on the website.[5] Previous work has shown the high correlation between frequently searched keywords and Wikipedia page views, suggesting that Wikipedia page views can a source to determining popular global web search trends (Yoshida et al. 2015). Additionally, Wikipedia is an interesting complementary data source, especially when traditional surveillance systems are not available in real time (Vilain et al. 2017).

Studies using Wikipedia

Wikipedia is the most popular free online encyclopedia maintained by volunteer experts in some domains (West, Weber, and Castillo 2012; Agarwal et al. 2020) around the world (Graham and Dittus 2022; Panciera, Halfaker, and Terveen 2009). Wikipedia readers can benefit from free access to information about current events and media coverage of a topic (Singer et al. 2017). For instance, a large percentage of Wikipedia users read about entertainment-related topics (e.g., TV series, movies, and biographies), history, health, and tech content (Lehmann et al. 2014).

[5] https://pageviews.wmcloud.org/langviews/

English articles are the most notable considering the number of collaborative interactions to create and edit them. However, Wikipedia is available in 288 other active languages (Bipat, McDonald, and Zachry 2018). Although many articles are available in many languages, some concepts are not represented or shared across languages (Miquel-Ribé and Laniado 2020, 2019). Other topics, such as some historical figures, appear in multiple languages depending on interactions between cultures (Eom et al. 2015). The number of edits and the complexity also vary according to the language. Overall, editors are less likely to edit complex topics in a second language, except in English since the level of complexity of English edits is the same regardless of the primary language of the editor (Kim et al. 2016). Based on the fact that some specific languages will be edited more often by editors who spoke the language as a primary language, our study uses the number of edits in one language as a proxy for the number of users from a region where people speak the language living in the city or country mentioned on the Wikipedia page.

The number of contributions on Wikipedia can also be affected by events in the real world. For instance, during COVID-19 mobility restrictions, the number of contributions to Wikipedia increased (Ruprechter et al. 2021), especially in languages associated with countries where the most severe mobility restrictions were implemented (Ribeiro et al. 2020). Twyman, Keegan, and Shaw (2017) showed how movements such as *Black Lives Matter* are also quickly reported and updated on Wikipedia. As a result of this process, Wikipedia works as a repository of knowledge about social movements as they unfold. In politics, Agarwal et al. (2020) showed that both edits and readership across Wikipedia pages dedicated to politicians in the UK are affected during election times. In finances, Moat et al. (2013) showed that the number of views on Wikipedia pages sign moves on the stock market. Finally, in public health, Vilain et al. (2017) used Wikipedia data to monitor the trends of seasonal diseases in France. In this sense, we believe that Wikipedia can be considered an interesting and important complementary data source, especially when traditional surveillance systems are not available in real time.

Finally, related to migration, Lucchini, Tonelli, and Lepri (2019) collected data on Wikipedia regarding historically notable individuals' movements to study important features of migration's behavior. However, the work is limited to content analysis and a specific set of notable individuals. Lerner and Lomi (2019) shows the network effects on the number of edits on migration-related topics on Wikipedia. Although the topic is migration, the study focuses on network effects on edits, while in our study we investigate the usefulness of Wikipedia to track and predict migration flows.To the best of our knowledge, this is the first study to use Wikipedia page views to track mass migration flows.

Data and Methods

The Ukrainian refugee crisis started on February 24th, 2022, after Russia's invasion of Ukraine. Ukrainian refugees primarily sought refuge in neighboring countries. If they sought refuge in a European Union (EU) member country, they could also move to other European countries due to the free movement policy within the EU. We compiled a list of European capitals by referring to the Wikipedia page on capitals in Europe.[6] In addition to that, given the high number of Ukrainian refugees in Poland, we also zoomed in our analysis into Poland.

Poland is one of the key host countries for Ukrainian refugees due to its proximity to Ukraine. The Polish government has implemented various measures to accommodate and support the displaced population. We specifically targeted the nineteen most populous cities in Poland.[7] The dataset encompassed information gathered from various sources, providing a comprehensive understanding of Ukrainian refugee flows across the continent.

Official statistics

European data We collected data from the United Nations High Commissioner for Refugees (UNHCR) to contrast it with the data collected on Wikipedia. We use this data to assess the level of association between the number of Ukrainian refugees across cities in Europe and the number of views in Ukrainian on Wikipedia pages dedicated to European cities. We leveraged the data from UNHCR[8] about the total number of refugees in European countries.

Polish data We collected data on the number of Ukrainian refugees crossing the border to Poland from UNHCR.[9] This dataset contains the daily number of Ukrainian refugees crossing the border from Ukraine to Poland from February 24th, 2022 to January 10th, 2023.

In addition to that, we also collected data from Poland's Data Portal (DANE)[10] regarding the number of Ukrainian refugees registered for temporary protection in Polish cities from April 2022 to August 2023. Each refugee registered for temporary protection in Poland receives a PESEL (*Powszechny Elektroniczny System Ewidencji Ludności*, in English: Universal Electronic System for Registration of the Population) number.

Wikipedia page views

We collected the data on the number of views on Wikipedia pages using the Wikimedia pageviews platform.[11] The platform provides daily counts for the number of views (i.e., every time the page is loaded) on each Wikipedia page across different languages since July 2015. We only consider user views, which include editors, anonymous editors, and readers. Views from search engine "web crawlers" or automated programs are not included.

[6] https://en.wikipedia.org/wiki/Category:Capitals_in_Europe

[7] Białystok, Bydgoszcz, Częstochowa, Gdańsk, Gdynia, Gliwice, Katowice, Kielce, Kraków, Łódź, Lublin, Poznań, Radom, Rzeszów, Sosnowiec, Szczecin, Toruń, Warszawa, Wrocław

[8] https://data.unhcr.org/es/situations/ukraine/location/10781. Accessed on October 12th, 2023.

[9] https://data.unhcr.org/es/situations/ukraine/location/10781

[10] https://dane.gov.pl/en

[11] https://pageviews.wmcloud.org/langviews/

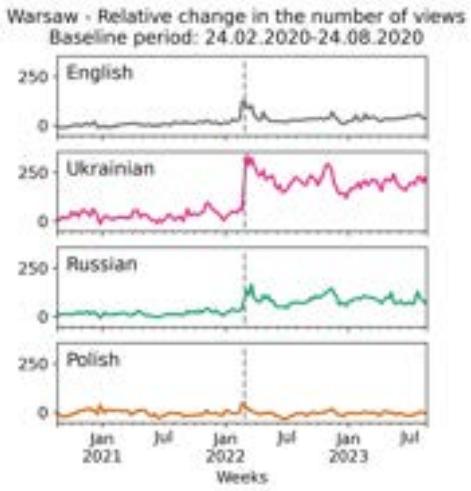


Figure 1: Relative change in the number of views on the Wikipedia page about Warsaw across different languages.

Wikipedia daily views We collected the daily number of views for the Wikipedia pages dedicated to each one of the European capitals and Polish cities across different languages since July 2015. We focused the analysis across some languages such as English, Ukrainian, Russian, and the official languages in each of the cities analyzed.

Wikipedia weekly views To reduce noise and provide a more stable representation of trends over time, we aggregated the daily data into weekly data. The aggregation was performed separately for each city. We focus our analysis on the period from August 24th, 2020, to August 24th, 2023, encompassing 18 months before and after the start of the war on February 24th, 2022.

After calculating the weekly data, we calculated the relative change in the number of views on the Wikipedia pages across different languages. We considered the average number of views during the period of February 24th, 2020 to August 24th, 2020 as a baseline. Figure 1 shows the relative change in the number of views on the Wikipedia page about Warsaw in English, Ukrainian, Russian, and Polish. The Ukrainian Wikipedia shows a high increase in the relative change right after the war started on February 24th, 2022. Figure 5 (in Appendix) shows the relative change in the number of views on the Wikipedia page dedicated to each one of the nineteen Polish cities in English, Ukrainian, Russian, and Polish.

Additionally, we computed the absolute increase in the number of views for each Wikipedia in each one of the languages English, Ukrainian, Russian, and Polish (for Polish cities only). The absolute increase in the number of views is given by the difference between the median weekly views on a Wikipedia page before and after the war started on February 24th, 2022.

Ethical considerations

We only collect publicly available data through the UNHRC website, Poland's Data Portal (DANE), and Wikimedia pageviews platform, following ethical guidelines (Rivers and Lewis 2014). Our study uses only aggregated data regarding the number of refugees in European countries and in some Polish cities, and views on Wikipedia pages. Regarding the Wikipedia data, for privacy reasons, the geographic location of readers on a per-page basis is not available. Finally, we do not attempt to identify users nor link any personal information to any particular user.

Case Study I: Ukrainian refugees in Europe

According to the United Nations High Commissioner for Refugees (UNHCR)[12], until April 2024, more than 5.9 million refugees from Ukraine have been recorded across Europe. The top countries used to cross the borders from Ukraine are Russia, Poland, Moldova, Romania, Slovakia, Hungary, and Belarus. In this first case study, we focus on the relationship between the number of Ukrainian refugees across all European countries and the increase in the number of views on Ukrainian Wikipedia pages dedicated to European capitals.

We created a ranking sorted by the European countries with more Ukrainian refugees based on the data collected from UNHCR about the total number of refugees in European countries. Similarly, we created a ranking sorting European capitals based on the increase in the number of views on their Wikipedia pages (see Figure 2). We calculated the Spearman's rank correlation between the ranking of Ukrainian refugees in Europe with a ranking of increase in the Wikipedia number of views. We observed that the increase in the number of views on Ukrainian Wikipedia pages dedicated to European capitals after February 24th, 2022 is highly correlated (Spearman's rank correlation = 0.7) with the current number of Ukrainian refugees in European countries.[13] The correlation is much lower between the current number of Ukrainian refugees in European countries and the increase in the number of views on Wikipedia dedicated to European capitals in Russian (0.28) or English (0.49).

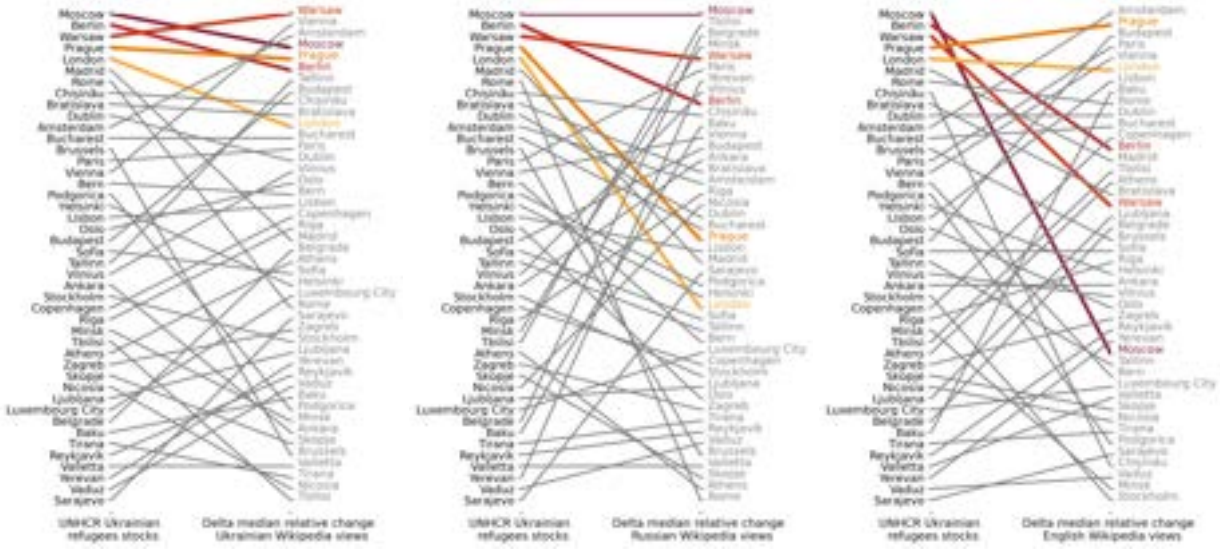
Case Study II: Ukrainian refugees in Poland

Since Poland is one of the top countries receiving Ukrainian refugees, our next set of analyses focuses on the number of views in some Wikipedia pages dedicated to Polish cities. First, we investigated the relationship between the number of Ukrainian refugees who have been assigned a PESEL number in Polish cities with the number of views on Wikipedia pages dedicated to Polish cities. Second, we compared the daily number of Ukrainian refugees crossing the border to Poland with the daily number of views on Wikipedia pages dedicated to Polish cities.

We created a ranking sorting the Polish cities by the total number of Ukrainian refugees who have been assigned a PESEL number based on the data collected from Poland's Data Portal (DANE). We also created a ranking sorting Polish cities according to the increase in the number of

[12] https://reporting.unhcr.org/operational/situations/ukraine-situation

[13] Data from October, 2023.



(a) *Ukrainian Wikipedia* (b) *Russian Wikipedia* (c) *English Wikipedia*

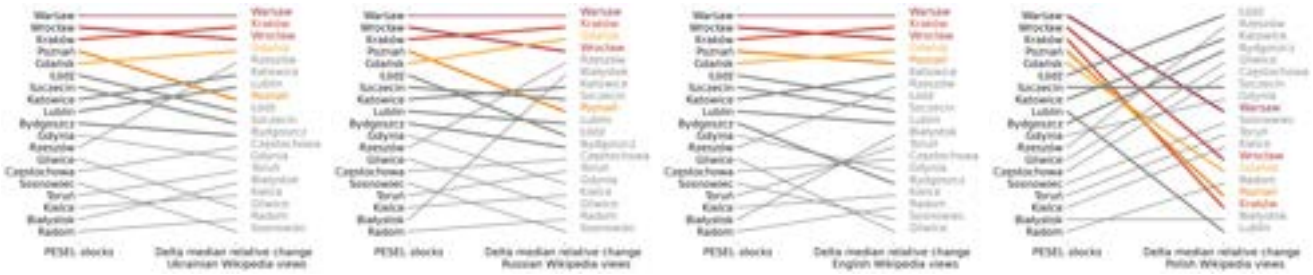
Figure 2: Correlation between the increase in the number of views on Wikipedia pages dedicated to European capitals after February 24th, 2022 and the current number of Ukrainian refugees in European countries. The top 5 countries with more Ukrainian refugees are highlighted in colors.

views on their Wikipedia pages (see Figure 3). We calculated the Spearman's rank correlation between the ranking of Ukrainian refugees who have been assigned a PESEL number with a ranking of increase in the Wikipedia number of views. We observed that the increase in the number of views on Ukrainian Wikipedia pages dedicated to Polish cities after February 24th, 2022 is highly correlated (Spearman's rank correlation = 0.87) with the number of Ukrainian refugees who have been assigned a PESEL number in Polish cities from April 2022 to August 2023. The correlation is also high considering the Russian (0.73) and English (0.83) Wikipedia and close to zero for the Polish (0.02) Wikipedia.

As a validation step, we correlate the number of views on Ukrainian Wikipedia dedicated to Polish cities with official statistics provided by the UNHCR regarding the number of refugees crossing the border from Ukraine to Poland since February 24th, 2022. Figure 6 (in Appendix) shows the daily number of views on Wikipedia pages dedicated to Polish cities in four different languages: English, Polish, Russian, and Ukrainian. For most of the cities, the daily number of views tends to be stable at the beginning of the times series for the Polish language. Similarly, the daily number of views on the English Wikipedia dedicated to Polish cities seems to be stable, with some non-regular picks over the years. The daily number of views in Russian, and especially in Ukrainian, increased dramatically in 2022, the most intense year in the war between Ukraine and Russia so far. Figure 7 (in Appendix) shows the correlation between the number of Ukrainian refugees crossing the border from Ukraine to Poland (since February 24th, 2022) and the number of views Wikipedia pages dedicated Polish cities across different languages: English, Polish, Russian and Ukrainian. In most of the Wikipedia pages dedicated to Polish cities, the strongest correlation occurs between the number of views in the Ukrainian language and the number of refugees in Poland. Figure 4 shows that 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.

Finally, we used Granger causality tests to determine if the daily number of views on Wikipedia pages dedicated to Polish cities predicts the number of Ukrainian refugees crossing the border to Poland. Granger causality is a statistical concept used to determine if one time series can help predict another time series. This test involves regressing the dependent variable (e.g., daily number of Ukrainian refugees crossing the border to Poland) on both its own lagged values and the lagged values of the independent variable (e.g., daily number of views on Wikipedia pages dedicated to Polish cities). The null hypothesis in this test is that the independent variable does not Granger-cause the dependent variable. If the p-value associated with this test is less than a chosen significance level (e.g., 0.05), we reject the null hypothesis and conclude that there is evidence of Granger causality. In this context, if the daily number of views on Wikipedia pages dedicated to Polish cities Granger-causes the number of Ukrainian refugees crossing the border to Poland, it implies that searching for information on Wikipedia might be an indicator of migration patterns or decisions to cross the border to Poland.

Figure 8 (in Appendix) shows the matrix of Granger causality tests. Each cell contains the f-score and p-value associated with the test if the variable in the column Granger-causes the variable in the row. Overall, we observed a bidirectional causality between the two variables, meaning that



(a) *Ukrainian Wikipedia* (b) *Russian Wikipedia* (c) *English Wikipedia* (d) *Polish Wikipedia*

Figure 3: Correlation between the increase in the number of views on Wikipedia pages dedicated to the 19 most populous cities in Poland in different languages after February 24th, 2022, and the number of Ukrainian refugees who have been assigned a PESEL number in Polish cities. The top 5 cities with more Ukrainian refugees who have been assigned a PESEL number are highlighted in colors.

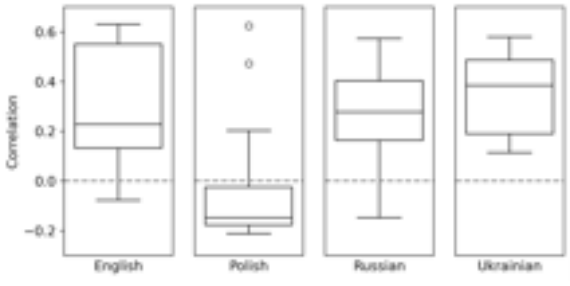


Figure 4: Correlation between the number of views on Wikipedia pages dedicated to the 19 most populous cities in Poland in different languages and the number of Ukrainian refugees who crossed the border to Poland.

the daily number of views on Wikipedia pages dedicated to Polish cities Granger-causes the number of Ukrainian refugees crossing the border to Poland and vice-versa. However, the f-scores are higher in the direction of the number of Ukrainian refugees crossing the border to Poland Granger-causing the daily number of views on Wikipedia pages dedicated to Polish cities. There is a delay in the increase of the number of views on Ukrainian Wikipedia pages dedicated to the nineteen most populous cities in Poland compared to the increase in the number of Ukrainian refugees who crossed the border to Poland. This result suggests that Ukrainian refugees and their relatives searched for more information about cities in Poland right after they crossed the border to Poland.

Discussion and Conclusion

There are many challenges involved in tracking and predicting migration flows, especially when these migration events are related to unexpected causes, such as conflicts, wars, and disasters. For instance, traditional data sources, such as surveys, require a lot of time, effort, logistic issues in including the target group on the move as well as high costs. An alternative to tracking, predicting, and assessing real-world changes faster than traditional data sources is via digital trace data.

In this study, we propose the use of Wikipedia as a new data source to study the online response to events in the real world, especially related to migration. In our case study, we focused on the high number of Ukrainian refugees in Europe due to the Russian invasion of Ukraine in 2022. We observed that the number of views on Ukrainian Wikipedia pages dedicated to European capitals and cities in Poland is in alignment with the number of Ukrainian refugees in Europe and Polish cities.

Our findings underscore the opportunities for the use of Wikipedia as a proxy for studying and predicting mass migration flows and contribute to the literature on the relationship between information networks and migration networks.

Implications To the best of our knowledge, our work is a first attempt to use Wikipedia views data to monitor migration events. We propose a methodology timely, cost-effective, reproducible, and scalable using Wikipedia as a new data source to monitor migration events in real-time. There are a couple of implications in this work.

First, from a political point of view, decision-makers would benefit from real-time estimates regarding big migration events. This implication itself is extremely important and could benefit especially countries where the political situation is unstable.

Second, a big question in demography or social sciences, in general, relates to predicting *When/Where is going to happen/be the next crisis or the next big wave of migrants?*. Our methodology could help to understand the space distribution of migrants after large-scale migration events. The methodology could be easily adapted to other contexts, and the data from Wikimedia itself could be incorporated as a new data source for social science projects in general.

Finally, our methodology repurposed the use of Wikipedia data to study an important real-world phenomenon. Given the real-world impacts of this project, Wikipedia also benefits from it.

Limitations and Future Work Besides all the positive implications, our work has also a couple of limitations. The most important one is the use of language as a proxy for the country of origin. However, it is important to notice that

some languages are broadly spoken in more than one country and could potentially bias the results. For instance, Russian is spoken in, at least, four countries as an official language. This limitation is not a big concern in the case of the Ukrainian language, since Ukrainian is concentrated around Ukraine.

In future work, the causal dimension of the relationship demonstrated in our study could be better investigated, and the data collection expanded to cover more cities affected differently by the Ukrainian refugee flows. Additionally, we would like to assess the impact of edits on Wikipedia pages on refugees seeking information about the place they are planning to move to. Connectivity to and networks at the destination are known to be a big factor in migration decisions. The dedication, size of the editors' community, and quality of the information provided by Wikipedia editors may work as a pull factor for certain destinations during the decision-making processes of refugees. Finally, Wikipedia data can be used to assess different stages and aspects of general migration processes: (i) views to certain pages can indicate the number of users who are likely to move to a city or country (pre-migration), (ii) edits reveal the size of the community in the city or country (migrant stocks), and (iii) edits in a specific language followed by edits in other languages for that same page reveals the community connection between those languages (migrant integration).

References

Agarwal, P.; Redi, M.; Sastry, N.; Wood, E.; and Blick, A. 2020. Wikipedia and Westminster: Quality and Dynamics of Wikipedia Pages about UK Politicians. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, 161–166. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-7098-1.

Alexander, M.; Polimis, K.; and Zagheni, E. 2019. The impact of Hurricane Maria on out-migration from Puerto Rico: Evidence from Facebook data. *Population and Development Review*, 617–630.

Anastasiadou, A.; Volgin, A.; and Leasure, D. R. 2024. War and mobility: Using Yandex web searches to characterize intentions to leave Russia after its invasion of Ukraine. *Demographic Research*, 50.

Avramescu, A.; and Wiśniowski, A. 2021. Now-casting Romanian migration into the United Kingdom by using Google Search engine data. *Demographic Research*, 45: 1219–1254.

Bipat, T.; McDonald, D. W.; and Zachry, M. 2018. Do We All Talk Before We Type? Understanding Collaboration in Wikipedia Language Editions. In *Proceedings of the 14th International Symposium on Open Collaboration*, OpenSym '18, 1–11. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5936-8.

Böhme, M. H.; Gröger, A.; and Stöhr, T. 2020. Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, 142: 102347.

Cesare, N.; Lee, H.; McCormick, T.; Spiro, E.; and Zagheni, E. 2018. Promises and pitfalls of using digital traces for demographic research. *Demography*, 55(5): 1979–1999.

Davidson, B. I.; Wischerath, D.; Racek, D.; Parry, D. A.; Godwin, E.; Hinds, J.; van der Linden, D.; Roscoe, J. F.; Ayravainen, L.; and Cork, A. G. 2023. Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, 7(12): 2054–2057.

Dekker, R.; Engbersen, G.; Klaver, J.; and Vonk, H. 2018. Smart refugees: How Syrian asylum migrants use social media information in migration decision-making. *Social Media+ Society*, 4(1): 2056305118764439.

Eom, Y.-H.; Aragón, P.; Laniado, D.; Kaltenbrunner, A.; Vigna, S.; and Shepelyansky, D. L. 2015. Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions. *PLOS ONE*, 10(3): e0114825. Publisher: Public Library of Science.

Felton, E. 2015. Migrants, refugees, and mobility: How useful are information communication technologies in the first phase of resettlement. *Journal of Technologies in Society*, 11(1): 1–13.

Golenvaux, N.; Alvarez, P. G.; Kiossou, H. S.; and Schaus, P. 2020. An LSTM approach to Forecast Migration using Google Trends. *arXiv preprint arXiv:2005.09902*.

González-Leonardo, M.; Neville, R.; Gil-Clavel, S.; and Rowe, F. 2024. Where have Ukrainian refugees gone? Identifying potential settlement areas across European regions integrating digital and traditional geographic data. *Population, Space and Place*, e2790.

Google. 2024. FAQ about Google Trends data. Online; accessed Mar. 2024.

Graham, M.; and Dittus, M. 2022. *Geographies of Digital Exclusion: Data and Inequality*. JSTOR.

Hsiao, Y.; Fiorio, L.; Wakefield, J.; and Zagheni, E. 2023. Modeling the bias of digital data: an approach to combining digital with official statistics to estimate and predict migration trends. *Sociological Methods & Research*, 00491241221140144.

Imperative, S. P. 2022. Ukraine Refugee Pulse. Online; accessed May. 2024.

International Organization for Migration. 2022. DTM Germany - Third Country Nationals arriving from Ukraine in Germany. Online; accessed May. 2024.

International Organization for Migration. 2023. DTM Romania "Surveys with refugees from Ukraine: needs, intentions and integration challenges". Online; accessed May. 2024.

Kim, S.; Park, S.; Hale, S. A.; Kim, S.; Byun, J.; and Oh, A. H. 2016. Understanding Editing Behaviors in Multilingual Wikipedia. *PLOS ONE*, 11(5): e0155305. Publisher: Public Library of Science.

Köksal, S.; Pesando, L. M.; Rotondi, V.; and Şanlıtürk, E. 2022. Harnessing the potential of Google searches for understanding dynamics of intimate partner violence before and after the COVID-19 outbreak. *European journal of population*, 38(3): 517–545.

Lazer, D.; Kennedy, R.; King, G.; and Vespignani, A. 2014. The parable of Google Flu: traps in big data analysis. *science*, 343(6176): 1203–1205.

Leasure, D. R.; Kashyap, R.; Rampazzo, F.; Dooley, C. A.; Elbers, B.; Bondarenko, M.; Verhagen, M.; Frey, A.; Yan, J.; Akimova, E. T.; et al. 2023. Nowcasting daily population displacement in Ukraine through social media advertising data. *Population and Development Review*, 49(2): 231–254.

Lehmann, J.; Müller-Birn, C.; Laniado, D.; Lalmas, M.; and Kaltenbrunner, A. 2014. Reader preferences and behavior on Wikipedia. In *Proceedings of the 25th ACM conference on Hypertext and social media*, 88–97. Santiago Chile: ACM. ISBN 978-1-4503-2954-5.

Lerner, J.; and Lomi, A. 2019. Let's Talk About Refugees: Network Effects Drive Contributor Attention to Wikipedia Articles About Migration-Related Topics. In Aiello, L. M.; Cherifi, C.; Cherifi, H.; Lambiotte, R.; Lió, P.; and Rocha, L. M., eds., *Complex Networks and Their Applications VII*, Studies in Computational Intelligence, 211–222. Cham: Springer International Publishing. ISBN 978-3-030-05414-4.

Leurs, K.; and Smets, K. 2018. Five questions for digital migration studies: Learning from digital connectivity and forced migration in (to) Europe. *Social Media+ Society*, 4(1): 2056305118764425.

Lin, A. Y.; Cranshaw, J.; and Counts, S. 2019. Forecasting us domestic migration using internet search queries. In *The world wide web conference*, 1061–1072.

Lucchini, L.; Tonelli, S.; and Lepri, B. 2019. Following the footsteps of giants: modeling the mobility of historically notable individuals using Wikipedia. *EPJ Data Science*, 8(1): 36. Number: 1 Publisher: Springer Berlin Heidelberg.

Miquel-Ribé, M.; and Laniado, D. 2019. Wikipedia Cultural Diversity Dataset: A Complete Cartography for 300 Language Editions. *Proceedings of the International AAAI Conference on Web and Social Media*, 13: 620–629.

Miquel-Ribé, M.; and Laniado, D. 2020. The Wikipedia Diversity Observatory: A Project to Identify and Bridge Content Gaps in Wikipedia. In *Proceedings of the 16th International Symposium on Open Collaboration*, OpenSym '20, 1–4. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8779-8.

Moat, H. S.; Curme, C.; Avakian, A.; Kenett, D. Y.; Stanley, H. E.; and Preis, T. 2013. Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports*, 3(1): 1801.

Panciera, K.; Halfaker, A.; and Terveen, L. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedinfs of the ACM 2009 international conference on Supporting group work - GROUP '09*, 51. Sanibel Island, Florida, USA: ACM Press. ISBN 978-1-60558-500-0.

Perrotta, D.; Johnson, S. C.; Theile, T.; Grow, A.; de Valk, H.; and Zagheni, E. 2022. Openness to migrate internationally for a job: evidence from LinkedIn data in Europe. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 759–769.

Ribeiro, M. H.; Gligorić, K.; Peyrard, M.; Lemmerich, F.; Strohmaier, M.; and West, R. 2020. Sudden Attention Shifts on Wikipedia During the COVID-19 Crisis. 12.

Rivers, C. M.; and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research*, 3(38): 38.

Robinson, V. 1998. The importance of information in the resettlement of refugees in the UK. *Journal of refugee studies*, 11(2): 146–160.

Ruprechter, T.; Horta Ribeiro, M.; Santos, T.; Lemmerich, F.; Strohmaier, M.; West, R.; and Helic, D. 2021. Volunteer contributions to Wikipedia increased during COVID-19 mobility restrictions. *Scientific Reports*, 11(1): 21505. Number: 1 Publisher: Nature Publishing Group.

Singer, P.; Lemmerich, F.; West, R.; Zia, L.; Wulczyn, E.; Strohmaier, M.; and Leskovec, J. 2017. Why We Read Wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*, 1591–1600. Perth Australia: International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4913-0.

Spyratos, S.; Vespe, M.; Natale, F.; Weber, I.; Zagheni, E.; and Rango, M. 2018. Migration data using social media: a European perspective.

The GSM Association. 2017. The Importance of Mobile for Refugees: A Landscape of New Services and Approaches. URL: https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2017/02/The-Importance-of-mobile-for-refugees_a-landscape-of-new-services-and-approaches.pdf. Online; accessed Apr. 2022.

Tjaden, J. 2021. Measuring migration 2.0: a review of digital data sources. *Comparative Migration Studies*, 9(1): 1–20.

Twyman, M.; Keegan, B. C.; and Shaw, A. 2017. Black Lives Matter in Wikipedia: Collective Memory and Collaboration around Online Social Movements. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 1400–1412. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4335-0.

Ulutürk, I.; Uysal, I.; and Varol, O. 2019. Refugee integration in Turkey: a study of mobile phone data for D4R challenge. In *Data for refugees challenge workshop*.

Vieira, C. C.; Fatehkia, M.; Garimella, K.; Weber, I.; and Zagheni, E. 2022. Using Facebook and LinkedIn Data to Study International Mobility.

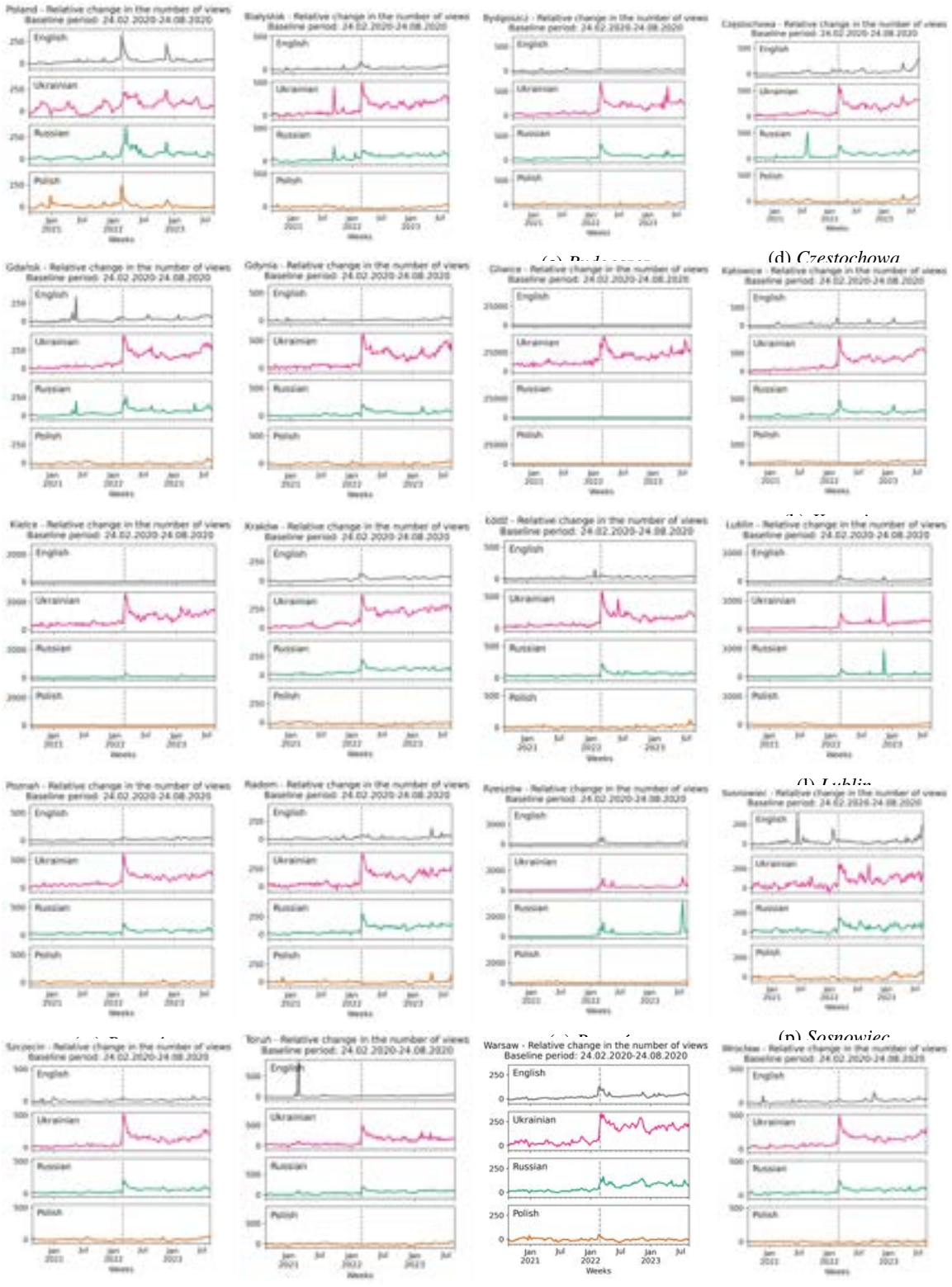
Vilain, P.; Larrieu, S.; Cossin, S.; Caserio-Schönemann, C.; and Filleul, L. 2017. Wikipedia: a tool to monitor seasonal diseases trends? *Online Journal of Public Health Informatics*, 9(1): e052.

West, R.; Weber, I.; and Castillo, C. 2012. A data-driven sketch of Wikipedia editors. In *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, 631. Lyon, France: ACM Press. ISBN 978-1-4503-1230-1.

Yoshida, M.; Arase, Y.; Tsunoda, T.; and Yamamoto, M. 2015. Wikipedia Page View Reflects Web Search Trend. In *Proceedings of the ACM Web Science Conference*, 1–2. Oxford United Kingdom: ACM. ISBN 978-1-4503-3672-7.

Zhu, T. J.; Fritzler, A.; and Orlowski, J. A. K. 2018. World Bank Group-LinkedIn Data Insights: Jobs, Skills and Migration Trends Methodology and Validation Results.

Appendix



(d) *Czestochowa*

(p) *Sosnowiec*

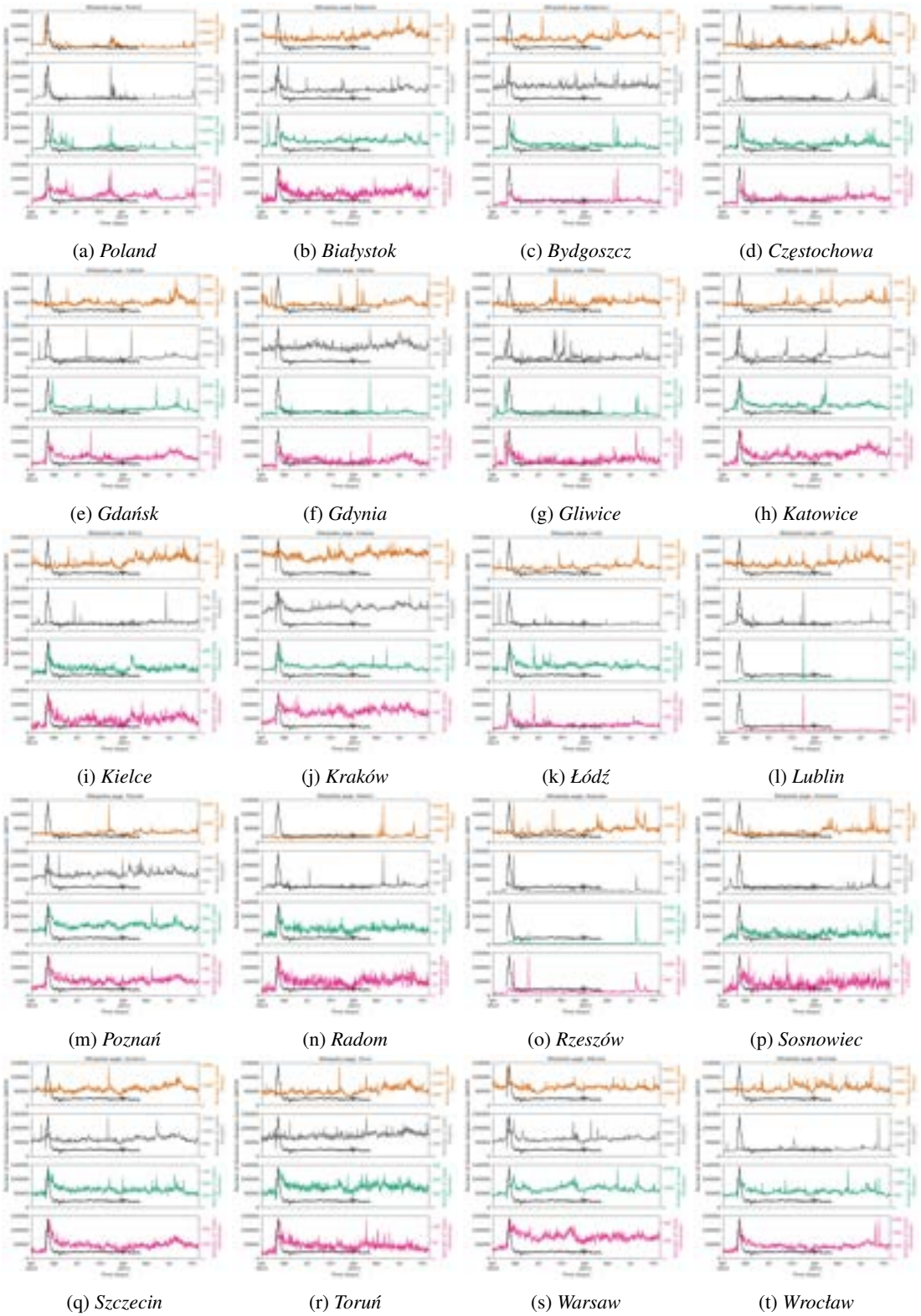
(q) *Szczecin*

(r) *Toruń*

(s) *Warsaw*

(t) *Wrocław*

Figure 5: Relative change in the number of views on Wikipedia pages dedicated to Poland and some of the most populous cities in Poland across different languages (i.e., English, Polish, Russian, and Ukrainian). Baseline period: 24.02.2020-24.08.2020.



(a) *Poland* (b) *Białystok* (c) *Bydgoszcz* (d) *Częstochowa*

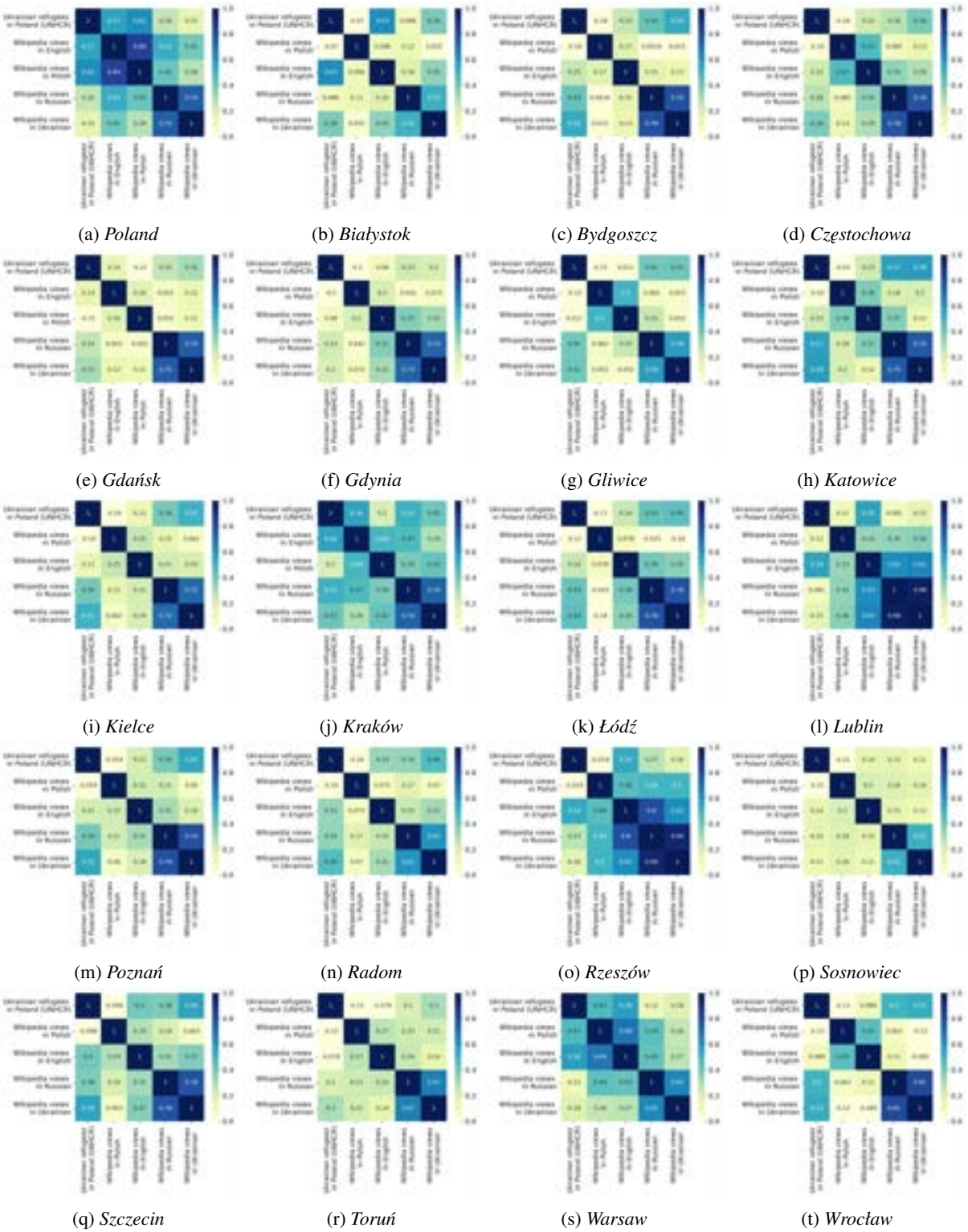
(e) *Gdańsk* (f) *Gdynia* (g) *Gliwice* (h) *Katowice*

(i) *Kielce* (j) *Kraków* (k) *Łódź* (l) *Lublin*

(m) *Poznań* (n) *Radom* (o) *Rzeszów* (p) *Sosnowiec*

(q) *Szczecin* (r) *Toruń* (s) *Warsaw* (t) *Wrocław*

Figure 6: Time series representing the daily number of Ukrainian refugees crossing the border from Ukraine to Poland (from February 24th, 2022 to March 3rd, 2023) and the daily number of views Wikipedia pages dedicated to some cities in Poland across different languages (i.e., English as a baseline, Polish, Russian, and Ukrainian).



(a) *Poland* (b) *Białystok* (c) *Bydgoszcz* (d) *Częstochowa*

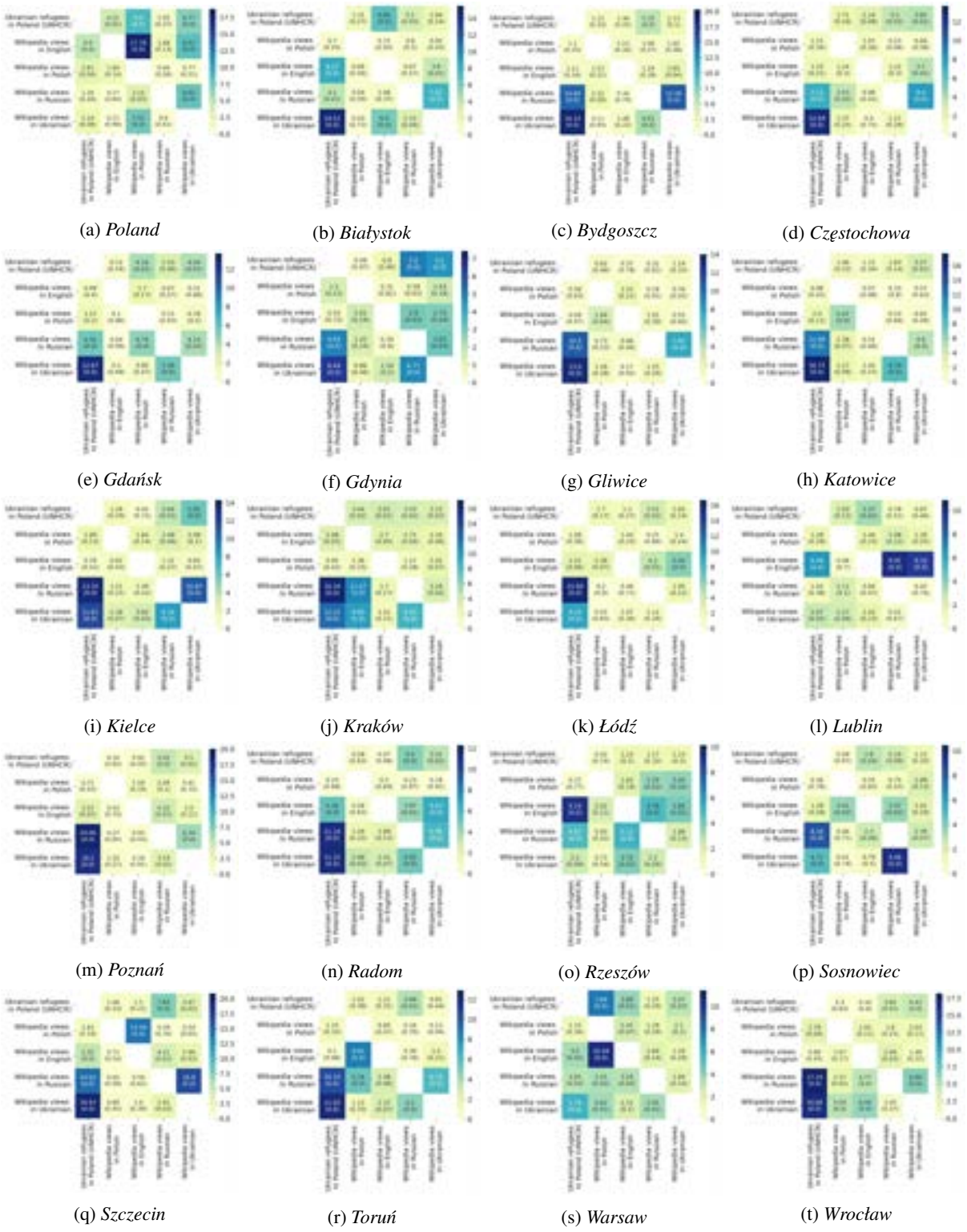
(e) *Gdańsk* (f) *Gdynia* (g) *Gliwice* (h) *Katowice*

(i) *Kielce* (j) *Kraków* (k) *Łódź* (l) *Lublin*

(m) *Poznań* (n) *Radom* (o) *Rzeszów* (p) *Sosnowiec*

(q) *Szczecin* (r) *Toruń* (s) *Warsaw* (t) *Wrocław*

Figure 7: Correlation between the official number of Ukrainian refugees crossing the border from Ukraine to Poland (from February 24th, 2022 to March 3rd, 2023) and the number of views Wikipedia pages dedicated to some cities in Poland across different languages (i.e., English as a baseline, Polish, Russian, and Ukrainian).



(a) *Poland* (b) *Białystok* (c) *Bydgoszcz* (d) *Częstochowa*

(e) *Gdańsk* (f) *Gdynia* (g) *Gliwice* (h) *Katowice*

(i) *Kielce* (j) *Kraków* (k) *Łódź* (l) *Lublin*

(m) *Poznań* (n) *Radom* (o) *Rzeszów* (p) *Sosnowiec*

(q) *Szczecin* (r) *Toruń* (s) *Warsaw* (t) *Wrocław*

Figure 8: Granger causality between the official number of Ukrainian refugees crossing the border from Ukraine to Poland (from February 24th, 2022 to March 3rd, 2023) and the number of views Wikipedia pages dedicated to some cities in Poland across different languages (i.e., English as a baseline, Polish, Russian, and Ukrainian).

Comparison between Wikipedia and Google Trends data

The growing literature on online information-seeking behavior and migration often relies on the data from online search engines such as Google (Böhme, Gröger, and Stöhr 2020; Avramescu and Wiśniowski 2021), and Yandex (Anastasiadou, Volgin, and Leasure 2024). Data provided by the Google Trends tool by Google is often preferred due to the widespread use of Google worldwide. While the Google Trends data is proven to be a useful indicator of interest and possible intention to move, the characteristics of the data pose limitations for advanced statistical analyses. Google does not disclose information on the volume of online searches but produces an index with a range of 0-100 normalized for online search popularity for the given query (keyword), location, and time period. Furthermore, this index is not based on the entire search data for the given parameters but on a sample large enough to represent the needed search data, yet the sample size is unknown to the users. Google Trends reports the daily search popularity index for a time period shorter than nine months. For longer periods, one can stitch together multiple datasets of nine months and rescale (Köksal et al. 2022), however it may bias the data. While proven useful and consistent despite these limitations, it must be acknowledged that Google Trends Index is representative of the required online search data, but "might not be a perfect mirror of search activity (Google 2024)."

In this study, we introduce the use of Wikipedia data as an indicator of migration and mobility. Wikipedia data bears certain advantages in comparison to Google Trends data. The most important of which is to provide the total number of page views instead of a normalized index that is more suitable for statistical analyses. In Google Trends, distinguishing between two different groups of people, such as host and migrant groups, at the same location is possible if the query words are in different languages or alphabets. Distinguishing by language creates issues when using city or province names as query words because they mostly remain the same across different languages and may leave the alphabetical difference as the only differentiation method. In contrast, Wikipedia pages are available in different languages, which may allow the differentiation between two groups easier; however, it provides only information on the language of the viewed page but not the location of the view. It must be underlined that differentiation by language would also be problematic for languages that are common second languages and/or native languages of multiple countries, such as English, Spanish, French, and Arabic. However, in the context of our case study, Polish and Ukrainian languages may be more easily attributed to the respective countries and their people.

In order to observe and demonstrate the potential advantages of Wikipedia data with respect to the Google Trends data, we collected Google Trends data [14] matching to the Wikipedia data in our study for a descriptive analysis. Figure 9 demonstrates a comparison between the Wikipedia data used in our study and the matching Google Trends data for the nineteen most populous Polish cities. We distinguish the online searches for Polish cities made by Ukrainians by setting the location as Ukraine. We take the relevant city as a topic (city) instead of the name of the city as a keyword to avoid the reporting of too much zero interest. To enable an easier visual comparison, we normalized the Wikipedia views to the same range as Google Trends data, i.e. 0-100. We then compare and contrast the changes in the information-seeking for Polish cities following the Russian invasion of Ukraine.

Looking at the descriptive analysis of these two data sources, we highlight two main points. First, even using the name of the city as the topic and not as the strict keyword, Google Trends data reports many zero values and noise in the data. This is not the case for the Wikipedia data, as it is not an index but based on the view counts. This creates an advantage for Wikipedia data over Google Trends data, which can be observed in the graphs for Białystok, Bydgoszcz, Częstochowa, Gdynia, Gliwice, Kielce, Radom, and Toruń (Figure 9). Second, for more populated big cities or for the country name (Poland), for which we have better quality (less noisy) Google Trends data, we do not observe important divergences between the patterns of Google Trends and Wikipedia data.

[14] Google Trends data was collected using the *gtrendsR* package on R and *pytrends* package on Python. We used both R and Python to accelerate the data collection.

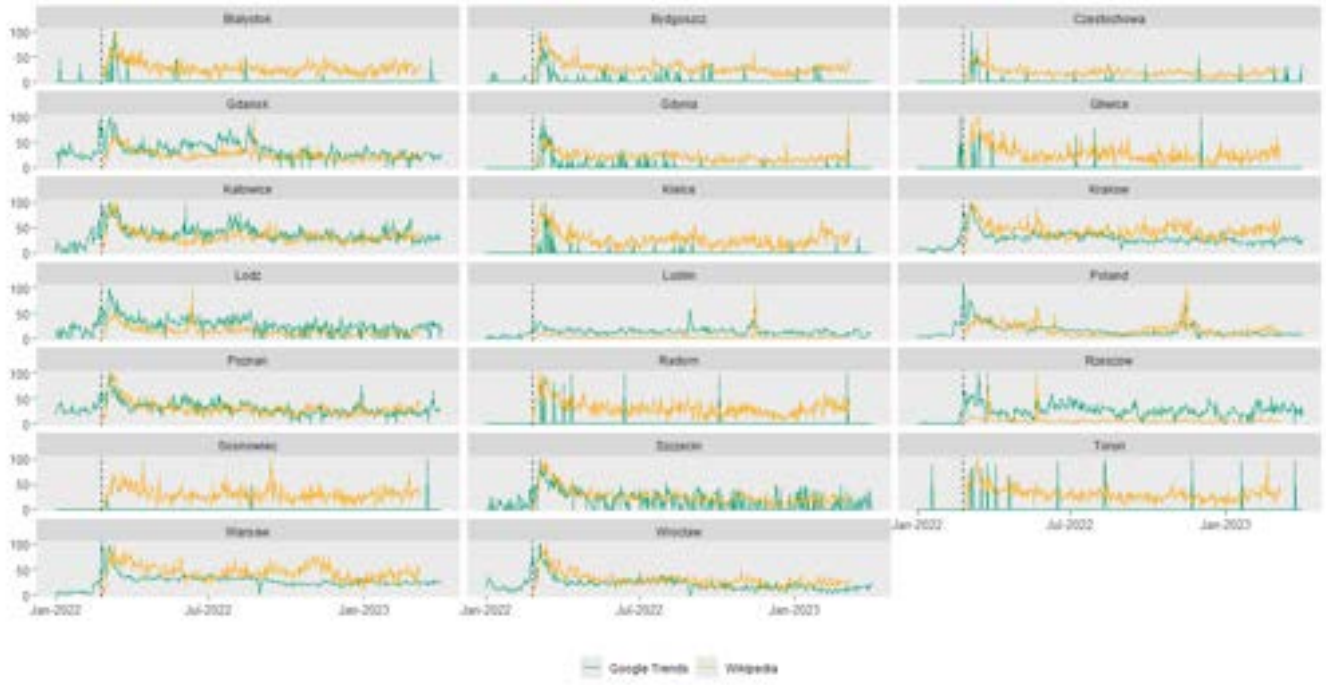


Figure 9: Comparison between the Google Trends Index (GTI) of daily searches on Google for Polish cities in Ukraine as topic and the daily views of Wikipedia pages for the Polish cities in Ukrainian language. To allow for the comparison between GTI and Wikipedia views, the latter has also been normalized to the 0-100 range for this figure. The GTI values are shown in green, while the Wikipedia views are shown in orange color. The GTI shows data from January 1st, 2022 to March 26th, 2023, while the Wikipedia views show data from February 24th, 2022 to March 7th, 2023. The vertical dashed line indicates the beginning of the Russian invasion of Ukraine, February 24th, 2022.